ISSN 2277-3061

Feature Selection by Using Discrete Imperialist Competitive Algorithm to Spam Detection

[1]Sorayya mirzapour kalaibar, [2]Seyed Naser Razavi

[1]Department of computer, Shabestar branch,Islamic Azad University, Shabestar,iran

[1]mirzapour.sy@gmail.com

[2]Dept. of Electrical Engineering, University of Tabriz, Tabriz, Iran

[2]razavi@iust.ac.ir

ABSTRACT

Spam is a basic problem in electronic communications such as email systems in large scales and large number of weblogs and social networks. Due to the problems created by spams, much research has been carried out in this regard by using classification techniques. Redundant and high dimensional information are considered as a serious problem for these classification algorithms due to their high computation costs and using a memory. Reducing feature space results in representing an understandable model and using various methods. In this paper, the method of feature selection by using imperialist competitive algorithm has been presented. Decision tree and SVM classifications have been taken into account in classification phase. In order to prove the efficiency of this method, the results of evaluating data set of Spam Base have been compared with the algorithms proposed in this regard such as genetic algorithm. The results show that this method improves the efficiency of spam detection.

Indexing terms/Keywords

feature selection; Imperialist competitive algorithm; classification; spam; data mining.

Academic Discipline And Sub-Disciplines

Computer Science and Engineering

SUBJECT CLASSIFICATION

Artificial Intelligence

1. INTRODUCTION

Email is one of the most general and quick as well as the cheapest communication tools affecting our life [2]. However one of the disadvantages of this service is increasing number of unfiltered email messages received by the users. Spam is completely a recognized problem experienced by everyone using email. Spam is an unwanted message that is directly or indirectly sent by the sender. It has no relation with the receiver, creates traffic, and involves economical loss. In contrast to spam email, there are legitimate emails. There are various purposes in sending spams such as economical purposes. Some of the spams are unwanted advertising and commercial message, 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 have been proposed in order to reduce processing overhead and to guarantee high detection rate [8].

Email classification has been considered as a serious problem for users and companies [3]. The tasks of email classifier have been divided into several subtasks involving data collection and presenting email message as well as email feature selection and feature dimensionality reduction [3]. The purpose of email classification is to distinguish spam and legitimate messages. The content and form of spams are continuously changing because, in this way, they cannot be detected by anti spams. Hence, this issue has been considered as a research. In data classification process, data set may involve irrelevant and random redundant information decreasing the accuracy of classification. Feature selection is one of the important steps in data mining and knowledge discovery. Finding the appropriate features with N number of features need to evaluate 2n possible subset.

2. LITERATURE SURVEY

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). Guyon and Elisseeff discussed various methods of feature selection and dimensionality reduction, and compared them [16]. 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. GA method was approved in terms of feature selection suitable for SVM classification. Data base of Spam Base was used in the experiments. The presented method had better results than main SVM [9]. Howley and et al. studied the effect of PCA on machine learning accuracy with high-dimensional data in various stages of pre-processing. The results show that using PCA method and classification can improve the classification accuracy in high-dimensional data [10]. 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. The results of experiments carried out in data set of Spam Base showed that, by using rough set, the accuracy has been improved considerably [11]. Ganster and et al. studied the advantages of dimensionality reduction in terms of concept representation to detect spams [12]. Li and et al. proposed the model of spam detection by using random features on the basis of feature selection. Also, they simultaneously proposed parameter optimization. Data set of Spam Base was used in the experiment. The results were summarized as optimization of RF parameters, detection of main features as number value, and spam detection with low process overhead and high detection coefficient. The obtained accuracy was 94.5% [8]. Parimala and et al. presented a new method for feature selection. This method was directed by feature package. The experiments show that the method of feature selection implemented has improved the classification accuracy of support vector machine [13]. Fagboula and et al. considered GA to select an appropriate subset of features, and they used SVM as a classifier. The purpose was to solve an inefficient problem of SVM, and it was considered in email data set involving high dimensions and more computation time. In order to improve the classification accuracy and computation time, some experiments were carried out in terms of data set of Spam assassin. The accuracy of SVM-GA method was improved in comparison to main SVM [2]. Patwadhan and Ozarkar presented random forest algorithm and partial decision trees for spam classification. The performance of this algorithm is better than other algorithms executed previously in terms of accuracy and time complexity. 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 [14].

