VOL. 11, NO. 22, NOVEMBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ENSEMBLE FEATURE SUBSET SELECTION TECHNIQUE IN SPAM DETECTION SYSTEM

Aida Mustapha[1,2] and Amir Rajabi Behjat[2]
[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, UPM Serdang, Selangor Darul Ehsan, Malaysia
E-mail: aidam@uthm.edu.my

ABSTRACT

In email spam detection, not only different parts and content of emails are important, but also the structural and special features of these emails have effective rule in dimensionality reduction and classification accuracy. Because spammers constantly change patterns of spamming messages using different advertising images and words to form new pattern features or attributes, feature subset selection and ensemble classification are necessary to address these issues. Recently, various techniques based on different algorithms have been developed. However, the classification accuracy and computational cost are often not satisfied. This study proposes a new ensemble feature selection techniques for spam detection, based on three feature selection algorithms: Novel Binary Bat Algorithm (NBBA), Binary Quantum Particle Swarm Optimization (BQPSO) Algorithm, and Binary Quantum Gravitational Search Algorithm (BQGSA) along with the Multi-layer Perceptron (MLP) classifier. The achieved results showed accuracy very near to 100% in email spam detection.

Keywords: binary bat algorithm, binary quantum particle swarm optimization, binary quantum gravitational search algorithm, multi-layer perceptron.

INTRODUCTION

High number of emails consumes bandwidth resources, as they are able to quickly block or extend storage space for large sites. From user point of view, spam emails waste valuable time for important communication (Lee *et al.*, 2010; Jindal and Liu, 2007). Automatic detection of spam emails is related to a classification problem, whereby the objective is to classify an email into spam or non-spam. In selecting the best features within the emails, various metaheuristic algorithms have been proposed in the past as they have shown to have high ability to optimize the feature sets in classification.

Mohammadand Zitar (2011) construed spam detection as a big challenge as detection systems attempts to separate spam and ham emails with the smallest fraction of misclassification (false positive). In addition, since spammers are constantly adapting and changing their features to bypass spam detection systems, there is a critical necessity to apply a spam filter based on robust classifiers rather than simplistic spam filters such as blacklisting and whitelisting to overcome the high growth of false positive (Puniskis *et al.*, 2006). Even though spam detection systems at present have achieved a reasonable level of accuracy, features change constantly because spammers repeatedly confuse the anti-spam filters and decrease the effective performance of the classifier. For example, text embedded in images, HTML layout and pattern of body of email are important defrauds applied by spammers (Carreras and Marquez, 2001; Wu *et al.*, 2005). One of the problems that decreases the performance of classifiers is high data dimensionality.

Recently, in order to increase the classification accuracy and overcome shortcomings of the individual classifiers and feature selection techniques, ensemble learning techniques have been proposed. Ensemble learning is made by several single models with a combined output (Wang, 2010; Fern and Givan, 2003). In real detection systems, incoming data is divided into different chunks instead of processing training data separately (Minku and Yao, 2012). Wang (2010) proposed a framework based on heterogeneous ensemble technique that combines various spam filters to improve classification accuracy and reliability. The work built three types of models; Naive Bayes (NB), Bayesian Network (BN), and Decision Tree (DT). The achieved result showed an accuracy of 94.44%, which was higher in comparison to previous studies based on Random Committee, Bagging, and Bayes Net.

In another work, a library-based ensemble classification is based on Neural Network (NN) and DT. Using this approach, a library assembles 2,000 different models of classifier before forming an ensemble technique. The achieved accuracy was 91.88% on the test data. Consequently, the ensemble detection system based on different ensemble models has increased the performance by 2.97% when compared to the best individual classifier such as the Support Vector Machine (SVM) (Carpinter and Hunt, 2006). Ying *et al.* (2010) proposed an ensemble technique based on DT, SVM, and back-propagation network using Gain Ratio feature selection method and achieved 91.78% accuracy with 14 relevant features. Ying *et al.* (2010) and Chharia *et al.* (2013) combined nine classifiers in a spam detection system to improve issues of other ensemble classifiers. When the result of weak classifier equals to the incorrect top classifier result, the ensemble system performance decreased. The system applied probability and rules instead of weights for each output and achieved 98.66% accuracy. In similar study, Yang *et al.* (2006) used a rule-based spam detection system using three naive Bayes classifiers as an ensemble classification system.

