ORIGINAL RESEARCH PAPER

Volume-7 | Issue-11 | November-2018 | PRINT ISSN No 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

SIGNIFICANCE OF DATA PRE-PROCESSING TECHNIQUES IN STOCK MARKET

Computer Science

Prof. Kainaz Bomi Sherdiwala Asst. Prof. in Z.F.Wadia Women's College & N.K. Jhota College of Commerce, Surat, Dept. of Computer Application

Dr. Samrat O. Khanna* Vidyanagar, Anand, Gujarat *Corresponding Author

ABSTRACT

In this paper, we have focused on the importance and necessity of pre-processing of raw data. Preprocessing helps to improve the quality of all types of data which turns out to be useful for Apriori algorithm in discovering Association Rules. Stock market data of past 7.3 years for the period between 01/01/2011 and 31/03/2018 is collected from NSE website. Apart from the stock prices, we have also collected data of NIFTY 50 index for the same period. Database was built for the same. Attributes like Company's symbol, Date, Open, High, Low and Close were considered for our research problem. This raw data will be pre-processed. This pre-processed will be used in Association Rule Mining (ARM) to generate high quality rules which will help us in discovering useful pattern from our dataset. The main aim of our research is to find co-relation between index value and price of the scripts. Therefore we have employed ARM in our study. Various preprocessing techniques such as data cleaning, data integration, binning was employed to make the quality of data better which can yield accurate results.

KEYWORDS

Data Cleaning, Integration, Binning Encoding

INTRODUCTION

Data pre-processing is a Data Mining technique in which raw data is converted into an appropriate format [1]. Data in real world is "dirty" which means it is incomplete, noisy and inconsistent [2]. The noisy data is very common feature of real world datasets especially in the field of stock market where the data is non-linear in nature. So majority of time is spent on data pre-processing. According to the surveys, it is found that the total effort spent on data pre-processing tasks in data mining application is as high as 60% to 80% [3]. The raw data which we have downloaded from NSE website have to be pre-processed. This pre-processed data will be used in Association Rule Mining to generate high quality rules which will help us in discovering useful pattern from our dataset. Various pre-processing techniques such as data cleaning, data integration, binning was employed to make the quality of data better which can yield accurate results.

1. DATA CLEANING

We have collected historic data from NSE website which consists of attributes like Symbol of the Company, Series, Date, Previous Close, Open Price, High Price, Low Price, Last Price, Close Price, Average Price, Total Traded Quantity and Turnover (in Lacs). As our research work requires only the closing values, we have taken into consideration only the Symbol of the Company, Date and Close Price for our study. Therefore, the remaining unnecessary attributes were eliminated. The table 4.1 shows data after eliminating unwanted attributes. There were few companies which had consecutive missing records. We have ignored those companies which comprised of many missing records that may lead to wrong results if considered. Missing values were substituted with the average of the adjacent 10 records of that column.

Table 4.1 Shows Data after Eliminating Unnecessary Attributes

Symbol	Date	Close Price
3IINFOTECH	03-Jan-11	62.2
3IINFOTECH	04-Jan-11	61.95
3IINFOTECH	05-Jan-11	62.05
3IINFOTECH	06-Jan-11	61.6
3IINFOTECH	07-Jan-11	58.75

Above table 4.1 shows the records of the script 3IINFOTECH after removing unwanted attributes.

2. DATA INTEGRATION

After data cleaning, the next step is to integrate all the files. There were 500 CSV files for each year having records of individual company from January 2011 to March 2018. We have also integrated all the files of NIFTY 50 index. We have integrated one thousand twenty five files as shown in table 4.2. All these files were merged into a single file for better management.

Table 4.2 Shows Data after Merging Files

Symbol	Date	Close Price
3IINFOTECH	28-Dec-15	4.9
3IINFOTECH	29-Dec-15	4.8
3IINFOTECH	30-Dec-15	4.85
3IINFOTECH	31-Dec-15	4.85
20MICRONS	04-Jan-11	47.55
20MICRONS	05-Jan-11	46.8
20MICRONS	06-Jan-11	46.4

According to the table 4.2, we have three columns namely Symbol which represents name of the script, Date which specifies the date of the transaction and Close Price which is the closing price of that symbol on that particular date. Above table shows records after merging two companies – 3IINFOTECH and 20MICRONS. In this way all the files of the different companies are merged into a file. After merging all the files into a single file, we have also integrated index files of all the years. The final merged file is shown in table 4.3.

