ISSN:0975-9646

Analysis of Fragment Mining on Indian Financial Market

Rajesh V. Argiddi^{*1} Dr.Mrs.S.S.Apte $^{\#2}$,

$^{\#}$ *Assit Prof. at Department of computer Science & Engineering*
Walchand Institute of Technology
Solapur, India

* *Professor at Department of computer Science & Engineering*
Walchand Institute of Technology
Solapur, India

***Abstract* — The previous work is carried out on sliding window approach for fragment mining rules which results in large & complex processing the data. In this paper we present an idea to find out association within inter-transaction with different windowing approach. These approaches first minimizes the huge input dataset using tumbling window approach and then apply fragment mining to generate rules among different transactions with window length. This experimental work find out effect of different windowing approaches and select the best windowing method which will best suited for processing huge amount of data with minimum complexity . We conclude that this approach is promising one and will be suitable for predictions and useful in stock trading platforms for proper investment in Indian Stock market depend on finance sector.**

Keywords*- sliding window, tumbling window, *stock market, fragment mining.

I. INTRODUCTION

The arrival of information technology in various fields of human life has tend to the large volumes of data storage in various structure like records, documents, images, sound recordings, videos, scientific data, and many new data structures. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD) many times called as data mining, aims at the discovery of useful knowledge from large collections of data. The important reason that attracted a great deal of attention towards discovery of useful information from large dataset is due to the perception of *"we are data rich but information poor"*. There is tremendous amount of data but we hardly able to convert them in to useful knowledge for effective decision making in business.

Data mining mostly known as Knowledge Discovery in Databases (KDD) as shown in fig.1 below, it is the non trivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

TYPES OF DATA MINING

Data mining systems can be categorized according to various basis the classification is as follows[2]:

1. Types on the basis of data source type used for mining:
This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

2. Types on the basis of data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

3. Types according to the kind of knowledge discovered in mining: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

4. Types of data mining according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

Financial Market

A financial market consists of two broad sectors: (a) Money Market and (b) Capital Market. While the money market handles with short-term transaction & the capital market deals the medium term and long-term transaction. Detailed structure is represented in the Fig1.

A stock market is a public market for the trading of industrial stocks, shares and derivatives at an agreed price. These include securities listed on a stock exchange as well as those traded privately. A stock market is also called as an equity market. It is an organized with a regulatory body and the members who trade in shares are registered with the stock market and regulatory body SEBI. A stock that is highly in demand will increase in price, whereas as a stock that is being heavily sold will decrease in price. Companies that are permitted to be traded in this market place are called "listed companies". Stock market gets investors together to buy and sell shares in companies. Share market sets prices according to supply and demand.

.

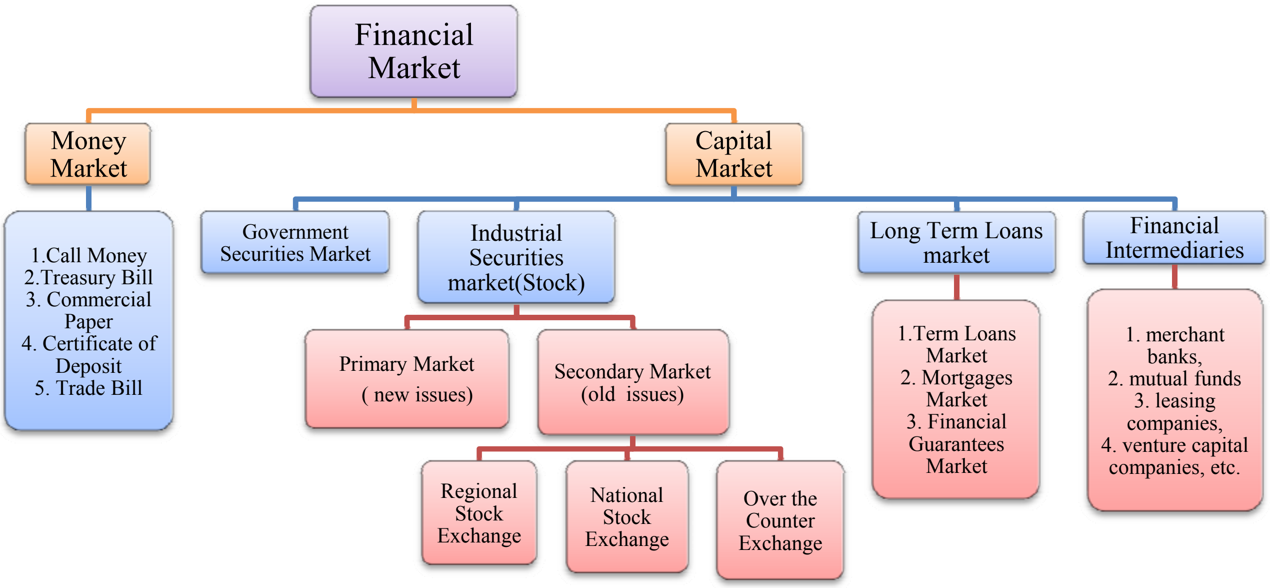


Figure 1. Overview of Financial Market

The stock market (Industrial Security) is one of the most important sources for companies to raise money. This allows businesses to be publicly raise additional capital for expansion by selling shares of ownership of the company in a public or private market. Stock markets specialize in bringing buyers and sellers of stocks and securities together

