INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

A Probabilistic Model Using Graph Based Sequential Pattern Mining Algorithm For Money Laundering Identification

G.Krishnapriya
Research Scholar, Bharathidasan University
Trichy, India
Dr. M. Prabakaran
Assistant Professor, Department Of Computer Science Government Arts College
Ariyalur, Tamil Nadu, India.

Abstract:
Money laundering an activity, which hides the source and origin of money in any banking or financial account of a country. The countries financial stability or financial growth depends on the overall amount in the banks and finance organizations. The money holds the financial stability of any country and leads to changes the countries money value in international market. In past decades the criminals started hiding the source of income and the source of money from where the amount transferred to the finance account, which is illegal towards the financial rule of any country. This causes the threat to the financial stability of the country, because those amounts can be washed at any point of time to some other country. Also terrorist and criminals makes this kind of transactions to finance their clients to encourage terrorism. We focus on identifying the behavior of transactions and account holder in managing their accounts. There has been various methodologies proposed using data mining techniques, but suffers to identify the origin of money. We propose a graph based sequential pattern mining technique and probability model to identify the account from where the transaction originated. We generate a graph with many numbers of nodes and vertices for each account using the transactional data set, and using the account graph we generate sequential patterns and transition paths. Using sequential patterns and transition paths we compute the probability model to identify the origin of amount.

Key words: *AML, ML, Pattern mining, data mining, IE*

1.Introduction

Money Laundering (ML) has become a symbol of the capability of the central government and the faith of the nation, as well as an important measure fighting organized crimes and forestalling the flooding of nation-crossed corruptions. However, the construction of an effective AML mechanism is just at its startup which is far from perfect. Most of relevant technical tools, the developing and applications of monitoring and analyzing system are nearly blank, while legislation work is relative coarse and primitive, given the opportunities for the launders to evade the detections. Money laundering threatens the economic and social development of countries. The threat is due to the injection of illegal proceeds into the legitimate financial system. Due to the high amount of transactions and the variety of money laundering tricks and techniques, it is difficult for the authorities to detect money laundering and prosecute the wrongdoers. Thus, it is not only the amount of transactions, but the ever changing characteristics of the methods used to launder money that are constantly being modified by the fraudsters, which makes this problem interesting to study.

Money laundering (ML) is a process of disguising the illicit origin of "dirty" money and makes them appear legitimate. It has been defined by Genzman as an activity that "knowingly engage in a financial transaction with the proceeds of some unlawful activity with the intent of promoting or carrying on that unlawful activity or to conceal or disguise the nature location, source, ownership, or control of these proceeds. Through money laundering, criminals try to convert monetary proceeds derived from illicit activities into “clean” funds using a legal medium such as large investment or pension funds hosted in retail or investment banks. This type of criminal activity is getting more and more sophisticated and seems to have moved from the cliché of drug trafficking to financing terrorism and surely not forgetting personal gain. Today, ML is the third largest “Business” in the world after Currency Exchange and Auto Industry. According to the United Nations Office on Drug and Crime, worldwide value of laundered money in a year ranges from $500 billion to $1 trillion and from this approximately $400-450 Billion is associated with drug trafficking. These figures are at times modest and are partially fabricated using statistical models, as no one exactly knows the true value of money laundering, one can only forecast according to the fraud that has already been exposed. Nowadays, it poses a serious threat not only to financial institutions but also to the nations. Some risks faced by financial institutions can be listed as reputation risk, operational risk, concentration risk and legal risk. At the society level, ML could provide the fuel for drug dealers, terrorists, arms dealers and other criminals to operate and expand their criminal enterprises. Hence, the governments, financial regulators require financial institutions to implement processes and procedures to prevent/detect money laundering as well as the financing of

terrorism and other illicit activities that money launderers are involved in. Therefore, anti-money laundering (AML) is of critical significance to national financial stability and international security.

Data mining applications are deployed in a wide range of business fields, especially in financial banking, telecommunication, and the World Wide Web that have to deal with extensive amount of data. Simple database querying is far from enough for information retrieval in those business areas. Data mining is used to extract more complex desired information. The desired information is usually presented as a pattern. Thus pattern recognition, although not equivalent to data mining, is usually the framework for data mining.

