Demšar, Janez

Statistical comparisons of classifiers over multiple data sets. (English) Zbl 1222.68184

J. Mach. Learn. Res. 7, 1-30 (2006).

Summary: While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
62G30 Order statistics; empirical distribution functions

Keywords:

comparative studies; statistical methods; Wilcoxon signed ranks test; Friedman test; multiple comparisons tests

Full Text: Link