2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery

DBSCAN clustering algorithm applied to identify suspicious financial transactions

Yan Yang*
School of Information science and engineering
Lanzhou University
Lanzhou, China
yyang2012@lzu.edu.cn

Bin Lian*
Anhui Economic Information Center
Hefei, China
lianb123@foxmail.com

Lian Li ,Chen Chen, Pu Li
School of Information science and engineering
Lanzhou University
Lanzhou, China
lilian@hfut.edu.cn ,{Chchen13,lipu12 }@lzu.edu.cn

***Abstract*—Money laundering refers to disguise or conceal the source and nature of variety ill-gotten gains, to make it legalization. In this paper, we design and implement the anti-money laundering regulatory application system(AMLRAS), which can not only automate sorting and counting the money laundering cases in comprehension and details, but also collect ,analyses and count the large cash transactions. We also adopt data mining techniques DBSCAN clustering algorithm to identify suspicious financial transactions, while using link analysis (LA) to mark the suspicious level. The presumptive approach is tested on large cash transaction data which is provided by a bank where AMLRAS has already been applied. The result proves that this method is automatable to detect suspicious financial transaction cases from mass financial data, which is helpful to prevent money laundering from occurring.**

Keywords—Money laundering; AML regulatory application system; DBSCAN clustering algorithm; Link Analysis (LA)

I. INTRODUCTION

Money laundering refers to disguise or conceal the source and nature of ill-gotten gains which are obtained by drug crimes, Mafia crimes, smuggling, terrorism, corruption bribery and other predicate offenses, then making it legalization by a variety of ways [1]. As money laundering is connected with a variety of predicate offenses, it is detrimental to the stability of country's political and economy, and even a threat to the security of international community.

In developed countries, anti-money laundering was carried out early. In 1970, the United States passed "Bank Secrecy Act" ,which reformed the traditional banking secrecy and established the basis for U.S. anti-money laundering regime[2]. The anti-money laundering information system which used currently in U.S.is FAIS (FinCEN Artificial Intelligence Systerm), was put into the use in 1993 and worked well in reporting the suspicious financial transactions[3]. The ScreenIT anti-money laundering system established by Australia's anti-money laundering transaction analysis and reporting centers is a transaction reporting screening system ,which can filter out highly suspicious transactions and then work for tax , police and other law enforcement agencies [3].

Compared to developed countries, anti-money laundering work started late in China, facing a complicated situation. In tradition, identifying suspicious financial transactions is completely manual, which is recognized by reading and analyzing thousands of textual documents to predict who is related to money laundering. So it is highly desirable to make this work automatable, and much work have been done on this issue. Among them, the authors Chenghu Zhang, Shi Li[4] proposed anti-money laundering system based on AI technology, which integrated a variety of technologies, including expert system, knowledge acquisition, artificial intelligence, statistical models, clustering, and visualization techniques to develop the computer-aided system supporting to detect suspicious financial transactions. The authors, Qi Chen, Ying-an Cui, Du-wu Cui[5], proposed anti-money laundering system based on Multi-Agent customer recognition, which introduces the pattern recognition to support identify anomalous. Author DeBinTan and Zao Chen [6] proposed applying data mining techniques in People's Bank regulatory systems to make sure its function of regulatory performs well.

This paper describes anti-money laundering regulatory application system and applies DBSCAN(Density-based spatial clustering of applications with noise) clustering algorithm to identify suspicious financial transactions. It is automatable to collect, analyze and count the large cash transactions and suspicious cash transactions, and clustering algorithm effectively reduces the prosecution time to detect anomalous.

The rest of the paper is organized as follows: section 2 is mainly about anti-money laundering regulatory application system. Section 3 explains the technology to identify outliers and section 4 describes the proposed approach in our work, while experiment and results are discussed in section 5. Section 6 concludes the paper and provides the future research direction.

II. AML REGULATORY APPLICATION SYSTEM(AMLRAS)

AML regulatory application System is aimed to support the work of bank to detecting suspicious financial transactions.

978-1-4799-6236-5/14 $31.00 © 2014 IEEE
DOI 10.1109/CyberC.2014.89

IEEE computer society

The system is able to effectively take advantage of large number of money laundering cases and show those information, while it can also automate to collect, analysis large cash transactions and then statistic the accounts which exceed some threshold.

