Customer Buying Behavior Analysis: A Clustered Closed Frequent Itemsets for Transactional Databases

Kaviha.N[1] and Karthikeyan.S[2]

[1] Research Scholar, Department of Computer Science, Karpagam University, Coimbatore,Tamilnadu,India

[2] Assistant Professor, Department of Information Technology, College of Applied Sciences, Oman

Abstract

The problem in data mining applications is the mining of frequent patterns. Though there has been various techniques, such as pattern discovery, association rule mining etc, these methods generates a large volume of frequent patterns and rules which are not useful for finding the essential patterns among them, from the database. Alternatively, CFIM technique produces a relatively lesser number of closed frequent patterns. Patterns are pruned before clustering and the clustered patterns help in predicting the customer purchasing behavior which in turn helps to maintain an inventory and focus the point of sale on transaction data, enhancing sales. This can be achieved by CF-CLUS algorithm which clusters the similar patterns from the generated closed frequent patterns. The distinguishing feature of CF-CLUS is the ability to reduce the number of scans of the database and works well when the minimum support is low and produce good results if the database is sparse. This led to efficient clustering of the results from the generated closed patterns.

Keywords: *Association Rule Mining, Pattern Discovery, Cluster Patterns and Market Basket Data, CFIM-Closed frequent itemset mining***.**

1. Introduction

Discovery of interesting relationship from the dataset is a challenging task in data mining. To perform this task, frequent item set mining was increasingly used. But the drawbacks of this method is, which generates voluminous frequent itemsets.It is very difficult to extract the interesting set from it. The above said disadvantage is overcome by CFIM [1] which produces closed frequent itemsets which are less when compared other frequent algorithms. After generating closed frequent itemsets, it is pruned to remove the unnecessary patterns/Itemsets in order to improve the clustering quality. Once patterns are pruned, clustering algorithm can be applied to cluster the similar patterns / Itemsets and which can be grouped as season wise. Clustering plays an important role as its helps in cross marketing, such as, increase the sales in season wise by updating the inventory, discount offers & store layout on the basis of the customer buying behavior in different seasons**.**

Consider the following Transaction Database,

Transaction	Items in Transactions
10	a, c,d,e,f
20	a,b,e
30	c,e,f
40	a,c,d,f
50	c,e,f

Table1: Transaction Database

The above transaction database consists of 5 transactions, each of which varying number of items.

Frequent Item set generations for the above transactions with minimum support is 2 are

{a},{b}{c},{d},{e},{f},{a,c},{a,d},{a,e},{a,f},{a,b}{c,d}, {c,e},{e,f},{d,f}{a,c,d},{a,c,f},{ce,f},{c,d,f}.{a,c,d,f}

But CFIM [1], generates only 6 closed item sets which are less as follows

{a,c,d,f},{c,e,f},{a,c},{c,f},{a},{e}

After generating closed patterns, the data base was transformed into the following,

Transaction ID	Closed Patterns
1	{a,c,d,f}
2	{c,e,f}
3	{a,c}
4	{c,f}
5	{a}
6	{e}

Table2: Transformed Transaction Database

From the above table2 similar patterns can be clustered as follows

C1={{a,c,d,f},{c,e,f},{c,f}}

C2={a,c,d,f},{a,c}

Like clustering the above transactional datasets, which is used to know the buying patterns of the customers in different seasons are identified.

Another Technique available for generating patterns is pattern discovery. It is widely used technique. The greatest disadvantage of using this technique is generates too many patterns and consumes so much time to cluster the patterns. The proposed cluster algorithm, CF-CLUS can be used to cluster the similar patterns as season wise effectively in order to categorize the buying behavior of the customer.

