DESIGN AND ANALYSIS OF SUBSPACE CLUSTERING ALGORITHMS AND THEIR APPLICABILITY

BY
JYOTI D. PAWAR

Abstract

Due to the rapid advancement in information technology, it has become very easy to capture data about almost every aspect of one's business or related field. The data captured is stored in various files or databases. Most of the time the mining of knowledge is carried out considering each of the files or databases independently. Hence, we cannot find the patterns or relationships that exist across attributes stored in different files or databases. In Subspace clustering, we try to find all the possible interrelationships that exist between the various data attributes by finding all the clusters that exist in the different subspaces of a very high dimensional dataset. The datasets that we deal with in subspace clustering contain a large number of attributes, are huge in size and most of the times contain many missing values.

In this thesis, based on the properties of very high dimensional huge data sets, and the requirements of the subspace clustering algorithms, an Attribute Oriented Storage Structure (AOSS) for storing very high dimensional huge data sets has been developed. Using the AOSS structure, the complexity of the function, to find the frequency count of the various candidate units in the datasets is reduced considerably. This fact is also proved by the experimental evaluation that has been carried out using
synthetic datasets. An algorithm to reduce the number of passes required over the dataset has been designed by using sampling technique and experimentally shown that it is efficient when we have to deal with huge datasets which cannot be loaded in main memory at one time. In order to efficiently find high-dimensional clusters in very high dimensional huge datasets, a depth-first approach instead of the currently used breadth-first method has been used to find the dense units in the datasets and it is extended to find clusters in datasets with attributes having varying threshold values. And finally, using the AOSS structure with this depth-first approach technique, proposed method to find the various clusters that exist within the clusters identified in the original datasets. This method can be very useful to do a through analysis of datasets in applications like census data analysis.

The AOSS structure along with the depth first method of finding the dense units is found to be very promising to make the design of the subspace clustering algorithms very efficient with respect to the space as well as the time factor to find high dimensional clusters in very high dimensional huge datasets.