IJICTR International Journal of Information & Communication Technology Research
ITRC
Volume 7- Number 4- Autumn 2015 (27-33)

A Clustering Based Feature Selection Approach to Detect Spam in Social Networks

Mohammad Karim Sohrabi
Department of Computer Engineering,
Semnan Branch, Islamic Azad University,
Semnan, Iran
Amir_sohraby@aut.ac.ir

Firoozeh Karimi
Department of Computer,
Semnan Branch, Islamic Azad University,
Semnan, Iran
firoozeh_karimi@yahoo.com

Received: April 24, 2015- Accepted: September 16, 2015

***Abstract*—In recent years, online social networks (OSNs) have been expanded with a lot of facilities and many users and enthusiasts have joined to OSNs. On the other hand, the proportion of low-value content such as spam is rapidly growing and releasing in the OSNs. Sometimes the spam advertising purposes, commercial purposes or spreading lies in the different mailing lists are placed and shipped in bulk to send for social network users. Spams not only damage the interests of users, usage time and bandwidth, but also are a threat to productivity, reliability and security of the network. In this paper, we present an online spam filtering system that can be deployed as a component of the OSN platform to inspect message generated by users in real time. Our filtering method is working on the basis of different features such as like, replay, hash tag, followers, and the existing URLs in the posts of Facebook social network. We employ three clustering algorithms for this purpose and we also use naïve Bayes and decision tree to detect spam from non-spam. We evaluate the system using 2000 wall posts collected from Facebook.**

Keywords-spam; spam detection; social networks; feature selection; clustering;

I. INTRODUCTION

Today, with the emergence of multiple social networks in the virtual world of fierce competition with each other, working with social networks to interact with each other is the way favored by the users, and a large number of opinion reviews are posted on the web [15]. Also, many users spend a lot of time on social networks such as Facebook, Twitter, and MySpace to share a significant amount of personal data. This information is shared, in addition to the connection between thousands of users in the world, favored by spammers as well. For example, spammers may be attracted to exploit their benefits users and also distort the relationship between them, lead them to malicious websites or even personal information to steal the identity of others [17]. So how to recognize and prevent the spread of spam mechanism which has increased dramatically in online social networks (OSN) are a very important issue is that if you do not pay attention to it due to the return of the network by the user [14].

Our work focuses on detection of spammers over one of the most popular OSN platforms, Facebook.

Facebook is now the largest social network in the world and of every 7 people 1 person is a Facebook member. Founded in February 2004, in May 2011 the number of Facebook users has reached more than 700 million users, about 70 percent of users are outside the united states [1].

Being one of the most prominent OSNs, Facebook is continuously under attack by spammers [2].

For this work, we first obtained the Facebook dataset from its members' profile information and thereafter we have performed pre-processing over it to obtain normalized set of features based on which the activities of spammers were studied. The key features which are extracted from the dataset are 'wall post like', 'shared wall post', 'URLs', 'comment like', 'replies', 'the size of a message', 'the duration', and 'hash tags'. After obtaining these features, we use feature selection with particle swarm optimization (PSO) to select the best features then using automatic clustering with differential evolution (DE) algorithm to detect Spam in the dataset.

II. Related works

The loss of privacy is a threat to social network users. In 2010, researchers found that the personal information of more than 100 million users of Facebook is accessible through search engines [3]. Users were facing with different threats such as spam and malware. Huber et al. in their study, the probability of an attack to Facebook, have proven and shown that with a little time and simple hardware resources, a large number of spams can be easily published on this network [4]. Abu-nimeh et al. operate a large-scale investigation in relation to malicious email and spam on Facebook [3]. The results of this study show that about 9 % of posts on Facebook are spam, and in about 3% of posts, the link is malicious. Leung et al. designed a system which has blocked spam based on credit obtained from the user's social relationships [10]. Wang has developed a system to detect spam messages on Twitter [13]. Relational followers and friends in the network have been studied in this research using the social graph. In this system, the policy of spam on Twitter is taken using the system to detect spam based on message content and graph-based aid.

