STOCK MARKET DIRECTION PREDICTION USING DATA MINING CLASSIFICATION

Pujana Paliyawan
Independent Researcher, Thailand
E-Mail: Pujana.P@gmail.com

ABSTRACT

The key of success in stock trading is to buy and sell stocks at the right time for the right price. “Buy Low, Sell High” sounds easy, but it is so difficult to carry out since the direction of stock market in the near future is almost unpredictable. With the advances in data mining, it has now become possible to predict the future market direction based on historical data. In this study, different approaches are used to predict the future market direction of the Stock Exchange of Thailand (SET). Time series forecasting is conducted and a suitable span of time for the stock market data is examined. A novel approach to predict future market direction has been introduced based on chart patterns recognition by using data mining classification. Models are built through different methods including neural network, decision tree, naïve Bayes and k-nearest neighbors. Results were obtained, compared and discussed in details. Important chart patterns to support decision making in stock trading had been found out. In order to visualize the result, a visualization technique is also introduced.

Keywords: Investment, stock, technical analysis, time series forecasting, classification, pattern recognition, analysis.

INTRODUCTION

“You can’t create a duplicate of yourself to increase your working time; so instead, you need to send an extension of yourself—your money—to work [1].”

Investment is the science and art of growing money by putting money to work. Stock market has been a center of attraction for the investors for a long period of time. It historically provided the highest returns of any financial asset which was close to 10% over the long term [2]. In stock market, it is possible to make multiple returns as well as to lose the principle and go bankrupt. The major key to success is to buy and sell the stock at the right time for the right price

Basically, an investor makes a decision by measuring of company financial statements; for example, earnings per share (EPS), net profit margin, price-to-book (P/B) ratio, and price-to-earnings (P/E) ratio [2,3].

P/E ratio is one most commonly used measurement for assessing stock prices, it has been used in deciding if a stock is undervalued and should be bought or a stock is overvalued and should be sold, as well as in estimating the payback period on a given investment. A stock with a lower P/E ratio seems to be cheaper and more attractive; however, since the stock price also reflects the growth and investors’ future expectation in a company, the stock that is traded undervalued may imply an unbright company future. [2,3]. Moreover, the study on factual analysis of each cycle’s winning stocks had shown that P/E ratio was not an important cause of the most successful stock moves and it was recommended not to buy a stock solely because the P/E ratio looks cheap [3].

Similar to P/E ratio, many stock measurements can change drastically based on uncontrollable factors such as interest rates, investor sentiment, or government actions [3,4]. Even though there are several measurements to assess stock prices, in fact, factors on which buying or selling depend on cannot be quantified and it is nearly impossible to correctly predict a price direction in the near future [5, 6].

Several studies have been conducted to find out the way to predict the future stock price and the market directions. Several methods from financial analysis as well as data analysis were applied. In recent years, neural network (NN) become a widely method used for stock forecasting. Unfortunately, even results from NN models were appreciated in term of accuracy; many of them were not put into practice. This problem happened due to inability of neural networks to explain its reasoning; results were opaque to human interpretation [7]. In addition, proposed systems as well as discovered knowledge were not reachable to retail investors.

In this study, time series forecasting models are constructed to forecast the future market index by using a back propagation neural network (a wildly used approach). Models are built in several workflows in order to examine a suitable span of time for the stock market data.

After that, a novel approach for predicting future market direction is proposed based on chart patterns recognition by using data mining classification. Classification methods [8,9] include neural network, decision tree, naive Bayes and k-nearest neighbours; the model aims to predict whether the market index will go up, go down or stay in the next 1, 6 and 21 days. In addition, important chart patterns are found out and a visualization technique for visualizing them is introduced. The case study is applied to the Stock Exchange of Thailand (SET) [10].

STOCK MARKET

Investor types

There are two major types of investors which are value investor and speculator [1,2,4,11].

Value investors view stocks as its original purpose, a share of ownership of the company. They look for stocks those are traded at discounts to book value, have high dividend yields and low P/E ratio. Concerning on safety of principal and an adequate return, they generally hold stocks for a long period of time, as long as there is no change in the business fundamental.

On the other hand, speculators optimize their profit by taking risky financial transactions. They buy and sell stocks dynamically to speculate on the volatility of the stock price. Concerning less on the business fundamental or the dividend yield, they focus on the short term trend of the price using news and technical analysis.

Technical analysis

Technical analysis [1,3,12,13] is a process of forecasting the future direction of prices or market indexes through the study on historical data. While fundamental analysis focuses on the financial statements and potentials of the business (e.g. the health of the company, management and competitive advantages), technical analysis mainly focuses on the chart and quantitative data (e.g. price and volume) with the belief that the history tends to repeat itself.

Technical analysts use either computational or non-computational analysis. Statistics or analytical methods can be performed on the past market data to analyze the price direction (e.g. RSI, MACD, DMI, Stochastic oscillator) and to estimate support & resistance (e.g. Fibonacci retracement). Graphical methods can also be used instead of complex computation, and it is widely used due to simplicity. One well-known method called “Candlestick pattern” is to predict the price direction based on patterns observing of a candlestick chart. Another method called “Elliott wave theory” is to predict the movement of the stock market by observing and identifying a repetitive pattern of waves.

Stock market prediction theories

1. **Efficient Market Hypothesis (EMH)**

Efficient Market Hypothesis was an idea developed in the 1965 by Fama [14,15]. EMH states that the price of a security will reflect the whole market information. As soon as there is any information indicating that a security is underpriced and therefore offers a profit opportunity, investors will immediately buy it and its price will rise up to a fair value.

Fama had introduced three forms of financial market efficiency: weak, semi-strong and strong. In weak-form efficiency, future prices cannot be predicted by analyzing prices from the past; hence the technical analysis was unreliable. In semi-strong-form efficiency, the price will rapidly adjust to publicly available new information; neither fundamental analysis nor technical analysis techniques are able to reliably produce excess returns. And in strong-form efficiency, the price will reflect all information including insider information; no one can earn excess returns.

EMH had been criticized for years. No one knows how much efficiency the market is, but one acceptable context is that the price of a security is already reflected by related information about that security.

2.

