Abstract

Modified Binary Cuckoo Search using rough set theory for Feature Selection

by Ahmed Fayez Alia

Feature Selection (FS) for classification is an important process to find the minimal subset of features from original data by removing the redundant and irrelevant features. This process aims to improve the classification accuracy, shorten computational time of classification algorithms, and reduce the complexity of classification model. Rough Set Theory (RST) is one of the effective approaches for feature selection, but it uses complete search to search for all combinations of features and uses dependency degree to evaluate these combinations. However, due to its high cost, complete search is not feasible for large datasets. In addition, RST, as it replies on the use nominal features, it cannot deal efficiently with mixed and numerical datasets [1]. Therefore, Meta-Heuristics algorithms especially nature inspired search algorithms have been widely used to replace the reduction part in RST. In addition other factors such as frequent values are used with dependency degree to improve the performance of RST for mixed and numerical datasets.

This thesis aims to propose a new filter feature selection approach for classification by developing a modified BCS algorithm, and a new objective function based on RST that utilizes distinct values to select the minimum number of features in an improved computational time yet without significantly reducing the performance of classification for nominal, mixed, and numerical datasets with different characteristics.

In the evaluation, our work and baseline approach are evaluated on sixteen datasets that are taken from the UCI repository of machine learning database. Also our work is compared with two known filter FS approaches (genetic and particle swarm optimization with correlation feature selection). Decision tree and naïve bays classification algorithms are used for measuring the classification performance of all approaches that are used in the evaluation. The results show our approach achieved best feature reduction for all mixed, all numerical, and most of nominal datasets compared to other approaches. Also our work achieved less computational time for all datasets compared to the baseline approach.