Journal of Machine Learning Research 9 (2008) 2677-2694 Submitted 10/07; Revised 4/08; Published 12/08

An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons

Salvador García SALVAGL@DECSAI.UGR.ES
Francisco Herrera HERRERA@DECSAI.UGR.ES
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, 18071, Spain

Editor: John Shawe-Taylor

Abstract

In a recently published paper in JMLR, Demšar (2006) recommends a set of non-parametric statistical tests and procedures which can be safely used for comparing the performance of classifiers over multiple data sets. After studying the paper, we realize that the paper correctly introduces the basic procedures and some of the most advanced ones when comparing a control method. However, it does not deal with some advanced topics in depth. Regarding these topics, we focus on more powerful proposals of statistical procedures for comparing $n \times n$ classifiers. Moreover, we illustrate an easy way of obtaining adjusted and comparable p-values in multiple comparison procedures.

Keywords: statistical methods, non-parametric test, multiple comparisons tests, adjusted p-values, logically related hypotheses

1. Introduction

In the Machine Learning (ML) scientific community there is a need for rigorous and correct statistical analysis of published results, due to the fact that the development or modifications of algorithms is a relatively easy task. The main inconvenient related to this necessity is to understand and study the statistics and to know the exact techniques which can or cannot be applied depending on the situation, that is, type of results obtained. In a recently published paper in JMLR by Demšar (2006), a group of useful guidelines are given in order to perform a correct analysis when we compare a set of classifiers over multiple data sets. Demšar recommends a set of non-parametric statistical techniques (Zar, 1999; Sheskin, 2003) for comparing classifiers under these circumstances, given that the sample of results obtained by them does not fulfill the required conditions and it is not large enough for making a parametric statistical analysis. He analyzed the behavior of the proposed statistics on classification tasks and he checked that they are more convenient than parametric techniques.

Recent studies apply the guidelines given by Demšar in the analysis of performance of classifiers (Esmeir and Markovitch, 2007; Marrocco et al., 2008). In them, a new proposal or methodology is offered and it is compared with other methods by means of pairwise comparisons. Another type of studies assume an empirical comparison or review of already proposed methods. In these cases, no proposal is offered and a statistical comparison could be very useful in determining the differences among the methods. In the specialized literature, many papers provide reviews on a specific topic and they also use statistical methodology to perform comparisons. For example, in a review of

ensembles of decision trees, non-parametric tests are also applied in the analysis of performance (Banfield et al., 2007). However, only the rankings computed by Friedman's method (Friedman, 1937) are stipulated and authors establish comparisons based on them, without taking into account significance levels. Demšar focused his work in the analysis of new proposals, and he introduced the Nemenyi test for making all pairwise comparisons (Nemenyi, 1963). Nevertheless, the Nemenyi test is very conservative and it may not find any difference in most of the experimentations. In recent papers, the authors have used the Nemenyi test in multiple comparisons. Due to the fact that this test posses low power, authors have to employ many data sets (Yang et al., 2007b) or most of the differences found are not significant (Yang et al., 2007a; Núñez et al., 2007). Although the employment of many data sets could seem beneficial in order to improve the generalization of results, in some specific domains, that is, imbalanced classification (Owen, 2007) or multi-instance classification (Murray et al., 2005), data sets are difficult to find.

Procedures with more power than Nemenyi's one can be found in specialized literature. We have based on the necessity to apply more powerful procedures in empirical studies in which no new method is proposed and the benefit consists of obtaining more statistical differences among the classifiers compared. Thus, in this paper we describe these procedures and we analyze their behavior by means of the analysis of multiple repetitions of experiments with randomly selected data sets.

On the other hand, we can see other works in which the *p*-value associated to a comparison between two classifiers is reported (García-Pedrajas and Fyfe, 2007). Classical non-parametric tests, such as Wilcoxon and Friedman (Sheskin, 2003), may be incorporated in most of the statistical packages (SPSS, SAS, R, etc.) and the computation of the final *p*-value is usually implemented. However, advanced procedures such as Holm (1979), Hochberg (1988), Hommel (1988) and the ones described in this paper are usually not incorporated in statistical packages. The computation of the correct *p*-value, or Adjusted *P*-Value (APV) (Westfall and Young, 2004), in a comparison using any of these procedures is not very difficult and, in this paper, we show how to include it with an illustrative example.

The paper is set up as follows. Section 2 presents more powerful procedures for comparing all the classifiers among them in a $n \times n$ comparison of multiple classifiers and a case study. In Section 3 we describe the procedures for obtaining the APV by considering the post-hoc procedures explained by Demšar and the ones explained in this paper. In Section 4, we perform an experimental study of the behavior of the statistical procedures and we discuss the results obtained. Finally, Section 5 concludes the paper.

