APPLICATION OF DATA MINING TOOLS FOR SELECTED SCRIPTS OF STOCK MARKET

K. S. Mahajan[1] and Dr. R. V. Kulkarni[2]

[1]Research student, Chh. Shahu Institute of Business Education and Research Center, Kolhapur, India
[2]Professor and HOD, Chh. Shahu Institute of Business Education and Research Center, Kolhapur, India

ABSTRACT

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions Therefore, Investment can be considered as one of the fundamental pillars of national economy. So, at the present time many investors look to find criterion to compare stocks together and selecting the best and also investors choose strategies that maximize the earning value of the investment process. Therefore the enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. Therefore research in data mining has gained a high attraction due to the importance of its applications and the increasing generation information. So, Data mining tools such as association rule, rule induction method and Apriori algorithm techniques are used to find association between different scripts of stock market, and also much of the research and development has taken place regarding the reasons for fluctuating Indian stock exchange. But, now days there are two important factors such as gold prices and US Dollar Prices are more dominating on Indian Stock Market and to find out the correlation between gold prices, dollar prices and BSE index statistical correlation is used and this helps the activities of stock operators, brokers, investors and jobbers. They are based on the forecasting the fluctuation of index share prices, gold prices, dollar prices and transactions of customers. Hence researcher has considered these problems as a topic for research.

KEYWORDS

Stock Market, Association Rules, Rule Induction Methods, Apriori Algorithm, Correlation, Data Mining.

1. INTRODUCTION

Data mining, the science and technology of exploring data in order to discover previously unknown patterns, is a part of the overall process of knowledge discovery in databases (KDD). In today's computer-driven world, these databases contain massive quantities of information. The accessibility of this information makes data mining important and necessary. Data mining often can improve existing models by finding additional, important variables, indentifying interaction terms and detecting nonlinear relationships.

DOI : 10.5121/ijdkp.2014.4405

Financial institutions such as stock markets produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Potential significant benefits of solving these problems motivated extensive research for years. Specifics of data mining in finance are coming from the need to accommodate specific efficiency criteria (e.g., the maximum of trading profit) to prediction accuracy, coordinated multiresolution forecast (minutes, days, weeks, months, and years), Be able to benefit from very subtle patterns with a short life time, and incorporate the impact of market players on market regularities , Impact of gold and US dollar prices on stock market and also to find association between different scripts of stock market which helps investors to earn more profit.

The techniques that are used in this project are:

1. Association rules
2. Apirori algorithm
3. Rule induction Method
4. Statistical Correlation

1.1 Association Rule:

Unlike the other data mining functions, association is transaction based. In transaction processing, a case consists of a transactions such as a market basket analysis. The collection of items in the transaction is a multi- record attributes.

Association rules are IF/THEN Statements.

Example: "if a customer purchases Infosys Ltd, Then customer also purchases Wipro Ltd with 60% confidence".

An association rule has two parts, an antecedent (if) and a consequent (then), an antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association Rule is created by analyzing data for frequent IF/THEN patterns & and using the criteria Support & Confidence to identify the most important relationships. Support and Confidence are two measures of association rule.

Association Rule take following form x=>y, where x and y are the sets of items. The goal is to discover all the rules that have the Support & Confidence greater than or equal to the minimum support and minimum confidence respectively.

Steps To Generate Association Rules:

1. Generate all possible association rules.
2. Compute the support and confidence of all possible association rules.

3. Apply two threshold criteria minimum support and minimum confidence to obtain association rule.

4. Minimum support and minimum confidence is taken as an average of all the calculated support and calculated confidence.

5. If the calculated support and confidence is greater than or equal to the minimum support and minimum confidence then these items are said to be associated with each other by association rule.

SUPPORT: The Support of a rule indicates how frequently the item in the rule occurs together.

Example: Dr.Reddy's lab and Cipla Ltd might appear together in 10% of the transaction.

Support is calculated as below:

Support (x=>y) = (Number of transaction Containing x&y) / (Total Number of transaction).

CONFIDENCE:

Confidence is the number of times the IF/THEN statements have been found to be true. The confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction.

Example: Dr.Reddy's lab might appear in 20 transactions, 10 of the 20 might also include Cipla Ltd.

Therefore Dr.Reddy's Lab implies Cipla Ltd with 67% confidence.

And Confidence is calculated as below:

Confidence(x->y) = [Support(x->y)] / [Support of x].

