Information Sciences 180 (2010) 2044–2064

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power

Salvador García [a,*], Alberto Fernández [b], Julián Luengo [b], Francisco Herrera [b]

[a] *Department of Computer Science, University of Jaén, Spain*
[b] *Department of Computer Science and Artificial Intelligence, University of Granada, Spain*

ARTICLE INFO

Article history:
Received 7 April 2009
Received in revised form 6 December 2009
Accepted 11 December 2009

Keywords:
Statistical analysis
Computational intelligence
Data mining
Nonparametric statistics
Multiple comparisons procedures
Genetics-based machine learning
Fuzzy classification systems

ABSTRACT

Experimental analysis of the performance of a proposed method is a crucial and necessary task in an investigation. In this paper, we focus on the use of nonparametric statistical inference for analyzing the results obtained in an experiment design in the field of computational intelligence. We present a case study which involves a set of techniques in classification tasks and we study a set of nonparametric procedures useful to analyze the behavior of a method with respect to a set of algorithms, such as the framework in which a new proposal is developed.

Particularly, we discuss some basic and advanced nonparametric approaches which improve the results offered by the Friedman test in some circumstances. A set of post hoc procedures for multiple comparisons is presented together with the computation of adjusted *p*-values. We also perform an experimental analysis for comparing their power, with the objective of detecting the advantages and disadvantages of the statistical tests described. We found that some aspects such as the number of algorithms, number of data sets and differences in performance offered by the control method are very influential in the statistical tests studied. Our final goal is to offer a complete guideline for the use of nonparametric statistical procedures for performing multiple comparisons in experimental studies.

1. Introduction

It is not possible to find one algorithm that is the best in behavior for all problems, as the “no free lunch” theorem suggests [50,51]. On the other hand, we know that we have available several degrees of knowledge associated with the problem, which we expect to solve, and there are clear differences when working on the problem without knowledge and having partial knowledge about it. This knowledge allows us to design algorithms with specific properties that can make them more suitable to the solution of the problem. Having the previous premise in mind, the question about deciding when an algorithm is better than another one is suggested. This question has given rise to the growing interest in the analysis of experiments in the field of computational intelligence (CI) [15] or the field of data mining (DM) [24,45]. This interest has brought in the use of statistical inference in the analysis of empirical results obtained by the algorithms. Inferential statistics show how well a

[*] Corresponding author. Tel.: +34 953 212802; fax: +34 953 212274.
E-mail addresses: sglopez@ujaen.es (S. García), alberto@decsai.ugr.es (A. Fernández), julianlm@decsai.ugr.es (J. Luengo), herrera@decsai.ugr.es (F. Herrera).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.12.010

sample of results supports a certain hypothesis and whether the conclusions achieved can be generalized beyond what was tested.

In some recent papers, the researchers have used statistical techniques to contrast the results offered by their proposals [33,37,46,48,53]. Due to the fact that statistical analysis is highly demanded in any research work, we can find recent studies that propose some methods for conducting comparisons among various approaches [11,12,22,43]. There are two main types of statistical test in the literature: parametric tests and nonparametric tests. The decision to use the former or the latter may depend on the properties of the sample of results to be analyzed. A parametric statistical test assumes that data comes from a type of probability distribution and makes inferences about the parameters of the distribution. For example, the use of the ANOVA test is only appropriate when the sample of results fulfills three required conditions: independency, normality and homoscedasticity [42,54]. In fact, if the assumptions required for a parametric test hold, the parametric test should always be preferred over a nonparametric one, in that it will have a lower Type I error and higher power. However, some studies involving CI algorithms in experimental comparisons show that these conditions are not easy to meet [21,23,47].

The analysis of results can be done following either one of two alternatives: single-problem analysis and multiple-problem analysis. The first one corresponds to the study of the performance of several algorithms over a unique problem case. The second one would suppose the study of several algorithms over more than one problem case simultaneously, assimilating the fact that each problem has a degree of difficulty and that the results obtained among different problems are not comparable. The single-problem analysis is well-known and is usually found in specialized literature [12]. Although the required conditions for using parametric statistics are not usually checked, a parametric statistical study could obtain similar conclusions to a nonparametric one. However, in a multiple-problem analysis, a parametric test may reach erroneous conclusions [11].

On the other hand, a distinction between pairwise and multiple comparison tests is necessary. The former are valid procedures to compare two algorithms and the latter should be used when comparing more than two methods. The main reason that distinguishes both kinds of test is related to the control of the family wise error, which is the probability of making one or more false discoveries (Type I errors) [42]. Intended pairwise tests, such as the Wilcoxon test [11], do not control the error propagation of making more than one comparison and they should not be used in multiple comparisons. If a researcher plans to make multiple comparisons using several statistical inferences simultaneously, then he/she has to account for the multiplicative effect in order to control the Family Wise Error Rate (FWER) [42]. Demšar [11] described a set of nonparametric test for performing multiple comparisons and he analyzed them in contrast to well-known parametric tests in terms of power, obtaining that the nonparametric tests are more suitable for use. He explained the Friedman test [18], the Iman–Davenport correction [30] and some post hoc procedures, such as Bonferroni–Dunn [14], Holm [28], Hochberg [25] and Hommel [29].

In this paper, we extend the set of nonparametric procedures for performing multiple statistical comparisons between more than two algorithms and we focus on the case in which a control treatment is compared against other treatments. In other words, we focus on the usual case in which a new CI or DM algorithm is proposed and the researcher is interested in comparing it to other similar approaches. Basic and advanced techniques for studying the differences among methods belonging to multiple comparisons will be described. The choice of the set of computational intelligence algorithms depends on their heterogeneity and performance obtained. This paper can be seen as a tutorial on the use of more advanced nonparametric tests and the case studies used require results provided by algorithms which present low and high degrees of differences among themselves. With respect to the choice of the tests, we have considered those that are not excessively complicated and well-known in statistics (although they are considered advanced procedures, all of them can be found in statistical books. However, they are almost unknown among non-statisticians). There are many other procedures similar to the ones described in this paper, but they do not offer significant differences with respect to the procedures already presented by Demšar [11] and in this paper. Thus, the choice of the tests may be influenced by a trade-off between their complexity and their differences in experimental power, taking into account that they are well-known in the statistics community.

Specifically, the paper will be focused on the following main topics:

- To present new nonparametric techniques which allow different types of comparison between various algorithms. Within this topic, the Multiple Sign-test [44] and the Contrast Estimation based on medians [13] will be introduced. The first is a basic procedure to conduct rapid comparison considering a control method. The second allows us to compute differences in performance based on medians among a set of algorithms.
- Two alternatives to the Friedman test will be discussed: The Friedman Aligned Ranks [26] and the Quade test [38]. They differ in the ranking computation procedure and they can offer better results depending on the characteristics of the experimental study considered.
- To extend the post hoc procedures described in [11] with the inclusion of four new procedures: Holland [27], Rom [41], Finner [17] and Li [34]. The computation of their adjusted *p*-values (APVs) will be included.
- To carry out an experimental analysis to estimate the power of all the procedures presented. It will be focused on detecting the advantages and inconveniences of each procedure, as well as to present a useful guideline for their use.

Fig. 1 schematizes the tests and procedures that are the object of study in this paper. Throughout the paper, all the procedures described will be illustrated by means of examples defined over a DM task of classification using CI techniques. Thus,

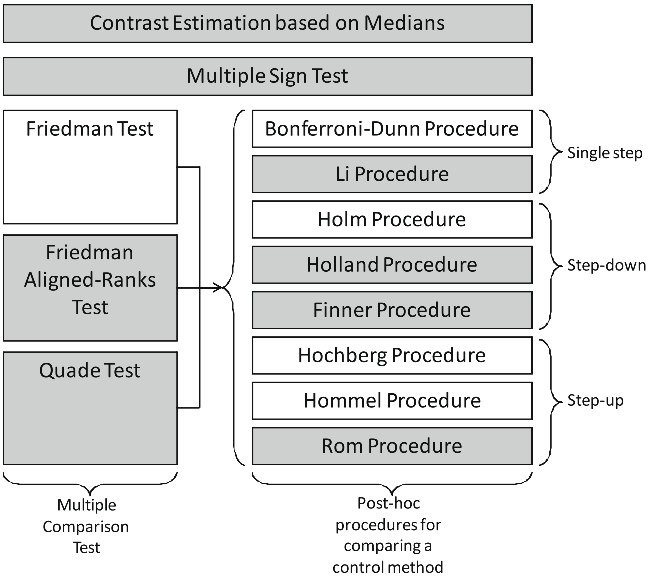


Fig. 1. Summarizing chart of procedures reviewed and presented in this paper. The procedures studied in [11] are in white and new ones are in grey.

several classifiers in a multiple-problem analysis will be compared by using the procedures presented in an experimental study.

In order to do that, the paper is organized as follows. In Section 2, we describe the set up of the experimental study: algorithms, data sets and parameters. Section 3 presents the basic nonparametric procedures and demonstrates their use in the experimental study. In Section 4, the two mentioned alternatives to the Friedman test are described. Section 5 enumerates a set of post hoc procedures suitable for detecting pairwise differences between two algorithms within a multiple comparison test. We carry out an experimental analysis in Section 6 to estimate the power and usefulness of the advanced nonparametric tests and post hoc procedures. Some criticisms and guidelines are given in Section 7. Finally, Section 8 concludes the paper. A URL of the software package which computes all the tests explained in this paper and the statistical table needed for the Multiple Sign-test is given in Appendix A.

2. Experimental framework

This section defines the set up of the experimental study. The classification data sets, validation and parameters are provided. We need to specify the experimental conditions used in this paper with respect to the parameters adopted by the algorithms, validation procedure used and classification data sets employed.

We have used 48 data sets,[1] which are specified in Table 1. For each data set, the name, number of examples, number of attributes (numeric and nominal) and number of classes are given. Some of them (24 data sets) have been used in the experimental study considered in each section, whereas all data sets have been used in the experimental analysis of power in Section 6. Data sets are from the UCI machine learning repository [3], all of them valid for classification tasks.

The algorithms used for our study and their parameters are specified in Table 2. Furthermore, these methods are included in the KEEL[2] software [2]. We have used 10-fold cross-validation and we have run the algorithms three times in order to obtain a sample of 30 results, which have been averaged, for each data set. Thus, the results, which will be further analyzed by the statistical techniques, correspond to average accuracies in test data.

Note that, throughout the paper, the experimental study that will be used as an example in the description of the statistical tests does not consider the complete set of classifiers and data sets. All of them will be used in the experimental analysis of power conducted in Section 6.

3. Basic nonparametric tests for performing multiple comparisons: Friedman test, Multiple Sign-test and Contrast Estimation based on medians

Frequently in CI, we are interested in detecting groups of differences among a set of results provided by various algorithms. In statistics, these groups are called *blocks* and they are usually associated with the problems met in the experimental

[1] Data sets marked with '*' have been sampled for being adapted to slow algorithms, such as Ant-Miner+.

[2] http://www.keel.es.

Table 1
Summary description for classification data sets.

