Development of a General Data Mining Methodology based on a Novel Weighted Hierarchical Adaptive Voting Ensemble Method and Its Applications

Huang, Yuchen

There are many machine learning methods developed for individual applications, and each has its strengths and weaknesses. However, none of these can be used for all databases and deliver satisfactory results in a reasonable amount of computation time. The purpose of this project is to develop a general data mining methodology that combines individual machine learning methods and achieves a higher accuracy than individual methods in a reasonable amount of computation time. This methodology is based on a novel Weighted Hierarchical Adaptive Voting Ensemble method, or WHAVE. It has four unique aspects. First, it uses statistical data processing to extract attribute Information Gain for preprocessing. Second, it applies a majority voting ensemble system with a novel weights formula that can be adjusted to yield the highest accuracy. Third, it employs a hierarchical ensemble algorithm to further improve accuracy and efficiency. Fourth, it is adaptive to newly updated databases and uses stopping criteria to search for the optimal ensemble. The WHAVE method was demonstrated in applications for breast cancer, heart disease detection and stock market prediction with accuracies of 99.8%, 96.7% and 95.2% respectively. Individual methods used to construct WHAVE include DNF, Decision Trees, Naive Bayes, kNN, Majority algorithm and SVM. A Python program was developed to implement the WHAVE method. Results show that WHAVE yielded highest accuracy compared to individual methods for all cases. WHAVE can function effectively on a large number of machine learning methods and databases in an adaptive way without risking high computation time.

Awards Won: