International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Improving Association Rule Mining By Defining A Novel Data Structure

Vinayak Suresh Shukla, Prof.Dr.Mrs.S.A.Itkar

[1]*Student,PES's Modern College Of Engineering,Pune 5.*
[2]*HOD,Computer Engineering Dept, PES's Modern College Of Engineering, Pune 5.*

Abstract - *In recent years, growth in digital data storage in rapidly increased due to ease of use and lower coast digital storage media. This data is high dimensional and heterogeneous in nature. The process of knowledge discovery is being affected due to high dimensional and heterogeneous data. This process can be abbreviated as association rule mining (ARM). Though, many association rule mining algorithms have been proposed in recent years to deal with large volume of data, the mining process under-performs when the data size is very large in terms of records. Hence the aim of this work is not to design a new algorithm for mining, but to design a new data structure to store data reliably .The original data is simplified, recognized and access time increased for that data, to meet up efficiency in terms of time and main memory requirements. Lower main memory requirements and faster data access are achieved by means of Shuffling, Inverted Index Mapping and Run Length Encoding. Hence the resulting data structure can be used along with the existing association rule mining algorithms to speed up mining and reducing main memory requirements, without changing original algorithms. This is further improved by replacing Run Length Encoding by Modified Run Length Encoding Algorithm for better memory utilization and efficiency of mining algorithms.*

Key Words: Association Rule Mining (ARM), Data Compression, Data Structure, Index Compression, Knowledge Discovery, Modified RLE,.

1.INTRODUCTION

In early years, digitization of all technical and non-technical fields has leaded to the production of large amount of digital data every day[1]. Storage of this large data is efficient since the cost of storage on media is less than early storage medias . Hence the cost of storage has negligible association with amount and heterogeneity of this data. But it has tremendously affected the mining process. Because of large amount of data, mining has become interesting but time consuming. Association rule mining[2] is well known and researched field for mining rules form given data. Many algorithms were introduced in order to speed up data analysis. Different strategies were used like reduce either number of candidate sets, number of transactions or number of comparisons of both. But any mining process with extremely high dimensional data [1]can be a hard process. So instead of changing existing algorithms or finding new one to speed up mining, it is better to introduce a new data structure [3]to store large data in compressed manner. So here we implemented a new compressed data structure which will store the original data in compressed manner without losing original content. This new compressed data structure is obtained by applying three sequential techniques on the original data. viz
Shuffling
Inverted index mapping
Run length encoding.

Then it is further improved by Modified Run Length Encoding
This new data structure will help in handling, speeding up access to data and fast computing when used with any of the existing mining algorithm.

2.REVIEW OF LITERATURE

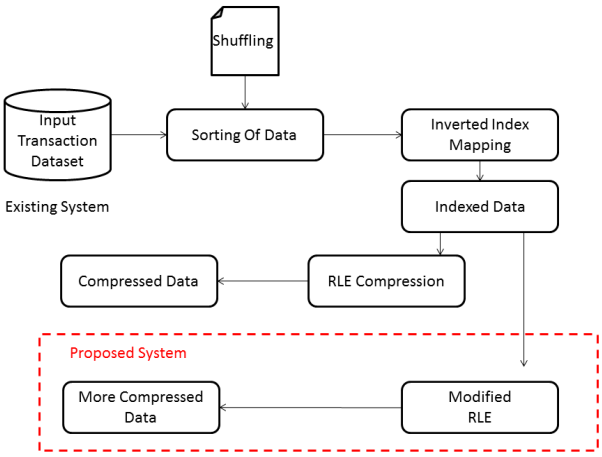
Association rule mining [2]is the most important technique used for mining the rules. This aims to extract strong and relative items within input data set. Basically it was designed for market basket analysis which would help shopkeepers arranging the sale items in order to grow the business.
Apriori[4] is the first algorithm for association rule mining. It is based on comprehensive search. It works in two steps as I. finding all frequent item sets from data set and II. Deriving association rules from those frequents item sets. If data set has k single items, and N number of transactions, then [2^k-1\] item sets can be generated. And of the data set size is huge, the processing is hard, because the complexity of above operation will be O(N X M X k) just to obtain candidate sets. The frequent pattern FP-growth[5] algorithm was introduced to reduce the number of transactions and related comparisons. Data needs to be scanned only once since FP-growth stored frequent items in a tree structure. But it suffers with large number if I/O and large number of memory required storing all sets.

