Ensemble Clustering Based on Feature Selection
Approach to Learning Relational Data
Kung Ke Shin, Rayner Alfred
Faculty of Computing and Informatics
Universiti Malaysia Sabah,
Jalan UMS, 88999, Kota Kinabalu, Sabah, Malaysia
keshincode9@gmail.com, ralfred@ums.edu.my.

Abstract. Many approaches have been developed to learn big relational data. One of the approaches used to learn big relational data is Dynamic Aggregation of Relational Attributes (DARA). The DARA algorithm is designed to summarize relational data with one-to-many relations. However, DARA suffers a major drawback when the cardinalities of attributes are very high because the size of the vector space representation depends on the number of unique values that exist for all attributes in the dataset. A feature selection process can be introduced to overcome this problem. In this work, a genetic algorithm based algorithm is used to select $k$ sets of features using a $k$-NN classifier. As, different set of selected features may produce different classification results, implementing a novel consensus feature selection method based on all these $k$ sets of the features is highly encouraged in order to achieve a good classification result. This can be achieved by introducing an ensemble technique into the framework by combining all the $k$ sets of selected features and producing the final consensus set of features selected. Ensembles are frequently used to improve the predictive accuracies of multiple classifiers by producing the final consensus result from multiple classifiers. In this work, an ensemble approach is introduced in a two-layered genetic algorithm-based feature selection in order to improve the classification performance in learning relational datasets. This work also investigates the effects of varying the number of features sets obtained from $k$-NN classifier and the proximity distance method used in the $k$-NN classifier on the classification performance in learning big relational data. Based on the results obtained from the experiments, it shows that the proposed method is able to improve the accuracies of classification tasks in which the $k$-NN classifiers with Euclidean distance as similarity measurements outperformed other classifiers.

Keywords: Relational Data Mining, $k$-Nearest Neighbours, Naïve Bayes, Classification, Ensembles, Feature Selection, Genetic Algorithm.