II. Related Work

One of the most important research areas in the field of Data mining is ARM. Association rules are used to identify relationships among a set of items in a transactional dataset. In the previous research, The problem of discovering association rules was first introduced in 1993 by *Agrawal et al*. and an algorithm called AIS was proposed for mining association rules. The associations‟ rules algorithm is used mainly to determine the relationships between items or features that occur synchronously in the database. For instance, if people who buy item X also buy item Y, there is a relationship between item X and item Y, and this information is useful for decision makers.

Lu, H.; Feng, L. & Han, *J*. Proposed multi-dimensional association rules from extended item-sets called E-Apriori and EH-Apriori algorithms. These Apriori inspired approaches make multiple passes over the database to find the set of frequent association rules. These algorithm uses the array & hashing methods which is not effective for large transaction since building hash table itself takes time.[3] To improve this, FITI algorithm was introduced. The First Intra Then Inter (FITI) algorithm is first invented by *"Tung, A.K.H., Lu, H., Han, J. and Feng, L."* [5] FITI is a more efficient than E-Apriori-like algorithm which initially finds the complete set of frequent intra-transaction item sets as a basis for transforming the database into a structure that aids subsequent mining of the inter-transaction item sets.

Wanzhong Yang also proposed one innovative technique to process the stock data named Granule mining technique, which reduces the size of the transaction data and generates the association rules[6].Granule mining finds interesting associations between granules in databases, where a granule is a attribute that describes common features of a set of transactions for a selected set of items. For example, a granule refers to a group of transactions that have the same attribute values. Granule mining extends the idea of decision tables in rough set theory into association mining. The attributes in an information table consist of condition attributes and decision attributes, with users requirements. Further, Prof.R.V.Argiddi has used this approach granule based mining as fragment based mining . As in granule mining, fragment based approach fragments the data sets into fragments for processing thereby reducing the input size of data sets fed to the algorithm. In contrast to granule mining, in fragment based mining the condition and decision attributes are summed for obtaining generalized association rules. The real time data has been processed based upon fragment based approach and generated the rules[7].

The Fragment mining process has two sub stages.

(1) Transform the transaction database into the form of a decision table;

(2) Divide attributes into tiers. i.e. Generate *C*-fragment and *D*-fragment based on users selected two industry categories.

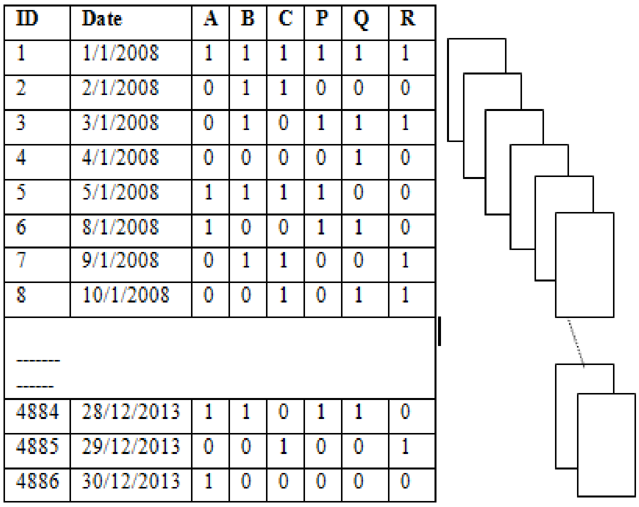
(3) Generate inter association rules between C-fragments and *D*-fragments.

III. Background

Sliding Window Vs Tumbling Window

In Sliding approach , window slids by m €{ 1,2,…n } with overlapping operational data while window size is 'n' .

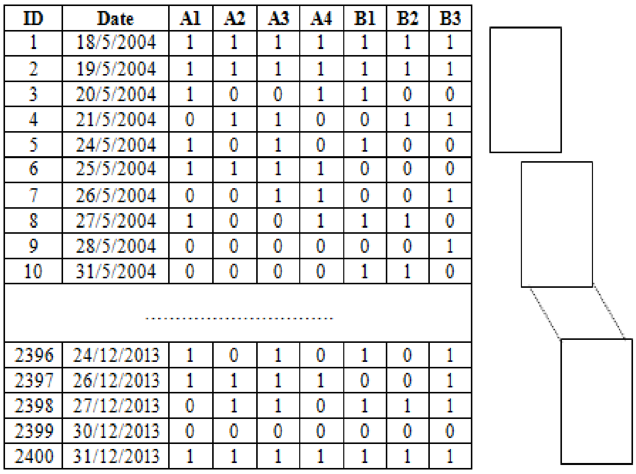
Figure 3.1 shows the overlapping sliding window approach on converted data set with window size 3 & slids by 1 .



