A Hybrid Feature Selection Method Based on Harmony Search

Yurou Zhang*,Jing Zhang*

* Chongqing Key Laboratory of network and information security technology, Chongqing University of Posts and Telecommunications,Chongqing 400065, P. R. China

Abstract: Spammer as the release of spam in social networks has seriously affected the other users' experience in social networking platform. Micro-blog has a huge amount of registered account, and a large number of users' behaviors have brought some difficulties for the detection of spammer. So the feature selection has become a primary problem of detecting spammer. This paper puts forward a hybrid feature selection method based on harmony search (HS)and ReliefF, so this method is called ReF-HS. Because HS is a heuristic algorithm, and ReliefF is a feature selection algorithm based on filter. This method considers not only the measure of single feature but also the correlation between the features. Compared with traditional method, this method converges quickly and is better to avoid the local optimal. Experimental results show that the feature subset selected by ReF-HS is small in size, and the selected feature subset leads to a higher accuracy rate of spam detection.

Key words: Harmony Search；ReliefF；Feature Selection

1.Introduction

Micro-blog as the representative of the service website has gradually becomes the main media which can generate information and diffuse the public opinions. Users through a variety of intelligent terminals post no more than 140 characters of the text in real time to share and discuss everything (including news, jokes, comments, and even their emotions, etc.) on micro-blog. Micro-blog brings people a variety of convenience, at the same time, it is also used by the attacker. The attackers create lots of false accounts and stole normal users' accounts to publish advertisements, pornography, phishing and other malicious information on social networking sites. These accounts are collectively called as spammer. Because there is trust relationship between users, the malicious information is more dangerous than the malicious information contained by traditional spam. These malicious behaviors cause a serious threat for the normal user's privacy information, account security and the experience of users. Malicious interaction[1], adding friends[2], like[3] and other acts for obtaining benefits are serious harmful to the reputation evaluation system of online social network and the users' trust relationship. In order to solve the security problem caused by spam, the research on the behavior of the account has become an important issue. However, micro-blog has a huge amount of registered accounts, the accounts have different forms of expression. And the performance of the account is a dynamic process, this produces a lot of behaviors. It has brought some difficulties for the detection of spammer, so the selection of the user's features has become a primary problem.

The main objective of feature selection (FS) is to determine a minimal feature subset which can represent the original data accurately. The detection spammer is essentially a problem of classification of two values. And the data without feature selection may be redundant or noisy, which may decrease the accuracy of classification. Typically, feature selection can be divided into three types: filter, wrapper and hybrid methods. The method based on filter is usually ranked or weighted, such as correlation measure, distance measure, information theory, etc. An obvious advantage of the filter is that it can quickly eliminate a large number of non critical noise features, and reduce the scale of the search of the optimized feature subset. The filter has high computational efficiency and good generality. But it does not guarantee that an optimal feature subset is selected. Especially when the feature and classifier are closely related, even if it can find an optimal subset which satisfies the condition, its scale will be quite large and contain some obvious noise features. The computing speed of wrapper is slower than the filter. But the scale of the optimized feature subset is relatively small, which is very advantageous to the identification of key features. At the same time, its accuracy is relatively high, but the generalization ability is poor. The hybrid method takes both the measurement of the single feature and the correlation between the features into account. In this paper, it focuses on the feature selection method based on the hybrid method.

A hybrid feature selection method based on harmony search is proposed in this paper. Harmony search (HS) [4] is a kind of meta heuristic algorithm, which imitates the music player's improvisation process. Harmony search algorithm has certain advantages compared to the traditional optimization technology, and has achieved a great success for a variety of optimization problems. ReliefF is a kind of feature selection algorithm based on filter, which is pre filter of features to get solution space. So this method is called ReF-HS. ReF-HS takes both the measurement of individual features and considers the correlation between features into account. Compared with the traditional methods, it is better to avoid falling into the local optimal. First of all, the ReliefF is used to select the features which have strong classification ability. This process can reduce the search space of HS and form the solution space that provides a range of options for musicians. And then the new harmony is obtained by HS. The fitness function is used to evaluate the new harmony, if achieves good scores, then it will replace the worst harmony of the harmony memory, instead, it will be discarded. The rest of this paper is as follows: the second part summarizes the current research work; third part gives a brief overview of the concept used in this paper; the fourth part is experiment which focuses on the data source, experimental setup and experimental results; the fifth part is the conclusion of the paper and propose future work.

