An Attribute Based Storage Method for Speeding up CLIQUE Algorithm for Subspace Clustering

Jyoti Pawar
DCST, Goa University,
GOA – INDIA.
jdp@unigoa.ac.in

P.R.Rao
DCST, Goa University,
GOA- INDIA.
pralhaadrao@rediffmail.com

Abstract

The subspace clustering algorithm CLIQUE finds all subspace clusters including overlapping clusters existing in high dimensional datasets. CLIQUE consists of three main steps namely –

(1) Identification of subspaces that contain clusters,
(2) Identification of clusters and
(3)Generation of the minimal description for the clusters obtained in step two.

In this paper, we have presented a method for speeding-up the first step of the CLIQUE algorithm. The proposed method is based on accessing the data from columns instead of rows. It is very efficient when there are many missing values in the high dimensional datasets given in the form of table. We have also proposed a depth-first method to find the maximal dense units, to further improve the performance of the first step.

1. Introduction

The subspace clustering algorithm *CLIQUE*[1] was the first clustering algorithm that found the clusters existing in the subspaces in high dimensional datasets. In this paper, we have used an attribute based storage method to access the values of the dataset and a depth-first method to find maximal dense units to identify the subspaces to speed-up the above stated step one of *CLIQUE*. In Subspace clustering, we find all the possible interrelationships that exist between all the attributes in the dataset. The dataset that we consider is a very high dimensional huge dataset and is likely to contain many missing values. We propose the **Attribute Oriented Storage Structure (AOSS)** for storing such datasets. Using AOSS, we store the information in such a way that all the records from the dataset need not be accessed to find the frequency counts of the candidate units. And at the same time we access the attribute information of only those attributes, which are present in the candidate unit whose frequency count is being found. In order to further improve the efficiency of the step one above, we used the maximal dense units to find the subspace clusters present in the dataset.

In this paper, we briefly present the results obtained using the proposed AOSS data representation and the MAximal Dense Unit GENeration (MADUGEN) algorithm. MADUGEN has been designed using a single threshold value for all attributes to find the maximal dense units present in a dataset. MADUGEN uses the AOSS method for data representation and is based on the GenMax[2] algorithm. The details of the AOSS data representation and the MADUGEN algorithm will appear in a full version of this paper.

2. Efficiency obtained using AOSS

The AOSS method of representation is useful in making the algorithm used to find the frequency count of candidate units efficient. An empirical evaluation of the CLIQUE algorithm using the AOSS structure with synthetic datasets was carried out. The performance of ROSCLIQUE(CLIQUE algorithm implemented using **R**ecord **O**riented **S**torage structure) v/s AOSSCLIQUE(CLIQUE algorithm implemented using **A**ttribute **O**riented **S**torage **S**tructure) was studied by varying the number of records with and without missing values, the total number of attributes and the dimension of the clusters. Using AOSS, the process of finding the one-dimensional dense units of all the attributes can be carried out in parallel. The Experimental Results obtained are shown in figure 1, figure 2 and figure 3.

3. Efficiency obtained using MADUGEN

The ROSCLIQUE and AOSSCLIQUE suffer when the dimensionality of the clusters increases. Using the MADUGEN algorithm to find the maximal dense units in step one along with the AOSS method has showed very good results. The results are shown in figure 4. AOMADUCLIQUE is implementation of CLIQUE using AOSS and MADUGEN algorithm.

10th International Database Engineering and Applications Symposium (IDEAS'06)
0-7695-2577-6/06 $20.00 © 2006 IEEE

Experimental Results

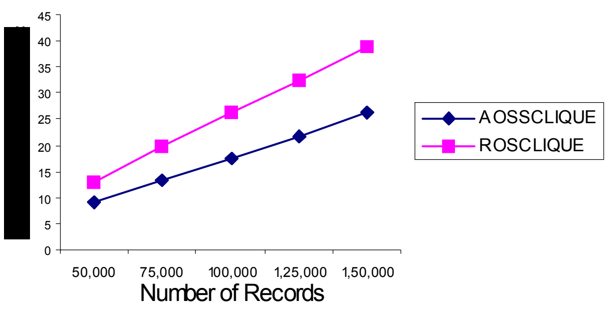

Figure 1. Scalability with the number of records (with missing values).

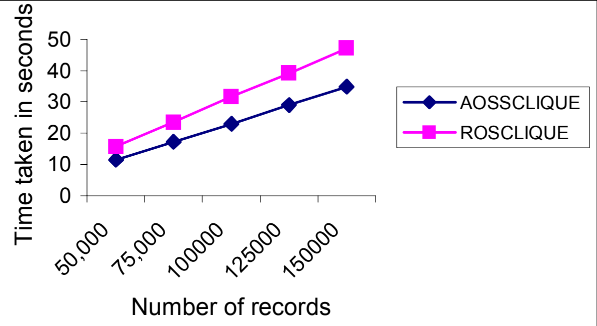

Figure 2. Scalability with number of records (without missing values)

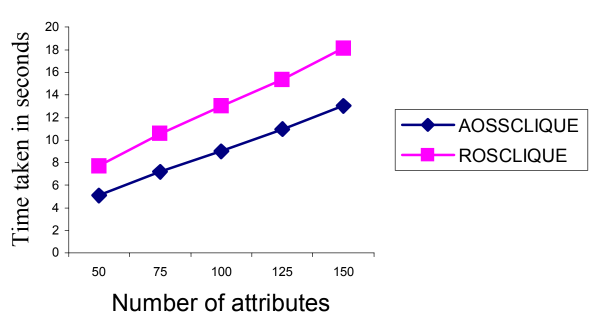

Figure 3. Scalability with the dimension of the data space.

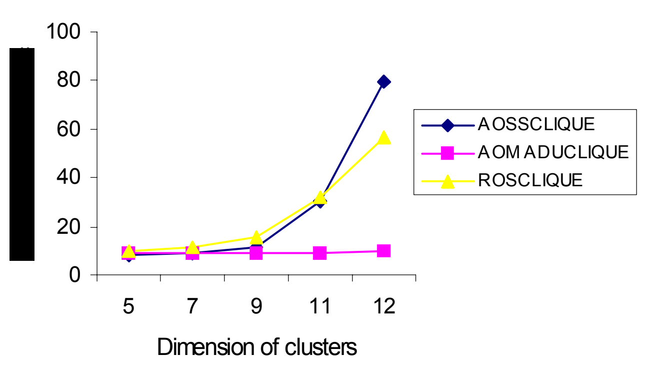

Figure 4. Scalability with the dimension of the data clusters.

4. Conclusion

In this paper, we have shown, by experimental study, that step one of the CLIQUE algorithm can be speeded up using AOSS data structure and by using maximal dense units. The performance gain is mainly due to the different approach used to find the frequency counts of the candidate units, which became possible due to AOSS.

5. References

[1] R. Agarwal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, 1998.

[2] M. J. Zaki, K. Gouda. GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets. Data Mining and Knowledge Discovery. 2004.