Predicting the Price of Tehran Stock Market Using Data Mining Algorithms

Mojtaba Zamani Noghabi1, Maryam Ashoori2, Mohsen Farshad3
1- Department of Electrical Engineering, Islamic Azad University, Gonabad, Iran. Email: noghab2@chmail.ir (Corresponding author)
2- Department of Engineering, Higher Educational Complex of Saravan, Saravan, Iran. Email: maryam.ashoori@gmail.com
3- Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. Email: mfarshad@birjand.ac.ir

Received: October 2014 Revised: October 2014 Accepted: November 2014

ABSTRACT:
Ability to predict direction of stock/index price accurately is crucial for market dealers or investors to maximize their profits. Data mining techniques have been successfully shown to generate high forecasting accuracy of stock price movement. Nowadays, instead of a single method, traders need to use various appropriate techniques to gain more information about the future of the markets. In this paper, three different techniques of data mining are discussed and applied to predict price of Tehran stock market. The approaches include decision trees (CART and CHAID) and Neural Network that execute on Saipa, Iran Khodro, Telecommunication, Mapna and Saderat Bank datasets. The aim of this paper is generating an effective predicting model to forecasting future price in Tehran stock market. Price prediction in stock market helps investors for exact and quick identification and more investment on valuable portion. Finally it causes portion basket of investors optimized. Results show that the Neural Network is suitable for producing industries and CART is suitable for services industries. Next, many rule induction that achieved CART modeling, discussed.

KEYWORDS: Price Prediction, Stock Market, Neural Network, CHAID, CART, Rule Induction.

1. INTRODUCTION
Many financial companies such as stock markets produce large datasets and are looking to find efficient ways to discover useful information about stocks and the market for investment decisions [1]. Financial market is a complex, non-stationary, noisy, chaotic, non-linear, non-parametric and dynamic system but it does not follow random walk process [2]. Due to that is extremely hard to model with any reasonable accuracy. Investors have been trying to find a way to predict stock prices and to find the right stocks and right timing to buy or sell [3]. Further, with the easy access to stock information and data, many private investors worldwide are interested in predicting stock prices and hope to maximize on the opportunities in the market and become rich[1].

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support companies and individuals in strategic planning and investment decision making [4]. Todays, large volume of available data is an opportunity but existing knowledge on these data bases are hidden and don't use . Therefore, data mining tools are used for hidden information discovery. Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical method [5].

Recently, data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area [3]. Data mining has been used in stock market to make predictions regarding trends and prices to gain maximum profits [6]. Data mining techniques have been successfully shown to generate high forecasting accuracy of stock price movement [2].

The main objective of this paper is to analyze the historical data available on stocks using decision
trees (CHAID\(^1\) and CART\(^2\)) and neural network
technique as the classification methods of data
mining in order to help investors to predict stock
prices. Data mining techniques have been used to
uncover hidden patterns and predict future trends
and behaviors in financial markets. The competitive
advantages achieved by data mining include
increased revenue, reduced cost, and much
improved marketplace responsiveness and awareness
[4].

In the first section of this paper, we introduce the
concepts of stock market and large dataset as
opportunity. Section 2 will introduce data mining and
section 3 includes the implementation data mining on
stock market data. Section 4 will present results and
discussion. Next, section 5 will prepare conclusions.

2. DATA MINING

Data mining is a creative process which requires
a number of different skills and knowledge. In the
market, there is still to some extent the expectation
that data mining is a push-button technology [7].
Data mining refers to extracting or mining knowledge
from large data stores or sets. Some of its functionalities
are the discovery of concept or class descriptions,
associations and correlations, classification, prediction,
clustering, trend analysis, outlier and deviation analysis,
and similarity analysis. Data classification can be done
in many different methods; one of those methods is the
classification by using Decision Tree. Decision trees
and artificial neural networks can be trained by using
an appropriate learning algorithm [3].

2.1. Decision Trees

Decision trees are a form of multiple variable
analyses and are powerful and popular tools for
classification and prediction. The attractiveness of
decision trees lies in their ease of interpretation,
relative power, robustness with a variety of data and
levels of measurement, and ease of use. Decision
trees attempt to find a strong relationship between
input values and target values in a group of
observations that form a data set. In contrast to
neural networks, decision trees represent rules and rules
can readily be expressed so that humans can understand
them or even directly used in database access language
like SQL so that records falling into particular category
may be retrieved [1].

2.2. Neural Network (NN)

A neural network model takes an input vector X and
produces and output vector Y. The relationship between
X and Y is determined by the network architecture.

