Investment Profit Folio Decisions based on CII Algorithm for Indian Stock Market

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

Sulabha S. Apte.
Computer Science Department
Walchand Institute of Technology
Solapur, India.

ABSTRACT
The globalization in market, foreign investment and effect of current news issues makes difficult for investor to take better decisions. This paper introduces a new algorithm CII and describes the process of finding best association rules which is promising one to forecast the market. This experimental work reduces the time required for processing huge stock data and extract best rules with minimum window size to proper investment in stock market.

Keywords
CII, Forecast Market, association rule

1. INTRODUCTION
Stock market is a dynamic and unpredictable area where investor is interested to invest money and get the maximum profit [1]. So this paper introduces new algorithm which generates some rules which are more accurate and effective than previous one. This section provides basic idea of stock market, primary data mining task and concentrate more on association rule mining.

The primary goal of data mining is prediction and description which are useful for stock market. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description concentrates on finding human-interpretable patterns describing the data. However, in the context of KDD, description tends to be more important than prediction [2]. This is in contrast to pattern recognition and machine learning applications where prediction is often the primary goal of the KDD process.

The goals of prediction and description are achieved by using the following primary data mining tasks:

1) Anomaly detection (Outlier/change/deviation detection) the identification of unusual data records, that might be interesting or data errors and require further investigation.

2) Association rule learning (Dependency modeling) Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

3) Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4) Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

5) Regression – Attempts to find a function which models the data with the least error.

6) Summarization – providing a more compact representation of the data set, including visualization and report generation.

1.1. Stock market
Stock market is a place where the companies and stack holder get revenue. People are trading in the company and it is regular source of income. There are plenty of sources people get the information to make the investment in stock market such as Books, Internet, news and also from previous experience. Many times market is in unpredictable conditions so proper investment in stock market is problem for investor. So this approach is suitable for prediction in different sectors of the stock market.

Stock market allows companies for publicly trade of the business, or raises the capital with selling the shares of company. It provides companies with access to capital and also for investors with a slice of ownership in the company get the profit based on company’s future performance.

There are different types of transaction based on trading sessions in the stock market

Intraday transaction-This is transaction within the day. The term intraday is used to describe the trade on markets during regular business hours, such as opposed stocks and ETFs. This is also called as short term investment.

Interday transaction-This is transaction within week or month. Here the investment is for long term. In our work we focus on Long term investment.

1.2. Association Rule mining

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules.

Association rules are often used to analyze sales transactions. For example, it might be noted that customers who buy bread at the grocery store often buy milk at the same time. In fact, association analysis might find that 85% of the checkout sessions that include bread also include milk. This relationship could be formulated as the following rule.

Bread implies milk with 85% confidence

This application of association modeling is called market-basket analysis. It is valuable for direct marketing, sales promotions, and for discovering business trends. Association modeling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for Web page personalization.

Association rule mining finds association among the transaction. This approach is best suited for stock market where there is association among different companies or companies from one sector is dependent on another. Through the observation we come to know that if share price of crude oil companies increases transportation company's decreases. Also banking share increases real estate decreases.

In this approach we are interested to find association rules in different companies from stock market within different days. So association rules are found in same day transaction and also among different day transaction. Association rules are categorized into two types that is intraday transaction-This approach finds association rules among same day transactions and Interday transaction means association among transactions within different day. Our focus is on inter-transaction association rule mining

2. RELATED WORK

Association rule mining, also viewed as frequent pattern mining, refers to finding all frequent item-sets and generating strong association rules from the frequent item-sets. Association rule mining can also be classified into intra and inter-transaction at the point of associations among transactions [5]. The traditional association mining is classified into single dimension and multidimensional, based on the data dimensions or predicates. There are two well known algorithms: the Apriori algorithm using candidate generation [2] and the frequent pattern tree algorithm without candidate generation [6] in the single dimensional Boolean association rule.

The property of the Apriori algorithm is that all nonempty subsets of a frequent item set must be also frequent [7]. It implies that no superset of any infrequent item set could be generated or tested. This property is used widely in data mining techniques. In particular, it is supported by the anti-monotone property. If a set cannot pass a test, all of its supersets fail the same test as well. The Apriori algorithm involves both joining and pruning. The Apriori algorithm generates length k+1 candidate item sets from length k frequent item sets and uses the Apriori property to prune the item sets. The association rule is generated from frequent item sets.

For multidimensional association mining, Han et al. [4] summarized the possible techniques in accordance with the corresponding treatments of quantitative attributes. Habitually, most current researchers on multidimensional association mining endeavor to use the existing efficient algorithms for single dimensional mining. Lee et al. [8] presented an approach for multidimensional constraints that checked constraints during FP-tree constructions. The approach first grouped products of the same cost and price into an item, and viewed the product table as a set of transactions.

3. BACKGROUND

3.1 Apriori

This algorithm is developed by Agarwal and Srikant in 1994[2], which provide Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item, Based on minimum support threshold.

The property of the Apriori [4] algorithm is that all non empty subsets of a frequent item set must be also frequent. It implies that no superset of any infrequent item set could be generated or tested. This property is used widely in data mining techniques. In particular, it is supported by the anti-monotone property. If a set cannot pass a test, all of its supersets fail the same test as well.

The algorithm attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the item-sets.

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

The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

3.2 FPT Frequent pattern tree

The FPT algorithm searches shorter frequent patterns instead of finding long frequent patterns. This algorithm often takes a long time to find complete frequent item sets.

In the first phase, there are a lot of item sets, including some that are extra and useless. In the second phase, support and confidence are two important measures to generate association rules. Support refers to the frequency of the item set while confidence reflects the certainty of association rules.

4. CII Algorithm

This method breaks though previous methods, such as Apriori, and demonstrates a clear concept for representing Inter-transaction association. The CII transaction algorithm is best approach in sliding windows [3]. This algorithm is based on candidate set generation by the joins. It inherits the advantage of the Apriori idea and can generate a complete set of frequent item sets.

The CII algorithm[9] assumes that the intervals (or the sliding window length) should be very small for the inter-transaction associations, because if there are many attributes in one sliding window, the candidate set can be too large for finding frequent item sets. However, there are different Intervals required in industry For example; the intervals could be a few days or even several months for a short-term investment of the stock market.

The CII algorithm has been the state of the art method for the sliding window based approach. The CII algorithm uses the sliding windows to separate attributes into tiers, and uses small sliding windows for certain tiers, based on user's constraints. In Figure 1.1, there are seven attributes in an information table. The sliding window in the CII algorithm includes all attributes; the number of intervals is as shown in Table I.

ID	C1	C2	C3	C4	D1	D2	D3
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
. . . .	. . . .						
500							

Table I: CII Algorithm with Sliding Window

The approach divides the attributes into two tiers: condition attributes (C1, C2, C3, C4) and decision attributes (D1, D2, D3).For the finding association rules we will consider both condition as well as decision attributes. In the above transaction table the sliding window size ω considered as a 4.

4.1 Data collection and preprocessing

In this experiment we collect the stock data of BSE for 40 companies from the YAHOO finance. Here we consider the date, High price, low price and opening and closing price of each company. If closing price is zero then here consider the closing price of last day. So data is preprocessed in unique format for processing.

Data is gathered for BSE as shown in the following Table II.

4.2 Working

After data collection and preprocessing we allow user to select the condition attribute and decision attribute. Here we consider the condition attribute as small scale companies and large scale companies as decision attribute.

Date	Open	High	Low	Close	Volume	Adj Close
05-03-2012	11.35	11.4	11	11.25	106600	11.25
06-03-2012	11.25	11.35	10.9	10.9	74200	10.9
07-03-2012	10.9	11	10.55	10.95	73700	10.95
09-03-2012	11.2	11.3	10.85	11.05	89400	11.05
12-03-2012	11.15	11.9	11.1	11.35	551000	11.35
13-03-2012	11.75	11.75	11.15	11.25	167800	11.25
14-03-2012	11.4	11.45	11.1	11.1	96500	11.1
15-03-2012	11.15	11.2	10.75	11	113100	11
16-03-2012	10.85	11.1	10.65	10.7	102900	10.7
19-03-2012	11.25	11.4	10.5	10.5	228000	10.5
20-03-2012	10.95	11.35	10.85	11.05	334100	11.05
21-03-2012	11.1	11.2	10.8	10.85	129100	10.85

Table II: Stock data from BSE

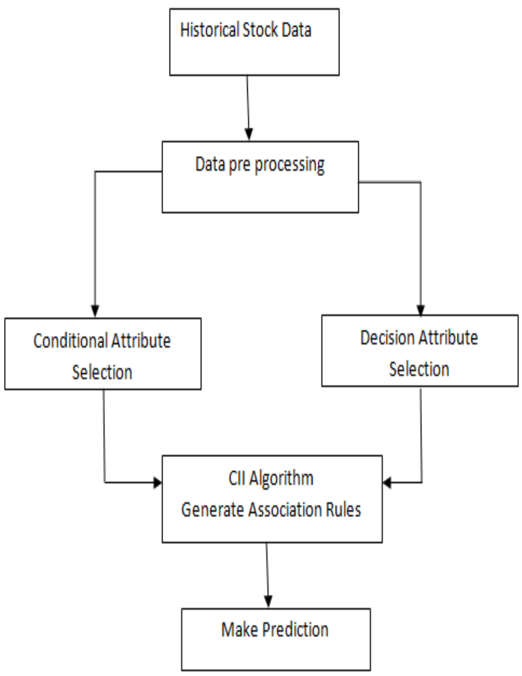

Fig I:Working Diagram of CII

Here in fig. I it is shown that both condition attribute and also decision attribute are used for finding association rules. Before this we allow user to select the window size.

If the user select window size as 4 and the rules are generated when Infosys UP then TCS also UP; it indicate if today Infosys UP then after 4 days TCS also UP.

4.3 ALGORITHM

Step 1. Preprocess Data with missing value

Step 2. Set all default values and also minimum confidence and minimum support

Step 3. Searching for most frequent values

Algorithm divides the generation of frequent inter-transaction item set I into two cases. k = 2 and k > 2 in order to control the implementation.

a) When k= 2, CII generates frequent inter-transaction two-item sets, L2. CII makes use of the hashing approach in previous research and refines the hash formula for the bucket number.

b) When k >2, CII uses the loop to generate the candidates of frequent inter-transaction item sets.

While(Lk≠Ø)

{

Generate candidate inter-transaction item sets, Ck;

Scan transformed database to update the count for Ck;

Let Lk = {c∈ Ck | support(c) ≥ mins up};

k++;

}

Step 4.Frequent item set generated based on join. Join are inter-transaction join and cross transaction join.

Step 5. Generate Association rules according to order of confidence value.

Step 6. Prediction on basis of rules

5. Result and Analysis

Here we consider stock data from BSE for 40 companies. We gathered data for last 5 years.

In this table III there are the KPIT, Mphasis, MahaSatyam, Zensar are the conditional attribute and TCS, Infosys, Wipro are the Decision attributes. On the basis of conditional attribute the decision attribute are dependent. Conditional attribute are all small scale companies and decision attribute are large scale companies. In first rule when TCS goes up after 5 days Infosys goes UP with 75 percentage confidence. In second rule if Infosys goes down then TCS goes down with 74 percent confidence value. When Infosys goes UP then TCS also goes UP with 74 percent confidence value. Rule four suggest that when Mphasis down TCS also down with 66 percent of confidence. Final rules shows that when Infosys down MahiSytm also down.

Here same experiment is carried out on different data set and we found that this result are effective than previous apriori algorithm. These results are more accurate when compare with future value.

Rule No.	Condition Attribute				Decision Attribute			Window Size	Confidence In Percent
	KPIT	Mphasis	MahiSytm	Zensar	TCS	Infosys	Wipro		
1	--	--	--	--	UP	UP	--	5	75
2	--	--	--	--	DOWN	DOWN	--	5	74
3	--	--	--	--	UP	UP		5	71
4	--	DOWN	--	--	DOWN	--	--	5	66
5	--	DOWN	--	--	--	DOWN		5	65