ID	Date	A	B	C	P	Q	R
1	1/1/2008	1	1	1	1	1	1
2	2/1/2008	0	1	1	0	0	0
3	3/1/2008	0	1	0	1	1	1
4	4/1/2008	0	0	0	0	1	0
5	5/1/2008	1	1	1	1	0	0
6	8/1/2008	1	0	0	1	1	0
7	9/1/2008	0	1	1	0	0	1
8	10/1/2008	0	0	1	0	1	1
------- ------							
4884	28/12/2013	1	1	0	1	1	0
4885	29/12/2013	0	0	1	0	0	1
4886	30/12/2013	1	0	0	0	0	0

Fig 3.1. Sliding Window for Data Processing

In Tumbling approach, window slids by its size i.e exactly by ‘n’ without overlapping operational data while window size is n. Figure 3.2 shows the non-overlapping windowing approach on converted data set with window size 3 & tumble by its size .



ID	Date	A1	A2	A3	A4	B1	B2	B3
1	18/5/2004	1	1	1	1	1	1	1
2	19/5/2004	1	1	1	1	1	1	1
3	20/5/2004	1	0	0	1	1	0	0
4	21/5/2004	0	1	1	0	0	1	1
5	24/5/2004	1	0	1	0	1	0	0
6	25/5/2004	1	1	1	1	0	0	0
7	26/5/2004	0	0	1	1	0	0	1
8	27/5/2004	1	0	0	1	1	1	0
9	28/5/2004	0	0	0	0	0	0	1
10	31/5/2004	0	0	0	0	1	1	0
....								
2396	24/12/2013	1	0	1	0	1	0	1
2397	26/12/2013	1	1	1	1	0	0	1
2398	27/12/2013	0	1	1	0	1	1	1
2399	30/12/2013	0	0	0	0	0	0	0
2400	31/12/2013	1	1	1	1	1	1	1

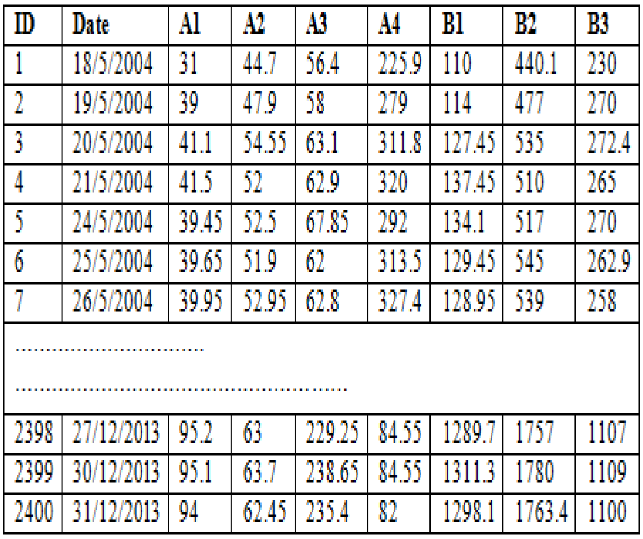
Fig 3.2. Tumbling Window for Data Processing

IV. METHODOLOGY

We propose data mining approach using association mining to solve the knowledge acquisition problems that are inherent in constructing and maintaining rule-based applications for business. Although there are an infinite number of possible rules by which we could trade, but only a few of them would have made us a profit if we had been following them. This study intends to find good sets of rules with minimal time & space complexity which would have made the most money over a certain historical period.

In FITI approach it is difficult to process an information table with large dataset and long intervals for inter transaction associations. This results into large amount of time and cost in processing the data. Fragment mining approach work on the length of window size & different windowing approaches used for finding association rules.

Let ID= {1, 2, 3,…, n} be a transaction database as shown in the Table 1.There are total 2400 rows are consisted in the above transaction table. We have considered about 10 years of Indian market data for this work.

