# Incremental Learning with Deep GMDH Neural Network for Data Stream Mining

Panida Lorwongtrakool
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
panidajlo@gmail.com

Phayung Meesad
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

***Abstract*— This research aims to create an incremental learning with Deep Group Method of Data Handling (GMDH) Neural Network for data stream mining. The signal derived from the e-nose and e-tongue consists of different sensors for water quality classification problem. The system mainly consists of three stages: 1) preparation of data stream input, 2) processing with Deep GMDH Neural Network by using Polynomial function to generate Partial Descriptions (PDs), estimate the coefficients of the PD and select the PD with the best predictive capability, 3) batch incremental learning by updating weight polynomial matrix for processing the next chunk of data. The results showed that the most accurate models with the maximum of layers had 10 layers and each layer had a maximum of 10 nodes, a batch size 800 and training data set 80% with accuracy 90.71%. Therefore, system can be applied to Data stream mining and monitoring system in the real environment.**

***Keywords—Increment learning, GMDH, Data stream Mining, Deep learning***

## I. INTRODUCTION

Nowadays, the machine learning techniques such as neural networks provide acceptable accuracy and are widely used for solving a lot of Data Stream Mining tasks[1]. However, the performance of conventional algorithms (such as ANN, SVM) depends on the design of network and features [2]. In addition, it is time-consuming when operated with high dimensional data or Big Data, and it is difficult to guarantee the global convergence. Thus, the shallow neural neurons are not suitable for data streams processing. [1] , Therefore, the deep network was introduced by Hinton [3].

In recent years, Computational Intelligence researcher are interested in deep neural networks (DNN), and it became practically feasible to solve this problem. Among a great number of possible deep neural networks' architectures, the deep networks based on GMDH are one of the most effective networks[4]. The networks are based on the group method of data handling [5] ,which automatically increases a number of layers for information processing to achieve the required accuracy of results.

The Group Method of Data Handling (GMDH) is also known as Polynomial Neural Networks, Abductive and Statistical Learning Networks. The GMDH was applied in a great variety of areas for deep learning and knowledge discovery, forecasting and data mining, optimization and pattern recognition. Inductive GMDH algorithms give possibility to find automatic interrelations in data, to select an optimal structure of model or network and to increase the accuracy of existing algorithms.

However, one relevant problem is "catastrophic forgetting" [6] that may occur when a network, trained with a large set of patterns, has to learn new input patterns, or has to be adapted to a different environment. The risk of catastrophic forgetting is particularly high when a network is adapted with new data that do not adequately represent the knowledge included in the original training data.[5] The solution for this problem is adding a new ability to classifiers. Having the incremental learning can be of great benefit by automatically including the newly presented patterns in the training dataset without affecting class integrity of the previously trained system. [7]

## II. METHODOLOGY

### A. Collection of Samples

Efficiency of the proposed algorithm was examined based on solving water quality classification problems. The samples of water were collected from Din Daeng water quality control plants. The data were collected by using e-nose and e-tongue consisting of sensors: MQ2, MQ3, MQ4, MQ5,MQ6, MQ7, MQ8, MQ9, MQ135, pH, EC, TDS, salinity, DO, temperature and turbidity. The samples were collected before water flew (inlet area) to the water quality control plants, and the data were collected at the area where water flew out (outlet area) by using a 800cc bottle to collect the data. In addition, a headspace was placed above the water surface, approximately 5 centimeters, to measure a response of the sensor to a smell of water sample.

### B. Data preparation

Data preparation: the data were prepared by cleaning and removing noisy data , outlier and normalize data to 0-1.

### C. Framework

The conceptual framework of algorithm is shown in Fig. 1.

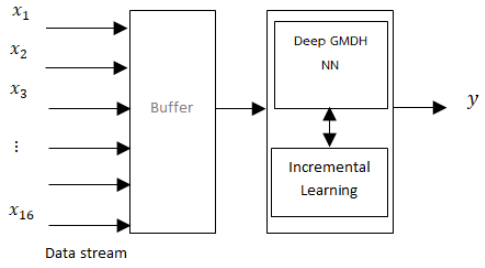


Fig. 1. The architecture of the proposed Incremental Learning with Deep GMDH Neural Network

978-616-368-065-5 ©2018 JSCI

**Algorithm** Incremental learning with Deep GMDH Neural Network.

*Input:load datastream($x_1, \ldots, x_n$)*

*1. Begin:*

*2. chunk_size ← 200,400,600 and 800(records)*

*3. for i←1 to chunk_size*

*4. set data.Inputs*

*5. set data.Targets*

*6. Innitial Train Data*

*7. Innitial Test Data*

*8. Calculate GMDH Network*

*9. Calculate FitPolynomial*

*10. Update weights(W)*

*11. for j←1 to numOf chunk_size*

*12.*

$$W_{x_i,L_{n(j)}new} \leftarrow \frac{(W_{x_i,L_{n(j)}old}(C_{J,new} - 1) + W_{x_i,L_{n(j)}new})}{C_{J,new}}$$

