The Application of Association Analysis in Mobile Phone Forensics System

Huan Li, Bin Xi, Shunxiang Wu, Jingchun Jiang, Yu Rao

The Application of Association Analysis in Mobile Phone Forensics System

Huan Li*,Bin Xi**,Shunxiang Wu***,Jingchun Jiang****,Yu Rao*****

*Department of Automation, Xiamen University
No.422, Siming South Road, Xiamen, Fujian,
361005,
China
1748468754@qq.com

**Department of Automation, Xiamen University
No.422, Siming South Road, Xiamen, Fujian
361005,
China
bxi@xmu.edu.cn

***Department of Automation, Xiamen University
No.422, Siming South Road, Xiamen, Fujian,
361005,
China
wsx1009@163.com

****Department of Automation, Xiamen University
No.422, Siming South Road, Xiamen, Fujian
361005,
China
945366381@qq.com

*****Department of Automation, Xiamen University
No.422, Siming South Road, Xiamen, Fujian
361005,
China
360584748@qq.com

ABSTRACT. As the connection between human life and smart phones becomes close, mobile devices store a large amount of private information. Forensic personnel can obtain information related to criminal through mobile phones, but the traditional mobile phone forensics system is limited to a simple analysis of the original information and can't find the hidden relationship between data. This article will introduce a method based on K-means clustering and association rule mining to improve the traditional forensics system. Through the cluster analysis of basic information, the relationship between suspects and their contactors can be explored. At the same time, association rules mining can be

used to analyze the behavior of suspects so as to predict the time and contact of each event. Help law enforcement agencies find evidence hidden behind the data and improve the efficiency of handling cases.

KEYWORDS: mobile phone forensics system, K-means clustering, association rules, Apriori algorithm

1 Introduction

With the rapid development of communication technology, smart mobile devices have become the main communication tools of the society. Many criminal activities using smart devices have also been brought out. It is common to forge information and transmit viruses. At the same time the digital forensics technology changed from computer to mobile device, especially mobile digital forensics technology based on Android and IOS operating systems. As mobile phone forensics plays an increasingly important role in handling of cases, extracting effective information from devices is particularly vital for improving the efficiency of police's work. However, the traditional method of forensics has not been applied to the needs of mobile devices, mass data, and intelligent analysis. Currently the digital forensics research should focus on the analysis of smart mobile devices and mass data, and provide the judicial department with a complete package for evidence collection, evidence analysis, and evidence reporting. In modern forensics systems, more hidden evidence can be found on the existing basis through clustering and association rules mining. Clustering is an unsupervised learning. By classifying similar objects into similar classes, all objects are divided into several categories. Entities within a cluster are similar, and the entities of different clusters are not similar .A class cluster is a collection of test points in the space, the distance between any two points in the same cluster smaller than the distance of any two points between different clusters . Using this method, the person associated with the suspect can be distinguished and the scope of the investigation can be narrowed. Association rule mining is a means of discovering hidden relationships in data sets. It can help us to extract valuable related information between transactions. Through the analysis of these information, the suspect's characteristics and daily life habits can be portrayed, thus providing more clues for investigators.

2 The analysis of intimacy based on clustering

In the traditional mobile phone forensics system, the user's basic information is simply listed, and the relationship between the contactors is hidden. In order to improve this, the clustering algorithm can be applied to the forensic work. So the forensic system becomes more effective.

2.1 Traditional clustering method

K-means clustering algorithm is one of the most classic and widely used clustering algorithm. It divided data objects into clusters based on similarity. The concert steps of the algorithm are as follows:

– Select initial cluster centroids from data sets randomly;

– Calculate the distance of point to the centre of each cluster, and assign it to the nearest one;

– Re-calculate the centre of every cluster, if no one has changed then go to step four, otherwise go to step two;

– Get the results of clustering.