A. Architecture of the system

The architecture of the system is shown as follow:

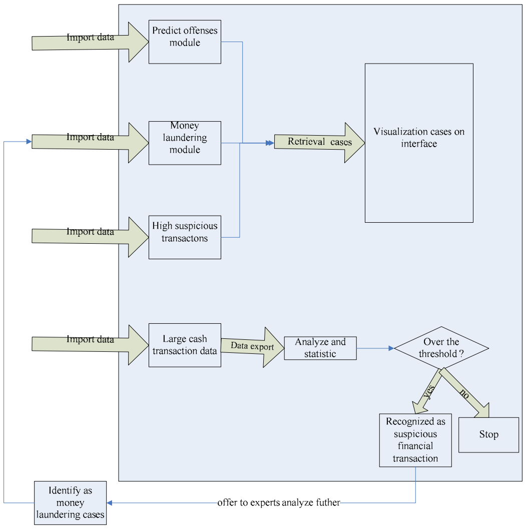

Import data
Predict offenses module
Import data
Money laundering module
Retrieval cases
Visualization cases on interface
Import data
High suspicious transactons
Import data
Large cash transaction data
Data export
Analyze and statistic
Over the threshold ?
yes
no
Recognized as suspicious financial transaction
Stop
Identify as money laundering cases
offer to experts analyze futher

Fig. 1. The architecture of AML regulatory application system

The system contains two main modules: money laundering cases module and large cash transaction module,which are introduced as follow.

As for money laundering case module,we loaded a large number of money laundering cases into the AMLRAS system. Users can retrieval and do statistics on the cases based on the visualized results. This module is convenient to be understood and analyzed money laundering cases.

Then it comes to large cash transaction module: Normalized large cash transaction data is imported into this module's AMLRAS system. Statistic module can be used to analyze the data beyond the threshold which is defined by anti-money laundering experts, and then count results visualized on interface in chart or histogram. Those abnormal phenomena will be offered to experts to do further study.

B. Functions of the System

The main functions of the system are divided into several steps ,specific description coming as follows:

Collect, count and retrieval money laundering cases or high suspicious : a large number of money laundering cases are imported into the AMLRAS system.Users can retrieve and do statistic on those cases to mine useful information which will be offered to AML experts.

Import, count large cash transaction data: large cash transaction data is imported into the AMLRAS system, then work out those accounts which are exceeding some threshold , tag them as anomalous and illustrate those large cash transactions or suspicious transaction accounts by diagram.

III. TECHNOLOGIES TO IDENTIFY SUSPICIOUS FINANCIAL TRANSACTIONS

Identifying suspicious financial transactions is a distributed, dynamic, and complex work, for it requires the good cooperation of many bank staffs and anti-money laundering experts. Commercial banks' staffs collect large cash transaction data in the first place, and report it to senior regulators. Anti-money laundering experts will analyze the data to detect real money laundering cases once they receive the request .

The primary method to monitor suspicious financial transactions is mainly consist of following four steps: (1) Filtering out the frequency and amount beyond a threshold accounts from the spreadsheet (excel). (2) Using statistical methods to identify the suspicious data beyond predetermined threshold value [7]. (3) Using data mining techniques to detect suspicious financial transactions. (4) Using correlation analysis method [8] to analyze the potential link between different accounts in different time points, in order to detect money laundering transaction chains.

Data mining is an efficient method to extract useful information from the large accounts data, that's why we apply it to deal with our finical data. Generally, these methods contain coming parts: 1) classification algorithm, such as neural networks [9], Bayesian [10], which requires the knowledge of classes before it applied, so its application is limited. 2) Clustering algorithm, which partitioned the data by their attributes that perform similar in transaction [4], recognizing the noise data as suspicious (outlier).

However, these methods have some flaws while applied in the financial sector. The data extracted from spreadsheet (excel) usually is not really suspicious data, because anomalous situations don't happen so frequently that most work we do turns out to be a waste of time and labor. Statistical method is suitable for low-dimensional data, but its application in financial data, which is high dimension, is limited [7]. Classification is a supervised learning method [4], which requires the knowledge of classes before. But it is difficult to collect the past data of money laundering transactions. Correlation analysis methods require the data at each point [8], which limits its application.