There are many forms of network architecture (inspired
by the neural architecture of the brain). The neural
network generally consists of at least three layers: one
input layer, one output layer, and one or more hidden
layers [8]. A typical neural network consists of several
neurons arranged in layers to create a network. Each
neuron can be thought of as a processing element that is
given a simple part of a task. The connections between
the neurons provide the network with the ability to learn
patterns and interrelationships in data [9]. The Figure 1
gives a simple representation of a neural network (a
multi-layer perceptron).

Fig. 1. Simple Representation of a Common Neural
Network

3. DATA MINING ON STOCK MARKET DATA
SETS

3.1. Data Understanding

This study was done by descriptive-cross sectional
method. It was done by Census sampling method and
contains 5 companies from 2001 to 2014. Table 1 shows
the exact time that company's data recorded in Tehran
stock market.

Table 1. Details of Company's Data

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Start Date</th>
<th>Finished Date</th>
<th>Records Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saipa</td>
<td>2001/06/05</td>
<td>2014/04/05</td>
<td>2727</td>
</tr>
<tr>
<td>Iran Khodro</td>
<td>2001/25/03</td>
<td>2014/04/05</td>
<td>2622</td>
</tr>
<tr>
<td>Telecommunication</td>
<td>2008/09/08</td>
<td>2014/04/05</td>
<td>1184</td>
</tr>
<tr>
<td>Mapna</td>
<td>2007/26/08</td>
<td>2014/04/05</td>
<td>1299</td>
</tr>
<tr>
<td>Saderat Bank</td>
<td>2009/09/06</td>
<td>2014/04/05</td>
<td>1057</td>
</tr>
</tbody>
</table>

Table 2 shows attributes and brief description of them.

Table 2. Stock Market Data Set

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Date</td>
</tr>
<tr>
<td>First Price</td>
<td>Numeric</td>
</tr>
<tr>
<td>High Price</td>
<td>Numeric</td>
</tr>
<tr>
<td>Low Price</td>
<td>Numeric</td>
</tr>
<tr>
<td>Close Price</td>
<td>Numeric</td>
</tr>
<tr>
<td>Value</td>
<td>Numeric</td>
</tr>
<tr>
<td>Volume</td>
<td>Numeric</td>
</tr>
<tr>
<td>Open</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

\(^1\) CHI-squared Automatic Interaction Detection
\(^2\) C&R Tree
3.2. Modeling

Selecting the appropriate algorithm with considering the under review data sets is essential. Therefore, authors surveyed available algorithms. Clementine contains a number of different algorithms for performing rule induction: C5.0, CHAID, QUEST, CART and Decision list. They are similar in that they can all construct a decision tree by recursively splitting data into subgroups defined by the predictor fields as they relate to the outcome. They differ in several ways that are important to users.

- Type of Output: C5.0, QUEST and Decision list only use symbolic output fields (of type set or ordered set) and Decision list only predicts a binary outcome. CART and CHAID support both symbolic and numeric outputs.
- Type of Split: CART and QUEST support only binary (two groups) splits while CHAID, C5.0 and Decision list support splits with more than two subgroups.
- Criterion for Selection of Predictor: the algorithms differ in the criterion used to drive the splitting. For C5.0 an information theory measures is used: the information gain ratio. When CART uses a dispersion measure (the Gini coefficient by default), CHAID and QUEST use a chi-square test and Decision list uses the statistical confidence.
- Handling of Missing Predictor Values: All algorithms allow missing values for the predictor fields, although they use different methods. C5.0 uses a fractioning method, CART and QUEST use substitute prediction fields, CHAID and Decision list make the missing values a separate category.
- Building Trees Interactively: CHAID, QUEST and CART support the ability to build the tree interactively, level by level, including selecting specific variables for a split. Decision list provides the ability to create one’s own rules.
- Growing Large Trees and Pruning: As Decision trees grow large and busy, the percentage of cases that passes through any given path in the tree decreases. QUEST, C5.0 and CART grow large trees and then prune them back, a method found to be effective. However, they differ in their pruning criteria.
- Rule sets: All the algorithms can represent a model as a rule set for a symbolic output. Rule sets can be easier to interpret than complex decision trees. However, a decision tree produces a unique classification for each data record, while more than one rule from a rule sets may apply to a record, which adds complexity [10].

With considering the above differences between decision trees, authors select the CART and CHAID decision trees because both of them support numeric outputs (output variable in this case is close price that it is numeric). Also these algorithms build the tree interactively, level by level.