In ensemble feature selection, the process selects a subset of relevant features in the original feature space in

order to increase performance of classifier. The advantages of feature selection is building fast and simple model according to small subset of selected features. Ensemble feature selection follows the principles of ensemble learning, where several feature selection techniques as classifiers are combined to produce a stable and robust ensemble feature selection. The robustness of feature selectors is significant when the dataset changes over time.

Saeys *et al.* (2008) applied an ensemble feature selection technique as a supervised learning method by combining the filter-based and wrapper-based methods. Therefore, each feature selector will produce its own results and the results are aggregated at the last step. This aggregating is done by weighted voting based on feature ranking. The result of this study showed better accuracy and performance in comparison to single method. However, this work was set to be improved by Attik (2006) based on the number of features. Attik (2006) proposed an ensemble feature selection technique that selects only relevant features and a subset of relevant features. The 17 relevant features selected based on OFSM-SRF feature selection algorithm using Multi-layer Perceptron (MLP). The achieved result showed a performance in classification near to 100%.

In similar studies based on ensemble feature selection using filter-based methods, multiple feature ranking such as document frequently, information gain and chi-square methods are combined for text classification. The combination of methods produced more than 80% precision result (Wang *et al.*, 2010). Meanwhile, Tsymbal *et al.* (2003) used the random subspaces based on a set of Bayesian classifiers with hill claiming feature selection algorithm. The results of this technique are evaluated on real world and synthetic data sets. The produced results of ensemble Bayesian classifier have uses 90.1% and 90.3% for dynamic voting with selection based on balance dataset.

In line with the previous works, this paper proposes a new ensemble feature selection techniques, focusing on new metaheuristic feature selection algorithms namely the Novel Binary Bat Algorithm (NBBA), Binary Quantum Particle Swarm Optimization (BQPSO) Algorithm, and Binary Quantum Gravitational Search Algorithm (BQGSA). The outputs are then fed into a Multi-layer Perceptron (MLP) classifier. The remaining of this paper keeps on as follows. The following section begins with the principles of Multi-Layer Perceptron (MLP). The next section presents the proposed ensemble learning techniques based on three metaheuristic algorithms followed by the details of the experimental results and analysis on the ROC curve. Finally the last section concludes the work and sets future research.

PRINCIPLES OF MULTI-LAYER PERCENTRON

Multi-Layer Perceptron (MLP) has been widely used due its ability to learn complex data structures and work fast with large amount of data. A set of small processing units build neurons of multi-layer back propagation that are arranged in different layers, namely input, hidden and output layers. In fact, these layers are organized to minimize appropriate error functions by a set of parameters such as mode of learning, information content, activation function, target values, input normalization, initialization, and learning rate. The error propagation is content of forward pass and backward pass, so forward pass fixes network weights and backward pass adjusts weights according to error-correction tools. Lastly, the actual results are compared by adjusting the weights during the learning process to accomplish the classification (Vafaie and De Jong, 1992; Perez *et al.*, 2011).

