COMPARING CLASSIFIERS EFFECTIVELY USING MCNEMAR'S TEST

By

Bhavani Shankar Yanamadala

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Approved:

_____

Prof. George Nagy
Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

Aug 2007
(For Graduation Dec 2007)

ABSTRACT

Classifiers are generally evaluated on the basis their performance on dataset(s). The performance measure used is predominantly based on the error rate, for example: difference in error rate, the difference in proportion, etc. We look at contingency tables for improving the evaluation of classifiers. If the error rate of classifier A is lower than that of classifier B, Test I is considered *better* than Test II if the probability that Test I indicates that A has lower error rate than B is higher than the corresponding probability for Test II. Some authors [**?**] [**?**] claim that contingency tables perform better. However, we also look at the conditions, under which the performance of the tests based on contingency tables is better. In order to do this, synthetic data sets are used to generate tables comparing the performance of classifier evaluation tests. Five tests are discussed and compared in this work. It is shown experimentally that McNemar's test for contingency tables performs better than the test for difference in two proportions. This is because of its ability to resolve better the cases where the difference in the error rate of the two pairs of classifiers is the same, but correlation is not. Type II error tests are also performed, and they indicate that this error is also dependent on correlation and sample size. Tables that reflect the performance of these tests under different conditions are also given.

The performance of three more tests which can only be run on multiple datasets is also discussed, and tables are generated to compare and analyze their performance. Two of those tests, namely, Paired t-test and Wilcoxon signed ranks test make use of ranking scheme based on error rate for evaluating classifiers. The performance of these tests with a ranking scheme based on McNemar's test is observed. We also discuss under what conditions this kind of simulation is valid, and also what kind of real world datasets show this sort of performance.