2 Related work

In recent years, researchers have made a lot of research on the application of feature selection to Spam detection or classification. Rajalaxmi et al.[5]employed a binary bat approach to select the best features. This method performs feature selection based on the echolocation behavior of bats. They collected real data from Facebook profile of different users containing normal and spam profiles, and used JRip classification algorithm to classify the profiles. The results show that the binary bat approach is better than other methods. Trivedi et al.[6]compared between various supervised feature selection methods such as Document Frequency (DF), Chi-Squared, Information Gain (IG), Gain Ratio (GR), ReliefF (RF), and One R (OR). Then they used Bayesian Classifier to classify the spam for validation of the results. Results of this study show that RF is the excellent feature selection technique among other in terms of classification accuracy and false positive rate. Dhawan et al.[7]presented Bayesian Classifier with Correlation Based Feature Selection which can key out relevant features as well as redundancy among relevant features without pair wise correlation analysis. Zhang et al.[8]proposed a novel spam detection method which used the wrapper-based feature selection method to reduce the false positive error of mislabeling legitimate user as spammer. Result of this research shows clearly that the proposed method is effective. Behjat et al.[9]applied a genetic algorithm (GA) to decrease the number of useless features in a collection of high-dimensional email body and subject. They employed LingSpam benchmark corpora as the dataset, and used a Multi-Layer Perceptron (MLP) to classify features. The results showed that a GA feature selector with the MLP classifier does not only decrease the data dimensionality but increase the spam detection rate as compared against other classifiers. Cervante et al.[10]proposed two new filter feature selection methods for classification problems. The first method is based on binary particle swarm optimization (BPSO) and the mutual information of each pair of features, and the other method is based on BPSO and the entropy of each group of features. Experimental results show that the two algorithms can effectively reduce the number of features, in most cases they can get a higher classification accuracy. The first algorithm usually can obtain a smaller feature subset, and the second algorithm can obtain a higher classification accuracy. In previous studies, many techniques were used to carry out Spam detection. A hybrid feature selection method based on harmony search is proposed to obtain the feature subset of smaller scale and improve the accuracy of Spam detection. Chen et al.[11] extracted 12 lightweight features for tweet representation. The results showed the streaming spam tweet detection was still a big challenge and a robust detection technique should take into account the three aspects of data, feature, and model. Abulaish et al.[12] presented an ensemble learning method for online social network security by evaluating the performance of some basic ensemble classifiers over novel community-based social networking features of legitimate users and spammers in online social networks. The experimental resulted reveal that the identified features were highly discriminative to identify spammers in online social networks. Dang et al.[13] Due to the evolving nature and hidden behavior of user, found spammer who hijacked the trending topics with a deliberate point of view which would affect people’s judgments and decisions seriously. They had to deal with two important issues to solve the problem of detecting this type of spammer groups. The experimental results demonstrated that our similarity measure keeped a leading performance in all evaluation metrics. Zheng et al.[14] proposed a supervised machine learning based solution for an effective spammer detection. They collected a dataset from Sina Weibo and extracted a set of feature from message content and users' social behavior, and applied into SVM (Support Vector Machines) based spammer detection algorithm. The experiment shows that the proposed solution is capable to provide excellent performance with true positive rate of spammers and non-spammers reaching 99.1% and 99.9% respectively. Mccord et al.[15] discussed some user based and content-based features that were different between spammers and legitimate users. Then, they used these features to facilitate spam detection. Their results based on the 100 most recent tweets also showed that spam detection based on their suggested features could achieve 95.7% precision and 95.7% F-measure using the Random Forest classifier. Zhao et al.[16] characterized two type of spammers on micro blogging platforms, namely advertised spammers and following spammers, and then presented preliminary approaches to detect these spammers. Specially, they introduced a new feature named duplication for spammer detection. In experimenting they ran several classification methods on the characterized features to test the effectiveness.