Example: Association Rules from BSE SENSEX, Here Researcher has selected sector wise scripts for the calculation of association between the same sector scripts:

Pharmaceuticals Sector:

From BSE SENSEX researcher has selected

Cipla Ltd, Dr.Reddy's Lab, SunPharma India Ltd, Glenmark Ltd, Orchid Chemicals Ltd to calculate association between these same sector scripts.

Here minimum support is the average of all the calculated support.

And the MINIMUM SUPPORT: sum of support / total number of scripts

=56/10

=5.6 %

So, minimum support is 5.6%

MINIMUM CONFIDENCE: sum of confidence / total number of scripts
= 350 / 10
=35%

So minimum confidence is 35%.

Researcher applied the above rule to calculate min.Support and min.Confidence to obtain result for other sector scripts.

Script1	Script2	Support	Confidence	Is in rule
Cipla	Dr.Reddy's Lab	10%	67%	TRUE
Cipla	Sun Pharma India ltd	9%	60%	TRUE
Cipla	Gelnmark Lab	1%	7%	FALSE
Cipla	Orchid Chem	4%	27%	FALSE
Dr.Reddy's Lab	Sun Pharma India ltd	12%	67%	TRUE
Dr.Reddy's Lab	Glenmark Lab	1%	6%	FALSE
Dr.Reddy's Lab	Orchid Chem	5%	28%	FALSE
Sun Pharma India ltd	Glenmark Lab	4%	18%	FALSE
Sun Pharma India ltd	Orchid Chem	7%	32%	FALSE
Glenmark Lab	Orchid Chem	3%	38%	FALSE

So from the above data analysis researcher can conclude that Cipla ltd and Dr.Reddy's Ltd go hand in hand and also Dr.reddy's lab And Sun Pharma India Ltd goes hand in hand. So researcher can say these scripts are strongly associated with each other.

2. APRIORI ALGORITHM:

Apriori is a classical algorithm and is designed to operate on databases containing transactions. The theory of Apriori algorithm is that "All nonempty subsets of a frequent item set must also be frequent."

Apriori principle can be shown as below:

For all(x, y) :(x belongs to y) => s(x)>=s(y)

i.e. support of an item set never exceeds the support of its subsets. This property is also known as monotone property of support. Algorithm is used to mine the frequent item sets.

Apriori Algorithm is as follows:

- Let K=1.
- Generate frequent item sets of length 1
- Repeat until no frequent item sets are identified.

Example:

Support count (Dr.Reddy's pharma lab Ltd) = No of transactions containing Dr. Reddy's Pharma ltd = 18.

3. RULE INDUCTION TECHNIQUE

Rule induction technique retrieves all interesting patterns from database.

In rule induction technique, the rule if of "if this then this". For example a rule that a stock market might find in their data collected from market transaction report would be: "if Reliance Industries Ltd script is purchased then Oil and Natural Gas Corporation is purchased".

or If Tata steel then SAIL

If Mahindra then Hindustan motors

In order for the rules to be useful there are two pieces on information that must be supplied as well as the actual rule:

Accuracy- How often is the rule correct?

Coverage- How often does the rule apply?

SCRIPT1	SCRIPT2	ACCURACY	COVERAGE	IS IN RULE
Cipla	Dr.Reddy's Lab	66%	15%	FALSE
Cipla	Sun Pharma India ltd	60%	15%	FALSE
Cipla	Gelnmark Lab	6%	15%	FALSE
Cipla	Orchid Chem	26%	15%	FALSE
Dr.Reddy's Lab	Sun Pharma India ltd	66%	18%	TRUE
Dr.Reddy's Lab	Glenmark Lab	5%	18%	FALSE
Dr.Reddy's Lab	Orchid Chem	27%	18%	FALSE
Sun Pharma India ltd	Glenmark Lab	18%	22%	FALSE
Sun Pharma India ltd	Orchid Chem	31%	22%	FALSE
Glenmark Lab	Orchid Chem	37%	8%	FALSE

From the above observation we conclude that Dr.Reddy's Lab and Sun Pharma India Ltd are associated with each other as these satisfies both the minimum accuracy= 34 and minimum coverage=16.

So the rule is true. So, when the customer purchases Dr.Reddy's lab customer will also go for Sun Pharma India Ltd with 66% Accuracy. So these are strongly associated with each other.

4. CORRELATION:

To find out the impact of fluctuating gold prices and BSE sensex and the impact of dollar prices and BSE sensex from 2008 to 2013 researcher has used a statistical formula coefficient of correlation.

