An Optimized Approach to Analyze Stock market using Data Mining Technique

Dattatray P. Gandhmal
Computer Science Department,
Walchand Institute of Technology, Solapur

Ranjeetsingh B. Parihar
Computer Science Department,
Walchand Institute of Technology, Solapur

Rajesh V. Argiddi
Assistant Proff. Computer Science Department,
Walchand Institute of Technology, Solapur

ABSTRACT

This paper basically deals with identifying frequent patterns from large amount of stock data. These frequent patterns are identified based on rise and fall of stock prices. We have two stages, in first stage we categorize the stock data based on zero growth, slow growth and fast growth using k-means algorithm. In second stage we use CIR algorithm to generate useful trends about the behavior of stock markets. The trend holds to interpret the present and predict the next stock price. Some item-set from sales data indicate market needs and can be used in forecasting which has great potential for decision making, market competition and strategic planning. The objective in this research is to identify or to predict the stock market from the viewpoint of investors. So the investors can invest their shares in the appropriate companies based on zero growth, slow growth and fast growth. These two stage mining process that is k-means and CIR algorithm can generate more useful item-set according to our analysis.

General Terms

Patten Recognition, Rule generation, Stock Price.

Keywords

Zero Growth (ZG), Slow Growth (SG), Fast Growth (FG), clustering.

1. INTRODUCTION

Data is very important for any organization or business process. Data which used to be measured in gigabytes or terabytes nowadays it has raised up to peta bytes, i.e. there has been a great amount of increase in the size of the database.

As the complexity involved in such kind of data is very large, so it is not at all practically possible to manually analyze or predict the data. Some clusters may be growing while others may be declining. Information generated is very useful for business decision making. Decision can take place on the basis of classification of Zero-Growth (ZG), Slow-Growth (SG) and Fast-Growth (FG) of the sale. However it is very useful in understanding stock market trends [1]. It is easy to turn cash into inventory, but the challenge is to turn inventory into cash. Only with help of Data mining it is possible to find out useful patterns and associations from the stock data.

Data mining consists of useful techniques such as Clustering and Association rules, these techniques can be used to predict the future trends based on the Item-sets [5]. Clustering is used to group similar item-sets while association is used to get generalized rules of dependent variables. Useful item-sets can be obtained from huge trading data using these rules.

Stock Prices are considered to be very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and in part because of the mix of known parameters (Previous Day's Closing Price, P/E Ratio etc.) and unknown factors (like Election Results, Rumors etc.)

Internet provides almost all possible information on the stock market worldwide through various useful websites as Google finance, CNN Money, Bloomberg, Financial Times, Thisismoney, Yahoo finance and many more. The reliability of the information depends on the reputation and the quality of the source sites.

In this research we have taken the original data sets of Bombay Stock Exchange (BSE) of different companies such as Infosys, TCS, TechM from yahoo finance and applied our two stage algorithm on this data set. When applying two stage algorithms to stock data, we are more interested in doing a Technical Analysis to see if our algorithm can accurately work for detecting patterns in the stock time series. This said two stage algorithm can also play a major role in evaluating and forecasting the performance of the company and other similar parameters helpful in fundamental Analysis.

Decision making for investors in stock market is considered to be one of the difficult tasks. There is a need for the study in data mining for shares selection with companies that has maximum growth rate. But our problem is to find out the companies which are reliable to invest in the shares with maximum profit. This is a useful approach to identify the company's growth rate based on some of the attributes, e.g. we can examine that the "Infosys having highest growth rate with maximum volume quantity because of frequent change in high and low values in stock market", and here we have basic property related to this example, i.e. company name, high value, low value, open, close, volume. Similarly we analyze different companies have different volumes based on their high and low values and predict the growth pattern. Thus on the basis of this scenario we can predict the reason of zero-growth, slow-growth and fast-growth items. Data mining techniques are best suited for the analysis of such type of classification, useful patterns extraction and predictions.

