Dimensionality Reduction in Data Summarization Approach to Learning Relational Data

Abstrak

Due to the growing amount of digital data stored in relational databases, more new approaches are required to learn relational data. The DARA algorithm is designed to summarize data and it is one of the approaches introduced in relational data mining in order to handle data with one-to-many relations. The DARA algorithm transforms data stored in relational databases into a vector space representation by applying the information retrieval theory. Based on the experimental results, the DARA algorithm is proven to be very effective in learning relational data. However, DARA suffers a major drawback when the cardinalities of attributes are very high because the size of the vector space representation depends on the number of unique values that exist for all attributes in the dataset. This paper investigates the effects of discretizing the magnitude of terms computed and applying a feature selection process that reduces the cardinalities of attributes of the relational datasets on the predictive accuracy of the overall classification task. This involves the task of finding the best set of relevant features used to summarize the data, in which the feature selection processed is performed based on the magnitude of terms computed earlier. Based on the results obtained, it shows that the predictive accuracy of the classification task can be improved by improving the quality of the summarized data. The quality of the summarized data can be enhanced by appropriately discretizing the magnitude of terms computed earlier and also appropriately selecting only a certain percentage of the attributes.