Expert Systems with Applications 36 (2009) 11944–11949

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Detecting stock-price manipulation in an emerging market: The case of Turkey

Hulisi Öğüt [a,*], M. Mete Doğanay [b], Ramazan Aktaş [a]

[a] *Department of Business Administration, TOBB University of Economics and Technology, Söğütözü Caddesi No: 43, Ankara 06560, Turkey*
[b] *Department of Business Administration, Çankaya University, Öğretmenler Caddesi No: 14, Balgat, Ankara 06530, Turkey*

ARTICLE INFO

Keywords:
Stock market
Manipulation
Data mining techniques
Multivariate statistical techniques

ABSTRACT

This paper aims to develop methods that are capable of detecting manipulation in the Istanbul Stock Exchange. We take the difference between manipulated stock's and index's average daily return, average daily change in trading volume and average daily volatility and used these statistics as explanatory variables. The data in post-manipulation and pre-manipulation periods are used as non-manipulated instances while the data in the manipulation period are used as manipulated instances. Test performance of classification accuracy, sensitivity and specificity statistics for Artificial Neural Networks (ANN) and Support Vector Machine (SVM) are compared with the results of discriminant analysis and logistics regression (logit). We found that the data mining techniques (ANN and SVM) are better suited to detect stock-price manipulation than multivariate statistical techniques (discriminant analysis, logistics regression) as the performances of the data mining techniques in terms of total classification accuracy and sensitivity statistics are better than those of multivariate techniques. We also found that unit change in difference between average daily return of manipulated stock and the index has the largest effect while unit change in difference between average daily change in trading volume of manipulated stock and index has the least effect on multivariate classifiers' decision functions.

1. Introduction

Manipulation is an important issue for both developed and emerging stock markets. Prices of the stocks must be determined by the market without any interference. Since the investors value a stock by taking all the relevant information into account, market prices determined without any interference reflect the common judgment of the investors about the value of the stocks and can be accepted as fair. One of the most important distortions of stock-prices is caused by manipulation. With this regard, manipulation can be defined as any interference to the market mechanism that prevents a fair price to prevail. Manipulation can also be defined as any action to artificially influence the price of a stock. If the investors believe that the prevailing prices do not reflect the true value of the stocks, they lose their confidence in the market. Furthermore, the most important consequence of manipulation is the lost incurred by the investors. Manipulators try to attract the investors to buy or sell a certain security by employing different methods. The main purpose of the manipulators is to make profit from this trading at the expense of the other investors.

There are different types of manipulation. Allen and Gale (1992) classify manipulation as action-based manipulation, information-based manipulation, and trade-based manipulation. Manipulators employ different methods to influence the price of the targeted stocks. Manipulators take actions such as closing a plant, not bidding in an auction, etc., to influence the price of the stocks in action-based manipulation. In this kind of manipulation, manipulators are the managers of the issuers whose actions have an impact on the stock-price. Because managers are restricted to take long or short positions in their firms' stocks in most of the countries, action-based manipulation is very rare nowadays. In information-based manipulation, manipulators spread rumors and false information to influence the stock-prices. In trade-based manipulation, manipulators engage in fraudulent trading to create an image of an active market. In fact, manipulators trade among themselves in order to artificially increase the price and volume of a stock for the purpose of attracting the other investors to buy the stock. Information-based manipulation and trade-based manipulation are the most common types of manipulation today.

Insider trading and financial information manipulation are other types of illegal trading that can be considered as manipulation. In insider trading, employees of the issuer use private information that has not been made public yet in order to gain personal benefit. In financial information manipulation, managers of the firm distort the information presented on the financial statements to give a false impression about the financial situation and the performance of their firms.

