ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 6, Issue 6, June 2018

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

An Efficient Multilevel Association Rule Mining Based On Unsupervised Learning

Prof. Alisha Patel[1]
CE/IT Dept.
CGPIT,UTU
Bardoli, Surat – India

Prof. Krishna Patel[2]
CE/IT Dept.
CGPIT,UTU
Bardoli, Surat – India

Abstract: ***Multilevel association rule mining is a well-developed field in data mining. The main goal is to find hidden pattern and information in or between levels of abstraction. Data mining is an analytic process for large data. It is mainly focuses on the CRM which is customer Relationship Management. Apriori algorithm is mostly used in the field of multilevel association rule mining. Generation of large number of candidate itemsets and multiple scanning of databases is main shortage of the Apriori algorithm. In this paper partial k-means clustering algorithm proposed .SOFM used random weight generation process which takes long training time. Hence, a different approach for clustering using maximum number of count is proposed for multi-level association rule mining. For the generation of frequent itemsets FP-Growth algorithm is introduced.***

Keywords: Multilevel association rule mining, Data mining, K-means, Clustering, Classification.

I. INTRODUCTION

Data mining is a subfield of computer science which is used to find out the hidden pattern or knowledge from the large amount of data. Data mining is a process of extracting information the dataset and transforms it into the useful knowledge. It is a subarea of artificial intelligence called knowledge discovery and machine learning. It is a well-researched technique of data mining. Association rule is the process of finding the relationship among the items of database. Agrawal first introduce association rule in 1993 to identify relationship between the items of database [8].It is used for decision making purpose and improvement of business [9]. The purchasing of one product when another product is purchased represented by an association rule. Database contains data which forms the hierarchical structure. Itemset may occur from any level in the hierarchy. For finding more specific and relevant knowledge multilevel association rule mining can helps.

II. RELATED WORK

Unsupervised learning identifies the pattern class information heuristically and reinforcement learning learns through trial and error interactions with its environment. . It used to identify hidden patterns in unlabelled input data. It has an ability to learn and organize information without providing an error signal to evaluate the potential solution [10]. Based on division of clustering multilevel association rule mining is introduced with internal threshold. Because of the internal threshold there is no need to set minimum support. By applying this approach efficient mining is done not only for single layer but also cross layer. Database scanning time and candidate itemset generation is reduced using this technique. By merging the SOFM and apriori algorithm for frequent itemset generation time is reduced and it is effective and feasible [1]. At lower level of the abstraction SC-BF algorithm is used for finding maximum frequent itemset which reduce the execution time. It will increase throughput. Concept of the reduce support is used with the progressive deepening method. Here data were combined at branch level so storing memory is decrease and reduce the execution time. Smaller set are generated with high quality of rules from dataset [2].

New approach for mining the frequent and infrequent association rules is introduced which helps user to generate rare item sets. For the development of multilevel propositional logic based knowledge discovery involve steps like data pre-processing, Propositional Logic for Coherent Rule Generation, Multi-Level concept hierarchy and Performance measure on generated rules. The proposed new mining algorithm can generate large item sets level by level and then derive concept multilevel association rules from transaction dataset. This algorithm can derive the multiple-level association rules under different propositional logic in a simple and effective way and reduced the risk associated with it [3]. MRA is stand for Multilevel Relationship Algorithm, containing three stages of working. In first two stages it performs MRA and after that in the third stage it performs Bayesian probability for the dependency & relationship among different shops. It generates different rules for learning. Pattern of sale can generate using MRA algorithm. MRA performs better and takes less time than Apriori algorithm [4]. By calculating initial centroids numbers of iterations are reduced and elapsed time is improved for the k-means clustering algorithm. In new improved k-means algorithm numbers of iterations are fixed. K-mean algorithm is efficient from basic K-mean algorithm in terms of iterations and elapsed time. Cluster quality and time complexity is also improved [5]. To reduce the execution time ranking method is used with k-means clustering algorithm. Author proposed ranking method for overall design of database for student data in order to form the clusters. Because of the ranking method execution time is less than the K-means algorithm. Because of the ranking method K-means algorithm generates better results [6]. Based on a SOFM and composite distance measure a new model of spatial clustering is introduced. According to their inter-distances spatial objects are divided into sub sets in SOFM. Land price samples are taking as dataset for the knowledge discovery process for spatial data using SOFM algorithm. SOFM clustering can detect spatial outlier which is based on composite distance statistic. The clustering method using composite distance is more flexible, and can implement many kinds of spatial clustering for different aims [7].

