Efficient Algorithm for Projected Clustering

Eric Ng Ka Ka, Ada Wai-chee Fu
Department of Computer Science and Engineering
The Chinese University of Hong Kong
kkng1, adafu@cse.cuhk.edu.hk

Abstract

With high dimensional data, natural clusters are expected to exist in different subspaces. We propose the EPC (Efficient Projected Clustering) algorithm to discover the sets of correlated dimensions and the location of the clusters. This algorithm is quite different from previous approaches and has the following advantages: (1) no requirement of the input of the number of natural clusters, and the average cardinality of the subspaces; (2) can handle clusters of irregular shapes; (3) produces better clustering results, compared to the best previous method to our knowledge; (4) high scalability. From experiments, it is several times faster than the previous method, while producing more accurate results.

We treat projected clusters as connected regions where data densities are significantly higher than their surrounding area with respect to their associated subspace. When a natural cluster exists in certain subspace, for example *XYZ*, the density within the area spanned by this cluster, with respect to subspace *XYZ*, should be higher than its neighboring area. In many cases, the projection of this cluster on each dimension *X*, *Y*, *Z* would also demonstrate a higher density. Hence we consider the projection of data points on each dimension.

We model the densities of the data objects on each dimension by constructing a **density estimation function** $\hat{f}(x)$ which is influenced only by the data values that are near to x [2]. In our implementation, we use a **1-d histogram** to build this density estimation function on each dimension.

With the 1-d histograms, we apply an adaptive approach to discover **dense regions** on each dimension, where data objects are densely located when projected at the dimension. These regions are potentially the 1-d projections of some natural clusters. Recursively, we locate regions of the histograms where densities exceed some threshold determined by the mean value and standard deviation of the histogram. The adaptive approach allows us to uncover clustering regions of different densities.

When the dense regions for all dimensions are uncovered, we can determine for each data object whether it is in the scope of any such region. When such information for all dimensions are combined for a point, we get a **signature** for the point. If there are many data points with similar signatures, they highlight a potential clustering region. After grouping data objects in similar dense regions together, we can discover the correlated dimensions, as well as the location of the clusters.

Experiments have been conducted to evaluate the returned projected clustering quality, in terms of percentage of correctly partitioned data objects, and the number of correctly discovered correlated dimensions on a synthetic data set. The data set is generated in the same way as [1]. The results are compared with the best previous projected clustering algorithm known to us, PROCLUS [1], the comparison is shown in the following table:

	EPC			PROCLUS		
Data Set	*C*	*P*	*CD/TD*	*C*	*P*	*CD/TD*
3000	4	82	19/19	2	72	6/10
6000	4	91	18/18	2	65	7/10
10000	4	73	21/21	3	88	7/14
30000	5	91	25/25	2	39	10/12
50000	3	93	15/15	2	80	6/10
65000	4	92	19/19	3	91	11/15
80000	4	91	16/18	4	88	16/20
100000	3	76	13/13	0	0	0/0
200000	5	90	25/25	2	76	8/12

C - No. of correctly discovered clusters
P - Percentage of correctly paritioned data objects
CD/TD - No. of Correctly Discovered Dimensions / Total No. of Correlated Dimension in the correctly discovered clusters

The scalability of EPC is tested by varying the number of data objects, the number of dimensions, and the number of natural clusters. Our experiment shows that the running time of EPC grows approximately linearly with all 3 factors. The algorithm is applicable for very large data sets in high dimensional space.

References

[1] Charu C. Aggarwal, Cecilia Procopiuc, Joel L. Wolf, Philip S. Yu, Jong Soo Park, *Fast Algorithms for Projected Clustering*, SIGMOD99.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

Proceedings of the 18th International Conference on Data Engineering (ICDE'02)
1063-6382/02 $17.00 © 2002 **IEEE**

IEEE COMPUTER SOCIETY