* Corresponding author. Tel.: +90 3122924218.
E-mail addresses: hogut@etu.edu.tr (H. Öğüt), mdoganay@cankaya.edu.tr (M. Mete Doğanay), raktas@etu.edu.tr (R. Aktaş).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.03.065

All types of manipulation harm the investors. Because manipulators have the investors believe that there are economic factors or private information behind the price changes. In reality, neither causes the price changes. Prices are artificially influenced by the fraudulent trading of the manipulators or rumors spread by the manipulators. When the prices revert to their economic values, investors incur losses and manipulators earn a profit at the expense of them. That is why manipulation is closely monitored in every stock market and charges brought against the manipulators when they are discovered. Since manipulation is a very important issue and it affects the investor confidence, it is of utmost importance to detect, deter and prevent it.

Many exchanges have devised methods to detect fraudulent trading and there are prison terms and fines to discourage manipulation. In spite of these measures, manipulation cases are still encountered in world stock exchanges. Allen (2006) in a recent conference gave examples of manipulation cases over the world and said: "manipulation is a very important topic for research particularly in US over the counter markets, European markets, and emerging markets". Manipulation is also a very important issue and a very important research topic in Turkey, which has an emerging stock market. Our paper aims to develop methods that are capable of detecting manipulation, especially the trade-based manipulation in the Istanbul Stock Exchange. As there are not many studies in this important research area, we believe that our paper makes modest contribution to this literature.

This paper is organized as follows: Section 2 summarizes the main result of prior studies. Section 3 gives information about methodologies used in this study. Data is presented in Section 4. The results of classifiers are discussed in Section 5 and last section concludes the paper.

2. Prior studies

Manipulation has drawn the attention of many researchers. Allen and Gale (1992) and Jarrow (1992) are the first researchers who studied manipulation. Allen and Gale studied the history of the stock-price manipulation and classified the manipulations as described above. Jarrow investigated, by constructing a model, whether a large trader could influence the price of a stock and generate a profit at no risk through manipulative trading. Other researchers studied whether profitable manipulation was possible in stock market and in other markets in different forms. Kumar and Seppi (1992) studied the possibility of manipulation in futures markets, Gerard and Handa (1993) investigated possible manipulations around seasoned equity offerings, Van Bommel (2003) developed a model for information-based manipulation in the stock market, Chakraborty and Yılmaz (2004) developed a model to show how informed insiders should trade in order to manipulate the market. Oprea, Hanson, and Porter (2005) developed an experimental model to study whether the manipulators could distort the prices in a prediction market.

Researches mentioned above can be categorized as theoretical studies. There are also empirical studies investigating whether the manipulators were successful in influencing the price and whether they had gained profit from their actions. Felixon and Pelli (1999) examined the closing-price manipulation; Khwaja and Mian (2003) investigated characteristics of the stock-price manipulation in Pakistan's main stock exchange. Both of the researches used unique data complied from the stock exchanges. Aggarwal and Wu (2006) developed a model to explain trade-based manipulation and tested the model by using data from US stock markets. Küçükoğlu (2004) examined closing-price manipulation in the Istanbul Stock Exchange (ISE), Turkey. Aktaş and Doğanay (2006) studied trade-based manipulations in the ISE. They applied the Aggarwal and Wu's model with minor modifications to trade-based manipulation cases in the ISE.

Although there are theoretical and empirical studies in the literature related to manipulation, studies that developed models to predict or detect manipulations are very rare. Prediction or detection models are widespread in other areas of finance. Models have been developed to predict financial failure and models have been developed to detect financial information manipulation. Explanatory variables in these multivariate statistical models (discriminant analysis, logistics regression, probit etc.) are financial ratios calculated by using the information contained in the financial statements. Besides multivariate statistical methods, fuzzy and neurofuzzy methods have also been used with the same variables to predict financial failure or detect financial information manipulation.

Few researchers have attempted to detect manipulation in different markets. Palshikar and Bahulkar (2000) are among the first who tackled this problem. They used a fuzzy temporal logic in which they specified trading patterns common to the manipulators. Pirrong (2004) used standard statistical techniques (regression analysis and error correction models) to detect manipulation in futures market. Author tried to determine whether price increase in one suspected manipulation case was a result of manipulation or normal economic conditions. Abrantes-Metz and Addanki (2007) developed an error-based model to detect manipulation in commodities market. They applied this model to Hunt Brother's silver manipulation case of 1979–80 and obtained successful results. Palshikar and Apte (2008) stated that many manipulative cases in the stock market involved collusion sets. They describe a collusion set as a group of traders who trade heavily among themselves. They tried to detect collusion sets by using graph clustering algorithms. Unlike other methods, the algorithms they used are unsupervised learning algorithms that can separate normal trading from fraudulent trading. They used synthetic databases where trading data based on different probability distributions were created and collusion sets of different sizes and characteristics were injected in.