*13. endfor j*

*14. endfor i*

*15. end*

## III. RESULTS

The performance of different models is evaluated by accuracy, RMSE and time. The results are shown in the Table 1

TABLE I. RESULTS OF MODEL PERFORMANCE TESTING

| **Maximum Number of Neurons in a Layer = 5<br>Maximum Number of Layers = 5** | | | | |
|---|---|---|---|---|
| **Chunk Size** | **Train(%)** | **Accuracy(%)** | **RMSE** | **Time(sec.)** |
| 200 | 50 | 83.51 | 0.36 | 23.56 |
| | 70 | 85.32 | 0.32 | 27.13 |
| | 80 | 85.64 | 0.31 | 26.49 |
| 400 | 50 | 83.51 | 0.34 | 11.69 |
| | 70 | 86.71 | 0.32 | 12.37 |
| | 80 | 84.96 | 0.33 | 12.39 |
| 600 | 50 | 87.53 | 0.33 | 8.15 |
| | 70 | 85.22 | 0.33 | 8.32 |
| | 80 | 84.33 | 0.33 | 7.54 |
| 800 | 50 | 83.67 | 0.34 | 4.99 |
| | 70 | 84.22 | 0.36 | 5.68 |
| | 80 | 85.75 | 0.33 | 5.35 |
| **Maximum Number of Neurons in a Layer = 10<br>Maximum Number of Layers = 10** | | | | |
| **Chunk Size** | **Train(%)** | **Accuracy(%)** | **RMSE** | **Time(sec.)** |
| 200 | 50 | 89.48 | 0.30 | 28.50 |
| | 70 | 88.63 | 0.30 | 32.63 |
| | 80 | 89.73 | 0.30 | 33.94 |
| 400 | 50 | 88.90 | 0.30 | 14.26 |
| | 70 | 89.08 | 0.30 | 17.59 |
| | 80 | 90.03 | 0.30 | 19.04 |
| 600 | 50 | 88.43 | 0.31 | 9.4 |
| | 70 | 88.00 | 0.29 | 10.69 |
| | 80 | 87.62 | 0.31 | 10.44 |
| 800 | 50 | 88.98 | 0.30 | 6.26 |
| | 70 | 89.37 | 0.29 | 7.60 |
| | 80 | **90.71** | 0.29 | 7.42 |
| **Maximum Number of Neurons in a Layer = 12** | | | | |
| **Maximum Number of Layers = 12** | | | | |
| **Chunk Size** | **Train(%)** | **Accuracy(%)** | **RMSE** | **Time(sec.)** |
| 200 | 50 | 89.33 | 0.29 | 38.12 |
| | 70 | 90.05 | 0.30 | 41.84 |
| | 80 | 88.78 | 0.29 | 41.00 |
| 400 | 50 | 89.07 | 0.29 | 20.20 |
| | 70 | 89.16 | 0.29 | 20.51 |
| | 80 | 90.23 | 0.29 | 25.55 |
| 600 | 50 | 89.25 | 0.28 | 12.10 |
| | 70 | 88.22 | 0.29 | 13.97 |
| | 80 | 89.83 | 0.29 | 14.34 |
| 800 | 50 | 89.28 | 0.30 | 8.68 |
| | 70 | 88.06 | 0.27 | 9.46 |
| | 80 | 88.57 | 0.30 | 9.35 |

According to the Table 1, it showed that the model with batch size was 800, the training dataset was 80%, the maximum number of neurons in a layer was 10, the maximum number of layer was 10, and the confirmed accuracy was up to 90.71%

It is noted that the number of neurons, number of layers and size of training dataset will affect the accuracy. An increase to these numbers will affect accuracy more. On the other hand, if there are too many or too few, it may cause a decrease in accuracy due to overfitting and underfitting, respectively.

According to the results, it can be concluded that the system can be applied to data stream mining and monitoring system in the real environment since the system has the ability to learn and is adaptive when faced with unseen data.

## REFERENCES


[1] Bodyanskiy, Y., et al. Fast learning algorithm for deep evolving GMDH-SVM neural network in data stream mining tasks. in 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). 2016.

[2] Liu, W., et al., *A survey of deep neural network architectures and their applications.* Neurocomputing, 2017. **234**: p. 11-26.

[3] Hinton, G.E., S. Osindero, and Y.-W. Teh, *A fast learning algorithm for deep belief nets.* Neural Comput., 2006. **18**(7): p. 1527-1554.

[4] Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural Networks, 2015. **61**: p. 85-117.

[5] Ivakhnenko, A.G., *Polynomial Theory of Complex Systems.* IEEE Transactions on Systems, Man, and Cybernetics, 1971. **SMC-1**(4): p. 364-378.

[6] French, R.M., *Catastrophic forgetting in connectionist networks.* Trends in Cognitive Sciences, 1999. **3**(4): p. 128-135.

[7] Tudu, B., et al., *Electronic nose for black tea quality evaluation by an incremental RBF network.* Sensors and Actuators B: Chemical, 2009. **138**(1): p. 90-95.