2. Related Works:

Many algorithms have been developed for finding frequent item sets. Such algorithms are AIS [2] Apriori [3], DHP [4], FP-Growth [5], Brin [6].The above said algorithms generates too many frequent item sets. In order to overcome the problem arises in frequent item sets, the following authors concentrated their research in generating of closed frequent item sets. Such as CHARM [7], CLOSE [8]. SUMMARY [9] developed by Jianyong Wang and Geoge Karypis for clustering the transactional Datasets. Transaction with same description can be grouped together to form a cluster and also which significantly reduce the search space. SCALE [10], CLOPE [11], ROCK [12] are some of the other algorithms which are used to cluster the transactional datasets. Another method proposed by Wong and Li [13] developed for the pattern generation and they uses the technique called Pattern Discovery. Pattern Discovery generated too many patterns which are difficult to find interesting rules from it. The authors suggested a method which helps in simultaneously clustering the discovered patterns and their associated data. Its usefulness lies in its ability to relate the patterns to the set of compound events. The method employs to cluster the patterns is hierarchical agglomerative approach. The Discover * e algorithm is used to generate patterns. The process employing this technique is highly time consuming. To reduce the time required to cluster patterns, it is suggested by the researchers that to cluster closed frequent item sets which is far less than the number of all Frequents Item sets ie's closed frequent item sets. The algorithm CFIM [1] generates the lesser number of frequent closed item sets in order to find the useful rules from it. This paper discusses the logic behind of using the CF-CLUS clustering algorithm to cluster the closed frequent item sets to improve the cluster quality in order to identify the buying patterns of the customers as season wise. The CF-CLUS clustering algorithm works very well when compared to other algorithms such as SUMMARY and Discover * e.

3. Proposed Methodology

The proposed methodologies which comprises the following,

- ✓ Generation of closed patterns
- ✓ Pruning of patterns
- ✓ Clustering similar patterns

The Transaction dataset contain millions of transactions and thousands of items, while each transaction usually contains more than tens of items. FIM algorithms generate too many patterns. But in turn CFIM [1] algorithm generates only the closed frequent patterns which are useful. The proposed clustering algorithm work very well with respect to size of the database.

