Association Rules Mining in the Stock Data

Mukesh Kumar and Arvind Kalia
1Dept. of Computer Sciences, Himachal Pradesh University SHIMLA
2Professor and Chairman, Dept. of Computer Sciences, Himachal Pradesh University SHIMLA

Abstract

Association rules mining is an important subject in the study of data mining. Data mining is the process of finding valid, useful and understandable pattern in data. Due to the large size of databases, importance of information stored, and valuable information obtained, finding hidden patterns in data has become increasingly significant. A time series data set consists of sequences of values or events that change with time. Time series data is popular in many applications, such as the daily closing prices of a share in a stock market. Stock data mining plays an important role to visualize the behavior of financial market. Association rule mining algorithms can be used to discover all item associations (or rules) in a dataset that satisfy user-specified constraints, i.e. minimum support and minimum confidence. Since only one minimum support is used for the whole database, it is implicitly assumed that all items are of the same nature and/or have similar frequencies in the data. Patterns are evaluated by means of generating itemsets using a predefined support and association rules with a higher confidence level. The pattern generated by the frequent itemset of size three is found to be same as being reflected by means of obtained association rules. The pattern so generated helps investors to build their portfolio and use these patterns to learn more about investment planning and financial market.

I. INTRODUCTION

Association is the discovery of association relationships or correlations among a set of items. Market Basket analysis can also help retailers to plan which items to put on sale at reduced rates. For a set of items available at the supermarket, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of the values assigned to these variables. The Boolean vector can be analyzed for buying patterns. These patterns can be represented in the form of association rules. Support and confidence are two measures of the rule interestingness. Association rules are considered interesting if both minimum support threshold and a minimum confidence threshold is satisfied. Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence [1].

Association rules mining is an important subject in the study of data mining. Data mining is an interdisciplinary field, having applications in diverse areas like bioinformatics, medical informatics, scientific data analysis, financial analysis, consumer profiling, etc. In each of these application domains, the amount of data available for analysis has exploded in recent years. A time series data set consists of sequences of values or events that change with time. Time series data is popular in many applications, such as the daily closing prices of a share in a stock market. Taking stock data as an example, it can be found the rules like “If stock A goes up and stock B goes up then stock C will goes up on the same day (5%, 75%)” with intra-transactional association rules, but it cannot be found the rules like “If stock A goes up on the first day and stock B goes up on the second day then stock C will goes up on the third day (5%, 75%)”. There is no time difference between the items in the intra-transactional association rules, so it cannot be used to predict the trend of time series.

Data mining is the process of finding valid, useful and understandable pattern in data. Due to the large size of databases, importance of information stored, and valuable information obtained, finding hidden patterns in data has become increasingly significant [16,21]. The stock market provides an area in which large volumes of data is created and stored on a daily basis, and hence an ideal dataset for applying data mining techniques [1, 5, 9]. Every investor wants to know or predict the trends of the stock trading. Some of the investors use fundamental data analysis, such as i.e. annual report to predict if there is any potential profit in its future stock trading. Others use the technical approach, such as various short-term, middle-term, and long-term indicators to decide when to buy or sell stocks. As an objective tool, data mining can be applied to discover the interesting behavior within a time series or the relationship among a set of time
series so that investors can collect more useful information from the already available but huge amount of data [9, 18]. For example, looking for repetitive patterns in a stock time series can be very useful for stock investors.

There are many types of pattern or knowledge that can be discovered from data. For example, association rules can be mined for market basket analysis, classification rules can be found for accurate classifiers, clusters and outliers can be identified for fraud detection [20]. Frequent pattern mining plays an important role in many data mining tasks, such as mining association rules, correlations, and sequential patterns and emerging pattern. Frequent itemset generation is the prerequisite for association rule mining. The association rule is an important kind of rule set, applicable to databases. Since it was firstly proposed in [1], the association rule has gained much attention thereafter. The association rule reflects the relationship or correlative relationship among itemsets in large amounts of data. The work of association rules mining can be decomposed in two phases first is frequent itemsets generation i.e. to find out all itemsets that sufficiently exceed the minimum support and second is rules construction i.e. from the frequent itemsets generate all association rules having confidence higher than minimum support [11, 24].

