Future Predictions in Indian Stock Market through Linguistic-Temporal Approach

Priti Saxena[1], Bhaskar Pant[2], R.H. Goudar[3], Smriti Srivastav[4], Varsha Garg[5] and Shreela Pareek[6]

Department of CSE, Graphic Era University, Dehradun, India

E-mail: [1]asmisaxena @gmail.com; [2]pantbhaskar2@gmail.com;[3] rhgoudar@gmail.com; [4]srivastava.smriti29@gmail.com; [5]varsha.garg10@gmail.com; [6]shreelapareek29@gmail.com

Abstract—*Stock market is an ever changing chaotic business area where prediction plays a major role. Prediction provides knowledgeable information regarding the current status of the stock price movement. Thus this can be utilised in decision making for customers in finalizing whether to buy or sell the particular shares of a given stock.*

Temporal data mining has been emerged as an interesting field in providing technologies for stock forecasting. As the data of the stock market is temporal in nature effective techniques can be applied to enhance prediction analysis.

Many researchers have focused on this prediction research area, but still the results are not very accurate.

The paper proposes a new approach of analysing the stock market and predicting in a hybrid form of linguistic-apriori concept. It provides accurate results in stock prediction which has a great impact in decision making with respect to the clients and knowledge discovery of various useful patterns for brokers. This approach provides the clients with the ease of getting information status of any stock price movement immediately.

Keywords: Linguistic, Apriori, Stock Market, Temporal Data, Frequent Pattern, Filtered Data.

I. INTRODUCTION

According toWilliam J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus 1991 [4]: "Knowledge discovery in databases, also known Data mining, is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Association rules mining is an essential aspect in the study of data mining. This problem is commenced [1] in 1993. Since then, it has been a centre of attraction for business, scientific research and management of enterprise, etc. The foremost goal of temporal data mining is to discover hidden relations between sequences which are temporal in nature. The detection of associations between sequences consist of major steps: the modeling and representation of data in appropriate form; the description of measures which are similar between sequences; implementing the models and representations to the real mining scenarios. Many authors have used a dissimilar approach to classify data mining problems and algorithms.

Predictive analytics uses historical data to formulate future predictions. These predictions rarely acquire the structure of absolute outcomes, and are described to show the behaviour that corresponds to the behaviour taking place in the future.

Prediction is one of the major scenario in temporal data mining in stock market analysis. However, this may be formulated as classification, clustering, association rule finding problems. However, we are still failing to discover significant applications that involve prediction of time series and that do not come under any of the previously described categories.

Stock market data plays a very crucial role in today's dynamic, challenging, versatile, expanding, demanding scenario for predictions.

The rapid process of digitization of data has emerged as the huge amount of data reservoir in the databases and data warehouses. It becomes difficult to find out the exact information from normal scan.

In real life applications, the numbers of frequent sets are large in number and as the result; the numbers of association rules are also very large. We select only the rules in which we are interested for stock price prediction in this context.

There are many data mining algorithms such A priori Algorithm, Partition algorithm, Pincer-Search Algorithm, Dynamic Item set Counting Algorithm, FP-Tree Growth etc are used for finding the discovery of frequent sets are related with association rules. The paper has applied Apriori algorithm to the dataset to find the frequent sets and with the help of the algorithm we are predicting the stock price variation for multiple companies for multiple number of days.

A. Related Work

The past researches, different data warehouse systems have presented different techniques to support data mining; Ahmed et al. [20] expressed the data warehouse backboned

978-1-4673-4603-0/12/$31.00 ©2012 IEEE

system integrated based data mining and OLAP techniques. This system makes use of a router to adopt the previous mining result stored in the data warehouse, accordingly avoiding processing large amounts of the raw data. [21]

The problem of discovering sequential patterns in temporal data has been emerged as a major research area problem ([1]; [7]; [10]; [8]; [2][9])and its importance is justified by the potential application domains where mining temporal sequential patterns is crucial issue, such as financial market (evolution of stock market shares quotations), retailing (evolution of clients purchases) etc.

The Apriori algorithm was first proposed by Agrawal et al. is a classical algorithm.It shows a prior knowledge of the item sets. The prior knowledge defines that any non-empty subset of a frequent itemset is also a frequent itemset. Apriori algorithm follows a level-wise and iterative approach, first generates the candidates and pass them to remove the non-frequent itemsets.

Following various considerations several approaches with the aim to find interesting changes in statistical measures of association rules have recently been proposed and the term *rule change mining* coined [14, 15, 16, 17]. The pruning methods are utilized to constrain the set of generated rules. Examples are *non-redundant rule sets* [18] and *informative rule sets* [19].

Many of previous studies approved an Apriori-like candidates generation-and-test approach.

