An Efficient Distributed SNP Selection Method for Porcine Breed Classification

Wanthanee Rathasamuth
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
rathasamuth.wan@gmail.com

Kitsuchart Pasupa
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kitsuchart@it.kmitl.ac.th

***Abstract*—In principle, a porcine Single Nucleotide Polymorphism (SNP—a specific piece of nucleotide in a DNA sequence) can be associated with a trait of an individual pig, like its meat quality or resistance to common diseases. It is most desirable to obtain a smallest number of most significant SNPs in genomic research and several computer classification algorithms have been used to find a small number of SNPs. This study proposed a vertically distributed feature selection method incorporating a modified binary flower pollination and a support vector machine classifier for selecting significant porcine SNPs. The proposed method was evaluated and compared against four baseline methods. It provided a mean number of 128.4 selected SNPs that resulted in 94.57% classification accuracy.**

***Index Terms*—Single nucleotide polymorphism, Feature selection, Flower pollination algorithm, Support vector machine**

I. Introduction

An important goal of researches on livestock rearing is to identify genetic variations underlying economically important traits such as traits in reproduction, in disease immunity, and in meat quality. A Single Nucleotide Polymorphism (SNP) is a specific piece of nucleotide in a DNA sequence associated with a certain trait of a living being. Association study and genomic prediction, two types of genetic variation study, have been performed using SNP array [1]. However, since the number of SNPs of a living being is very large and the detection process cannot be automated fully even with the help of a computer system, a genetic variation study requires a lot of time and money. Therefore, in a study related to feature selection, reduction of the number of features (SNPs) can help make an investigation feasible. Recently, a lot of feature selection methods have been used in microarray data classification [2], in SNP data classification [3]–[6], and in proteomic data analysis [7] and other bioinformatics tasks. A paper by Jović *et al*. [8] presents a review of feature selection techniques used in several fields of study including bioinformatics.

Since the data in a genomic, proteomic, SNPs, and microarray study are high-dimensional, they present a challenge to computer researchers. In the machine learning field [9], [10], feature reduction techniques have been widely used to discard duplicated or unrelated features. The reduced number of features not only improves the classification accuracy and reduces the model's learning time, but also alleviates an overfitting problem. An overfitting problem occurs when a model with a large number of features has been successfully trained on a training data set but performs poorly on a test data set.

Three types of feature selection methods have been widely used: filter, wrapper, and embedded methods [8], [10]. As mentioned, feature selection is essential in machine learning. Therefore, a chosen feature selection method must be suitable for the machine learning task, especially a task with high dimensional data. Filter methods select features based on performance measure. Features are ranked independently of data classifier algorithm. After the best feature subset was found, the features in it are sent to a classifier algorithm. On the other hand, wrapper methods select features based on a classifier algorithm such as support vector machine (SVM) [3]–[5], [9], neural network [9], and nearest neighbor [9]. The process of finding the best feature subset is repeated in many iterations, and so wrapper methods require more processing time than filter methods. Embedded methods, however, rank features while the classifier algorithm is running—the feature selection process is embedded in the algorithm.

Our aim was to develop a better feature selection method that could select a small number of porcine SNPs for classification. Even though feature selection by filter methods is easy to perform and rapid, the subset of selected features is not evaluated by a classifier algorithm, so the features are not confirmed whether they provide a good classification accuracy or not. Paper [4] presented a hybrid information gain [IG] and binary flower pollination algorithm (BFPA), a wrapper method, to successfully classify porcine SNPs. The selected features from BFPA were evaluated by an SVM classifier before they were included in the final subset of selected features. This evaluation by a classifier before screening out features had an important advantage of not screening out any significant SNPs even if they were lower ranks. That work [5] combined IG and genetic algorithm (GA) to select significant SNPs. Whereas the GA step in the method was capable of select significant SNPs that were in the lower ranks, the IG step in could screen out some SNPs that were significant, which was not our objective. Therefore, in this study, we used a distributed feature selection (DFS) method that incorporated a wrapper method, BFPA, to enable every significant feature to be selected as well as to reduce the processing time (by

distributing the computation into parallel streams).

