ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PARAMETER MAPPING OF FEATURE SELECTION VIA TAGUCHI METHOD ON EMAIL FILTERING

[1]NOORMADINAH ALLIAS, [1]MEGAT NORULAZMI MEGAT MOHAMED NOOR,
[2] MOHD NAZRI ISMAIL
[1]Universiti Kuala Lumpur, Department of Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia
[2]National Defence University of Malaysia, Faculty of Defence Science and Technology, 57000 Kuala Lumpur, Malaysia
E-mail: [1]noormadinah@yahoo.com , [1] megatnorulazmi@miit.unikl.edu.my , [2]m.nazri@upnm.edu.my

ABSTRACT

In spam filtering, the use of machine learning as a filtering method is prone to a high dimensionality of features space. In order to overcome the problem, a lot of feature selection methods have been introduced. Besides, the number of features used as an input to machine learning classifier is still high, thus it will delay the delivery of incoming emails to user's inbox. Therefore, two stages of feature selection by using Taguchi method to reduce a high dimensionality of features and obtained a good result are introduced. Firstly, we used Gini Index to reduce a high dimensionality and selecting the best subset of features, while Taguchi method is applied to assist Gini Index and PSO-SVM in selecting the best combination of parameter settings. Apart from this, the impact of population size on different classifier is investigated as it brings a high impact on the classifier performances. This method is trained and tested on Ling-spam email dataset. The experimental result shows that a hybrid Gini PSO- SVM feature selection with Taguchi method is able to produce a good classification result even when the population size is less than 10.

Keywords: S*pam Email, High Dimensionality, Feature Selection, Orthogonal Array, and Population Size.*

1. INTRODUCTION

Unsolicited bulk emails or most preferably called as spam has become as one of the common scenarios and concerned issues in today's community. However, as email has become one of the most important communication tools, the flooding of the spam emails bring a lot of headaches and damages due to its ability to limit the size of mail storage, wasted the network bandwidth and also becomes as a transport to carry viruses.

As a result from the above problems, a lot of works have been carried out in order to cope with the issue; either by using legislation, hand craft rules, or by using machine learning method. However, legislation method of CAN- SPAM Act has been criticized due to the fact that it provides guidance for spammers to send more spam as long as they follow the rules correctly. On the other hand, hand craft rules method such as blacklist and white list are prone to the misclassification of the legitimate emails because humans can make mistakes as well.

Obviously, spammers tend to change their spamming methods by employing obfuscation techniques. Alternatively, in regards of machine learning, the process of filtering spam email has become automated, due to its capability of extracting knowledge of the contents within the emails. However, in machine learning, filtering spam emails is related to text categorization problems. This is because filtering spam by using machine learning usually requires the content of the email messages to be presented as a vector of space model; thus, it can even lead to tens of thousands number of features as well as high dimensionality of features space.

Furthermore, since not all classifier can scale well on a high dimensionality, there is a need to reduce the high dimensionality of features as it can degrade the classifier performance. Consequently, in order to cope with this problem, the dimensionality reduction methods have been proposed as an aim to reduce the number of features and the dimensionality of data. In the last several years, several studies have focused on dimensionality reduction by improving the feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

selections. However, the method has a few weaknesses: (i) a number of features used after dimension reduction has been still high. Thus, it causes delay in email delivery to the users' mailbox. (ii) There is no prior work that applies the filter and wrapper methods, which map feature selections by choosing and matching for the best combination of parameter settings as they require a lot of time to test the parameters one by one.

As a consequence, in this research, a dimensionality reduction method is presented by using two stages of feature selection, which is a hybrid of filter and wrapper types of feature selection. However, as the wrapper feature selection used in this research having more than two parameters and in regards of difficulties to choose the best design of parameter setting between both of the filter and wrapper, Taguchi method is implemented to assist in finding the best mapping of parameter settings between these two feature selection. The remainder of this paper is organized as follows: Section 2 presents previous research works. Methodology is described in Section 3. Section 4 presents the performance evaluation used for comparing the achieved results. Section 5 describes the experimental results. Finally, Section 6 discusses conclusion and future works.

2. PREVIOUS WORK

Numerous types of feature selection have been widely used in spam filtering. By far, information gain has become as the most commonly used of feature selection to reduce a high dimensionality of feature space in spam filtering [1]. In a research by L. Zhang, et al.,[2] the experimental results obtained by researchers on PU1, Ling-spam, Spam Assassin and ZHI1 spam dataset shows that the sensitivity of feature selection method varies greatly from classifier to classifier. Result from the experiment shows that X^2- test and IG performed a better result than DF, especially when feature set is small.