Data set	# Ex.	# Atts.	# Num.	# Nom.	# Cl.	Data set	# Ex.	# Atts.	# Num.	# Nom.	# Cl.
Abalone*	418	8	7	1	28	Monks	432	6	6	0	2
Adult*	624	14	6	8	2	Mushrooms*	813	22	0	22	2
Australian	690	14	8	6	2	Newthyroid	215	5	5	0	3
Balance	625	4	4	0	3	Nursery*	1296	8	0	8	5
Breast	286	9	0	9	2	Pageblocks*	548	10	10	0	5
Bupa	345	6	6	0	2	Penbased	1099	16	16	0	10
Car	1728	6	0	6	4	Pima	768	8	8	0	2
Cleveland	297	13	13	0	5	Optdigits	1912	63	63	0	10
Contraceptive	1473	9	6	3	3	Ring	740	20	20	0	2
Crx	125	15	6	9	2	Satimage*	643	36	36	0	7
Dermatology	366	33	1	32	6	Segment	2310	19	19	0	7
Ecoli	336	7	7	0	8	Spambase*	460	55	55	0	2
Flare	1389	10	0	10	6	Splice*	319	60	0	60	3
German	1000	20	6	14	2	Shuttle*	2175	9	9	0	7
Glass	214	9	9	0	7	Tae	151	5	5	0	3
Haberman	306	3	3	0	2	Tic-tac-toe	958	9	0	9	2
Hayes-roth	133	4	4	0	3	Thyroid*	720	21	6	15	3
Heart	270	13	6	7	2	Twonorm*	740	20	20	0	2
Ion	351	34	34	0	2	Vehicle	846	18	18	0	4
Iris	150	4	4	0	3	Vowel	990	13	11	2	11
Led7digit	500	7	0	1	10	Wine	178	13	13	0	3
Letter*	2000	16	16	0	26	Wisconsin	683	9	9	0	2
Lymphography	148	18	3	15	4	Yeast	1484	8	8	0	10
Magic*	1902	10	10	0	2	Zoo	101	17	0	17	7

Table 2
Parameter specification for the algorithms employed in the experimentation.

PDFC [8]	C = 100.0, d = 0.25, tolerance = 0.001, epsilon = 1.0E−12, PDRFtype = Gaussian
NNEP [35]	Hidden nodes = 4, Transfer = Product unit, Generations = 200
IS-CHC+1-NN [6,20,16]	Population size = 50, Number of evaluations = 10,000, α equilibrate factor = 0.5 Percent of change in restart = 0.35, 0–1 probability in restart = 0.25 0–1 probability in diverge = 0.05, Number of neighbours = 1, Distance function = Euclidean
FH-GBML [31]	Number of fuzzy rules = 20, Number of rule sets = 200, Crossover probability = 0.9, Mutation probability = 1/d (d = dimensionality), Number of replaced rules: All rules except the best-one (Pittsburgh-part, elitist approach), number of rules/5 (Michigan-part), Number of generations = 1000 generations, Don't care probability = 0.5, Probability of the application of the Michigan iteration = 0.5
GASSIST-ADI [4,5]	Threshold in Hierarchical Selection = 0, Iteration of Activation for Rule Deletion Operator = 5, Iteration of Activation for Hierarchical Selection = 24, Minimum Number of Rules before Disabling the Deletion Operator = 12, Minimum Number of Rules before Disabling the Size Penalty Operator = 4, Number of Iterations = 750, Initial Number of Rules = 20, Population Size = 400, Crossover Probability = 0.6, Probability of Individual Mutation = 0.6, Probability of Value 1 in Initialization = 0.90, Tournament Size = 3, Possible size in *micro-intervals* of an attribute = {4,5,6,7,8,10,15,20,25}, Maximum Number of Intervals per Attribute = 5, $p_{split} = 0.05,\ p_{merge} = 0.05$, Probability of Reinitialize Begin = 0.03, Probability of Reinitialize End = 0, Use MDL = true, Iteration MDL = 25, Initial Theory Length Ratio = 0.075, Weight Relaxation Factor = 0.90, Class Initialization Method = cwinit, Default Class = auto,
DT-GA [7]	Confidence = 0.25, Instances per leaf = 2, Threshold S to consider a Small Disjunct = 10, Number of Gen. = 50, Number of chrom. = 200, Crossover Prob. = 0.8, Mutation Prob. = 0.01
Ant-Miner+ [36]	Number of ants = 3000, Max number of uncovered examples = 10 Min number of samples by rule = 10, Max iterations without converge = 10 Max pheromone = 0.999, Min pheromone = 0.1, $\alpha = 1,\ \beta = 2$
EvRBF [40]	Neurons rate = 0.1, Variation Rate = 0.15, Population size = 100, Tournament size = 2 Replacement rate = 0.1, Max generations = 100, Crossover rate = 0.9, Mutator rate = 0.1

study. For example, in a multiple data set comparison of classification, each block corresponds to the results offered over a specific data set. When referring to multiple comparisons tests, a block is composed of three or more subjects or results, each one corresponding to the performance evaluation of the algorithm over the data set.

In nonparametric statistics, the most well-known procedure for testing the differences between more than two related samples is the Friedman test. The related samples in classification are the performances of the methods measured across the same data sets. The Friedman test will be introduced in Section 3.1. It considers that the null hypothesis being tested is that all methods obtain similar results with nonsignificant differences. If we want to obtain what classifiers are better/

worse than our proposal, we have to use a post hoc procedure, which will be described in Section 5. The Friedman test could be bypassed and directly applied to a post hoc procedure, because the latter only requires that the final rankings are computed. However, from a statistical point of view, it is more correct to conduct the Friedman first and after the post hoc procedure. We have to take into account that the rankings are the required data to conduct post hoc pairwise tests and they are obtained as part of the Friedman procedure, so it is wiser to conduct it fully and complement it after if needed.

There is a rapid procedure, not very powerful but easy to use, to compare all classifiers with a control that will be described in Section 3.2. Finally, the researcher frequently wishes to estimate the differences between two classifiers over multiple data sets. A procedure for estimating the contrast between classifiers medians will be discussed in Section 3.3.

3.1. Friedman test and Iman–Davenport extension

The Friedman test [18,19] (Friedman two-way analysis of variances by ranks) is a nonparametric analogue of the parametric two-way analysis of variance. The objective of this test is to determine if we may conclude from a sample of results that there is difference among treatment effects. The first step in calculating the test statistic is to convert the original results to ranks. Thus, it ranks the algorithms for each problem separately, the best performing algorithm should have the rank of 1, the second best rank 2, etc., as shown in Table 3. In case of ties, average ranks are computed.

Let r_i^j be the rank of the jth of k algorithms on the ith of n data sets. The Friedman test needs the computation of the average ranks of algorithms, $R_j = \frac{1}{n}\sum_i r_i^j$. Under the null hypothesis, which states that all the algorithms behave similarly and thus their ranks R_j should be equal, the Friedman statistic

$$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{1}$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, when n and k are big enough (as a rule of a thumb, $n > 10$ and $k > 5$). For a smaller number of algorithms and data sets, exact critical values have been computed [42,54].

In [30], Iman and Davenport showed that Friedmans χ_F^2 presents a conservative behavior and proposed a better statistic

$$F_F = \frac{(n-1)\chi_F^2}{n(k-1)\chi_F^2} \tag{2}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. See Table A10 in [42] to find the critical values. Furthermore, the p-value could be computed through normal approximations [1]. If the null hypothesis is rejected, we can proceed with a post hoc test, which will be explained in Section 5.

Table 3
Comparison of accuracy among the four algorithms selected in the experimental study. The ranks in the parentheses are used in the computation of the Friedman test.

Data set	PDFC	NNEP	IS-CHC + 1NN	FH-GBML
Adult*	0.752 (4)	0.773 (3)	0.785 (2)	0.795 (1)
Breast	0.727 (2)	0.748 (1)	0.724 (3)	0.713 (4)
Bupa	0.736 (1)	0.716 (2)	0.585 (4)	0.638 (3)
Car	0.994 (1)	0.861 (3)	0.880 (2)	0.791 (4)
Cleveland	0.508 (4)	0.553 (2)	0.575 (1)	0.515 (3)
Contraceptive	0.535 (2)	0.536 (1)	0.513 (3)	0.471 (4)
Dermatology	0.967 (1)	0.871 (3)	0.954 (2)	0.532 (4)
Ecoli	0.831 (1)	0.807 (3)	0.819 (2)	0.768 (4)
German	0.745 (1)	0.702 (4)	0.719 (2)	0.705 (3)
Glass	0.709 (1)	0.572 (4)	0.669 (2)	0.607 (3)
Haberman	0.722 (4)	0.728 (2)	0.725 (3)	0.732 (1)
Iris	0.967 (1)	0.947 (4)	0.953 (3)	0.960 (2)
Lymphography	0.832 (1)	0.752 (3)	0.802 (2)	0.691 (4)
Mushrooms*	0.998 (1)	0.992 (2)	0.482 (4)	0.910 (3)
Newthyroid	0.963 (1.5)	0.963 (1.5)	0.954 (3)	0.926 (4)
Penbased*	0.982 (1)	0.953 (2)	0.932 (3)	0.630 (4)
Ring*	0.978 (1)	0.773 (4)	0.834 (3)	0.849 (2)
Satimage*	0.854 (1)	0.787 (3)	0.841 (2)	0.779 (4)
Shuttle*	0.965 (3)	0.984 (2)	0.995 (1)	0.947 (4)
spambase*	0.924 (1)	0.887 (2)	0.861 (3)	0.804 (4)
Thyroid*	0.929 (3)	0.942 (1)	0.931 (2)	0.921 (4)
Vehicle	0.837 (1)	0.643 (2)	0.602 (3)	0.554 (4)
Wine	0.972 (1)	0.956 (2)	0.944 (3)	0.922 (4)
Wisconsin	0.958 (4)	0.959 (3)	0.964 (1.5)	0.964 (1.5)
Average rank	1.771	2.479	2.479	3.271

The procedure is illustrated by the data from Table 3, which compares the four algorithms considered as the global experimental study in this paper: PDFC, NNEP, IS-CHC + 1NN and FH-GBML. Average ranks by themselves provide a useful comparison of the algorithms. On average, PDFC ranked the first with rank 1.771; NNEP and IS-CHC + 1NN ranked the second and the third, with equal ranks 2.479; and the last is FH-GBML with rank 3.271. The Friedman test proves whether the measured average ranks are significantly different from the mean rank $R_j = 2.5$ expected under the null hypothesis:

$$\text{(Friedman)} \quad \chi_F^2 = \frac{12 \cdot 24}{4 \cdot 5}\left[(1.771^2 + 2.479^2 + 2.479^2 + 3.271^2) - \frac{4 \cdot 5^2}{4}\right] = 16.225 \tag{3}$$

$$\text{(Iman—Davenport)} \quad F_F = \frac{23 \cdot 16.225}{24 \cdot 3 - 16.225} = 6.691 \tag{4}$$

With four algorithms and 24 data sets, F_F is distributed according to the F distribution with $4 - 1 = 3$ and $(4 - 1) \cdot (24 - 1) = 69$ degrees of freedom. The p-value computed by using the $F(3, 69)$ distribution is 4.97×10^{-4}, so the null hypothesis is rejected at a high level of significance.

3.2. Multiple Sign-test

As discussed at the beginning of this section, it would be useful to determine which of the other algorithms exhibit a different performance from the control one.

The following procedure, proposed in [39,44], allows us to compare all of the other algorithms with a control labeled algorithm 1. The technique, an extension of the familiar sign test [11], carries out the following steps:

1. Represent by x_{i1} and x_{ij} the performances of the control and the jth classifier in the ith data set.
2. Compute the signed differences $d_{ij} = x_{ij} - x_{i1}$. In other words, pair each performance with the control and, in each data set, subtract the control performance from the jth classifier.
3. Let r_j equal the number of differences, d_{ij}, that have the less frequently occurring sign (either positive or negative) within a pairing of an algorithm with the control.
4. Let M_1 be the median response of a sample of results of the control method and M_j be the median response of a sample of results of the jth algorithm. Apply one of the following decision rules:
   - For testing $H_0 : M_j \geqslant M_1$ against $H_1 : M_j < M_1$, reject H_0 if the number of plus signs is less than or equal to the critical value of R_j appearing in Table A.1 in Appendix A for $k - 1$ (number of algorithms excluding control), n and the chosen experimentwise error rate.
   - For testing $H_0 : M_j \leqslant M_1$ against $H_1 : M_j > M_1$, reject H_0 if the number of minus signs is less than or equal to the critical value of R_j appearing in Table A.1 for $k - 1,\ n$ and the chosen experimentwise error rate.