A. K-means Algorithm [11] [12]

K-means is a well-known partitioning method which clusters data by separating samples in n groups of equal variance. Number of cluster should be to be specified for this algorithm. Every Cluster is formed by calculating the centroid of each group. It assigns each object to the group with the closest centroid. This algorithm execute in three steps. First choose the random k as number of the cluster. Then assign object to that centroid. In the second step calculate distance of all data points to centroid. In the last step data items are move in appropriate cluster. This process will continue until change occurs. In k-means clustering algorithm initial clusters k are based on randomly selected centroids.

B. Self-Organizing Feature Map (SOFM)[1]

It is a type of ANN (Artificial Neural Network) using unsupervised learning to produce a low-dimensional (typically two-dimensional), discredited representation of the input space of the training samples, which is known as a map. Self-Organizing Maps are using a neighborhood function to preserve the topological properties of the input space and different to other artificial neural network. These components are called nodes or neurons. Weight is associated with each node of the same dimension as the input data vectors and a position in the map space. There are usually two layers. Input layer's all neurons are connected with output layer and it is one dimensional. Output layer is one or multidimensional. The basic arrangement of nodes is a two-dimensional regular spacing in a rectangular or hexagonal grid. Network training convergence speed is faster .It is a two layer feed forward neural network. The output layer is also called competitive layer.

III. PROPOSED WORK

Multilevel association rule mining with counting strategy is proposed for large dataset. Algorithm is work in two steps:

1. Find maximum occurrence of item belonging to a particular category for one transaction.
2. Generate different clusters k for finding frequent item set
3. Apply FP-Growth algorithm for each cluster to find frequent item set

From literature survey we found that the success of conventional feature map formation, trained by basic SOFM algorithm, is critically relying on how the main parameter of the algorithm like weight initialization, learning rate and the neighborhood function are selected. Process of trial and error normally determines them. From all literature we determine it usually requires a huge amount of iterations and selections of the different parameters. For SOFM initial weights would have been more appropriate. Distinct different approaches have been proposed to replace the conventional SOFM algorithm.

Conventional SOFM algorithm has to update the weight vectors side-by-side for winning neuron as well as corresponding neighbor. K-means algorithm requires less computational resources than conventional SOFM. Traditional K-means algorithm works with Euclidean distance. It usually finds distance between two objects which is takes more time for creating Euclidean space. Partial K-means algorithm is introduced for cluster formation. In [1] they used cumulate algorithm with SOFM. Basically cumulate algorithm is extended from the Apriori algorithm. A new different approach for generating clusters is defined with FP growth for multi level association rule mining.

A. PROPOSED METHOD AND ALGORITHM

SOFM is an unsupervised learning and clustering algorithm of high-dimensional input data domains onto the elements of a two dimensional array. Its goal is to find out a set of centroids and generally initialized using random values for the weights. For selection of particular weight the process have to goes under the whole scanning of the database. It takes more time for selection of particular weight. A different maximum counting approach is proposed for clustering counting of number of occurrence.

