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

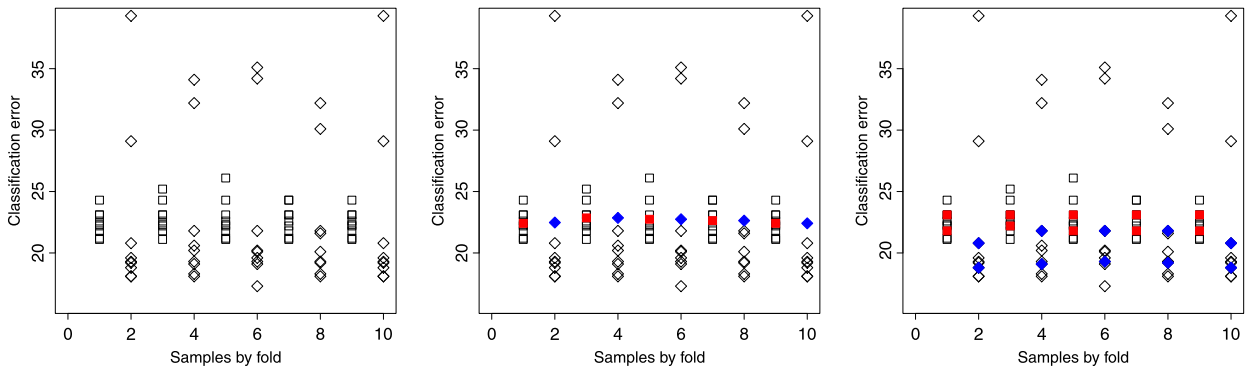

Fig. 1. 10-cv based comparison to two stochastic algorithms. Left: 10 repetitions of each algorithm. Center: solid red and blue symbols mark sample means of each fold. Right: solid symbols mark interquartile ranges of the same folds. (For interpretation of the references to color in this figure legend,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

For deciding whether the differences between interquartile ranges of diamonds and squares in Fig. 1 could have happened by chance or not, a statistical test for interval data must be used. Different extensions of statistical tests to interval-valued data have been proposed (see [3] for a discussion about this subject). The generalization used in this paper for paired tests is described in the remaining of this section. Multiple comparison tests will be addressed in Section 3.

Let $([q^a_{1-}, q^a_{1+}], \ldots, [q^a_{k-}, q^a_{k+}])$ and $([q^b_{1-}, q^b_{1+}], \ldots, [q^b_{k-}, q^b_{k+}])$ be interval-valued measurements of the typical accuracy of two classifiers a and b in k folds. Let $x^a = (x^a_1, x^a_2, \ldots, x^a_k)$ and $x^b = (x^b_1, x^b_2, \ldots, x^b_k)$ be two vectors of k real numbers each. Lastly, let the test being generalized be defined by a function $p(x^a, x^b)$ that maps each pair (x^a, x^b) to the probability of the null hypothesis being false (*p-value*), given that x^a and x^b are the test errors of either classifier at each fold.

Given two vectors of intervals

$$\left([q^a_{1-}, q^a_{1+}], \ldots, [q^a_{k-}, q^a_{k+}]\right) \tag{1}$$

and

$$\left([q^b_{1-}, q^b_{1+}], \ldots, [q^b_{k-}, q^b_{k+}]\right) \tag{2}$$

the p-value of the extended test is defined as the interval $[p^-, p^+]$, where

$$p^-(x^a, x^b) = \inf\left\{p(x^a, x^b) \mid x^a_i \in [q^a_{i-}, q^a_{i+}],\ x^b_i \in [q^b_{i-}, q^b_{i+}]\right\} \tag{3}$$

$$p^+(x^a, x^b) = \sup\left\{p(x^a, x^b) \mid x^a_i \in [q^a_{i-}, q^a_{i+}],\ x^b_i \in [q^b_{i-}, q^b_{i+}]\right\} \tag{4}$$

Observe that determining p^- and p^+ requires solving two constrained non-linear optimization problems with $2k$ variables and $2k$ interval restrictions each.

3. Two proposals of bootstrap tests for making multiple comparisons

As mentioned in the introduction, a multiple comparison procedure is needed for comparing series of executions of different algorithms. Friedman's test is often used because normality is not assumed in rank tests [5]. But replacing measurements by their ranks has the same effect as if the sample size is reduced by 3% for very large samples and much more for smaller ones [11]. In addition to this, Friedman's test requires that the distribution of the differences scores between any pair of levels is continuous and symmetrical in the population. This assumption is required to ensure that the test evaluates difference in medians rather than other characteristics of the distribution [16].

Bootstrap tests make less restrictive assumptions [7], nonetheless their use in combination with cross-validation is not common. In this section two permutations-based bootstrap test are proposed that can be applied to single factor repeated measures designs, either with scalar or interval-valued data.

3.1. Test Bootstrap-A for multiple comparisons of algorithms with scalar data

Let e_{adfr} be the test error of the a-th algorithm in the d-th dataset, f-th fold and r-th repetition. Let n_d, n_a, n_f and n_r the number of datasets, algorithms, folds and repetitions in the experimental setup. Let

$$\hat{F}_{a\cdots}(x) = \frac{1}{n_d n_r n_f} \#\left\{(d, r, f) \mid e_{adfr} \leqslant x\right\} \tag{5}$$

be the sample cumulative distribution function (cdf) of the outcome of the a-th algorithm, and let

$$\hat{F}_{\cdots}(x) = \frac{1}{n_a n_d n_r n_f} \#\left\{(a, d, r, f) \mid e_{adfr} \leqslant x\right\} \tag{6}$$

be the sample cdf of the prior distribution of the test error. Let $F_{a\cdots}$ and $F_{\cdots}$ be the corresponding population cdfs.

If the differences between the algorithms were not significant, the expectations obtained with respect to $F_{\cdots}$ and wrt $F_{1\cdots}, \ldots, F_{n_a\cdots}$ should not be significantly different. The null hypothesis of the test will then be expressed as "the expectations

$$e_a = \int x \, \mathrm{d}F_{a\cdots}, \quad a = 1, \ldots, n_a \tag{7}$$

do not depend on the algorithm index a".

Following [11], this problem can be solved with a bootstrap test, obtained via rearrangements of the sample. This requires four steps:

1. Choice of test statistic that best discriminates between the primary hypothesis and the alternative hypothesis.
2. The value of this statistic is determined for the set of observations before rearrangement of their labels.
3. A rearrangement distribution is generated by computing the value of the test statistic for each rearrangement.

4. The value of the statistic obtained at step 2 is compared with the set of possible values generated at step 3. If the original value of the test statistic lies in the tails of the rearrangement distribution favoring the alternative hypothesis, the primary hypothesis is rejected.