There are mainly two approaches to mining relational information: logic-based approaches and graph-based approaches. Logic-based approaches involve logical definition of a data pattern which is usually more complicated than that of graph-based approaches. They allow recursions and variables in defining a data pattern, which are not easy to implement with graph-based approaches. With certain limitations, graph-based data mining is however more data-driven.

Banks collect detailed personal information from their clients as well as information such as the relations between clients and the association between clients and certain companies. Data mining systems are used to extract interesting and valuable information from the large database. Relations among clients and companies, accompanied with monetary transaction histories, could be used as effective indication of suspicious money laundering activities.

First and foremost, we have to define a pattern in a visualized and human-readable way in the front end. The pattern is to be defined by an end user, therefore, the representation of a pattern need to be intuitive but also expressive. It has to be intuitive enough so as to make it easier for users to convey their ideas with the pattern. And it has to be expressive enough so that whatever the users want to express can be represented by the pattern. We use graphs to represent patterns. Graphs are expressive in describing relations between objects, and it is easy to understand as well. The pattern information encapsulated by the graph would be sent to the backend and get processed and interpreted. How the patterns are interpreted and processed is the core part of the backend. Finally, the backend would return the whole set of data that matches the user-defined pattern.

2.Related Works

A lot of research work has been done on graph data mining. Graph data mining is the task of finding novel, useful, and understandable patterns in a graph representation of data. In a lot of works, graph data mining is used for finding frequently occurring structures, such as in molecular biology, people are interested in finding certain structures comprising of some elements. Diane and Lawrence et al. developed the Subdue system to find frequent patterns. Subdue system is the process of incrementally compressing frequently occurring substructures into units until reaching the pattern occurring frequency. Mohammed J. Zaki proposed a Tree Miner algorithm for finding frequent structures [5]. This algorithm performs DFS to find frequent sub trees, and this algorithm uses strings to encode the tree structures.

Active Learning via Sequential Design with Applications to Detection of Money Laundering [6] , is proposed an active learning method using Bayesian sequential designs to identify the suspicious accounts. The method uses a combination of stochastic approximation and D-optimal designs to judiciously select the accounts for investigation. The sequential nature of the method helps to identify the suspicious accounts with minimal time and effort.

A framework on developing an intelligent discriminating system of anti-money laundering [8], proposes a four layer model to identify money laundering. Different layers play different roles during the analyzing procedure. Data of Transaction layer and Account Layer are submitted from the root bank branches and have composed the fundamental sources. Only isolated intelligence may be derived from the perspectives of both inner layers. Organization layer and Link layer provide perspectives to take a comprehensive and aggregate discriminating and analyzing procedure to all data involved by multiple banks, areas and departments, to check, contrast, mine, judge and derive in all those data collected from varied channels. The later layers have much more advantages during macro situation judgment and relevant cases investigation.

Money Laundering Detection using Synthetic Data [9], they present an analysis of the difficulties and considerations of applying machine learning techniques to this problem. We discuss the pros and cons of using synthetic data and problems and advantages inherent in the generation of such a data set. They using a case study and suggest an approach based on Multi-Agent Based Simulations (MABS).

[10] ,Anti-Money Laundering: enhancing effectiveness with anomaly detection, rules, and link analysis is proposed. It uses Association rules approaches to address these issues. Instead of relying on a set of externally-defined strictures, Association rules methods use patterns and associations in the data to create the rules. Using a dataset of millions of transactions, the models mine transaction data to find signals – the patterns that occur frequently for true cases of money laundering, but are rarely present for legitimate customer transactions. The "rules" are based on these patterns of associations, whether it is a pattern involving a single transaction or a group of related transactions, or a single person, or a group of people. What's more, the modeling techniques can assign an "incursion probability" to different patterns. This enables cases to be more effectively prioritized, saving effort and more accurately identifying true criminal cases. The models train on historical Suspicious Activity Reports (SARs). First, the process applies an algorithm to efficiently find large numbers of rules that might define suspicious activity. Then, the model systematically prunes and refines this broad set of rules to arrive at a set of highly focused guidelines that provide maximum separation between true cases of money laundering and normal transaction activity. Once these data-driven rules have been identified and implemented, they can also be combined with an anomaly detection model for even more accurate detection performance.