2. MOTIVATION AND RELATED WORK

Researchers in the field of data mining always try to find innovative techniques so as to improve the performance of the extraction methods used in data mining as they usually use history of the different transactions done in finding the data as it will be useful for future use. This data collection can be used by them to predict the customer behavior and their interests L.K.Soon et al [9], compared the execution performance of numerical and symbolic representation of using data in term of similar search. M. C. Lo [10] he considered a model for inventory decision support system[IDSS] in which ordering quantity, ordering cost, safety factor, lead time and backorder discounts are decision variables, the algorithm is applied to fined the optimal solution for the case where the lead time demands to follow a general distribution. J. ting et al he proposed a technique based trading data mining approach for intra-stock mining which usually perform concentrates on finding most appearing items for the stock time series data and inter trading mining which used to discover the different strong relationship among the several stocks. L. K. Soon et al [9] generated a list of stocks which are influential to Kuala Lumpur Composite Index (KLCI), and then produce classification rules, which he denotes the inter-relationships among the stocks in terms of their trading performance with respect to KLCI. Further Aurangzeb Khan; Kairullah Khan used the Most Frequent Item set rule to generate patterns on super market and trading data [4]. In the current years of development in the field of data mining, it is considered that the partitioned clustering technique is well suited for clustering a large document dataset due to their relatively low computational requirements and increase in the gradual performance of the system. The time factor complexity of the partitioning technique is almost linear, because of which it is widely used. The best known partitioning clustering algorithm is the K-means algorithm and its variants [8]. As this algorithm is simple, straightforward and is based on the firm foundation of analysis of variances. In addition to the K-means algorithm, several algorithms, other algorithms such as Particle Swarm Optimization (PSO) [11] is another computational intelligence method that has already been applied to image clustering and other low dimensional datasets. For this work we have used clustering algorithm for clusters and CIR for item-set association among the cluster. The classification of similar objects into different groups, or the partition of data into subsets called clusters. The data in each subset share some common trait-often proximity accordingly to some defined distance measure.

3. METHODOLOGY

In this research we have proposed an algorithm that mines huge amount of stock data that identifies or helps in predicting growth rate of company's shares for the investors. In first stage we divide stock data in three different clusters which are categorized using k-means algorithm on the basis of volume quantity and fluctuation in high and low price of the shares that is Zero Growth (ZG), Slow Growth (SG) and Fast Growth (FG). In second stage we have proposed Consistent Item set Rule (CIR) algorithm to find out the shares that has maximum profit for the investors. CIR is similar to Apriori algorithm [7] that provides the strong association among the Item-set. Some results obtained from this experimental analysis from stock data, we observed that the combination of k-mean and association of CIR can generate more useful item-set from large market data.

3.1 Proposed Architecture

Our proposed approach is a two stage model. First we generate clusters using K-Mean algorithm, and then CIR is implemented for finding shares with maximum profit. First we will explain stage 1.The block diagram of the whole process is given in the figure 1.

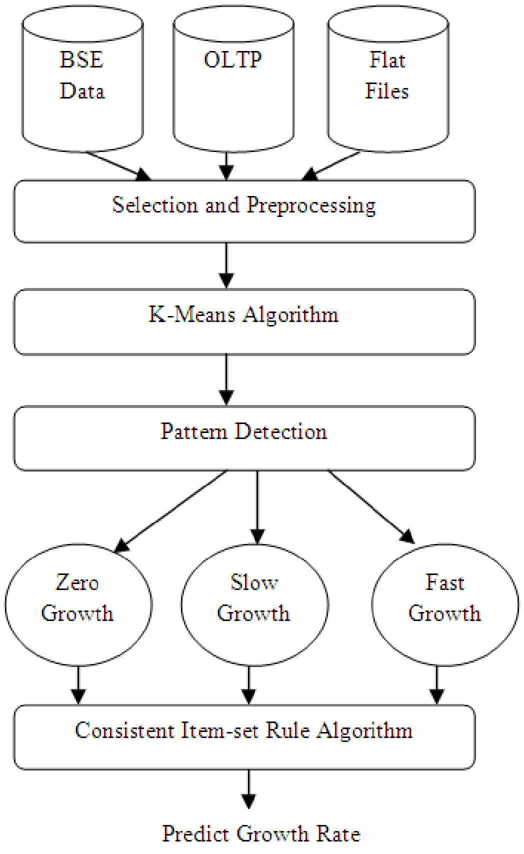

Fig 1: Block Diagram

3.2 K-Means

One of the algorithms that are used in data mining to classify the data is K-means algorithm. Nearness is usually measured by some sort of distance; the most commonly being used is the Euclidean distance [6].

$$dist(i,j) = \sqrt{\sum_{k=1}^{i} \left(x_{ik} - x_{jk}\right)^2} \qquad (1)$$

The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.

So, the better choice is to place them as much as possible far away from each other. This algorithm aims at minimizing an objective function, in this case a squared error function.

