Discretization numerical data for relational data with one-to-many relations

Abstract

Problem statement: Handling numerical data stored in a relational database has been performed differently from handling those numerical data stored in a single table due to the multiple occurrences (one-to-many association) of an individual record in the non-target table and non-determinate relations between tables. Numbers in Multi-Relational Data Mining (MRDM) were often discretized, after considering the schema of the relational database. Study the effects of taking the one-to-many association issue into consideration in the process of discretizing continuous numbers. Approach: Different alternatives for dealing with continuous attributes in MRDM were considered in this study, namely equal-width (EWD), Equal-Height (EH), equal-weight (EWG) and Entropy-Based (EB). The discretization procedures considered in this study included algorithms that were not depended on the multi-relational structure of the data and also that are sensitive to this structure. A new method of discretization, called the entropy instance-based (EIB) discretization method was implemented and evaluated with respect to C4.5 on the two well-known multi-relational databases that include the Mutagenesis dataset and the Hepatitis dataset for Discovery Challenge PKDD 2005. Results: When the number of bins, b, is big (b = 8), the entropy-instance-based discretization method produced better data summarization results compared to the other discretization methods, in the mutagenesis dataset. In contrast, for the hepatitis dataset, the entropy-instance-based discretization method produced better data summarization results for all values of b, compared to the other discretization methods. In the Hepatitis dataset, all discretization methods produced higher average performance accuracy (%) for partitional clustering technique, compared to the hierarchical technique. Conclusion: These results demonstrated that entropy-based discretization can be improved by taking into consideration the multiple-instance problem. It was also found that the partitional clustering technique produced better performance accuracy compared to the one produced by hierarchical clustering technique.