In line with their research work, years ahead, Méndez, et al., [3] had analyzed the strengths and weaknesses of the mainly used feature selection methods; including Document Frequency (DF),

Information Gain (IG), Mutual Information (MI), and X^2- test (CHI) in spam filtering domain. The selected feature selection methods have been analyzed in conjunction with Naïve Bayes (NB), Boosting Trees, Support Vector Machines (SVM) and ECUE models on a Spam Assassin dataset. A finding from this work shows that DF, IG and CHI are better than MI. On the other hand, IG achieved a little bit better precision while CHI test is slightly superior in recall measure.

Similar to a previous study, from a research by Tiago A. Almeida, et al.,[4], researchers compared the most popular feature selection method; Document Frequency (DF), Information Gain (IG), Mutual Information (MI), X^2- test (CHI) and Odd Ratio (OR) on six well known Enron corpora with variations of Naive- Bayes anti- spam filter. A result finding shows that IG and OR are the most aggressive dimension reduction methods with the four variants of Naive Bayes, especially with MV Bernoulli NB which achieves the best performance, instead of MI which produces poor categorization results.

From the research work performed by [2-4] in comparing the effectiveness of filter types feature selection, it can be said that Information Gain outperforms any other feature selection methods throughout different classifier and corpus. Thus, IG can be considered as a good benchmark for comparing the effectiveness of proposed feature selection methods later on.

In contrast with research work done by the above mentioned researchers who use a common type of feature selection, Wang, et al [5] has taken a different approach by investigating the use of different local search optimization feature selection; Hill Climbing (HC),Simulated Annealing(SI), and Threshold Accepting(TA) as a part of feature selection algorithm to reduce the high dimension in spam filtering. Tested with K- Nearest Neighbour (KNN) on PU1 corpus, the experimental result shows that the proposed strategies not only reduce the dimensional space, but also improve the performance of the filter; whereby simulated annealing achieved the best performance of accuracy by 95.5%. However, as filtering spam becoming more challenging task; spammers change their spamming method to avoid spam filtering engine [6] , a high dimensionality of features still become as the main concern.

Hence, more sophisticated dimension reduction methods have been introduced to cope with the changes.

In a research by Gang et.al [7], researchers proposes a new fuzzy adaptive multi-population

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

genetic algorithm (FAMGA) to reduce high dimensionality and also find the best feature subset to classify the spam e-mails, by using multiple subpopulations in genetic algorithm and with the assistance of two fuzzy controllers. Tested on PU1 and Ling-spam dataset, experiment on Ling-spam dataset shows the least features, 452 features, results show that FAMGA combining SVM surpasses other methods in predictive rate and the number of selected features by producing 98.32% precision and 97.88% recall.

In a research by Hao et.al [8], researchers introduces a novel spam filtering framework (NSFF) by using a fuzzy adaptive particle swarm (FAPSO) to find an optimal feature subset. The method tested with PU1, Ling-Spam and Spam Assassin corpora. Experiment run based on Ling-spam dataset shows that with the least number of features, the NMFS method is able to defeat standard particle swarm optimization, standard genetic algorithm, Information gain, Gini index, and CHI with the least number of features, 483; thus produced 97.83% precision and 97.15% recall.

Based on research work done by [5, 7, 8], it can be understood that evolutionary feature selection has started to receive an attention by researchers in spam filtering domain, due to a good classification results. However, a selection of features by using [7, 8] method are still high, thus it is not a good thing for spam filtering tools as it will delay the email delivery to reach user's inbox; as a large number of attribute will take classifier times [9].

3. PREVIOUS WORK

To carry out the study, the experiment with Ling-Spam [10] email dataset by using a Rapid Miner 5.3.008 software on AMD A6- 3420M APU with Radeon (tm) HD Graphic 1.5 GHz with 8 GB RAM is conducted. The experiment is run with 10- fold stratified cross validation. The Ling-spam corpus used for the experiment consists of 2412 Linguist messages and 481 spam messages from the Linguist list.

