APPLYING ASSOCIATION RULE MINING FOR INDIRECT DEPENDENCIES ELIMINATION

[1] Dr.J.S Kanchana, [2] S.Anusuya Devi, [3] J.Jenila,[4] S.Kiruthika,[5] A.Kanaga Durga

[1]Associate Professor, [2]UG Student, [3] UG Student, [4]UG Student, [5]UG Student

[1]Department of Information Technologyy,

K.L.N College of Engineering, Pottapalayam, Sivagangai, Tamil Nadu, India.

Abstract : Data Mining is defined as a progress used to extract usable data from a larger set of any raw data. It examine a data patterns in large batches of data using two or more software. It is an application of data mining, businesses can learn about their customers and develop more effective strategies related to various business functions. This paper defines the association rules to describe the indirect dependences. First, an algorithm is proposed to identify a product, sets of items, and characteristics that are given to the transactional database. There some indirect dependences, which refer to the relationship between discontinuous activities in business activities. To recognize anomaly detection for business process is the hindmost goal.

IndexTerms **- Clustering, Association rule, Indirect dependency.**

I. INTRODUCTION

Data mining is the process of exploration and analysis of large data to discover meaningful patterns and rules. It can be considered a discipline under the data science field of study and differs from predictive analytics because it describes historical data, while data mining aims to predict future outcomes. Web mining is the application of data mining techniques to discover patterns from the World Wide Web. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information.

Clustering is a important in data analysis and data mining applications. It is the task of grouping a set of objects , that objects in the same group are more similar to each other than to those in other groups . Partitioning a data is the centroid based clustering; the value of k-mean is set. similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. The advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. It also helps in classifying documents on the web for information discovery.

K-means clustering is one of the popular unsupervised machine learning algorithms. The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible k-means clustering is a method of vector quantization. It is a prototype-based clustering technique because the prototype in terms of a centroid which is calculated to be the mean of a group of points and is applicable to objects in a continuous n-dimensional space.

Association rules are computed from item sets, which are made up of two or more items. Association rules are typically created from rules well-represented in data. Association analysis is about discovering relationship among large amount of data sets. This rule was created by searching data for frequent if-then patterns and using the criteria support and confidence to recognize the most important relationships. Support determines how often a rule is applicable to the data set while confidence determines how frequently items in B appear in transactions that contain A. Third metric, called lift, can be used to compare confidence with expected confidence. In previously, many algorithms were developed by research team for Boolean and Fuzzy association rule mining such as Apriori, FP-tree.

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. It is one of the classical algorithms in data mining. Apriori algorithm (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996) is a data mining method which outputs all frequent item sets and association rules from given data.

II. RELATED WORK:

Mining of frequent item sets is an important phase in association mining which discovers frequent item sets in transactions database. It is the core in many tasks of data mining that try to find interesting patterns from datasets, such as association rules, episodes, classifier, clustering and correlation, etc. Many algorithms are proposed to find frequent item sets, but all of them can be catalogued into two classes: candidate generation or pattern growth. Apriori is a representative the candidate generation approach. It generates length (k+1) candidate item sets based on length (k) frequent item sets. The frequency of item sets is defined by counting their occurrence in transactions. FP-growth, is proposed by Han in 2000, represents pattern growth approach, it used specific data structure (FP-tree), FP-growth discover the frequent item sets by finding all frequent in 1-itemsets into condition pattern base , the condition pattern base is constructed efficiently based on the link of node structure that association with FP-tree. FP-growth does not generate candidate item sets explicitly.

Yan-hua WANG Xia FENG 'The Optimization of Apriori Algorithm Based on Directed Network', (2009 Third International Symposium on Intelligent Information Technology Application) this paper gives an experiment to analyze and compare the difference between the two (Apriori algorithm and proposes an improved algorithm based on the directed network) algorithms and the result shows that the improved algorithm promotes the efficiency of computing. In this paper, algorithm improved that based on directed network.

Yiwu Xie, Yutong Li, Chunli Wang, Mingyu Lu" The Optimization and Improvement of the Apriori Algorithm", Through the study of Apriori algorithm we discover two aspects that affect the efficiency of the algorithm. One is the frequent scanning database, the other is large scale of the candidate item sets. Therefore, Apriori algorithm is proposed that can reduce the times of scanning database, optimize the join procedure of frequent item sets generated in order to reduce the size of the candidate item sets. In this paper It not only decrease the times of scanning database but also optimize the process that generates candidate item sets.

