Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm

Zhao, Lei
Baylor University
Email: cxhdy@foxmail.com

Wang, Lin
Japan Advanced Institute of Science and Technology
Email: linwang@jaist.ac.jp

***Abstract*—In this paper we present a novel data miming approach to predict long term behavior of stock trend. Traditional techniques on stock trend prediction have shown their limitations when using time series algorithms or volatility modelling on price sequence. In our research, a novel outlier mining algorithm is proposed to detect anomalies on the basis of volume sequence of high frequency tick-by tick data of stock market. Such anomaly trades always inference with the stock price in the stock market. By using the cluster information of such anomalies, our approach predict the stock trend effectively in the really world market. Experiment results show that our proposed approach makes profits on the Chinese stock market, especially in a long-term usage.**

***Keywords*—Stock trend prediction, data mining, cluster analysis, stock market, anomaly**

I. Introduction

Financial time series change dynamically and selectively. Such time series are obviously difficult to predict because the problem is nonlinear, non-stationary and have a lot of noises[4]. Stock price is a kind of time series in financial domain. The approach to predict stock trend in the future has become one of the most import issues by using data mining techniques. However, prediction is difficult from the principle of the efficient market hypothesis [2] that if the market is an efficient market then the stock price will follow a random work pattern. In addition, a stationary prediction strategy is also not possible if the market is efficient because investors will soon discover such strategies and those successful forecasting rules will lead to self-destruct [3]. A lot of researchers devote their time to study such random walks by time series modeling [5], volatility modeling[6] and even artificial intelligence modelling[4]. But those algorithms are all on the basis of the stock price itself which has random property.

In this paper, we turn back our attention to the distribution on volume in the high frequency tick-by-tick data in the market. The trading volume will follow some random distribution because in the efficient market hypothesis the market always follows a random walk. Therefore, we assume that if the volume is not so random anymore that there are some anomalies in the distribution. At that point the market is not efficient and this means the stock price is not a random walk anymore so a long term predicting strategy is possible. Here, we want to study whether using the detected anomalies from historical financial time series data can predict stock trend effectively or not.

Our contribution are as follows:

1. We first propose using anomalies on distribution of trading volume to predict upward trend of stock prices.
2. We use tick-by-tick data instead of time series data on stock price in a novel outlier mining algorithm.
3. We select 200 stocks randomly in our experiment. The result shows that using anomalies can predict the upward trend of stock prices effectively.

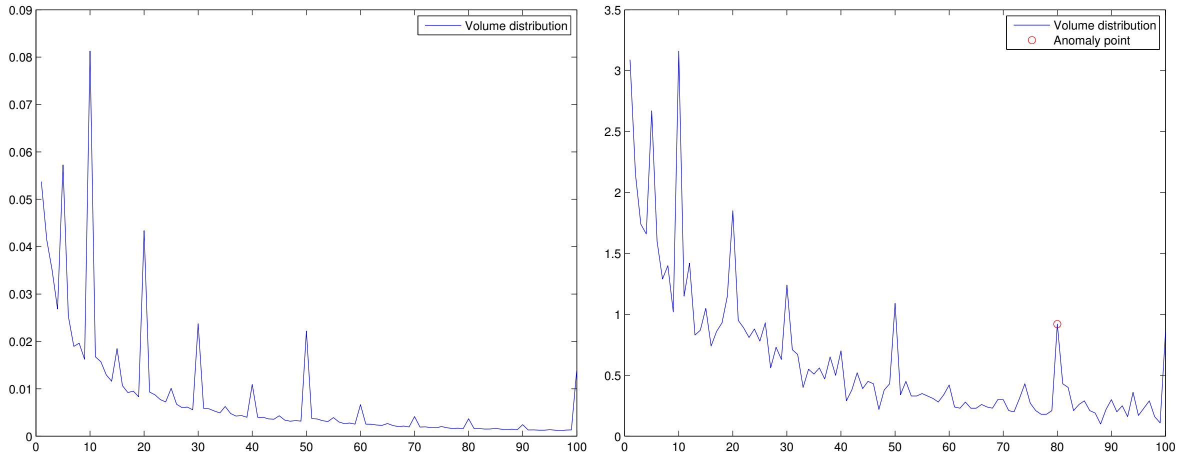
The rest of the paper is organized as follows. Section 2 introduces the motivation and provides an example to illustrate the problem. Section 3 introduces our approach and explain the outlier algorithm. Section 4 evaluates our approach by applying the method to the data to get the metrics. Section 5 gives some related works of our subject. Section 6 concludes this paper.