Our study differs from other studies in this literature at least one of the following reasons. First, our focus is on predicting and detecting stock-price manipulation. Second, we used real world data while prior studies used synthetic data in detecting stock-price manipulation. Moreover, our explanatory variables are different from the variables used in other studies about manipulation.

3. Methodology

The data used in this research is used first at Aktaş and Doğanay (2006) study to see whether the manipulators were successful in gaining profits at the expense of the other investors. They applied Aggarwal and Wu (2006) model with minor modifications to answer the research questions related to characteristics of stock-price manipulation.

Aggarwal and Wu's model is a four-period model. The first period is the pre-manipulation period. Manipulators engage in fraudulent trading during the second period. The fraudulent trading of the manipulators increases the trading volume of the stock and inflates the price artificially. Information seekers (positive feedback traders) observe the volume and price changes in the second period and they think that there are economic fundamentals or private information behind these increases. In fact, the main purpose of the manipulators is to lure the information seekers to enter the market. When information seekers enter the market during the third period, their extra demand causes the price to increase further and manipulators sell their stocks at a higher price to the information seekers and realize their gains. The second and the third periods are called manipulation period. Price of the stock returns to its normal value during the fourth period. The fourth period is called the post-manipulation period. Although it is possible that manipulators try to deflate the price, we observe that all manipulation cases in our dataset try to inflate the price. In information-based manipulations, manipulators try to increase (or decrease) the price of a stock during the second period by spreading rumors. According to this model, we expect that price (as a result return), volume, and volatility increase during the manipulation period and drop after the manipulation. Volume is used as proxy for the number of information seekers. As explained above, manipulators try to attract as many information seekers as possible and when the information seekers enter the market, their extra demand further increases the price. So, we expect that higher volume during the manipulation period (more information seekers) creates higher price. There must be disagreement and uncertainty about the true value of the stock so that the manipulators can attract information seekers. Volatility represents this uncertainty and we expect that as volatility increases manipulators earn a higher return. All of these expectations were tested in Aktaş and Doğanay (2006) by using a unique data set and the results were consistent with the expectations. Motivated by these results, we try to develop models in this paper to detect possible stock-price manipulations by observing price (return), volume, and volatility of the targeted stocks. Since a unique data set related to manipulation cases is available, multivariate statistical models and supervised learning algorithms can be used. These algorithms are presented next.

3.1. Multiple Discriminant Analysis (MDA)

The basic idea underlying discriminant analysis is to determine whether groups differ with regard to the mean of a variable. For this purpose, ratio of the between-groups variance to within-group variance is computed. If this ratio is high, it means that group means of the variable are significantly different. These variables are the best discriminators among priori defined groups and using linear combination of these variables, discriminant functions maximizing degree of separation between two groups are generated for the estimation of group membership. MDA assumes that the data (for the variables) represent a sample from a multivariate normal distribution and covariance/variance matrices for every group are equal (homogenous).

3.2. Logistic regression

Logistic regression models the relationship between explanatory variable and independent variable by making logistic transformation of odds ratio. Odds ratio for binary output represents how frequently one class is observed relative to other class. Iterative method is used to determine coefficients of the explanatory variables. This method aims to minimize logarithmic sums of predicted probabilities and it converges when this summation becomes close to zero or it does not change from one iteration to another. Unlike other methods used in this study, it is possible to estimate probabilities from the logistic regression by using the following equation

$$\text{Probability} = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + a_2 x_2 \ldots + a_n x_n)}}$$

where a_0, $a_1 \ldots a_n$ are the parameters and x_1, $x_2 \ldots x_n$ are the inputs.

