International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 6, December 2019, pp. 5479~5485
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp5479-5485

Comparative between optimization feature selection by using classifiers algorithms on spam email

Ghada Rawashdeh, Rabiei Mamat, Zuriana Binti Abu Bakar, Noor Hafhizah Abd Rahim
Department of Computer Science, University Malaysia Terengganu, Malaysia

Article Info

Article history:

Received Dec 7, 2018
Revised Apr 18, 2019
Accepted Jun 11, 2019

Keywords:

Classifiers algorithm
Email spam
Meta-heuristic
Optimization feature selections

ABSTRACT

Spam mail has become a rising phenomenon in a world that has recently witnessed high growth in the volume of emails. This indicates the need to develop an effective spam filter. At the present time, Classification algorithms for text mining are used for the classification of emails. This paper provides a description and evaluation of the effectiveness of three popular classifiers using optimization feature selections, such as Genetic algorithm, Harmony search, practical swarm optimization, and simulating annealing. The research focuses on a comparison of the effect of classifiers using K-nearest Neighbor (KNN), Naïve Bayesian (NB), and Support Vector Machine (SVM) on spam classifiers (without using feature selection) also enhances the reliability of feature selection by proposing optimization feature selection to reduce number of features that are not important.

Corresponding Author:

Ghada Rawashdeh,
Department of Computer Science,
University Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia.
Email: ghada_rwashdeh@yahoo.com

1. INTRODUCTION

Currently, there is no exact definition for spam, however, spam is often referred to as unsolicited email, but it is not all unsolicited e-mails are spams. Spam could also be said to unsolicited commercial e-mail [1], but unfortunately, not all advertising materials are spam. Even though most e-mail users are aware of what spam represents, how spam and spamming are defined is still not clear. Supervised learning is the machine learning task of generating a mapping from labeled or supervised training data to a class of predictions or output [2]. A major aspect of supervised learning is a classification task whose aim is to establish a function from known as labels or classes from input objects to output values. The set of labeled examples is called a training set. Then a classifier is used alongside the training set to generate a mapping from examples to labels. Subsequently, the trained classifier could be used for classification.

Text classification is prone to several challenges due to a large number of features in the dataset. The applicability of the existing classification techniques in these datasets is restricted due to the huge number of features [3]. Several IR techniques such as Stemming, Feature selection (FS) and Stop-words Removal have applied for feature space dimension reduction. The FS techniques like Chi-Square Statistic (CHI), Mutual Information (MI), and Information Gain (IG) are employed for feature dimensionality reduction via the elimination of irrelevant features for a given category [4-7].

Nature-inspired meta-heuristics have found applicability in several fields, including in computer science, and data mining [8, 9]. For instance, the genetic algorithm (GA) has been employed by [10] as an FS method using Multi-Layer Perceptron (MLP) as the classifier ‘bag of the word’ as the extraction method. The performance of these techniques was benchmarked against Neural Network (NN), Naïve Bayes (NB, and support vector machine (SVM) and found to outperform the other classifiers. The performance of different

classifiers on spam classification has been compared by [11] using 'bag of the word' as an extraction technique without FS. From the results, NB performed better than SVM and tree-based J48.

This research is significant as it aims to improve the performance of spam classifiers and determine the spams available in our datasets with improved performance. It also determines the most suitable optimization to address the weakness of the spam classifiers in a large amount of feature. In addition, this research incorporates suitable classifiers with optimization feature selection algorithm. This is important because the current techniques of spam classifiers are still not effective. With the rapid increase in internet usage, the technology for automatic classification of a huge amount of email information has come to play a very significant role. A wide range of emails is increasingly disseminated almost daily, making it more difficult to detect the emails. Another issue of developing TC system is how to handle the huge amount of features, which can easily reach many people [4, 12].

