International Journal of Security and Its Applications
Vol. 10, No. 1 (2016), pp.343-354
http://dx.doi.org/10.14257/ijsia.2016.10.1.31

An Improved Kernel Clustering Algorithm for Mixed-Type Data in Network Forensic

Min Ren[1, 2], Peiyu Liu[1, 3,*], Zhihao Wang[1], Lin Lü[1]

[1] *School of Information Science and Engineering, Shandong Normal University, Shandong, China*
[2] *School of Mathematic and Quantitative Economics, Shandong University of Finance and Economics, Shandong, China*
[3] *Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Shandong, China*
[1] *sdkeylab@163.com,* [2] *rm_sd@163.com*
[*] *Corresponding Author*

Abstract

Clustering algorithm is a common analysis technology for network forensics, which, lacking of any prior knowledge, can effectively find out the invasions by analyzing the collected real-time communication data flowing through the network. This paper proposed an improved dynamic kernel clustering algorithm for mixed numeric and categorical network communication data. First, centroid prototype based on the mean and distribution centroid was put forward to represent the cluster center. Then by using Gaussian kernel function, the paper introduced a new dissimilarity measure between the data object and the centroid prototype in combination with the significance of different categorical values. On this basis, the objective function was defined, which took into account both the compact degree in a cluster and the discrete degree among the clusters. After that an improved kernel clustering algorithm was designed. In the process of clustering, centroid prototype and the value of the clustering parameter dynamically updated for a better description of the characteristics of clusters' change. Finally, in order to verify the feasibility and effectiveness of the algorithm, the paper further applied it to network forensics, and the experimental results showed that the method could mine the intrusion behavior more accurately.

Keywords*: Kernel Clustering; Gaussian Kernel; Mixed-type Data; Network Forensics*

1. Introduction

Network forensic uses various techniques to analyze the evidence obtained from network intrusion events and network crimes, aiming to find the invaders and their locations, as well as to explain the specific content and process of their invasions. According to the forensic time, network forensics can be divided into post forensics and real-time forensics. In general, real-time forensics often cooperates with Intrusion Detection System (IDS).

Communication packets through the network are important evidence objects of network forensics. These data are high-dimensional and mixed with the features of numeric and categorical data. Without prior knowledge, it is a very difficult and complex task to cluster information that can reflect the objective fact of the invasion from such a dataset. Clustering is one of the most popular and suitable techniques widely used for the analysis of network forensic. It is unsupervised and can divide the objects of a dataset into classes or clusters, resulting in a high similarity among objects of the same cluster, whereas a huge diversity occurs between different clusters.

ISSN: 1738-9976 IJSIA

Gaussian kernel function nonlinearly maps the samples of the input space to a high-dimensional feature space, enlarges the difference between data, and has certain robustness to noise and boundary. Thus, this paper, by using it, presented an improved dynamic kernel clustering algorithm. The main contributions are as follows: 1. The mean and distribution centroid respectively stand for the numeric and categorical attribute of the cluster prototype. On the basis of Gaussian kernel function, a new dissimilarity measure is defined, which calculates quantitative evaluation value of categorical attributes according to their semantics. 2. A new objective function is used in conjunction with the clustering algorithm, considering the maximum similarity of the inner cluster and the minimum similarity of the outer clusters. 3. The clustering prototype and the value of the clustering parameter are dynamically changed in the process of clustering so as to better reflect the changes of characteristics of clusters, improve the clustering accuracy and achieve the fast convergence.

The rest of the paper is organized as follows. In next section, we discuss some clustering algorithms for mixed numeric and categorical data. Section 3 describes the improved kernel clustering algorithm in details. In Section 4, we present the experiments and results to demonstrate the advantages of the proposed algorithm in Network Forensic. Finally, conclusions and future work are given in Section 5.

2. Related Work

Traditional clustering algorithms mainly deal with the numeric data, such as K-means [1], BRICH [2], CURE [3] and so on. With the expansion of application fields, researchers also did some works on the clustering algorithms for the data with categorical attributes, such as K-modes [4], ROCK [5], COOLCAT [6] and so on. These algorithms can only apply to single-type data. However, most datasets are combined with both numeric attributes and categorical attributes, so the clustering problem of data with mixed attributes has caused wide attention at home and abroad. In particular, clustering algorithms for mixed-type data are mainly divided into two types.