III. PROPOSED WORK:

To reduce the time while scanning some transactions and improve accuracy, Improved Apriori algorithm is taken over. To be more efficient and less time consuming, Improved Apriori algorithm is used. To avoid scanning the database repeatedly. To improve joining efficiency, pruning is done.

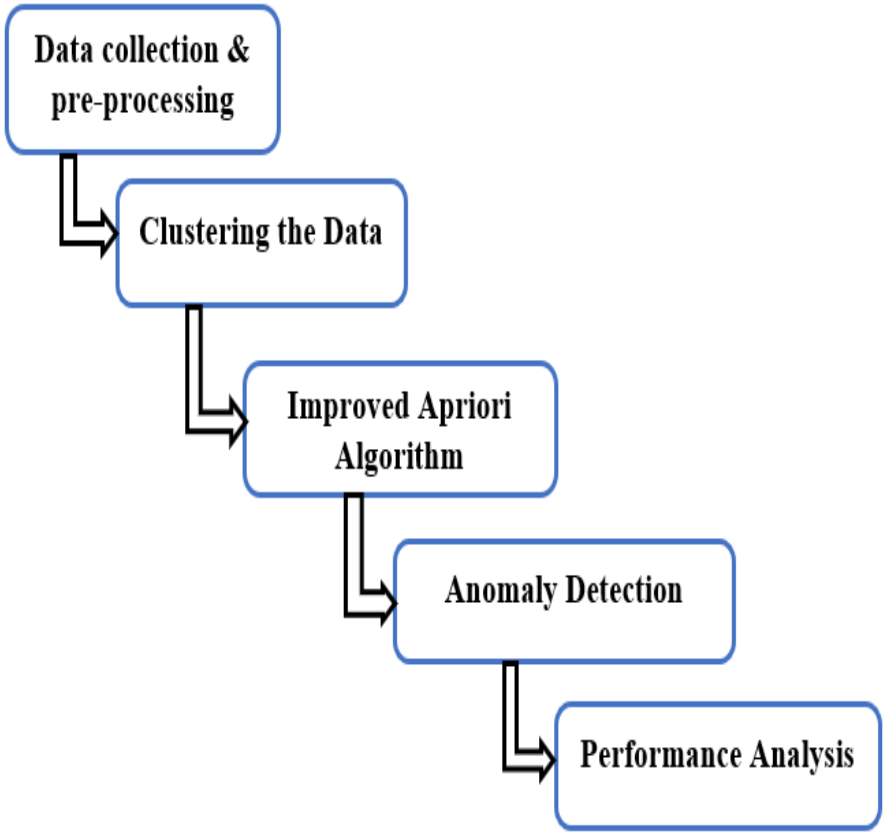


Fig -1: Improved Apriori Framework

3.1 Data Collection

Data collection is the systematic approach to gathering and measuring information from a variety of sources. A data set is a collection of discrete items of related data that may be accessed individually managed as a whole entity. To remove noisy and unwanted data.

	A	B	C	D	E	F
1	Bread	Milk				
2	Bread	Diapers	Beer	Eggs		
3	Milk	Diapers	Beer	Cola		
4	Bread	Milk	Diapers	Beer		
5	Bread	Milk	Diapers	Cola		
6	Bread	Milk				
7	Bread	Cola	Beer	Milk		
8	Milk	Bread	Beer	Cola		
9	Bread	Milk	Diapers	Beer		
10	Bread	Beer	Diapers	Diapers		
11						
12						
13						

Fig -2: Grossaries Data set

3.2 K-Means Clustering

Cluster is a group of objects that reside to the same class. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster. Clustering can therefore be developed as a multi-objective optimization problem. K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. K-means algorithm is an iterative algorithm that tries to split the dataset into non-overlapping subgroups. K-means clusterings used in search engines, market segmentation, statistics and even astronomy. Cluster centre is represented by mean value of the object in the cluster. Importance of Clustering is dividing the population or data points into a number of batches such that data points in the same batches are more similar to other data points in the same batches than those in other batches.

