University of Sfax
Faculty of Sciences of Sfax

EMB
IEEE Tunisia Chapter

3rd International Conference on

Bayesian Networks and Applications

Previously known as: Journées sur les Réseaux Bayésiens et Applications

October 14 – 16, 2016
Sousse – Tunisia

2016 ICBNA



Proceedings

Invited Talks

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Discrete exponential Bayesian networks

Afif MASMOUDI

ABSTRACT

Our work aims at developing or expliciting bridges between Bayesian networks (BNs) and Natural Exponential Families, by proposing discrete exponential Bayesian networks as a generalization of usual discrete ones. We introduce a family of prior distributions which generalizes the Dirichlet prior applied on discrete Bayesian networks, and then we determine the overall posterior distribution. Subsequently, we develop the Bayesian estimators of the parameters, and a new score function that extends the Bayesian Dirichlet score for BN structure learning. Our goal is to determine empirically in which contexts some of our discrete exponential BNs (Poisson deBNs) can be an effective alternative to usual BNs for density estimation.

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Multivariate Dispersion Models & Applications : On characterizations of multiple stables-Tweedie models

Kokonendji Célestin C.

ABSTRACT

Dispersion models are introduced to extend Normal model into specific analyses such as regression, extreme, geometric sum and count; and, why not for a Bayesian network? The basic family of dispersion models is the (univariate) Tweedie model which generalizes the stable distributions. As particular cases of multivariate Exponential Dispersion models, we consider first disturbances of Gaussian random vector by some Tweedie components, namely Normal stables-Tweedie (NST) models, and then disturbances of NST for getting multiple stables-Tweedie (MST) models. After a global scheme of characterizations of (exponential) dispersion models by (generalized) variance functions, we show several results on NST and MST models. For instance, NST and MST models are classified by their variance functions. Only NST are characterized by some of their assoiciated pseudo-orthogonal polynomial functions and also by their generalized variance functions (or determinant of variance functions) through explicit solutions of the corresponding Monge-Ampère equations.

Accepted Articles

Bayesian network for constructing probabilistic ontologies

Emna hlel
MIRACL Laboratory, Sfax University, Technology Center
BP 242 – 3021, Sakiet Ezzit, Sfax, Tunisia
emnahlel@gmail.com

Salma Jamoussi and Abdelmajid Ben Hamadou
MIRACL Laboratory, Sfax University, Technology Center
BP 242 – 3021, Sakiet Ezzit, Sfax, Tunisia
salma.jamoussi@isimsf.rnu.tn and
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract— During the past years, the ontologies are widely used for representing knowledge of most real world domains. They provide a definition of concepts, relationships, etc. Thanks to these elements, they are used to model the reality (real world applications). However, this world includes inaccuracies and imperfections which cannot be represented by classical or ontologies (COs). Probabilistic ontologies (POs) have come to remedy this defect. This paper is part of this framework in which we have proposed a novel method to construct probabilistic ontologies. For this aim, we have used Bayesian network which is a probabilistic model allowing to represent the knowledge of domain on a formal theoretical basis. Indeed, in this paper, we have presented a way to discover probabilistic relationships between a list of instances of an OWL ontology by using Bayesian network and how we can model these relationships in ontologies.

Keywords—Bayesian network, Bayesian inference, probabilistic ontology, uncertain.

I. INTRODUCTION

During the past years, the ontologies are widely used for representing knowledge of most real world domains. They provide a definition of concepts, relationships, and other features related to modeling knowledge of complex domains [1]. Despite that the ontologies have become standard for representing knowledge in many applications; however they are not able to deal with uncertainty that is a ubiquitous aspect of most real world problems [2]. Modeling uncertainty is a big challenge. Several methods for modeling uncertainty in ontologies have recently started emerging. Generally, these methods are based on mathematical techniques of uncertainty: the probabilistic theory, the fuzzy logic approach and the Dempster-Shafer theory [3, 4, 5]. The Probability theory has been proven to be one of the most powerful approaches to deal with uncertainty and it is a natural choice for representing the uncertain and probabilistic knowledge [6]. In this work, we restrict our attention to approaches based on the probabilistic theory for representing uncertain and probabilistic knowledge and more specifically on the Bayesian network (BN). The latter is one of the best models for representing the knowledge on a formal theoretical basis. It aims to represent knowledge of a particular system (complex and simple) in the form of a graph that is intuitive, clear, readable and understandable by a non-specialist. This model has powerful techniques: BN structure learning, BN parameters learning, Bayesian inference (BI), etc [7, 8].

In this work, we have proposed a novel method to build probabilistic ontologies. Indeed, in this paper, we have presented a way to discover probabilistic relationships between a list of instances of an OWL ontology by using Bayesian network and how we can model these relationships in ontologies. The remainder of this paper is organized as follows. The sections 2 and 3 present the Bayesian networks and the major existing works. In section 4, we propose our method for constructing probabilistic ontologies. Finally, we conclude by summarizing our work and listing points for future work.

II. BAYESIAN NETWORK

In the recent years, BNs have been widely applied with success in various fields as medical service performance analysis [9], gene expression analysis [10], breast cancer prognosis [11], etc. This probabilistic model is one of the best models for representing the knowledge on a formal theoretical basis [7, 8]. It aims to represent knowledge of a particular system (complex and simple) in the form of a graph that is intuitive, clear, readable and understandable by a non-specialist. Indeed, it represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG): the nodes represent variables and arcs represent probabilistic dependencies between them. Mathematically, a bayesian network $B=(G, C)$ is defined by:

- $G=(X, E)$ is a directed acyclic graph whose nodes are associated with a set of random variables $X=\{X_1, ..., X_n\}$ and E is the set of arcs that represent dependencies between these nodes,
- $C=\{P(X_i|Pa(X_i))\}$ is a set of probabilities where each node X_i is conditionally dependent of the state of its parent $Pa(X_i)$ in G.

We distinguish two methods of learning: BN parameters learning (when the BN structure is given) and BN structure learning. The BN structure learning determines an appropriate graph from the observed data. The BN parameters learning determines the best set of parameters of BN (probability distribution: conditional and marginal probability) taking into account the observed data. Moreover, one of the powerful techniques of BN is Bayesian inference which consists to propagate one or more certain information (values established by certain variables) to deduce how this intervenes on the

probabilities of other network variables. Mathematically, the inference in a bayesian network is the calculation of $P(U|\varepsilon)$ (with U is the set of variables and ε is the new information on one or more variables U), that is to say calculating the posterior probability of the network knowing that ε [12]. It has a wide range of practical applications, for example tracking aircraft based on radar data, building a bibliographic database based on citation lists, analyzing a list of symptoms to infer the illness of a patient, etc. Different algorithms exist to perform inference on BN: "loop cutset conditioning" [13], "algorithm LS" [14], "algorithm of Shenoy-Shafer" [15], and algorithm "lazy propagation" [16], etc. The most popular inference algorithm for multiple connected networks was proposed by Lauritzen and Speigelhalter [14]. This algorithm involves the extraction of an undirected triangulated graph of the BN, and the creation of a tree whose vertices are the cliques of the triangulated graph. This tree is called junction tree. The conditional probabilities are then computed by transmitting messages in this tree.

III. Related work

None of the existing semantic web languages such as RDF/RDFS, SHOE, OWL and etc provide a means for representing uncertain and probabilistic knowledge of real world domains. Different probabilistic approaches for extending these languages, especially OWL, with the ability to support uncertainty are explored in literature. Indeed, several Bayesian-based approaches to model uncertainty in ontologies have been proposed: BayesOWL [6], OntoBayes [17] and PR-OWL [18]. Various works provide a comparative study on those approaches such as [19]. BayesOWL [6] is a proposal to represent the uncertainty in OWL ontologies through BN in order to facilitate Ontology Mapping in the semantic web. The representation of probabilistic knowledge in BayesOWL is performed via additional language markups, which can be simply viewed as an upper ontology [17]. BayesOWL is used to estimate the degree of the overlap or inclusion between two concepts in terms of conditional probabilities of the form $P(C|D)$ where C and D are two classes. This degree expresses the probability that an instance of D is also an instance of C. It is true that this work is the first major research effort published in the field of probabilistic extension of Semantic Web language (OWL). However, there are several limitations considering this approach. Firstly, its application potential is very limited: Ontology Mapping in the semantic web. In addition, it cannot represent probabilistic information about any relations of ontology, except the subsumption relation. Indeed, the authors of [6] have only focused on the concepts taxonomy. On the contrary, an ontology is not only a taxonomy, it is also a model which includes a list of components such as proprieties, instances, non-taxonomic relations, etc.

OntoBayes [17] is an ontology-driven uncertainty model, which integrates probabilistic models (BNs) into OWL ontologies for preserving their advantages. It was developed as an extension which enables OWL ontologies to represent BNs. Indeed, the authors of [17] have proposed an upper ontology, called Ontology OntoBayes, for representing random variables, dependencies between them and probabilities associated to these variables. In other words, this ontology allows describing the qualitative and quantitative representation of a BN (the essentials of a BN). By using OntoBayes ontology, the users can write down probabilistic models that correspond to BNs [19].

Probabilistic OWL (PR-OWL) [18] is a PO approach that is implemented on the basis of first-order logic. It is a probabilistic extension which enables OWL ontologies to represent MEBNs (Multi-Entity Bayesian Networks) [20]. It provides a number of new OWL constructs for constructing POs. Indeed, it is an upper ontology that describes first-order probabilistic models (MEBNs). In order to write a PO by using this upper ontology, the users can import this ontology into an editor of ontology such as Protégé[1]. After importing it, they start the step of construction of domain-specific concepts by using the PR-OWL definitions to represent uncertainty. However, there are several limitations of this probabilistic approach. Firstly, the construction of fundamental elements of model PR-OWL (such as Mfrags and all their elements) for representing the uncertainty is performed with a manual way while the whole process is error-prone and tedious [21]. Moreover, this probabilistic approach uses the MEBN for representing the uncertainty; however, the community MEBN is not wide enough to be considered as an emerging standard for representing uncertainty [21].

We have noted that these probabilistic extensions of semantic web language OWL (OntoBayes, BayesOWL and PR-OWL) have some common points. They provide Bayesian extensions which enable OWL ontologies to represent the essential components of the corresponding probabilistic model (BN, MEBN) by using new major OWL elements (classes and properties). However, for constructing PO based on one of these extensions it is necessary that the ontologist has known the foundation of the corresponding probabilistic model to identify the different ontology elements that describe the quantitative and qualitative representation of this model (as states, variables, etc). This requires an enormous effort by ontologist for doing so. On the contrary, in this paper, we have tried to propose a new method (semi-automatic) of PO construction by using the BN only for calculating the probabilities which represent the uncertain aspect without modeling the different components of BN in OWL ontology. In other words, the modeling of uncertain knowledge in ontologies is totally independent of the representation provided by BN. Moreover, in the most proposals of extensions of semantic web languages for supporting uncertainty [17, 18], the representation of the probabilistic and uncertain knowledge into OWL ontologies is performed via additional language markups. More specifically, these proposals extend the OWL model with a list of major elements for modeling uncertainty (modeling the fundamental elements of corresponding probabilistic model). On the contrary, in this paper, we have tried to extend the language OWL only with minor statistical extensions that do not require an OWL

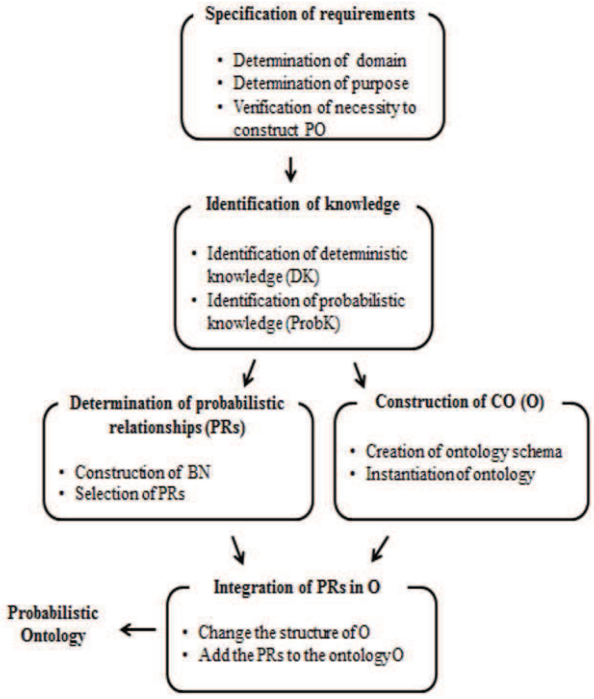
[1] http://protege.stanford.edu/

philosophy for representing the probabilities obtained with the help of BN. In other words, we have extended the OWL ontologies with minimal changes for supporting the uncertainty.

IV. PROPOSED METHOD FOR BUILDING PROBABILISTIC ONTOLOGY

During the past years, the ontologies are widely used for representing knowledge of most real world domains. They provide a definition of concepts, relationships, and other features related to modeling knowledge of particular domain [1]. Thanks to these elements, they are used to model the reality (real world applications). However, this world includes inaccuracies and imperfections which cannot be represented by classical ontologies. POs have come to remedy this defect [18]. We can define the PO simply as a CO enriched with uncertain and probabilistic knowledge. Indeed, POs augment COs with the ability to represent the uncertainty [22, 23]. This paper is part of this framework in which we have proposed a method for constructing probabilistic ontologies. Indeed, we have presented a way to discover probabilistic relationships between a list of instances of an OWL ontology by using Bayesian network and how we can model these relationships in ontologies. The process of this method includes five steps, as shows in figure 1, which are: specification of requirements, identification of knowledge of domain (deterministic and probabilistic), construction of classical ontology, determination of probabilistic relationships and integration of these relationships in the classical ontology obtained previously. As a case study, we have tried to construct a PO which describes a system of scientific documentation (research themes, authors, etc).

Fig. 1. Process of construction of probabilistic ontologies



A. Specification of requirements

This phase determines the domain and the purpose of ontology: It is important to be clear identified the purpose (goal) of the ontology. In addition, the ontologist must verify the necessity of the creation of PO through research of uncertainties and inaccuracies in the field of study.

B. Identification of knowledge of domain

We have distinguished between two types of knowledge: deterministic (classical) knowledge and probabilistic (uncertain) knowledge. In our case study, the classical knowledge is a list of concepts (*Theme*, *Author*, etc) and its instances, list of relations (*co-author*, *characterized*, etc). Moreover, the relation "*be-interested*" (between the instances of *Author* and the instances of *Theme*) and the relation "*be-connected*" (between the instances of the concept *Theme*) are considered as probabilistic relationships (probabilistic knowledge).

C. Construction of classical ontology

This phase consists to build a CO, named O, which represents the deterministic knowledge previously mentioned. There are several methodologies for developing ontologies in literature [24, 25]. Generally, to develop an ontology one should follow a list of steps. Firstly, we identify the concepts (classes), their properties and the concepts hierarchy, etc. Also, it is necessary to specify the relationships among the concepts in detail by giving a name, source concept, target concept, cardinality, etc. Then, we create the instances or individuals of the ontology.

In our case study, we construct the classical ontology, named O, which is composed of three concepts which are: "*Author*", "*Key-Word*" and "*Theme*". Each research theme is characterized by a keyword list. Each author has from 1 to M coauthors. In addition, O consists of two relationships which are "*have-word*", "*co-author*". The relation "*have-word*" expresses that a theme is characterized by a set of keywords. The relationship "*co-author*" indicates that an author may have a co-author list. After that, we create the different instances of the obtained ontology (population of ontology). The ontology O includes a list of instances: $A=\{A_0, ..., A_i, ...A_n\}$ which represent the instances of the concept *Author*, $T=\{T_0, ..., T_j, ..., T_m\}$ which represents the instances of the concept *Theme* and $W=\{W_0, ..., W_y, ..., W_k\}$ which represents the instances of the concept *Key-Word*. So, the obtained ontology represents the deterministic knowledge of the interest domain.

D. Determination of probabilistic relationships

The aim of this phase is to determine the probabilistic relationships between the different instances of the ontology O. For doing this, we have used the BN. Indeed, this phase includes three steps: construction of BN graph, BN parameters learning and selection of probabilistic relationships.

In our case study, for determining the probabilistic relationships ("*be-interested*") between the instances of the concept *Theme* and the concept *Author* and the probabilistic relationships ("*be-connected*") between the instances of *Theme*,

we have followed this process. Firstly, we have constructed the BN graph from the observed data. We can divide the nodes of this graph in two sets: $A=\{A_0, ..., A_i, ...A_n\}$ which represent the instances of *Author* and $T=\{T_0, ..., T_j, ..., T_m\}$ which represents the instances of *Theme*. Each node of the obtained BN graph has two values true and false.

After determining the BN graph, the process of BN parameters learning can be started. The latter is an essential step in a BN construction. This is done by specifying a conditional probability distribution for each node of BN graph taking into consideration the observed data. It can be performed with a simple statistical or Bayesian learning (if the database is complete). In our case, we used an estimator with complete data. This estimator is based on a statistical approach. It involves estimating the probability of an event by the frequency of occurrence of the event in the database:

$$P(X_i = x_k | pa(X_i) = x_j) = \frac{N_{i,j,k}}{\sum N_{i,j,k}} \quad (1)$$

where $N_{i,j,k}$ is the number of events in the database for which the variable X_i is in the state x_k and its parents are in the configuration x_j. This approach is called Maximum Likelihood Estimation (MLE) [20].

After that, we have selected the probabilistic relationships between the instances of the ontology O based on the results of the inference Bayesian. For each element T_j of T, we have followed this process to determine the probabilistic relations "*be-connected*" between this theme and the other themes ($T\backslash\{T_j\}$):

- Determination of probabilities of BN variables: Given the observation of the variable of the obtained BN which represents the theme T_j (T_j=true), this step determines the probabilities of the BN variables which represents the set of themes by applying the Bayesian inference (we have interested only to the probabilities associated to the set $T/\{T_j\}$). This step us provides a set of probabilities $ProbT=\{P_0, ..., P_u, ..., P_m\}$, with P_u is equal to:

$$P_u=P(T_u="true"|T_j="true") \quad (2)$$

  where T_u belongs to T, u is from 1 to m and P_u measures the strength of connection between the theme T_j and the theme T_u and it allows estimating the degree of dependency between them. For example, the probability P(imagery="true"|multimedia= "true")=0.9 means that the theme "*imagery*" is strongly connected to the theme "*multimedia*" with a probability equal to 0.9.

- Selection of the probabilistic relationships ("*be-connected*"): In this step, we have selected only the probabilistic relationships between the theme T_j and the other themes that have the highest probability values by using a threshold which is fixed empirically. It us provide a list of probabilistic relationships $RT=\{R_0, ... R_u, ..., R_f\}$ and the set $ProbRT=\{P_0, ..., P_u, ..., P_f\}$ where R_u is a probabilistic relationship between the T_j and T_u and it is associated with a probabilistic value P_u which represents the strength of connection between these two themes.

For each element a_i of A, we have followed the same process to determine the probabilistic relations "*be-interested*" between this author and the list of themes T:

- Determination of probabilities of BN variables: Given the observation of the variable of the obtained BN which represents the author a_i (forcing the value of the variable a_i to true), this step determines the probabilities of the BN variables which represents the set of themes. It us provides a set of probabilities $ProbA=\{P_0, ..., P_j, ..., P_m\}$, with P_j is equal to:

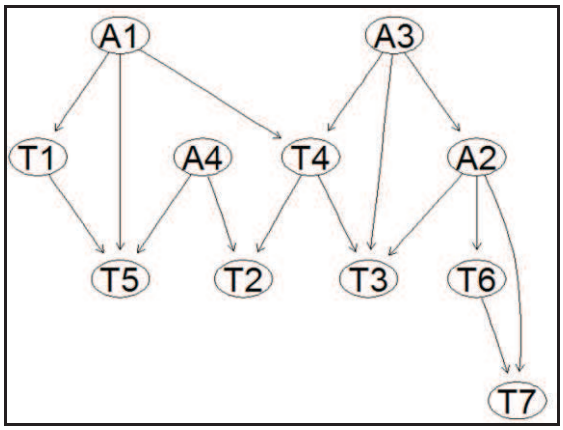
$$P_j=P(T_j="true"|a_i="true") \quad (3)$$

  where T_j belongs to T, j is from 1 to m and P_j depicts the probability of having the theme T_j when we see that the author a_i is present.

- Selection of the probabilistic relationships ("*be-interested*"): In this step, we have selected only the probabilistic relationships between the author a_i and the themes that have the highest probability values by using a threshold which is fixed empirically. This step us provide a list of probabilistic relationships $RA=\{R_0, ... R_j, ..., R_p\}$ and the set $ProbRA=\{P_0, ..., P_j, ..., P_p\}$ where R_j is a probabilistic relationship between the author a_i and the theme T_j and it is associated with P_j.

For better explanation, we can take an example. Imagine this situation: let G be a graph of BN which includes a set of authors (A) and a set of themes (T). The figure 2 shows this structure.

Fig. 2. An example of BN graphe.



For selecting the probabilistic relationships between the author A1 and their themes, we have followed this procedure. Given the observation of the BN variable which represents the author a_i, what are the probabilities of the other BN variables which represent the themes? After calculating the probabilities of these variables by applying Bayesian inference, the next step consists to select the probabilistic relationships ("*be-interested*") between the author A1 and the list of themes

which have the high probabilities (with *threshold=0.45*). These themes are T_1 (P_1=0.62), T_4 (P_4=0.6), T_5 (P_5=0.58), T_2 (P_2=0.5), T_3 (P_3=0.48). So, the probabilistic relationships between the instance A1 of *Author* and the instances of *Theme* are: R_1 (between A_1 and T_1), R_4 (between A_1 and T_4), R_5 (between A_1 and T_5), R_2 (between A_1 and T_2) and R_3 (between A_1 and T_3).

E. *Integration of the probabilistic relationships in the classical ontology*

For modeling the probabilistic relationships, we have performed some change in the structure of the classical ontology O which is previously constructed. Indeed, we have used relations N-ary for representing the probabilistic relationships. In our case study, the relationship "*be-connected*" is represented as an N-ary relation, which is characterized by a data-Property named "*ProbT*". For representing this relationship in OWL ontology, we have added to O a new concept "*Theme-Theme*" which is characterized by "*ProbT*" data-Property and two ObjectProperty: "*have-theme1* " and "*have-theme2*". "*ProbT*" expresses the probability that the theme T_j is connected to another theme T_s: $P(T_s="true"|T_j="true")$. The value of this property for each pair of instances (*Theme, Theme)* is determined from the previous phase. For adding the probabilistic relationships "*be-interested*" between the instances of *Author* and the instances of *Theme*, we have followed the same procedure.

V. CONCLUSION AND PERSPECTIVES

In this paper, we have proposed a new method for constructing probabilistic ontologies. Indeed, we have presented a way to discover probabilistic relationships between a list of instances of an OWL ontology by using Bayesian network and how we can model these relationships in ontologies. This method includes five steps which are: specification of requirements, identification of knowledge of domain (deterministic and probabilistic), construction of classical ontology, determination of probabilistic relationships and integration of these relationships in the classical ontology obtained previously.

As perspectives for this work, we will try to propose a method for representing another probabilistic component (such as probabilistic instance) of probabilistic ontology. In addition, we will try to create a tool allowing to guide the users for building a probabilistic ontologies.

REFERENCES

[1] T. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human and Computer Studies, pp.907-928, 1995.

[2] J.E. Santos and J.C Jurmain, "Bayesian Knowledge-driven Ontologies: Intuitive Uncertainty Reasoning for Semantic Networks". In International Conference on Systems, Man, and Cybernetics (SMC), pp. 856–863, 2011.

[3] L. A. Zadeh, "Fuzzy sets". Information and Control, pp. 338–353, 1965.

[4] P.K. Maji, R. Bismas and A.R. Roy, "Soft Set Theory". Computers & Mathematics with Applications, pp. 555-562, 2010.

[5] S. Russell and P. Norvig, "Artificial Intelligence, A Modern Approach", Engle-wood, NJ: Prentice Hall, 1995.

[6] Z. Ding, "BayesOWL: A Probabilistic Framework for Uncertainty in Semantic Web", thesis, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore, USA, 2005.

[7] F. V. Jensen, "An introduction to Bayesian networks". UCL Press, 1996.

[8] A. Ben Mrad, V. Delcroix, S. Piechowiak, P. Leicester and M. Abid, "An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence- Uncertain evidence in Bayesian networks". In Appl. Intell, pp. 802-824, 2015.

[9] S. Acid, L.M. J. de Campos, Fernandez-Luna, S. Rodriguez, J. Rodriguez and J. Salcedo, "A Comparison of Learning Algorithms for Bayesian Networks: A Case Study Based on Data from: An Emergency Medical Service". In Artificial Intelligence in Medicine, pp. 215–232, 2014.

[10] Friedman N., M. Linial and I. Nachman, "Using Bayesian Networks to Analyze Expression Data". In Journal of Computational Biology, pp. 601–620, 2000.

[11] D.E. Holmes and L.C. Jain, "Innovations in Bayesian Networks: Theory and Applications". In Computational Intelligence, 2008.

[12] K. Korb, and Nicholson, A. "Bayesian artificial intelligence", 2nd Chapman and Hall, London, CRC Press, 2010.

[13] P. Judea, Fusion, "propagation and structuring in belief networks". In Artificial Intelligence, pp.241– 288, 1986.

[14] S. L. Lauritzen and D. J. Spiegelhalter, "Local computation with probabilities and graphical structures and their application to expert systems". In J. Royal Statistical Society, pp.157–224, 1988.

[15] P. Shenoy and G. Shafer, "Axioms for probability and belief-function propagation". In Uncertainty in Artificial Intelligence, pp.169–198, 1990.

[16] A. L. Madsen and F. V. Jensen, "Lazy propagation in junction trees". In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 362–369, 1998.

[17] Yang Y., "A Framework for Decision Support Systems Adapted to Uncertain Knowledge". Von der Fakultat fur Informatik, der Universitat Fridericiana zu Karlsruhe (TH), thesis, 2007.

[18] Costa P.C.G., K.B. Laskey, "PR-OWL: A Framework for Probabilistic Ontologies". In Proceedings of the Conference on Formal Ontologies and Information Systems, Amsterdam, pp. 237–249, 2006.

[19] L. Predoiu and H. Stuckenschmidt, "Probabilistic Models for the Semantic Web: Survey". In Web Technologies: Concepts, Methodologies, Tools, and Applications. Book, University of Victoria, pp. 1896-1928, Germany, 2010.

[20] K.B. Laskey, "MEBN: a language for first-order bayesian knowledge bases". In Artif. Intell., vol. 172, no. 2-3, pp. 140–178, 2008.

[21] A. S. Foni, C.W. Wahyu and B. Novita, "Handling Uncertainty in Ontology Construction Based on Bayesian Approaches: A Comparative Study Ginting". In 4th International Conference on Soft Computing, Intelligent Systems and Information Technology, Indonesia, pp. 11–14, 2015.

[22] E. Hlel, S. Jamoussi and A. Ben Hamadou, "A probabilistic ontology for the prediction of author's Interests". In International conference ICCCI (Springer), pp. 492-501, 2015.

[23] E. Hlel, S. Jamoussi, A. Ben Hamadou, "Intégration d'un réseau bayésien dans une ontologie". In Actes des 25 journées francophones IC2014, pp.295-297, 2014.

[24] K.R. Stephen and P. Adam,"A framework for constructing cognition ontologies using WordNet, FrameNet, and SUMO". In Journal Cognitive Systems Research, 2015.

[25] A. Gómez-Pérez, L.M. Fernández and O. Corcho, "Ontological Engineering with examples from the areas of knowledge management", e-commerce and the semantic web. Springer Science & Business Media, 2006.

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

BAYESIAN NETWORK MODELING: a CASE STUDY of EXCHANGE RATE ANALYSIS

Sahar Charfi 1[1,*], Afif Masmoudi 2[2], Salah Ben Hamad 3[3]

Keywords: Bayesian Networks, Exchange rate

ABSTRACT

Explaining exchange rate dynamics and identifying its determinants still a difficult task for both econometrics and financial analysts. How does the association between financial-economic factors and exchange rate vary depending on different context is the main question that this paper attempts to answer. For this reason, this study uses an additive Gaussian Bayesian Network in order to evaluate interactions between different determinants of exchange rate and hence providing a powerful predicting tool. We take TND/USD exchange rate as a case study. The data set spans over the period 2002-2015.

The application of GBN tools has led to interesting results: significant interactions between all variables, inflation rate and interest rate of both Tunisia and US countries are the most responsible of TND/USD exchange rate fluctuations and TND/USD affect directly the Tunisia international reserves.

[1,3] Finance
[2] Applied Mathematics

* corresponding autor: charfisahar@live.fr

Heuristic and metaheuristic approaches for scheduling on single machine

Selt Omar

E -mail: selt.omar@yahoo.fr

Department of mathematics- University of M'sila , Algeria

Abstract. In this paper, an approach for scheduling problems of *n* tasks on *single* machine with three unavailability periods is proposed. This problem is strongly NP-complete which makes finding an optimal solution looks impossible task. In this frame, we suggested a novel heuristic in which availability periods are filled with the highest weighted tasks. To improve the performance of this approach, we used, on one hand, different diversification strategies with the aim of exploring unvisited regions of the solution space, and on the other hand, two well-known neighborhoods (neighborhood by swapping and neighborhood by blocs). The computational experiment was carried out on single machine with different availability zone. It must be noted that tasks movement can be within one zone or between different zones. Note that all data in this problem are integer and deterministic. The weighted sum of the end dates of tasks constitutes the optimization performance criterion in the problem treated in this paper.

.**Keywords**: Scheduling, metaheuristic, single machine ,NP-complete, unavailability periods.

1. Introduction. A scheduling problem under machines availability constraints has been studied by many authors. For example $P_m\,//N-\boldsymbol{C}\,//C_{\max}$ has been studied by Lee [9,10,11], Schmidt [15] and Yun-Chia *et al*([18]

The tabu search is a metaheuristic originally developed by Glover [5], Glover and Hanafi[4]) and independently by Hancen[6].

This method combines a local search procedure with a certain number of rules and mechanism which allows surmounting the obstacle of local optima without cycling. Toward furthermore, it proved hight efficiently in resolution of the problems NP-complet and approximate more the optimal solution.

The scheduling problem of a single machine with minimization of the weighted sum of the end dates of tasks. without unavailability constraint is optimally resolved by using the WSPT (weighted shortest processing time) rules. The case of several machines isstudied by manyauthorslikeBelouadeh[3], Sadfi[7] and Haouari[13].

Zribi*et al*. [19] have studied the problem $1//N-\boldsymbol{C}\,//\sum_{j=1}^{n}w_jC_j$ and have comparedtwo exact methods: one is the Branch and Bound, the otheris the integerprogramming.They have concludedthatBranch and Boundmethod have better performance and itallowedresolving instances of more than 1000 tasks.

Another studysAdamu and Adewunmi[1] have studied the problem $P_m\,//\sum_{j=1}^{n}w_j\left(U_j+V_j\right)$, they proposed some metaheuristics for

scheduling on parallel identical machines to minimize weighted number of early and tardy jobs.

In 2013, they carried out a comparative study of different (a genetic algorithm, particle swarm optimization and simulated annealing with their hybrids) metaheuristics for identical machines.

Selt and zitouni[16] have studied the problem $P_M//N\text{-}C//\sum w_i C_i$ they carried out a comparative study heuristic and metaheuristic for three identical parallel machines

2. Problem statements.

The objective is to determinethe input sequenceoftaskson the machineasthe weighted sum ofenddatesoftasks($\sum w_j C_j$) to be minimal.
It must be noted that there is $(n!)$ possibility to assign n tasks to the machine I Sakarovitch,[14]

3 Neighborhoods structure.

Neighborhood determination constitutes the most important stage in metaheurstic methods elaboration. In the following part, we use twowell-known Neighborhoods, (neighborhood by swap) and (neighborhood by block).It must be noted that tasks movement can be within one period or between different periods.

4.Tabu list structure.

Thetabu method is based on the principle that consists in maintaining in memory the last visited solutions and in forbidding the return to them for a certain number of iterations. The aim is to provide sufficient time to the algorithm so it can leave the local optimum. In other words, the tabu method conserves in each stage a list L of solutions (Tabu's) which it is forbidden to pass-by temporarily. The necessary space for saving a set of solutions tabus in the memory is indispensable.

The list, that we propose, contains the found solutions sequences. After many tests, a dynamic size list, which varies according to the search amelioration state, is conceived. The initial size of this list is considered to be $\frac{3\sqrt{N}}{2}$ where N the tasks number is. After that, during the search, when 5 successive iterations pass without amelioration of solution, the list is reduced to a number inferior or equal to $\sqrt{N}$. On the other hand, when 5 successive iterations pass and the solution is ameliorated, the list is increased to a number superior or equal to $2\sqrt{N}$. The Tabu list is consequently dynamic and its size varies within the interval $[\sqrt{N}, 2\sqrt{N}]$. The decrease or the increase of list size must always be done at the end of the list.

5. Formulations mathematics

This problem is formulated as an integre liner programming model:

$$\min\left(\sum_{j=1}^{n} w_j C_j\right)$$

Such that

$$T_z \le \sum_{\substack{i=1,2,..,m \\ j=1,2,..,n \\ z=1,2,..,\alpha}} p_j X_{j_z} \le S_z \qquad (1)$$

$$C_j = \sum_{\substack{j=1,2,..,n \\ z=1,2,..,\alpha}} p_j X_{j_z} + T_z \qquad (2)\ (P)$$

$$\sum_{\substack{j=1,2,..,n \\ z=1,2,..,\alpha}} X_{j_z} = 1 \text{ and } X_{j_z} \in \{0,1\} \qquad (3)$$

6 .Heuristic (H_1)

An initial solution is always necessary. For this reason, we suggest in this part the following heuristic: assigned the (best) task h where $\left(\frac{p_h}{w_h} = \min_{j \in J}\left\{\frac{p_j}{wj}\right\}\right)$ to the machine I, based on two principles justified by the two following propositions:

Proposition 3. In an optimal scheduling, it is necessary to schedule the tasks. in each availability period of the machine according to the order SWPT.

Proof. It results directly by adjacent task exchange like used by Smith (1956) for the corresponding periods.

Proposition 4. It is not useful to let the machine (idle) if a task can be assigned to this machine.

Notations

We denote by:

$J = \{1, 2, ..., n\}$: The set of tasks.

p_h : Execution time of the task h

I_{NA} : No task assigned to the machine I.

α : Number of availability periods.

$Z = \{1, 2, ..., \alpha\}$: Availability periods.

S_z ($z \in Z$) : The beginning of the unavailability period of the machine I

.

T_z $(z \in Z)$: The end of the unavailability period of the machine I

σ_z $(z \in Z)$: The set of partial sequences assigned to the machine I.

$\sigma_z = \sigma^{(1)} \cup \sigma^{(2)} \cup ... \cup \sigma^{(\alpha)}$.

$J_z\ (z \in Z) = \left\{ \begin{matrix} j \ / \ j \text{ task assigned to the machine I} \\ \text{with } T_z \le C_j \le S_z \end{matrix} \right\}$.

C_z $(z \in Z)$: Execution time of the task $j \in J_z$.

ALGORITHM

Initialization

$J = \{1,2,...,n\}, I = \{1\}; Z = \{1,2,...,\alpha\}; I_{NA} = I$, $\sigma = \varphi, f_\sigma = 0,$
$z = 1, C_z = 0$ and $T_1 = 0$.

Sort task $h \in J$ in increasing order according to the criterion p_h / w_h in a list L_1

Sort task $h \in J$ in increasing order according to the criterion p_h in a list L_2

While

$(L_1 \neq \varphi$ and $z \le \alpha)$ **do**

Begin

Set $p_{h_1} = p_h / w_h$ from the top list of L_1.

$p_{h_2} = p_h$ from the top list of L_2.

Determine the task $h \in J$ such that

$S_z - C_z >= \min(p_{h_1}, p_{h_2})$

Endif

Begin

Assigned the task h to the machine I

Delete the task h from the two lists L_1 and L_2

Compute $C_z = \sum_{j \in J_z} p_j + T_z$;

Determine $\sigma_z = \sigma_z \cup \{h\}$ and $f_\sigma = f_\sigma + w_h C_z$;

Set $J = J \setminus \{h\}$

End

Else

Begin

Set $z = z + 1$; $I_{NA} = I$;

End

Endif

End

RECAPITULATED TABLE

In table 1 below one use the followingabbreviations:

N by swap :Tabusearch by Swapping

Am perc:Percentage of improvement of the initial costs

N by blo:Tabusearch by block

In co by Heuristic : Initial cost by heuristic

7 .Experimentation and results

7.1.Data generation

The heuristic were tested on problems generated with 700 tasks similar to that used in previous studies (Adamu and Abass, 2010), (Baptiste et al, 2000), (Ho and Chang,1995), (M'Hallah and Bulfin, 2005); for each task j an integer processing time p_j was randomly generated in the interval $(1,99)$ with a weight randomly w_j chosen in interval $(1,10)$.

Number of tasks	N by swap	Am perc	N by blo	Am perc	In co by Heuristic
n = 150	96151	2%	101614	1,17%	**117409**
	155672	3%	124821	6,5%	**189178**
	74340	22%	60770	26%	**96165**
n=350	815645	15%	939556	3%	**967359**
	595982	31%	924034	6,2%	**986806**
	927050	04%	102202	1,4%	**116201**
N=700	2703709	1%	2362804	5%	**2809706**
	2683403	**2%**	**2291407**	**6%**	**2836902**

Table(1) - (Results with $\alpha = 3$)

8. Results

The results listed in table (01) show clearly that the tabu method based on neighborhood by block presents the best (lowest) costs compared with tabu method based on neighborhood by swapping. This is due to the fact that the first neighborhood ensures a faster tasks movement besides that the search space is richer with optimal partial sequences in each availability periods. This can also be explained by the nature of used neighborhoods, besides the left shifting of other tasks in the swapping neighborhood. The results show that execution time obtained by the two neighborhoods is acceptable.
On the other hand, the heuristic amelioration rate between the two neighborhoods is remarkable (FIG1. FIG2. FIG4). It is also noted that the cost amelioration rate of the proposed tabu search heuristic is situated between 1% and 26%.

7 .Conclusion

In this paper, a metaheuristic polynomial approach (Tabu search) as solution for tasks scheduling problem with single machine and unavailability periods is presented., by considering that the tabu list is dynamic and its size varies according to amelioration state of the solution. According to the curried out tests, it can be concluded that the proposed approach ensure better results (heuristic amelioration costs up to 26%). It must be noted that the neighborhood by block presents the best costs (FIG3. FIG4) with an acceptable execution time.

Reference

[1] A.adamu, MO. and Adewunmi, A..Comparative study of metaheuristics for identical parallel machines. J. Eng.Technol. Res. 5(7),(2013),: 207-216.

[2] A.adamu, MO. and Adewunmi, A. Metaheuristics for scheduling on parallel machine to minimize weighted number of early and tardy jobs. Int. J. Phys. Sci. 7(10) ,(2012);1641-1652.

[3] H.Belouadah, , Posner, ME. and Potts, CN. Scheduling with relates dates on a single machine to minimize total weighted completion time. Discrete Appl. Math..(1992), 36:213-231.

[4] F. Glover, and Hanafi, S. Tabu Search and Finite Convergence, Special Issue on Foundations of heuristics in Combinatorial Optimization. Discrete Appl. Math..(2002),119 :3-36.

[5] F. Glover,.Futurepaths for integer programming and links to artificial intelligence, Comput. Open Res,(1986),13: 533-549.

[6] P. Hansen, The steepest ascent mildest descent heuristic for combinatorial programming. In: Proceedings of the Congress on

[7] M. Haouari, and Ladhari, T. Branch and bound-based local search method for the flow shop problème. J. Oper. Res. Soc..(2003), 54:1076-1084.

[8] J.Ho, , and Chang, YL. Minimizing the number of tardy jobs for m paralell machines. Eur. J. Oper. Res.(1995),84: 334-355.

[9] C.Y.Lee,. Machine scheduling with an availability constraints. J. Global Optim. 1996 ,9:395-416.

[10] C.Y,Lee. Minimising the makespan in two machines flow shop scheduling problem with availability constraints.Oper. Res. Lett. ,. 1997,20:129-139.

[11] C.Y.Lee, Two machines flow shop scheduling problem with availability constraints. European J. Oper.Res.1999.114:420-429.

[12] R.M'Hallah, , and Bulfin, RL. Minimizing the weighted number of tardy jobs of paralell processors. Eur. J. Oper. Res. 2005.160 :471-484.

[13] C.Sadfi,.Problèmes d'ordonnancement avec minimisation des encoursThèse PhD. Institut National Polytechnique de Grenoble, France.2002

[14] Sakarovitch, M..Optimisation combinatoire: Programmation discrète. Hermann, France.1984

[15] G.Schmidt, Scheduling with limited machine availability. European J. Oper. Res. 2000,121:1-15.

[16] O .Selt and Zitouni,R.A comparative study of heuristic and metaheuristic for three identical parallel machines.cjpas..2014V 8,No.3147-3153

[17] WE.Smith,. Various optimizes for single-stage production. Naval Res. ,Logist.1956, 3:59-66.

[18] L .Yun-Chia,., H., Yu-Ming. and Chia-Yun, T Metaheuristics for drilling operation scheduling in Taiwan PCB industries. Int. J. Prod. Econ. 2013,141(1):189-198.