The result of this algorithm is mainly dependent on the initial cluster centroid, so the choice of the initial cluster centroid may have a great impact on the final result. Once the initial centre is chosen improperly, it maybe lead to an incorrect result. At the same time, due to the existence of isolated points, it will affect the calculation of the cluster centre after completion of the iteration. If choosing noise data far from the data-intensive area, it will cause the formed cluster centre to deviate from the real data-intensive area and reduce the accuracy of clustering.

2.2 An improved k-means algorithm

2.2.1. The choice of initial cluster centre

The traditional algorithm does not consider the distribution characteristics of the sample when selecting the initial point, so there will be a large deviation in the random selection, which leads to the clustering results do not meet expectations. In this paper, we will determine the classification of the categories and the choice of the initial centre point according to the sample distribution characteristics. As shown in the Figure 1, It's obvious that K=3, and then calculate the average of these three cluster.

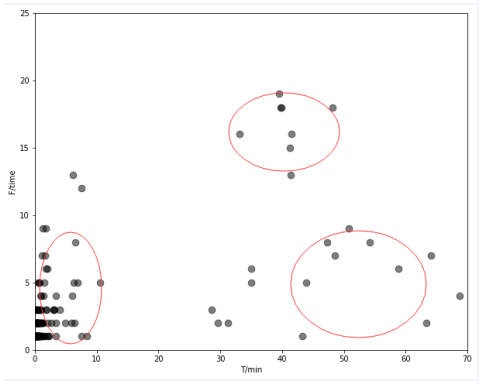

Figure 1. *The initial number of clustering*

2.2.2. The pre-processing of the isolated points

Due to the sensitivity of K-means to isolated points, the number of data in different cluster may vary greatly, and it's one of factors that affect the final results. To solve the problem, this paper proposes a method based on KNN algorithm to reduce the impact of isolated points. KNN algorithm is used to detect isolated points as shown in Figure 2. For numerical data, it is a relatively common algorithm. However, in large-scale data sets, it has a large amount of calculations and high algorithm complexity.

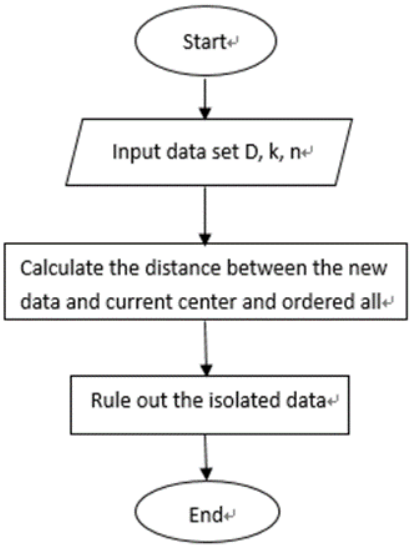

Figure 2. *The detection of the isolated points*

3 An improved association analysis algorithm

Association rule mining is a kind of data mining technology. It can discover the relationship between items or attributes in a data set. These relationships are unknown in advance and cannot be directly derived from the operation of the database. In general, correlation analysis is used to find meaningful relationships hidden in large data sets. The links found can be expressed in the form of association rules or frequent item-sets. These rules can be used to assist people in market operations, decision support, and business management.

3.1 The basic concepts of association rule mining

Suppose that D={d1,d2...dn} is a set of all transactions, N represents the total number of transactions, and I={i1,i2,...im} is a set of all items in the data set, A and B are a subset of I, and λ (A)=|{ di| A⊆ di ,di∈D}| indicates the number of transactions that include A, A→B support:

$$s(A \rightarrow B) = \frac{\lambda(A \cup B)}{N} \qquad [1]$$

s(A→B) determines how often a rule can be used for a given dataset. If s(A→B) is greater than the initially set threshold, (A→B) is a frequent set.

Based on frequent sets can calculate the corresponding confidence:

$$c(A \rightarrow B) = \frac{\lambda(A \cup B)}{\lambda(A)} \qquad [2]$$

c(A→B) determines the reliability of the rule. If c(A→B) meets the minimum confidence requirement, the rule is considered to be reliable.

