AN EVOLUTIONARY BASED FEATURES CONSTRUCTION METHODS FOR DATA SUMMARIZATION APPROACH

GRANT NO: RAG0007-TK-2012

Rayner Alfred, Suraya Alias and Chin Kim On

PERPUSTAKAAN
UNIVERSITI MALAYSIA SABAH

Final Report

Faculty of Computing and Informatics
Universiti Malaysia Sabah
Malaysia
2015

AN EVOLUTIONARY BASED FEATURES CONSTRUCTION METHODS FOR DATA SUMMARIZATION APPROACH

Abstract

Coral reefs are on course to become the first ecosystem that human activity will eliminate entirely from the Earth, a leading United Nations scientist claims. It is predicted that this event will occur before the end of the present century, which means that there are children already born who will live to see a world without coral. Coral reefs are important for the immense biodiversity of their ecosystems. They contain a quarter of all marine species. This research addresses the question whether a data summarization approach can be utilized to predict the survival of Coral Reefs in Malaysia by identifying the survival factors for these Coral Reefs. A data summarization approach is proposed due to its capability to learn data stored in multiple tables. In other words, this research will discuss the application of genetic algorithm to optimize the feature construction process from the Coral Reefs data to generate input data for the data summarization method called Dynamic Aggregation of Relational Attributes (DARA). The DARA algorithm will be applied 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.This research proposes novel feature construction methods, called Variable Length Feature Constructionwithout Substitution (VLFCWOS) and Variable Length Feature Construction with Substitution(VLFCWS), in order to construct a set of relevant features in learning relational data. These methods are proposed to improve the descriptive accuracy of the summarized data. In the process of summarizing relational data, a genetic algorithm is also applied and several feature scoring measures are evaluated in order to find the best set of relevant constructed features. In this work, we empirically compare the predictive accuracies of classification tasks based on the proposed feature construction methods and also the existing feature construction methods. The experimental results show that the predictive accuracy of classifying data that are summarized based on VLFCWS method using Total Cluster Entropy combined with Information Gain (CE-IG) as feature scoring outperforms in most cases.

UMS
UNIVERSITI MALAYSIA SABAH