A Parameter-Free Hybrid Clustering algorithm used for Malware Categorization

ZhiXue Han
Department of Computer Science
Xiamen University
Xiamen, China
Jessehan@yahoo.cn

Shaorong Feng*, Yanfang Ye
Department of Computer Science
Xiamen University
Xiamen, China
yeyanfang@yahoo.com.cn ,
shaorong@xmu.edu.cn

Qingshan Jiang
Software School
Xiamen University
Xiamen, China
qjiang@xmu.edu.cn

***Abstract*—Nowadays, numerous attacks made by the malware, such as viruses, backdoors, spyware, trojans and worms, have presented a major security threat to computer users. The most significant line of defense against malware is anti-virus products which detects, removes, and characterizes these threats. The ability of these AV products to successfully characterize these threats greatly depends on the method for categorizing these profiles of malware into groups. Therefore, clustering malware into different families is one of the computer security topics that are of great interest. In this paper, resting on the analysis of the extracted instruction of malware samples, we propose a novel parameter-free hybrid clustering algorithm (PFHC) which combines the merits of hierarchical clustering and K-means algorithms for malware clustering. It can not only generate stable initial division, but also give the best K. PFHC first utilizes agglomerative hierarchical clustering algorithm as the frame, starting with N singleton clusters, each of which exactly includes one sample, then reuses the centroids of upper level in every level and merges the two nearest clusters, finally adopts K-means algorithm for iteration to achieve an approximate global optimal division. PFHC evaluates clustering validity of each iteration procedure and generates the best K by comparing the values. The promising studies on real daily data collection illustrate that, compared with popular existing K-means and hierarchical clustering approaches, our proposed PFHC algorithm always generates much higher quality clusters and it can be well used for malware categorization.**

Keywords-Malware categorization; Parameter-Free Hybrid Clustering (PFHC); K-means; Hierarchical clustering

I. INTRODUCTION

Numerous attacks made by the malware, such as viruses, backdoors, spyware, Trojan Horses, and worms [1, 2, 18] have presented a major security threat to computer users. Nowadays, the most significant line of defense against malware is anti-virus products [19] which detects, removes, and characterizes these threats [3, 4]. The ability of these AV products to successfully characterize these threats is greatly depends on the method for categorizing these profiles of malware into groups effectively. Therefore, clustering malware into different families is one of the computer security topics that are of great interest. There are a few attempts on automatically malware categorization using clustering algorithms [5, 15]. In [5], the authors simply construct relationships between malware using hierarchical clustering algorithm [6]. In [15], the authors computed a set of centorid models by the K-means algorithm under Manhattan Distance as the similarity metric representing different families. However, hierarchical clustering lacks global objective function and k-means algorithm suffers shortcoming that there is no efficient and universal method for identifying the initial partitions and the number of clusters K. In this paper, resting on the analysis of the extracted instruction of malware samples, we propose a novel parameter-free hybrid clustering algorithm (PFHC) which combines the merits of hierarchical clustering and K-means algorithms for malware clustering. It can not only generate stable initial division, but also give the best K. PFHC first utilizes agglomerative hierarchical clustering algorithm as the frame, starting with N singleton clusters, each of which exactly includes one sample, then reuses the centroids of upper level in every level and merges two nearest clusters, finally adopts K-means algorithm for iteration to achieve an approximate global optimal division. PFHC evaluates clustering validity of each iteration procedure and generates the best K by comparing the values. The promising studies on real daily collection of the extracted instruction of malware samples from the anti-virus laboratory of Kingsoft Corporation illustrate that, compared with popular existing K-means and hierarchical clustering approaches, our proposed PFHC algorithm always generates much higher quality clusters and it can be well used for malware categorization.

The rest of the paper is organized as follows: Section 2 discusses the related work, Parameter-Free hybrid clustering algorithm is proposed in Section 3, the experimental results and conclusion are discussed in Section 4 and 5.

II. RELATED WORK

Malware categorization has increasingly become an urgent and complex task. The work in [5] is a significant attempt on automated malware categorization. The authors use pairwise single-linkage hierarchical clustering method to construct a tree structure relationship between malware, and then extract meaningful clusters by cutting links with "inconsistency coefficient" higher than a user-specified threshold. Because it measures the distance between two clusters by the two closest

The work of Z. Han, Y. Ye and Q. Jiang is supported by the Guangdong Province Foundation under Grant 2008A090300017. Shaorong Feng is the corresponding author.

samples in different clusters, it is especially sensitive to noise and outliers. Though there are other agglomerative clustering algorithms based on the different definitions for distance between two clusters [7], including complete linkage, group average linkage, median linkage, centroid linkage, and Ward's method, classical hierarchical clustering algorithms have their innate drawbacks. They lack global objective function, and are not capable of correcting possible previous misclassification as the mergers are ultimate. Besides, there is no a satisfactory mechanism to extract meaningful groups from the generated tree.