The performance of three classifiers, namely: NB, J48, and IB1 have been compared by [13] using 'bag of the word' as an extraction technique and NB was found as the best classifier. Furthermore, the FS performance of Pearson correlation, Mutual Information, Chi-square, and Symmetric Uncertainty has been compared and Chi-square was found as the better classifier. Another study compared the performance of SVM, AdaBoost, and Random Forests (RF) as FS techniques using 'bag of the word' as the extraction technique. The study found SVM as the best classifier. Two FS techniques (Information Gain and Chi-Square) have also been compared using SVM as a classifier. The comparison result showed the use of Chi-Square as the FS method and RF as the classifier to be better than using SVM with FS.

Additionally, six FS techniques (SVM, NB, Optimal Document Frequency-based Feature Selection (ODFFS), Term Frequency-based Feature Selections (TFFSs), and a hybrid method (HBM)) has been compared by [14]. They suggested the Feature Subset Evaluating Parameter Optimization (FSEPO) for parameter optimization. From the observed performance of the classifiers, NB performed better than other classifiers. This is one of the works that informed the current study to use optimization feature selection for spam optimization.

[15] employed feature selections of SVM using bag of word as extraction to identify spam classification. According to the study, the best type of SVM was Gaussian Kernel which performed better than Polynomial Kernel and Linear Kernel of SVM classifier. On the other hand, Kernel-Penalized outperformed Recursive Feature Elimination (RFE), Fisher, Kernel-Penalized, and Feature Selection ConcaVe (FSV).

The current research is consistent with the research conducted by [11] which revealed that NB was better without feature selection. However, [11] used tree-based J48 while our research used KNN with NB and tested the classifier with OFS. This research is also in line with [13] who found that NB is the best classifier. Nevertheless, our research disagrees with [10] who compared NB with the genetic algorithm as feature selection, but our research focusses on the effect of OFS on the classifiers.

The presence of the crossover and mutation operators in the GA makes it seem similar to biological evolution process of chromosomes. The chromosomes are assessed based on their fitness function to select parents as they provide a solution to problems. The selection of the parents preceded new population generation; hence, it is a key process which can affect the GA convergence. The convergence speed of different selection schemes was first studied by [16].

The Particle Swarm Optimization (PSO) was developed based on inspiration from the flocking of birds or insects when determining the optimal solution. Thus, it is a swarm-based meta-heuristic optimization technique. The PSO is prone to certain problems such as premature convergence despite its accuracy [17]. This differentiates PSO from the rest of the mathematical frameworks. Additionally, PSO can accurately determine the global optimal position for single-peak-search problems. Meanwhile, there is a tendency of the PSO being trapped in a local optimum when faced with complex multi-peak-search problems. There is no mutation operator in the PSO as obtainable in the GA, but it can still achieve the best results based on the interaction between the particles.

Harmony Search (HS) was presented by [18] as a meta-heuristic that imitates the process of music improvisation, a situation where musicians aim to achieve a perfect state of harmony by improvising the pitches of their instruments. The performance of the HS algorithm has been proven through its application to different problems. Additionally, the capability of HS to explore an entire search space has been demonstrated by its explorative power. Besides, the explorative power of the HS has been ensured owing to the evolution of the expected population variance over generations. It takes the HS a long time to converge to a globally optimal partition. The HS optimization can be adopted to identify a global near-optimal solution because it is good at avoiding local optimal solution convergence [19]. Thus, HS has been proven effective in several optimization problems [20-22]. The HS technique, unlike the conventional optimization methods, can yield several merits as briefly discussed below.

The HS algorithm demands lesser mathematical computation and has no need for the initial value settings for decision variables. Derivative information is not needed in the HS algorithm because it depends on stochastic random searches. The HS algorithm generates a new vector by considering all the existing vectors, unlike the other methods, like the GA, which consider only the two-parent vectors. Therefore, the HS is a flexible algorithm although several problems still need to be solved as it applies more control parameters techniques.