Training or learning step of MLP focuses on numeric attributes that have a limited domain $a_i \epsilon v_{i_1}, v_{i_2}, \cdots, v_{i_k}$ where $v_i$is the number of likely values for attribute a_i. In addition, the training process covered set *NV* training patterns (x_p, t_p) where *P* is related to the pattern number and x_p. answers to the *N*-dimensional input vector of the p^{th} training pattern. Moreover, *YP* answers to the *M*-dimensional output vector from the trained network for thepattern. For decreasing the analysis and handling the amount of hidden units and output units, transferring the value of one to a vector component denoted by $x_p N + 1$is a necessary need. Input and output neurons set linear activations by encoded input values. The input to the j^{th} hidden unit $net_p(j)$is shown in Equation 1(Tretyakov, 2004; Carpinteiro *et al.*, 2006).

$$net_p(j) = \sum_{k=1}^{N+1} w_{hi}\,(j,k), x_k \tag{1}$$

where $1 \le j \le N_h$. Output activation for the p^{th} training pattern, $O_p(j)$is stated in Equation 2:

$$O_p(j) = f(net_p(j)) \tag{2}$$

where the sigmoid function is produced by the nonlinear activation function as shown in Equation 3.

$$net_p(j) = \frac{1}{1 + e^{net_p(i)}} \tag{3}$$

Based on Equation 1 and Equation 2, the *N*input units are signified by *K* and $W_{hi}(J,K)$marks the connected weights of the K^{th} input unit to the J^{th} hidden unit. Additionally, cross validation controls training performance. Thus, every time the total error increases during testing process cross validation will be stopped. Learning rate is reduced by 50% when the number of error increases. Motion is stopped to the end of training if total error does not decrease. In general, the MLP performance is measured by the mean square error (MSE) detailed by Equation 4.

$$E = \frac{1}{N}\sum_{p=1}^{N_v} E_p = \frac{1}{N}\sum_{p=1}^{N_v}\sum_{i=1}^{M}[t_p(i) - y_p(i)]^2 \tag{4}$$

where in Equation 5:

$$E_p = \sum_{i=0}^{M} [t_p(i) - y_p(i)]^2 \quad (5)$$

E_p corresponds to the error for the p^{th} pattern and t_p is the desired output for the p^{th}. This also allows the i^{th} calculation of the napping error for the ith output unit to be expressed by Equation 6.

$$Y_p(i) = \sum_{k=1}^{N+1} W_{oi}(i,k) X_p(K) + \sum_{j=1}^{N_h} W_{oi}(i,j) O_p(j) \quad (6)$$

This equation signifies the weight from the input nodes to the output nodes and denotes the weight from the hidden nodes to the output nodes (Ruan and Tan, 2010).

ENSEMBLE LEARNING TECHNIQUES

In the proposed ensemble feature selection technique, three metaheuristic algorithms are used for feature selection; Novel Binary Bat Algorithm (NBBA), Binary Quantum Particle Swarm Optimization (BQPSO) algorithm, and Binary Quantum Gravitational Search Algorithm (BQGSA). The three algorithms NBBA, BGSA, and BQPSO are combined them to create better results in comparison to individual feature selection algorithm. Figure-1 illustrates the processes in the proposed ensemble learning technique.

In NBBA, the algorithm uses the tanh function instead of the sigmoid function that will discard the difference between the big value of velocity toward positive and negative values. BQPSO and BQGSA adopts the concepts of quantum computing. The BQPSO algorithm motivates particles to have a quantum behavior instead of classical behavior. The BQPSO is chosen to find the best location of each particle in the search space. As compared to the Binary Particle Swarm Optiomization (BPSO) algorithm, BQPSO is fast to search the good solution and also reduces dimensionality. Similar to the BQPSO algorithm, BQGSA also follow the principles of quantum computing with the base of Gravitational Search Algorithm (GSA).

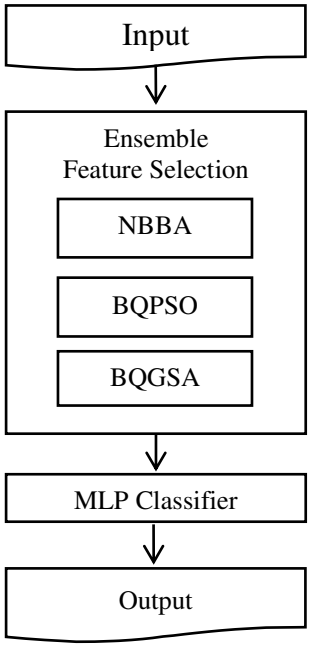

Figure-1. Proposed ensemble learning technique.