3. **Behavioral finance**

Behavioral finance [16] explains why and how markets might be inefficient. It is to study the influence of psychological, social, cognitive, and emotional factors on the security and the markets. The historical patterns and the current market circumstance can psychologically affect the investor, and their respond will lead to a predictable market trend.

METHODOLOGIES

Time Series Forecasting (TSF)

Time series forecasting [8,17] is a process of analyzing time series data and predicting the future outcome. It uses statistical techniques to model and explain time-dependent series of data points. TSF is a widely used technique in financial forecasting.

Back Propagation Neural Network (BPNN)

Neural network (NN) [8] is a computational model that is capable of estimation (TSF) and pattern recognition (Classification). It is robust with respect to noisy and erroneous data, and is able to learn and adapt to the environment. Neural network is applicable to the problem that algorithm is indefinable or exhaustive search is infeasible. NN is in consideration as long as the optimal solution is the first priori, not the interpretable result.

In this work, the neural network is multi-layer perceptron (MLP) and uses a back propagation (BP) algorithm in its learning.

Data mining classification

1. **Decision tree (D-Tree)**

Decision tree [8,9] is a classification method which yields output as a flowchart-like tree structure. The result from D-Tree is highly interpretable, but the outcome must be represented in categorical data. In this work, D-Tree algorithm called “J48” is applied to classify future stock market direction.

2. **Naïve bayes**

Naïve Bayes [8,9] is a simple probabilistic classifier based on Bayes’ theorem, with a naive assumption of independence between every pair of features.

3. **k-Nearest neighbors (KNN)**

KNN [8,9] is a non-parametric lazy learning algorithm that predicts class of the object based on the k closest training examples in the feature space. An object is classified by a majority vote of its neighbors; the object will be assigned to the class most common amongst its k nearest neighbors (k = 15 in this study).

Evaluation methods for TSF

There were three methods used for evaluation of forecasting accuracy in this study, which are mean absolute error (MAE), root mean squared error (RMSE) and direction accuracy (DAC) [17].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - y_j\right| \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}_j - y_j\right)^2} \qquad (2)$$

$$DAC = \frac{Count\lfloor Sign(y_j - y_{j-1}) == Sign(\hat{y}_j - \hat{y}_{j-1})\rfloor}{n} \qquad (3)$$

$\hat{y}$ = predicted value of j
y = actual value of j

RELATED WORKS

In 2000, Kim and Han [6] mentioned that NN a limitation in patterns recognition since the stock market data had tremendous noise and complex dimensionality. Previous studies tried to optimize the controlling parameters of NN, yet they were less concern on reducing dimensionality and eliminating of irrelevant patterns. So they focused on feature discretization in NN by using genetic algorithms. The proposed model (GA approach to feature discretization (GAFD), outperformed other two conventional models, which were the linear transformation with BPNN (BPLT) and the linear transformation with NN trained by GA (GALT).

In 2001, Gencay and Min Qi [18] conducted a study to improve prediction accuracy in pricing and hedging derivative securities (S&P 500 index call options). It was mentioned that NN was accurate and computationally efficient, but it had a tendency to be overfitted when the data contain irrelevant information or noise. Bayesian regularization, early stopping, and bagging were applied for preventing overfitting. They found that feed forward networks with Bayesian regularization generated smaller pricing and delta-hedging errors than baseline NN models.

In 2003, Cao and Tay [19] conducted a study on financial time series forecasting by using support vector machine (SVM). The case study was applied to the Chicago Mercantile Market. Results from SVM, BPNN, and radial basis function neural network (RBFNN) were compared; they found that BPNN was likely to minimize the training error rather than the generalization error. It did not converge to global solutions and stuck in the local minima. Setting the training time also needed much care, otherwise it would not fully learn the complexity required for prediction and would result in overfitting. On the other hand, performances from SVM and RBFNN were similar and both of them were better than BPNN; they were robust to overfitting and provided a better generalization performance.

In 2009, Sutheebanjard and Premchaiswadi [20] proposed a method for predicting future SET index using a function that its coefficients were found by evolution strategies. They conducted experimental studies on short-term and long-term prediction, and then measure the result using MSE and Mean Absolute Percentage Error (MAPE) measures. The proposed method showed lower errors than well-known approaches.

In 2010, Sutheebanjard and Premchaiswadi [21] applied BPNN for TSF of SET index. Input data consisted of six attributes which were the SET index, the Dow Jones index, the Straits Times index, the Nikkei index, the Hang Seng index, Minimum Loan Rate from the Bank of Thailand and gold price data from Gold Trader Association. The result showed that BPNN returns an MSE of 243.68 and MAPE of 1.96% on the test data.

In 2011, Omidi et al. [22] studied on the performance of NN models in the future stock price prediction. Instead of the stock price, the stock prices increment ratio was used. Models were built through different techniques, and then results were compared by using MSE. They also provided logical results to support decision making, by generating suggestions for stock trading (i.e. buy, sell or do no action).

Olatunji et al. [23] proposed NN models to predict stocks' closing price in the Saudi stock market. Their model used only the closing price, and a training data set covered historical data with a large span of time (7 – 17 years approximately, 2132 – 3894 records). The optimal window of past days for prediction was investigated. NN models were built with different parameters (number of neuron, learning rate); results were compared by using correlation coefficient and RMSE.

In 2012, Iuhasz et al. [24] developed a hybrid system to compute stock price liquidity, predict next day price, trigger buy and sell signals, and forecast the direction of the market trend. The system includes 4 neural network methods along with technical and fundamental analysis. The case study was applied to the Bucharest Stock Exchange Market indexes.

Abhishek et al. [25] predicted a stock price by using BPNN model constructing by one year historical data. The network was successful in prediction the trends of stock market.

Yang, Gao and Zhao [26] proposed a system to support decision making in short-term stock trading. They improve traditional BPNN by using conjugate gradient method. The improved model had a faster convergence speed and higher precision in short-term stock decision.

In 2013, Khirbat, Gupta and Singh [27] constructed TSF models using BPNN. The optimal NN architecture was acquired through empirical studies on the number of epoch and neuron in the hidden layer.

