SNP selection for Porcine breed classification by a hybrid information gain and genetic algorithm

Wanthanee Rathasamuth*, Kitsuchart Pasupa†, and Sissades Tongsima‡
*Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
†National Center for Genetic Engineering and Biotechnology (BIOTEC),
National Science and Technology Development Agency (NSTDA), Pathum Thani 12120, Thailand
Email: *rathasamuth.wan@gmail.com, †kitsuchart@it.kmitl.ac.th, ‡sissades@biotec.or.th

***Abstract*—Single Nucleotide Polymorphism (SNP) is a variability of DNA sequence that connects to a unique trait of an organism. A good SNP selection can provide a good porcine breed that grows fast with high yield. SNP selection can be done by a computerized feature selection method and classification technique. At present, an effective classification model can only handle a small number of features efficiently. Too large a number may cause an over-fitting problem in the classification. Therefore, SNPs or features need to be reduced to an optimum number for an effective porcine SNP analysis. This paper proposes an approach to reducing the number of features in porcine SNP analysis to an optimum number by a hybrid of Information Gain (IG) and genetic algorithm (GA) techniques. A performance test demonstrated that this approach was able to select a minimum number of features (at 1.51% of the total number of features) that provided an average classification accuracy of 94.02%, as compared to 95.28% provided by the total number of features.**

***Index Terms*—Feature selection, Bioinformatics, Machine learning, Single Nucleotide Polymorphisms.**

I. Introduction

China might be the first country that has started to selectively bred wild pigs 5,000 years ago. Pig breeding in Thailand was heavily influenced by the Chinese immigrants to this country. Today, pigs are an important economic animal in Thailand, hence selecting the right breed for the geographical location is a very important issue. Diverse physical traits of pigs are the results of the differences in DNA base sequences which are called single nucleotide polymorphism (SNP). A thorough porcine SNP analysis can determine the SNPs that provide good growth and reproduction. The issue is that there are millions of SNPs for a single organism, and so a manual SNP analysis by an expert is out of the question, not to mention the huge amount of other kinds of resources needed. Today, a good way to address this issue is to use bioinformatics, an integration of computer science, biology, mathematics, and engineering. Machine learning [1] has been applied to genomics, proteomics, microarray, and system biology for classification of genes. In [1], several classification techniques for bioinformatics are presented such as support vector machine (SVM), decision tree, neural networks, Bayesian classifiers, and nearest neighbors. Since these classification techniques cannot effectively support too large a number of features that may cause a commonly encountered over-fitting problem–high accuracy when used with training dataset but low accuracy with testing dataset, reducing the number of features into a subset of optimum features can make a classification attempt successful. Papers that deal with this issue are such as [2] which is a review of feature selection applying to bioinformatics. The paper reports three types of feature selection: filter methods such as i-test, and information gain (IG); wrapper methods such as genetic algorithms (GA) and other nature inspired algorithms; and embedded methods such as random forest, and decision tree. In [3], a review of several nature-inspired algorithms that were used to perform feature selection was presented. The main point was how to increase selection efficiency and reduce prediction error. The most common problems found were too large a number of features and too small a training dataset. These are some of the challenges in doing a classification analysis.

The conceptual frameworks of feature selection by filter, wrapper, and embedded methods are quite different. Simple and efficient, filter methods select features that have high index values, independent of the classification method. Wrapper methods rely on the classification method to select an optimal subset of features that provide high classification accuracy–each round of feature evaluation includes a classification step; therefore, a large number of features results in high computation time. Embedded methods are not very different from wrapper methods but include a feature reduction step that reduces their computation time. Wrapper and Embedded methods rely on a classification step to select an optimal subset of features hence the selected features can facilitate better learning of training dataset but their predictions may suffer from an over-fitting problem. On the other hand, filter methods tend to have less over-fitting problem.

This paper proposes using a hybrid feature selection technique that combines IG with GA (IG+GA) for the purpose of selecting the best porcine SNPs for classification of pig breeds. The next section presents a description of datasets, followed by experimental framework and its results.

II. Datasets

The dataset used in this study consists of SNP data from 677 pig samples of 22 breeds, 356 samples from the dataset of porcine colonization of the Americas [4] and 321 samples from the dataset of the Project of Porcine Breed Improvement by Selection according to Whole Genome SNP Data supported

978-616-455-375-0 ©2018 JSCI

by the National Center for Genetic Engineering and Biotechnology (BIOTEC), for a total of 16,579 SNPs. All of the data had been through data cleansing already. However, there were some missing values that were estimated by a single imputation method that replaced the values with the mode of the whole individual feature data. The combined dataset was random-seeded into 10 datasets; each dataset contains a training set (80% of the data) and a test set (20% of the data).