III. Proposed approach

The preliminary step for the detection of spammers in any OSN is data collection and necessary preprocessing dataset to convert it into a form which can be used by the learning algorithms.

A. Dataset description

To develop a dataset for training and testing of classification systems, we have manually identified a set of spam and non-spam comment from Facebook wall posts. This dataset contains 2000 comments between December 2014 and October 2015. We need labeled spam and legitimate comments for training and evaluating. Since more than 80 percent of spams were containing malicious links, at last we have 600 comments labeled as spam.

B. Feature Identification

The Facebook wall of a user is a place where her/his friends or other Facebook users can interact by posting messages and useful links. Users can also like and comment on the wall posts. According to Facebook statistics published in September 2011, about 2 billion wall posts on Facebook are liked or commented in a single day [18]. Since spam or non-spam messages behavior is different, various features which we have used to detect spam accounts include:

Hash tags (Wall posts and Comment): A hash tag is a type of label or metadata tag used on social network and micro blogging services which make it easier for users to find messages with a specific theme or content. Users create and use hash tags by placing the hash character # in front of a word or unspaced phrase, either in the main text of a message or at the end. Searching for that hash tag will then present each message that has been tagged with it. [6]

The hash tags for the spammer a lot more attention from the users and allows more visibility in their comments or Wall Posts.

Replies: Spammers replies to a large number of wall posts in order to get noticed by many users. This pattern can be used in the detection of spam.

Comment: This, similar to 'like', is quite self-explanatory. The 'comment' function allows you to post a comment on things the same as you would by 'like' it. Again, similarly to 'liking', comments made are as public as the place you're posting them to – not private (and, after all, this is the internet, so don't post things you wouldn't be happy with the entire world seeing).

Commenting on things is a great way to engage with people and businesses. How much of the comment of a post more that post has a comment on spam is more.

Spam Words: An account with spam words in almost every wall posts can be considered to be a spam account.

Likes (Wall posts and Comments): The 'like' button is a feature of facebook social networking service, which users can use to like contents such as status updates, comments, photos, links shared by friends, and advertisements. This feature may appear differently on mobile web applications. A "Like Box" also allows Facebook page owners to see how many users and which of their friends like the page. Likes posts and comments spammers are much lower than normal users.

URLs: URLs are the links which direct to some other page on the browser. With the development of URL shorteners, it has now become easy to post malicious links on any OSN. This is because URL shorteners hide the source of the link, thereby making it difficult for the detection algorithms [6]. More than 80 percent were spam containing malicious links.

Share: Share means that users are sharing this photo, video, note, etc. with everyone that are friends with or with a "custom" group. The high number of sharing a wall post represents the important of being more and more hits from the post. Therefore, this number is above the share can be an important factor for a large number of spam for the post.

Average time interval: Known as the "bursty" property, most spam campaigns involve coordinated

action by many accounts within short periods of time [19]. The effect is that messages from the same campaign are densely populated in the time period when the campaign in active. Consequently, if we compute the intervals between the generation times of consecutive messages in each cluster, the spam clusters are expected have shorter intervals than the legitimate clusters [5].

Followers: Followers are the users who follow a particular user.

Friends: Friends are the users who are friend with other users.

The size of a comment: Comment size is a popularly used feature, because spam comments have variable sizes, a group of them only contains URL links and a group of them who are advertising have a big size and remainder contains comments with URLs. Comments that have more than 5 lines and were also contains URLs are as spam in these data sets.

C. Feature selection

Feature selection plays an important role in classification, since it can shorten the learning time, simplify the learning classifiers, and improve the classification performance. Since there may be complex interaction among features, it is generally difficult to find the best feature subset [20]. In order to solve this problem, various methods have been proposed.

Generally, these methods can be classified into two categories: wrapper and filter approaches [21].