In [5], the paper is analysing the long-run relationship among seven prominent stock indices using Wavelet theory concept. In paper [11] a statistical approach is applied to distinguish trend, semi-stable and stable rules with respect to their histories of confidence and support. The paper [12] proposes an approach which is based on statistical tests in order to find derivative rule change histories and marks the respective rules as redundant. The approaches studied so far are lacking in terms of low complexity as the data is in enormous amount which needs a complex procedure and takes long duration of time. In this paper the linguistic approach is used which reduces the amount of temporal data and aprori is applied on small data set that does not require much time.

II. Proposed Architechture

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A model has been applied in this paper. The model generates the linguistic forms from the crisp data values of the stock market by analysis of various companies. Then MFP's (Most frequent pattern) are discovered from transforming the stock prices into linguistic form (High, Low, Medium) under the specified attributes.

1. First step is to collect data from real stock market. The collected data is of 15min duration time.
2. The numeric deviation of open and close price of the stock data is transformed into the linguistic approach.

Apriori is applied on linguistic data to find out most frequent patterns. The overall architecture is shown in Fig1. The data is taken from the NSE and BSE stock market of India. One of the leading stock market business analyst, Motilal Oswal firm has been approached to get the day to day transaction on the data.

A. Problem Scenario

The frequent problem that a broker faces nowadays is the summarization of the data of the stocks and providing the on the spot information about stock status.

The task of association rule mining is to discover the interesting hidden pattern from the given database by discovering new inter and intra-sectional relation between item sets. Apriori algorithm concept is applied to find out the frequent item sets in a database. A frequent item sets is a set of items appears at least in a pre-specified number of transactions. These are further use to generate association rules. The paper proposes the application of apriori algorithm on datasets available from the stock market.

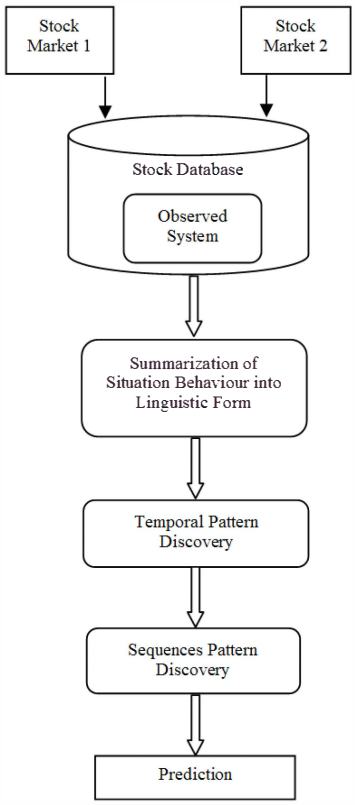

FIG. 1: THE BASIC ARCHITECTURE OF THE SYSTEM

B. Most Frequent Pattern

Apriori algorithm follows a "bottom up" approach, where frequent subsets are extended one item at a time a step known as candidate generation step, and groups of candidates are tested against the data.

The frequent set can be determined by the following rule. Let T be the transaction database and σ be the user specified minimum support, then the item set X€A is said to be frequent ser in T with respect to σ if S(X) r>= σ. The process of frequent item sets generation is as follows:

Let I_0 denotes a list of stock open prices and I_c denotes the list of close prices for a given company.

A set of $X_0=\{I_1,I_2.....I_K\} \leq$ is called an item set or a k-item set, if it contains a number of price values of stocks.

A transaction over I_0 is the coupling of T_0= ($t_0$id, I) where $t_0$id is the transaction identifiers & I_0 is an item set.

A transaction T_0 is said to be the support of an X_0 c I_0; if X_0 c I.

A transaction database DB over I (I_0 c I, I_c c I) is a set transactions performed over I.

The support of an item set X_0 in DB is the number of transactions in DB which supports X_0 or X_c:

Support(X_0, DB) = | {tid |(tid,I_0) € DB,X_0 c I_0}|

The frequency of an item set X_0 in DB is the equals to the probability of X occurring in transaction T_0 € DB.

Frequency(X_0, DB) =P(X_0) = Support(X_0, DB)

Where

|DB|=Support ({}, DB).

An item set is called frequent if its support count is greater than or equal to a pre-specified support threshold value.

Support threshold – SDB with 0≤SDB≤|DB|.

The implementation for phase-1 is as follows:

Input: Data of stock market and specified attributes

Output: Linguistic values of stock attribute data variables or terms used:

TABLE 1: LIST OF VARIABLES USED

S. No.	*Name*	*Function*
1	Counter	Used in conversion from 15-min duration to 1 hr
2	i	variables that runs through 15min values
3	j	runs for values for number of days
4	tot_openprice	Total of all values of open price
5	tot_closeprice	Total of all values of close price
6	Openprice (i)	List of open price values
7	Closeprice (i)	List of close price values
8	avg_openprice (i)	Average of list of values for given no of days
9	avg_closeprice (i)	Average of list of values for given no of days
10	numeric_dev (i)	Change in the prices from close to open
11	minimum_dev	Minimum value in the deviation list
12	maximum_dev	Maximum value in the deviation list
13	med_dev	Medium value of the deviation price list
14	dev_ling(i)	Linguistic list of deviated values

III. IMPLEMENTATION PHASE

First Phase shows the transformation of stock prices from numeric to linguistic form and the second phase applies Apriori to this form. The working is shown in tableI.