Another well-known clustering algorithm for malware categorization is Squared Error-Based partitioning clustering, such as K-means [8] and K-medoids [7,9] which assigns a set of samples into clusters using an iterative relocation technique [7]. A cluster is represented by one of its real sample (called medoids) or by the mean of its samples (called centroid) in K-medoids and K-means methods respectively. They are very simple but effective and widely used in many scientific and industrial applications. However, they suffer common criticism that there is no efficient and universal method for identifying the initial partitions and the number of clusters K. In addition, they are sensitive to outliers and noise and not suitable to explore non-spherical shape clusters.

In order to overcome the drawbacks of the above algorithms, researchers had introduced some interesting hybrid clustering methods [10, 11, 12 and 13]. In [10], the proposed approach combines hierarchical clustering and k-means methods, which runs hierarchical clustering at first, and then decides the location and the number of initial seeds by calculating and finding a big jump in the value of R-squares (RS) and centroid distance (CD), finally runs k-means with the initial seeds generated by hierarchical clustering. This method can be useful in correcting possible previous misclassification in hierarchical clustering. However it still has some drawbacks: (1) the mechanism of deciding the location and the clusters number lacks theoretical support and usually works poor; (2) it can be used in numerical data only as the definition of centroids of k-means. The jobs on [11, 12] tried hard to design an algorithm which did not need to specify clusters number K and could handle outliers. Their methods are similar to that in [10], except that they use a user-defined percentage to stop the hierarchical clustering process. The authors in [13] adopt an opposite way to design a two level hybrid clustering algorithm. The user can select k-means or hierarchical clustering techniques to get a small set of prototype vectors (cluster means) and then hierarchical clustering is used in second stage [16].

However, the existing hybrid methods execute the two singleton algorithms in two stages. That is using hierarchical clustering method to generate the initial seeds for K-means. The merger is too simple, inadequacy and lacking a prominent mechanisms to explore proper K. Those may result in a poor performance for malware categorization. In order to address these challenges, in this paper, resting on the analysis of the extracted instruction of malware samples, we propose a novel parameter-free hybrid clustering algorithm (PFHC) which combines the merits by cross-executing the hierarchical clustering and K-means clustering algorithms for malware categorization. PFHC can not only generate stable initial division, but also give the best K. It first utilizes agglomerative hierarchical clustering algorithm as the frame, starting with N singleton clusters, each of which exactly includes one sample, then reuses the medoids of upper level in every level and merges two nearest clusters, finally adopts K-means algorithm for iteration to achieve an approximate global optimal division. Meanwhile PFHC evaluates clustering validity in each iteration procedure and generates the best K by comparing the values.

III. PARAMETER-FREE HYBRID CLUSTERING ALGORITHM

In this paper, resting on the analysis of the extracted instruction of malware samples, we propose a novel parameter-free hybrid clustering algorithm which combines the merits of hierarchical clustering and K-means algorithms for malware clustering. That is to categorizing a set of malware profiles into families automatically.

A. Problem description

In this section, we will introduce the definitions of PFHC algorithm, including distance measure, medoids representing clusters, and a validation criterion FS.

- Definition 1 Cosine Similarity [7] is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them. Cosine similarity is used as our distance measure, because of its independent of data length. It is defined as follow:

$$S_{ij} = \cos\alpha = \frac{x_i^T x_j}{|x_i||x_j|} \tag{1}$$

- Definition 2 Cluster medoid [7, 9] is defined as the item in the cluster whose average dissimilarity to all other items in the same cluster is minimal. In order to make the algorithm can deal with non-numerical data; we use medoid replacing the centroid in K-means.

- Definition 3 Validation Criteria FS [14] is defined as follow:

$$V_{FS} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}^{m}(\| x_k - v_i \|^2 - \| v_i - \bar{v} \|^2) \tag{2}$$

Here, we use medoid v_i as the representative of a cluster, and $\bar{v}$ as the medoid of all data. Fukayama and Sugno[14] have introduced FS for evaluating k-partitions by exploiting the concepts of the compactness within each cluster and distances of the clusters representatives. The optimal partition is obtained by minimizing FS with respect to K = 2, 3, … , N (the number of all samples).

B. PFHC algorithm

Based on the definition given above, we propose a novel Parameter-Free Hybrid Clustering algorithm (PFHC) which combines the merits of hierarchical clustering and k-means. In order to clustering automatically without being specified the

clusters number K, we implement the validation Criteria FS to evaluate clustering validity of each iteration procedure and generate the best K by comparing the values.