A filter approach relies primarily on general characteristics of a data set to evaluate and select feature subsets without considering a special learning approach [22]. A wrapper approach employs a classification algorithm to evaluate feature subsets, and adopts a strategy to seek for optimal subsets. Since the wrapper approach considers a classifier within the search process, this approach gets often better result than the filter one [23].However, this approach still suffers from a variety of problems, such as local convergence. A meta-heuristic is a high-level problem-independent algorithmic framework that provides a set of strategies to develop heuristic algorithm [24, 25]. Recently many researchers attempt to use meta-heuristic technique to tackle the feature selection problem, such as genetic algorithms [26], ant colony optimization, simulated annealing [28].

D. Feature Selection Algorithm

Particle swarm optimization (PSO) is a relatively recent meta-heuristic algorithm, which is inspired from the behavior of bird flocks. Due to its advantages in simplicity and fast convergence, it has been used in the feature selection problem and many other complicated problems [20].

In this paper, particle swarm optimization (PSO) meta-heuristic algorithms for feature selection we have used up the 13 listed properties for total data number of features that we find the greatest impact on their clustering to identify the spam. This is a simulation of a multi-objective optimization.

Normalizing the dataset using Multi-PSO algorithm (Figure 1), is in fact, the implementation of a single objective algorithm for solving a problem of multi-objective simulation. How does it work this way, the algorithm begins by selecting a feature of his work, and at each step by adding a number of features, the error rate is calculated and then displayed.

```
data = LoadData();
nx = data.nx;
BestSol = cell(nx,1);
S = cell(nx,1);
BestCost = zeros(nx,1);
for   nf = 1:nx
begin
   disp(['Selecting ' num2str(nf) ' feature(s) ...']);
   results = RunPSO(data,nf);
   disp(' ');
   BestSol{nf} = results.BestSol;
   S{nf} = BestSol{nf}.Out.S;
   BestCost(nf) = BestSol{nf}.Cost;
 end
```

Fig.1. The Multi-PSO Algorithm Method

In Figure 2, we see that all the answers are not satisfactory, observed the error rate four features of the features less than 5 features, In fact adding feature fifth made worse results. Thus, it has five characteristics point must be removed from the final answer. If we have less memory and more time, we must be choose the lesser number of features but if the aimed at reducing the error rate certainly will not be this way.

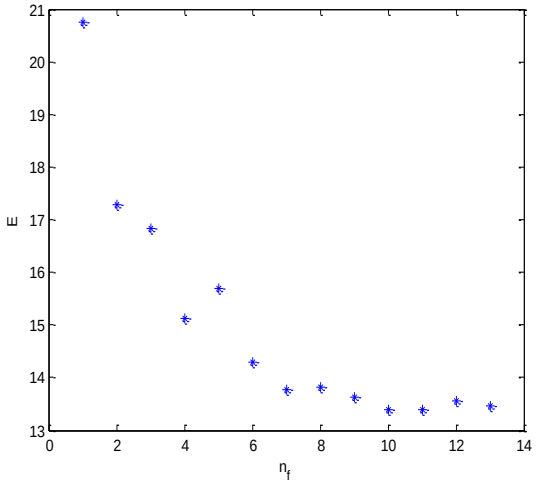

Fig.2. Feature selection error rate

Also in this place we can see this episode that in some cases the increase the number of features to witness the results worse, like sample error rate by five features compared with the error rate by four features. We can the sixth feature then the answer as namghlob and the rest of the answer are recessive and can pass them. But whichever we choose is sure optimum operation and in fact is a kind of multi-objective optimization has to be simulated. According to the results obtained from the algorithm is expressed and having less error rate and having the proper number of features that can be simultaneously both haft less and

have more time also from between the raised Thirteen features , seven features selection to continue.

After determining the appropriate number of features we can evaluate the r rate for these selected features. For this purpose, we investigate the process with both continuous and discrete coding techniques and their associated algorithms, such as particle swarm optimization (PSO), differential evolution (DE), ant colony algorithm (ACO), and simulated annealing (SA).