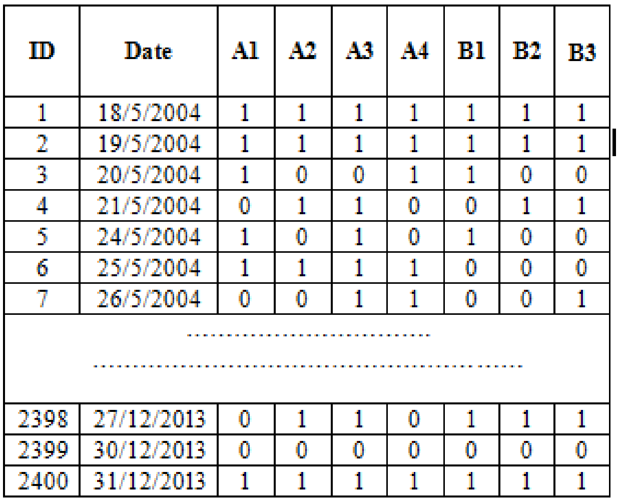


ID	Date	A1	A2	A3	A4	B1	B2	B3
1	18/5/2004	31	44.7	56.4	225.9	110	440.1	230
2	19/5/2004	39	47.9	58	279	114	477	270
3	20/5/2004	41.1	54.55	63.1	311.8	127.45	535	272.4
4	21/5/2004	41.5	52	62.9	320	137.45	510	265
5	24/5/2004	39.45	52.5	67.85	292	134.1	517	270
6	25/5/2004	39.65	51.9	62	313.5	129.45	545	262.9
7	26/5/2004	39.95	52.95	62.8	327.4	128.95	539	258
....								
2398	27/12/2013	95.2	63	229.25	84.55	1289.7	1757	1107
2399	30/12/2013	95.1	63.7	238.65	84.55	1311.3	1780	1109
2400	31/12/2013	94	62.45	235.4	82	1298.1	1763.4	1100

Table I. Sample Market Data

In this Table I. $A_{1,}$ $A_{2,}$ $A_{3,}$ A_4 , $B_{1,}$ $B_{2,}$ B_3 is the shares from Bank sector from BSE that represent Allahabad Bank ,Andhra Bank , BoB ,Federal Bank, Axis Bank ,SBI, ICICI respectively. Here we are interested to find out to find out how Finance sector affects Bank, Automobiles, Energy, IT shares etc . Here share price refers only for the open price at the transaction data , in which fragment mining algorithm with varying size of window ώ & its approach.

Our main aim is to identify the best windowing approach for processing large dataset of the transaction table and increase the performance to find out best association rules.

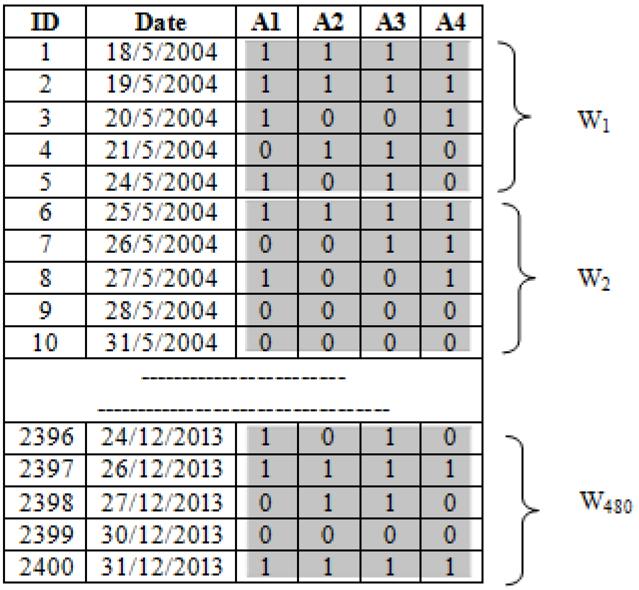


ID	Date	A1	A2	A3	A4	B1	B2	B3
1	18/5/2004	1	1	1	1	1	1	1
2	19/5/2004	1	1	1	1	1	1	1
3	20/5/2004	1	0	0	1	1	0	0
4	21/5/2004	0	1	1	0	0	1	1
5	24/5/2004	1	0	1	0	1	0	0
6	25/5/2004	1	1	1	1	0	0	0
7	26/5/2004	0	0	1	1	0	0	1
....								
2398	27/12/2013	0	1	1	0	1	1	1
2399	30/12/2013	0	0	0	0	0	0	0
2400	31/12/2013	1	1	1	1	1	1	1

Table II. Convert Table

In above Table II,we have represented increase in share price by 1, otherwise decrease in price is represented it by 0. For that purpose we compared opening price of shares for two consecutive days. ID1 represents transaction one and ID 2 represent the transaction two. we have divided the attributes into condition fragments and decision fragments. Let C be the condition fragments where $C = \{A_1, A_2, \ldots, A_m\}$ and let D be the decision attributes where $D = \{B_1, B_2, \ldots, B_n\}$.