Al-Radaideh et al. [28] presented an approach to help deciding the right time for buying and selling stocks based on knowledge extracted from the historical data. The predictive model was developed as a D-Tree classifier with CRISP-DM methodology. The classifier analysed 5 predictive attributes which are previous day close price and current day open, minimum, maximum and close prices of the stock. All attributes were transformed into discrete values (Positive, Negative and equal) by comparing the targeted price with the previous day closing price. The predicted outcome is called “Action,” which is a suggestion whether the stock should be bought or be sold. On the training dataset, Action is derived from how biggest brokers dealt with the above mentioned stocks in the past.

TIME SERIES FORECASTING

TSF models are constructed to forecast the future stock market index by using traditional BPNN. The input data is historical records of SET index. The lag length is determined by literature reviews and experiments. Five days lag length was suggested by Cao and Tay [19], and Thomason [29]. Through experiments, this number has been proven and the lag length is set at 5 in this study. So the model uses 5 previous days’ indices to predict the index in the next day.

Input data

The input data contained 2 attributes which are date and closing price. Several models are built by using data sets with a different span of time ranging from 20 years (20Y: 08/03/1994 – 08/03/2014) to 1 years (1Y: 08/03/2013 – 08/03/2014). The bigger data set (the wider span of time) tends to help the model in learning general market behaviors; on the other hand, using too old data may result in obsolete patterns recognition. This experiment is to find a suitable span of time for stock market data.

Results

The evaluation is performed by using MAE, RMSE and DAC. The training times are set at 500 and 5,000 epochs. Results are shown in Table-1. The models are more accurate when the number of epochs is high. MAE and RMSE are obviously lower on bigger data sets, indicates that a historical data set with a wider span of time can decrease the error rate of forecasting.

Table-1. Forcasting accuracy rate.

	Instance	500 Epochs			5000 Epochs		
		MAE	RMSE	DAC	MAE	RMSE	DAC
20Y Data	4895	25.5122	33.83	51.16	8.19	11.70	53.20
10Y Data	2445	22.14	28.27	50.76	20.89	27.18	50.96
5Y Data	1221	31.5113	39.184	51.6872	31.35	38.51	51.85
1Y Data	245	30.19	37.59	52.30	28.00	36.07	52.72

Figure-1 and Figure-2 show SET index forecasting on the most recent year (08/03/2013 – 08/03/2014), the model in Figure 1 is built on the 20Y data set over 5,000 epochs, when the model in Figure 2 is built on the 1Y data set over 500 epochs. The first model is the optimal model with the highest precision and the lowest error rate (MAE, RMSE). On the other hand, the relative between predicted values and actual values are clearly seen in the second model; the model cannot keep up with the market trend on the last 80 records where there is a political crisis in Thailand.

Anywise, the overall DAC is achieved at an average of 50% approximately, indicating that TSF by using traditional BPNN is not effective at SET market direction prediction (because randomizing whether the market will go up or go down already has 50-50 outcome probability).

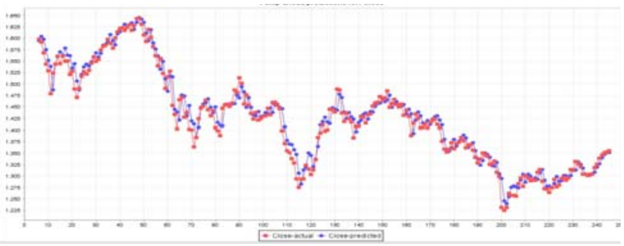


Figure-1. The most recent year forecasting by using model built on 20 years historical data.

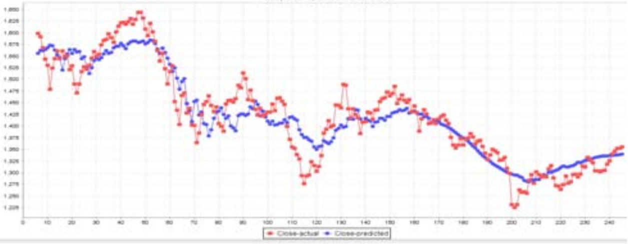


Figure-2. The most recent year forecasting by using model built on 1 year historical data.

PREDICTION OVER GRAPH PATTERNS

A novel approach for predicting future market direction is proposed by using data mining classification. The concept is derived from technical analysis—a chart pattern on previous days can be used to predict the future outcome. The workflow overview is shown in Figure-3.

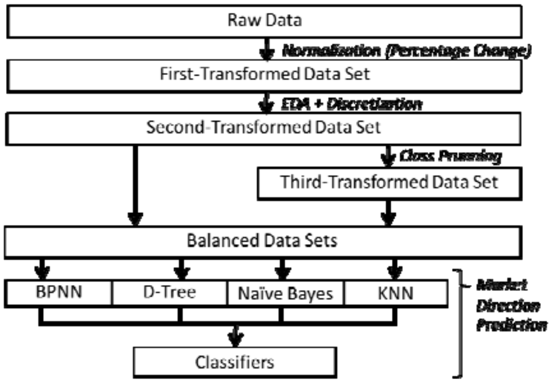

Raw Data
Normalization (Percentage Change)
First-Transformed Data Set
EDA + Discretization
Second-Transformed Data Set
Class Pruning
Third-Transformed Data Set
Balanced Data Sets
BPNN
D-Tree
Naïve Bayes
KNN
Classifiers
Market Direction Prediction

Figure-3. Workflow overview.

Data transformation I

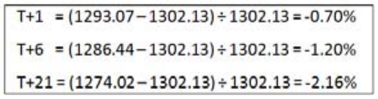
Twenty years historical data is transformed into a new data set, which consists of 8 attributes (5 predictor attributes and 3 target attributes). Predictor attributes are past 5 days’ indices and target attributes are the next 1, 6, and 21 stock business days’ indices (approximately to the next day, the next week, and the next month). Let’s T is an index on a given date, 8 attributes are calculated as percentage changes from those dates' indices to T as in equation (4).

The result is shown in Table-2. Each record is calculated as an example in Figure-4 and a result can be represented as in Figure-5. Note that the first 5 records and the last 31 records will be cut out.

$$Change\%_{T+k} = \frac{Closing\ index_{T+k} - Closing\ index_{T}}{Closing\ index_{T}} \quad (4)$$

Table-2. First-transformed data set.