It is proposed that these steps are implemented as follows:

1. The test statistic is the sample mean.
2. The value before rearrangement is a vector of n_a components $\hat{e}_a$. These are the expected test errors of the algorithms wrt cdfs $\hat{F}_{a\cdots}$:

$$\hat{e}_a = \frac{\sum_{i=1}^{n_f}\sum_{j=1}^{n_r}\sum_{k=1}^{n_d} e_{akij}}{n_f n_r n_d} \tag{8}$$

3. Let $\{\pi_d\}_{d=1,\ldots,n_d} = \{(\alpha_{1,d},\ldots,\alpha_{n_a,d})\}_{d=1,\ldots,n_d}$ be a family of permutations of the indices $1,\ldots,n_a$, and let

$$\hat{e}_a^* = \frac{\sum_{i=1}^{n_f}\sum_{j=1}^{n_r}\sum_{k=1}^{n_d} e_{\alpha_{a,k}kij}}{n_f n_r n_d} \tag{9}$$

the value of the test statistics for the rearrangement given by $\{\pi_d\}_{d=1,\ldots,n_d}$. The rearrangement distributions of the values $\hat{e}_a^*$ are numerically approximated by bootstrap estimation.
4. If the value $\hat{e}_a$ belongs to the tails of the distribution of $\hat{e}_a^*$ for any a, the null hypothesis is rejected and the index a marks the algorithms whose expected error is different than the average. The tails of the distribution of $\hat{e}_a^*$ must be determined so that their probability mass is lower than the significance level of the test, adjusted for simultaneous n_a tests.

In case the null hypothesis is rejected, the post-hoc tests for comparing pairs of algorithms can be defined by particularizing the same test: let $\pi_d^{(2)} = (\alpha_{1,d}^{(2)}, \alpha_{2,d}^{(2)})$ be a permutation of the pair of indices (a,b), and let

$$\hat{e}_b^{(2)} = \frac{\sum_{i=1}^{n_f}\sum_{j=1}^{n_r}\sum_{k=1}^{n_d} e_{\alpha_{2,k}^{(2)}kij}}{n_f n_r n_d} \tag{10}$$

If the value $\hat{e}_a$ belongs to the tails of the distribution of $\hat{e}_b^{(2)}$, the null hypothesis "the test errors of algorithms a and b are the same" is rejected. The tails of the distribution of $\hat{e}_b^{(2)}$ are determined as before.

3.2. Test Bootstrap-B for multiple comparisons of algorithms with interval data

The interval-valued bootstrap test proposed in this section will be called Bootstrap-B. Let $[e_{-adfr}, e_{+adfr}]$ be the interval-valued error of the a-th algorithm in the d-*th* dataset, f-th fold and r-th repetition. In the first place, each group of r repetitions of an algorithm over the same fold is consolidated into a confidence interval $[q_{-adf}, q_{+adf}]$. For scalar problems, $e_{-adfr} = e_{+adfr}$ and $q_{-adf} = q_{+adf}$ is a robust central tendency measure summarizing the n_r repetitions of the algorithm.

Let $[\hat{F}_{-a\cdot\cdot}(x), \hat{F}_{+a\cdot\cdot}(x)]$ be the sample cdf of the outcome of the a-th algorithm [4],

$$\hat{F}_{-a\cdot\cdot}(x) = \frac{1}{n_d n_f}\#\left\{(d,f) \;\middle|\; q_{+adf} \leqslant x\right\} \tag{11}$$

$$\hat{F}_{+a\cdot\cdot}(x) = \frac{1}{n_d n_f}\#\left\{(d,f) \;\middle|\; x \in [q_{-adf}, q_{+adf})\right\} + \hat{F}_{-a\cdot\cdot}(x) \tag{12}$$

and let $[\hat{F}_{-\cdots}(x), \hat{F}_{+\cdots}(x)]$ be the sample cdf of the prior distribution of the test error,

$$\hat{F}_{-\cdots}(x) = \frac{1}{n_a n_d n_f}\#\left\{(a,d,f) \;\middle|\; q_{+adf} \leqslant x\right\} \tag{13}$$

$$\hat{F}_{+\cdots}(x) = \frac{1}{n_a n_d n_f}\#\left\{(a,d,f) \;\middle|\; x \in [q_{-adf}, q_{+adf})\right\} + \hat{F}_{-\cdots}(x). \tag{14}$$

Let also $[F_{-a\cdot\cdot}(x), F_{+a\cdot\cdot}(x)]$ and $[F_{-\cdots}(x), F_{+\cdots}(x)]$ be the corresponding population cdfs. The null hypothesis of the test will then be expressed as "the set of expectations

$$[q_{-a}, q_{+a}] = \left\{\int x\,\mathrm{d}F \;\middle|\; F(x) \in \left[F_{-a\cdot\cdot}(x), F_{+a\cdot\cdot}(x)\right] \text{ for all } x\right\}, \quad a = 1,\ldots,n_a \tag{15}$$

do not depend on the algorithm index a".

Extending [3], a rearrangement bootstrap problem will be defined for comparing a mix of scalar or interval data-based algorithms, with the following premises:

1. The test statistic is the sample Aumann mean [19].
2. The value before rearrangement is a vector of n_a intervals $[\hat{q}_{-a}, \hat{q}_{+a}]$:

$$\hat{q}_{-a} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{-aki}}{n_f n_d} \tag{16}$$

$$\hat{q}_{+a} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{+aki}}{n_f n_d} \tag{17}$$

3. Let $\{\pi_d\}_{d=1,\ldots,n_d} = \{(\alpha_{1,d}, \ldots, \alpha_{n_a,d})\}_{d=1,\ldots,n_d}$ be a family of permutations of the indices $1, \ldots, n_a$, and let

$$\hat{q}^*_{-a} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{-\alpha_{a,k}ki}}{n_f n_d} \tag{18}$$

$$\hat{q}^*_{+a} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{+\alpha_{a,k}ki}}{n_f n_d} \tag{19}$$

   the value of the test statistics for the rearrangement given by $\{\pi_d\}_{d=1,\ldots,n_d}$. The rearrangement distributions of the values $\hat{q}^*_a$ are numerically approximated by bootstrap estimation, as before.
4. If the interval $[\hat{q}_{-a}, \hat{q}_{+a}]$ belongs to the tails of the distribution of $[\hat{q}^*_{-a}, \hat{q}^*_{-a}]$ for any a, the null hypothesis is rejected and these indices a mark the algorithms whose expected error is different than the average. In other words, let $q^*_{-adf}(s)$ and $q^*_{+adf}(s)$ be the results of evaluating expressions 18 and 19 in the s-th bootstrap resample, and let n_s the number of these resamples. Then,

$$\hat{F}^*_{-a}(x) = \frac{1}{n_s} \#\left\{s \mid q^*_{+adf}(s) \leqslant x\right\} \tag{20}$$

$$\hat{F}^*_{+a}(x) = \frac{1}{n_s} \#\left\{s \mid x \in \left[q^*_{-adf}(s), q^*_{+adf}(s)\right)\right\} + \hat{F}^*_{-a}(x) \tag{21}$$