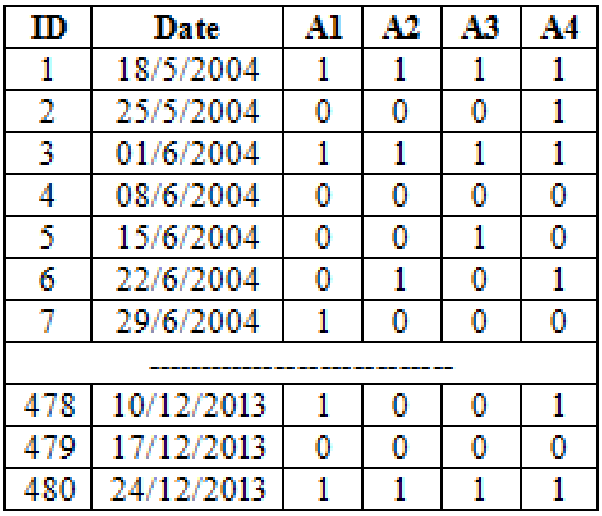
Now, A Tumbling window for transaction table II is a block of ὤ continuous intervals along time dimensions. In table III the transaction table is form of continuous Tumbling windows and window length is fixed as 5 applied on C-fragments first . So, We get the compressed C-Fragments for further processing of fragment mining as shown in Table IV. Note that Data size is reduced from 2400 to 480.



ID	Date	A1	A2	A3	A4	
1	18/5/2004	1	1	1	1	
2	19/5/2004	1	1	1	1	
3	20/5/2004	1	0	0	1	W_1
4	21/5/2004	0	1	1	0	
5	24/5/2004	1	0	1	0	
6	25/5/2004	1	1	1	1	
7	26/5/2004	0	0	1	1	
8	27/5/2004	1	0	0	1	W_2
9	28/5/2004	0	0	0	0	
10	31/5/2004	0	0	0	0	

2396	24/12/2013	1	0	1	0	
2397	26/12/2013	1	1	1	1	
2398	27/12/2013	0	1	1	0	W_{480}
2399	30/12/2013	0	0	0	0	
2400	31/12/2013	1	1	1	1	

Table III. Slide Window by its size on C-Fragments



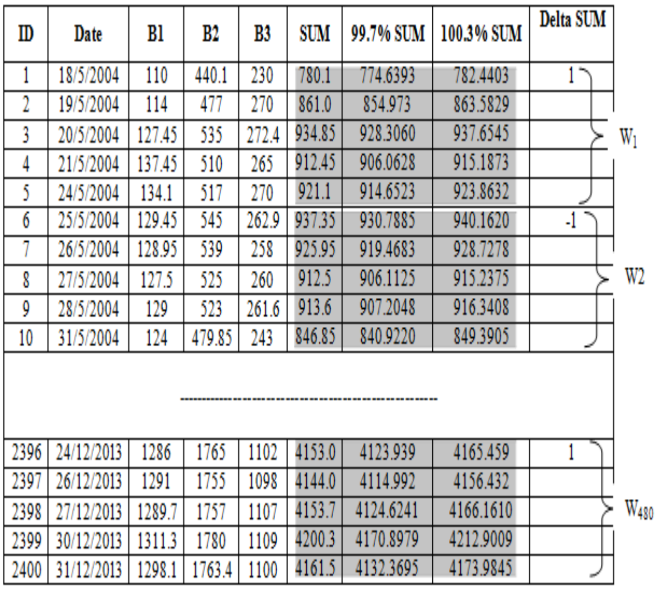
ID	Date	A1	A2	A3	A4
1	18/5/2004	1	1	1	1
2	25/5/2004	0	0	0	1
3	01/6/2004	1	1	1	1
4	08/6/2004	0	0	0	0
5	15/6/2004	0	0	1	0
6	22/6/2004	0	1	0	1
7	29/6/2004	1	0	0	0

478	10/12/2013	1	0	0	1
479	17/12/2013	0	0	0	0
480	24/12/2013	1	1	1	1

Table IV. Compressed C-Fragments Data

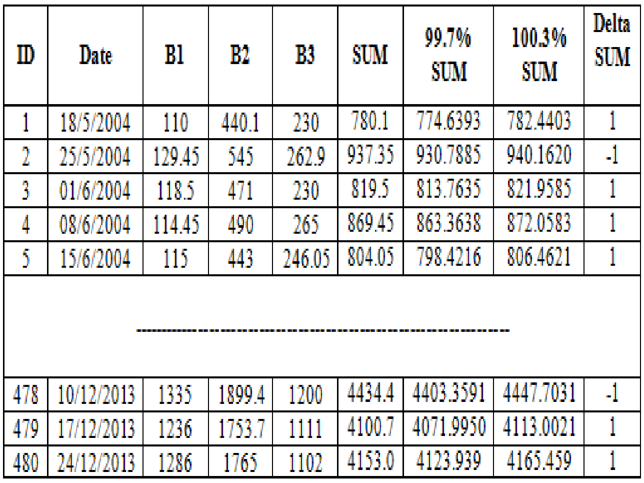
Again, A Tumbling window for transaction table I is a block of ὤ continuous non-overlapping intervals along time dimensions. In table V, the transaction table is form of continuous Tumbling windows and window length is fixed as 5 applied on aggregation of D-Fragments. So, we get the compressed D-Fragments for further processing of fragment mining as shown in Table VI while D-fragments are also reduced from 2400 to 480.