To overcome the existing drawback, like large memory requirements and long computational times, many evolutionary association rule mining algorithms were proposed [6],[7],[8],[9]. But, even after considering mining rules from different views, working with high-dimensional data was hard.

So as a solution, it is found that creating a new data structure in order to reduce the original data set size would be advantageous, and that can be used for speeding up the mining, with current algorithms without changing their original functioning.

The data structure introduced in the paper is meant to reduce the data set size and produce faster data access. Despite this guarantees the data validity and doesn't change the original data values. The conversion of data into new data structure comprises of three steps viz. 1. Shuffling, 2. Inverted Index Mapping and finally and 3. Data Compression. The Modified Run Length Encoding technique helps in[10] more compression of input data than the original Run Length Encoding algorithm, that eventually helps in improving performance of ARMs.

3.SYSTEM ARCHITECTURE / SYSTEM OVERVIEW



As the ultimate goal of this work is to device a new data structure which will reduce the data size and speed up the operating.

So as shown in figure the new data structure is obtained by performing three functions on original data. Viz shuffling, inverted index mapping and run length encoding. Each of these will work as follows.

1.Shuffling: Similarity between records is the characteristic of specific data set i.e. two or more records share similar feature values. This characteristic is used to sort records among data using Hamming Distance (HD). HD can be used as similarity metric in order to minimize number of changes in the feature values from record to record. Sorted data is compressed by merging features in subsequent records. Lower the HD between two records implies higher probability of clustering those into same group. Original Data Set is input for this function which is modified to sorted data after completing this function.

2.Inverted Index Mapping: This takes the sorted data as an input and aims to construct the efficient structure i.e. index based structure. Attribute values are indexed based on satisfied transaction and then indexed attribute values can be joined with consecutive transactions to share same values. This inverted index methodology assigns key-value pair to each list of attribute, and each attribute has pointer to transaction index satisfied by that attribute value.

3.RLE encoding: This step is responsible for data compression and faster data access. This takes inverted index structure as an input and groups consecutive indices for each attribute into a 2-tuple that defines a starting index and the total number of consecutive transactions containing the same feature value {index, displacement}.

4.Modified RLE: The modified run length encoding scheme gives a major improvement in compression ratio for any kind of data. Analysing the input data is the first and core step. Data is analysed to highlight if there are any largest numbers of sequences that may increase the number of bits to represent the length of each run. In this method if input data contains nearest value with its adjacent data then both value are considered as same data.

The resulting data structure is now compressed without losing values.

4.SYSTEM ANALYSIS

1.Dataset Information Here, we have used different datasets like Adult and Mushroom each having variant number of Instances and attribute.

2.Software and Hardware Requirements The above system is implemented using RAM of 1GB and above, Speed of 1.1 GHz. The processor used is intel - CORE i3 and above. A Hard Disk of 20 is used. Operating System Windows 7 with Java 1.8 is used. Development Environment considered here is Netbeans 8.0.2.

3.Performance Parameters

1.Data Size: After analyzing universal compressors like Zip we state that the proposed structure represents the data into a compressed structure according to the resulting file size.

2.Loading Time: as the data size is now compressed, loading time for compressed data will be less than that of loading the original data.

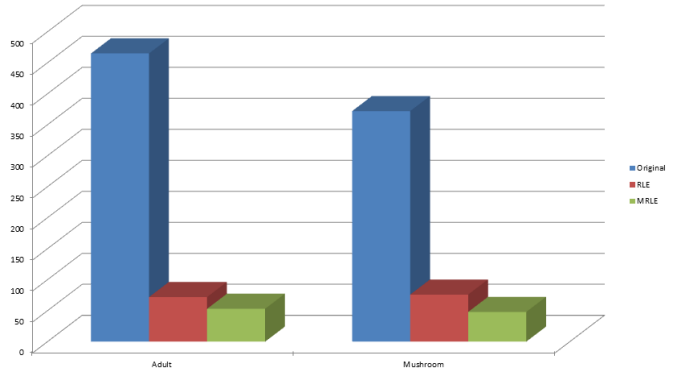
3.Memory Requirements: Analyzing the Adult data set of 465 KB, it only requires 53 KB of main memory by using the new data structure, hence processing the compressed data will ultimately require less memory space.