II. Motivation

Stock markets are changing all the time and prediction of stock trend is a significant issue in the modern financial market. However, according to the efficient market hypothesis [2], the market price will follow a random walk and a permanent prediction strategy is not possible. An interesting issue is that for some trading price that market is not efficient anymore in the real word, so it breaks the efficient market hypothesis. Therefore, the data of stock price will not be so random and prediction of stock trend becomes possible. A traditional way to predict the stock trend is using the data mining techniques on the basis of stock prices. Unfortunately, the data of stock price have many noises [1] and for noisy data people always build stochastic volatility models to make predictions whose efficiency is low.

In the above non-efficient case, when we analyze the volume data there are always anomalies in the distribution of trading volumes. Insider trading and market manipulation [7], [8] are the two key anomalies in stock market. Insider trading is the trades on the basis of non-public information by insiders, such as the directors, employees and officers [9], [10]. Market manipulation is the trades or actions that attempt to affect the fair and free operation of the stock market and create false or misleading appearance of a stock [11]. The anomalies will severely impair the stock market and obviously will in fact have long term influence on the stock prices. Thus, anomalies have the long term predictability on the stock trend, in our method we will utilize these anomalies to get rid of the effect of price noises to predict market trend. In this paper we will limit our scope on the upward trend prediction because an upward trend usually means stable and long term arbitrage opportunities.

CFP1552Z-CDR/15 $31.00 © 2015 IEEE
DOI 10.1109/BDCloud.2015.19



(a) Volume distribution of all the stocks

(b) Volume distribution of "Shenzhen development bank"

Fig. 1: Anomaly found in price 10.12 of Shenzhen development bank

Fig. 1 shows one example of anomaly. The left part of Fig. 1 is the trading volume distribution of all the stocks in the market at price 10.12. Compared with right part of the figure, which is the volume distribution of stock with the name "Shenzhen development bank" at price 10.12, we can find there's an anomaly on the volume 80 which is an outlier of the distribution and marked in circle.

In our approach, we detect all the anomalies and mark them on the price sequence. After that it is easy to predict that the stock trend changed dramatically when our approach clusters such anomalies. For example, in Fig. 2, the anomalies are marked with '+' on the price sequence. The horizontal axis is the index of the trade point in tick-by-tick data and the vertical axis is the stock price. We can see after the anomalies marked with '+' there's an obvious upward trend in the stock price. Next Section will introduce our approach in details.

III. Approach

An diagram of overview of our approach is shown in Fig. 3. We first fetch the data from data source then make a preprocessing to the data, after that we transform the high frequency data to a ratio matrix and then feed it into the outlier algorithm to find anomalies. We can then make predictions according to the position of the anomalies and evaluate the result.

A. Data Preprocessing

The data we use in this paper is the high frequency tick-by-tick trading data. Tick-by-tick data is a kind of format used frequently in financial industry. This data records each trade for every stock in the market, if there's 1000 trades for a specific stock then there will be 1000 records for that stock on that day, so for a relatively long period the data size can be very big. One record of the tick-by-tick data is defined as:

$$R = \{t, p, c, v, a, b\}$$

TABLE I: This table shows some data

Time	Price	Change	Volume	Amount	Bs
15:00:19	10.77	–	1785	1923500	b
14:57:01	10.77	–	1	1077	s
14:56:55	10.77	–	10	10770	b
14:56:52	10.77	-0.01	186	200322	s
14:56:46	10.78	–	94	101332	b
14:56:43	10.78	0.01	20	21560	b
14:56:43	10.77	-0.01	75	80775	s

For the record R each field have the means:

t	time of the trade
p	price of the trade
c	change of the price
v	volume of the trade
a	amount of the trade
b	buy or sell signal of the trade

One example of our tick-by-tick data is shown in Table I:

Once we have the data in hand, we can start to pre-process the data into the ratio matrix for later use. The following steps is what we need to do for this process.