Table 4.3 Shows data after merging closing price of shares and closing price of index files

Symbol	Date	Close Price	Close Index
20MICRONS	29-Dec-15	4.8	7928.95
20MICRONS	30-Dec-15	4.85	7896.25
20MICRONS	31-Dec-15	4.85	7946.35
3IINFOTECH	03-Jan-11	46.65	6157.6
3IINFOTECH	04-Jan-11	46.05	6146.35
3IINFOTECH	05-Jan-11	47.75	6079.8

According to the table 4.3, now we have four columns namely Symbol, Date, Close Price of the scripts and Close Index which represents the closing value of the index on that particular date. Above table shows records after merging two companies and the closing index value. In this way all the files of the different companies are merged into a single file along with the closing index values of all the 7.3 years.

3. DERIVED ATTRIBUTE VALUE

We simply cannot compare the closing price of the current day with the previous day. So we have to compute closing price movement. Closing price movement is calculated as a Percentage change on the close price of the previous day record with respect to the current day record for each day of all the companies. We have named this derived attribute as per_change_stock.

Table 4.4 Shows Derived Attribute per_change_stock and per_change_index for each Company

Symbol	Date	Close Price	per_change_stock	Close Index	per_change_index
20MICRONS	16-Dec-15	4.9	-5.49	7750.9	0.65
20MICRONS	17-Dec-15	5	2.04	7844.35	1.21
20MICRONS	18-Dec-15	5.05	1.00	7761.95	-1.05
20MICRONS	21-Dec-15	5.1	0.99	7834.45	0.93
20MICRONS	22-Dec-15	5.05	-0.98	7786.1	-0.62
20MICRONS	23-Dec-15	5.05	0.00	7865.95	1.03
20MICRONS	24-Dec-15	4.9	-2.97	7861.05	-0.06
20MICRONS	28-Dec-15	4.9	0.00	7925.15	0.82
20MICRONS	29-Dec-15	4.8	-2.04	7928.95	0.05
20MICRONS	30-Dec-15	4.85	1.04	7896.25	-0.41
20MICRONS	31-Dec-15	4.85	0.00	7946.35	0.63
3IINFOTECH	03-Jan-11	46.65	0.00	6157.6	-22.51
3IINFOTECH	04-Jan-11	46.05	-1.29	6146.35	-0.18
3IINFOTECH	05-Jan-11	47.75	3.69	6079.8	-1.08

As shown in Table 4.4, we have calculated percentage change (per_change_stock) for the scripts: 3IINFOTECH and 20MICRONS and also calculated percentage change (per_change_index) for the closing index values. We have calculated these attributes for all the scripts.

4. Data Binning (Discretization)

Data Binning can be used to obtain reduced representation of database without loss of original data. We have used equal-width binning technique to construct bin values for percentage change in closing price of stock and closing price of index value. For data reduction, we have replaced the percentage change value with its corresponding bin value. Using equal-width binning technique, we have created bin values by considering a gap of 0.50. If the percentage change is between 0 and 0.50, then its bin value will be 0.50 and if the percentage change is between 0.50 and 1, then its bin value will be 1 and so on. Bin values are represented in table 4.5. We have calculated bin values for negative values also. They are represented in table 4.6. In this way, the data is reduced without loss of important information. Bin values have been calculated for the percentage change in index value also. In this way we have carried out data binning with the help of discretization technique.

Table 4.5 Shows Calculated Bin Value for Percentage Change (Positive)

PERCENTAGE CHANGE	BIN VALUE
0 TO 0.50	0.5
0.50 TO 1	1
1 TO 1.50	1.5
1.50 TO 2	2
-----	---

Table 4.5 shows bin values calculated on the basis of percentage change in value of the script that is per_change_stock for positive values.

Table 4.6 Shows Calculated Bin Value for Percentage Change (Negative)

PERCENTAGE CHANGE	BIN VALUE
0 TO -0.50	-0.5
-0.50 TO -1	-1
-1 TO -1.50	-1.5
-1.50 TO -2	-2
-----	---

Table 4.6 shows bin values calculated on the basis of percentage change in value of the script that is per_change_stock for negative values. We have considered a gap of 0.50. It is not feasible to consider gap lesser than 0.50 as we have huge amount of data in our database. In the same way considering a gap more than 0.50 would also not yield accurate results because there would be large variation in the generated association rules making it difficult to analyse the results. So we have decided to maintain gap of 0.5% as it would generate good quality of rules. Also a variety of rules with different bin values would be useful in making decisions regarding a particular script with regards to the index value. It provides guidance for the investors that if there is some percentage change in the index value what will be its effect on the price of a particular script.