II. Related works

This section briefly describes research works on DFS technique as well as their application in our work.

Papers [11] and [12] presented a vertically distributed feature selection technique for high-dimensional microarray data. Bolón-Canedo *et al*. [11] presented a distributed filter method for improving the classification accuracy on microarray data as well as to reduce processing time. Each vertical partition had a number of samples and a number of features that was half of it. The features in each partition was ranked and selected into a best feature subset by a classification algorithm. Potharaju and Sreedevi [12] have used symmetric uncertainty in primary feature selection, discarding unrelated features. The outcome of the correlation-based feature subset selection determined the number of features contained in each partition. Finally, they used multi-layer perceptron to evaluate the feature subset of each partition and find the best feature subset. Feature selection can be vertically or horizontally, Morán-Fernández *et al*. [13] suggested DFS based on complexity measure for partitioned data. They used several filter methods on 11 datasets. A horizontally distributed method was suitable for a data set that have a small number of features but a sufficiently large number of samples. On the other hand, a vertically distributed method was suitable for data that have a large number of features but a small number of samples.

Our contribution was in proposing a vertically distributed feature selection method that incorporated a modified BFPA algorithm [4] and an SVM classifier.

III. Methodology

This study used a vertically distributed feature selection method with modified BFPA [4] and SVM classifier to select porcine SNPs. This method was suitable to our considered data set that had a small number of samples but a large number of features. We followed Morán-Fernández *et al*. [13] in the use of this method. This section describes binary flower pollination algorithm, support vector machine, then our proposed method.

A. Binary flower pollination algorithm

The original FPA [14] was inspired by the pollination process of flower plants. There are two major forms of pollination: abiotic and biotic. Pollination is classified as cross-pollination or self-pollination. Cross-pollination is pollination of flowers from different plants, whereas self-pollination is pollination of flowers from the same plant. For cross-pollination, FPA acts like pollinators traveling over a long distance, moved in random walk according to Lévy distribution. It can be considered a global pollination.

Yang [14], the originator of FPA, described the pollination behaviors of FPA as follows: biotic and cross-pollination were considered as a global pollination process with pollen-carrying pollinators performing Lévy flights; local pollination is abiotic self-pollination; Flower constancy can be considered as a reproduction probability that is proportional to the similarity between the two flowers involved; Switching between local pollination and global pollination is controlled by a switching probability $p \in [0, 1]$ which is slightly biased toward local pollination.

Local pollination takes place when a generated random number is less than a switch probability p represented by (1).

$$x_i^{t+1} = x_i^t + \epsilon(x_k^t - x_l^t), \tag{1}$$

where x_i^t is an individual x_i of the population in iteration t; x_k^t and x_l^t are pollens from different flowers k and l of the same plant species; ϵ is a generated random number $\in [0, 1]$.

Global pollination is expressed as (2).

$$x_i^{(t+1)} = x_i^t + \alpha L(\lambda)(g_* - x_i^t), \tag{2}$$

Lévy distribution is given by (3).

$$L(\lambda) = \frac{\lambda \cdot \Gamma(\lambda) \cdot \sin(\lambda)}{\pi} \cdot \frac{1}{s^{1+\lambda}}, s > 0, \tag{3}$$

where $L(\lambda)$ is a Lévy flight distribution; $\Gamma(\lambda)$ is the standard gamma function, valid for large steps $s > 0$; λ = 1.5; α is a scaling factor for controlling step size; s is step size; and g_* is the current best individual.