Anomaly detection is a method to find "unknown" patterns, more fully protecting financial institutions from current and future money laundering attempts. Anomaly detection uses sophisticated adaptive models to look through transactions, spotting unusual activities. Anomaly detection models analyze customer behavior and transaction behavior, distinguishing normal patterns from abnormalities that may indicate a high-risk activity. Capable of detecting all types of money laundering activities, an anomaly

detection system can keep up with the changing face of this class of crime by discovering new classes of fraud and other suspicious activities just as they emerge.

The third lever is link analysis, which identifies and measures how closely an account is connected with known suspicious accounts. It looks at the flows and connections of transactions to spot suspicious patterns and questionable connections. For example, money that originates at one account, travels through several other accounts, and converges on a single endpoint is more likely to indicate money laundering activity. Often these accounts share common characteristics, such as phone numbers or work addresses. Using link analysis, even subtle connections between accounts can help identify and prioritize anomalous transactions and suspicious relationships.

SAS Anti-Money Laundering [11] combines context-specific intellectual property and SAS Foundation technologies to support all key areas of a complete AML solution – including suspicious activity monitoring, customer due diligence and watch-list filtering – and all steps involved in AML processes.

Applied a discretisation process on their datasets to build clusters. They firstly discretise the whole timeline into difference time instances. Hence, each transaction is viewed as a node in one-dimensional timeline space. They project all transactions of customers to the timeline axis by accumulating transactions and transaction frequency to form a histogram. They create clusters based on segments in the histogram. A local and a global correlation analyzing are then applied to detect suspicious patters. This approach improves firstly the complexity by reducing the clustering problem to a segmentation problem [9]. Furthermore, it is more or less appropriate for analyzing individual behaviors or group behaviors by their transactions to detect suspicious behaviors related to "abnormal" hills in their histogram. However, as we have to analyse many customers with many transactions with a variety of amounts for a long period, it is difficult to detect suspicious cases, as there are very few or no "peak hills" in the histogram. Firstly, another global analysis is needed and we can then apply this method for further analysis in this case.

A Heuristics Approach for Fast Detecting Suspicious Money Laundering Cases in an Investment Bank [12], Firstly, transaction datasets are divided into two groups according to two kinds of investors: individual and corporate. Secondly, we refine the parameter $\Delta 1$ so that it is the proportion between the redemption value in the time τk and the maximum of the subscription values from time $\tau k\text{-}l$ to τk instead of the proportion between the redemption value and the subscription value in the time τk as in our previous solutions. Besides, the parameter l in $\tau k\text{-}l$ is adjustable and is defined by AML experts. It normally varies from 3 to 5. For instance, in the Table II, $\Delta 1$ of the customer A01 at the week 33 is not the proportion between the redemption value and the subscription value in week 33 but now is the maximum of subscription values from week 30 to week 33 (l=3 in this case). We apply this first heuristics because of AML experts' experience: the relevant subscriptions of a redemption activity in suspicious cases are normally not only in the current investigation term (week, month...) but also in its short previous term (two, three weeks or two three month ago). By applying heuristics in the clustering process as well as the suspicious screening, we only need to perform the clustering algorithm one time to determine the suspicious group. Briefly, these heuristics help to improve the running time of clustering process.

Besides, data mining techniques (DM) [14] have been proven to be well suited for identifying trends and patterns in large datasets. Therefore, DM techniques are expected to be applied successfully in the area of AML. Nevertheless, there is still little research concerning this bias especially a DM framework/solution for supporting AML experts in their daily tasks. Recently, these are some AML approaches based on DM that have been proposed and discussed in literature. Most of these approaches try to recognize ML patterns by different techniques such as support vector machine [15], correlation analysis [16], histogram analysis [16]… They aim to provide techniques for detecting a variety of ML by exploring a massive dimensionality of datasets including customers x accounts x products x geography x time. However, these approaches are more or less appropriate for the cash world and not scaled well for investment activities due to the lack of good methods in choosing parameters and they still have performance issues. In [17][18],they proposed a new solution basing on a combination of clustering and classification techniques for analyzing ML patterns in an international investment bank. Customer behavior in investment activities is complicated because it is influenced by many factors. We also show that by choosing suitable dimensions, simple DM techniques can be applied together to detect suspicious ML cases in investment activities. In this solution, the same clustering algorithm is repetitively executed to analysis transactions depending on the characteristic of each transaction datasets. Hence, in this paper, we present a one-step clustering approach basing on some heuristics from AML experts to improve the performance of our previous solution in the term of running time.

