A Random Length Feature Construction Method for Learning Relational Data using DARA

Abstrak

In learning relational data, DARA (Dynamic Aggregation of Relational Attributes) algorithm transforms a relational data model representation into a vector space model representation. This data transformation is required in order to summarize or cluster data stored in relational databases in which a target record stored in a target table has a one-to-many relationship with non-target records stored in a non-target table. The descriptive accuracy of the summarized data performed by DARA is highly influenced by the representation of records stored in non-target tables that are associated with records stored in target table. This is important because when this summarized data is fed as input data for the classification task, the predictive accuracy of the classification task will also be affected. This paper proposes novel feature construction methods, called Variable Length Feature Construction without Substitution (VLFCWOS) and Variable Length Feature Construction with Substitution (VLFCWS), in order to construct a set of relevant features in learning relational data. These methods are proposed to improve the descriptive accuracy of the summarized data. In the process of summarizing relational data, a genetic algorithm is also applied and several feature scoring measures are evaluated in order to find the best set of relevant constructed features. In this work, we empirically compare the predictive accuracies of classification tasks based on the proposed feature construction methods and also the existing feature construction methods. The experimental results show that the predictive accuracy of classifying data that are summarized based on VLFCWS method using Total Cluster Entropy combined with Information Gain (CE-IG) as feature scoring outperforms in most cases