Figure 3 shows the result of running PSO algorithm with seven features for different number of replications. After 100 replications, the error rate is 13/59 using a random key with selecting seven features, including 13, 6, 7, 3, 8, 4, and 9, respectively.

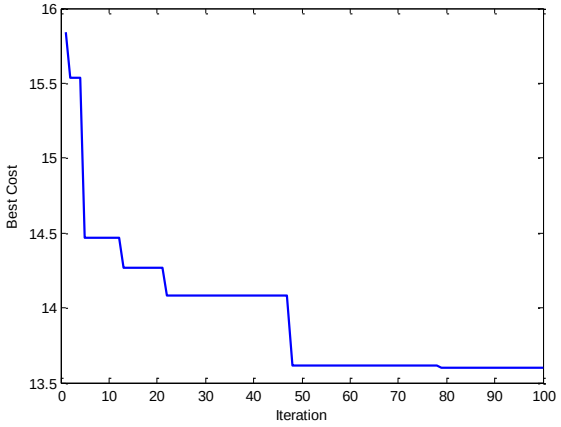

Fig..3. The best result of PSO algorithm

Figure 4 shows the result of running DE algorithm with seven features for different number of replications. After 100 replications, the error rate is 13/99 using a random key with selecting seven features, including 11, 1, 6, 5, 10, 13, and 9, respectively.

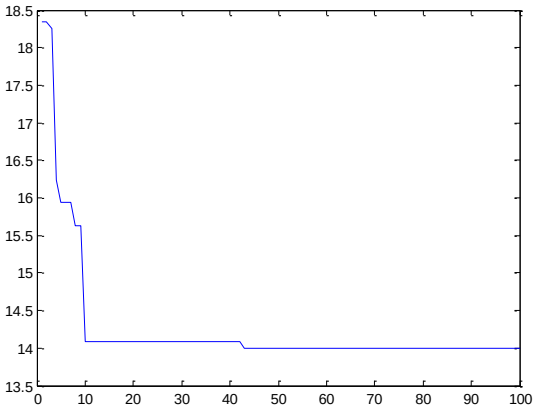

Fig..4. The best result of DE algorithm

In figure 5, the result of ACO algorithm with seven features for different number of replications is shown. After 100 replications, the error rate is 13/78 with selecting thirteen features including 4, 2, 6, 13, 11, 3, 1, 8, 12, 10, 5, 9, and 7, respectively.

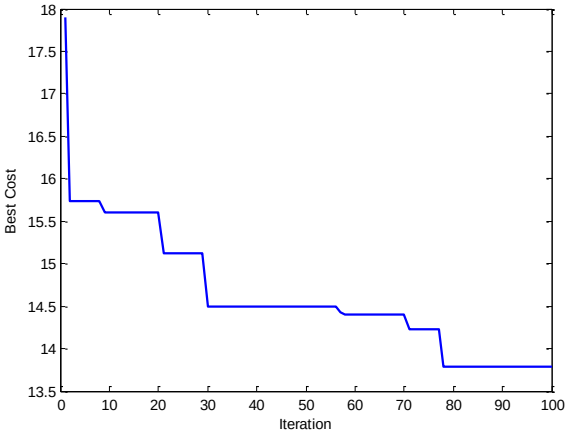

Fig..5. ACO algorithm Best Cost

Figure 6 shows the result of running SA algorithm with seven features for different number of replications. After 100 replications, the error rate is 13/64 with selecting thirteen features, including 13, 2, 7, 9, 8, 3, 6, 1, 10, 5, 12, 11, and 4, respectively.

