Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Predicting Movement of Stock on The Basis of Daily Fluctuation Using Data Mining

[1]Dr. Rahul G. Thakkar, [2]Mr. Vimal Patel, [3]Mr. Hardik Desai
[1]Assistant Professor, ASPEE Agribusiness Management Institute, Navsari Agricultural University, Navsari, Gujarat, India
[2]Assistant Professor, College of Agriculture, Navsari Agricultural University, Navsari, Gujarat, India
[3]Assistant Professor, Naran Lala College of Professional & Applied Sciences, Navsari, Gujarat, India

Abstract - This research paper is based on decision tree (induction algorithm) which generates classification rules that will help in knowing next day trend of stocks. The research paper provides a glimpse of the market and trading tips. We have used classification rule generation method of Data Mining.
This research paper predict the next day trend of stock based on daily price movement of the stock (Open_price, High_price, Low_price, Close_price) as compare to that of with previous day price movement of the stock (Open_price, High_price, Low_price, Close_price). We have considered those classification rules which have accuracy more than ninety.

Keywords: Agriculture Market, Data Mining, Binning, Decision Tree, Classification Rules, Support, Confidence, Accuracy.

I. INTRODUCTION

Stock Market
It is an exchange place or a market that facilitates the trading of stocks. In India, the most preferable exchanges or markets are the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE).

Stock
A stock is a partial ownership in a company or an industry, with rights to share in its profits. When an investor buys a stock of a company, he is called a shareholder or a stockholder of that company.

Data Mining
Discover hidden values from the huge database. It is a powerful technology with a great potential to focus on the most important information in data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.
For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. In addition, they could make sure beer and diapers were sold at full price on Thursdays.
How does data mining work?
While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

Decision trees
Tree shaped structures that represent sets of decisions. Specific decision tree methods include Induction (ID3) Technique, Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). ID3, CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset for the prediction.
Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent *rules*. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. Decision tree programs construct a decision tree from a set of training cases.
Constructing decision trees
Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees. Decision tree programs construct a decision tree *T*

from a set of training cases. J. Ross Quinlan originally developed ID3 at the University of Sydney. He first presented ID3 in 1975 in a book, *Machine Learning*, vol. 1, no. 1. ID3 is based on the Concept Learning System (CLS) algorithm.
ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the m (where m = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.
The central focus of the decision tree growing algorithm is selecting which attribute to test at each node in the tree. For the selection of the attributes with the most inhomogeneous class distribution the algorithm uses the concept of entropy. Each discovered pattern should have measure of certainity associated with it that assesses the validity or “trustworthiness” of the pattern. A certainty measure for rules the form “A=>B” is confidence. The support of a pattern refers to the percentage of task-relevant data tuples for which the pattern is true.

II. STEPS INCLUDED IN GENERATING CLASSIFICATION RULES

Building Database

A basic requirement for the system is to get the stock historical data on the daily basis having all data regarding open, high, low, close of each and every stock which are listed in NSE. www.nseindia.com provides historical data. So we used it as a main source for the data.

Table1: Raw data directly fetched from the www.nseindia.com database.

Company_Na...	Ser...	Open_Bhav	High_Bhav	Low_Bhav	Close_Bhav	Last_Bhav	Prev_Close	Tot_trde_qty	Tot_trde_vol	Timestamp
20THCENFIN	EQ	14.450	14.500	14.100	14.100	14.100	14.000	1600.000	22697.500	12/29/1998
21STCENMGM	EQ	0.600	0.600	0.200	0.200	0.200	0.400	1200.000	280.000	12/29/1998
AARTIDRUGS	EQ	13.950	13.950	13.500	13.500	13.500	14.000	600.000	8195.000	12/29/1998
AARTIIND	EQ	27.050	27.600	27.050	27.550	27.550	26.650	800.000	21975.000	12/29/1998
ABANLLOYD	EQ	32.500	33.300	32.500	33.300	33.300	33.000	4500.000	148605.000	12/29/1998
ABB	AE	530.000	530.000	530.000	530.000	530.000	518.300	100.000	53000.000	12/29/1998
ABBOTTLAB	EQ	530.000	560.000	523.000	560.000	560.000	533.500	1650.000	890345.000	12/29/1998
ABGHEAVY	EQ	22.250	22.800	22.000	22.800	22.800	22.850	2000.000	44715.000	12/29/1998
ABSIND	EQ	35.650	36.800	35.000	36.700	36.700	36.750	2050.000	72642.500	12/29/1998

