Unsupervised Fraud Detection in Time Series data

Zakia Ferdousi† and Akira Maeda‡
† Graduate School of Science and Engineering, Ritsumeikan University
‡ Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University
1-1-1, Noji-Higashi, Kusatsu, Shiga, 525-8577, Japan
E-mail: † laboni23@yahoo.com , ‡ amaeda@is.ritsumei.ac.jp

Abstract: Fraud detection is of great importance to financial institutions. This paper is concerned with the problem of finding outliers in time series financial data using Peer Group Analysis (PGA), which is an unsupervised technique for fraud detection. The objective of PGA is to characterize the expected pattern of behavior around the target sequence in terms of the behavior of similar objects, and then to detect any difference in evolution between the expected pattern and the target. The tool has been applied to the stock market data, which has been collected from Bangladesh Stock Exchange to assess its performance in stock fraud detection. We observed PGA can detect those brokers who suddenly start selling the stock in a different way to other brokers to whom they were previously similar. We also applied t-statistics to find the deviations effectively.

Keywords: Outlier Detection, Fraud Detection, Time Series Data, Data Mining, Peer Group Analysis.

1. Introduction

With the expanded Internet and the increase of online financial transactions, financial services companies have become more vulnerable to fraud. Detecting the frauds means identifying suspicious fraudulent transfers, orders and other illegal activities against the company. Outlier detection is a fundamental issue in data mining, specifically in fraud detection. Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [1, 2], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [3]. The identification of outliers can lead to the discovery of useful knowledge and has a number of practical applications in areas such as credit card fraud detection, athlete performance analysis, voting irregularity analysis, severe weather prediction etc. [4, 5, 6]. Peer Group Analysis (PGA) is an unsupervised method for monitoring behavior over time in data mining [7]. Unsupervised methods do not need the prior knowledge of fraudulent and non-fraudulent transactions in historical database, but instead detect changes in behavior or unusual transactions. An advantage of using unsupervised methods over supervised methods is that previously undiscovered types of fraud may be detected. Supervised methods are only trained to discriminate between legitimate transactions and previously known fraud.

2. Stock Market Analysis

2.1 Stock Fraud & The Manipulators

Stock fraud usually takes place when brokers try to manipulate their customers into trading stocks without regard for the customers' own real interests. Stock fraud can be at a company level, or can be committed by a single stockbroker. Stock fraud can also vary in size from multi-million deals to penny stocks, but it consistently involves the intentional disregard for the financial situation of the customers and with personal profits. The key principle of stock fraud is that the investor's interests are secondary to the financial gain the broker can make.

Corporate insiders, brokers, underwriters, large shareholders and market makers are likely to be manipulators.

Table 1: Types of People Involved in Manipulation

Year	Broker	Insider	Market maker	Under writer	Share holder	Total Cases
1990	17	9	0	6	3	25
1991	3	3	0	1	1	4
1992	11	2	2	2	0	12
1993	2	0	0	0	0	2
1994	1	1	0	0	1	1
1995	8	8	7	0	7	9
1996	1	2	0	0	2	2
1997	10	10	1	0	8	11
1998	5	3	0	0	3	7
1999	7	5	1	1	7	11
2000	12	8	2	5	6	28
2001	14	17	1	0	7	30
Total	91	68	14	15	45	142
Total %	**64.08**	**47.89**	**9.86**	**10.56**	**31.69**	**-**

Table 1 reports the occurrence of ‘potentially informed’ people who are involved in manipulation cases from 1990 to 2001 in United States. ‘Insider’ denotes corporate executives

and directors. 'Shareholder' denotes large shareholders with 5% or more ownership in the manipulated stock. More than one type of person may be involved in any case [8].