Here, Now form the covering set of C-fragments & generate the inter transaction association among C-fragments & D-fragments as shown in table VII.



ID	Date	B1	B2	B3	SUM	99.7% SUM	100.3% SUM	Delta SUM	
1	18/5/2004	110	440.1	230	780.1	774.6393	782.4403	1	
2	19/5/2004	114	477	270	861.0	854.973	863.5829		
3	20/5/2004	127.45	535	272.4	934.85	928.3060	937.6545		W_1
4	21/5/2004	137.45	510	265	912.45	906.0628	915.1873		
5	24/5/2004	134.1	517	270	921.1	914.6523	923.8632		
6	25/5/2004	129.45	545	262.9	937.35	930.7885	940.1620	-1	
7	26/5/2004	128.95	539	258	925.95	919.4683	928.7278		
8	27/5/2004	127.5	525	260	912.5	906.1125	915.2375		W2
9	28/5/2004	129	523	261.6	913.6	907.2048	916.3408		
10	31/5/2004	124	479.85	243	846.85	840.9220	849.3905		
--									
2396	24/12/2013	1286	1765	1102	4153.0	4123.939	4165.459	1	
2397	26/12/2013	1291	1755	1098	4144.0	4114.992	4156.432		
2398	27/12/2013	1289.7	1757	1107	4153.7	4124.6241	4166.1610		W_{480}
2399	30/12/2013	1311.3	1780	1109	4200.3	4170.8979	4212.9009		
2400	31/12/2013	1298.1	1763.4	1100	4161.5	4132.3695	4173.9845		

Table V. Slide Window by its size on D-Fragments



ID	Date	B1	B2	B3	SUM	99.7% SUM	100.3% SUM	Delta SUM
1	18/5/2004	110	440.1	230	780.1	774.6393	782.4403	1
2	25/5/2004	129.45	545	262.9	937.35	930.7885	940.1620	-1
3	01/6/2004	118.5	471	230	819.5	813.7635	821.9585	1
4	08/6/2004	114.45	490	265	869.45	863.3638	872.0583	1
5	15/6/2004	115	443	246.05	804.05	798.4216	806.4621	1
--								
478	10/12/2013	1335	1899.4	1200	4434.4	4403.3591	4447.7031	-1
479	17/12/2013	1236	1753.7	1111	4100.7	4071.9950	4113.0021	1
480	24/12/2013	1286	1765	1102	4153.0	4123.939	4165.459	1

Table VI. Compressed D-Fragments Data

In below Table IV represents the compressed C-fragments & Table VI represents the corresponding D-fragments. We considered window length as five for finding inter-transaction rules. A Tumbling window W can be viewed as a block of non-overlapping transactions with fixed intervals of its size , which is called maxspan. All items in the sliding window can be viewed as extended items. Hence, an inter-transaction item set refers to a set of extended items.

V. EXPERIMENTS AND RESULTS

We have collected last 10 years data of different Indian stock sectors from Yahoo Finance & divided that dataset into training & testing set as shown in table VII.

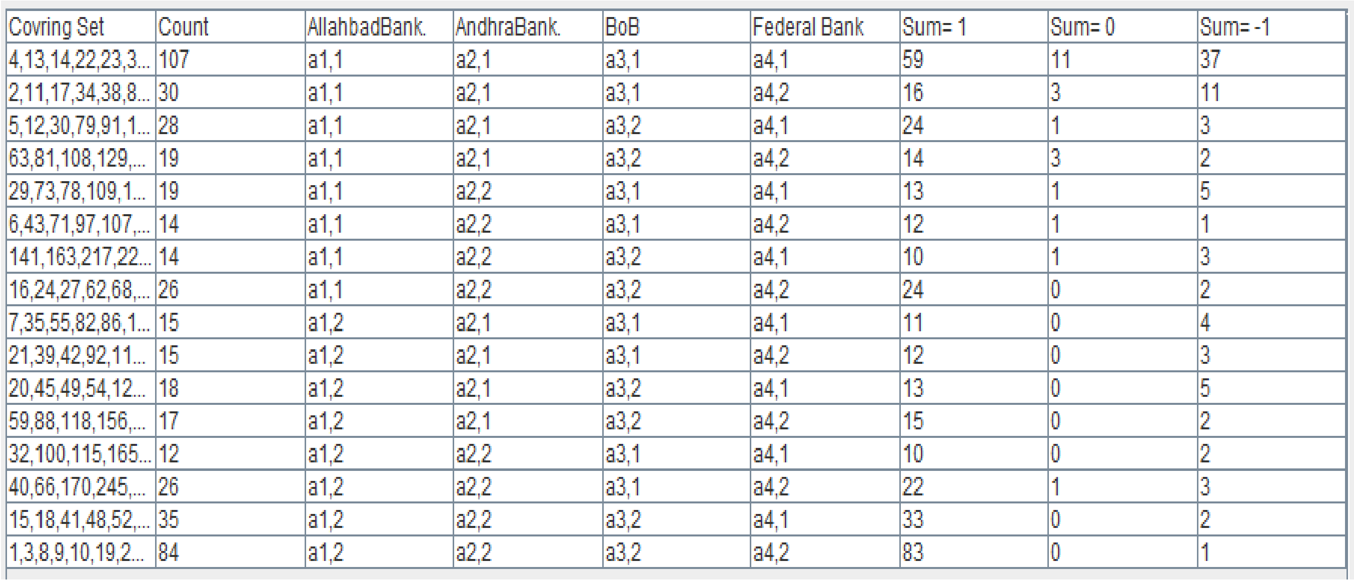
Data Set	From	To
Training Set	3 Jan 2004	30 Dec 2013
Testing Set	3 Jan 2014	20 Aug 2014