3.3. Artificial Neural Networks (ANN)

Artificial Neural networks (ANN) consist of three layers: input layer, one or more hidden layers and output layer. In each layer, there are neurons and neurons in the one layer are connected to all neurons in the next layer at the fully connected networks. Each connection has an associated weight and weighted combination of connected neuron's output becomes the input for next layer neuron. Transformation function takes this input and converts into the output value of the neuron. The final computed output is compared with the actual value of the output and difference is used to recalculate the weights by different methods. Among them, back-propagation algorithm is the most popular one and it adjusts the weights from output nodes to inner nodes based on the difference between actual and computed output.

3.4. Support Vector Machine

Based on neural network and statistical theory, Support Vector Machine (SVM) is used for classification, regression and density estimation. Besides bankruptcy prediction (Min & Lee, 2005), SVM has a number of other successful applications in business such as marketing (Cui & Curry, 2005), customer loan evaluation (Li, Shiue, & Huang, 2006), customer churn prediction (Coussement & Van den Poel, 2008). We will briefly discuss SVM technique as more detailed discussion can be found in Burges (1998) and Vapnik (1995). Let's define training examples as $[x_i, y_i]$ where $x_i \in R^n$ is the input vector, n is the dimension of input vector and $y_i \in [-1, 1]$ is the output vector. SVM finds optimal hyper-plane that separates one class from the other by using quadratic programming technique. Optimal hyper-plane minimizes misclassification error and maximizes margin between hyper-plane and nearest point. The nearest points are called Support Vectors. The quadratic programming can be written mathematically as,

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$
$$\text{subject to} \quad y_i(w\phi(x_i) + b) + \xi_i - 1 \geqslant 0 \quad \xi_i \geqslant 0$$

where $\phi(x_i)$ maps input data into high dimensional feature space, w is weight vector, b is the bias term, C is the penalty for the error term andξ_i is the slack variable. (Vapnik, 1995). Lagrange multiplier technique is used as a solution procedure of this quadratic programming formulation. Once optimal hyper-plane that separates one class from the other is constructed, classification decision is given by the following equation,

$$f(y) = sign\left(\sum_{i=1}^{N} y_i a_i K(x, x_i) + b\right)$$

where sign is the sign function, a_i is the parameter and $K(x_i,x_j) = \Theta(x_i)^T \Theta\ (x_j)$ is Kernel function. Computational complexities occur when $\phi(x_i)$ maps training data into high dimensional feature space. For this reason, kernel function is used in the classification function as it makes implementation easier by considering only inner product rather than high dimensional feature space. Four popular examples of kernel function are given below.

Linear kernel function: $(K(x_i, x_j) = x_i^T x_i)$; polynomial kernel function $(K(x_i, x_j) = (\gamma x_i^T x_i + r)^d)$; radial basis function $(K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2))$; sigmoid kernel function $K(x_i, x_j) = \tanh\{\gamma x_i^T x_j + d\}$ where d, r and γ are constants (Hsu, Chang, & Lin, 2004).

4. Data

We used the same dataset in Aktaş and Doğanay (2006). Information about manipulation cases were obtained by examining the case summaries published in Capital Markets Board (CMB) Weekly Bulletins and the time period for manipulation cases covers from January 1995 to March 2004. For our analysis, we collected daily prices and trading volumes of the manipulated stocks and ISE National-All Shares Index. ISE National-All Shares Index is composed of all National Market companies except investment trusts and it is used as market benchmark. If the manipulated stock was issued by an investment trust, we used ISE Investment Trust Index as benchmark. There are a few number of manipulation cases occurred in the second national market and ISE Second National Market Index was used as benchmark for these cases.

We identify two hundred seventy seven manipulation cases by examining CMB Weekly Bulletins. Other than nine information-based manipulation cases, all of these cases are trade-based manipulations. When cases with insufficient data were eliminated, we ended up with 222 manipulation cases.