#	Date	Close	T-5	T-4	T-3	T-2	T-1	T+1	T+6	T+21
1	8/3/2537	1368.03								
2	9/3/2537	1358.56								
3	10/3/2537	1332.39								
4	11/3/2537	1319.41								
5	14/3/2537	1296.55								
6	15/3/2537	1302.13	5.06%	4.33%	2.32%	1.33%	-0.43%	-0.70%	-1.20%	-2.16%
7	16/3/2537	1293.07	5.06%	3.04%	2.04%	0.27%	0.70%	-2.04%	-0.48%	-0.83%
8	17/3/2537	1266.64	5.19%	4.17%	2.36%	2.80%	2.09%	-1.56%	1.06%	1.06%
9	18/3/2537	1246.84	5.82%	3.99%	4.43%	3.71%	1.59%	-3.43%	-0.04%	3.75%
10	21/3/2537	1204.12	7.68%	8.14%	7.39%	5.19%	3.55%	2.74%	4.11%	6.12%
11	22/3/2537	1237.12	5.25%	4.52%	2.39%	0.79%	-2.67%	3.99%	0.83%	2.16%
12	23/3/2537	1286.44	0.52%	-1.54%	-3.08%	-6.40%	-3.83%	0.03%	-3.61%	-0.91%
13	24/3/2537	1286.81	-1.57%	-3.11%	-6.43%	-3.86%	-0.03%	-0.52%	-4.22%	-1.22%
14	25/3/2537	1280.07	-2.60%	-5.93%	-3.36%	0.50%	0.53%	-2.64%	-6.52%	-1.05%
15	28/3/2537	1246.34	-3.39%	-0.74%	3.22%	3.25%	2.71%	0.58%	-3.56%	0.98%
16	29/3/2537	1253.61	-1.32%	2.62%	2.65%	2.11%	-0.58%	-0.50%	-2.30%	-0.10%
17	30/3/2537	1247.33	3.14%	3.17%	2.62%	-0.08%	0.50%	-0.59%	-1.87%	-0.53%
18	31/3/2537	1239.99	3.78%	3.23%	0.51%	1.10%	0.59%	-0.60%	-0.38%	-1.39%
19	1/4/2537	1232.53	3.86%	1.12%	1.71%	1.20%	0.61%	-2.92%	4.03%	-0.37%
20	4/4/2537	1196.59	4.16%	4.77%	4.24%	3.63%	3.00%	0.45%	9.67%	3.27%
21	5/4/2537	1201.95	4.30%	3.78%	3.16%	2.54%	-0.45%	1.90%	6.00%	2.56%
22	7/4/2537	1224.74	1.84%	1.25%	0.64%	-2.30%	-1.86%	-0.06%	4.71%	2.81%
23	8/4/2537	1224.00	1.31%	0.70%	-2.24%	-1.80%	0.06%	0.92%	4.58%	5.46%
24	11/4/2537	1235.26	-0.22%	-3.13%	-2.70%	-0.85%	-0.91%	3.80%	4.73%	5.47%
25	15/4/2537	1282.16	-6.67%	-6.26%	-4.48%	-4.54%	-3.66%	2.35%	-0.33%	3.38%
26	18/4/2537	1312.33	-8.41%	-6.67%	-6.73%	-5.87%	-2.30%	-2.92%	-3.70%	1.68%
27	19/4/2537	1274.02	-3.87%	-3.93%	-3.04%	0.64%	3.01%	0.66%	0.05%	5.04%


T+1 = (1293.07 – 1302.13) ÷ 1302.13 = -0.70%

T+6 = (1286.44 – 1302.13) ÷ 1302.13 = -1.20%

T+21 = (1274.02 – 1302.13) ÷ 1302.13 = -2.16%

Figure-4. Calculation of the 6th record.

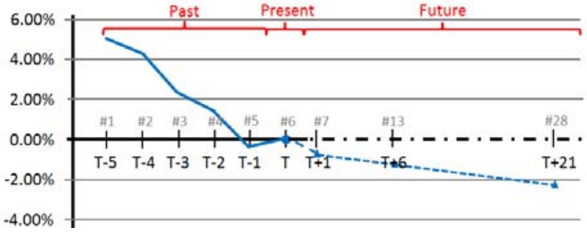


Figure-5. The graph representing the 6th record.

Exploratory data analysis (EDA)

In order to perform classification, target attributes must be discretized into categorical data. Before that, exploratory data analysis is conducted to understand more on the distribution of the data set. The results are shown Table-3 and Figure-6.

Table-3. Exploratory data analysis of target attributes.

	T+1	T+6	T+21
Mean	0.0001	0.0010	0.0042
Median	0.0000	0.0027	0.0067
SD	0.0164	0.0444	0.0904
Min	-0.1484	-0.2438	-0.3763
Max	0.1202	0.2404	0.2776

*** 1 = 100%*

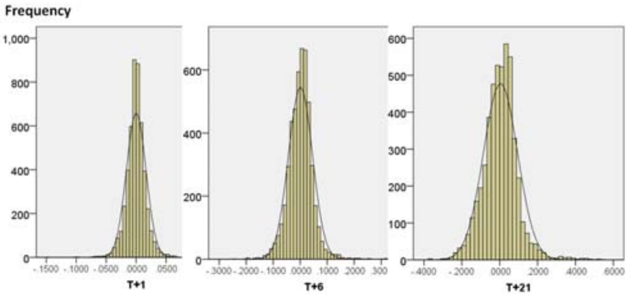


Figure-6. The normal distribution curve of target attributes.

The distribution of percentage changes for all future indices (T+1, T+6, T+21) are nearly normal. Five classes are declared to represent the future market direction including dramatically go down (Down2), go down (Down1), stay (Stay), go up (Up2) and dramatically go up (Up2). Threshold values are determined using SD as shown in Table-4.

Table-4. Threshold values indicating direction.

	Direction				
	Down2	Down1	Stay	Up1	Up2
T+1	(-∞, -1.6%]	(-1.6%, -0.8%]	(-0.8%, 0.8%)	[0.8%, 1.5%)	[1.6%, ∞)
T+6	(-∞, -4.4%]	(-4.4%, -2.2%]	(-2.2%, 2.2%)	[2.2%, 4.4%)	[4.4%, ∞)
T+21	(-∞, -9.0%]	(-9.0%, -4.5%]	(-4.5%, 4.5%)	[4.5%, 9.0%)	[9.0%, ∞)

Data Transformation II

From the first-transformed data set, values of T+1, T+6 and T+21 are replaced with Direction class labels. The result is shown in Table-5 and each record can be visualized as shown in Figure-7.