Second Phase is as follows: Now we have the final linguistic tables for different companies merged into the single.

Input: Datasets (DS) in linguistic form from different companies.

Output: Frequent pattern.

```
/*running the loop for n number of days*/
1.  for j ←1 to mdays
2.   counter ←0
3.  for counter ←1 to n
4.   i ←1
5.   counter←counter+1
6.   for i←1 to 4 /* 15- min to 1 hour duration */
7.   tot_openprice=openprice(i)+tot_openprice
8.   tot_closeprice=closeprice(i)+tot_closeprice
9.   next (end of inner loop)
10.  counter←counter+4 /* for every next hour */
11.  avg_openprice(counter)←tot_openprice/4
12.  avg_closeprice(counter)←tot_closeprice/4
13.  next (end of inner loop)
14.  for i←1 to counter
15.  numeric_dev(i)←avg_closeprice-avg_openprice
16.  next(end of outer loop)
17.  minimum_dev←numeric_dev(1)
18.  maximum_dev←numeric_dev(1)
19. for i←2 to counter
20.  if (numeric_dev(i)<minimum_dev)
21.  then
22.  minimum_dev← numeric_dev(i)
23.  end if
24.  if (numeric_dev(i)>maximum_dev)
25.  then
26.  maximum_openprice← numeric_dev(i)
27.  end if
28. next[end of loop]
29.  med_dev←(maximum_dev+minimum_deve)/2;
30. for i←1 to counter /* using linguistic list of
deviated prices */
31.  if (numeric_dev(i)<=minimum_dev)
32.  linguistic_dev(i)= “low”
33.  else if(numeric_dev(i)>minimum_dev &&
numeric_dev(i)<=med_dev)
34. linguistic_dev(i)= “medium”
35.  else linguistic_dev(i)= “medium”
36. endif
37.  next(end of inner loop)
38.  next(end of main loop)
```

Join Step: Cik is generated by joining Lik-1with itself
•Pruning Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset is removed

APRIORI ALGO
Step1: i = 1
Find frequent set L_i from C_i of all candidate itemsets
Form C_{i+1} from L_i; i = i + 1
Repeat 2-3 until C_k is empty

Step 2: scan *list* and count each itemset in C_i ,
if it's greater than minimumSupport, it
is considered as frequent

Candidate generation step
Step 3:For j=1, C_1 = all 1-itemsets.
For j>1, generate C_j from L_{j-1} as follows:
The join step
C_j = j-2 way join of L_{j-1} with itself
If both $\{a_1, \ldots, a_{j-2}, a_{j-1}\}$ & $\{a_1, \ldots, a_{j-2}, a_j\}$ are in L_{j-1}, then add $\{a_1, \ldots, a_{j-2}, a_{j-1}, a_j\}$ to C_j
(keep items sorted).
The prune step
Remove all$\{a_1, \ldots, a_{j-2}, a_{j-1}, a_j\}$ if it contains a non-frequent (j-1) subset

IV. Experiment and Results

Demonstration Is As Follows:

The raw data is taken from Table 2.

Table 2: Stock Market Data

STOCK MARKET DATA		
COAL India Ltd.		
Time	*Open Price*	*Close Price*
9:00-9:15	354.2	362.1
9:15-9:30	363.25	364.8
...		
...		
3:15-3:30	363.35	363.45
3:30-3:45	363.45	363.45

STEP 1: The data is pre-processed into the linguistic form for multiple companies for a given number of days. The output is as follows in Table 3.

Table 3: Stock Market Linguistic Form

Final linguistic Table of Stock Data			
Transaction	*Coal India Ltd*	*ICICI*	*HDFC*
T1	H	L	M
T2	H	M	M
T3	M	L	M
T4	L	L	H
T5	H	M	L
T6	H	M	L
T7	M	M	L

Step 2: The Apriori is applied to generate the first candidate set of items. The result is as follows:

Item Value	*Support Count*
L	5
M	7
H	9

Step 3: The join operation together with pruning results into step 2 and 3.

Item Set	*Support Count*
HL	4
HM	4
LM	5

Results

The result the final status in the form of low price variation, medium price variation and high price variation in the stock price movement.

The result is as follows:

H L M—Support Count IS 3

3 means a threshold value greater than the minimm threshold value.