The technique is illustrated by means of the global experimental study considered in the paper. Table 4 depicts the computation of this procedure.

Suppose we choose a level of significance $\alpha = 0.05$ and let our hypotheses be $H_0 : M_j \geqslant M_0$ and $H_1 : M_j < M_0$; that is, our control algorithm PDFC is better than the remaining classifiers. Reference to Table A.1 for $(k - 1) = 3$ and $n = 24$ reveals that the critical value of r_j is 6. Since the number of pluses in the pairwise comparison between the control and IS-CHC + 1NN and FH-GBML is less than 6, then PDFC has a better performance than them. However, the null hypothesis cannot be rejected in the pairwise comparison among PDFC and NNEP, so we conclude that they perform similarly.

3.3. Contrast Estimation based on medians

Using the data resulting from the run of various classifiers over multiple data sets in an experiment, the researcher could be interested in the estimation of the difference between two classifiers' performance. A procedure for this purpose has been proposed in [13] and it assumes that the expected differences between performances of algorithms are the same across data sets. We assume that the performance is reflected by the magnitudes of the differences between the performances of the algorithms. Consequently, we are interested in estimating the contrast between medians of samples of results considering all pairwise comparisons. It obtains a quantitative difference computed through medians between two algorithms over multiple data sets, but the value obtained will change when using other data sets in the experiment.

We can proceed as follows:

1. For every pair of k algorithms in the experiment, we compute the difference between the performances of the two algorithms in each of the n data sets. In other words, we compute the differences

   $$D_{i(uv)} = x_{iu} - x_{iv}$$

   where $i = 1, \ldots, n;\ u = 1, \ldots, k$; and $v = 1, \ldots, k$. We form performance pairs only for those in which $u < v$.
2. We find the median of each set of differences and call it Z_{uv}. We call Z_{uv} the *unadjusted estimator* of $M_u - M_v$. Since $Z_{vu} = Z_{uv}$, we have only to calculate Z_{uv} for the case where $u < v$. There are $k(k - 1)/2$ of these medians. Also note that $Z_{uu} = 0$.

Table 4
Comparison of accuracy between the control algorithm PDFC and the rest of algorithms selected in the experimental study. The signs in the parentheses are used in the computation of the Multiple Comparison Sign test.

Data set	PDFC	NNEP	IS-CHC + 1NN	FH-GBML
	1 (Control)	2	3	4
Adult*	0.752	0.773 (+)	0.785 (+)	0.795 (+)
Breast	0.727	0.748 (+)	0.724 (−)	0.713 (−)
Bupa	0.736	0.716 (−)	0.585 (−)	0.638 (−)
Car	0.994	0.861 (−)	0.880 (−)	0.791 (−)
Cleveland	0.508	0.553 (−)	0.575 (+)	0.515 (+)
Contraceptive	0.535	0.536 (+)	0.513 (−)	0.471 (−)
Dermatology	0.967	0.871 (−)	0.954 (−)	0.532 (−)
Ecoli	0.831	0.807 (−)	0.819 (−)	0.768 (−)
German	0.745	0.702 (−)	0.719 (−)	0.705 (−)
Glass	0.709	0.572 (−)	0.669 (−)	0.607 (−)
Haberman	0.722	0.728 (+)	0.725 (+)	0.732 (+)
Iris	0.967	0.947 (−)	0.953 (−)	0.960 (−)
Lymphography	0.832	0.752 (−)	0.802 (−)	0.691 (−)
Mushrooms*	0.998	0.992 (−)	0.482 (−)	0.910 (−)
Newthyroid	0.963	0.963 (=)	0.954 (−)	0.926 (−)
Penbased*	0.982	0.953 (−)	0.932 (−)	0.630 (−)
Ring*	0.978	0.773 (−)	0.834 (−)	0.849 (−)
Satimage**	0.854	0.787 (−)	0.841 (−)	0.779 (−)
Shuttle*	0.965	0.984 (+)	0.995 (+)	0.947 (−)
Spambase*	0.924	0.887 (−)	0.861 (−)	0.804 (−)
Thyroid*	0.929	0.942 (+)	0.931 (+)	0.921 (−)
Vehicle	0.837	0.643 (−)	0.602 (−)	0.554 (−)
Wine	0.972	0.956 (−)	0.944 (−)	0.922 (−)
Wisconsin	0.958	0.959 (+)	0.964 (+)	0.964 (+)
Number of minuses		16	18	20
Number of pluses		7	6	4
r_j		7	6	4

Table 5
Differences between pairs of performances in each data set for different pairs of algorithms.

Data set	$D_{i(12)}$	$D_{i(13)}$	$D_{i(14)}$	$D_{i(23)}$	$D_{i(24)}$	$D_{i(34)}$
Adult*	−0.021	−0.033	−0.043	−0.012	−0.022	−0.010
Breast	−0.021	0.003	0.014	0.024	0.035	0.011
Bupa	0.020	0.151	0.099	0.131	0.078	-0.053
Car	0.133	0.114	0.203	−0.019	0.071	0.089
Cleveland	−0.045	−0.067	−0.007	−0.021	0.039	0.060
Contraceptive	−0.001	0.022	0.064	0.023	0.065	0.042
Dermatology	0.096	0.014	0.436	−0.083	0.339	0.422
Ecoli	0.024	0.012	0.063	−0.012	0.039	0.051
German	0.043	0.026	0.040	−0.017	−0.003	0.014
Glass	0.137	0.040	0.101	−0.097	−0.036	0.062
Haberman	−0.006	−0.003	−0.010	0.004	−0.003	−0.007
Iris	0.020	0.013	0.007	−0.007	−0.013	−0.007
Lymphography	0.080	0.031	0.141	−0.049	0.061	0.110
Mushrooms*	0.006	0.515	0.087	0.509	0.081	−0.428
Newthyroid	0.000	0.010	0.038	0.010	0.038	0.028
Penbased*	0.029	0.049	0.352	0.020	0.323	0.302
Ring*	0.205	0.145	0.130	−0.061	−0.076	−0.015
Satimage*	0.067	0.012	0.075	−0.054	0.008	0.062
Shuttle*	−0.019	−0.030	0.018	−0.011	0.038	0.048
Spambase*	0.037	0.063	0.120	0.026	0.083	0.057
Thyroid*	−0.013	−0.001	0.008	0.011	0.021	0.010
Vehicle	0.194	0.235	0.282	0.041	0.089	0.047
Wine	0.016	0.028	0.050	0.011	0.034	0.023
Wisconsin	−0.001	−0.006	−0.006	−0.005	−0.005	0.000

3. We compute the mean of each set of unadjusted medians having the same first subscript and call the result m_u; that is, we compute

$$m_u = \frac{\sum_{j=1}^{k} Z_{uj}}{k}, \quad u = 1,\ldots,k$$

Table 6
Contrast Estimation based on medians among all algorithms of the experimental study.

	PDFC	NNEP	IS-CHC + 1NN	FH-GBML
PDFC	0.00000	0.02257	0.01976	0.05955
NNEP	−0.02257	0.00000	−0.00281	0.03698
IS-CHC − 1NN	−0.01976	0.00281	0.00000	0.03979
FH-GBML	−0.05955	−0.03698	−0.03979	0.00000

4. The estimator of $M_u - M_v$ is $m_u - m_v$, where u and v range from 1 through k. For example, the difference between M_1 and M_2 is $m_1 - m_2$.

The procedure is illustrated by using the global experimental study considered in this paper. From the performances of the four classifiers considered we compute, for each pair of algorithms, the differences of performances in each data set. Table 5 shows this computation.

From the differences in Table 5, we find that the six medians are $Z_{12} = 0.02,\ Z_{13} = 0.018,\ Z_{14} = 0.064,\ Z_{23} = -0.006,\ Z_{24} = 0.038$ and $Z_{34} = 0.035$. We now compute the following averages for M_1 and M_2:

$$m_1 = \frac{0 + 0.02 + 0.018 + 0.064}{4} = 0.026 \tag{5}$$

$$m_2 = \frac{-0.02 + 0 + (-0.006) + 0.038}{4} = 0.003 \tag{6}$$

Our estimate of $M_1 - M_2$ is $m_1 - m_2 = 0.026 - 0.003 = 0.023$. In other words, the difference in accuracy between PDFC and NNEP classifiers estimated over the median in multiple data sets is equal to 0.023. Table 6 shows all the estimators among the four algorithms.

As we can see in Table 6, PDFC method always obtains a positive difference value with respect to the other three methods compared, indicating to us that it is the best performing method. FH-GBML is the worst one and IS-CHC − 1NN can be considered better than NNEP. These conclusions are drawn from the results offered by the table, but they cannot be considered definitive. This procedure is especially useful to estimate how far a method in performance is over other, but it cannot provide a probability of error associated with the rejection of the null hypothesis.

4. Advanced nonparametric tests for performing multiple comparisons: Friedman Aligned Ranks and the test of Quade

As we have seen before, nonparametric statistics can be used over real data through ranking the data. This transformation to ranks can be made in different ways; i.e. the Friedman test uses sets of ranks whose treatments are ranked separately in each data set. In this section, we explain two modifications to improve, in certain circumstances, the application of the Friedman test in an experimental analysis. The first one is the use of aligned ranks, which will be described in Section 4.1, whereas in Section 4.2 we will present another alternative based on weighted rankings, the Quade test.

4.1. Friedman Aligned Ranks

The Friedman test is based on n sets of ranks, one set for each data set in our case; and the performances of the algorithms analyzed are ranked separately for each data set. Such a ranking scheme allows for intra-set comparisons only, since inter-set comparisons are not meaningful. When the number of algorithms for comparison is small, this may pose a disadvantage. In such cases, comparability among data sets is desirable and we can employ the method of aligned ranks [26].

In this technique, a value of location is computed as the average performance achieved by all algorithms in each data set. Then, it calculates the difference between the performance obtained by an algorithm and the value of location. This step is repeated for algorithms and data sets. The resulting differences, called aligned observations, which keep their identities with respect to the data set and the combination of algorithms to which they belong, are then ranked from 1 to kn relative to each other. Then, the ranking scheme is the same as that employed by a multiple comparison procedure which employs independent samples; such as the Kruskal–Wallis test [32]. The ranks assigned to the aligned observations are called aligned ranks.

The Friedman Aligned Ranks test statistic can be written as

$$T = \frac{(k-1)\left[\sum_{j=1}^{k}\widehat{R}_{.j}^2 - (kn^2/4)(kn+1)^2\right]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n}\widehat{R}_{i.}^2} \tag{7}$$

where $\widehat{R}_{i.}$ is equal to the rank total of the ith data set and $\widehat{R}_{.j}$ is the rank total of the jth algorithm.

The test statistic T is compared for significance with a chi-square distribution for $k-1$ degrees of freedom. Critical values can be found at Table A3 in [42]. Furthermore, the p-value could be computed through normal approximations [1]. If the null hypothesis is rejected, we can proceed with a post hoc test. We illustrate the Friedman Aligned Ranks technique by means of the global experimental study considered in this paper. Table 7 displays the aligned observations and the aligned ranks in the parentheses considering the known four algorithms and 24 data sets.