Table-5: Second-transformed data set.

#	T-5	T-4	T-3	T-2	T-1	T+1	T+6	T+21
6	5.06%	4.33%	2.32%	1.33%	-0.43%	Stay	Stay	Stay
7	5.06%	3.04%	2.04%	0.27%	0.70%	Down2	Stay	Stay

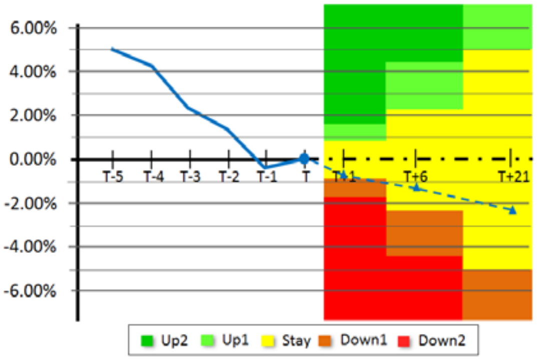
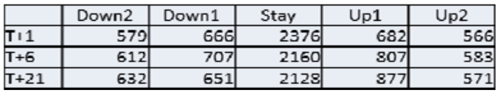


Figure-7. The graph representing the 6th record.

In second-transformed data set, there are 4,869 records and numbers of records for each class are shown in Table-6. The researcher then sorts data randomly and created 3 balanced data sets for predicting T+1, T+6 and T+21 separately. Each data consist of 2,500 records (500 for each class).

Table-6. Numbers of records for each class.

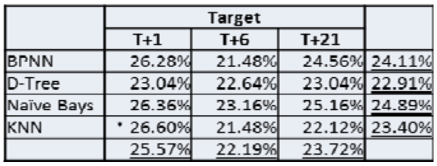


	Down2	Down1	Stay	Up1	Up2
T+1	579	666	2376	682	566
T+6	612	707	2160	807	583
T+21	632	651	2128	877	571

Market direction prediction (over 5 Classes)

From the second-transformed data set, the researcher performs data mining classifications by using several methods including BPNN (5,000 epochs), D-Tree, naïve Bayes and KNN (k = 15). Classifiers are built through 4 methods to predict 3 targets attributes; there are 12 classifiers in total. Training and testing are done using the 10-fold cross validation. The result is shown in Table-7.

Table-7. Classification accuracy rate of market direction prediction (5 classes).



	Target			
	T+1	T+6	T+21	
BPNN	26.28%	21.48%	24.56%	24.11%
D-Tree	23.04%	22.64%	23.04%	22.91%
Naïve Bays	26.36%	23.16%	25.16%	24.89%
KNN	* 26.60%	21.48%	22.12%	23.40%
	25.57%	22.19%	23.72%	

In overall, prediction seems to be most accurate when predicting tomorrow’s market index. The most accurate method is KNN for T+1 prediction with 26.60% accuracy, which is better than 5-class randomization by 33.33%. However, KNN will become the worst method when predicting T+6 and T+21. The following accurate methods are Naïve Bays and BPNN; both achieve the closed rate of accuracy.

Data transformation III

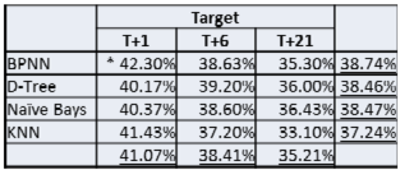
Another experiment is performed by pruning class labels from 5 classes to 3 classes. The process is done by merging Down1 with Down2 into Down, and merging Up1 and Up2 into Up. The result is called the “Third-transformed data set”.

Again, three balanced data sets are created for predicting T+1, T+6 and T+21. Each data consist of 3,000 records (1,000 for each class).

Market direction prediction (over 3 Classes)

The results from classification over 3 classes are shown in Table-8. The highest rate of accuracy rate is achieved when predicting tomorrow’s market index by using BPNN. The highest rate of accuracy is 42.30%, which is better than 3-class randomization by 26.91%.

Table-8. Classification accuracy rate of market direction prediction (3 classes).



	Target			
	T+1	T+6	T+21	
BPNN	* 42.30%	38.63%	35.30%	38.74%
D-Tree	40.17%	39.20%	36.00%	38.46%
Naïve Bays	40.37%	38.60%	36.43%	38.47%
KNN	41.43%	37.20%	33.10%	37.24%
	41.07%	38.41%	35.21%	

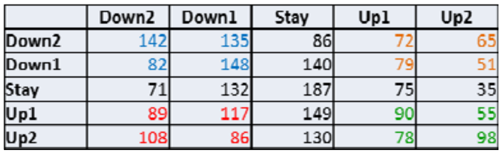
RESULTS AND DISCUSSION

According to the finding, an input of 5 previous days’ market indices will be most effectively used when predicting the next day’s index. KNN is considered preferable only in T+1 prediction, indicating that a type of instance-based learning (lazy learning) should be applied only if the prediction is lying in the very near future.

The overall performance of BPNN, D-Tree, and KNN are closed. To determine the optimal method, confusion matrixes are further looked into.

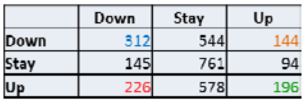
If the model is used for supporting decision making in stock trading; an investor will buy when the market index is predicted to go up, and will sell when the market index is predicted to go down. According to Table-9 and Table-10, the green color represents Right Buying, which results in capital gain, the red color represents Wrong Buying, which results in capital loss, the blue color represents Right Selling, and the orange color represents Opportunity Loss.

Table-9. Confusion matrix of KNN in T+1 prediction over 5 classes (Horizontal = Predicted, Vertical = Actual).



	Down2	Down1	Stay	Up1	Up2
Down2	142	135	86	72	65
Down1	82	148	140	79	51
Stay	71	132	187	75	35
Up1	89	117	149	90	55
Up2	108	86	130	78	98

Table-10. Confusion matrix of BPNN in T+1 prediction over 3 classes (Horizontal = Predicted, Vertical = Actual).