B. FLOW OF WORK

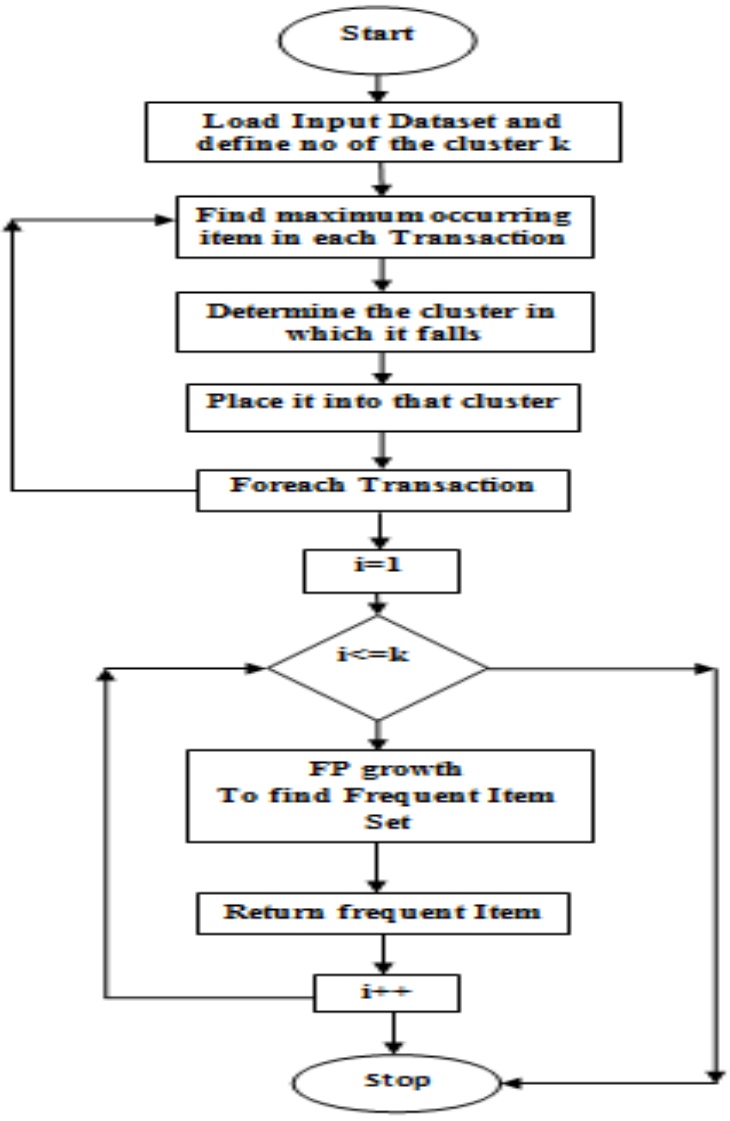

Fig.1 Flow of proposed work

IV. RESULT ANALYSIS

A. Computational Environment

Processor: Intel Core i3-2310M

RAM: 4 GB

Development Tool: Eclipse (Indigo)

Dataset:

Retail dataset is used for the frequent itemset generation. The retail dataset was provided by Tom Brijs and contains the retail market basket data from an anonymous Belgian retail store.

B. Analysis of experimental results

Eclipse Indigo is used for the generation. Total ten predefine clusters are there. In that five are main which are 1205,1305,1405,1505 and 1605. 1205 determines bakery product. While 1305 determines body products, 1405 dairy products, 1505 food products and 1605 stationary product. Remaining five clusters are 1705,1805,2005,2305 and 2505. When no of count are same and maximum for dairy and stationary products that transaction will fall in to 2505 cluster. Likewise 2305 for body and stationary products, 2005 for bakery and stationary products, 1805 for bakery and dairy products and 1705 for bakery and body products. From the dataset first transaction is taken and then on it clustering process is applied. For first transaction maximum count is calculated for particular product and then put that transaction into that particular cluster. Likewise for all transactions this process is repeats. From these process finally ten clusters we get. By applying fp growth algorithm frequent itemsets will generate.