   For an adjusted signification level α, the test is rejected if any of the following conditions are met:

$$\hat{F}^*_{-a}(\hat{q}_{-a}) > 1 - \frac{\alpha}{2} \tag{22}$$

$$\hat{F}^*_{+a}(\hat{q}_{+a}) < \frac{\alpha}{2} \tag{23}$$

In case the null hypothesis is rejected, the post-hoc tests for comparing pairs of algorithms can be defined, as was done in the preceding case, by particularizing the test: let $\pi^{(2)}_d = (\alpha^{(2)}_{1,d}, \alpha^{(2)}_{2,d})$ be a permutation of the pair of indices (a, b), and let

$$\hat{q}^{(2)}_{-b} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{-\alpha^{(2)}_{2,k}ki}}{n_f n_d} \tag{24}$$

$$\hat{q}^{(2)}_{+b} = \frac{\sum_{i=1}^{n_f} \sum_{k=1}^{n_d} q_{+\alpha^{(2)}_{2,k}ki}}{n_f n_d} \tag{25}$$

If the value $[\hat{q}_{-a}, \hat{q}_{+a}]$ belongs to the tails of the distribution of $[\hat{q}^{(2)}_{-b}, \hat{q}^{(2)}_{+b}]$, the null hypothesis "the set of test errors of algorithms a and b are the same" is rejected. This happens when any of the following conditions are met:

$$\hat{F}^{(2)}_{-b}(\hat{q}_{-a}) > 1 - \frac{\alpha}{2} \tag{26}$$

$$\hat{F}^{(2)}_{+b}(\hat{q}_{+a}) < \frac{\alpha}{2} \tag{27}$$

where

$$\hat{F}^{(2)}_{-b}(x) = \frac{1}{n_s} \#\left\{s \mid q^{(2)}_{+b}(s) \leqslant x\right\} \tag{28}$$

$$\hat{F}^{(2)}_{+b}(x) = \frac{1}{n_s} \#\left\{s \mid x \in \left[q^{(2)}_{-b}(s), q^{(2)}_{+b}(s)\right)\right\} + \hat{F}^{(2)}_{-b}(x) \tag{29}$$

4. Numerical results

Numerical experiments are provided that are not in disagreement with 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.

Experiments related to items 1 and 2 are based on synthetic data. Item 3 will be supported by standard machine learning benchmarks.

4.1. Claims 1 and 2

In this section, power and type I error of ANOVA, Friedman, Bootstrap-A and Bootstrap-B tests are estimated by the fraction of correct and wrong conclusions taken by these tests when confronted with synthetic classification problems with known statistical properties.

Let $x(\omega)$ be a set of features measured on an object $\omega \in \Omega$, whose class is denoted as class(ω). Let $A(x(\omega))$ be the output of a classification algorithm, and let

$$e_A = P\left\{\omega \in \Omega \mid A\big(x(\omega)\big) \neq \text{class}(\omega)\right\} \tag{30}$$

be the expected error of this classifier. Let also T be a test set comprising n_t objects, $T = \{\omega_1, \ldots, \omega_{n_t}\}$. The fraction of misclassifications in T is

$$\hat{e}_A(T) = \frac{1}{n_t} \#\left\{\omega \in T \mid A\big(x(\omega)\big) \neq \text{class}(\omega)\right\} \tag{31}$$

and $\hat{e}_A(T)$ is an estimator of e_A. In a k-fold cv based experimental design, classifiers are learned from k training sets and tested in k independent test sets. The experimental measurement of the performance of the algorithm A on a given dataset is a vector comprising k different estimations

$$\big(\hat{e}_A(T_1), \ldots, \hat{e}_A(T_k)\big) \tag{32}$$

for k independent test sets $T_1, \ldots, T_k$.

The simulation of these estimations will be different for a deterministic algorithm (the outcome of the learning process is uniquely determined by the training set) or a stochastic algorithm (the outcome of the learning process is determined by both the training set and a random seed). Both are described below.

4.1.1. Deterministic algorithms

Assuming that the probability of misclassifying an instance is e_A, a random variable Y_A following a binomial distribution models the number of errors in the test set T:

$$Y_A \to \text{B}(n_t, e_A) \tag{33}$$

thus the fraction of errors is

$$\hat{e}_A(T_i) = \frac{1}{n_t} Y_A \tag{34}$$

4.1.2. Stochastic algorithms

For stochastic algorithms, the probability of committing an error is higher if the learning algorithm is trapped in a local minimum. Let $A^{(r)}$ be the r-th repetition of the algorithm being simulated, let $p_{A^{(r)}}$ be the probability that $A^{(r)}$ is trapped in a local minimum, and let $e^*_{A^{(r)}}$ be the average fraction of misclassifications committed in this case. Let $Y_{A^{(r)}}$ be a random variable with binomial distribution, as before:

$$Y_{A^{(r)}} \to B(n_t, e_{A^{(r)}}) \tag{35}$$

and let $Z_{A^{(r)}}$ be a random variable with Bernoulli distribution,

$$Z_{A^{(r)}} \to B(1, p_{A^{(r)}}) \tag{36}$$

Assuming that $Y_{A^{(r)}}$ and $Z_{A^{(r)}}$ are independent, the test error of $A^{(r)}$ is modeled as follows:

$$\hat{e}_{A^{(r)}}(T_i) = Z e^*_{A^{(r)}} + \frac{1}{n_t}(1 - Z_{A^{(r)}}) Y_{A^{(r)}} \tag{37}$$

Table 1
Theoretical errors and simulated sample errors for $\Delta p = 0.03$.