Table: III Results with Association Rules

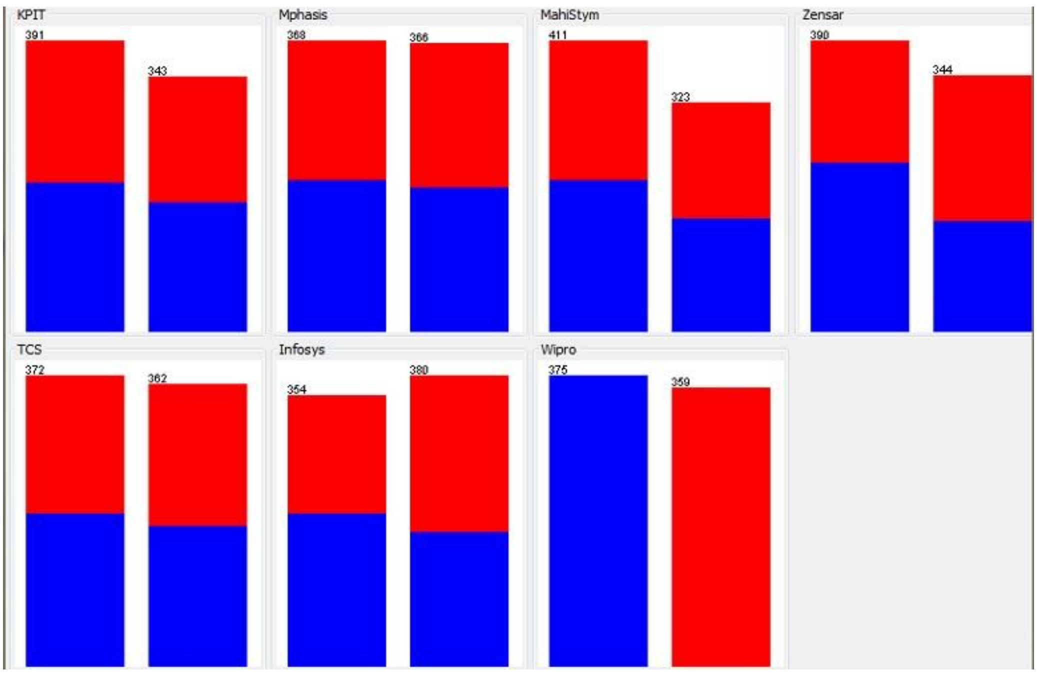

Fig II: Graphical Views for Association Rule Analysis

In the above fig II Red color is indicate UP and Blue color indicate down. So total 7 companies are represented with their performance. The setup is carried out with use of data mining tool Weka.

6. Conclusion

This current work of CII algorithm is carried out to mine inter-transaction association rule rather than intra-transaction. The proposed method proves it is better than previous methods and its performance is acceptable for real life industry. The evaluation on both efficiency and effectiveness shows that CII based approach has potential than previous used methods

7. Future Scope

This method is more suitable for different research areas like business intelligence, web applications and there is scope in different fields with industrial and commercial requirements.

CII based approach is more suitable for finding inter-transaction association rules on single item based pattern, so we can extend our work for multi-dimensional association rule mining.

8. REFERENCES

[1] **Eugene F Fama**: "The Behavior of stock Market Prices", The Journal of Business, Jan 1995

[2] R**. Agrawal, R. Srikant**: Fast algorithms for mining association rules, in: Proceedings of International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 487–499.

[3] **A.K.H. Tung, H. L., J. Han, and L. Feng (2003):** "Efficient mining of inter-transaction association rules." IEEE Transactions on Knowledge and Data Engineering 15(1): 43-56.

[4] **Agraw, R., Imielinski, T., Swami, A.,** "Mining association rules between sets of items in large database", Proceedings of ACM-SIGMOD, Montreal, Canada, 1993

[5] **Antony Tung,Jaiwei Han,Ling Feng** : Beyond Intra-transaction Association Analysis: Mining Multi-dimensional Intertransaction Association Rules (Lu, Feng, Han, ACM journal, 30 pages)

[6] **J. Han, J. Pei, and Y. Yin (2000):** Mining frequent patterns without candidate generation. The 2000 ACM SIGMOD international conference on Management of data table of contents, USA.

[7] **A.K.H. Tung, H. L., J. Han, and L. Feng (2003):** "Efficient mining of inter-transaction association rules." IEEE Transactions on Knowledge and Data Engineering 15(1): 43-56.

[8] **A. J. T. Lee and C. Wang (2007)**: "An efficient algorithm for mining frequent inter- transaction patterns." Information Sciences 177(17): 3453-3476.

[9] **Ranjeetsingh B.Parihar and Rajesh V.Argiddi (ITBI 2011):** "Combined Intra-Inter transaction based approach for mining Association among the sectors in Indian Stock Market" In ITBI 2011). Page No.26