In this paper, patterns are evaluated by means of generating itemsets using a predefined support and association rules with a higher confidence level. By identifying individual stocks in generated pattern and their similar behavior or category one can build its portfolio to have better returns. One can gain insight into the underlying pattern, which is helpful in further analysis, such as stock market forecasting. One of the main challenges in stock data mining is to find the effective knowledge representation of the stock dataset [2, 10, 13]. As such, the main objective of this study is to explore the suitability and a comparative study of the performance of itemsets and association rules from a stock dataset. Our experimental results demonstrate notable similarity in both types of patterns as well as categorization of stocks.

II. RELATED WORKS

Association rule discovery has been an active area for database researchers for a number of years. The problem of association rule mining was first introduced in [1]. The paper identified the discovery of frequent itemsets as a key step in association rule mining. In [2], the authors presented the basic algorithm, called Apriori, that quickly became the seed of several other data mining algorithms. However, recent works have demonstrated promising results of representing financial data symbolically. However it is argued [12, 17] that symbolic relational data mining is more suitable in incorporating background knowledge. Their proposed methodology outperforms numeric financial data in generating IF-Then rules. Sequential and non-sequential association rule mining (ARM) were used to perform intra and inter-stock pattern mining [14, 22], where each stock is represented symbolically based on its performance with respect to a user-defined threshold. There have been numerous strategies applied to the problem of discovering frequent itemsets. In [5], the authors apply significance testing to associations, essentially searching for correlation using a chi-squared test. Another approach involves sampling [23], in which a subset of the dataset is analyzed for likely associations. These associations from the sample are then verified in the complete dataset.

An innovative counting scheme is employed in [3], which begins counting larger itemsets based on partially reading the dataset. The number of passes is reduced through the data, since for large datasets, the I/O costs are significant. Research in feature reduction and product hierarchies [6, 19] is aimed at controlling the potential exponential blowup of candidate itemsets. Separate and distinct approaches presented utilizing SQL to discover itemsets [7, 16]. In [15], the authors described results using multiple minimum support levels and their goal was to discover rare itemsets. Although discovery of frequent itemsets to generate pattern for creating portfolio is the main concern and the generation of association rules exhibiting matching pattern so that an investor can plan their investment strategies are our main attempt to carry out the required task.

III. DATA PRE-PROCESSING

For the purpose of this study, the stocks dataset of thirteen years period i.e. from Jan. 1996 to Dec.2008 of NSE stock exchange that amounts to 3252 days was used [25]. The trading of the stock market within a day is recorded in a single text file. Each line represents the trading information of a stock. Before data is scrutinized
and processed for problem, it must be thoroughly inspected, cleaned and selected. Since even the best predictor will fail on bad data, data quality and preparation is crucial. Also, since a predictor can exploit only certain data features, it is important to detect which data preprocessing/presentation works best. Data cleaning fills in missing values, smoothes noisy data, handles or removes outliers, resolves inconsistencies. Missing values can be handled by a generic method [6]. The noisy data is very familiar property of real-world datasets, so majority of time was spent in these experiments on data pre-processing. The noise exists in the stock dataset can be generally classified into three main categories, which are duplicate records, inconsistent stock codes and incomplete files. Certain files contain duplicate records. Thus a thorough checking is applied on all the files to ensure only unique records on the same date. In most of these files, stocks are represented by stock codes. For example, Tata Motors is represented as Telco and in some cases it is represented as TataMotors. TataSteel is represented as Tisco where in some files it is as TataSteel, which causes lots of inconsistencies. This inconsistency gets more complicated when the listed companies change their names or when companies merge. In that case data is partly available with old name and remaining data is with new name. Lot of time is spent on tackling such inconsistencies [6, 8]. Missing values are observed when certain stocks or indices record null values on certain days [4,13]. For the purpose of this study, only stocks which have at most ten null values are selected. Methods include skipping the whole instance with a missing value, or filling the missing value with the mean/new ‘unknown’ constant, or using inference, e.g. based on most similar instances. The existing null values in all the selected records were filled by using the average of its first left and first right non-null values.

IV. RESEARCH METHODOLOGY

Time series data is difficult to manipulate, but when they are treated as symbols (item units) instead of data points, interesting patterns can be discovered and it becomes an easier task to mine them [6,8]. Thus, it is suggested to convert the basic unit into symbols, i.e., numeric-to-symbolic conversion. The numeric-to-symbolic conversion transforms the available features (e.g. Open, High, Low, Close prices) of a financial instrument into a string of symbols. In other words, the numeric data sequences from each stock time series are interpreted and a unique symbol is then used to label them individually. Such a conversion process can be extended to granulate the numerical data into different time granularities and it provides a large collection of symbol strings, hopefully at various time granularities, which can then be used for different applications.