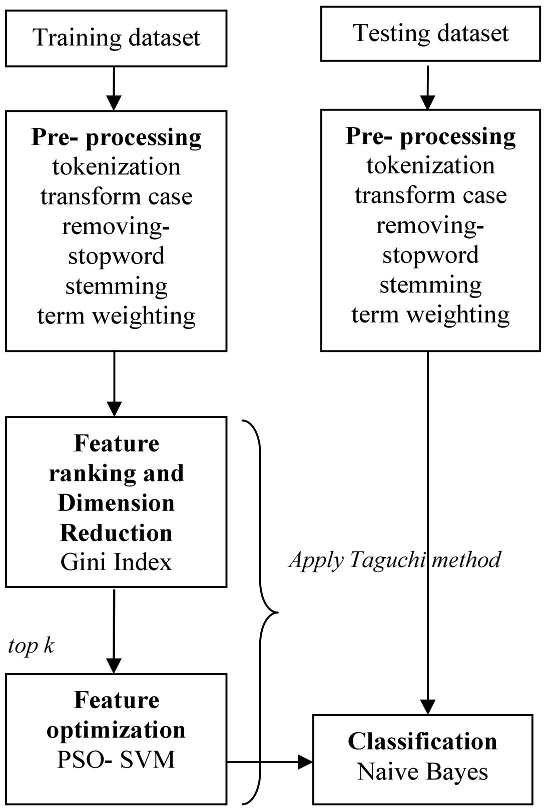

Figure 1: The Structure of Proposed Feature Selection Method

3.1 Pre- processing

Tokenization is used to extract the words in the message body. After a tokenization process, all of the words will be transformed into lowercase. Then, the unnecessary words that occur in many messages are eliminated by using stopword removal. E.g.; "is", "a", "the". Porter algorithm is used as a stemming method to reduce the words into a root form. E.g.; "cooking", "cooked" to "cook". Finally, all of the documents are converted and represented in a form of vector space model using a TF- IDF algorithm. This representation of features are very important as they constitute the basis for most classification algorithms [11]. After the final stage of pre- processing, the features ranking and dimension reduction are performed on the next stage.

3.2 Feature Ranking and Dimension Reduction

Gini index is used to select and rank the attributes accordingly to how important they are. The attributes are ranked in decreasing order. Attribute with the highest weighting score is selected; varying from 25 features until 125

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features. All of the selected features later on will become as input for feature optimization in PSO-SVM.

3.3 Feature Optimization

3.3.1 Particle swarm optimization (PSO)

Particle swarm optimization is an optimization method which work based on the movement of bird flocking or fish schooling [12]. In PSO, all particles in a search space find their best solution by adjusting their own flying experience or the other social flying experience. Instead of current position and current velocities, each particle modifies and keeps track of their position based on personal best (pbest) and global best (gbest).

Later on, the inertia weight is used in order to balance between the local and global search. A large inertia weight facilitates global search while the small inertia weight facilitates the local search. The value of the inertia weight is linearly decreased from large to small throughout the iteration in order to have more global search ability at the beginning and more local search ability near the end of iteration [13]. However, in this research, PSO with SVM is integrated for more optimization.

3.3.2 Support vector machine (SVM)

Support Vector Machine (SVM) method is a new and very popular technique for data categorization that is used in the machine learning community. SVM has been proven to be very effective in the field of text categorization because it can handle the high dimensional data by using kernels [14].

3.4 Taguchi Method

Even though there are a lot of feature selections and classifier available nowadays, a wise choose of parameter settings need to be given more attention, as many factors need to be put into consideration

when dealing with multiple parameters in feature selection; especially feature selection which has more than two types of parameters. Moreover, it will become a headache for researchers in order to find and map the parameters between two or more types of feature selection as they require more time to test and find the best combination of parameters. Therefore, Taguchi method [15] has been identified as a good approach for determining the best combination of inputs to produce a good result.

By referring to parameter setting from [13, 16], a Taguchi method is applied to assist PSO-SVM in selecting the best combination settings for the value of population, iteration, inertia and *k* from Gini Index. Firstly, to implement the Taguchi method, there is a need to identify the number of control factors and levels. After the identification, all of the factors and levels are inserted into Minitab statistical software, to produce an appropriate Orthogonal Array (OA) which consists of the best mapping of Gini index and PSO- SVM. Through this method, only 25 experiments need to be conducted instead of 625 numbers of experiments.

Table 1: An Orthogonal Array (OA) Mapping between Attributes and PSO Parameter Setting.

