A Hybrid Gini PSO-SVM Feature Selection: An Empirical Study of Population Sizes on Different Classifier

Noormadinah Allias [1] Megat NorulAzmi Megat Mohamed Noor [1] Mohd. Nazri Ismail [2] Kim de Silva [1]

[1] *Department of MIIT*
Universiti Kuala Lumpur
50250, Kuala Lumpur, Malaysia
noormadinah@yahoo.com,
megatnorulazmi@miit.unikl.edu.my

[2] Faculty of Defence Science and Technology
National Defence University of Malaysia
Kem Perdana Sungai Besi
57000 Kuala Lumpur
m.nazri@upnm.edu.my

***Abstract*— A performance of anti-spam filter not only depends on the number of features and types of classifier that are used, but it also depends on the other parameter settings. Deriving from previous experiments, we extended our work by investigating the effect of population sizes from our proposed method of feature selection on different learning classifier algorithms using Random Forest, Voting, Decision Tree, Support Vector Machine and Stacking. The experiment was conducted on Ling-Spam email dataset. The results showed that the Decision Tree with the smallest size of population is able to give the best result compared to NB, SVM, RF, stacking and voting.**

Keywords-component; swarm size, Taguchi method, orthogonal array, learning algorithms

I. INTRODUCTION

In the era of technology, e-mails have become an important communication tool for business users and home users. Unfortunately, this technology has been misused by irresponsible people sending junk mails. Junk mails or spam are able to cause a lot of headaches to e-mail users, such as loss of business deals, thus bringing to financial loss. In addition to that, junk emails are not also limit the size of mailbox storage by causing a delay of another emails, but they also able to become as a transport to carry viruses and worms [1].

To filter spam emails, a lot of filtering methods have been introduced, including machine learning and non-machine learning. Even though machine learning seems to be the most promising and successful method in eradicating the issue, however, a high dimensional issue in machine learning has become a big hurdle for the classifier itself due to the degradation of the classification results.

Although a lot of feature selection methods have been introduced to reduce a high dimensionality of feature space, its focus which is more on the influence of attribute size selected as features, has neglected the other parameter setting that is able to contribute a high impact on the classifier's performance instead of the size of features. For example in evolutionary feature selection, instead of Genetic Algorithm(GA), Particle Swarm Optimization(PSO) is among the most selected algorithms used in dimension reduction as an optimization method.

PSO is selected as an optimization method compared with GA because it is easy to be implemented as only parameters setting need to be adjusted. Another reason to use PSO is because the information sharing mechanism in PSO is significantly different. In GAs, chromosomes share information with each other while in PSO, only gBest (or lBest) gives out the information to others. As a result, it only looks for the best solution.

Thus, in this paper we present a comparative study of different population sizes from our previous proposed feature selection method based on the Support Vector Machine (SVM), Decision Tree(DT), Random Forest(RF), stacking, and voting with the aim to investigate how the different sizes of population is able to give an impact on different classifiers. The classifiers selected are based on the fact that they are the most employed learning algorithms used for a comparative study [2, 3].

The remainder of this paper is organized as follows. Part II presents related work of previous research in the spam filtering domain. Part III describes the methodology used to do the experiment. Part IV discusses the performance measurement used to evaluate the influence of the population sizes on different learning classifiers. Part V focuses the experimental results and the conclusion and future employment is presented in the final section; Part VI.

II. RELATED WORKS

To the best of our knowledge, there have been not many studies conducted to investigate the effect of population sizes of feature selection in a spam filtering domain. Moreover, there is no work of implementing Taguchi method with feature selection in this field.

A. Previous research work

W.Gang et al. [4] proposes a new fuzzy adaptive multi-population genetic algorithm (FAMGA), to automatically find the best subset of features for email classification. Their

978-1-4799-3251-1/13 $31.00 © 2013 IEEE
DOI 10.1109/AIMS.2013.24

method had utilised multiple subpopulation which ran independently, and each of them was controlled by the fuzzy controller to adjust the crossover rate and the size of each population. Tested on the Ling-Spam email dataset, the result given by the FAMGA showed that they were able to reach until 98.32% and 97.88% for precision and recall when tested on SVM classifier. Whereas, for Naive Bayes, they only managed to achieve 97.34% and 95.54% precision and recall respectively.