1 Prepare tick-by-tick data

2 Fix a price for all the stocks

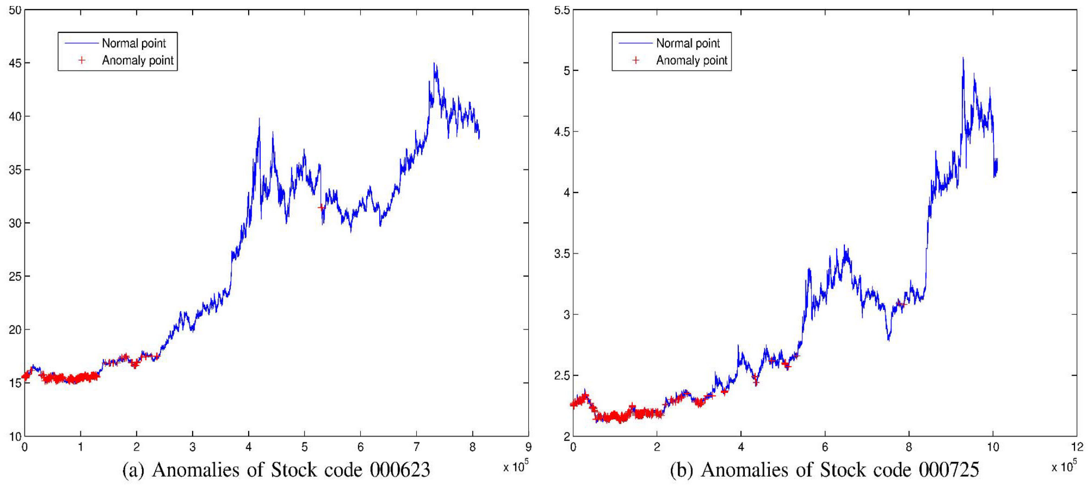
3 Fix a price for one specific stock

4 Make ratio matrix for step 2 and 3 for later use

Here's the explanation for each step:

Step 1: For each stock collect all the tick-by-tick data for all the trading day we want into a single matrix of records T. Each row of matrix T is a record R we defined earlier.

Step 2: We define a new vector by the following means from the matrix T:



(a) Anomalies of Stock code 000623

(b) Anomalies of Stock code 000725

Fig. 2: Anomalies of different stocks

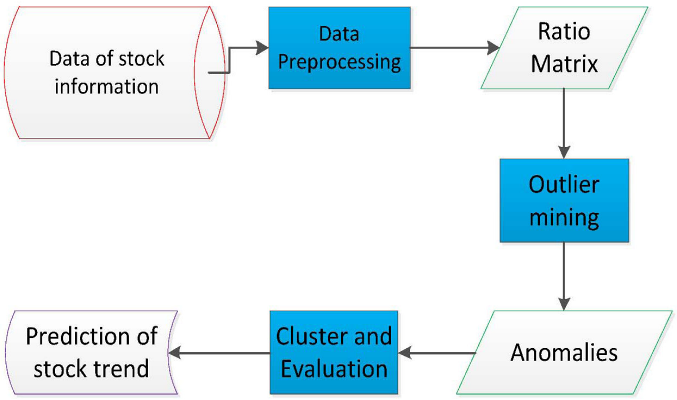


Fig. 3: An overview of our approach

$$T_s^v(p)$$

Where v means we only keep the volume column in T, s means the data is from a specific stock and p means we select the data at a specific price p. Then for all the stock in the market this vector becomes $T^v(p)$. This conclude step 2 and 3.