Clustering, an unsupervised learning method based on techniques on grouping customers, performs the similar kind of transactions into a single cluster and categorizes small-size clusters as anomalous or outliers. Clustering method can be classified as four categories: division based, distance based, density based and grid based[11]. The density based on method can detect random shape cluster, declaring areas with less data points as anomalous. As the data set used in this paper is records of personal large cash transactions in 2013, it is not considered as public data and has already been imported into AMLRAS system. It is difficult to obtain past data of transactions related to money laundering. So the density based clustering method was applied in this paper to detect suspicious financial transactions.

Density-based clustering method declares that the high-density area is divided by low-density area, which is regarded as anomalous. This paper presents density-based clustering algorithm DBSCAN to classify large data of cash transactions, to detect the noise points which are recognized as suspicious transactions, and link analysis (LA) was used to mark the suspicious level. DBSCAN clustering algorithm is a density set clustering method, which can cluster the dataset into all kinds of irregular shape. So it is effective to detect the noise data point. The approach was tested on large cash transaction data set. The marked suspicious accounts were submitted to anti-money laundering experts and it is helpful to prevent money laundering.

IV. Proposed approach

This section explains the presented method to identify anomalous transaction. It is divided into three steps: (1)Data pre-processing; (2) DBSCAN clustering algorithm; (3) Link analysis(LA). Each of three steps is explained below.

A. Data pre-processing

Data pre-processing includes two parts: 1) normalize the data format and enter those data into AML regulatory application system. 2) extract key attributes according to data pattern and personal experience.

B. DBSCAN algorithm

- DBSCAN (Density-based spatial clustering of applications with noise) is a density-based clustering algorithm, which finds a number of clusters starting from the estimated density distribution of corresponding nodes. Its cluster is random shape, which declares areas with less number of data points as anomalous. It defines the density as the min-points(Minpts) included in neighborhood of ε .
- Principle of DBSCAN algorithm

  DBSCAN requires two parameters: ε (Eps) and the minimum number of points required to form a cluster (MinPts). It starts with an arbitrary starting point that has not been visited. This point, ε -neighborhood, is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε -environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ε -neighborhood is also part of that cluster. Hence, all points that are found within the ε -neighborhood are added, consisting of their dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise. Table I shows the process of DBSCAN algorithm.

TABLE I. DBSCAN ALGORITHM

```
• Input: a database containing n objects, the radius Eps, the minimum number MinPts;
•Output: all the generated clusters and noises.
•0:Repeat
•1: draw an unprocessed point from the database;
•2: IF extracted point is a core point THEN find all reachable from the point density objects form a cluster;
•3: ELSE points are extracted edge points (non-core object), out of this cycle, to find the next point;
• 4: UNTIL all points have been processed.
```

By reasonable select parameter Eps and MinPts, using DBSCAN algorithm to analysis financial transaction data, noise point is recognized as anomalous.

C. Link Analysis (LA)

The extracted noise data points from DBSCAN clustering algorithm are temporarily considered as anomalous, and link analysis method is used to mark those data's suspicious level. The analyze steps are as follows:

- Set Pi = {Pi1, Pi2, Pi3, ..., Pin}   where i=1,2,3,…,k (1)   Pin represents the attribute1, attribute 2, attribute 3, ..., attribute n of i-th record.

The data set Pi is clustered by DBSCAN algorithm of each attribute, including attribute 1, attribute 2, attributes 3, attribute 4, ..., attribute n , then obtain a collection of noise data points set Zj, as follows,

- Zj = {Z1, Z2, Z3, ..., Zn} where j = 1,2,3,4, ...,n   (2)

  Zj represents the noise data set clustered by DBSCAN algorithm using every attribute.
- Link analysis:

  $$S1=Z1 \wedge Z2 \wedge Z3 \wedge Z4 \wedge \ldots \wedge Zn \qquad (3)$$

  S1 represents the first suspicious level transaction accounts, which are the intersection of noise data set obtained by DBSCAN clustering algorithm corresponding to n attributes.

  $$S2=Z1 \wedge Z2 \text{ or } Z3 \wedge Z4 \text{ or } Z1 \wedge Z3 \text{ or } Z2 \wedge Z4 \ldots \qquad (4)$$

  S2 represents the second suspicious level transaction accounts which are the intersection of noise data set corresponding to 2 attributes. Those transactions are identified as the second suspicious level transaction accounts. It is hypothesis that there have similar characteristic of attribute 1 and attribute 2, thus the intersection of data set Z1 and Z2 is extracted.

