Research Paper Computer Science

Implementation of Data Mining Techniques for Extrapolative Stock Prices Using Crisp-Dm Methodology

Dr. Kishor Atkotiya Associate Professor& Head of the Department of Computer Science, J H Bhalodia Women's College Rajkot

Malam Bindu K. Assistant Professor, Department of Computer Science, Shri K.M. Savjani College Veraval,

ABSTRACT

Estimating stock and stock price return is an important financial subject that has attracted researcher's attention for many years. It involves an assumption that fundamental information publicly available in the past has some predictive relationships to the future stock returns. This paper tries to help the investors in the stock market to decide the better timing for buying or selling stock based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on decision tree classifier, Classification, Clustering and Association rule which are these data mining techniques. To Extrapolative of Stock prices, the CRISP-DM methodology is used over real historical data of National Stock Exchange (NSE).

KEYWORDS Data Mining, Decision Tree, CRISP-DM methodology.

1. INTRODUCTION

The stock market is essentially a non-linear, non parametric system 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. To achieve those objectives, and according to some research used the techniques of fundamental analysis, where trading rules are developed based on the information associated with macroeconomics, industry, and company. Some authors are said that fundamental analysis assumes that the price of a stock depends on its intrinsic value and expected return on investment. Analyzing the company's operations and the market in which the company is operating can do this. Consequently, the stock price can be predicted reasonably well. Most people believe that fundamental analysis is a good method only on a long-term basis. However, for short- and medium term speculation, fundamental analysis is generally not suitable.

Recently, data mining techniques and artificial intelligence techniques like decision trees and Association rule, and artificial neural networks have been applied to this area. 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 departure analysis, and similarity analysis. Data classification can be done in many different methods; one of those methods is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which they may be reached.

Today, the grand challenge of using a database is to generate useful rules from raw data in a database for users to make decisions, and these rules may be hidden deeply in the raw data of the database. Traditionally, the method of turning data into knowledge relies on manual analysis; this is becoming impractical in many domains as data volumes grow exponentially. The problem with predicting stock prices is that the volume of data is too large and huge. This paper uses one of the data mining methods; which are the classification approach on the historical data available to try to help the investors build their decision on whether to buy or sell that stock in order to achieve profit.

The main objective of this paper is to analyze the historical data available on stocks using decision tree technique as one of the classification methods of data mining in order to help investors to know when to buy new stocks or to sell their stocks.

Analyzing stock price data over several years may involve a few hundreds or thousands of records, but these must be selected from millions. The data that will be used in this paper to build the decision tree will be the historical prices of three listed companies in National Stock Exchange over two years of time.

2. THE CRISP-DM METHODOLOGY USED IN CLASSIFICATION MODEL:-

Data mining methodology is designed to ensure that the data mining effort leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodologies have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements. To build the model that analyses the stock trends using the decision tree technique, the CRISP-DM (Cross-Industry Standard Process for data mining) is used. This model consists of the following six steps:

- Understanding the reason and objective of mining the stock prices and Understanding the reason and objective of building the model.
- Understanding the collected data and how it is structured.
- Preparing the data that is used in the classification model.
- Selecting the technique to build the model.
- Evaluating the model by using one of the well known evaluation methods.
- Deploying the model in the stock market to predict the best action to be taken, either selling or buying the stocks.

The main reason and objective of building the model is to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on one of the data mining techniques; the decision tree classifiers.

2.1 Understanding the collected data

FoxPro and Excel Data sheet of National Stock Exchange (NSE) contains the historical prices of more than 1,582 companies listed from on its platform in the year 2015. As the amount of such data is very large and complicated, the decision was tak-

en to choose three companies listed in the exchange. These Companies are "20MICRONS", "3IINFOTECH" and "3MINDIA". The period that was selected is from April 2010 to May 2012, which presented the current and actual status of the market at that period of time.

Table 1 shows the 6 attributes selected with their descriptions and their possible values.
Table 1: Attribute Description

Attribute	Description	Possible Values
Previous	Previous day close price of the stock	Positive, Negative, Equal
Open	Current day open price of the stock	Positive, Negative, Equal
Min	Current day minimum price of the stock	Positive, Negative, Equal
Max	Current day maximum price of the stock	Positive, Negative, Equal
Last	Current day close price of the stock	Positive, Negative, Equal
Action	The action taken by the investor on this stock	Buy, Sell

2.2 Preparing the data

At the beginning, when the data was collected, all the values of the attributes selected were continuous numeric values. Data transformation was applied by generalizing data to a higher-level concept so as all the values became discrete. The criterion that was made to transform the numeric values of each attribute to discrete values depended on the previous day closing price of the stock. If the values of the attributes open, min, max, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the values of the attributes mentioned above were less than the value of the attribute previous, the numeric values of the attributes were replaced by Negative. If the values of those attributes were equal to the value of the attribute previous, the values were Replaced by the value Equal.

Table 2: Sample of historical data before selecting relevant attributes and before generalization

Previous	Open	Max	Min	Last	Action
45.82	45.99	46	45.41	45.67	Sell
45.67	45.68	45.68	45.2	45.3	Buy
45.3	44.8	45.3	44.41	44.9	Buy
44.9	44.8	44.9	44.3	44.87	Sell
44.87	44.87	45.55	44.85	45.3	Buy

Table 3: Sample of historical data after selecting attributes and after generalization.