[19] N.Zribi, ,Kacem, I., El-Kamel, A. and Borne. P. Minimisation de la somme des retards dans un jobshop flexible. Revue e-STA (SEE), 2005. 2(2):2

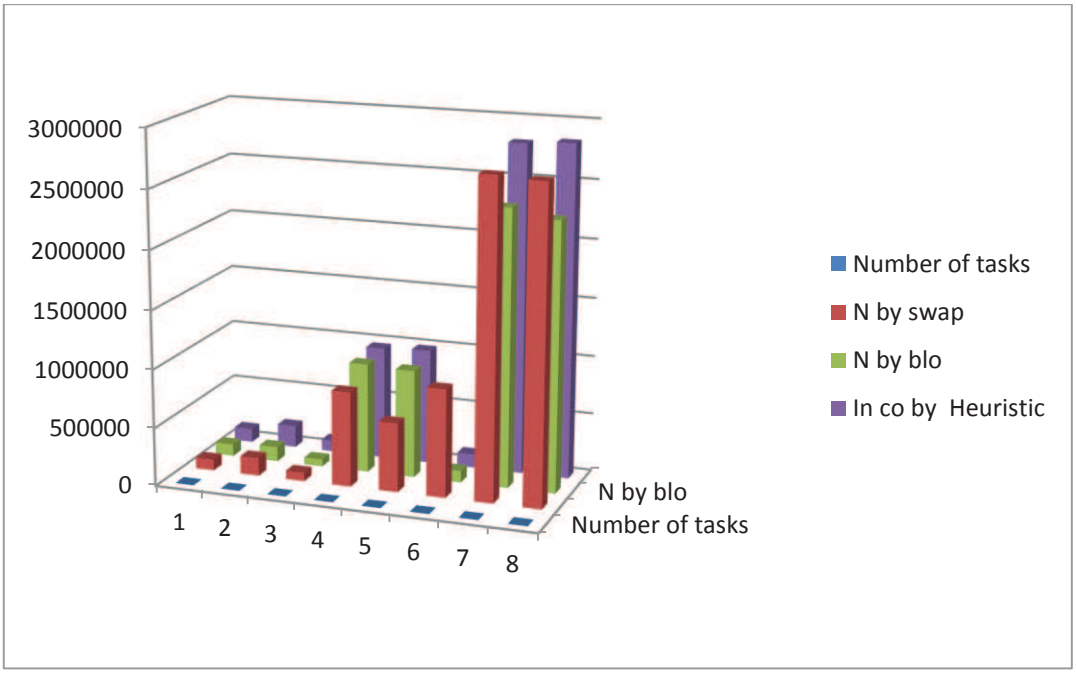


FIG1 .Percentage of heuristic cost amelioration based on metaheuristic for n=700

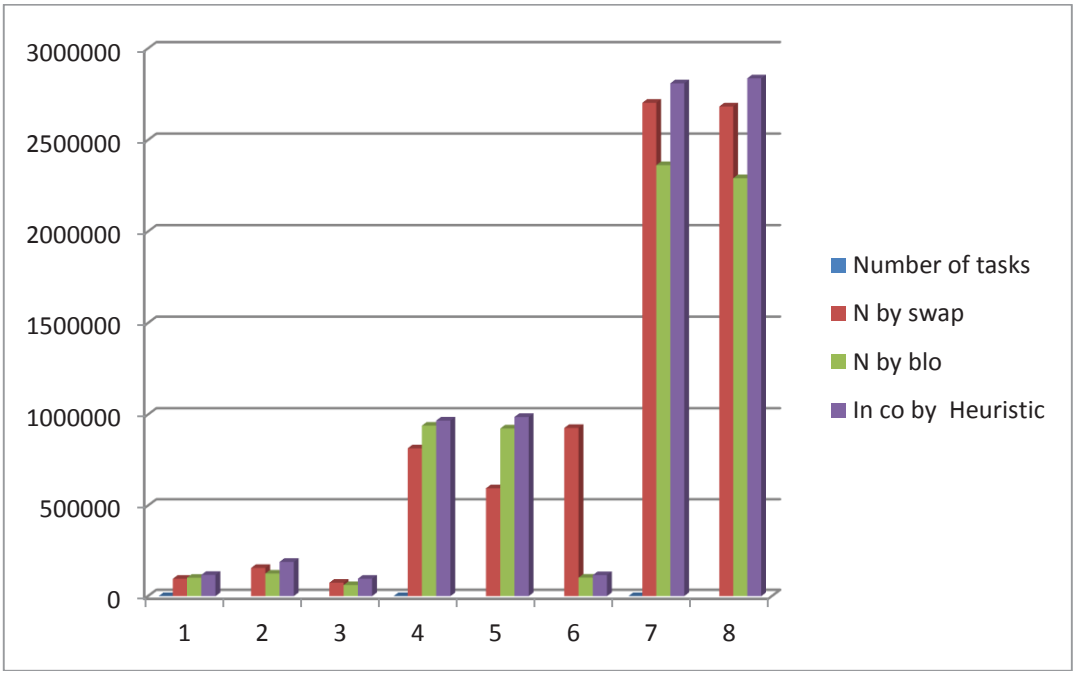


FIG2 . Comparison of heuristic and metaheuristic for n=700

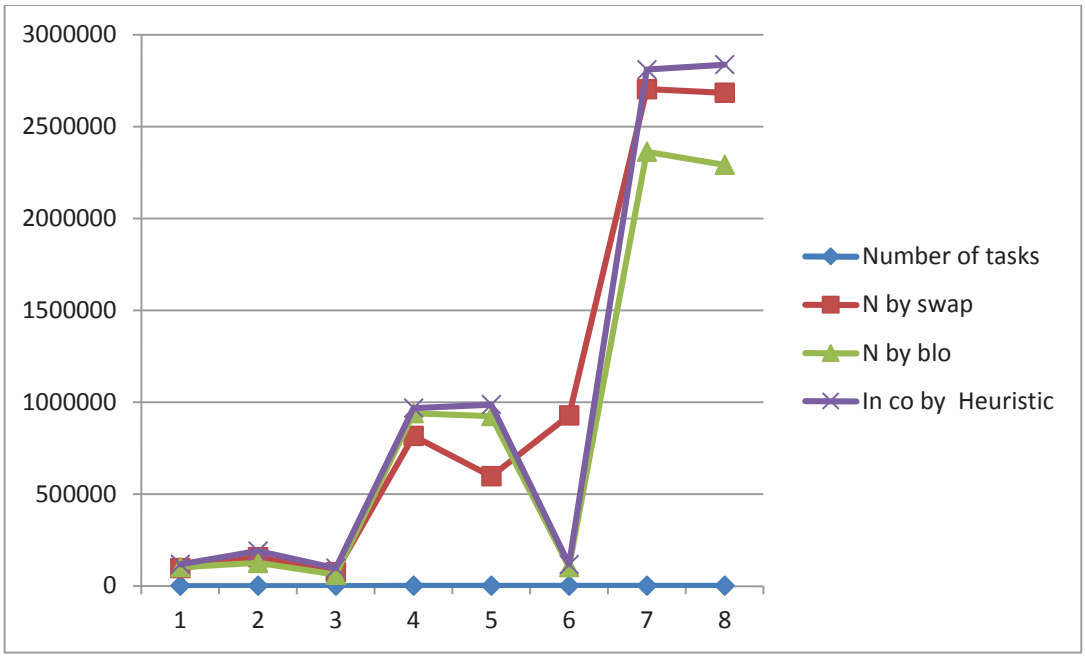


FIG3. Comparison of heuristic and metaheuristic for n=700

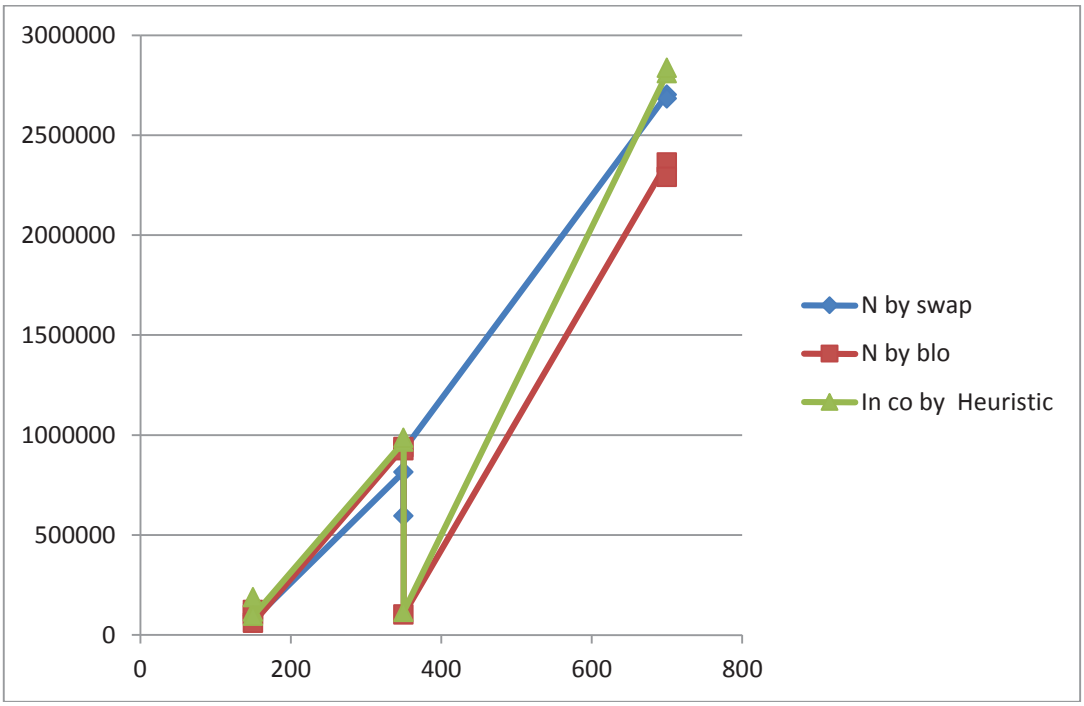


FIG4. Comparison of heuristic and metaheuristic for n=700

Bayesian network's tool for modeling and diagnosis online

Salma Chaieb
Faculté des Sciences de Monastir
Avenue de l'environnement 5019 Monastir, Tunisie
salma.chaieb2@yahoo.com

Ali Ben Mrad
CES Lab, Ecole Nationale d'Ingénieurs de Sfax (ENIS),
Université de Sfax, 3038, Sfax, Tunisie
Ali.BenMrad@univ-valenciennes.fr

***Abstract*— Due to the relatively clear semantics and ability to handle complex systems, Bayesian networks tend to be increasingly used for managing uncertainty. They are proving to be an innovative and attractive reasoning tool in many fields, in particular, the web. In this context, they have attracted a lot of attention from theorists and system developers and they were adopted in various web applications. Several software are used to model Bayesian networks such as BNT, Probayes, Hugin, Bayesia and Genie. However, few software authorize the manipulation of Bayesian networks online. In our work, we implement a dynamic website serving to create and manipulate a Bayesian network.**

***Keywords*— Bayesian network; web; modeling; learning; inference; decision-make support; diagnosis.**

I. INTRODUCTION

Bayesian networks (BN) [1], [2], [3] are by far one of the most reliable and coherent formalisms for problem solving as well as knowledge management. They are the result of a convergence between statistical methods and artificial intelligence technologies. These models proved their efficiency in various fields of application due to their convivial structure, inferential power, as well as their natural way of representing uncertainty using probabilities.

The tools, libraries and software necessary for constructing and manipulating BNs are still evolving in the direction where they develop conviviality in their interfaces in order to better represent a BN. However, most available BNs engine are not adapted for online exploiting.

The objective of this work is to conceive and implement a tool for modeling and diagnosing using Online Bayesian networks. This application will be hosted on a dynamic website which will allow users creating and manipulating BN online. Our contribution consists in formalizing the proper tools for our application using the Matlab toolbox BNT (Bayesian Network Toolbox) [4].

II. CASE STUDIES

Bayesian networks tend to be more and more used for the development of several applications. They are increasingly popular methods for modelling uncertain and complex domains such as decisional analysis, telecommunications and expert systems. At best, they provide a robust and mathematically coherent framework for the analysis of this kind of problems. They emerge from artificial intelligence research and have been applied to a wide range of problems, ranging from text analysis [5] to problems in medical diagnosis [6] and the evaluation of scientific evidence [7]. They are also increasingly used in the modelling and management of web applications.

Bayesian networks are not a household name in the web context. However, they are gaining popularity in this field and are likely to establish their position as one of the standard methods of analysis especially in problems dominated by uncertainty. With the growth on the concern about context-aware applications, Hong et al [16] have proposed a context-aware messenger application that exploits dynamic Bayesian networks to automatically infer a user's context and shares contextual information to enrich electronic communication. Recently, a goal-directed, context-sensitive, Bayesian control strategy for active sensing, termed C-DAC (Context-Dependent Active Controller), was proposed by Ahmad and Yu [8] to focus sensory and cognitive resources on the behaviorally most relevant stimuli and events in the environment. Graf and Liu [9] have suggested an automatic approach for identifying students' learning styles in LMSs (learning management systems). The approach is based on inferring students' learning styles from their behavior in an online course and was developed for LMSs in general. Horvitz, Kadie, Paek and Hovel [10] have created a computing and communication systems that sense and reason about human attention by fusing together information from multiple streams. In other research work, Tlemsani and Abdelkader [11] have proposed a web application which aims to design and implement a system of automatic recognition of isolated characters online. In the same context, Philippot, Belaïd and Belaïd [12] have proposed a method of recognition of handwritten forms online using an electronic pen clip type. This method identifies the class of the form based on the positioning of the filled fields.

III. MODELING

A. Tools

The growing research in Bayesian networks stimulated an equivalent development in tools for acquiring and processing graphical models, of which is the BNT that offer various

functionalities to allow users to manipulate Bayesian Networks. Although BNT is found on a solid algorithmic basis that is both rich and open source, it is still not adapted for online modeling of BN it also lucks a graphical interface that would help creating BN.

In this section we will present the details of implantation of our application. In this project we created a dynamic website with graphical interfaces that are flexible that allows users to manipulate Bayesian networks online and perform diagnosis in several fields without acquiring any additional software on their machines.

B. *Defining the user's needs*

It consists in creating an online Bayesian network from a database of samples. The user selects a database from his personal space and specifies the characteristics of its content (separator, missing value, ...). Once the database is uploaded, the user can introduce the structure and estimate the parameters by choosing one of the two options suggested, either manually or by learning.

The second method relies on the conceived system to learn the model by inserting the necessary parameters and returning the graph and the results of the learning of the parameters. The corresponding results are displayed on a specialized web interface and consulted by the user.

Once the BN structure is created and the probability tables are defined, the user adds the necessary observations to adjust the system to a new situation, resulting in generating inference. We have used the BNT to generate the structure, parameters and the inferences.

The user can exploit the diagnosis in numerous fields, he is asked to provide additional information through the enquiry handed over by the platform. This information is processed by the BNT in order to output diagnosis results.

C. *Building the graphical tool*

Now that we established on overview of the application and the development environment. We can focus on the making and functioning of the application.

The current work presents a tool for online modeling and diagnosing by Bayesian networks. In this paper, we present two parts. The first consist in modeling the BN online using learning algorithms to recognize the structure and the parameters provided by the BNT. The second part illustrates one example of application that validates the efficacy of our website.

1) Modeling a Bayesian Network.

The application provides the users with an online graphical interface allowing the modeling of Bayesian Networks that communicates with distance server on which Matlab software is installed.

The process of creating a BN in the application possesses the following characteristics:

- Manual: The user is free to manipulate node, arcs and probabilities table of each node.
- By learning: We apply learning algorithms in order to identify the structure and the parameters (probability tables) of a BN.
- Hybrid: This approach combines both manual input of the network structure and automatic gathering of the probabilities from a database.

We are currently interested in creating BNs by learning. In the following, we will detail the main interfaces to give an idea about the functioning of this part of the application.

a) Importing a database of samples.

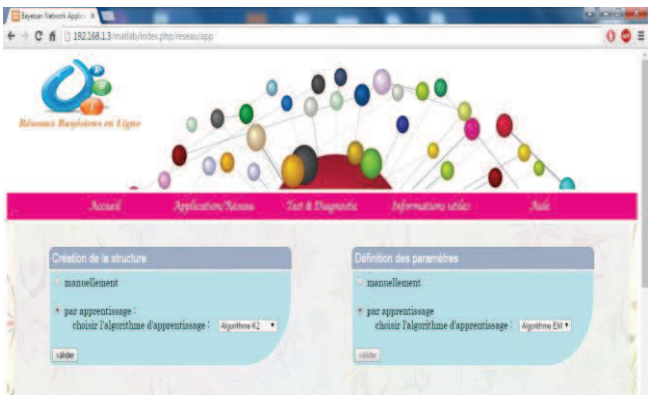
The user uploads the database from his personal space, indicates the set of characters he uses to separate fields, and then specifies if the database contains a header line for variables definition and also if it contains missing values. Once the user's options are collected, the database is saved on a Matlab server. Fig. 1 illustrates the specification of file format and samples database loading.

Fig. 1. Interface "Load database of samples".

b) Creating a Bayesian network

The examples database is now saved in the Matlab server, now the database can be exploited using the learning algorithms. Many algorithms can be used to infer the structure and parameters of BN [2]. In our work, we used the K2 and the EM algorithm. An interface appears to help creating the structure and parameters using the suggested algorithms (Fig. 2).

Fig. 2. Interface "Learning".

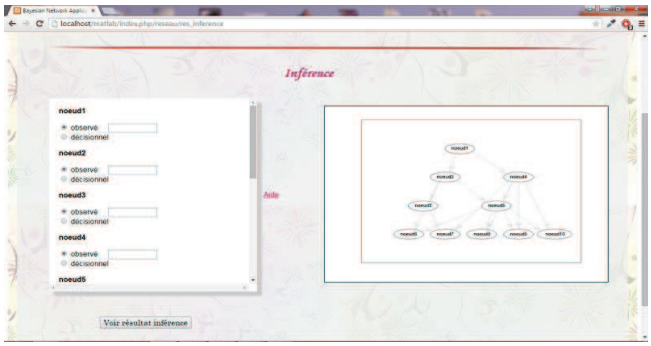


The results are displayed on the same graphical interface in a graph representing the connections between nodes of the BN created and also the probability tables. In addition, our application can convert the network to many formats compatible with different BN software and libraries.

c) Entering observations

Entering new observations consists in adjusting the model to a new situation in which the information is available. It guides the system in calculating the a posteriori probabilities distribution. This way the user can indicate the nature of each node (observed or not) by referring to the structure of the displayed network. For each node, he can add an observation as depicted in Fig. 3.

Fig. 3. Interface "Learning results".



The inference output is a table and a bar chart containing the *a posterior* probabilities of each node.

2) Case study: Medical diagnosis

In this section, we will operate on a well know example in the BN literature: The "visit to Asia" BN [2].

"Visit to Asia" database contains eight nodes; if one visits Asia recently he risks tuberculosis infection, Smoking is a risk factor for lungs cancer and bronchitis. Lungs radiology diagnosis cannot recognize lungs cancer and tuberculosis and it also cannot detect the presence of dyspnea. According to these facts we need the knowledge database to allow determining the probabilities of positive diagnosis for each disease.

In order to perform such diagnosis the user has to choose the appropriate category « Health », thereby an interface appears with a single choice question.

After inserting the observations, the application outputs a chart indicating the probability of positive diagnoses for each disease and provides the corresponding results in percentage (Fig. 4).

Fig. 4. Interface "Diagnosis results".

IV. Conclusion

In this paper we provided an overview of the application "Modeling and diagnosis by online Bayesian networks". This application is an online platform allowing knowledge modeling using Bayesian networks. Once the network is established, an inference component can process the available observations. The application can find solutions and help decision making using the diagnosis available in various fields.

Looking ahead, we are interested on improving our website. Thereby this application can serve as a starting point for later study. A number of issues remain to be addressed to refine the development of our application. On the one hand we aim to host your website to be public in order to test the well-functioning of our application in terms of response time and expected results. On the other hand it will be possible to implement a graphical tool that allows the user to create his Bayesian network manually. We also suggest vary the choice of algorithms used for both learning and inference and expand our testing on other bases examples and benchmarks.

We believe that it is important to continuously verify the adequacy of the model as we develop it. The functions and needs will be studied in more detail. Our application will be incrementally modified and improved in order to win the challenge of quality and better adapt to the environment and users.

References

[1] Jensen,F. , "An introduction to Bayesian Networks", Springer, New York, 1996.

[2] P.Naïm, P.Wuillemin, P.Leray, O.Pourret, ET A.Becker, "Réseaux bayésiens", 3ème Edition Eyrolles, 2004.

[3] Pearl J., " Probabilistic reasoning in intelligence systems : networks of plausible inference", Morgan Kaufmann, 1988.

[4] Murphy, K, The bayes net toolbox for matlab. "Computing science and statistics", 33(2), 1024-1034, 2001.

[5] Dong, A., Agogino, A.M., "Text analysis for constructing design representations", Artif. Intell. Eng. 11, 65–75, 1997.

[6] Kahn Jr., C.E., Roberts, L.M., Shaffer, K.A., Haddawy, P., "Construction of a Bayesian network for mammographic diagnosis of breast cancer". Computers Biol. Med. 27 (1), 19–29., 1997.

[7] Garbolino, P., Taroni, F., "Evaluation of scientific evidence using Bayesian networks". Forensic Sci. Int. 125, 149–155, 2002.

[8] Ahmad, S., & Yu, A., "Active sensing as bayes-optimal sequential decision making". arXiv preprint arXiv:1408.2056, 2014.

[9] Graf, S., & Liu, T. C., "Supporting teachers in identifying students' learning styles in learning management systems: an automatic student

modelling approach.", Journal of Educational Technology & Society, 12(4), 3, 2009.

[10] Horvitz, E., Kadie, C., Paek, T., & Hovel, D., "Models of attention in computing and communication: from principles to applications.", Communications of the ACM, 46(3), 52-59, 2003.

[11] TLEMSANI, Redouane et BENYETTOU, Abdelkader. "Application des réseaux bayésiens dynamiques à la reconnaissance en-ligne des caractères isolés". In : 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications. 2007.

[12] Philippot, E., Belaïd, Y., & Belaïd, A, "Classification de formulaires manuscrits en-ligne à l'aide de réseaux bayésiens". In Colloque International Francophone sur l'Ecrit et le Document-CIFED, (pp. 95-110), March 2010.

[13] Bessière, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K, "Bayesian programming". CRC Press, 2013.

[14] DEMPSTER, Arthur P., LAIRD, Nan M., et RUBIN, Donald B. Maximum likelihood from incomplete data via the EM algorithm.

[15] Journal of the royal statistical society. Series B (methodological), p. 1-38, 1977.

[16] Hong, J. H., Yang, S. I., & Cho, S. B., "ConaMSN: A context-aware messenger using dynamic Bayesian networks with wearable sensors.", Expert Systems with Applications, 37(6), 4680-4686, 2010.

[17] DEMPSTER, Arthur P., LAIRD, Nan M., et RUBIN, Donald B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), p. 1-38, 1977.

[18] Roberto G. Cowell, A. Philip Dawid, Stefenl L. Lauritzen, and David.J. Spiegelhalter, "Probabilistic Networks and Expert Systems". Springer, 1999.

[19] David Heckerman, A tutorial on learning with bayesian network. In Michael I. Jordan, editor, "Learning in Graphical Models", pages 301–354. Kluwer Academic Publishers, Boston, 1998.

[20] Paul Krause, "Learning probabilistic networks", 1998.

[21] Judea Pearl., "Causality : Models, Reasoning, and Inference", Cambridge University Press, Cambridge, England, 2000.

[22] Peter Spirtes, Clark Glymour, and Richard Scheines, "Causation, prediction, and search.", Springer-Verlag, 1993.

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Tweedie Hidden Markov Random Field and the Expectation-Method of moments and Maximization algorithm for Brain MR Image Segmentation

Mourad Zribi[1], Mouna Zitouni[2]*, Afif Masmoudi[3]

Keywords: EM algorithm-Markov Random Field-Method of moments- Segmentation-Tweedie

ABSTRACT

In this paper, a segmentation algorithm of brain Magnetic Resonance Imaging (MRI) is presented. The proposed method is based on Tweedie Hidden Markov Random Field (THMRF) processing and the Expectation- Method of moments and Maximization algorithm (EMM). THMRF is a Markov Random Field (MRF) whose state sequence cannot be observed directly and whose spatial information is encoded through the mutual influences of neighboring sites. EMM algorithm is obtained by the combination of both EM (Expectation-Maximization) algorithm and the method of moments. It is an iterative technique for determining the maximum likelihood estimates (MLEs) of the mean parameter μ and the estimator of dispersion parameter λ: The algorithm has been validated on synthetic data and tested on real images. The proposed method of segmentation can be detect the brain tumor.

[1] Laboratoire d'Informatique Signal et Image de la Cote d'Opale (LISIC-EA 4491), ULCO, 50 Rue Ferdinand Buisson BP 719, 62228 Calais Cedex France.
[2] Laboratory of Probability and Statistics.
* corresponding author: zitounimounaa@yahoo.fr

A loose hierarchical Bayesian network for clustering high-dimensional data

Hasna Njah
MIRACL laboratory
FSEGS - University of Sfax
Sfax, Tunisia
njah.hasna@gmail.com

Salma Jamoussi
MIRACL laboratory
ISIMS - University of Sfax
Sfax, Tunisia
salma.jamoussi@isimsf.rnu.tn

Walid Mahdi
College of Computers and Information Technology
Taif University – Saudi Arabia
wmahdifr@gmail.com

***Abstract* — With the emergence of high-dimensional data in various application domains, the accurate clustering of instances becomes a challenging task due to the problem of the curse of dimensionality. We propose a new loose hierarchical Bayesian network model for instances clustering. This new model takes into account the complex relationships that exist among the features. It uses the latent variables for ensuring the convergence to an optimal partition of the instances. We propose a generalized algorithm for learning the proposed Bayesian network model. Our method escapes the over-fitting problem of the Bayesian network learning with a limited number of instances. The experimental results show the efficiency of our proposed model as well as the proposed learning algorithm on benchmark datasets.**

Keywords — Bayesian network; latent variable; clustering; high-dimensional data

I. INTRODUCTION

The exponential growth of the automatic information processing has led to the development of data acquisition techniques such as the sensors, the computer-assisted management systems, the activity log applications, etc. The collected data is, therefore, consolidated so as to form high-dimensional data and it is used for modeling real-life problems. Mining the knowledge behind the expression of such data yields to rich results that support decision-making tasks. In the context of this paper, we are interested in finding groups of individuals that have similar behaviors given a set of features. This is referred to *Cluster analysis* or *Clustering* of instances. There is a variety of clustering approaches that have been applied on many datasets and have shown competitive performances. Accordingly, new trends of clustering are focusing on the enhancement and the adaptation of the classic algorithms so as to take into account the high-dimensionality of the used datasets.

Actually, high dimensional data is characterized by a high number of features (variables) and a very low number of instances. Affected by the *Curse of dimensionality* problem [1], clustering these instances with such data-expression engenders two main challenges. From the one hand, the distance between each pair of individuals becomes meaningless, given that an instance is expressed over a large vector. Therefore, the value of the found distance does not reflect the degree of similarity between two given instances. From the other hand, it is likely that the features of the used dataset are correlated. With such assumption, it is possible that two different subsets of features lead to different partitions of instances [2]. Consequently, it is primordial to examine the dependency's relations that exist among the whole set of features before running the instances clustering routine. The aim of extracting these relations is to abstract the information among correlated features.

In this article, we opt for the Bayesian Network modeling [3] for tackling the above-mentioned problems. It is defined as a graphical and a probabilistic model where the features are represented by a set of nodes and corresponding random variables. The dependencies among the features are modeled through directed edges between pairs of nodes so as to form an acyclic graph. The probability distribution of each random variable is computed based on its corresponding node's parents in the graph. We use this Bayesian network model for representing the features in a given high-dimensional dataset. Since these features are very likely to be correlated, we propose to introduce a finite set of *latent variables* in order to represent these complex dependencies among the features. Recall that the latent variable is a hidden variable that controls the high dependencies between groups of features. Learning a Bayesian network from a high-dimensional data can be easily affected by the curse of dimensionality problem. In fact, it is necessary that the number of instances has to be very high as compared to the number of features; which is not the case of high-dimensional data. Therefore, the introduction of the latent variables is useful for ensuring the building of a generalized Bayesian network.

We propose a new clustering model based on a loose hierarchical Bayesian network with latent variables. Our proposed model has a tree-like structure where the leaf nodes represent the observed set of features of a given training dataset and the inner nodes represent the latent variables. The aim of using the latent variables is to provide a sub-partition of the instances based on a given set of highly dependent observed features. It is possible that certain features, of the training dataset, are dependent to various groups. Hence, our proposed hierarchical structure is loose, i.e. supports multiple parents. The root of the obtained Bayesian network is therefore a global latent variable that enables to cluster the instances given the learnt latent variables. Accordingly, our proposed algorithm for learning this loose hierarchical Bayesian network with latent variables proceeds on three steps: feature fuzzy clustering, latent variable modeling and global clustering.

II. Related Work

Clustering high dimensional-data has been handled by different ways. The most intuitive point of view consists on applying a classical clustering algorithm preceded by a dimension reduction method. Features' selection methods were used for assessing the observed set of features, ranking them and selecting the most discriminative ones [4]. These methods are simple and easily implemented. However, it is not guaranteed that the obtained sub-set of features is efficient for an optimized partition of instances. Feature extraction methods, such as the Principal Component Analysis [5] and Latent Semantic Analysis [6], were also applied. These methods mainly focus on eliminating information redundancy through the correlated features. They look for a low-dimensional representation of the training data in order to perform a given classical clustering algorithm. Experimentally, these methods have shown a great performance as compared to the feature selection methods. However, the transformation of the feature space, through these matrices' manipulation methods, does not guarantee an optimal instance partition. It has been shown mathematically and experimentally that vectors with the highest eigenvalues do not necessary contain the optimal discriminative information [7].

Furthermore, the most distinguished related work to Bayesian network modeling for instances clustering is the *Autoclass* model [8]. It consists on creating a simple hierarchical Bayesian network structure where the leaf-nodes are the set of observed features and the root node is a latent variable that clusters the instances. It is assumed, in this model, that the features are independent one one-to-another; they have a common parent which is the latent variable. This Bayesian clustering algorithm has shown remarkable ability to not only automatically find the number of clusters but also provide a relevant instances partition. Unfortunately, *Autoclass* does not take into account the dependencies between the features. This leads to a deterioration of its clustering performance with high-dimensional data. Some attempts were trying for enhancing this hierarchical model by inserting more latent variables. A famous example is the strict hierarchical model proposed by [9]. It allows the modeling of the complex relations between the observed features through stacks of latent variables. This model was successfully applied for a Bayesian classification context and was not adapted for the clustering's finality. Moreover, it was remarkably overtaken by the semi-hierarchical Bayesian structure [10] because this latter used a loose latent architecture for modeling complex relationships between the observed features.

III. Proposed Method

Our contribution is an instance clustering model based on a loose hierarchical Bayesian network structure. It deals with high-dimensional data. It is based on the assumption that different sub-sets of features lead to different instances clustering. It also uses the latent variables for abstracting the groups of highly dependent features. An example of the proposed loose Bayesian network for instances clustering is given in Fig.1. where the training dataset contains 8 observed features $\{F_1, F_2, \dots, F_8\}$. These features are grouped, according to the dependency between them, into 3 clusters. Each cluster is represented by a latent variable LV_1, LV_2 and LV_3. Each of these latent variables gives a partial partition of the instances according to its corresponding selected cluster of the observed variables. For example, the latent variable LV_1 allows to cluster the instances by using the observed features F_1, F_2 and F_3. Similarly, LV_2 allows to cluster the instances by using the observed features F_3, F_4, F_5 and F_6. Finally, the three obtained latent variables are used for learning an enclosing latent variable, represented as GLV, which provides a global partition of the instances.

Accordingly, we propose a new clustering model based on Bayesian network modeling and latent variable learning. Our proposed algorithm starts by applying a "Features fuzzy clustering" module in order to group similarly expressed features. Subsequently, a latent variable is learnt in order to abstract each of the found clusters of features. Finally, this same module, the "Latent variable learning", is applied on the whole set of the found latent variables in order to provide the global partition of the instances given all the observed features.

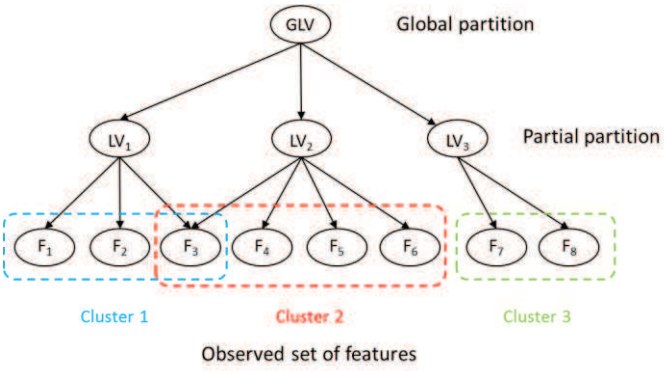


Fig. 1. Example of the loose hierarchical Bayesian network model

A. Features fuzzy clustering

The first step of our proposed algorithm consists on mining overlapped groups of highly-dependent features of the used training dataset. In that manner, even the most complex relationships between the observed features will be modeled in the obtained Bayesian model for instances clustering. It was shown experimentally in recent studies that the graphical clustering approach is adequate for clustering the features in high-dimensional data [10, 12]. Besides, the classical distances, such as the Euclidean distance and Manhattan distance, are not suitable for measuring the degree of dependency between two given variables. Thus, we use the Mutual Information (1) for measuring the quantity of the flow of information among a pair of features/variables noted V_1 and V_2.

$$MI(V_1, V_2) = \sum_{v_1 \in V_1} \sum_{v_2 \in V_2} P(v_1, v_2) log \frac{P(v_1, v_2)}{P(v_1)P(v_2)} \quad (1)$$

Consequently, our clustering algorithm starts by computing the Mutual Information between all couples of features. These values are used, subsequently, for building a weighted dependency-based graph. We obtain a complete graph where the nodes represent the features and the weighted undirected edges represent the degree of dependency between them. We propose

the use of an intuitive thresholding function (2) in order to eliminate the low-weighted edges. We define α as a coefficient that determines the proportion of significant values of the Mutual Information. The value of α can be set manually depending on the category of the studied problem.

$$\min(MI(V_i,V_j)) + [\alpha * (\max(MI(V_i,V_j)) - \min(MI(V_i,V_j)))] \quad (2)$$

Having the graph that models the dependency relationships between the observed variables, we look for the highly connected cliques in order to represent them by a latent variable. Two overlapped cliques are represented by two latent variables (See the example of the clusters 1 and 2 in Fig.1).

B. Latent variable learning

Once the fuzzy clusters of features are identified, each group of highly dependent observed variables is represented by a latent variable. In this step, we address more precisely the problem of estimating the cardinality (i.e. the number of states) of the latent variable. The cardinality represents the number of clusters in the corresponding sub-partition. Actually, there is a tradeoff between candidate values of the latent variable's cardinality. In fact, low values of the cardinality lead to information loss and high values lead to complex parameter estimation.

We use the Equilibrium Criterion [11] for estimating the optimal cardinality of each latent variable. This criterion ensures a statistic balance of the latent variable. In fact, the optimal cardinality of this hidden variable is reached when the Log Likelihood (LL) of the partial structure is approximated to the overall Mutual Information (MI) between the latent and observed variables. This is formally written as in (3) where a given latent variable LV^k has k states and represents a cluster that contains n_c observed features ($F_1, F_2, \dots, F_{nc}$).

$$k^* \approx arg\min_{k\geq 2} |MI\left(LV^k|F_1,F_2,\dots,F_{n_c}\right) - LL(LV^k|F_1,F_2,\dots,F_{n_c})| \quad (3)$$

Once the cardinality of the latent variable is determined, we proceed to learning the parameters of this hidden variable. To do so, we adopted the same principle that was used in [10] for finding the values of the latent variables. Then, we apply the Maximum Likelihood Estimation [13] algorithm for computing the probability distribution of the latent variable given its representative set of observed features. When all the latent variables are fully learnt, we repeat these steps on the obtained latent variables in order to learn a last latent variable that does the global clustering (See the example of Fig.1 where GLV is the parent of LV_1, LV_2 and LV_3).

C. Loose hierarchical Bayesian Network learning

As it is shown in the example of Fig.1, the structure of our proposed instances' clustering model is a loose hierarchical Bayesian network. It is described as a three-levelled tree. The first level is the leaves of the tree which represent the observed features of the used training dataset. The second level is composed of a set of latent variables. Each of them provides a possible partition of the instances according to its corresponding sub-set of observed features. Finally, in third level contains the root of this Bayesian network's structure. This root is a latent variable that provides a global partition of the instances. This global instances' partition is based on the partial partitions provided by each of the latent variables learnt in the second level of the tree.

Having a cluster of n_q features ($F_1, F_2, \dots, F_{nq}$) and a corresponding latent variable LV_q, the partial partition of the instances given LV_q is determined as $P\left(LV_q\middle|F_1,F_2,\dots,F_{n_q}\right)$. The computation of this quantity is done in (4) through the application of the Bayes theorem.

$$P\left(LV_q\middle|F_1,F_2,\dots,F_{n_q}\right) = \frac{P\left(F_1,F_2,\dots,F_{n_q}\middle|LV_q\right)P(LV_q)}{P(F_1,F_2,\dots,F_{n_q})} \quad (4)$$

The prior probability of $P\left(F_1,F_2,\dots,F_{n_q}\middle|LV_q\right)$ is determined through the parameter estimation step [13], the probability distribution of the latent variable $P(LV_q)$ is already inferred from the values determination step [10] and the probability distribution of the observed features of the cluster $P(F_1,F_2,\dots,F_{n_q})$ is computed based on the used training dataset.

These partial partitions are very useful for converging to a "good" clustering. Unfortunately, the high-dimensional data is subject to the curse of dimensionality problem [1]. This means that the number of instances is very limited as compared to the number of features. Consequently, the learnt clustering model for instances can easily fall in the over-fitting problem. This gets more accentuated by the fact that the Bayesian network needs a rich training dataset in terms of instances for ensuring a generalized model. To tackle this problem, we adopt the principle of ensemble learning which has shown a great performance for escaping the over-fitting problems with the Bayesian networks [14]. We use all the partial partitions provided by each latent variable of the second level of the proposed hierarchy. These partitions are combined so as to provide a final instances clustering that optimizes as much as possible the discrimination between the instances. Correspondingly, the global latent variable (GLV), which is the root of the proposed structure, combines the results of the latent variables in the second level of the hierarchy. Its probability distribution is based indirectly on the original set of features and it is determined as in (5) where c is the number of the found overlapped clusters of features.

$$\mathrm{P(GLV|LV_1, LV_2, \dots, LV_c)} = \prod_{\mathrm{i=1}}^{c} \mathrm{P(GLV|LV_i)} \quad (5)$$

The posterior probability of the global latent variable, GLV, given a latent variable LV_i of the second level of the proposed hierarchy is also computed in (6) based on the Bayes theorem.

$$P(GLV|LV_i) = \frac{P(LV_i|GLV)P(GLV)}{P(LV_i)} \quad (6)$$

The prior probability of a given latent variable LV_i given the global latent variable is also inferred through the parameter learning step [13] that we adopted in our proposed algorithm. Likewise, the probability distribution of the global latent variable P(GLV) is inferred by the values determination process

[10]. Finally, the probability distribution of LV_i is, indeed, computed based on the corresponding observed set of features. Therefore, $P(LV_i)$ is computed based on the observed features in its corresponding cluster (4).

IV. Results and Discussion

The perquisite of our proposed loose hierarchical Bayesian network model for instances clustering consists on (i) the use of a fuzzy feature clustering in order to take into account the complex relationships between the observed set of features, (ii) the automatic determination of the number of clusters and (iii) the robustness to the low number of instances in the high-dimensional data. In this section, we assess the ability of our algorithm to estimate the optimal number of clusters and we evaluate the quality of the obtained partition while using our new proposed model for instances clustering.

A. Estimation of the number of clusters

For certain application domains, the number of the clusters of instances, in a given training dataset, is unknown. The cluster analysis process, therefore, is required to automatically look for the optimal number of groups of instances. We propose to compare our adopted method for estimating this number Equilibrium Criterion (EC) [11] with some of the most known clustering algorithms that offer this functionality. The compared algorithms are:

- The EAST algorithm [15]. It uses the Hill-climbing routine in order to find the optimal cardinality of the LV. The algorithm adds or removes the states of the LV so as to get a simple hierarchical model (i.e. structure and parameter). The optimal model is determined by a modified version of the BIC score. The root of the obtained structure is a latent variable that clusters the wholes set of instances.
- The density based algorithm DBSCAN [16]. As its name implies (Density-based spatial clustering of applications with noise), DBSCAN is a density based algorithm. It finds the clusters based on the density distribution of the dataset. Two given instances are merged when they present a low distance value. Therefore, the number of highly dense found clusters is the optimal number of clusters of the corresponding dataset.
- The X-means [17] variation of the K-means algorithm [18]. It runs the K-means clustering algorithm with different number of clusters and return the partition that corresponds to the optimal number of cluster. The determination of such partition is done through an intern clustering validation index such as the within cluster inertia.
- Ensemble clustering [14]. The idea is quite similar to X-means. It consists on running K-means with different numbers of clusters and choosing the optimal partition through a set of internal clustering validation indices (Calinski–Harabasz, Simplified Silhouette, Dunn's, Davies–Bouldin, Xie–Beni and S_DBW).

We run these algorithms on 5 benchmark datasets: Iris. Ionosphere, Glass, Wine and Vote. We report, in TABLE 1, the obtained number of clusters when applying each of the above mentioned algorithms. We indicate in the second column the real number of clusters in each dataset. The last line of the table contains the score of each clustering algorithm (i.e. the number of times the algorithm succeeds in finding the real number of clusters). We remark that the EC method for estimating the optimal cardinality of the latent variable, hence the optimal number of instances' clusters, and the X-means algorithm succeed in pioneering the other algorithms (EAST, DBSCAN and Ensemble learning). As the X-means finds the optimal number of clusters after several runs of K-means, we consider that it is not suitable for high-dimensional datasets where the number of features is very high. Therefore, the EC is more straightforward and is promising in terms of running-time optimization.

B. Partition assessment

In this subsection, we evaluate the quality of the obtained partition while executing our proposed instances clustering algorithm based on the loose hierarchical Bayesian network with latent variables. We apply 4 different clustering algorithms: K-Means [18], as a partitioning method, the Hierarchical Clustering (HC), as a hierarchical method, DBSCAN [16], as a density-based method, and the Expectation-Maximization (EM) algorithm as a model-based method. In order to ensure a fair comparative study, we initialize the above mentioned algorithms with the correct number of the instances' clusters. We apply our method as well as these 4 clustering algorithms on relatively large benchmark datasets. We note, in TABLE. II, the number of features and the number of classes of each of the chosen datasets.

We record the cluster to class error rate of each obtained partition as compared to the original cluster assignment of the used training dataset. This is called the cluster to class evaluation and it is the number of times an instance is incorrectly attributed to a given cluster as reference to its original class. We note, in TABLE III, the obtained results of the comparison.

TABLE I. Comparison of the number of clusters estimation

Dataset	Number of clusters	EC	EAST	DBSCAN	X-MEANS	Ensemble clustering
Iris	3	3	1	3	3	2
Ionosphere	2	2	Out-of-memory	1	4	2
Glass	6	5	1	1	6	6
Cancer	2	2	2	1	2	3
Vote	2	2	2	13	4	4
Score		**4/5**	**2/5**	**1/5**	**4/5**	**3/5**

TABLE II. DATASETS DESCRIPTION

Dataset	Number of features	Number of clusters
Glass	9	6
Labor	16	2
Ionosphere	35	2
Soybean	35	19

TABLE III. CLUSTER TO CLASS ERROR RATES COMPARISON

Dataset	Our method	K-means	HC	DBSCAN	EM
Glass	0.27	0.57	0.63	0.55	0.57
Labor	0.22	0.22	0.35	0.21	0.21
Ionosphere	0.24	0.28	0.35	0.27	0.25
Soybean	0.35	0.91	0.84	0.75	0.33

We remark that our method tops up the other clustering algorithms in terms of the provided partition. In fact, the lower is the Cluster-to-class rate, the more adequate are the found clusters. This means that the obtained partition resembles as much as possible to the original partition of classes. Also, the low values of this score indicate that our method is more likely to ensure the correct affectation of each instance to its corresponding cluster. Furthermore, we remark that our method is suitable for multiple class problems (i.e. datasets with a high number of classes).

V. CONCLUSIONS AND FUTURE WORK

We proposed a new method for clustering the instances of a high-dimensional dataset. Our proposed clustering model is a loose hierarchical Bayesian network with latent variables. Our proposed model is characterized by the use of a fuzzy feature clustering in order to take into account the complex relationships between the observed set of features. It is also distinguished by the possibility to the automatically determine the number of clusters of a given dataset. The use of the latent variable, in this clustering model, is beneficent. From the one hand, it abstracts the complex relations between the features. From the other hand, it provides a partial partition of the instances while using a sub-set of features. Our proposed method is, therefore, robust to the over-fitting problem caused by the low number of instances in the high-dimensional data. The experimental results validated these advantages of our proposed instances' clustering model. As future work, we propose to enhance our proposed algorithm so as to be able to cluster the instances of extremely high-dimensional datasets (i.e. with thousands of features). We also opt for applying our method for real-life problems.

REFERENCES

[1] Bellman, Richard E., and Stuart E. Dreyfus. Applied dynamic programming. Princeton university press, 2015.

[2] Assent, Ira. "Clustering high dimensional data." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.4 (2012): 340-350.

[3] Jensen, Finn V. An introduction to Bayesian networks. Vol. 210. London: UCL press, 1996.

[4] Dash, Manoranjan, and Huan Liu. "Feature selection for clustering." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2000.

[5] Brubaker, S. Charles. "Robust PCA and clustering in noisy mixtures."Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2009.

[6] Song, Wei, and Soon Cheol Park. "A novel document clustering model based on latent semantic analysis." Semantics, Knowledge and Grid, Third International Conference on. IEEE, 2007.

[7] Chang, Wei-Chien. "On using principal components before separating a mixture of two multivariate normal distributions." Applied Statistics (1983): 267-275.

[8] ACS, RI. "Bayesian classification (autoclass): Theory and results." (1996).

[9] Zhang, Nevin L., Thomas D. Nielsen, and Finn V. Jensen. "Latent variable discovery in classification models." Artificial Intelligence in Medicine 30.3 (2004): 283-299.

[10] Njah, H., Jamoussi, S., and Mahdi, W. (2015, November). "Semi-hierarchical naïve Bayes classifier". Neural Networks (IJCNN), 2016 IEEE, International Joint Conference on (pp. 958-965).

[11] Njah, H., Jamoussi, S., Mahdi, W. and Masmoudi, A. "A new equilibrium criterion for learning the cardinality of latent variables." Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on. IEEE, 2015.

[12] Song, Qinbao, Jingjie Ni, and Guangtao Wang. "A fast clustering-based feature subset selection algorithm for high-dimensional data." IEEE transactions on knowledge and data engineering 25.1 (2013): 1-14.

[13] Redner, R.A. and Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2), pp.195-239.

[14] Njah, H., and Jamoussi, S. "Weighted ensemble learning of Bayesian network for gene regulatory networks." Neurocomputing 150 (2015): 404-416.

[15] Chen, Tao, et al. "Model-based multidimensional clustering of categorical data." *Artificial Intelligence* 176.1 (2012): 2246-2269.

[16] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[17] Pelleg, Dan, and Andrew W. Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters." *ICML*. Vol. 1. 2000.

[18] MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967.

Ant-B-learn : Learning Bayesian networks structures by Ant Colony Optimization

Marzougui Boutheina
Faculty of Economics and Management
Department of Computer Science
Applied to Management
Email: bouthaina_marzougui@yahoo.fr

Mariem Gzara
MIRACL: Multimedia InforRmation systems and Advanced Computing Laboratory
BP 1030, Sfax 3018, TUNISIA
Higher School of Computer Science and Mathematics
at the University of Monastir
Avenue de la Korniche - B.P. 223 - Monastir - 5000, Tunisia

Salma Jamoussi
MIRACL: Multimedia InforRmation systems and Advanced Computing Laboratory
BP 1030, Sfax 3018, TUNISIA
High School of Computer Sciences and Multimedia
University of Sfax

***Abstract*—This paper proposes a new adaptation of the ant colony optimization metaheuristic to learn Bayesian networks structures. The Ant based algorithm for Bayesian networks structures learning (Ant-B-Learn) is a score based algorithm where ants are constructive heuristics with both forward and backward behavior. Experimental results on real data sets and by comparison with other heuristic based algorithm have shown the effectiveness of the Ant-B-Learn.**

I. Introduction

As a method of reasoning under uncertainty Bayesian networks (BNs) have become so popular within the artificial intelligence probability and uncertainty community. In many practical applications the BNs is unknown and there is a need to learn it from data. This problem is known as the BN learning problem, informally its stated as follows: given training data and prior information (e.g. expert knowledge, causal relationships) estimate the graph topology (network structure) and then learning network parameters. Learning the BN structure is considered as a harder problem than learning the BN parameters because once the BN structure has been learned, the parameters are usually estimated (in the case of discrete variables) using the relative frequencies of all combinations of variable states as exemplified in the data. Learning the structure from data by considering exhaustively all possible structures is not feasible in most domains, regardless of the size of the data, since the number of possible structures grows exponentially with the number of nodes. Hence, structure learning requires methods that, in a large space of candidate models, identify a solution that approached the best the optimal model. Metaheuristics are stochastic search methods in a space of hypothesis that proceed by sampling an objective function until converging to an optimal or a near optimal solution. The Ant Colony Optimisation (ACO) metaheuristic has been successfully applied to a large number of difficult problems and it is well adapted to solve problems modeled by graphs. In this paper, we present a new algorithm for solving the problem of the structure learning of BNs from data based on the Ant Colony Optimization (ACO) metaheuristic. The Ant-B-Learn algorithm has demonstrated its efficiency on benchmark data sets when compared with other metheuristics based algorithms. The reminder of the paper is organized as follows. In section two we remind basic notions related to the Bayesian Networks theory and the problem of structure learning of Bayesian networks from data. In section three, we give a brief overview of related works. In section four, we describe the basic ideas of Ant Colony Optimisation metaheuristic. In section five, we detail how we have adapted the ACO to learn Bayesian networks structures. In section six, we give and comment the experimental results. Finally, we end with a general conclusion and some perspectives

II. Basic Notions Related to Bayesian Networks

Bayesian networks are graphical structures for representing the probabilistic relationships among a large number of variables and doing probabilistic inference (implication) with those variables [1]. So they are a useful tool in the representation of uncertain knowledge which is a human reasoning.

