A Clustering Approach to Generalized Pattern Identification Based on Multi-instanced Objects with DARA

Abstrak

Clustering is an essential data mining task with various types of applications. Traditional clustering algorithms are based on a vector space model representation. A relational database system often contains multirelational information spread across multiple relations (tables). In order to cluster such data, one would require to restrict the analysis to a single representation, or to construct a feature space comprising all possible representations from the data stored in multiple tables. In this paper, we present a data summarization approach, borrowed from the Information Retrieval theory, to clustering in multi-relational environment. We find that the data summarization technique can be used here to capture the typical high volume of multiple instances and numerous forms of patterns. Our experiments demonstrate a technique to cluster data in a multi-relational environment and show the evaluation results on the mutagenesis dataset. In addition, the effect of varying the number of features considered in clustering on the classification performance is also evaluated.