Rodrigues *et al*. [15] presented a BFPA for feature selection. Each element in an individual flower of BFPA was assigned a binary value. The initial population was assigned binary elements. However, as (1) and (2) were applied, the individual flowers or solution vectors became vectors of continuous elements. Hence, (4) was applied to those elements, converting them back to binary elements.

$$x_i^j(t) = \begin{cases} 1 & , S(x_i^j(t)) > r \\ 0 & , Otherwise \end{cases}, \tag{4}$$

where $x_i^j(t)$ is pollen j of flower x_i in iteration t; r is a random number $\in [0, 1]$; and S is a sigmoid function.

We employed a switch probability, as shown in (5), where σ is a cut-off-point-finding threshold [4], because assigning a 0 or 1 value to each element of a solution vector according to this cut-off point was more efficient than assigning one of these values by computing (4),

$$x_i^j(t) = \begin{cases} 1 & , x_i^j(t) \geq \sigma \\ 0 & , Otherwise \end{cases}, \tag{5}$$

In addition to that employment, the proposed method included a GA bit-flip mutation [4], expressed by (6), that modified basic GA and improved feature selection efficiency as well as provided only a small number of features. Namely, a pollen $x_i^j(t)$ calculated from (1) through (5) had a much greater chance to be flipped to a value of 0 than to a value of 1. This modification resulted in an SNP having a greater chance of not getting selected than getting selected unless it was really significant.

$$x_i^j(t) = \begin{cases} 1 & , r \leq P_m \\ 0 & , Otherwise \end{cases}, \tag{6}$$

where P_m is a mutation probability, and r is a random number $\in [0, 1]$.

B. Support vector machine

SVM is an effective, supervised learning classifier for problems with high dimensions [9]. The idea behind SVM is to put data into a feature space then determine the hyperplane with the highest margin that separates data points into two classes in that space. The data points from which the hyperplane is constructed are called support vectors. SVM can have one of many kinds of kernel functions. This study used a linear kernel. The mathematical expression for the linear kernel is as expressed by (7).

$$k(x, x') = x^T x', \tag{7}$$

where $k(x, x')$ is a kernel function; x and x' are porcine SNP samples. The SVM that we used had a hyperparameter C that balanced training error and model's complexity.

C. Proposed method

The proposed method combined a distributed feature selection method with a modified BFPA method (see Algorithm 1 in [4]) as well as a new population creation step in modified BFPA. The conceptual framework of the proposed method is shown in Fig. 1. It should be noted that in a primary trial, we used only the distributed modified BFPA and found that the number of selected features were still too high compared to the number in [4] which had achieved with IG+modified BFPA. Therefore, we added a new population creation step to the modified BFPA in order for the method to select a fewer number of features.

The new population creation step came after a best new solution was discovered. Its fitness value was checked whether it was identical to the values achieved in the last four consecutive iterations. If so, a new individual was randomly generated under a specified P_m. This individual would replace a randomly selected original individual in the population. Further steps followed the modified BFPA procedure. As the assigned number of generations was reached, the method stopped and provided the best feature subset.

After the entire data set had been partitioned into training data sets and test data sets, the indices of all features in a training data set were randomized in order to make sure that no features would be systematically selected because of some biases from the ordering of the features. The features in the training data set were divided into several groups where each group had the same number of features as the number of samples. We followed Morán-Fernández *et al*. [13] in such division of features into groups. In their work, each group of features contained a number of features that was a half of the number of samples. This assignment suited their objective of selecting the smallest number of features, like the objectives of most genomic classification works that used a filter method. Paper [3] used a wrapper method that already selected around 50% of features in the initialization step. In addition, a feature subset in our work would not contain any