For the manipulated stocks, we computed average daily returns, average daily change in trading volume and the average daily volatility (standard deviation) statistics for the manipulation, pre-manipulation and post-manipulation period. We also computed the same statistics of the appropriate indices for the same periods. Based on these data, three inputs are generated. These are (i) the difference between average daily return of manipulated stock and average daily return of the index (X_1), (ii) the difference between average daily change in trading volume of manipulated stock and average daily change in trading volume of index (X_2), (iii) the difference between average daily volatility of manipulated stock and average daily volatility of index (X_3). As these input statistics measure stock's deviation from the non- manipulated case (index), higher deviation is an indication of anomaly. For each manipulated stock, the output attribute takes the value of 1 in the manipulation period and 0 in the pre-manipulation and post manipulation periods. Thus, we have 222 data instances belonging to the manipulation period and the 444 data instances belonging to the pre and post-manipulation periods as there are 222 manipulated firms.

5. Results

We used MATLAB for the implementation of ANN and SVM and SAS for the implementation of MDA and logistic regression classifiers. We divide the data into two parts: test and training. Training and test dataset has 100 and 122 instances from the manipulation period out of 300 and 366 total data instances, respectively. We used the same set of training and test data in both datasets across techniques in order to compare the performances of the classifiers.

5.1. ANN models

The performance of ANN classifier depends on the hidden layers and number of neurons in the hidden layers. Unfortunately, there is no agreed standard on how to choose these numbers. We limited the number of hidden layers to one and we used 5-fold cross validation technique to determine the number of neurons in the hidden layer. In an n-fold cross validation technique, training data is divided into n equal parts at first. Model is developed in $n - 1$ parts and it is tested in the remaining part. This process is repeated until all subset of data is tested with the model developed by the remaining part of the training data. Once all subset of data is tested, model is retrained with the parameters having highest average classification accuracy and performance of ANN classifier is validated on the test data. The advantage of the cross validation technique is to prevent over fitting of training data (Hsu et al., 2004). In the ANN model, hidden and output layer use sigmoid activation functions and quasi-Newton back-propagation is used as training method. The numbers of epochs tried in this study are 100, 200 and 300. We varied number of nodes in the hidden layers from 5 to 25 by an interval of 5. Table 1 shows cross validation accuracy of training data (CV) and classification accuracy of test data (CA) for the ANN classifier. As it can be seen, when the number of the neurons in the hidden layer is 25 and numbers of epoch are 100, best cross validation accuracy (85.3%) is obtained. Model retrained with these parameters achieved 79.2% classification accuracy on the test data.

5.2. SVM model

We choose radial basis function (RBF) as a kernel function for Support Vector Machine. Although there is no established procedure for determining best kernel function, the advantages of using RBF over other kernel functions are following: (i) Lin and Lin (2003) shows that the performance of sigmoid kernel is similar to RBF for certain parameters. (ii) Keerthi and Lin (2003) found that linear kernel with parameter C has the same performance as RBF kernel with parameters C'. (iii) Unlike linear kernel, RBF kernel can nonlinearly map input space into higher dimensional feature space (Hsu et al., 2004).

Two parameters are needed to determine the RBF kernel function: penalty parameter of the error (C) and kernel parameter (γ). There are many techniques proposed for choosing C and γ parameters to obtain better performance from SVM classifier on test data. We used grid search technique as its implementation is fairly simple and computational time required for this search is not so much different from other heuristic or advanced methods (Hsu et al., 2004). Grid search technique explores parameter space with the combination of (C, γ) using cross validation technique and parameter pairs having best cross validation is used on test data.

In the implementation stage of grid search technique, we chose exponential sequence of $C = \{2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}\}$ and $\gamma = \{2^{9}, 2^{7}, 2^{5}, 2^{3}, 2^{1}, 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}, 2^{-9}\}$ for the parameters. Next, we tried every combination of C and γ parameters in the training data using 5-fold cross validation as in ANN. Once parameter pair having the best cross validation accuracy is identified, SVM classifier is retrained with these parameters and its performance is validated on the test data.