Based on the above analysis, association rule mining can be divided into two steps:

– Generate frequent item sets that meet the minimum support requirements;

– Extract the rules that satisfy the minimum confidence requirement from frequent item-sets.

3.2 Apriori algorithm and its improvement

Apriori algorithm is an algorithm used for association mining. The principle is: If an item-set is frequent, all its subsets must be frequent. The Apriori algorithm uses a layer-by-layer iterative method to generate frequent sets. Starting from the 1-item set, pruning is based on the support degree to find a frequent 1-item set L1. A candidate 2-item set C2 is generated from L1, and then C2 is pruned by support to obtain a frequent 2-item set L2. In sequence, C3 is obtained from L2, and L3 is obtained from C3 pruning until no new frequent item-sets are generated. After that, rule extraction is performed for all frequent item-sets one by one: Initially, the extraction rule contains

only one item that satisfies the requirement for confidence, and then uses a back piece rule to generate two back items... Until Lk items are created. However when a candidate K-item set is first generated from a frequent (K-1)-entry set, a (K-1)-item set pairwise join method is used, so the number of candidate set entries grows exponentially, and this part of the data will taking up a lot of memory affects the performance of the machine.

Secondly, for the newly generated candidate K-item set, it is necessary to re-scan the database to count its support count, so as to determine the frequent K-item sets. Assume that the number of transactions in the data set is N and the number of candidate K-item sets is M. For frequent K-item sets, it is necessary to scan the MN database, which will consume a lot of time during the entire rule generation process, which will seriously affect the efficiency of the algorithm. In view of the above deficiencies, this article adopts an improved Apriori algorithm to improve the mining efficiency, uses a vertical structure to represent database data, marks each item contained in the transaction with 1 and uncontained items with 0, thus the original transaction database. Table 1 is converted to Boolean matrix Table 2.

TID	Item_sets
1	I1,I2,I5
2	I2,I4
3	I1,I3
4	I3,I5
5	I1,I2,I3,I4

Table 1. *Transaction data*

ID	I1	I2	I3	I4	I5	Length
1	1	1	0	0	1	3
2	0	1	0	1	0	2
3	1	0	1	0	0	2
4	0	0	1	0	1	2
5	1	1	1	1	0	4
	3	3	3	2	2	

Table 2. *Boolean data*

4 Experiment and results

4.1 The results of the intimacy analysis

The improved K-means algorithm is used to cluster the call records in the forensics system. After clustering, the results of classifying some clusters based on experience are as follows:

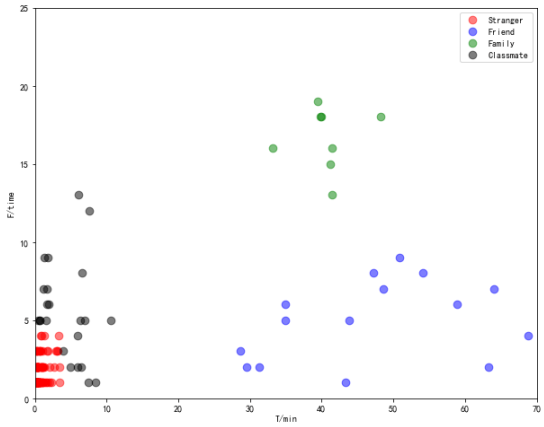

Figure 3. *The result of clustering*

The abscissa indicates the call duration, the ordinate indicates the number of calls, the red dots indicate strangers, the black dots indicate classmates, the green dots indicate family members, and the blue dots indicate friend relationships.

After comparing the experimental results with the real situation, it was found that the accuracy of the clustering of family and friends is high, and the accuracy of the relationship between strangers and classmates is relatively low.

4.2 The results of association rule mining

In the mining of association rules, the chat records of social software are used as the original data set, and the improved Apriori algorithm is used to extract the rules that satisfy the confidence requirement.