W.Hao et al. [5] had proposed a novel fuzzy adaptive particle swarm optimization(NFSS) in spam mail detection. In their proposed method, they had used 45 particles and 450 iterations and tested with SVM classifier on the Ling-Spam email dataset. The result from the NFSS showed 97.83% precision and 97.15% recall. However, none of the methods above had focused and investigated the effect of population sizes on different classifier.

B. Taguchi Method

Introduced by Dr.G.Taguchi in 1985, Taguchi method is a robust design approach that used many ideas from statistical experimental design for evaluating and implementing improvements in products, processes, and equipment [6]. Taguchi method implemented the Orthogonal Array(OA) as a way to reduce the number of experiments and to obtain good experimental results.

Instead of orthogonal array, Taguchi method used Taguchi design analysis which aims to identify the most significant factor that is able to determine which factor contributes more to the public presentation of classification issues. Therefore, in order to produce a good experimental result and determine the contribution factor, a selection of the best combination of parameter settings became a very tedious work and a challenging task.

Thus, by referring to the parameter settings in [7, 8], Taguchi method is applied to assist Gini PSO-SVM in selecting the best combination of the parameter settings while Taguchi analysis design is used later, after obtaining all of the experimental results. Hence, the value of k from Gini Index, population, iteration and inertia factors in PSO-SVM have been identified as the parameter with the varies value.

C. PSO swarm sizes

Particle Swarm Optimization(PSO) is an optimization algorithm introduced by Kennedy and Eberhart [9]. It operated based on social behavior of bird flocking. In PSO, the swarm size or population size is used as a possible solution to identify promising regions of the search space.

Previously in our proposed method, through Taguchi design analysis, we figured out that population sizes are the most significant factor influencing the performance of the classification results. Extended from this work, we ran another experiment by investigating the effect of population sizes from our proposed method of feature selection on different algorithm learning classifiers by using Random Forest, Voting, Decision Tree, Support Vector Machine and Stacking.

The objective of testing different populations with various classifier algorithms is due to the following reasons, previous literature showed that most of the researcher had used a swarm size range between 10 to 50 [10], thus, in this paper, we tried to investigate whether population sizes below than 10 are able to give a better classification result or not since we only use population sizes lower than 10, especially in the area of spam domain. Secondly, we wanted to determine which classifier worked better either with the lowest or highest population size obtained from the previous experiment.

We run an experiment to evaluate the effect of the population sizes mentioned above. The explanation for it can be found below.

III. EXPERIMENTAL PROCEDURE

We conducted the experiments with Ling-Spam [11] email dataset by using the Rapid Miner 5.3.008 software on AMD A6- 3420M APU with Radeon (tm) HD Graphic 1.5 GHz with 8 GB RAM with 10- fold stratified cross validation. The Lingspam corpus consists of 2412 Linguist messages and 481 spam messages from the Linguist list.

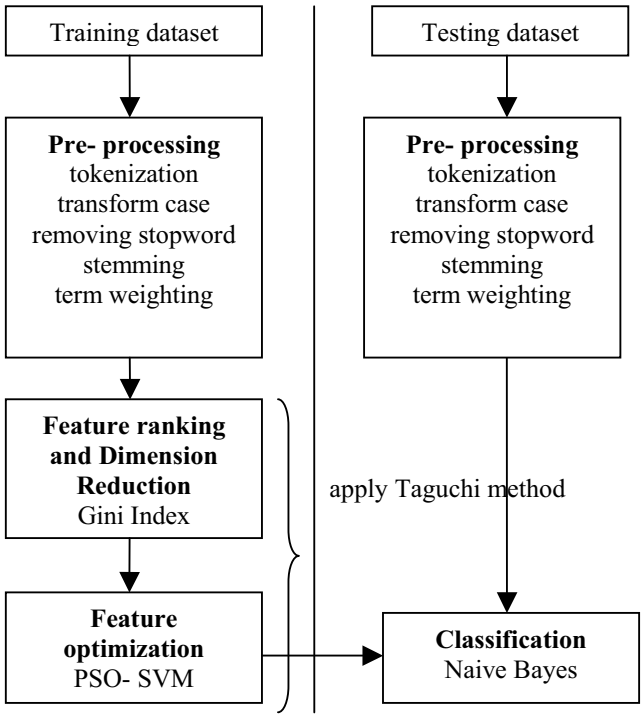

Figure 1: The structure of proposed feature selection method

Figure 1 shows our proposed feature selection method that has been evaluated during our previous experiment.