The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{k} \left\| x^{(j)} - c_j \right\|^2 \qquad (2)$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$

The steps of the algorithm:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.3 Consistent Item-Set Rule

Association rule mining is one of the most important and well defines technique for extract correlations, sequential item-set, associations or causal structures among sets of items in the transaction databases or other repositories. Association rules are widely used in various areas such as risk management, telecomm, market analysis, inventory control, and trading data [1]. Apriori algorithm [7] for strong association among the item-set is highly recommended. In this work we proposed a new algorithm CIR that is more efficiently generates sequential item-set and strong association between them. For this purpose a property matrix containing counted values of corresponding properties of each product has been used as shown in Figure 2.

```
Input: Data Set (DS): //Stock data from BSE
Output: Maximum CIR Matrix //Consistent Item-set Rule with maximum volume
Start
CIR = Ø
For each Company $C_i$ in DS
    a.  For each company calculate volume ($V_i$)
        $V_i$ =Calculate ($C_i$)
        Find company having maximum volume
    b.  Generate CIR
        CIR = Maximum ($V_i$)
        Next [Repeat]
End
```

Fig 1: CIR Matrix Algorithm

4. DATA SELECTION AND PREPROCESSING

Data in its original format is never or cannot be used directly with data mining. Initial data needs to be SELECTED AND preprocessed i.e. transformed, aggregated and integrated such that mining process can effectively perform on it. There is a need to process the data before it used in the knowledge discovery (KDD) process.

company_name, open, high, low, close, volume, Date

AccenTech, 29.7, 29.7, 29.7, 29.7, 0, 3/12/2001

VMFSoft, 6.4, 6.4, 5, 5. 1, 1400, 1/12/2010

Relictech, 7.85, 10.1, 7.33, 8.99, 2200, 1/12/2010

LOGIX, 32.5, 34.25, 26.75, 28.85, 7300, 1/12/2010

Patintlog, 36, 38.6, 29.5, 32.35, 8600, 1/12/2010

Infosys, 3049.8, 3454, 3032, 3435.1, 824200, 1/12/2010

TECHM, 655,712,623.15, 702.7,157200, 1/12/2010

ELNET, 53.3, 56.85, 49.25, 55.9, 2900, 1/12/2010

BALTE, 3, 3, 3, 3, 3, 0, 3/12/2010

One of the key issue is data quality, 50% to 80% data mining experts spend their time in analyzing quality of data. In this case the collected data in Fig 3 was cleaned by using SQL Server Data Transformation Services, and then removed noise from the transformed data.

The processed data are shown in the Table 1.

company_ name	open	high	low	close	Volume	Date
AccenTech	29.7	29.7	29.7	29.7	0	3/12/2001
VMFSoft	6.4	6.4	5	5. 1	1400	1/12/2010
Relictech	7.85	10.1	7.33	8.99	2200	1/12/2010
LOGIX	32.5	34.25	26.75	28.85	7300	1/12/2010
Patintlog	36	38.6	29.5	32.35	8600	1/12/2010

5. TECHNICAL ANALYSIS

In technical analysis we will first apply K-means algorithm on the selected and preprocessed data to form three different clusters such as Zero Growth (ZG), Slow Growth (SG), and Fast Growth (FG) and further we will use the Consistent Item-set Rule algorithm (CIR) to generate the shares which has maximum volume. We have applied this algorithm on the real data set of Bombay stock exchange which is downloaded from Yahoo Finance. This downloaded data is then sorted on the basis of different companies and provided this data as input to our two stage algorithm. Next part we will elaborate separately ZG, SG and FG to understand the concept clearly.

So the process has two stages as shown below:

5.1 Stage One

As explained before in stage one, we will apply K-Means algorithm to form three clusters Zero Growth (ZG), Slow Growth (SG), Fast Growth (FG) based on the volume of the company share using high and low price. The table shown consists of attributes company_name, open, high, low, close, volume, date. The attribute volume is very important because on the basis of the shares of the volume we have divided them into three clusters.

Cluster 1 (Zero Growth): This cluster consists of those companies' share that has zero volume rates i.e. those company's shares that are not purchased by anyone.

company_name	open	high	low	close	Volume	Date
AccenTech	29.7	29.7	29.7	29.7	0	1/1/2001 to 3/12/2010
AEGISLOG	9.25	9.25	9.25	9.25	0	1/1/2001 to 1/11/2001
Icesoft	0.45	0.45	0.45	0.45	0	4/1/2010 to 1/12/2010
PFLINFOTC	22.9	22.9	22.9	22.9	0	3/1/2005 to 1/6/2005
SPLTechno	143	143	143	143	0	31/3/2003 to 1/12/2010
TCLTech	1.3	1.3	1.3	1.3	0	1/1/2007 to 3/3/2008
Softrakt	0.3	0.3	0.3	0.3	0	3/2/2003 to 1/12/2010