The single-based approaches are the oldest but simplest meta-heuristics that perturb a single solution at each iteration. As defined by [23]a local search is an "algorithmic method for searching a given space of candidate solutions, starting from an initial candidate solution to and iteratively moves to a candidate solution from its direct neighborhood, based on local information, until a termination condition is satisfied."

In this technique, the single-based which is employed for addressing a given problem initializes with the help of an initial candidate solution before considering the neighbors each time with respect to the current solution as a likely alternative for the improvement of the objective function of any given solution until the desired search condition is achieved. Furthermore, a new solution would be accepted from the neighborhood if it is better than the existing possible solutions in the candidate sets. A critical example of this can be drawn from Hill climbing [24] Simulated Annealing [25] and Tabu Search [26], as studies of single-based approaches.

Simulated annealing is a technique for solving combinatorial optimization problems [27], such as the minimization of the functions of numerous variables. Owing to the fact that several real-world problems can be considered as optimization problems, interests in general techniques for addressing such problems have increased. One of such techniques with an unusual lineage is Simulated annealing which was inspired by the statistical mechanics of annealing in solids. A better understanding of such physics problem can be achieved by considering the process of coercing a solid into a low energy state (typically referring to a greatly ordered state, such as a crystal lattice). In the simulated annealing methods, the analogous set of "controlled cooling" operations are used for non-physical optimization problems; in effect, the process of transforming a poor, unordered solution into a highly desirable optimized solution. It provides an effective physical analogy for solutions to optimization problems. It can also reshape mathematical insights from the physics perspective into the perspective of the real optimization problem.

Therefore, previous studies focused on the weaknesses of spam classifiers [23, 24]. However, the present is envisioned to improve the performance of spam classifiers. The faults of spam classifiers are characterized by being insufficient to handle the huge volume of relevant emails and efficiently detect the new spam email, which is regarded as the main problem in spam classification. This problem is highly dependent upon the features used in processing spam classifiers and selecting the huge amount of feature.

Another problem is related to the ambiguity of the effect of optimization feature selection on multiple classifier algorithms which are commonly used in the previous works namely, such as K-NN, SVM, and NB. Some of the previous studies used some techniques to reduce the number of features but did not discover whether the reduction could affect the performance of the spam classifiers. Meta-heuristic optimization is used to determine the optimal solution between possible multi-solutions. With regard to this research, it is imperative to examine the performance and optimization of spam email.

2. RESEARCH METHOD

This study aims to develop a solution method for a given problem; this is the standard research methodology in computer science. The adopted methods in this research consist of different phases, including Groundwork, Induction, Improvement, Evaluation, and Quality Comparison [28-30].

In this study, the groundwork relates to the problem of spam classification which was addressed by reviewing the existing literature on each problem. The aim of the literature review is to detect the weaknesses in the existing spam classifiers. This section consists of problem identification and detection and identification case. also, Induction Phase in This research focuses on Pre-processing and construction of spam classifiers.

The improvement phase mainly aims to improve the spam classification quality in two perspectives: first, to extract the most significant terms in the email in order to increase the relevant terms and avoid irrelevant terms in emails. Secondly, it increases the effectiveness of reducing the number of feature selection obtained in the previous phase by maximizing the similarity of the value of accuracy function as much as possible, which is the performance of selected features. This was achieved through the application of the proposed framework for spam classification on the investigated problem to improve the quality by the terms and reduce the number of feature for classifiers by using (GA, HS, SA, PSO). This research proposed two different steps to improve spam classifiers. Firstly, the three classifiers in this step will categorize the new email into spam or non-spam. The output of this step will address the problem in spam

classifiers which the features effect on the performance. Secondly, the effect of optimization feature selection on the improvement of classifiers which optimization feature selection used here to decrease the number of features as the Figure 1.

