Journal of Computer and System Sciences 80 (2014) 88–100

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Bootstrap analysis of multiple repetitions of experiments using an interval-valued multiple comparison procedure

José Otero [a], Luciano Sánchez [a,*], Inés Couso [b], Ana Palacios [c]

[a] Universidad de Oviedo, Computer Science Department, Spain
[b] Universidad de Oviedo, Statistics Department, Spain
[c] Universidad de Granada, Computer Science Department, Spain

ARTICLE INFO

Article history:
Received 23 July 2012
Received in revised form 5 December 2012
Accepted 14 March 2013
Available online 21 March 2013

Keywords:
Cross validation
Statistical comparisons of algorithms
Tests for interval-valued data

ABSTRACT

A new bootstrap test is introduced that allows for assessing the significance of the differences between stochastic algorithms in a cross-validation with repeated folds experimental setup. Intervals are used for modeling the variability of the data that can be attributed to the repetition of learning and testing stages over the same folds in cross validation. Numerical experiments are provided that support the following three claims: (1) Bootstrap tests can be more powerful than ANOVA or Friedman test for comparing multiple classifiers. (2) In the presence of outliers, interval-valued bootstrap tests achieve a better discrimination between stochastic algorithms than nonparametric tests. (3) Choosing ANOVA, Friedman or Bootstrap can produce different conclusions in experiments involving actual data from machine learning tasks.

1. Introduction

The most common experimental setup for comparing multiple machine learning algorithms is k fold cross-validation. Data sets are broken into k disjoint subsets of approximately equal size. For each fold, a subset is removed, the system trained on the remaining data and tested on the held-out subset. The training sets overlap, but all test sets are independent [22].

Cross validation is often combined with a single factor repeated measures experimental design [5]. This is a design with one response variable, where each experimental unit is measured multiple times in this variable. In the context of this contribution, experimental units are the algorithms being compared. The values of the response variable are the averages of the k test values obtained for each pair (algorithm, dataset) with the cross-validation setup. The significance of differences between algorithms is assessed with repeated-measures ANOVA or its nonparametric equivalent, the Friedman test [5]. Multiple comparisons tests are accompanied by post-hoc tests that assess the relevance of paired differences between algorithms [6,9,10].

Algorithms whose output depends only on training and test sets are called *deterministic*, and those that also depend on a random seed are called *stochastic* [17]. For comparing stochastic algorithms, the variability added by the random seed must be accounted for by repeating each fold a number of times. In this case the single factor repeated measures experimental design cannot be applied. There are designs considering multiple independent observations per cell [14], but according to [5] they cannot be applied to this problem because repeating training/test episodes breaks the independence assumption of the test values, thus analyzing the variance of the repetitions of folds in cross validation is a yet unresolved problem.

* Corresponding author.
E-mail address: luciano@uniovi.es (L. Sánchez).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2013.03.009