Step 4: For the volume vectors $T_s^v(p)$ and $T^v(p)$, they need to be lesser than a certain upper bound because if the volume is too big then the occurrence of them is rather rare and can be considered as noises. By this means we can save the calculation time significantly for the size of data is really big. In this paper the upper limit is 1000. Then we define a new price vector:

$$P = \{unique\ prices\}$$

P is consist of all the unique prices of the data we use. Then $p = P_i$ correspond to a specific price. With these quantities we can define the ratio matrix needed for the algorithm:

$$M_{si}^{\ v} = \frac{N(T_s^v(P_i))}{N(T_s^{(v<1000)}(P_i))}$$

and

$$M_i^v = \frac{N(T^v(P_i))}{N(T^{(v<1000)}(P_i))}$$

where i is the row index and v is volume and also the column index. N is the counting function that count the length of the T vectors. So the numerator is the length of the volume vector T of a specific volume number v at the price P_i and the denominator is the length of the vector for all the volume number at the same price P_i. And the same as T vectors, $M_{si}^{\ v}$ is the ratio matrix for a specific stock s while M_i^v is for all the stocks. A typical ratio matrix M looks like following:

Algorithm 1 : Anomaly price and volume finding algorithm

```
Input: M(i,:), Ms(i,:), prices
Output: Anomaly price, Anomaly volume
pseq:= unique(prices)
for (int i=1; i<length(pseq); i++) do
    pi:= pseq(i)
    theoryseq:=M(i,:)
    actualseq:=Ms(i,:)
    difference:=actualseq-theoryseq
    k:=find(difference>0.8)
    if k is not empty then
        an anomaly is found on price pi and volume number k
    end if
end for
return
```

Algorithm 2 : Anomaly location finding algorithm

```
Input: Anomaly price, Anomaly volume, T
Output: Anomaly position
price:=Anomaly price
volume:=Anomaly volume
index:=find(T(:,2)==price and T(:,4)==volume);
for (int i=1; i<=3; i++) do
    dif:=diff(index)
    indexDiff:=find(dif<5)
    index:=index(indexDiff+1)
end for
if index is not empty then
    An anomaly location is found
end if
return
```

$$M = \begin{pmatrix} 0.06 & 0.01 & 0.04 & 0.02 & 0.01 & 0.01 & 0.02 & \cdots \\ 0.04 & 0.01 & 0.05 & 0.06 & 0.02 & 0.01 & 0.01 & \cdots \\ 0.08 & 0.01 & 0.03 & 0.02 & 0.07 & 0.02 & 0.01 & \cdots \\ 0.01 & 0.01 & 0.03 & 0.05 & 0.03 & 0.07 & 0.02 & \cdots \end{pmatrix}$$

Each row corresponds to a price in the price sequence P while each column corresponds to a volume number in volume sequence T. Have all the data prepared we can start the explanation of our algorithm.

B. Algorithm

We propose an outlier miming algorithm to detect the anomalies of high frequency trading data in this paper. The detail of this algorithm is show in Algorithm 1.

In *Algorithm* 1, M is the matrix M_i^v while Ms is the matrix M_{si}^v, so iterate through index s we can find anomalies for all the stocks. What we did is we first retrieve the ratio sequence from M_i^v for the price P_i and compared it with M_{si}^v for the same price to find the difference of this two ratio sequence. If there's a value in the difference exceeds some certain limit then that value is considered to be an anomaly and the corresponding volume number and price is recorded. A typical anomaly record can be defined as following:

$$A = \{s, p, v\}$$

where s is the stock index,p is the anomaly price and v is the anomaly volume number. After all the anomalies of a stock are found we define another algorithm to locate those anomalies on the time of the trade. The implement of such algorithm is shown in *Algorithm* 2.