Level	k	Population	Iteration	Inertia
1	25	1	10	0.9
2	25	3	20	0.8
3	25	5	30	0.7
4	25	7	40	0.6
5	25	9	50	0.5
6	50	1	20	0.7
7	50	3	30	0.6
8	50	5	40	0.5
9	50	7	50	0.9
10	50	9	10	0.8
11	75	1	30	0.5
12	75	3	40	0.9
13	75	5	50	0.8
14	75	7	10	0.7
15	75	9	20	0.6
16	100	1	40	0.8
17	100	3	50	0.7
18	100	5	10	0.6
19	100	7	20	0.5
20	100	9	30	0.9
21	125	1	50	0.6
22	125	3	10	0.5
23	125	5	20	0.9
24	125	7	30	0.8
25	125	9	40	0.7

3.5 Impact of Population Size on Different Classifiers

Upon completion of the full array of the experiments on Ling-spam benchmark dataset, the result is analyzed to determine which factor brings high influence of the result via Taguchi Analysis Design. 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.

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: A Response Table of Signal to Noise Ratio: Larger is Better

Level	k	Population	Iteration	Inertia
1	39.3389	39.4761	39.4553	39.3920
2	39.558	39.5006	39.4706	39.3587
3	39.3133	39.2435	39.3915	39.3306
4	39.4892	39.3347	39.3443	39.3758
5	39.4850	39.4273	39.3204	39.5250
Delta	0.1760	0.2571	0.1502	0.1944
Rank	3	1	4	2

Table 2 depicts a response table of Signal to Noise Ratio of the experimental result. Based on the table, population size has brought the biggest impact on Signal to Noise ratio of a proposed hybrid Gini PSO-SVM feature selection, whereby the population size obtains the highest Delta = 0.2571, Rank 1. The inertia weight having Delta = 0.1944, Rank 2 and number of features (*k*) Delta = 0.1760, Rank 3 has become as the second and third factors influence the classification performance respectively, while the number of iteration brings a less impact on the classification performance with Delta = 0.1502, Rank 4. From the result of the finding above, another experiment is conducted to determine the influence of population size on different classifiers as showed in Table3.

Table 3: Population Table Map between number of k, Iteration and Inertia Based on Different Classifiers

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

In the experiments, the population sizes are varied, while the parameter settings are maintained as shown in Table 3 for *k* value, iteration and inertia as resulted from previous experiment, these settings have contributed for the best result obtained. The objective of testing different populations with various classifier algorithms is due to the following reasons; previous literature shows that most of the researchers use a swarm size range between 10 to 50 [17]. Nevertheless, the use of a small population sizes has not been addressed in spam filtering domain. Besides, as mentioned by [18, 19], a large population size is required to handle a high dimensionality.

Therefore, in this experiment, the investigation is carried out to find whether a small population size which is below than 10 is able or not to give better classification result. Secondly, it is important to determine which classifier works better either with the lowest or highest population size obtained from the previous experiment. As a result, a comparative study of different population sizes is proposed using 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 [20, 21]. As a consequence, the experiment for the training and testing operational set up is just the same as the one indicated in Figure 1, except for the data tabulated in Table 3.

4. PERFORMANCE MEASURE

The results from this experiment are measured based on precision, recall and F1- measure.

4.1 Precision

Precision measures how many messages classified as spam are truly spam and it also reflects the amount of legitimate e-mails mistakenly classified as spam [22].

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

4.2 Recall

Recall measures the percentage of spam that can be filtered by an algorithm or model [22].

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

4.3 F1- measure

F1- measure combines recall with precision. Hence, a good classifier is assumed to have a high F1- measure, which indicates that classifier performs well with respect to both precision and recall [6, 23].

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

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. RESULT ANALYSIS

5.1 Experiment Comparison with Standard Feature Selection Method

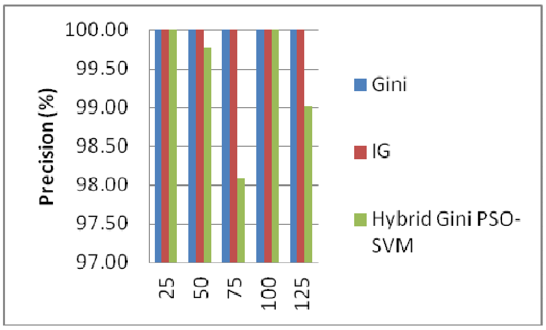

Figure 2: Precision Performance for Hybrid Gini PSO-SVM With Taguchi Vs. Gini Index, And Information Gain

Figure 2 depicts the precision results of Hybrid Gini PSO-SVM with Taguchi, Gini index and Information gain. Based on the figure above, Gini index and Information gain obtains 100% precision with the smallest number of features, 25 features. The results seem stable for both of the feature selection method even though the number of attributes is increased until the maximum number of 125 features.