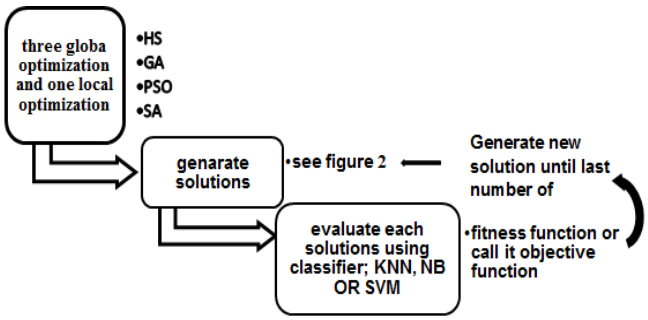

Figure 1. The state of art evaluates the global and local optimization feature selection on three classifiers

Figure 1 showed the methodology of our experimentation, consisting of the application of the algorithms and performance evaluation based on the performance measures as shown in section 3. A-Application of Algorithms: The algorithms discussed above are global search (generates more than one solution) like GA, HS, PSO, and SA is local search (generate just one solution), are applied with the use classifier algorithm (KNN, SVM, NB). The split between training and test data set is 50% training and 50% testing Cross-validation.

1 - Initialization (Parameters Initialization, random Initialization of feature selection)
2 - Evaluate the fitness function of each initial solution and take the best
3 - Update solution FOR EXAMPLE
  – in GA (crossover operator) Select two individual and swap a solution of gene between the feature
  – in SA Generate candidate solution Y based on mutation operator: Select one individual and mutate the feature in it.
4 - IF NEW SOLUTION(Y) better than OLD SOLUTION (X)-swap- ELSE- go to the step3
5 - Is the stopping criteria satisfied-STOP-ELSE- go to the step3

The algorithm used some representations to code the whole F of the features set in a vector of length m, where m represents the number of features as depicted in Figure 2. In this vector, each element is a label for features to be selected or dropped. These solutions are exemplified in Figure 2. In this case, 57 features in s1{1, 4, .., and57} were selected while the others were dropped {2, 3, 5, 6,….}, and so on. S1 is solution one, S2 is solution two, and F is the number of features in the dataset.

F	1	2	3	4	5	6	…….....	57	cost
S1	1	0	0	1	0	0	.....	1	45
S2	0	1	0	0	1	0	.....	0	17

Figure 2. Two solutions for the dataset has 57 features or terms

3. RESULTS AND DISCUSSION

In assessing the quality of classifiers quality, three forms of measures are used, namely: f-measurement, accuracy, and error rate [25]. In the current study, accuracy is used as an external quality measure, which is one of the most commonly used measures in text mining.

3.1. Performance measures (accuracy)

In the classification of problems, the evaluation measures are generally defined from a matrix with the exact number of examples that have been correctly and wrongly classified for each class (called 'confusion matrix'). Table 1 shows the confusion matrix of a binary classification problem with just 2 classes (positive and negative).

Table 1. Confusion matrix

	Predicted Class	
True class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The accuracy rate (ACC) is the commonest evaluation measure used in practice. The percentage of its correct predictions is used to evaluate classifier's effectiveness. It is computed thus:

$$ACC = ((TP+TN)/ (TP+TN + FP+FN)) *100 \quad (1)$$

3.2. Performance measures (F-measurement)

The F-measure is a combination of precision and recall ideas from information retrieval. Here, each class is considered as the result of an email and each class is regarded as the desired set of emails or spam. The recall and precision for each email j and class i is calculated as follows:

$$\text{Recall (i,j)} = \frac{n_{ij}}{n_i}. \quad (2)$$

$$\text{Precision (i,j)} = \frac{n_{ij}}{n_j}. \quad (3)$$

Here, n_{ij} represents the number of emails having the class label i in class j, and n_i refers to the number of emails having the class label i. Finally, n_j is the number of emails in class j. The calculation of the F-measure of email j and class i is presented as follows:

$$\text{F (i,j)} = \frac{2\text{Recall (i,j)Precision (i,j)}}{\text{Recall (i,j)+Precision (i,j)}}. \quad (4)$$

The overall F-measure value is calculated by considering the weighted average of all the F-measure values as follows:

$$F=\sum_i \frac{n_i}{N} maxF(i,j). \quad (5)$$

Thus, the F-measure value is observed to be in the range of (0,1), where larger values represent a higher classifier quality.