2. Comparison of Multiple Classifiers: Performing All Pairwise Comparisons

In the paper Demšar (2006), referring to carrying out comparisons of more than two classifiers, a set of useful guidelines were given for detecting significant differences among the results obtained and post-hoc procedures for identifying these differences. Friedman's test is an omnibus test which can be used to carry out these types of comparison. It allows to detect differences considering the global set of classifiers. Once Friedman's test rejects the null hypothesis, we can proceed with a post-hoc test in order to find the concrete pairwise comparisons which produce differences. Demšar described the use of the Nemenyi test used when all classifiers are compared with each other. Then, he focused on procedures that control the family-wise error when comparing with a control classifier, arguing that the objective of a study is to test whether a newly proposed method is better than the existing

ones. For this reason, he described and studied in depth more powerful and sophisticated procedures derived from Bonferroni-Dunn such as Holm's, Hochberg's and Hommel's methods.

Nevertheless, we think that performing all pairwise comparisons in an experimental analysis may be useful and interesting in different cases when proposing a new method. For example, it would be interesting to conduct a statistical analysis over multiple classifiers in review works in which no method is proposed. In this case, the repetition of comparisons choosing different control classifiers may lose the control of the family-wise error.

Our intention in this section is to give a detailed description of more powerful and advanced procedures derived from the Nemenyi test and to show a case study that uses these procedures.

2.1 Advanced Procedures for Performing All Pairwise Comparisons

A set of pairwise comparisons can be associated with a set or family of hypotheses. Any of the post-hoc tests which can be applied to non-parametric tests (that is, those derived from the Bonferroni correction or similar procedures) work over a family of hypotheses. As Demšar explained, the test statistics for comparing the i-th and j-th classifier is

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6N}}},$$

where R_i is the average rank computed through the Friedman test for the i-th classifier, k is the number of classifiers to be compared and N is the number of data sets used in the comparison.

The z value is used to find the corresponding probability (p-value) from the table of normal distribution, which is then compared with an appropriate level of significance α (Table A1 in Sheskin, 2003). Two basic procedures are:

- Nemenyi (1963) procedure: it adjusts the value of α in a single step by dividing the value of α by the number of comparisons performed, $m = k(k-1)/2$. This procedure is the simplest but it also has little power.

- Holm (1979) procedure: it was also described in Demšar (2006) but it was used for comparisons of multiple classifiers involving a control method. It adjusts the value of α in a step down method. Let $p_1, ..., p_m$ be the ordered p-values (smallest to largest) and $H_1, ..., H_m$ be the corresponding hypotheses. Holm's procedure rejects H_1 to $H_{(i-1)}$ if i is the smallest integer such that $p_i > \alpha/(m-i+1)$. Other alternatives were developed by Hochberg (1988), Hommel (1988) and Rom (1990). They are easy to perform, but they often have a similar power to Holm's procedure (they have more power than Holm's procedure, but the difference between them is not very notable) when considering all pairwise comparisons.

The hypotheses being tested belonging to a family of all pairwise comparisons are logically interrelated so that not all combinations of true and false hypotheses are possible. As a simple example of such a situation suppose that we want to test the three hypotheses of pairwise equality associated with the pairwise comparisons of three classifiers $C_i, i = 1, 2, 3$. It is easily seen from the relations among the hypotheses that if any one of them is false, at least one other must be false. For example, if C_1 is better/worse than C_2, then it is not possible that C_1 has the same performance as C_3 and C_2 has the same performance as C_3. C_3 must be better/worse than C_1 or C_2 or the two classifiers at the same time. Thus, there cannot be one false and two true hypotheses among these three.

Based on this argument, Shaffer proposed two procedures which make use of the logical relation among the family of hypotheses for adjusting the value of α (Shaffer, 1986).

- Shaffer's static procedure: following Holm's step down method, at stage j, instead of rejecting H_i if $p_i \leq \alpha/(m-i+1)$, reject H_i if $p_i \leq \alpha/t_i$, where t_i is the maximum number of hypotheses which can be true given that any $(i-1)$ hypotheses are false. It is a static procedure, that is, $t_1, ..., t_m$ are fully determined for the given hypotheses $H_1, ..., H_m$, independent of the observed p-values. The possible numbers of true hypotheses, and thus the values of t_i can be obtained from the recursive formula

$$S(k) = \bigcup_{j=1}^{k} \{ \binom{j}{2} + x : x \in S(k-j) \},$$

  where $S(k)$ is the set of possible numbers of true hypotheses with k classifiers being compared, $k \geq 2$, and $S(0) = S(1) = \{0\}$.

- Shaffer's dynamic procedure: it increases the power of the first by substituting α/t_i at stage i by the value α/t_i^*, where t_i^* is the maximum number of hypotheses that could be true, given that the previous hypotheses are false. It is a dynamic procedure since t_i^* depends not only on the logical structure of the hypotheses, but also on the hypotheses already rejected at step i. Obviously, this procedure has more power than the first one. In this paper, we have not used this second procedure, given that it is included in an advanced procedure which we will describe in the following.