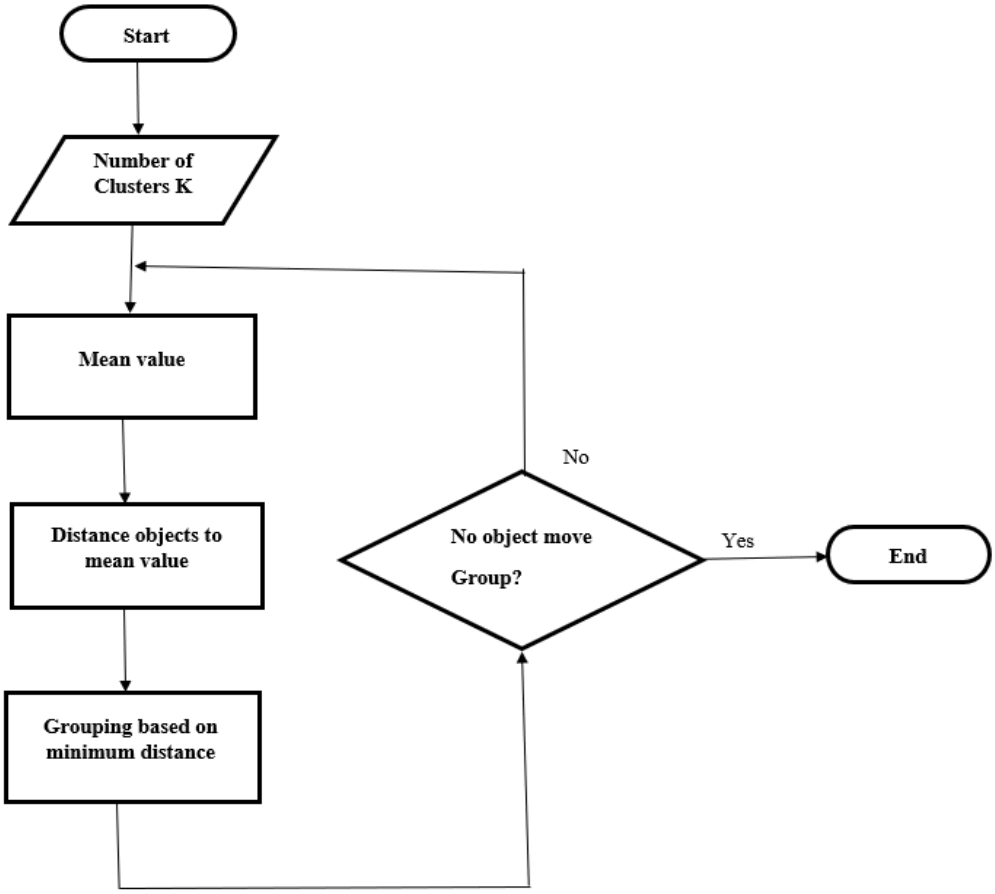


Fig -3: K-means clustering

3.3 Improved Apriori Algorithm

Association rule mining algorith is a rule-based machine learning method for evaluate a interesting relations between variables in large databases. Improved Apriori algorithm proceeds by identifying the frequent individual items in the database .It are devised to operate on a database containing a lot of transactions, for instance. It is more efficient and less time consuming. Minimum-Support is a parameter supplied to the Apriori algorithm in order to prune candidate rules by specifying a minimum lower bound for the Support measure of resulting association rules. There is a corresponding Minimum-Confidence pruning parameter as well. Applications of association rule mining are stock analysis, web log mining, medical diagnosis, customer market analysis bioinformatics etc.