In contrast, the precision results of a hybrid Gini PSO- SVM feature selection with Taguchi method varies from the minimum attributes, 25 features until it reaches the 125 maximum number of features as it is influenced by the size of population used, iteration and inertia, instead the number of attributes. However, it obtains the best 100% precision performance with a 100 attributes by using population size = 9, inertia= 0.9 and iteration = 30.

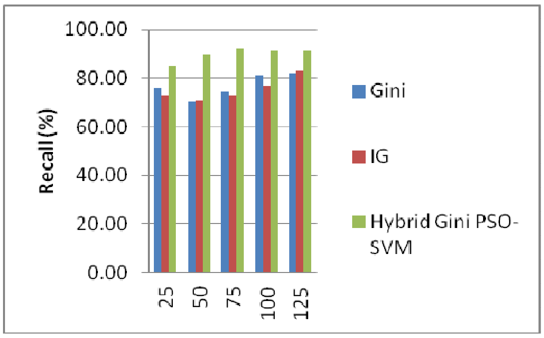

Figure 3: Recall performance for Hybrid Gini PSO-SVM with Taguchi vs. Gini Index, and Information Gain

In Figure 3, the recall results shows that Information gain and a hybrid Gini PSO-SVM feature selection with Taguchi method achieves higher recall result as compared to Gini index. Besides, the proposed hybrid Gini PSO- SVM feature selection with Taguchi method amplifies the recall result of a single stage of Gini index feature selection by obtaining 91.36% recall with 100 attributes by using population size = 9, inertia= 0.9 and iteration = 30.

Based on the Figure 3, one stage of Gini index feature selection achieves its best recall result with 125 attributes, which is 82.3%. In regards with Information gain, this feature selection method achieves 82.73% recall by using the maximum number of 125 attributes. Obviously, this indicates that the total number of spam messages correctly filtered by the proposed method is higher as compared to both of the well- known feature selection methods.

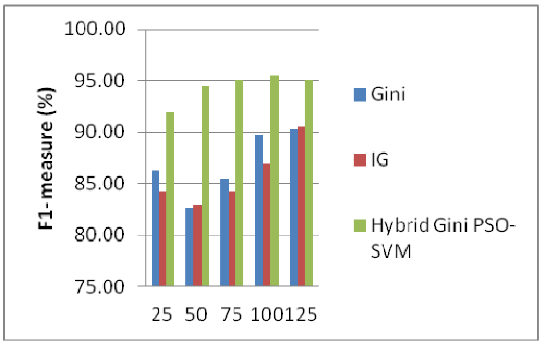

Figure 4: F1-measure performance for a Hybrid Gini PSO-SVM with Taguchi vs. Gini Index, and Information Gain

To define either the filtering system is good or not, the F1- measure is used as an indicator. As depicted in Figure 4, the F1- measure result of a hybrid Gini PSO-SVM feature selection with Taguchi is superior compared to Information gain and Gini index. When the selected number of features reached 100 attributes by population size = 9, number of iteration = 30 and inertia weight = 0.9, the F1-measure value for a hybrid Gini PSO-SVM with Taguchi is the highest (95.48%). However, Gini index and Information gain only manages to achieve 90.30% and 90.55% respectively using 125 attributes.

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.2 Comparison with Other Researcher Works

Based on the comparison of experiments with a well- known feature selection method, it is obviously showed the hybrid Gini PSO-SVM feature selection with Taguchi method achieves the best result with 100 attributes, population size = 9, number of iteration = 30 and inertia weight = 0.9. In addition to that, the comparison is not only conducted between the research work and the well-known feature method, but it also compares the work of a hybrid Gini- PSO-SVM features selection with Taguchi to the other previous research works.