Dataset	A_1	A_2	A_3	A_4	A_5	A_1	A_2	A_3	A_4	A_5	n_t
1	**0.20**	0.37	0.40	0.49	0.44	0.22	0.41	0.33	0.37	0.42	12
2	0.31	0.37	0.40	**0.20**	0.44	0.34	0.34	0.42	0.23	0.46	13
3	0.31	0.37	0.40	**0.20**	0.44	0.39	0.32	0.47	0.17	0.49	15
4	0.31	0.37	0.40	0.49	**0.20**	0.35	0.41	0.37	0.52	0.17	19
5	0.31	**0.20**	0.40	0.49	0.44	0.34	0.23	0.32	0.45	0.39	12
6	0.31	**0.20**	0.40	0.49	0.44	0.36	0.18	0.38	0.48	0.46	18
7	0.31	0.37	**0.20**	0.49	0.44	0.28	0.37	0.19	0.39	0.46	19
8	0.31	0.37	**0.20**	0.49	0.44	0.38	0.34	0.26	0.47	0.36	16
9	**0.20**	0.37	0.40	0.49	0.44	0.22	0.33	0.41	0.55	0.49	16
10	0.31	0.37	0.40	**0.20**	0.44	0.38	0.35	0.46	0.16	0.48	10
11	0.31	**0.20**	0.40	0.49	0.44	0.39	0.15	0.42	0.57	0.44	12
12	0.31	0.37	0.40	**0.20**	0.44	0.28	0.49	0.33	0.24	0.52	11
13	**0.20**	0.37	0.40	0.49	0.44	0.13	0.42	0.38	0.46	0.43	16
14	0.31	0.37	0.40	**0.20**	0.44	0.35	0.33	0.32	0.15	0.50	13
15	0.31	0.37	**0.20**	0.49	0.44	0.31	0.32	0.16	0.40	0.46	17
16	0.31	0.37	0.40	**0.20**	0.44	0.31	0.32	0.43	0.17	0.50	14
17	0.31	0.37	0.40	**0.20**	0.44	0.35	0.31	0.36	0.26	0.41	17
18	0.31	0.37	**0.20**	0.49	0.44	0.31	0.40	0.19	0.49	0.44	19
19	0.31	**0.20**	0.40	0.49	0.44	0.32	0.21	0.38	0.43	0.37	13
20	0.31	**0.20**	0.40	0.49	0.44	0.35	0.24	0.39	0.55	0.43	17
21	0.31	0.37	0.40	**0.20**	0.44	0.31	0.42	0.40	0.33	0.43	19
22	0.31	0.37	0.40	0.49	**0.20**	0.25	0.35	0.37	0.48	0.17	12
23	0.31	**0.20**	0.40	0.49	0.44	0.33	0.21	0.46	0.46	0.47	16
24	0.31	0.37	**0.20**	0.49	0.44	0.34	0.29	0.16	0.49	0.47	11
25	**0.20**	0.37	0.40	0.49	0.44	0.17	0.36	0.35	0.38	0.42	12
26	0.31	0.37	**0.20**	0.49	0.44	0.30	0.39	0.17	0.52	0.44	13
27	0.31	0.37	0.40	**0.20**	0.44	0.21	0.42	0.38	0.14	0.46	10
28	0.31	0.37	0.40	**0.20**	0.44	0.34	0.39	0.48	0.18	0.48	13
29	0.31	0.37	0.40	0.49	**0.20**	0.24	0.43	0.46	0.51	0.18	18
30	0.31	0.37	**0.20**	0.49	0.44	0.32	0.43	0.22	0.51	0.42	13
31	0.31	**0.20**	0.40	0.49	0.44	0.41	0.19	0.46	0.49	0.44	14
32	0.31	0.37	0.40	**0.20**	0.44	0.31	0.37	0.39	0.25	0.41	15
Avg.	0.30	0.33	0.36	0.39	0.42	0.31	0.33	0.35	0.38	0.42	14.5

4.1.3. Experimental setup and results for claim 1

Five algorithms $A_1, \ldots, A_5$ and 32 datasets are simulated. 5-fold cross validation with 30 repetitions is used. A_3 and A_4 are deterministic, A_1, A_2 and A_5 are stochastic. In this first experiment, none of the stochastic algorithms converges to a suboptimal solution, $p^{(r)}_{A_i} = 0$.

For each test set T_d, one algorithm j_d is assigned the theoretical error $e^{(r)}_{A_{j_d}}(T_d) = 0.20$. The remaining algorithms were assigned a higher value such that the average of the theoretical errors of the algorithms for each dataset is $\frac{1}{32}\sum_d e^{(r)}_{A_i}(T_d) = 0.30 + (i-1)\cdot\Delta p$, $i = 1, \ldots, 5$. For each value $\Delta p = 0, 0.005, 0.01, \ldots, 1$, 100 simulations were made (see Table 1 for an example of theoretical errors and simulated sample means for $\Delta p = 0.03$).

In Fig. 2, power and type I errors are plotted for ANOVA (dotted line), Friedman (dashed line) and Bootstrap-A (solid line). The contents of this figure are:

1. Left part: Power of the tests, estimated by the fraction of times the combination of multiple comparisons test and post-hoc tests correctly detected that an algorithm was better than other. Horizontal axis is Δp, vertical axis is the power. Bootstrap-A is more powerful than a Friedman test followed by Wilcoxon post-hoc tests and Hochberg adjustment, as claimed. In turn, Friedman's test is better than ANOVA followed by t-tests.
2. Right part: Type I error of the tests, estimated by the fraction of times the combination of multiple comparisons test and post-hoc tests wrongly concluded that an algorithm was better than a preferable alternative. The horizontal axis is Δp, vertical axis is the error. Notice that the significance level is 0.95 thus it is expected that this error is 0.05 (marked with the horizontal dotted line).

In Table 2, numerical values plotted in Fig. 2 are given and in Table 3 a detail of the column for $\Delta p = 0.03$ is provided. The number of significant and correct comparisons (labeled "Sig OK"), not significant ("No Sig") and significant but wrong conclusions ("Sig Err") were obtained for each pair of algorithms being compared.

4.1.4. Experimental setup and results for claim 2

As done in the preceding section, five algorithms $A_1, \ldots, A_5$ and 32 datasets are simulated. 5-fold cross validation with 30 repetitions is used. A_3 and A_4 are deterministic, A_1, A_2 and A_5 are stochastic. In this second experiment, stochastic

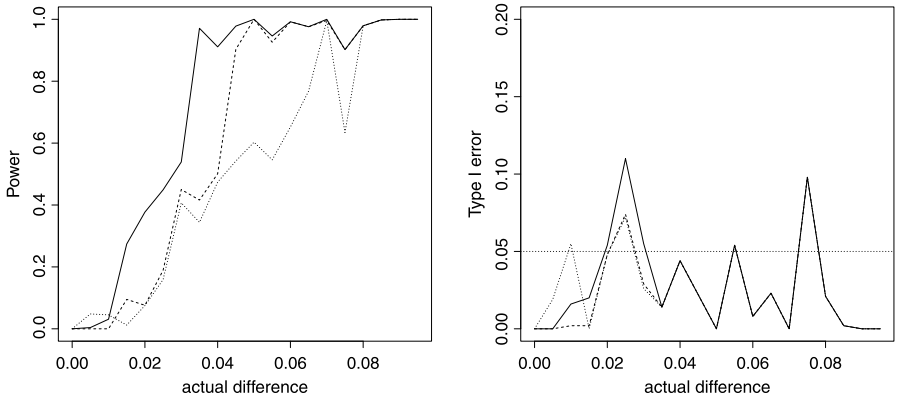

Fig. 2. Left: Average power of post-hoc tests a function of distance. Right: Average type I error of post-hoc tests as a function of distance. Solid line: Bootstrap-A. Dotted line: ANOVA + t-test. Dashed line: Friedman + Wilcoxon. Horizontal dotted line in the right part: expected type I error (0.05).

Table 2
Numerical data plotted in Fig. 2. Column "MC" contains the number of simulations where the multiple comparisons test detected a relevant difference. Columns "PH" count how many post-hoc tests found existing differences between each pair of algorithms ("Sig OK"), found non-existing differences ("Sig Err") or did not find differences ("No Sig").