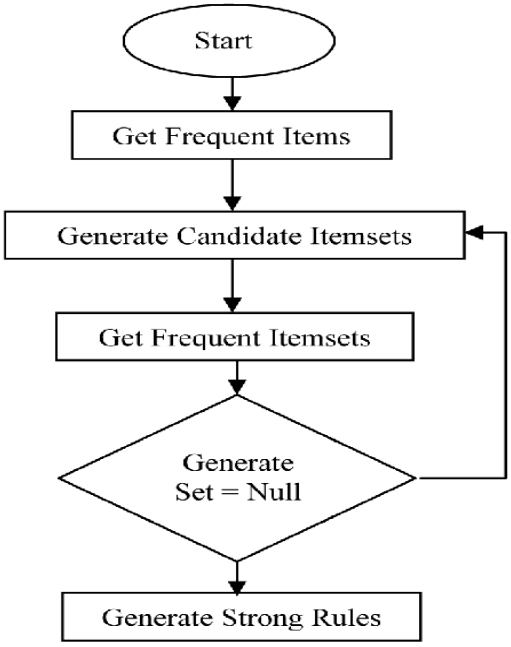


Fig -4: Apriori algorithm

Pseudo code for Improved Apriori algorithm

Input: D, σ
Output: F(D, σ)
1: C1 := {{i} | i∈ I}
2: k := 1
3: while Ck 6= {} do
4: // Compute the supports of all candidate itemsets
5: for all transactions (tid, I) ∈ D do
6: if transaction time = Frequent set expected time then
7: for all candidate itemsets X ∈Ck do
8: if X ⊆ I then
9: X.support++
10: end if
11: end for
12: end if
13: end for
14: // Extract all frequent itemsets
15: Fk := {X | X.support ≥ σ}
16: // Generate new candidate itemsets
17: for all X, Y ∈Fk, X[i] = Y [i] for 1 ≤ i ≤ k − 1, and X[k] < Y [k] do
18: I = X ∪ {Y [k]}
19: if ∀J ⊂ I, |J| = k : J ∈Fk then
20: Ck+1 := Ck+1∪ I
21: end if
22: end for
23: k++
24: end while

3.4 Anomaly Detection:

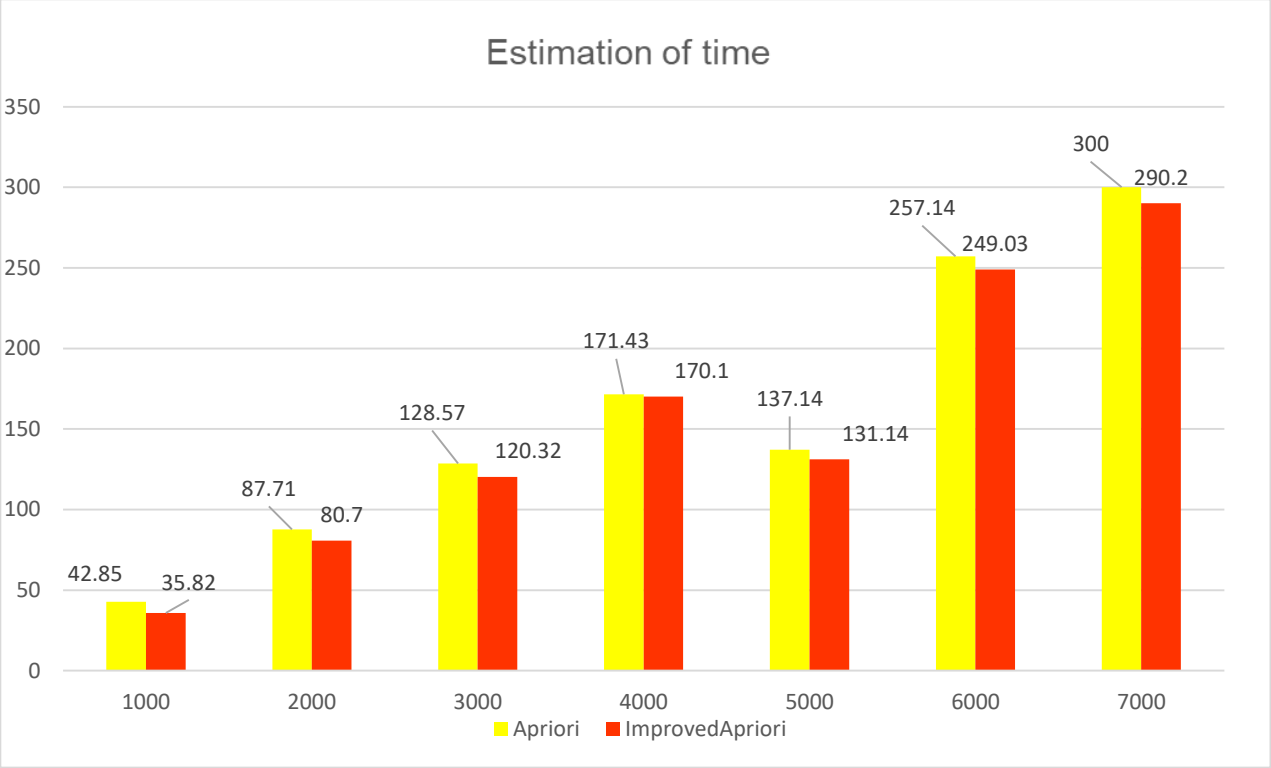
An Outlier is a rare chance of occurrence within a given data set Outliers can be of two kinds:

• Univariate outliers can be found when looking at a distribution of values in a single feature space.
• Multivariate outliers can be found in a n-dimensional space (of n-features).