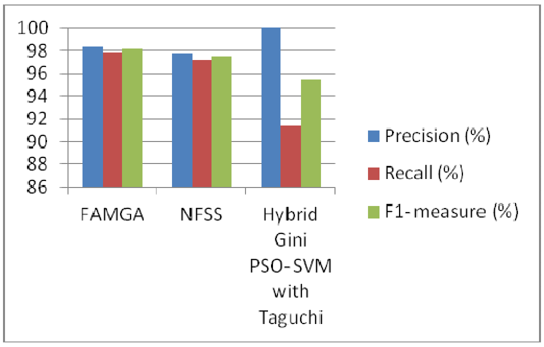

Figure 5: Comparison work with the other researchers

In Figure 5, the proposed method is compared with fuzzy adaptive multi-population genetic algorithm (FAMGA) and a novel spam filtering framework (NSFF) on the same Ling-spam benchmark dataset. Based on the figure, FAMGA achieves 98.32% precision and 97.88% recall and 98.10% F1-measure with 454 attributes. Whereas the NFSS method achieves 97.83% precision, 97.15% recall and 97.49% F1- measure with 484 attributes.

Although the FAMGA and NFSS method obtain a higher recall result as compared to the hybrid Gini PSO-SVM feature selection with Taguchi method, , in terms of precision criteria, it records 100% precision. Furthermore, from the result, it shows that all of the legitimate emails are managed to be correctly classified as ham, whereby the FAMGA and NFSS method have misclassified some of the legitimate emails as spam. In the area of spam filtering, misclassification of the legitimate emails as spam is unacceptable as it brings more damaged to the users. Besides, the number of features used is the lowest, which is 100 attributes to obtain 100% precision, 91.36% recall and 95.48% F1- measure as compared to FAMGA and NFSS which uses 454 and 483 numbers of features respectively.

5.3 Evaluation of the Impact of Population Sizes on Different Classifier

From the experimental results, Random Forest and voting methods produce the worst classification results; whereby the results of precision, recall and F1-measure for all populations from both classifiers are unknown. Hence, due to the bad results, only the best performance of population sizes is displayed, while, the discussion only is made on the best result obtained by each classifier.

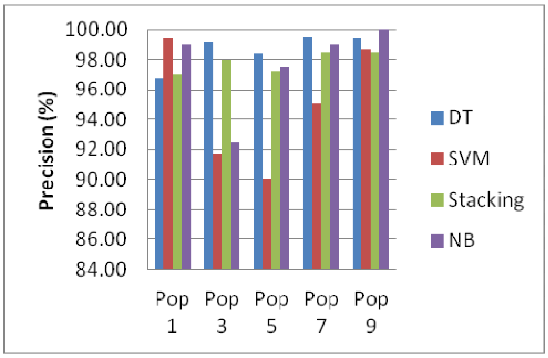

Figure 6: Precision result produces by each classifier vs. number of populations

As showed in Figure 6, the proposed method with NB produced 100.00% of precision when population size is equal to 9. Decision Tree (DT) obtains the second highest with 99.47% precision using 7 population size, followed by SVM, which is 99.43% with 1 population size. Whereas, stacking only manages to obtain 99.46% of precision with 9 population size.

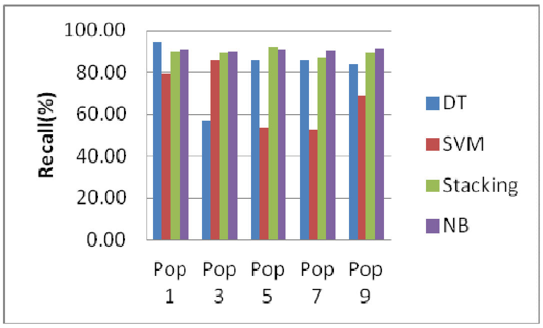

Figure 7: Recall result produce by each classifier vs. number of populations

In Figure 7, Decision tree (DT) with 1 population size produces the highest recall result 94.55%, followed by Stacking respectively; 92.27% with 5 population size. However, Naive Bayes (NB) is only able to produce 91.36% of

Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

recall result with the maximum number of population size, while SVM used 1 population size to obtain the best result of recall; 79.09%.