Cluster 2 (Slow Growth): This cluster consist that company that has average volume rate.

company_name	open	high	low	close	Volume	Date
JetKingQ	137.95	143.25	129.5	134.25	1700	12/1/2010
TechFor	22.2	23.85	15.85	22.45	1800	12/1/2010
RelicTech	16.5	17.3	12.75	14.1	2000	12/1/2010
CSSTech	21	22.4	18.45	22	2800	12/1/2010
EINet	53.3	56.85	49.25	55.9	2900	12/1/2010

Cluster 3 (Fast Growth): This cluster consists those company that has fast volume rate i.e. those company's shares that are purchased by many people.

company_name	open	high	low	close	Volume	Date
Patni	468	505	426.9	476.65	128200	12/1/2010
LNT	1950.05	2064	1918.7	1979.05	139700	12/1/2010
TCS	1067.55	1179	1050	1165.05	152900	12/1/2010
TechM	655	712	623.15	702.7	157200	12/1/2010
Infosys	3049.8	3454	3032	3435	842400	12/1/2010

5.2 Stage Two

In this stage our proposed algorithm is used to generate a property matrix of the company with maximum profit based on volume of shares as shown in the figure.

Here we form the matrix on the basis of Trading Price, Growth Rate and volume. Trading price can be categorized into high, low price of share. Growth Rate is clustered into three types that are zero growth, slow growth and fast growth. Volume is divided into 3 fields based on time period that is 3 months, 6 months and 1 year. So based on these attributes we can easily identify the growth rate of companies.

In the following table we have considered some of the companies share data such as AccenTech, Infosys, RelicTech, TechFor, SPLTechno, TechM, EINet.From that matrix returns the companies which has the highest growth rate.

Company Name		**AccenTech**	**Infosys**	**RelicTech**	**TechFor**	**SPLTechno**	**TechM**	**EINet**
Trading Price	**High**	29.7	3454	17.3	23.85	143	712	56.85
	Low	29.7	3032	12.75	15.85	143	623.15	49.25
Growth Rate	**Zero**	Yes	No	No	No	Yes	No	No
	Slow	No	No	Yes	Yes	No	No	Yes
	Fast	No	Yes	No	No	No	Yes	No
Volume	**3 Months**	0	3345200	23600	1900	0	967000	44700
	6 Months	0	6594900	33000	11700	0	2340100	153500
	12 Months	0	12334600	49300	28100	0	4237600	277400

6. CONCLUSION AND FUTURE WORK

This paper presents solution to pattern discovery which provide maximum profit to investors. This two stage clustering association rule allow to analyze the huge stock data and classify that data with growth rate and generate the item set with maximum profit. The experiment result is applicable for the investors to invest their money in share market in a proper manner. Some of the limitations of this algorithm are that it requires proper data with specific attribute from stock data. Data is in unstructured format so we will try to implement it properly which provide more efficiency and accuracy.

7. REFERENCES

[1] Eugene F. Fama "The Behavior of Stock Market Prices", The Journal of Business, Jan 1965.

[2] Hua Yuan1 Junjie Wu "Mining Maximal Frequent Patterns with Similarity Matrices of Data Records", Beihang University, Beijing, China.

[3] Marcello Braglia, Andrea Grassi, Robert Montanari "Multi attribute Classification method for spare parts inventory management".

[4] Aurangzeb Khan, Khairullah khan "Frequent Patterns Minning of Stock Data Using Hybrid Clustering Association Algorithm", University Technology PETRONAS.

[5] Gebouw D, B-3590 Diepenbeek, Belgium "Building an Association Rules Framework to Improve Product Assortment Decisions" 2004

[6] J.C Gower "Euclidean Distance"

[7] Jiawan Han, Micheline Kamber "Data Mining Concepts and Techniques" 2nd edition 2004

[8] Artigan, J. A. Clustering Algorithms. Ohn Wiley and Sons, Inc., New York, NY. 1975.

[9] L.K. Soon and Sang Ho Lee, "Explorative Data Mining on Stock Data"

[10] M.Cheng Lo, "Decission support system for the integrated inventory model" with general distribution demand. Information technology journal 6(7) PP.1069-1074, 2007.

[11] Kennedy J., Eberhart R. C. and Shi Y., 2001. Swarm intelligence, Morgan Kaufmann, New York.