On different sizes of dataset by applying SOFM with cumulate and new proposed algorithm comparison is given in fig.1 and fig. 2.Fig. 2 shows the proposed work in terms of relationship time and no of transactions. Results are measures on 1000,5000,10000 and 15000 of transactions. Proposed work will take less time than the existing one. Time gap is increase between them.

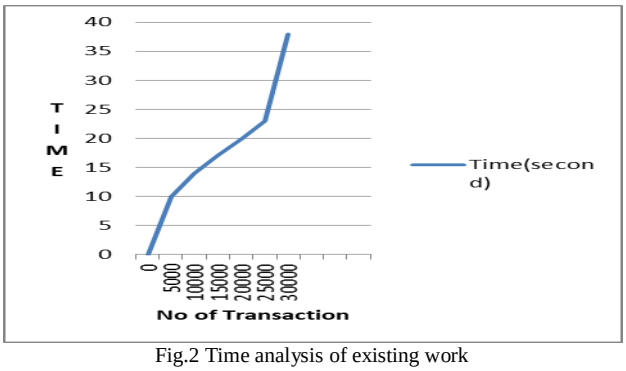

Fig.2 Time analysis of existing work

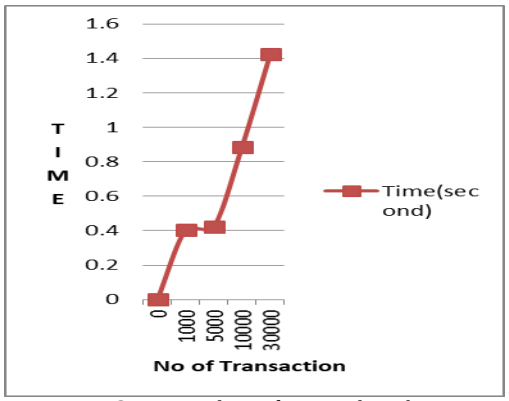

Fig3. Time analysis of proposed Work

Table 1. Time Analysis of Proposed Work

No of transactions	Time(ms)
1000	0.404
5000	0.422
10000	0.886
30000	1.422

V. CONCLUSION

Traditional multilevel association rule mining based on the apriori algorithm is not accepted for the large database because of the multiple generations of candidate keys and scanning time. Random weight generation is main weakness of the traditional SOFM algorithm. In that we have to perform trial and error method for the weight initialization. In this we proposed partial K-means algorithm which does not contain Euclidean distance. So using this approach time parameter will be reduced. Using FP-Growth algorithm we get the frequent item-sets.

References

1. Huang QingLan, Duan LongZhen, "Multi-level association rule mining based on clustering partition", IEEE(2013).
2. Pratima Gautam, Dr. K. R. Pardasani , "Algorithm for Efficient Multilevel Association Rule Mining", IJCSE(2010).
3. S.Prakash, M.Vijayakumar, " A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy", IJCA(2011)
4. Deepak Vidhate, Dr. Parag Kulkarni, "To improve Association Rule Mining using New Technique: Multilevel Relationship Algorithm towards Cooperative Learning" ,IEEE(2014).
5. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro, Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research(2012).
6.] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur, "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining", IJARCET (2012).
7. Limin Jiao, Yaolin Liu, "Knowledge Discovery By Spatial Clustering Based On Self-Organizing Feature Map And A Composite Distance Measure", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2008).
8. Jiawei Han, Yongjian Fu, "Mining Multi-level Association Rules in Large
9. Nikunj H. Domadiya, Udai Pratap Rao , "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database", IEEE(2012).
10. R.Sathya,Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification",IJARAI(2013).
11. Osama Abu Abbas, "Comparisons between Data Clustering Algorithm", IAJIT(2008).
12. Amanpreet Kaur Toor, Amarpreet Singh , "Analysis of Clustering Algorithm Based on Number of Clusters, Error Rate, Computation Time and Map Topology on Large Data Set", IJETTCS(2013).