Cross validation accuracies of SVM classifier are shown in Table 2. Best cross validation accuracy is obtained when combination of penalty parameter (C) and kernel parameter (γ) is $(2\wedge5, 2\wedge3)$. Once model is retrained with these parameters, SVM achieved 79.2% classification accuracy on test data. Corresponding classification accuracies are presented in Table 3.

Table 1
Cross validation and classification accuracies of Neural Network.

Number of Epoch	100		200		300	
Hidden layer	CV	CA	CV	CA	CV	CA
5	0.823	0.801	0.820	0.801	0.820	0.792
10	0.813	0.779	0.843	0.787	0.840	0.790
15	0.827	0.795	0.813	0.792	0.823	0.784
20	0.667	0.667	0.667	0.667	0.667	0.667
25	0.853	0.792	0.840	0.790	0.843	0.779

Table 2
Cross validation accuracies for support vector machine classifier.

C	2∧9	2∧7	2∧5	2∧3	2∧1	2∧−1	2∧−3	2∧−5	2∧−7	2∧−9
γ										
2∧−9	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
2∧−7	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
2∧−5	0.667	0.667	0.703	0.673	0.667	0.667	0.667	0.667	0.667	0.667
2∧−3	0.677	0.817	0.820	0.800	0.750	0.667	0.667	0.667	0.667	0.667
2∧−1	0.800	0.837	0.817	0.823	0.800	0.740	0.670	0.667	0.667	0.667
2∧1	0.813	0.827	0.833	0.813	0.823	0.797	0.740	0.670	0.667	0.667
2∧3	0.793	0.823	0.847	0.813	0.823	0.830	0.793	0.737	0.670	0.667
2∧5	0.773	0.807	0.840	0.830	0.820	0.823	0.827	0.793	0.743	0.670
2∧7	0.773	0.790	0.813	0.830	0.827	0.820	0.823	0.827	0.793	0.743
2∧9	0.743	0.783	0.807	0.817	0.837	0.820	0.827	0.823	0.827	0.793

Table 3
Classification accuracies of support vector machine classifier.

C	2∧9	2∧7	2∧5	2∧3	2∧1	2∧−1	2∧−3	2∧−5	2∧−7	2∧−9
γ										
2∧−9	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
2∧−7	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
2∧−5	0.667	0.667	0.738	0.700	0.667	0.667	0.667	0.667	0.667	0.667
2∧−3	0.719	0.790	0.773	0.792	0.740	0.675	0.667	0.667	0.667	0.667
2∧−1	0.787	0.795	0.779	0.773	0.792	0.740	0.675	0.667	0.667	0.667
2∧1	0.790	0.801	0.795	0.773	0.781	0.792	0.740	0.678	0.667	0.667
2∧3	0.790	0.798	0.792	0.781	0.781	0.784	0.792	0.740	0.678	0.667
2∧5	0.792	0.792	0.801	0.784	0.776	0.779	0.784	0.792	0.740	0.678
2∧7	0.784	0.795	0.795	0.779	0.781	0.773	0.779	0.784	0.792	0.740
2∧9	0.757	0.784	0.790	0.792	0.781	0.781	0.779	0.779	0.784	0.792

5.3. Logistic regression and MDA

For logit classifier and MDA, we used stepwise selection method to determine which variables are used in the final model. First, relevant statistics (*F* statistics for discriminant analysis and chi-square for the logistic regression classifier) for each variable is computed and the largest of them enter into the model. Then, the procedure computes the statistics except the chosen variables and determines the largest one. If this variable's statistics are greater than threshold value, this variable enters into the model. Each selection step is followed by one elimination step. Thus, if variable is selected into the model, it does not necessarily stay in the final model. This process stops till no further variable is selected into the model or the last variable entered and exited is the same.