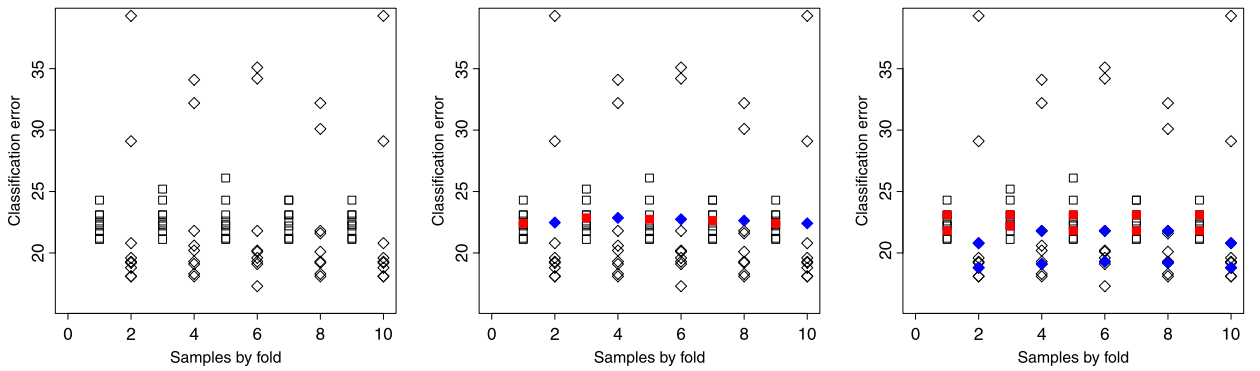

Fig. 1. 10-cv based comparison to two stochastic algorithms. Left: 10 repetitions of each algorithm. Center: solid red and blue symbols mark sample means of each fold. Right: solid symbols mark interquartile ranges of the same folds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this paper it is proposed that intervals are used for describing the part of the variability of the data that can be attributed to the repetition of learning and testing stages over the same sets. Each group of non-independent repetitions will be consolidated into a single interval-valued measure of the response variable, thus the single factor repeated measures design can still be applied. The drawback of the proposal is the need of extending the experimental design and statistical tests to interval-valued data [8]. In this respect, extending ANOVA or Friedman tests to interval data would be feasible, but involves an optimization task that is computationally costly. On the contrary, there exist efficient algorithms for the particular case of bootstrap tests for interval data [3]. This raises the question about whether bootstrap tests improve ANOVA or Friedman tests for this particular problem. It will be shown that the answer is positive, thus a new bootstrap test is introduced that allows for assessing the significance of the differences between stochastic algorithms in a cross-validation with repeated folds experimental setup.

The structure of this paper is as follows: in Section 2 the interval representation is introduced, and the general procedure for extending paired tests to interval data recalled. In Section 3 the proposed bootstrap tests are defined for point and interval data. In Section 4 a numerical analysis is included where the following three conclusions are supported by data: (1) Boostrap tests can be more powerful than ANOVA or Friedman test for comparing multiple classifiers. (2) In the presence of outliers, interval-valued bootstrap tests achieve a better discrimination between stochastic algorithms than nonparametric tests. (3) Choosing ANOVA, Friedman or Bootstrap can produce different conclusions in experiments involving actual data from machine learning tasks. The paper concludes in Section 5, with the concluding remarks and future work.

2. Interval-valued representations and statistical tests

Consider the example shown in Fig. 1. Test errors after 100 executions of two stochastic algorithms are plotted. Results of the first algorithm are drawn with squares, and those of the second are drawn with diamonds. The experimental setup is 10-cv with 10 repetitions. Horizontal axis are folds, and the vertical axis represents the classification error of each training/test pair.

Repetitions of the 'square' algorithm form compact clouds, but some executions of the 'diamond' algorithm were trapped in local minima. Average errors of both are the same (see Fig. 1, central part) but the typical error of the diamonds is better, as shown in the interquartile ranges in the rightmost part of the same figure. Different facts can be tested with this data:

- If the null hypothesis is *average accuracies of algorithms are the same*, both algorithms seem to be similar. However, the experimental design is not adequate for drawing this conclusion. The sample mean is not a good estimator of the test error of the diamond algorithm, because different repetitions for the same fold are not independent, as mentioned in the introduction. For instance, should the data set contain one instance that disrupted the learning algorithm, this instance would be a part of the training set in ninety percent of the experiments, heavily biasing the error estimate. It is a well-known fact that cross validation should not be applied to algorithms that are not stable with respect to the data set, i.e. to algorithms for which a small change in the training set triggers large deviations in the test error [15]. Stochastic algorithms are unstable in the sense that if they converge to local minima, large changes in the test error may occur without modifying the training set.
- If the null hypothesis is *typical accuracies of algorithms are the same*, then the diamond algorithm is better. "Typical accuracy" can be understood either as median, censored mean or interquartile range, to name some robust estimates. The percentage of repetitions that must be kept and discarded for obtaining a robust estimate can be estimated with additional experiments about the convergence ratio of the learning algorithm. Intervals are arguably more informative than punctual estimations for this purpose. Some authors claim that they allow for better modeling of asymmetrical distributions [18]. For instance, the smallest intervals covering at least 10% of repetitions of each algorithm could be used for describing the typical range of accuracies. Centers of these intervals provide information about the mode of the distribution of the repetitions. Their widths inform about the dispersion of the same distribution.