A number of basic countermeasures against money laundering have been proposed, including basic statistical analysis which constrains the amount of the transactions as well as restricting their frequency [19]. Other methods that complement these basic security measures are based on checking every customer against a black list originating from previous investigated cases and a white list to e.g. avoid mistakes when faced with persons with the same name. Unfortunately, these and other methods have proved to be insufficient [20].

Several machine learning techniques have been used for detecting fraud, and more specifically money laundering, [22]. From the point of view of machine learning, the application is interesting, due to the successful classification rate (high True Positives and low False Positives) that the classification model can achieve compared to other methods such as simple rule based detection that compares transactions against fixed thresholds.

Data mining based methods have also been used to detect fraud [21, 23, 24]. This leads to the observation that machine learning algorithms can identify novel methods of fraud by detecting those transactions that are different (suspicious) in comparison with the benign transactions. This problem in machine learning is known as novelty detection. Supervised learning algorithms have been used on a synthetic data set to prove the performance of outlier's detection [1].

3.Proposed Method

We propose a new probabilistic model using graph based sequential pattern mining algorithm for the detection of money laundering in financial sectors like banks and financial institutions and etc.. The proposed model contains five stages namely data cleaning, account graph construction, pattern mining, probability calculation, fraud detection.

3.1.Data Cleaning

Initially the large transactional data set from the banking sector is retrieved. The record set which contains noisy or incomplete data is removed. The cleaned data set is converted to computing forms, because various banks uses various form of customer data. For example few banks uses customer id in form of variable length of characters and few others uses only numbers and so on. All those information's are mapped to a single processing form. The converted transactional data sets are taken to the next stage of money laundering identification process.

3.1.1.Algorithm

Step 1: start
Step 2: read transactional data set T_s , Attribute set A_s.
Step 3: for each transaction T_i from T_s.
For each A_i from A_s
If T_i € A_i then
Else
$T_s = ø(T_i(T_s))$.
End
End
Step 4: end

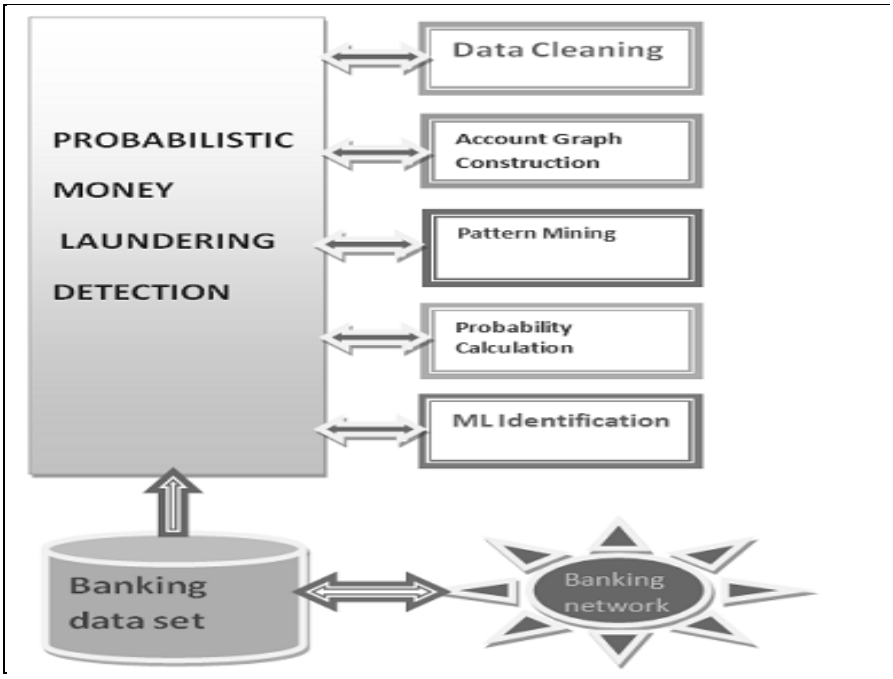

Figure 1: Proposed System Architecture

3.2.Account Graph Construction

Using the cleaned data set, for each account we construct the account graph. Each account has many numbers of recipients and each recipient has further recipient and so on. We restrict our self for two levels. Each graph has some edges E and vertex V, edges connects vertices. Starts from a single account, we generate a node and we identify the set of accounts linked to that particular account and we connect them with vertices. Here we represent each account as edge and there is a connection between two accounts only if they are linked ie if there is any transaction between them.
We construct this graph if there is any huge transaction or regular transaction with more capital.