2.2 Why Stock Fraud Detection is Necessary

Several fraud detection methods are available for the fields like credit card, telecommunications, network intrusion detections etc. But stock market fraud detection area is still behind. Most of the stock market researches are about the prediction of stock price. Since stock market enhances the economic development of a country greatly, this field has a vital need for efficient security system. Also the amount of money involved in stock market is huge. So, appropriate fraud detection system is essential. For example, in Australia, 63 per cent of people's superannuation, namely their retirement savings, is invested in securities. Investment in stock market is high in almost all the countries. If we don't protect against the ability of people to manipulate those securities, then implicitly, we're open to attack, or we're allowing open to attack a country's very wealth. Indeed. It is a very real threat, a threat that very few people really, are acknowledging. Stock fraud may not be very frequent but when it occurs the amount of loss is abundant. Outlier detection in stock market transactions will not only prevent the fraud but also alert the stock markets and broking houses to unusual movements in the markets. The recent incident in Tokyo Stock Exchange (Japan) can be a proper example to understand the impact of unusual movements in the market. On Dec. 8, a trader of Mizuho Securities mistakenly typed a sell order for 610,000 J-Com shares for 1 yen each, rather than the intended one J-Com share for 610,000 yen, which brought the firm a loss of some 40 billion yen. The incident has seriously blemished investor confidence in the world's second largest bourse. Thus stock fraud detection has become a vital research issue in current situation.

3. Our Contribution

First we analyzed how the fraud cases occur in stock market by the thorough technical reviews and from the practical experiences with stock markets. Most of the occurrences are due to artificial price increases, internet-rumor, artificial supply restriction & demand creation, insider trading and short selling. To investigate for the effective stock fraud detection method we generalize the possible outliers more specifically. The following two cases are the most important which have to mine first to detect stock fraud:

- Identify seller IDs whose sell quantity rise up suddenly.
- Identify seller IDs whose sell quantity fall suddenly.

Manipulators are mostly involved in selling rather than buying. So identification of the seller brokers with abnormal sell quantity is the most vital need for stock fraud detection.

We simulate the PGA tool in various situations and illustrate its use on a set of stock market transaction data. We evaluated the performance of PGA over Stock fraud detection. We found that this tool is quite efficient for the above cases. PGA was initially proposed for credit card fraud detection by Bolton & Hand in 2001[7]. We applied the tool in our research by changing some parameters. Our intention is to modify PGA to fit for stock market fraud detection and also to increase its effectiveness.

The statistics used here to compare stock selling within the accounts is the mean quantity of stock sells over the time window. We also demonstrated t-statistics to find the deviations more effectively.

4. Related Work

Outlier detection in time series has recently received considerable attention in the field of data mining. It is also important for fraud detection.

Unsupervised fraud detection methods have been researched in the detection of computer intrusion (hacking). Here profiles are trained on the combinations of commands that a user uses most frequently in their account. If a hacker gains illegal access to the account then their intrusion is detected by the presence of sequences of commands that are not in the profile of commands typed by the legitimate user. Qu, Vetter et al. (1998) use probabilities of events to define the profile [9], Lane and Brodley (1998) [10], Forrest et al (1996) [11] and Kosoresow and Hofmeyr (1997) [12] use similarity of sequences that can be interpreted in a probabilistic framework.

The neural network and Bayesian network comparison study (Maes *et al*, 2002) uses the STAGE algorithm for Bayesian networks and back propagation algorithm for neural networks in credit transactional fraud detection. Comparative results show that Bayesian networks were more accurate and much faster to train, but Bayesian networks are slower when applied to new instances [13].

The Securities Observation, News Analysis, and Regulation (SONAR) (Goldberg *et al*, 2003) uses text mining, statistical regression, rule-based inference, uncertainty, and fuzzy matching. It mines for explicit and

implicit relationships among the entities and events, all of which form episodes or scenarios with specific identifiers. It has been reported to be successful in generating breaks the main stock markets for insider trading (trading upon inside information of a material nature) and misrepresentation fraud (falsified news) [14].

Yamanish et al. [15] reduce the problem of change point detection in time series into that of outlier detection from time series of moving-averaged scores. Ge et al. [16] extend hidden semi markov model for change detection. Both these solutions are applicable to different data distributions using different regression functions; however, they are not scalable to large size datasets due to their time complexity.