The first multivariate statistical technique that we use is logistics regression (logit) model. Logit model can be expressed as follows:

$$F(Z_i) = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{j}\beta_{ji}*x_{ji}\right)}}$$

where $F(Z_i)$ is a probability function that stock is manipulated. We obtained following result from the estimation of the logit model.

$$Z = -1.5966 + 142.7X_1 + 11.0423X_3 \tag{1}$$

All of the variables are statistically significant at the 10% level. As a cutoff value, we use 0.5 meaning that if $F(Z) < 0.5$, the stock is classified as being manipulated. Otherwise, instance is classified as non-manipulated stock. Using Eq. (1) and cutoff value of 0.5, we can conclude that stock is manipulated if

$$Z = -1.5966 + 142.7X_1 + 11.0423X_3 < 0 \tag{2}$$

Otherwise, we decided that stock is not manipulated. This relationship tells us that probability of manipulation increases as the differences between average daily returns of the individual stock and average daily returns of index (X_1) and the difference between average daily volatility of the individual stock and average daily volatility of index (X_3) increase.

The second multivariate statistical model used is multiple discriminant analysis (MDA). MDA generates two linear functions belonging to two classes. Two scores for each class are computed and instance is assigned to the class having the highest score. We estimated the following discriminant functions for non-manipulation (Z_0) and manipulation cases (Z_1)

$$Z_0 = -0.54 - 24.17X_1 + 0.07X_2 + 10.38X_3 \tag{3}$$

$$Z_1 = -2.25 + 85.51X_1 + 0.15X_2 + 19.73X_3 \tag{4}$$

To obtain more meaningful result, we subtract Z_0 in Eq. (3) from Z_1 in Eq. (4) and we get following decision function.

$$Z_1 - Z_0 = -1.71 + 109.68X_1 + 0.08X_2 + 9.45X_3 \tag{5}$$

If the sign of Eq. (5) is negative, MDA classifies the instance as being manipulated. Otherwise, the instance is classified as being non-manipulated. Eq. (5) says that when the differences between average daily returns of the individual stock and average daily returns of index (X_1), the difference between average daily change in trading volume of the individual stock and average daily change in index (X_2) and the difference between average daily volatility of the individual stock and average daily volatility of index (X_3) are not positive, instance is classified as being non-manipulated. When X_1, X_2 and X_3 are positive and large enough, instance is classified as being manipulated stock. Eq. (5) also implies that X_1 has the largest effect and X_2 has the least effect on the result for the unit change in variables. We would like to note that values of the intercept, X_1 and X_3 in Eqs. (5) and (2) are close to each other. This shows that these two results are consistent with each other.

We also compare the performance of classifiers on test and training data and these results are summarized in Table 4. We have found that SVM and ANN achieve the best classification accuracies

Table 4
The performance of classifiers.

	Training data	Test data
Logit	0.830	0.773
MDA	0.800	0.784
ANN	0.857	0.792
SVM	0.863	0.792

Table 5
The performance statistics of the classifiers.

	Specificity	Sensitivity
ANN	0.88	0.65
SVM	0.88	0.65
Logit	0.89	0.54
Discriminant	0.92	0.51

on test data while Logit classifier has the worst performance. However, test performance comparison of the classifiers based on total classification accuracy may be misleading if errors have different misclassification costs. For this reason, we compare test performance of the classifiers based on the following statistics defined below.

Specificity: number of correctly classified non-manipulated instances/number of total non-manipulated instances.

Sensitivity: number of correctly classified manipulated instances/number of total manipulated instances.

Two types of error may be encountered in our case as we have two classes: non-manipulated class and manipulated stock. Type-1 error occurs when the model classifies manipulated stock as non-manipulated stock and Type-2 error occurs when the model classifies non-manipulated stock as manipulated stock. If investor classifies manipulated stock as non-manipulated stock, the values of the investment decrease and investor incurs loss when manipulation period ends. When non-manipulated stock is classified as manipulated stock, investor may miss a profitable investment opportunity. But cost of this type of error may be alleviated by investing in other alternatives. Consequently, higher sensitivity statistics is more important than specificity and total classification accuracy statistics in our context as it gives information about percentage of correctly classified manipulated stock.