Table 4.6 Shows Calculated Bin Value for Percentage Change (Negative)

PERCENTAGE CHANGE	BIN VALUE
0 TO -0.50	-0.5
-0.50 TO -1	-1
-1 TO -1.50	-1.5
-1.50 TO -2	-2
-----	---

Table 4.6 shows bin values calculated on the basis of percentage change in value of the script that is per_change_stock for negative values. We have considered a gap of 0.50. It is not feasible to consider gap lesser than 0.50 as we have huge amount of data in our database. In the same way considering a gap more than 0.50 would also not yield accurate results because there would be large variation in the generated association rules making it difficult to analyse the results. So we have decided to maintain gap of 0.5% as it would generate good quality of rules. Also a variety of rules with different bin values would be useful in making decisions regarding a particular script with regards to the index value. It provides guidance for the investors that if there is some percentage change in the index value what will be its effect on the price of a particular script.

Table 4.7 Shows Calculated Bin Values considering a gap of 0.5%

Symbol	per_change_stock	bin_stock	per_change_index	bin_index
20MICRONS	0	0	0	0
20MICRONS	-0.40193	-0.5	-0.1827	-0.5
20MICRONS	0.161421	0.5	-1.08276	-1.5
20MICRONS	-0.72522	-1	-0.51893	-1
20MICRONS	-4.62662	-5	-2.37507	-2.5

Table 4.7 shows calculated bin values for both per_change_stock and per_change_index taking into consideration a gap of 0.5%.

5. DATA ENCODING

We have to convert the data in the format which is acceptable to our proposed algorithm. With the help of binning, we have calculated bin values and then it is combined with the script symbol. For faster processing, we have encoded our data by generating unique codes. Code_stock is the combination of the name of the script and its corresponding bin value. In the same way code_index is the combination of the letters "In" and its corresponding bin value.

For the purpose of our research work, we require only two attributes code_stock and code_index. So we will eliminate the other attributes. Now, we do not require these attributes because on the basis of the symbol, bin_stock and bin_index, we have encoded our data. Our final table will look like this (table 4.8).

Table 4.8 Shows final representation of data

code_stock	code_index
20MICRONS_0.0	In_0
20MICRONS_-0.5	In_-0.5
20MICRONS_0.5	In_-1.5
20MICRONS_-1.0	In_-1
20MICRONS_-5.0	In_-2.5

Table 4.8 shows the representation of the final data which will serve as an input to our proposed algorithm. After applying pre-processing techniques to our dataset, we have derived the above final data. The next step is to select appropriate algorithm for generating rules to obtain associations between the index value and the price of the scripts which are not the part of NIFTY 50 index.

CONCLUSION

Data pre-processing is very important and useful for data mining. The key issue is to maintain the most significant data while throwing may unimportant data. The over processing of the data may lead to unpredictable results. One should always start with minimal pre-processing and incrementally add more pre-processing while evaluating the result.

REFERENCES

1. Thakkar R, Patel V. & Kayasth M. (2014, November). Model to Predict Stock Price With respect To Day of The Week. International Journal of Advanced Research in Computer Science and Software Engineering, 4(11).
2. Malik J., Goyal P. & Sharma A. (2009, August). A Comprehensive Approach towards Data Preprocessing Techniques & Association Rules.
3. Jermyn P., Dixon M. & Read B. Preparing Clean Views of Data for Data Mining.
4. Jasdeep Singh Malik, Prachi Goyal and Mr.Akhilesh K Sharma, "A Comprehensive Approach towards Data Preprocessing Techniques and Association Rules".
5. Rahul Thakkar and Rustom Morena entitled "Preprocessing of Intraday Stock Market Data", ISSN - 2229-4333, IJCST Vol. 5, Issue 2, April - June 2014.
6. Shichao Zhang, Chengqi Zhang and Qiang Yang, "Data Preparation For Data Mining", Applied Artificial Intelligence, 2003.
7. Jasdeep Singh Malik, Prachi Goyal, Mr.Akhilesh K Sharma, "A Comprehensive Approach towards Data Preprocessing Techniques & Association Rules".
8. Maria N. Moreno, Saddys Segrera, Vivian F. Lopez and M. Jose Polo entitled "Improving the Quality of Association Rule by Preprocessing Numerical Data".
9. Jasdeep Singh Malik, Prachi Goyal and Mr Akhilesh K Sharma, "A Comprehensive Approach towards Data Preprocessing Techniques & Association Rules".