Actual Δp	MC			PH								
	Sig			t-test			Wilcoxon			Bootstrap-A		
	AOV	Fried	Boot-A	Sig OK	No Sig	Sig Err	Sig OK	No Sig	Sig Err	Sig OK	No Sig	Sig Err
0	0	6	4	0	1000	0	0	1000	0	0	1000	0
0.005	0	69	2	48	933	19	0	1000	0	4	996	0
0.01	2	100	30	45	900	55	0	998	2	31	953	16
0.015	64	100	99	12	988	0	95	903	2	274	706	20
0.02	100	100	100	75	877	48	76	876	48	377	569	54
0.025	100	100	100	159	769	72	189	737	74	449	441	110
0.03	100	100	100	407	567	26	450	521	29	539	406	55
0.035	100	100	100	344	642	14	416	570	14	971	15	14
0.04	100	100	100	474	482	44	501	455	44	911	45	44
0.045	100	100	100	542	436	22	903	75	22	978	0	22
0.05	100	100	100	603	397	0	1000	0	0	1000	0	0
0.055	100	100	100	546	400	54	926	20	54	946	0	54
0.06	100	100	100	652	340	8	992	0	8	992	0	8
0.065	100	100	100	768	209	23	976	1	23	976	1	23
0.07	100	100	100	996	4	0	996	4	0	1000	0	0
0.075	100	100	100	634	268	98	902	0	98	902	0	98
0.08	100	100	100	979	0	21	979	0	21	979	0	21
0.085	100	100	100	998	0	2	998	0	2	998	0	2
0.09	100	100	100	1000	0	0	1000	0	0	1000	0	0
0.095	100	100	100	1000	0	0	1000	0	0	1000	0	0
1	100	100	100	1000	0	0	1000	0	0	1000	0	0
Avg				0.56 Pow		0.025 T1 Err	0.67 Pow		0.022 T1 Err	**0.76** Pow		0.027 T1 Err

algorithms can converge to a suboptimal solution with probability $p_{A_i}^{(r)} = 0.1$. The expected error of suboptimal classifiers is $e^*_{A^{(r)}}(T_d) = 0.75$.

For each test set T_d, one algorithm j_d was assigned a theoretical error $e^{(r)}_{A_{j_d}}(T_d) = 0.20$ and the remaining algorithms were assigned an error such that $\frac{1}{32}\sum_d e^{(r)}_{A_i}(T_d) = 0.30 + (i-1)\cdot\Delta p$, $i = 1,\ldots,5$. For each value $\Delta p = 0, 0.005, 0.01, \ldots, 1$, 100 simulations were made. The number of samples of the test partitions was chosen at random between 10 and 20. The interval estimation of the dispersion of the repetitions is estimated by a confidence interval, centered in the median and covering 10% of data.

In Fig. 3, power and type I errors are plotted for ANOVA (dotted line), Friedman (dashed line) and Bootstrap-A (solid line). The contents of this figure are:

1. Left part: Power of the tests is estimated by the fraction of times the combination of multiple comparisons test and post-hoc tests detected that an algorithm was better than other. Horizontal axis is Δp, vertical axis is the power. Bootstrap-B is more robust than Friedman and achieves better discrimination, as claimed. For instance, differences as high as 0.05 were considered as not significant in 69% of simulations by ANOVA + t-test, 46% by Friedman Test and

Table 3
Detail of Table 2 for $\Delta p = 0.03$.

Actual $\Delta p = 0.030$	(Hochberg pv adjust)						Bootstrap-A		
	ANOVA + t-test			Friedman + Wilcoxon					
	Sig OK	No Sig	Sig ERR	Sig OK	No Sig	Sig ERR	Sig OK	No Sig	Sig ERR
A_1 vs. A_2	0	100	0	0	100	0	0	100	0
A_1 vs. A_3	0	100	0	0	100	0	0	100	0
A_1 vs. A_4	0	100	0	0	100	0	0	100	0
A_1 vs. A_5	0	100	0	0	100	0	0	100	0
A_2 vs. A_3	35	45	20	38	39	23	49	2	49
A_2 vs. A_4	55	45	0	61	39	0	98	2	0
A_2 vs. A_5	55	45	0	61	39	0	98	2	0
A_3 vs. A_4	83	16	1	97	2	1	99	0	1
A_3 vs. A_5	84	16	0	98	2	0	100	0	0
A_4 vs. A_5	95	0	5	95	0	5	95	0	5
Avg.	40.7	56.7	2.6	45	52.1	2.9	53.9	40.6	5.5

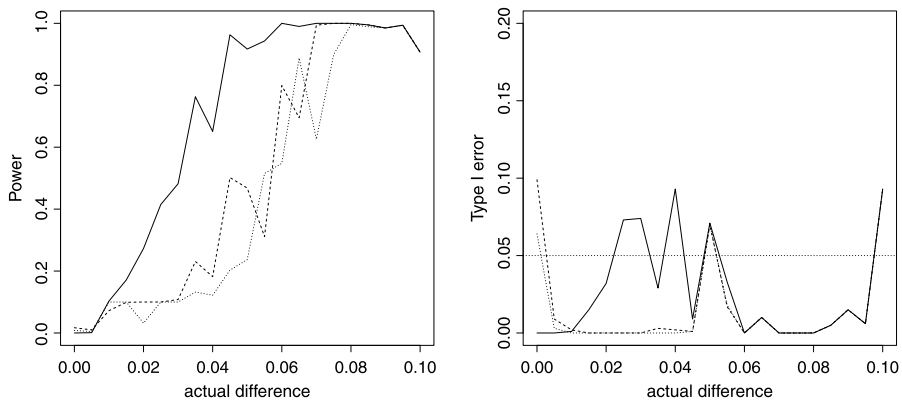

Fig. 3. Data with 10% of outliers. Solid line: Bootstrap-B. Dotted line: ANOVA + t-test. Dashed line: Friedman + Wilcoxon. Left: Average power of post-hoc tests a function of the differences between the theoretical errors of the classifiers. Right: Estimation of type I error of post-hoc tests as a function of the theoretical differences.

only in 1% by Bootstrap-B. Observe also that, in the presence of outliers, Friedman's test is not always better than ANOVA followed by t-tests.

2. Right part: Type I error of the tests, estimated by the fraction of times the combination of multiple comparisons test and post-hoc tests wrongly concluded that an algorithm was better than other. It is expected that this error is 0.05. For Bootstrap-B, not conclusive results were regarded as not significant.

In Table 4, numerical values plotted in Fig. 2 are given and in Table 5 a detail of the column for $\Delta p = 0.03$ is provided. The number of significant and correct comparisons (labeled "Sig OK"), not significant ("No Sig"), significant but wrong conclusions ("Sig Err") and (only for Bootstrap-B) not conclusive ("Inc") were obtained for each pair of algorithms being compared. Post-hoc tests were assigned the outcome "not significant" whenever the corresponding multiple comparisons test was not conclusive, as mentioned before.