	Down	Stay	Up
Down	312	544	144
Stay	145	761	94
Up	226	578	196

Probabilities of each outcome are calculated by using formula 5 – 8; the results were shown in Table-11.

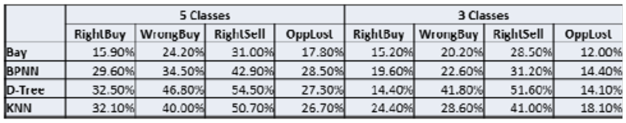
$$\text{Right Buying\%} = \frac{Up_{Predict} \cap Up_{Actual}}{Up_{Predict}} \quad (5)$$

$$\text{Wrong Buying\%} = \frac{Up_{Predict} \cap Down_{Actual}}{Up_{Predict}} \quad (6)$$

$$\text{Right Selling\%} = \frac{Down_{Predict} \cap Down_{Actual}}{Down_{Predict}} \quad (7)$$

$$\text{Oppotunity Loss\%} = \frac{Down_{Predict} \cap Up_{Actual}}{Down_{Predict}} \quad (8)$$

Table-11. Probabilities on T+1 prediction.



	5 Classes				3 Classes			
	RightBuy	WrongBuy	RightSell	OppLost	RightBuy	WrongBuy	RightSell	OppLost
Bay	15.90%	24.20%	31.00%	17.80%	15.20%	20.20%	28.50%	12.00%
BPNN	29.60%	34.50%	42.90%	28.50%	19.60%	22.60%	31.20%	14.40%
D-Tree	32.50%	46.80%	54.50%	27.30%	14.40%	41.80%	51.60%	14.10%
KNN	32.10%	40.00%	50.70%	26.70%	24.40%	28.60%	41.00%	18.10%

According to Table-11, the probabilities of Right Buying are totally lower than the probabilities of Wrong Buying in all models; the approach is not significantly useful to support decision making on buying stocks.

On the other hand, by comparing Right Sell with Opportunity Loss, the approach is very effective at supporting decision making on selling stocks. The investor can take this advantage in "Shorting"; to speculate on the downturn of the market. D-Tree with 3 classes is considered as the optimal model; when the model encourages an investor to perform shorting, there is 51.60% chance that an investor will gain benefits with only 14.10% chance of loss (and the left 34.30% is a chance of a draw).

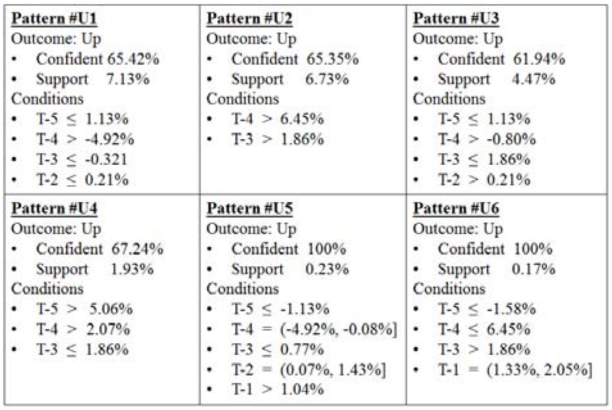
CHART PATTERNS DISCOVERY

It is considered that buying and selling stocks frequently is less meaningful than doing it once at the right time. Instead of using the model all the time and relying on the overall rate of accuracy, it is better to use the model when the confidence of prediction is high. With this principle, the overall rate of accuracy is not the matter; the emphasis is put on discovering valuable parts of the model.

One great advantage of D-tree is that the result is interpretable. So we will drill down into the model and find valuable patterns which could be easily used in a decision making.

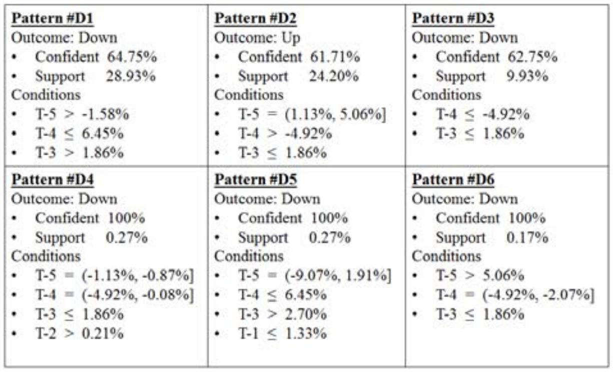
Patterns discovery

Examples of discovered patterns are shown in Figure-8 – 10. Confident is the accuracy rate of a pattern calculated by dividing the number of correctly classified instances with the number of total classified instances. Support is the frequency of occurrence calculated by dividing the number of total classified instances with the number of total instances (#instances = #records = 3,000).



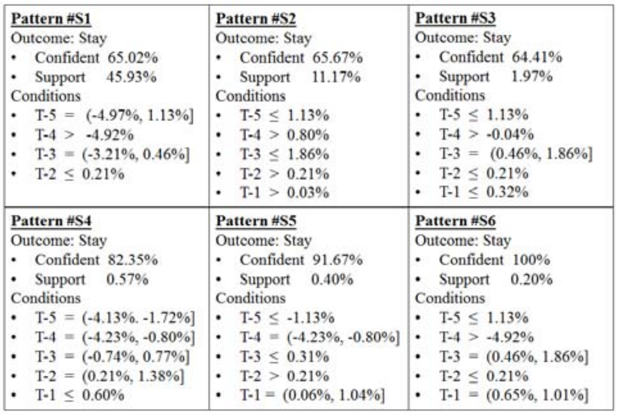
Pattern #U1 Outcome: Up • Confident 65.42% • Support 7.13% Conditions • T-5 ≤ 1.13% • T-4 > -4.92% • T-3 ≤ -0.321 • T-2 ≤ 0.21%	**Pattern #U2** Outcome: Up • Confident 65.35% • Support 6.73% Conditions • T-4 > 6.45% • T-3 > 1.86%	**Pattern #U3** Outcome: Up • Confident 61.94% • Support 4.47% Conditions • T-5 ≤ 1.13% • T-4 > -0.80% • T-3 ≤ 1.86% • T-2 > 0.21%
Pattern #U4 Outcome: Up • Confident 67.24% • Support 1.93% Conditions • T-5 > 5.06% • T-4 > 2.07% • T-3 ≤ 1.86%	**Pattern #U5** Outcome: Up • Confident 100% • Support 0.23% Conditions • T-5 ≤ -1.13% • T-4 = (-4.92%, -0.08%] • T-3 ≤ 0.77% • T-2 = (0.07%, 1.43%] • T-1 > 1.04%	**Pattern #U6** Outcome: Up • Confident 100% • Support 0.17% Conditions • T-5 ≤ -1.58% • T-4 ≤ 6.45% • T-3 > 1.86% • T-1 = (1.33%, 2.05%]

Figure-8. Bullish patterns.



