

Anomaly Detection Using Unsupervised Profiling Method in Time Series Data

Zakia Ferdousi¹ and Akira Maeda²

¹Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1, Noji-Higashi, Kusatsu, Shiga, 525-8577, Japan

laboni23@yahoo.com,

²Department of Media Technology, College of Information Science and Engineering, Ritsumeikan University, 1-1-1, Noji-Higashi, Kusatsu, Shiga, 525-8577, Japan

amaeda@is.ritsumeai.ac.jp

Abstract. The anomaly detection problem has important applications in the field of fraud detection, network robustness analysis and intrusion detection. This paper is concerned with the problem of detecting anomalies in time series data using Peer Group Analysis (PGA), which is an unsupervised technique. 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 differences in evolution between the expected pattern and the target. The experimental results demonstrate that the method is able to flag anomalous records effectively.

Keywords: Anomaly Detection, Data Mining, Peer Group Analysis, Unsupervised Profiling, Time Series Data.

1 Introduction

With the expanded Internet and the increase of online financial transactions, financial services companies have become more vulnerable to fraud. 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.

2 Stock Market Analysis

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

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 lagging. 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. For example, in Australia, 63 per cent of people's retirement savings is invested in securities. Investment in stock market is high in almost all the countries. If we do not protect against the ability of people to manipulate those securities, then implicitly, we are open to attack, or we are allowing open to attack a country's 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 arises 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.

3 Our Contribution

First we analyzed how fraudulent cases occur in stock market through the thorough technical reviews and from the practical experiences of stock markets. The following

two cases are the most important criteria, which we aim to mine first to detect the stock fraud:

- To identify broker accounts whose sell quantity rise up or fall down suddenly.
- To identify broker accounts whose trade volume rise up or fall down suddenly.

We simulate the PGA tool in various situations and illustrate its use on a set of stock market transaction data. PGA was initially proposed for credit card fraud detection by Bolton & Hand in 2001[7], where it considered only the spending amount of each account. But using one attribute is not enough to flag an account as a fraud. An effective and practical fraud detection method needs to incorporate more information. We tried to overcome the problem by including more attributes within the outlier detection process by PGA. We applied PGA over two attributes and then we performed a comparative analysis between those two observations. We flagged the accounts as suspicious based on the knowledge discovered from the comparative analysis. Thus the results of outliers mining become more realistic and effective than the traditional PGA. We also demonstrated *t*-statistics to find the deviations more effectively.

4 Related Work

Outlier detection in time series database has recently received considerable attention in the field of data mining. Qu, et al. uses probabilities of events to define the profile [8], Lane and Brodley [9], Forrest et al. [10] and Kosoresow and Hofmeyr [11] use similarity of sequences that can be interpreted in a probabilistic framework.

The neural network and Bayesian network comparison study [12] 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. The Securities Observation, News Analysis, and Regulation (SONAR) [13] 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. Yamanishi et al. [14] 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. [15] 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 (fig. 1).

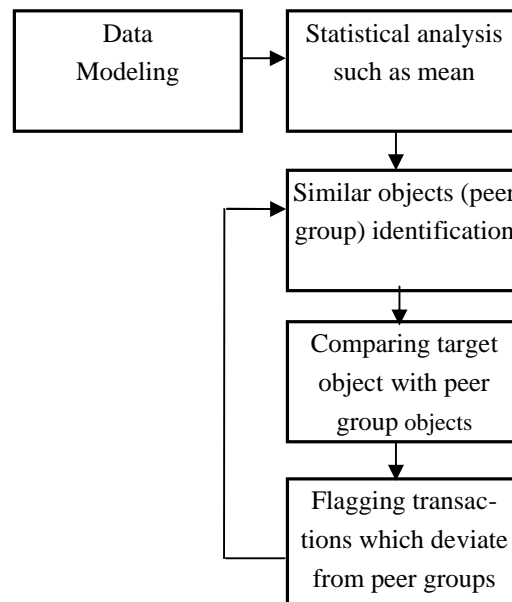


Fig 1. Overview of PGA

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 quantity sold can be aggregated per week or the number of phone calls can be counted per day.

After the data modeling process, some statistical analyses 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 can be chosen based on the particular data set. 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. The processes from the peer group identification to the account flagging are repeated as long as the proper result is found.

5.2 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.

A point may not be seen as unusual when compared with the whole set of points but may display unusual properties when compared with its peer group. This is the most significance feature of PGA.

5.3 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 $PG_i(t_j) = \{\text{some subset of observations } (\neq \mathbf{x}_i), \text{ which show behavior similar to that of } \mathbf{x}_i \text{ at time } t_j\}$. Then $PG_i(t_j)$ is the peer groups 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.