5. Peer Group Analysis

5.1 Overview

The Following processes are involved in PGA**.**

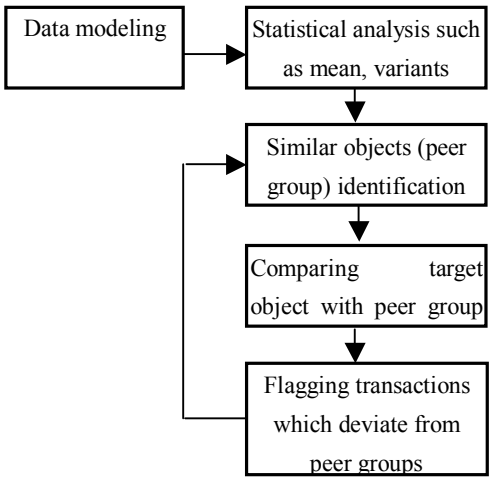

Figure 1: Process Flow of PGA

Peer group analysis (PGA) is a term that have been coined to describe the analysis of the time evolution of a given object (the *target*) relative to other objects that have been identified as initially similar to the target in some sense (the *peer group*).

- Since PGA finds anomalous trends in the data, it is reasonable to characterize such data in balanced form by collating data under fixed time periods. For example, the total sell quantity can be aggregated per week or the number of phone calls can be counted per day.
- After the proper data modeling some statistical analysis are required. Mean or variance can be appropriate. In our research we used weekly mean of stock transactions.
- Then the most important task of PGA method is to identify peer groups for all the target observations (objects). Member of peer groups are the most similar objects to the target object. In order to make the definition of peer group precise, we must decide how many objects, *npeer*, it contains from the complete set of objects. The parameter *npeer* effectively controls the sensitivity of the peer group analysis. Of course, if *npeer* is chosen to be too small then the behavior of the peer group may be too sensitive to random errors and thus inaccurate. The length of time window for calculating the peer group has been chosen arbitrarily here. We used 5 weeks for our experiments.
- Peer groups are summarized at each subsequent time point and the target object is then compared with its peer group's summary.
- Those accounts deviate from their peer groups more substantially are flagged as outliers for further investigation.
- These processes repeat from the peer group identification to the account flagging as long as proper result received.

5.2 How PGA Works

PGA detects individual objects that begin to behave in a way distinct from objects to which they had previously been similar. Each object is selected as a target object and is compared with all other objects in the database, using either external comparison criteria or internal criteria summarizing earlier behavior patterns of each object. Based on this comparison, a peer group of objects most similar to the target object is chosen. The behavior of the peer group is then summarized at each subsequent time point, and the behavior of the target object compared with the summary of its peer group. Those target objects exhibiting behavior most different from their peer group summary behavior are flagged as meriting closer investigation.

5.3 Significance of PGA

The approach of PGA is different in that a profile is formed based on the behavior of several similar users where current outlier detection techniques over time include profiling for single user. The most distinguishing feature of PGA lies in its focus on local patterns rather than global models; a sequence may not evolve unusually when compared with the whole population of sequences but may display unusual properties when compared with its peer group. That is, it may begin to deviate in behavior from

objects to which it has previously been similar.

5.4 Definition of Peer Groups

Based on [7], Let us suppose that we have observations on *N* objects, where each observation is a sequence of *d* values, represented by a vector, $\mathbf{x}_i$, of length *d*. The *j*th value of the *i*th observation, x_{ij}, occurs at a fixed time point t_j.

Let $PG_i(t_j)$ = {Some subset of observations ($\neq \mathbf{x}_i$) which show behavior similar to that of $\mathbf{x}_i$ at time t_j }. Then $PG_i(t_j)$ is the peer group of object *i*, at time *j*.

The parameter ***npeer*** describes the number of objects in the peer group and effectively controls the sensitivity of the peer group analysis. The problem of finding a good number of peers is akin to finding the correct number of neighbors in a nearest-neighbor analysis.

5.5 Peer Group Statistics

Let S_{ij} be a statistic summarizing the behavior of the *ith* observations at time *j*. Once we have found the peer group for the target observation $\mathbf{x}_i$ we can calculate peer group statistics, P_{ij}. These will generally be summaries of the values of S_{ij} for the members of the peer group. The principle here is that the peer group initially provides a local model, P_{i1}, for S_{i1}, thus characterizing the local behavior of $\mathbf{x}_i$ at time t_1, and will subsequently provide models, P_{ij}, for S_{ij}, at time t_j, j>1. If our target observation, S_{ik}, deviates 'significantly' from its peer group model P_{ik} at time t_k, then we conclude that our target is no longer behaving like its peers at time t_k. If the departure is large enough, then the target observation will be flagged as worthy of investigation.

