Discovery of Cluster Patterns and Its Associated Data Simultaneously

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

[1] Karpagam University,Coimbatore
nkavee1@gmail.com
[2] Department of Information Technology,
College of Applied Science ,Oman
skaarthi@gmail.com

Abstract. Discovering automatic patterns from the databases are most useful information and great demand in science and engineering fields. The effective pattern mining methods such as pattern discovery and association rule mining have been developed and used in various applications. The existing methods are unable to uncover the useful information from the raw data. Discovering large volume of patterns is easy .But finding the relationship between the patterns and associated data are very difficult and further analyzing the patterns are also complex task. In this Paper, we presented a new algorithm which generates closed frequent patterns and its associated data simultaneously. Here the relationship between the patterns and its associated data are made explicit. The experiment result has been included.

Keywords: Clustering, Association Rule Mining, Cluster Patterns.

1 Introduction

Pattern Discovery is a useful tool for categorical data analysis. The patterns produced are easy to understand. Hence it is widely used in business and commercial applications. PD typically produces an overwhelming number of patterns. The scope of each pattern is very difficult and time consuming to comprehend. There is no systematic and objective way of combining fragments of information from individual patterns to produce a more generalized form of information. Since there are too many patterns, it is difficult to us e them to further explore or analyze the data. To address the problems in Pattern discovery , We proposed a new method that simultaneously clusters the discovered patterns and their associated data. One important property of the proposed method was each cluster pattern was explicitly associated with a corresponding data cluster. To effectively cluster patterns and their associated data, several distance measures are used. Once a distance measure is defined, existing clustering methods can be used to cluster patterns and their associated data. After clusters are found, each of them can be further explored and analyzed individually.

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 224–227, 2012.

2 Related Work

Many algorithms are developed to find the frequent patterns. The first proposed algorithm for association rule mining was AIS algorithm. AIS Algorithm [1] used the Candidate Generation to generate frequent itemsets. The Main drawback of this algorithm is generating too many candidate itemsets. The next algorithm is Apriori Algorithm [2] was developed to find the frequent patterns. It used the Breadth –First Strategy to count the support of itemsets and used the candidate generation function. Wong and Li[3] have proposed a method that simultaneously clusters the discovered patterns and their associated data. In which pattern induced data clusters is introduced. It relates patterns to the set of compound events containing them and makes the relation between patterns and their associated data explicit. Pattern induced data clusters defined are constants. That is each attribute has only one value in the cluster. Since each pattern can induce a constant cluster, the number of constant clusters is overwhelming. To reduce the number, it is desirable to merge clusters. Let us say two clusters I (i), I (j) are two clusters. The merged data cluster of I (i) and I (j) is the union of their matched samp les and matched attributes. When two data cluster are merged, the corresponding patterns including them are simultaneously clustered. The author used the hierarchical agglomerative app roach to clusters the patterns. To Generate pattern, the author used Discover *e Algorithm. The main drawback of the algorithm is speed and Pattern pruning was not done. Rather than clustering all frequent item sets, cluster closed frequent itemsets which could be much fewer than the number of all frequent item sets. Our works does exactly that.

3 The Clustering Algorithm

In this work, modified k- Means algorithm has been used.

3.1 Modified K-Means Clustering

Let D={D(j) | j=1,n} be a data set having *K* clusters, *C* = *{ci|i* = 1, *K}* be a set of *K* centers And *Sj* = *{d* (*j*) |*d* (*j*) is member of cluster k*}* be the set of samples that belong to the *j* th cluster. Conventional *K* Means algorithm minimizes the following function which is defined as an objective function

$$Cost(D,C)=\sum_{j=1}^{n} dist(d^{(j)},c_k)$$

Where dist($d^{(j),}$ c_k) measures the Euclidean distance between a points d $^{(J)}$ and its cluster center c_k. The k-means algorithm calculates cluster centers iteratively as follows:

1. Initialize the centers in c_k using random sampling;
2. Decide membership of the points in one of the K clusters according to the minimum distance from cluster center Criteria;

$$c_k = \frac{\sum_{d^{(j)} \in S_k} d^{(j)}}{|S_k|}$$

3. Calculate new c_k centers as:
   Where $|S_k|$ is the number of data items in the k^{th} cluster.
4. Repeat step 2 and 3, till there is no change in cluster centers.