III. Experimental Framework

Among all features, some features may not exert any significant influence on the constructed learning model, hence they can only waste computation time; therefore, it is essential to select only significant features that strongly influence the learning model that can make accurate prediction. The experiments in this study tested the comparative feature-selection performances of IG, GA, and IG+GA when used with a linear-kernel SVM classifier. In this study, the hyperparameter of SVM, C, was specified to be in the range of $10^{-6} \sim 10^{6}$, and five-fold cross-validation was used to obtain an optimal parameter to construct an acceptably-accurate model.

IG is a feature selection technique of the filter type [2] that calculates feature indexes from the relationships between features. It has been successfully applied to selection of text, microarray, and SNPs yielding a small set of significant features. GA is another feature selection technique that can reduce the number of features to a small subset of significant features, but it has a problem of getting trapped at local optimums. IG+GA was proposed by a previous study [5] and demonstrated to provide a smaller subset of significant features that were able to give more accurate predictions than IG or GA alone. In this study, we used IG to rank the significant levels of features, then used an elbow method to find the cut-point for feature selection. This cut-point determined the number of genes of each chromosome to be constructed in GA. For instance, if the cut-point was 300, the number of constructed genes in each GA chromosome would be a number no higher than 300. The average cut-point we found by IG from our datasets was 409, so we roughly specified the number of genes in GA as 400. GA was used to select features within the range of the first 400 ranked features obtained from IG. The population size of GA was 20 chromosomes. The selection method in GA was a roulette wheel method. The crossover step was a multi-point crossover with a crossover probability of 0.8. The mutation step was a bit-flip mutation, and the maximum number of generations was 10.

To remedy the problem of solutions getting trapped at local optimums, we set the mutation probability of a '1 to 0' flip to 0.3 and that of '0 to 1' flip at 0.7. This strategy of setting different mutation probabilities for different kinds of bit-flipping was proposed in [6] for avoiding local optimums.

IV. Experimental Results

Results of the reduction of number of features by IG, GA, and IG+GA are presented in Table I. In the same table, accuracies in porcine breed classification by using the features from these 3 methods and from using the whole original features are presented.

TABLE I
Comparative results of average classification accuracies and the number of selected SNP.

Methods	Accuracy (%)	#SNP
All Features	95.28 ± 1.23	$16,579$
IG	93.96 ± 0.60	409 ± 0.09
GA	94.80 ± 1.40	$2,357.9 \pm 3.20$
IG+GA	94.02 ± 1.19	250.1 ± 0.01

It can be seen that every approach to feature selection that we had attempted gave nearly the same value of average classification accuracy: the using of the whole original SNPs approach yielded an accuracy of 95.28% while the using of the features from GA selection alone yielded the highest accuracy at 94.80%; that from IG alone yielded 93.96%; and that from IG+GA yielded 94.02%. The average percentages of the number of SNPs reduced were 85.78%, 97.53%, 98.49% achieved by GA, IG, and IG+GA, respectively. The approach with GA alone was more accurate but used a larger number of SNPs than the IG alone and IG+GA approaches, while the IG+GA approach yielded nearly the same accuracy as that provided by GA but used significantly fewer SNPs.

V. Conclusion

SNP selection for porcine breed classification can be done by several feature selection methods and classifiers. This study used IG alone, GA alone, and IG+GA approaches to select an optimal set of SNPs for SVM to classify and found that the IG+GA approach not only was able to reduce the highest number of features, at 98.49%, but also provided a classification accuracy that was very nearly equal to the highest accuracy provided by the GA alone approach.

References

[1] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[3] H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 142–148.

[4] W. Burgos-Paz, C. A. Souza, H. J. Megens, Y. Ramayo-Caldas, M. Melo, C. Lemús-Flores, E. Caal, H. W. Soto, R. Martínez, L. A. Álvarez, L. Aguirre, V. Iñiguez, M. A. Revidatti, O. R. Martínez-López, S. Llambi, A. Esteve-Codina, M. C. Rodríguez, R. P. M. A. Crooijmans, S. R. Paiva, L. B. Schook, M. a. M. Groenen, and M. Pérez-Enciso, "Porcine colonization of the Americas: A 60k SNP story," *Heredity*, vol. 110, no. 4, pp. 321–330, 2013.

[5] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm," in *Proceeding of the 2012 International Conference on Computer Science and Electronics Engineering*, 2012, pp. 355–358.

[6] G. Mahdevar, J. Zahiri, M. Sadeghi, A. Nowzari-Dalini, and H. Ahrabian, "Tag SNP selection via a genetic algorithm," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 800–804, 2010.