First of all we convert the numeric representation of the data to symbolic one. For symbolic representation we make use of the Open, High, Low and Close prices to carry out the numeric-to-symbolic conversion [12]. Here, one of the challenges being faced is to determine an appropriate number of symbols that is representative and also flexible enough for different time series. If the number of symbols is too many, then the occurrence of each symbol would be infrequent, making the mining process and the subsequent prediction task difficult. Even the rule can be generated with high confidence, say 100%, the pattern may not happen again and hence the rule is useless. On the other hand, if the number of symbols is too few, the support of each symbol would increase but the confidence may not be high enough and the interestingness of the mined rules is questionable.

So, in this paper, only one feature is taken, i.e., the price movement consisting of three values/possibilities:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>up</td>
<td>(close-open)/close &gt; threshold</td>
</tr>
<tr>
<td>down</td>
<td>(open-close)/close &gt; threshold</td>
</tr>
<tr>
<td>neutral</td>
<td>(close-open)/close = threshold</td>
</tr>
</tbody>
</table>

\[ e.g., \text{for open=100, close=101.25 and threshold=1\%}, \text{a up feature value will be generated.} \]
\[ \text{for open=100, close=98.9 and threshold=1\%}, \text{a down feature value will be generated.} \]
\[ \text{for open=100, close=99.3 and threshold=1\%}, \text{a neutral feature value will be} \]
The Threshold is a user-defined parameter.

A. Association Rules

The association rules try to discover association or correlation relationships among a large set of data items. They identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form X \implies Y where X and Y are disjoint conjunctions of attribute-value pairs. The confidence of the rule is the conditional probability of Y given X, \( \text{Probability}(Y|X) \), and the support of the rule is the prior probability of X and Y, \( \text{Probability}(X \text{ and } Y) \).

\begin{align*}
\text{Support} (A \implies B) &= P (A \cup B). \\
\text{Confidence} (A \implies B) &= P (B|A). 
\end{align*}

The support count aims at alienating the item sets which occur infrequently within the data set and hence are irrelevant in the final associations. The confidence establishes the intensity of the association whether it is weak, moderate or strong.

B. Itemset

An itemset is a set of items. Each item is an attribute value. In the portfolio example, an itemset contains a set of stocks such as Reliance Capital, Reliance, TataSteel. Each itemset has a size, which is the number of items contained in the itemset. The size of itemset [Reliance Capital, Reliance, TataSteel] is 3. Frequent itemsets are those itemsets that are relatively popular in the dataset. The popularity threshold for an itemset is defined using support.

C. Support

Support is used to measure the popularity of an itemset. Support of an itemset \([A, B]\) is made up of the total number of portfolios that contain both Stock A and Stock B.

\[ \text{Support} (\{A, B\}) = \text{Number of Transactions (A, B)}. \]

D. Confidence

Confidence is a property of an association rule. The confidence of a rule \( A \Rightarrow B \) is calculated using the support of itemset \([A, B]\) divided by the support of \([A]\). It is defined as follows:

\[ \text{Confidence} (A \Rightarrow B) = \frac{\text{Support} (A, B)}{\text{Support} (A)}. \]

Here the problem encountered when the appearance frequencies were different among different data items. (1) If the minimum support was set very high, the association rules supported by data items having low appearance frequencies were lost; (2) To discover all association rules supported by data items having low appearance frequencies and those having high appearance frequencies at the same time, the minimum support must be set very low. However, a very low minimum support value resulted in “combination explosions” and produced too many rules. This happened because these association rules having high appearance frequencies were relevant to each other in all possible ways. Therefore, most of these rules were meaningless. In view of this an optimum value of support and confidence level is chosen so that meaningful rules might be drawn and a meaningful pattern may be emerged.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>PARAMETERS SETTINGS FOR OUR STUDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Value</td>
</tr>
<tr>
<td>Minimum Support</td>
<td>0.15</td>
</tr>
<tr>
<td>Minimum Confidence</td>
<td>0.60</td>
</tr>
<tr>
<td>Maximum Itemset Size</td>
<td>3</td>
</tr>
<tr>
<td>Minimum Itemset Size</td>
<td>2</td>
</tr>
</tbody>
</table>

V. EXPERIMENTAL RESULTS AND OBSERVATIONS

The frequent itemsets are presented in table 2 and table 3. Every frequent itemset represents a portfolio, and the size of it equals the size of an itemset, it is two in Table 2 and it is three in table 3. Each of these stocks is a blue chip stock (a stock of a well-established company).