4.2. Claim 3

In this section, an experimentation is designed to check whether the state-of-the-art fuzzy classification algorithm FURIA [13] is better than a selection of classical classifiers in imbalanced classification problems. FURIA, Linear Discriminant Analysis (LDA) [21], Nearest Neighbor (1NN) [21], Multilayer Perceptron (NNET) [12] and C4.5 [20] were applied to 64 imbalanced classification problems taken from KEEL repository [1]. Their performances were measured both by the classification error and by the area under the ROC curve (AUC) [2]. The average results of 30 repetitions of each pair (algorithm, dataset) are shown in Table 6.

According to the results, FURIA has the highest fraction of correct classifications. FURIA and 1NN are tied if the performance is measured with AUC. For assessing the relevance of the differences, three different sets of tests were applied to the data: (1) ANOVA + paired t-tests, (2) Friedman + Wilcoxon and (3) Bootstrap-B with confidence intervals with mass 10%. The p-values of the three tests are shown in Table 7.

Observe that ANOVA and Friedman's tests show a strong relevance of the differences in AUC, however post-hoc tests are needed to show the fact that LDA and not FURIA is responsible of this result (LDA is significantly worse than the mean).

Table 4
Numerical data plotted in Fig. 3. Column "MC" contains the number of simulations where the multiple comparisons test detected a relevant difference. Columns "PH" count how many post-hoc tests found existing differences between each pair of algorithms ("Sig OK"), found non-existing differences ("Sig Err"), did not find differences ("No Sig") or were inconclusive ("Inc").

Actual Δp	MC				PH t-test			PH Wilx			PH Boot-B			
	Sig		Boot-B		Sig OK	No Sig	Sig Err	Sig OK	No Sig	Sig Err	Sig OK	No Sig	Sig Err	Inc
	AOV	Fried	Sig	Inc										
0	93	55	13	0	8	928	64	17	884	99	0	968	0	32
0.005	82	90	6	41	5	992	3	10	981	9	1	958	0	41
0.01	78	100	61	38	100	900	0	72	926	2	103	886	1	10
0.015	99	100	100	0	100	900	0	99	901	0	171	718	15	96
0.02	100	100	100	0	32	968	0	100	900	0	273	567	32	128
0.025	100	100	100	0	100	900	0	100	900	0	415	144	73	368
0.03	100	100	100	0	100	900	0	108	892	0	482	101	74	343
0.035	100	100	100	0	132	868	0	231	766	3	763	15	29	193
0.04	100	100	100	0	122	878	0	182	816	2	651	4	93	252
0.045	100	100	100	0	203	796	1	502	497	1	963	0	9	28
0.05	100	100	100	0	237	693	70	468	462	70	917	0	71	12
0.055	100	100	100	0	515	467	18	311	672	17	943	0	33	24
0.06	100	100	100	0	548	452	0	799	201	0	1000	0	0	0
0.065	100	100	100	0	888	102	10	695	295	10	990	0	10	0
0.07	100	100	100	0	626	374	0	994	6	0	1000	0	0	0
0.075	100	100	100	0	899	101	0	1000	0	0	1000	0	0	0
0.08	100	100	100	0	996	4	0	1000	0	0	1000	0	0	0
0.085	100	100	100	0	989	6	5	995	0	5	995	0	5	0
0.09	100	100	100	0	985	0	15	985	0	15	985	0	15	0
0.095	100	100	100	0	994	0	6	994	0	6	994	0	6	0
0.1	100	100	100	0	907	0	93	907	0	93	907	0	93	0
					0.47		0.011	0.53		0.011	**0.73**		0.028	0.075
Avg					Pow		T1 Err	Pow		T1 Err	Pow		T1 Err	

Table 5
Detail of Table 3 for $\Delta p = 0.03$.

Actual $\Delta p = 0.030$	(Hochberg pv adjust)						Bootstrap-B			
	ANOVA + t-test			Friedman + Wilcoxon			Sig OK	No Sig	Sig ERR	Inc
	Sig OK	No Sig	Sig ERR	Sig OK	No Sig	Sig ERR				
A_1 vs. A_2	0	100	0	0	100	0	4	20	0	76
A_1 vs. A_3	0	100	0	0	100	0	4	20	0	76
A_1 vs. A_4	0	100	0	0	100	0	4	20	0	76
A_1 vs. A_5	0	100	0	0	100	0	4	20	0	76
A_2 vs. A_3	0	100	0	0	100	0	19	7	61	13
A_2 vs. A_4	0	100	0	0	100	0	74	6	7	13
A_2 vs. A_5	0	100	0	0	100	0	81	6	0	13
A_3 vs. A_4	0	100	0	4	96	0	92	2	6	0
A_3 vs. A_5	0	100	0	4	96	0	100	0	0	0
A_4 vs. A_5	100	0	0	100	0	0	100	0	0	0
Avg.	10	90	0	10.8	89.2	0	48.2	10.1	7.4	34.3

Bootstrap-B provides more information: it correctly shows that algorithms LDA and 1NN are responsible of the differences in AUC (the quality of 1NN is better than the mean, LDA is inferior and the data is inconclusive for FURIA). If the performance of the classifier is measured by the test error, FURIA and NNET are both different than the average because FURIA is better and NNET is worse.

For ANOVA and Friedman, the setup in [10] is followed and the best ranked classifier for AUC (FURIA) has been compared to its alternatives and the results shown in Table 8. The only disagreement between the tests is in FURIA vs. C4.5 (boldfaced in the table). In Fig. 4 density functions of the distributions of values of AUC of FURIA, C4.5 and their paired differences are displayed, thus the similarity between these algorithms can be judged. Observe that the mode of both algorithms is the same and therefore the mode of their difference is zero. The typical performance of both is typically the same, however there are a small number of datasets for which FURIA performed better than C4.5. After Hochberg adjust, a paired t-test between FURIA and C4.5 does not reject that both algorithms have the same AUC, but Friedman's test reject the hypothesis at 99% level. An interval-valued bootstrap test estimate a p-value between 0.03 and 0.40, thus the test is inconclusive, meaning that the dispersion of the results is too high and a decision cannot be taken.

Table 6
AUC and test error of 5 machine learning algorithms in 64 datasets.

