LEARNING RELATIONAL DATA USING CLUSTERING ENSEMBLES TECHNIQUES

GRANT NO: RAG0006-TK-2012

Tan Soo Fun, Rayner Alfred and Chin Kim On

Final Report

Faculty of Computing and Informatics
Universiti Malaysia Sabah
Malaysia
2015

LEARNING RELATIONAL DATA USING CLUSTERING ENSEMBLES TECHNIQUES

Abstract

Biodiversity data related to endangered species in Sabah are normally stored in relational databases in which data are stored in multiple tables. A data summarization approach to knowledge discovery in structured biodiversity datasets is often limited due to the complexity of the database schema. Since most of these data are stored in multiple tables, designing a suitable data summarization method for each individual table that is associated with the target table is required in order to get the best result in summarizing the overall data stored in a multi-relational environment. In short, the ever-growing amount of digital data stored in relational databases resulted in the need for new approaches to extract useful information from these databases. One of those approaches, the DARA algorithm, is designed to transform data stored in relational databases into a vector space representation utilising information retrieval theory. The DARA algorithm has shown to produce improvements over other state-of-the-art approaches. However, the DARA suffers a major drawback when the cardinality of attributes in relations are very high. This is because the size of the vector space representation depends on the number of unique values of all attributes in the dataset. This issue can be solved by reducing the number of features generated from the DARA transformation process by selecting only part of the relevant features to be processed. Since relational data is transformed into a vector space representation (in the form of $TF\text{-}IDF$), only numerical values will be used to represent each record. As a result, discretizing these numerical attributes may also reduce the dimensionality of the transformed dataset. When clustering is applied to these datasets, clustering results of various dimensions may be produced as the number of bins used to discretize these numerical attributes is varied. From these clustering results, a final consensus clustering can be applied to produce a single clustering result which is a better fit, in some sense, than the existing clusterings. In this study, an ensemble DARA clustering approach that provides a mechanism to represent the consensus across multiple runs of a clustering algorithm on the relational datasets is proposed.