Peer Review
Reviewing Paper : CLIQUE, Automatic Subspace Clustering of High Dimension Data for Data Mining Applications

Aditya Patil Bhasker Goel Kartikeya Saxena and V Anirudh

Department of Computer Science, Indian Institute of Technology, Guwahati
*Corresponding author: {aditya18, bhasker18, kartikey18a, anirudh18}@iitg.ac.in

1 Summary

- CLIQUE is a grid-based subspace clustering algorithm. It automatically identifies dense cluster in sub-spaces of a high dimensional data space. CLIQUE consists of three major steps, i.e., identification of subspaces that contain clusters, identification of clusters and generation of minimal description for the clusters. The algorithm first determines 1-dimensional dense units by making a pass over the data by creating histograms and counting the points contained in each unit. The algorithm proceeds level-by-level in a priori fashion. CLIQUE selects dense units only in interesting subspaces. The basic idea underlying the MDL principle is to encode the input data under a given model and select the encoding that minimizes the code length. CLIQUE takes $O(c^k + m \times k)$ for some constant c. The two popular variants of CLIQUE, namely, ENCLUS and MAFIA improve the subspace selection criterion of the original algorithm. ENCLUS uses an entropy criteria and only those subspaces spanned by attributes A_1, ..., A_p with entropy $H(A_1, ..., A_p) < \omega$ (a threshold) are selected for selected clustering. On the contrary, MAFIA constructs adaptive grids to improve subspace clustering and exploits parallelism to handle massive data sets.
- They have presented an incremental version of CLIQUE which identifies clusters in the maximum dimensionality space with pretty good accuracy. A count of number of points per unit hyper-rectangle is maintained. For insertion of points, if the newly inserted point was added to an already dense unit then nothing is done otherwise if it forms a newly dense unit, then that unit is combined to its neighbourhood dense unit's cluster or (to an independent cluster if neighbourhood dense unit not found). For deletion, if after deletion of a point, the unit is still dense then nothing is done otherwise DFS is run on the neighbouring dense unit's cluster to consider the potential split situations.
- Incremental algorithm achieves a significant reduction in time of execution (maximum speed up 99%) without much drop in accuracy (80.76% on average) or consumption of memory (25% more on average).

2 Key Strengths of the Project

- The code architecture is properly organized and documented.
- The suggested incremental algorithm is seen to perform much faster than the static algorithm and also without much loss of accuracy.
- The suggested algorithm performs both insertions and deletions of points.

3 Key Weaknesses of the Project

- The algorithm fails to merge two or more already existing clusters. This is the main reason for loss of accuracy in the suggested algorithm.
- Clusters are formed in horizontal and vertical directions, but are not formed in diagonal direction.

4 Suggested Improvements in the Project

- Testing was performed only on two datasets one synthetic and one real. Moreover synthetic dataset was too small to make reasonable conclusion on time and memory consumption. Could include some more datasets.
- The algorithm makes assumptions like "Boundary points are not considered for deletion" on the points to be deleted. Algorithm can be improved to handle these cases.

- Same data set could be analysed for different threshold and number of interval values.
- Incremental algorithms were already proposed for CLIQUE. Comparison with other incremental algorithms can be formed.
- More accuracy measures could be explored for the suggested incremental algorithm.

5 Suggested Improvements in the Report

- More information in terms of graphic visualization could be provided.
- Accuracy is not mentioned for all datasets. The tables could be more robust and descriptive.
- Very little description is given for the static algorithm. Could explain the underlying terminologies of the algorithm with more details.
- The tables and figures were not referenced anywhere in the report.
- Keywords should be after abstract.
- No explanation about the accuracy measure for testing.

6 Overall rating for the report on the scale of 0 to 10

All things considered the report was good. We think the overall rating for the report should be **9**.
Link to report: Group 4