A Bayesian network consists of the following:

- There is a set of variables and a set of directed edges between variables.
- Each variable has a finite set of mutually exclusive states.
- The variable together with the directed edges form an acyclic directed graph (DAG).
- To each variable A with parents $B_1, , B_n$ there is attached a conditional probability table $P(A|B_1, , B_n)$.
- If A has no parents, then the table reduces to the unconditional probability table $P(A)$.

The DAG of a Bayesian network model is a compact graphical representation of the dependence and independence properties of the joint probability distribution represented by the model. Note that if A has no parents, then the table reduces to the unconditional probability table $P(A)$. For the DAG in Fig.1., the prior probabilities $P(A)$ and $P(B)$ must be specified.

To exploit information in data bases, it is necessary to transform them into probabilistic graphical models. Learning a Bayesian network from data involves tow subtasks: learning the structure of the network (i.e. determining what depends on what) and learning the parameters (i.e. the strength of the dependencies). Structure learning is a hard task, because of all difficulties related to it, we mention for example:

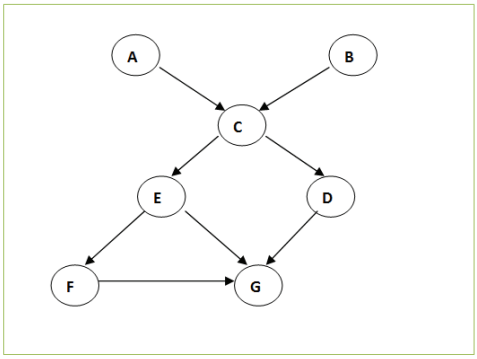


Fig.1. A directed acyclic graph (DAG). The probabilities to specify are $P(A)$,$P(B)$,$P(C|A,B)$,$P(E|C)$,$P(D|C)$,$P(F|E)$, and $P(G|D,E,F)$.

- Node ordered requirement: many BN-learning algorithms require additional information, notably an ordering of the nodes to reduce the search space.
- Computational complexity: the space of all Bayesian network structures is extremely large because the number of different structures$f(n)$, grows more than exponentially in the number n of nodes.

There are basically two approaches used for learning the structure of Bayesian networks; constraint-based and score-based [2]. The constraint based methods establish a set of conditional independence statements holding for the data, and use this set to build a network with d-separation properties corresponding to the conditional independence properties determined. The Score-based learning assigns a number (a score) to each Bayesian network structure. The objective of score based learning is to resolve the problem of the exhaustive search by ideas like restriction at the tree space, nodes ordering and Greedy search [3].

III. Related works

Metaheuristics such as Genetic Algorithms and Ant Colony Optimization were applied to solve the problem of the structure learning of Bayesian networks. The first adaptation of the genetic algorithm to the problem is due to Larranage [4]. A Bayesian network structure is a one dimensional array that concatenates the lines of the adjacency matrix of the DAG of the BN. Later, Larranage and al [5] propose four hybrid algorithms to overcome the drawbacks of their first one. They impose an order to matrix elements, so if c_{ij} designs the existence of an arc between variable i and variable j, the inequality $i > j$ is originated in the assumed ancestral order between the variables. Because of the inequality the crossover and the mutation operators to be used are closed operators. And that helps to generate valid models.

In J. Lee and al [6] a BN structure is represented as a pair of chromosomes (ordering chromosome and the connection chromosome):. The ordering chromosome consists of indices of variables. If there are n root nodes (nodes with no parents), the first n genes of the chromosome are the indices of the root nodes. The next genes are the indices of the children nodes of the root nodes without overlaps. This structure is repeated until the whole variables show up. The connection chromosome denotes the connectivity matrices which defines the dependency relation between the variables. The genetic operations are only applied to the first chromosome. But the second one is also automatically changed to accommodate the change of the first chromosome.

The first adaptation of the ACO metaheuristic for learning Bayesian Networks structures was called ACO-B[7] and was proposed by Luis M. de Campos and all. ACO-B is a scoring based algorithm which main idea is: each ant constructs a BN structure by adding one edge at a time while starting with an empty graph while taking into account the pheromone intensity of an edge and the heuristic information which is based on measuring the k2 score during the process of the construction of the BN structure in a feasible solution space. In their approach, JI Jun-Zhong and all [8] combine the ideas of two basic approaches to learn a BN structure with ACO. The first approach on which they are based is learning structure of a BN by means of performing effective conditional independence tests on sample data and distinguishing all connected relationship among nodes. The second approach is performed by considering a learning model of a BN structure to learn it using score functions.

In this paper we learn BN by an adaptation of the ACO. The Ant-B-Learn allows forward and backward decisions in the ants construction heuristic.

IV. Basic ideas of Ant Colony Optimisation metaheuristic

Ant Colony Optimization was introduced by Marco Dorigo and his colleagues in the early 1990s. The first ACO algorithm was called Ant System (AS)[9] and it was aimed to solve the travelling salesman problem (TSP). The TSP is a problem of a salesman who wants to find a shortest path starting from one city, passing once by a given set of cities and then return to departure city. For Dorigo and his colleagues they modeled the TSP problem as a graph (N, E, w) where N is the set of cities, E is the set of edges between cities and w is the set of edges weight that corresponds to the distance between two cities. They consider a colony of m ants where m is a parameter of the algorithm. When the ant a_k is positioned on the city i, it has to choose the next city to visit among $\Gamma(k,i)$ cities. In order to satisfy the constraint that an ant visits all the n cities once, when a tour is completed; the list $\Gamma(k,0)$ is composed of all the cities initially and every time an ant visit a city i it is removed from Γ. The ant a_k moves from its current city i to the city j with the transition probability defined as:

$$p_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha . [\eta_{ij}]^\beta}{\sum_{l \in \Gamma(k,i)} [\tau_{il}(t)]^\alpha . [\eta_{il}]^\beta} & if \text{ j} \in \Gamma(k,i) \\ 0 & \text{otherwise} \end{cases}$$

(1)

$\tau_{ij}(t)$ is a function of the trail intensity which gives information about how many ants in the past have chosen that same edge (i,j); the second, the visibility η_{ij} which says that the closer j to i is the more desirable it is.

The parameters α and β are two desirability measures that control the relative importance of trail versus visibility.When the ant completes a tour, the value of the tour cost L_k to minimize is computed and the values $\Delta\tau_{ij}^k$ are updated.

At the start, initial values τ_0 for trail intensity are set on edges. Then, the trail intensity is updated at each iteration of the ACO according by ants traversing the arcs and a global pheromone update is performed by the best ant. This process is iterated until the tour counter reaches the maximum (user- defined) number of cycles, or all ants make the same tour. We call this last case stagnation behavior because it denotes a situation in which the algorithm stops searching for alternative solutions.

V. ANT-B-LEARN : AN ANT BASED ALGORITHM FOR BAYESIAN NETWORKS LEARNING

The proposed ant colony optimization algorithm for Bayesian network structure learning (Ant-B-Learn) is a score based algorithm. The Ant-B-Learn algorithm is a constructive heuristic where each ant constructs a complete solution to the problem while maximizing the bayes metric score function. The behavior of an ant is function of the global knowledge of the colony and of some trivial heuristic that might be used by the ant if it is left at its own.

In previous works based on the ant colony metaheuristic to solve this problem, each ant starts from an empty graph and adds at each step one edge to its partial solution. Thus, the ants havent the possibility to turn back on decisions made on the previous iterations. In the Ant-B-Learn algorithm, ants have the possibility to make a backtracking by deleting an edge from the current graph if this deletion increases the score of the BN.

In the next sub-sections, we will describe in details the Ant-B-Learn algorithm.

A. Ant constructive heuristic

To construct its own solution each ant $k(k = 1, 2, ., m)$ starts from the empty graph G_0 (arcs-less DAG) and proceeds either by adding or by deleting (after adding some arcs) an arc at one time until no score gain is possible.

We denote by G_h the current graph modeling an incomplete BN structure. G_h is a graph with all nodes $X_i \in X$, exactly h arcs and no directed cycles. Suppose there are a feasible candidate directed arcs to add and d feasible candidate directed arcs to delete. The addition of an arc is forbidden if it already exists in the current graph or it leads to a cycle formation. The deletion of an arc is not feasible if it leads to a reduction of the score value. By mean of heuristic information and pheromone intensity the ant selects one arc a_{ij} and add it to the graph G_h, or select an existing arc a_{ef} and remove it from the graph G_h. Thus, the new state is denoted as $G_{h+1} = G_h \cup \{a_{ij}\}$ in the case of arc addition and $G_{h-1} = G_h - \{a_{ef}\}$ in the case of graph deletion. A modification of the state of the graph G_h is allowed only if it improves the score value. Once there is no way to make the score of a Bayesian network structure higher by adding or deleting an arc, the construction process is ended and the ant gets its solution G_g. The figure Fig.2. below shows the construction mechanism.

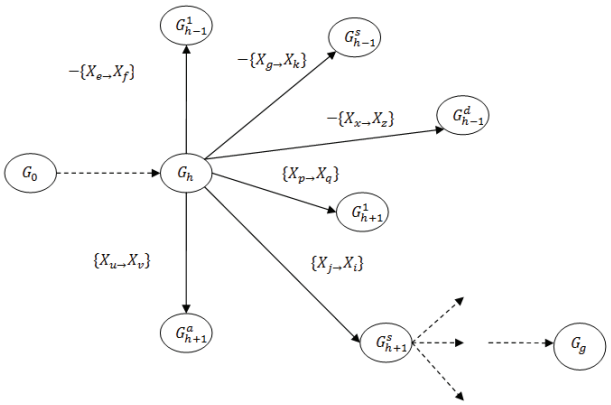


FIG.2. The construction process of a BN for an ant

Let $\Gamma(G_h, +)$ be the set of feasible arcs that can be added to the partial graph G_h . The ant selects from $\Gamma(G_h, +)$, the arc (i, j) that maximizes the quantity $[\tau_{rl}(t)].[\eta_{rl}](t)]^\beta$ where $(r, l) \in \Gamma(G_h, +)$. Let $\Delta_{score}(G_h, +, (i, j))$ be the increase of the score value due to the addition of the arc (i, j) to the graph G_h.

Let $\Gamma(G_h, -)$ be the set of feasible arcs that can be deleted from the partial graph G_h. The ant selects from $\Gamma(G_h, -)$, the arc (i, j) that maximizes the quantity $[\eta_{rl}(t)]^\beta/[\tau_{rl}(t)]$ where $(r, l) \in \Gamma(G_h, -)$ and let $\Delta_{score}(G_h, -, (i^{'}, j^{'}))$ be the increase of the score value due to the deletion of the arc $(i^{'}, j^{'})$ from the graph G_h.

An ant k could whether proceeds by adding or deleting an arc according to a probabilistic transition rule that it selects a directed arc a_{ij} to add or select a directed arc $a_{i^{'}j^{'}}$ to delete, from the current candidate arcs to add or to delete. The decision rule (3.5) applied by an ant at the decision point where it has a current partial graph G_h (which was an empty graph at first decision point) is the following:

$$\text{(u,v)=}\begin{cases} argmax(\Delta_{score}(G_h,+,(i,j)),\Delta_{score}(G_h,-,(i^{'},j^{'})) & \text{if q} \le q_0 \\ \text{where} \\ (i,j) = argmax_{r,l\in\Gamma(G_h,+)}\{[\tau_{rl}(t)].[\eta_{rl}(t)]^\beta\} \\ (i^{'},j^{'}) = argmax_{r,l\in\Gamma(G_h,-)}\{\frac{[\eta_{rl}(t)]^\beta}{[\tau_{rl}(t)]}\} \\ \text{I,J} & \text{otherwise} \end{cases} \tag{2}$$

The term $\eta_{rl}(t)$ represents the heuristic information of the directed arc a_{rl} and $\tau_{rl}(t)$ is the quantity of the pheromone deposited on the arc a_{rl}. The parameter β is a weighted coefficient which controls the importance of the heuristic information $\eta_{rl}(t)$ versus the pheromone rate $\eta_{rl}(t)$ to influence the selection of arc. If β equals 0 only the pheromone rate influences the decision of the ant. The parameter $q_0(0 < q_0 \le 1)$ determines the relative importance of exploitation versus exploration, $q(q \in [0, 1])$ is a random number.

- First case we have $q \le q_0$ so :
  If $\Delta_{score}(G_h, +, (i, j))$ is higher than $\Delta_{score}(G_h, -, (i^{'}, j^{'}))$ then the arc (i, j) is added otherwise a deletion of the arc $(i^{'}, j^{'})$ is performed.
- Second case $q > q_0$ so:
  I and J are a pair of nodes randomly selected to be add according to the probability in (3).

$$p^k_{ij}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha.[\eta_{ij}(t)]^\beta}{\sum_{r,l\in\Gamma(G_h,+)}[\tau_{rl}(t)]^\alpha.[\eta_{rl}]^\beta} & if \text{ (i,j)} \in \Gamma(G_h,+) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Where the parameter α depicts the relative importance of the pheromone $\tau_{rl}(t)$ left by real ants.

B. Local pheromone update

Initially, the pheromone intensity of every directed arc is a constant value τ_0, i.e.,$\tau_{ij}(0) = \tau_0$. When an ant completes the construction of its solution, it performs a local update of the pheromone intensity on all directed arcs selected in this solution. If the arc a_{ij} belongs to the current solution then the pheromone level of the corresponding arc is changed in the following way:

$$\tau_{ij}(t + 1) = (1 - \psi)\tau_{ij}(t) + \psi\tau_0 \tag{4}$$

Where τ_0 is a constant related with the initial solution, $0 < \psi \le 1$ is a parameter that controls the pheromone evaporation.

C. Global pheromone update

At each iteration of the algorithm, the colony updates the best solution obtained so far by means of the bayes metric, and performs the global updating rule for each arc of the current best solution.

$$\tau_{ij} = (1-\rho)\tau_{ij} + \rho\Delta\tau_{ij} \quad (5)$$

With :

$$\Delta\tau_{ij} = \begin{cases} \frac{1}{|f(G^+:D)|'} & \text{if } a_{ij} \in G^+ \\ \tau_{ij} & \text{otherwise} \end{cases} \quad (6)$$

Where $0 < \rho \leq 1$ is also a parameter of the pheromone evaporation, and $f(G^+ : D)$ is the metric value of the best solution G^+.

D. The algorithm

The main difference between our algorithm and the previous ones which are based on the ACO is that in the Ant-B-Learn algorithm, an ant has the possibility to return back on previous decisions and it has the possibility to delete an already added arc in the previous iterations.

The Ant-B-Learn algorithm executes a number of NC iterations where at each iteration the m ants of the colony construct independently their solutions and the best one will be picked by mean of their score, the algorithm is given below.

Algorithm 1 Ant-B-Learn Algorithm

procedure BEGIN PROCEDURE
  1) Initialization
  Initialize m τ_0 and NC
  2) Search Procedure
  NC time iteration
  For $l = 1$ **to** NC **do**:

    i) For $k = 1$ to m do:
    $G_k = AntSearchProcedure()$;
    Perform local pheromone updating
    ii) $G^+_{(l)} = argmax_k f(G_k : D)$;
    iii) If $(f(G^+_{(l)} : D) \geq f(G_k : D))$ then $G^+ = G^+_{(l)}$;
    iv) Perform global pheromone updating ;

  EndFor
  return G^+
end procedure

E. The score metric

As the learning goal is to achieve the best BN structure whose score is the maximum, the heuristic information function of a directed arc can be interpreted as the greatest increase produced in f score when the arc is added or deleted from the graph. The heuristic information will be calculated by mean of Bayes [10] metric which is a decomposable score.

According to the decomposability of metric Bayes the operator that adds an arc $X_j \to X_i$ to the current G_h will bring:

$$\begin{aligned} &f(G_{(h+1)} : D) - f(G_h : D) = \\ &f(X_i, \Pi(X_i) \cup (X_j)) - f(X_i, \cup(X_i)) \end{aligned} \quad (7)$$

The operator that deletes an arc $X_e \to X_f$ to the current G_h will bring:

$$\begin{aligned} &f(G_{(h-1)} : D) - f(G_h : D) = \\ &f(X_f, \Pi(X_f) - X_e) - f(X_f, \Pi(X_f)) \end{aligned} \quad (8)$$

The Bayes metric of a Bayesian network structure G_D for a database D is:

$$\begin{aligned} &Q_{Bayes}(G_S, D) = \\ &P(G_S)\Pi_{i=0}^{n}\Pi_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_i)}\Pi_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \end{aligned} \quad (9)$$

Where

- $P(G_S)$ is the prior on the network structure (taken to be constant hence ignored in the weka implementation).
- $\Gamma(.)$ the gamma-function.
- D database.
- G_S network structure.
- r_i $(1 \leq i \leq n)$ cardinality of x_i.
- q_i the cardinality of the parent set of x_i in G_S (the number of different values to which the parent of x_i can be instantiated).
- N'_{ij} and N'_{ijk} represent choices of priors on counts restricted by $N'_{ij} = \Sigma_{k=1}^{r_i} N'_{ijk}$.
- N_{ij} $(1 \leq i \leq n, 1 \leq j \leq q_i)$ number of records in D for which $pa(x_i)$ takes its jth value.
- N_{ijk} $(1 \leq i \leq n, 1 \leq j \leq q_i, 1 \leq k \leq r_i)$ number of records in D for which $pa(x_i)$ takes its jth value and for which x_i takes its kth value.So $N_{ij} = \Sigma_{k=1}^{r_i} N_{ijk}$

VI. EXPERIMENTAL RESULTS

To evaluate the performance of our approach, we have performed the experimental evaluation on the Weka software on benchmark data sets. The table give for each considered data set the number of attributes; their type, the number of instances and the number of classes.

Data set	Attributes	Type	Instances	Classes
Weather	5	Numeric	14	2
Iris	5	Numeric	150	3
Glass	10	Numeric	214	7
prima-diabets	9	Numeric	764	2
german-credit	21	Nominal	1000	2
breast-cancer	10	Nominal	286	2
Vote	17	Nominal	435	2

TABLE I: data sets

The parameters of the algorithm are fixed after many simulation as follow: $\alpha = 0.75$, $\beta = 0.9$, $\tau_0 = 0.7$, $q_0 = 0.9$, $\psi_0 = 0.75$ and $\rho_0 = 0.75$

A. Evaluation measures

To evaluate the performance of a learner, we use some known measures which may either be maximized or minimized.

- **Correctly and Incorrectly Classified Instances:** shows the rate of test instances that were correctly and incorrectly classified.
- **The Kappa coefficient:** measure the degree of agreement of two or more judges which are the classifier and the actual class of the example. The agreement / disagreement between the two judges can be read directly into the matrix of confusion measure whose value is even larger than the matrix is diagonal. The Kappa coefficient is calculated as follows:

$$kappa = \frac{P_0 - P_e}{1 - P_e} \quad (10)$$

Where P_0 : The proportion of the sample on which both judges agree (i,e the main diagonal of the confusion matrix). And $P_e = \frac{\sum_i p_i p_{.i}}{n^2}$
where

- p_i: sum of the elements of the line i
- $p_{.i}$: sum of the elements of the column i
- n : sample size[11]

- **Mean Absolute error:** For each example, we calculate the difference between the probability (calculated by the classification) for an example of belonging to his true class, and its initial probability of belonging to the class that has been fixed in the set of examples (in general, this probability is 1). It then divides the sum of these errors by the number of instances in the set of examples [11].

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \tag{11}$$

where

- Actual target values: $a_1 a_2 \ldots a_n$
- Predicted target values: $p_1 p_2 \ldots p_n$

- **Precision and Recall:** In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance [11].

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

- **F-Measure :** A measure that combines precision and recall is the harmonic mean of precision and recall [Wikipedia]. This amount is used to group into a single number performance classification (for a class) as regards the Recall and Precision [11]:

$$F_{Measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{14}$$

- **ROC-Area :** Receiver Operating Characteristic (ROC) curves measure the capability of a binary test or classifier to correctly distinguish between positive and negative instances.

B. Discussion of the results

First we test Ant-B-Learn (A-B-L) algorithm and compare the results obtained by our algorithm with those of the K2, Genetic Search (GS), Hill Climber (HC) , Simulated Annualing (SA), Tabu Search (TS) which are learning algorithms in weka. Then we compare the results with an ACO algorithm (named Ant-ADD) which only add arcs in their process of structure construction. For 10 runs of every algorithm, we give the mean of different evaluations measures to compare our results to other approaches. Despite they show great results for small data sets Genetic Algorithm couldnt handle with a big number of attributes. Ant-B-Learn shows good results compared to K2 and other algorithms in weka. Based on the table of results, we can clearly see that Ant-B-Learn shows best results for most datasets in weka, in terms of values of Correctly Classified Instances and kappa static. It also has the lower values for the error in general; an algorithm which has a lower error rate will be preferred as it has more powerful classification capability and ability. The Asia database is a large real database which is used to test the performance of approaches; it contains 6058 instances and 8 attributes. The results on this data set confirm the efficiency of the Ant-B-Learn algorithm.

Data sets	Alg	CCI	Kappa	MAE	P	FM	RA
Weather	A-B-L	42.85%	-0.365	0.503	0.351	0.386	0.333
	Genetic Search	64.28%	0.186	0.396	0.629	0.632	0.689
	K2	42.85%	-0.365	0.516	0.351	0.386	0.333
	HillClimber	42.85%	-0.365	0.516	0.351	0.386	0.333
	SimulatedAnnualing	64.28%	0.186	0.381	0.629	0.632	0.778
	TabuSearch	42.85%	-0.365	0.516	0.351	0.386	0.333
Iris	A-B-L	93.33%	0.9	0.058	0.934	0.933	0.979
	Genetic Search	94%	0.91	0.049	0.94	0.94	0.979
	K2	92.66%	0.89	0.04	0.927	0.927	0.98
	HillClimber	94.66%	0.92	0.0387	0.947	0.947	0.98
	SimulatedAnnualing	92.66%	0.89	0.052	0.927	0.927	0.979
	TabuSearch	93.33%	0.9	0.0417	0.934	0.933	0.931
breast-cancer	A-B-L	74.47%	0.25	0.37	0.733	0.702	0.635
	Genetic Search	-	-	-	-	-	-
	K2	70.27%	0	0.418	0.494	0.58	0.471
	HillClimber	68.93%	0.2601	0.3784	0.691	0.695	0.597
	SimulatedAnnualing	73.77%	0.247	0.375	0.718	0.699	0.602
	TabuSearch	69.93%	0.2601	0.3784	0.691	0.695	0.597
german-credit	A-B-L	70%	0.151	0.38	0.662	0.659	0.677
	Genetic Search	-	-	-	-	-	-
	K2	72.1%	0.25	0.346	0.698	0.698	0.742
	HillClimber	72%	0.2678	0.3516	0.7	0.703	0.728
	SimulatedAnnualing	71.7%	0.2758	0.3447	0.7	0.704	0.73
	TabuSearch	68.7%	0.1257	0.3613	0.645	0.648	0.708
prima-diabets	A-B-L	74.86%	0.434	0.307	0.749	0.747	0.809
	Genetic Search	-	-	-	-	-	-
	K2	74.47%	0.41	0.32	0.737	0.737	0.792
	HillClimber	73.56%	0.3947	0.3188	0.728	0.729	0.794
	SimulatedAnnualing	74.34%	0.423	0.313	0.739	0.74	0.8
	TabuSearch	73.56%	0.3947	0.3188	0.728	0.729	0.794
Glass	A-B-L	72.89%	0.632	0.105	0.731	0.725	0.874
	Genetic Search	-	-	-	-	-	-
	K2	70.56%	0.59	0.103	0.695	0.876	0.847
	HillClimber	70.56%	0.5946	0.1071	0.708	0.685	0.868
	SimulatedAnnualing	72.42%	0.6258	0.1101	0.727	0.727	0.878
	TabuSearch	70.56%	0.5946	0.1071	0.708	0.685	0.868
Vote	A-B-L	93.56%	0.864	0.104	0.936	0.936	0.972
	Genetic Search	-	-	-	-	-	-
	K2	94.71%	0.88	0.063	0.947	0.947	0.987
	HillClimber	94.25%	0.8794	0.0666	0.943	0.943	0.98
	SimulatedAnnualing	94.94%	0.8936	0.066	0.95	0.949	0.99
	TabuSearch	94.48%	0.8841	0.0691	0.945	0.945	0.97
Asia	A-B-L	85.787%	0.713	0.238	0.861	0.861	0.862
	Genetic Search	85.737%	0.712	0.238	0.861	0.859	0.856
	K2	85.176%	0.699	0.242	0.853	0.861	0.857
	HillClimber	84.912%	0.693	0.249	0.850	0.857	0.857
	SimulatedAnnualing	84.912%	0.693	0.249	0.850	0.857	0.863
	TabuSearch	84.912%	0.693	0.249	0.850	0.859	0.857

TABLE II: A Comparison between Ant-B-Learn algorithm and other algorithms in weka for different data sets

VII. CONCLUSION AND PERSPECTIVES

The Ant-B-Learn algorithm merges a local score measure with pheromone intensity on arcs to find best solutions. Dealing with an empty graph, ants proceed by adding arcs and then iteratively whether by adding or deleting arcs in the aim to construct a structure with maximized score. The empirical results show good results first when comparing the performance of our approach to approaches in weka like HillClimber and K2 algorithm, second to a developed algorithm which mainly just adds arcs without a possibility of removing arcs when constructing a solution. We have done also many tests to fix parameters related to ACO metaheurstic because these parameters influence the quality of the results. Evaluation measures in weka give us a clear evaluation of our approach, because it gives us the opportunity to test it on small and large data sets and to compare it with other approaches. Our future work is applying our algorithm to some real-life data mining problems, and improving the approach either by applying a new score or extending this approach to more complicated problems like incomplete data sets hidden variable, and

multi-relational data.

REFERENCES

[1] Richard E. Neapolitan, *Learning Bayesian Networks*, 1rd ed. Northeastern Illinois University Chicago Illinois, 2003.

[2] Finn V. jensen and Thomas D. Nielsen, *Bayesian Networks and Decision Graphs*, 3rd ed. Springer , Series: Information Science and Statistics February 8, 2007.

[3] Patrick Naim, Pierre-Henri Wuillemin , Philippe Leray, Olivier Pourret et Anna Becker, *Rseaux Baysiens*, 3rd ed Eyrolles. 2008.

[4] Pedro Larranaga, Mikle Poza, Yosu Yurramendi, Roberto H. Murga, and Cindy M.H. Kuijpers, *Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters*. IEEE 9 SEPTEMBER , 1996.

[5] Pedro Larranaga, Roberto Murga, Mikel Poza and Cindy Kuijpers , *Structure Learning of Bayesian Networks by Hybrid Genetic Algorithms*. Departement of Computer Science and Artificial Intelligence University of the Basque Country, P.O. Box 649,20080 San Sebastian, Spain.

[6] Jaehun Lee, Wooyong Chung, and Euntai Kim, *A New Genetic Approach to Structure Learning of Bayesian Networks*. IInternational Journal of Control Automation and Systems (Impact Factor: 1.07). 04/2010; 8(2):398-407. DOI: 10.1007/s12555-010-0227-3.

[7] Luis M. de Campos, Juan M. Fern andez-Luna, Jose A. Gamez, Jose M. Puerta, *Ant colony optimization for learning Bayesian networks*. International Journal of Approximate Reasoning (Impact Factor: 1.98). 11/2002; DOI: 10.1016/S0888-613X(02)00091-9, (2002).

[8] JI Jun-Zhong, ZHANG Hong-Xun HU Ren-Bing, LIU Chun-Nian , *A Bayesian Network Learning Algorithm Based on Independence Test and Ant Colony Optimization*. ACTA AUTOMATICA SINICA Vol.35, No.3 March, 2009.

[9] Marco Dorigo, Vittorio Maniezzo Alberto Colorni, *The Ant System: Optimization by a colony of cooperating agents*, 3rd ed. IEEE Transactions on Systems, Man, and CyberneticsPart B, Vol.26, No.1, 1996, pp.1-13.

[10] Remco R.Bouckaert, *Bayesian Networks Classifier in Weka for Version 3-5-7*. The university of WAIKATO May 12,2008.

[11] http://stackoverflow.com/questions/2903933/how-to-interpret-weka classification

Deep Learning For Concept Extraction

Amal Bouraoui Salma Jamoussi Abdelmajid Ben Hamadou

Multimedia InfoRmation system and Advanced Computing Laboratory (MIRACL)
Sfax-Tunisia

***Abstract*— In recent years, deep neural networks have won numerous contests in various natural language processing tasks. It tries to mimic the human brain, which is capable of processing and learning from the input data and solving different kinds of complicated tasks well without require neither any hand craft features nor separately tuned components. Word embeddings resulting from neural networks models have been shown to be a great asset for a large variety of NLP tasks such as sentiment analysis and computing semantic similarity. In this paper, we explore some well known deep learning models proposed in the NLP field. Then, we review some contributions based on word2vec, well known neural network model. Finally, we present an application for concepts extraction based on word2vec as words representations method and K-means as clustering algorithm.**

Keywords—deep learning; word embedding; semantic similarity; word2vec; natural language processing, concept extraction

I. INTRODUCTION

Natural language processing (NLP) aims to process text with computers in order to analyze it, extract information and eventually to represent the same information differently. We may want to associate categories to parts of the text (e.g. POS tagging or sentiment analysis), structure text differently (e.g. parsing), or computing semantic similarity of words in order to determine a list of concepts or similar words or sentences or documents. The use of neural networks for NLP applications is attracting huge interest in the research community and they are systematically applied to all NLP tasks.

In recent years we have witnessed a resurgence in the field of neural networks which it is based on the representation of the data through multiple layers. This topic is referred to as deep learning. Recent studies have shown that deep learning methods perform very well on various natural language processing tasks. In most NLP approaches, documents or sentences are represented by a sparse bag-of-words representation. There is now a surge of interest in single word vector spaces which goes beyond this by adopting a distributed representation of words, by constructing a vector space representation for each word, called a word embedding. These distributed representations have proven the capacity to encapsulate the semantic and syntactic structure of language while remaining a fixed-length real-valued vector, allowing for easy usage and extension into a variety of tasks. An important step was the introduction of continuous representations of words. These word embeddings are now the state-of-the-art in NLP. Among successful works, we cite the study of Collobert and Weston [1], Turian et al., 2010 [2], Collobert et al. [3] and Mikolov et al. [4]. Reference [4] introduce the word2vec model, an efficient method for learning high quality vector representations of words. We focus on this model as it represents an interesting state-of-the-art for distributed semantic representation in the NLP field and we describe some contributions based on this neural network method. At the end of this paper, we present our application to extract the semantic concepts of the considered basedata. To attend our goal, we use the word2vec model to represent words and K-means as unsupervised classifier to build up concepts.

The remainder of this paper is organized as follows: Section 2 introduces the concept of deep learning, some of its architectures, word embeddings and semantic similarity. We summarize some relevant previous works for natural laguage processing in Section 3. Word2vec and some of its based work is presented in Section 4. In section 6, we present our application for concept extraction. Section 5 is conclusion.

II. BACKGROUND

A. Deep Learning

Deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations, with higher levels representing more abstract concepts. A central idea of deep learning is referred to as greedy layerwise unsupervised pre-training, which is to learn a hierarchy of features one level at a time [5]. The features learning process can be purely unsupervised and is trying to learn a new transformation at each level to be composed with the previously learned transformations. The greedy layerwise unsupervised pre-training [6-8] is based on training each layer with an unsupervised learning algorithm, taking the features produced at the previous level as input for the next level. Each iteration of unsupervised feature learning adds one layer of weights to a deep neural network. It is then straightforward to extracted features either as input to a standard supervised machine learning classifier or as initialization for a deep supervised neural network. The most attractive quality of deep networks is that they can perform well without any external hand-designed resources or time-intensive feature engineering.

The architecture of a deep neural network is drawn in Figure 1.

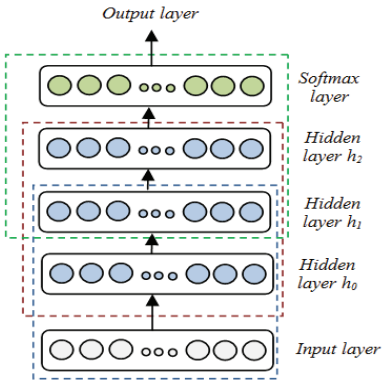


Fig. 1. The architecture of deep neural networks

B. Deep architectures

There are several deep architectures, but DBNs, auto-encoders and RNNs have been successfully used for numerous tasks.

Deep Belief Networks are probably amongst the most famous and basic kind of deep neural network architectures. It is a generative probabilistic model with one visible layer at the bottom and many hidden layers up to the output. Each hidden layer unit learns the statistical representation via the links to the lower layers. The more higher the layers are the representations and they are the more complex [9]. Deep Belief Networks are stacks of Restricted Boltzmann Machines (RBMs) followed by fine tuning. RBM is a two-layer network, which can be trained reasonably efficiently in an unsupervised fashion.

Autoencoders are simple learning networks which aim to transform inputs into outputs with the least possible amount of distortion. Stacked auto-encoders are stacked auto-encoders and they are trained bottom up in unsupervised fashion, followed by a supervised learning phase to train the top layer and fine-tune the entire architecture.

Recurrent Neural Networks (RNNs), have connections that have loops, adding feedback and memory to the networks over time. This memory allows this type of network to learn and generalize across sequences of inputs rather than individual patterns. In fact, new information is added in every layer (every network iteration), and the network can pass this information on for an indefinite number of network updates.

C. Word embeddings

Distributed word representations are a critical component of many natural language processing applications such as information retrieval, search query expansions, or representing semantics of words. It is common to represent words as indices in a vocabulary, but this fails to capture the rich relational structure of the lexicon. In recent years, vector-based models, also referred to as word embeddings, have proven to do much better in this regard. They consist at mapped each word of the corpus text to a set of numbers in a high-dimensional space. These word embeddings have shown the capacity to encapsulate the semantic and syntactic structure of language while remaining a fixed-length real-valued vector. They encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space.

D. Semantic Similarity

Understanding the meaning of a piece of text is a fundamental challenge of natural language processing. Humans can compute semantic similarity due to their background knowledge about words and their interpretation ability. However, for a machine to "understand" a piece of text and compute semantic similarity of words which have multiple meanings is still remained as an ambiguous task. It must be noted that the meaning of each word is depend on the words in its context and humans can interpret the meaning of a word according to its context. The main challenge refers to machines and how they deal with natural language and interpret concepts. Distributed representations of words in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words.

III. Deep Learning for NLP

Neural network based deep learning methods have recently been shown to perform well on various natural language processing tasks. A number of methods have been explored to train and apply word embeddings using continuous models for language. In the well-known work on NLP, Collobert and Weston [1] developed and employed a convolutional neural network architecure to learn embeddings in an unsupervised manner through a contrastive estimation technique. They define a unified architecture for natural language processing that solve simulateously a number of classic problems including part-of-speech tagging, chunking, named entity tagging, semantic role identification and similar word identification.

In [3], Collobert et al., provide a comprehensive review on ways of applying unified neural network architectures and related deep learning algorithms to solve NLP problems from "scratch". The proposed system learn automatically internal representations or word embedding from vast amounts of mostly unlabeled training data while performing a wide range of NLP tasks. It was proposed to use a time-delay neural network which ran over words and a decoder layer calculating sentence-level likelihood using transition matrix at the top to find the best tag sequence at the sentence level.

In [10], a recurrent neural network architecture is successfully applied to sentiment analysis for semantic compositionality. It was applied with local context to build a deep architecture. Despite missing global context, the network was proven to be capable of successful merging of natural language words based on the learned semantic transformations of their original features.

In [11], wu et al. introduce a deep semantic embedding networks which is a supervised learning algorithm that computes semantic representation for text documents by respecting their similarity to a given query. Unlike other methods that use single layer learning machines, deep semantic embedding networks maps word inputs into a low dimensional semantic space with deep neural network, and achieves a

highly nonlinear embedding to model the human perception of text semantics. Through discriminative fine tuning of the deep neural network, the proposed network is able to encode the relative similarity between relevant/irrelevant document pairs in training data, and hence learn a reliable ranking score for a query-document pair [11].

More recently, Mikolov et al. [4] introduce word2vec, a novel word embedding procedure. Their model offers two architectures based on neural network language model [1,12], continuous bag-of-words (CBOW) and Skip-gram. This model has proved effective in several tasks in natural language processing. We will describe this neural network and its alternative models in the next section.

IV. WORD2VEC AND ITS ALTERNATIVE MODELS

Word2vec was designed to handle a large amount of texts and inferring a linear structure that models semantic and syntactic relations linking words. As mentioned above, this method offers two artificial neural networks architecture: the CBOW architecture and Skip-gram architecture. In CBOW model, a word is predicted from its surrounding words whereas for Skip-gram, multiple surrounding words are predicted from one input word. It built for each word a context window which corresponds to the *n* previous words and *n* next words for a central word. Within this window, all words are treated equally. To avoid computing a full softmax over the entire vocabulary, hierarchical softmax can be applied on a Huffman tree representation of the vocabulary, which saves calculations, at the potential loss of some accuracy. An additional strategy to get better embeddings is negative sampling, where, instead of only using the words observed next to one another in the training data as positive examples, random words are sampled from the corpus and presented to the network as negative examples. As result for this model, words whose meanings are related are generally closer in terms of euclidean distance than between unrelated words.

An alternative way of getting word embeddings, called GloVe (Global Vectors), is proposed in [13]. Rather than being based on language models it is based on global matrix factorisation. Only a word-word co-occurrence matrix is used. GloVe avoids the large computational cost, but training directly on the non-zero elements in it. As a cost function, the model uses a weighted least squares variant. The weighting function has two parameters, an exponent and a maximum cut-off value that influence the performance.

As well, Le and Mikolov [14] learn a dense representation for documents using a simplified neural language model, inspired by the word2vec. Their algorithm aims to calculate paragraph vectors, by adding an explicit paragraph feature to the input of the neural network. In [14], a convolutional neural network, built on top of word2vec word embeddings, is proposed for matching natural language sentences. This network can nicely combine the hierarchical sentence modeling through layer-by-layer composition and pooling, and the capturing of the rich matching patterns at different levels of abstraction.

Another work based on word2vec word embeddings is presented in [15]. In this work, a convolutional neural network is trained on top of word2vec word embeddings. The method is only evaluated on sentence classification tasks, not on semantic similarity.

In [16], Mikolov et al. present another contribution for learning representations of phrases. The idea is that many phrases have a meaning that is not a simple composition of the meanings of its individual words. To learn vector representation for phrases, they find words that appear frequently together, and infrequently in other contexts. Word2phrase was based on the skip-gram word2vec model.

V. APPLICATION TO CONCEPT EXTRACTION

In this section, we aim to automatically identify semantic concept related to an application; a touristic basedata. To build up the appropriate concepts, the corpus words has to be gathered in several classes, each class will represent one semantic concept. To do this, we used the unsupervised classification method K-means. In fact, the K-means will decide which words to group in the same class according to their representation. Thus, semantically significant words representations constitutes a key step in the building up of the list of concepts. As word2vec has proven its efficiency in various NLP tasks, we employed this neural network for determining the words representation. The used corpus contains 821 words in English. We considered 8 concepts. To evaluate results, we calculate the F-measure well known measure which is designed as follows:

$$F = \frac{2 \times R_{ij} \times P_{ij}}{R_{ij} + P_{ij}}$$

where R_{ij} and P_{ij} refer to the Recall and Precision measures. These measures are defined as:

$$R_{ij} = \frac{n_{ij}}{N_i}$$

$$P_{ij} = \frac{n_{ij}}{N_j}$$

Where n_{ij} is the number of words present in both concepts C_i and C_j. N_i and N_j represent the total number of words of t-he concept C_i and C_j.

We achieved an F-measure rate equals to *92.35%*. This result is so satisfactory. It is noted that the K-means based word2vec representations finds a coherent list of concepts which are related to the considered application. The considered reference concepts were prepared manually.

TABLE I. SOME OBTAINED CONCEPTS EXAMPLE

Concept	Group of words
Concept 1	Favourites, preferred, chosen, appreciated, adored, liked, loved
Concept2	Tunis, Djerba, Sousse, Paris, Hammemet, Londres
Concept3	Dining, mall, shopping, shops, buy, pay, money, visit, discover
Concept4	Discover, travels, flights

VI. Conclusion

Learning good semantic vector representations is a challenging and on going area of research in natural language processing. Recently, deep learning models have been proven to be one of the most powerful methods in tackling this problem. In this paper, we present deep learning, word embeddings and semantic similarity concepts and review some relevant deep neural networks for natural language processing. Word2vec is one of neural networks models which has proven its efficiency in a variety of natural language processing and be the most used as state-of-the art. In order to extract concepts of our application, we have firstly employed this model to construct words representations. Then, we used the K-means to classify to words according to their representation. We give good results that demonstrate the efficiency of using K-means based word2vec representation for extracting concepts.

References

[1] R. Collobert and J. Weston, " A unified architecture for natural language processing: Deep neural networks with multitask learning," In Proceedings of the 25th international conference on Machine learning, pp. 160–167, ACM, 2008.

[2] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semisupervised learning," In ACL, 2010.

[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," The Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," In Advances in Neural Information Processing Systems, pp. 3111–3119, 2013.

[5] Y. Bengio, A. Courville, et tHE IDEAal. , "Representation learning: A review and new perspectives," 2013.

[6] G. E. Hinton, S. Osindero, et al., "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, pp. 1527–1554, 2006.

[7] Y. Bengio, P. Lamblin, et al., "Greedy layerwise training of deep networks," Advances in neural information processing systems, vol. 19, 2007.

[8] Y. Bengio, "Learning Deep Architectures for AI," Found. Trends Mach. Learn. , vol. 2, pp. 1–127, 2009.

[9] H. Lee, R. Grosse, R. R., and Ng, A, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," In Proceedings of the 26th International Conference on Machine Learning, 2009.

[10] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts, " Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013.

[11] H. Wu, M. R. Min, B. Bai, "Deep semantic embedding," SMIR@ SIGIR, pp. 46–52, 2014.

[12] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. , "Neural probabilistic language models," In Innovations in Machine Learning. Springer, 2006.

[13] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation," 2014.

[14] Q. V. Le and T. Mikolov, " Distributed representations of sentences and documents," In Proceedings of the 31st International Conference on Machine Learning, 2014.

[15] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," In NIPS 2014, 2014.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," In: Advances in Neural Information Processing Systems 26 (NIPS 2013), 2014.

A Survey of Machine Learning Techniques Applied in Big Data

Ahmed Masmoudi, Lobna Bouchaala and Mounir Ben Ayed
REGIM: Research Group on Intelligent Machines
University of Sfax, National School of
Engineers (ENIS) Sfax, Tunisia
E-mail: {ahmed1.masmoudi@gmail.com; lobna.bouchaala@yahoo.fr; mounir.benayed@ieee.org}

***Abstract*— Big Data has become a tempting solution for companies that aim at building competitiveness through the use of their customer data, suppliers, products, processes, machines, etc. Big Data can create efficient challenging solutions in health, security, government and more; and usher in a new era of analytics and decisions. Knowledge discovery in Big Data comprises a set of methods and techniques used to identify, create, represent, predict, and enable creating experience that can constitute a real immaterial capital. However, to explore significant meaning to the perpetual tsunami of data, Big Data needs Data Mining. In turn, to broaden the scope of its targeted analyses, Data Mining requires Big Data. Thus, there is a complementary relation between these two major concepts. This paper presents a state of art where we try to make a comparison of Machine Learning techniques applied in Big Data and the best performing technique.**

Keywords—Big Data; data mining; Machine Learning; Bayesian network;

I. INTRODUCTION

Processing huge amounts of data stored using computers to extract relevant information has become the stake of the twenty first century's data processing worldwide. It is necessary to have a model making a link between the different observations and reality, even when observations are incomplete and inaccurate. We seek to find relevant relationships between variables or groups of variables.

Big Data relate data in large volumes, complex, from different sources and of different types. With the rapid development of networks, data storage, and data collection capacity, Big Data is now expanding in numerous scientific and technical areas, including physics, biology and biomedicine. Large-scale data sets are collected and studied in numerous domains, from engineering sciences to social networks, commerce, biomolecular research, and security [1]. The issue of Big Data research is typically highlighted with 3 V that denote: volume, variety and velocity. However, other characteristics should not be neglected as the versatility, accuracy, and recovery of data and information. So, we can take advantage of Big Data by knowledge discovery to understand or predict, hence the role of Machine Learning.

The remaining sections of the paper are organized as follows. We discuss in section II the basic concepts of Big Data, knowledge discovery in Big Data and Machine Learning, and we show the specify of Machine Learning in Big Data. In section III, we make a comparative study between the main techniques of Machine Learning according to the knowledge discovery in Big Data. In section IV, we discuss the advantages of Bayesian network in the context of Big Data. The last section presents conclusions drawn from this state of art.