Cleaning

We have checked for the missing values. If missing values are found then past 10 trading day prices are taken for that particular field and average is taken of that prices to fill out the missing value. Care is taken that a new price is within the high and low prices of that day. While calculating percentage change for open_price, High_price, Low_price and Close_price as compare to that of previous day prices a special care is taken if the price of a stock is after a bonus or a split. Sorting is performed on the basis of company name.

Table2: Raw data after performing cleaning task.

Company_name	Trade_date	Open_price	High_price	Low_price	Close_price	Total_traded_q...	Total_traded_vol
AARTIDRUGS	12/29/1998	13.950	13.950	13.500	13.500	600.000	8195.000
AARTIIND	12/29/1998	27.050	27.600	27.050	27.550	800.000	21975.000
ACC	12/29/1998	980.000	1063.800	975.000	1063.800	306250.000	312998902.000
AEGISCHEM	12/29/1998	10.100	10.250	10.100	10.200	2100.000	21370.000
AGRODUTCH	12/29/1998	11.500	12.000	11.000	12.000	10200.000	119335.000

Calculating percentage change

To generate a decision tree, we need a percentage change for open_price, High_price, Low_price and Close_price as compare to that of previous day prices.

Deciding valuation

Based on the percentage change of Close_price a valuation of previous day record is decided, valuation are fairly_valued, under_valued or over_valued. The valuation is decided on the basis of following criteria:

If percentage change of a close price is

- >=5% then valuation of previous day record for the same company stock is “under_valued”
- Between -5% and 5% then valuation of previous day record for the same company stock is “fairly_valued”
- <=-5% then valuation of previous day record for the same company stock is “over_valued”

Table3: Data after deriving Valuation attribute.

Company_Na...	Open_Bhav	High_Bhav	Low_Bhav	Close_Bhav	Tot_Trde_Vol	Timestamp	Splited	Valuation
3IINFOTECH	-9.233	-6.843	-8.393	-7.795	-59.291	10/7/2008	F	OverValued
3IINFOTECH	-18.966	-14.858	-11.988	-10.183	-9.860	10/8/2008	F	OverValued
3IINFOTECH	-6.383	-9.804	-14.729	-5.561	3.991	10/10/2008	F	OverValued
3IINFOTECH	-18.182	14.783	-6.494	16.648	-7.043	10/13/2008	F	UnderValued
3IINFOTECH	47.778	12.500	45.972	4.078	32.973	10/14/2008	F	UnderValued
3IINFOTECH	-4.135	-10.606	-4.853	-4.851	-55.320	10/15/2008	F	OverValued
3IINFOTECH	-1.961	2.542	-7.000	3.922	109.768	10/16/2008	F	UnderValued

Binning

Binning is done on each and every field of database for each company. Binning value will replace the original value which is calculated by applying sorting on each and every attribute of each company. Total number of values in each bin is calculated on the basis of total number of records for a company divided by ten. It means that we allow maximum ten bins.

Table4: Data after performing binning on Open_Bhav, High_Bhav and Low_Bhav.

Company_Na...	TimeStamp	Open_Bhav	High_Bhav	Low_Bhav	Tot_Trde_Vol	Valuation
3IINFOTECH	10/7/2008	-6.878	-5.784	-7.435	-66.713	OverValued
3IINFOTECH	10/8/2008	-6.878	-5.784	-7.435	-14.105	OverValued
3IINFOTECH	10/10/2008	-6.878	-5.784	-7.435	-1.101	OverValued
3IINFOTECH	10/13/2008	-6.878	6.488	-7.435	-1.101	UnderValued
3IINFOTECH	10/14/2008	7.829	6.488	8.166	49.190	UnderValued
3IINFOTECH	10/15/2008	-6.878	-5.784	-7.435	-49.970	OverValued
3IINFOTECH	10/16/2008	-1.794	2.812	-7.435	109.542	UnderValued
3IINFOTECH	10/17/2008	7.829	0.872	8.166	-66.713	OverValued

Rule generation

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given database. A node is created and labeled with the attribute, branching are created for each value of the attribute and the samples are partitioned accordingly.