Again, average ranks by themselves provide a fair comparison of the algorithms. On average, PDFC ranked the first with rank 29.313; NNEP and IS-CHC + 1NN ranked the second and the third, with ranks 46.729 and 47.063, respectively; and the last is FH-GBML with rank 70.896. The Friedman Aligned Rank test checks whether the measured sum of aligned ranks are significantly different from the total aligned rank $\widehat{R}_j = 1164$ expected under the null hypothesis:

$$\sum_{j=1}^{k} \widehat{R}_{.j}^2 = 703.5^2 + 1121^2 + 1129.5^2 + 1701.5^2 = 5,923,547 \quad (8)$$

$$\sum_{i=1}^{n} \widehat{R}_{i.}^2 = 199^2 + 207^2 + 198^2 + \cdots + 199^2 = 926,830 \quad (9)$$

$$T = \frac{(4-1)[5,923,547 - (4\cdot 24^2/4)(4\cdot 24+1)^2]}{\{[4\cdot 24(4\cdot 24+1)(2\cdot 4\cdot 24+1)]/6\} - (1/4)\cdot 926,830} = 18.837 \quad (10)$$

With four algorithms and 24 data sets, T is distributed according to the chi-square distribution with $4 - 1 = 3$ degrees of freedom. The p-value computed by using the $\chi^2(3)$ distribution is 2.96×10^{-4}, so the null hypothesis is rejected at a high level of significance.

4.2. Quade test

The Friedman test considers all data sets to be equal in terms of importance. An alternative to this could take into account the fact that some data sets are more difficult or the differences registered on the run of various algorithms over them are larger. The rankings computed on each data set could be scaled depending on the differences observed in the algorithms' performances. The Quade test conducts a weighted ranking analysis of the sample of results [38].

The procedure starts finding the ranks r_i^j in the same way as the Friedman test does. The next step requires the original values of performance of the classifiers x_{ij}. Ranks are assigned to the data sets themselves according to the size of the sample range in each data set. The sample range within data set i is the difference between the largest and the smallest observations within that data set:

$$\text{Range in data set}: \ i = \max_j\{x_{ij}\} - \min_j\{x_{ij}\} \quad (11)$$

Obviously, there are n sample ranges, one for each data set. Assign rank 1 to the data set with the smallest range, rank 2 to the second smallest, and so on to the data set with the largest range, which gets rank n. Use average ranks in case of ties. Let $Q_1, Q_2, \ldots, Q_n$ be the ranks assigned to data sets $1, 2, \ldots, n$, respectively.

Table 7
Aligned observations of the four algorithms selected in the experimental study. The ranks in the parentheses are used in the computation of the Friedman Aligned Ranks test.

Data set	PDFC	NNEP	IS-CHC – 1NN	FH-GBML
Adult*	–0.024 (74)	–0.003 (56)	0.009 (39)	0.019 (30)
Breast	–0.001 (51)	0.020 (29)	–0.004 (59)	–0.015 (68)
Bupa	0.068 (11)	0.047 (16)	–0.084 (90)	–0.031 (81)
Car	0.112 (7)	–0.020 (72)	–0.002 (53)	–0.091 (92)
Cleveland	–0.030 (80)	0.016 (32)	0.037 (19)	–0.023 (73)
Contraceptive	0.022 (28)	0.022 (26)	–0.001 (50)	–0.043 (85)
Dermatology	0.136 (4)	0.040 (17)	0.123 (5)	–0.299 (95)
Ecoli	0.025 (24)	0.001 (48)	0.013 (33)	–0.038 (84)
German	0.027 (22)	–0.016 (69)	0.001 (47)	–0.013 (67)
Glass	0.069 (10)	–0.068 (88)	0.030 (21)	–0.032 (82)
Haberman	–0.005 (61)	0.002 (46)	–0.002 (54)	0.005 (41)
Iris	0.010 (38)	–0.010 (66)	–0.003 (58)	0.003 (42)
Lymphography	0.063 (13)	–0.017 (71)	0.032 (20)	–0.078 (89)
Mushrooms*	0.152 (2)	0.146 (3)	–0.363 (96)	0.065 (12)
Newthyroid	0.012 (34.5)	0.012 (34.5)	0.002 (45)	–0.026 (76)
Penbased*	0.108 (8)	0.078 (9)	0.058 (14)	–0.244 (94)
Ring*	0.120 (6)	–0.085 (91)	–0.025 (75)	–0.010 (65)
Satimage*	0.038 (18)	–0.028 (79)	0.026 (23)	–0.036 (83)
Shuttle*	–0.008 (62)	0.012 (36)	0.022 (27)	–0.026 (77)
Spambase*	0.055 (15)	0.018 (31)	–0.008 (63)	–0.065 (87)
Thyroid*	–0.001 (52)	0.011 (37)	0.000 (49)	–0.010 (64)
Vehicle	0.178 (1)	–0.016 (70)	–0.057 (86)	–0.105 (93)
Wine	0.024 (25)	0.007 (40)	–0.004 (60)	–0.027 (78)
Wisconsin	–0.003 (57)	–0.002 (55)	0.003 (43.5)	0.003 (43.5)
Total	703.5	1121.5	1129.5	1701.5
Average rank	29.313	46.729	47.063	70.896

Finally, the data set rank Q_i is multiplied by the difference between the rank within data set i, r_i^j, and the average rank within data sets, $(k+1)/2$, to get the product S_{ij}, where

$$S_{ij} = Q_i\left[r_i^j - \frac{k+1}{2}\right] \tag{12}$$

is a statistic that represents the relative size of each observation within the data set, adjusted to reflect the relative significance of the data set in which it appears.

For convenience and to establish a relationship with the Friedman test, we will also use rankings without average adjusting:

$$W_{ij} = Q_i\left[r_i^j\right] \tag{13}$$

Let S_j denote the sum for each classifier, $S_j = \sum_{i=1}^{n} S_{ij}$ and $W_j = \sum_{i=1}^{n} W_{ij}$, for $j = 1, 2, \ldots, k$. Next we must to calculate the terms:

$$A_2 = n(n+1)(2n+1)(k)(k+1)(k-1)/72 \tag{14}$$

$$B = \frac{1}{n}\sum_{j=1}^{k} S_j^2 \tag{15}$$

The test statistic is

$$T_3 = \frac{(n-1)B}{A_2 - B} \tag{16}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. The table of critical values for the F-distribution is given in [42, Table A10]. Moreover, the p-value could be computed through normal approximations [1]. If $A_2 = B$, consider the point to be in the critical region of the statistical distribution and calculate the p-value as $(1/k!)^{n-1}$. If the null hypothesis is rejected, we can proceed with a post hoc test, which will be explained in Section 5.

We show an example of the use of the Quade technique by means of the global experimental study considered in this paper. Table 8 displays the aligned observations and the aligned ranks in the parentheses considering the known four algorithms and 24 data sets.

Average ranks T_j could be compared with the ranks obtained by the Friedman test. In this case, PDFC ranked the first with rank 1.393; NNEP and IS-CHC + 1NN ranked the second and the third, with ranks 2.537 and 2.592, respectively; and the last

Table 8
Comparison of accuracy among the four algorithms selected in the experimental study. The ranks in the parentheses are used in the computation of the Quade test. S_{ij} and W_{ij} are displayed in this order.

Data set	Sample range	Rank Q_i	PDFC	NNEP	IS-CHC − 1NN	FH-GBML
Adult*	0.043	8	0.752 (12)(32)	0.773 (4)(24)	0.785 (−4)(16)	0.795 (−12)(8)
Breast	0.035	5	0.727 (−2.5)(10)	0.748 (−7.5)(5)	0.724 (2.5)(15)	0.713 (7.5)(20)
Bupa	0.151	18	0.736 (−27)(18)	0.716 (−9)(36)	0.585 (27)(72)	0.638 (9)(54)
Car	0.203	19	0.994 (−28.5)(19)	0.861 (9.5)(57)	0.880 (−9.5)(38)	0.791 (28.5)(76)
Cleveland	0.067	13	0.508 (19.5)(52)	0.553 (−6.5)(26)	0.575 (−19.5)(13)	0.515 (6.5)(39)
Contraceptive	0.065	12	0.535 (−6)(24)	0.536 (−18)(12)	0.513 (6)(36)	0.471 (18)(48)
Dermatology	0.436	23	0.967 (−34.5)(23)	0.871 (11.5)(69)	0.954 (−11.5)(46)	0.532 (34.5)(92)
Ecoli	0.063	11	0.831 (−16.5)(11)	0.807 (5.5)(33)	0.819 (−5.5)(22)	0.768 (16.5)(44)
German	0.043	7	0.745 (−10.5)(7)	0.702 (10.5)(28)	0.719 (−3.5)(14)	0.705 (3.5)(21)
Glass	0.137	16	0.709 (−24)(16)	0.572 (24)(64)	0.669 (−8)(32)	0.607 (8)(48)
Haberman	0.010	2	0.722 (3)(8)	0.728 (−1)(4)	0.725 (1)(6)	0.732 (−3)(2)
Iris	0.020	3	0.967 (−4.5)(3)	0.947 (4.5)(12)	0.953 (1.5)(9)	0.960 (−1.5)(6)
Lymphography	0.141	17	0.832 (−25.5)(17)	0.752 (8.5)(51)	0.802 (−8.5)(34)	0.691 (25.5)(68)
Mushrooms*	0.515	24	0.998 (−36)(24)	0.992 (−12)(48)	0.482 (36)(96)	0.910 (12)(72)
Newthyroid	0.038	6	0.963 (−6)(9)	0.963 (−6)(9)	0.954 (3)(18)	0.926 (9)(24)
Penbased*	0.352	22	0.982 (−33)(22)	0.953 (−11)(44)	0.932 (11)(66)	0.630 (33)(88)
Ring*	0.205	20	0.978 (−30)(20)	0.773 (30)(80)	0.834 (10)(60)	0.849 (−10)(40)
Satimage*	0.075	14	0.854 (−21)(14)	0.787 (7)(42)	0.841 (−7)(28)	0.779 (21)(56)
Shuttle*	0.048	9	0.965 (4.5)(27)	0.984 (−4.5)(18)	0.995 (−13.5)(9)	0.947 (13.5)(36)
Spambase*	0.120	15	0.924 (−22.5)(15)	0.887 (−7.5)(30)	0.861 (7.5)(45)	0.804 (22.5)(60)
Thyroid*	0.021	4	0.929 (2)(12)	0.942 (−6)(4)	0.931 (−2)(8)	0.921 (6)(16)
Vehicle	0.282	21	0.837 (−31.5)(21)	0.643 (−10.5)(42)	0.602 (10.5)(63)	0.554 (31.5)(84)
Wine	0.050	10	0.972 (−15)(10)	0.956 (−5)(20)	0.944 (5)(30)	0.922 (15)(40)
Wisconsin	0.006	1	0.958 (1.5)(4)	0.959 (0.5)(3)	0.964 (−1)(1.5)	0.964 (−1)(1.5)
Sum of ranks S_j			−332	11	27.5	293.5
Average ranks $T_j = \frac{W_j}{n(n+1)/2}$			1.393	2.537	2.592	3.478

is FH-GBML with rank 3.487. The Quade test checks whether the measured sum of weighted ranks S_j are significantly different from 0, expected under the null hypothesis:

$$A_2 = 24(24+1)(2\cdot 24+1)4(4+1)(4-1)/72 = 24,500 \tag{17}$$

$$B = \frac{1}{24}[(-332)^2 + 11^2 + 27.5^2 + 293.5^2] = 4068.479 \tag{18}$$

$$T_3 = \frac{23\cdot 4068.479}{24,500 - 4068.479} = 21.967 \tag{19}$$

With four algorithms and 24 data sets, T_3 is distributed according to the F distribution with $4-1=3$ and $(4-1)\cdot(24-1)=69$ degrees of freedom. The p-value computed by using the $F(3,69)$ distribution is 4.28×10^{-10}, so the null hypothesis is rejected at a high level of significance.