The first type makes use of attribute conversion to unify the type of attributes, and then traditional single-type clustering algorithms can be applied to converted data [7]. There are two approaches of attribute conversion [8]. One is that categorical values are converted to numeric values. Binary encoding is a common technology, which transforms categorical values to a set of 0 and 1 binary values [9]. However, the conversion increases the calculation and space complexity due to the higher dimension of data. Besides, the conversion in references [7, 10] depended on the frequency of each categorical value in the domain. But it did not consider the significance of categorical values. For example, two values of a categorical attribute with the same frequency will be mistaken for their maximum similarity. Another problem is that numeric values are discretized and converted to categorical values [11]. In many cases, the discretization leads to the loss of information [12]. In addition, it also results in a boundary problem that two close values near a discretization boundary may be assigned to different ranges [8].

Secondly, clustering algorithm for mixed-type data is redesigned, which deals with the numerical and categorical attributes respectively. For example, Huang [13] proposed the famous K-Prototypes, combining with partitional clustering algorithm K-means and K-modes. The prototype of the algorithm was made up of the mean of the numeric attributes and the mode of the categorical attributes in the cluster. Euclidean distance was used for the dissimilarity measure between numeric values, and the simple matching for that between categorical values, *i.e*. 0 for identical values and 1 for different values. And, a parameter was introduced to control the weight of categorical attributes in the clustering process. Adopting Goodall similarity measure, Li and Biswas [14] proposed the Similarity Based Agglomerative Clustering algorithm (SBAC), which gave the greater weight to the feature value of categorical attribute that appeared less. And for numeric

attribute it considered not only the difference of its feature value, but also the uniqueness of pairs of its values. Hsu *et al*. [8] presented a clustering algorithm named CAVE which used different methods for numeric and categorical attributes to measure the similarity, *i.e*. variance for numeric values and entropy with distance hierarchies for categorical values. Subsequently, Hsu *et al*. [15] proposed an incremental clustering algorithm by combining the thought of the adaptive resonance theory network and conceptual distance hierarchy. Later, in references [12, 16], Hsu *et al*. put forward a GMixSOM algorithm by using the self-organizing map and the distance level to solve the clustering problem of multivariate mixed-type data. Ahmad and Dey first proposed an algorithm of K-means type [7] to process the mixed-type data, which defined a new objective function to improve the cost function of k-prototypes and calculated the distances between the categorical values by using the co-occurrence of the values, and then put forward a subspace clustering algorithm [17] of K-means type for the data with high-dimensional mixed attributes. Zheng *et al*. [18] proposed the EKP algorithm by introducing evolutionary algorithm framework. Plant and Bohm [19] put forward a top-to-bottom hierarchy clustering algorithm named INCONCO. Ji *et al*. [20] proposed an improved k-prototypes algorithm for mixed data. The algorithm introduced the distribution centroid for representing the prototype of categorical attributes and a new dissimilarity measure taking into account the significance of different categorical attributes. Cheung *et al*. [21] proposed a clustering framework based on the concept of object-cluster similarity and defined a unified similarity metric for both numerical and categorical attributes. Liu [22] put forward a clustering algorithm based on average mutual information. Entropy was used to quantify the category characteristics of the parameter and measure the similarity and the difference between category characteristics according to the average mutual information.

So far, there have been lots of clustering algorithms for the mixed numeric and categorical data. But, none of them can be applied to all mixed-type datasets. As for specific data, it is required to find a matching algorithm according to realistic needs [23]. Therefore, this paper proposed an improved kernel clustering algorithm for the mixed-type network communication data with unique characteristics in order to obtain the better clustering result.

3. Proposed Algorithm

3.1. Notations

Let $D = \{X_1, X_2, \cdots, X_N\}$ denote the dataset of N objects and each object consists of n attributes $A_1, A_2, \cdots, A_t, A_{t+1}, \cdots, A_n$, in which $A_1, A_2, \cdots, A_t$ and $A_{t+1} \cdots, A_n$ are respectively the numerical attributes and the categorical attributes. The i^{th} data object is represented as $X_i = \{X_{i1}, X_{i2}, \cdots, X_{it}, X_{i(t+1)} \cdots, X_{in}\}$, with X_{ij} being its j^{th} attribute and x_{ij} being the corresponding value. The number of clusters is set as k , the set of clusters is $C = \{C_1, C_2, \cdots, C_k\}$, the corresponding set of cluster prototypes is $V = \{V_1, V_2, \cdots, V_k\}$, and the l^{th} prototype is described as $V_l = \{v_{l1}, v_{l2}, \cdots, v_{lt}, c_{l(t+1)}, \cdots, c_{ln}\}$.