Description of algorithm is given below:

Create a node N;
If samples are all of the same class, C then
    Return N as a leaf node labeled with the class C;
If attribute-list is empty then
    Return N as a leaf node labeled with the most common class in samples;
Select test-attribute, the attribute among attribute-list with highest info. Gain;
Label node N with test-attribute;
For each known value ai of test-attribute
    Grow a branch from node N for the condition test-attribute= ai
    Let si be the set of samples in samples for which test-attribute= ai
    If si is empty then
        Attach a leaf labeled with the most common class in samples;
    Else
        Attach the node returned by the algorithm

Tree Pruning

We have selected pre-pruning approach where a tree is "pruned" by halting its construction early (by deciding not to further split or partition the subset of training samples at a given node). Upon halting, the node becomes a leaf. The leaf holds the most frequent class among the subset samples or the probability distribution of those samples.

Support & Confidence

The decision tree can be converted to classification IF_THEN rules by tracing the path from the root node to each leaf node in the tree. We have calculated support and confidence for each classification rule that is that is converted into IF_THEN rule in the following manner.

Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong.

Support

The rule A=>B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain AnB (i.e., both A and B). This is taken to be the probability, P (AnB).

For each rule generated by ID3 Technique we have calculated support. The rule holds in the training data set with support s, where s is the percentage of transactions in training data set that contains both IF and THEN part. This is taken to be the probability that both occur. We have considered minimum support of twenty.

Confidence

The rule A=>B has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the probability, P (B|A).

For each rule generated by ID3 Technique we have calculated confidence. The rule holds in the training data set with confidence c, where c is the percentage of transactions in training data set that contains IF part that also contains THEN part. We have considered minimum confidence of eighty.

Accuracy (Hold-Out method)

We have used hold-out method for determining accuracy in which two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, the accuracy of which is estimated with the test set.

Table 5: Classification rules generated having minimum support of twenty, minimum confidence of eighty and minimum accuracy of 90 are given below:

Classification Rule	Valuation	Accuracy
Open_bhav >= 1 AND Open_bhav < 2 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	93.174
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 15	UnderValued	92.996
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 10 AND High_bhav < 15	UnderValued	92.949
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	92.795
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	92.659
Open_bhav >= -2 AND open_bhav <- 1 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	92.557
Open_bhav >= 1 AND open_bhav < 2 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	92.105
Open_bhav >= 5 AND open_bhav < 10 AND High_bhav >= 0 AND High_bhav < 1 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	91.971
Open_bhav >= 1 AND open_bhav < 2 AND High_bhav >= 10 AND High_bhav < 15	UnderValued	91.818
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= 10 AND High_bhav < 15	UnderValued	91.603
Open_bhav >= -1 AND open_bhav < 0 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	91.331
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 0 AND High_bhav < 1 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	90.946
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	90.827
Open_bhav >= 5 AND open_bhav < 10 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -5 AND Low_Bhav <- 2	OverValued	90.426
Open_bhav >= -1 AND open_bhav < 0 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	90.249
Open_bhav >= 1 AND open_bhav < 2 AND High_bhav >= 0 AND High_bhav < 1 AND Low_Bhav >= -10 AND Low_Bhav <- 5	OverValued	90.013
Open_bhav >= 1 AND open_bhav < 2 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -5 AND Low_Bhav <- 2	OverValued	89.876
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	89.838
Open_bhav >= 5 AND open_bhav < 10 AND High_bhav >= 0 AND High_bhav < 1 AND Low_Bhav >= -5 AND Low_Bhav <- 2	OverValued	89.773
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >= -5 AND Low_Bhav <- 2	OverValued	89.733
Open_bhav >= 1 AND open_bhav < 2 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	93.174
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 15	UnderValued	92.996
Open_bhav >= 2 AND open_bhav < 5 AND High_bhav >= 10 AND High_bhav < 15	UnderValued	92.949
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= -1 AND High_bhav <	OverValued	92.795