Table VII. INPUT Dataset

Fragment Mining Algorithm on Automobile sector *with tumbling window length as three*

In this method, we have collected last 10 years data of Indian Bank sector from Yahoo Finance and converted that into a tabular format and applied Fragment Mining algorithm[2] on that data set with window size 5 & slides by its size. Fragment mining approach is explained in previous paper in detail[7].

Data after data preprocessing & Fragment Mining algorithm:



Covring Set	Count	AllahbadBank.	AndhraBank.	BoB	Federal Bank	Sum= 1	Sum= 0	Sum= -1
4,13,14,22,23,3...	107	a1,1	a2,1	a3,1	a4,1	59	11	37
2,11,17,34,38,8...	30	a1,1	a2,1	a3,1	a4,2	16	3	11
5,12,30,79,91,1...	28	a1,1	a2,1	a3,2	a4,1	24	1	3
63,81,108,129,...	19	a1,1	a2,1	a3,2	a4,2	14	3	2
29,73,78,109,1...	19	a1,1	a2,2	a3,1	a4,1	13	1	5
6,43,71,97,107,...	14	a1,1	a2,2	a3,1	a4,2	12	1	1
141,163,217,22...	14	a1,1	a2,2	a3,2	a4,1	10	1	3
16,24,27,62,68,...	26	a1,1	a2,2	a3,2	a4,2	24	0	2
7,35,55,82,86,1...	15	a1,2	a2,1	a3,1	a4,1	11	0	4
21,39,42,92,11...	15	a1,2	a2,1	a3,1	a4,2	12	0	3
20,45,49,54,12...	18	a1,2	a2,1	a3,2	a4,1	13	0	5
59,88,118,156,...	17	a1,2	a2,1	a3,2	a4,2	15	0	2
32,100,115,165...	12	a1,2	a2,2	a3,1	a4,1	10	0	2
40,66,170,245,...	26	a1,2	a2,2	a3,1	a4,2	22	1	3
15,18,41,48,52,...	35	a1,2	a2,2	a3,2	a4,1	33	0	2
1,3,8,9,10,19,2...	84	a1,2	a2,2	a3,2	a4,2	83	0	1

Table VII. Final Decision Table for Bank Sector

Output Data:

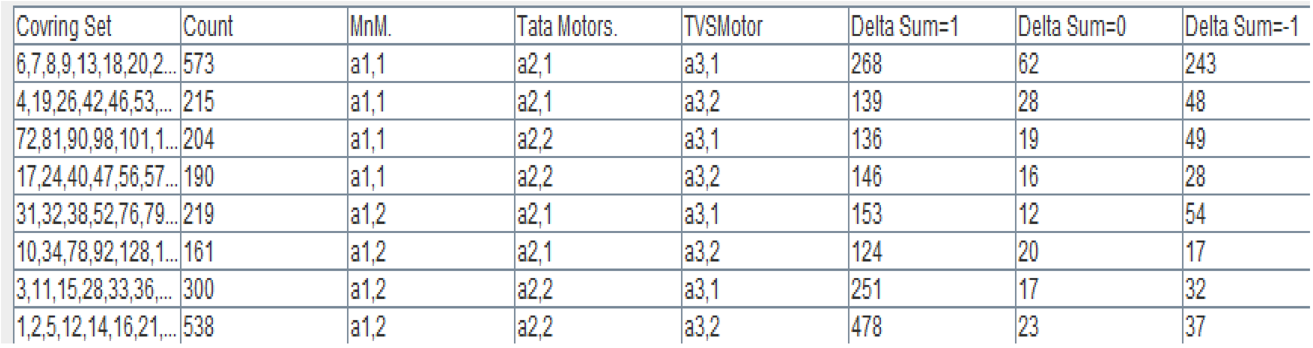
1. AllahbadBank(↓),AndhraBank(↓),BoB(↓),Federal Bank(↓) = AxisBank ,SBI, ICICI (↑)
Confidence=98.8