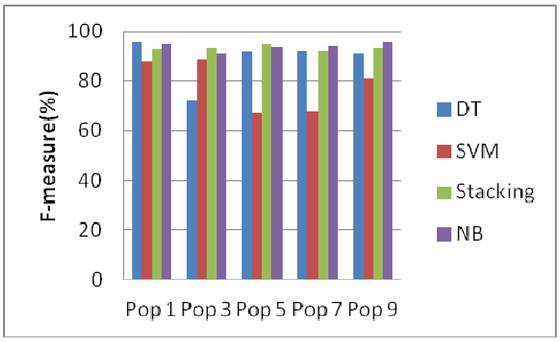

Figure 8: F1-measure result produce by each classifier vs. number of populations

Even though Naive Bayes (NB) outperformed the Decision Tree (DT) classifier by obtaining 100.00% precision result, on the whole, it is the Decision Tree is to be regarded as the best classifier with F1-measure 95.63% as compared to NB which just manages to achieve only 95.48%. In addition to that, the population size used in the Decision Tree is the smallest among the population size used by another classifier to get the best results. In this experimental result, DT only uses 1 size of population, while NB uses 9 population sizes. Stacking is only able to produce 94.66% of F1-measure with 5 population sizes, while SVM is the worst classifier with the value of the F1 - measure 88.47% and 3 population sizes.

6. CONCLUSION AND FUTURE WORKS

In this study, two stages of feature selection methods which combine a filter and wrapper types of feature selection via a hybrid Gini PSO- SVM feature selection based on Taguchi method in spam filtering is proposed. Filter types of Gini index feature selection is used to rank the attributes in decreasing order and select the top *k* number of attributes as an input to PSO-SVM wrapper feature selection. The Taguchi method is employed to the value of *k*, population size, inertia, and iteration by using an orthogonal array which is able to help to find the best combination of parameter setting.

Even though this research led to a good classification results, but it encountered a few limitation during the experiments as it is difficult to use the same algorithms used by other researchers in Rapid Miner, as the software did not support the algorithm in order to make a comparison to see which feature selection method is better. The best thing to do is by using the same Lingspam email dataset. However, even though Lingspam email dataset used in this research are relatively old, however, it is still been used until today as a comparison of recent research.

A comparison is made to identify how the proposed method perform, in terms of precision, recall and F1- measure by using well- known feature selection algorithm and method proposed by [7, 8] as a benchmark. The result shows that the proposed method is higher than the FAMGA and NFSS method in terms of precision with the lowest number of attributes, which are 100 features as compared to 484 features and 454 features used by NFSS and FAMGA methods. Comparing with Information gain and Gini index, the Hybrid Gini PSO-SVM with Taguchi obtains better recall result. Thus, it indicates the proposed method is competent to filter more spam effectively as compared to a single stage of Information gain and Gini index.

As for the experiment based on population sizes with different classifiers, it is found that by using less than 10 population sizes, the proposed method is still able to produce a good classification results. Besides, it is also figured that Naive Bayes (NB) works well with the highest population sizes, which is 9 in this experiment, while Decision Tree (DT) works well with the lowest size of population (population size = 1).

Therefore, it confirms the finding that is figured out previously, which indicates that population sizes are the most important factors that influence the performance depending on the types of classifier used instead of iteration, inertia, the number of attributes.

From results discussed, this research added into two types of contribution; methodological contribution and experimental contribution. In methodological contribution, the proposed method enable the researcher in saving their time in finding and selecting the best parameter, especially for hybrid types of feature selection which have a lot of parameter setting value. In term of experimental contribution, this research showed that it able to correctly identify all of the legitimate emails as ham and at the same time amplified the recall result as compared to a single Gini index of feature selection.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

For future work, there is a plan to implement a combination of PSO-GA feature selection and to apply Taguchi method in enlightening the best parameters mapping between the two types of feature selection on the spam filtering domain.

REFRENCES:

[1] Guzella, T.S. and W.M. Caminhas, "A Review of Machine Learning Approaches to Spam Filtering." *Expert Systems with Applications*, Volume 36, , 2009, pp. 10206-10222.

[2] Zhang, L., J. Zhu, and T. Yao, "An Evaluation of Statistical Spam Filtering Techniques", *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol 3, 2004, pp. 243-269.

[3] Méndez, J., Fdez-Riverola, F., Díaz, F., Iglesias, E., Corchado, J. & Perner, P., *A Comparative Performance Study of Feature Selection Methods for The Anti-Spam Filtering Domain.* Advance in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining., Springer Berlin / Heidelberg, , 2006, pp. 106-120.

[4] Almeida, T.A., A. Yamakami, and J. Almeida, "Probabilistic Anti-Spam Filtering With Dimensionality Reduction", in *Proceedings of the 2010 ACM Symposium on Applied Computing,* Sierre, Switzerland, , 2010, pp. 1-5.

[5] Wang, R., A.M. Youssef, and A.K. Elhakeem. "On Some Feature Selection Strategies for Spam Filter Design", *in Canadian Conference on Electrical and Computer Engineering.CCECE '06*, Ottawa, Ont, , 2006, pp.2186- 2189.