II. BACKROUND AND DEFINITIONS

In this section, we first broach the Big Data environment and its main characteristics and then the background of Knowledge discovery in Big Data.

A. Era of Big Data

Actually, Big Data does not have a unique definition; different works have defined it in different ways. Gartner [2] defines Big Data as high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization. As for Jacobs, he specifies in [3] that Big Data is data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time. To simplify the meaning of Big Data, the concept is defined within the concept of the 3V model, which is presented by the IBM researchers [4] as illustrated in Figure 1.

- Volume (size): Actually, it is very normal to have Terabytes and Petabytes of storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be revaluated always. According to [5], in 2011, digital information has grown nine times in volume in just 5 years and its amount in the world will reach 35 trillion gigabytes by 2020 [6].

- Velocity (streaming data): Social media explosion has changed our vision to data. People reply more and more on social media to update them with the latest happening. The data movement is now real time and the update window has minimized to fractions of the seconds.

- Variety (structured and unstructured data): The real world has data in many different formats and this is one among the important challenges we need to face with Big Data. Data can be stored in multiple structured formats like excel, database, access. It can also be stored in unstructured formats

like videos, SMS, pdf or something we can have not thought of.

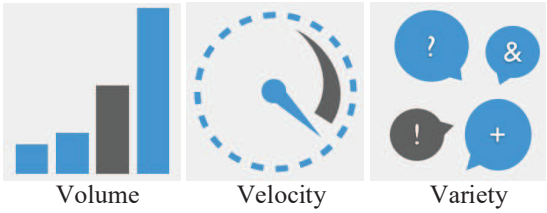


Figure 1. The 3V model defining Big Data

B. *Knowledge discovery and data mining in Big Data*

The knowledge extraction and data mining is to make sense of large amounts of data, from a certain area, massively captured and stored by businesses today. Indeed, the real value is not in the acquisition and storage of data, but rather in our ability to extract useful reports and find trends and interesting correlations to support decisions made by the makers of companies and scientists. This extraction uses a range of techniques, methods, algorithms and tools origins statistics, artificial intelligence, databases, etc. However, before attempting to extract useful knowledge from data, it is important to have a clear procedure and understand the overall approach. In fact, just knowing data analysis algorithms and apply them on hand in the data is not sufficient for the conduct of a data mining project. Certainly, a blind application of data mining methods on data in hand can lead to the discovery of knowledge incomprehensible or even useless for the end user [7].

It is mainly for this reason that the activity of knowledge extraction and data mining was quickly organized in the form of a process called Knowledge Discovery in Databases (KDD). This process is presented as a complex, non-trivial, consisting of several iterative steps, and requires a permanent interactivity by the expert user. The process provides a roadmap to be followed by practitioners in the planning and implementation of knowledge extraction projects from data.

C. *Machine Learning*

The Machine Learning (ML) is a set of statistical tools or geometric and computer algorithms that automate the construction of a prediction function f from a set of observations called the training set [8]. It is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics [9–12]. Because of its implementation in a wide range of applications, Machine Learning has covered almost every scientific domain, which has brought great impact on science and society [13]. It has been used on a variety of problems, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems [14].

ML is therefore a fundamental data-driven approach. It is deemed helpful to build decision support systems that adapt to the data and for which no treatment algorithm is listed. We find the following terms to describe ML:

- ML can create systems that learn their own rules.
- ML retrieves a program from the data.
- ML model learns by itself associations and similarities from different data sets.

Generally, the field of Machine Learning is divided into three subdomains: supervised learning, unsupervised learning, and reinforcement learning [15].

Briefly, supervised learning requires training with labelled data which has inputs and desired outputs.

In contrast with the supervised learning, unsupervised learning does not require labelled training data and the environment only provides inputs without desired targets. Reinforcement learning enables learning from feedback received through interactions with an external environment.

D. *Specificity of Machine Learning for Big Data*

Machine Learning for Big Data requires specific methods because it must take into account that the 3 V characterizing the term "Big Data" leaves open the crucial question of what the big challenge in the "Big Data" is.

The first interpretation suggested by the "V" = volume of 3V: what is big is the volume of data. So a large volume of data requires techniques that have the power to be able to overcome this difficulty.

The second interpretation suggested by the "V" = variety: what is great is the number of parameters that characterize each observation in some contexts. That Big Data number of parameters could number in thousands. So, we need to find the solution to address the heterogeneity of data.

The third interpretation, suggested by the "V" = velocity: what is great is the frequency at which data are generated, captured and shared. For this criteria, we must find the right technical solution that can handle high-speed data.

So, we have 3 essential critical issues of Machine Learning techniques for Big Data from three different perspectives, as described in Figure. 2, including learning for large scale of data, learning for different types of data, learning for high speed of streaming data [16].

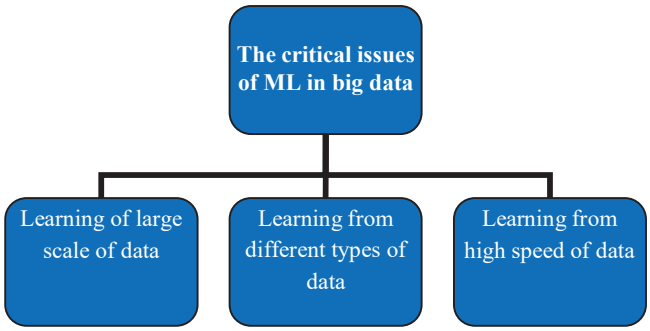


Figure 2. The critical issues of Machine Learning for Big Data

III. COMPARATIVE STUDY

In this section, we compare the main ML techniques that can be applied to Big Data from the perspective of knowledge extraction.

A. Comparison criteria

we analyse and compare the techniques according to 6 criteria:

- **Discrete variables**
- **Continuous variables**
- **Large classes**
- **Large number of variables**
- **Missing data**
- **Time**
- **Mixed data**

B. Main Machine Learning techniques

1) Decision tree

Decision trees [17] (also known as classification trees) are probably one of the most intuitive and frequently used data mining techniques. From an analyst's point of view, they are easy to set up and from a business user's point of view they are easy to interpret. Classification trees, as the name implies, are used to separate a data set into classes belonging to the response variable. Usually, the response variable has two classes: Yes, or No (1 or 0). If the response variable has more than two categories, then variants of the developed decision tree algorithm may be applied. In either case, classification trees are used when the response or target variable is categorical in nature.

2) Bayesian networks

Judea Pearl has pioneered Bayesian Networks, from its work on the 80 probabilistic reasoning in intelligent systems [18]. Bayesian networks are models that come from the marriage between graph theory and probability theory. They use the basic rule Bayes theorem, hence their name. Indeed, a Bayesian network is described by a graph and a set of parameters. The graph is shown as nodes connected to each other by directed arcs. The parameters are all the probabilities of each graph node to other nodes conditionally.

3) Support Vector Machine

Support Vector Machine (SVM) [19] formulate the classification problem as a quadratic optimization problem related to maximizing the maximum margin.
This choice is justified by the statistical learning theory, which shows that the border separating maximum margin has the smallest generalization error.
The margin is the distance between the border of separation and the nearest samples (support vectors).
In the non-linearly separable case, the key idea is to transform the representation space of input data in a larger description space (possibly infinite), wherein it is likely that there is a linear separator through a core function, the kernel functions for transforming a scalar product in a large space in a single point assessment of a function (kernel trick).

4) Neural network

A neural network [20] is a computation model whose schematic operation is based on the operation of biological neurons. Each neuron is a weighted sum of its inputs (or synapses) and returns a value depending on its activation function. This value can be used either as one input of a new layer of neurons, whether as a result it is up to the user to interpret (class, calculation results, etc.).
The learning phase of a neural network adjusts the weight associated with each synapse input (also referred to as synaptic coefficient).
It is a long process that must be repeated for each structural change to the database being processed.

5) Deep Learning

Deep learning [25] (also known as deep structured learning, hierarchical learning or deep machine learning) is a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations.
Deep learning is part of a broader family of machine learning methods based on learning representations of data.

a. Comparison assessment

We present in this section a comparative assessment based on the previously defined criteria. Table I show a comparison between various Machine Learning techniques from the point of view of extraction of knowledge that can be applied on Big Data. The adopted representation is as follows:

• Each line corresponds to a characteristic, which can be an advantage, or the inclusion of a specific problem.

• If the technology allows to take into account this problem, or has this advantage, a + sign is placed in the box.

• A ++ sign is placed in the box for the best technique from the considered characteristic point of view.

TABLE I. A COMPARAISON OF MACHINE LEARNING TECHNIQUES APPLIED IN BIG DATA

Knowledge discovery	Decision tree	Bayesian networks	SVM	Neural network	Deep Learning
Discrete variables	++	++	+	++	++
Continuous variables	+	++	+	++	++
Large classes	+	+	-	+	+
Large number of variables	+	+	+	+	+
Missing data	+	++	+	+	+
Time	+	+	+	+	+
Mixed data	-	++	-	+	+

• A - sign is placed in the box if the technology does not take into account this criterion.

From Table 1 [26] [27] [28], we notice that the Bayesian networks have more advantages over other techniques but the comparison between the Bayesian networks, neural network and deep learning was very difficult because these 3 techniques are very near in the point of view of the answer to the criteria of Big Data. Indeed, respecting the crucial questions of big data, Bayesian networks meet more than other machine learning techniques. Compared with neural network and deep learning, Bayesian networks treated more mixed data (criterion variety of Big Data), both Bayesian networks can find solutions for the missing data because certainly in the big data there is a large enough volume so we will certainly find the missing data. However, certain characteristics associated with Big Data pose challenges for modifying and adapting Bayesian Network to address those issues.

1) ADVANTAGES OF BAYESIAN NETWORKS

Several significant advantages of Bayesian networks can be argued [20,21,22]. First, they provide the ability to collect and merge the knowledge of various kinds in the same model, they are flexible in regards to missing information. Second, Bayesian networks allow investigators to use their domain expert knowledge in the discovery process for Big Data, while other techniques rely primarily on coded data to extract knowledge. Also, Bayesian network models can be more easily understood than many of the other techniques via the use of nodes and arrows. Third, researchers in the domain of Big Data can easily encode domain expert knowledge through the use of these graphical diagrams, and thus more easily understand and interpret the output of the Bayesian network. Bayesian networks are also superior in capturing interactions among input variables. In some situations, decision trees may appear to produce more accurate classifications because they consider only relationships between output and input variables. However, ability to capture the relationships among input variables has tremendous value in exploring data. Bayesian network models can produce relatively accurate prediction even in the situation where complete data are not available. Last, because Bayesian networks can incorporate domain knowledge into statistical data, Bayesian networks are less influenced by small sample size [23]. They can be also more easily understood than many of the other techniques via the use of nodes and arrows. These represent the variables of interest and the relationships of variables, respectively.

IV. CONCLUSIONS

Big Data are now rapidly expanding in different science and engineering domains. Learning from these massive data is expected to bring significant opportunities and transformative potential for various sectors. However, most traditional Machine Learning techniques are not inherently efficient or scalable enough to handle the data with the characteristics of large volume, different types, high speed, uncertainty and incompleteness, and low value density. In response, Machine Learning needs to reinvent itself for Big Data processing. This paper began with a brief review of some definitions of Big Data, knowledge discovery, data mining and Machine Learning, followed by a comparative study of some Machine Learning techniques. Then, a discussion about the best technique that can be applied with Big Data was made. Anyway, an absolutely appropriate technique does not exist, so we need to update the old techniques in order to join the era of Big Data or invent a new technique.

REFERENCES

[1] A. Sandryhaila and JMF. Moura, "Big Data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure". IEEE Signal Proc Mag 31(5), 2014, pp 80–90.

[2] M. A. Beyer and D. Laney, "The Importance of 'Big Data': a Definition," Stamford, CT: Gartner, 2012.

[3] A. Jacobs, "The Pathologies of Big Data," Communications of the ACM, vol. 52,2009, pp. 36-44.

[4] P. Russom, "Big Data Analytics", TDWI Best Practices Report, Fourth Quarter, 20011.

[5] J. Gantz and D. Reinsel, "Extracting value from chaos" MC, Hopkinton, 2011.

[6] J. Gantz, D. Reinsel, "The digital universe decade are you ready". EMC, Hopkinton, 2010.

[7] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "Advances in Knowledge Discovery and Data Mining". MIT Press, 1996.

[8] P. Lemberger, M. Batty, M. Morel and J-L. Raffaëlli, "Big Data et Machine Learning: Manuel du data scientist". Management des systèmes d'information, Éditeur Dunod,2015, pp. 110-111.

[9] TM. Mitchell, Machine Learning (McGraw-Hill, New York, 1997)

[10] S. Russell and P. Norvig, "Artificial intelligence: a modern approach". Prentice-Hall,Englewood Cliffs, 1995.

[11] V. Cherkassky and FM. Mulier, "Learning from data: concepts, theory, and methods". John Wiley & Sons, New Jersey, 2007.

[12] TM. Mitchell, "The discipline of Machine Learning". Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

[13] C. Rudin and KL. Wagstaff, "Machine Learning for science and society". Mach Learn 95-1, 2014, pp. 1–9.

[14] CM. Bishop, "Pattern recognition and Machine Learning". Springer, New York,2006.

[15] B. Adam, IFC. Smith and F. Asce, "Reinforcement learning for structural control". J Comput Civil Eng 22-2, 2008, pp. 133–139.

[16] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of Machine Learning for Big Data processing". EURASIP Journal on Advances in Signal Processing,2016, pp. 1-16.

[17] J. R. Quinlan, "Simplifying decision trees". International Journal of Man-Machine Studies 27-3, 1987, pp. 221.

[18] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers", San Mateo, 1988.

[19] J. Zheng, F. Shen, H. Fan, J. Zhao, "An online incremental learning Support Vector Machine for large-scale data". Neural Comput Appl 22-5,2013 pp. 1023–1035.

[20] X. Dong, Y. Li, C. Wu, Y. Cai, "A learner based on neural network for cognitive radio". In Proceedings of the 12th IEEE International Conference on Communication Technology (ICCT), Nanjing 2010, pp. 893–896.

[21] D. Ckerman, "Bayesian networks for data mining". Data Min Knowl Disc 1997;1: pp 79–119.

[22] RG. Cowell, AP. Dawid, SL. Lauritzen, DJ. Spiegelhalter, "Probabilistic networks and expert systems". New York: Springer; 1999.

[23] DE. Heckerman, "Learning Bayesian networks: The combination of knowledge and statistical data". MSR-TR-94-09. 1995. Redmond, WA, Microsoft Research.

[24] EL. Eisenstein and F. Alemi, "A comparison of three techniques for rapid model development: an application in patient risk-stratification". Proc/AMIA Annu Fall Symp 1996:443–7.

[25] L. Deng and D. Yu,"Deep Learning: Methods and Applications", Foundations and Trends in Signal Processing. 7 (3-4):2014 pp. 1–199.

[26] M. M. Najafabadi, F. Villanustre, T. M Khoshgoftaar, Seliya, N. R. Wald, & E. Muharemagic, "Deep learning applications and challenges in big data analytics", Journal of Big Data 2(1), 1 , 2015.

[27] P. Naïm, P. H. Wuillemin, P. Leray, O. Pourret, & A. Becker. « Réseaux bayésiens", Paris: Eyrolles, 2008.

[28] I. H. Witten, & E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005.

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Singular Gaussian Graphical Models: Structure Learning

Khalil MASMOUDI[1,*], Afif MASMOUDI[1]

Keywords: Gaussian Graphical Models, Structure Learning, Singular covaraince matrix, Conditional independence

ABSTRACT

The goal of this paper is to introduce Singular Gaussian graphical models and their conditional independence properties. Assuming a GMRF (Gaussian Markov Random Field) with a singular or ill-conditioned covariance matrix, we construct the graph structure from the covariance matrix's pseudo-inverse. The proposed approach ensures graph stability when the covariance matrix is ill-conditioned through projecting data on a smaller subspace.

[1] Laboratory of Probability and statistics, Faculty of Science of Sfax, University of Sfax.

Limit of risks ratios of shrinkage estimators for the multivariate normal mean with unknown variance

Abdenour Hamdaoui
Department of Mathematics, Oran (U.S.T.O) University, 31000, Laboratory of Statistics and Random Modelings of Tlemcen University, Algeria
Email: abdenour.hamdaoui@yahoo.fr

Djamel Benmansour
Department of Environnement and Ecology, Laboratory of Statistics and Random Modelings of Tlemcen University, 13000, Algeria
Email: dbenmansour@yahoo.fr

***Abstract*—We study the estimation of the mean θ of a multivariate normal distribution $X \sim N_p(\theta, \sigma^2 I_p)$ *in* $\Re^p$, σ^2 is unknown and estimated by the chi-square variable $S^2 \sim \sigma^2 \chi_n^2$. In this work we are interested in studying bounds and limits of risk ratios of shrinkage estimators to the maximum likelihood estimator, when n and p tend to infinity provided that $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c$. We recall that the risk ratios of shrinkage estimators to the maximum likelihood estimator, has a lower bound $B_m = \frac{c}{1+c}$, when n and p tend to infinity provided that $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c$. We show that if the shrinkage function $\psi(S^2, \|X\|^2)$ satisfies some conditions, the risk ratios of shrinkage estimators $\delta = (1 - \psi(S^2, \|X\|^2)\frac{S^2}{\|X\|^2})X$, which did not inevitably minimax, to attain the limiting lower bound B_m. When the dimension is moderate, we give sufficient conditions for a shrinkage estimator to dominate the maximum likelihood estimator, establishing a minimaxity result. We deduce that the James-Stein estimator is minimax, and for any estimator dominate it, his risk ratio attain this lower bound B_m (in particularly its positive-part version). We graph the corresponding risk ratios for estimators of James-Stein δ^{JS}, its positive-part δ^{JS+} and estimators defined in selected examples, for diverse values of n and p.*

Keywords*—James-Stein estimator ; multivariate gaussian random variable; non-central chi-square distribution; shrinkage estimator; quadratic risk.

I. Introduction

The estimation by shrinkage estimators, of the mean θ of a multivariate normal distribution $N_p(\theta, \sigma^2 I_p)$ in $\Re^p$, has experienced many development since the papers of C. Stein [11], W. James [8] and C. Stein [12]. In these works one estimates the mean θ by shrinkage estimators deduced from the empirical mean estimator, which are better in quadratic loss than the empirical mean estimator.

More precisely, if X represents an observation or a sample of multivariate normal distribution $N_p(\theta, \sigma^2 I_p)$, the aim is to estimate θ by an estimator δ relatively at the quadratic loss function :

$$L(\delta, \theta) = \|\delta - \theta\|_p^2$$

where $\|.\|_p$ is the usual norm in $\Re^p$. We associate his risk function :

$$R(\delta, \theta) = E_\theta(L(\delta, \theta)).$$

W. James, and C. Stein [8], introduced a class of estimators improving $\delta_0 = X$, when the dimension of the space of the observations $p \geq 3$, denoted by :

$$\delta_j^{JS} = \left(1 - \frac{(p-2)}{(n+2)} \frac{S^2}{\|X\|^2}\right) X \ , \ j = 1, ..., p . \qquad (1.1)$$

where $S^2 \sim \sigma^2 \chi_n^2$ is the estimate of σ^2.

A.J. Baranchik [1] proposed the positive-part of James-Stein estimator dominating the James-Stein estimator when $p \geq 3$,

$$\delta_j^{JS+} = \max\left(0, \left(1 - \frac{(p-2)}{(n+2)} \frac{S^2}{\|X\|^2}\right)\right) X , \qquad (1.2)$$

G. Casella and J.T. Hwang [5] studied the case where σ^2 is known $(\sigma^2 = 1)$ and showed that if the limit of the ratio $\frac{\|\theta\|^2}{p}$, when p tends to infinity is a constant $c > 0$, then

$$\lim_{p \to +\infty} \frac{R(\delta_{JS}, \theta)}{R(X, \theta)} = \lim_{p \to +\infty} \frac{R(\delta_{JS}^{+}, \theta)}{R(X, \theta)} = \frac{c}{1+c}.$$

L. Sun [9] has considered the following model : $(y_{ij}/\theta_j, \sigma^2) \sim N(\theta_j, \sigma^2)$ $i = 1,...,n$, $j = 1,...,m$ where $E(y_{ij}) = \theta_j$ for the group j and $var(y_{ij}) = \sigma^2$ is unknown.

The James-Stein estimators are written in this case :

$$\delta^{JS} = (\delta_1^{JS},...,\delta_m^{JS}),$$

where

$$\delta_j^{JS} = \left(1 - \frac{(m-3)}{(N+2)}\frac{S^2}{T^2}\right)(\bar{y}_j - \bar{y}) + \bar{y}, \ j = 1,...,m,$$

and $S^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_j)^2$, $T^2 = n\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2$,

$\bar{y}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}$ and $\bar{y} = \frac{\sum_{j=1}^{m} \bar{y}_j}{m}$, $N = (n-1)m$.

He shows that for any estimator of the form

$$\delta^{JS} = (\delta_1^{JS},...,\delta_m^{JS})$$

where

$$\delta_j^{\psi} = (1 - \psi(S^2, T^2))(\bar{y}_j - \bar{y}) + \bar{y}, \ j = 1,...,m,$$

if $\lim_{m \to +\infty} \frac{\sum_{j=1}^{m}(\theta_j - \bar{\theta})^2}{m} = q$ exists,

then $\lim_{m \to +\infty} \frac{R(\delta^{\psi}, \theta)}{R(X, \theta)} \geq \frac{q}{q + \frac{\sigma^2}{n}}$

and $\lim_{m \to +\infty} \frac{R(\delta^{JS}, \theta)}{R(X, \theta)} = \lim_{m \to +\infty} \frac{R(\delta^{JS+}, \theta)}{R(X, \theta)} = \frac{q}{q + \frac{\sigma^2}{n}}$.

Namely $\frac{q}{q + \frac{\sigma^2}{n}}$ constitutes a lower bound for the ratio $\lim_{m \to +\infty} \frac{R(\delta^{\psi}, \theta)}{R(\delta_0, \theta)}$ and is equal to $\lim_{m \to +\infty} \frac{R(\delta^{JS}, \theta)}{R(\delta_0, \theta)}$.

L. Sun [9] also shows that this bound is attained for a class of estimators defined by :

$$\delta_j = (1 - \psi(S^2, T^2)\frac{S^2}{T^2})(\bar{y}_j - \bar{y}) + \bar{y}, \ j = 1,...,m$$

where ψ satisfies certain conditions. This bound is also attained for any estimator dominating the James-Stein estimator, in particular the positive-part version of the James-Stein estimator.

Finally, we note that if n tends to infinity then the ratio $\frac{q}{q + \frac{\sigma^2}{n}}$ tends to 1, and thus the risk of the James-Stein estimator is that of δ_0 (When n and m tend to infinity).

A. Hamdaoui and D. Benmansour [7], considered the following classe of shrinkage estimators $\delta_{\psi} = \delta_{JS} + l\psi(S^2, \|X^2\|)X$, which is introduced in D. Benmansour, T. Mourid [3]. The authors showed that if $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p} = c(> 0)$ then the risk ratios $\frac{R(\delta_{\psi}, \theta)}{R(X, \theta)}, \frac{R(\delta^{JS}, \theta)}{R(X, \theta)} and \frac{R(\delta^{JS+}, \theta)}{R(X, \theta)}$ attain the lower bound $B_m = \frac{c}{1+c}$ when n and p tend to infinity provided that $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p} = c.$

When the dimension p is moderate, A.C. Brandwein and W.E. Strawderman [4] considered the following model $(X, U) \sim f(\|X - \theta\|^2 + \|U\|^2)$, where $\dim X = \dim \theta = p$ and $\dim U = k$. The classical example of this model is, of course, the normal model of density $\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{p+k} e^{-\frac{\|X-\theta\|^2}{2\sigma^2}}$. They showed that the estimator $\delta = X + \left\{\frac{\|U\|^2}{k+2}\right\}g(X)$ dominate X, so that δ is minimax, provided the function g satisfies certain conditions.

Y. Maruyama [10] has also studied the minimaxity of shrinkage estimator when the dimension of parameter's space is moderate. Then he considered the following model : $z \sim N_d(\theta, I_d)$ and the so called $\bar{\pi}$ -norm given by :

$\|z\|_p = \left\{ \sum_{i=1}^{i=d} |z_i|^p \right\}^{\frac{1}{p}}$, $p > 0$. He stydied the minimaxity of shrinkage estimators defined as follows : $\hat{\theta}_\phi = (\hat{\theta}_{1\phi},...,\hat{\theta}_{d\phi})$ with : $\hat{\theta}_{i\phi} = \left(1 - \phi(\|z\|_p) \Big/ \phi\left(\|z\|_p^{2-\alpha} |z_i|^\alpha\right)\right) z_i$ where $0 \le \alpha \le (d-2)/(d-1)$, $p > 0$.

Note that the risk functions of these estimators are calculated relatively to the usual quadratic loss function defined at above.

In this work we adopt the model $X \sim N_p(\theta, \sigma^2 I_p)$ and independently of the observations X, we observe $S^2 \sim \sigma^2 \chi_n^2$ an estimator of σ^2. Note that $R(X,\theta) = p\sigma^2$, is the risk of the maximum likelihood estimator. We generalize the results given in the papers of D. Benmansour and A. Hamdaoui [2], and A. Hamdaoui and D. Benmansour [7]. Then we give different conditions for a general class of shrinkage estimator as given in the paper of A. Hamdaoui and D. Benmansour [7] so that, if $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c,$ then the risk ratio attain the lower bound $B_m = \frac{c}{1+c}$ when n and p tend to infinity.

When the dimension is moderate, we give sufficient conditions for a shrinkage estimator to dominate the maximum likelihood estimator, establishing a minimaxity result. Then we deduce another proof which shows that the James-Stein estimator is minimax, and for any estimator dominate it, its risk ratio attain this lower bound B_m (in particularly its positive-part version).

In the following we denote the general form of a shrinkage estimator as follows :

$$\delta_\phi = (1 - \phi(S^2, \|X\|^2))X \quad , \qquad (1.3)$$

In Section 1, we recall somme results obtained in the paper of A. Hamdaoui and D. Benmansour [7]. The authors showed, that under condition $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c,$ the risk ratio of shrinkage estimators δ_ϕ given in (1.3) , to the maximum likelihood estimator X, has a lower bound B_m, when n and p tend to infinity. The second result indicates that under the same condition $\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c,$ the risk ratio of James-Stein estimator δ^{JS}, to the maximum likelihood estimator X, tends to the value $\frac{c}{1+c}$ when n and p tend simultaneously to infinity.

In Section 2 we give the main results of this paper. In the first we considered the general classe of shrinkage estimators $\delta_\psi = (1 - \psi(S^2, \|X^2\|)\frac{S^2}{\|X\|^2})X$, then we show that if the function ψ satisfies some conditions which is different from the one given in A. Hamdaoui and D. Benmansour [7], the risk ratios of shrinkage estimators δ_ψ, to attain the limiting lower bound B_m, provided that

$$\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c, .$$

In second part of this section, we studie the minimaxity of shrinkage estimators given in (3.1), when p is moderate.We deduce another proof of the minimaxity of James-Stein estimator δ^{JS}, and we note that any shrinkage estimator dominate it (in particulary its positive-part δ^{JS+}), has a risk ratio attain the limiting lower bound B_m, provided

$$\lim_{p \to +\infty} \frac{\|\theta\|^2}{p\sigma^2} = c \, .$$

Finally, we graph the corresponding risks ratios for estimators of James-Stein δ^{JS}, its positive-part δ^{JS+}, and estimators defined in selected examples for diverse values of n and p .

References

[1] A.J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution", *Stanford Univ.* Technical Report (51), 1964.

[2] D. Benmansour and A. Hamdaoui, "Limit of the Ratio of Risks of James-Stein Estimators with Unknown Variance". Far East Journal of Theorical Statistics, 36(1), 2011, 31-53.

[3] D. Benmansour, T. Mourid, "Etude d'une classe d'estimateurs avec rétrécisseur de la moyenne d'une loi gaussienne", Annales de l.ISUP, Fascicule 51, 2007, 1-2.

[4] A.C. Brandwein, and W.E. Strawderman, "Stein Estimation for Spherically Symmetric Distributions : Recent Developments", Statistical Science, 27(1), 2012, 11-23.

[5] G. Casella and J.T. Hwang, "Limit expressions for the risk of James-Stein estimators", The canadian Journal of Statistics, 10(4), 1982, 305-309.

[6] D. Fourdrinier, I. Ouassou and W. E. Strawderman, "Estimation of a mean vector under quartic loss". Journal of Statistical Planning and Inference, 138, 2008, 3841-3857 .

[7] A. Hamdaoui and D. Benmansour, "Asymptotic properties of risks ratios of shrinkage estimators", Hacettepe Journal of Mathematics and Statistics, 44(5), 2015, 1181-1195.

[8] W. James, and C. Stein, "Estimation of quadratique loss", Proc 4th Berkeley Symp,Math. Statist.Prob., Univ of california Press, Berkeley. 1, 1961, 361-379.

[9] L. Sun, "Risk Ratio and Minimaxity in Estimating the Multivariate Normal Mean with unknown Variance", Scandinavian Journal of Statistics, 22(1), 1995, 105-120.

[10] Y. Maruyama, " $\bar{\pi}$-norm based James-Stein estimation with minimaxity and sparsity", Statistics Theory (math.ST) arXiv, 1402-0302, 28 May 2015.

[11] C. Stein, "Inadmissibilty of the usual estimator for the mean of a multivariate normal distribution", Proc 3th Berkeley Symp,Math.Statist.Prob. Univ of california Press,Berkeley, 1, 1956, 197-206.

[12] C. Stein, "Estimation of the mean of a multivariate normal distribution". Ann. Statis. 9(6), 1981, 1135-1151.

[13] Strawderman, W. E (1973) "Proper bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances". Ann of Statis. Vol. 1,No.6, 1189-1194 .

.

Multiple Supervised classifiers Fusion For Arabic Polarity Detection

Amel Ziani
Computer science department, Lri laboratory: Information Research Laboratory
Badji Mokhtar University
Annaba, Algeria
z_amel1911@live.fr

Cheick Abdoul Kadir A Kounta
Computer science department, Labged laboratory: Electronic documents control laboratory
Badji Mokhtar University
Annaba, Algeria
abdaty11@hotmail.fr

Nabiha Azizi
Computer science department, Labged laboratory: Electronic documents control laboratory
Badji Mokhtar University
Annaba, Algeria
azizi@labged.net

Djamel Zenakhra
Computer science department, Labged laboratory: Electronic documents control laboratory
Badji Mokhtar University
Annaba, Algeria
dzenakhra@gmail.com

Soraya Cheriguene
Computer science department, Labged laboratory: Electronic documents control laboratory
Badji Mokhtar University
Annaba, Algeria
cheriguene@labged.net

***Abstract*— Sentiment analysis deals with determining the sentiment polarity, subjective, objective, or neutral of a text. Recently, it has attracted great interest both in academia and in industry due to its useful potential applications. One of the most promising applications is analysis of Arabic newspapers reader's opinions. In this paper, we examine how classifiers work while doing opinion mining over Arabic reviews. Most research efforts in the area of subjectivity detection deal with English texts. Some new works deal with other languages, but in Arabic, there are a few works in this area. We explore how different classifiers nature can affect the precision of the general Arabic opinion mining system. We experiment the proposed approach with Naïve Byes, Multinomial Naïve Byes, Support Vector Machine and Neural Networks in order to combine them to generate the final decision. The main steps of this study are based primarily on corpus construction, Statistical features extraction and then subjectivity detection by the proposed hybrid approach. Experiments results based on 1000 comments collected from Arabic Algerian news web sites are very encouraging;**

Keywords—Opinion polarity detection; Classifiers fusion; Naïve Byes; Multinomial Naïve Byes; SVM(Support Vector Machine); Neural Networks;

I. INTRODUCTION

Subjectivity detection seeks to identify whether the given text expresses opinions (subjective) or reports facts (objective). Such a task of distinguishing subjective information from objective is useful for many natural language processing applications. This new area of research is becoming more and more important mainly due to the growth of social media where users continually generate contents on the web in the form of comments, opinions, emotions, etc.

Arabic is a challenging language for a number of reasons. It has a very complex morphology as compare to English language. This is due to the unique nature of Arabic language. This work is primarily concerned with the task of extracting relevant characteristics and subjectivity detection in Arabic sports reviews. Unfortunately, not much work has been done on Arabic sentiment analysis and opinion mining. The authors in [1] applied sentiment analysis techniques to identify and classify document level opinions in text crawled from English and Arabic web forums. In [2] they proposed a method for identifying the polarity of nonEnglish words using multilingual semantic graphs. They applied their method to Arabic and Hindi. Also in [3] they annotated a corpus of Modern Standard Arabic (MSA) news text for subjectivity at the sentence level.

The approach is based on novel paradigm of hybrid supervised machine learning techniques. An important extracted statistical characteristics set (such as the number of subjective words, emotionality, reflexivity… etc.) are defined and analyzed in section (2.2). We noticed that no one can predict the rate of a classifier in any field specially the Arabic opinion classification, that's why we had to analyze and test the most performed supervised techniques in the data mining, which are: Naïve byes, Multinomial Naïve Byes, Support Vector Machine and Neural Networks.

But the complex morphology of the Arabic language and its difficulties in the polarity detection task require many supervised techniques to be fused. Combination of multiple classifier decisions is a powerful method for increasing classification rates in difficult Arabic NLP problems. To achieve better classification rates, it has been found that in many applications, it is better to fuse multiple relatively simple classifiers than to build a single sophisticated classifier.

The paper presents the research we conducted toward implementing an Arabic polarity detection system based on a hybrid approach fusing multiple supervised techniques. It starts from a brief state of the system phases. Then, we describe the used dataset and the features extraction phase. The results of our experiments are then presented and discussed. Finally, we conclude with a summary of the most important results provided by this study along with some possible extension of work.

II. THE PROPOSED SYSTEM ARCHITECTURE

The following architecture explains our proposed system; we start first by explaining the needed pre-processing phase to the input data. Then provide a short over view of the selected features and later we will explain the classification phase using the four supervised techniques: Naïve Byes, Multinomial Naïve Byes, Support Vector Machine and Neural Networks.

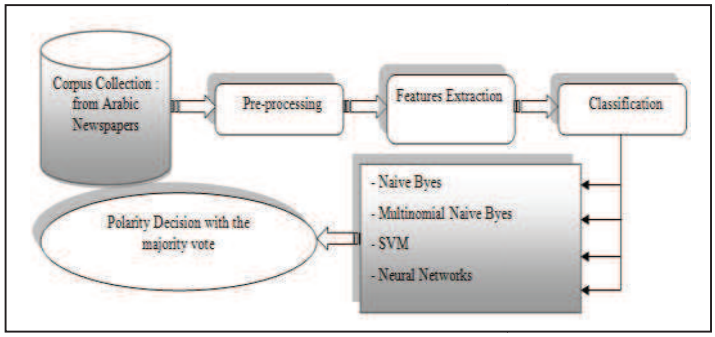


Fig. 1. The proposed process for subjectivity detection.

1 The corpus collection and the pre-processing

Although Arabic is considered one of the top 10 languages mostly used on the Internet based on the ranking carried out by the Internet World State [4] rank in 2010 and it is spoken by hundreds of millions of people, there exist limited annotated resources for sentiment analysis such as labeled corpora, and polarity lexica.

To generate the input data we have extracted the reviews from different Arabic Algerian news web sites. The corpus comprises almost 1000 reviews in Arabic, of which 350 of them are considered as subjective reviews, the second 350 are objective opinions and the rest of them are neutral.

The selection of the web sites was based on the quality of the language used, because many sites use slang, and that makes understanding difficult for many Arabic speakers. Most of Arabic dialects can be understood in different Arabic countries except some specific cases. Therefore, we carried out an in-depth analysis of these web sites to ensure that the dialects used in all comments were understandable by Arabic native speakers. Thus, this process involved collecting reviews from several Arabic sports sites, cleaning and annotating all of them.

TABLE I. EXEMPELS OF ARABIC REVIEWS.

Polarity	Reviews in English language	Reviews in Arabic language
Subjective	Good game and the players were good I enjoyed this interesting game.	المباراة جيدة واللاعبون كانوا جيدين لقد استمتعت كثيرا بهذه المباراة الشيقة.
Objective	Yes certainly Breeze was stupid when he thought of abandoning the struggling valiant star Di Maria.	نعم وبكل تأكيد بريز كان أحمق عندما فكر بالتخلي عن النجم المكافح الصنديد دي ماريا
Neutral	He is free to say whatever he wants	انه حر في قول ما يريد.

2 Features extraction

Feature extraction is a basic and essential phase for Sentiment detection process. Therefore, it is important to convert the Arabic review text into a feature vector, so as to process text in a much efficient manner. In text domain, effective feature selection is a must in order to make the learning task effective and accurate. But in text classification, it's a group of statistical features.

The selected features that we used in our system for Arabic polarity detection can be summarized as follow:

$$\text{Number of sentences} = \sum sentences \quad (1)$$

$$\text{Number of positive words} = \sum PositiveWord \quad (2)$$

$$\text{Number of negative words} = \sum \text{NegativeWord} \quad (3)$$

$$\text{Number of neutral words} = \sum \text{NeutralWord} \quad (4)$$

$$\text{Sum of polarity words} = \sum PositiveWord + \sum NegativeWord + \sum NeutralWord \quad (5)$$

$$\text{Average of positive polarity words} = \frac{\sum PositiveWord}{\text{Sum of polarity words}} \quad (6)$$

$$\text{Average of negative polarity words} = \frac{\sum NegativeWord}{\text{Sum of polarity words}} \quad (7)$$

$$\text{Average of neutral polarity words} = \frac{\sum Neutral\ Word}{\text{Sum of polarity words}} \quad (8)$$

Number of predicates = $\sum Predicates$ (9)

Number of Adjectives = $\sum Adjectives$ (10)

Number of Adjectives = $\sum Adverbs$ (11)

Average of predicates = $\frac{\sum Predicates}{\sum Predicates+\sum Adjectives+\sum Adverbs}$ (12)

Average of adjectives = $\frac{\sum Adjectives}{\sum Predicates+\sum Adjectives+\sum Adverbs}$ (13)

Average of adverbs = $\frac{\sum Adverbs}{\sum Predicates+\sum Adjectives+\sum Adverbs}$ (14)

- Emotionalism: The researchers exploited the presence of the adverbs and adjectives in a document as an indicator permitting to determine the opinions. We calculate the emotionalism of a document with counting the number of the adverbs, adjectives and predicates.

Emotionalism= $\frac{\sum Predicates+\sum Adjectives+\sum Adverbs}{\sum noun+\sum verbs}$ (15)

- Reflexivity: The reviewers uses a lot of reflexivity pronouns as «I/me أنا, I am personally أنا شخصيا ». For example, use of «ي» in « رأيي»« I think that », « من وجهة نظري », «my point of view »…etc, all these sentences make reference to an opinion of review, and therefore, we include the measure of the reflexivity. All documents contain a large number of these words will be more subjective. This measure is expressed by Ref (d):

$$Ref(d) = \frac{|\{w \cap w' \backslash w \in d, w' \in R\}|}{|R|+|A|} \quad (16)$$

  - ✓ *d:* document
  - ✓ *R:* reflexivity list
  - ✓ |R|: the number of reflexivity pronouns in *d* from *R*
  - ✓ |A|: the number of addressing pronouns in *d* from *A*

- Addressing: Most reviews contain some addressing words like: « you أنت, you انتم, yourself نفسك, yourselves أنفسكم, he هو, she هي, them هم, himself نفسه, herself نفسها, themselves أنفسهم», because the reviewers write their opinions while addressing them to others.

$$Add(d) = \frac{|\{w \cap w' \backslash w \in d, w' \in A\}|}{|R|+|A|} \quad (17)$$

  - ✓ *d:* document
  - ✓ *A:* Addressing list
  - ✓ |R|: the number of reflexivity pronouns in *d* from *R*
  - ✓ |A|: the number of addressing pronouns in *d* from *A*

3 Classification

The subjectivity detection is a binary classification task where an opinionated document is labeled with an overall subjective or objective. The input to the Sentiment Classifier can be opinionated or sometimes not. So the results of the classification phase are three classes: subjective, objective and neutral.

To accomplish this phase and to obtain the perfect results we set as a goal to test and to analyze different classifiers in order to decide which one is the most performed in the opinion mining field and with Arabic language specially. There for, the most four known classifiers are used in this experiment: Naïve Byes, Multinomial Naïve Byes, SVM (Support Vector Machine) and Neural Networks.

The classifiers may be of different nature, but it uses the same feature space and generates the same classes. So the combination of multiple classifiers can ameliorate the performance of the system and increase the ratings. There for a fusion strategy is used which is the majority vote to point to the importance of all the classifiers despite their performance individually as a decision maker in the obtained hybrid classifier.

III. EXPERIMENTS AND DISCUSSION

In order to evaluate our approach, we applied widely used measurements such as Precision, Recall, F measure, and Accuracy [5].

The terms TrueSubjective (TS counts the subjective opinions correctly classified), FalseSubjective (FS counts the subjective opinions incorrectly classified), FalseObjective (FO counts the objective opinions incorrectly classified), TrueObjective (TO counts the objective opinions correctly classified), TrueNeutral (TN counts the neutral opinions correctly classified), FalseNeutral (FN counts the neutral opinions incorrectly classified), were gathered to calculate the previous measures, according to the following formulas:

$$Precision = \frac{TS}{TS+FS} \quad (18)$$

$$Recall = \frac{TS}{TS+FO+FN} \quad (19)$$

$$Fmeasure = 2\ \frac{Precision.Recall}{Precision+Recall} \quad (20)$$

$$Accuracy = \frac{TS+TO+TN}{TS+FS+TO+FO+TN+FN} \quad (21)$$

To highlight the importance, and to clarify the effect of the fusion of multiple supervised techniques in the Arabic subjectivity detection results, we carried out two experiments to calculate the previous measures. The first experiment applied on the four classifiers Naïve Byes, Multinomial Naïve Byes, SVM and Neural Networks, the second experiment consists on combining these four classifiers in order to vote for the final decision. Fusion of classifiers is a promising research area, which allows the overall improvement of the system recognition performance. Table 2 shows the results of the experiments.

Table 2. The results of the experiment

Classifiers	Precision	Recall	Fmeasure	Accuracy
Naïve Byes	0.763	0.609	0.635	60.886%
Multinomial Naïve Byes	0.688	0.648	0.657	64.822%
SVM	0.727	0.597	0.723	**72.332%**
Neural Networks	0.926	0.913	0.914	**91.304%**
Fusion	0.979	0.933	0.953	**97.300%**

After testing the four classifiers on our Arabic dataset, we noticed that the Neural Networks gives the best results for subjectivity detection with an accuracy of 91.30%. Despite that, we can't generalize this by saying that the Neural Networks are the best performing classifier for Arabic subjectivity detection for the reason of the sensibility of the classifiers with the training data. The augmentation of the dataset can be helpful for the classifier and sometimes can be perturbation for it. To overcome this problem, we lead another experiment with the fusion of all these classifiers and it increased the general accuracy indeed. So our proposed hybrid approach improved the system performance for the subjectivity detection in Arabic sports reviews.

IV. CONCLUSION

This work has considered analyzing different classifiers for subjectivity detection in Arabic text. A dataset, which consists of 1000 reviews, was collected and labeled manually. Unfortunately, working with Arabic adds more difficulties than the languages that derive from Latin, because it implies the solving of different types of problems such as the short vowels, al-hamzah, prefixes, suffixes, etc. Therefore, we have tried to combine multiple classifiers from different nature in order to increase the rate of the polarity detection system. Although obtained results with individual classifiers were promising, we have shown that the hybrid classifier and the combination technique improved on the performances achieved individually.

Certainly there are many ways that this work can and will be improved. Firstly, the size of the dataset is rather small and if we want to make solid conclusions then we definitely need big datasets and different domains. Secondly, we could enrich feature vector with another sort of morphological primitives using natural language processing. We must also study the semi supervised techniques.

References

[1] A. Abbasi, H. Chen and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums" ACM Trans, 2008.

[2] A. Hassan, A. Abu-Jbara, R. Jha and D. Radev, "Identifying the semantic orientation of foreign words,". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, Stroudsburg, PA, USA. Association for Computational Linguistics, pp 592– 597, 2011.

[3] M. Abdul-Mageed and M. Diab, "Subjectivity and sentiment annotation of modern standard arabic newswire,". In Proceedings of the 5th Linguistic Annotation Workshop, Portland, Oregon, USA, June. Association for Computational Linguistics, pp 110–118, 2011.

[4] http://www.internetworldstats.com

[5] Y. Yang, "An evaluation of statistical approaches to text categorization". Journal of Information Retrieval, 1(1/2), pp 67–88, 1999.

Medical Data Classification System based on bayesian networks Selection Ensemble

Soraya Cheriguene[1,2], Nabiha Azizi[1,2], Nadir Farah[1,2], Amel Ziani[3]
[1]Labged laboratory: Computer science department,
[2]Badji Mokhtar University of Annaba, PO box 12, Annaba, 23000, Algeria
[3] Computer science department, Lri laboratory: Information Research Laboratory,
Annaba University, Algeria
cheriguene@labged.net, azizi@labged.net

***Abstract*— Computer-Aided Diagnosis (CAD) systems have become an emerging research field in medical diagnosis. The classification is considered as the last and the decisive step in a CAD system. It exploits the results of feature extraction step to classify the new and unseen instances. In this paper, we propose a new multiple bayesian network classifier system (MCS) for medical detection system. Classifier selection is an important issue in MCS which aims to remove redundant and irrelevant members from the initial classifier set and increase the performance ensemble. Accuracy and diversity are two important parameters referring to the performance of an ensemble system. Inspiring from the minimum Redundancy Maximum relevance (mRMR) feature selection algorithm, we propose a two-stage classifier set selection based on relevance and diversity measures. Experiments were carried out on six data sets from UCI Machine Learning Repository and Ludmila Kuncheva Collection. The experimental results are encouraging and validate the effectiveness of the proposed classifiers selection method and also the effectiveness of the bayesian network ensemble.**

Keywords—medical data classification; classifiers selection; diversity measures; relevance; MRMR method, bayesian network.