Table -1: Data Compression

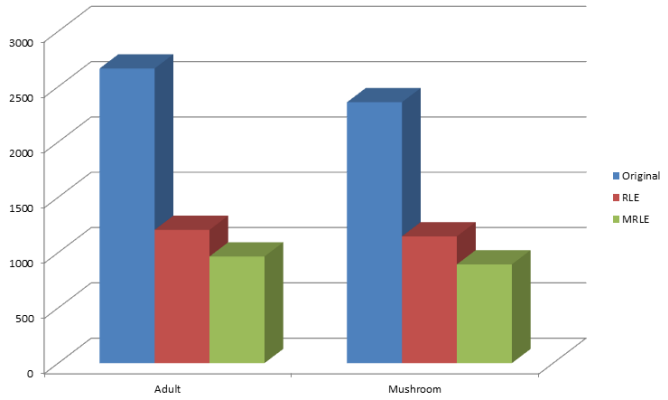
Data Structure	Adult	Mushroom
Original	465	372
RLE	72	68
MRLE	53	46

Table-2: Time Required to process Apriori Algorithm

Dataset	Original	RLE	MRLE
Adult	2663	1204	964
Mushroom	2358	1146	893



Graph 1



Graph 2

4.Results:

The data sets Adult and mushroom is compressed using shuffling, mapping and RLE and then Modified RLE compression techniques. Then the result are obtained after applying Apriori algorithm on compressed datasets. The observations and results are then plotted in graph in order to compare size and time required to store and process respectively for both original and compressed data structures.

From the graph and table it can be observed that, the compressed data set reduce the size required to store and time required to process.

5. CONCLUSIONS

Association rule mining is being much interesting due to rapid growth of volume of data which. But the mining process is being time consuming for such data using traditional mining algorithms. Hence data compression technique can improve the performance of existing algorithms without changing their original schema. The data compression technique used above compresses the data to be used by mining algorithms and hence improves the performance in terms of data size, loading time/processing time and main memory requirement efficiently. The mining process can be further improved in above terms using MP-Graph for better utilization of memory space and efficiency when large data sets are used.

ACKNOWLEDGEMENT

Every orientation work has a mark of many people and it is the responsibility of author to convey an intense gratitude for the same. I feel colossal pleasure to express deep sense of gratitude and indebtedness to my guide Prof. (Ms) Dr. S. A. Itkar, for constant uplifting and virtuous guidance. I also express my sincere gratitude to the Department of Computer Engineering and Library of my college. Last but not the least; I am grateful to my friends and my parents for their best wishes..

REFERENCES

[1]V. Marx, "Biology: The big challenges of big data," Nature, vol. 498, no. 7453, pp. 255–260, 2013.

[2]R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487–499.

[3]José María Luna, Alberto Cano, Mykola Pechenizkiy, " Speeding-Up Association Rule Mining With Inverted Index Compression", IEEE TRANSACTIONS ON CYBERNETICS, 2016.

[4]R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in Advances in Knowledge Discovery and Data Mining. Menlo Park, CA, USA: Amer. Assoc. Artif. Intell., 1996, pp. 307–328.

[5]J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Min. Knowl. Disc., vol. 8, no. 1, pp. 53–87, 2004.

[6]B. Goethals and M. J. Zaki, "Advances in frequent itemset mining implementations: Report on FIMI'03," ACM SIGKDD Explor. Newslett., vol. 6, no. 1, pp. 109–117, 2004.

[7]C. Lucchese, S. Orlando, and R. Perego, "DCI closed: A fast and memory efficient algorithm to mine frequent closed itemsets," in Proc. IEEE ICDM Workshop Frequent Itemset Min. Implement. (FIMI), Brighton, U.K., 2004, pp. 20–28.

[8]J. M. Luna, J. R. Romero, C. Romero, and S. Ventura, "On the use of genetic programming for mining comprehensible rules in subgroup discovery," IEEE Trans. Cybern., vol. 44, no. 12, pp. 2329–2341, Dec. 2014.

[9]C.Borgelt,"Efficient Impleme -ntations of Apriori and Eclat," in Proc. 1st IEEE ICDM Workshop Frequent Item Set Min. Implement. (FIMI), Melbourne, FL, USA, 2003, pp. 90–99.

[10] S. Joseph, N. Srikanth, J. E. N. Abhilash, "A Novel Approach of Modified Run Length Encoding Scheme for High Speed Data Communication Application ," International Journal of Science and Research Volume, 2 Issue 12, December 2013