Expert Systems with Applications 39 (2012) 8986–8992

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Stock fraud detection using peer group analysis

Yoonseong Kim, So Young Sohn *

Department of Information & Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, Republic of Korea

ARTICLE INFO

Keywords:
Stock price manipulation
Peer group analysis
Unsupervised data mining
Anomaly detection

ABSTRACT

This study proposes a method to detect suspicious patterns of stock price manipulation using an unsupervised data mining technique: peer group analysis. This technique detects abnormal behavior of a target by comparing it with its peer group and measuring the deviation of its behavior from that of its peers. Moreover, this study proposes a method to improve the general peer group analysis by incorporating the weight of peer group members into summarizing their behavior, along with the consideration of parameter updates over time. Using real time series data of Korean stock market, this study shows the advantage of the proposed peer group analysis in detecting abnormal stock price change. In addition, we perform sensitivity analysis to examine the effect of the parameters used in the proposed method.

1. Introduction

Stock price manipulation has been traditionally classified into three different categories (Allen & Gale, 1992). First, stock prices can be manipulated by actions that change the actual or perceived value of the assets (*action-based manipulation*). Second, manipulation can occur when false information or false rumors are released (*information-based manipulation*). Finally, traders attempt to manipulate stock prices simply by buying and then selling, without making any publicly observable actions to alter the value of the firm or releasing false information to change the price (*trade-based manipulation*). While the eradication of action and information based manipulation has been fairly successful, trade-based manipulation is still difficult to detect. This has been exacerbated since the online trading system was adopted. Moreover, it has been reported that trade-based manipulation has shown more diverse patterns and thereby the rules and patterns derived from historical data regarding previous attempts at manipulation may quickly become outdated.

This study proposes a method to detect a suspicious symptom of stock price manipulation only using the change of the stock market data itself. For this, we apply an unsupervised learning technique, peer group analysis, which detects individual objects (*target*) that begin to behave significantly different from the other objects to which they had previously been similar (*peer group*). That is, we detect the abnormal behavior of a target by comparing it with its peer group members and measuring the deviation of its behavior from the peer group. The advantage of this approach is that we can find local outliers which cannot otherwise be detected when compared to the whole population. Based on the general peer group analysis, this study moreover proposes a method to improve the general technique by using weighted means as a summarizing statistic of peer group as well as updating the weights over time. Using time series data of stock prices of companies listed in the Korean stock market, we examine the application of the peer group analysis to the detection of stock price manipulation. Finally, we conduct a sensitivity analysis to examine the effect of the parameters used in the proposed method.

The remainder of this paper is structured as follows. Section 2 reviews related literature and Section 3 describes peer group analysis. Section 4 applies peer group analysis to the real stock price data and performs sensitivity analysis for the parameters used in the peer group analysis. Finally, Section 5 concludes the paper.

2. Literature review

Various detection systems or stock watch systems to monitor abnormal stock price changes have been developed to detect stock price manipulations. The National Association of Securities Dealers (NASD) in the United States developed the Advanced Detection System (ADS) that has been used to monitor trades and quotations in the NASDAQ stock market since mid-1997 (Kirkland et al., 1999; Senator, 2000). The ADS uses two pattern matchers, a rule matcher and a time-sequence matcher, to detect patterns or practices of regulatory concern. The rule pattern matcher detects predefined suspicious behaviors, while the sequence matcher finds temporal relationships between events which exist in a potential violation pattern. In addition to ADS, NASD also developed Securities Observation, News Analysis, and Regulation (SONAR) to monitor stock markets for potential insider trading and fraud through misrepresentation (Goldberg, Kirkland, Lee, Shyr, & Thakker, 2003). SONAR

* Corresponding author. Tel.: +82 2 2123 4014; fax: +82 2 364 7807.
E-mail address: sohns@yonsei.ac.kr (S.Y. Sohn).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.025

applies several AI and statistical techniques such as NLP text mining, statistical regression, rule-based inference, and fuzzy matching. In the similar manner, the stock exchange markets in other countries must have been operating their own watch system, although the detailed nature of the systems is not open to the public for security reasons.

In addition to the surveillance of stock price changes, anomaly detection techniques have been applied to various fields such as network intrusion detection (Naiman, 2004; Scott, 2004), fault detection (Chen, Martin, & Montague, 2009; Martins, Pires, & Amaral, 2011; Yiakopoulos, Gryllias, & Antoniadis, 2011), financial fraud detection (Juszczak, Adams, Hand, Whitrow, & Weston, 2008). Comprehensive summaries of these methods are provided by Patcha and Park (2007). In the study, the authors roughly classify anomaly detection methods into three categories: statistical anomaly detection, data-mining based methods, and machine learning based techniques. They also summarized the advantages and disadvantages of each technique. Among them, data mining-based techniques encompass the algorithms and methodologies to automatically explore anomalous patterns or deviations hidden in large amount of data. Thus, they can eliminate the manual and ad hoc elements from the process of building an anomaly detection system. Those techniques can be divided into two types in terms of whether the fraudulent event is identified in the past data: supervised and unsupervised. Supervised algorithms examine all previously-labeled events to determine whether or not new events are fraudulent by calculating a risk score. Popular supervised data mining techniques are neural networks (Kirkos, Spathis, & Manolopoulos, 2007; Oliveira & Meira, 2006; Syeda, Zhang, & Pan, 2002; Viaene, Dedene, & Derrig, 2005), decision trees (Kirkos et al., 2007; Wang, Fan, Yu, & Han, 2003), case-based reasoning (Wheeler & Aitken, 2000), *k*-nearest-neighbor (Nikulin, 2006), and association rules (Sanchez, Vila, Cerda, & Serrano, 2009). Though the supervised methods generally show high detection accuracy, they are not effective when there is only a very small percentage of anomalies and frauds. Also, because the methods are based only on known patterns in the past data, they can be ineffective for detecting new patterns.