Dataset	AUC					Test error				
	FURIA	LDA	1NN	NNET	C4.5	FURIA	LDA	1NN	NNET	C4.5
ecoli147vs2356	0.85	0.81	0.82	0.83	0.84	0.96	0.96	0.95	0.94	0.94
ecoli34vs5	0.84	0.86	0.87	0.86	0.81	0.95	0.94	0.96	0.95	0.95
glass	0.78	0.54	0.81	0.68	0.73	0.81	0.64	0.85	0.72	0.72
ecolivs1	0.99	0.98	0.96	0.96	0.98	0.99	0.99	0.96	0.96	0.96
leddigit02456789vs1	0.90	0.90	0.91	0.89	0.88	0.97	0.96	0.96	0.96	0.96
yeastvs4	0.85	0.83	0.85	0.85	0.83	0.95	0.96	0.96	0.95	0.95
ecoli67vs35	0.85	0.75	0.83	0.81	0.85	0.96	0.92	0.95	0.93	0.93
glass6vs5	0.95	0.68	0.94	0.98	0.99	0.98	0.92	0.97	0.99	0.99
wisconsin	0.97	0.95	0.95	0.94	0.95	0.97	0.96	0.96	0.95	0.95
ecoli1vs5	0.84	0.89	0.87	0.88	0.81	0.95	0.97	0.97	0.96	0.96
ecoli234vs5	0.84	0.89	0.87	0.86	0.79	0.96	0.96	0.97	0.95	0.95
pima	0.74	0.72	0.65	0.64	0.70	0.78	0.77	0.68	0.70	0.70
glass146vs2	0.52	0.49	0.63	0.56	0.64	0.92	0.91	0.87	0.84	0.84
glass15vs2	0.52	0.48	0.61	0.53	0.50	0.89	0.87	0.87	0.83	0.83
iris	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
ecoli147vs56	0.79	0.80	0.87	0.86	0.75	0.96	0.96	0.97	0.96	0.96
glass	0.83	0.70	0.78	0.66	0.82	0.87	0.76	0.79	0.73	0.73
yeast359vs78	0.60	0.61	0.67	0.62	0.59	0.91	0.92	0.88	0.85	0.85
clevelandvs4	0.71	0.78	0.59	0.69	0.66	0.93	0.94	0.89	0.92	0.92
yeast2579vs368	0.89	0.89	0.87	0.82	0.84	0.97	0.97	0.96	0.93	0.93
yeast	0.71	0.63	0.64	0.67	0.67	0.80	0.76	0.70	0.75	0.75
vehicle1	0.74	0.71	0.59	0.60	0.66	0.82	0.80	0.70	0.76	0.76
vehicle2	0.98	0.96	0.92	0.76	0.95	0.99	0.97	0.93	0.88	0.88
vehicle3	0.75	0.70	0.61	0.58	0.67	0.84	0.80	0.73	0.76	0.76
ecoli146vs5	0.78	0.87	0.87	0.83	0.78	0.95	0.97	0.98	0.96	0.96
yeast256vs3789	0.73	0.72	0.76	0.69	0.66	0.93	0.93	0.91	0.89	0.89
ecoli46vs5	0.84	0.89	0.87	0.87	0.81	0.96	0.97	0.97	0.95	0.95
ecoli	0.84	0.92	0.90	0.86	0.81	0.97	0.98	0.98	0.97	0.97
glass123vs456	0.88	0.88	0.94	0.83	0.92	0.92	0.93	0.96	0.90	0.90
ecoli01vs235	0.77	0.86	0.79	0.83	0.77	0.94	0.96	0.94	0.95	0.95
vehicle0	0.95	0.93	0.91	0.79	0.93	0.97	0.95	0.94	0.89	0.89
yeast1vs7	0.57	0.56	0.64	0.64	0.59	0.94	0.93	0.91	0.89	0.89
ecoli0267vs35	0.80	0.85	0.78	0.82	0.76	0.94	0.96	0.95	0.94	0.94
ecoli1	0.87	0.85	0.80	0.82	0.86	0.91	0.88	0.86	0.87	0.87
haberman	0.59	0.55	0.55	0.56	0.58	0.74	0.74	0.68	0.71	0.71
shuttlec0vsc4	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
glass04vs5	0.99	0.89	0.95	0.99	0.99	0.99	0.96	0.99	1.00	1.00
glass4	0.85	0.59	0.94	0.93	0.79	0.97	0.93	0.98	0.97	0.97
newthyroid2	0.94	0.81	0.99	0.93	0.95	0.98	0.94	1.00	0.97	0.97
ecoli0346vs5	0.86	0.86	0.90	0.88	0.82	0.97	0.95	0.98	0.96	0.96
newthyroid1	0.97	0.84	0.98	0.90	0.95	0.98	0.95	0.99	0.97	0.97
pageblocks13vs4	1.00	0.75	0.86	0.95	1.00	1.00	0.96	0.97	0.99	0.99
ecoli0347vs56	0.79	0.81	0.87	0.87	0.79	0.95	0.95	0.95	0.96	0.96
ecoli2	0.86	0.82	0.92	0.85	0.86	0.94	0.91	0.95	0.91	0.91
glass016vs5	0.84	0.59	0.84	0.82	0.89	0.97	0.95	0.96	0.97	0.97
segment0	0.99	0.98	0.99	0.99	0.98	1.00	0.98	0.99	1.00	1.00
yeast05679vs4	0.70	0.73	0.69	0.69	0.68	0.92	0.94	0.90	0.89	0.89
ecoli067vs5	0.84	0.86	0.84	0.86	0.77	0.96	0.96	0.95	0.96	0.96
glass6	0.88	0.91	0.93	0.88	0.81	0.96	0.96	0.97	0.94	0.94
shuttlec2vsc4	0.95	0.94	1.00	1.00	1.00	0.99	0.98	1.00	1.00	1.00
vowel0	0.96	0.86	1.00	0.99	0.97	0.99	0.95	1.00	1.00	1.00
yeast1458vs7	0.51	0.50	0.57	0.55	0.50	0.95	0.96	0.94	0.90	0.90
yeast3	0.88	0.83	0.80	0.83	0.86	0.95	0.94	0.93	0.93	0.93
ecoli3	0.80	0.84	0.75	0.71	0.73	0.93	0.93	0.91	0.89	0.89
glass016vs2	0.54	0.49	0.59	0.55	0.62	0.89	0.90	0.88	0.84	0.84
glass5	0.90	0.64	0.84	0.88	0.90	0.99	0.96	0.97	0.98	0.98
glass2	0.58	0.49	0.61	0.54	0.67	0.93	0.91	0.88	0.84	0.84
pageblocks0	0.94	0.79	0.87	0.90	0.92	0.98	0.95	0.96	0.97	0.97
yeast2vs8	0.77	0.77	0.74	0.74	0.50	0.98	0.98	0.96	0.96	0.96
yeast4	0.61	0.61	0.67	0.65	0.60	0.97	0.96	0.96	0.95	0.95
yeast1289vs7	0.55	0.53	0.56	0.59	0.62	0.97	0.97	0.95	0.94	0.94
yeast5	0.90	0.80	0.85	0.84	0.88	0.99	0.98	0.98	0.98	0.98
ecoli0137vs26	0.75	0.84	0.84	0.82	0.75	0.98	0.98	0.98	0.98	0.98
yeast6	0.75	0.70	0.78	0.73	0.78	0.98	0.98	0.97	0.97	0.97
Avg	**0.81**	0.77	**0.81**	0.80	0.80	**0.94**	0.93	0.93	0.92	0.93

Table 7
p-Values of different multiple comparisons tests for data in Table 6.