To measure the departure of the target observation from its peer group we calculate its standardized distance from the peer group model; the example we use here is a standardized distance from the centroid of the peer group based on a *t*-statistic. The centroid value of the peer group is given by the equation:

$$\boldsymbol{P}_{ij} = \frac{1}{npeer}\left(\sum_{p \in P_i(t_1)} S_{pj}\right); \quad j \geq 1,\ p \neq i.$$

where $P_i(t_1)$ is the peer group calculated at time t_1. The variance of the peer group is then

$$\boldsymbol{V}_{ij} = \frac{1}{(npeer-1)} \sum_{p \in P_i(t_1)} (S_{pj} - P_{ij})(S_{pj} - P_{ij})'$$

Where $j \geq 1,\ p \neq i$.

The square root of this can be used to standardize the difference between the target S_{ij} and the peer group summary P_{ij}, yielding

$$\mathbf{T}_{ij} = (S_{ij} - P_{ij}) / \sqrt{V_{ij}}$$

6. Experiments

Table 2: Parameters Used in Experimental Setup

Symbol	Meaning
d	Total number of weeks
N	Number of target objects
npeer	Number of peer group member
w	Length of time window

6.1 Experimental Data

Our data set consists of 3 months real data from 06/01/2005 to 08/31/2005 for the daily stock amount sold for each of 143 brokers, which has been collected from Bangladesh Stock Exchange (Dhaka). Since stock frauds are more frequent in emerging stock markets rather than developed stock markets, data from Bangladesh stock market is very suitable for the simulation. The total number of transaction is 340,234.

Here we set, d = 14 weeks, N = 143. The length of time window, w = 5, but varied *npeer* to take values *npeer* = 13 and *npeer* = 26.

Using peer group analysis, we can detect those brokers who suddenly start selling the stock in a different way to other brokers to whom they were previously similar. A sample of stock market data is shown below:

Table 3: Stock Market Transaction

ID	Date	Stock	Seller	Buyer	Quantity
002205	6/1/05	11102	30	184	10
002206	6/1/05	11102	30	194	5
002207	6/1/05	11102	30	178	5
002208	6/1/05	11102	134	178	5
002209	6/1/05	11102	134	184	10
002210	6/1/05	11102	134	184	10

6.2 Experimental Results

For comparison purpose, we simulated PGA over stock transactions many times by changing the number of peers. The following plots illustrate the power of PGA to detect local anomalies in the data. The vertical axis shows cumulative stock sold as weeks pass on the horizontal axis.

The sold quantity of the target observation is represented by a red line and the sold quantity of the peer group by green lines; sold quantity from a sample of the remaining accounts is represented by blue lines.