[6] Panigrahi, P.K. "A Comparative Study of Supervised Machine Learning Techniques for Spam E-mail Filtering", *in 2012 Fourth International Conference on.Computational Intelligence and Communication Networks (CICN*, Mathura , 2012, pp. 506- 512.

[7] Gang, W., Yuan-Ning, L., Xiao-Dong, Z., Hui-Ling, C. & Liu, Z. "A New Fuzzy Adaptive Multi-Population Genetic Algorithm Based Spam Filtering Method", in *2nd International Conference on Information Engineering and Computer Science (ICIECS)*, 2010, Wuhan, pp.1-4.

[8] Hao, W., Hong-Zuo, L., Gang, W., Hui-Ling, C. & Xiao-Kui, L. "A Novel Spam Filtering Framework Based on Fuzzy Adaptive Particle Swarm Optimization", *in International Conference on Intelligent Computation Technology and Automation (ICICTA), 2011*, Shenzen, Guangdong, pp.38-41.

[9] Delany, S., P.d. Cunningham, and L. Coyle, "An Assessment of Case-Based Reasoning for Spam Filtering", *Artificial Intelligence Review*, Vol.24, , 2005, pp. 359-378.

[10] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G. & Spyropoulos, C. D, "An Evaluation of Naive Bayesian Anti-Spam Filtering." in *Proceedings of the Workshop on Machine Learning in the New Information Age 11th European Conference on Machine Learning (ECML 2000)*, Spain , 2000, pp. 9-17.

[11] Silva, C. and B. Ribeiro, *Background on Text Classification, in Inductive Inference for Large Scale Text Classification,* Springer Berlin Heidelberg, 2010 p. 3-29.

[12] Kennedy, J. and R. Eberhart. "Particle Swarm Optimization". *in Proceedings IEEE International Conference on Neural Networks, 1995*, Perth, WA, 1995, pp. 1942- 1948.

[13] Shi, Y. and R.C. Eberhart, "Empirical Study of Particle Swarm Optimization" *in Proceedings of the 1999 Congress on Evolutionary Computation, 1999. CEC 99.* IEEE, Washington, DC.

[14] Meng, J., H. Lin, and Y. Yu, "A Two-Stage Feature Selection Method For Text Categorization," *Computers & Mathematics with Applications*, Vol. 62, 2011, pp. 2793-2800.

[15] Taguchi, G., "Quality Engineering (Taguchi methods) For The Development of Electronic Circuit Technology", *IEEE Transactions on Reliability*, Vol. 44, 1995, pp. 225-229.

[16] Unler, A. and A. Murat, "A Discrete Particle Swarm Optimization Method for Feature Selection in Binary Classification Problems", *European Journal of Operational Research*, Vol. 206, , 2010, pp. 528-539.

[17] Hassan, R., Cohanim, B., De Weck, O. & Venter, G, "A Comparison of Particle Swarm Optimization and The Genetic Algorithm" in *Proceedings of the 1st AIAA multidisciplinary design optimization specialist conference*, Austin, Texas, 2005, pp.18-21.

[18] Zhang, B., F. Chen, and L. Bai, "An Improved Partical Swarm Optimization Algorithm", *Journal of Computers*, Vol. 6, 2011, pp. 2460.

[19] Huang, T. and A.S. Mohan, "A Microparticle Swarm Optimizer For The Reconstruction of Microwave Images". IEEE Transactions on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Antennas and Propagation, Vol. 55, , 2007, pp. 568-576.

[20] Almeida, T.A. and A. Yamakami, "Content-Based Spam Filtering", in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, 2010, pp. 1-7.

[21] Crain, J., L. Opyrchal, and A. Prakash. "Fighting Phishing with Trusted Email" in *International Conference on Availability, Reliability, and Security, 2010. ARES '10*, Krakow, , 2010 pp. 462- 467.

[22] Yuanchun, Z. and T. Ying, "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering". *IEEE Transactions on Information Forensics and Security*, Vol. 6, 2010, pp. 486-497.

[23] Yang, J., Liu, Y., Liu, Z., Zhu, X. & Zhang, X, "A New Feature Selection Algorithm Based on Binomial Hypothesis Testing For Spam Filtering", *Knowledge-Based Systems*, Vol. 24, 2011, pp. 904-914.