I. INTRODUCTION

Over the recent years, computer-aided diagnosis (CAD) systems have become an emerging research field in medical diagnosis and receive growing attention from medicine industry [1]. The basic concept of CAD system is to assist clinicians in their medical diagnoses using computer result as a second opinion [2]. The success of such system is due to its speed of diagnosis, efficiency, effectiveness, consistency and ability to deliver reliable solutions to assist the step of glaucoma detection. CAD system is a suite of phases that have to be executed one after the other; it generally consists of three main stages: image acquisition, feature extraction and the classification phase [3].

The classification is considered as the last and the decisive step in a CAD system. It exploits the results of feature extraction step to classify the new and unseen instances. Numbers of CAD systems have been developed for medical detection system and have mostly make use of single classifier [4,5]. Therefore, in this paper, we focus on employing multiple classifier systems (MCS) using bayesian network as base classifier for improving the accuracy of medical data classification.

MCSs are very efficient technique, mostly because of the fact that the classification performance of an ensemble often outperforms their base models [6-8]. The diversity of individual classifiers, classifier selection and the combination rule for the outputs of these classifiers are three main issues in classifier ensemble [9]. Many researchers have demonstrated that there is not a clear accuracy gain in an ensemble built from a set of identical base classifiers [10]. The ideal situation is that the initial base classifiers should make uncorrelated errors, and the final classification error can be minimized by the combination of multiple base classifiers. In others words, it is important to assure the diversity between the members in the generation phase of the ensemble [11]. Generally, the diversity in classifier ensemble is achieved by manipulating the training set using randomly selection in order to generate multiple hypotheses like bagging and boosting methods [12, 13] or by manipulating the feature set, i.e. Random Subspaces (RSS) [18-16].

In this paper we focus on Random Subspace method using Bayesian Network as base learner to construct the initial classifier ensemble. The bayesian network is one of the widely used models in medical data classification [16-18]. Bayesian networks are popular decision support models because they inherently model the uncertainty in the data. They are a successful marriage between probability theory and graph theory. They allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data [17]. The RSS create various component classifiers using different random subsets of features to train them. In each pass, such a selection is made and a subspace is fixed. Then all samples are projected to this subspace, and a classifier is trained by using the projected training samples [15].

Classifier selection is an important and difficult problem which aims to decrease the computational complexity of the machine learning algorithm, remove redundant and irrelevant members from the initial classifier set and increase the

performance ensemble. Classifier selection can be divided into two general categories: static and dynamic. In the first type, the ensemble members are defined during the training phase once and used for the classification of unseen patterns, while in the second type, they are defined during the classification based on training performances and also various parameters of the actual sample to be classified [19]. Accuracy and diversity are two important parameters referring to the performance of an ensemble system [20]. The ideal situation is when individual classifiers are the most accurate where the probability of correct classification for the recognition objects is the greatest, but are possibly different from each other at the same time. The individual classifiers must be both diverse and accurate [21-24]. The key issue in classifier ensemble selection is how to use these two parameters to find optimal classifier subset.

Recently, various approaches have been developed for gene selections which achieve promising results. Ding and Peng have been developed Minimum Redundancy Maximum Relevance method (mRMR) which aims to maximize the relevancy of a gene subset while minimizing the redundancy among the genes to find the optimal subset of multiple genes [25-27].

Inspiring from this philosophy of selection, we propose in this paper a new MCS to classify medical diagnostic data. At first stage an initial pool of classifiers are trained using the Random Subspace method with the aim of assuring initial diversity among base classifiers. In the next step, we propose a two-stage classifier subset selection scheme using modified version of mRMR algorithm. The proposed selection approach: Maximum Relevance Maximum diversity (MRMD), attempts to choose an optimal subset of classifiers by selecting in the first stage the accurate classifiers that have the highest relevance and eliminating redundant ones in the second stage while increasing the amount of diversity within the selected ensemble. The decisions provided by the selected classifiers are merged using majority voting rule to produce the final classification result. Majority voting is widely combination method whose effectiveness has been proven empirically and theoretically [16]. Majority voting is the most popular voting method. The concept of voting is both simple to implement and appealing [29]. Here, given a new pattern x, each classifier votes for one specific class, and the final output class label is the one that receives the highest number of votes.

The remainder of this paper is organized as follows. In Section 2, we give an outline of the MRMR method. Section 3 our classifier selection method is presented in detail. Several comparative experiments with UCI classification data sets and Ludmila Kuncheva Collection of real medical data are demonstrated in Section 4. Finally section 5 concludes this paper.

II. MRMR METHOD

The mutual information is a quantity that measures the mutual dependence between two random variables X and Y. In this case, information is thought of as a reduction the uncertainty associated of a random variable due to the knowledge of the other random variable [27]. Thus, the more mutual information between X and Y, the less uncertainty there is in X knowing Y or Y knowing X. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;\ Y)\ =\ \sum_{x\in X}\sum_{y\in Y} p(x;\ y)\log\frac{p(x;y)}{p(x)\,p(y)} \qquad (1)$$

Where *p(x, y)* is the joint probability distribution function of *X* and *Y*, and *p(x)* and *p(y)* are the marginal probability distribution functions of *X* and *Y*, respectively. In the continuous case, we replace summation by a double integral.

Using the concept of mutual information, the MRMR method proposed by Peng et al. [25] aims at selecting candidate genes with both the maximum relevance for the target concerned and the minimum redundancy among the genes themselves. Given g_i, which represents the gene *i* in *S,* g_j, which represents the gene *j*, and the class label *c*, The redundancy p and the pertinence r are defined as :

$$Per(i) = \frac{1}{|S|}\sum_{g_i\in S} I(g_i, \mathrm{c}) \qquad (2)$$

$$Red(i) = \frac{1}{|S|}\sum_{g_i,g_j\in S} I(g_i, g_j) \qquad (3)$$

To obtain the gene *i* with maximal relevance for the target c and minimal redundancy relative to the others genes in S, Eqs. (2) and (3) are combined into the score function [33]:

$$Score(i) = Per(i)/Red(i) \qquad (4)$$

III. PROPOSED APPROACH

As mentioned before, the object of the classifier selection process is to choose appropriate set of members according to selection measures and selection algorithm with the aim to reduce the number of ensemble members. In practice, accuracy and diversity of base learners are two important factors to achieve better classification performance in ensemble learning. In this paper, we propose a two-stage classifier subset selection inspired from the mRMR feature selection algorithm which based on the mutual information and diversity measures to control the balance between the accuracy and diversity among the base classifiers.

The main idea of the proposed algorithm MRMD takes as input a pool of classifiers $\mathrm{C} = \{\mathrm{c}_1, \mathrm{c}_2, \dots, \mathrm{c}_\mathrm{N}\}$ with size N. The classifiers outputs must be presented with a binary matrix of size N *L (N is the number of the objects and L is the number of classifiers), when each column represents the output of classifier c_p for all samples N, $\mathrm{c}_\mathrm{p} = [\mathrm{c}_{\mathrm{p},1}, \mathrm{c}_{\mathrm{p},2}, \dots \mathrm{c}_{\mathrm{p,N}}]^{\mathrm{T}}$, with $\mathrm{c}_{\mathrm{p,i}}$ is the output of the classifier p for the instance i. the vector Ω represents the class of the instances, $\Omega = \{\omega_1, \omega_2, \dots \omega_\mathrm{M}\}$.

In the first stage, each classifier's pertinence is evaluated on the validation set by calculating the relevance using mutual information as per Equation 5 and then the highest scorer classifier is selected and added to new subset. Next a looping is performed for the remaining classifiers. At each iteration, we calculate the relevance of all possible combinations ensemble between the output classifier and the remaining classifiers and we select highest scorer set comparing with the others sets and the current selected set. The algorithm stops when the maximum iteration number threshold is reached.

$$Per(c_p) = \frac{1}{|L|}\sum_{c_p \in C} I(c_p, \Omega) \quad (5)$$

In the second stage, the disagreement measure is used to accomplish the classifier selection in the reduced classifier set. In this step, we evaluate the classifiers by calculating the diversity of each one using disagreement measure as per Equation 6 and remove the members with weak diversity from the ensemble. The selection process is repeated until the ensemble size meeting requirement.

$$Div(i) = \frac{1}{|C|}\sum_{c_i, c_j \in S} Dis(c_i, c_j) \quad (6)$$

Where $dis(c_i, c_j)$ is the disagreement measure between the classifiers c_i, c_j in *C*.

Disagreement measure (DS) represents the number of times that one of the classifiers was incorrect and the other correct. It can thus be defined for two classifiers a and b as :

$$Dis(i,j) = \frac{N^{10}+N^{01}}{N^{00}+N^{01}+N^{10}+N^{11}} \quad (7)$$

Where N^{00} is the number of patterns that both classifiers wrongly classified; in contrast, N^{11} stands for the number of patterns that both classifiers correctly classified; N^{10} is the number of patterns classified correctly by classifier D_i but not by D_j; likewise, N^{01} is the total of patterns classified correctly by classifier D_j, but not by D_i. The diversity increases with increasing values of the disagreement measure in the range from 0 to 1 [28].

IV. Experimentation

A. Experimentation setup

In order to evaluate the performance of the new proposed method, we have chosen six binary medical datasets from the UCI machine learning data repository [29] and Ludmila Kuncheva Collection of real medical data [30]. These datasets represent a real-world medical data collected from human patients and the attributes are similar to the one used by the pathologist. In order to minimize the influence of variability in the training set, 5-fold cross validation was on the five datasets. In detail, each dataset was partitioned into five subsets with similar sizes and distributions. Then, the union of 3 subsets was for generating individual classifiers, a subset is used as the validation set for competence and diversity measures while the remaining subset is used as the test set.

Our experiments are divided into two parts. First, we present the results of the overall classifiers using different learner base, we compare the results of the bayesian ensemble using the Naïve-Bayes with other classifier ensemble as Multi-Layer Percepetron (MLP), J48, SVM and LADTree. The base classifiers are taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.4. The parameters used for each algorithm in this study were set at the default settings.

In the second part, we compare the performance of the proposed selection method using the Naïve-Bayes as base learner against four multiple classifier selection systems: (1) SB - this system selects the best classifier in the ensemble with the highest accuracy; (2) MRMR – this system selects the best classifiers in the pool with the high score using MRMR method; (3) DIV (DS/WCEC) – this system defines the competence of the classifier subsets according to diversity measures and next the ensemble of divers classifiers is selected; (4) PER – this system is based on selecting the best pertinent classifiers.

All the learning and combination methods used in this study were conducted using Java Language using WEKA tools and an initial pool of 30 classifier are trained using the Random Subspace method with the aim of assuring initial diversity among base classifiers. The parameters used for each algorithm in this study were set at the default settings. The size of selected classifiers is fixed at 5 in all methods.

TABLE I. Datasets used for classification

Dataset	#Attribute	#Simple	#Positive	#Negative
Diabetes (D1)	8	768	268	500
Echocardiogram (D2)	132	12	24	50
Heart disease (D3)	13	270	120	150
Parkinson (D4)	22	195	48	147
Respiratory (D5)	17	85	45	40
Wisconsin breast cancer (D6)	9	699	458	241

B. Experimentation results

TABLE II. Random Subspace results

Dataset	MLP	J48	ADTree	SVM	Naïve-B
D1	**76,69**	74,49	**76,69**	65,62	75,52
D2	82,59	82,92	81,48	81,02	**85,18**
D3	83,51	81,29	82,58	79,35	**83,96**
D4	77,94	78,46	78,46	74,87	**81,02**
D5	**92,94**	89,41	91,76	91,54	**92,94**
D6	96,15	95,85	96,85	95,99	**97,14**
Average	84,97	83,73	84,63	81,39	**85,96**

The classification accuracy of the Random Subspace classifier ensembles are presented in Table 2. The accuracy refers to the percentage of correct classification on testing data. As depicted in this table, the average performance of the bayesian network ensemble is better than those of MLP, J48, ADTree and SVM on five datasets: Naïve-B ensemble is on average 0.99% higher than MLP ensemble, 2.23% higher than J48, 1.33% higher than ADTree and 4.57 higher than SVM. The only exception was D5 dataset when the accuracy has significantly enhanced using MLP ensemble. Also, the SVM ensemble shows a lower accuracy compared to the other ensembles.

TABLE III. ACCURACY OF DIFFERENT CLASSIFIER SELECTION METHOD USING BAYESIAN NETWORK

Dataset	SB	mRMR	DIV	PER	MRMD
D1	72,36	78,62	74,96	78,02	**79,92**
D2	75,61	78,51	87,94	87,98	**88,51**
D3	77,03	84,07	85,92	84,07	**87,03**
D4	84,01	89,23	87;69	89,23	**90,76**
D5	88,23	91,76	90,58	94,11	**95,29**
D6	96,28	97,00	96,42	97,28	**98,42**

TABLE IV. SPECIFICTY OF DIFFERENT CLASSIFIER SELECTION METHOD USING BAYESIAN NETWORK

Dataset	SB	mRMR	DIV	PER	MRMD
D1	32,09	53,67	48,48	52,81	**54,97**
D2	48,00	64,00	74,54	74,54	**80,00**
D3	82,23	89,07	90,13	**90,78**	**90,78**
D4	59,52	85,71	76,19	**83,33**	**83,33**
D5	90,47	91,17	94,11	**97,05**	**97,05**
D6	96,66	97,24	96,66	97,24	**98,44**

Table 3 presents the classification accuracy of the different classifier set selection method using the Bayesian network. The best result for each database is bolded. As depicted in this table, the proposed method (MRMD) provides better performance than the others method in the six datasets: D1 with **79.92%,** D2 with 88.51%, D3 with 87.03%, D4 with 90.78, D5 with 95.23% and D6 with 98.29%

Beside the accuracy evaluation, it is important to observe the specificity, sensitivity, AUC of the medical diagnostic system. Sensitivity is defined as the number of true positive classifications divided by all positive classifications. Specificity represents the true negative rate and it is calculated by the division of true negative classifications by true negative and false positive classifications .The area under curve (AUC) is another widely used metric for evaluating the classifiers performances. It equals to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It takes on values from 0 to 1. The higher the value of AUC, the better the classification algorithm is .AUC value is calculated from the area under the ROC curve. ROC curves are usually plotted using true positives rate versus false positives rate, as the discrimination threshold of classification algorithm is varied [28].

Given the ratio of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), specificity and sensitivity measures can be calculated with the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} * 100\% \quad (8)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} * 100\% \quad (9)$$

TABLE V. SENSITIVITY OF DIFFERENT CLASSIFIER SELECTION METHOD USING BAYESIAN NETWORK

Dataset	SB	mRMR	DIV	PER	MRMD
D1	89,73	89,92	86,32	**89,73**	88,80
D2	92,59	87,65	**97,46**	**97,46**	93,82
D3	70,33	77,11	76,27	79,66	**82,03**
D4	90,84	90,19	90,84	90,84	**92,81**
D5	86,04	92,15	88,23	92,15	**94,11**
D6	96,28	97,00	96,42	97,28	**98,42**

TABLE VI. AUC OF DIFFERENT CLASSIFIER SELECTION METHOD USING BAYESIAN NETWORK

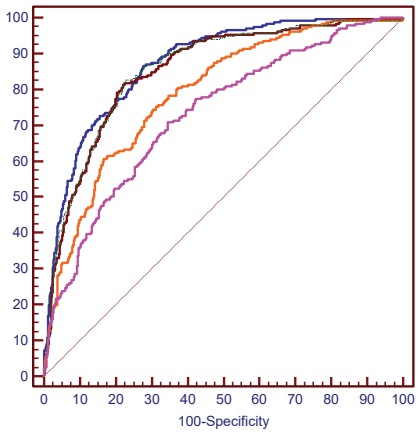
Dataset	SB	mRMR	DIV	PER	MRMD
D1	73,00	85,20	78,90	85,60	**87,20**
D2	84,90	90,80	70,20	71,40	**93,40**
D3	89,60	93,00	93,70	94,20	**95,40**
D4	91,90	94,50	84,30	94,50	**96,50**
D5	98,70	97,80	95,70	97,90	**96,40**
D6	99,00	99,10	99,20	99,30	**99,60**

In terms of Sensitivity and Specificity and As Table 4 and Table 5 are examined, MRMD have has offered better results than the others ensemble learning. Based on Naïve-Bayes ensembles, the proposed approach MRMD has the highest rate of sensitivity over the six datasets, and also the PER ensemble has the highest rate over D3, D4 and D5 datasets. On the other hand the MRMD has the best sensitivity results over four datasets: D3 with 82.03%, D4 with 92.81%, D5 with 94.11% and D6 with 98.42%. For D1 and D2 datasets the PER ensemble outperforms the others classifier selection systems.

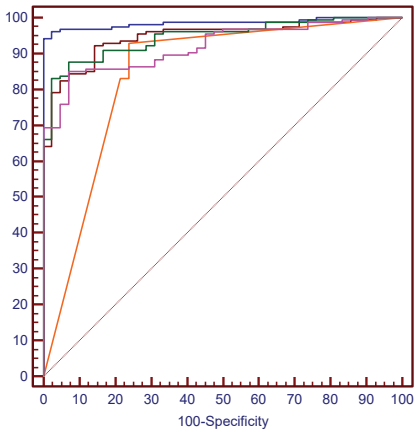
In Table 6, a comparison of the areas under the ROC curve for the proposed approach is listed. The receivers operating characteristic (ROC) curve of the given medical datasets are shown in Fig. 1.

V. CONCLUSION

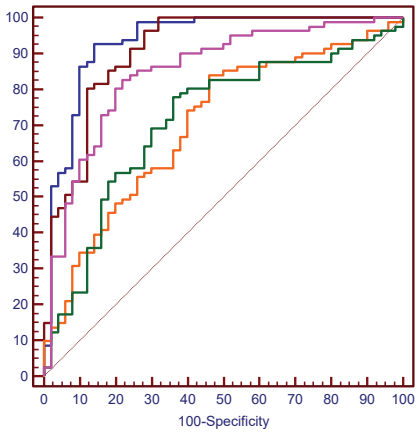
Computer-aided diagnosis is an emerging research field that has been extensively studied. A crucial task for a CAD system is discovering efficient machine learning method. In this paper, we have presented a novel approach for constructing ensembles for medical data classification problems based on classifier set selection and mRMR method. The proposed method uses the bayesian network as classifier base. The main contribution of this paper is the two-stage classifier set selection scheme that use both accuracy and diversity as a means to select classifiers. Experiments conducted on six data sets from UCI Machine Learning Repository and Ludmila Kuncheva Collection have confirmed that the classifier subset selection step looks really promising in improving the use of multiple classifiers. During the experimentation, we also looked at using diversity measure, together with mean base classifier relevance, as the selection criterions is better than the results achieved by each one separately.



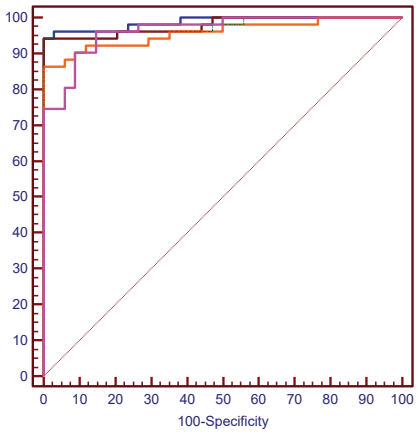
(Diabetes)



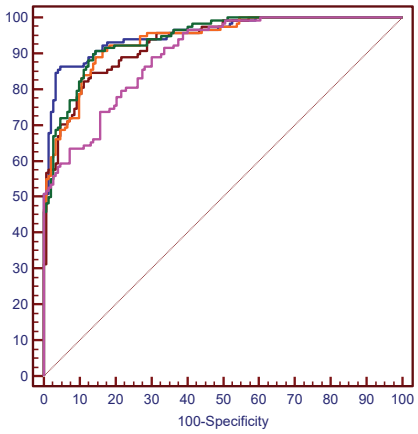
(Parkinsons)



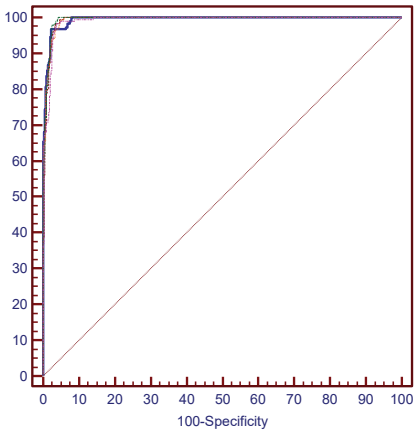
(Echocardiogram)



(Respiratory)



(Heart-Statlog)



(W Breast Cancer)

Fig. 1. Roc curves of the classifier selection methods on six data sets

REFERENCES

[1] M. Daliri, "Automated diagnosis of Alzheimer Disease using Scale Invariant feature transform in magnectic resonance Images, " Journal of medical systems, pp. 995-1000. 2011.

[2] S. Lahmiri and M. Boukadoum, "Alzheimer's disease detection in brain magnetic resonance images using multiscale fractal analysis," *ISRN Radiol.*, vol. 2013, p. 627303, 2013.

[3] S. Zhang, I. Cohen, M. Goldszmidt, J. Symons, and A. Fox, "Ensembles of models for automated diagnosis of system performance problems," *Proc. Int. Conf. Dependable Syst. Networks*, no. July, pp. 644–653, 2005.

[4] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6844–6852, 2015.

[5] S. Cheriguene, N. Azizi, N. Zemmal, N. Dey, N. Farah, and H. Djellali, "Optimized Tumor Breast Cancer Classification Using Combining Random Subspace and Static Classifiers Selection Paradigms," *Appl. Intell. Optim. Biol. Med. Intell. Syst. Ref. Libr.*, vol. 96, pp. 159–180, 2016.

[6] L. I. Kuncheva, *Combining Pattern Classifiers*, vol. 47, no. 4. 2004.

[7] E. Board, "Multiple Classifier Systems," pp. 1–24, 2010.

[8] L. Chen and M. S. Kamel, "A generalized adaptive ensemble generation and aggregation approach for multiple classifier systems," *Pattern Recognit.*, vol. 42, no. 5, pp. 629–644, 2009.

[9] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, and T. Ing Ren, "META-DES: A dynamic ensemble selection framework using meta-learning," *Pattern Recognit.*, vol. 48, no. 5, pp. 1925–1935, 2015.

[10] L. Zhang, W. Da Zhou, and F. Z. Li, "Kernel sparse representation-based classifier ensemble for face recognition," *Multimed. Tools Appl.*, vol. 74, no. 1, pp. 123–137, 2013.

[11] L. I. Kuncheva, "That elusive diversity in classifier ensembles," *Lect. Notes Comput. Sci.*, pp. 1126–1138, 2003.

[12] L. Breiman, "Bagging predictors," *MachineLearning*, 1996, pp. 123–140.

[13] Y. Freund, R. Schapire, "Experiments with a new boosting algorithm". Proceedings 13th International Conference on Machine Learning, 1996, pp. 148–156.

[14] T. Ho, "The random subspace method for constructing decision forests," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998, pp. 832–844.

[15] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. A. Larson, "POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis," *Inf. Process. Manag.*, vol. 51, no. 4, pp. 458–479, 2015.

[16] H. Borchani, C. Bielza, P. Martí nez-Martín, and P. Larrañaga, "Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: An application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39)," *J. Biomed. Inform.*, vol. 45, no. 6, pp. 1175–1184, 2012.

[17] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. 184–190, 2006.

[18] C. E. Kahn, L. M. Roberts, K. A. Shaffer, and P. Haddawy, "Construction of a Bayesian network for mammographic diagnosis of breast cancer," *Comput. Biol. Med.*, vol. 27, no. 1, pp. 19–29, 1997.

[19] A. S. Britto, R. Sabourin, and L. E. S. Oliveira, "Dynamic selection of classifiers - A comprehensive review," *Pattern Recognit.*, vol. 47, no. 11, pp. 3665–3680, 2014.

[20] M. Aksela and J. Laaksonen, "Using diversity of errors for selecting members of a committee classifier," *Pattern Recognit.*, vol. 39, no. 4, pp. 608–623, 2006.

[21] M. Aksela and J. Laaksonen, "Using diversity of errors for selecting members of a committee classifier," *Pattern Recognit.*, vol. 39, no. 4, pp. 608–623, 2006.

[22] C.-Y. Chiu and B. Verma, "Effect of Varying Hidden Neurons and Data Size on Clusters, Layers, Diversity and Accuracy in Neural Ensemble Classifier," *2013 IEEE 16th Int. Conf. Comput. Sci. Eng.*, pp. 455–459, 2013.

[23] Y. Bi, "The impact of diversity on the accuracy of evidential classifier ensembles," *Int. J. Approx. Reason.*, vol. 53, no. 4, pp. 584–607, 2012.

[24] Lam, L, and Suen, C.Y, "A theoretical analysis of the application of majority voting to pattern recognition", *IEEE Transaction on Systems, Man, and Cybernetics*, Volume 2, 9-13 Oct. 1994, Page(s):418-420.

[25] H. Peng , F. Long and C. Ding , "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy" , *IEEE Trans. Pattern Anal. Mach. Intell.* , vol. 27 , no. 8 , pp.1226 -1238 , 2005.

[26] A. El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, 2011.

[27] E. G. Learned-miller, "Entropy and Mutual Information," pp. 1–4, 2013.

[28] X. Ma and X. Sun, "Sequence-based predictor of ATP-binding residues using random forest and mRMR-IFS feature selection," *J. Theor. Biol.*, vol. 360, pp. 59–66, 2014.

[29] C.L. Blake, C.J Merz, UCI Repository of Machine Learning Databases[http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science,1998

[30] L. Kuncheva, "Ludmila kuncheva collection," 2004. [Online]. Available: http://pages.bangor.ac.uk/~mas00a/activities/real_data.html.

Multi-Dialect Arabic continuous speech Recognition with Dynamic Bayesian Networks

Elyes Zarrouk, Yassine Benayed and Faïez Gargouri
MIRACL, Multimedia InfoRmation system and Advanced Computing Laboratory
Higher Institute of Computer Science and Multimedia, ISIMS, University of Sfax, Tunisia.

ABSTRACT

***Abstract*—Such as Hidden Markov Models (HMM), Dynamic Bayesian networks (DBN) suffers from a discriminatory ability especially on speech recognition even if their progress is huge. In order to ameliorate the results of recognition systems, we apply Support Vectors Machine (SVM) as an estimator of posterior probabilities since they are characterised by a high predictive power and discrimination. Moreover, they are based on a structural risk minimization (SRM) where the aim is to set up a classifier that minimizes a bound on the expected risk, rather than the empirical risk. In this paper, we describe the use of the hybrid model SVM/DBN for Arabic triphones-based continuous speech recognition. The best results are obtained with the proposed system SVM/DBN when we achieve 93.02% as the best average recognition rate of multi-dialect Arabic continuous speech compared to 73.24%, 80.83% and 88.13 % respectively with triphones mixture-Gaussian HMM, SVM/HMM and DBN systems.**

Keywords— Dynamic Bayesian networks; hybrid models; supports vectors machine; Hidden Markov models, Arabic continuous speech.

I. INTRODUCTION

A typical ASR system operates with the help of five basic modules: feature extraction for signal parameterization, acoustic models, language models, pronunciation model and decoder.

Traditionally statistical models, such as Gaussian mixture models, have been used to represent the various modalities for a given speech sound.

Bayesian Networks are a particular type of Graphical Models, providing a general and flexible framework to model, factor, and compute joint probability distributions among random variables in a compact and efficient way. DBN are a useful tool for representing complex stochastic processes. Recent developments in inference and learning in DBNs allow their use in real-world applications is the first successful application of DBNs to a large scale speech recognition problem. Investigation of the learned models indicates that the hidden state variables are strongly correlated with acoustic properties of the speech signal.

In this paper, we present the performance of the hybridization of Supports Vectors machine with DBN Arabic triphones-based continuous speech recognition.

In literature, several works were concerned with DBN based automatic speech recognition.

In [1] the authors presented a methodology for modeling of speech without any priori hypothesis is carried out on the dependencies between the observed values and hidden processes on the floor. The approach is technically very interesting because all the effort of calculation is realized in the learning phase. This approach is used for a spot of recognition of isolated words. In [2], Zweig describes the probabilistic procedure for recognition of sequences of observations of a Word through the HMM signal that are a special case of the DBN. Subsequently, a brief description on the basic concepts of Bayesian Networks was presented. Then, Bayesian architecture was designed for speech recognition. The results obtained are too encouraging reports to those obtained with the HMMs. In [3] the authors used the DBN framework to construct acoustic models that are capable to learn the dependency structure of the hidden and observed speech process. In [4], new ideas to improve automatic speech recognition have been proposed that make use of context user information such as gender, age and dialect. In this thesis a basic speech recognition system was built using Gaia to test if speech recognition is possible using Gaia anddbns.dbn models were designed for the acoustic model, language model and training part of the speech recognizer. Experiments using a small data set proved that speech recognition is possible using Gaia. In [5], the paper presents a Bayesian network model that uses an additional variable to represent the State of joints. The strongest point in this system is that it uses measurement of joints along the learning data, will not need to know these values in the recognition phase. Isolated words recognition results show that the system combining the acoustic variables and measures of joints is more effective than a sound system.

Several works were tried to solve many problems related to ASR using HMM by many extensions of DBN models.

In [6], the authors present the effectiveness of DBN compared to HMM for multi-dialect isolated words recognition with MSA corpora which motivated us to apply SVM as an estimator of emission probabilities hoping ameliorating the recognition rates like it was obtained with SVM/HMM by report HMM and MLP/HMM on [7].

In [9], the authors' presents the effectiveness of SVM/HMM for the recognition of Arabic continuous speech recognition based triphones recognition comparing to HMM and MLP/HMM.

The main contribution of this paper consists in the definition of a new approach able to ameliorate automatic Arabic speech recognition. We make a comparative study between HMM, SVM/HMM, DBN and SVM/DBN for continuous automatic speech recognition based triphones modeling.

This paper is organized as follows: Section 1 displays generally the architecture of an ASR. Section 2, introduces the definition of graphic models. Section 3 describes the proposed hybrid model SVM/DBN in section 4. Experimental results are presented in section 5.

II. AUTOMATIC SPEECH RECOGNITION

A system of automatic speech recognition (ASR) is to transcribe a voice message into a text. The main applications using ASR systems are automatic transcriptions, indexing multimedia documents and man-machine dialogue. Systems of automatic speech recognition of a continuous current are based on a statistical approach which [10] has proposed formalization, resulting from the information theory.

From acoustic observations X, the goal of a recognition engine is to find the sequence of words W most likely among all possible sequences. This sequence must maximize the following equation [12]:

$$\hat{W} = \operatorname{argmax}_W P(W|X) \quad (1)$$

$$\hat{W} = \operatorname{argmax}_W \frac{P(X|W)P(W)}{P(X)} \quad (2)$$

$P(X) = \sum_W P(X|W)P(W)$ does not depend on a particular value of W and can be released from the calculation of the argmax:

$$\hat{W} = \operatorname{argmax}_W P(X|W)P(W) \quad (3)$$

Where the term P (W) is estimated by using the language model. P (X|W) corresponds to the probability given by the acoustic models. This approach allows integrating the same decision-making process, the acoustic and linguistic information.

III. GRAPHIC MODELS FOR AUTOMATIC SPEECH RECOGNITION

A. *Hidden Markov models*

The speech signal can be likened to a series of units. In the context of Markov ASR, the acoustic units are modeled by HMM which are typically left-right tristate.

At each state of the Markov model there is a probability distribution associated modeling the generation of acoustic vectors via this state [11].

. An HMM is characterized by several parameters:

-N: the number of states of the model.

-$A = \{a_{ij}\} = \{P(q_{t=1}|q_{t-1=i})$ is the matrix of transition probabilities on the set of states of the model.

- $B = \{b_k(x_t)\} = P(x_t|q_{t=k})$ is the matrix of emission probabilities of the observations X_t for the state q_k.

-π is the initial distribution of states, $P(q_{i=0})$.

B. *Dynamic Bayesian networks*

In recent years, probabilistic of Bayesian networks [13] have emerged as the primary method for representing and manipulating probabilistic information. A Bayesian network provides a means of encoding the dependencies between a set of random variables (RV). The RVs and dependencies are represented as the nodes and edges of a directed acyclic graph. A Bayesian network exploits missing edges (implying conditional independence) to factor the joint distribution of all RVs into a set of simpler probability distributions. A dynamic Bayesian network consists of instances of a Bayesian network repeated over time, with dependencies across time.

The BNs formalism consists in associating a directed acyclic graph to the joint probability distribution (JPD) $P(X)$ of a set of Random variables

$$X[t] = \{X_1[t], \ldots, X_n[t]\} \quad (4)$$

The nodes of the graph represent random variables, and arrows encode conditional independences that are assumed in the JPD. The separation properties of the graph involved the set of all conditional independence relationships which are named the Markov properties. BN is completely defined by a graph structure S and a set of parameters θ of conditional probabilities of variables given their parents. Indeed, the JPD can be expressed in a form that factored is

$$P(X[t]) = \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i[t] | \prod_i) \quad (5)$$

Where $\prod_i$ denotes parents X_i in S.

A DBN consists of instances of a Bayesian network repeated overtime, with temporal dependency arcs linking the instances The structure and parameters are assumed to repeat for each time slice (i.e., the process is assumed to be stationary), so the conditional probabilities associated with Xi[t], t:1..T, are tied. In fact, DBNs can be seen as "unrolling" a one-frame network for T time steps [14] and adding time-dependencies, in effect creating a BN of size N x T. The required conditional probabilities may be stored either in tabular form or with a functional representation.

IV. HYBRID MODEL SVM/DBN

A. *Supports vectors machine*

SVMs are a new statistical learning techniques initiated by V. Vapnick in 1995 [15]. The success of this method is justified by the solid theoretical foundation that underpins it. They can address a variety of problems including classification. SVM is a method well suited to deal with high dimension data such as text and images [10]. Since their introduction in the field of pattern recognition, several studies have demonstrated the effectiveness of these techniques primarily in image processing.

Classifiers are typically optimized based on some form of risk minimization. Empirical risk minimization is one of the most commonly used techniques where the goal is to find a parameter setting that minimizes the risk:

$$R_{emp}(\alpha) = \frac{1}{m} \sum_{i=1}^{1} |y_i - f(x_i, \alpha)| \quad (6)$$

Where α is the set of adjustable parameters y_i, x_i are the expected output and given input, respectively. The parameter m is the number of couples (x_i, y_i) of input data. However, minimizing R_{emp} does not necessarily imply the best classifier possible [16]. The form of two-class problem learning is an example of structural risk minimization where the aim is to learn a classifier that minimizes

a bound on the expected risk, rather than the empirical risk [10] .SVM is based on this Structural risk.

The classification of data depends on nature of separation of data. There are cases of linearly separable data and nonlinearly separable data. With SVM a discriminative hyperlane with maximal border is searched when classes are linear separable. With constant intra classes variation classification confidence grows with increasing interclass distance. The former are the simplest SVM because they can easily find a linear separation.

$$\varphi(x): \mathbb{R}^n \rightarrow \mathbb{R}^m \tag{7}$$

And an optimal linear discrimination function will be finding in this new space. This transformation increases the linear separation.

For the nonlinear SVM, we are in front with very high dimension of the feature space $\mathbb{R}^m$. So $\varphi(x_i)\varphi(y_i)$must not be calculated explicitly, but can be expressed with reduced complexity with kernel functions.

$$K(x_i, y_i) = \varphi(x_i)\varphi(y_i) \tag{8}$$

We do not need to know, how the new feature space $\mathbb{R}^m$ looks like. We only need the kernel function as a measure of similarity such as Polynomial-kernel, Linear-kernel, Sigmoid-kernel and Radial-Basis Function kernel.

B. Hybrid model SVM/DBN

Here we have used SVM to estimate posterior probabilities in the training phase and the recognition phase [7][9]. First, we train one SVM for every sub-task signal which means one versus all. Every triphone is a separate class. The function f (x_i) that describes the separation plane measures the distance of the element x_i to the margin. The inclusion of the element x_i on one of the classes depends of sign (f (x_i)). Also, the distance is far from the margin it has a higher probability of belonging to the class

1) Emission probabilities with SVM

After choosing and applying the kernel function the conditional probability $P(x \mid class_j)$ is generating when a general model is summarized by minimizing the number of support vectors and supports the maximum data [16]. We need to calculate the likelihoods that the input vector x is given the classj of the appropriate phoneme j $P(class_j \mid x_i)$. We apply the Bayes rule to obtain those HMM emission probabilities:

$$P(class_j \mid x_i) = \frac{P(x_i \mid class_j)\, P(class_j)}{P(x_i)} \tag{13}$$

- $P(class_j \mid x_i)$ is the likelihood of the input vector x_i is given the class of the triphone j.
- $P(class_j)$ is the prior probability of the triphone.
- $P(x_i)$ is the priori probability of acoustic vector x_i

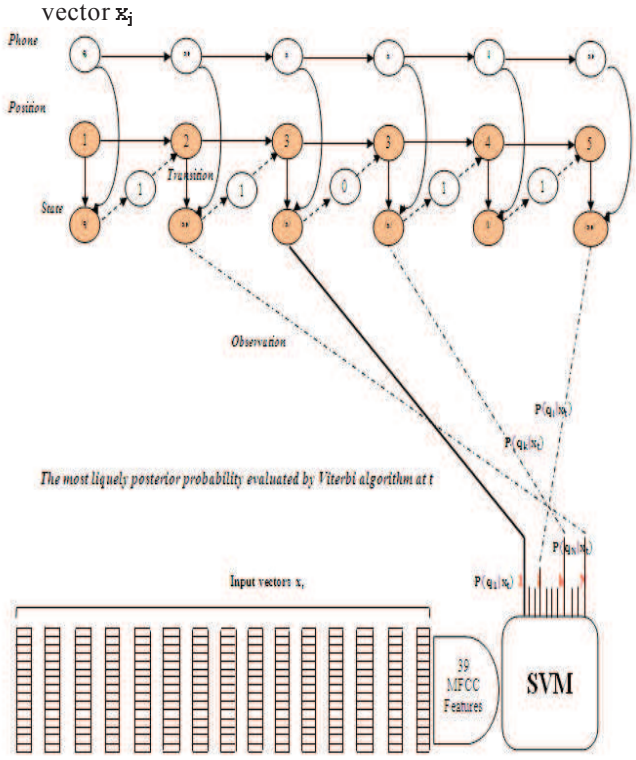


Fig. 1. General architecture of SVM/DBN hybrid model applying to the recognition of the Arabic word "qāl"

2) The decoding phase

For each triphone we attribute a DBN described on [8]. We consider that all the states are combined on one DBN Given a sequence of observations $X = x_1, x_2, \ldots, x_L$ and a DBN M with N states we wish to find the maximum probability state path $Q = q_1, q_2, \ldots q_N$.

Let $\delta_j(t)$be the probability of the most probable path ending in state j at time t:

$$\delta_j(t) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = j, x_1 x_2 \ldots x_t \mid M)$$

So $\delta_j(t) = P(q_j \mid x_t)$ which is the probability estimated by the SVM kernel function from the observation x_t

We have to determinate finally:

$$q_{optimal} = \max_{1 \le j \le N} [\delta_j(N)] \tag{14}$$

At the end we choose the highest probability endpoint, and then we backtrack from there to find the highest probability path. We obtain a sequence of states that represents the observations' sequence X.

The Viterbi algorithm is the best solution to this problem is to estimate the posterior probability P (Q | X). The state sequence most likely at the time t depends only t and the most likely sequence to t-1.

V. EXPERIMENTATIONS

In our experimentations, we have used multi dialect Arabic speech parallel corpus. It is designed to encompass four main dialects; Modern Standard Arabic (MSA), Gulf, Egypt and Levantine dialects. Parallel prompts were written for the four main dialects, which involved 1291 recordings for MSA and 1069 recordings for other dialects. The recordings were conducted with the consent of 52 participants. 32 speech hours were obtained. After the segmentation stage, a total number of 67,132 speech files were resulted [19].

The main domain in this work is travel and tourism. The intention behind focusing on a single specific domain rather than attempting to undertake more general work was to try and control the volume of data.

To split this domain into parts, eight general sections were done. Four sections directly related to travel and tourism: Restaurant (A), Hotel (B), Transport (C) and Street and shopping (D); and the remaining four are necessary for tasks related to this area: Days and times (E), Currency (F), Global cities (G) and Numbers (H). We will use those alphabetic abbreviations to divide every section to the authors on the results tables.

The following table presents the speakers count, the utterance counts and the phonemes counts of the four main dialects [19].

TABLE I. Description of MSA multi-dialect corpora

	Speaker count	Utterance count	Total utterance	Phonemes count	Total Phonemes
MSA	12	2790	33.480	15.505	186.060
Gulf	12	2531	30.372	11.920	171.480
Egyptian	20	2702	54.040	15.088	301.760
Levantine	8	2643	21.144	11.828	119.184
Total	52	10.666	160.180	59.781	778.484

This work was done with the graphic models tookit (GMTK). GMTK is a freely-available toolkit written in C++ that is designed for DBN-based speech recognition [14]. It has many desirable features, such as sparse, continuous observation distributions, switching parents, beam search and generalized EM training. It supports smoothing and Viterbi inference using the online Frontier algorithm. GMTK has a number of features, including a language for specifying structures and probability distributions, logarithmic space exact training and decoding procedures, the concept of switching parents, and a generalized EM training method which allows arbitrary sub-Gaussian parameter tying. Taken together, these features endow GMTK with a degree of expressiveness and functionality that significantly complements other publically available packages. We have developed our systems with the best parameters obtained after many tests and experimentations. HTK is used for modeling the phoneme or the triphone HMM. For each Arab phoneme or triphone we assigned a left-right HMM with 5 states. The emission probabilities are modeled by 64 mixtures of Gaussians. Maximum-likelihood parameter estimation was used to train HMM on data using the Viterbi algorithm.

For SVM/HMM model we have used the same architecture and parameters obtained in [9]. Also for the use of DBN system we applied the architecture presented by zweig [2] and stephenson [5] and the parameters done on [6]. Now we will present all the recognition rates obtained for each Arabic dialect by the eight sections from A to H.

When reporting the performance of a speech recognition system, sometimes word accuracy (WAcc) is used instead:

$$WAcc = \frac{\text{number of recognized words}}{\text{number of all the words}} \qquad (15)$$

TABLE II. Recognition rates with HMM, SVM/HMM, DBN and SVM/DBN of MSA-dialect continuous speech.

	HMM	SVM/HMM	DBN	SVM/DBN
A	73.81	78.32 %	87.84%	94.30%
B	68.96	77.25 %	88.07%	93.43%
C	73.7 %	76.65 %	86.65%	92.93%
D	69.61	78.84 %	88.33%	91.45%
E	68.7 %	83.93 %	86.84%	93.92%
F	68.7 %	78.56 %	87.69%	94.02%
G	73.7 %	78.94 %	89.03%	93.42%
H	72.21	75.84 %	88.44%	94.21%

TABLE III. Recognition rates with HMM, SVM/HMM, DBN and SVM/DBN of Gulf-dialect continuous speech.

	HMM	SVM/HMM	DBN	SVM/DBN
A	71.7 %	77.53 %	86.81%	91.26%
B	75.21	83.98%	87.04%	92.34%
C	71.64	78.64%	86.26%	95.02%
D	71.16	78.83 %	88.93%	94.31%
E	66.7 %	81.9 %	89.32%	93.79%
F	73.29	77.83 %	89.39%	92.09%
G	70.69	79.73%	89.86%	97.93%
H	71.27	78.94 %	90.18%	94.72%

TABLE IV. Recognition rates with HMM, SVM/HMM, DBN and SVM/DBN of Egyptian-dialect continuous speech.

	HMM	SVM/HMM	DBN	SVM/DBN
A	73.87	85.44 %	87.92%	92.22%
B	71.29	84.16 %	89.14%	93.45%
C	73.32	82.7 %	87.28%	91.90%
D	70.68	83.29 %	89.29%	93.23%
E	73.54	80.69 %	88.93%	92.09%
F	71.13	86.88 %	90.23%	91.29%
G	73.09	81.56 %	89.21%	93.49%
H	73.81	82.87 %	89.07%	92.97%

TABLE V. Recognition rates with HMM, SVM/HMM, DBN and SVM/DBN of Levantine-dialect continuous speech.

	HM	SVM/HMM	DBN	SVM/DBN
A	76.81	78.09 %	86.72%	93.02%
B	76.61	82.93 %	88.14%	91.64%
C	74.43	83.32 %	85.20%	90.29%
D	75.18	81.39%	87.22%	92.32%
E	80.68	82.12 %	88.23%	93.19%
F	87.81	80.34%	85.98%	92.32%
G	79.96	82.77%	87.20%	91.04%
H	70.48	82.34 %	89.96%	93.27%

The results presented on the previous tables' shows the good effectiveness of applying DBN on the recognition of continuous multi dialect Arabic speech comparing to HMM standards and SVM/HMM. As it is seen the SVM/DBN behaves very well comparing to all the others systems for the four dialects. We see that the recognition rates obtained for all domains in all Arabic dialects by SVM/DBN are the better results for the four systems.

The following resumes the total of average of the recognition rates obtained for the MSA, Gulf, Egyptian and Levantine dialect.

