International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 839 - 843, 2014, ISSN:- 2319–8656

Spam filtering by using Genetic based Feature Selection

Sorayya Mirzapour Kalaibar
Department of Computer, Shabestar Branch
Islamic Azad University
Shabestar, Iran

Seyed Naser Razavi
Computer Engineering Department
Faulty of Electrical and Computer Engineering,
University of Tabriz, Iran

Abstract: Spam is defined as redundant and unwanted electronical letters, and nowadays, it has created many problems in business life such as occupying networks bandwidth and the space of user's mailbox. Due to these problems, much research has been carried out in this regard by using classification technique. The resent research show that feature selection can have positive effect on the efficiency of machine learning algorithm. Most algorithms try to present a data model depending on certain detection of small set of features. Unrelated features in the process of making model result in weak estimation and more computations. In this research it has been tried to evaluate spam detection in legal electronica letters, and their effect on several Machin learning algorithms through presenting a feature selection method based on genetic algorithm. Bayesian network and KNN classifiers have been taken into account in classification phase and spam base dataset is used.

Keywords: Email spam, feature selection, genetic algorithm, classification.

1. INTRODUCTION

Nowadays, e-mail has been widely considered as one of the fastest and most economical forms of communication .Thus, the e-mail is prone to be misused. such misuse is posting unsolicited, unwanted e-mails known as spam or junk e-mails[1]. Spam has been considered as a serious problem. Many Internet Service Providers (ISPs) receive more than billion spam messages per day. Much of these e-mails are filtered before end users can access them. Content-Based filtering is a key technological method for e-mail filtering. The spam e-mail contents usually contain common words called features. Frequency of occurrence of these features inside an e-mail gives an indication that the e-mail is a spam or legitimate [2,3,4]. There are various purposes in sending spams such as economical purposes. Some spams are unwanted advertising and commercial messages, while others deceive the users to use their private information (phishing), or they temporarily destroy the mail server by sending malicious software to the user's computer. Also, they create traffic, or distribute immoral messages. Therefore, it is necessary to find some ways to filter these troublesome and annoying emails automatically. In order to detect spams, some methods such as parameter optimization and feature selection methods have been proposed in order to reduce processing overhead and to guarantee high detection rate [16].The spam filtering is high sensitive application of text classification (TC) task. The main problem in text classification tasks which is more serious in email filtering is existence of large number of features. For solving the issue, various feature selection methods are considered. They extract one, and offer it as input to classifier[5]. In this paper, we incorporate genetic algorithm to find an optimal subset of features of the spam base data set. The selected features are used for classification of the spam base.

2. LITERATURE REVIEW

Features selection approaches are usually employed to reduce the size of feature set, and to select a subset of original features. Over the past years, the following methods have been considered to select effective features such as the algorithms based on population to select important features, and to remove irrelevant and redundant features such as genetic algorithm (GA), particle swarm optimization (PSO), and ant colony algorithm (ACO). Some algorithms are developed to classify and filter e-mails. The RIPPER algorithm [6] is a rule-based algorithm used for filtering e-mails. Drucker, et. al. [7] proposed an SVM algorithm for spam categorization. Sahami, et. al. [8] proposed Bayesian junk E-mail filter using bag-of-words representation and Naïve Bayes algorithm. Clark, et. al. [9] used the bag-of-words representation and ANN for automated spam filtering system. Branke, J. [10] discussed how genetic algorithm can be used to assist designing and training. Riley. J. [11] described a method of utilizing genetic algorithms to train fixed architecture feed-forward and recurrent neural networks. Yao. X. and Liu. Y. [12] reviewed different combinations between ANN and GA, and used GA to evolve ANN connection weights, architectures, learning rules, and input features. Wang and et al. presented feature selection incorporation based on genetic algorithm and support vector machine based on SRM to detect spam and legitimate emails. The presented method had better results than main SVM [13]. Zhu developed a new method based on rough set and SVM in order to improve the level of classification. Rough set was used as a feature selection to decrease the number of feature and SVM as a classifier [14]. Fagboula and et al. considered GA to select an appropriate subset of features, and they used SVM as a classifier. In order to improve the classification accuracy and computation time, some experiments were carried out in terms of data set of Spam assassin [15]. Patwadhan and Ozarkar presented random forest algorithm and partial decision trees for spam classification. Some feature selection methods have been used as a preprocessing stage such as Correlation based feature selection, Chi-square, Entropy, Information Gain, Gain Ratio, Mutual Information, Symmetrical Uncertainty, One R and Relief. Using above mentioned methods resulting in selecting more efficient and useful features decrease time complexity and increase accuracy [17].