V. Design of Experiment and results

The above method is tested on real data sets. Our source data consists of 100,000 copies of large cash transaction data offered by a financial institution. Cash transactions have a close relationship with financial crime, such as financial fraud, money laundering, tax evasion, insider trading etc. Besides, the warning of the cash transaction risk by computer science is a blank in the management of financial agencies. Therefore, it is very important to detect suspicious financial transactions in large cash transactions. As the trading accounts data for person or for company has same format and large cash transaction plays important role in personal financial transaction, the large cash transactions of personal are used as experimental data set.

Each data set contains 9 attributes, which are No.,name, ID number, account, monthly deposit frequency, monthly deposit amount, monthly withdrawal frequency and monthly withdrawal amount, transaction net.

Each data set is identified as 9-dimensional feature vector X,

X={x1,x2,x3,…,x9},

Then 4 key attributes are selected to form the training sample Y,Y is the input data set of DBSCAN clustering algorithm.

Y={x5,x6,x7,x8}={y1,y2,y3,y4}

A. Data processing

First of all, we need to standardize the data format by removing unnecessary redundant information for import into the system. In large cash transactions, the transaction amount and frequency play an important role, so we choose monthly deposit frequency, monthly deposit amount, monthly withdrawal frequency and monthly withdrawal amount as key attributes for further clustering analysis.

B. Application of DBSCAN clustering algorithm

The selected four principal attributes monthly deposit frequency, monthly deposit amount, monthly withdrawal frequency and monthly withdrawal amount are labeled as 1,2,3,4. It is inevitable to set the parameter of cluster radius Eps and minimum number of cluster point MinPts in DBSCAN clustering algorithm , so we select the optimal minimum radius Eps and number of clusters points MinPts of every attributes as: attributes 1,2,4 select Eps = 20 and MinPts = 1000; attribute 3 Select Eps = 10 and MinPts = 1000. The testing results are shown in figure 3, where -1 represents the noise data points, 1 represents the core data points, 0 represents the boundary data points. Table Ⅱ shows the noise data points and the number of noise data clustered by DBSCAN algorithm. The number of noise data obtained by DBSCAN clustering algorithm of each attribute are 20, 51, 10 and 83.

TABLE II. THE NOISE DATA SET OBTAINED BY DBSCAN CLUSTERING ALGORITHM

attributes	**Noise data points**	**number of noise data**
monthly deposit frequency	Z1={3599 , 6509，9644，19234，35956，40718，44859，47091，67190，68607，68657，68863，72405，72429，89296，97207，97659，97825，98554，98996}	20
monthly deposit amount	Z2={2446，3593，5133，5509，6507，9585，12587，4015，16914，19133，25884，26022，26023，26626，29208，29988，32638，32655，32656，32657，32658，33970，35435，35813，36756，43031，44044，44859，46992，47879，53134，55288，55309，56637，57081，60032，60958，72390，72391，72392，73476，74066，81756，82054，84846，84847，86772，89988，91505，95480，99161}	51
Monthly withdrawal frequency	Z3={7084，54418，79686，88225，89296，90262，91442，91954，92293，99729}	10
monthly withdrawal amount	Z4={5509 ，8697，9309，9312，9585，12775，14015，16914，17832，18515，19602，21479，21531，21575，26022，26023，26371，26372，26625，26628，26670，27026，29208，30001，31412，31415，31652，33073，33499，33585，37093，37236，37544，39790， 42806，43031，46994，47030，47880，48196，49786，50528，50771，50810，50830，53044，53508，54418，55288，55309，55404，56530， 56612，56637，56826，57081，57133，57381，61657，61854，61855，63333，64577，66080，70668，72324，72369，72561，72562，74377，75288，75597，75730，75898，85559，87597，88101，90153，93665，94837，96313，96614，99148}	83

C. Application of link analysis(LA)

Once clusters of four attributes are obtained, the noise data points are divided into 4 sets Z1, Z2, Z3 and Z4. Link analysis(LA) is used to get intersection of each two noise data sets which have the same characteristic, on the condition of the intersection of data set Z1 and Z2, Z3 and Z4, Z1 and Z3 and Z2 and Z4 . Those appeared in both noise data sets are identified as second suspicious level transaction, while the intersection of four noise data points is recognized as first suspicious level financial transaction.