3.2.1.Algorithm

Step 1**:** start
Step 2: read transactional set , initialize graph set G_s.
Step 3: identify unique source account S_a
Step 4: add S_a to account set Ac_s
Step 5: for each S_a from Ac_s
Create graph $G_{i..}$
Construct root node n with node value S_a.
$G_s = \sum G_s + G_i$.

End
Step 6: for each S_a from Ac_s
Identify recipient set R of S_a .
For each recipient R_i of R
Create an edge to the graph node with value S_a.
end
End
Step6: stop.

3.3.Pattern Mining
Once we construct the graph, we identify the paths of transition and amount transferred. We compute N (O) number of transition paths using the graph. N represents the total number of available paths in the graph and O represents the total number of nodes in the graph set G. A transition path is one, if there is a huge amount transfer between the nodes in chain. We collect all those transition paths and available paths of combination to compute the probability of money laundering. The transition paths are the patterns to compute probability.

3.3.1.Algorithm
Step1: start
Step2: read transactional data set T_s.
Step3: identify unique source account set Sa.
Step3: for each account Sa_i
Compute overall amount transferred ota = $\sum Ta(T_i)_n$. // T_a-amount transfered
If ota>th then //th-a threshold amount
Identify set of transition Tr_s between S_{i-j} and path from G_i .
Add tr_s to transition set Tr.
End
End
Step4:stop.

3.4.Probability Calculation And Money Laundering Identification
Using the computed patterns, we identify the pattern which has more strike value. The strike value represents the frequency of money transfer. And also it is not necessary that the criminals transfer the amount on a single path regularly. We compute probability for each out going nodes and each combination. To identify ML we use , the financial sector norms and rules for fund transfer. The government listed many rules for the transfer of amount from domestic and international banking. Based on the rules specified we sort the probability values and we select few accounts to monitor the fund transfer and so on.

3.4.1.Algorithm
Step1: start
Step2: read computed transition pattern Tr_s , initialize suspected transition set ST_s.
Step3: for each transition path Tr_i
Compute transition frequency tf=nt/tn.
Nt – number of occurrence of Tr_i in Tr_s.
tn – total number of transaction from Tr_s.
compute pt=tn×log(tf).
Pt-probability of transition in a transition path Tr_i.
If(pt>=0.7) then
Add transition path Tr_i to suspect set ST_s.
$ST_s=\sum ST_s+Tr_i$.
End
End
Step4: forward ST_s for monitoring.
Step5: stop.

4.Result And Discussion
The proposed methodology generates efficient transition patterns to identify money laundering. The generated transition set contains set of money transfer transition paths which shows the set of travel sequence of amount transfer. The criminal may transfer amount in different paths as little amount towards different destination accounts which could merge in a single account. The proposed method identifies the criminal activities in better manner to identify the money laundering efficiently.

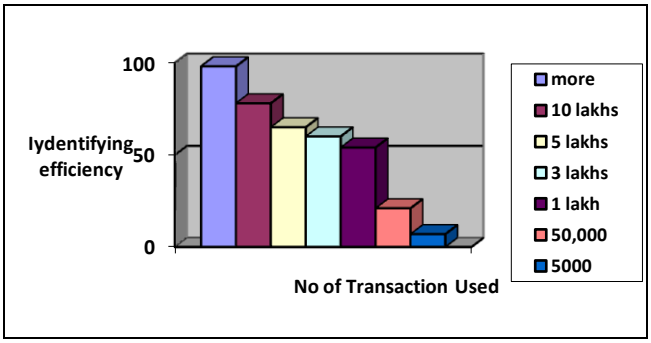

Figure 2: Shows The Efficiency Of Identifying Money Laundering

The Figure 2 shows the efficiency of identifying money laundering with respect to number of transaction used. It is clear that the efficiency is increased if the size of transaction is increased. The proposed methodology produces efficient result by increasing the size of transaction.