3. GENETIC ALGORITHMS

A genetic algorithm (GA) is one heuristic techniques that are based on natural selection from the population members, and tries to find high-quality solutions to large and complex optimization problems. This algorithm can identify and exploit regularities in the environment, and converges on solutions (it can also be regarded as locating the local maxima) that were globally optimal [18]. This method is very effective and widely used to find-out optimal or near optimal solutions to a wide variety of problems. The genetic algorithm repeatedly modifies the population of individual solutions. At each step, the genetic algorithm tries to select the best individuals. Now, "parent" population genetic algorithm creates "children" constituting next generation. Over successive generations, the population evolves toward an optimal solution. The genetic algorithm uses three main rules at each step to create next generation: a. Select the individuals, called parents that contribute to the population at the next generation. b. Crossover rules that combine two parents to form children for the next generation. c. Mutation rules, apply random changes to individual parents to form children

4. FEATURE SELECTION

Features selection approaches are usually employed to reduce the size of feature set, and to select a subset of the original features. We use the proposed genetic algorithms to optimize the features that significantly contribute to the classification.

4.1. Feature Selection Using Proposed Genetic Algorithm

In this section, the method of feature selection by using the proposed genetic Algorithm has been presented. The procedure of the proposed method has been stated in details in the following section.

4.1.1. Initialize population

In the genetic algorithm, each solution to the feature selection problem is a string of binary numbers called chromosome. In this algorithm, initial population is generated randomly. IN feature representation is considered as a chromosome, and if the value of chromosome [i] is 1, the ith feature is selected for classification, while if it is 0, then these features will be removed [19,20]. Figure 1 shows feature presentation as a chromosome.

Chromosome:

F_1	F_2	F_3	...	F_{n-1}	F_n
1	0	1	...	1	0

Figure 1. Feature Subset: $\{F_1, F_3, \ldots, F_{n-1}\}$

In this research, we used weighted F-score to calculate the fitness value of each chromosome. The algorithm starts by randomly initializing a population of N number of initial chromosome.

4.1.2 Cross over

The crossover is the most important operation in GA. Crossover, as its name suggests, is a process of recombination of bit strings via the exchange of segments between pairs of chromosomes. There are various kinds of crossover. In one point of cross-over, a bit position is randomly selected that should be changed. In this process, a random number is generated. This number (less than or equal to the chromosome length) is the crossover position [21]. Here, one crossover point is selected, binary string from beginning of chromosome to the crossover point is copied from one parent, and the rest is copied from the second parent [22].

4.1.3. Proposed mutation

In mutation, it can be ensured that all possible chromosomes can maintain good gene in the newly generated chromosomes. In our approach, Mutation operator is a two-steps process, and is a combination of random and substitution mutation operator. Also it occurs on the basis of two various mutation rates. In mutation operator, substitution step is considered with the probability of 0.03. In each generation, the best chromosome involving better features and higher fitness is selected, and it substitutes for the weakest chromosome having lesser fitness than others. In this stage, the better chromosome transfers the current generation to next generation, and it follows rapid convergence of algorithm. Otherwise, it enters the second mutation step with probability of 0.02. This step changes some gens of chromosome randomly by inverting their binary cells. In fact, the second one is considered to prevent reducing exploration capability of search space to keep diversity in other chromosomes. Generally, mutation probability is equal to 0.05.