The test results as followings:

（1）Z1∧Z2={44859}

Z3∧Z4={54418}

Z1∧Z3={89296}

Z2∧Z4={5509 9585 14015 16914 26022 26023 29208 43031 55288 55309 56637 57081}

（2）Z1∧Z2∧Z3∧Z4={ ϕ }

It can be seen that there are 15 second suspicious level accounts, no first suspicious level account. The detailed information of 8 accounts are extracted for further analysis, table Ⅲ shows the details of 8 suspicious transaction accounts extracted by LA.

TABLE III. DETAILS OF EIGHT SUSPICIOUS TRANSACTION ACCOUNTS EXTRACTED BY LA

No.	account	deposit frequency	deposit amount	withdrawal frequency	withdrawal amount	Suspicious level
26022	*****423	19	1542	18	1287.1	2
26023	*****232	9	960	13	1567	2
29208	*****232	5	962	11	1233	2
43031	*****013	40	1139.7	33	1139.7	2
44859	*****898	344	2315 . 6	0	0	2
54418	*****055	0	0	170	844	2
55288	*****630	56	1558	89	1868	2
89296	*****054	1345	239.15	1216	112.36	2

It can be seen that the illustrated 8 suspicious financial transactions not only contain the high frequency and amount accounts, but also contain low frequency and amount accounts, such as no. 29208 and no. 54418 transaction accounts.

Then the extracted 15 suspicious financial transactions were submitted to corresponding local financial institutions to read their accounts details. The flow of reporting suspicious financial transactions is illustrated in fig.2.

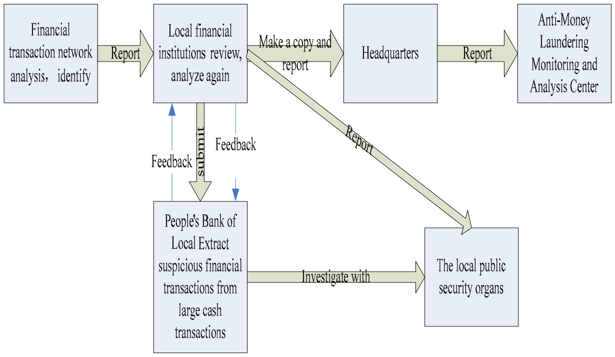

Financial transaction network analysis，identify
Report
Local financial institutions review, analyze again
Make a copy and report
Headquarters
Report
Anti-Money Laundering Monitoring and Analysis Center
submit
Feedback
Feedback
Report
People's Bank of Local Extract suspicious financial transactions from large cash transactions
Investigate with
The local public security organs

Fig. 2. Flowchart of reporting suspicious financial transaction

The suspicious accounts were observed continuously when every account is received by Local financial institutions. Then, 2 local commercial banks responded that there were 3 customers' identities inconsistent with their large cash transactions. Then their transaction details were sent to the local People's bank and experts in AML who analyzed the above trading accounts further, they concluded that some accounts went with deposits or withdrawals by large amount in a short period, which does not match the customers' economic status. Moreover, some accounts existing in duplicate accounts required to be monitored further. Among which one account had high deposit and withdraw frequency in continuous several months, and the transaction amounts were relatively larger, so it owner was recognized as high suspicious customer. Another customer deposited multiplied funds into one account by several times, then withdrew all of the cash at one time, thus this customer was recognized as high suspicious, too. The local public security organs will assist people's bank to investigate further if necessary.

The results prove that our method is not only effective to reduce the time of detecting suspicious financial transaction, but also useful to mine anomalies from mass large cash transactions data. It is contributed to prevent money laundering cases occur.

VI. CONCLUSION

Anti-money laundering is a complex and difficult task. It is uneasy to detect anomalous from mass financial transactions. This paper presents anti-money laundering regulatory application system to help collect and count suspicious financial transactions. Besides, the paper proposed an approach connected with the DBSCAN clustering algorithm and link analysis (LA) method. We apply the approach to actual large cash transaction data and obtain anomalous effectively. The result proved it works well. However, first level suspicious financial transaction is not founded in this experiment, by the reason that it only contains 100,000 large cash transaction which expand to 2 month. If the data sets expand to one year or more, it will get more effective results. In the further, we will aim to apply the method in our system to help prevent occurrence of the money laundering cases effectively.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Fo undation of Gansu Province of China under Grant No.145RJZ A126.