In Bergmann and Hommel (1988) was proposed a procedure based on the idea of finding all elementary hypotheses which cannot be rejected. In order to formulate Bergmann-Hommel's procedure, we need the following definition.

Definition 1 *An index set of hypotheses $I \subseteq \{1, ..., m\}$ is called* exhaustive *if exactly all H_j, $j \in I$, could be true.*

In order to exemplify the previous definition, we will consider the following case: We have three classifiers, and we will compare them in a $n \times n$ comparison. We will obtain three hypotheses:

- H_1 = C_1 es equal in behavior than C_2.
- H_2 = C_1 es equal in behavior than C_3.
- H_3 = C_2 es equal in behavior than C_3.

and eight possible sets S_i:

- S_1: All H_j are true.
- S_2: H_1 and H_2 are true and H_3 is false.
- S_3: H_1 and H_3 are true and H_2 is false.

- S_4: H_2 and H_3 are true and H_1 is false.
- S_5: H_1 is true and H_2 and H_3 are false.
- S_6: H_2 is true and H_1 and H_3 are false.
- S_7: H_3 is true and H_1 and H_2 are false.
- S_8: All H_j are false.

Sets S_1, S_5, S_6, S_7 and S_8 can be possible, because their hypotheses can be true at the same time, so they are exhaustive sets. Set S_2, basing on logically related hypotheses principles, is not possible because the performance of C_1 cannot be equal to C_2 and C_3, whereas C_2 has different performance than C_3. The same consideration can be done to S_3 and S_4, which are not exhaustive sets.

Under this definition, it works as follows.

- Bergmann and Hommel (1988) procedure: Reject all H_j with $j \notin A$, where the *acceptance set*

$$A = \bigcup \{I : I \textit{ exhaustive}, \ min\{P_i : i \in I\} > \alpha/|I|\}$$

  is the index set of null hypotheses which are retained.

  For this procedure, one has to check for each subset I of $\{1,...,m\}$ if I is exhaustive, which leads to intensive computation. Due to this fact, we will obtain a set, named E, which will contain all the possible exhaustive sets of hypotheses for a certain comparison. A rapid algorithm which was described in Hommel and Bernhard (1994) allows a substantial reduction in computing time. Once the E set is obtained, the hypotheses that do not belong to the A set are rejected.

  Figure 1 shows a valid algorithm for obtaining all the exhaustive sets of hypotheses, using as input a list of classifiers C. E is a set of families of hypotheses; likewise, a family of hypotheses is a set of hypotheses. The most important step in the algorithm is the number 6. It performs a division of the classifiers into two subsets, in which the last classifier k always is inserted in the second subset and the first subset cannot be empty. In this way, we ensure that a subset yielded in a division is never empty and no repetitions are produced. For example, suppose a set C with three classifiers $C = \{1,2,3\}$. All possible divisions without taking into account the previous assumptions are: $D_1 = \{C_1 = \{\}, C_2 = \{1,2,3\}\}, D_2 = \{C_1 = \{1\}, C_2 = \{2,3\}\}, D_3 = \{C_1 = \{2\}, C_2 = \{1,3\}\}, D_4 = \{C_1 = \{1,2\}, C_2 = \{3\}\}, D_5 = \{C_1 = \{3\}, C_2 = \{1,2\}\}, D_6 = \{C_1 = \{1,3\}, C_2 = \{2\}\}, D_7 = \{C_1 = \{2,3\}, C_2 = \{1\}\}, D_8 = \{C_1 = \{1,2,3\}, C_2 = \{\}\}$. Divisions D_1 and D_8, D_2 and D_7, D_3 and D_6, D_4 and D_5 are equivalent, respectively. Furthermore, divisions D_1 and D_8 are not interesting. Using the assumptions in step 6 of the algorithm, the possible divisions are: $D_1 = \{C_1 = \{1\}, C_2 = \{2,3\}\}, D_2 = \{C_1 = \{2\}, C_2 = \{1,3\}\}, D_3 = \{C_1 = \{1,2\}, C_2 = \{3\}\}$. In this case, all the divisions are interesting and no repetitions are yielded. The computational complexity of the algorithm for obtaining exhaustive sets is $O(2^{n^2})$. However, the computation requirements may be reduced by means of using storage capabilities. Relative exhaustive sets for $k-i$, $1 \leq i \leq (k-2)$ classifiers can be stored in memory and there is no necessity of invoking the *obtainingExhaustive* function recursively. The computational complexity using storage capabilities is $O(2^n)$, so the algorithm still requires intensive computation.

An example illustrating the algorithm for obtaining all exhaustive sets is drawn in Figure 2. In it, four classifiers, enumerated from 1 to 4 in the C set, are used. The comparisons or hypotheses are denoted by pairs of numbers without a separation character between them. This illustration does not show the case in which the set $|C_i| < 2$, for simplifying the representation. When $|C_i| < 2$, no comparisons can be performed, so the *obtainExhaustive* function returns an empty set E.