It should be noted that the Quade test offers a different way of ranking computation. It ranks the data sets in order of importance, so the importance can be defined. For example, instead of using the differences between the best and the worst method, we could use the differences between our proposal and the best performing method (obviously, if our proposal is the one with the minimum mean error or whatever other measure is used, we should take the one which has the second lowest error). Another example is to justify an order of importance among the data sets before beginning the analysis of results.

5. A candidate set of post hoc tests: *p*-values and adjusted *p*-values

This section is devoted to presenting a set of post hoc procedures which can be used after the null hypothesis of equivalence of rankings is rejected through the Friedman and Iman–Davenport extension, Friedman Aligned Ranks or Quade tests, to explain the usefulness of APVs and the procedures to compute depending on the post hoc test and to show an example of their use. It is organized as follows:

- Section 5.1 explains the method of conducting pairwise comparisons that involve the control algorithm within multiple comparison tests and lists a set of post hoc tests. It also introduces the meaning and usefulness of using APVs and provides the necessary computations to obtain them.
- Section 5.2 illustrates the use of the post hoc test under the Friedman, Friedman Aligned Ranks and Quade tests considering the experimental study followed in this paper.

5.1. Post hoc procedures

This paper is focused on the comparison between a control method, which is usually the proposed method, and a set of algorithms used in the empirical study. This set of comparisons is associated with a set or family of hypotheses, all of which are related to the control method. Any of the post hoc tests is suitable for application to nonparametric tests working over a family of hypotheses. The test statistic for comparing the ith algorithm and jth algorithm depends on the main nonparametric procedure used:

- Friedman test: In [10] we can see that the expression for computing the test statistic in Friedman test is

  $$z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6n}} \tag{20}$$

  where $R_i,\ R_j$ are the average rankings by Friedman of the algorithms compared.
- Friedman Aligned Ranks test: Since the set of related rankings is converted to absolute rankings, the expression for computing the test statistic in Friedman Aligned Ranks is the same as that used by the Kruskal–Wallis test [32,10]

  $$z = (\widehat{R}_i - \widehat{R}_j) \Big/ \sqrt{\frac{k(n+1)}{6}} \tag{21}$$

  where $\widehat{R}_i,\ \widehat{R}_j$ are the average rankings by Friedman Aligned Ranks of the algorithms compared.
- Quade test: In [9], the test statistic for comparing two algorithms is given by using the t-student distribution, but we can easily obtain the equivalent in a normal distribution $N(0,1)$ [42]

  $$z = (T_i - T_j) \Big/ \sqrt{\frac{k(k+1)(2n+1)(k-1)}{18n(n+1)}} \tag{22}$$

  where $T_i = \frac{W_i}{n(n+1)/2},\ T_j = \frac{W_j}{n(n+1)/2}$, and W_i and W_j are the rankings without average adjusting by Quade of the algorithms compared. In fact, T_i and T_j compute the correct average magnitudes.

In statistical hypothesis testing, the p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. It is a useful and interesting datum for many consumers of statistical analysis. A p-value provides information about whether a statistical hypothesis test is significant or not, and it also indicates something about "how significant" the result is: the smaller the p-value, the stronger the evidence against the null hypothesis. Most importantly, it does this without committing to a particular level of significance.

When a p-value is considered in a multiple comparison, it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons belonging to the family. If one is comparing k algorithms and in each comparison the level of significance is α, then in a single comparison the probability of not making a Type I error is $(1-\alpha)$, then the probability of not making a Type I error in the $k-1$ comparison is $(1-\alpha)^{(k-1)}$. Then the probability of making one or more Type I error is $1-(1-\alpha)^{(k-1)}$. For instance, if $\alpha = 0.05$ and $k = 10$ this is 0.37, which is rather high.

One way to solve this problem is to report adjusted p-values (APVs) which take into account that multiple tests are conducted. An APV can be compared directly with any chosen significance level α. We recommend the use of APVs due to the fact that they provide more information in a statistical analysis.

The z value in all cases is used to find the corresponding probability (p-value) from the table of normal distribution $N(0,1)$, which is then compared with an appropriate level of significance α (Table A1 in [42]). The post hoc tests differ in the way they adjust the value of α to compensate for multiple comparisons.

Next, we will define a candidate set of post hoc procedures and we will explain how to compute the APVs depending on the post hoc procedure used in the analysis, following the indications given in [52,49]. The notation used in the computation of the APVs is as follows:

- Indexes i and j each correspond to a concrete comparison or hypothesis in the family of hypotheses, according to an incremental order of their p-values. Index i always refers to the hypothesis in question whose APV is being computed and index j refers to another hypothesis in the family.
- p_j is the p-value obtained for the jth hypothesis.
- k is the number of classifiers being compared.

The procedures of p-value adjustment can be classified into:

- one-step
  - The Bonferroni–Dunn procedure (Dunn–Sidak approximation) [14]: it adjusts the value of α in a single step by dividing the value of α by the number of comparisons performed, $(k-1)$. This procedure is the simplest but it also has little power.Bonferroni APV_i: $\min\{v; 1\}$, where $v = (k-1)p_i$.
- step-down
  - The Holm procedure [28]: it adjusts the value of α in a step-down manner. Let $p_1, p_2, \ldots, p_{k-1}$ be the ordered p-values (smallest to largest), so that $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_{k-1}$, and $H_1, H_2, \ldots, H_{k-1}$ be the corresponding hypotheses. The Holm procedure rejects H_1 to H_{i-1} if i is the smallest integer such that $p_i > \alpha/(k-i)$. Holm's step-down procedure starts with the most significant p-value. If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on. As soon as a certain null hypothesis cannot be rejected, all the remaining hypotheses are retained as well.-Holm APV_i: $\min\{v; 1\}$, where $v = \max\{(k-j)p_j : 1 \leqslant j \leqslant i\}$.
  - The Holland procedure [27]: it also adjusts the value of α in a step-down manner, as Holm's method does. It rejects H_1 to H_{i-1} if i is the smallest integer so that $p_i > 1-(1-\alpha)^{k-i}$.Holland APV_i: $\min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{k-j} : 1 \leqslant j \leqslant i\}$.
  - The Finner procedure [17]: it also adjusts the value of α in a step-down manner, as Holm's or Holland's method do. It rejects H_1 to H_{i-1} if i is the smallest integer so that $p_i > 1-(1-\alpha)^{(k-1)/i}$.Finner $APV_i : \min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{(k-1)/j} : 1 \leqslant j \leqslant i\}$.
- step-up
  - The Hochberg procedure [25] it adjusts the value of α in a step-up manner. Hochberg's step-up procedure works in the opposite direction, comparing the largest p value with α, the next largest with $\alpha/2$, the next with $\alpha/3$, and so forth until it encounters a hypothesis it can reject. All hypotheses with smaller p-values are then rejected as well. Hochberg APV_i: $\max\{(k-j)p_j : (k-1) \geqslant j \geqslant i\}$.
  - The Hommel procedure [29] is more complicated to compute and understand. First, we need to find the largest j for which $p_{n-j+k} > k\alpha/j$ for all $k = 1, \ldots, j$. If no such j exists, we can reject all hypotheses, otherwise we reject all for which $p_i \leqslant \alpha/j$.
    Hommel APV_i: see algorithm at Fig. 2.
  - The Rom procedure [41]: Rom developed a modification to Hochberg's procedure to increase its power. It works in exactly the same way as the Hochberg procedure, except that the α values are computed through the expression

$$\alpha_{k-i} = \left[\sum_{j=1}^{i-1} \alpha^j - \sum_{j=1}^{i-2} \binom{i}{k} \alpha_{k-1-j}^{i-j}\right] \Big/ i \tag{23}$$

where $\alpha_{k-1} = \alpha$ and $\alpha_{k-2} = \alpha/2$.

Rom $APV_i : \max\{(r_{k-j})p_j : (k-1) \geqslant j \geqslant i\}$, where r_{k-j} can be obtained from Eq. (23) ($r = \{1, 2, 3, 3.814, 4.755, 5.705, 6.655, \ldots\}$).

- two-step.
  - The Li procedure [34]: Li proposed a two-step rejection procedure.
    - *Step 1:* Reject all H_i if $p_{k-1} \leqslant \alpha$. Otherwise, accept the hypothesis associated to p_{k-1} and go to Step 2.
    - *Step 2:* Reject any remaining H_i with $p_i \leqslant (1 - p_{k-1})/(1 - \alpha)\alpha$.

    Li $APV_i : p_i/(p_i + 1 - p_{k-1})$

5.2. Experimental study

Tables 9–11 show the results in the final form of APVs for the experimental study considered in this paper. As we can see, this example is suitable for observing the difference of power among the test procedures. Also, these tables can provide information about the state of retention or rejection of any hypothesis, comparing its associated APV with the level of significance fixed at the beginning of the statistical analysis.

The statistical tests are conducted once, assuming that the results come from an aggregation based on the means of several repetitions. In our case, we perform 3 repetitions of 10 cfv, so each result belonging to the sample analyzed by the statistical tests actually represents the mean of 30 runs of the algorithm in question. In order to do it, we use 3 different seeds in stochastic algorithms, keeping the same cross-validation partitions. The scope of this paper is focused on the analysis by statistical tests of multiple data sets and we assume that we have available a result from each pair of data set-algorithms. The providence of the results is not our concern, while we understand that it is a controversial question in the literature.

First of all, we can observe that the Bonferroni procedure obtains the highest APV. Theoretically, the step-down procedures usually have less power than step-up ones and the Li procedure seems to be the multiple comparison test with highest power. On the other hand, referring to a comparison between multiple comparison nonparametric tests; that is, Friedman, Friedman Aligned Ranks and Quade; we can see that Quade's procedure is the one which obtains the lowest unadjusted p-value in this example.

The results offered by an experimental study could lead to erroneous conclusions. To better analyze the differences in behavior among all the procedures, we will conduct an experimental study in the next section, where the procedures will be analyzed, taking into account more than one scenario and set of parameters which configure the comparison between classifiers.

6. Experimental analysis: power of the multiple comparisons tests

The power of a statistical test is the probability that the test will reject a false null hypothesis. As power increases, the chances of a Type II error decrease. The probability of a Type II error is referred to as the false negative rate [42]. In this section, we show an actual estimation of the power of the presented procedures through the experiments in which we repeatedly compared the classifiers over a number of randomly chosen data sets, recording the number of equivalence hypothesis rejected and APVs. We follow a similar method used in [11,22].

The classifiers, data sets, parameters and validation procedure used are those described in Section 2. As was mentioned, we measured the performance of each classifier by means of its accuracy in test data by using 3 repetitions of 10-fold cross-validation. When comparing two classifiers, a subgroup of the complete group of data sets considered was randomly selected so that the probability for the data set i being chosen was proportional to $1/(1 + e^{-kd_i})$, where d_i is the (positive or negative) difference of the two classifier accuracies in that data set and k is the bias through which we can regulate the differences

```
Set $APV_i = p_i$ for all $i$.
For each $j = k-1, k-2, ..., 2$ (in that order)
    Let $B = \emptyset$.
    For each $i, i > (k-1-j)$
        Compute value $c_i = (j \cdot p_i)/(j + i - k + 1)$.
        $B = B \cup c_i$.
    End for
    Find the smallest $c_i$ value in $B$; call it $c_{min}$.
    If $APV_i < c_{min}$, then $APV_i = c_{min}$.
    For each $i, i \leq (k-1-j)$
        Let $c_i = min(c_{min}, j \cdot p_i)$.
        If $APV_i < c_i$, then $APV_i = c_i$.
    End for
```

Fig. 2. Algorithm for calculating APVs based on Hommel's procedure.