On the other hand, unsupervised learning techniques do not require known forms of fraudulent events or anomalies. While supervised techniques identify suspicious patterns by comparing the behavior with the prior knowledge, unsupervised methods trace the change of behavior itself and detect the abnormal action by comparing it with the normal state. It is therefore better to apply unsupervised learning techniques to finding novel patterns of fraud or anomaly which cannot be detected by supervised ones. Various unsupervised techniques and their application to the detection of anomaly and fraud have been extensively studied in past research (Amini, Jalili, & Shahriari, 2006; Cortes & Pregibon, 2001; Giacinto, Perdisci, Del Rio, & Roli, 2008; Yamanishi & Takeuchi, 2002).

Among wide range of anomaly detection methods introduced in the above, this study applies peer group analysis to detect suspicious behavior for stock price manipulation. The peer group analysis detects anomalies of an object by comparing its behavior with that of other objects which had been similar to the target object before. Like other unsupervised learning methods, peer group analysis has been also widely applied to various fields such as credit card transactions (Bolton & Hand, 2002; Weston, Hand, Adams, Whitrow, & Juszczak, 2008), business transactions (Tang, 2006), and stock transactions (Ferdousi & Maeda, 2006). Our choice of an unsupervised learning method, peer group analysis, is based on the fact that it is almost impossible to define all patterns of stock price manipulation and they are unceasingly evolving and become diverse to avoid the surveillance of the detection system. In addition, we choose peer group analysis as a research method since the price of stocks in the same industry or closely associated one another show similar patterns of price change. That is, we assume that the anomalous pattern of a target stock price change is more accurately identified in comparison with the normal behavior of its peer group members. Ferdousi and Maeda (2006) also applied peer group analysis to the stock fraud detection and observed suspicious cases of fraud using stock exchange data. However, they do not take into account the change of peer groups over time. Some members in the same peer group may gradually exhibit distinct behavior from that of other members. In this case, it is more desirable to split the peer group into two different peer groups or reorganize all the peer groups according to newly observed behavior. To address this issue, we propose a method to continuously update the behavior of peer groups over time.

3. Peer group analysis

The objective of peer group analysis is to characterize the expected pattern of behavior around the target sequence by monitoring the behavior of similar objects, and then to detect any differences between the expected pattern and the target. Peer group analysis can be divided into two stages: (1) building peer groups, and (2) detecting anomalous behavior in the constructed peer groups. In the following subsections, we give a detailed explanation about how to build peer groups and how to apply them to the detection of anomalous patterns.

3.1. Peer group membership

Objects that have a similar pattern of change over time are classified as a peer group. Let us assume that we have m time series, each of which has a sequence of D values for each time point from $t = 1$ to D. That is, the observations of time series i can be represented by a vector $\mathbf{x_i}$ with a length of D. The observation of time series i at time t is denoted as x_{it}.

Peer groups are built based on the similarity between time series. To measure this similarity, we first subdivide time series into N non-overlapping windows and calculate the average of the D/N observations within each window for the purpose of smoothing time series data. That is, time series i is represented by the sequence of

$$y_{in} = \frac{1}{D/N}\sum_{j=1}^{D/N} x_{i,,\frac{D}{N}(n-1)+j}, \quad \text{for } n = 1, 2, \ldots, N. \tag{1}$$

Then, the dissimilarity between two time series $\mathbf{y_i}$ and $\mathbf{y_{i'}}$ is measured by the Euclidean distance of the N subdivided values:

$$d_{ii'} = \sqrt{(\mathbf{y_i} - \mathbf{y_{i'}})(\mathbf{y_i} - \mathbf{y_{i'}})^T} \quad \text{where} \quad \mathbf{y_i} = [y_{i1,}y_{i2,}, \ldots, y_{iN}]. \tag{2}$$

Once we have measured the similarity between time series, we can build the peer groups. For each time series $\mathbf{y_i}$, we sort the remaining time series in the order of increasing distance (decreasing similarity), yielding $y_{i,\pi(1),n}, \ldots, y_{i,\pi(m-1),n}$ where $y_{i,\pi(j),n}$ is the value of the nth window of the jth closest peer group member of time series i. Since there are a total number of m time series, a time series $\mathbf{y_i}$ can have at most $m - 1$ peer group members. If we fix the size of each peer group as k, then the peer group of the time series $\mathbf{x_i}$ is simply the first k time series in this ordering.

3.2. Peer group summary

Once we have identified the peer group members for a target time series ($\mathbf{x_i}$), we can summarize the behavior of the peer group at time t as one value (P_{it}). The mean of peer group members is one of the most typical measures for P_{it}. That is