3.2. Centroid Prototype

The centroid prototype is utilized in this paper to measure the center trends metric of the mixed-type data. For numeric attributes, the mean of the numeric value of the samples in the cluster C_l will be regarded as its centroid, that is

$$v_{lj} = \frac{1}{N_l}\sum_{i=1}^{N} u_{li} x_{ij},\ l = 1,2,\cdots,k,\ j = 1,2,\cdots,t \tag{1}$$

where $u_{li} = \begin{cases} 1, & X_i \in C_l \\ 0, & X_i \notin C_l \end{cases}$, with $N_l = \sum_{i=1}^{N} u_{li}$ being the number of data objects in cluster C_l.

For categorical attributes, the centroid of the cluster C_l is represented by distribution centroid used in [20]. The value domain of the j[th] categorical attribute is assumed as $Dom(A_j) = \{a_{j1}, a_{j2}, \cdots, a_{js}\}$, with s being the number of values. Then the centroid of the j[th] categorical attribute in cluster C_l is denoted as $c_{lj} = \{(a_{j1}, \omega_{lj1}), (a_{j2}, \omega_{lj2}), \cdots, (a_{js}, \omega_{ljs})\}$, and the value pair (a_{jm}, ω_{ljm}), $m = 1,2,\cdots,s$, shows the distribution of the m[th] value of the j[th] categorical attribute in the cluster C_l. Therefore,

$$\omega_{ljm} = \sum_{i=1}^{N} \eta(x_{ij}) \tag{2}$$

and

$$\eta(x_{ij}) = \begin{cases} \frac{u_{li}}{N_l}, & x_{ij} = a_{jm} \\ 0, & x_{ij} \neq a_{jm} \end{cases} \tag{3}$$

ω_{ljm} satisfies $0 \le \omega_{ljm} \le 1$ and $\sum_{m=1}^{s} \omega_{ljm} = 1$.

3.3. Dissimilarity Measure

Assume that ϕ is a nonlinear mapping function, x and y are the samples in the same dataset, $\phi(x)$ and $\phi(y)$ are the mappings of x and y to a particular feature space Φ, and then the distance between x and y in the feature space can be expressed as

$$d_\phi(x,y) = \|\phi(x) - \phi(y)\| = \sqrt{\phi(x)\phi(x) - 2\phi(x)\phi(y) + \phi(y)\phi(y)} = \sqrt{K(x,x) - 2K(x,y) + K(y,y)} \tag{4}$$

$K(x,y)$ is an inner product kernel function. Gaussian kernel is a commonly used one with an infinite-dimensional feature space, in which limited samples must be linearly separable and the problem of linear inseparability can be solved effectively. Gaussian kernel is defined as

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right) \tag{5}$$

where $\sigma = \sigma_0 d$ is the width of Gaussian kernel, d is the diameter of the dataset, and σ_0 is the parameter of the width of the kernel. And then

$$d_\phi(x,y) = \|\phi(x) - \phi(y)\| = \sqrt{2 - 2K(x,y)} \tag{6}$$

Based on centroid prototype and the significance of the categorical value, this paper proposed a new dissimilarity measure between the object and the prototype. If $X_i \in D$ and $V_l \in V$, their distance is given by

$$M\_Dist(X_i, V_l) = \sqrt{2 - 2K(X_i, V_l)} \tag{7}$$

where

$$K(X_i, V_l) = \exp\left(- \frac{\sqrt{\sum_{j=1}^{t} d\left(x_{ij}, v_{lj}\right)} + \gamma \sum_{j=t+1}^{n} d\left(x_{ij}, c_{lj}\right)}{\sigma^2} \right)^2 \tag{8}$$

$$d\left(x_{ij}, v_{lj}\right) = \left(x_{ij} - v_{lj}\right)^2 \tag{9}$$

$$d\left(x_{ij}, c_{lj}\right) = \sum_{m=1}^{s} \vartheta\left(x_{ij}, a_{jm}\right) \tag{10}$$

and

$$\vartheta\left(x_{ij}, a_{jm}\right) = \begin{cases} 0 & x_{ij} = a_{jm} \\ \omega_{ljm} & x_{ij} \neq a_{jm} \end{cases} \tag{11}$$

The parameter γ is used to measure the role of categorical attributes in the clustering process. In particular, when $\gamma = 0$, categorical attributes do not work, while the larger the value of γ is, the greater the role of categorical attributes is.