Table 6. Average of Recognition rates with HMM, SVM/HMM, DBN and SVM/DBN systems of multi-dialect continuous speech.

			DBN	SVM/DB
MSA	71.17 %	78.54%	87.86 %	93.46%
Gulf	71.45 %	79.67%	88.47 %	93.93%
Egyp-	72.59 %	83.44%	88.88 %	92.58%
Levan-	77.74 %	81.66%	87.33 %	92.13%
Aver-	**73.24 %**	**80.83%**	**88.13 %**	**93.02%**

We conclude from table 6 that the hybrid model SVM/DBN obtains a good effectiveness and performance comparing to DBN system experimented on the same conditions of learning and testing.

As illustrated in the previous tables, the recognition rates of the multi-dialect Arabic triphones-based continuous speech obtained by the system of HMM are the lowest.

As it is seen, DBN behave well, although the hybrid system SVM/DBN seems more efficient than the last one. It obtains the best recognition rate for the 4 dialects. That's why we evaluate the gain obtained by SVM/DBN by reporting HMM standards, SVM/HMM and DBN.

Thus, compared to the performance of the others systems we realize that there is an improvement of the recognition rates of Arabic triphones-based continuous speech difference between them i.e. the recognition rates of Arabic triphones-based continuous speech with SVM/DBN are bigger than those obtained by the HMM standards, SVM/HMM and DBN which proves the effectiveness of the hybridization of the SVM with DBN.

VI. CONCLUSION

In this paper, we have presented a hybrid ASR system SVM/DBN applied to the recognition of Arabic triphones-based continuous speech. To improve the performance of the SVM/DBN model, we have presented a comparison of the recognition rate of Arabic multi-dialect triphones-based continuous speech using consecutively: HMM standards, SVM/HMM, DBN and SVM/DBN which is our proposed work. The results of multi-dialect Arabic triphones-based continuous speech recognition obtained by the hybrid model SVM/DBN compared to those obtained with the others systems showed a good effectiveness and performance. In fact, the best results are obtained with the proposed system SVM/DBN where we have achieved 93.02% as an average of recognition rates of 4 Arabic dialects. The speech recognizer was evaluated with multi-dialect Arabic continuous speech corpora MSA corpus and performs at 93.03% as an average of recognition rates of 4 Arabic dialects compared to 73.24%, 80.83% and 88.13 % respectively with triphones mixture-Gaussian HMM, SVM/HMM and DBN systems.

REFERENCES

[1] M.Deviren and K.Daoudi. "Structural learning of Dynamic Bayesian networks in speech recognition".in Eurospeech 2001

[2] G.Zweig . Speech recognition with dynamic Bayesian networks. Ph.D. thesis. university of California. Berkley. 1998

[3] M.Deviren and K.Daoudi. "Continuous speech recognition using structural learning of dynamic Bayesian networks" in EUSIPCO 2002

[4] R.V.Lisdonk. Automatic speech recognition using dynamic Bayesian networks.Ph.D. thesis. Delft uinversity of technology . 2009

[5] T.A.Stephenson. H.Bourlard. S.Bengio and A.C. Morris. "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables". in the 6th International Conference on Spoken Language Processing (ICSLP '00). volume 2. pages 951-954. Beijing. October 2000

[6] E.Zarrouk. Y.Ben Ayed and F.Gargouri. " Dynamic Bayesian networks for Multi-Dialect Arabic isolated words recognition". on proceedings of International conference on Artificial Intelligence and Pattern recognition (AIPR). 2014 Kualambur Malysia. November 2014. .pp159-166.

[7] E.Zarrouk and Y.Ben Ayed. "Hybrid SVM/HMM model for the Arab phonemes recognition". The International Arab Journal of Information Technology (IAJIT) Volume 13 issue no 5. September 2016.pp 574-582.

[8] E.Zarrouk. Y.Ben Ayed and F.Gargouri. "Graphical models for the recognition of Arabic continuous speech based triphones modeling". On 16th IEEE/ACIS International Conference of Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 . pp480-485.

[9] E.Zarrouk. Y.Ben Ayed and F.Gargouri. Hybrid continuous speech recognition systems by HMM. MLP and SVM: a comparative study"

International Journal Of Speech Technology (IJST). September 2014, Volume 17, issue 3, pp 223-233.

[10] Vapnik V.. "Estimation of Dependences Based an Empirical Data." Nouka. Moscow. English translation. Springer Verlog. New York. 1979.

[11] Rabiner L-R., Juang B-H., "Fundamentals of Speech Recognition", Prentice-Hall, 1993.Hermansky H. and Cox L.. "Perceptual linear predictive (PLP) analysis-resynthesis". Proceedings of Eurospeech'91. Genova; pp329-332. 1991.

[12] Markel J. D. and Gray. A. H. JR. "Linear Prediction of Speech". Berlin: Springer-Verlag. 1976

[13] N.Friedman. K.Murphy. and S.Russell. " Learning the structure of dynamic probabilistic networks". Proceedings of the UAI'98. Madison. Wisconsin.139-147

[14] M. A.Peot and R. D.Shachter. Fusion and Propagation with Multiple Observations in Belie Networks. Artificial Intelligence. 48(3). pp299-318.

[15] S. Connel. A Comparison of Hidden Markov Model Features for the Recognition of Cursive Handwriting. Computer Science Department. Michigan State University. MS Thesis (1996).

[16] K. Almemanand M. Lee and A. Almim an, "Multi Dialect Arabic Speech Parallel Corpora", In The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, 12-14 Feb. 2013, IEEE, 2013.pp1-6

Bayesian network classification: Application to Epilepsy type prediction using PET scan data

Badih Ghattas
Department of Mathematics and Statistic
Aix Marseilles University, France
Email: badih.ghattas@univ-amu.fr

Kamel Jebreen
Department of Mathematics and Statistic
Aix Marseilles University, France
Email: jebreen20@yahoo.com

***Abstract*—Different types of Bayesian networks may be used for supervised classification. We combine such approaches together with feature selection and discretization and we show that such combination gives rise to powerful classifiers. A large choice of data sets from the UCI machine learning repository are used in our experiments and an application to Epilepsy type prediction based on PET scan data confirms the efficiency of our approach.**

1. Introduction

Bayesian networks are powerful graphical models for representing the joint distribution of a random vector $\mathbf{X}$. They have been extended to answer the classification task where one wishes to predict the label of a class variable $Y \in \{1, ..., J\}$ having observed a set of explanatory variables $\mathbf{X} = (X_1, ..., X_p)$.

The simplest Bayesian network classifier is the naive Bayes approach (NB) where the components of $\mathbf{X}$ are assumed to be independent given Y. Tree augmented naive Bayes (TAN) [4] is a direct extension of NB where each variable X_i, $j = 1, , , p$, may depend on at most one other variable than Y.

Other approaches use more general and complex unrestricted Bayesian networks for classification. *Unrestricted Bayesian networks* build Bayesian networks over the joint set (Y, $\mathbf{X}$) and classify any instance by estimating the posterior probability $P(Y|\mathbf{X})$ using the network. *Multinet Bayesian networks* build multiple Bayesian networks over the observations corresponding to each label of Y. This gives an estimation of $P(\mathbf{X}, Y)$ and using Bayes rule one may compute $P(Y|\mathbf{X})$.

Feature selection approaches aim to reduce the dimension of the data keeping only the important variables for the classification. Different feature selection approaches have been suggested in the litterature [15], [19]. Such approaches may be embedded in the process of classification, or done as an independent preprocessing. We use the later approaches based on random forests variable importance.

Bayesian networks for continuous data are strongly based on the Gaussian assumption and very sensitive to it. When this assumption is not true it is common either to use Cox- Box transformations or to discretize the data [4].

In this work, we show that combining feature selection and discretization through Bayesian network classifiers gives a powerful approach when compared to other classical machine learning methods for classification like random forests (RF), support vector machines (SVM) and CART (classification and regression trees [1]). The method giving the best result over a wide choice of experiments is the multinets approach.

To illustrate the efficiency of our approach we apply it to PET scanning data (Positron Emission Tomography) obtained for **fifty four** patients suffering from **four** different types of epilepsy. Each of the **thirty seven** variables in the data corresponds to the signal intensity measured at a Region of interest (ROI) in the brain. It is supposed that the connectivity whether it exists or not between the ROIs is very important to identify the class of epilepsy. Graphical models used for classification of such datasets coming from MRI images have shown to be very powerful [18].

This work will be presented as follows: Section 2 give a brief introduction of classification using Bayesian networks: Naive Bayes, tree augmented naive Bayes, unrestricted and multinets Bayesian networks. Section 3 describes the discretization method we used and the feature selection based on random forests. Section 4 describes the experiments and results. Finally, section 5 considers the application to Epilepsy type prediction from PET scan data.

2. Classification using Bayesian network

A Bayesian network denoted $B = (G, \Theta)$ is a directed acyclic graph G with parameter Θ. Each node of G corresponds to a variable X_i, $i = 1, ..., p$. The structure of graph G represents the dependencies between the variables and $\Theta = (\theta_1, ..., \theta_p)$ are the conditional probabilities of each variable X_i over $\pi(X_i)$, the set of its parents in the graph. A Bayesian network assumes that the joint probability of $\mathbf{X} = (X_1, ..., X_p)$ is factorized as follows:

$$P(X_1, ..., X_p) = \prod_{i=1}^{p} P(X_i|\pi(X_i)) \quad (1)$$

Learning a Bayesian network is done through the estimation of the structure of the graph G and the set of parameters Θ from the data set.

Structure learning for Bayesian networks may be done using various algorithms. Some of them aim to maximize a score over the structure (BIC score [22], MDL score [14], [20], [25]) and others identify the existence of connections and their directions using conditional independence tests for instance. In our experiments we used BIC score for learning the structure of Bayesian networks.

The parameters of Bayesian networks may be learned either by maximum likelihood estimation (MLE) or by Bayesian estimation.

In our work, we used MLE for estimating the parameters for the continuous attributes. For example, for a continuous node X_1 having two parents X_2 and X_3, the local parameters for X_1 are defined by the conditional Gaussian model

$$X_1|X_2, X_3 \sim N(\mu_{X_1|X_2,X_3}, \sigma^2_{X_1|X_2,X_3}).$$

Thus, the parameters estimates are the same induced by fitting a linear regression model of X_1 over its parents X_2 and X_3.

For a discrete node, the corresponding conditional distribution is assumed to be multinomial with parameter θ. The Bayesian Maximum a Posteriori approach is used to estimate the parameters that is, the estimate of θ is the one maximizing

$$P(\Theta|\mathbf{X}) \propto P(\mathbf{X}|\Theta)P(\Theta)$$

where $P(\Theta|\mathbf{X})$ is the posterior distribution, $P(\mathbf{X}|\Theta)$ is the multinomial likelihood function and $P(\Theta)$ is the prior distribution taken to be a Dirichlet distribution.

In classification, a discrete variable Y taking values $y \in \{1, ..., J\}$ is also available. One may construct a Bayesian network over $(\mathbf{X}, Y)$ putting a constraint over the graph structure such that Y has no parents.

Equation 1 becomes:

$$P(X_1, ..., X_p, y) = P(y) \prod_{i=1}^{p} P(X_i|\pi(X_i)) \qquad (2)$$

Once a Bayesian network is learned from the data, it may be used to compute any marginal or conditional probability over a subset of the variables in the network. This is called *inference* from the Bayesian network. We present now the different Bayesian networks classifiers used later in our experiments.

2.1. Naive Bayes

Naive Bayes classifier (NB) [16], [17] is a restricted Bayesian network which makes the assumption that the predictor variables are conditionally independent given the class variable (see figure 1). Under this assumption the joint probability distribution is given by:

$$P(y, X_1, ..., X_p) = P(y) \prod_{i=1}^{p} P(X_i|y)$$

and the class prediction for Y may be computed using

$$\underset{y}{argmax}\ P(y|X_1, ..., X_p) = \underset{y}{argmax}\ P(y)P(X_1, ..., X_p|y)$$

where $P(X_1, ..., X_p|y)$ is inferred from the network. The number of parameters in NB is much lower than in unrestricted Bayesian networks.

2.2. Tree Augmented Naive Bayes (TAN)

TAN [4] is another type of restricted Bayesian network classifier which takes into account the correlation between predictor variables. It allows each variable in the network to have at most one other parent than Y. So, the factorization for joint probability distribution under TAN assumption is given by

$$P(X_1, .., X_p, y) = P(y)P(X_q|y) \prod_{i=1, i\neq q}^{p} P(X_i|y, \pi(X_i))$$

where X_q denotes the root node which is X_3 in figure 2.

Chow and Liu algorithm [2] is updated by [4] to learn the structure of TAN taking into account the class variable by computing the conditional mutual information between each pair of predictor variables given the class as follows:

$$I(X_i, X_j|Y) = \sum_{x_i} \sum_{x_j} \sum_{y} P(x_i, x_j, y) \log \frac{P(x_i, x_j|y)}{P(x_i|y)P(x_j|y)}$$

for $i, j = 1, .., p$ and $i \neq j$. Learning TAN structure starts from a complete undirected graph with edges having weight equal to $I(X_i, X_j|Y)$. Kruskal algorithm [13] is used to obtain the maximum spanning tree from the structure. Finally, the class variable is set to be the parent of all predictor variables and the second parent is chosen randomly among the predictor variables to determine the direction of the edges which agree with TAN structure restriction. More about Bayesian networks see [9], [10], [12], [23]

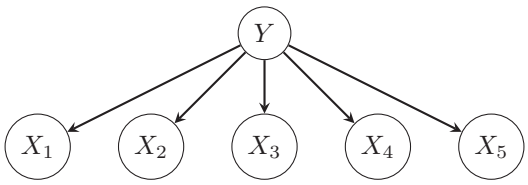


Figure 1: Naive Bayes classifier

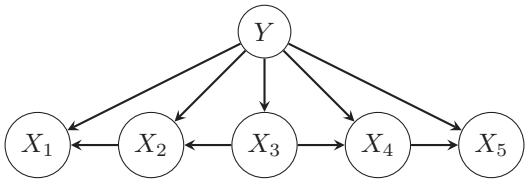


Figure 2: TAN classifier

2.3. Unrestricted Bayesian networks (UBN) and Multinet (MN)

Unrestricted Bayesian networks (UBN) may be used in at least two other ways for classification [2], [4]. First,

a UBN may be learned from a data set using both input and class variable $(\mathbf{X}, Y)$. Once the UBN is learned, the posterior probabilities $P(Y|\mathbf{X})$ may be inferred by likelihood weighting sampling [5], [24]. This consists on having a large sample from the networks and the probabilities by estimating frequencies.

Multinets Bayesian networks consists of estimating J Bayesian networks separately for each label of Y. Each used to estimate the conditional probabilities $P(\mathbf{X}|Y)$. Bayes rule is then used to estimate the posterior probabilities

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{\sum\limits_{x,y} P(\mathbf{X}, Y)P(Y)}$$

.

The idea behind this approach is that the interactions between the input variables $\mathbf{X}$ may be different according to the value of Y [4]

3. Discretization and Feature selection using random forests

Bayesian networks for continuous data are strongly based on the Gaussian assumption. In real data sets, this assumption rare holds and variables may be of both kinds continuous and discrete. That is why we choose to discretize the data.

We also apply a feature selection approach in order to reduce the data dimensions, reduce the complexity cost of models estimation and avoid noise due to irrelevant variables.

3.1. Discretization

We used the ReliefF measure [11] for variables quality estimation to discretize the continuous variables [21]. ReliefF algorithm selects an instance R_i randomly and then searches for the k nearest instances having the same class as R_i called nearest hits $H_j(Y)$ and searches for the k nearest neighbors having different class than R_i called nearest misses $M_j(Y)$. This process is repeated m times to update the quality estimation for the variables, where m is a user defined parameter. The quality estimation for the variable X_i at every iteration is: $W[X_i] = W[X_i] - \frac{-\sum\limits_{j=1}^{k} \text{diff}(X_i, R_i, H_j)}{m.k} + \sum\limits_{Y \neq class(R_i)} \left[\frac{P(Y)}{1-P(class(R_i))} \sum\limits_{j=1}^{k} \text{diff}(X_i, R_i, M_j(Y)) \right] /(m.k)$ where $W[X_i]$ is initialized to zero, $\forall\ i = 1, .., p$, and $\text{diff}(X_i, t_1, t_2)$ computes the difference between the values of the attribute X_i for instances t_1 and t_2. We use the measures

$$\text{diff}(X_i, t_1, t_2) = \left\{ \begin{array}{ll} 0, & value(X_i, t_1) = value(X_i, t_2) \\ 1, & value(X_i, t_1) \neq value(X_i, t_2) \end{array} \right\},$$

if the variables are nominal.

And

$$\text{diff}(X_i, t_1, t_2) = \frac{|value(X_i, t_1) - value(X_i, t_2)|}{max(X_i) - min(X_i)},$$

if the variables are continuous.

To discretize the attribute X_i, the ReliefF measure is used with a greedy search algorithm (see algorithm 1) to find the split point that maximizes the heuristic measure $W[X_i]$ [21]. At each iteration the algorithm searches for the new split point maximizes the heuristic estimate of the discretized variable [11].

Algorithm 1 Greedy algorithm searching for the split points that maximize the ReliefF measure.

```
Best Discretization = {}
Set of split point = {}
repeat m times
    if Set of split points is best so far then
        Best discretization =Set of split points
    end if
until the heuristic search is worse than the previous step.
```

3.2. Feature selection using random forests

Random forests designed among the most known and powerful classification models. They combine a large number of trees trained over bootstrap samples of the original data set.

Random forests are particularly attractive because they offer a very original variable importance measure which has been widely analyzed in the literature and proved to be very efficient in a large number of situations [3].

We use variable importance assessed by random forests in order to select the best subset of variables for the classification task. This is done following the idea given by [7] and summarized in algorithm 2.

For each variable we compute its importance by averaging over 100 runs of random forests. Variables are ranked in the decreasing order of their importance and introduced sequentially in an embedded increasing random forests model. The accuracy of each model is estimated by ten cross-validation and the optimal number of important variables to retain is the one corresponding to the most accurate model.

4. Experimental methodology and results

In this section, we compared the efficiency of Bayesian network classifiers (NB, TAN, unrestricted Bayesian networks (UBN) and multinets Bayesian networks (MN) to other classical methods like SVM, random forests (RF) and decision trees (CART). Support vector machines depend on two parameters β and $C = cost$ which are the kernel parameter and the constant of the regularization term in the Lagrange formulation respectively. These parameters are tunned and compared with their default values, $\beta = 1/p$

Algorithm 2 Random forests feature selection.

let D be data set and p is the number of features.
for (i=1:100) **do**
  $VI = \frac{\sum_{i=1}^{100} VI_i}{100}$, where VI is the **V**ariables **I**mportance vector,
end for
Order VI descending: $X^{(1)}, ..., X^{(k)}$
Partition D through stratified cross validation: $D_1, ..., D_{10}$.
for (j=1:10) **do**
  for (k=1:p) **do**
    $M_j^k = f(X^{(1)}, ..., X^{(k),D_{-j}})$
    $Error_j^k = Test(M_j^k, D_j)$
    $Error^k = \frac{1}{10} \sum_{j=1}^{10} Error_j^k$
  end for
end for
$kopt = Argmin_k \{Error^k\}$, where $kopt$ is the threshold point.

and $C = 1$ to choose the best performance. The range of β and C are chosen respectively to be $10^{-6:-1}$ and $10^{1:4}$. For random forests we choose the default values of the parameters suggested in R packages.

The experiments are done over thirteen data sets from the machine learning UCI repository. As mentioned in previous sections the structure is learned using BIC score, and learning parameters are computed using MLE for continuous data sets and using Bayesian estimation for discrete data sets. A short description of these data sets is given in table 1. First, all the explanatory variables are discretized using the ReliefF measure. Important variables are computed using random forests by fixing the number of tree to `ntree=5001` to ensure the stability of variables importance and it is averagely computed over 100 times as shown in algorithm 2.

To assess the accuracy of the classifiers we compute the average misclassification errors (**MCE**) using five folds cross-validation. Cross-validation is run fifty times in each case and the average over these runs is reported.

Data sets	# Instances	# Variables	# classes	# Important variables in discrete case	# Important variables in continuous case
toys	100	50	2	8	4
breast	683	9	2	6	7
glass	214	9	6	9	7
wine	178	13	3	9	8
vehicle	846	18	4	12	8
pima	768	8	2	3	6
satimage	4435	36	6	35	32
segment	2310	19	7	19	7
vowel	990	10	11	10	10
waveform	5000	40	3	21	30
landsat	6435	36	6	36	28
pendigits	10992	16	10	16	16
letter	20000	16	26	16	15

TABLE 1: Data sets description.

Table 1 gives a summary of the data sets used in the experiments including the number of instances, number of variables, number of labels, and the number of important variables retained by feature selection for both the discretized and continuous versions of data sets. Table 2 and Table 3 give the missclassification errors of all the compared models for the original continuous data sets and for their discretized version respectively. In both cases MCE are reported for each data set and its reduced version by feature selection.

Data sets	SVM	CART	RF	NB	UBN	MN
toys	1.50	12.45	6.35	7.39	3.07	40.94
toys-R	0.04	12.14	1.40	2.40	1.19	0.28
breast	3.05	5.31	2.84	3.82	4.60	4.65
breast-R	3.02	5.27	2.95	3.36	4.61	4.62
glass	28.66	31.06	20.89	61.23	45.03	50.28
glass-R	28.15	30.53	20.53	55.89	45.84	51.25
wine	1.79	11.93	1.98	2.69	1.16	0.76
wine-R	1.71	11.58	2.10	2.90	1.29	1.28
vehicle	14.82	31.87	24.86	54.24	15.59	16.01
vehicle-R	21.20	32.06	25.22	50.11	25.52	25.97
pima	22.9	25.67	23.56	24.62	25.05	26.06
pima-R	23.29	25.79	24.10	24.36	24.54	24.99
satimage	8.14	18.96	8.82	20.35	14.54	14.58
satimage-R	8.33	18.91	8.86	20.37	14.54	14.57
segment	3.13	8.09	2.15	20.29	11.84	85.72
segment-R	2.64	8.15	1.62	11.26	7.70	7.56
vowel	1.14	39.83	4.00	33.24	16.29	14.95
vowel-R	1.14	39.83	4.00	33.24	16.29	14.95
waveform	13.61	26.55	14.39	20.01	14.71	14.78
waveform-R	13.36	26.55	14.23	20.02	14.66	14.63
landsat	7.77	18.87	8.24	20.39	14.75	14.60
landsat-R	8.32	18.88	8.30	20.33	14.47	14.51
pendigits	0.38	10.15	0.14	12.01	1.00	24.99
pendigits-R	0.38	10.15	0.14	12.01	1.00	24.99
letter	2.51	60.01	3.70	41.57	12.69	13.94
letter-R	2.61	51.70	3.18	32.60	10.22	10.11

TABLE 2: Experimental results with five CV (averaged over fifty runs). Data are continuous and "R" denotes the reduced data set by feature selection.

Table 4 shows the differences of MCE between the continuous and the discrete case, for both reduced and not reduced data sets. Whereas discretization does not contribute to increase the performance of the classical machine learning approaches, its contribution for BN classifiers mainly, Naive Bayes and Multinet are very significant in most of the cases.

To see clearly whether the feature selection procedure gives rise to better models, we compute the difference of MCE before and after feature selection. These differences are reported in table 5 and table 6 for the continuous and discrete case respectively. We can see that despite of the unrestricted BN approach, in most cases we gain in accuracy when performing feature selection. The gain is higher in general for the discretized version of the data sets.

5. Application of Bayesian classifier on PET scan data

Epileptic patients are followed by brain PET scan imaging. The images are segmented according to predefined anatomical regions (variables) in the brain. Thirty seven regions were considered for fifty four patients followed at the Timon hospital, Marseilles, France. Each patient belongs to one of the four categories of Epilepsy: BILATERAL, LATERAL, MESIAL and PLUS. The distribution of these labels in our sample are: sixteen, seven, seventeen and fourteen respectively. For more details for this data set see [8].

The aim is to predict the Epilepsy category using the intensity measure of the thirty-seven regions of interest

Data sets	SVM	CART	RF	NB	UBN	MN	TAN
toys	2.23	12.07	3.63	2.51	6.21	2.08	7.15
toys-R	2.67	11.99	3.29	2.33	4.35	2.53	4.00
breast	2.21	5.31	2.73	2.49	3.03	2.95	3.13
breast-R	2.20	5.30	2.37	2.47	2.98	2.68	3.19
glass	22.57	30.90	20.49	26.91	34.60	23.97	22.50
glass-R	22.57	30.90	20.49	26.91	34.60	23.97	22.50
wine	1.49	9.90	1.82	0.99	1.62	2.09	2.23
wine-R	1.67	9.94	1.63	1.12	1.67	1.89	1.79
vehicle	28.84	34.64	29.86	40.62	34.30	29.07	29.35
vehicle-R	26.65	34.27	27.50	38.97	37.64	29.05	29.53
pima	23.34	24.83	23.64	24.88	24.41	24.60	25.40
pima-R	23.14	23.79	21.82	22.30	22.29	22.33	23.65
satimage	10.4	19.06	10.43	19.91	19.01	14.00	13.70
satimage-R	10.46	19.05	10.43	19.93	19.72	13.95	13.46
segment	4.48	11.23	4.47	9.58	6.74	6.53	5.97
segment-R	4.48	11.23	4.47	9.58	6.74	6.53	5.97
vowel	13.64	45.51	13.00	35.01	50.85	26.98	24.71
vowel-R	13.64	45.51	13.00	35.01	50.85	26.98	24.71
waveform	17.36	26.93	18.14	20.43	20.52	21.15	21.14
waveform-R	17.05	26.93	17.77	20.39	20.96	21.11	20.51
landsat	10.00	19.19	9.87	20.26	18.77	13.66	13.38
landsat-R	10.00	19.19	9.87	20.26	18.77	13.66	13.38
pendigits	1.52	20.60	1.77	14.24	6.41	3.09	5.17
pendigits-R	1.52	20.60	1.77	14.24	6.41	3.09	5.17
letter	8.03	53.36	7.59	33.92	24.34	18.20	23.14
letter-R	8.03	53.36	7.59	33.92	24.34	18.20	23.14

TABLE 3: Experimental Results with five CV (averaged over fifty runs). Data are discretized and "R" denotes the reduced data set by features selection.

Data sets	SVM	CART	RF	NB	UBN	MN
toys	−0.73	0.38	2.72	4.88	−3.14	38.86
toys-R	−2.63	0.15	−1.89	0.07	−3.16	−2.25
breast	0.84	0.00	0.11	1.33	1.57	1.70
breast-R	0.82	−0.03	0.58	0.89	1.63	1.94
glass	6.09	0.16	0.40	34.32	10.43	26.31
glass-R	5.58	−0.37	0.04	28.98	11.24	27.28
wine	0.30	2.03	0.16	1.70	−0.46	−1.33
wine-R	0.04	1.64	0.47	1.78	−0.38	−0.61
vehicle	−14.02	−2.77	−5.00	13.62	−18.71	−13.06
vehicle-R	−5.45	−2.21	−2.28	11.14	−12.12	−3.08
pima	−0.44	0.84	−0.08	−0.26	0.64	1.46
pima-R	0.15	2.00	2.28	2.06	2.25	2.66
satimage	−2.26	−0.10	−1.61	0.44	−4.47	0.58
satimage-R	−2.13	−0.14	−1.57	0.44	−5.18	0.62
segment	−1.35	−3.14	−2.32	10.71	5.10	79.19
segment-R	−1.84	−3.08	−2.85	1.68	0.96	1.03
vowel	−12.5	−5.68	−9.00	−1.77	−34.56	−12.03
vowel-R	−12.5	−5.68	−9.00	−1.77	−34.56	−12.03
waveform	−3.75	−0.38	−3.75	−0.42	−5.81	−6.37
waveform-R	−3.69	−0.38	−3.54	−0.37	−6.30	−6.48
landsat	−2.23	−0.32	−1.63	0.13	−4.02	0.94
landsat-R	−1.68	−0.31	−1.57	0.07	−4.30	0.85
pendigits	−1.14	−10.45	−1.63	−2.23	−5.41	21.90
pendigits-R	−1.14	−10.45	−1.63	−2.23	−5.41	21.90
letter	−5.52	6.65	−3.89	7.65	−11.65	−4.26
letter-R	−5.42	−1.66	−4.41	−1.32	−14.12	−8.09

TABLE 4: The difference of MCE between continuous and discrete data.

Data sets	SVM	CART	RF	NB	UBN	MN
toys	1.46	0.31	4.95	4.99	1.88	40.66
breast	0.03	0.04	−0.11	0.46	−0.01	0.03
glass	0.51	0.53	0.36	5.34	−0.81	−0.97
wine	0.08	0.35	−0.12	−0.21	−0.13	−0.52
vehicle	−6.38	−0.19	−0.36	4.13	−9.93	−9.96
pima	−0.39	−0.12	−0.54	0.26	0.51	1.07
satimage	−0.19	0.05	−0.04	−0.02	0.00	0.01
segment	0.49	−0.06	0.53	9.03	4.14	78.16
vowel	0.00	0.00	0.00	0.00	0.00	0.00
waveform	0.25	0.00	0.16	−0.01	0.05	0.15
landsat	−0.55	−0.01	−0.06	0.06	0.28	0.09
pendigits	0.00	0.00	0.00	0.00	0.00	0.00
letter	−0.10	8.31	0.52	8.97	2.47	3.83

TABLE 5: Difference between MCE without and with feature selection for continuous data. Positive values mean that models are more accurate with feature selection.

(ROI). A particular focus in this application is about the complex connectivity of the ROI in the brain. Statistical models should take into account this connectivity. As it may be seen in table 7 and 8 the models fitted over the continuous data set show very poor performance.

We run here the classical methods together with the

Data sets	SVM	CART	RF	NB	UBN	MN	TAN
toys	−0.44	0.08	0.34	0.18	1.86	−0.45	3.15
breast	0.01	0.01	0.36	0.02	0.05	0.27	−0.06
glass	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wine	−0.18	−0.04	0.19	−0.13	−0.05	0.20	0.44
vehicle	2.19	0.37	2.36	1.65	−3.34	0.02	−0.18
pima	0.20	1.04	1.82	2.58	2.12	2.27	1.75
satimage	−0.06	0.01	0.00	−0.02	−0.71	0.05	0.24
segment	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vowel	0.00	0.00	0.00	0.00	0.00	0.00	0.00
waveform	0.31	0.00	0.37	0.04	−0.44	0.04	0.63
landsat	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pendigits	0.00	0.00	0.00	0.00	0.00	0.00	0.00
letter	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 6: Difference between MCE without and with feature selection for discrete data. Positive values mean that models are more accurate with feature selection.

Bayesian classifiers in the previous section using the same discretization and feature selection approaches. Since the sample size is very small and as before the MCE , We report using five folds cross-validation as well as using leave one out (LOO) approach.

Data sets	SVM	CART	RF	NB	UBN	MN	TAN
Epilepsy	31.87	45.67	34.24	43.59	55.22	70.31	–
Epilepsy-R	22.58	43.94	26.10	29.74	36.68	50.52	–
Epilepsy-D	22.03	39.03	20.87	26.75	31.45	26.58	24.25
Epilepsy-D-R	20.25	36.14	18.11	22.03	23.47	17.65	26.03

TABLE 7: Experimental Results with five CV (averaged over fifty runs) for Epilepsy data set, "R" denoted the reduced data set by features selection and "D" denotes the discrete data set.

Data sets	SVM	CART	RF	NB	UBN	MN	TAN
Epilepsy	29.63	42.59	31.48	44.44	59.26	70.37	–
Epilepsy-R	22.22	42.59	25.93	27.78	33.33	50.00	–
Epilepsy-D	24.07	46.30	20.37	25.93	24.07	24.07	24.07
Epilepsy-D-R	18.52	42.59	18.52	22.22	22.22	16.67	27.78

TABLE 8: Experimental Results with LOO for Epilepsy data set, "R" denoted the reduced data set by features selection and "D" denotes the discrete data set.

For all models, feature selection applied to the Epilepsy data decreases significantly their MCE. SVM has the best performance with MCE equal to 31.87% and 22.58% over non-reduced and reduced data respectively in their continuous version.

Except for SVM, MCE are reduced very significantly when the data set is discretized. Finally, using feature selection, MCE is yet reduced reaching 17.65% for multinets Bayesian network classifier (whereas with the original continuous data set it had 70.31%). LOO estimation for MCE are yet lower but they show the same patterns.

6. Conclusion

In this paper, we have shown that Bayesian networks classifiers are very accurate models when compared to other classical machine learning methods. Discretizing input variables often increases the performance of Bayesian networks classifiers, so does a feature selection procedure. This is probably due to the fact that discrete Bayesian networks are less sensitive to the underlying distribution of the data and

are easier to estimate in low dimensions. One specific advantage of Bayesian networks classifiers is that they directly estimate the distribution of the data and take into account the high order interactions between the variables. The models may be also graphically presented.

Future work aims to use a more specific discretization approach preserving the dependencies structure within the original data and including latent variables accounting for hidden clusters in the data.

7. Acknowledgment

The authors are grateful to Eric Guedj from the nuclear medicine service of La Timone hospital for the data he provided. This work was partilay supported by the ECOS-SUD project N^0 $U14E02$.

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

[2] Chow, C. and Liu, C. (1968). *Approximating discrete probability distributions with dependency trees*. IEEE Transactions on Information Theory **14**(3): 462-467.

[3] Diaz-Uriarte, R., Alvarez de Andres, S. (2006) *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 7:3.

[4] Friedman, N. (1997) *Bayesian Network Classifiers*. Machine Learning. **29**: 131–163.

[5] Fung, R., and Chang, K.-C. (1989). *Weighting and integrating evidence for stochastic simulation in Bayesian networks.* Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence, 209–220.

[6] Geiger, D. and Heckerman, D. (1996). *Knowledge representation and inference in similarity networks and Bayesian multinets*. Artificial Intelligence **82**: 45-74.

[7] Ghattas, B. and Oppenheim, G. (2011). *Etude de faisabilit: Modles globaux pour la mise au point moteur. Rapport technique Renault*.

[8] Guedj, E., Bonini, F., Gavaret, M., Trebuchon, A., Aubert, S., Boucekine, M., Boyer, L., Carron, R., McGonigal, A. and Bartolomei, F. (2015). *18FDG-PET in different subtypes of temporal lobe epilepsy: SEEG validation and predictive value*. Epilepsia **56**(3): 414-21.

[9] Heckerman, D, Geiger, D. Chickering, D. (1995) *Learning Bayesian networks: the combination of knowledge and statistical data*. Mach Learn **20(3)**: 197-243. Available as Technical Report MSR-TR-94-09.

[10] Koller, D. and Friedman, N. (2009). probabilistic graphical models principles and techniques. MIT press.

[11] Kononenko,I. (1994). *Estimating attributes: analysis and extension of Relief*. ECML-94 Proceedings of the European conference on machine learning on Machine Learning, 171–182.

[12] Korb, K. and Nicholson, AE. (2010). *Bayesian Artificial Intelligence*. Chapman and Hall.

[13] Kruskal, J. B. (1956). *On the shortest spanning subtree of a graph and the traveling salesman problem*. Proceedings of the American Mathematical Society **7**(1): 48-50.

[14] Lam, W. and Bacchus, F. (1994). *Learning Bayesian belief networks. An approach based on the MDL principle.* Computational Intelligence. **10**: 269–293.

[15] Langley, P. and Sage, S. (1994). *Induction of selective Bayesian classifiers*. In: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence. 399-406.

[16] Maron, .M and Kuhns, J. (1960). *On relevance, probabilistic indexing, and information retrieval.* Journal of the Association for Computing Machinery. **7**: 216-244.

[17] Minsky, M. (1961). *Steps toward artificial intelligence. Transactions on Institute of Radio Engineers* **49**: 8-30.

[18] Mumford, J.A. and Ramsy, J.D. (2014) *Bayesian network for fMRI: Aprimer.* NeuroImage, **86**, 537–582.

[19] Pazzani, M. and Billsus, D. (1997). *Learning and revising user profiles: the identification of interesting web sites*. Machine Learning. **27**: 313-331.

[20] Rissanen, .J. (2007). *Information and complexity in statistical models*. Springer, New York.

[21] Robnik,M. and Kononeko,I. (1995). Discretization of continuous attributes using ReliefF. In proceeding of ERK-95.

[22] Schwarz, G. (1978). *Estimating the dimension of a model.* Annals of Statistics **6**: 46–464.

[23] Scutari, M. and S. and Denis, J. (2015). *Bayesian Networks with Example in R*. Chapman and Hall.

[24] Shachter, R., and Peot, M. (1989). *Simulation approaches to general probabilistic inference on belief networks*. Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence, 22–234.

[25] Suzuki, J. (1993). *A construction of Bayesian networks from databases based on an MDL scheme. In D. Heckerman and A. Mamdani (Eds.)*, Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence. (pp.266-273). San Francisco, CA: Morgan Kaufman.

Bayesian methods for opinion dynamics modeling in social networks

Ons Abid, Salma Jammousi
Multimedia InfoRmation systems and Advanced
Computing Laboratory, MIRACL
Sfax University, Tunisia
abid.ons26@gmail.com, salma.jammoussi@isimsf.rnu.tn

Yassine Ben Ayed
Multimedia InfoRmation systems and Advanced
Computing Laboratory, MIRACL
Sfax University, Tunisia
yassine.benayed@gmail.com

***Abstract*— how do opinions form and evolve in a social network is an interesting subject and has attracted much attention of researchers from various disciplines. One of the ways of thinking, opinion can be formed through a process called informational influence, where a user forms his/her opinion, on various issues, according to the information they obtained or observed from certain number of agents in his friendship neighbourhood. In this paper, we consider the problem of modeling how users form and update opinions based on their neighbours' opinions, their decisions, and their privates' information.**

The main goal of this paper is to explore various work of modeling opinion dynamics in social networks where we focus on the implications of the form of learning (e.g., Bayesian vs. non-Bayesian). In this study, we try to give a serve as guidelines for scientists, practitioners and developers who intend to design new methods in this area. Non-Bayesian models

Keywords—Opinion dynamics; Bayesian models; Non-Byaesian models; Social Networks

I. INTRODUCTION

In recent years, both social network and social media have become ubiquitous in our daily life and allow hundreds of millions of Internet users worldwide to produce and consume content. From where, they have become a global favorite for discussion of topics, ideas and news: about the latest breaking news, political issues sports events, new products, life style or celebrities. . .

Online social networks have offered an incredible platform for information exchange and have proved to be very powerful in many situations, such as Facebook during the 2010 Arab spring [1] or Twitter during the 2008 U.S. presidential elections [2] for instance. Various contents can be exchanged between Internet users such as photos, videos and opinions, concerning many issues.

A lot of research have been done during the past decade for understand how opinions change following to social influence. This last lies at the heart of individual's opinion formation because users may form or update their opinion about a particular topic by learning from the information and opinions that their friends/neighbours share. It is crucial to study how opinions are formed and are evaluated over time and how opinions change following the social interactions.

So, in addition to this introductory section, the paper is organized in three distinct but complementary sections:

- In section two, we present the key components in opinion formation proposed by Acemoglu and Ozdaglar [4].
-An overview on the state of the art of opinion dynamics modeling methods will be presented in the third section. We highlight two alternative approaches of social learning and opinion formation. In the first sub-section, we start with Non-Bayesian models of communication learning. We study the evolution of opinions in a society where agents, instead of performing Bayesian updates, apply a simple learning rule to incorporate the views of individuals in their social. In the second sub-section, we turn to Bayesian social learning by a set of agents observing others' actions.
- Conclusion is given in final section.

II. OPINION DYNAMICS

Since a long time, opinion dynamics is an active area of research among statistical physicists, mathematicians and computer scientists. The first researcher, who started in this area, proposes to create a simple and intuitive model of interactions between people and information. This model was proposed by a French psychologist in [37]. Before, due to limitations in communication, conventional methods do not treat the dynamics of opinion on large groups of people. Thanks to the development of the Internet people, we can exchange information and ideas more freely and frequently. Therefore, a calculation model will be a necessary element in the process of the dynamics of opinion.

Building a model of opinion dynamics that is consistent with existing social theories and capable of handling large problem scalability is a challenging issue. For reasons of simplification and extraction of the core parts of a difficult problem, Acemoglu and Ozdaglar [4] proposed to divide the formation of opinion into three key components: Prior opinions, Method of information processing and Information source.

A. Prior opinions:

Opinion dynamics is a continuous procedure which contains a sequence of different opinions. Any model of opinion formation has to start with some types of initial opinions (priors), determined by the knowledge base, which can be personal or public.

B. Information sources:

One critical component in opinion dynamics is information sources. By interacting with sources of information, a person can update their knowledge bases and his/her opinion based on new information that he/she receives. This might come from observing others' actions and experiences, or from communicating with others.

C. Method of information processing:

It presents how the individual will combine her initial opinions and the information he/she receives. There are models that similarly combine priors and information to yield a new opinion and other models use the Bayes rule. Various factors are responsible for the way in which agents adopt the opinions of others such as: relationship closeness [3] to personal and opinion similarity [4]...

III. Literature reviews of opinion dynamics modeling methods

In everyday life, people form opinions over various issues such as economic, political, social...

However, the relevant information for such problems is not often concentrated in any source or body of sufficient knowledge. Instead, the data are dispersed throughout a vast network, where each individual observes only his/her personal experience. For learning from other people's experiences, each agent will be motivated to communicate with others or to observe actions of others. Two significant models of information aggregation in networks have been used in the literature namely the Bayesian learning and the Non-Bayesian learning. The fundamental issue, in this context, is to whether a formal learning model is used. Bayesians use Bayes' theorem as a formal learning model whereas non- Bayesians do not appear to use a formal learning model. In effect, non-Bayesians are learning informally.

In this section, we provide an overview of research on opinion dynamics models in social networks. We discuss both Bayesian and non-Bayesian models of social learning.

A. Non-Bayesian Learning

In non-Bayesian approaches, people start to form their opinions by specifying simple rules of thumb. The work provided by French [5] allows him to become the pioneer of this model family. Actually, individuals, in their works, form their new opinions by averaging other people's opinions with whom they have directly communicated (at a time from their social neighborhoods: friends, coworkers or peers). After that, several Non-Bayesian approaches were developed around the same idea: "opinions evolve dynamically over time as a function of their neighbors' opinions". For instance, DeGroot Model [6] proposes to replace the simple average function in French [5] with a weighted mean in order to assess opinion pooling of a dialogue among experts.

The literature has considered several non-Bayesian. Such as Friedkin-Johnsen model, [7],[8] where some level of stubbornness has been added to each agent. The latter is supposed to adhere to its initial opinion or prejudice to some degree. In other words, the agent is stubborn never forget their prejudices, and thus remain persistently influenced by exogenous conditions under which those prejudices were formed. A significant extension of the classical Friedkin-Johnsen model has been proposed by [9], represent the dynamics of agents' opinions on two or more topics, and those topic-specific opinions are interdependent. Another dynamical model [10] named DeGroot-friedkin model, proposes to combine the averaging rule by DeGroot to describe the dynamics of opinions over a single issue and the reflected appraisal mechanism by Friedkin to describe the dynamics of individuals' self-appraisal and social power. The presence of stubborn individuals has received increasing attention and it did not touch only FJ model and its variations but it was considered in a lot of others works [11], [12], [13]. In [11] and [12] carefully the effects of stubborn individuals are investigated in a randomized gossiping process. In [13], the opinion formation process is regarded as a local interaction game and the concept of the stubbornness of an individual regarding his/her initial opinion is introduced.

B. Bayesian learning

The most widely-used approach for dealing with uncertainty and more recently, for opinion formation is the Bayesian approach. The Bayesian learning approaches assume that "individuals would update their beliefs optimally given an underlying model of the world" [14]. It has the advantage of relying on well-understood techniques from probability theory for update the opinions of agents.

Consider a situation in which an individual is trying to form an opinion about some underlying state θ. The state could correspond to some variable (economic variable, social variable, political variable…).

After observing some evidence or receiving some information s, each agent would first have to start with some priors. This is captured by y a function $P(\theta)$, which gives the prior belief of the individual about the likelihood of each possible value of θ.

Let us consider s the information that the individual will receive or this could correspond to some observation concerning θ.

The Bayesian approach then posits that the individual will update his/her prior after observing s according to the Bayes rule, in particular $P(\theta|s)=(P(s|\theta) \times P(\theta))/ P(s)$.

The individual can compute the probability that the true state is θ given the signal s, $P(\theta|s)$, if only if he/she knows $P(s|\theta)$ implies that she has an understanding of what types of signals to expect when the true value is θ, $P(\theta)$ implies that she has priors on each possible value of θ, and P(s).

In most Bayesian models, it is assumed, that individuals have beliefs about $P(\theta)$ and $P(s|\theta)$ that coincide with the true data generating process and with each other's beliefs, and in fact, there is common knowledge that they all share the same

priors. The only uncertainty is about the specific value of θ; there is no uncertainty or doubt about the underlying model of the world, and this plays a central role in the implications of Bayesian models.
There are two models of Bayesian social learning by a set of agents observing others' actions or communicating with each other over a social Network. The literature on learning by communication in groups is somewhat smaller than the observational learning literature. But there are several non-Bayesian models of communication. In this section, we will present some Bayesian models of learning with observation and then with communication

C. Models of Bayesian learning

A sizable literature focuses on Bayesian models of observational learning for several reasons: First, it is difficult to control and measure exactly what is (or is not) communicated by various agents in a more general communication model. So, most models propose to focus on the mechanics of the learning process by restricting communication to observable actions. Second, action models correspond more with models of Bayesian learning on networks. This is already an appropriate comparison.