As per the companies evaluated:
1. There is a high price movement in the coal India Ltd.
2. There is low price movement for ICICI.
3. There is medium price movement in the HDFC.

V. Conclusion

As per the final result, it has become evident that the approach is a valuable approach from the broker's view it is very easy to keep track of the current situation (status) of the stocks in the market irrespective of the time period.

From the client's point of view it is very to decide whether to hold , buy or sell any particular stock based on its price movement information provided by the broker in the market. On subsequent evaluation we find that hybrid based approach is proved as a promising one for extracting some association rules of predictive nature from Indian stock market which could be used for prediction or recommendations in stock trading platforms. We have presented the result implementation of apriori mining algorithm. The approach is a valuable approach as it provides immediate information about the status of the stock, or multiple stocks together, for irrespective number of days.

V. FUTURE ENHANCEMENT

The *apriori*-based approach in the candidate generation concept, is one of the most costliest operations of association rule mining. In order to avoid costly database scan, we can perform the same with FP growth. The same algorithm can also be applied on different datasets. In future FP growth technique can be applied on other Stock Market fields such as Banking sectors, Insurance sector efficiently.

REFERENCES

[1] Agrawal, R., Srikant, R.: Fast Algoritms for Mining Association Rules. VLDB (1994) 487-499.

[2] Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. ACM SIGMOD Int. Conf. on Management of Data (2000) 1-12.

[3] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., Hsu, M.: FreeSpan: Frequent pattern-projected sequential pattern mining. ACM SIGKDD (2000) 355-359.

[4] [FSSU96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth,and Ramasamy Uthurusamy. Advances in Knowledge Discovery and DataMining. AAAI Press / The MIT Press 1996..

[5] r.sahu p.b.sanjeev:co-integration of stock markets using wavelet theory and data mining.

[6] Agrawal, R. and Srikant, R. Mining Sequential Patterns. In Proceedings of the International Conference on Data Engineering. Taipei, Taiwan, pp. 3–14, 1995.

[7] Agrawal, R. and Srikant, R. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the Fifth Int. Conference on Extending Database Technology. Avignon, France, pp. 3–17, 1996.

[8] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., and Hsu, M.-C. Freespan: frequent pattern-projected sequential pattern mining. In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, pp. 355–359, 2000.

[9] Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q.,and Dayal, U. Multi-dimensional sequential pattern mining. In Proceedings of the Tenth International Conference on Information and Knowledge Management. Atlanta, USA, pp. 81–88, 2001.

[10] Zaki, M. J. Spade: an efficient algorithm for mining frequent sequences. Machine Learning Journal vol. 42, pp. 31–60, 2001.

[11] B. Liu, Y. Ma, and R. Lee. Analyzing the interestingness of association rules from the temporal dimension. In *Proc.IEEE ICDM 2001*, pages 377–384, San Jose, CA, 2001.

[12] Mirko B¨ottcher Martin Spott Detlef Nauck Intelligent Systems Research Centre, BT Research and Venturing Adastral Park, Orion Bldg. pp1/12, Ipswich, IP5 3RE, UK: Detecting Temporally Redundant Association Rules.

[13] B. Liu, W. Hsu, and Y. Ma. Discovering the set of fundamental rule changes. In *Proc. ACM SIGKDD 2001*, pages 335–340, San Francisco, CA, 2001.

[14] R. Agrawal and G. Psaila. Active data mining. In *Proc. KDD 1995*, pages 3–8, Montreal, Canada, 1995.

[15] B. Liu, Y. Ma, and R. Lee. Analyzing the interestingness of association rules from the temporal dimension. In *Proc.IEEE ICDM 2001*, pages 377–384, San Jose, CA, 2001.

[16] M. Spiliopoulou, S. Baron, and O. G¨unther. Efficient monitoring of patternss in data mining environments. In *Proc. ADBIS 2003*, pp. 253–265, Dresden, Germany, 2003. Springer.

[17] W.-H. Au and K. Chan. Mining changes in association rules: a fuzzy approach. *Fuzzy Sets and Systems*, 149(1):87–104, 2005.

[18] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9(3):223–248, 2004.

[19] J. Li, H. Shen, and R. Topor. Mining informative rule set for prediction. *Journal of Intelligent Information Systems*, 22(2):155–174, 2004.

[20] K. M. Ahmed, N. M. El-Makky, and Y. Taha (1998). Effective data mining: a data warehouse-backboned architecture. The 1998 conference of the Centre for Advanced Studies on Collaborative research, Toronto.

[21] R. S. Monteiro, G. Zimbrão, H. Schwarz, B. Mitschang, and J. M. Souza (2005). “Building the Data Warehouse of Frequent Itemsets in the DWFIST Approach.” Foundations of Intelligent Systems 3488: 294-303.