3.4. 0bjective Function

The clustering criterion is tantamount to minimize the following objective function.

$$J = \sum_{l=1}^{k} \left(\frac{1}{N_l} \sum_{i=1}^{N} u_{li} M\_Dist\left(X_i, V_l\right) \right) - \frac{2}{k(k-1)} \sum_{l=1}^{k} \sum_{l'=1}^{l-1} D\left(V_l, V_{l'}\right) \tag{12}$$

The first item in the formula is the sum of the average distances of the inner clusters, which is used to measure the compact degree in the cluster. The second is the average distance between any two centroid prototypes to measure the dispersion degree between the clusters. $D\left(V_l, V_{l'}\right)$ denotes the dissimilarity between two centroid prototypes, which is still based on Gaussian kernel, and is defined as

$$D\left(V_l, V_{l'}\right) = \sqrt{2\left(1 - K\left(V_l, V_{l'}\right)\right)} = \sqrt{2\left(1 - \exp\left(- \frac{\sqrt{\sum_{j=1}^{t} d\left(v_{lj}, v_{l'j}\right)} + \gamma \sum_{j=t+1}^{n} d\left(c_{lj}, c_{l'j}\right)}{\sigma^2} \right)^2\right)} \tag{13}$$

But, if the j[th] attribute is categorical, then

$$d\left(c_{lj}, c_{l'j}\right) = \sqrt{\sum_{m=1}^{s} \left(\omega_{ljm} - \omega_{l'jm}\right)^2} \tag{14}$$

3.5. Improved Kernel Clustering aAgorithm

This paper proposed an improved kernel clustering algorithm for mixed-type dataset. The algorithm is a dynamic method. In the process of clustering, cluster prototypes and the diameter of the dataset change dynamically, which can better reflect the characteristics of the clusters. The algorithm is described as follows:

Step 1. Normalize the samples in the dataset;

Step 2. Input the initial cluster centroid prototypes;

Step 3. For each sample, repeat Step 3.1 to 3.3;

Step 3.1 Calculate the distance between the sample and the cluster centroid prototype, and classify the sample to the closest cluster;

Step 3.2 Update the cluster centroid prototypes;

Step 3.3 Update the diameter of the dataset;

Step 4. Calculate the objective function;

Step 5. Repeat Step 3 to Step 4, until the objects in each clusters no longer change or the maximum iteration is reached;

Step 6. Output k clusters.

The initial cluster prototypes are selected randomly from objects in the dataset. When it is converted to centroid prototypes, the numeric attribute remains unchanged, but the categorical one is represented as $c_{lj} = \{(a_{j1},\omega_{lj1}),(a_{j2},\omega_{lj2}),\cdots,(a_{js},\omega_{ljs})\}$. If the categorical value is a_{jm}, then $\omega_{ljm}=1$, otherwise $\omega_{ljm}=0$.

4. Experiment and Result Analysis

4.1. Dataset and Data Normalization

The KDD CUP 1999, a benchmark evaluation dataset which is commonly recognized and widely used in the field of intrusion detection, consists of about 7 million link records, including normal data and four kinds of attack data, such as DoS (Denial of Service), Probe, R2L (Remote-to-local) and U2R (User-to-root). They are obtained by Lincoln Laboratory of MIT that conducts proper treatment and feature selection on the 9-week tcpdump data provided by DARPA. Each record has 42 attributes. Except that the last attribute is used to describe whether it is a normal link or an intrusion behavior, it has a total of 34 numeric attributes and 7 categorical attributes, which are extracted from a primitive network connection. Our experimental dataset consists of 24,573 pieces of data randomly selected from the 10% of the training subset with a total of 494,021 pieces of data and 22 types of attacks. We only select Nomal, Probe and DoS categories with 8 behaviors, because U2R and R2L attacks are relatively few. The data type and distribution of the experimental dataset are shown in Table 1.