3 Related concepts

3.1 ReliefF

ReliefF algorithm is based on the Relief algorithm proposed by Kononenko[17]. The ReliefF algorithm calculates the weight value of the feature by Euler distance. The value of the feature weight value indicates the ability of each feature to distinguish the different samples. The same with ReliefF algorithm, ReliefF algorithm is relatively simple and high efficiency. It can select the features who have a strong classify ability. In addition, the ReliefF algorithm overcomes the problem of noise data and data missing. Each weight in ReliefF algorithm updates is as follows:

$$W(A)=W(A)-\sum_{j=1}^{k}\frac{diff(A,R_i,H_j)}{mk}+\sum_{C\neq class(R_i)}\left[\frac{P(C)}{1-P(class(R_i))}\sum_{j=1}^{k}\frac{diff(A,R_i,M_j(C))}{mk}\right]$$

m represents the number of sampling. $M_j(C)$ represents the jth nearest neighbor samples from the different category C. $class(R_i)$ represents the R_i's category. $P(C)$ represents the proportion of the target sample which belongs to category C to the total number of samples. Function $diff(A,R_i,R_j)$ is used to calculate the distance between the sample instance R_i and R_j about a feature. Specific formula is as follows:

$$diff(A,R_i,R_j)=\begin{cases}\dfrac{|R_i(A)-R_j(A)|}{\max(A)-\min(A)}, & A\ \ is\ \ \text{continuous}\\ 0 & ,A\ \ is\ \ \text{discrete}\ \ and\ \ R_i(A)=R_j(A)\\ 1 & ,A\ \ is\ \ \text{discrete}\ \ and\ \ R_i(A)\neq R_j(A)\end{cases}$$

It can see that for the feature A the smaller the distance between the two samples of the same class is or the greater the distance between the two samples of different classes is, and the more favorable for classification the feature is. Specific ReliefF algorithm pseudo code as shown in Table 1.

Table 1 ReliefF algorithm pseudo code

```
Algorithm 1. ReliefF Algorithm
Input: training data set $D$, frequency in sampling $m$, the number of the nearest
        neighbors $k$
Output: feature weight vector $W$
Procedure begin
        set all feature weight to be 0
        for $i=1$ to $m$ do
            randomly select a sample $R$
              find $k$ nearest samples $H$ of $R$ from the same sample set
            for each class $C \neq class(R)$, find $k$ nearest samples $M$ from the
                different sample set
```

```
6.              for  $A=1$  to  $N$  do
7.              update each feature weight
                $W(A)=W(A)-\sum_{j=1}^{k}\frac{diff(A,R_i,H_j)}{mk}+\sum_{C\neq class(R_i)}[\frac{P(C)}{1-P(class(R_i))}\sum_{j=1}^{k}\frac{diff(A,R_i,M_j(C))}{mk}]$
8.           end for
9.       end for
Procedure end
```

3.2 Harmony search

Harmony search is a improvisation process of musicians. Music harmony is a kind of beautiful sound combination which comes from the aesthetic view. The music performance is to seek for an optimal state that is determined by the aesthetic evaluation. When applied to the optimization problem, the musicians represent the objective function of the decision variable, and HS is used as a heuristic algorithm to find the optimal state that is determined by the value of the objective function. Geem et al.[4] took the musician analogy to the decision variables of the optimization problem. The note that a musician plays is the value of every decision variable. Harmony memory consists of the harmony which is played by the musician, or is a storage space for the solution vector. In particular, the harmony memory is a two-dimensional matrix, in which the row vector representation harmony (the solution vector), and the number of rows represents the size of the harmony memory. Each column stores the harmony that is played by different musicians, that is, the harmony threshold of each musician. The following is the specific iterative steps and algorithm description:

(1) initialization of five variables

The parameters of the HS includes: Harmony memory Size(HMS), Harmony memory considering rate(HMCR), Pitch adjusting rate(PAR), tone tuning bandwidth bw, the maximum number of iterations.