5. RESULTS SIMULATION

In order to investigate the impact of our approach on email spam classification, spam base data set downloaded from the UCI Machine Learning Repository has been used [23]. Data set of Spam base involving 4601 emails was proposed by Mark Hopkinsand his colleagues. In This data set that is divided into two parts, 1 shows spam, and zero indicates non-spam. This data set involves 57 features with continuous values. In simulation of the proposed method, training set involving 70% of the main data set and two experimental sets have been separately considered for feature selection and classification. Each one involves 15% of the main data set. After performing feature selection by using the training set, the test set was used to evaluate the selected subset of features. The evaluation of overall process was based

on weighted f-score which is a suitable measure for the spam classification problem. The performance of spam filtering techniques is determined by two well-known measures used in text classification. These measures are precision and recall [24, 25]. Here four metric have been used for evaluating the performance of proposed method such as precision, accuracy, recall and F1 score. These metrics are computed as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$F1 = \frac{2 \times \pi \times \rho}{(\pi + \rho)} \quad (3)$$

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \quad (4)$$

Where:
TP_i = the number of test samples that have been properly classified in c_i class.
FP_i = the number of test samples that have been incorrectly classified in c_i class.
TN_i = the number of test samples belonging to c_i class, and have been correctly classified in other classes.
FN_i = the number of test samples belonging to c_i class, and have been incorrectly classified in other classes.
The methods of Bayesian network and K nearest neighbors algorithm (KNN) have been used for classification. The executed program and the obtained average have been compared 8 times to investigate the performance of each classifier. The results obtained from the proposed method of feature selection have been compared without considering feature selection. The obtained results show that when the parameters are presented in tables 1, the best performance is observed in terms of GAFS.

Table 1: the parameters of feature selection by using genetic algorithm

Initial population	80
Mutation rate1	0.03
Mutation rate2	0.02
Crossover	0.7
Generations	100

6. RESULT EVALUATION

In this section, the results of experiments have been presented to evaluate efficiency of the proposed method. The results of comparing classifier Bayesian network and KNN have been presented in table 2. In addition, figure 2 shows graphical diagram of the effects of the proposed method on reduction of redundant features. According to the results obtained in terms of Bayesian network classification, the proposed method has increased the classification accuracy, and at the same time, it has removed significant number of features. Also, although KNN classifier has removed some features, it has reached the accuracy that is near to the accuracy obtained before selecting the feature. The results obtained for three other criteria have been demonstrated in table 3. As it can be observed in table, Bayesian classifier has been considerably optimized in all three evaluation criteria, while KNN classifier has reached to the precision near to the previous precision. These results indicate that feature selection by GA technique improves email spam classification. GA FS and all features by using mentioned classifiers have been compared in terms of accuracy, number of selected feature, recall, precision and F score of spam class

Table 2: comparing feature selection methods in terms of accuracy

Algorithms / classifier	All Feature	GA FS
Bayesian network	0.891	0.918
KNN (N=1)	0.9	0.891

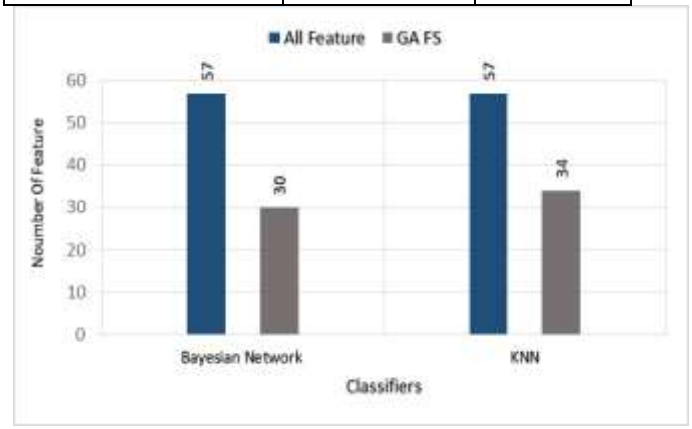

Figure 3: column graph of comparing the number of selected features

Table 3: comparing feature selection methods

classifier / measures	KNN(N=1)		Bayesian network	
	All Feature	**GA FS**	**All Feature**	**GA FS**
precision	0.892	0.886	0.89	0.935
recall	0.871	0.860	0.851	0.869
F1 score	0.882	0.871	0.87	0.900

7. CONCLUTION

In this paper, the proposed GA based feature selection method has been presented and evaluated by using data set of Spam Base. The results obtained from proposed method were compared with position without feature selection. The obtained results show that, with regard to the number of removed features, the proposed method has accuracy comparable with the methods that lack feature selection. In addition, in Bayesian network classifier, better results have been obtained compared to KNN classifier and all evaluation criteria have been considerably improved. Therefore, the proposed method has considerable effect on features selection and increasing the accuracy. We can use parameter optimization in this work. Also, the proposed algorithm can be combined with other classification algorithms in the future.