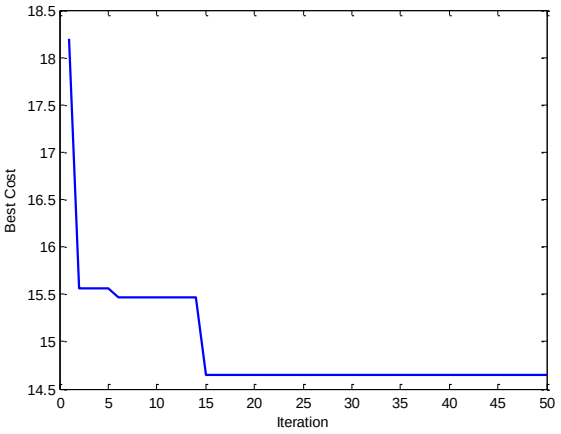

Fig..6. SA algorithm Best Cost

Given that the ACO and SA algorithms are continuous coding on their disposal and not based on random keys every thirteen feature based on more effective in order from left to right as the show output.

Therefore, based on the results of the error rate, the characteristics chosen by the PSO algorithm with an error rate of less use.

These features include: size of a comment, wall-posts likes, comments likes, replies, URLs, comments and shares.

In [5], some features like cluster size, average time interval, words and the average number of words per message, average URL number per message and unique URL number have been intended. According to the survey carried out in the received comments on the Facebook wall posts, we can conclude that these features are not applicable to achieve the correct percentage of spam detection. For example, the size of spam messages in many cases is short and this feature will be very useful in the case of spam emails.

Other features such as location-based features and density IP network transmitters and transmitter service

port status are also suggested. In practice, this feature because it requires large facilities and the unavailability of some of these cases is not possible. About average time interval can be seen by examining comment spammer in the various time intervals to spread their spam messages as spam email attachment to a lack of time. With all these interpretations of these features with particle swarm algorithm has been tested with these features also examine the error rate .

Error rate based on the characteristics listed in Figure 7 is displayed on the error rate is achieved 16/83. Therefore, in determining the effective features to achieve optimum error rate 13/59.

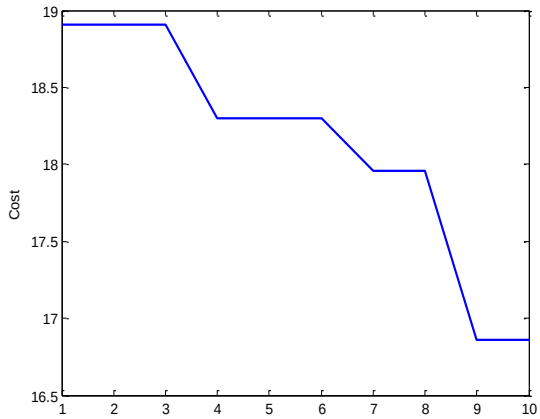

Fig.7. PSO algorithm Best Cost

After optimization and selection of convenient features, the data are prepared for clustering. In the next section we will discuss the clustering and the results thereof.

E. Clustering

Clustering is one of the crucial unsupervised learning techniques for dealing with massive amounts of heterogeneous information. The aim of clustering is to group a set of data objects into a set of meaningful sub-classes, called clusters which could be disjoint or not [7].

Clustering is a fundamental tool in exploratory data analysis with practical importance in a wide variety of applications such as data mining, machine learning, pattern recognition, statistical data analysis, data compression, and vector quantization [8].

Data clustering algorithms can be hierarchical or partitioned [10]. Within each of the types, there exists a wealth of subtypes and different algorithms for finding the clusters. In hierarchical clustering, the output is a tree showing a sequence of clustering, with each cluster being a partition of the data set [9]. Hierarchical algorithms can be agglomerative or divisive. Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Hierarchical algorithms have two basic advantages [12].

First, the number of classes need not be specified a priori, and second, they are independent of the initial conditions. However, the main drawback of hierarchical clustering techniques is that they are static; that is, data points assigned to a cluster cannot move to another cluster. In addition to that, they may fail to separate overlapping clusters due to lack of information about the global shape or size of the clusters. Partitioned clustering algorithms, on the other hand, attempt to decompose the data set directly into a set of disjoint clusters. They try to optimize certain criteria. The criterion function may emphasize the local structure of the data, such as by assigning clusters to peaks in the probability density function, or the global structure. Typically, the global criteria involve minimizing some measure of dissimilarity in the samples within each cluster while maximizing the dissimilarity of different clusters. The advantages of the hierarchical algorithms are the disadvantages of the partitioned algorithms, and vice versa. An extensive survey of various clustering techniques can be found in [16].