(2) initialization of the harmony memory

Harmony memory is filled with randomly generated solution vectors x^1, x^2, $\cdots$, x^{HMS}, and the form of harmony memory is as follows:

$$HM=\left[\begin{array}{c|c} x^1 & f(x^1) \\ x^2 & f(x^2) \\ \vdots & \vdots \\ x^{HMS} & f(x^{HMS}) \end{array}\right]=\left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_N^1 & f(x^1) \\ x_1^2 & x_2^2 & \cdots & x_N^2 & f(x^2) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1^{HMS} & x_2^{HMS} & \cdots & x_N^{HMS} & f(x^{HMS}) \end{array}\right]$$

(3) produce a new harmony

In the HS algorithm, the influence of a new harmony generation of has three main factors: HMCR, PAR, random selection. The value range of HMCR is $[0,1]$. The first variable $x_1^{'}$ of the new solution has HMCR probability to be selected from any value in HM, and it has 1-HMCR probability to be selected from any value outside of the HM (and within the range of the variable). The way it produces is as follows:

$$x_i^{'}=\begin{cases} x_i^{'}\in(x_i^1,x_i^2,\cdots x_i^{HMS}), & if \quad rand<HMCR \\ x_i^{'}\in X_i, & otherwise; \quad i=1,2,\cdots,N \end{cases}$$

If the value of HMCR is low, musicians will continue to search for other regions of the solution space, while a higher value of HMCR will restrict the musicians.

If the new harmony $x_1^{'}$ comes from the HM, the PAR parameter is used to tune the tone, the specific way is as follows:

$$x_i^{'}=\begin{cases} x_i^{'}+rand*bw, & if \quad rand<PAR \quad \text{(Continuous)} \\ x_i^{'}(k+m), & m\in\{-1,1\},if \quad rand<PAR,\text{（Discrete）} \\ x_i^{'}, & otherwise, i=1,2,\cdots,N \end{cases}$$

bw is an arbitrary distance bandwidth, $rand$ is a random number in [0,1].

（4）Update harmony memory

Compared with the worst solution in the initial HM, if the new harmony has the better solution of fitness function, the new harmony replaces the worst harmony in the HM. Whether the current iteration number has reached the maximum number of iteration or not. If not, then repeat the process of 3 and 4 until the maximum number of iteration to reach.

3.3 ReF-HS algorithm specific description

According to the evaluation method of feature subset, it can be divided into two types: Filter and Wrapper. They each have their own advantages and disadvantages, and complement each other. Harmony search is one kind of feature selection based on Wrapper, and ReliefF is one kind of feature selection based on Filter. So here a feature selection algorithm based on ReliefF algorithm and harmony search algorithm is proposed. The specific process is as follows:

(1) Filtering: using the ReliefF algorithm to select the features which have strong classification ability to reduce the size of the search space. The purpose of using the ReliefF algorithm is that it removes the features which are weak for classifying or irrelevant, reduces the dimension of the original feature set initially and improves the efficiency of the algorithm.

(2) Initialization: initializing the five parameters of the harmony search: HMS, HMCR, PAR, bw, the maximum number of iterations. According to (1), the solution space of the harmony search algorithm is obtained by the ReliefF algorithm , which provides a range of choices for each selector. The solution space randomly generates a harmony memory.

(3) Creates a new subset: randomly generating a variable $rand_1$ which value is $[0,1]$ and comparing it with the HMCR above. If $rand_1$ is less than HMCR, then getting a set of harmony from the initialization of the harmony memory, otherwise getting a set of harmony in the solution space above. In the end, a set of harmony is gotten. If this set of harmony is obtained from the harmony memory, it needs to tune the set of harmony. A variable $rand_2$ whose value is chosen from $[0,1]$ is generated randomly generated. If $rand_2$ is less than PAR, it needs to use formula (5) to get a new set of harmony, otherwise, do not.