In *Algorithm* 2, T is the matrix of tick-by-tick records R introduced in step 1 of last subsection. This program first locate the anomaly trades in the record matrix T and use a simple method to determine if they are clustered on the row index in matrix T, and if they are then a cluster of anomalies is found and the following trend of the stock can be evaluated.

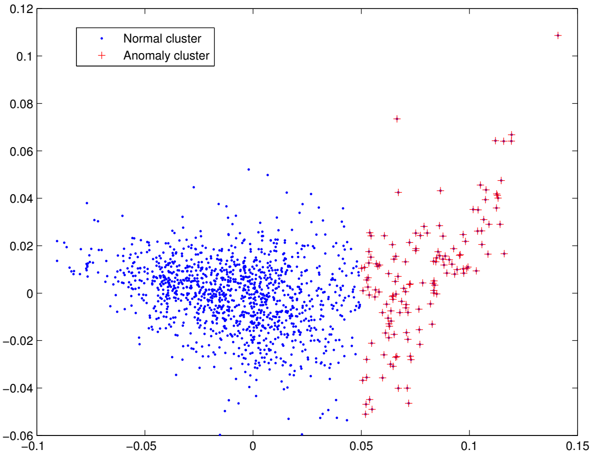


Fig. 4: Clustering of stock code 000623

IV. Experiment Results

The experiments on real exchange data have shown that our approach is more effective in prediction than the traditional data mining algorithm and the predictability of our approach is satisfactory. The experiment data are from Chinese stock exchange with the time range 03-31-2014 to 04-30-2015, which include 272 trading days. The size of the data set is 7.1 GB.

A. Clustering

In this part of experiment, we use k-means clustering algorithm to the rows of matrix M_{si}^v and check if any cluster exist. One of the typical result shows in Fig. 4.

Fig. 4 shows the clustering on the ratio matrix. The points are grouped into two clusters, one is marked in dot and one is marked in '+'. The cluster on the left represent the major part of the trading thus is the normal cluster, the cluster on the right represents the trading that different from normal then is considered to be the anomaly cluster. Recall that each row in matrix M_{si}^v correspond to a price P_i in the unique price sequence, so each cluster of rows of M_{si}^v is also a cluster of price.

Fig. 5 shows the histogram of the prices corresponding to the cluster who is marked with '+' in Fig. 4. The horizontal axis in Fig. 5 is the price and the vertical axis is the occurrences of that price. In Fig. 5, we observe that the price is clustered around 15. It is coincide with the anomaly cluster of stock 000623 in Fig. 2a.

This means our outlier mining algorithm is coherent with traditional cluster algorithms. But our algorithm is more effective in prediction that for other algorithms we need all the information in the data before we perform the algorithm, which means in order to found the anomaly cluster we also need the

