Optimizing feature construction process for dynamic aggregation of relational attributes

Abstract

Problem statement: The importance of input representation has been recognized already in machine learning. Feature construction is one of the methods used to generate relevant features for learning data. This study addressed the question whether or not the descriptive accuracy of the DARA algorithm benefits from the feature construction process. In other words, this paper discusses the application of genetic algorithm to optimize the feature construction process to generate input data for the data summarization method called Dynamic Aggregation of Relational Attributes (DARA). Approach: The DARA algorithm was designed to summarize data stored in the non-target tables by clustering them into groups, where multiple records stored in non-target tables correspond to a single record stored in a target table. Here, feature construction methods are applied in order to improve the descriptive accuracy of the DARA algorithm. Since, the study addressed the question whether or not the descriptive accuracy of the DARA algorithm benefits from the feature construction process, the involved task includes solving the problem of constructing a relevant set of features for the DARA algorithm by using a genetic-based algorithm. Results: It is shown in the experimental results that the quality of summarized data is directly influenced by the methods used to create patterns that represent records in the (n×p) TF-IDF weighted frequency matrix. The results of the evaluation of the geneticbased feature construction algorithm showed that the data summarization results can be improved by constructing features by using the Cluster Entropy (CE) genetic-based feature construction algorithm. Conclusion: This study showed that the data summarization results can be improved by constructing features by using the cluster entropy genetic-based feature construction algorithm.