EXPERIMENTS AND RESULTS

In this research, the experiments were performed using the Intel Pentium IV processor with 2.7GHz CPU, 4GB RAM, and Windows 7 Operating System. The development environment was MATHWORK_R2010b. The classification experiment used the Multi-Layer Perceptron (MLP), trained with the measurement vectors of 5424 spam and non-spam emails. A total of 1524 instances were available for testing. The classification performance used the ensemble feature subset selection with an MLP classifier assessed by the LingSpam (available at http://nlp.cs.aueb.gr/software.html) and SpamAssassin (available at: http://spamassassin.apache.org/publiccorpus) datasets as shown in Table 1.The evaluation metrics used are accuracy, recall and precision. The results are presented in Table 2 and Table 3.

Table-1. Dataset and the number of class.

No.	Dataset	Size	No. Features
1	LingSpam	6,954	180
2	SpamAssassin	6,954	120

Table-2. Accuracy, precision, recall for LingSpam.

Evaluation Metric	NBBA	BQPSO	BQGSA
Accuracy	99.99	99.00	99.13
Recall	100.00	81.90	95.84
Precision	100.00	99.00	98.93

Table-3. Accuracy, precision, recall for SpamAssassin.

Evaluation Metric	NBBA	BQPSO	BQGSA
Accuracy	99.77	98.99	99.39
Recall	99.43	98.05	96.10
Precision	100.00	99.10	98.79

In addition to the accuracy, recall and precision, this experiment evaluated the AUC value based on different number of features acd on NBBA, BQPSO, and BQGSA with an MLP classifier. Figure-2 displays the highest average rate of ensemble feature selection accuracy based on the ROC curve using different number of features.

Since each feature selection algorithm in ensemble technique produces various relevant features, the generated accuracy rates of algorithms are aggregated. The combination of different set of relevant features reduced the number of irrelevant features. Based on Figure-3, the highest AUC value obtained as 99.83% and 14.66%, 99.96% and 12.45%, 98.49% and 16.12%, 98.99% and 13.45% by ensemble technique for SpamAssassin respectively as compared to BGSA as an individual feature selection algorithm.

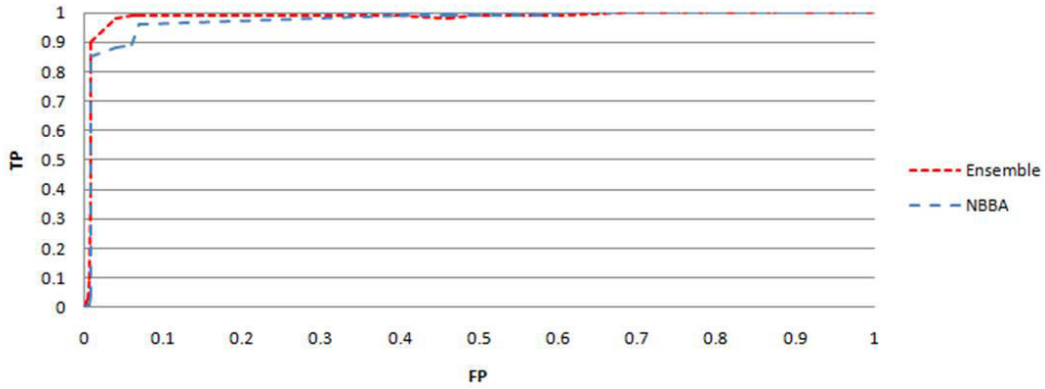

Figure-2. ROC curve of the proposed ensemble feature selection technique.