An edge connecting two boxes represents an invocation of this function. In each box, the list of classifiers given as input and the first initialization of the E set are displayed. The main edges, whose starting point is the initial box, are labeled by the order of invocation. Below the graph, the resulting E subset in each main edge is denoted. The final E will be composed by the union of these E subsets. At the end of the process, 14 distinct exhaustive sets are found: $E = \{(12,13,14,23,24,34),(23,24,34),(13,14,34),(12,14,24),(12,13,23),$ $(12),(13),(14),(23),(24),(34),(12,34),(13,24),(23,14)\}$.

Table 1 gives the number of hypotheses (m), the number (2^n) of index sets I and the number of exhaustive index sets (n_e) for k classifiers being compared.

Function obtainExhaustive($C = \{c_1, c_2, ..., c_k\}$: list of classifiers)

1. Let $E = \emptyset$
2. $E = E \cup \{$set of all possible and distinct pairwise comparisons using $C\}$
3. If $E == \emptyset$
   4. Return E
5. End if
6. For all possible divisions of C into two subsets C_1 and C_2, $c_k \in C_2$ and $C_1 \neq \emptyset$
   7. $E_1 = obtainExhaustive(C_1)$
   8. $E_2 = obtainExhaustive(C_2)$
   9. $E = E \cup E_1$
   10. $E = E \cup E_2$
   11. For each family of hypotheses e_1 of E_1
      12. For each family of hypotheses e_2 of E_2
         13. $E = E \cup (e_1 \cup e_2)$
      14. End for
   15. End for
16. End for
17. Return E

Figure 1: Algorithm for obtaining all exhaustive sets

The following subsections present a case study of a $n \times n$ comparison of some well-known classifiers over thirty data sets. In it, the four procedures explained above are employed.

2.2 Performing All Pairwise Comparisons: A Case Study

In the following, we show an example involving the four procedures described with a comparison of five classifiers: C4.5 (Quinlan, 1993); One Nearest Neighbor (1-NN) with Euclidean distance,

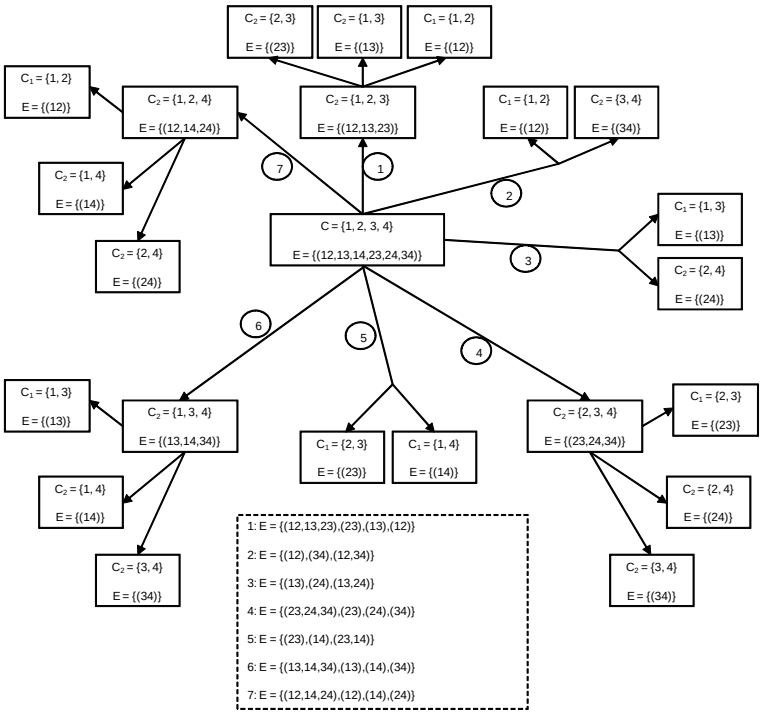


Figure 2: Example of the obtaining of exhaustive sets of hypotheses considering 4 classifiers

k	$m = \binom{k}{2}$	2^m	n_e
4	6	64	14
5	10	1024	51
6	15	32768	202
7	21	2097152	876
8	28	$2.7 \cdot 10^8$	4139
9	36	$6.7 \cdot 10^{10}$	21146

Table 1: All pairwise comparisons of k classifiers

NaiveBayes, Kernel (McLachlan, 2004)[1] and, finally, CN2 (Clark and Niblett, 1989).[2] The parameters used are specified in Section 4. We have used 10-fold cross validation and standard parameters for each algorithm. The results correspond to average accuracy or $1 - class\_error$ in test data. We have used 30 data sets.[3] Table 2 shows the overall process of computation of average rankings.