3. BASIC RULES IN ORIGINAL IMPERIALIST COMPETITIVE ALGORITHM

ICA is an optimization algorithm base on population, and has been recently introduced to deal with various types of optimization problems [15]. This algorithm has been derived from mathematic modeling of imperialistic competition. It has been proposed by Atashpaz and Loucas in 2007 for the first time. Imperialist competitive algorithm, like other population-based algorithms, begins with random initial population, and each one is called a country. Some of the best population elements are selected as an imperialist. The remaining population is taken into account as a colony. Imperialists consider these colonies through a special procedure called assimilation, and it depends on their power. Total power of each empire depends on two parts; namely, the imperialist country and its colonies. In mathematics, this dependency is modeled by defining the empire power as the sum of imperialist country power plus the percent of colonies power average. Imperialist competition begins by forming an initial empire. The empire that cannot be successful in imperialist competition, and cannot increase the power is removed from imperialist competition stage. Therefore, the survival of each empire depends on its power to take the empire colonies on competitor and to possession it.

This competition reduces the power of weaker empires and increases the power of the powerful ones. Any empire that cannot compete with other empires and increase its power or at least prevent decreasing it, will gradually collapse. As a result, after some iterations, the algorithm converges and only one imperialist remains and all other countries are colonies of it [6].

4. THE PROPOSED METHOD

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

4.1. Initialize The Empires

In n-dimensional problem, the country is 1×n array. In feature selection problem, each country is considered by using binary values. Hence, each country is created and initialized randomly by 0, 1 values. In feature representation as a country, if the value of country [i] is 1, the ith feature is selected for classification, while if it is 0, then these features will be removed[7]. Figure 1 shows feature presentation as a country.

Country:

1	0	1	...	1	0

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

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

Figure1: representation of the countries

The power of each country is computed by F-measure related to spam class. F-measure is one of the most common criteria in machine learning and information retrieval[1], and its statements will be presented later.

N number of initial country is randomly created to begin the algorithm. Nimp number of the best member of population is selected as the imperialist, and each one belongs to an empire. In order to divide the initial colonies among the imperialists, some colonies are devoted to each imperialist according to the power and the type of problem. By considering the power of all countries, relative power of each imperialist is computed as follows, and then colonies divided among imperialists.

$$NP_n = \frac{P_n}{\sum_{i=1}^{N_{imp}} P_i} \quad (1)$$

According to computed power, the number of initial colonies is equal to the number of imperialist.

$$NC_n = round\{NP_n \times N_{col}\} \quad (2)$$

Where NCn is the initial number of empire colonies, and Ncol stands for the number of colony countries in population of initial countries. Through considering NCn for each empire, these initial colony countries are selected according to the proposed method and are considered in nth imperialist.

4.2. The Proposed Method To Devote The Colonies To Imperialists

According to (2) equation, in order to devote colonies to empires, some colonies are selected and devoted to imperialists. They are randomly selected in imperialist competition algorithm. In the real world, imperialists usually develop their own colonies by seizing neighboring countries. Imperialists try to seize the colonies that have more benefits. Therefore, we have presented a method on the basis of distance criterion to devote a colony to an imperialist. In this method, the colonies that do not have much distance with the imperialist move toward it. Here, distance criterion means the power. Hence, colonies and imperialists are ranked on the basis of their power, and the more powerful imperialist seizes the more powerful colonies. In this way, other colonies are devoted to the related imperialist according to their power. Imperialist competition algorithm begins with considering the initial mode of all empires. The evolution procedure is considered in a loop, and it continues until the halting condition is satisfied.

4.3. Assimilation

After creating the initial empire, imperialist countries improve their own colonies. For this purpose, the colonies move toward their own imperialists on the basis of assimilation policy.

The original version of ICA is performed in terms of continuous optimization problems. Since feature selection is a discrete problem. The following operator has been taken into account [7]:

For each imperialist and colonies:

- A binary string is created, and random binary values are devoted to each cell.
- Imperialist cells corresponding with position 1 in a binary string are devoted to that position of colony.

4.4. Revolution

When the process of assimilation occurs, the countries experience the revolution process. In this algorithm, revolution makes sudden changes in the social-political features of a country. In fact, the purpose of revolution is to maintain diversity and to prevent creation of local minimums. In proposed revolution operator, at first, a country is randomly selected. Among the features of the selected country, two features are randomly selected, and values of the features in the range of these two points are reversed on the basis of binary possibility. The revolution operator occurs in the colonies with a fixed rate. In order to obtain revolution rate, various values were investigated, and the best one with value of 0.1 was determined.

4.5. Updating Imperialists

After assimilation operator and revolution, the power of countries is computed. If there is more powerful colony, then the colony position is replaced by the imperialist position.

4.6. Computing The Total Power Of Empire

Total power of empires is computed according to follow:

$$TP_n = power(imperialist_n) + \xi\, mean\{power(colonies\ of\ empire_n)\} \tag{3}$$

Where TP_n is total power of nth empire, and ξ is a positive number whose range is usually between zero and one or closer to zero.

4.7. Imperialist Competition

In imperialist competition, all empires attempt to seize the colonies of other empires. Imperialist competition occurs among empires. Each empire that cannot increase its own power, and loses the power of its own competition, will be removed. This means that the more powerful empires takeover the weakest colonies belonging to the weakest empires. Each empire has a likelihood of possessing the mentioned colony.