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>FREQUENT ITEMSETS OF SIZE TWO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent Itemsets</td>
<td>Size</td>
</tr>
<tr>
<td>Reliance Capital, Reliance, TataSteel</td>
<td>3</td>
</tr>
<tr>
<td>Reliance Capital, Reliance, TataSteel</td>
<td>3</td>
</tr>
</tbody>
</table>
The eight itemsets of size two comprise of only five stocks (Abb, Hll, Cipla, Ranbaxy and Sunpharma) and the interesting observation is that out of these five stocks three are of the same category i.e. Pharma while other two are of FMCG and diversified. The four other itemsets of size three comprise of only six individual stocks (Reliance Capital, Reliance, Sail, Tata Steel, Tata motor and L&T). Two of them belong to steel, one is of finance, one is of auto and two belong to diversified. Out of six, these two set of companies belong to the same group like (Reliance, Reliance Capital) and (Tata steel, Tata Motor). Further it is imperative to note that there is no common stock in these two groups of itemsets of size two and size three.

The same phenomenon is observed in the association rules, as shown in Table 4. Many association rules are generated with confidence as .60, but only top ten association rules are considered. The interpretation of the formula, “Stock A -> Stock B” is explained as those who purchase Stock A are likely to purchase Stock B with the possibility of confidence, and Stock B is the frequent partner (FP) stock of Stock A. For instance, in Table 4, stock Reliance Capital is the FP stock of the union of Sail and Tata Steel in first row of association rules. These ten rules contain six individual stocks and in each rule there is always one stock of steel as well as one
stock of finance may be in antecedent part or consequent part. Reliance Capital is occurring at four places in consequent side, whereas steel stocks either Sail or Tata Steel are occurring at 5 places in consequent side and Reliance is occurring at one place in consequent side. The interesting aspect is that the stocks appeared in the frequent itemset pattern of size 3 is exactly the same that appeared in the association rules, so the pattern generated by the frequent itemset of size 3 is same as being reflected by means of association rules. So it may be concluded that there is a similarity in the pattern generated by frequent itemsets and association rules.

VI. Conclusion and Further Work

The pattern is presented about how a portfolio is built and how an investor can make use of this to decide about more investment plans and help understanding the finance market.

In this paper, the association rules algorithm is applied to a real time dataset, and a group of association rules and frequent itemsets are obtained. First itemsets are mined for a given minimum support and based on these itemsets obtained, the association rules are computed for a given minimum confidence. The pattern so generated helps investors to build their portfolio and using these patterns investment strategies may be planned. Further it is observed that a portfolio is safely built with major large cap companies as it is evident that mainly pattern recommends (Reliance and Reliance Capital) or (Tata Steel and Tata Motor). It is also discovered that a portfolio may be built with one finance company and one Steel company, which will definitely give better returns in long term. Further multiple minimum support may be specified based on the different characteristics of association rules in order to discover them. Using this strategy the problem of infrequent data items can be solved without creating an amount of invalid rules implicated in frequent data items. By extending the traditional association rules models, this work can set different minimum support for every data item. Thus it may overcome the problem that the traditional methods with the single minimum support can not completely reflect the different appearing frequencies and natures of different data items.

REFERENCES


7. Han, J., and Fu, Y. “Discovery of multiple-level association rules from large databases”. In VLDB’95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland (1995),


9. Keogh, E. “A Decade of Progress in
Indexing and Mining Time Series Data". Tutorial of the 32nd International Conference on Very Large Data Bases, Seoul, Korea. 2006.


13. Liu, B., Hsu, W., and Ma, Y. “Mining association rules with multiple minimum supports”. In Knowledge Discovery and Data Mining, 1999, pp. 337-341.

14. Meo, R., Psaila, G., and Ceri, S. “A new sql-like operator for mining association rules”. In VLDB’96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India.


23. Toivonen H. “Sampling Large Databases for Association Rules”. In Proceedings of the 22nd International Conference on Very Large Data Bases, 1996, Bombay, India.,


25. http : www.nseindia.com