The total number of chat records in this experiment is 3541. Each record contains 5 attributes: app_name, account, friend_account, friend_nickname, and send_time. The total number of mining rules is 93. Table 3 shows the filtered rules.

No	Rule	Confidence
1	21--->WeChat	0.957
2	veter**--->QQ	0.756

3	Adam**、silv**--->10	0.960
4	QQ、363****56、16--->283****06	0.718
5	WeChat、Adam**、Xia**--->18	0.704
…	…	…

Table 3. *Association rules*

– The first rule indicates that users usually use WeChat to contact people around 9:00 pm;

– The second rule indicates that users and nicknames veter** are usually contacted by QQ;

– The third rule shows that Adam** and Silv** generally communicate during the day;

– The fourth rule indicates that users with an account number of 363****56 generally use QQ in the afternoon and users with an account number of 283****06 to contact.

– The fifth rule states that the user whose account is Adam** normally uses WeChat in the afternoon to contact the user whose account is Xia**.

Through the excavation of rules, criminal investigators can better understand the suspect's daily life habits and characteristics, so as to discover important clues hidden in the mobile phone, speed up the process of handling cases.

5 Conclusions

In this paper, the improved K-means algorithm and Apriori algorithm are applied to the mobile forensics system, thus making up for the insufficiency of traditional forensics equipment for data association analysis. According to the intimacy, different contactors are clustered, and then the clustering results are properly classified to obtain the relationship between different people. In the mining of association rules, high-confidence rules reflect the user's daily habits and characteristics, while low confidence indicates that the relationship between transactions is not tight, which means that there may be abnormal activities, so these two aspects All need attention. The new forensic system can provide judicial personnel with more valuable information, help them to further understand the suspects and improve the efficiency of handling cases.

References

1. Li Zhi. Design and implementation of a mobile digital forensic system. Diss. Xiamen University, 2017.
2. Yang Zeming, Liu Baoxu, and Xu Yusheng. "*Research Status and Development Trend of Digital Forensics.*" Scientific Research Information Technology and Application 6.1 (2015): 3-11.
3. Pollitt, Mark. "*A History of Digital Forensics.*" Advances in Digital Forensics VI - Sixth IFIP WG 11.9 International Conference on Digital Forensics, Hong Kong, China, January 4-6, 2010, Revised Selected Papers DBLP, 2010:3-15.
4. Gao Chuankai. "*Analysis of Computer Forensics Technology.*" Information Systems Engineering 2 (2017): 19-19.
5. Li, Jun Tao, Y. H. Liu, and Y. Hao. "*The improvement and application of a K-means clustering algorithm.*" IEEE International Conference on Cloud Computing and Big Data Analysis IEEE, 2016:93-96.
6. Zhang, Tao. *Association Rules*. Knowledge Discovery and Data Mining. Current Issues and New Applications. Springer Berlin Heidelberg, 2000:245-256.
7. Yu H, Wen J, Wang H, et al. *An Improved Apriori Algorithm Based On the Boolean Matrix and Hadoop*. Procedia Engineering, 2011, 15(1):1827-1831.
8. Zhang, Tian, C. Yin, and L. Pan. "*Improved clustering and association rules mining for university student course scores.*" International Conference on Intelligent Systems and Knowledge Engineering 2017:1-6.
9. Chen, Xiukuan, et al. *Application of Non-redundant Association Rules in University Library*. Advances in Computational Intelligence. Springer Berlin Heidelberg, 2009:423-431.
10. Chen, Zhimin, W. Song, and L. Liu. "*The application of association rules and interestingness in course selection system.*" IEEE, International Conference on Big Data Analysis IEEE, 2017:612-616.
11. Raj, K. Antony Arokia Durai, and P. Padma. "*Application of Association Rule Mining: A case study on team India.*" International Conference on Computer Communication and Informatics IEEE, 2013:1-6.
12. Wang, Hua, P. Liu, and H. Li. "*Application of improved association rule algorithm in the courses management.*" IEEE International Conference on Software Engineering and Service Science IEEE, 2014:804-807.
13.