Friedman (1937) and Iman and Davenport (1980) tests check whether the measured average ranks are significantly different from the mean rank $R_j = 3$. They respectively use the χ^2 and the F statistical distributions to determine if a distribution of observed frequencies differs from the theoretical expected frequencies. Their statistics use nominal (categorical) or ordinal level data, instead of using means and variances. Demšar (2006) detailed the computation of the critical values in each distribution. In this case, the critical values are 9.488 and 2.45, respectively at $\alpha = 0.05$, and the Friedman's and Iman-Davenport's statistics are:

$$\chi^2_F = 39.647, F_F = 14.309.$$

Due to the fact that the critical values are lower than the respective statistics, we can proceed with the post-hoc tests in order to detect significant pairwise differences among all the classifiers. For this, we have to compute and order the corresponding statistics and p-values. The standard error in the pairwise comparison between two classifiers is $SE = \sqrt{\frac{k(k+1)}{6N}} = \sqrt{\frac{5 \cdot 6}{6 \cdot 30}} = 0.408$. Table 3 presents the family of hypotheses ordered by their p-value and the adjustment of α by Nemenyi's, Holm's and Shaffer's static procedures.

- Nemenyi's test rejects the hypotheses [1–4] since the corresponding p-values are smaller than the adjusted α's.
- Holm's procedure rejects the hypotheses [1–5].
- Shaffer's static procedure rejects the hypotheses [1–6].
- Bergmann-Hommel's dynamic procedure first obtains the exhaustive index set of hypotheses. It obtains 51 index sets. We can see them in Table 4. From the index sets, it computes the A set.[4] It rejects all hypotheses H_j with $j \notin A$, so it rejects the hypotheses [1–8].

  Bergmann-Hommel's dynamic procedure allows to clearly distinguishing among three groups of classifiers, attending to their performance:

- Best classifiers: C4.5 and NaiveBayes.
- Middle classifiers: 1-NN and CN2.
- Worst classifier: Kernel.

1. Kernel method is a bayesian classifier which employs a non-parametric estimation of density functions through a gaussian kernel function. The adjustment of the covariance matrix is performed by the ad-hoc method.
2. NaiveBayes and CN2 are classifiers for discrete domains, so we have discretized the data prior to learning with them. The discretizer algorithm is Fayyad and Irani (1993).
3. Data sets marked with '*' have been subsampled being adapted to slow algorithms, such as CN2.
4. We have considered that each classifier follows the order: 1 - C4.5, 2 - 1-NN, 3 - NaiveBayes, 4 - Kernel, 5 - CN2. For example, the hypothesis 13 represents the comparison between C4.5 and NaiveBayes.

	C4.5	1-NN	NaiveBayes	Kernel	CN2
Abalone*	0.219 (3)	0.202 (4)	0.249 (2)	0.165 (5)	0.261 (1)
Adult*	0.803 (2)	0.750 (4)	0.813 (1)	0.692 (5)	0.798 (3)
Australian	0.859 (1)	0.814 (4)	0.845 (2)	0.542 (5)	0.816 (3)
Autos	0.809 (1)	0.774 (3)	0.673 (4)	0.275 (5)	0.785 (2)
Balance	0.768 (3)	0.790 (2)	0.727 (4)	0.872 (1)	0.706 (5)
Breast	0.759 (1)	0.654 (5)	0.734 (2)	0.703 (4)	0.714 (3)
Bupa	0.693 (1)	0.611 (3)	0.572 (4.5)	0.689 (2)	0.572 (4.5)
Car	0.915 (1)	0.857 (3)	0.860 (2)	0.700 (5)	0.777 (4)
Cleveland	0.544 (2)	0.531 (4)	0.558 (1)	0.439 (5)	0.541 (3)
Crx	0.855 (2)	0.796 (4)	0.857 (1)	0.607 (5)	0.809 (3)
Dermatology	0.945 (3)	0.954 (2)	0.978 (1)	0.541 (5)	0.858 (4)
German	0.725 (2)	0.705 (4)	0.739 (1)	0.625 (5)	0.717 (3)
Glass	0.674 (4)	0.736 (1)	0.721 (2)	0.356 (5)	0.704 (3)
Hayes-Roth	0.801 (1)	0.357 (4)	0.520 (2.5)	0.309 (5)	0.520 (2.5)
Heart	0.785 (2)	0.770 (3)	0.841 (1)	0.659 (5)	0.759 (4)
Ion	0.906 (2)	0.359 (5)	0.895 (3)	0.641 (4)	0.918 (1)
Led7Digit	0.710 (2)	0.402 (4)	0.728 (1)	0.120 (5)	0.674 (3)
Letter*	0.691 (2)	0.827 (1)	0.667 (3)	0.527 (5)	0.638 (4)
Lymphography	0.743 (3)	0.739 (4)	0.830 (1)	0.549 (5)	0.746 (2)
Mushrooms*	0.990 (1.5)	0.482 (5)	0.941 (3)	0.857 (4)	0.990 (1.5)
OptDigits*	0.867 (3)	0.098 (1)	0.915 (2)	0.986 (1)	0.784 (4)
Satimage*	0.821 (3)	0.872 (2)	0.815 (4)	0.885 (1)	0.778 (5)
SpamBase*	0.893 (2)	0.824 (4)	0.902 (1)	0.739 (5)	0.885 (3)
Splice*	0.799 (2)	0.655 (4)	0.925 (1)	0.517 (5)	0.755 (3)
Tic-tac-toe	0.845 (1)	0.731 (2)	0.693 (4)	0.653 (5)	0.704 (3)
Vehicle	0.741 (1)	0.701 (2)	0.591 (5)	0.663 (3)	0.619 (4)
Vowel	0.799 (2)	0.994 (1)	0.603 (4)	0.269 (5)	0.621 (3)
Wine	0.949 (4)	0.955 (2)	0.989 (1)	0.770 (5)	0.954 (3)
Yeast	0.555 (3)	0.505 (4)	0.569 (1)	0.312 (5)	0.556 (2)
Zoo	0.928 (2.5)	0.928 (2.5)	0.945 (1)	0.419 (5)	0.897 (4)
average rank	2.100	3.250	2.200	4.333	3.117