REFERENCE

[1] GOWEDER, A. M., RASHED, T., ELBEKAIE, A., & ALHAMMI, H. A. (2008). An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. Paper presented at the Proceedings of the 2008 International Arab Conference on Information Technology.

[2] Bruening, P.(2004). Technological Responses to the Problem of Spam: Preserving Free Speech and Open Internet Values. First Conference on E-mail and Anti-Spam.

[3] Graham, P.(2003). A Plan for Spam. MIT Conference on Spam.

[4] William, S., et. al. (2005). A Unified Model of Spam Filtration, MIT Spam Conference, Cambridge.

[5] GOWEDER, A. M., RASHED, T., ELBEKAIE, A., & ALHAMMI, H. A. (2008). An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. Paper presented at the Proceedings of the 2008 International Arab Conference on Information Technology.

[6] Cohen, W. (1996). Learning Rules that Classify E-mail, In AAAI Spring Symposium on Machine Learning in Information Access, California.

[7] Drucker, H., et. al.(1999) Support Vector Machines for Spam Categorization, In IEEE Transactions on Neural Networks.

Sahami, M., et. al.,(1998). A Bayesian Approach to Filtering Junk E-Mail, In Learning for Text Categorization, AAAI Technical Report, U.S.A.

[8] Riley. J. (2002). An evolutionary approach to training Feed-Forward and Recurrent Neural Networks", Master thesis of Applied Science in Information Technology, Department of Computer Science, Royal Melbourne Institute of Technology, Australia.

[9] Clark, et. al. (2003). A Neural Network Based Approach to Automated E-Mail Classification, IEEE/WIC International Conference on Web Intelligence.

[10] Branke, J. (1995). Evolutionary algorithms for neural network design and training, In Proceedings 1st Nordic Workshop on Genetic Algorithms and its Applications, Finland.

[11] Yao. X., Liu. Y. (1997). A new evolutionary system for evolving artificial neural networks", IEEE Transactions on Neural Networks.

[12] Wang, H.-b., Y. Yu, and Z. Liu. (2005) SVM classifier incorporating feature selection using GA for spam detection, in Embedded and Ubiquitous Computing–EUC 2005., Springer. p. 1147-1154.

[13] Zhu, Z. (2008). An email classification model based on rough set and support vector machine. in Fuzzy Systems and Knowledge Discovery.

[14] .Temitayo, F., O. Stephen, and A. Abimbola. (2012). Hybrid GA-SVM for efficient feature selection in e-mail classification. Computer Engineering and Intelligent Systems. 3(3): p. 17-28.

[15] Stern, H. (2008) A Survey of Modern Spam Tools. in CEAS. Citeseer.

[16] Ozarkar, P. and M. Patwardhan. (2013). INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor. 4(3): p. 123-139.

[17] Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.

[18] Vafaie H, De Jong K. (1992). Genetic algorithms as a tool for feature selection in machine learning. In Proceedings of Fourth International Conference on Tools with Artificial Intelligence (TAI '92). 200-203.

[19] Yang J, Honavar V. (1998). Feature subset selection using a genetic algorithm. Intelligent Systems and their Applications, IEEE, 13(2):44-49.

[20] Shrivastava, J. N., & Bindu, M. H. (2014). E-mail Spam Filtering Using Adaptive Genetic Algorithm. International Journal of Intelligent Systems & Applications, 6(2).

[21] Karimpour, J., A.A. Noroozi, and A. Abadi. (2012). The Impact of Feature Selection on Web Spam Detection. International Journal of Intelligent Systems and Applications (IJISA), 4(9): p. 61.

[22] UCI repository of Machine learning Databases. (1998). Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/~mlearn/MLRepository.html, Hettich, S., Blake, C. L., and Merz, C. J.

[23] Liao, C., Alpha, S., Dixon.P. (2004). Feature Preparation in Text Categorization, Oracle Corporation.

[24] Clark, et. al. (2003). A Neural Network Based Approach to Automated E-Mail Classification, IEEE/WIC International Conference on Web Intelligence