Bikchandani, Hirshleifer and Welch (1992) [15] and Banerjee (1992)[16] started the literature on learning in situations in which individuals are Bayesian and observe past actions. They analyze a sequential decision model in which each decision maker looks at the decisions made by previous decision makers in taking her own decision. This is rational for her because these other decision makers may have some information that is important for her. In these models of social learning, Banerjee [16] and Bikhchandani, Hirshleifer, and Welch [15], assume a sequence of agents (individuals), indexed by $n \in N$, sequentially make a single decision each. These decisions of all agents are public information. Thus, at date 1, agent 1 chooses an decision a1, based on his private information; at date 2, agent 2 observes the decision chosen by agent 1 and chooses an decision a2 based on his private information and the information revealed by agent 1's decision; at date 3, agent 3 observes the decisions chosen by agents 1 and 2 and chooses an action a3 ...; and so on.

The payoff of agent n depends on an underlying, payoff-relevant state of the world, θ, and on her decision. So, Agent n obtains payoff 1, $\mu_0(x_n, \theta)$, if $x_n=\theta$ and payoff 0 otherwise. To simplify notation, we assume that both values of the underlying state are equally likely, so that $P(\theta = 0) = P(\theta = 1) = 1/2$.

Agent n knows his/her signal s_n and the decisions of previous agents x_1, x_2 ...., $x_{(n-1)}$. He/She chooses action 1 if $P(\theta=1| x_1, x_2 ...., x_{(n-1)}) > P(\theta=0| x_1, x_2 ...., x_{(n-1)})$. If $x_1, x_2 <> \theta$, then $x_n<>\theta$ for all agents.

Bikchandani, Hirshleifer and Welch (1992) [15] say that:
An information cascade occurs when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behaviour of the preceding individual without regard to his own information
After observing two individuals, the third individual is faced with one of three situations: (1) Both predecessors have adopted (2) both have rejected (3) one has adopted and other rejected.
Under these assumptions, Bikchandani et al. [15], calculate the unconditional ex-ante probability of UP cascade, No cascade or a DOWN cascade after two individuals. They calculate also the probability of ending up in the correct cascade after observing two agents
Annamaria and Andrea's [17] model proved that the first k ($k<n$) individuals are not allowed to observe previous decisions, whereas the entire history of decisions is commonly known to the last n–k individuals.

The first k individuals observe only their own signal, and follow their private information for make a decision
In contrast, the N–k individuals should base their decision on both their own signal and all past decisions, thereby choosing the most frequently observed action. Compared to Bikhchandani et al.'s model [15], and Banerjee (1992)[16], this model allows aggregating information in a more accurate way. They present another way to calculate the probability of each event on the basis of precise statistical laws rather than on the basis of earlier possible biased actions as [15] did.

Smith and Sørensen (2000) [18] have shown that, herd behaviour arises infinite time with probability one. Once the proportion of agents choosing a particular action is large enough, the public information in favour of this action outweighs the private information of any single agent. So each subsequent agent _ignores_ his/her own signal and "follows the herd". So, Smith and Sørensen propose to provide a most comprehensive and complete analysis of this environment which they generalize the environment to include a richer set of private signals which they introduced the notion of unbounded private signals. The main result is that when each individual observes all past actions and private beliefs are unbounded, information will be aggregated and the correct action will be chosen asymptotically.
Other work on social learning includes Celen and Kariv in 2004[19], who study Bayesian learning when each individual observes his/her immediate predecessor, Callander and Horner 2009[20], who show that it may be optimal to follow the action of agents that deviate from past average behaviour.

This literature is voluminous and most of papers, among others, [21], [22], [23], [24], [25], [26] maintain the assumption that all past actions are observed.
Using language from the analysis of networks, we can say that they focus on "the full observation network topology".
A key impediment to information aggregation in the most models of Bayesian learning from observations of past actions is the fact that actions do not reflect all of the information that an individual has, where individuals ignore their own information and copy the behaviour of others.
We can note that observational learning is important in many situations, but a large part of information exchange in practice is through communication because we tend to talk to people in our social network. This type of learning is not well captured

by observational models, and instead requires a model of communication.

Several papers in the literature study communication, though typically they use non-Bayesian rule (for example [7] [8] [9] …)

The complexity of updating when individuals share their ex post beliefs, is a major problem faced by most non-Bayesian models. In these non-Bayesian approaches such as [6] and [28], people continue to use the simple rules of thumb, which treats all information as 'new' for updating opinions. In this situation, updating opinions with repetitive information may establish an extreme form of duplication of information.

To overcome this problem, [29] propose to label the information sent between agents. This lets not confuse between new information and previously communicated information. For example, at t=2, agent 1 communicates again with agent 2 if he/she has updated her beliefs with somebody. Else, when agent 2 repeats her original signal, his/her message will not be recorded as an additional piece of information by agent 1.

IV. Conclusion

Most social decisions rely on the information agents gather through communication with friends, neighbours, and by observing the others' actions. In this paper, we have provided an overview of work on modeling and evolving opinion dynamics in social networks where we focus on the implications of the form of learning (Bayesian and non-Bayesian), the sources of information (observation and communication).

V. References

[1] Philip N Howard, Aiden Du_y, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. Opening closed regimes: what was the role of social media during the arab spring? Available at SSRN 2595096, 2011.

[2] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3-4):248_260, 2009.

[3] Robert Axelrod. The dissemination of culture a model with local convergence and global polarization. Journal of con_ict resolution, 41(2):203_226, 1997.

[4] Paul M Sniderman and Sean M Theriault. The structure of political argument and the logic of issue framing. Studies in public opinion: Attitudes, nonattitudes, measurement error, and change, pages 133_65, 2004.

[5] John RP French Jr. A formal theory of social power. Psychological review, 63(3):181, 1956.

[6] Morris H DeGroot. Reaching a consensus. Journal of the American Statistical Association, 69(345):118_121, 1974.

[7] Noah E Friedkin and Eugene C Johnsen. Social influence networks and opinion change. Advances in group processes, 16(1):1_29, 1999.

[8] Noah E Friedkin and Eugene C Johnsen. Social influence network theory: A sociological examination of small group dynamics, volume 33. Cambridge University Press, 2011.

[9] Sergey E Parsegov, Anton V Proskurnikov, Roberto Tempo, and Noah E Friedkin. Novel multidimensional models of opinion dynamics in social networks. arXiv preprint arXiv:1505.04920, 2015.

[10] Peng Jia, Anahita MirTabatabaei, Noah E Friedkin, and Francesco Bullo. Opinion dynamics and the evolution of social power in influence networks. SIAM Review, 75(3):367_397,2015.

[11] Daron Acemoglu, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. Opinion fluctuations and disagreement in social networks. Mathematics of Operations Research, 38(1):1_27, 2013

[12] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary opinion dynamics with stubborn agents. ACM Transactions on Economics and Computation, 1(4):19, 2013.

[13] Javad Ghaderi and R Srikant. Opinion dynamics in social networks: A local interaction game with stubborn agents. In 2013 American Control Conference, pages 1982_1987. IEEE, 2013.

[14] Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. Dynamic Games and Applications, 1(1):3_49, 2011.

[15] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy, pages 992_1026, 1992.

[16] Abhijit V Banerjee. A simple model of herd behavior. The Quarterly Journal of Economics, pages 797_817, 1992.

[17] Annamaria Fiore, Andrea Morone, et al. Is playing alone in the darkness sufficient to prevent informational cascades? Max Planck Inst. for Research into Economic Systems, Strategic Interaction Group, 2005.

[18] Lones Smith and Peter Sørensen. Pathological outcomes of observational learning. Econometrica, 68(2):371_398, 2000.

[19] Bogaçhan Çelen and Shachar Kariv. Observational learning under imperfect information. Games and Economic Behavior, 47(1):72_86, 2004.

[20] Steven Callander, Johannes Horner, et al. The wisdom of the minority. WP Northwestern University, 2006.

[21] Ivo Welch. Sequential sales, learning, and cascades. The Journal of finance, 47(2):695_732, 1992.

[22] In Ho Lee. On the convergence of informational cascades. Journal of Economic theory, 61(2):395_411, 1993.

[23] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. The Review of Economic Studies, 78(4):1201_1236, 2011.

[24] Ilan Lobel, Munther Dahleh, Daron Acemoglu, and Asuman E Ozdaglar. Bayesian learning in social networks. NYU, Stern, Center for Digital Economy Research Vol, 2009.

[25] Lones Smith and Peter Norman Sorensen. Rational social learning by random sampling. Available at SSRN 1138095, 2013.

[26] Manuel Mueller-Frank. A general framework for rational learning in social networks. Theoretical Economics, 8(1):1_40, 2013.

[27] David Bindel, Jon Kleinberg, and Sigal Oren. How bad is forming your own opinion? Games and Economic Behavior, 92:248_265, 2015.

[28] Peter M DeMarzo, Je_rey Zwiebel, and Dimitri Vayanos. Persuasion bias, social inuence, and uni-dimensional opinions. Social Influence, and Uni-Dimensional Opinions (November 2001). MIT Sloan Working Paper, (4339-01), 2001.

[29] Daron Acemoglu, Kostas Bimpikis, and Asuman Ozdaglar. Dynamics of information exchange in endogenous social networks. Theoretical Economics, 9(1):41_97, 2014.

Decision Support System Based on Dynamic Bayesian Networks:

Application to optimize the irrigation scheduling

Mohamed Ali Fourati; Anas Kamoun
Renewable Energies and Electric Vehicles Laboratory
University of Sfax, National School of
Engineers (ENIS) Sfax, Tunisia
mohamed.ali.fourati@gmail.com

Mounir Ben Ayed
REGIM: Research Groups in Intelligent Machines
University of Sfax, National School of
Engineers (ENIS) Sfax, Tunisia
mounir.benayed@ieee.org

***Abstract*—In the field of water resources management, the prediction and early determination of climatic measures is one of the most promising ways to control irrigation tasks and optimize the utility profits. In fact, the irrigation system is subject of unpredictable and external factors especially related to climate and atmosphere (temperature, rains, winds …) that have to be rightly taken in the quantities decision making.**

In this study, we will focus on a set of meteorological observations (temperature, humidity, pressure, radiation …) which reflect the temporal reasoning within a decision making process. We propose to set up a process of knowledge extraction based on Bayesian networks from multidimensional, temporal and progressive data. This system has to make dynamic decision on temporal data in order to assure the classification of new climatic states and the prediction of specific measures (such rains) that may affect the water amount to be delivered for irrigation. We are interested to the Dynamic Decision Support Systems in water resources domain (IDSS).

The objective of this work is to reveal the utility of Dynamic Bayesian Networks in the optimization of the water consumption and performing of the irrigation scheduling

Keywords— prediction, meteorological observations, irrigation scheduling, knowledge extraction, dynamic Bayesian networks, decision support systems

I. INTRODUCTION

Nowadays, water resource has been considered an essential production factor in agriculture, for crops [1]. In fact, with the increasing water scarcity, there is a need to optimize water use, mainly for irrigation purposes [2].

The task of water conserving in irrigated agriculture can be supposed as a predictive problem [3]. The reason is that some factors related to climate and atmosphere (temperature, rains, winds …) are well integrated in the decision making process of the irrigation quantities determination and then have to be rightly exploited in order to reduce losses, optimize the consumption and improve the utility profits [4]

Given the hazardous and temporal aspects of the decision process, a predictive approach should be adopted to assure the estimation of specific measures (such rains) that may affect the water amount to be delivered for irrigation. We propose to set up a process of knowledge extraction based on Bayesian networks from multidimensional, temporal and progressive data.

Literature works had reported various predictive solutions for the resolution of the main irrigation problems. Useful studies of [5, 6] have proved efficiency of Fuzzy techniques to estimate irrigation water needs. Other contributions of [7, 8, 9] proposed to apply Artificial Neural Networks to estimate the soil moisture distribution in order to control irrigation amounts determination.

In this paper, we propose to adopt the Dynamic Bayesian Networks technique for classification of climatic states relatively to available parameters. The objective of this work is to reveal the utility of Dynamic Bayesian Networks in the optimization of the water consumption and the improvement of agricultural efficiency under conditions which water is a key limiting factor in crop production.

The paper is structured as two main sections. The first one concerns a detailed technical presentation of the dynamic Bayesian Networks. The second section describes the empirical use case analysis followed by relative results and main conclusion.

II. BAYESIAN NETWORKS

A. Bayesian process and decision making

The uncertainty fact is well supported by the decision making process. While solutions are generated, potential events have to be estimated based on probabilistic measures. Various approaches have been proposed in the literature in order to select best estimations. Decision trees are the common adopted presentations to resolve such problems [10]. Others as Maximum Expected Utility aims to select the alternative of maximum score utility [11]. The Bayesian theory allows to present probabilistic reports between observed variables which is well adopted to uncertainty resolution.

The Bayesian networks present a probabilistic formalism of reasoning proposed by [12] and lately developed by [13]. The Bayesian graphical modelling presents a probabilistic distribution between a large collections of random variables. They are currently used in artificial intelligence and automatic

learning that's avoiding the manipulation of large data structures and limiting the number of estimated variables [14, 15].

A Bayesian network is a causal oriented acyclic graph (DAG) that allows presenting dependencies between random variables. It is considered as a combination between probabilistic and graphical theories that adopts the conditional probability distribution between interconnected nodes. [13, 16]

The conditional probability distribution in a Bayesian network is referred to the Bayes' theorem. Let X and Y two sets of random variables, the Bayes' theorem determines the conditional probability of X given Y based on the probabilities of X, Y and Y given X, with the following formula:

$$P(X|Y)=[P(Y|X)*P(X)]|P(Y) \quad (1)$$

Given:

- There is no period after the "et" in the Latin abbreviation "et al."
- P (X) and P (Y) are a priori probabilities respectively of X and Y
- P (X|Y) is a posteriori probability of X given Y
- If Y is observed, P (Y|X) is named as the likelihood function of X

This formula can be expressed anywhere as: posteriori Probability = likelihood * priori Probability

Therefore, the probability distribution (or inference) is determined by the multiplication of local conditional local probabilities of each node of the network. Let considering a random variable Xi and Pa(Xi) the parents' nodes of Xi, the probability distribution of S = {X1 … Xn} is expressed as:

$$P(X1 \dots XN) = \prod_{(1 \to N)} P(Xi|Pa(Xi)) \quad (2)$$

The estimation of the a priori distribution assures the parameter learning of a Bayesian network by the determination of the conditional probability table (CPT) for every node of the graph. Every CPT is presented with the set {P (Xi | Pa (Xi) = j) 1<=i<=n} where Pa (Xi): Xi's parents' probability

B. Static structural learning of Bayesian network

In most cases, Bayesian network structure is delivered a priori from an expert due to the complexity of this problem (NP-complete problem). This complexity is related to the size of the research space that grows exponentially with the size of the Bayesian network (number of nodes).

In order to resolve the complexity problem, several methods of structural learning have been developed that aim at limiting the research space of the DAG (Directed Acyclic Graph) to more restricted subspace. These methods are based on:

- conditional dependency research between variables
- score calculating functions

The idea of constraint-based structure learning is to identify and perform the properties of conditional independence among the variables in the data and then to build a network that exhibits the observed dependencies and independencies and satisfies constraints. Various algorithms have been proposed to resolve the issue as Statistical hypothesis test (X^2 and G2 tests) and SGS variations as PC & IC [17, 18]

The second approach of score considers learning as an optimization problem. The optimization approach computes the optimal structure with the highest statistically motivated and maximized score. They are exploited by several algorithms for structure learning such Maximum Weighted Spanning Tree (MWST), K2, algorithm, Greedy search algorithm (GS), Simulated Annealing algorithm (SA).

C. Dynamic Bayesian networks

Static Bayesian networks cannot support temporal relations between random variables. In that case, the built structure is supposed as weak model that may lead to volatile prediction and wrong alerts. It is considered then, as an essential task to build correspondence between anterior and posterior temporal events and evaluate Bayesian static to dynamic models

For the reason, [19, 20] had proposed to integrate a dynamic stochastic process into the classical structural and parameter Bayesian learning. Every variable is associated to temporal slice noted Xt. It consists on periodic unroll of two temporal slices of the network '2TBN'. A dynamic Bayesian network (DBN) is special Bayesian Network BN that uses probabilistic transition between time slices. The objective of DBN is to predict the happening of future event given its related past observations.

The DBN formalize the probability distribution of a set of temporal variables X[t] = {X1[t] … XN[t]} as:

$$P(X1 \dots XN) = \prod_{(i:1 \to N)} \prod_{(t:1 \to T)} P(Xi[t]|Pa(Xi)[t]) \quad (2)$$

Where: Pa (Xi[t]) parents of Xi[t]

In the dynamic context, most of the structure learning methods for DBN are based on the usual score-based algorithms (such as greedy search (GS)). Learning techniques of DBNs are extensions from those of static RBs, decomposed in two independent phases:

- intra-slice: Learning the static BN structure with a static dataset corresponding to X [t = 0]
- Inter-slice: Learning the transition variable structure with another "static" dataset corresponding to all the transitions X[t] ∪ X[t + 1] since for each node in slice t, we must choose its parents from slice t-1

III. EXPERIMENTAL STUDY

We aim in this experiment to predict the precipitation likelihood of one given day in order to decide whether to irrigate or not. The application supports dynamic decisions on temporal data in order to assure the classification of new climatic states and predict of rains' measures that may affect the water amount to be delivered for irrigation.

A. *Data and Variables*

We dispose a meteorological data base that contains hourly measures of temperature, humidity, rains, air pressure, wind speed and direction and the solar radiation of 2015 year. These temporal variables consist on a large collection of time series. They are a set of hourly sequentially recorded values and a succession of couples < (v1, t1), (v2, t2)…, (vi, ti)…> where vi is a real number or a vector of values taken at a moment ti. These temporal data have complex structure of two-dimensional presentation that every measure corresponds to one time slice. We dispose 8723*7 records (climatic measure) to be prepared for the knowledge extraction. They are irregular and noisy, but they are pertinent and highly relevant for the classification task. The following table I describes the useful temporal variables.

TABLE I. TEMPORAL VARIABLES

Acronyms	Description
T	air temperature
H	air humidity
SR	global solar radiation
WS	wind speed
WD	wind direction
AP	air pressure
R	quantity of rains

Our objective is to generate the DBN where nodes are extracted from these variables. The learning phase of our dynamic knowledge model is followed by a Bayesian inference process that allows us to represent relationships between these observed variables in a probabilistic way which is well adapted to the uncertainty inherent to the rains' prediction issue

B. *Learning of temporal structure*

Our Bayesian system is developing and learns at every slice during the period test. It has the same structure at every time (Time-invariant) where variables are interdepended with intra-slices arcs. We have adopted the naïve Bayes classifier for the static structure since all variables are independent and directly connected to the result node (predicted rains quantities). The application of the Bayes classifier generates a causal graph of interdependent variables as shown in the figure 1

The inter-slice arcs present the temporal interdependence of the same variable between two successive and different time-slices. Thus, the system can learn a result at given time 't' as it automatically depends to its anterior result at time 't-1'. The figure 2 shows the evolution of the Bayesian network structure at instants t and t+1

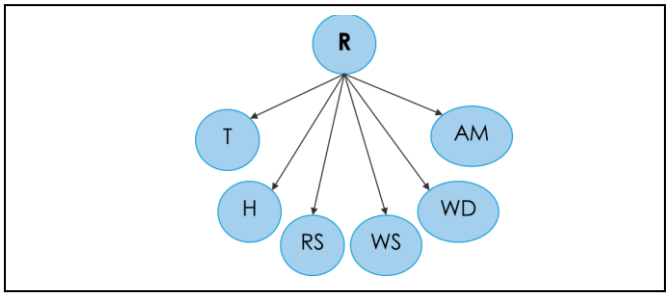


Fig. 1. Structure of naïve bayesian network at time t

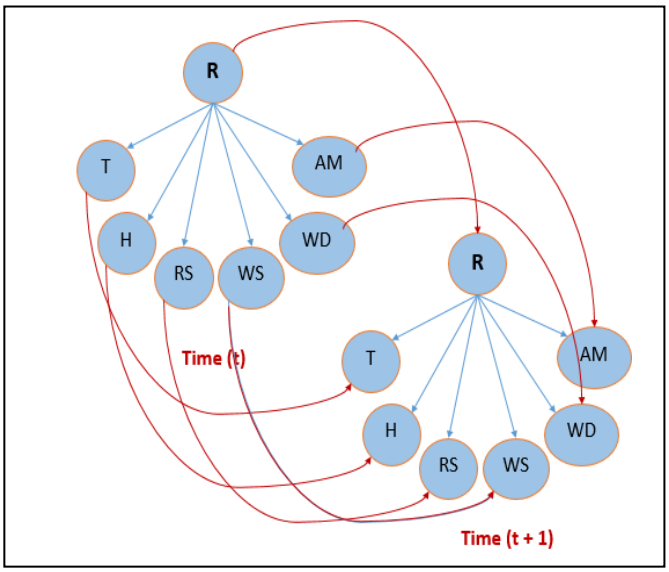


Fig. 2. Structure of dynamic baayesian network at times t and t+1.

It remains to unroll every two time-slices during the period test, calculate the probability of every event existing in the future given its relative past observations Pr (Xt|O1, O2 …… Ot). The probability distribution result is given by equation (2), where T is the test period, N the number of variables at every time-slice, X the predicted target (rains quantities) and Pa(X) the parents of the target node according to the Bayesian network

C. *Results*

A successful application of already detailed algorithm have resulted in a dynamic prediction using a data base of 5721 hours where 5328 rows are chosen for the knowledge model construction and the rest are designed to evaluate the model performance. Results are reliable at 90% compared with the existing results which is considered promising.

The following table shows an hourly forecast of the rain attenuation during the period between 29/10/2015 17:00 and 30/10/2015 09:00

TABLE II. SAMPLE OF OBSERVED VS. PREDICTED MEASURES

Time	Real Rain Measures	Real Probabilities	Bayesian Probabilities	Predicted Rains Measures
…	…	…	…	…
29/10/2015 17:00	0	0,985230980594	0,648749387	0
29/10/2015 18:00	0	0,985230980594	0,966421694	0
29/10/2015 19:00	0	0,985230980594	0,894991903	0
29/10/2015 20:00	0	0,985230980594	0,940918719	0
29/10/2015 21:00	0	0,985230980594	0,971600025	0
29/10/2015 22:00	1	0,015799416109	0,000000000	1

Time	Real Rain Measures	Real Probabilities	Bayesian Probabilities	Predicted Rains Measures
…	…	…	…	…
29/10/2015 23:00	1	0,015799416109	0,000000000	1
30/10/2015 00:00	1	0,015799416109	0,000000000	1
30/10/2015 01:00	1	0,015799416109	0,000000000	1
30/10/2015 02:00	1	0,015799416109	0,000000000	1
30/10/2015 03:00	0	0,985230980594	0,943974628	0
30/10/2015 04:00	0	0,985230980594	0,925619401	0
30/10/2015 05:00	1	0,015799416109	0,000000000	1
30/10/2015 06:00	0	0,985230980594	1,000000000	0
30/10/2015 07:00	0	0,985230980594	0,988095381	0
30/10/2015 08:00	1	0,015799416109	0,000000000	1
30/10/2015 09:00	0	0,985230980594	0,928247828	0
…	…	…	…	…

A comparison between the generated observed vs. predicted results are given by the confusion matrix in table III and figure fig. 3.

TABLE III. CONFUSION MATRIX

		Predicted		
		Rainy	***Not Rainy***	***Total***
Observed	***Rainy***	28	2	30
	Not Rainy	6	2854	2860
	Total	34	2856	2890

According to the elaborated results presented by the confusion matrix in table x, the classification rate was correct to 90%, the positive and negative capacities of prediction are respectively equal to 0,82 and 0.99, and the Sensibility and Specificity correspond to 0.93 and 0.99

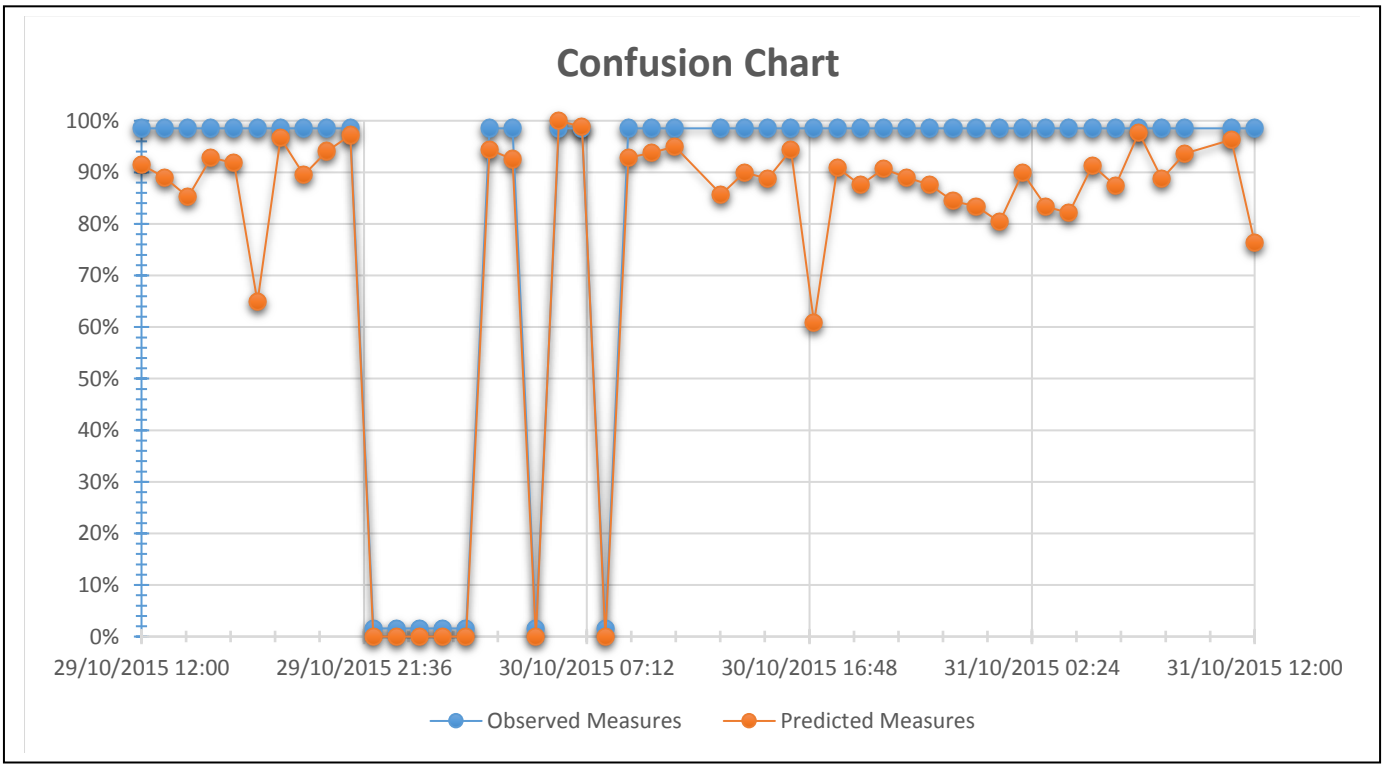


Fig. 3. Confusion chart of observed vs. predicted measures

IV. Conclusion and perspectives

In this paper, we had implemented Dynamic Bayesian Networks as a decisional tool for the prediction of the rain measure based on temporal climatic factors (hourly measurements during one season) in order to more conserve the irrigation water resources. Probabilistic given results were reliable at 90% and have proven the efficiency of the developed applied model of knowledge construction and inference.

As perspectives, we aim to extend the application of the DBN technique by integrating larger parameters as soil and crop characteristics and to carry out the technique in real-time context in order to improve the quantitative and qualitative decisions at each observation and at every moment.

Acknowledgment

This work has been performed as a part of MOBIDOC-PASRI program (pasri.tn) supported by the UE committee and the General Direction of Scientific Research of Tunisia with collaboration of the national meteorological subdivision of Sfax (meteo.tn) represented by Mr. Fekri Ghorbel

References

[1] H. Darouich, J.M.Gonçalves, A.Muga, L.S.Pereira, "Water saving vs. farm economics in cotton surface irrigation: An application of multicriteria analysis" J. Agriculture and Water Management, vol. 115, 2012 pp. 223-231

[2] C.M.G. Pedras, L.S. Pereira, J.M. Gonçalves, "MIRRIG: A decision support system for de-sign and evaluation of micro irrigation systems," J. Agriculture and Water Management, vol. 96, 2009, pp. 691-701

[3] J. Brossier, B. Dent, « Gestion des exploitations et des ressources rurales : entreprendre, négocier, évaluer. Farm and Rural Ressources management : New Context, New Constraints, New Opportunities," SAD N° 31 Introduction Brossier Dent pp 11-25, 1998

[4] C. Bontemps, S. Couture, P. Favard, "Demande estimation of water irrigation under incertitude," J. Rural economy, 2003

[5] V. S. Rahangadale, D. S. Choudhary, "On Fuzzy Logic based Model for Irrigation Controller using Penman-Monteith equation," Proceedings published in International Journal of Computer Applications (IJCA), 2011

[6] F. Touati, M. Al-Hitimi, K. Benhmed, R. Tabish, "A fuzzy logic based irrigation system enhanced with wireless data logging applied to the state of Qatar," J. Computers and Electronics in Agriculture, vol:98, pp:233-241, 2013

[7] P. Marti, M. Gasque, P.G. Altozano, "An artificial neural network approach to the estimation of stem water potential from frequency domain reflectometry soil moisture measurements and meteorological data," J. Computers and Electronics in Agriculture, vol:91, pp:75-86, 2012

[8] M. Dursun, S. Ozden "An efficient improved photovoltaic irrigation system with artificial neural network based modeling of soil moisture distribution – A case study in Turkey," J. Computers and Electronics in Agriculture, vol:102, pp:120-126, 2014

[9] E. Giustia, S. Marsili-Libelli, "A Fuzzy Decision Support System for irrigation and water conservation in agriculture," J. Environmental Modelling & Software, vol:63, pp: 73–86, 2015

[10] S. Holtzman, "Intelligent Decision Systems," ed. Wesley, 1989

[11] G. Itzhak, D. Schmeidler, "Maximum Expected Utility with a Non-Unique Prior," Journal of Mathematical Economics, vol. 18, 1989, pp. 141-153

[12] D. Speigelhalter, S. Lauritzen, "Local computations with probabilities on graphical structures and their application to expert systems" J. Royal statistical Society, vol. 50, 1988, pp. 157–224.

[13] F. Jensen, "An Introduction to Bayesian Networks," ed. Springer, 1996

[14] M.I. Jordan. "Graphical models," J. Statistical Science, vol. 19, 2004, pp. 140–155

[15] R.E. Neapolitan, "Learning Bayesian Networks," ed. Pearson, 2004

[16] L. Lauritzen, G. Cowell, P.A. Dawid, D.J. Spiegelhalter, "Probabilistic Networks and Expert Systems," ed. Springer-Verlag New York: Academic, 1999

[17] T. Verma, J. Pearl, "Equivalence and synthesis of causal models," Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, San Francisco, pp. 220–227, 1991

[18] P. Spirtes, C. Glymour, R. Scheines, "Causation, prediction, and search," ed. pringer-Verlag, 1993

[19] K.P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Phd, University of California, Berkeley, 2002

Safety and Risk Analysis in Oil and Gas Industry Using Bayesian Networks

Zerrouki Hamza
University of Batna 2
Occupational Health and Safety Institute
Batna, Algeria
Email: Hamza.zerrouki@univ-batna.dz

Smadi Hacene
University of Batna 2
Occupational Health and Safety Institute
Batna, Algeria

***Abstract*—Chemical industries, including oil and gas industry, can be exposed to undesired events that may cause terrible accidents. These accidents must be controlled and reduced. To this end, numerous risk analysis management approaches have been aimed at reducing the risks to a tolerable level to avoid the catastrophic accident. This reduction is achieved by implementing several layers of protection, including organizational and technical barriers. In this paper, a risk analysis is established based on Bayesian networks to investigate the safety barriers of a gas facility. Firstly, the causes, the consequences and the protective barriers of the possible accidents are identified using Hazard Operability (HAZOP) analysis. Then, the barriers of the system are modeled using Bayesian networks, the results are compared with the risk tolerance criteria to specify the dangerous consequences. Finally, some measures of prevention are proposed to improve the safety of the process.**

***Keywords*—Risk analysis, HAZOP, Bayesian Networks, Safety Barriers.**

I. Introduction

Over the last 70 or more years, the oil and gas industry witnessed great development from all sides and have seen significant interest from researches. Due to this development, many catastrophic incidents have occurred among them; the Bhopal Gas Tragedy of 1984 which killed or maimed over 20 000 persons. The Macondo Blowout (the Gulf of Mexico, 2010), the accident cost the lives of 11 men and millions of dollars, in 2004 an LNG release from unit 40 in Skikda (Algeria) caused 27 fatalities and 74 injuries[1]. The impact of the accidents in this industry exceed the human and the economies lose, it causes a huge damage to the environment. To avoid these consequences, it is important to learn from past accidents to build a real picture of the front risks and to understand the mechanisms of accidents [2]. The latter helps to develop accident prevention and control strategies. To this end, risk analysis is performed to identify the possible causes and consequences of the accidents. A risk analysis can be defined as an exercise, which includes both qualitative and quantitative determination of risk and its multidimensional impacts. In qualitative analysis, several techniques can be used such as Hazard Operability (HAZOP) and Failure mode effect analysis (FMEA) [3]. The quantitative analysis including Fault tree analysis (FTA), Event tree analysis (ETA) and Bayesian networks are used to determine the frequency of occurrence of an accidental event. In addition, some methods classified as semi-quantitative such as the layer of protection analysis (LOPA). The latter used particularly to identify the safety barriers in the process (e.g. alarm, safety valve). In this paper, a risk analysis of gas facility in Hassi RMel (south Algeria) is performed. Firstly, HAZOP analysis is used to identify the causes, the consequences of possible accidents and the protective Barriers implemented to mitigate the consequences. Secondly, BN analysis is established to investigate the safety barriers of the gas facility, the outcomes from the BN analysis are compared with the risk acceptance criteria. This latter defined as upper limits of acceptable risks that help to make a decision about outcomes of the accidents. Finally, some measures of protections and preventions are proposed to improve the safety of the facility.

This paper is organized as follows; a brief description of the methods used in risk analysis is presented in section 2. Section 3 devoted to the description of the case study also, we applied the risk analysis approaches on the process while the discussion of the results and the improvement measure proposed are given in section 4. The conclusion is presented in section 5.

II. Qualitative and Quantitative Approaches

This section gives a brief description of HAZOP analysis and Bayesian networks analysis.

A. Hazard and operability (HAZOP) analysis

HAZOP is a qualitative approach that is used for hazard identification and assessment. It studies the deviation from normal conditions to hazards which allow its user to make intelligent guesses in the identification of hazard and operability problems [3]. HAZOP study is considered by a group of multi-disciplinary experts to identified abnormal causes and consequences for all possible deviations using different design and operation documents such PI&Ds (Piping and instrumentation diagrams) and PFD (process flow diagrams). HAZOP should be complemented by a quantitative technique (fault tree, event tree) to give deterministic values to the consequences and the safety barriers.

B. Bayesian Networks (BN) method

BNs have been successfully applied in a wide range of domains such as safety and reliability domains, in this paper, BNs are defined on discrete random variables.

Bayesian networks or Bayesian belief networks use graphical models, the DAGs, to depict any probabilistic relationships among a set of variables [4]. In BNs, each node of the DAG represents a variable and the arcs between variables indicate direct probabilistic relations between the connected nodes [5]. These arcs are directed from the parent or cause node to the child or effect node [6], and provide a compact representation of joint probability distributions.

Within BN models, a BN is defined to be a pair of variables $\{(V, E), P\}$, where $\{V, E\}$ are the nodes and the edges of a Directed Acyclic Graph (DAG) respectively, and P is a probability distribution over V discrete random variables $V = \{X_1, X_2, ..., X_n\}$ that are assigned to the nodes. While the edges E represent the causal probabilistic relationship among the nodes [7].

According to the conditional independence, resulting from the d-separation concept [4], and the chain rule. BNs represent the joint probability distribution P(X) of variables $X = \{X_1, X_2, ..., X_n\}$ of any Bayesian networks as:

$$P(X) = \prod_{i=1} P\left(X_i / parents\left(X_i\right)\right) \tag{1}$$

Where $P(X_i)$ are the parents of X_i in the BN and P(X) reflects the properties of the BN [8]. BNs can update the prior probability of any events given new information (posterior probability), called evidence M taking advantage of Bayes theorem:

$$P(X/M) = \frac{P(X, M)}{P(M)} = \frac{P(X, M)}{\sum_x (X, M)} \tag{2}$$

III. Case Study

A. Description of the process

The module processing plant 2 (MPP2) in south Algeria is a set of facilities with different functions (distillation columns, pumps, turbo-expander, heat exchangers...) as shows Fig. 1, these facilities are used to extract heavy hydrocarbons (condensate and GPL) of raw gas coming from the wells, and also produce various treaties gas (sales gas and reinjection).

The biphasic raw gas coming from 39 producer wells met in 9 collectors then gather in manifolds arrives towards the Boosting whose role is to compress the raw gas from 93 kg/cm^2 to 120 kg/cm^2. In order to recover the maximum of liquid hydrocarbons, then it is distributed on three identical trains (A, B, C) with same capacity of 20 million m^3/j to the diffuser D001 at a pressure of 120 kg/cm^2 and a temperature of 58°C.

After the diffuser, the raw gas is cooled in the cooling tower E101 to 40°C, the cooled gas passes through admission separator D101 where the gas separates from the water and liquid hydrocarbons, the water goes towards the evaporation basin. The gas originating from D101 with a pressure of 118 kg/cm^2 passes through the heat exchangers of gas/gas E103 and E102 where it cooled to -9°C. the gas passes through the Joule-Thomson valve (PRCV108) where it undergoes a first isenthalpic expansion up to 100 kg/cm^2 and a temperature of -13°C before reaching the high-pressure separator D102.

To avoid the formation of hydrates that may block the exchangers, injecting a glycol solution containing 80% mass injected at the heat exchangers E102 and E103. The D102 separator separates the gas again, the solution of MEG (monoethylene glycol) and the liquid condensate. Having absorbed water, MEG solution is sent under pressure to the glycol regenerator section.

The gas from the D102 undergoes isentropic expansion in the turbine of "Turbo-Expender K101", this last recover the energy, which occurs when high-pressure gas passes through the turbine to reduce its pressure at 67 kg/cm^2 and a temperature of -35°C before passing through the cold separator D103. The gas cooled from D103 passes through the gas / gas exchanger E102 A/F calendar side to cool the raw gas, and is heated itself to a temperature of 43°C, then it is compressed to 74 kg/cm^2 on the compressor K101 and directed to the sales gas pipeline.

The liquid hydrocarbons from the admission separator D101 are relaxed at the condensate separator rich D105 at 33 kg/cm^2 and 42°C. The hydrocarbons recovered at D102 and D103 pass to the low-pressure separator D104 at a temperature -40°C and a pressure 34 kg/cm^2.

We are focusing on the separator tank D101, which is the first station of the gas in the facilities. The rest of the facilities will not be considered in this study.

B. Application of the methodology

To identify the potential scenarios, HAZOP (Hazard and Operability) is performed. The HAZOP technique was applied in the risk analysis intended to identify the causes and consequences (effects) of deviations in the process. Table 1 shows the potential scenarios that could lead to explosion, process shutdown and environmental impacts and the different barriers that mitigate these consequences.

To keeps the internal pressure limited to the design values the pressure indicator of the Alarm PICAH139 opens the safety valve PICA-139V when the pressure in the admission separator D101 rises (104.8 kg/cm^2). In a case of the safety valve PICA-139V failure and the pressure inside D101 rises above 105.4kg/cm^2, the LNG is discharged to the flare. To this end, the signal from the high-pressure sensor PIC-144 opens the valves PCV-101 automatically. Also, a high-pressure alarm PICAH 139 alerts the operator to close the manual valve XV-920 and take appropriate actions. The different components related to the accident scenarios in the process are depicted in Table 2.

The Bayesian network was constructed for the accident scenario of the processing system. As can see in Fig. 2 the relationship between components of the process is modeled using Hugin software [9]. In the beginning, the initiating event and its frequency should be identified. Then the various

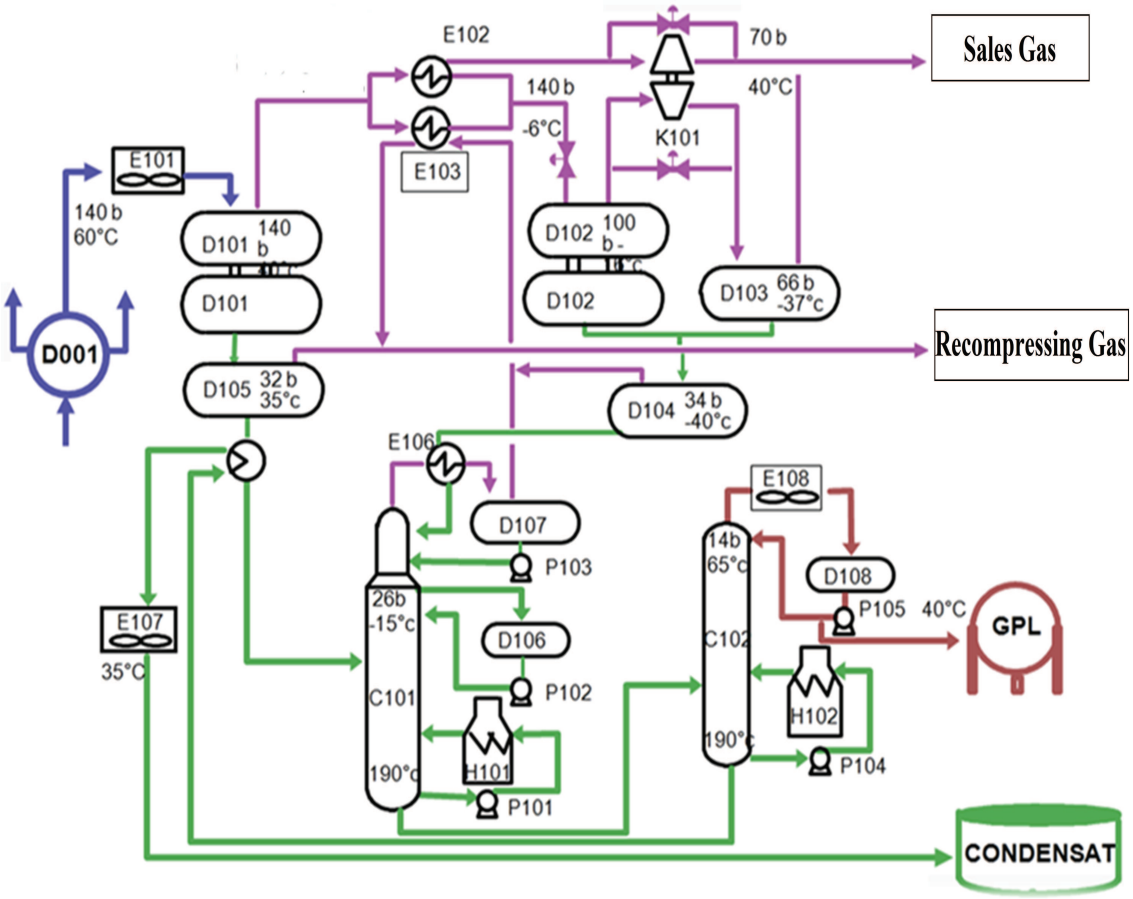


Fig. 1. Simplified plant block diagram of the process

TABLE I
HAZOP FROM ACCIDENT SCENARIOS RELATED TO THE PROCESS

No	Guide-word	Element	Deviation	Causes	Consequences	Protective Barriers
1	High	pressure	High pressure	Failure of the safety valve PCV139 (stuck open)	pressure increase in the D101, rupture of the admission separator D101, fire, explosion, Flash fire, Pool fire, Vapor cloud explosion (VCE)	high pressure alarm PZAH102 human operator high pressure alarm PICAH139 human operator Pressure indicators Pressure relief valves (PCV-101 and HXC-102) manual valve XV-920 Emergency evacuation system
2	High	Level	High level	Operator fail to manipulate the manual valve XV 920 (stay opened)	Very high pressure inside the admission separator D101 can lead to explosion and process shutdown	High level Alarm LZH103 on the D101A: human operator close the valve PICV 139 high pressure alarm PICAH 139 in case of pressure increase in the D101 (due to a rise in level) level indicators
				Failure of the safety valve PCV139 (stuck open)	High level in the D101 drive to HP Flare, formation of frost in the exchanger tubes (E102) with potential breaking of tubes, damage of the calendar with release gas into the atmosphere.	High level alarms LICAH101/102: human operator (control room) actuate the liquid line with bypass valves LICV 101A from D101 to D105 or from LICV 101B to D003A (off-spec) level indicators

TABLE II
DIFFERENT COMPONENTS RELATED TO THE ACCIDENT SCENARIOS IN THE PROCESS AND THEIR OCCURRENCE PROBABILITIES

Number	Component	Probability of failure
1	Alarm: PZAH102/ Human operator	0.1
2	Safety Valve PCV139	3.81E-6
3	Alarm: PICAH139 /Human operator	0.1
4	Pressure relief valve PCV101	0.01
5	Pressure relief valve HXC102	0.01

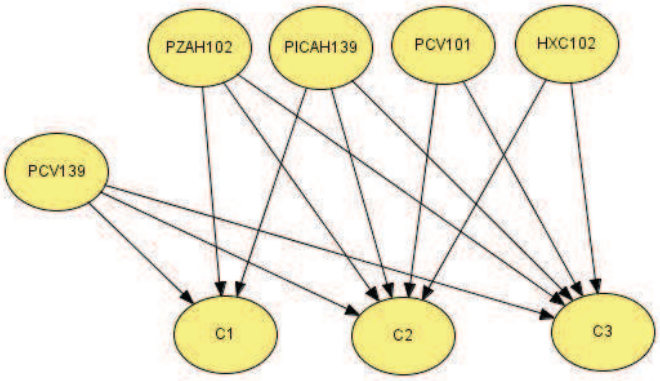


Fig. 2. Bayesian network model for the accident scenario

barriers implemented in the system need to be recognized. Finally, the residual risk posed by the scenario needs to be checked with that of the tolerable limit and one needs to make sure the residual risk is within the tolerable limit [10]. In SONATRACH Company, the maximum tolerable frequency for accident scenarios has defined to be 10^{-5}/year. If the risk level exceeds the tolerable limit, additional measures of security must be proposed to improve the safety of the process.