Test	p-value – AUC				
ANOVA	0.00023				
Friedman	0.0020				
	FURIA	LDA	1NN	NNET	C4.5
Bootstrap-B (10%)	[0.07, 0.3]	[0.005, 0.005]	[0.03, 0.03]	[0.4, 1]	[1, 1]
Test	p-value – Test error				
ANOVA	≈ 0				
Friedman	≈ 0				
	FURIA	LDA	1NN	NNET	C4.5
Bootstrap-B	[0.0001, 0.0001]	[1, 1]	[0.7, 0.7]	[0.0001, 0.03]	[0.8, 0.8]

Table 8
p-Values of post-hoc tests. Cases where the selection of the test influences the difference have been marked.

	FURIA vs. LDA	FURIA vs. 1NN	FURIA vs. NNET	FURIA vs. C4.5
	AUC			
ANOVA	0.01	0.98	0.24	**0.12**
Friedman	0.02	0.93	0.71	**0.01**
Bootstrap-B	[0.00, 0.02]	[0.50, 1.00]	[0.60, 1.00]	**[0.03, 0.40]**
	Test error			
ANOVA	0.01	0.01	≈ 0	≈ 0
Friedman	0.0004	0.002	≈ 0	≈ 0
Bootstrap-B	[0.0001, 0.0001]	[0.0001, 0.003]	[0.0001, 0.0001]	[0.0001, 0.0001]

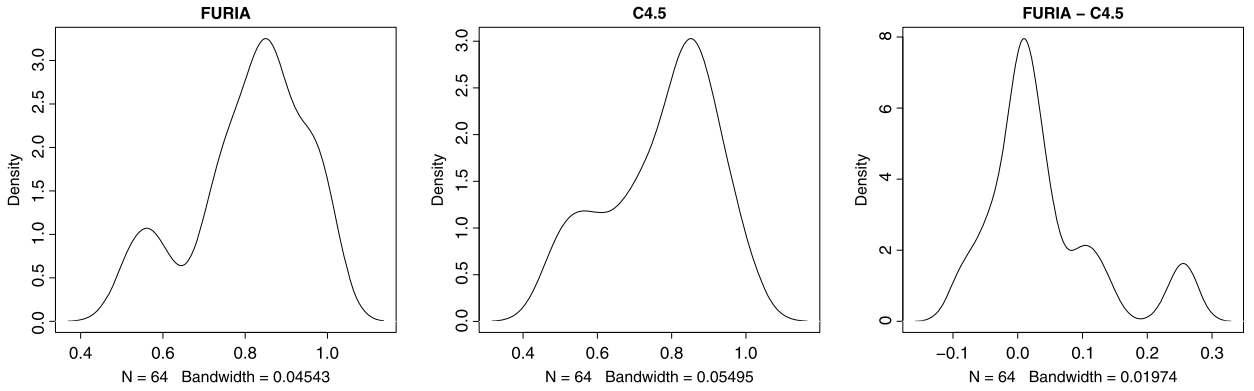

Fig. 4. Density function of the distribution of values of AUC of FURIA, C4.5 and their paired differences.

5. Concluding remarks and future work

In deterministic algorithms, the variability of the test error in the different folds of cross validation is originated on the random selection of the test sets. In stochastic algorithms, the chance that the algorithm converges to a suboptimal solution introduces a second source of uncertainty in the estimation of its performance, that cannot be properly accounted with a single factor repeated measures experimental design. In this study it is proposed that a confidence interval for a robust central tendency measure of the repetitions (median, mode, censored mean) is used instead of the mean when modeling the repetitions of a fold. A new interval-valued statistical test (Bootstrap-B) has been proposed, and it has been shown that in the presence of outliers, its power can be better than that of Friedman's test. In addition to this, in future works the following properties will be explored:

- The new test can be applied to learning algorithms that produce interval-valued estimations of the test error. Up to our knowledge, this is the first proposal of a mixed experimental design that allows for multiple comparisons between a combination of algorithms for scalar and interval-valued data.
- Incomplete tables of results can be tackled. Missing values in an experimentation could possibly be replaced by an interval spanning the range of errors.

A symmetric redefinition of the bootstrap post-hoc tests will also be considered in the future. Lastly, the use of a family of confidence intervals (a fuzzy set) for describing the variability attributable to repetitions of folds will be analyzed. This

representation might remove the need for determining the best width for the intervals with additional experiments, as was proposed in this contribution.

Acknowledgment

This work has been funded by Spanish Ministry of Economy and Competitiveness, grant TIN2011-24302.

References

[1] J. Alcalá, L. Sánchez, S. García, M.J. Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, KEEL: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2008) 307–318.
[2] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recogn. 30 (7) (1997) 1145–1159.
[3] I. Couso, L. Sánchez, Mark-recapture techniques in statistical tests for imprecise data, Internat. J. Approx. Reason. 52 (2) (2011) 240–260.
[4] I. Couso, L. Sánchez, P. Gil, Imprecise distribution function associated to a random set, Inform. Sci. 159 (1–2) (2004) 109–123.
[5] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[6] J. Derrac, S. Garcia, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm Evol. Comput. 1 (1) (2011) 3–18.
[7] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Amer. Statist. 37 (1) (1983) 36–48.
[8] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, L. Ginzburg, Experimental uncertainty estimation and statistics for data having interval uncertainty, Tech. Rep., Sandia National Laboratories, 2007.
[9] S. Garcia, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Inform. Sci. 180 (10) (2010) 2044–2064.
[10] S. Garcia, F. Herrera, An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons, J. Mach. Learn. Res. 9 (2008) 2677–2694.
[11] P.I. Good, Permutation, Parametric, and Bootstrap Tests of Hypotheses, Springer Ser. Statist., Springer-Verlag, 2004.
[12] S.O. Haykin, Neural Networks and Learning Machines, 3rd ed., Prentice Hall, 2008.
[13] J. Hühn, E. Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction, Data Min. Knowl. Discov. 19 (2009) 293–319.
[14] G.K. Kanji, 100 Statistical Tests, Sage Publications, Inc., 2006.
[15] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, vol. 14, 1995, pp. 1137–1145.
[16] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Amer. Statist. Assoc. 47 (260) (1952) 583–621.
[17] P. Larrañaga, B. Calvo, R. Santana, Machine learning in bioinformatics, Brief. Bioinform. 7 (1) (2006) 86–112.
[18] D.D. Margineantu, T.G. Dietterich, Bootstrap methods for the cost-sensitive evaluation of classifiers, in: ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 2000, pp. 582–590.
[19] I.S. Molchanov, Theory of Random Sets, Springer-Verlag, 2005.
[20] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., 1993.
[21] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 2008.
[22] S. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, Data Min. Knowl. Discov. 1 (3) (1997) 317–328.