Chapter 6

Discussion

We have developed an Attribute Oriented Storage Structure (AOSS), for effective and efficient subspace clustering of very high dimensional huge datasets. In this chapter, we first summarize the major characteristics of the very high dimensional huge datasets. We then explain the efficiency of the AOSS method for use in subspace clustering and give some extensions and applications of subspace clustering using the AOSS method.

6.1 Characteristics of the AOSS method

We have developed the new class of AOSS method for effective and efficient subspace clustering of very high dimensional huge datasets. We, summarize the major characteristics of AOSS based methods here.

- AOSS based methods adopt a divide-and-conquer methodology and partition the data sets consisting of various records from the high dimensional data space into independent AOSS Attribute tables for each attribute belonging to the data set. In general, subspace clustering has to search a very huge very high dimensional data
space. Divide-and-conquer methodology enables the subspace clustering algorithms to focus on reduced subsets of records within each AOSS Attribute table in the first pass and then again a much smaller subset of records belonging to the various dense units by applying the divide-and-conquer strategy on the AOSS Attribute tables. This process automatically focuses on only the relevant set of attributes and the relevant set of records belonging to the various units while finding their selectivity. Hence it eliminates the processing of irrelevant records as well as irrelevant attributes of a particular unit, thereby saving a lot of processing time. It also helps in processing some set of units, which are independent with respect to attributes and units in parallel.

- AOSS based methods also save on the main memory space requirements as those records and attributes not contributing to any dense units are automatically pruned and are not loaded in memory in subsequent processing of the high-dimensional candidate units for determining their selectivity.

- AOSS based-methods MADUGEN and MADUGENMT for finding the dense units use a depth-first search algorithm and eliminate the requirement to find the selectivity of all the \(2^{k-1}\) subsets of a k-dimensional dense unit before finding the selectivity of the k-dimensional unit, a feature which is most common in the level-wise (apriori based) algorithms.
6.2 Extensions and Applications of AOSS based methods

We have shown that the AOSS based methods are effective and efficient in subspace clustering for finding the dense units belonging to the various subspaces of a high dimensional dense unit. The AOSS method however is also applicable to mining other kinds of knowledge and solving some other interesting high dimensional data processing problems. In this section, we discuss some examples.

6.2.1 Mining Multi-dimensional Sequential Patterns from high-dimensional data

Sequential pattern mining, which finds the set of frequent subsequences in sequence databases, is an important data-mining task and has broad applications. Mining of sequential patterns from very high dimensional datasets which is a common requirement in the emerging new applications like protein classification, keyword extraction from text documents, etc is a very interesting task and can greatly benefit from the AOSS structure.
6.2.2 Mining Closed Association Rules from high-dimensional data

In order to reduce the generation of redundant association rules the use of the closed frequent itemsets has been proposed by Pasquier in [26a]. The AOSS can be used for the extraction of the frequent closed itemsets from the high dimensional datasets in an efficient manner.

6.2.3 Categorization of high dimensional datasets using subspace clustering

Many real datasets like collection of documents on various subjects, can be viewed as having very high dimensionality and missing dimensional values. Subspace clustering based on the AOSS can be used here, to find efficiently all the clusters some of which may be overlapping.

6.2.4 Mining long Sequences from high-dimensional data

Applications like protein classification, and other bio-informatics applications require effective and efficient mining of long sequences from their high dimensional
data sets. Since, AOSS based methods are efficient for the depth-first search, it can be used to extract the long sequences efficiently from the high dimensional datasets.