Table 2: Computation of the rankings for the five algorithms considered in the study over 30 data sets, based on test accuracy by using ten-fold cross validation

i	hypothesis	$z = (R_0 - R_i)/SE$	p	α_{NM}	α_{HM}	α_{SH}
1	C4.5 vs. Kernel	5.471	$4.487 \cdot 10^{-8}$	0.005	0.005	0.005
2	NaiveBayes vs. Kernel	5.226	$1.736 \cdot 10^{-7}$	0.005	0.0055	0.0083
3	Kernel vs. CN2	2.98	0.0029	0.005	0.0063	0.0083
4	C4.5 vs. 1NN	2.817	0.0048	0.005	0.0071	0.0083
5	1NN vs. Kernel	2.654	0.008	0.005	0.0083	0.0083
6	1NN vs. NaiveBayes	2.572	0.0101	0.005	0.01	0.0125
7	C4.5 vs. CN2	2.49	0.0128	0.005	0.0125	0.0125
8	NaiveBayes vs. CN2	2.245	0.0247	0.005	0.0167	0.0167
9	1NN vs. CN2	0.327	0.744	0.005	0.025	0.025
10	C4.5 vs. NaiveBayes	0.245	0.8065	0.005	0.05	0.05

Table 3: Family of hypotheses ordered by p-value and adjusting of α by Nemenyi (NM), Holm (HM) and Shaffer (SH) procedures, considering an initial $\alpha = 0.05$

Size 1	Size 2	Size 3	Size 4	Size ≥ 6
(12)	(12,34)	(12,13,23)	(12,13,23,45)	(12,13,14,15,23,24,25,34,35,45)
(13)	(13,24)	(12,14,24)	(12,14,24,35)	(12,13,14,23,24,34)
(23)	(14,23)	(13,14,34)	(12,34,35,45)	(12,13,15,23,25,35)
(14)	(12,35)	(23,24,34)	(13,14,25,34)	(12,14,15,24,25,45)
(24)	**(13,25)**	(12,15,25)	(13,15,24,35)	(13,14,15,34,35,45)
(34)	(15,23)	(13,15,35)	(13,24,25,45)	(23,24,25,34,35,45)
(15)	(12,45)	(23,25,35)	(14,15,23,45)	
(25)	(13,45)	(14,15,45)	(14,23,25,35)	
(35)	(23,45)	(24,25,45)	(15,23,24,34)	
(45)	(14,25)	(34,35,45)		
	(15,24)			
	(14,35)			
	(24,35)			

Table 4: Exhaustive sets obtained for the case study. Those belonging to the *Acceptance* set (A) are typed in bold.

In Demšar (2006), we can find a discussion about the power of Hochberg's and Hommel's procedures with respect to Holm's one. They reject more hypothesis than Holm's, but the differences are in practice rather small (Shaffer, 1995). The most powerful procedures detailed in this paper, Shaffer's and Bergmann-Hommel's, work following the same method of Holm's procedure, so it is possible to hybridize them with other types of step up procedures, such as Hochberg's, Hommel's and Rom's methods. When we apply these methods by using the logical relationships among hypothesis in a static way, they do not control the family-wise error (Hochberg and Rom, 1995). In opposite, when applying these methods by detecting dynamical relationships, they control the family-wise error. In Hochberg and Rom (1995), several extensions were given in this way. Furthermore, a small improvement of power in the Bergmann-Hommel procedure described here can be achieved when using Simes conjecture (Simes, 1986) in the obtaining of A set (see Hommel and Bernhard, 1999, for more details).

3. Adjusted P-Values

The smallest level of significance that results in the rejection of the null hypothesis, the p-value, 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 important, it does this without committing to a particular level of significance.

When a p-value is within a multiple comparison, as in the example in Table 3, it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons belonging to the family. One way to solve this problem is to report APVs which take into account that multiple tests are conducted. An APV can be compared directly with any chosen significance level α. In this paper, we encourage the use of APVs due to the fact that they provide more information in a statistical analysis.