F. Clustering Validity Indexes

Cluster validity indexes correspond to the statistical mathematical functions used to evaluate the results of a clustering algorithm on a quantitative basis. Generally, a cluster validity index serves two purposes. First, it can be used to determine the number of clusters, and second, it finds out the corresponding best partition. One traditional approach for determining the optimum number of classes is to repeatedly run the algorithm with a different number of classes as input and then to select the partitioning of the data resulting in the best validity measure [11].

Ideally, a validity index should take care of the two aspects of partitioning.

- Cohesion: The patterns in one cluster should be as similar to each other as possible. The fitness variance of the patterns in a cluster is an indication of the cluster's cohesion or compactness.
- Separation: Clusters should be well separated. The distance among the cluster centers (may be their Euclidean distance) gives an indication of cluster separation [9].

DB Index: This measure is a function of the ratio of the sum of within cluster scatter to between cluster separation, and it uses the clusters and their sample means [9]. DB index is defined as fraction of distance within clusters to the distance between the clusters.

As powerful and fast method of differential evolution (DE) algorithm for solving optimization problems is presented in continuous space search algorithm is one of the newest methods. One advantage of this algorithm is a memory that appropriate information solutions in the current population keeps. The advantage of this algorithm is the selection operator. In this algorithm, all questions have an equal chance to be selected as one of the parents.

In contrast to most of the existing clustering techniques, the proposed algorithm requires no prior knowledge of the data to be classified. Rather, it determines the optimal number of partitions of the data "on the run." Superiority of the new method is demonstrated by comparing it with two recently developed partitioned clustering techniques and one popular hierarchical clustering algorithm [9].

IV. EXPRIMENTS AND RESULTS

In this section, we represent the result of using data clustering technique to spam detection according to the selected features of the index DB code (figure 8) and DE algorithm.

```
function [DB, out] = DBIndex(m, X)
    k = size(m,1);
    % Calculate Distance Matrix
    d = pdist2(X, m);
    % Assign Clusters and Find
      Closest Distances
    [dmin, ind] = min(d, [], 2);
    q=2;
    S=zeros(k,1);
    for i=1:k
     if sum(ind==i)>0
      S(i)=
     (mean(dmin(ind==i).^q))^(1/q);
     else
      S(i)=10*norm(max(X)-min(X));
     end
    end
    t=2;
    D=pdist2(m,m,'minkowski',t);
    r = zeros(k);
    for i=1:k
     for j=i+1:k
      r(i,j) = (S(i)+S(j))/D(i,j);
      r(j,i) = r(i,j);
     end
    end
    R=max(r);
    DB = mean(R);
    out.d=d;
    out.dmin=dmin;
    out.ind=ind;
    out.DB=DB;
    out.S=S;
    out.D=D;
    out.r=r;
    out.R=R;
end
```

Fig. 8. The DB index code

Using the DE algorithm and DB index, the test data are clustered into two 'spam' and 'non-spam' clusters with error rates of 0.02, which is shown in Figure 9.

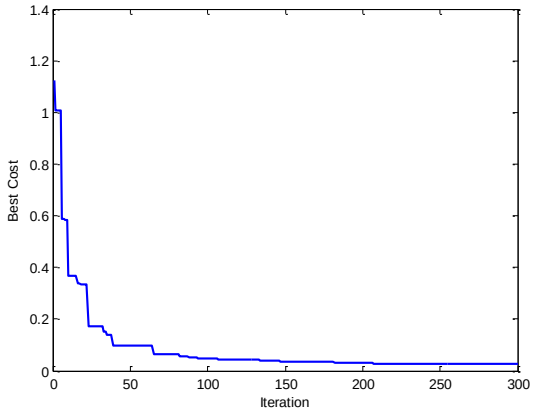

Fig. 9. DE Algorithm Best Cost

The output of clustering with implemented heuristics algorithm is depicted in figure 10.

