Application of Cluster-Based Local Outlier Factor Algorithm in Anti-Money Laundering

Gao Zengan

Post Doctoral Station of Theoretical Economics
China Center for Anti-Money Laundering Studies
Fudan University
Shanghai, P. R. China

School of Economics and Management
Southwest Jiaotong University
Chengdu, P. R. China
E-mail address: gaozengan133@163.com

Abstract—**Financial institutions' capability in recognizing suspicious money laundering transactional behavioral patterns (SMLTBPs) is critical to anti-money laundering. Combining distance-based unsupervised clustering and local outlier detection, this paper designs a new cluster-based local outlier factor (CBLOF) algorithm to identify SMLTBPs and use authentic and synthetic data experimentally to test its applicability and effectiveness.**

Keywords-clustering; outlier detection; local outlier factor (LOF); suspicious money laundering transactional behavioral patterns (SMLTBPs); anti-money laundering (AML)

I. INTRODUCTION

Anti-money laundering (AML) in financial industry is based on the analysis and processing of Suspicious Activity Reports (SARs) filed by financial institutions (FIs), but the very large number of SARs usually makes financial intelligence units' (FIUs') analysis a waste of time and resources simply because only a few transactions are really suspicious in a given amount [1], so financial AML is far from a real-time, dynamic, and self-adaptable recognition of suspicious money laundering transactional behavioral patterns (SMLTBPs). Literature review finds that artificial intelligence [2], support vector machine (SVM) [3], outlier detection [4], and break-point analysis (BPA) [5] are used to improve FIs' ability in processing suspicious data, various approaches to novelty detection on time series data are examined in [6], outlier detection methodologies are surveyed by [7], and a data mining-based framework for AML research is proposed in [8] after a comprehensive comment is made on relative studies. But the effectiveness and efficiency of SMLTBP identification remains a hot spot for research since the passage of the USA Patriot Act and the creation of the U.S. Department of Homeland Security signaled a new era in applying information technology and data mining in detecting money laundering and terrorist financing [9].

As SMLTBP recognition is short of training data, the number of clusters is usually unknown, and the result of clustering is always changing dynamically [10, 11], this paper designs a cluster-based local outlier factor (CBLOF) algorithm to help FIUs concentrate on a desirable number of SMLTBPs having a proper degree of suspiciousness as determined by their actual needs and resources endowments. Following the introduction, Section II describes the design of the algorithm, Section III is about the experimental process, and Section IV ends the paper with a suggestion for future research.

II. ALGORITHM DESIGN

The CBLOF algorithm combines distance-based unsupervised clustering and local outlier [12] detection, and clustering is for the purpose of pre-processing data for the consequent anomaly identification.

A. Clustering

As far as the nature of money laundering (ML) is concerned, the chosen clustering algorithm should be able to generate the number of clusters automatically (with no need for pre-establishment) and all the clusters are to be ranked according to the number of the components in each. Thus we propose the following procedures:

1) Start with any object (say p) in a dataset and create a cluster. The initial cluster is supposed to be C_1.

2) Choose any other object q, calculate its distance to the existing clusters C_1, C_2, C_3, ..., C_i and denote it by $distance(q, C_i)$, and then figure out the minimal distance value $distance(q, C_{min})$.

3) Let the threshold be ε. If $distance(q, C_{min}) \le \varepsilon$ holds and "q has never been clustered" satisfies, add q to the cluster C_i which is assumed to be nearest to q when compared with all

The research is supported by the National Social Science Foundation of China (No. 08BGJ013).

978-1-4244-4639-1/09/$25.00 ©2009 IEEE

the other known clusters. Conversely, if $distance(q, C_{min}) > \varepsilon$, implying that q has not yet been clustered into any category, build a new cluster C_j and embed q into it. Nevertheless, suppose there is a cluster C_m outside C_i, if $distance(q, C_m) \le \varepsilon$, then integrate C_m and C_i into one cluster, that is, if we have $distance(q, C_m) \le \varepsilon$ and $distance(q, C_i) \le \varepsilon$ simultaneously, then clusters C_m and C_i are merged into one.

4) Repeat Steps 2) and 3) until all the objects have been clustered.