Table 1. The Data Type and Distribution of the Experimental Dataset

Category	Behavior	Number of samples selected
Normal	normal	4,801
Probe	ipsweep	63
	portsweep	57
	satan	84
DoS	back	98
	neptune	5,455
	smurf	13,943
	teardrop	72
Total		24,573

In this paper, numeric attributes are normalized [10]. Let $\bar{x}_j$ be the mean of the j^{th} attribute value, then

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{15}$$

Its standard deviation s_j is calculated by the formula below:

$$s_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)^2} \tag{16}$$

Then, the numeric value x_{ij} can be normalized as

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{17}$$

After the normalization, the numeric value is mapped to the space of normal value.

4.2. Evaluation Method

The Clustering Accuracy, which is a kind of widely used evaluation standard proposed by Huang and Ng [24, 25], is used in this paper to evaluate the quality of the clustering results. The Clustering Accuracy r is defined as follows:

$$r = \frac{\sum_{i=1}^{k} b_i}{N} \tag{18}$$

Here b_i is the number of the objects which co-occur in the ith cluster and the ith real cluster, and N is the number of the objects in the dataset. According to this measurement, the higher the Clustering Accuracy is, the better the clustering result of the algorithm is. When $r = 1$, the clustering result of the algorithm on the dataset is totally accurate.

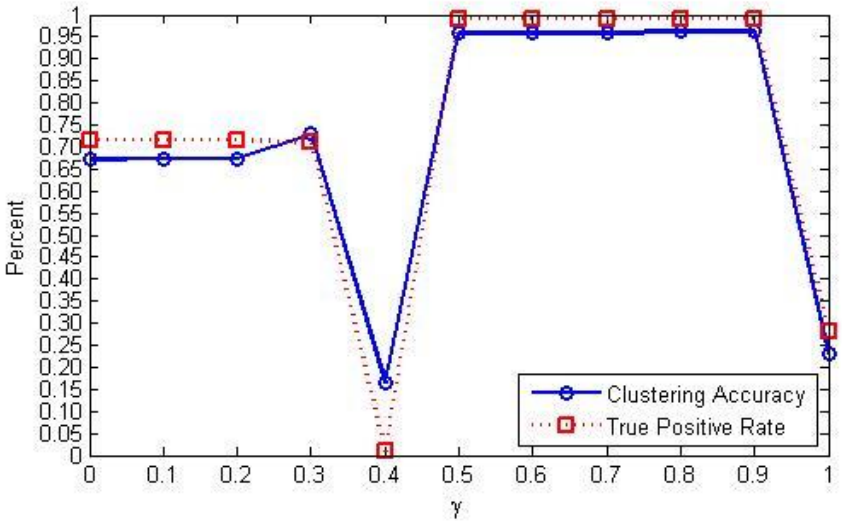
In order to validate the effectiveness of our method in network forensics, this paper also calculates True Positive rate (TP).

$$TP = \frac{number\ of\ invasions\ detected\ correctly}{total\ number\ of\ invasions} \times 100\% \tag{19}$$

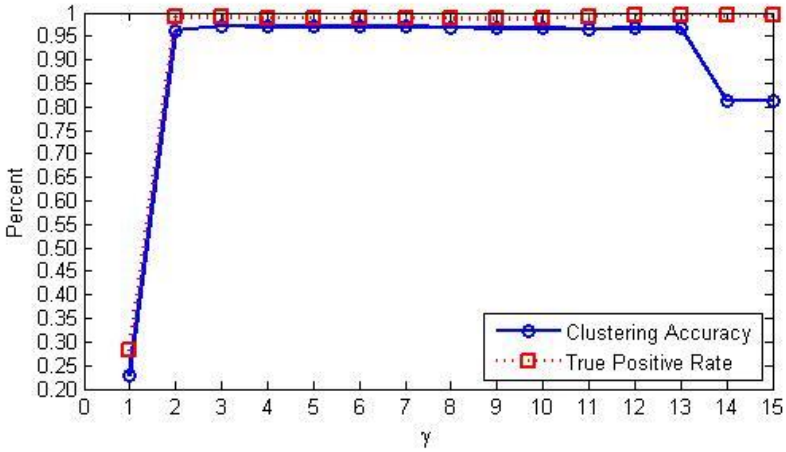
4.3. Experimental Results

The range of parameter σ_0 is generally between 0.02 and 0.2, and for high-dimensional dataset (*e.g.*, USPS), it is often limited between 0.1 and 0.2 [26]. This paper uses the default value, namely, $\sigma_0 = 0.05$. Let $k = 8$ be the actual number of clusters.