In the following, we will explain how to compute the APVs depending on the post-hoc procedure used in the analysis, following the indications given in Wright (1992) and Hommel and Bernhard (1999). We also include the post-hoc tests explained in Demšar (2006) and other for comparisons with a control classifier. The notation used in the computation of the APVs is the following:

- Indexes i and j correspond each one to a concrete comparison or hypothesis in the family of hypotheses, according to an incremental order by 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 j-th hypothesis.
- k is the number of classifiers being compared.
- m is the number of possible comparisons in an all pairwise comparisons design; that is, $m = \frac{k \cdot (k-1)}{2}$.
- t_j is the maximum number of hypotheses which can be true given that any $(j-1)$ hypotheses are false (see the description of Shaffer's static procedure in Section 2.1).

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

- one-step.
  - Bonferroni APV_i: $min\{v;1\}$, where $v=(k-1)p_i$.
  - Nemenyi APV_i: $min\{v;1\}$, where $v=m\cdot p_i$.
- step-up.
  - Hochberg APV_i: $max\{(k-j)p_j : (k-1)\geq j\geq i\}$.
  - Hommel APV_i: see algorithm at Figure 3.
- step-down.
  - Holm APV_i (using a control classifier): $min\{v;1\}$, where $v=max\{(k-j)p_j : 1\leq j\leq i\}$.
  - Nemenyi APV_i: $min\{v;1\}$, where $v=m\cdot p_i$.
  - Holm APV_i (using it in all pairwise comparisons): $min\{v;1\}$, where $v=max\{(m-j+1)p_j : 1\leq j\leq i\}$.
  - Shaffer static APV_i: $min\{v;1\}$, where $v=max\{t_j p_j : 1\leq j\leq i\}$.
  - Bergmann-Hommel APV_i: $min\{v;1\}$, where $v=max\{|I|\cdot min\{p_j, j\in I\} : I\ exhaustive, i\in I\}$.

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

Figure 3: Algorithm for calculating APVs based on Hommel's procedure

Table 5 shows the results in the final form of APVs for the example considered in this section. As we can see, this example is suitable for observing the difference of power among the test procedures. Also, this table can provide information about the state of retainment or rejection of any hypothesis, comparing its associated APV with the level of significance previously fixed.

i	hypothesis	p_i	APV_{NM}	APV_{HM}	APV_{SH}	APV_{BH}
1	C4.5 vs .Kernel	$4.487 \cdot 10^{-8}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$	$4.487 \cdot 10^{-7}$
2	NaiveBayes vs .Kernel	$1.736 \cdot 10^{-7}$	$1.736 \cdot 10^{-6}$	$1.563 \cdot 10^{-6}$	$1.042 \cdot 10^{-6}$	$1.042 \cdot 10^{-6}$
3	Kernel vs .CN2	0.0029	0.0288	0.023	0.0173	0.0115
4	C4.5 vs .1NN	0.0048	0.0485	0.0339	0.0291	0.0291
5	1NN vs .Kernel	0.008	0.0796	0.0478	0.0478	0.0319
6	1NN vs .NaiveBayes	0.0101	0.1011	0.0506	0.0478	0.0319
7	C4.5 vs .CN2	0.0128	0.1276	0.0511	0.0511	0.0383
8	NaiveBayes vs .CN2	0.0247	0.2474	0.0742	0.0742	0.0383
9	1NN vs .CN2	0.744	1.0	1.0	1.0	1.0
10	C4.5 vs .NaiveBayes	0.8065	1.0	1.0	1.0	1.0

Table 5: APVs obtained in the example by Nemenyi (NM), Holm (HM), Shaffer's static (SH) and Bergmann-Hommel's dynamic (BH)

4. Experimental Framework

In this section, we want to determine the power and behavior of the studied procedures through the experiments in which we repeatedly compared the classifiers on sets of ten randomly chosen data sets, recording the number of equivalence hypothesis rejected and APVs. We follow a similar method used in Demšar (2006).

The classifiers used are the same as in the case study of the previous subsection: C4.5 with minimum number of item-sets per leaf equal to 2 and confidence level fitted for optimal accuracy and pruning strategy, naive Bayesian learner with continuous attributes discretized using Fayyad and Irani (1993) discretization, classic 1-Nearest-Neighbor classifier with Euclidean distance, CN2 with Fayyad-Irani's discretizer, star size = 5 and 95% of examples to cover and Kernel classifier with $sigmaKernel = 0.01$, which is the inverse value of the variance that represents the radius of neighborhood. All classifiers are available in KEEL software (Alcalá-Fdez et al., 2008).[5]