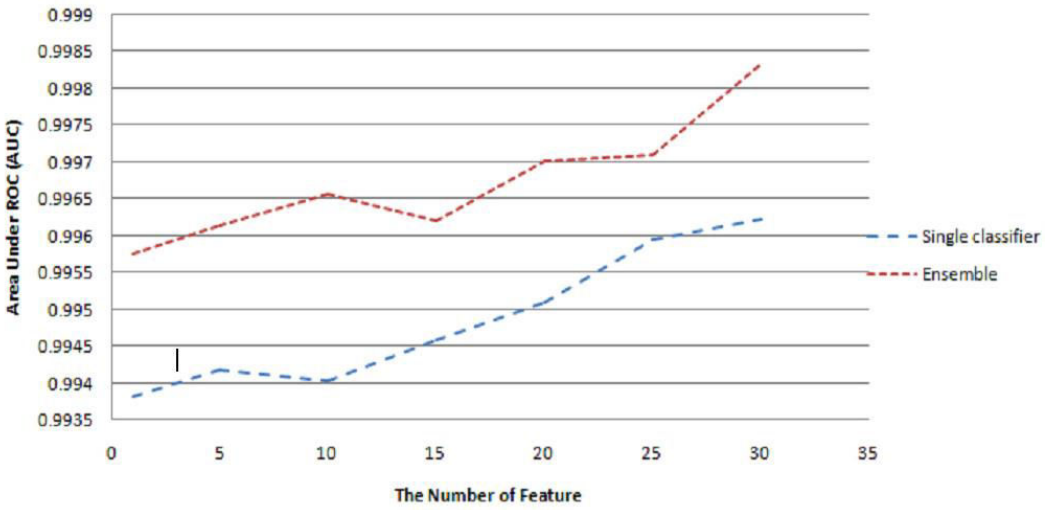

Figure-3. AUC under ROC curve of the proposed ensemble feature selection technique.

Figure-2 and Figure-3 displayed the highest accuracy and the lowest FPR produced by ensemble technique based on the AUC values as reflected by the ROC curve. This graph showed that the highest accuracy produced by ensemble feature selection technique is increased when BQPSO, BQGSA and NBBA algorithms as the best feature selectors in speed and performance are combined on two benchmark datasets. The highest result in terms of high accuracy and low FPR are produced as 99.53% and 0.2%, 99.67% and 0.2%, 99.03% and 0.02%, 98.13% and 0.1% for LingSpam and SpamAssassin.

Consequently, the achieved results of ensemble technique show the FPR together with accuracy rate to produce an integrated experimental result. In fact, the highest average of AUC value and accuracy rate of ensemble feature selection technique is greater than individual methods and selected 32 and 11 relevant selected features for LingSpam and SpamAssassin respectively.

CONCLUSIONS

Ensemble techniques in the feature selection and classification process increase the performance of spam detection system by combining and aggregating the achieved results, hence overcoming the inherent weakness within individual technique. The literature reviews showed that the performance of NBBA, BQPSO, and BQGSA were higher than other heuristic algorithms. Thus, these algorithms built ensemble feature selection and classifier ensemble techniques in the second phase. In this phase, an ensemble feature selection technique based on BQPSO, BQGSA and NBBA algorithms selected a set of relevant features in the spam detection system which the increase of performance indicated that this system overcome individual feature selection techniques such as low accuracy, trapping to local optimum and selection of irrelevant features.

ACKNOWLEDGEMENT

This project is sponsored by the Short Term Grant Scheme at Universiti Tun Hussein Onn Malaysia.

REFERENCES

Lee, S.M., Kim, D.S., Kim, J. H., and Park, J.S. 2010. Spam Detection using Feature Selection and Parameters Optimization. In Proceedings of the 10th IEEE International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp. 883-888.

Jindal, N. and Liu, B. 2007. Analyzing and Detecting Review Spam. InProceeding of the 7th IEEE International Conference on Data Mining, pp. 547-552.

Mohammad, A.H. and Zitar, R.A. 2011. Application of Genetic Optimized Artificial Immune System and Neural Networks in Spam Detection. Applied Soft Computing, 11(4), pp.3827-3845.