We first analyzed the influence of the different parameter γ in clustering. Select the same initial clustering prototypes, change the value of the parameter γ, and calculate the corresponding Clustering Accuracy and TP respectively, as shown in Figure 1. In Figure 1 (a), when γ is between 0.5 and 0.9, implying that categorical attributes are less import than numerical ones, Clustering Accuracy reaches 95.7% - 95.7% and TP 99.1% - 99.2%. But the defect is that Clustering Accuracy of back attack is 0. And when γ is between 2 and 13, as depicted in Figure 1 (b), implying that categorical attributes are relatively more important than numerical ones, Clustering Accuracy can reach 96.2% - 97.3% and TP reaches 99.1% -99.7%. Clustering Accuracy of back attack gradually increases, especially when $\gamma = 12$, it can reach 100%, but compared with those when $\gamma < 12$, Clustering Accuracy of normal accuracy slightly reduces. Given that our forensics aim is to find out the attack events, we make $\gamma = 12$.



(a)



(b)

Figure 1. Training of Parameter γ

Figure 2 and Figure 3 describe the dynamic change of the objective function value and the diameter of the dataset in each iteration. Obviously, in the third iteration, the algorithm has been converged.

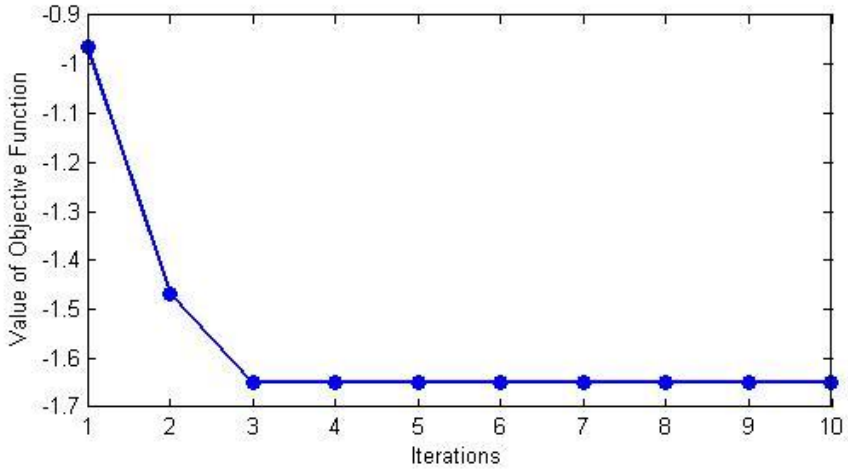


Figure 2. Value of Objective Function in each Iteration

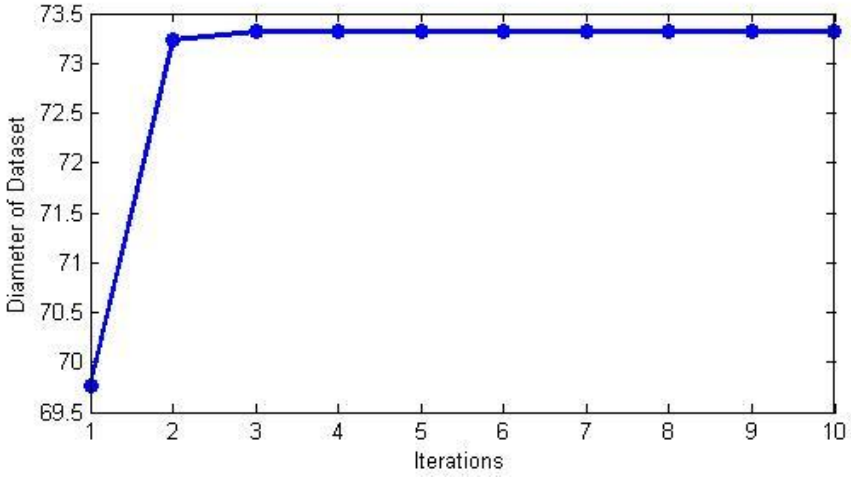


Figure 3. Diameter of Dataset in each Iteration

Then we used the same initial clustering prototypes and compared the proposed algorithm with the classical K-means and K-Prototypes. For K-means algorithm, the categorical attributes were normalized according to the occurrence frequency of their values. For K-Prototypes, the value of parameter γ is set as 8 through experiments. The comparison of Clustering Accuracy is shown in table 2. As can be seen, the proposed algorithm can achieve the highest accuracy in most cases, especially the Clustering Accuracy of ipsweep and portsweep.