Pattern #D1 Outcome: Down • Confident 64.75% • Support 28.93% Conditions • T-5 > -1.58% • T-4 ≤ 6.45% • T-3 > 1.86%	**Pattern #D2** Outcome: Up • Confident 61.71% • Support 24.20% Conditions • T-5 = (1.13%, 5.06%] • T-4 > -4.92% • T-3 ≤ 1.86%	**Pattern #D3** Outcome: Down • Confident 62.75% • Support 9.93% Conditions • T-4 ≤ -4.92% • T-3 ≤ 1.86%
Pattern #D4 Outcome: Down • Confident 100% • Support 0.27% Conditions • T-5 = (-1.13%, -0.87%] • T-4 = (-4.92%, -0.08%] • T-3 ≤ 1.86% • T-2 > 0.21%	**Pattern #D5** Outcome: Down • Confident 100% • Support 0.27% Conditions • T-5 = (-9.07%, 1.91%] • T-4 ≤ 6.45% • T-3 > 2.70% • T-1 ≤ 1.33%	**Pattern #D6** Outcome: Down • Confident 100% • Support 0.17% Conditions • T-5 > 5.06% • T-4 = (-4.92%, -2.07%] • T-3 ≤ 1.86%

Figure-9. Bearish patterns.



Pattern #S1 Outcome: Stay • Confident 65.02% • Support 45.93% Conditions • T-5 = (-4.97%, 1.13%] • T-4 > -4.92% • T-3 = (-3.21%, 0.46%] • T-2 ≤ 0.21%	**Pattern #S2** Outcome: Stay • Confident 65.67% • Support 11.17% Conditions • T-5 ≤ 1.13% • T-4 > 0.80% • T-3 ≤ 1.86% • T-2 > 0.21% • T-1 > 0.03%	**Pattern #S3** Outcome: Stay • Confident 64.41% • Support 1.97% Conditions • T-5 ≤ 1.13% • T-4 > -0.04% • T-3 = (0.46%, 1.86%] • T-2 ≤ 0.21% • T-1 ≤ 0.32%
Pattern #S4 Outcome: Stay • Confident 82.35% • Support 0.57% Conditions • T-5 = (-4.13%. -1.72%] • T-4 = (-4.23%, -0.80%] • T-3 = (-0.74%, 0.77%] • T-2 = (0.21%, 1.38%] • T-1 ≤ 0.60%	**Pattern #S5** Outcome: Stay • Confident 91.67% • Support 0.40% Conditions • T-5 ≤ -1.13% • T-4 = (-4.23%, -0.80%] • T-3 ≤ 0.31% • T-2 > 0.21% • T-1 = (0.06%, 1.04%]	**Pattern #S6** Outcome: Stay • Confident 100% • Support 0.20% Conditions • T-5 ≤ 1.13% • T-4 > -4.92% • T-3 = (0.46%, 1.86%] • T-2 ≤ 0.21% • T-1 = (0.65%, 1.01%]

Figure-10. Sideways patterns.

According to Figure-8, Pattern #U1 and #U2 are closed in terms of confident and support. But #U1 is considered more complex and more specific, as it has 2 more conditions than #U2. These patterns has frequency of occurrence at about 7%, which is high; once it occurs, an investor will has 65% chance of getting benefits by following this prediction.

Compared to #U1 and #U2, Pattern #U5 is rare, but every single occurrence of this pattern will lead to the bull market with 100% confident.

Pattern #S1 is an example of sideway pattern, it predicts that the market will experience neither an uptrend nor a downtrend. #S1 has 4 conditions and seems a little bit complex, but the support at 45.93% indicating that this pattern occurs almost every other day.

Patterns visualization

A visualization technique is introduced for visualizing discovered patterns. Pattern #U2, #U5, #D2 and #S1 can be visualized as shown in Figure 11. The colored areas represent the possibility area of input and output. Blue areas are past five days' indices and a purple area is the next day's predictive index. This visualization can be used by matching pattern as shown in Figure-12.

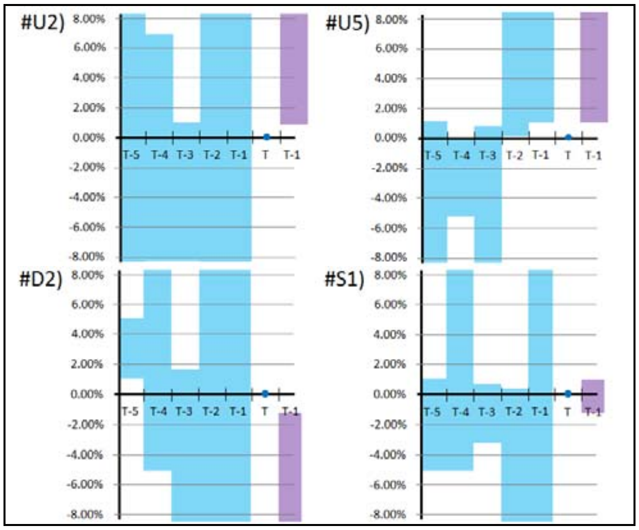


Figure-11. Visualized patterns.

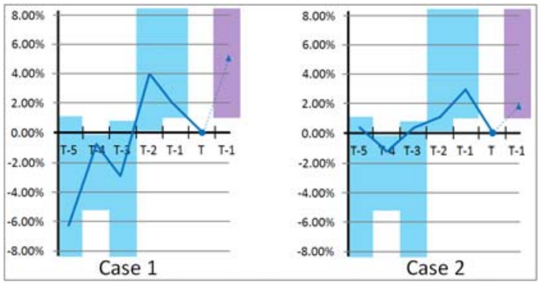


Figure-12. Cases those match pattern #U5.