After that, Clementine 12.0 software used for generating the model with aim of predicting the stock market price. For this prediction, authors select the close price as output variable and other variables as input. After that, CART, CHAID and neural network was executed on available data sets. Neural network used for modeling because of previous research done using neural network and authors want to compare the accuracy of decision tress against neural network. Also, 70 percent of data considered as train and 30 percent as test.

4. RESULTS AND DISCUSSION

It is clear, while the model accuracy did not determine, judge about reliability of model was not possible [11]. Table 3 (appendix) shows the model accuracy for CART, CHAID and neural network algorithms. In all algorithms, for numeric output fields, model accuracy shows the following information:

- Minimum Error. Shows the minimum error (difference between observed and predicted values).
- Maximum Error. Shows the maximum error.
- Mean Error. Shows the average (mean) of errors across all records. This indicates whether there is a systematic bias (a stronger tendency to overestimate than to underestimate or vice versa) in the model.
- Mean Absolute Error. Shows the average of the absolute values of the errors across all records. Indicates the average magnitude of error, independent of the direction.
- Standard Deviation. Shows the standard deviation of the errors.
- Linear Correlation. Shows the linear correlation between the predicted and actual values. This statistic varies between −1.0 and 1.0. Values close to +1.0 indicate a strong positive association, such that high predicted values are associated with high actual values and low predicted values are associated with low actual values. Values close to −1.0 indicate a strong negative association, such that high predicted values are associated with low actual values, and vice versa. Values close to 0.0 indicate a weak association, such that predicted values are more or less independent of actual values.
- Occurrences. Shows the number of records used in the analysis [12].

Table 3 shows neural network has good results on Saipa, Iran Khodro and Mapna datasets and CART has good results on Telecommunication and Saderat Bank. In Neural network's modeling, model has 7 neurons in input layer, 3 neurons in hidden layer and 1 neuron in output layer. CART is a decision tree algorithm and creates rule induction as bellow for price prediction in Telecommunication.
If $1082 < \text{high price} \leq 2059.5$ and $1679 < \text{low price}$ and $\text{first price} \leq 1875$ Then close price is $1816.28$.

If $2059.5 < \text{high price} \leq 2450.5$ and $\text{low price} \leq 2197.5$ and $2123.5 < \text{open}$ Then close price is $2200.758$.

Figure 2 to Figure 5 confirm achieved results in Table 3.

Estimated accuracy for Saipa, Iran Khodro, Telecommunication, Mapna and Saderat Bank are $99.659$, $99.847$, $99.861$, $99.661$ and $99.923$ orderly. Finally, results of modeling shows one algorithms is not effective for all datasets. Due to, researchers must be select the appropriate algorithm for predicting the stock market price.

Wei and et.al use a hybrid ANFIS model based on $n$-period to forecast TAIEX stock [13] and Lee and Lim used neural network for proposing financial time-series forecasting in stock [14]. References [8], [15], [16], [17] used neural network for predicting prices in market or stock. Also, Hargreaves and Hao used CHAID decision tree and neural network for predicting the stock performance [1]. Approximately all of researchers used neural network techniques for prediction in stock but results show neural network is not appropriate for all aspects in stock market. Due to, authors try to show decision trees have good output on stock and can apply for predicting.

5. CONCLUSION

Many financial companies such as stock markets produce large datasets and are looking to find efficient ways to discover useful information about stocks and the market for investment decisions. Data mining tools are used for hidden information discovery from large datasets but the main point in data mining is selecting the appropriate algorithm because the results of this research show that one algorithm did not suitable for predicting stock market price in all industries. Results show the neural network had good output on producing industries such as Saipa, Iran khodro and Mapna and CART had good output on services industries such as Telecommunication and saderat Bank.
6. APPENDIX

Table 3. Models Accuracy for Available Datasets

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Sajda</th>
<th>Iran Khodro</th>
<th>Telecommunication</th>
<th>Mapna</th>
<th>Soderat Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm Name</td>
<td>CART</td>
<td>CHAID</td>
<td>NN</td>
<td>CART</td>
<td>CHAID</td>
</tr>
<tr>
<td>Minimum Error</td>
<td>Train</td>
<td>Test</td>
<td>Min</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>3486.327</td>
<td>4043.327</td>
<td>1.00</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Maximum Error</td>
<td>Train</td>
<td>Test</td>
<td>Min</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>4626.765</td>
<td>2975.765</td>
<td>1.00</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Mean Error</td>
<td>Train</td>
<td>Test</td>
<td>Min</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>163.00</td>
<td>272.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>Train</td>
<td>Test</td>
<td>Min</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>-0.14</td>
<td>-0.14</td>
<td>1.00</td>
<td>0.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>

REFERENCES