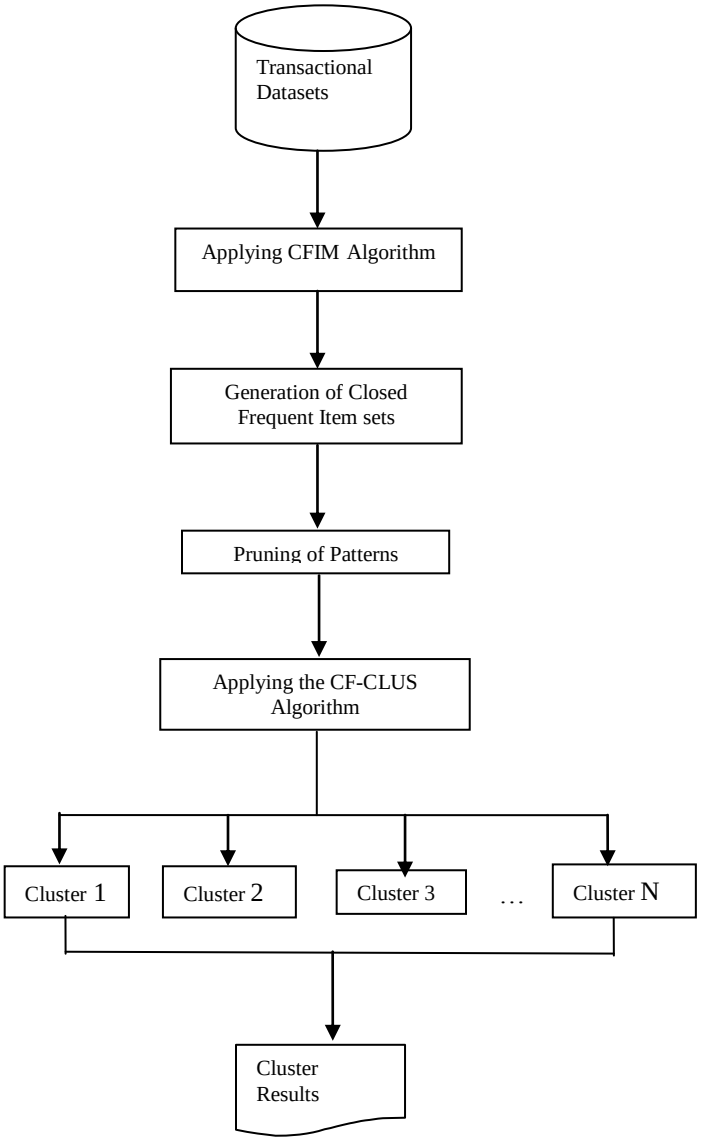

FIG 1: Clustering Closed Item sets

Clustering is an important data mining technique that groups similar data items. The technique of clustering the transactional data sets is very much useful in order to identify the buying behavior of the customers. The above

flowchart represents the proposed work of the system. The proposed clustering algorithm need to first mine a large set of closed frequent item sets in order to identify the most promising ones that can be used for clustering. Once the closed patterns are generated, Pruning can be done. Here pruning is the unnecessary patterns can be removed. Finally CF-CLUS algorithm can be applied to cluster the closed patterns in order to identify the buying behavior of the customers by season wise.

4. CF-CLUS ALGORITHM:

The proposed algorithm is as follows,

Input: a Transactional database, Minimum Threshold Value

Output: C1, C2, C3…Ck

Begin

Partition the database as P1, P2, P3, and P4 by Season wise

For all t_i £ TDB

Generate closed frequent patterns using CFIM [1] Algorithm.

Cl-Closed Itemset

//Scan the database for finding similar itemsets.

While not end of the database do

{

Read the next transaction t_i;

- Calculate the each cluster's minimum support.

- Move the similar k-itemsets to Cluster 1, Cluster 2...Cluster K according to the length of the Itemsets as follows .The Maximum Length of the Itemset is 4

Cluster[i]= Itemset2;

Cluster[j]=Itemset3;

Cluster[k]=Itemset4;

Repeat the process till the end of the database.

}

5. EXPERIMENTAL RESULTS:

Mushroom Dataset:

We had taken the Mushroom Dataset from the UCI Machine learning data mining repository for clustering. It consists of 8,124 records with two classes; 4,208 edible Mushrooms and 3,916 Poisonous Mushrooms. By treating the value of each attributes as items of transactions, we converted all the 22 categorical attribute to Transactions. From the transformed informations, we applied CFIM [1] algorithm which generates lesser number of closed patterns. Preprocessing can be carried out in order to clustering. Unwanted patterns can removed from the generated patterns. CF-CLUS can be applied for clustering which generated the following output.

Table 3: clustering of the mushroom database.

Minimum support	No of cluster	quality	Time
10	654	99.6	0.79
20	428	99.4	0.71
30	376	98.9	0.65
40	287	99.6	0.45
50	230	99.9	0.33
60	212	99.5	0.25

From the table of data , we can see that CF-CLUS algorithm has the highest clustering quality about 99.6% accuracy and it takes the minimum run time about 0.79 seconds for the different minimum values of support.

Transactional data set (Real data set)

We had taken the pazhamuthir super market data set which consists of 1, 00,000 transactions. The transactions can be partitioned into season wise such as summer, autumn, winter, spring .The original database attributes are trans _id, item_name, no_of_qty, price, date.The transaction database can be transformed as trans _id, closed _ itemset. Each partitioned data base can be applied the proposed algorithm CF-CLUS which can group the similar item set into clusters from the generated closed patterns. Consider the Partitioned summer transaction data base which groups the similar item sets into clusters. By means of clustering in this way, one can improve the business by way of organizing the item sets in the store and increase the sale by giving the offers such as discount. CF-CLUS algorithm is implemented using Java Language with weka tool. The no of clusters and time taken for clustering closed item sets for various seasons can be tabulated in table4. The result is performed using Pentium(R) Dual – Core CPU with the speed of 3.00GHZ and 2GB of RAM. For better understanding, Clustering results for various seasons are represented in the chart in fig 2.