CONCLUSIONS

Models for SET index forecasting have been constructed based on TSF by using traditional BPNN. It is found that a historical data set with a wider span of time can decrease the error rate of forecasting. Anywise, TSF is considered not effective at SET market direction.

A new approach for predicting the stock market direction is proposed by using data mining classification. Models are built through several techniques and results are discussed in details. The optimal model is achieved by D-Tree over 3-class classification; the model is highly effective in predicting market downturn. When it suggests an investor to perform shorting, there is 51.60% chance that an investor will gain benefits, with only 14.10% chance of loss.

To discover important patterns, a D-Tree model is drilled down into. Eighteen example patterns are presented, which any investor can practically use it to support decision making in trading. The visualization technique for visualizing these patterns is also provisioned.

ACKNOWLEDGEMENTS

The researcher thanks to Asst. Prof. Dr. Chakarida Nukoolkit and Asst. Prof. Dr. Bunthit Watanapa from King Monkut's University of Technology Thonburi and Dr.Anan Osothsilp for their valuable technical discussions. And thanks to Dr. Vivarad Phonekeo for the English proof-reading.

REFERENCES

[1] Investopedia. Available : http://www.investopedia.com [August 8, 2014].

[2] Kelly, J. (2010). The Neatest Little Guide to Stock Market Investing. Plume Publishers, Revised Edition, pp.51-75.

[3] O'Neil, W.J. (2013). How To Make Money In Stocks. McGraw-Hill Publishers, 2nd Edition, pp.18-21, 45-78.

[4] Graham, B. and Dodd, D. (2008). Security Analysis. McGraw-Hill Publishers, 6th Edition, pp.61-111.

[5] Achelis, S.B. (2013). Technical Analysis from A to Z. McGraw-Hill Publishers, 2nd Edition, pp.1-400.

[6] Kim, K. and Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert System Application, Vancouver, Canada, pp.125–132.

[7] Burke, L. and Ignizio, J.P. (1997). A practical overview of neural networks. Journal of Intelligent Manufacturing, 8, pp.157-165.

[8] Han, J., Kamber, M., and Pei, J. (2012), Data Mining Concepts and Techniques. Morgan Kaufmann Publishing, 3rd Edition.

[9] The University of Waikato. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available : http://www.cs.waikato.ac.nz/ml/weka/ [August 8,

2014].

[10] The Stock Exchange of Thailand. Available : http://www.set.or.th/en/index.html [August 8, 2014].

[11] Graham, B., Zweig, J. and Buffett, W. (2006). The Intelligent Investor. HarperCollins Publisers, Revised Edition, pp.18-34.

[12] Murphy, J.J. (1999). Technical Analysis Of The Financial Markets. NYIF, Revised Edition, pp.1-596.

[13] Astrophysics Group, the University of Cambridge. Technical Analysis. Available : http://www.mrao.cam.ac.uk/~mph/Technical_Analysis.pdf [August 8, 2014].

[14] Fama, E. (1965). The Behavior of Stock-Market Prices. Journal of Business, vol.38, no.1.

[15] Hill, McGraw. Chapter 8: The Efficient Market Hypothesis. Available : http://highered.mcgraw-hill.com/sites/dl/free/007338240x/773409/Sample_Chapter_8_New.pdf [August 8, 2014].

[16] Dehnad, K. (2011). Behavioral Finance and Technical Analysis. Journal of Financial Transformation, Capco Institute, vol.32, pp.107-111.

[17] Pentaho Data Mining Community Documentation. Time Series Analysis and Forecasting with Weka. Available : http://wiki.pentaho.com/display/datamining /Time+Series+Analysis+and+Forecasting+with+Weka [August 8, 2014].

[18] Gencay, R. and Qi, M. (2001). Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. Neural Networks, IEEE Transactions, vol.12, no.4, pp.726-734.

[19] Cao, L.J. and Tay, F.E.H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. Neural Networks, IEEE Transactions, vol.14, no.6, pp.1506-1518.

[20] Sutheebanjard, P. and Premchaiswadi, W. (2010). Stock Exchange of Thailand Index Prediction Using Back Propagation Neural Networks. Computer and Network Technology (ICCNT), 2nd International Conference, pp.377-380.

[21] Sutheebanjard, P. and Premchaiswadi, W. (2009). Factors analysis on Stock Exchange of Thailand (SET) index movement. ICT and Knowledge Engineering, 7th International Conference, pp.69-74.

[22] Omidi, A., Nourani, E. and Jalili, M. (2011). Forecasting stock prices using financial data mining and Neural Network. Computer Research and Development (ICCRD), 3rd International Conference, vol.3, pp.242-246.

[23] Olatunji, S.O., Al-Ahmadi, M.S., Elshafei, M., and Fallatah, Y.A. (2011). Saudi Arabia stock prices forecasting using artificial neural networks. Applications of Digital Information and Web Technologies (ICADIWT), 4th International Conference, pp.81-86.

[24] Luhasz, G., Tirea, M. and Negru, V. (2012). Neural Network Predictions of Stock Price Fluctuations. Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 14th International, pp.505-512.

[25] Abhishek, K., Khairwa, A., Pratap, T. and Prakash, S. (2012). A stock market prediction model using Artificial Neural Network. Computing Communication & Networking Technologies (ICCCNT), 3rd International Conference, pp.1-5.

[26] Yang, Y., Gao, S. and Zhao, Y. (2012). A novel stock decision model based on BP neural network. Intelligent Control and Information Processing (ICICIP), 3rd International Conference, pp.317-321.

[27] Khirbat, G., Gupta, R. and Singh, S. (2013). Optimal Neural Network Architecture for Stock Market Forecasting. Communication Systems and Network Technologies (CSNT), International Conference, pp.557-561.

[28] Al-Radaideh, Q.A., Assaf, A.A. and Alnagi, E. (2013). Predicting Stock Prices Using Data Mining Techniques. The International Arab Conference on Information Technology (ACIT'2013).

[29] Thomason, M. (1999). The practitioner methods and tool. J. Comput. Intell. in Finance, vol.7, no.3, pp.36-45.