0 AND Low_Bhav >= -10 AND Low_Bhav <- 5		
Open_bhav >= 0 AND open_bhav < 1 AND High_bhav >= 5 AND High_bhav < 10	UnderValued	92.659
Open_bhav **<=** ***-5*** **AND High_bhav <=** ***-5***	OverValued	91.897
Open_bhav >= -1 AND open_bhav < 0 AND High_bhav >= -1 AND High_bhav < 0 AND Low_Bhav >=0	FairlyValued	90.879
Open_bhav >= -1 AND open_bhav < 0 AND High_bhav <= 0 AND High_bhav < 0 AND Low_Bhav >= 0	FairlyValued	91.361

III. CONCLUSION

“Technical approach” is developed for the prediction of next day trend of stocks. For the daily traders it’s interesting if one can know the next day movement before one day. Our classification rules helps in predicting next day movement of the stock. So that before buying or selling shares we may assure our profit or loss percent.

REFERENCES

[1] Alor-Hernandez, G., Gomez-Berbis, J. M., Jimenez-Domingo, E., Rodríguez-González, A., Torres-Niño, J., “AKNOBAS: A Knowledge-based Segmentation Recommender System based on Intelligent Data Mining Techniques”. Computer Science and Information Systems, Vol. 9, No. 2, 2012, pp: 713-740.

[2] Han, J., Kamber, M., “Data mining concepts and techniques 2nd edithion”. Morgan Kaufman, 2006, pp: 227-378.

[3] Agrawal R., Imielinski T. “Mining associations between sets of items in large databases”. Proceedings of the ACM SIGMOD International Conference on Management of Data. pp: 207-216.

[4] Konda, S., “Web Data Mining Based Business Intelligence and Its Applications”. IJCST, Vol. 4, No. 4, 2013, pp: 112-116.

[5] Aher, B., “Association Rule Mining in Data Mining”. IJCST, Vol. 4, No. 3, 2013.

[6] Nagabhushanam, D., Naresh., N., “Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient’s Data”. IJCST, Vol. 4, No. 4, 2013, pp: 262-265.

[7] Surya, K., Priya, K., “Exploring Frequent Patterns of Human Interaction Using Tree Based Mining”. IJCST, Vol. 4, No. 4, 2013.

[8] Rajnikanth, J., “Database Primitives, Algorithms and Implementation Procedure: A Study on Spatial Data Mining”. IJCST, Vol. 4, No. 2, 2013.

[9] Quinlan, J. R., “C4.5: Program for machine learning. CA”. Morgan Kaufmann, San Francisco, 1992

[10] Clark, P., Niblitt, T., “The CN2 induction algorithm. Machine Learning”. Vol. 3, No. 4, 1989, pp: 261-283.

[11] Cohen, W.(1995). “Fast effective rule induction”. Proceedings Twelth International Conference on Machine Learning. Pp: 115-123

[12] Duda, R., Hart, P., “Pattern classification and scene analysis”. Wiley, New York, 1973

[13] Ardakani, H. D., Hajizadeh, E., Shahrabi, J., “Application of Data Mining Techniques in stock markets: A survey”. Journal of Economics and International Finance, Vol. 2, No. 7, 2010, pp: 109-118.

[14] Lopez, V. F., Moreno, M. N., Polo, M. J., Segrera, S., “Improving the Quality of Association Rules by Preprocessing Numerical Data”. II Congreso Espanola Informatics, 2007, pp: 223-230.

[15] Chung, F. L., Fu, T. C., Ting, J. (2006), “Mining of Stock Data: Intra-Stock and Inter-Stock Pattern Associative Classification”. Proceedings International Conference on Data Mining, pp: 30-36.

[16] Frank, E., Witten, I. H., (1998). “Generating accurate rule sets without global optimization”. Proceedings Fifteenth International Conference on Machine Learning, pp: 141-151.

[17] http://www.nseindia.com

[18] http://www.bazaartrend.com

[19] http://www.traderji.com

[20] http://marketlive.in

[21] http://equitymaster.com