5) Rank all the clusters in the decreasing number of their components involved.

B. Outlier Detection

An outlier is a point that deviates so much from surrounding "normal" points as to arouse suspicion that it was generated by a different mechanism. After clustering, all the samples have been categorized into mutually exclusive clusters ranked as per the number of their components. As most transactions in an account are usually normal or legal, the clusters generated from above are divided into Large Category (LC) and Small Category (SC) in this paper, with the former being supposed to represent normal transactional behavioral patterns free of ML suspicion and the latter, on the contrary, for anomalous patterns worth notice.

For the clustering result of dataset D, let $C = \{c_1, c_2, \cdots, c_k\}$ and $|c_1| > |c_2| > \cdots |c_k|$. Given any two parameters α and β, we have:

$$|c_1| + |c_2| + \cdots + |c_b| \ge |D| * \alpha \quad (1)$$

$$C_b / C_{b+1} \ge \beta \quad (2)$$

where $c = \{c_i \,|\, i \le b\}$ for LC, that is, $LC = \{c_i \,|\, i \le b\}$ and $SC = \{c_j \,|\, j > b\}$. While (1) represents the majority of the objects in the dataset, (2) indicates that the number of LC components is greatly different from that of SC components.

Furthermore, the points in SC are all outliers when compared with those in LC [13, 14]. But for AML research, seasonal industries and some special industries must be exempted because abnormal phenomena in a particular period can never be treated as ML red flags. So the paper will study *n* number of data points with top local outlier factor (LOF) values because they are more of ML suspicion. Also, this can effectively improve AML pertinence.

In the light of the local outlier definition in [12], LOF can be employed to measure the deviant degree of SC points from LC, i.e., how far the transactional behavioral patterns represented by the points in SC deviate from the normal or legitimate patterns, where LOF value is determined by the number of the components in the clusters sample data belong to and the distance from sample data to the nearest LC.

Given a point *o* in the dataset, its LOF value is:

$$LOF(o) = |c_i| * min[distance(o, c_i)] \quad (3)$$

where $o \in c_i, c_i \in SC, c_j \in LC$. The higher the LOF value is, the farther the point *o* deviates from the normal transactional behavioral patterns. Once the LOF value is fixed for each object, we can get to know how suspicious the transactional behavioral patterns are in the given account. Rank the data points as per their LOF values, we can get a feature-oriented ordering of SMLTBPs to help FIs choose *n* number of objects as they like for a detailed exploration and finally determine what to file to FIUs under the restraints of investigation resources in labor, capital, and technology, etc.

III. EXPERIMENTAL PROCESS

The experimental process mainly includes extracting research variables, preparing data samples, actualize the algorithm, and discussing the experimental results.

A. Choose and Define the Features to Be Studied

We are more interested in the transactional behavioral attributes like amount and frequency than in the account owner's subjective characters, thus transaction amount, transaction amount deviation coefficiency, and transaction frequency (i.e., withdrawal frequency and deposit frequency) are chosen to be research variables with the following definitions:

Definition 1: Transaction amount (Ta_i) is the total amount of all the transactional segments or subsequences, that is,

$$Ta_i = \sum_{j=1}^{n} ta_{ij} \quad (4)$$

where, Ta_i is the transaction amount of the *i*th transactional segment of a given account, Ta_{ij} is the amount of the *j*th transaction in the *i*th segment (and so and so forth hereinafter). Transaction amount is a critical criterion for us to determine whether a transaction is suspicious or not since a large cash transaction is viewed as a special kind of suspicious transaction in this research.

Definition 2: Transaction amount deviation coefficiency (Tad_i) is the ratio between transaction amount variation (Ts_i^2) and average transaction amount ($\overline{Ta_i}$), that is,

$$Tad_i = Ts_i^2 / \overline{Ta_i} = \frac{n_i}{n_i - 1} \bullet \frac{\sum_{j=1}^{n_i} (ta_{ij} - \overline{Ta_i})^2}{\sum_{j=1}^{n} ta_{ij}} \quad (5)$$

Tad_i is used to measure the degree of equalization of transaction amount (i.e., structuring) which means a large cash transaction, either deposit or withdrawal, is purposefully divided into several transactions of a nearly equal amount in order to exempt from filing Currency Transaction Reports (CTRs) as required by the authority. The less the Tad_i value is, the more equalized the transaction amount is, and the more suspicious the transaction is as far as CTR regulations are concerned.

