Abstract

Research Title : Data Mining Techniques for Automatical Disease Analysis

Author : Mr. Nivet Chirawichitchai

Year : 2010

.....

The objective of this research was to find the data mining techniques to create a model of efficiency in the automated analysis of disease classification for medical dataset by experiments with 7 algorithms, including Naïve Bayes, Multilayer Perceptron, Radial Basis Function Network, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Ripper. Comparative study of feature selection methods with Correlation-based Feature Subset Selection (CFS) and Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) including test algorithms on Single learning and Multiple learning by increasing the efficiency, and enhance the classification by Bagging and Boosting.

From the experimental results of this research showed all models were built with efficiency in the classification of the disease in up to 80% when the feature is not reduced. And when sorting accuracy (Accuracy) by type of dataset found that information Hypothyroid dataset on Decision Tree for the best performance is 99.57%, Leukemia dataset on Naive Bayes or Support Vector Machine for best performance is 98.61%, Breast-w dataset on Support Vector Machine for best performance is 96.99%, Lymphography dataset on Support Vector Machine for best performance is 86.48%, Hepatitis dataset on Radial Basis Function or K-Nearest Neighbor for best performance is 85.80%, Heart-c dataset on Support Vector Machine for best performance is 84.15%, Heart-statlog dataset on Radial Basis Function or Support Vector Machine for best performance is 84.07% respectively. The models that has an acceptable quality level and to develop software for the automatic diagnosis.

Optimization with Multiple learning methods using Bagging and Boosting algorithm that result is increased accuracy for some datasets. The observations that proportion of samples in each class must be similar or equal. If the distribution of samples in each class is very different proportions. Bagging and Boosting the performance does not improve.

Reducing features with Correlation-based Feature Subset Selection (CFS) and Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) resulted in performance accuracy to similar non-reducing features. But to reduce the dimensions of the features, saving resources of the computer system and the time to learn to build models as well.