Minimum Support	Seasons							
	Summer		autumn		winter		Spring	
	No of clusters	Time Taken (in sec)	No of clusters	Time taken (in sec)	No of clusters	Time taken (in sec)	No of clusters	Time taken (in sec)
10	956	17	816	15	926	16	719	14
20	1058	28	972	26	998	25	836	16
30	1001	30	986	27	1108	31	978	20
40	1185	38	1085	34	1124	36	1120	33
50	1278	45	1178	40	1256	42	1178	35

Table 4: Clustering results.

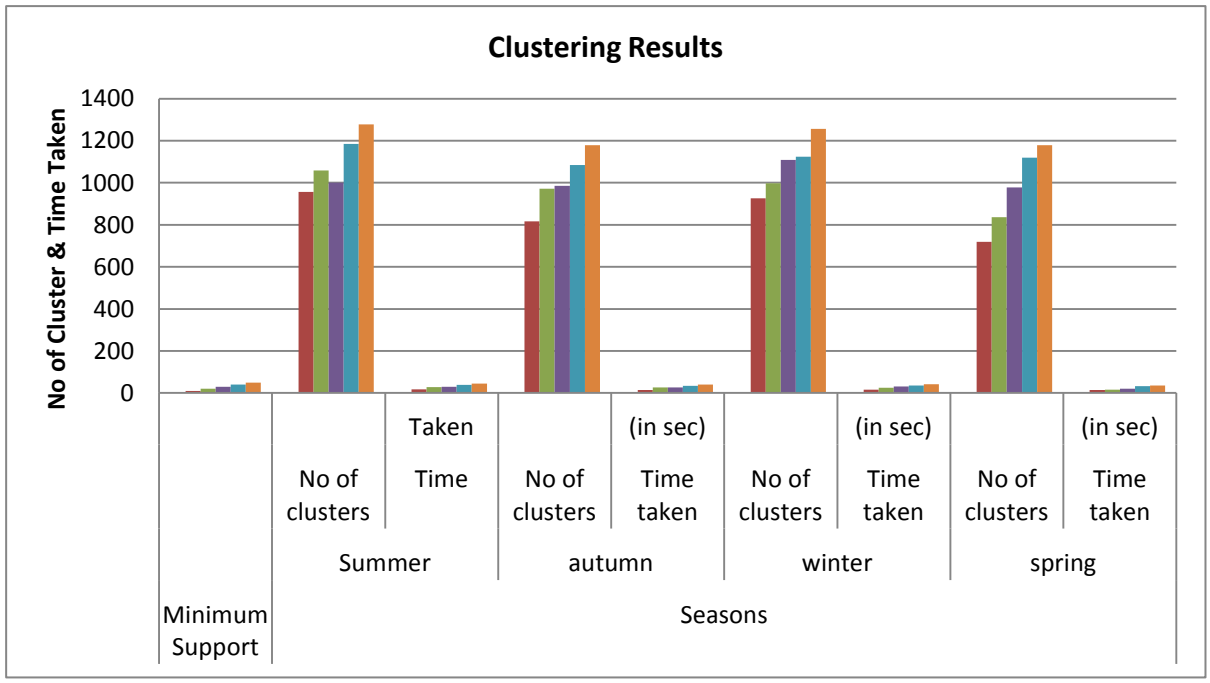

Fig2: Clustering Results for Various Season

6. Conclusion

In this paper, we implemented a novel algorithm for clustering the similar closed item sets as season wise from the transactional data base. In the proposed work, database can be partitioned according to season wise for clustering. Proposed algorithm CF- CLUS generates the useful lesser number of clusters in order to increases the sales in particular season. The unique feature of CF-CLUS is the ability to reduce the number of scans of the database and works well when the minimum support is low and we obtained good results if the database is sparse.