The main idea of PFHC algorithm is as follow: First, it utilizes agglomerative hierarchical clustering algorithm as the frame, starting with N singleton clusters, each of which exactly includes one sample. The iteration procedure in each level is reusing the medoids of upper level and merging the two most similar clusters whose medoids are nearest to each other. Then it gets an initial partition, so it can run K-means (medoids) for iteration to achieve an approximate global optimal division using these reliable seeds (medoids). At the end of the iteration in this layer, we can evaluate the clustering validity in this layer using FS validation criteria, which is used to evaluate the clustering quality. Again we can iterate the merger and iteration procedures in other succeed layers, meanwhile calculate the corresponding clustering validities until all samples being the same cluster. At the end of this algorithm, by comparing all those validity values, we can achieve the best clusters number and corresponding clusters result.

Compared with the existing hybrid algorithms above [10, 11, 12 and 13], our proposed PFHC algorithm is parameter-free with the implement of cluster validation Criteria FS. In addition, the initial seeds are generated in a much better mechanism, so the clustering result is stable and reliable. As the cross-execution of the two clustering methods, PFHC always generates much higher quality clusters.

The outline of parameter-free hybrid clustering algorithm used for Malware Categorization is described as follows:

Input: The data set D

Output: The best K and data clusters

Algorithm:

1) Set each sample as a singleton cluster
2) Set each sample as its own medoid
3) For K=N-1 to 1
4) Merge two clusters with closest medoids
5) Generate the new medoids of the merged clusters
6) Repeat
7) For all samples
8) Assign it to the nearest cluster (medoid)
9) End for
10) Update the medoids of clusters
11) Until no medoids of clusters changed
12) Calculate the validity value in this layer
13) Compare and keep the best K and corresponding clusters until now
14) End for

IV. EXPERIMENT AND ANALYSIS

We use three daily malware samples obtained from the anti-virus laboratory of Kingsoft Corporation as our data set, named D_1, D_2 and D_3. There are 711, 1277 and 1210 malware samples respectively. And the families for all samples are pre-marked by the malware analysts.

Resting on the analysis of the extracted instruction of malware samples, we use the frequency of extracted instructions as the feature of malware sample and conduct two sets of experiments on the three daily sets. First, we evaluate the ability of PFHC generating the proper K. Second, we evaluate the performance of our clustering algorithm compared with other classical clustering methods using Micro-average and Macro-average measures [17]. Both of the experiments are conducted under the environment of Windows XP OS plus Intel Core Duo 1.66GHz CPU and 1GB of RAM.

A. Generating the proper cluster number

As described above, determining the clusters number K is an important problem. Most existing clustering algorithms need to be specified K. In our proposed algorithm, we use FS [14] described above as the validation criteria and conduct the clustering process to generate the clusters number and corresponding clusters.

We apply PFHC on the three daily sample set D1, D2 and D3 respectively, and the results are shown in Table 1:

TABLE I. GENERAGED CLUSTERS NUMBER K

Data	Comparison between real and generated K		
	|D|	***Real K***	***Generated K***
D_1	711	46	50
D_2	1159	35	40
D_3	1200	50	52

(|D|: is the number of samples in data set D; K: is the number of families)

The experimental results illustrate that the clusters number generated by PFHC is close to the real clusters number pre-marked by the malware analysts. As the analysts mark the samples by their own empirical knowledge, the slightly difference between the real clusters number and generated clusters number is in an allowable range. So PFHC can generate a good clusters number for automatically malware categorization.

B. Comparison of different clustering methods

In order to evaluate the performance of our clustering algorithm, we compare PFHC with popular existing K-means and hierarchical clustering approaches, and another Two-level hybrid clustering method TLHC proposed in [10]. TLHC runs hierarchical clustering at first, and then decides the location and the number of initial seeds by calculating and finding a big jump in the value of R-squares (RS) and centroid distance (CD), finally runs k-means with the initial seeds generated. In addition, we also implement FS in hierarchical clustering algorithm to promote its performance. In this section, we measure the clustering performance of different algorithms using Micro-average and Macro-average measures [15]. Results shown in Table 2:

TABLE II. COMPARTION DIFFERENT ALGORITHMS

Data	PFHC		TLHC	
	Micro-F1	***Macro-F1***	***Micro-F1***	***Macro-F1***
D_1	0.791275	0.841077	0.677567	0.757598
D_2	0.853996	0.895881	0.74803	0.818381

Data	PFHC		TLHC	
	Micro-F1	***Macro-F1***	***Micro-F1***	***Macro-F1***
D_3	0.824664	0.885506	0.787029	0.862023
Data	**K-means**		**Hierarchical clustering**	
	Micro-F1	***Macro-F1***	***Micro-F1***	***Macro-F1***
D_1	0.603655	0.679984	0.743095	0.799706
D_2	0.673412	0,710243	0.828655	0.883034
D_3	0.604879	0.702585	0.708328	0.774655

(Micro-F1, is the Micro average measure; Macro-F2 is the Macro average measure)

The experimental results illustrate that: (1) PFHC can always generate much higher quality clusters than other three existing algorithms; (2) PFHC is stable and reliable by comparing the results of different days.