Later, we extended this experiment by investigating the effect of population sizes of proposed feature selection method on different algorithm learning classifier based on the data analyzed by Taguchi. The data is represented as below:

TABLE I : THE MOST SIGNIFICANT FACTOR INFLUENCED THE PERFORMANCE OF CLASSIFICATION RESULT OBTAIN FROM TAGUCHI METHOD.

Level	K	Population	Iteration	Inertia
1	92	93.45	93.17	92.62
2	92.09	93.74	93.39	92.1
3	91.83	91.1	92.56	91.96
4	93.62	91.91	92.05	92.48
5	93.56	92.9	91.32	93.96
Delta	1.78	2.64	1.46	2
Rank	3	1	4	2

Data from Table I is obtained after the process of analyzing Taguchi's design from a previous study conducted [12]. From Table 1, it can be seen that the populations has the largest effect on the classifier's performance and the iteration has the smallest effect on the classifier's performance.

By referring to Table I, we ran another experiment based only on the population while maintaining the best number of k, iteration and inertia gathered from a previous experiment. The population sizes varied from values of 1 until 9 as shown in Table II below:

TABLE II: POPULATION TABLE MAP BETWEEN NUMBER OF K, ITERATION AND INERTIA

Classifier	K	Population					Iteration	Inertia
RF	100	1	3	5	7	9	30	0.9
DT	100	1	3	5	7	9	30	0.9
SVM	100	1	3	5	7	9	30	0.9
Voting	100	1	3	5	7	9	30	0.9
Stacking	100	1	3	5	7	9	30	0.9

IV. PERFORMANCE MEASUREMENT

The objective of this experiment is to investigate how the population is able to influence the classifier's performance thus giving the best classification results. The results from this experiment are measured based on F1-measure, precision and recall.

Precision measure how many messages, classified as spam, are truly spam. It reflects the amount of legitimate e-mails mistakenly classified as spam. The higher the spam precision is, the fewer legitimate emails have been mistakenly filtered [13]. It is defined as follows:

$$P = \frac{TP}{TP+FP} \quad (1)$$

Recall measures the percentage of spam that can be filtered by an algorithm or model. High spam recall ensures that the filter can protect the users from spam effectively [13].

$$R = \frac{TP}{TP+FN} \quad (2)$$

Where TP, FP, and FN represent the number of true positive, false positive and false negative respectively.

F1- measure is defined as the harmonic mean of precision and recall. A good classifier is assumed to have a high F1- measure, which indicates that the classifier performs well with respect to both precision and recall [14].

$$\text{F1- measure} = \frac{2*PR}{P+R} \quad (3)$$

V. RESULT AND DISCUSSION

TABLE III: POPULATION RESULT ON DIFFERENT CLASSIFIER

	NB	SVM	DT	Stacking
Population	9	3	1	5
Precision(%)	100.00	91.71	96.74	97.17
Recall(%)	91.36	85.45	94.55	92.27

From the experimental results, Random Forest and voting methods produce the worst classification results; whereby the result of precision, recall and F-measure for all populations from both classifiers is unknown.

Due to the bad results, hence we only displayed the best performance of populations sizes only. As can be seen in Table III, our proposed method with NB produced 100.00 percent of precision and 91.36 percent recall when population size is 9.

When the population size is 5, stacking produces 97.17 percent precision and 92.27% of recall result. Precision result of Decision tree (DT) is 96.74 percent when population size is 1, while recall is 94.55 percent. However, SVM produced the lowest precision and recall result; 91.71 percent and 85.45 percent compared with the other learning classifier.

Unfortunately, the results precision and recall is not the only key index in determining the overall effectiveness of the classifier's result. Thus, the result from F-measure in Figure 2 is used as it evaluates the harmonic mean between precision and recall.