7. References

1] N.Kavitha and S.KarthikeyaN," An Efficient Algorithm for Closed Frequent Item sets and its Associated Data", International Journal of Computer Applications (0975 – 8887) Volume 49– No.14, July 2012.

2] Rakesh Agrawal, Tomaz Lmielinski and Arun Swami," Mining association rules between sets of items in large databases", Proc of ACM SIGMOD Conference on Management of Data, Washington, 1993.

3] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 1995 Int"l Conf. Data Eng. (ICDE "95), pp. 3-14, Mar. 1995.

4] J.S.Park, M.Chen, P.S.Yu," An effective hash based algorithm for mining association rules", Proc. of ACM SIGMOD International Conference on Management of Data, May 1995.

5] Jiawei Han, Ian Pei and Yiwen Yin,"Mining Patterns without Candidate Generation", Proc.of 2000, International Conference on Management of Data, May 16-18, 2000 Dallas, Texas, USA.

6] S. Brin, R. Motwani, and R. Silverstein, "Beyond Market Basket: Generalizing Association Rules to Correlations," Proc. ACM-1997.
7] M.J. Zaki and C.-J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemsets Mining," Proc. Second SIAM Int"l Conf. Data Mining, Apr. 2002.
8] Nicolas Pasquier, Yves Bastide, Rafik Taouil, Lotfi Lakhal," Efficient mining of association rules using closed itemset lattices", Information System, Vol 24, 1999.
9] Jianyong Wang and George Karypis,"SUMMARY:Efficiently Summarizing Transactions for Clustering,22 june 2004.
10] Hua Yan, Keke Chen, Ling Liu, Zhang Yi1, " SCALE: A Scalable Framework for Efficiently Clustering Transactional Data", February 3, 2009.
11] Yiling Yang, Xudong Guan Jinyuan You,"CLOPE: A Fast Effective Clustering Algorithm for Transactional Data, SIGKDD'02 July 23-26, 2002, Edmonton, and Alberia, Canada.
12] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim,"ROCK: A robust Clustering Algorithm for Categorical Attributes",Stanford university,pp 512-521,in Proc.of 15th Int.Conf on Data Engineering, Mar 1999.
13] A.K.C. Wong, Fellow, IEEE, and Gary C.L.Li "Simultaneous pattern and data clustering for pattern cluster analysis" IEEE Trans.Knowledge and Data Eng., vol.20, no. 7, pp. 911-923, JULY 2008.
14] M.J. Zaki, "Mining Non-Redundant Association Rules," Data Mining and Knowledge Discovery, vol. 9, no. 3, pp. 223-248, 2004 .
15] Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques" Elsevier Publishers, 2001 Edition.
16] K.Wong,C.Xu,and B.Liu."Clustering Transaction Using Large Items ",ACM CIKM International Conference on Information aand Knowledge Management,Pages 483-490,Nov1999.

Mrs.N.Kavitha, Assistant professor, completed M.C.A.M.Phil.; she is having 10 yrs of Teaching Experience. Currently, pursuing Doctorate Degree at Karpagam University, Coimbatore. Her Area of Specialization is Data Mining. She has presented various papers in various national Conferences. Presently, she has published various papers in national /international level journals. Her other area of specialization includes networking, software engineering and web technologies. She is an active member in CSI Coimbatore Chapter.

Karthikeyan S. received the Ph.D. Degree in Computer Science and Engineering from Alagappa University, Karaikudi in 2008. He is working as Assistant Professor in the Department of Information Technology, College of Applied College of Applied Sciences, Sohar, Sulatanate of Oman. He has published more than 35 papers in National/International Journals. His research interests include Cryptography and Network Security.