(4) Update harmony memory

In order to evaluate the new feature subset, the fitness function of this paper considers two factors: the classification accuracy and the dimension of the feature subset. So the fitness function is designed as follows:

$$f(X) = \delta Acc(X) + \varphi(1 - \frac{|X|}{N})$$

Among them, the dimension of the feature subset is represented by $|X|$. N is the dimension of the original feature set, δ and φ are the control parameters, which measures the proportion between the classification accuracy and the size of the feature subset. In this paper, it is considered that classification accuracy is more important, so $\delta = 0.9$, $\varphi = 0.1$. $Acc(X)$ is the classification accuracy, the average accuracy is obtained by using the Naive Bayes(NB) on the test set.

Compared with the worst solution in the initial HM, if the new harmony has the better solution of fitness function, the new harmony replaces the worst harmony in the HM. Whether the current iteration number has reached the maximum number of iteration or not. If not, then repeat the process of 3 and 4 until the maximum number of iteration to reach.

The flow chart of the ReF-HS algorithm is shown in the following figure.

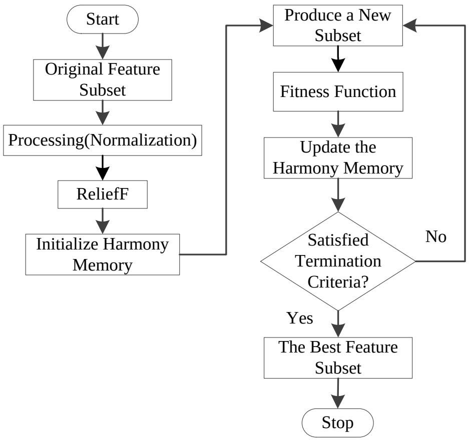

Start
Original Feature Subset
Processing(Normalization)
ReliefF
Initialize Harmony Memory
Produce a New Subset
Fitness Function
Update the Harmony Memory
Satisfied Termination Criteria?
No
Yes
The Best Feature Subset
Stop

Figure 1 the flow chart of the ReF-HS algorithm

3.4 Micro-blog Users' Feature Analysis

The study found that the related information of accounts in micro-blog platform mainly includes basic information, historical posts, friends, personal interest, etc.. In this paper, the features is extracted from the state of the accounts, the historical posts. The state of the account comes from basic account information. This reflects the basic state of the account which includes the number of followers, follow, the number of posts, account level, account age, etc.. The historical posts refers to the feature extracted from publishing or forwarding historical posts of the accounts, which can reflect personal habits and the quality of posts. This includes the frequency of posts, the proportion of the original posts, the proportion of the spam forwarding, the average number of posts forwarded and the number of comments, etc. In the end, the 17 features are selected as shown in Table 2.

Table 2 users' feature

Category	**Feature**	**Number**
The state of the accounts feature	The number of followings	f_1
	The number of followers	f_2
	The number of posts	f_3
	Reputation	f_4
	The number of friends	f_5
	VIP	f_6
	Authentication	f_7
The historical posts feature	The frequency of posts	f_8
	The average number of like	f_9
	The average number of comments	f_{10}
	The average length of posts	f_{11}
	The proportion of the original posts	f_{12}
	Content similarity	f_{13}
	The average number of @	f_{14}
	The average number of pictures	f_{15}
	The average number of URL	f_{16}
	The average number of topic	f_{17}

4 Experimental

4.1 Data Set Description

API provided by sina can only download the latest posts of authorized users, and these information for Spammer detection is far from enough. In order to solve this problem, the web crawler program is used to crawl the ID, the number of posts, the number of follower, nearly 50 posts, the number of forwarding, the comments and so on. This experiment randomly selects 4500 users. and the data set is marked. the marked rules are as follows: (1) to publish pornographic pictures or information, or publish content with URL pointing to pornographic sites. (2) release a large number of advertising information, or publish content with URL pointing to the shopping site. (3) released at least one harmful link or phishing network. In the end, according to the above rules there are 830 spammer to be found, and the number of normal users is 3670.