Table 9
Adjusted *p*-values for the Friedman test (PDFC is the control method).

i Algorithm	1 FH-GBML	2 IS-CHC + 1NN	3 NNEP
Unadjusted p	5.69941×10^{-5}	0.05735	0.05735
p_{Bonf}	1.70982×10^{-4}	0.17204	0.17204
p_{Holm}	1.70982×10^{-4}	0.11469	0.11469
p_{Hoch}	1.70982×10^{-4}	0.05735	0.05735
p_{Homm}	1.70982×10^{-4}	0.05735	0.05735
p_{Holl}	1.70973×10^{-4}	0.11141	0.11141
p_{Rom}	1.70982×10^{-4}	0.05735	0.05735
p_{Finn}	1.70982×10^{-4}	0.08477	0.08477
p_{Li}	6.04577×10^{-4}	0.05735	0.05735

Table 10
Adjusted *p*-values for the Friedman Aligned Ranks test (PDFC is the control method).

i Algorithm	1 FH-GBML	2 IS-CHC + 1NN	3 NNEP
Unadjusted p	2.32777×10^{-7}	0.02729	0.03032
p_{Bonf}	6.98332×10^{-7}	0.08188	0.09097
p_{Holm}	6.98332×10^{-7}	0.05459	0.05459
p_{Hoch}	6.98332×10^{-7}	0.03032	0.03032
p_{Homm}	6.98332×10^{-7}	0.03032	0.03032
p_{Holl}	6.98332×10^{-7}	0.05384	0.05384
p_{Rom}	6.98332×10^{-7}	0.03032	0.03032
p_{Finn}	6.98332×10^{-7}	0.04066	0.04066
p_{Li}	2.40057×10^{-7}	0.02738	0.03032

Table 11
Adjusted *p*-values for the Quade test (PDFC is the control method).

i Algorithm	1 FH-GBML	2 IS-CHC + 1NN	3 NNEP
Unadjusted p	6.43747×10^{-4}	0.02163	0.02843
p_{Bonf}	1.93124×10^{-4}	0.06490	0.08528
p_{Holm}	1.93124×10^{-4}	0.04326	0.04326
p_{Hoch}	1.93124×10^{-4}	0.02843	0.02843
p_{Homm}	1.93124×10^{-4}	0.02843	0.02843
p_{Holl}	1.93112×10^{-4}	0.04280	0.04280
p_{Rom}	1.93124×10^{-4}	0.02843	0.02843
p_{Finn}	1.93124×10^{-4}	0.03227	0.03227
p_{Li}	6.62538×10^{-4}	0.02178	0.02843

between the classifiers. With $k = 0$, the selection is purely random because its value is 0.5, hence the probability of selection of each data set is 50% (the order of data set presentation is assumed to be random each time). As k is higher, the selected data sets are favorable to a particular classifier, increasing the probability of selection for those data sets in which the classifier in question obtains better accuracy than the other one selected. For each k, 1000 choices of data sets and statistical analysis are made. Depending on the characteristics of the study, we will choose two algorithms from the eight considered in this paper.

Three studies will be carried out in order to analyze the properties of the procedures presented.

6.1. Analysis of the power of nonparametric multiple comparisons tests

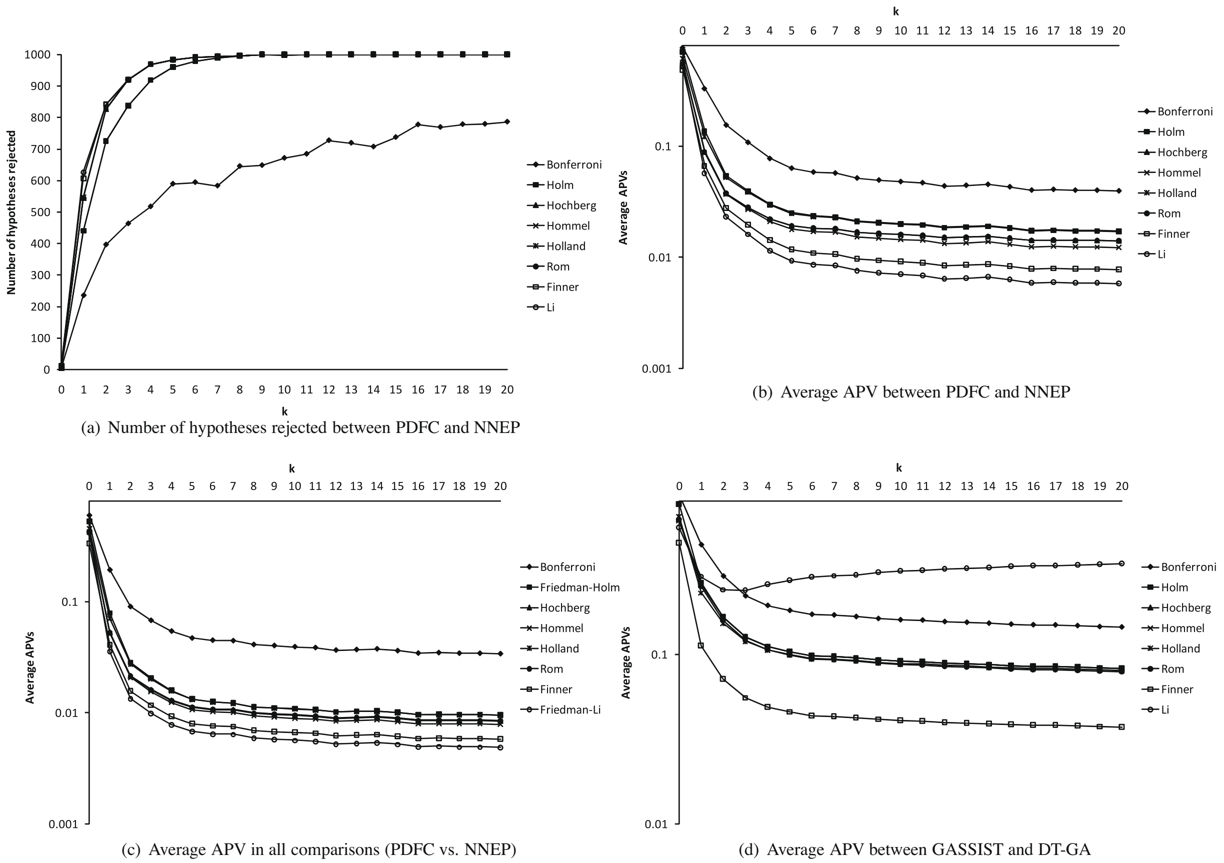
The first study corresponds to the analysis of the multiple comparisons tests used in this paper: Friedman, Friedman Aligned Ranks and Quade tests. Fig. 4 illustrates this study. Our interest is to detect differences in the behavior of these tests when an important factor that takes part in a multiple comparison analysis changes. This factor will be the number of algorithms. We study the effect and power of the tests mentioned when eight algorithms participate in the multiple comparison and when only four methods do so. The choice of data sets is again based upon PDFC and NNEP, and the number of data sets selected is 10 when comparing the 8 methods and 6 when comparing 4 methods. The reason for doing this is to avoid promoting the rejection of hypotheses due to the elevated number of data sets with respect to the number of classifiers. The post hoc procedures used are Holm and Li.

Fig. 4a displays the average APVs in all comparisons when 8 classifiers take part in the multiple comparison. It is appreciable that the Quade test has the best performance in terms of power and the Friedman Aligned Ranks is the test that has an inferior power. However, in Fig. 4b, we can see that the Friedman test present a more conservative behavior, in the case when only four algorithms take part in the multiple comparison. These graphics check an interesting feature remarked upon in [10], where the author indicates that the power of this test depends on the number of treatments (classifiers in our case) per block (data set), so the Friedman-AR test suffers more than the other two tests when the number of algorithms is high. We have empirically showed that the use of 4 algorithms in the comparison allows us to improve the power of the Friedman Aligned Ranks test. On the other hand, we have noticed that the Quade test performs better than the Friedman and Friedman Aligned Ranks test in this study. We will see later that the Quade test is not always the best choice.

6.2. Analysis of the power of the post hoc procedures

The second study we have carried out is illustrated in Fig. 3. In this case, we employ all the classifiers described in the paper and we carry out the procedure explained above by choosing, at each step, 10 of the 48 data sets. The methods chosen for guiding the data set choice are PDFC and NNEP, except in the last study (Fig. 3d). The Friedman test is then used to compute the rankings. The purpose of this study is to detect and quantify the differences in power observed among the 8 post hoc procedures explained in Section 5.

Fig. 3a represents the number of hypotheses rejected between the two methods considered (at a level of significance $\alpha = 0.05$) and Fig. 3b their associated average APV. It perfectly illustrates the power of the post hoc procedures when the data sets are chosen agree that the differences between the two methods are greater; that is, data sets that favor the accuracy of the PDFC method. The power of the Bonferroni–Dunn procedure is the lowest and far from the rest of procedures. The Holm and Holland tests behave very similarly among themselves, which is also the case for Hochberg and Rom. However, the second ones, which are step-up procedures, have more power than the first, step-down methods. The next, in terms of power, is the Hommel test. The two best procedures in terms of power are Finner and Li and we should note that the distance between Hommel and Finner is very noticeable. It seems that the Li test is the most powerful of them.



(a) Number of hypotheses rejected between PDFC and NNEP

(b) Average APV between PDFC and NNEP

(c) Average APV in all comparisons (PDFC vs. NNEP)

(d) Average APV between GASSIST and DT-GA

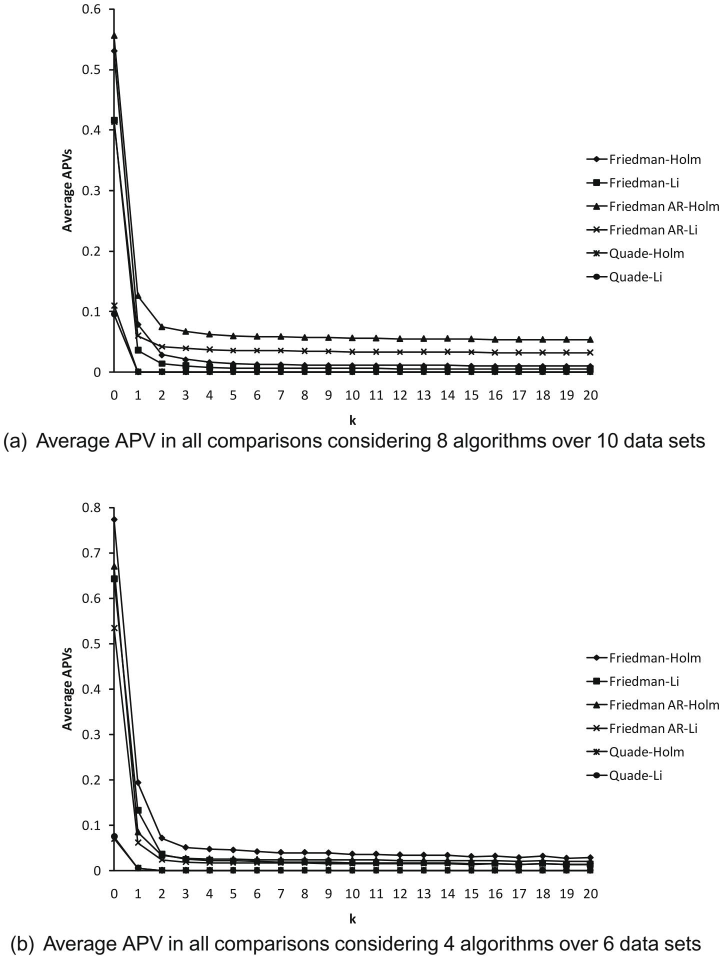
Fig. 3. Comparison between all post hoc tests.