Puniskis, D., Laurutis, R., and Dirmeikis, R. 2006. Artificial Neural Nets for Spam Email Recognition. Elektronika ir Elektrotechnika, 5(69), pp.73-76.

Carreras, X. and Marquez, L. 2001. Boosting Trees for Anti-Spam Email Filtering. arXiv preprint cs/0109015.

Wu, C.-T., Cheng, K.-T., Zhu, Q., and Wu, Y.-L. 2005. Using Visual Features for Anti-Spam Filtering. In Proceedings of the IEEE International Conference on Image Processing, pp.III-509.

Wang, W. 2010. Heterogeneous Bayesian Ensembles for Classifying Spam Emails. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), pp.1-8.

Fern, A. and Givan, R. 2003. Online Ensemble Learning: An Empirical Study. Machine Learning, 53(1-2), pp.71-109.

Minku, L.L. and Yao, X. 2012. DDD: A New Ensemble Approach for Dealing with Concept Drift. IEEE Transactions on Knowledge and Data Engineering, 24(4), pp. 619-633.

Carpinter, J. and Hunt, R. 2006. Tightening the net: A Review of Current and Next Generation Spam Filtering Tools. Computers and Security, 25(8), pp. 566-578.

Ying, K.-C., Lin, S.-W., Lee, Z.-J., and Lin, Y.-T. 2010. An Ensemble Approach Applied to Classify Spam E-mails. Expert Systems with Applications, 37(3), pp. 2197-2201.

Chharia, A. and Gupta, R. 2013. Email Classifier: An Ensemble using Probability and Rules. In Proceedings of the 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 130-136.

Yang, Z., Nie, X., Xu, W., and Guo, J. 2006. An Approach to Spam Detection by Naive Bayes Ensemble based on Decision Induction. In Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, pp. 861-866.

Saeys, Y., Abeel, T. and Van de Peer, Y. 2008. In Machine Learning and Knowledge Discovery in Databases. Machine Learning and Knowledge Discovery in Databases, Springer, pp. 313-325.

Attik, M. 2006. Using Ensemble Feature Selection Approach in Selecting Subset with Relevant Features. In Proceedings of the Advances in Neural Networks, Lecture Notes in Computer Science, 3971, pp.1359-1366.

Wang, H., Khoshgoftaar, T.M., and Napolitano, A. 2010. A Comparative Study of Ensemble Feature Selection Techniques for Software Defect Prediction. In Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 135-140.

Tsymbal, A., Puuronen, S., and Patterson, D.W. 2003. Ensemble Feature Selection with Simple Bayesian Classification. Information Fusion, 4(2), pp. 87-100.

Vafaie, H. and De Jong, K. 1992. Genetic Algorithms as a Tool for Feature Selection in Machine Learning. In Proceedings of the Proceedings of the Fourth International Conf. on Tools with Artificial Intelligence, pp. 200-203.

Perez, F.M., Gimeno, F.J.M., Jorquera, D.M.M., Abarca, J.A.G.M., Morillo, H.R., and Fonseca, I.L. 2011. Network Intrusion Detection System Embedded on a Smart Sensor. In Proceedings of the IEEE Industrial Informatics, 58.

Tretyakov, K. 2004. Machine Learning Techniques in Spam Filtering. In Proceedings of the Data Mining Problem-Oriented Seminar, pp.60-79.

Carpinteiro, O.A., Lima, I., Assis, J.M., de Souza, A.C.Z., Moreira, E.M. and Pinheiro, C.A. 2006. A Neural Model in Anti-spam Systems. In Proceedings of the Artificial Neural Networks ICANN 2006, Lecture Notes in Computer Science, 4132, pp. 847-855.

Ruan, G. and Tan, Y. 2010. A Three-Layer Back-Propagation Neural Network for Spam Detection using Artificial Immune Concentration. Soft Computing, 14(2), pp. 139-150.