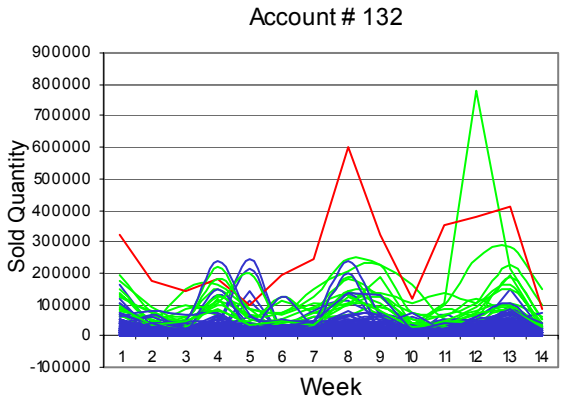

Figure 2: PGA Over Stock Transactions, account # 132 when npeer = 13

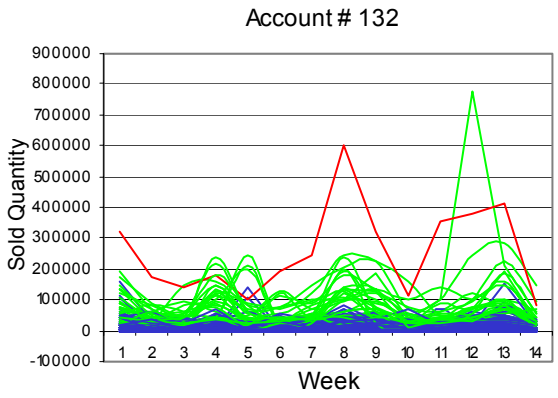

Figure 3: PGA Over Stock Transactions, account # 132 when npeer = 26

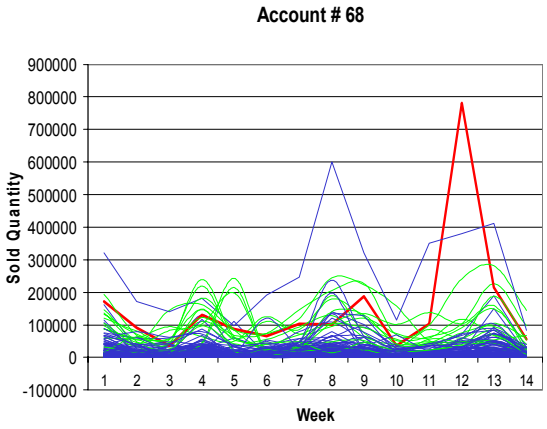

Figure 4: PGA Over Stock Transactions, account # 68 when npeer = 13

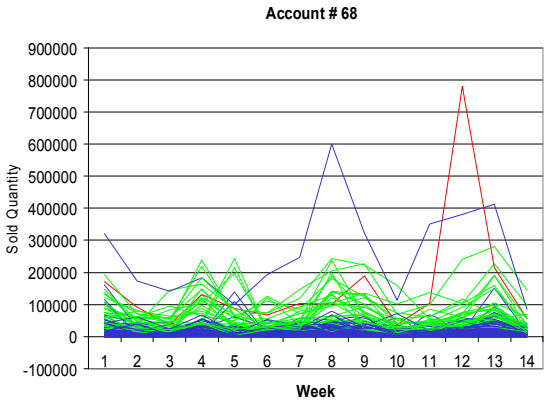

Figure 5: PGA Over Stock Transactions, account # 68 when npeer = 26

We also measured the departure of the target observation from its peer group. If the departure is large enough then the target observation will be flagged as worthy of investigation. For this purpose we calculated its standardized distance from the peer group model; the following results shown here are the standardized distances from the centroid of the peer group based on a t-statistic [17].

Table 4: Departure of Some Broker Accounts

Account No.	T-Score
132	**5.65768366**
68	**2.1516554**
99	1.74654872
129	1.61005567
164	1.20917806
3	0.778209479
7	0.587235098
52	0.216076926
34	-0.44583502
65	-1.3929922

7. Discussions

Figure 2 shows an account (132) flagged since it has the highest suspicious score in 8th week. Figure 3 also shows account (132) but here *npeer* is increased to 26. The behavior of this account varied largely from its peers almost in every week even though number of peers was increased. According to the suspicious score calculated by t-statistics (Table 4), this account (132) is the most suspicious one. This is an outlier but it may not be a fraud case. Since the behavior of this account is different to its peer groups from the beginning, so may be it is the general nature of this particular

broker. This indicates the need for negative alarm reduction. But this information is also necessary for proper knowledge discovery of such stock transactions.

Figure 4 shows an account (68) flagged as having the highest suspicious score at 12th week whereas most peers have very little spending in this week. This could be a possible fraud case since the behavior of this account was quite similar to its peer groups for all the weeks except the sudden rise on 12th week. Figure 5 shows account (68) where *npeer* is 26. Here we got very interesting findings. The behavior of this account has not been affected by the increase of *npeer*, which makes this account more suspicious.

Comparisons with lower *npeer* may result outliers in some cases. But if the same target account is compared with higher *npeer* then the situation may revert. The account can be found as normal. Because higher *npeer* means the target account is being compared with more number of objects. Thus the behavior of the target account may not be much different from its peer. In our experiment, we determined the proper value of *npeer* by comparing with the total number of objects. We have about 143 objects. So, taking npeer as 26 is quite perfect for the method.