4.2 Experimental Methods

In order to validate the hybrid feature selection method based on harmony search, the method is compared with Harmony Search(HS), Genetic Algorithm (GA) and Particle Swarm optimization (PSO) . All methods will use the features of Table 2 for feature selection. The parameter setting for all methods is just like this: (1) ReF-HS: the number of sampling is 1400, the number of the nearest neighbor is 10, memory size is 100, max iteration is 200, HMCR is 0.9, PAR is 0.3、bw is 0.01; (2) HS: memory size is 100, max iteration is 200, HMCR is 0.9, musician is 10、PAR is 0.3、bw is 0.01; (3) PSO: c_1 is 0.8, c_2 is 1.2, max generation is 2000, particles is 100;(4)GA: crossover rate is 1.0, mutation rate is 0.1, max generation is 2000, population size is 60.

The proportion of normal users and spammers in the data set is $5:1$. Four feature selection algorithms are run on the data set, and the feature subset size and running time are recorded. The results of feature selection are combined with the learning algorithm classifier for training. Finally, the test is carried out on the test set and the results are recorded by micro-blog. At the same time, in order to obtain reliable results, the ten fold cross validation method is used to verify the classification performance. The learning algorithms used in the experiments are logical regression(Logistic Regression, LR), random forest(RF), decision tree(DT) C4.5 and support vector machine(SVM).

4.3 Experimental results and discussion

For the evaluation of a feature selection algorithm, it is usually from three aspects: (1) feature subset size. (2) operation efficiency of the algorithm. (3) the effect on classifier. In this paper, the algorithms are compared from these aspects.

(1) feature subset size

The result of feature selection by ReF-HS, HS, GA and PSO is shown as table 3. The minimum feature subset's scale obtained by ReF-HS algorithm is 7, The minimum feature subset's scale obtained by HS algorithm is 9, The minimum feature subset's scale obtained by PSO algorithm is 13, The minimum feature subset's scale obtained by GA is 9. So ReF-HS algorithm to obtain the smallest feature subset and the result of PSO dimensionality reduction is the worst.

Table 3 the result of feature selection by ReF-HS, HS, PSO and GA

Algorithm	the features
ReF-HS	$\{f_1, f_4, f_5, f_{12}, f_{13}, f_{14}, f_{15}\}$
HS	$\{f_1, f_4, f_5, f_9, f_{10}, f_{12}, f_{13}, f_{14}, f_{15}\}$
PSO	$\{f_1, f_3, f_4, f_5, f_6, f_7, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}\}$
GA	$\{f_1, f_3, f_4, f_5, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}\}$

(2) operation efficiency of the algorithm

In order to compare the computational time of the algorithm, all algorithms ran 10 times, and the computation time of ReF-HS, HS, PSO and GA are shown in Fig. 2.

