ABSTRACT

Rule-based classification system (*RB_C*) has been widely used in many real world applications because of the easy interpretability of rules. *RB_C* mines a collection of rule via knowledge which is hidden in dataset in order to accurately map new cases to the decision class. In the real world, the number of attribute of dataset could be very large due the capability of database technology to store much information. Following that, the large dataset may contain thousands of relationship and it will likely provide more knowledge since the interrelationship between data will give more description. Furthermore, it is also have the possibility to have most number of rules that contain unnecessary rule or redundancies in the model. Theoretically, a good set of knowledge should provide good accuracy when dealing with new cases. Besides accuracy, a good rule set must also has a minimum number of rules and each rule should be short as possible. It is often that a rule set contains smaller quantity of rules but they usually have more conditions. An ideal model should be able to produces fewer, shorter rule and classify new data with good accuracy. Consequently, the quality and compact knowledge will contribute manager with a good decision model. Because of that, the search for appropriate data mining approach which can provide quality knowledge is important. Rough classifier (R_C) and decision tree classifier (DT_C) are categorized as *RB_C*. The purpose of this study is to investigate the capability of R_C and DT_C in generating quality knowledge which leads to the good accuracy. To achieve that, both classifiers are compared based on four measurements that are accuracy of the classification, the number of rule, the length of rule, and the coverage of rule. Five dataset from UCI Machine Learning namely United States Congressional Voting Records, Credit Approval, Wisconsin Diagnostic Breast Cancer, Pima Indians Diabetes Database, and Vehicle Silhouettes are chosen as data experiment. All datasets were mined using R_C toolkit namely ROSETTA while C4.5 algorithm in WEKA application was chosen as DT_C rule generator. The experimental results indicated that both classifiers produced good classification result and had generated quality rule in different types of model – higher accuracy, fewer rule, shorter rule, and higher coverage. In term of accuracy, R_C obtained higher accuracy in average while DT_C significantly generated lower number of rule than R_C. In term of rule length, R_C produced compact and shorter rule than DT_C and the length is not significantly different. Meanwhile, R_C has better coverage than DT_C. Final conclusion can be decided as follows “If the user interested at a variety of rule pattern with a good accuracy and the number of rule is not important, R_C is the best solution whereas if the user looks for fewer nr, DT_C might be the best choice”