As you can notice, the most powerful empire does not take possession of the weakest colony of the weakest empire, but it will be more likely to possess the mentioned colony. Considering the total power of empire, The possession probability of empiren is computed as follows:

$$P_{emp_n} = \frac{TP_n}{\sum_{i=1}^{N_{imp}} TP_i} \tag{4}$$

With regard to the possibility of seizing each empire, the competition colony is considered by using roulette wheel on the basis of the power of empires.

4.8. Elimination The Powerless Empires

As it was already mentioned, in imperialist competitions, weak empires gradually lose their own colonies. Figure 2 shows this issue. In this case, imperialist of these empires are devoted to powerful empire in the form of a colony.

ISSN 2277-3061

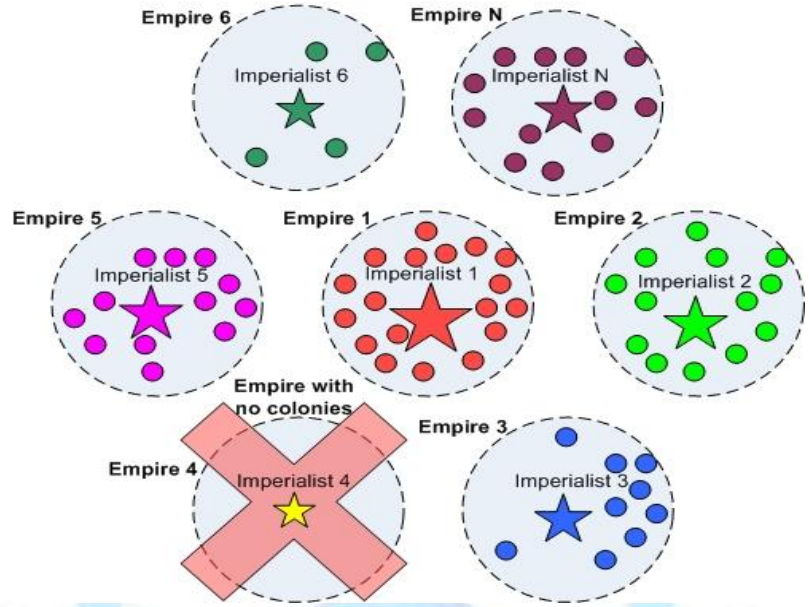

Figure 2: falling weak empire

5. RESULTS SIMULATION

The data of the spam email problem in this paper is downloaded from the UCI Machine Learning Repository [18]. Data set of Spam base involving 4601 emails was proposed by Mark Hopkins, and his colleagues. This data set is divided into two parts. In this set, 1813 emails (39.4%) were determined as spam, and 2788 emails (60.6%) were determined as non-spam. 1 shows spam, and zero indicates non-spam. This data set involves 57 features with continuous values. The last feature shows class label involving discrete value. 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. 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:

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$F1 = \frac{2 \times \pi \times \rho}{(\pi + \rho)} \tag{7}$$

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \tag{8}$$

In above mentioned equation, TP_i shows the number of test samples that have been properly classified in c_i class. FP_i shows the number of test samples that have been incorrectly classified in c_i class.

TN_i demonstrates the number of test samples belonging to c_i class, and have been correctly classified in other classes.

FN_i shows the number of test samples belonging to c_i class, and have been incorrectly classified in other classes.

The methods of support vector machine and C4.5 have been used for classification. The executed program and the obtained average have been compared 10 times to investigate the performance of each classifier. In support vector machine, the best value has been considered for penalty coefficient of 3.5. the results obtained from the proposed method of feature selection have been compared with genetic algorithm. Also, this method has been compared without considering feature selection. The obtained results show that when the parameters are presented in tables 1 and 2, the best performance is observed in terms of GAFS and ICAFS.

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

Initial population	80
Mutation	0.05
Crossover	0.7
Generations	100

Table 2: the parameters of feature selection by using DICA

The number of country	50
number of imperialists	3
Revolution rate	0.1
ξ	0.01
The number of decades	200

6. RESULT EVALUATION

In this section, the results of experiments have been presented to evaluate the efficiency of proposed method and their comparison with other methods. After executing the proposed algorithm in the samples, the obtained results are collected in two tables. Table 3 shows the comparisons of GA FS and DICA FS algorithm and all features by using C4.5 and SVM classifiers in terms of Accuracy and the number of selected feature. In table 4, these algorithms have been compared in terms of recall, precision and F score of spam class. As it is observed in table 3, the Accuracy of proposed method, DICA FS, is more than the method of feature selection based on genetic algorithm and the method considered without feature selection. The number of selected features is equal. In addition, in comparing two classifiers of C4.5 algorithm, better results were presented in comparison to SVM.