3.3. Data sets used in the experiments

This study uses a spam email dataset that is publicly available in the UCI Machine Learning Repository, i.e. SPAM E-mail Database. It contains 57 attributes and 4601 emails, with 1813 emails being spam while the rest (2788) are normal emails. The dataset is multivariate with real integer attributes.

As in the Table 2, there are three classifiers (K-NN, NB, SVM) tested in Matlab using email spam dataset. This research relied on the accuracy result and F-measurement as mentioned before. We found that the SVM algorithm is the best classifier, followed by KNN and NB. Without using an optimization feature selection method, the number of features achieved was 57.

Table 2. The result of three classification algorithm without feature selection

Classifier Algorithm	Accuracy	f-Measurement	Elapsed Time	Number of features before selected
K-NN	79.096045	0.729517	11.666530 seconds	57
NB	51.977401	0.616453	1.697521 seconds.	57
SVM	91.0039	0.887439	2.427657 seconds	57

Table 3 shows the results of the effect by using popular three classifiers on spam email. OFS was used to reduce the number of features. After testing the algorithms on MATLAB three of them were identified as global optimization while one was observed as local optimization. It is observed the performance for all of these algorithms achieved Good result through using one dataset on SVM classifier and all of them reduce the number of features.

Table 3. Results of the effect of using OFS on three classification algorithms

Optimization Feature Selection	Classifier Algorithm	Accuracy	f-Measurement	Number of Feature before selected	Number of features after selected
GA	KNN	0.877010	0.843039	57	32
	NB	0.548023	0.635088	57	48
	SVM	0.906128	0.880531	57	32
HS	KNN	0.786180	0.745868	57	20
	NB	0.481530	0.599261	57	25
	SVM	0.808779	0.744186	57	31
PSO	KNN	0.787484	0.736388	57	48
	NB	0.597526	0.684321	57	48
	SVM	0.922208	0.905640	57	50
SA	KNN	0.851369	0.808939	57	24
	NB	0.6012584	0.713258	57	21
	SVM	0.928547	0.918932	57	22

4. CONCLUSION

This paper classified emails based on three classifiers: SVM, KNN, and NB. These classifiers were evaluated using MATLAB to separate spam from the email dataset. Each email was identified as spam (1) or not spam (0), reflecting the attributes of the email dataset for spam classification. SVM gave the most accurate result in the experiment. K-NN classifier also showed good results, but NB reported poorer results compared with SVM or K-NN classifier. Nevertheless, while using OFS, we reduced the number of features. As a result, all optimization feature selection algorithm gave a good performance. Much work needs to be done in the future. Future research may use the hybridization between global search and local search algorithm and use more datasets.

REFERENCES

[1] J. A. Zdziarski, "*Ending spam: Bayesian content filtering and the art of statistical language classificatio*," No Starch Press, 2005.

[2] N. Spirin and J. Han, "Survey on web spam detection: principles and algorithms," *ACM SIGKDD explorations newsletter*, vol. 13, pp. 50-64, 2012.

[3] J. C. Bansal, *et al.*, "Artificial bee colony algorithm: a survey," *International Journal of Advanced Intelligence Paradigms*, vol. 5, pp. 123-59, 2013.

[4] S. Al-Harbi, *et al.*, *"Al-Rajeh A. Automatic Arabic text classification,"* Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, 2008.

[5] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, pp. 1-47, 2002.

[6] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, pp. 1289-305, 2003.