Definition 3: Withdrawal/deposit frequency is the ratio of

the number of withdrawal/deposit transfers to the aggregated frequency of transactions.

Denote the withdrawal frequency and the deposit frequency of the *i*th transactional segment of a given account by Tfw_i and Tfd_i, respectively, and denote the total frequency of transactions by *tf*, we have the following formulas:

$$\begin{aligned} tf &= tfw_i + tfd_i \\ Tfw_i &= tfw_i / tf \\ Tfd_i &= tfd_i / tf \end{aligned} \tag{6}$$

Analyzing withdrawal frequency and deposit frequency can identify two novel capital flows within a short time frame: one is centralized capital in-transfers followed by decentralized capital out-transfers, and the other is decentralized capital in-transfers followed by centralized out-transfers. Also, this analysis can compensate for the research of [5].

B. Prepare Data Samples

Just like the authors of [6], we are most interested in data patterns that deviate from the normal operational data. So historical transaction records are to be transformed into several segments or subsequences of neighboring single transactions, with one segment (subsequence) representing one behavioral pattern, and the transactional data embedded in SMLTBPs are just the suspicious objects we hope to find out.

For each feature as above mentioned, calculate its feature value for each segment and take the feature vectors composed of feature values as research samples.

In this research, we have collected from 108 accounts of one commercial bank 34,303 authentic transactional data from January 1 through October 30, 2006 out of which the account data of 25 firms in 4 industries sharing similar turnover scale and transactional frequency are taken as experimental samples. Meanwhile, twenty segments of synthetic data are generated by the mechanism in Figure 1 [15, 16] to test the applicability of the algorithm in detecting abnormal objects. Each segment of artificial simulation data is employed twice.

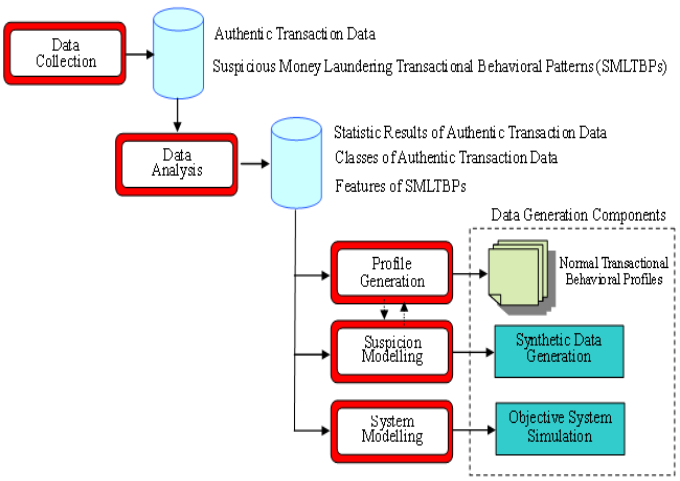

Figure 1. Synthetic Data Simulation Method

TABLE 1. SYNTHETIC DATA SAMPLES AFTER PRE-PROCESSING DATA

Location of Data (Account, Sample No.)	Ta_i	Tad_i	Tfw_i	Tfd_i
(1, 5), (3, 5)	0.411904	0.027473	0.133	0.867
(3, 29), (18, 5)	0.720696	0.763975	0.973	0.007
(2, 13), (19, 19)	0.512147	0.017695	0.87	0.13
(20, 10), (21, 17)	0.357504	0.010843	0.67	0.33
(20, 14), (21, 6)	1.302777	0.851729	0.53	0.47
(2, 3), (4, 17)	3.200864	0.884096	0.87	0.13
(6, 7), (7, 4)	2.958707	0.832994	0.067	0.977
(6, 13), (7, 5)	2.890457	0.828636	0.93	0.07
(12, 29), (13, 20)	2.983899	0.009847	0.73	0.27
(12, 11), (13, 22)	2.980798	0.002417	0.8	0.2
(5, 2), (8, 9)	3.822738	1.049567	0.47	0.53
(5, 20), (8, 12)	3.527972	1.004896	0.47	0.53
(9, 5), (14, 2)	3.557559	1.015748	1	0
(9, 29), (14, 20)	3.529352	1.049067	0.53	0.47
(15, 2), (22, 6)	3.538852	1.008545	0.53	0.47
(10, 2), (11, 14)	1.434711	0.444943	0.4	0.6
(10, 20), (11, 9)	1.424323	0.453174	0.933	0.067
(16, 12), (17, 20)	1.429347	0.435606	0.027	0.973
(16, 22), (17, 24)	1.429204	0.317638	0.47	0.53
(23, 7), (25, 6)	1.425056	0.235013	0.33	0.67