5.Conclusion

We analyze various methodologies to identify money laundering crime. We identify that all methods have scalable in accuracy and efficiency. We proposed a probabilistic model using graph based sequence pattern mining algorithm, which produces good results. We further move our research to next level using graph based behavior analysis to identify the money laundering crime.

6.References

1. Diane J. Cook, Lawrence B. Holder, Jeff Coble and Joseph Potts. (2005). Graph-based Mining of Complex Data. Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, Part I, pp.75-94.
2. Tao Jiang and Ah-Hwee Tan. (2005). Ontolopy-Assisted Mining of RDF Documents. Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, Part II, pp.231-252.
3. Carlo Batini, Monica Scannapiece. (2006). Data Quality (Concepts, Methodologies and Techniques). First Edition, Springer, 2006.
4. Vicenc Torra. (2003). Trends in Information fusion in Data Mining. Information Fusion in Data Mining, Springer, 2003, pp. 1-6.
5. Mohammed J. Zaki. (2005). TreeMiner: An Efficient Algorithm for Mining Embedded Ordered Frequent Trees. Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, Part I, pp.123-152.
6. J. Kingdon. AI Fights Money Laundering, IEEE Transactions on Intelligent Systems, 2004.
7. J. Tang. A Framework on Developing an Intelligent Discriminating System of Anti Money Laundering, International Conference on Financial and Banking, Czech Rep., 2005
8. Nhien An Le Khac an investigation into Data Mining approaches for Anti Money Laundering, 2009 International Conference on Computer Engineering and Applications IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore.
9. Edgar Alonso Lopez-Rojas , Money Laundering Detection using Synthetic Data, The 27th annual workshop of the Swedish Artificial Intelligence Society (SAIS), 14–15 May 2012, Örebro, Sweden.
10. SAS Anti-Money Laundering.
11. Z. Zang, J.J. Salermo and P. S. Yu, Applying Data mining in Investigating Money Laundering Crimes, SIGKDD'03, August 2003, Washington DC, USA. pp: 747-752.
12. R. Jain, R. Kasturi and B.G. Schunck, Machine Vision, Prentice Hall, 1995.
13. J. Han and M. Kamber, Data Mining: Concept and Techniques. Morgan Kaufmann publishers, 2nd Eds., Nov. 2005.
14. J. Tang, J. Yin, Developing an intelligent data discriminating system of anti-money laundering based on SVM, Proceedings of the Four International Conference on Machine Learning and Cybernetics, Guangzhou, Aug. 2005: pp.3453-3457.
15. Z. Zang, J.J. Salermo and P. S. Yu, Applying Data mining in Investigating Money Laundering Crimes, SIGKDD'03, August 2003, Washington DC, USA. pp: 747-752.
16. N-A. Le-Khac, S. Markos, M. O'Neill, A. Brabazon and M-T. Kechadi, An Efficient Search Tool for an Anti-Money Laundering Application of an Multi-National Bank's Dataset, The 2009 International Conference on Information and Knowledge Engineering, July 13-16, 2009 (IKE 2009), LA, USA.
17. N-A. Le-Khac, S. Markos and M-T. Kechadi, Towards a new Data Mining-based approach for Anti Money laundering in an international investment bank. a NY, USA (to appear).
18. R.J. Bolton and D.J. Hand. Statistical fraud detection: A review. Statistical Science, 17(3):235{249, 2002.
19. Dan Magnusson. The costs of implementing the anti money laundering regulations in Sweden. Journal of Money Laundering Control, 12(2):101{112, 2009.
20. Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining- based fraud detection research. Arxiv preprint arXiv:1009.6119, 2010.
21. Agus Sudjianto, Sheela Nair, Ming Yuan, Aijun Zhang, Daniel Kern, and Fernando . Statistical Methods for Fighting Financial Crimes. Technometrics, 52(1):5{19, February 2010.
22. Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, and Chao-Hsien Chu. A Review of Data Mining-Based Financial Fraud Detection Research. In 2007 International Conference on Wireless Communications, Networking and Mobile Computing, pages 5514{5517. Ieee, September 2007.
23. ZM Zhang and JJ Salerno. Applying data mining in investigating money laundering crimes. discovery and data mining, (Mlc):747, 2003.