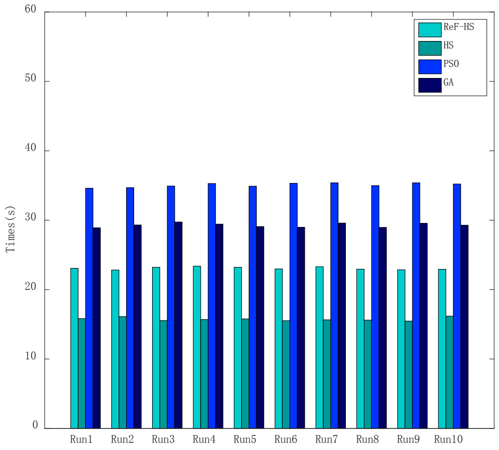
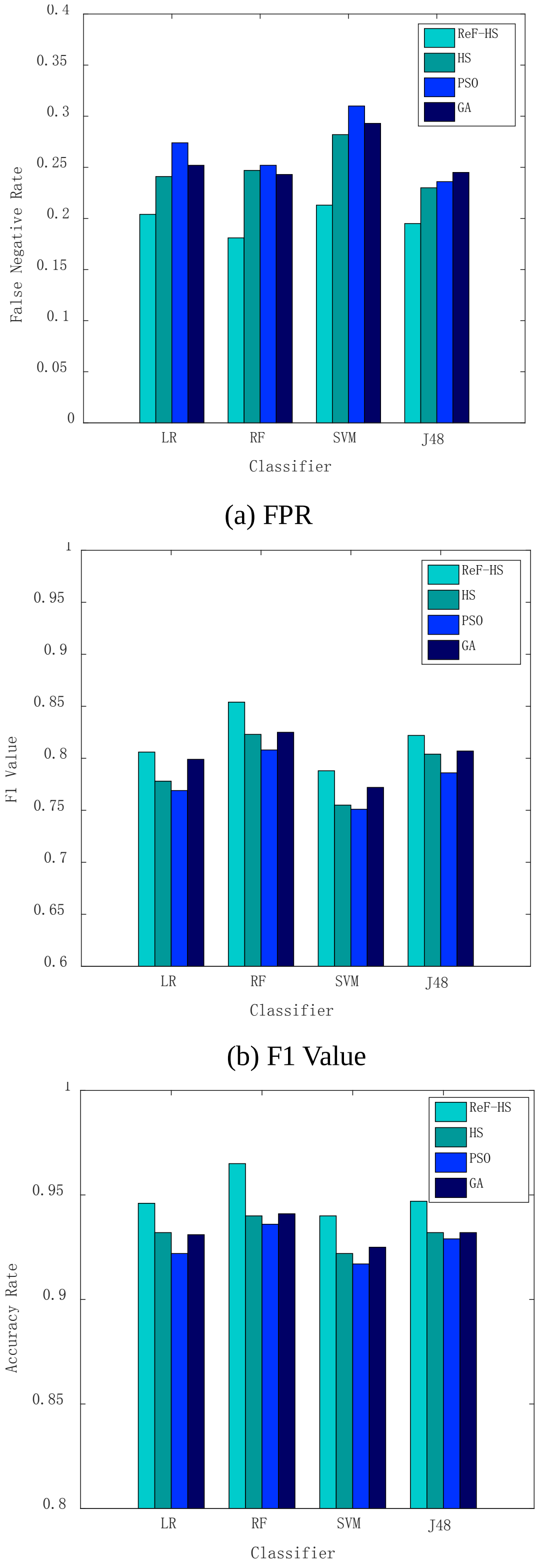


Figure 2 algorithms' run time comparison

The average time for ReF-HS, HS, PSO and GA are 23.07s, 15.73s, 35.07s and 29.28s. The running time of ReF-HS algorithm is obviously lower than that of PSO and GA. This is because although ReF-HS has one more step, but ReliefF regarded as the profiler runs fast and reduces searching space of the HS algorithm, while other algorithms are in search of the original feature space. The running time of ReF-HS algorithm is higher than that of HS algorithm, but in practical application, ReliefF algorithm can be run in advance. In this way, the running time of ReF-HS algorithm and HS algorithm are approximately .

(3) the effect on classifier

In this paper, it uses the accuracy rate(AR), false positive rate (FPR) and the comprehensive evaluation index (F1 value) to measure the classifier's ability to detect spammer. The experimental results are shown in figure 3.



(a) FPR

(b) F1 Value

(c) AR

Figure. 3 the performance of classifiers with four feature selection algorithms

From figure 3.6, it can see that, the spammer detection based on HS, based on PSO and based on user GA, after the feature selection, data still have redundant information, especially the reduction effect of PSO algorithm is poor and has the most redundant information. So that spammer detection accuracy rate and the F1 value are low, and the false positives rate is high. In contrast, for the ReF-HS algorithm proposed by this paper, since the introduction of the heuristic information of ReliefF, HS is conducive to quickly determine the optimal feature subset, and the obtained feature subset is small, when detecting spammers, because of the less redundant data set, the accuracy rate and the F1 value are high, and the false positives rate is low,