As per the Regulations for Financial Institutions to File Currency Transaction Reports and Suspicious Transaction Reports of the People's Bank of China, ten days is accepted as the standard to segment transactional data. After pre-processing the experimental data, segmenting the subsequences, and extracting the feature values, we obtain 696 experimental samples of 25 accounts, of which only 40 synthetic data samples are listed in Table 1 due to the limit of the paper.

C. Actualize the Algorithm

Do experiment with the CBLOF algorithm on the sample set. As global outlier detection cannot mine all the outliers [12], give LOF value to each sample, and then identify *n* number of samples with the highest LOF values for further investigation and final reporting.

D. Experimental Results and Discussions

Let clustering threshold $\varepsilon = 0.15$ and categorization parameters $\alpha = 75\%$ and $\beta = 4$, we will first of all standardize the dimensions of data samples, and then program with C++ language, cluster, categorize LC and SC, and compute LOF values of transaction segments. Once more only a part of the experimental results are shown in Table 2 due to the limit of the paper, where only the five samples with top LOF values are listed for each account. They are the five transactions with the highest degree of suspiciousness, as well.

TABLE 2. PART OF THE SMLTBP RECOGNITION EXPERIMENTAL RESULTS

A/C	Sample	LOF Value	A/C	Sample	LOF Value	A/C	Sample	LOF Value
1	(1, 17)	5.982922	5	(5, 2)	0.668178	9	(9, 29)	2.353695
	(1, 32)	0.464220		(5, 13)	0.439318		(9, 5)	1.88573
	(1, 1)	0.407418		(5, 20)	0.415933		(9, 19)	0.255343
	(1, 22)	0.392810		(5, 6)	0.300100		(9, 23)	0.245357
	(1, 7)	0.382421		(5, 23)	0.280535		(9, 14)	0.207123
2	(2, 13)	0.325883	6	(6, 7)	0.984965	10	(10, 2)	1.668178
	(2, 3)	0.313209		(6, 13)	0.873594		(10, 20)	1.415933
	(2, 20)	0.302324		(6, 27)	0.71211		(10, 13)	0.439318
	(2, 14)	0.234307		(6, 29)	0.397222		(10, 6)	0.3001
	(2, 18)	0.221585		(6, 8)	0.233948		(10, 23)	0.280535
3	(3, 5)	0.885730	7	(7, 5)	4.982922	11	(11, 9)	1.54074
	(3, 29)	0.353695		(7, 4)	4.529272		(11, 1)	0.521822
	(3, 19)	0.255343		(7, 1)	0.407442		(11, 2)	0.394264
	(3, 21)	0.245357		(7, 7)	0.382421		(11, 7)	0.238956
	(3, 14)	0.207123		(7, 6)	0.308618		(11, 15)	0.223159
4	(4, 17)	1.27074	8	(8, 9)	1.839707	12	(12, 29)	1.460348
	(4, 16)	0.25807		(8, 12)	0.325883		(12, 3)	0.379584
	(4, 11)	0.174356		(8, 3)	0.313209		(12, 1)	0.377692
	(4, 14)	0.158745		(8, 20)	0.302324		(12, 18)	0.358329
	(4, 26)	0.152971		(8, 25)	0.209284		(12, 4)	0.285378

What is more, all the anomalous samples in the 25 accounts are ranked in the decreasing order of their LOF values. Particularly, 33 of 40 artificial samples are found with relatively higher LOF values and they even rank the top five of their accounts, which proves the effectiveness of the CBLOF algorithm in identifying suspicious data.