5.4 Peer Group Statistics

Let S_{ij} be a statistic summarizing the behavior of the i th 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 1:

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

where $PG_i(t_1)$ is the peer group calculated at time t_1 . The variance of the peer group can be calculated by the equation 2:

$$V_{ij} = \frac{1}{(npeer - 1)} \sum_{p \in PG_i(t_1)} (S_{pj} - P_{ij})^2. \quad (2)$$

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 equation 3:

$$T_{ij} = (S_{ij} - P_{ij}) / \sqrt{V_{ij}}. \quad (3)$$

6 Experiments

Table 1 shows the list of parameters used in the experiments.

Table 1: 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 6/01/2005 to 08/31/2005 for the daily stock sell quantity and the number of transactions for each of 143 brokers. The total number of transaction is 340,234. This data has been collected from the Dhaka Stock Exchange (Bangladesh).

Here we set, $d = 14$ weeks, $N = 143$. The length of time window, $w = 5$, but varied $npeer$ to take values $npeer = 13$ and $npeer = 26$. A sample of stock market data has shown in table 2:

Table 2: Stock Market Transactions

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

6.2 Experimental Results

For comparison purpose, we simulated PGA over stock transactions many times by changing the number of peers. Here we have shown some of the results, which are more interesting. 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 sale quantity of the target observation is represented by thick

black line and the sale quantities of the peer group are represented by black dotted lines. The graph of number of transactions has shown in the same manner.

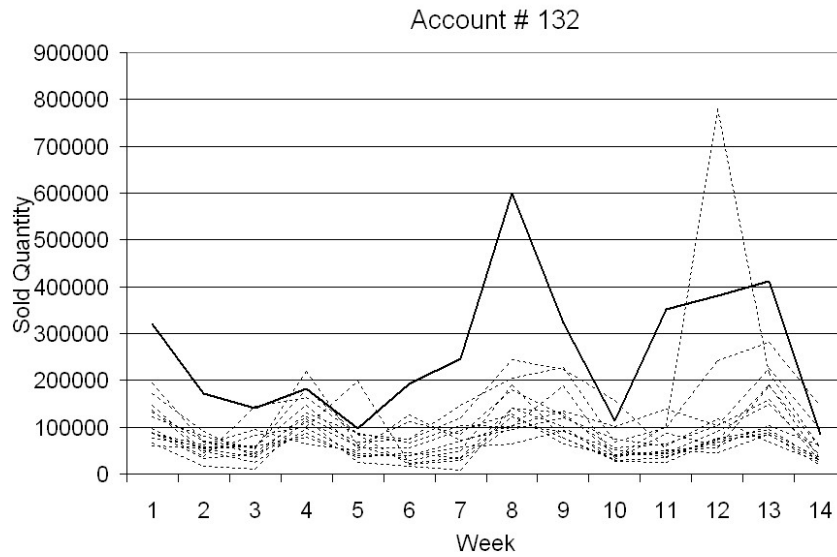


Fig 2. PGA Over Sell Quantity, Account # 132 $n_{peer} = 13$

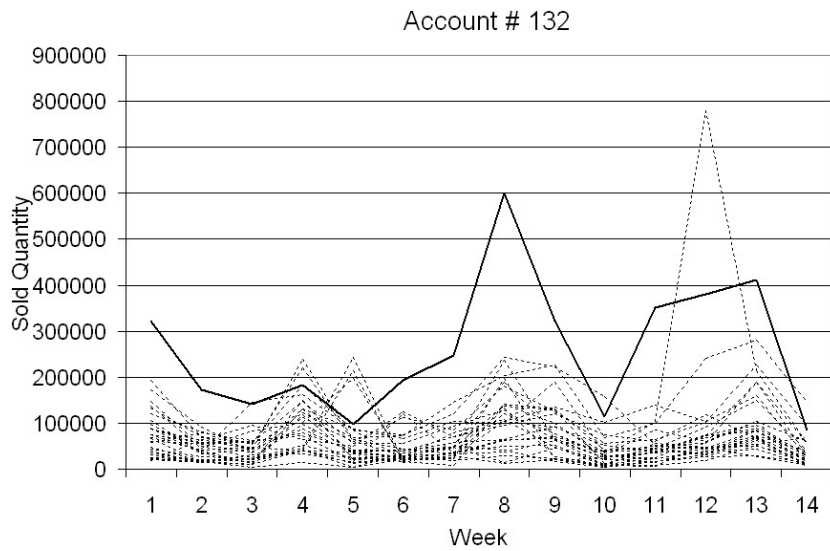


Fig 3. PGA Over Sell Quantity, Account # 132 $n_{peer} = 26$.