The specificity and sensitivity statistics of classifiers are reported in Table 5. We would like to note that there is a tradeoff between these statistics since data mining techniques (SVM and ANN) have done much better in terms of sensitivity statistics while the performance of multivariate techniques are better for specificity statistics. Thus, we can conclude that the data mining techniques are better suited to detect stock-price manipulation than multivariate statistics techniques as the performances of the data mining techniques in terms of total classification accuracy and sensitivity statistics are better than those of multivariate techniques.

6. Conclusion

The aim of this research is to detect stock-price manipulation by using three inputs generated from manipulated stock information. For this purpose, we use two popular data mining techniques (Support Vector Machine and Artificial Neural Network) to estimate a suitable model and compare their performance with those of two mostly used multivariate techniques (MDA and logit model). We found that the data mining techniques have done a better job than multivariate statistical techniques in detecting stock-price manipulation as the performance of the data mining techniques in terms of total classification accuracy and sensitivity statistics are superior to those of multivariate techniques. However, data mining technique has one disadvantage over multivariate statistics techniques: Although it is not possible to quantify the relationship between output and inputs through data mining classifiers, one can derive decision function using multivariate statistics techniques.

References

Abrantes-Metz, R. M., & Addanki, S. (2007). Is the market being fooled? *An error-based screen for manipulation, working paper*.

Aggarwal, R. K., & Wu, G. (2006). Stock market manipulations. *Journal of Business, 79*(4), 1915–1953.

Aktaş, R., & Doğanay, M. (2006). Stock-price manipulation in the Istanbul stock exchange. *Eurasian Review of Economics and Finance, 2*, 21–28.

Allen, F. (2006). Market manipulation: Past, present and future. *EFA meetings, Philadelphia, USA*.

Allen, F., & Gale, D. (1992). Stock-price manipulation. *Review of Financial Studies, 5*, 503–529.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Chakraborty, A., & Yılmaz, B. (2004). Informed manipulation. *Journal of Economic Theory, 114*, 132–152.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Cui, D., & Curry, D. (2005). Predictions in marketing using the support vector machine. *Marketing Science, 24*(4), 595–615.

Felixon, K., & Pelli, A. (1999). Day end returns-stock manipulation. *Journal of Multinational Financial Management, 9*, 95–127.

Gerard, B., & Handa, V. (1993). Trading and manipulation around seasoned equity offerings. *Journal of Finance, 48*, 213–245.

Oprea, R., Hanson, R., & Porter, D. (2005). *Information aggregation and manipulation in an experimental market, working paper*. George Mason University Interdisciplinary Center for Economic Science.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2004). *A practical guide to support vector classification*. Technical report, Department of Computer Science and Information Engineering, National Taiwan University.

Jarrow, R. A. (1992). Market manipulation, bubble, corners, and short squeezes. *Journal of Financial and Quantitative Analysis, 27*, 311–336.

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation, 15*(7), 1667–1689.

Khwaja, A. I., & Mian, A. (2003). *Trading in phantom markets: Price manipulation in an emerging stock market, working paper*.

Küçükoğlu, G. (2004). Intra-day stock returns and close-end price manipulation in the Istanbul stock exchange. In *Proceedings of the sixth international conference in economics, Middle east Technical University, Ankara, Turkey*.

Kumar, P., & Seppi, D. J. (1992). Futures manipulation with cash settlement. *Journal of Finance, 47*, 1485–1502.

Lin, H.-T. & Lin, C.-J. (2003). *A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods*. Technical report, Department of Computer Science, National Taiwan University.

Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications, 30*(4), 772–782.

Min, J., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*, 603–614.

Palshikar, G. K., & Bahulkar, A. (2000). Fuzzy temporal patterns for analyzing stock market databases. In *Proceedings of the international conference on advances in data management* (pp. 135–142). Pune, India: Tata-McGraw Hill.

Palshikar, G. K., & Apte, M. M. (2008). Collusion set detection using graph clustering. *Data Mining and Knowledge Discovery, 16*, 135–164.

Pirrong, C. (2004). Detecting manipulations in futures markets: The ferruzzi soybean episode. *American Law and Economics Review, 6*, 28–71.

Van Bommel, J. (2003). Rumors. *Journal of Finance, 58*, 1499–1519.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.