Aggregating Multiple Instances in Relational Database Using Semi-Supervised Genetic Algorithm-based Clustering Technique.

Abstrak

In solving the 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. In this paper, we propose a genetic semi-supervised clustering technique as a means of aggregating data in multiple tables for the 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 paper, 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.