2. AllahbadBank(↑),AndhraBank(↑),BoB(↓),Federal Bank(↑) = AxisBank ,SBI, ICICI (↑)
Confidence=85.71

3. AllahbadBank(↓),AndhraBank(↑),BoB(↑),Federal Bank(↓) = AxisBank ,SBI, ICICI (↑)
Confidence=80.8

4. AllahbadBank(↓),AndhraBank(↓),BoB(↓),Federal Bank(↑) = AxisBank ,SBI, ICICI (↑)
Confidence=94.28

5. AllahbadBank(↑),AndhraBank(↑),BoB(↑),Federal Bank(↑) = AxisBank ,SBI, ICICI (↑)
Fitness value=55.14



Covring Set	Count	MnM.	Tata Motors.	TVSMotor	Delta Sum=1	Delta Sum=0	Delta Sum=-1
6,7,8,9,13,18,20,2...	573	a1,1	a2,1	a3,1	268	62	243
4,19,26,42,46,53,...	215	a1,1	a2,1	a3,2	139	28	48
72,81,90,98,101,1...	204	a1,1	a2,2	a3,1	136	19	49
17,24,40,47,56,57...	190	a1,1	a2,2	a3,2	146	16	28
31,32,38,52,76,79...	219	a1,2	a2,1	a3,1	153	12	54
10,34,78,92,128,1...	161	a1,2	a2,1	a3,2	124	20	17
3,11,15,28,33,36,...	300	a1,2	a2,2	a3,1	251	17	32
1,2,5,12,14,16,21,...	538	a1,2	a2,2	a3,2	478	23	37

Table VII. Final Decision Table For automobile sector

Fragment Mining Algorithm on Bank sector *with Tumbling window length as five*

In this method, we have collected last 10 years data of Indian Automobile sector from Yahoo Finance and converted that into a tabular format and applied Fragment Mining algorithm on that data set with window size 3 & slides by its size.

Output:
1. MnM(↑) ,Tata Motors(↓), TVSMotor(↓) = HeroMotorCorp , MarutiSuzuki (↑)
Confidence: 79.27
2. If MnM (↓), if Tata Motors (↑), if TVSMotor (↑) = HeroMotorCorp , MarutiSuzuki (↑)
Confidence: 69.86
3. If MnM (↓), if Tata Motors (↑), if TVSMotor (↓) = HeroMotorCorp , MarutiSuzuki (↑).
Confidence: 76.68

VI. CONCLUSION

The aim of this paper is to reduce the processing time & space of fragment rule mining algorithm that mines fragmented rules by presenting fast and scalable algorithm for discovering useful rules in large databases. For this we presented data mining approach for Fragmented item-sets . association mining methods, are usually accurate, but have very large and meaningless results. In recent years lots of work has been carried out using genetic algorithm for mining association rules. Future work includes applying genetic approach to fragment mining to get more optimized rules

REFERENCES

[1] Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[2] Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[3] Lu, H.; Feng, L. & Han, J. (2000). Beyond intra-transaction association analysis: mining multi-dimensional inter-transaction association rules. ACM Transactions on Information Systems, Vol. 18, Issue 4, (October 2000), pp. 423-454.

[4] Yang, Wanzhong and Li, Yuefeng and Xu, Yue “Granule Based Intertransaction Association Rule Mining” In Proceedings 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) 1, pages pp. 337-40, Patras, Greece.

[5] Tung, A.K.H., Lu, H., Han, J. and Feng, L “Efficient mining of intertransaction association rules” IEEE Transactions on Knowledge and Data Engineering, in 2003” .

[6] Wanzhong Yang, July 2009. “Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining”, Queensland University of Technology.

[7] R.V.Argiddi , S.S.Apte ,“ study of association rule mining in fragmented item-sets for prediction of transactions outcome in stock trading systems”, IJCET, volume 3, issue 2, july- september (2012), pp. 478-486.

[8] J.R. Quinlan, “Discovering Rules by induction from large collection of examples” in D.Michie(ed):Expert systems in the microelectronic sage, Edinburgh University press, 1979.

[9] Quinlan J.R., “Induction on decision tree,” Machine Learning, vol. 1, pp. 819106, 1986.

[10] R.V Argiddi, S.SApte “ Future Trend Prediction of Indian IT Stock Market using Association Rule Mining of Transaction data” IJCA-2012.

[11] Dr.G.Manoj Someswar, B. Satheesh, G.Vivekanand, “Finance Mining – Analysis Of Stock Market Exchange For Foreign Using Classification Techniques.” IJERA Vol. 2, Issue 4, June-July 2012, pp.717-723 717