Fig. 3c shows the average APV considering all comparisons between the control method PDFC and the rest of classifiers. The results are similar to those commented in the previous graphic. However, the Li test is not always the less conservative procedure. In Fig. 3d we can see a case in which the Li test performs even worse than Bonferroni–Dunn. This case represents the average APV between GASSIST and DT-GA, two methods whose results obtained are very similar in terms of accuracy. Note that the Finner methods is not able to reject the null hypothesis with a level of significance $\alpha = 0.05$ when $k = 20$. This fact indicates to us that the Li test loses effectiveness when the classifiers to be compared are very similar in behavior. More specifically, the Li procedure is notably superior when the largest p-value is anticipated to be less than 0.5 [34].

6.3. Analysis of the stability of the Quade test

Finally, we want to study the stability of the Quade test with respect to the choice of data sets. We refer to stability as the property of reporting similar results (number of rejections at a certain significance level or average APV) when the samples to be analyzed do not substantially change. In this study, it can be accomplished by keeping $k = 0$ in all repetitions. Fig. 5 shows the results of this study. In this case, as leading classifiers, we have used the two best: PDFC and GASSIST. PDFC is much better than GASSIST in many data sets. The value of k remains unaltered throughout the experiment, so the choice of data sets is always random and actually the graphics show the run of 1000×21 statistical tests and it represents the average APV every 1000 runs. Our objective is to show that the Quade test is much more dependent on the choice of data sets than the remaining tests.

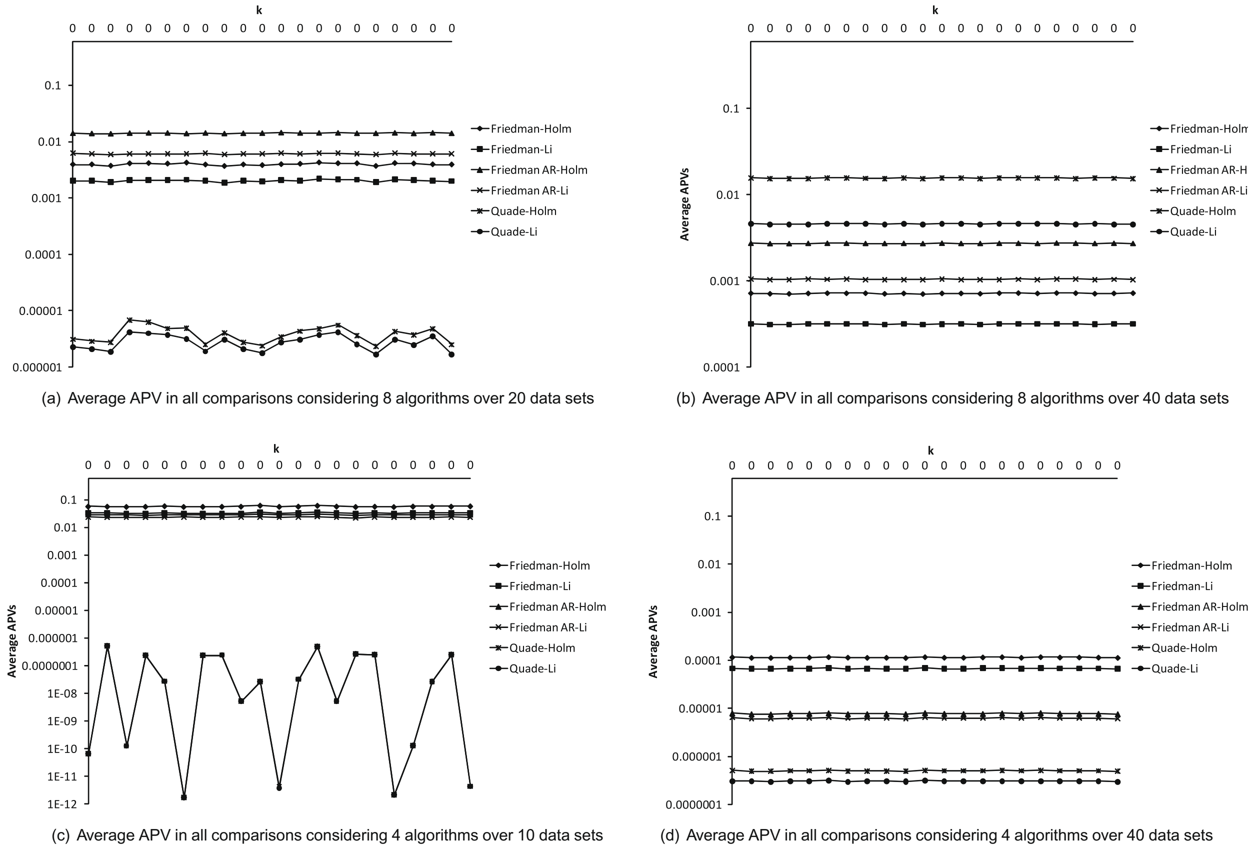
Fig. 5a and b illustrates the case which works with 8 classifiers. In the first one, 20 data sets are randomly chosen, whereas in the second one, 40 data sets are chosen. Each point in the graphics represents the average APV over 1000 runs (or data set choices) as we have mentioned before. If the number of data sets is low with respect to the total number of them, and the control method presents significantly better results in some data sets than in the remaining algorithms, the behavior of the test is very dependent on the data set choice. The Quade test benefits very much from high performance values in a subset of



(a) Average APV in all comparisons considering 8 algorithms over 10 data sets

(b) Average APV in all comparisons considering 4 algorithms over 6 data sets

Fig. 4. Comparison between Friedman, Friedman Aligned Ranks and Quade (PDFC vs. NNEP).



(a) Average APV in all comparisons considering 8 algorithms over 20 data sets

(b) Average APV in all comparisons considering 8 algorithms over 40 data sets

(c) Average APV in all comparisons considering 4 algorithms over 10 data sets

(d) Average APV in all comparisons considering 4 algorithms over 40 data sets

Fig. 5. Comparison of stability between Friedman, Friedman Aligned Ranks and Quade (PDFC vs. GASSIST).

data set, and the power registered could be much higher than Friedman as we can see in Fig. 5a. However, when the number of data sets to be chosen is high with respect to the original set of data sets, the Quade test is the one that obtains the lowest power (Fig. 5b). Note that this procedure also ranks the data sets among themselves, so the data sets in which the control method presents the best results, influence the process even more than in other nonparametric tests.

Fig. 5c and d illustrates the same study but now considering 4 classifiers instead of 8. Centering on this case, the effects of superiority showed by the Quade test are stressed because in Fig. 5c the number of random data sets chosen is only 10. Notice the instability of the Quade test in terms of power. When the number of data sets chosen is increased to 40, we obtain a similar result to before, but the power of the procedures follows the opposite order: Quade's power > Friedman Aligned Ranks' power > Friedman's power. In relation to the study conducted in Fig. 4, we observe that the Quade test performs better than the Friedman and even Friedman Aligned Ranks when the number of treatments (classifiers in our case) per block (data set) is low [9]. We have showed that in real circumstances, the Quade test is more powerful than Friedman Aligned Ranks test when the number of algorithms to be analyzed is low, but we have to be cautious when using the Quade test, because it is very dependant on the choice of data sets and the differences reported among algorithms could be excessive due to the fact that the data sets chosen could benefit the computation ranking procedure.

7. Summary and suggestions

This section is dedicated to give some considerations on the use of the nonparametric and post hoc tests presented in this paper. Their characteristics as well as suggestions on some of their aspects and details of the multiple comparisons tests are enumerated:

- As we have suggested, multiple comparison tests must be used when we want to establish a statistical comparison of the results reported among various algorithms. This paper focuses on procedures that work with a control method, that is, a method to be compared against a set of algorithms. It could be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. In this paper we present three alternatives: the Friedman test with the Iman–Davenport extension, the Friedman Aligned

Ranks test and the Quade test. Once one of these tests rejects the hypothesis of equivalence of medians, the detection of the specific differences among the algorithms can be made with the application of post hoc statistical procedures, which are methods used for specifically comparing a control algorithm with two or more algorithms.

- An alternative to directly performing a comparison between a control algorithm and a set of algorithms is the Multiple Sign-test. It has been described in this paper and an example of its use has been provided. We have shown that this procedure is rapid and easy to apply, but it has low power with respect to more advanced techniques. We recommend its use when the differences reported by the control algorithm with respect to the rest of methods are very clear for a certain performance metric.
- Another interesting procedure presented in this paper is related to the Contrast Estimation based on medians between two samples of results. The Contrast Estimation in nonparametric statistics is used for computing the real differences between two algorithms, considering the median measure the most important. Taking into account that the samples of results in CI experiments rarely fulfill the needed conditions for a safe use of parametric tests, the computation of nonparametric contrast estimation through the use of medians is very useful. For example, a paper could provide, apart from the average values of accuracies over various data sets reported by the classifiers compared, the contrast estimation between them over multiple data sets, which is a safer metric in multiple-problem environments.
- Apart from the well-known Friedman test, we can use two alternatives which differ in the ranking computation. Both, the Friedman Aligned Rank test and the Quade test, can be used under the same circumstances as the Friedman test. We have studied their relative power and we have shown that they perform better than Friedman when the number of algorithms is low, at not more than 4 or 5 algorithms. The differences in power between Friedman Aligned Ranks and Quade are unknown [10,9], but we encourage the use of these tests when the number of algorithms to be compared is low.
- As we have described, the Quade test adds to the ranking computation of Friedman's test a weight factor computed through the maximum and minimum differences in a data set. This implies that those algorithms that obtain further positive results in diverse data sets could benefit from this test. The use of this test should be regulated because it is very sensitive to the choice of data sets. If a researcher decided to include a subgroup of an already studied group of data sets where in most of them the proposal obtained good results, this test would report excessive significant differences. On the other hand, for specific problems in which we are interested in quantifying the real differences obtained between methods, the use of this test can be justified. We recommend the use of this procedure under justified circumstances and with special caution.
- In relation to the post hoc procedures analyzed, we have seen the contrast of power among them. Actually, the differences of power between the methods are rather small, with some exceptions. The Bonferroni–Dunn test should not be used in spite of its simplicity, because it is a very conservative test and many differences may not be detected. Five procedures – Holm, Hochberg, Hommel, Holland and Rom – have a similar power. Although Hommel and Rom are the two most powerful procedures, they also are the most difficult to be applied and to be understood. A good alternative is to use the Finner test, which is easy to comprehend and offers better results than the remaining tests, except the Li test in some cases.
- The Li test is even simpler than the Finner, Holm or Hochberg tests. This test needs to check only two steps and to know the greatest unadjusted p-value in the comparison, which is easy to obtain. In [34], the author declares that the power of his test is highly influenced by the p-value and when it is lower than 0.5, the test will perform very well. We have shown in the experimental analysis that depending on the classifiers compared, this test has the highest or the lowest power, depending on the circumstances. We recommend that it be used with care and only when the differences between the control algorithm and the rest seem to be high in the performance measure analyzed.
- Finally, we want to remark that the choice of any of the statistical procedures presented in this paper for conducting an experimental analysis should be justified by the researcher. The use of the most powerful procedures does not imply that the results obtained by his/her proposal will be better. The choice of a statistical technique is ruled by a trade-off between its power and its complexity when it comes to being used or explained to non-expert readers in statistics.

8. Conclusions

In this paper, we have studied the use of nonparametric statistical techniques in the analysis of the behavior of computational intelligence algorithms for data mining classification tasks, analyzing the use of multiple comparisons procedures that use a control method.

