Summarizing relational data using semi-supervised genetic algorithm-based clustering techniques

Abstract

Problem statement: In solving a classification problem in relational data mining, traditional methods, for example, the C4.5 and its variants, usually require data transformations from datasets stored in multiple tables into a single table. Unfortunately, we may loss some information when we join tables with a high degree of one-to-many association. Therefore, data transformation becomes a tedious trial-and-error work and the classification result is often not very promising especially when the number of tables and the degree of one-to-many association are large. Approach: We proposed a genetic semi-supervised clustering technique as a means of aggregating data stored in multiple tables to facilitate the task of solving a classification problem in relational database. This algorithm is suitable for classification of datasets with a high degree of one-to-many associations. It can be used in two ways. One is user-controlled clustering, where the user may control the result of clustering by varying the compactness of the spherical cluster. The other is automatic clustering, where a non-overlap clustering strategy is applied. In this study, we use the latter method to dynamically cluster multiple instances, as a means of aggregating them and illustrate the effectiveness of this method using the semi-supervised genetic algorithm-based clustering technique. Results: It was shown in the experimental results that using the reciprocal of Davies-Bouldin Index for cluster dispersion and the reciprocal of Gini Index for cluster purity, as the fitness function in the Genetic Algorithm (GA), finds solutions with much greater accuracy. The results obtained in this study showed that automatic clustering (seeding), by optimizing the cluster dispersion or cluster purity alone using GA, provides one with good results compared to the traditional k-means clustering. However, the best result can be achieved by optimizing the combination values of both the cluster dispersion and the cluster purity, by putting more weight on the cluster purity measurement. Conclusion: This study showed that semi-supervised genetic

algorithm-based clustering techniques can be applied to summarize relational data with more effectively and efficiently.