REFERENCES

[1] Reuter P, Truman E M. Chasing dirty money: The fight against money laundering[M]. Peterson Institute, 2004.

[2] Masciandaro D. Money laundering: the economics of regulation[J]. European Journal of Law and Economics, 1999, 7(3): 225-240.

[3] Tanzi V. Money laundering and the international financial system[R]. International Monetary Fund, 1996.

[4] Zhang ChengHu, Li Shi. Design AML system based on AI technology[J]. Financial Computer of China. 2005 (3): 44-47.

[5] Chen Qi, Cui Yingan, CUI Duwu. Based on Multi-Agent design AML Customer Identification System [J].

[6] Tan Debin, Chen Zao. Based on data mining technology design Bank's anti-money laundering systems [J]. Financial Computer of China. 2003 (7): 33-35.

[7] Larik A S, Haider S. Clustering based anomalous transaction reporting[J]. Procedia Computer Science, 2011, 3: 606-610

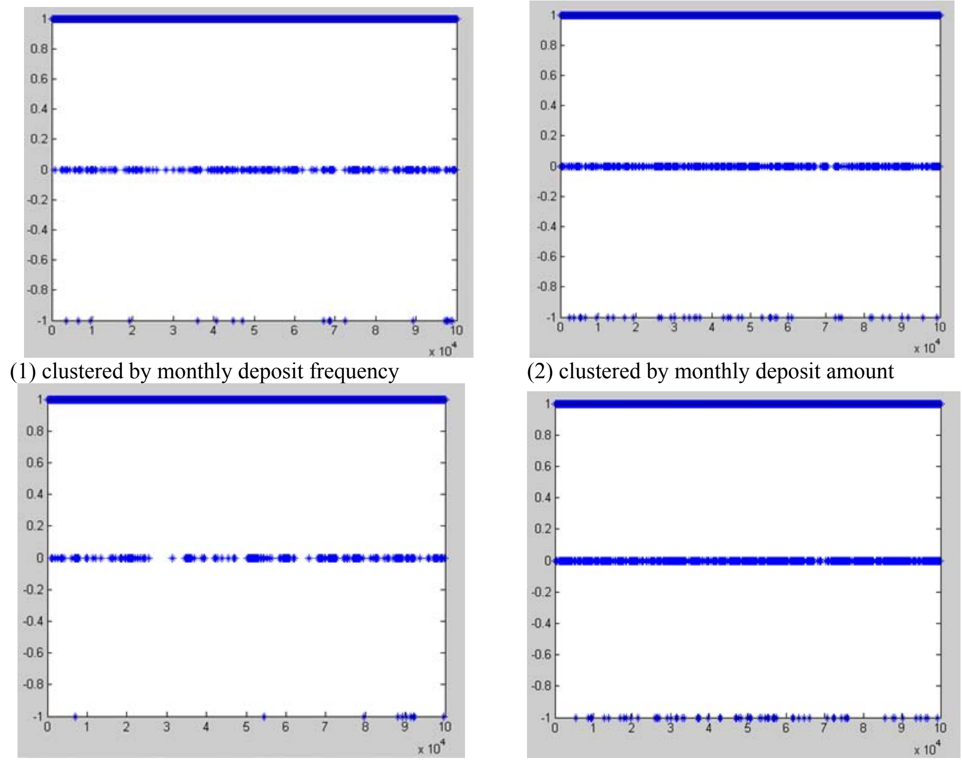
[8] Zhang Z M, Salerno J J, Yu P S. Applying data mining in investigating money laundering crimes[C]//Proceedings of the ninth ACM SIGKDD

international conference on Knowledge discovery and data mining. ACM, 2003: 747-752.

[9] Senator T E, Goldberg H G, Wooton J, et al. Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions[J]. AI magazine, 1995, 16(4): 21.

[10] Lv L T, Ji N, Zhang J L. A RBF neural network model for anti-money laundering[C]//Wavelet Analysis and Pattern Recognition, 2008. ICWAPR'08. International Conference on. IEEE, 2008, 1: 209-215.

[11] Han J, Kamber M, Pei J. Data mining: concepts and techniques[M]. Morgan kaufmann, 2006.



(1) clustered by monthly deposit frequency

(2) clustered by monthly deposit amount

(3) clustered by monthly withdrawal frequency

(4) clustered by monthly withdrawal amount

Fig. 3. clustered results of four attributes by DBSCAN clustering algorithm