[7] D. Fragoudis, *et al.*, "Best terms: an efficient feature-selection algorithm for text categorization," *Knowledge and Information Systems*, vol. 8, pp. 16-33, 2005.

[8] A. Singh, *et al.*, *"Online Mining of data to generate association rule mining in large databases,"* 2011 International Conference on Recent Trends in Information Systems, pp. 126-131, 2011.

[9] P. Parveen and G. Halse, "Spam mail detection using classification," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, 2016.

[10] S. DeepaLakshmi and T. Velmurugan, "Empirical study of feature selection methods for high dimensional data," *Indian Journal of Science and Technology*, vol. 39, 2016.

[11] D. Liu, *et al.*, "Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry," *Food and Bioprocess Technology*, vol. 7, pp. 307-23, 2014.

[12] S. Eyheramendy, *et al.*, "On the naive bayes model for text categorization," 2003.

[13] S. Maldonado and G. L'Huillier, "SVM-based feature selection and classification for email filtering," *Pattern recognition-applications and methods*, *Springer*, Berlin, Heidelberg, pp. 135-148, 2013.

[14] W. H. Steeb, "*The nonlinear workbook: Chaos, fractals, cellular automata, genetic algorithms, gene expression programming, support vector machine, wavelets, hidden Markov models, fuzzy logic with C++*," World Scientific Publishing Company, 2014.

[15] K. S. Lee and Z. W. Geem, "A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice," *Computer methods in applied mechanics and engineering*, vol. 194, pp. 3902-33, 2005.

[16] Z. W. Geem, *et al.*, "Application of harmony search to vehicle routing," *American Journal of Applied Sciences*, vol. 2, pp. 1552-7, 2005.

[17] Z. W. Geem, "Novel derivative of harmony search algorithm for discrete design variables," *Applied mathematics and computation*, vol. 199, pp. 223-30, 2008.

[18] Z. W. Geem, *"Music-inspired harmony search algorithm: theory and applications,"* Springer, 2009.

[19] M. Steinbach, *et al*., "A comparison of document clustering techniques," *KDD workshop on text mining*, vol. 400, pp. 525-526, 2000.
[20] N. K. Jhankal and D. Adhyaru, "Comparative analysis of bacterial foraging optimization algorithm with simulated annealing," *Int. J. Sci. Res.(IJSR)*, vol. 3, pp. 10-3, 2014.
[21] S. Abdullah, *et al.*, *"A constructive hyper-heuristics for rough set attribute reduction*," 2010 10th International Conference on Intelligent Systems Design and Applications, pp. 1032-1035, 2010.
[22] A. Hatamlou, *et al*., "A combined approach for clustering based on K-means and gravitational search algorithms," *Swarm and Evolutionary Computation*, vol. 6, pp. 47-52, 2012.
[23] F. Hutter, *et al.*, "Efficient stochastic local search for MPE solving," *IJCAI*, pp. 169-174, 2005.
[24] N .Jozefowiez, *et al*., "An evolutionary algorithm for the vehicle routing problem with route balancing," *European Journal of Operational Research*, vol. 195, pp. 761-9, 2009.
[25] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," *The Journal of Physical Chemistry A*, vol. 101, pp. 5111-6, 1997.
[26] F. Glover, "Tabu search—part I," *ORSA Journal on computing*, vol. 1, pp. 190-206, 1989.
[27] S. S. Shreem, *et al.*, "Hybridising harmony search with a Markov blanket for gene selection problems," *Information Sciences*, vol. 258, pp. 108-21, 2014.
[28] M. Mahdavi and H. Abolhassani, "Harmony K-means algorithm for document clustering," *Data Mining and Knowledge Discovery*, vol. 18, pp. 370-91, 2009.
[29] G. Fanelli, *et al*., "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, pp. 437-58, 2013.
[30] C. Szegedy, *et al*., *"Going deeper with convolutions,"* Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.