Project Title: Multiple Classifier Algorithm Comparison

Overview

Building a robust classifier is an important task in data mining. There exist several algorithms that can be used for training a classification model. However, not all of them will have the same performance. Given the nature of a data set, one can identify the set of classifier algorithms that can be applied but it is not trivial to pick the best performing method. While performance metrics are helpful to some extent, they do not indicate whether the results of the best performing classifier are significantly different than the rest of the classifiers.

Hypothesis testing provides a solution to estimate the statistical significance of a classifier and allows one to compare the performance of several classifiers over multiple data sets. In the process, a hypothesis test formulates a null hypothesis and an alternative hypothesis. The null hypothesis is tested given the evidence (such as mean of classification errors etc.) by calculating a statistic (for example, t statistic, F statistic). Based on the value of the statistic, a p-value is calculated which indicates whether we reject the null hypothesis or fail to reject it.

This project gives an opportunity to understand hypothesis testing and its application to derive useful information about the performance of classifier algorithms. In the process, you will be required to apply different statistical tests that will evaluate the performance of, 1) two classifier algorithms on a single data set, 2) more than two classifier algorithms on a single data set and 3) comparison of two algorithms over multiple data sets.

Datasets

The experiments will be conducted on a built-in data set in R and five data sets from the UCI repository.

Procedure

For the first task, you will be using the iris data set to build two classification models using C5.0 decision trees (see C5.0 package) and Support Vector Machines (see kernlab package). The training and test sets will be divided using 10-fold cross validation and the error percentages for each fold will be calculated. In the next step, you need to perform a K-Fold Cross-Validated Paired *t*-Test to compare the mean of the error percentages of the two classifier algorithms.

In the second task, you now compare the performance of four classifier algorithms on the Breast Cancer data set. In addition to C5.0 decision trees and Support Vector Machines, you will be using Naïve Bayes (see e1071 package) and Logistic Regression (see stats package) to build classification models. To compare their performance, you will be applying the *analysis of variance* (ANOVA) test.

In the last part, you will compare performance of two classifier algorithms over five data sets collected from the UCI repository. To do this, you will be applying the Wilcoxon Signed Rank Test. For all the three tasks, you are supposed to perform cross-validation to compare the performance of a classifier.

Note: It is recommended that you read the pdf of Chapter 19 Design and Analysis of Machine Learning Experiments provided by Dr. Samatova.

Q&A:

Think about the following questions:

- What is the null hypothesis and the alternative hypothesis in each of the statistical test?
- What did you conclude after performing the tests? State which hypothesis seemed favorable given the evidence.
- Briefly explain why signed rank test is more useful than ANOVA test while comparing two classifier algorithms on multiple data sets. (Hint: See Section 19.13 of Chapter 19 Design and Analysis of Machine Learning Experiments)