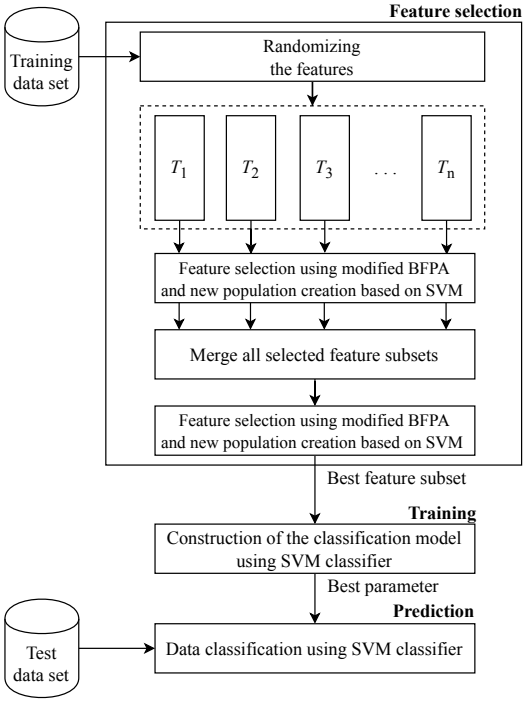

Fig. 1. Conceptual framework of the distributed feature selection method.

identical features to the ones in any other feature subsets. A training data set T was composed of n partitions (n is a round-down integer of the total number of features divided by the number of samples in the training data set). In other words, $T = \{T_1, T_2, T_3, \ldots, T_n\}$, where each T was selected by the modified BFPA. The fitness of every individual in the population was evaluated by five-fold cross-validation to find an optimum C parameter for SVM classifier. The objective function was as in (8),

$$Fit(P) = \left(\frac{sf}{tf} \times w_1\right) + \left(\frac{1}{Acc + \rho} \times w_2\right), \tag{8}$$

where $Fit(P)$ is the fitness value of an individual in population P; sf is the number of selected features in population P; tf is the total number of features; Acc is classification accuracy, ranging from 0-100; and ρ is a small real number; w_1 and w_2 are weights.

The feature subsets from every T were then combined and selected by the modified BFPA for the last time, and the final and best feature subset was generated. They were inputted into the training process to find the best parameter for SVM, the classifier algorithm. That best parameter was used to check the validity of the whole model, modified BFPA+SVM, on the test data set in the prediction phase. To conclude, we ran the proposed method 10 times—one time for each pair of the partitioned training and test data sets.

IV. Data Set

The entire porcine data set used in this study had 676 samples of 21 breeds with 10,210 SNPs [4], [5]. It is available for

download at https://github.com/dsmlr/th-vn-us-swine. In this data set, the data have already been processed through a quality control procedure with a PLINK computer program. Some missing values were mode estimated by a single imputation method.

V. Experimental setup

Ten training datasets and ten test datasets were constructed from random seeds of the porcine SNP dataset. Each one of the training datasets contained 80% of the entire porcine SNP data set, while each of the test sets contained 20%. The parameter settings of modified BFPA were as follows: population size of 30 individuals; number of generations of 200; mutation probability of 0.3; σ of 0.7; α of 1; p of 0.8; C ranging from 10^{-6} to 10^{6}; w_1 of 0.01; and w_2 of 0.99.

VI. Results

The results of SNP selection and swine breed classification are presented in this section. We compared five feature selection methods—original BFPA, distributed original BFPA, distributed BFPA+mutation, distributed modified BFPA, and the proposed method. The results, the mean number of selected SNPs obtained from every method and the mean classification accuracy obtained by using the features selected by each method, are shown in Table I. The distributed original BFPA provided highest classification accuracy (96.67%), followed by distributed BFPA+mutation (96.36%), original BFPA (95.66%), distributed modified BFPA (94.96%), and the proposed method (94.57%). Even though the other four methods provided a slightly higher classification accuracy, the proposed method was able to select the smallest number of SNPs that we aimed to obtain. The distributed original BFPA selected a mean number of 2,199.00 SNPs; the distributed BFPA+mutation selected a mean number of 428.00 SNPs; the original BFPA selected a mean number of 4,993.60 SNPs; the distributed modified BFPA selected a mean number of 312.20 SNPs On the other hand, the proposed method selected a mean number of 128.40 SNPs. Therefore, the proposed method fulfilled our aim better than all of the other methods while providing a sufficient classification accuracy.

TABLE I
Average numbers of selected porcine SNPs and associated accuracy values provided by five selection methods.

Method	Accuracy (%)	#SNP
Original BFPA	95.66	4,993.60
Distributed original BFPA	96.67	2,199.00
Distributed BFPA+mutation	96.36	428.00
Distribute modified BFPA	94.96	312.20
Proposed method	94.57	128.40

VII. Conclusion

We propose a porcine SNPs classification method by a distributed feature selection method incorporating a modified Binary Flower Pollination Algorithm (BFPA) and a support vector machine classifier. We also introduced a new population creation step into the modified BFPA for the solutions to avoid getting trapped at local minima. The proposed method was evaluated and compared against four other methods: original BFPA, distributed original BFPA, distributed BFPA+mutation, and distributed modified BFPA. The proposed method was able to select the minimum mean number of features, 128.4 SNPs that provided a mean accuracy of 94.57%.

References

[1] A. Al-Chalabi, L. H. van den Berg, and J. Veldink, "Gene discovery in amyotrophic lateral sclerosis: implications for clinical management," *Nature Reviews Neurology*, vol. 13, no. 2, pp. 96–104, 2017.

[2] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Advances in Bioinformatics*, vol. 2015, no. 198363, 2015.

[3] W. Rathasamuth, K. Pasupa, and S. Tongsima, "Selection of a Minimal Number of Significant Porcine SNPs by an Information Gain and Genetic Algorithm Hybrid," *Malaysian Journal of Computer Science*, vol. 32, pp. 79–95, 2019.

[4] W. Rathasamuth and K. Pasupa, "A Modified Binary Flower Pollination Algorithm: A Fast and Effective Combination of Feature Selection Techniques for SNP Classification," in *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2019, pp. 1–6.

[5] K. Pasupa, W. Rathasamuth, and S. Tongsima, "Discovery of significant porcine SNPs for swine breed identification by a hybrid of information gain, genetic algorithm, and frequency feature selection technique," *BMC Bioinformatics*, 2020.

[6] E. Thamwiwatthana, K. Pasupa, and S. Tongsima, "Selection of SNP Subsets for Severity of Beta-thalassaemia Classification Problem," in *Proceeding of the 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018), 10-13 December 2018, Bangkok, Thailand*, 2018, pp. 1–7.

[7] M. Lualdi and M. Fasano, "Statistical analysis of proteomics data: A review on feature selection," *Journal of Proteomics*, vol. 198, pp. 18–26, Apr. 2019.

[8] A. Jović, K. Bogunovi´c, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1200–1205.

[9] 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," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[10] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[11] V. Bolón-Canedo, N. Sánchez-Maroño Universidade da Coruña, and A. Alonso-Betanzos, "Distributed feature selection: An application to microarray data classification," *Applied Soft Computing*, vol. 30, pp. 136–150, 2015.

[12] S. P. Potharaju and M. Sreedevi, "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance," *Clinical Epidemiology and Global Health*, vol. 7, no. 2, pp. 171–176, 2019.

[13] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, "Centralized vs. distributed feature selection methods based on data complexity measures," *Knowledge-Based Systems*, vol. 117, pp. 27–45, 2017.

[14] X.-S. Yang, "Flower Pollination Algorithm for Global Optimization," in *Proceedings of the 11th International Conference on Unconventional Computation and Natural Computation (UCNC 2012), Orléan, France, September 3-7, 2012*, ser. Lecture Notes in Computer Science, vol. 7445. Springer, Berlin, Heidelberg, 2012, pp. 240–249.

[15] D. Rodrigues, X.-S. Yang, A. N. de Souza, and J. P. Papa, "Binary Flower Pollination Algorithm and Its Application to Feature Selection," in *Recent Advances in Swarm Intelligence and Evolutionary Computation*, ser. Studies in Computational Intelligence, 2015, pp. 85–100.