In practical application, the flagged accounts will simply be noted as meriting more detailed examination, which has to be done definitely by human.

The process of calculating the peer groups and t-scores can be run every minute in a real-time manner. Using over 340,234 transactions gives an indicator of the performance of PGA on large data sets.

8. Conclusions and Future Work

In this paper, we tried to mention the necessity of stock market fraud detection since the area has lack of proper researches. We have demonstrated the experimental results of PGA tool in an unsupervised problem over real stock market data sets with continuous values over regular time intervals. The visual evidences have been shown through graphical plots that peer group analysis can be useful in detecting observations that deviate from their peers. We also applied t-statistics to find the deviations effectively.

We aim to proceed by incorporating other information, other than simply the quantity sold, into the outlier detection process (PGA) to increase the effectiveness of the fraud detection system. The following cases of possible outliers have to be investigated:

- Identify buyer IDs whose buy quantity rise up suddenly.
- Identify seller/buyer IDs who suddenly starts a large volume of trade.
- Identify stock IDs if trade volume or trade quantity increases suspiciously.
- Identify stock IDs with sudden raise or fall in price or having same buyer and seller.

We will develop necessary methods to minimize the negative alarms since all outlier are not frauds.

We have intention to integrate some other effective methods with PGA. We will also apply our strategy on other more applications, such as banking fraud detection.

References

[1] Barnett V. and Lewis T. *Outliers in Statistical Data*. John Wiley, New York, 3rd edition, (1994).

[2] Johnson R. *Applied Multivariate Statistical Analysis*. Prentice Hall, (1992).

[3] Victoria J. Hodge, Jim Austin: A Survey of Outlier Detection Methodologies. Artif. Intell. Rev. 22(2): 85-126 (2004).

[4] Knorr E. and Ng. R. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proc. 24th VLDB Conference*, 392–403, 24–27, (1998).

[5] Ruts I. and Rousseeuw. P. Computing Depth Contours of Bivariate Point Clouds. In *Computational Statistics and Data Analysis, 23:153–168*, (1996).

[6] Johnson T. and Kwok I. and Ng. R. Fast Computation of 2-Dimensional Depth Contours. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 224–228, (1998).

[7] Bolton, R. J. and Hand. D.J. "Unsupervised Profiling Methods for Fraud Detection," *Credit Scoring and Credit Control VII, Edinburgh, UK, 5-7 Sept* (2001).

[8] Rajesh K. Aggarwal & Guojun Wu. Stock Market Manipulation — *Theory and Evidence,* Working paper, Univ. of Michigan. *March 11,(2003).*

[9] Qu, D., Vetter B. M., Wang F., Narayan R., Wu S. F., Hou Y. F., Gong F. and Sargor C. Statistical Anomaly Detection for Link-State Routing Protocols. *Proceedings. Sixth International Conference on Network Protocols*, 62-70, (1998).

[10] Lane, T. and Brodley C. E. Temporal Sequence learning and data reduction for Anomaly Detection. *Proceedings of the 5th ACM conference on Computer and Communication Security, New York, 150-158 (1998).*

[11] Forrest, S., Hofmeyr S., Somayaji A. and Longstaff T. A sense of self for unix processes. *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, Los Alamitos, CA, (1996).

[12] Kosoresow, A. P. and Hofmeyr S. A. Intrusion Detection via System Call Traces. *IEEE Software* 14(5), 24-42, (1997).

[13] Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. Credit Card Fraud Detection using Bayesian and Neural Networks. *Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies,* (2002).

[14] Goldberg, H., Kirkland, J., Lee, D., Shyr, P. & Thakker, D. The NASD Securities Observation, News Analysis \& Regulation System (SONAR). *Proc. Of IAAI03,* (2003).

[15] Yamanishi K. and Takeuchi. J. Ichi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 676–681, 2002.

[16] Ge. X. *Segmental semi-markov models and applications to sequence analysis*. PhD thesis, (2002).

[17] Hand D.J., Mannila H., and Smyth P. Principles of Data Mining, MIT Press (2001).