We have presented some basic techniques for performing multiple comparisons of performance results between a proposed method and a set of algorithms. Among them, we explained the Multiple Sign-test, which is a very interesting test for carrying out rapid empirical comparisons and the Contrast based on medians, which can be used to obtain the exact differences between two algorithms over various case problems.

In addition, we have presented two advanced alternatives to the Friedman test, namely the Aligned Friedman test and Quade test. They differ in the computation of the rankings in the set of results and they provide certain advantages depending on the properties of the experimental study. In addition, a set of post hoc procedures for detecting the differences between two algorithms that belong to the multiple comparison have been described and analyzed in terms of power and we have given a recommendation as to their use.

Table A.1
Critical values of minimum r_j for comparison of $m = k - 1$ algorithms against one control in n data sets. *Source:* A.L. Rhyne, R.G.D. Steel, Tables for a treatments versus control multiple comparisons sign test, Technometrics 7 (1965) 293–306.

n	Level of significance	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
5	0.1	0	0	–	–	–	–	–	–
	0.05	–	–	–	–	–	–	–	–
6	0.1	0	0	0	0	0	–	–	–
	0.05	0	0	–	–	–	–	–	–
7	0.1	0	0	0	0	0	0	0	0
	0.05	0	0	0	0	–	–	–	–
8	0.1	1	1	0	0	0	0	0	0
	0.05	0	0	0	0	0	0	0	0
9	0.1	1	1	1	1	0	0	0	0
	0.05	1	0	0	0	0	0	0	0
10	0.1	1	1	1	1	1	1	1	1
	0.05	1	1	1	0	0	0	0	0
11	0.1	2	2	1	1	1	1	1	1
	0.05	1	1	1	1	1	1	0	0
12	0.1	2	2	2	2	1	1	1	1
	0.05	2	1	1	1	1	1	1	1
13	0.1	3	2	2	2	2	2	2	2
	0.05	2	2	2	1	1	1	1	1
14	0.1	3	3	2	2	2	2	2	2
	0.05	2	2	2	2	2	2	1	1
15	0.1	3	3	3	3	3	2	2	2
	0.05	3	3	2	2	2	2	2	2
16	0.1	4	3	3	3	3	3	3	3
	0.05	3	3	3	3	2	2	2	2
17	0.1	4	4	4	3	3	3	3	3
	0.05	4	3	3	3	3	3	2	2
18	0.1	5	4	4	4	4	4	3	3
	0.05	4	4	3	3	3	3	3	3
19	0.1	5	5	4	4	4	4	4	4
	0.05	4	4	4	4	3	3	3	3
20	0.1	5	5	5	5	4	4	4	4
	0.05	5	4	4	4	4	4	3	3
21	0.1	6	5	5	5	5	5	5	5
	0.05	5	5	5	4	4	4	4	4
22	0.1	6	6	6	5	5	5	5	5
	0.05	6	5	5	5	4	4	4	4
23	0.1	7	6	6	6	6	5	5	5
	0.05	6	6	5	5	5	5	5	5
24	0.1	7	7	6	6	6	6	6	6
	0.05	6	6	6	5	5	5	5	5
25	0.1	7	7	7	7	6	6	6	6
	0.05	7	6	6	6	6	6	5	5
30	0.1	10	9	9	9	8	8	8	8
	0.05	9	8	8	8	8	8	7	7
35	0.1	12	11	11	11	10	10	10	10
	0.05	11	10	10	10	10	9	9	9
40	0.1	14	13	13	13	13	12	12	12
	0.05	13	12	12	12	12	11	11	11
45	0.1	16	16	15	15	15	14	14	14
	0.05	15	14	14	14	14	13	13	13
50	0.1	18	18	17	17	17	17	16	16
	0.05	17	17	16	16	16	16	15	15

For a better understanding, all the procedures described in this paper have been applied to an example of experimental study of classification over multiple data sets.

Acknowledgements

This work was supported in part by the Spanish Ministry of Science and Innovation (MICINN) under Project TIN-2008-06681-C06-01. J. Luengo holds a FPU scholarship from Spanish Ministry of Education and Science. The authors are very grateful to the anonymous reviewers for their valuable suggestions and comments to improve the quality of this paper.

Appendix A. Source code of the procedures and table for Multiple Comparison Sign test

The source code, written in JAVA, that implements all procedures described in this paper, is available at http://www.sci2s.ugr.es/keel/controlTest.zip. The program supports data inputs in CSV format and outputs a LATEX document. A complete description of nonparametric tests together with software for their use can be also found in the web site available at: http://www.sci2s.ugr.es/sicidm/.

References

[1] M. Abramowitz, Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables, Dover Publications, 1974.

[2] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del 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 to data mining problems, Soft Computing 13 (3) (2009) 307–318.

[3] A. Asuncion, D. Newman, UCI machine learning repository, 2007. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

[4] J. Bacardit, J.M. Garrell, Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach learning classifier system. In: T. Kovacs, X. Llora, K. Takadama, P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), Revised Selected Papers of the International Workshop on Learning Classifier Systems 2003–2005, vol. 4399 of LNCS, Springer, 2007, pp. 59–79.

[5] J. Bacardit, D. Goldberg, M. Butz, Improving the performance of a pittsburgh learning classifier system using a default rule. In: T. Kovacs, X. Llora, K. Takadama, P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), Revised Selected Papers of the International Workshop on Learning Classifier Systems 2003–2005, vol. 4399 of LNCS, Springer, 2007, pp. 291–307.

[6] J.R. Cano, F. Herrera, M. Lozano, Using evolutionary algorithms as instance selection for data reduction in KDD an experimental study, IEEE Transactions on Evolutionary Computation 7 (6) (2003) 561–575.

[7] D. Carvalho, A. Freitas, A hybrid decision tree/genetic algorithm method for data mining, Information Sciences 163 (1–3) (2004) 13–35.

[8] Y. Chen, J. Wang, Support vector learning for fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems 11 (6) (2003) 716–728.

[9] W.J. Conover, Practical Nonparametric Statistics, John Wiley and Sons, 1999.

[10] W.W. Daniel, Applied Nonparametric Statistics, Duxbury Thomson Learning, 1990.

[11] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[12] T.G. Dietterich, Approximate statistical test for comparing supervised classification learning algorithms, Neural Computation 10 (7) (1998) 1895–1923.

[13] K. Doksum, Robust procedures for some linear models with one observation per cell, Annals of Mathematical Statistics 38 (1967) 878–883.

[14] O.J. Dunn, Multiple comparisons among means, Journal of the American Statistical Association 56 (1961) 52–64.

[15] A.P. Engelbrecht, Computational Intelligence: An Introduction, Wiley, 2007.

[16] L.J. Eshelman, The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination, in: G.J.E. Rawlings (Ed.), Foundations of Genetic Algorithms and Classifier Systems, Morgan Kaufmann, 1991, pp. 265–283.

[17] H. Finner, On a monotonicity problem in step-down multiple test procedures, Journal of the American Statistical Association 88 (1993) 920–923.

[18] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 674–701.

[19] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1940) 86–92.

[20] S. García, J.R. Cano, E. Bernadó-Mansilla, F. Herrera, Diagnose of effective evolutionary prototype selection using an overlapping measure, International Journal of Pattern Recognition and Artificial Intelligence 23 (8) (2009) 1527–1548.

[21] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing 13 (10) (2009) 959–977.

[22] S. García, F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.

[23] S. García, D. Molina, M. Lozano, F. Herrera, A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization, Journal of Heuristics 15 (2009) 617–644.

[24] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2005.

[25] Y. Hochberg, A sharper Bonferroni procedure for multiple tests of significance, Biometrika 75 (1988) 800–803.

[26] J.L. Hodges, E.L. Lehmann, Ranks methods for combination of independent experiments in analysis of variance, Annals of Mathematical Statistics 33 (1962) 482–497.

[27] B.S. Holland, M.D. Copenhaver, An improved sequentially rejective Bonferroni test procedure, Biometrics 43 (1987) 417–423.

[28] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70.

[29] G. Hommel, A stagewise rejective multiple test procedure based on a modified Bonferroni test, Biometrika 75 (1988) 383–386.

[30] R.L. Iman, J.M. Davenport, Approximations of the critical region of the friedman statistic, Communications in Statistics 9 (1980) 571–595.

[31] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy GBML approaches for pattern classification problems, IEEE Transactions on System, Man and Cybernetics B 35 (2) (2005) 359–365.

[32] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American Statistical Association 47 (1952) 583–621.

[33] Z. Lei, L. Ren-hou, Designing of classifiers based on immune principles and fuzzy rules, Information Sciences 178 (7) (2008) 1836–1847.

[34] J. Li, A two-step rejection procedure for testing multiple hypotheses, Journal of Statistical Planning and Inference 138 (2008) 1521–1527.

[35] F. Martínez-Estudillo, C. Hervás-Martínez, P. Gutiérrez, A. Martínez-Estudillo, Evolutionary product-unit neural networks classifiers, Neurocomputing 72 (1–3) (2008) 548–561.

[36] R. Parpinelli, H. Lopes, A. Freitas, Data mining with an ant colony optimization algorithm, Transactions on Evolutionary Computation 6 (4) (2002) 321–332.

[37] I. Partalas, G. Tsoumakas, E.V. Hatzikos, I.P. Vlahavas, Greedy regression ensemble selection: theory and an application to water quality prediction, Information Sciences 178 (20) (2008) 3867–3879.

[38] D. Quade, Using weighted rankings in the analysis of complete blocks with additive block effects, Journal of the American Statistical Association 74 (1979) 680–683.

[39] A.L. Rhyne, R.G.D. Steel, Tables for a treatments versus control multiple comparisons sign test, Technometrics 7 (1965) 293–306.
[40] V. Rivas, J. Merelo, P. Castillo, M. Arenas, J. Castellano, Evolving RBF neural networks for time-series forecasting with EvRBF, Information Sciences 165 (3–4) (2004) 207–220.
[41] D.M. Rom, A sequentially rejective test procedure based on a modified Bonferroni inequality, Biometrika 77 (1990) 663–665.
[42] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, 2006.
[43] D. Shilane, J. Martikainen, S. Dudoit, S.J. Ovaska, A general framework for statistical performance comparison of evolutionary computation algorithms, Information Sciences 178 (14) (2008) 2870–2879.
[44] R.G.D. Steel, A multiple comparison sign test: treatments versus control, Journal of American Statistical Association 54 (1959) 767–775.
[45] P.N. Tan, Introduction to Data Mining, Pearson, 2006.
[46] C.-J. Tsai, C.-I. Lee, W.-P. Yang, A discretization algorithm based on class-attribute contingency coefficient, Information Sciences 178 (3) (2008) 714–731.
[47] S. Tsumoto, Contingency matrix theory: statistical dependence in a contingency table, Information Sciences 179 (11) (2009) 1615–1627.
[48] A. Ulaş, M. Semerci, O.T. Yildiz, E. Alpaydin, Incremental construction of classifier and discriminant ensembles, Information Sciences 179 (9) (2009) 1298–1318.
[49] P.H. Westfall, S.S. Young, Resampling-based Multiple Testing: Examples and Methods for *p*-Value Adjustment, John Wiley and Sons, 2004.
[50] D.H. Wolpert, The supervised learning no-free-lunch theorems. In: Proceedings of the Sixth Online World Conference on Soft Computing in Industrial Applications, 2001.
[51] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Transactions on Evolutionary Computation 1 (1) (1997) 67–82.
[52] S.P. Wright, Adjusted *p*-values for simultaneous inference, Biometrics 48 (1992) 1005–1013.
[53] Y. Yang, G. Webb, Discretization for Naive-Bayes learning: managing discretization bias and variance, Machine Learning 74 (3) (2009) 39–74.
[54] J.H. Zar, Biostatistical Analysis, Prentice Hall, 1999.