ABSTRACT

Methods to reduce the number of attributes and discretization are two important data pre-processing steps before the data can be used for classification activity. Web documents contain enormous number of attributes as compared to other type of data. Ant-Miner algorithm is also still lacking in efficiency, accuracy and rule simplicity because of the local minima problem. Therefore, the Ant-Miner algorithm needs to be improved by taking into consideration of the accuracy and rule simplicity criteria so that it could be used to classify Web documents data sets or any large data sets.

The best attribute selection method for Web texts categorization is the combination of correlation-based evaluation with random search as the search method. However, this attribute selection method will not give the best performance in attributes reduction. Using Classifier-based attribute subset selection will reduce more attributes, but sacrifice the performance of the classifier.

A hybrid ant colony optimization with simulated annealing algorithm to discover rules from data is proposed. The simulated annealing technique will minimize the problem of low quality discovered rule by an ant in a colony. The best rule for a colony will then be chosen and later the best rule among the colonies will be included in the rule set. The best rule for a colony will then be chosen and later the best rule among the colonies will be included in the rule set. The rule set is arranged in decreasing order of generation. Thirteen data sets which consist of discrete and continuous data were used to evaluate the performance of the proposed algorithm in terms of accuracy, number of rules and number of terms in the rules. Experimental results obtained from the proposed algorithm are comparable to the results of the Ant-Miner algorithm in terms of rule accuracy but are better in terms of rule simplicity.