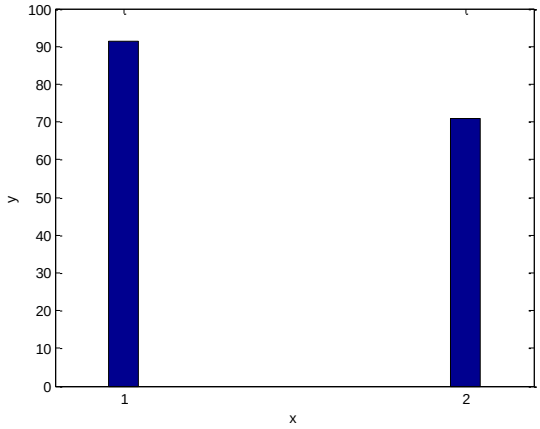

Fig. 10. Percentage of spam detection

Figure 10 shows that the proposed method can detect 71/8 % spam and 91/2% non-spam correctly.

V. CONCLUSION

In this paper, feature selection and simulation of Multi-PSO algorithm were used to solve a multi-objective problem with the number of features to achieve lower error rate in detecting spam and non spam comments. The proposed method specified the number of features and used the PSO algorithm to extract particles. Furthermore, we used a novel DE algorithm with index DB for clustering and combined it with our PSO-based method of spam detection to attain the acceptable results of 71/8 percent in spam detection rate.

As future work, using more features and a number of other indices, such as CS index, can be considered and other optimization algorithms such as GA, SA, and ABC can be used to improve clustering efficiency for spam detection.

REFERENCES

[1] https://wikipedia.org

[2] Zeus botnet targets facebook http://blog.appriver.com/2009/10 zeus-botnet-targets-facebook.html

[3] S. Abu-Nimeh, T. M-Chen and O. Alzubi, "Malicious and Spam Posts in Online Social Networks", IEEE Computer Society,Vol. 44, pp 23-28, 2011.

[4] M. Huber, M. Mulazzani, G. Kitzler, S. Goluch, and E. Weippl, "Friend-in-the-middle Attacks", Exploiting Social Networking Sites for Spam Internet computing, pp 28-34, 2011.

[5] H. Gao, Y. Chen, and K. Lee, "Towards Online Spam Filtering in Social Networks" Evanston, IL, USA, 2014.

[6] A. Gupta, R. Kaushal, "Improving Spam Detection in Online Social Networks", Indira Gandhi Delhi Technical University for Woman, Delhi, 2015.

[7] R. Forsati, A. Keikha, and M. Shamsfard, "An Improved Bee Colony Optimization Algorithm with an Application to Document Clustring", Neurocomputing, 159-9-26,2015.

[8] X. Xiaohua, L. Lin and H. Ping and P. Zhoujin and C. Ling, "Improving Constrained Clustering via Swarm Intelligence" ,Neurocomputing, 116 - 317–325 ,2013.

[9] S. Das, A. Abraham and S. Member and A. Konar, "Automatic Clustring Using an Improved Differential Evolution Algorithm", IEEE Trans on Systems,Man,and Cybernetics-Part :Systems and Humans,Vol.38,No.1,January 2008.

[10] Y. Leung, J. Zhang, and Z. Xu, "Clustering by Scale-Space Filtering", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22, No. 12, pp. 1396–1410, Dec. 2000.

[11] M. Halkidi ,M. Vazirgiannis, "Clustering Validity Assessment: Finding the Optimal Partitioning of a Dataset" , in Proc. IEEE ICDM, San Jose, CA, pp. 187–194,2001.

[12] H. Frigui and R. Krishnapuram, "A Robust Competitive Clustering Algorithm with Applications in Computer Vision", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 21, no. 5, pp. 450–465, May 1999.