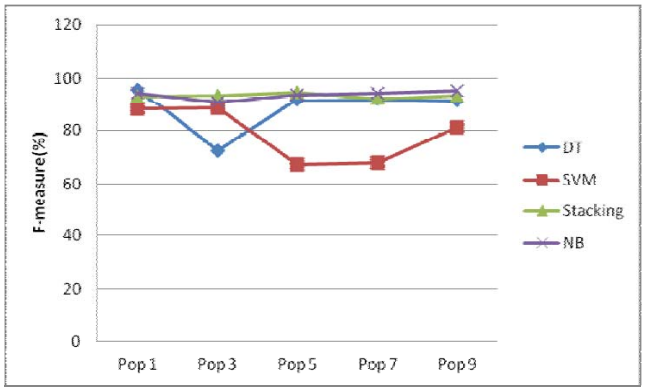

Figure 2: F-measure performance of the DT, SVM, NB, and stacking based on population sizes

Even though Naive Bayes (NB) outperformed the Decision Tree (DT) classifier by obtaining 100.00% precision result, on the whole, the Decision Tree became the best classifier with F-measure 95.63% compared with NB which just managed to achieve only 95.48%.

In addition to that, the population size used in the Decision Tree was the smallest among the population size used by another classifier to get the best results. In this experimental result, DT only used 1 size of population, while NB used 9 population sizes.

Stacking was only able to produce 94.66% of F-measure with 5 population sizes, while SVM is the worst classifier with the value of the F - measure 88.47% and 3 population sizes.

VI. Conclusion and Future Work

In this paper, we compared different population sizes of Hybrid Gini PSO-SVM on different classifiers. An experiment with Ling-Spam email dataset showed that even by using the smallest size of the population, it is still able to produce a good result. Thus, in this experiment, Decision Tree became the best classifier to work out with the smallest size of population compared with Naive Bayes, SVM, stacking, voting and Random Forest.

As a conclusion, the size of the attributes, iteration, and inertia are not the only factors to determine a good result performance, as the appropriate size of a population is the main sources of influence that need to be considered for the effectiveness of the classification performance. For future studies, we would like to replace the PSO-SVM with GA-SVM and investigate the impact of population sizes on different classifiers.

Acknowledgment

The authors are grateful to the reviewers who have given their thoughtful comments to improve the quality of this paper.

References

[1] J. Crain, et al., "Fighting Phishing with Trusted Email," *Proc. Availability, Reliability, and Security, 2010. ARES '10 International Conference on*, pp. 462-467.

[2] T.A. Almeida and A. Yamakami, "Content-based spam filtering," *Proc. Neural Networks (IJCNN), The 2010 International Joint Conference on*, IEEE, pp. 1-7.

[3] I. Koprinska, et al., "Learning to classify e-mail," *Information Sciences*, vol. 177, no. 10, 2007, pp. 2167-2187.

[4] W. Gang, et al., "A New Fuzzy Adaptive Multi-Population Genetic Algorithm Based Spam Filtering Method," *Proc. Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on*, pp. 1-4.

[5] W. Hao, et al., "A Novel Spam Filtering Framework Based on Fuzzy Adaptive Particle Swarm Optimization," *Proc. Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, pp. 38-41.

[6] W.-C. Hsu and Y. Tsan-Ying, "E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters," *Proc. Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, 2009, pp. 764-767.

[7] Y. Shi and R.C. Eberhart, "Empirical study of particle swarm optimization," *Proc. Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, IEEE, 1999.

[8] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *European Journal of Operational Research*, vol. 206, no. 3, pp. 528-539; DOI http://dx.doi.org/10.1016/j.ejor.2010.02.032.

[9] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proc. Neural Networks, 1995. Proceedings., IEEE International Conference on*, IEEE, 1995, pp. 1942-1948.

[10] R. Hassan, et al., "A COPMARISON OF PARTICLE SWARM OPTIMIZATION AND THE GENETIC ALGORITHM," 2004.

[11] J.K. I. Androutsopoulos, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering" In Potamias, G., Moustakis, V. and van Someren, M. (Eds.)," *Proc. Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)* 2000, pp. 9-17.

[12] M.N.M.M.N. Noormadinah Allias, Mohd Nazri Ismail, "A Hybrid Gini PSO-SVM Feature Selection Based on Taguchi Method: An Evaluation on Email Filtering,"unpublished.

[13] Z. Yuanchun and T. Ying, "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 2, 2010, pp. 486-497; DOI 10.1109/tifs.2010.2103060.

[14] J. Meng, et al., "A two-stage feature selection method for text categorization," *Computers & Mathematics with Applications*, vol. 62, no. 7, pp. 2793-2800; DOI http://dx.doi.org/10.1016/j.camwa.2011.07.045.