Account # 68

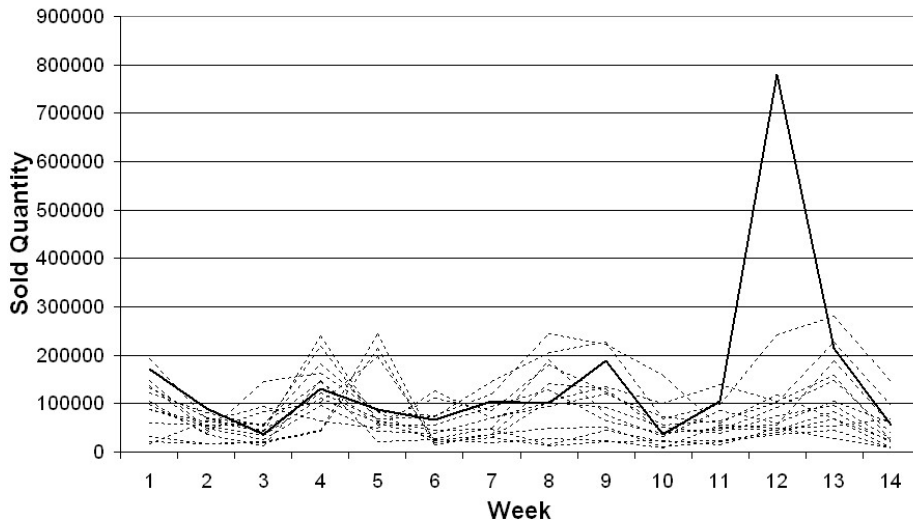


Fig 4. PGA Over Sell Quantity, Account # 68 $npeer = 13$

Account # 68

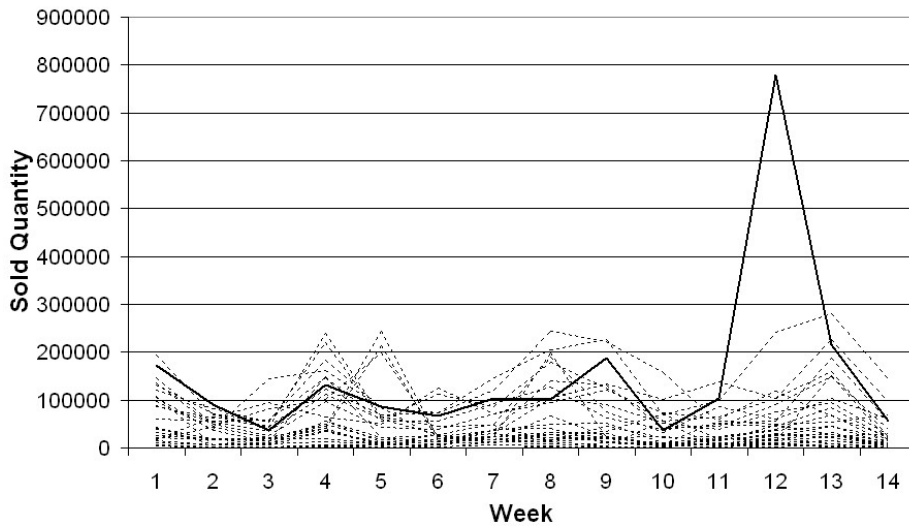


Fig 5. PGA Over Sell Quantity, Account # 68 $npeer = 26$

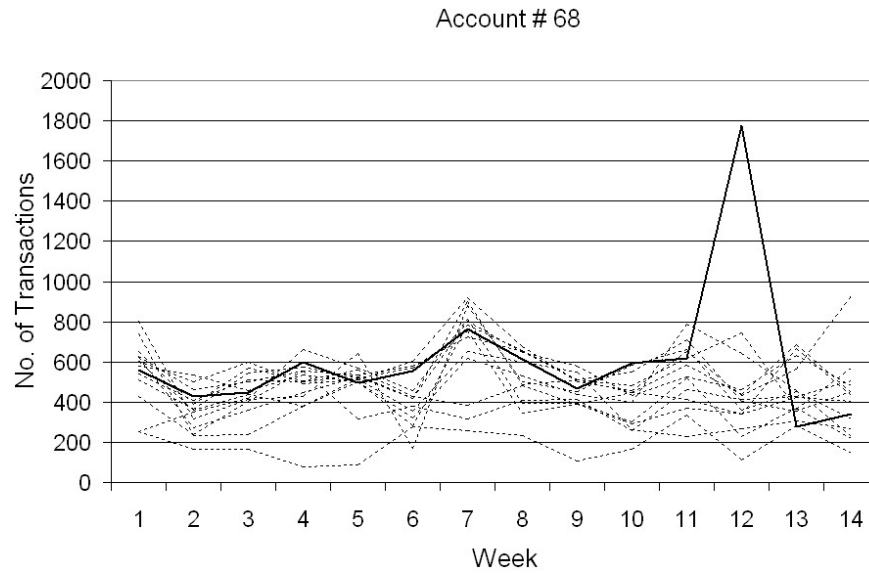


Fig 6. PGA Over Number of Stock Transactions, Account # 68 $n_{peer} = 13$

We have 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 have calculated its standardized distance from the peer group model. Table 3 shows the standardized distances from the centroid of the peer group based on a t-statistic [16].