Instead of using centers found by (2) every time, our proposed algorithm calculates the cluster centers that are quite close to the desire cluster centers. The proposed algorithm, first divides the data set D into K subsets according to some rule as associated with data space patterns , then chooses cluster centers for each subset.

4 The Closed Frequent Itemset:

Itemset X is closed if none of the proper supper sets of X have the same s support. For mining frequent closed itemset, we proposed an efficient algorithm (CFIM) for mining all the closed frequent itemsets. We first described the algorithm in general terms, independent of the implementation details. We then showed how the algorithm can be implemented efficiently. The proposed algorithm simultaneously explores both the itemset space and tidset space using the IT-tree, unlike previous methods which typically exploit only the itemset space. The proposed algorithm(CFIM) used a novel search method, based on the IT - pair properties , that skips many levels in Tree to quickly converge on the itemset closures , rather than having to enumerate many possible subsets . The algorithm starts by initializing the prefix class P of nodes to be examined, to the frequent single items and their tidsets.The main computation is performed which returns the set of closed frequent itemsets *C*FI.

5 Experimental Results

We had taken the iris database from UCI Machine learning database repository for finding frequent closed item sets. It consists of 160 samples, 4 attributes and 3 classes (Setosa, Versicolor and Virginica). The classes Versicolor and Virginica are highly overlapped while the class Setosa is linearly separable from the other two. The algorithms are implemented using java programming language. It generates the 56 closed itemsets if the minimum support is 2 and average execution time is 0.011 secs. It generates the 48 closed itemsets if the minimum support is 3 and average execution time is 0.002 secs . If the existing algorithm, apriori which generates 90 frequent item sets when the minimum support is 2 and average execution time is 0.016 secs. Ariori generates 85 frequent itemsets if the minimum support is 3 and average execution is 0.003 secs .

The comparison of Charm and Apriori is shown in the fig 1.

By seeing the chart, the proposed out performs the result. Apriori generates all the frequent itemsets. But proposed algorithm produced only the closed Frequent Patterns. The execution speed is faster when compared to Apriori.

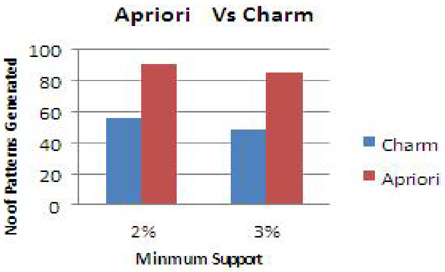

Fig. 1. Comparison of Apriori Vs Charm

6 Conclusion

This paper has proposed a method for clustering patterns and their associated data. The effectiveness of the above divide-and-conquer approach lies in the proposed clustering method. One important property of the proposed method is that each pattern cluster is explicitly associated with a corresponding data cluster. To effectively cluster patterns and their associated data, several distance measures are used.

References

1. Agrawal, Srikant, R.: Algorithms for Mining Frequent Itemsets. In: Proc. of the ACM SIGMOD Conference on Management of Data (1993)
2. Agrawal, Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th Int'l Conf. Very Large Data Bases (VLDB 1994), pp. 487–499 (1994)
3. Wong, A.K.C., Li, G.C.L.: Simultaneous pattern and data clustering for pattern cluster analysis. IEEE Trans. Knowledge and Data Eng. 20(7), 911–923 (2008)
4. Pei, J., Han, J., Mao, R.: CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In: Proc. ACM SIGMOD Int'l Workshop Data Mining and Knowledge Discovery (May 2000)
5. Pei, J., Han, J., Wang, J.: CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. In: Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (August 2003)
6. Zaki, M., Hsiao, C.: CHARM: An efficient algorithm for closed itemset mining. In: SDM 2002 (April 2002)