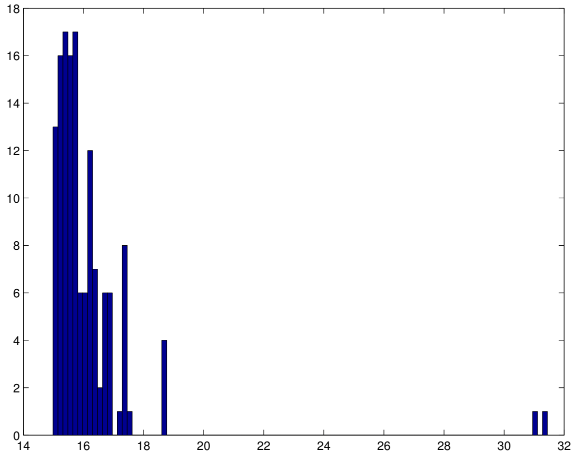


Fig. 5: Price histogram of the anomaly cluster for stock code 000623

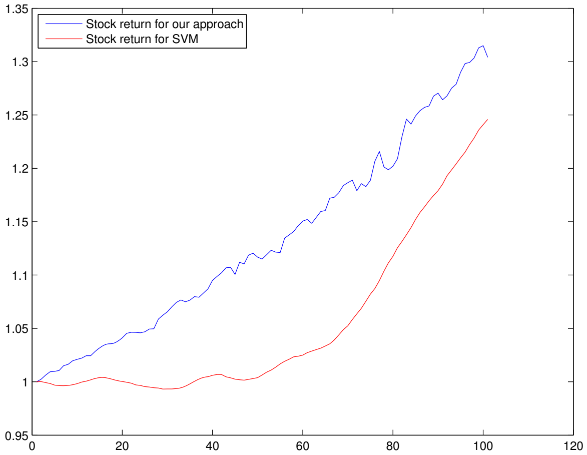


Fig. 6: Average return of 200 stocks

normal data in our dataset, thus it is not possible to make prediction if the normal data happens later than anomaly data while for our algorithm such issue doesn't exist that we can make prediction no matter where the anomaly data is.

B. Evaluate for prediction

We randomly chose 200 stocks in Chinese Shenzhen stock market and found 111 of them have the behavior of cluster of anomalies. We measured the average return of the stock after the anomaly cluster for 100 days and plot it in Fig. 6. For comparison we also measured the average return for the same set of stocks in the meaning of support vector machine [13] as in the related works section. The horizontal axis of Fig. 6 is the index of trading day starting from 03-31-2014, the vertical axis is the mean return of this 200 stocks compared with the price on the day 03-31-2014. We see that the upward trend is very obvious in the figure. We also measured the successful rate of our prediction. Successful rate is defined to be:

$$Successful\ Rate = \frac{number\ of\ correct\ predictions}{number\ of\ all\ predictions}$$

here a correct prediction means the stock return is bigger than 1.

Fig. 7 shows the successful rate. The horizontal axis of Fig. 7 is the index of trading day starting from 03-31-2014, the vertical axis is the successful rate of the 200 stocks on each day compared with the price on the day 03-31-2014. We observe that as time goes by the successful rate goes higher and close to 1 at last, which means almost all the stocks changed their trend after the anomaly clusters. The results in Fig. 6 and Fig. 7 show that the predictability of our approach is satisfactory.

V. Related Works

A. Neural Network Approaches

There are many researches using artificial neural networks (ANNs). A lot of successful trials have shown that ANN can be a powerful tool for time series forecasting and modeling [12]. However, too many factors required to be tuned would affect the ANN performance. It is a challenge to design the sampling schema, choose training and testing datasets and select the effective factors for improving the prediction performance and it is difficult to define the structures of the models such as the hidden layers, the neurons, etc. Zhang et al. [12] presented a piecewise nonlinear model to analyzing stock market tick data. They proposed Prop NN, which can improve the predictability of stock price. They claimed that it is significantly better than the basic BPN model. But as many of the other machine learning algorithms, ANN suffers from the problem of overfitting. It can not discriminate between useful information and noisy information and many of the time the noise level is too high that what the algorithm did is actually make a fitting on the noise, in this case the prediction on useful data is impossible. For our algorithm there's no such issue that we can simply ignore the noise and only pick up the useful information which is the anomaly volume, this will make the analysis much easier.

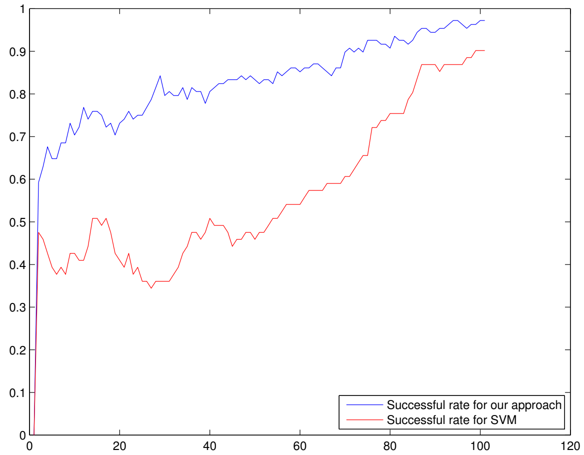


Fig. 7: Successful Rate of 200 Stocks

B. SVM based approaches

Support vector machine proposed by Boser et al [13] is attracting more attention these years. It is used as a clustering algorithm at first, derived from the structural risk minimization principle [14] and by separating the decision hyperplane it can also be used in classification and regression analysis, and can help users make well-informed business decisions. Wang et al. [15] showed that the K-means SVM (KMSVM) algorithm can speed up the response time of classifiers by decreasing the number of support vectors while maintaining a compatible accuracy to SVM. But the situation is the same of ANN algorithms that if the noise level is high then it is impossible to make prediction.