[13] W. .A. Award and S. M. Elseuofi, "Machine Learning Methods for Spam E-mail Classification" ,International Journal of Computer Science & Information Technology LJCA(LJCSTT), Vol 16, No1, pp. 39-45, 2011.

[14] X. Zheng, Z. Zeng and Z. Chen and Y. Yu and Ch.Rong, "Detecting Spammers on Social Networks", Neurocomputing 159- 27-34, 2015.

[15] D. H. Fusilier, M. M Gomez, P. Rosso and R. G. Cabrera, "Detecting Positive and Negative Deceptive Opinion Using PU-Learning", Information Processing and Management 51-433-443, 2015.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A Review", ACM Comput. Surv., Vol. 31, No. 3, pp. 264–323, 1999.

[17] A. Heydari, M.A Tavakoli and N. Salim and Z. Heydari, "Detection of Review Spam:A Survey",Computer 42-3634-3642,2015.

[18] F.Ahmad , M.Abulaish, "A Generic Statistical Approach for Spam Detection in Online Social Networks" ,Computer Communications,36(10-11),Elsevier,pp,1120-1129,2013.

[19] X.Yu, F.A. chan and K. Panigrahy and R. Hulten and G. Andosipkov, "Spamming Botnets:Signatures and Characteristics" , In proc. of SIGCOMM, 2008.

[20] Z. Yong, G. Wei and Z. Wan-qiu, "Feature Selection of Unreliable Data Using An Improved Multi-Objective PSO aAlgorithm", Neurocomputing 171-1281-1290,2016.

[21] H. W. Roberto, D. C. George and F. C. Renato, "A Global-Ranking Local Feature Selection Method for Text Categorization", Expert Systems with Applications 39 (17) 12851–12857, 2012.

[22] C.T. Su, H.C. Lin, "Applying Electromagnetism-like Mechanism for Feature Selection" , Inf. Sci. 181 (5) 972–986 ,2011.

[23] M. A. Esseghir, G. Goncalves and Y. Slimani, "Adaptive Particle Swarm Optimizer for Feature Selection", in: Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning, LNCS 6283, pp. 226–233,2011.

[24] J. Sun, J. M. Garibaldi and N. Krasnogor, and Q. Zhang, "An Intelligent Multi-restart Mematic Algorithm for Box Constrained Global Optimisation", Evolu. Compu. 21(1)107-147,2012.

[25] J. Sun, Q. Zhang and X. Yao, "Meta-heuristic Combining Prior Online and Offline Information for the Quadratic Assignment Problem",IEEE Transaction on Cybern. 44 pp. 3429-444, 2014.

[26] I. S. Oh, J. S. Lee and B. R. Moon, "Hybrid Genetic Algorithms for Feature Selection", IEEE Trans. Pattern Anal. Mach. Intell. 26 (1) 1424–1437, 2004.

[27] C. L. Huang, "ACO-based Hybrid Classification System with Feature Subset Selection and Model Parameters Optimization", Neurocomputing 73 (1–3) 438–448, 2009.

[28] S. W. Lin, Z. J. Lee, S. C. Chen, and T. Y. Tseng, "Parameter Determination of Support Vector Machine and Feature Selection Using Simulated Annealing Approach", Applied Soft Computing 8 (4) 1505–1512, 2008.

Mohammad Karim Sohrabi received Ph.D. degree in Software Engineering from Computer Eng. Faculty, Amirkabir University of Technology (Polytechnic of Tehran), Tehran, Iran, in 2012. He is an assistant professor of Engineering Faculty, Semnan Branch, Islamic Azad University, Semnan, Iran. He has more than 40 journal and conference papers within the scope of his research area.

Firoozeh Karimi is a M.Sc. student of Software Engineering at Semnan branch, Islamic Azad University, Semnan, Iran. She received the B.S. degree in Software engineering from Computer Engineering Faculty, Tehran-North Branch, Islamic Azad University, Tehran, Iran, in 2012. Her research interests include software engineering, data mining, text mining and recommender systems.

This Page intentionally left blank.