Table 2. Clustering Accuracy

Behavior	K-means	K-prototype	Proposed algorithm
normal	58.72%	84.98%	86.82%
ipsweep	74.60%	74.60%	88.89%
portsweep	63.16%	78.95%	89.47%
satan	73.81%	76.19%	73.81%
back	0%	100%	100%
neptune	99.95%	99.98%	100%

smurf	99.61%	100%	100%
teardrop	0%	100%	100%
Total	90.77%	96.87%	97.28%

In order to analyze the overall detecting results, the paper calculated TP of 2 categories without considering the specific invasion behavior, while all kinds of behaviors were classified as a major category. Table 3 is the experimental results, which show good effects of the proposed algorithm in intrusion detection. Not only did TP of Dos attack reach 100%, but also that of Probe attack improved significantly.

Table 3. True Positive Rate

Category	K-means	K-prototype	Proposed algorithm
Probe	71.08%	76.47%	82.84%
Dos	98.84%	99.99%	100%
Total	98.55%	99.75%	99.82%

5. Conclusion and Future Work

Network forensics technology has become one of the most important topics of the international network security. The improved kernel clustering algorithm proposed in the paper is feasible for the application of network forensics and can obtain better clustering results. However, its biggest flaw is that initial clustering prototypes have a great impact on the clustering result. Besides, the fuzziness of objects in the cluster is not taken into account and the number of clusters in the algorithm needs to be manually set. So these are part of our future research work.

Acknowledgements

This work is partially supported by National Natural Science Foundation (61373148, 61502151), National Social Science Fund (12BXW040), Shandong Province Natural Science Foundation (ZR2012FM038, ZR2014FL010), Shandong Province Outstanding Young Scientist Award Fund (BS2013DX033), Science Foundation of Ministry of Education of China (14YJC860042) and Project of Shandong Province Higher Educational Science and Technology Program (No.J15LN02). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which improve the presentation greatly.

References

[1] S. P. Lloyd, “Least Square Quantization in PCM”, IEEE Transactions on Information Theory, vol. 28, no. 2, **(1982)**, pp. 129-137.
[2] T. Zhang, R. Ramakrishnan and M. Livny, “BIRCH: An Efficient Data Clustering Method for Very Large Databases”, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, USA, **(1996)** June 103-114.
[3] S. Guha, R. Rastogi and K. Shim, “CURE: An Efficient Clustering Algorithm for Large Databases”, Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, Washington, USA, **(1998)** June 73-84.
[4] Z. Huang, “A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining”, Proceedings of Research Issues on Data Mining and Knowledge Discovery, Arizona, USA, **(1997)** May 1-8.
[5] S. Guha, R. Rastogi and K. Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes”, Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, **(1999)** March 512-521.
[6] D. Barbará, Y. Li and J. Couto, “COOLCAT: An Entropy-based Algorithm for Categorical Clustering”, Proceedings of the 11th International Conference on Information and Knowledge Management, Virginia, USA, **(2002)** November 582-589.