VI. Conclusion

In this paper, starting from the efficient market hypothesis we found a way to locate the anomaly trade data among the high frequency tick-by-tick data by comparing the distribution of volume sequence between the market and the specific stock. By making the volume distribution matrix of all the stocks in the market and any individual stock we can discover the difference between them and if that difference is bigger than a certain limit then an anomaly is found. We found that clusters of anomalies always predict an upward trend of the stock price. A traditional algorithm for cluster analysis is also possible to find the anomalies but our algorithm is more practical in that it is more effective in making predictions. We tested our novel outlier mining algorithm and found that it is consistent with k-means clustering algorithm. Finally the average return and successful rate is tested against our algorithm and the prediction about this two quantities is correct and satisfactory.

References

[1] Antoniou A, Vorlow C E. Price clustering and discreteness: is there chaos behind the noise?[J]. Physica A Statistical Mechanics & Its Applications, 2005, 348:389.

[2] Malkiel B G. The Efficient Market Hypothesis and Its Critics[J]. Journal of Economic Perspectives, 2003, 17(1):pgs. 59-82.

[3] Timmermann A, Granger C W J. Efficient market hypothesis and forecasting[J]. International Journal of Forecasting, 2004, 20(3):15C27.

[4] P. K. Padhiary and A. P. Mishra, Development of improved artificial neural network model for stock market prediction, International Journal of Engineering Science and Technology, Vol. 3, 2011, pp. 1576-1581.

[5] Amihud Y. Illiquidity And Stock Returns: Cross-Section And Time-Series Effects[J]. Social Science Electronic Publishing, 2002, 5:31-56. DOI:http://dx.doi.org/10.1016/S1386-4181(01)00024-6.

[6] Stein E 1, Stein J 2. Stock Price Distributions with Stochastic Volatility: An Analytic Approach[J]. Review of Financial Studies, 1991, volume 4(4):727-752(26).

[7] F. Allen and G. Gorton. Stock price manipulation, market microstructure and asymmetric information. European Economic Review, pages 624C630, 1992.

[8] M. Minenna. Insider trading abnormal return and preferential information: Supervising through a probabilistic model. Journal of Banking and Finance, pages 59C86, 2003.

[9] L. Cheng, M. Firth, T. Leung, and O. Rui. The effects of insider trading on liquidity. Pacific-Basin Finance Journal, pages 467C483, 2006.

[10] B. Cornell and B. Sirri. The reaction of investors and stock prices to insider trading. Journal of Finance, pages 1031C1059, 1992.

[11] K. Felixson and A. Pelli. Day end returns: Stock price manipulation. Journal of Multinational Financial Management,pages 95C127, 1999.

[12] G. Zhang, B. E. Patuwo, and M. Y. Hu, Forecasting with artificial neural networks: The state of the art, International Journal of Forecasting, Vol. 14, 1998, pp. 35-62.

[13] B. E. Boser, I. M. Guyon, and V. N. Vapnik, A training algorithm for optimal margin classifiers, in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, 1992, pp. 144-152.

[14] K.-J. Kim, Financial time series forecasting using support vector machines, Neurocomputing, Vol. 55, 2003, pp. 307-319.

[15] J. Wang, X. Wu, and C. Zhang, Support vector machines based on K-means clustering for real-time business intelligence systems, International Journal of Business Intelligence and Data Mining, Vol. 1, 2005, pp. 54-64.