5 Conclusions

It can be seen from the experimental results that the ReF-HS algorithm performs well in the classification accuracy, the feature subset size and the algorithm running time. Compared with the HS, the classification accuracy and the feature subset size are all excellent, but the algorithm runs a little slower. The harmony memory solution space has an important influence on the performance of HS algorithm. For PSO and GA, the initial population also affects the performance of the algorithm, without pre filter it makes the search space large, the algorithm is not efficient, and easy to fall into local optimum. Thus affecting the performance of the algorithm. General ReF-HS algorithm can get higher classification accuracy and the calculation of efficiency is higher than the other search heuristic algorithm. For micro-blog with a large data, ReF-HS has high competitiveness.

The actual running time of the whole feature selection process is mainly decided by two main factors: the number of iterations and the evaluation method of the subset. So through the study of this paper, the total number of iterations is predefined, and an optimal subset may be found in the early search process. Building a better stop rule will improve the situation and save the running time of the algorithm. This will be the direction of research in the future.

Reference

[1] Stringhini G, Wang G, Egele M, et al. Follow the green: growth and dynamics in twitter follower markets//Proceedings of the 2013 conference on Internet measurement conference. Barcelona, Spain, 2013: 163-176.

[2] Xue J, Yang Z, Yang X, et al. VoteTrust: leveraging friend invitation graph to defend against social network sybils//Proceedings of the 32nd IEEE International Conference on Computer Communications. Turin, Italy, 2013: 2400-2408.

[3] De Cristofaro E, Friedman A, Jourjon G, et al. Paying for likes?: understanding facebook like fraud using honeypots//Proceedings of the 2014 Conference on Internet Measurement Conference. Vancouver, Canada, 2014: 129-136.

[4] 1 Geem ZW, Kim JH, Loganathan GV. A new heuristic optimization algorithm: harmony search. Simulation, 2001, 76(2):60−68.

[5] Rajalaxmi, R R. Binary Bat Approach for Effective Spam Classification in Online Social Networks. Australian Journal of Basic & Applied Sciences8.18(2014):383-388.

[6] Trivedi S K, Dey S. A Comparative Study of Various Supervised Feature Selection Methods for Spam Classification[C]//

International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.

[7] Dhawan S, Devi M. Spam Detection in Social Networks Using Correlation Based Feature Subset Selection. International Journal of Computer Applications Technology and Research Volume 4. 2015:629 - 632,

[8] Zhang Y, Wang S, Phillips P, et al. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 2014, 64:22-31.

[9] Behjat A R, Mustapha A, Nezamabadi-Pour H, et al. GA-based feature subset selection in a spam/non-spam detection system. International Conference on Computer and Communication Engineering. 2012:675-679.

[10] Cervante L, Xue B, Zhang M, et al. Binary particle swarm optimisation for feature selection: A filter based approach. 2012:1-8.

[11] Chen C, Zhang J, Xie Y, et al. A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection[J]. IEEE Transactions on Computational Social Systems, 2016, 3(1):42-42.

[12] Abulaish M, Bhat S Y. Classifier Ensembles Using Structural Features For Spammer Detection In Online Social Networks[J]. Foundations of Computing & Decision Sciences, 2015, 40(2):89-105.

[13] Dang Q, Zhou Y, Gao F, et al. Detecting cooperative and organized spammer groups in micro-blogging community[J]. Data Mining & Knowledge Discovery, 2016:1-33.

[14] Zheng X, Zeng Z, Chen Z, et al. Detecting spammers on social networks[J]. Neurocomputing, 2015, 42(1):27-34.

[15] Mccord M, Chuah M. Spam Detection on Twitter Using Traditional Classifiers.[J]. 2011, 6906:175-186.

[16] Jie Zhao, Yan Liu, Shuhan Liu. Towards Spammer Detection in Microblogging platforms.[J].2016,Vol.9, No.10, pp:239-250.

[17] Kira K, Rendell L. A. The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of Ninth National Conference on Artificial Intelligence, 1992. 129-134.