Table 3: 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

7 Discussions

Fig 2 shows account132 is flagged since it has the highest suspicious score in the 8th week. Fig 3 shows account 132 where $npeer$ is increased to 26. These simulations were conducted according to the traditional PGA but with a slight change in parameters $npeer$ and the time window. 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 3), 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 from its peer group from the beginning, may be it is the general nature of this particular broker. Even introduction of more $npeer$ is not enough to decide weather it is a fraud case or not. The account's behavior is still far away from its peers.

Fig 4 shows account 68 is flagged since it has a clear sudden rise in 12th week whereas most peers have very little sales in this week. This could be a possible fraud case since the behavior of this account was quite similar to its peer group for all the weeks except the sudden rise in the 12th week. Fig 5 shows account 68 where $npeer$ is 26. Here we got very interesting findings. The behavior of this account also has not been affected by the increase of $npeer$ but it makes the account more suspicious. According to our proposed method in section 3, we extended the investigation for the suspicious accounts in fig 6. We included another attribute into the outlier mining process. The main idea is to evaluate more information before flagging the fraudulent account, where traditional PGA considers only one attribute to flag an account.

We considered the number of transactions as well as trade volume as another indicator of stock fraud. Fig 6 shows PGA over number of stock transactions for the same account 68. It is obvious from the figure 6 that the number of transactions has suspiciously increased in the 12th week.

So, from fig 4, 5 and 6 we can do a comparative analysis about account 68. The discovered knowledge here is:

- Account 68 has sudden rise for both the sale quantity and the number of transactions in 12th week.
- The behavior of this account was similar with its peer group's sale quantity and number of transactions in other weeks.

Now we can flag the account as an outlier or possible fraud case more confidently, because both the observations have same findings. The process of calculating the peer groups and t-scores can be run every minute in a real-time manner. Thus the outlier mining process becomes more effective than the one-attribute observation process.

Here we have demonstrated the results with the suitable *npeer* for our data set. In practical application, the flagged accounts will simply be noted as meriting more detailed examination. 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.

In future, we aim to investigate for whether PGA can identify labeled fraudulent objects or not from a real fraud data set. To make the stock fraud detection more effective we will mine the following cases of possible outliers:

- To identify stock IDs and buyer IDs in case of trade volume or trade quantity increases suspiciously.
- To identify stock IDs with sudden raise or fall in price or having same buyer and seller.

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, network intrusion detection etc.

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.* (2004), 22(2): 85-126
4. Knorr E. and Ng R.: Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proceedings of 24th VLDB Conference*, (1998).
5. Ruts I. and Rousseeuw P.: Computing Depth Contours of Bivariate Point Clouds. *Computational Statistics and Data Analysis*, (1996) 23:153–168.

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* (1998), 224–228. AAAI Press.
7. Bolton, R. J. and Hand D. J.: Unsupervised Profiling Methods for Fraud Detection, Credit Scoring and Credit Control VII, Edinburgh, UK (2001)
8. 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. In *Proceedings of the Sixth International Conference on Network Protocols* (1998), 62-70.
9. Lane, T. and Brodley C. E.: Temporal Sequence Learning and Anomaly Detection. In *Proceedings of the 5th ACM Conference on Communications Security*. (1998), 150- 158.
10. Forrest, S., Hofmeyr S., Somayaji A. and Longstaff T.: A sense of self for unix processes. In *Proceedings of the IEEE Symposium on Security and Privacy*, (1996).
11. Kosoresow, A. P. and Hofmeyr S. A.: Intrusion Detection via System Call Traces. *IEEE Software*,(1997), 14(5), 24-42.
12. Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B.: Credit Card Fraud Detection using Bayesian and Neural Networks. In *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies* (2002).
13. Goldberg, H., Kirkland, J., Lee, D., Shyr, P. & Thakker, D.: The NASD Securities Observation, News Analysis & Regulation System (SONAR). In *Proceedings of IAAI*, (2003).
14. Yamanishi K. and Takeuchi J.: A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (2002), 676–681.
15. Ge X.: *Segmental semi-markov models and applications to sequence analysis*, PhD thesis, Chair-Padhraic Smyth, (2002).
16. Hand D.J., Mannila H., and Smyth P: Principles of Data Mining, MIT Press (2001).