IV. CONCLUSION

Making a good use of the advantages of both distance-based unsupervised clustering and local outlier detecting, the CBLOF algorithm can effectively identify the synthetic data suspicious of ML transactions with a high processing speed and a satisfactory accuracy. Needing neither prior samples to serve as training data nor the number of clusters to be designated in advance can solve the problem that AML research is always in short of case data. In particular, the algorithm is self-adaptable to the evolution of ML methods and can recognize SMLTBPs that haven't been detected before, which is quite beneficial in saving limited investigation resources and preventing FIs from filing defensive SARs [17].

However, only a few transactional behavioral features of amount and frequency are studied in this paper, so relative subjective characters of the account owner remains open to our future research.

REFERENCES

[1] H. G. Goldberg and R. W. H. Wong. "Restructuring transactional data for link analysis in the FinCEN AI System". In: Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis, Menlo Park, C. A.: AAAI Press, 1998.

[2] H. G. Goldberg and T. E. Senator. "Restructuring database for knowledge discovery by consolidation and link formation". In: Proceedings of the First International Conference on Knowledge Discovery in Database (KDD-95), Menlo Park, Calif.: AAAI Press, 1995: 136-141.

[3] J. Tang and J. Yin. "Developing an intelligent data discriminating system of antimoney laundering based on SVM". In: Proc. of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 2005: 3453-3457.

[4] R. C. Watkins, K. M. Reynolds, R. Demara, M. Georgiopoulos, A. Gonzalez, and R. Eaglin. "Tracking dirty proceeds: exploring data mining technologies as tools to investigate money laundering". Police Practice and Research, 2003, 4 (2): 163-178.

[5] R. J. Bolton and D. J. Hand. "Unsupervised profiling methods for fraud detection". Conference on Credit Scoring and Credit Control VII, Edinburgh, UK, 2001: 5-7.

[6] J. Anstey, D. Peters, and C. Dawson. "Discovering novelty in time series data". In: Proceedings of the 15th Annual Newfoundland Electrical and Computer Engineering Conference, Canada, 2005.

[7] V. Hodge and J. Austin, "A survey of outlier detection methodologies." Artif. Intell. Rev., 2004, 22 (2): 85-126.

[8] Z. Gao and M. Ye. "A framework for data mining-based anti-money laundering research". Journal of Money Laundering Control, 2007, 10 (2): 170-179.

[9] J. S. Zdanowicz. "Detecting money laundering and terrorist financing via data mining". Communications of the ACM, 2004, 47 (5): 53–55.

[10] J. Han, Y. Huang, and N. Cercone. "Data mining: an overview from database perspective". IEEE Transaction on Knowledge and Data Engineering, 1996, 8 (6): 1-40.

[11] J. Han and M. Kamber. Data Mining: Concepts and Technique. Morgan Kaufmann Publishers, 2001.

[12] M. M. Breuning, H. P. Kriegel, T. Ng. Raymond, and J. Sander. "LOF: identifying density-based local outliers". In: Proc. ACM SIGMOD 2000. Int. Conf. on Management of Data. Dallas, TX, 2000: 93-104.

[13] Z. He, X. Xu, and S. Deng. "Discovering cluster-based local outliers". Pattern Recognition Letters. 2003, 24 (9-10): 1641-1650.

[14] M. F. Jiang, S. S. Tseng, and C. M. Su. "Two-phase clustering process for outliers detection". Pattern Cognition Letters, 2001, 22: 691-700.

[15] E. Lundin, H. Kvarnström, and E. A. Jonsson. "Synthetic fraud data generation methodology". In: Lecture Note in Computer Science, ICICS2002. Laboratories for Information Technology, Singapore, Springer Verlag, 2002: 265-277.

[16] E. Barse, H. Kvarnström, and E. Jonsson. "Synthesizing, test data for fraud detection systems". In: Proceedings of the 19th Annual Computer Security Applications Conference. 2003: 384-395.

[17] Z. Gao. "Comments on anti-money laundering suspicious activity report regime". In: Proceedings of 2006 International Conference on Public Administration. X. Zhu and S. Zhao, Eds. Chengdu: University of Electronic Science and Technology of China Press, 2006: 970–975.