In order to perform the calculations and determine the consequences of the scenario, the failure probabilities of the barriers and the other components are given in Table 2 (Data derived from [11]).

IV. RESULTS AND DISCUSSION

The results of the BN model in Figure 2 are presented in Table 3. From HAZOP analysis three principal consequences are distinguished; pressure increase in the D101, release gas into the atmosphere and explosion and process shutdown. In addition, a new node is added in the BN model called safe situation indicate that the system working in normal conditions. The four probabilities of possible consequences are calculated using BN and presented in Table 3. Compared with the risk acceptance criteria, which described previously, the most dangerous consequences are C1 and C2 with probability 7.24E-5 and 3.05E-4, respectively. Furthermore, BN can use for both qualitative and quantitative assessment. The qualitative phase identifies by a network presentation while the quantitative analysis is represented by conditional probability tables associated with each node in the network. Also, BN

TABLE III
FREQUENCIES OF THE CONSEQUENCES

Number	Consequences	Frequencies of scenarios	
		Prior	*Posterior*
1	pressure increase in the D101	7.24E-5	0.19
2	release gas into the atmosphere	3.05E-4	0.802
3	explosion and process shutdown	3.06E-6	0.008
4	Safe situation	0.99961945	0

has the ability to perform diagnostic (posterior) and predictive (prior) analysis [12]. In the right side of Table 3, the prior and the posterior probabilities of different component and consequences are calculated, the posterior probability is estimated considering the failure of the safety valve PICA-139V.from Table 3 we can conclude that the released gas into the atmosphere (C2) has the highest increase which can be explained that C2 have a deep impact in the human factor and the environment. In addition, in the presence of ignition source C2 can cause multiple consequences such as fire, explosion Flash fire, Pool fire, Vapor cloud explosion (VCE). some recommendations can be proposed for the case study , another safety valve must be implemented redundancy with the safety valve PICA-139V also; the process should be supplemented with safety-instrumented systems (SIS) to decrease the frequencies of the consequences

V. CONCLUSION

The study shows the suitability of Bayesian networks in the Risk analysis of process system; it can perform both qualitative and quantitative analysis. Moreover, BN used diagnostic (posterior) and predictive (prior) analysis to calculate the probabilities of the components and the possible consequences. The posterior probability present a good comprehension of the most dangerous consequence. BN-HAZOP analysis has proven their efficiency in the field of gas and oil industry due to the ease of application and the results obtained.

REFERENCES

[1] S. Chettouh, R. Hamzi, and K. Benaroua, "Examination of fire and related accidents in Skikda Oil Refinery for the period 2002–2013," J. Loss Prev. Process Ind., vol. 41, pp. 186–193, 2016.

[2] F. I. Khan and S. . Abbasi, "Reply to comments on 'Major accidents in process industries and an analysis of causes and consequences,'" J. Loss Prev. Process Ind., vol. 14, no. 1, p. 85, 1999.

[3] F. I. Khan and S. . Abbasi, "Techniques and methodologies for risk analysis in chemical process industries," J. Loss Prev. Process Ind., vol. 11, no. 4, pp. 261–277, 1998.

[4] J. Pearl and a. Hasman, Probabilistic reasoning in intelligent systems: Networks of plausible inference, vol. 28, no. 3. San Francisco, California: MORGAN KAUFMANN PUBLISHERS, INC., 1991.

[5] P. Weber and L. Jouffe, "Complex system reliability modelling with Dynamic Object Oriented Bayesian Networks (DOOBN)," Reliab. Eng. Syst. Saf., vol. 91, no. 2, pp. 149–162, 2006.

[6] D. Hanea and B. Ale, "Risk of human fatality in building fires: A decision tool using Bayesian networks," Fire Saf. J., vol. 44, no. 5, pp. 704–710, 2009.

[7] A. Bobbio, L. Portinale, M. Minichino, and E. Ciancamerla, "Improving the analysis of dependable systems by mapping Fault Trees into Bayesian Networks," Reliab. Eng. Syst. Saf., vol. 71, no. 3, pp. 249–260, 2001.

[8] Finn V. Jensen and Thomas D. Nielsen, Bayesian Networks and Decision Graphs, Second edi., vol. 1. New York: Springer, 2007.

[9] HUGIN, “HUGIN Expert software version 8.1,” 2015. [Online]. Available: http://www.hugin.com.
[10] P. R. Kannan, “Bayesian networks: Application in safety instrumentation and risk reduction,” ISA Trans., vol. 46, no. 2, pp. 255–259, 2007.
[11] F. . Lees, Loss Prevention in the Process Industries, Third edit. Butterworths, London, 2005.
[12] J. Bhandari, R. Abbassi, V. Garaniya, and F. Khan, “Risk analysis of deepwater drilling operations using Bayesian network,” J. Loss Prev. Process Ind., vol. 38, pp. 11–23, 2015.

Modeling of Soil Penetration Resistance Using Multiple Linear Regression (MLR)

[a*]Ben Salah Nahla, [b]Anis Elaoud, [c]Hanen Ben Hassen, [a]Afif Masmoudi, [b]Sayed Chehaibi

[a] Faculty of Sciences of Sfax
[b] Higher Institute of Agronomic Siences of Chott Mariem
[c] School of Engineering Rural Equipment Medjez Elbeb

[a*] Email: *bensalahnahla@yahoo.fr*

Abstract

Using of mechanization can increase crop productivity but their inadequate implementation also has adverse effects on agricultural soils. Latter can be generated by increasing the penetration resistance and as a result the phenomenon of compaction. This problem becomes more acute with increasing contact wheel / ground. For this reason, the purpose of this study was to evaluate the efficiency with which soil penetration resistance is estimated using a proposed model based on moisture content, density, tractor weight, number of passes, and the wheel inflation pressure. Experimental works were analyzed statistically and the penetration resistance was modeled using multiple linear regressions (MLR).

Key words: Compaction / Soil / Resistance to penetration / Modeling / MLR.

Accepted Posters

Integral inequalities for k-fractional continuous random variables

Mohamed HOUAS
Laboratory FIMA, UDBKM, University of Khemis Miliana, Algeria

Abstract: The integral inequalities have many applications in differential equations,numerical quadrature, probability and statistical problems. For details, we refer to [1, 2] and the references therein. Moreover, the study of fractional type inequalities is also of great importance. Barnett and al [2] established several integral inequalities for the expectation and the variance of a random variable having a probability density function. For several results concerning the probability inequalities we refer the interested reader to [1, 2, 3, 4, 6, 7]. Recently, Dahmani [3] introduced new concepts on fractional continuous random variables, Dahmani presented several integral inequalities for the fractional dispersion and the fractional variance functions of a continuous random variable. In this work, we use the $k-$Riemann–Liouville fractional integral to develop some new integral inequalities for continuous random with a probability density function defined on some finite real interval.

Keywords: Riemann-Liouville integral. $k-$fractional integration. integral inequalities. random variable.

References

[1] Barnett, N.S., Cerone, P., Dragomir, S.S., Roumeliotis, J, Some inequalities for the expectation and variance of a random variable whose PDF is n-time differentiable. J. Inequal. Pure Appl. Math. 1(21), 1-29 (2000).

[2] Barnett, N.S., Cerone, P., Dragomir, S.S., Roumeliotis, J, Some inequalities for the dispersion of a random variable whose PDF is defined on a finite interval. J. Inequal. Pure Appl. Math. 2(1), 1-18 (2001).

[3] Dahmani, Z.: Fractional integral inequalities for continuous random variables. Malaya J. Mat. 2(2), 172-179 (2014).

[4] M. Houas, Some inequalities for k-fractional continuous random variables, J. Advan. Res. In Dynamical and Control Systems. 7 (2015), no.4, 43-50.

[5] M. Houas, Some new saigo fractional integral inequalities in quantum calculus. Accepeted.

[6] M. Houas, Some integral inequalities involving Saigo fractional integral operators. Accepeted.

[7] Mubeen, S, $k-$Fractional Integrals and Application. Int. J. Contemp. Math. Sciences. 7(2), 89-94 (2012).

Study the modeling to estimate of band gap energy and urbach energy of ZnO:X (X = Co, V and F) thin films

Said Benramache *
Material Sciences Department, Faculty of Sciences,
University of Biskra, Biskra 07000, Algeria

Boubaker Benhaoua
Physic Laboratory of Thin Films and Applications
LPCMA, University of El-oued, Algeria

***Abstract*—Investigation of new calculate depends on the controlled optical properties of nanomaterials. ZnO is one of the most important semiconductor materials for its semiconducting characteristics. This paper focus to present a new approach to calculate the optical gap and Urbach energies, these correlations based on experimental data were published previously in the literatures. The thin films were deposited at different precursor molarities by ultrasonic and spray pyrolysis techniques. The models proposed to calculate the band gap and/or the Urbach energies of undoped and Co, V and F doped ZnO films were studied. The relation between experimental data and theoretical calculation with molarities suggests that the band gap and/or the Urbach energies are predominantly estimated by these energies and doping. The measurements by these proposals models are in qualitative agreements with the experimental data, which the correlation coefficients values were found higher than 0.9 in all calculation.**

***Keywords*—** *ZnO; Thin film; Semiconductor doping; Correlation.*

I. INTRODUCTION (*HEADING 1*)

Zinc oxide is an oxide of group II metal Zinc that belongs to P63mc space group. Zinc is in the transition metal row which has 3d10 moments and hence it does not have any unpaired electron orbiting around the nucleus [1–5]. Zinc oxide (ZnO) has a wurtzite (WZ) structure; this is a hexagonal crystal structure [5–7]. It has been reported that ZnO has very most important semiconductor material due to its wide band gap (3.37 eV) and large exciton binding energy (60 meV) at room temperature [8,9].

Zinc oxide based coatings are of much interest in science and technology due to their interesting applications such as in microelectronic devices, light emitting diodes, thin films, antireflection coatings, transparent electrodes in solar cells, gas sensors surface acoustic wave devices, varistors, spintronic devices and lasers [10–17].

ZnO thin films can be produced by several techniques such as reactive evaporation and thermal annealing [18], molecular beam epitaxy (MBE) [19], magnetron sputtered techn-ique [20], pulsed laser deposition (PLD) [21], the low-temperature solution method [22], electrodeposition [23], the sol-gel technique [24], chemical vapor deposition, electroche-mical deposition [25] and spray pyrolysis [26], have been reported to prepare thin films of ZnO. Among these, we will focus more particularly in this paper on the spray techni-que because of its simplicity and suitability for large-scale production, it has several advantages in producing nanocrystalline thin films, such as, relatively homogeneous composition with fine and porous microstructure, a simple deposition on glass substrate because of the low substrate temperatures involved, easy control of film thickness [27].

The aim of this work to study the development of estimation of the optical gap energy Eg and Urbach energy Eu in an undoped and a Co, V and F doped ZnO thin films by varying the precursor molarities and doping level of doped films.

II. EXPERIMENTAL AND METHODS

In this study, the undoped and Co, V and F doped ZnO samples were deposited on glass substrates using the ultrasonic spray and spray pyrolysis technique. The some films were deposited at a substrate temperature of 350 °C. The optical parameters such as the band gap energy and the Urbach energy of undo-ped and Co, V and F doped ZnO thin films were taken from our previous papers, where studied the effect of precursor molarity, doping level on structural, electrical and optical parameters of ZnO thin films. However, the optical parameters of the thin films were also taken from [28–37] studied the effect of precursor molarity, doping level and substrate temperature on undoped and Co, V and F doped ZnO thin films. The data is obtained from publications, most of which used zinc as a precursor (see Table 1). From these data it can be derive-ed that the differences shown in Tables 2 to 3 can be partially ascribed to differences in the deposition circumstances, such as reactor geometry, substrate temperature, annealing tempe-rature, deposition time, flows, concentration of ZnO solution, etc. Nevertheless, the data suggests that across different authors, a common trend can be discerned. The variations of optical gap energy and Urbach energy of undoped and doped ZnO thin films varied in the form nonlinear (see Tables 2 and 3). The model proposed of thin films with precursor molarity and doping concentration is discussed.

* Corresponding author. Tel.: +213667697692.
E-mail address: saidbenramache07@gmail.com ***(S. Benramache)***.

In this study, we will show the evolution of the precursor molarity and doping level on the Urbach energy and band gap energy, we tried to establish correlations for each model proposed. In our calculations, the band gap energy can be calculated from precursor molarity and Urbach energy of undoped and doping level for ZnO thin films; the ZnO exhibit a single crystals exhibit n-type semiconductor with a high crystallinity.

III. THEORETICAL CALCULATIONS

A. *Undoped ZnO thin films*

Firstly, for undoped ZnO thin films, we have used the relationship in the form nonlinear to calculate the optical gap energy from the Urbach energy, precursor molarity, and the following relationship is evaluated in this step:

$$E_g = a\left(\frac{M}{E_u}\right)^b \tag{1}$$

where E_u is the Urbach energy, E_g is the band gap energy correlate and M is the precursor molarity (see Table 2), and a and b are empirical constants as $a \approx 3.28711$ and $b \approx 0.0184683$

B. *doped ZnO thin films*

For the Al doped ZnO films, we have studied the correlation with doping level of Co, V and F doped thin films, these letters were deposited at different precursor molarities of 0.01, 0.02, 0.05, 0.1 and 0.2 M, to perform the correlation in this step the optical parameters was measured with doping level. In this section the optical gap energy was estimated with undoped ZnO thin film (0 wt. %). The correlation can be written in another form is expressed as:

$$\begin{cases} \underline{if \quad X_0 \succ 0} \\ E_g = \left(3.28711 \times (1+AX_0)\right) \times \left(\dfrac{M}{E_u}\right)^{\left(0.0184683 \times (1+BX_0)\right)} \\ \underline{if \quad X_0 = 0} \\ E_g = \left(3.28711\right) \times \left(\dfrac{M}{E_u}\right)^{0.0184683} \end{cases} \tag{2}$$

where A and B are empirical constants and X_0 is the concentration of doped ZnO films. Therefore, were related on the dopant elements are presented in the Table 3.

C. *The Urbach energy evaluated*

The Urbach energy of undoped and doped ZnO thin films was also correlated as the following relationships:

TABLE I. THE PARAMETERS CONDITIONS WERE USED IN THIS PAPER.

Method	ultrasonic Spray or spray pyrolysis
Oxide	zinc oxide thin films
Zn reactants	Zn acetate or $ZnCl_2$
Solvents	ethanol– methanol– water
Substrate	glass
Molarity (M)	0.02, 0.05, 0.075, 0.1, 0.125
Substrat Temperature (C°)	350, 450
dopant's	Cobalt, Vanadium, Fluorine
[X]/[Zn] (%)	0 to 15

$$\begin{cases} \underline{if \quad X_0 \succ 0} \\ E_u = epx\left(\ln M + \dfrac{1}{b(1+BX_0)} \ln \dfrac{a\ (1+AX_0)}{E_g}\right) \pm \Delta E_u \\ \underline{if \quad X_0 = 0} \\ E_u = epx\left(\ln M + \dfrac{1}{b} \ln \dfrac{a}{E_g}\right) \pm \Delta E_u \end{cases} \tag{3}$$

where a, b and A, B are empirical constants relates the undoped and doped, respectively. These constants were measured in the section 3.1 and 3.2. The resulting errors (ΔE_u) were measured from Eq. (2) as described in the following formula:

$$\Delta E_u = \frac{1}{(b\ (1+BX_0\ ))} \frac{\Delta E_g}{E_g} E_u \tag{4}$$

D. *The relative error measurement*

The relative error value was measured between the experimental data and correlate values by the following relationship

$$\varepsilon = \left|(X_{Exp} - X_{Corr}) / X_{Exp}\right| \times 100 \tag{5}$$

where $X_{(e)}$ and $X_{(c)}$ are the experimental and correlate values, respectively, ε is the relative error.

IV. RESULTS AND DISCUSSION

In the present study an attempt is made to correlates optical gap energy with Urbach energy of undoped and Co, V and F doped ZnO thin films by varying the precursor molarity and doping concentrations. In the Figure 1 present the results of calculations for the scaled parameter values according to Eq. (2) and Eq. (3) also are presented in Table 2 with X = 0 mol.l-1, respectively. In the Figure 1 we investigated the estimation values of the optical gap energy and Urbach energy as a function of the samples numbers of undoped ZnO thin films.

TABLE II. SUMMARY RESULTS OF EXPERIMENTAL DATA, THE CORRELATE OPTICAL GAP ENERGY, THE CORRELATE URBACH ENERGY AND RELATIVE ERRORS FOR THE UNDOPED ZNO THIN FILMS.

Undoped ZnO thin films $a \approx 3.28711$ and $b \approx 0.0184683$									
S.N.	M ($mol.l^{-1}$)	T (°C)	$E_{g\ (Exp.)}$ (eV)	E_u (eV)	$E_{g\ (Corr.)}$ (eV)	Error (%)	$E_{u\ (Corr.)}$ (eV)	Error (%)	Ref.
1	0.05	350	3.08	0.9221	3.115	1.136	0.938	1.72	[28]
2	0.075	350	3.22	0.3186	3.200	0.590	0.281	11.80	[28]
3	0.1	350	3.37	0.085	3.297	2.166	0.069	18.82	[28]
4	0.125	350	3.15	0.1757	3.266	3.682	0.201	14.39	[28]
5	0.1	350	3.10	0.2734	3.212	3.612	0.279	2.05	[29]
6	0.1	350	3.267	0.108	3.273	0.183	0.108	0	[30]
7	0.02	350	3.19	0.08	3.204	0.439	0.071	11.25	[31]
8	0.1	350	3.25	0.064	3.314	1.969	0.074	15.62	[32]
9	0.1	350	3.304	0.1139	3.279	0.757	0.101	11.32	[33]
10	0.1	350	3.317	0.0983	3.288	0.874	0.097	1.32	[34]
11	0.1	350	3.27	0.17	3.255	0.458	0.165	2.94	[35]
12	0.1	350	3.25	0.209	3.243	0.215	0.203	2.87	[36]
13	0.1	350	3.23	0.490	3.192	1.176	0.444	9.39	[37]

TABLE III. SUMMARY RESULTS OF EXPERIMENTAL DATA, THE CORRELATE OPTICAL GAP ENERGY, THE CORRELATE URBACH ENERGY AND RELATIVE ERRORS FOR THE AL DOPED ZNO THIN FILMS.

Co doped ZnO thin films with 0.1 $mol.l^{-1}$ [38] $A \approx 1.31409$ and $B \approx 47.53595$							
[Co]/[Zn] (%)	T (°C)	$E_{g\ (Exp.)}$ (eV)	E_u (eV)	$E_{g\ (Corr.)}$ (eV)	Error (%)	$E_{u\ (Corr.)}$ (eV)	Error (%)
0	350	3.250	0.209	3.243	0.215	0.203	2.87
1	350	3.295	0.183	3.276	0.576	0.169	7.67
2	350	3.362	0.108	3.364	0.059	0.106	1.85
3	350	3.300	0.210	3.305	0.152	0.211	0.47

V doped ZnO thin films with 0.01 mol.l-1 [39] $A \approx 5.85427$ and $B \approx 83.96820$							
[V]/[Zn] (%)	T (°C)	$E_{g\ (Exp.)}$ (eV)	E_u (eV)	$E_{g\ (Corr.)}$ (eV)	Error (%)	$E_{u\ (Corr.)}$ (eV)	Error (%)
1	–	3.17	0.1766	3.157	0.410	0.182	3.05
2	–	3.23	0.1534	3.208	0.681	0.162	5.61
3	–	3.25	0.1551	3.234	0.492	0.167	7.67
4	–	3.20	0.1579	3.248	1.500	0.167	5.76
5	–	3.22	0.1900	3.203	0.527	0.193	1.58

F doped ZnO thin films with 0.2 mol.l-1 [40] $A \approx 0.25053$ and $B \approx 2.76142$							
[F]/[Zn] (%)	T (°C)	$E_{g\ (Exp.)}$ (eV)	E_u (eV)	$E_{g\ (Corr.)}$ (eV)	Error (%)	$E_{u\ (Corr.)}$ (eV)	Error (%)
0	450	3.285	0.5757	3.224	1.857	0.569	1.16
5	450	3.295	0.4983	3.265	0.910	0.479	3.87
10	450	3.296	0.4387	3.307	0.333	0.425	3.12
15	450	3.295	0.7299	3.297	0.061	0.716	1.90

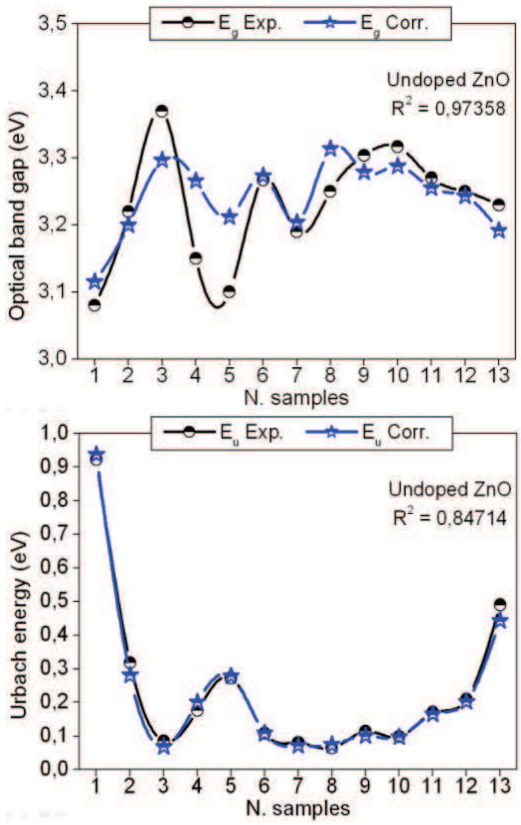


Fig. 1. Summary results of experimental data, the correlate optical gap energy and Urbach energy in the Undoped ZnO thin films.

As can be seen, in the some points the optical energy gap is inversely proportional to the Urbach energy. We obtained by the proposed equations a measurement were in are in qualitative agreements with the experimental data. The correlation coefficient increased with the calculation of the optical gap energy, which the maximum agreement of the estimation was found to be minimum error.

As shown in Figures 2 and 3 (see Table 3), significant estimation was found between the optical gap energy values and the Urbach energy values, respectively. The measurements were investigated with Co, V and F doped ZnO thin films as a function of Co, V and F concentrations. The measurement in the optical gap energy values and the Urbach energy values of Co, V and F doped films also investigated by Eq. (2) and Eq. (3), respectively, which found in qualitative agreements with the experimental data, also the correlation coefficient increased with the calculation of the optical energy, which the maximum agreement of the estimation was found to be minimum relative error. The latter can be calculated from relationships $\left|(E_{gExp} - E_{g\,Corr})/E_{g\,Exp}\right| \times 100$ and $\left|(E_{u\,Exp} - E_{u\,Corr})/E_{u\,Exp}\right| \times 100$, the correlation coefficients for this correlation are also presented in Figures 1 and 2, which related to the relative errors and doping via:

$$R = 1 - \frac{\sum_{i}^{N} \varepsilon_i}{N} \qquad (6)$$

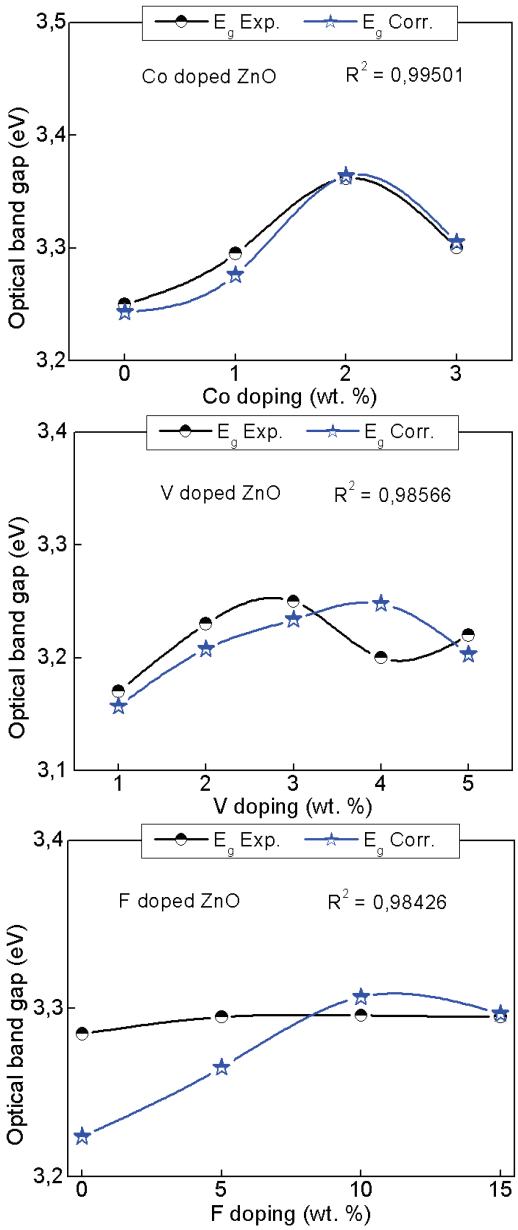


Fig. 2. Summary results of experimental data, the correlate optical gap energy in the Co, V and F doped ZnO thin films.

where N is the number of measurement and ε is the relative error. The correlation coefficients are presented in Figures 1 and 2, it can be seen that the Co, V and F doped ZnO thin films have a good and achieved results obtaining from increasing the correlation coefficients, this is the approach we have adopted in the enhance of band gaps energy and less disorder of ZnO thin films after doping. The maximum enhanced of the correlation coefficients values were estimated for the optical gap energy values (see Figure 3).

In the Figure 4, we obtained that the relative errors of all calculation are smaller than 4 % for the estimation of optical gap energy and 20 % in the Urbach energy, it were confirmed that these models are suitable for calculation of optical properties with varying the growth parameters. The decreases

in the relative errors of undoped to doped films (see Figure 5) can be explained by the good transparency and the Adhesion between the films and the substrate, which can observed in the less defects and less disorder.

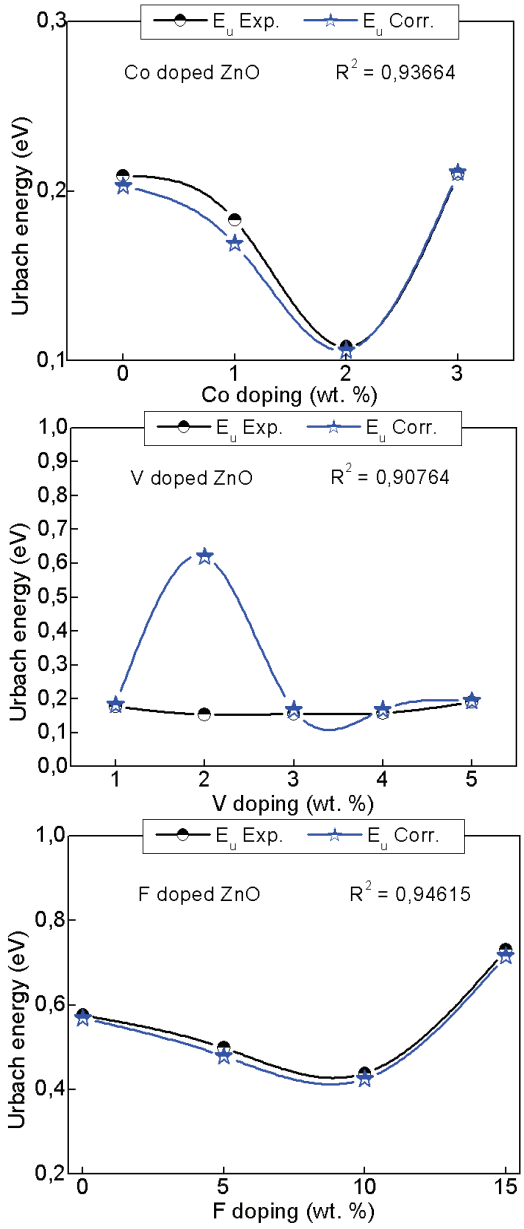


Fig. 3. Summary results of experimental data, the correlate Urbach energy in the Co, V and F doped ZnO thin films.

In our calculations the optical properties for characterizing the undoped and Co, V and F doped ZnO thin films; Stoichiometric the doped ZnO films are highly transparency and good optical band gap. We have estimated the optical band gap and the Urbach energies of the undoped and Co, V and F doped ZnO thin films by varying the precursor molarities and doping concentrations; it are predominantly influenced by the transition tail width of undoped and Co, V and F doped films. The correlation between the optical properties and the experimental conditions was investigated.

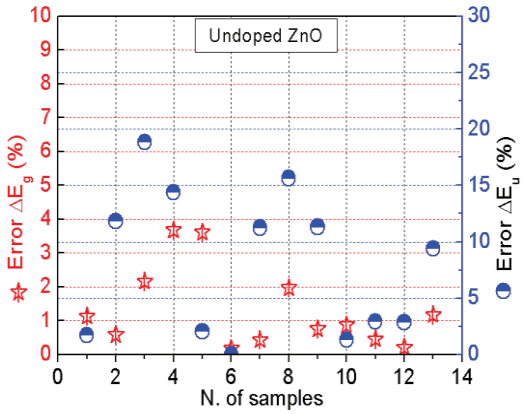


Fig. 4. The relative errors variation as a function of the samples numbers for the optical gap energy and Urbach energy in the Undoped ZnO thin films.

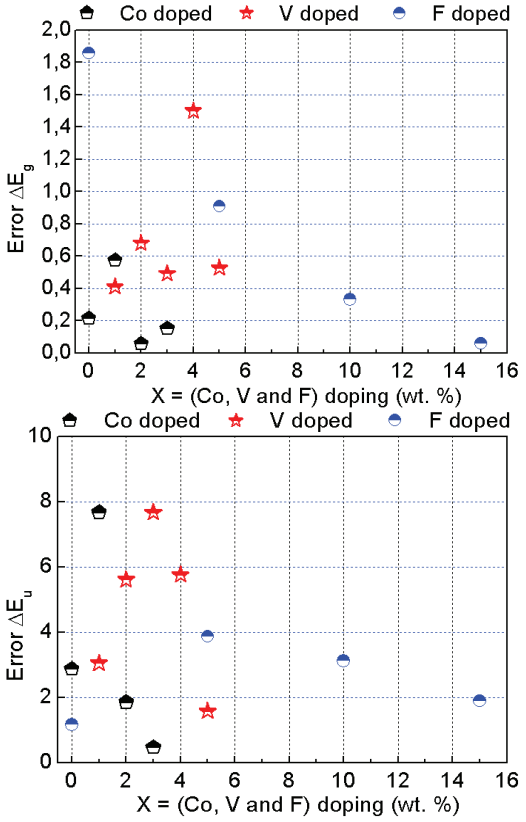


Fig. 5. The relative errors variation as a function of the samples numbers for the optical gap energy and Urbach energy in the Co, V and F doped ZnO thin films.

V. Conclusion

In summary, the undoped and Co, V and F doped ZnO thin films were deposited on glass substrates using the ultrasonic spray and spray pyrolysis technique. The model proposed to calculate the band gap and the Urbach energies of undoped and

Co, V and F doped ZnO thin films were investigated. The relation between the experimental data and theoretical calculation with precursor molarities suggests that the band gap and/or the Urbach energies are predominantly estimated by the band gap and/or the Urbach energies and the concentration of ZnO solution. The measurements by these proposals models are in qualitative agreements with the experimental data, the correlation coefficients values were varied in the rand 0.84–0.99, so that the relative errors of all calculation are smaller than 10 %. The best estimated results were obtained after doping; which the relative errors values are smaller than 2 % for the estimation of optical gap energy and 10 % in the Urbach energy, it were confirmed that these models are suitable for calculation of optical properties with varying the growth parameters, respectively. This is the approach we have adopted in the enhance of band gaps energy and less disorder of ZnO thin films after doping. Stoichiometric the Co, V and F doped ZnO films are highly transparency and good optical band gap.

REFERENCES

[1] Y. Dai, Y. Zhang, Z.L. Wang, Solid State Commun. 126, 629 (2003)

[2] J.W. Sun, Y.M. Lu, Y.C. Liu, D.Z. Shen, Z.Z. Zhang, B.H. Li, J.Y. Zhang, B. Yao, D.X. Zhao, X.W. Fan, J. Phys. D Appl. Phys. 41, 155103 (2008).

[3] L.W. Zhong, J. Phys. Condens. Matter 16, 829–858 (2004).

[4] J. Balaji et al. International Journal of Nanoscience 10, 787 (2011).

[5] W.L. Wang, L.Li, K.J. Liao, J. Zhang, R.J. Zhang, F.F. Yang and G.Z. Fu , International Journal of Modern Physics B 19, 651 (2005).

[6] F. Benharrats, K. Zitouni, A. Kadri, B. Gil, Superlattices and Microstructures, 47, 592 (2010).

[7] S. Benramache, A. Rahal, B. Benhaoua, Optik, 124, 663 (2013).

[8] H. Bhadane, E. Samuel, D. Kumar Gautam, Surface Review and Letters, 21, 1450046 (2014).

[9] H. Eshghi and Y. Arjmand, Modern Physics Letters, B 26, 1250176 (2012).

[10] B. Janfeshan, M.A. Ahrevar, K. Ahmadi, International Journal of Modern Physics B 22, 3289 (2008).

[11] N. Tabet, M. Faiz and A.A. Oteibi, International Journal of Nanoscience 6, 23 (2007).

[12] S.J. PEARTON et al. Brief Reports and Reviews 2, 201 (2007).

[13] K. Miyajima, A. Amamoto, T. Goto, H.J. Ko, T. Ao, International Journal of Modern Physics B, 15, 3614 (2001) 3611-3614.

[14] Y. Lv et al. International Journal of Modern Physics B 20, 3635 (2006).

[15] J. Kennedy and A. Markwitz, International Journal of Modern Physics B 20, 4655 (2006).

[16] Z.H. Khan et al. International Journal of Nanoscience 9, 423 (2010).

[17] X.Q. Wei, B.Y. Man, Y.T. Wang and H.Z. Zhuang, International Journal of Modern Physics B 21, 1775 (2007).

[18] S. Rahmane, M.A. Djouadi, M.S. Aida, N. Barreau, B. Abdallah, N. Hadj Zoubir, Thin Solid Films, 519, 5 (2010).

[19] D. Vernardou, G. Kenanakis, S. Couris, A.C. Manikas, G.A. Voyiatzis, M.E. Pemble, E. Koudoumas, N. Katsarakis, Journal of Crystal Growth, 308, 105 (2007).

[20] Y.D. Ko, K.C. Kim, Y.S. Kim, Superlattices and Microstructures, 51, 933 (2012).

[21] E.F. Keskenler, G. Turgut, S. Dogăn, Superlattices and Microstructures, 52, 107 (2012).

[22] C.C. Ting, C.H. Li, C.Y. Kuo, C.C. Hsu, H.C. Wang, M.H. Yang, Thin Solid Films, 518, 4156 (2010).

[23] R.E. Marotti, P. Giorgi, G. Machado, E.A. Dalchiele, Solar Energy Materials & Solar Cells, 90, 2356 (2006).

[24] J. Ramesh, G. Pasupathi, R. Mariappan, V. Senthil Kumar, V. Ponnuswamy, Optik, 124, 2023 (2013).

[25] S. Benramache, B. Benhaoua, N. Khechai, F. Chabane, Matériaux & Techniques, 100, 573 (2012).

[26] S. Benramache, B. Benhaoua, Superlattices and Microstructures, 52, 1062 (2012).

[27] K. Subbulakshmi, R. Pandeeswari and B. G. Jeyaprakash, Superlattices and Microstructures, 65, 219 (2013).

[28] S. Benramache, O. Belahssen, A. Guettaf, A. Arif, Journal of Semiconductors, 34, 113001 (2013).

[29] A. Gahtar, S. Benramache, B Benhaoua, F. Chabane, Journal of Semiconductors, 34, 073001 (2013).

[30] B. Benhaoua, A. Rahal, S. Benramache, The Structural, Superlattices and Microstructures, 68, 38 (2014).

[31] F. Chouikh, Y. Beggah, M. S. Aida, Journal of Materials Science: Materials Electronic, 22, 499 (2011).

[32] N. Zebbar, Y. Kheireddine, K. Mokeddem, A. Hafdallah, M. Kechouane, M.S. Aida, Materials Science in Semiconductor Processing, 14, 229 (2011).

[33] S. Ilican, Y. Caglar, M. Caglar, F. Yakuphanoglu, Physica E 35, 131 (2006).

[34] A. Rahal, S. Benramache, B. Benhaoua, Engineering Journal, 18, 81 (2014).

[35] B.L. Zhu, X.H. Sun, X.Z. Zhao, F.H. Su, G.H. Li, X.G. Wu, J. Wu, R. Wu, J. Liu, Vacuum, 82, 495 (2008).

[36] S. Benramache, B. Benhaoua, Superlattices and Microstructures, 52, 807 (2012).

[37] A. Hafdallah, F. Yanineb, M.S. Aida, N. Attaf, In doped ZnO thin films, Journal of Alloys and Compounds, 509, 7267 (2011).

[38] S. Benramache, B. Benhaoua, H. Bentrah, Journal of Nanostructure Chemistry, 3, 54 (2013).

[39] A. Mhamdi, A. Boukhachem, M. Madani, H. Lachheb, K. Boubaker, A. Amlouk, M. Amlouk, Optik, 124, 3764 (2013).

[40] F. Yakuphanoglu, Y. Caglar, S. Ilican, M. Caglar, Physica, B 394, 86 (2007).

A mollification method for a Cauchy problem for the Laplace equation

Djenaoui meriem
Djenaoui.m@hotmail.com
Saidouni sherif
cherifsaidouni@gmail.com

Abstract:

In this paper, we consider a Cauchy problem for the Laplace equation in the strip region o< ξ <1; ψ 2 R, where the Cauchy data is given at ξ = o and the solution is sought in the interval o < ξ · 1. The problem is ill-posed in the sense that the solution (if it exists) does not depend continuously on the data, and a small error in the data can destroys the numerical solution. To solve the problem numerically, a molli¯cation method is considered. Error estimate between the exact solution and its approximation is given. The choices of molli¯cation parameter for a priori and a posteriori are discussed respectively. Numerical examples show that the method works e®ectively.

Key words: Cauchy problem for Laplace equation; Ill-posed problem; Molli¯cation method.

Using Bayesian networks and Naïve Bayes classifier for the prediction of Karenia selliformis occurrences and blooms in the Gulf of Gabès, Tunisia

Wafa Feki-Sahnoun, Hasna Njah, Asma Hamza, Ahmed Rebai, Malika Bel Hassen

Abstract

Bayesian networks are one of the most powerful tools in the design of expert systems located in an uncertainty framework. However, normally their application is determined by the discretization of the continuous variables. In this paper the Bayesian Networks (BN) and the naïve Bayes (NB) models are developed. They are used to identify the most impacting physical (salinity, temperature and tide amplitude), and meteorological parameters (evaporation, air temperature, insolation, rainfall, atmospheric pressure and humidity) on the occurrence and abundance of *Karenia selliformis* using 10-year study of this toxic species at 15 stations along the Gulf of Gabès coast. Therefore, the BN model show that the relationship between salinity and *Karenia selliformis* is more apparent when we focus on the species concentrations and that the bloom occurrences can be predicted based on salinity. The NB model shows that the species occurred mainly in the southern and the central part of the Gulf of Gabès. The shift to the highest salinity level, associated with reduced tide is the most favorable conditions for the species blooms. The cause-and-effect relationships between the physical environment and the *Karenia selliformis* blooms were discussed evoking possible processes leading to blooms.

Keywords—Bayesian network, Naïve Bayes; *Karenia selliformis*; physico-meteorological parameters; Gulf of Gabès.

Bayesian network modeling: A case study of credit scoring analysis of Consumer Loan's default Payment

Lobna Abid[a], Afif Masmoudi[b], Sonia Ghorbel-Zouari[c]

[a] *University of Sfax, Department of Economic Development, Faculty of Economic and Management of Sfax, Tunisia*
lobnabid@yahoo.fr

[b] *Laboratory of probability and Statistics. Faculty of Sciences of Sfax. B.P.1171,CP 3000 Tunisia*
Afif.Masmoudi@fss.rnu.tn

[c] *University of Sfax, Department of Economic Development, Faculty of Economic and Management of Sfax,* Tunisia
sonia.zouari@hotmail.com

Abstract

The current paper deals with the issue of predicting customers' default payment. The Bayesian network credit model is applied for the prediction and classification of personal loan customers with regard to credit worthiness. Relying on information taken from credit experts and using K2 algorithm for learning structure, we constructed the dependency conditional relations between variables explaining default payments. Then, the parametric learning is adopted to detect conditional probabilities of customers' default payment. The parameters are estimated on the basis of real personal loan data obtained from a Tunisian Commercial bank.

The Bayesian network analysis revealed that customers' age, gender, type of credit, professional status, the monthly repayment burden, and credit duration have an important predictive power for the detection of customers' default payment.

Therefore, our findings serve to provide an effective decision support system for banks to detect and alleviate the rate of bad borrowers through the use of a Bayesian Network model.

Keywords:

Bayesian Network; credit scoring; Tunisian Commercial Bank; Consumer credit

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Characterization of the q-exponential distribution

Boutouria Imen[1], Bouzida Imed [2], Masoudi Afif [3]

Keywords: q-calculus, q-gamma, q-beta, q-exponential, q-distribution

ABSTRACT

In this paper we determine the q-beta distribution of the second kind by making a variable change at the constant of normalization, then we generalized the q-gamma distribution.

Based on the q-gamma distribution, we defined the q-exponential distribution. We used the definition of the new q-derivative, we characterized the q-exponential distribution.

We used the Jackson integral and these properties, we introduced the definition of the q-mean, q-variance and q-moments then we calculated for some q-distribution.

[1] imen.boutouria@gmail.com
[2] imed.bouzida@gmail.com
[3] afif.masmoudi@fss.rnu.tn

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Viterbi Algorithm for DNA Sequence Alignment

Afif Masmoudi, Rahma Abid, and Hanen Ben Hassen

Keywords: pairwaise sequence alignment-Baum-Welsh algorithm-Viterbi algoritm

ABSTRACT

Progressive algorithms are widely used heuristics for the production of alignments among multiple nucleic-acid or protein sequences. We present here a new method for pairwaise sequence alignment that combines an HMM approach, a progressive alignment algorithm, and a probabilistic evolution model describing the character substitution process.

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Câblage d'un réseau modélisé par un graphe non orienté

Aymen Hassine 1[1,*], Afif Masmoudi 1[2] , Abdelaziz Ghribi 2

Keywords: Arbre couvrant minimal; Câblage d'un réseau.

Résumé

Dans ce travail, on se propose de décomposer un graphe non orienté **G** en un nombre minimal de sous graphes connexes, dont chacun d'eux (noté ***C***) vérifie les deux contraintes suivantes:

- **Attenuation:** $\exists s_0 \in C / dist(s_0, s) \leq R_{\max}$ **,** $\forall s \in C.$
- **Capacité:** $Card(C) \leq N_{\max}$ **.**
  Pour ce faire, on utilise l'algorithme de Kruskal pour déterminer l'arbre couvrant minimal assurant le câblage minimal entre les sommets de **G**. Dans le cadre du dimensionnement d'un réseau de fibre optique, nous avons proposé une méthodologie qui consiste à déterminer le câblage "optimal".

[1] Unité de probabilité et statistique de Sfax

[2] Laboratoire de recherche Physique, Mathématiques et Applications

Aymen Hassine: aymen_hassine@yahoo.fr

[1] Unité de probabilité et statistique de Sfax

[2] Laboratoire de recherche Physique, Mathématiques et Applications

Aymen Hassine: aymen_hassine@yahoo.fr

INTERNATIONAL CONFERENCE ON BAYESIAN NETWORKS AND APPLICATIONS
SOUSSE, OCTOBER 14TH-16TH, 2016

Impact of a priori MS/MS intensity distributions on database search for peptide identification

Hatem Loukil, Mohamed Tmar

ABSTRACT

Many database search methods has been developed for peptide identification throughout a large peptide data set. Most of these approaches attempt to build a decision function that allows the identification of an experimental spectrum. This function is either built starting from similarity measures to the database peptides to identify the most similar one to a given spectrum, or by applying useful learning techniques considering the database itself as a training data. In this paper, we propose a peptide identification method based on a similarity measure for peptide-spectrum matches (PSMs). Our method takes into account peak intensity distributions in its probabilistic scoring model to rank peptide matches. The main goal of our approach is to highlight the relationship between peak intensities and peptide cleavage positions and to show its impact on peptide identification. To evaluate our method, a set of experiments have been undertaken into two high mass accuracy data sets. The obtained results shows the effectiveness of our approach.

ICBNA 2016 aims at bringing together a wide range of researchers, practitioners and graduate students whose work is related to the Bayesian networks and their applications in real-world problems. This edition intends to cover other related topics.

Contact

E-mail : internationalconferencebna@gmail.com
Website : http://sites.ieee.org/tunisia-embs/icbna-2016