IV. EXPERIMENTEL RESULTS

Comparision between apriori and improved apriory for reduce the time complexity. In Apriory algorithm First 1000 records to be processed at the time of 42.85 seconds , next the data sets can be increased and that time of process also increased. Then improved apriori algorithm is used for the 1000 sets ,2000 sets upto 7000 data sets to slightly decrease the time of execution .

Transaction	Execution time in secs	
	Apriori	Improved Apriori
1000	42.85	35.82
2000	87.71	80.7
3000	128.57	120.32
4000	171.43	170.1
5000	137.14	131.14
6000	257.14	249.03
7000	300	290.2



V. CONCLUSIONS

In this paper, an algorithm is proposed to construct models with indirect dependencies. We have prepared in this paper an association rule mining for apriori algorithm. And we have used in an K means algorithm for Clustering. And it split a given anonymous data set. Then in order to solve the problem of difficult to ordering the sequence items of the products. Association rule generated using improved Apriori Algorithm and it's reduced the time to collect the sequence of items. The results show that proposed algorithm is effective. Besides, models mined by this algorithm have higher precision than other algorithms mine. This method can be used to construct real process models, which helps people to further understand the system operation, thereby discovering system bottlenecks and optimizing the process. Our future work will focus on indirect dependencies in multiple parallel structures and other complex structures.

VI. ACKNOWLEDGEMENT

We would like to thank all academic staff in our university for supporting us in each research. Project work at the end of term always becomes a great deal and requires great amount of work. but sometimes guidance and co –operation of other people directly or indirectly help to temp this problem.

We are thankful to our professor Dr.J.S Kanchana, who guided and taught data mining. Without her the core understanding of course work could not have been possible. we are also extending our acknowledgement to all who helped as to generate 1000 tuples of database. Last but not least, we are thankful to everybody who directly or indirectly helped us in this project.

VI. REFERENCES

1. Huiming Sun, Yuyue Du, Liang Qi and Zhaoyang He "A Method for Mining Process Models with Indirect Dependencies via Petri Nets" in IEEE Access, pages 81211 - 81226, July.2019.

2. M.Ranjanisindu, Ms.K. Jenifer "An improved apriori algorithm for Frequent item set Mining accuracy" in IJIRSET Vol. 8, Issue 5, 2019.

3. Swee Chuan Tan "Improving Association Rule Mining Using Clustering Based Discritization of Numerical Data", in IEEE International Conference 2018.

4. Yiwu Xie, Yutong Li, Chunli Wang, Mingyu Lu" The Optimization and Improvement of the Apriori Algorithm", in IEEE International Conference 2008.

5. Amritpal Kaur, Vaishali Aggarwal and Shashi Kant Shankar "An efficient algorithm for generating association rules by using constrained item sets mining" in 2016 IEEE International conference.

6. Yasir Ali , Amjad Farooq , Talha Mahboob Alam , Muhammad Shoaib Farooq , Mazhar Javed Awan , Talha Imtiaz Baig "Detection of Schistosomiasis Factors Using Association Rule Mining", in IEEE Access Vol. 7 2019.

7. Gehao Sheng , Huijuan Hou , Xiuchen Jiang , Yufeng Chen " A Novel Association Rule Mining Method of Big Data for Power Transformers State Parameters Based on Probabilistic Graph Model" IEEE Transactions on Smart Grid Vol. 9, 2018.

8. José María Luna , Francisco Padillo , Mykola Pechenizkiy , Sebastián Ventura "Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data", in IEEE Transactions on Cybernetics Vol. 48,2018.

9. Mahsa Salehi ; Christopher Leckie ; James C. Bezdek ; Tharshan Vaithianathan ; Xuyun Zhang " Fast Memory Efficient Local Outlier Detection in Data Streams", in IEEE Transactions on Knowledge and Data Engineering Vol..28, 2016

10. M. U. Munir, M. Y. Javed, S. A. Khan, " A hierarchical k -means clustering based fingerprint quality classification ", Neurocomputing, vol. 85, pp. 62-67, May 2012.

11. J. M. Luna, A. Cano, M. Pechenizkiy, S. Ventura, "Speeding-up association rule mining with inverted index compression", IEEE Trans. Cybern., vol. 46, no. 12, pp. 3059-3072, Dec. 2016