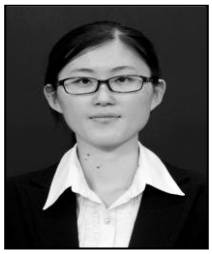
[7] A. Ahmad and L. Dey, "A K-mean Clustering Algorithm for Mixed Numeric and Categorical Data", Data & Knowledge Engineering, vol. 63, no. 2, **(2007)**, pp. 503-527.
[8] C. C. Hsu and Y. C. Chen, "Mining of Mixed Data with Application to Catalog Marketing", Expert Systems with Applications, vol. 32, no. 1, **(2007)**, pp. 12-23.
[9] H. Ralambondrainy, "A Conceptual Version of the K-means Algorithm", Pattern Recognition Letters, vol. 16, no. 11, **(1995)**, pp. 1147-1157.
[10] Y. Li, B. X. Fang, L. Guo and Z. H. Tian, "Supervised Intrusion Detection Based on Active Learning and TCM-KNN Algorithm", Chinese Journal of Computers, vol. 30, no. 8, **(2007)**, pp. 1464-1473.
[11] G. David and A. Averbuch, "SpectralCAT: Categorical Spectral Clustering of Numerical and Nominal Data", Pattern Recognition, vol. 45, no. 1, **(2012)**, pp. 416-433.
[12] C. C. Hsu, S. H. Lin and W. S. Tai, "Apply Extended Self-organizing Map to Cluster and Classify Mixed-type Data", Neurocomputing, vol. 74, no. 18, **(2011)**, pp. 3832-3842.
[13] Z. Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values", Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, **(1997)** February 21-34.
[14] C. Li and G. Biswas, "Unsupervised Learning with Mixed Numeric and Nominal Data", IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 4, **(2002)**, pp. 673-690.
[15] C. C. Hsu and Y. P. Huang, "Incremental Clustering of Mixed Data Based on Distance hierarchy", Expert Systems with Applications, vol. 35, no. 3, **(2008)**, pp. 1177-1185.
[16] W. S. Tai and C. C. Hsu, "Growing Self-Organizing Map with Cross Insert for Mixed-type Data Clustering", Applied Soft Computing, vol. 12, no. 9, **(2012)**, pp. 2856-2866.
[17] A. Ahmad and L. Dey, "A K-means Type Clustering Algorithm for Subspace Clustering of Mixed Numeric and Categorical Datasets", Pattern Recognition Letters, vol. 32, no. 7, **(2011)**, pp. 1062-1069.
[18] Z. Zheng, M. Gong, J. Ma, L. Jiao and Q. Wu, "Unsupervised Evolutionary Clustering Algorithm for Mixed Type Data", Proceedings of the 2010 IEEE Congress on Evolutionary Computation, Barcelona, Spain, **(2010)** July 1-8.
[19] C. Plant and C. Böhm, "INCONCO: Interpretable Clustering of Numerical and Categorical Objects", Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, **(2011)** August 1127-1135.
[20] J. Ji, T. Bai, C. Zhou, C. Ma and Z. Wang, "An Improved K-prototypes Clustering Algorithm for Mixed Numeric and Categorical Data", Neurocomputing, vol. 120, **(2013)**, pp. 590-596.
[21] Y. M. Cheung and H. Jia, "Categorical and Numerical Attribute Data Clustering Based on a Unified Similarity Metric without Knowing Cluster Number", Pattern Recognition, vol. 46, no. 8, **(2013)**, pp. 2228-2238.
[22] J. S. Liu, "Clustering with Mixed Condition Attributes Based on Average Mutual Information", Computer Science, vol. 42, no. 3, **(2015)**, pp. 261-265.
[23] F. Weng, Q. Jiang, L. Chen and Z. Hong, "Clustering Ensemble Based on the Fuzzy KNN Algorithm", Proceedings of Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Qingdao, China, **(2007)** July 1001-1006.
[24] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, vol. 2, no. 3, **(1998)**, pp. 283-304.
[25] Z. X. Huang and M. K. Ng, "A Fuzzy K-modes Algorithm for Clustering Categorical Data", IEEE Transactions on Fuzzy Systems, vol. 7, no. 4, **(1999)**, pp. 446-452.
[26] R. N. Ma, X. L. Wang and J. D. Ding, "Multilevel Core-Sets Based Aggregation Clustering Algorithm", Journal of Software, vol. 24, no. 3, **(2013)**, pp. 490-506.

Authors

Min Ren, She received the Bachelor degree in Shandong Finance Institute, China, in 2001 and the Master degree in Shandong Normal University, China, in 2006. Now she is an associate professor in Shandong University of Finance and Economics, China, and a PhD candidate in Shandong Normal University, China. Her main research interests include network forensic and clustering algorithm.

Peiyu Liu, He received the Master degree in East China Normal University, China, in 1986. Now he is a professor and doctoral supervisor in Shandong Normal University. His main research interests include network information security and data mining.

Zhihao Wang, He received the Bachelor degree in Jining Medical University, China, in 2010 and the Master degree in Shandong Normal University, China, in 2013. Now he is pursuing the PhD degree in Shandong Normal University, China. His main research interests include intelligent information security and evolutionary algorithm.



Lin Lü, She received the Bachelor degree in Shan Dong Normal University, China, in 2014. Now she is pursuing the Master degree in Shan Dong Normal University, China. Her main research interest is network information security.