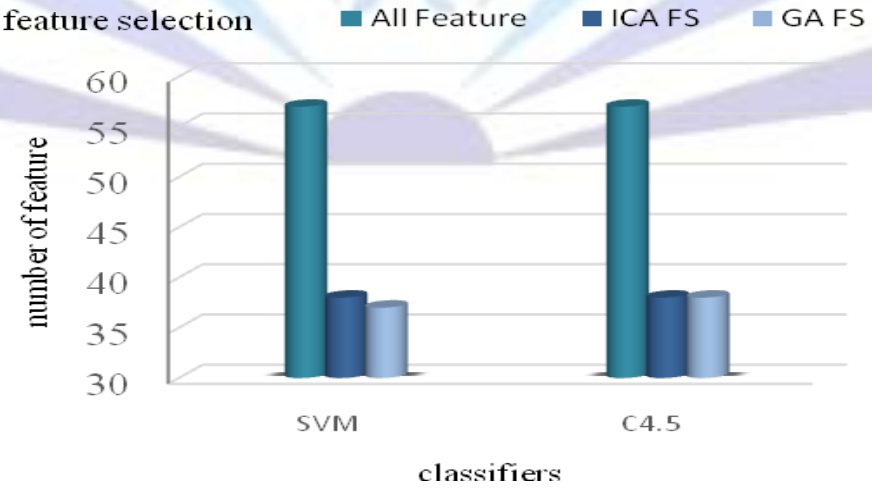

Figure 3: column graph of comparing the number of selected features

Table 3: comparing feature selection methods in terms of accuracy and the number of features

classifier / Algorithms	C4.5		SVM	
	Number of features	Accuracy	Number of features	Accuracy
All Feature	57	0.92	57	0.846
DICA FS	38	0.925	38	0.913
GA FS	38	0.918	37	0.863

Table 4: the classification results by using the methods of feature selection

Classifier / Algorithms	C4.5			SVM		
	F1 score	Recall	Precision	F1 score	Recall	Precision
All features	0.906	0.895	0.917	0.826	0.851	0.802
DICA FS	0.912	0.906	0.919	0.899	0.909	0.89
GA FS	0.905	0.896	0.911	0.841	0.864	0.839

7. CONCLUTION

In this paper, the method of feature selection has been presented on the basis of Discrete Imperialist Competitive Algorithm. The proposed method was evaluated by using data set of Spam Base. The results obtained from DICA were compared with genetic algorithm and position without feature selection. The obtained results show that the proposed method has accuracy comparable with other methods. In addition, other evaluation criteria have been considerably improved. The proposed algorithm can be combined with other classification algorithms in the future. Also, it can be used for classification in ensemble classifiers.

8. REFRENCES

[1] Han J, Kaber M, Pei J. Data Mining, Concepts and Techniques. 3rd edn, Morgan Kaufman, 2011.

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

[3] Awad, W. and S. ELseuofi, MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION. International Journal of Computer Science & Information Technology, 3(1), 2011.

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

[5] Hoanca, B., How good are our weapons in the spam wars? Technology and Society Magazine, IEEE, 25(1): p. 22-30, 2006.

[6] Atashpaz-Gargari, E. and C. Lucas. Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. in Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. 2007. IEEE.

[7] Mousavi Rad, S., F. Akhlaghian Tab, and K. Mollazade, Application of Imperialist Competitive Algorithm for Feature Selection: A Case Study on Bulk Rice Classification. International Journal of Computer Applications, 40, 2012.

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

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

[10] Howley, T., et al., The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge-Based Systems, 19(5): p. 363-370, 2006.

[11] Zhu, Z. An email classification model based on rough set and support vector machine. in Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on IEEE, 2008.

[12] Janecek, A., et al., On the Relationship Between Feature Selection and Classification Accuracy. Journal of Machine Learning Research-Proceedings Track,. 4: p. 90-105, 2008.

[13] Parimala, R. and R. Nallaswamy, A Study of Spam E-mail classification using Feature Selection package. Global Journal of Computer Science and Technology. 11(7), 2011.

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

[15] E. A. Gargari, F. Hashemzadeh, R. Rajabioun, and C. Lucas, "Colonial competitive algorithm: a novel approach for PID controller design in MIMO distillation column process," International Journal of Intelligent Computing and Cybernetics, vol. 1, pp. 337-355, 2008.

[16] Guyon, I. and A. Elisseeff, An introduction to variable and feature selection. The Journal of Machine Learning Research, 3: p. 1157-1182, 2003.

[17] Mousavirad, S. and H. Ebrahimpour-Komleh.Feature selection using modified imperialist competitive algorithm. in Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on. 2013. IEEE,2013.

[18] “UCI repository of Machine learning Databases”,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.,1998.