The mathematical formula for computing r is:

$$r = n \sum xy - (\sum x)(\sum y) / \sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}$$

Where x and y are the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y.

If x and y are results of measurements that contain measurement error, the realistic limits on the correlation coefficient are not −1 to +1 but a smaller range.

The value of r is such that -1 < r < +1. The + and - signs are used for positive linear correlation and negative linear correlations, respectively.

1. **Positive correlation:** If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase.

2. **Negative correlation:** If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

3. **No correlation:** If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

4. Note that r is a dimensionless quantity; that is; it does not depend on the units employed.

5. A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.

6. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 are generally described as weak. These values can vary based upon the “type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Impact of gold prices on stock market:

According to Indian scenario, Indian culture and tradition majority of the Indian women would like to invest in gold because of their tradition and their liking for gold. This leads to invest in gold because of its nature of keeping value, low risk, and as India is having parallel economy there are no any rules or fix criteria for investing in gold. So Indian people feel more beneficial to invest in gold therefore, Gold is having more impact on Equity Market.

EXAMPLE: Table shows the correlations between GOLD and BSE SENSEX from January 2008 to august 2013

SR.NO	YEAR	CORRELATION	RESULT
1	2008	-0.78	Negatively correlated
2	2009	0.48	Weakly correlated
3	2010	0.84	Strongly correlated
4	2011	-0.81	Negatively correlated
5	2012	0.76	Positive correlated
6	2013	0.99	Strongly correlated

Table shows the correlation between US DOLLAR and BSE SENSEX from January 2008 to august 2013:

Currency market launched in 1999. 4000cr is daily turnover of the exchange because of which currency market became strong. And the more popular and effective currency is US Dollar. As compare to Equity, Dollar fluctuates slowly and more effective to the investors which leads to investors to prefer investing in currency market. Dollar is not only important for investment purpose but every countries financial strategy for planning to balance their currency with dollar for good economic results which leads to become dollar stronger. Therefore dollar is having more impact on equity market from last few years.

SR.NO	YEAR	CORELATION	RESULT
1	2008	-0.95	Negatively correlated
2	2009	-0.86	Negatively correlated
3	2010	-0.50	Negatively correlated
4	2011	-0.93	Negatively correlated
5	2012	0.27	Weakly correlated
6	2013	0.98	Strongly correlated

5. CONCLUSION

An Association between selected scripts of Indian Stock Market and a correlation between Indian Gold Prices and BSE SENSEX INDEX and Dollar Prices and BSE SENSEX INDEX has been described. In this paper Data Mining Tools such as Association Rule, Apriori Algorithm, and Rule Induction Methods are used for Association of Indian Stock market in order to find out which scripts are much associated with each other. Results shows that sector wise scripts are much associated with each other which helps investors, brokers, jobbers for investment decision. Statistical Correlation result shows that, Gold Prices and Dollar Prices has an impact on Indian Stock Market.

REFERENCES

[1] Alex Berson and Stephen j. Smith, "Data Warehousing, Data Mining, and OLAP", MC Graw Hill, 1997.
[2] A. D. Devale and Dr. R. V. Kulkarni, "Application Of Data Mining Techniques In Life Insurance", International Journal Of Data Mining and Knowledge Management Process Vol.2. No.4, July 2012.
[3] Arun. K. Pujari, "Data Mining Techniques", Universities Press (India) PVT Ltd, 2001.
[4] C.R. Kothari, "Research Methodology: Methods and Techniques", New Age International (p) Ltd, 2004.
[5] Chengqi Zhang, Shichao Zhang, "Association Rule Mining: Models and Algorithm, Springer, 2002.
[6] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, "Efficient Mining of Association Rules in Distributed Databases", IEEE, 1996.
[7] J. Date, "An Introduction to Database Systems", Addition Wesley longman, Seven Edition, 2000.
[8] J.K. Sharma, "Business Statistics" Pearson Education, 2008.
[9] Ken Orr, "Data Warehousing Technology", Copyright. The Ken Or Institute, 1997.
[10] Krzysztof J. Cios, Witold Pedryez and Roman W. Surniarski, "Data Mining Methods for Knowledge Discovery", Kluwer Academic Publishers 1998 Second Printing 2000.
[11] L. M. Bhole, "Financial Institutions and Markets: Structure, Growth and Innovation, MC Graw Hill, 2006.
[12] Ming-Syan chen, Jiawei Han and Philip S. Yu, "Data Mining: An Overview From a Database Perspective", IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.
[13] NSE's Certification In Financial Markets, National Stock Exchange of India ltd