Previous	Open	Max	Min	Last	Action
Positive	Positive	Positive	Negative	Negative	Sell
Negative	Positive	Positive	Negative	Negative	Buy
Negative	Negative	Equal	Negative	Negative	Buy
Negative	Negative	Equal	Negative	Negative	Sell
Negative	Equal	Positive	Negative	Positive	Buy

2.3 Building the model

After the data has been prepared and transformed, the next step was to build the classification model using the decision tree technique. The decision tree technique was selected because the construction of decision tree classifiers does not require any domain knowledge, thus it is appropriate for exploratory knowledge discovery. Also, it can handle high dimensional data. Another benefit is that the steps of decision tree induction are simple and fast. Generally, decision tree accuracy is considered good. The decision tree method depends on using the information gain metric that determines the most useful attribute. The information gain depends on the entropy measure. When the decision tree model was applied on the data of the three companies using the WEKA software version 3.5, the root attribute for both 20MICRONS and 3IINFOTECH Company was the Open, while the attribute last was the root for the decision tree of the 3MINDIA company. As the process of building the tree goes on, all the remaining attributes were used to continue with this process. After building the complete decision tree, the set of classification rules were generated by following all the paths of the tree. Both the ID3 and C4.5 algorithms were used in building the decision trees and the pruning technique was used in the C4.5 algorithm in order to reduce the size of the produced decision trees.

Table 4: Summary of the number of the classification rules

Company	Number of classification rules without pruning	Number of classification rules with pruning
20MICRONS	28	7
3IINFOTECH	35	11
3MINDIA	24	5

Figure 1. Decision Tree for the 20MICRONS

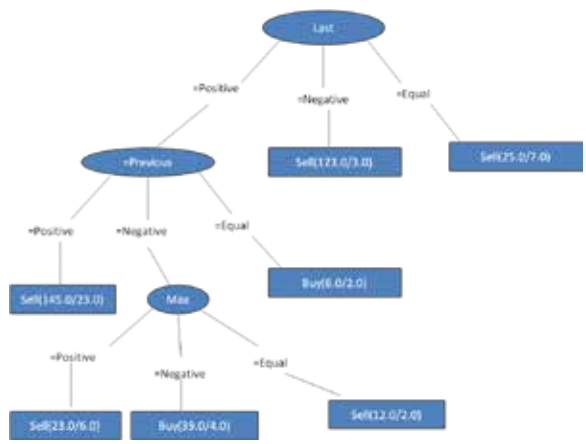
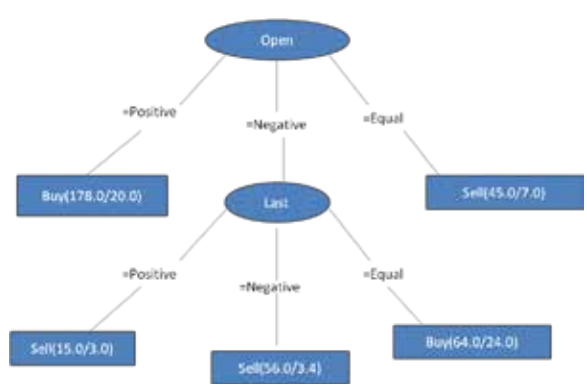


Figure 2. Decision Tree for the 3IINFOTECH



Figure 3. Decision Tree for the 3MINDIA



2.4 Deploying the model

The classification rules that were generated from the decision tree model can be used and integrated in a system that predict the best action and timing for the investors, either to buy or sell the stocks on that day.

3. RESULTS AND CONVERSATION

In order to evaluate the model, the WEKA software was used to calculate the accuracy of the classification model. Two evaluation methods were used, the K-Fold Cross Validation (K-CV) where K=10 folds and the percentage split method where66% of the data was used for training and the remainder for testing.

Table 5 Classification accuracy using ID3 & C4.5 classification methods and using 10-CV

Company	Classification Method	Total Instances	Correctly Classified	Accuracy %
20MICRONS	ID3	492	230	42.78
	C4.5		240	48.78
3IINFOTECH	ID3	504	253	52.76
	C4.5		268	53.17
3MINDIA	ID3	505	259	58.33
	C4.5		260	51.48

Table 6 Classification accuracy using ID3 & C4.5 classification methods and using Holdout Holdout 66%

Company	Classification Method	Total Instances	Correctly Classified	Accuracy %
20MICRONS	ID3	175	70	45.98
	C4.5		81	46.28
3IINFOTECH	ID3	173	86	44.34
	C4.5		93	53.75
3MINDIA	ID3	173	86	54.35
	C4.5		92	53.17

4. CONCLUSION

This paper presents a proposal to use the decision tree classifier on the historical prices of the stocks to create decision rules that give buy or sell recommendations in the stock market. Such proposed model can be a helpful tool for the investors to take the right decision regarding their stocks based on the analysis of the historical prices of stocks in order to extract any predictive information from that historical data. The results for the proposed model were not perfect because many factors including but not limited to political events, general economic conditions, and investors' expectations influence stock market.

REFERENCES

[1] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) "An effective application of decision tree to stock trading", Expert Systems with Applications, 31, pp. 270-274. | | [2] Wang, Y.F., (2003) "Mining stock price using fuzzy rough set system", Expert Systems with Applications, 24, pp. 13-23. | | [3] Murphy, J.J., (1999) Technical Analysis of the Financial Markets: a Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance. | | [4] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., and Wirth R., (2000). "CRISPDM 1.0: Step-by-step data mining guide".