V. CONCLUSION

In this paper, we have proposed a novel parameter-free hybrid clustering algorithm using for malware categorization. Our algorithm is stable and reliable method to achieve initial seeds and has a good mechanism to explore good clusters number. In addition, the PFHC algorithm can automatically and effectively categorize a set of malware profiles in to different families and perform better than other clustering methods, such as hierarchical clustering and k-means approaches.

ACKNOWLEDGMENT

The authors would like to thank the members in the Anti-virus Lab at Kingsoft Corporation for their helpful discussions and suggestions.

REFERENCES

[1] G. McGraw and G. Morrisett, "Attacking malicious code: A report to the infosec research council," IEEE Software, vol 17, pp. 33–41, September 2000.

[2] M.Christodorescu, S. Jha, S. Seshia, D. Song, and R. Bryant, "Semantics-aware malware detection," In Proceedings of the 2005 IEEE Symposium on Security and Privacy, pp. 32–46, May 2005.

[3] E. Filiol, "Malware pattern scanning schemes secure against blackbox analysis," Journal in Computer Virology, vol 2, pp. 35–50, May 2006.

[4] N. Idika and AP. Mathur, "A Survey of Malware Detection Techniques, " Purdue University, 2007.

[5] M. Bailey, J Oberheide, J Andersen, Z. Morley Mao, F. Jahanian, and J. Nazario, "Automated classification and analysis of internet malware," In Proceedings of the 10th Symposium on Recent Advances in Intrusion Detection RAID, pp. 178-197, August 2007.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning Data Mining, Inference, and Prediction," New York: Springer , pp.79-114, 2001.

[7] R. Xu and II Donald Wunsch, "Survey of Clustering Algorithms," IEEE transactions on neural networks, vol. 16, pp. 645-678, May 2005.

[8] J. Hartigan and M. Wong, "A k-means clustering algorithm," Applied Statistics, vol. 28, pp. 100-108, 1979.

[9] L. Kaufman and P.J. Rousseeuw PJ, "Finding groups in Data: An Introduction to Cluster Analysis," New York: Wiley, March 1990.

[10] S. Kwon and C. Han, "Hybrid Clustering Method for DNA Microarray Data Analysis," Genome Informatics, vol. 13, pp. 258-259, 2002

[11] Chen, B. Tai, P.C. Harrison, and R. Yi Pan, "Novel Hybrid Hierarchical-K-means Mehod (H-K-means) for Microarray Analysis," Computational Systems Bioinformatics conference, pp. 105-108, 2005.

[12] Chen, B. Jieyue He, Pellicer, and S. Yi Pan, "Protein Sequence Motif Super-Rule-Tree (SRT) Structure Constructed by Hybrid Hierarchical K-means Clustering Algorithm," Conf. BIBM 2008, pp. 98-103, November 2008.

[13] EY. Cheu, C. Keongg, and Z. Zhou, "On the two-level hybrid clustering algorithm," In International conf. on artificial intelligence in science and technology, pp. 138–142, 2004.

[14] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," Proc. 5th Fuzzy Syst. Symp, pp. 247-250, 1989.

[15] S.J. Stolfo, K. Wang, and W.J. Li, "Fileprint analysis for Malware Detection," WORMS 2005, June 2005.

[16] T. Kohonen, "Self-organizing maps: Optimization Approaches," In proceedings of the International conf. on artificial neural Networks, pp. 981-990, 1991.

[17] A. Ozgur, L. Ozgur, and T. Gungor, "Text categorization with class-based and corpus-based keyword selection," In Proceedings of the 20th International Symposium on Computer and Information Sciences, vol. 3733, pp. 607–616, 2005.

[18] Y. Ye, D.Wang, T. Li, and D. Ye, "IMDS: Intelligent malware detection system," In Proccedings of ACM International onference on Knowlege Discovery and Data Mining, pp. 1043-1047, 2007.

[19] Y. Ye, D.Wang, T. Li, D. Ye and Q. Jiang, "An intelligent PE-malware detection system based on association mining." Journal in Computer Viorology, 2008.