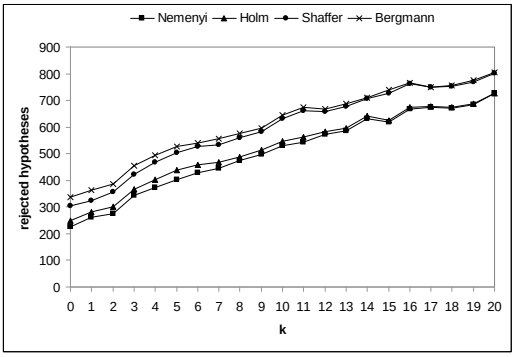
For performing this study, we have compiled a sample of fifty data sets from the UCI machine learning repository (Asuncion and Newman, 2007), all of them valid for a classification task.[6] We measured the performance of each classifier by means of accuracy in test by using ten-fold cross validation. As Demšar did, when comparing two classifiers, samples of ten data sets were 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 in the classification accuracies on that data set and k is the bias through which we can regulate the differences between the classifiers. With $k = 0$, the selection is purely random and as k is being higher, the selected data sets are favorable to a particular classifier.

In comparisons of multiple classifiers, samples of data sets have to be selected with the probabilities computed from the differences in accuracy of two classifiers. We have chosen C4.5 and 1-NN, due to the fact that we have found significant differences between them in the study conducted before (Section 2.2) which involved thirty data sets. Note that the repeated comparisons done here only involve ten data sets each time, so the rejection of equivalence of two classifiers is more difficult at the beginning of the process.

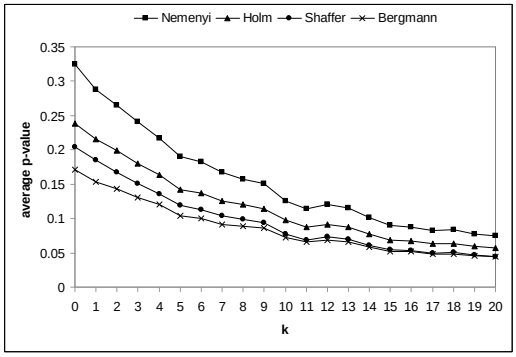
5. It is also available at `http://www.keel.es`.

6. The data sets used are: abalone, adult, australian, autos, balance, bands, breast, bupa, car, cleveland, dermatology, ecoli, flare, german, glass, haberman, hayes-roth, heart, iris, led7digit, letter, lymphography, magic, monks, mushrooms, newthyroid, nursery, optdigits, pageblocks, penbased, pima, ring, satimage, segment, shuttle, spambase, splice, tae, thyroid, tic-tac-toe, twonorm, vehicle, vowel, wine, wisconsin, yeast, zoo.

Figure 4 shows the results of this study considering the pairwise comparison between C4.5 and 1-NN. It gives an approximation of the power of the statistical procedures considered in this paper. Figure 4(a) reflects the number of times they rejected the equivalence of C4.5 and 1-NN. Obviously, the Bergmann-Hommel procedure is the most powerful, followed by Shaffer's static procedure. The graphic also informs us about the use of logically related hypothesis, given that the procedures that use this information have a bias towards the same point and those which do not use this information, tend to a lower point than the first. When the selection of data sets is purely random ($k = 0$), the benefit of using the Bergmann-Hommel procedure is appreciable. Figure 4(b) shows the average APV of the same comparison of classifiers. As we can see, the Nemenyi procedure is too conservative in comparison with the remaining procedures. Again, the benefit of using more sophisticated testing procedures is easily noticeable.



(a) Number of hypotheses rejected in pairwise comparisons



(b) Average APV in pairwise comparisons

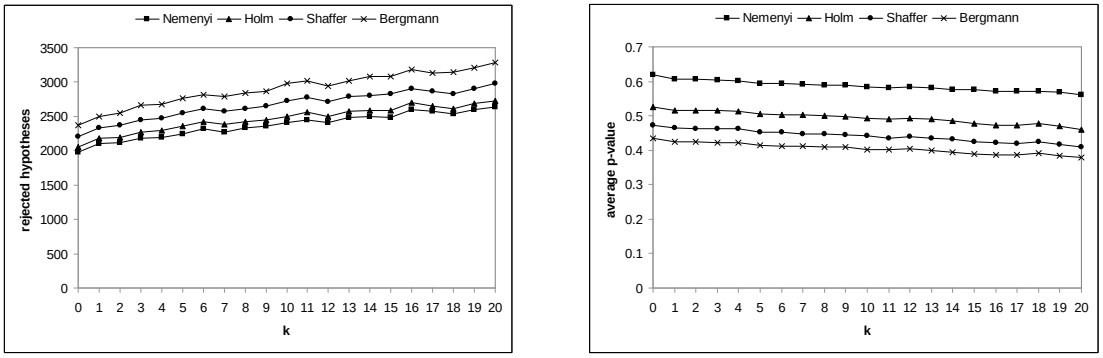
Figure 4: C4.5 vs. 1-NN

Figure 5 shows the results of this study considering all possible pairwise comparisons in the set of classifiers. It helps us to compare the overall behavior of the four testing procedures. Figure 5(a) presents the number of times they rejected any comparison belonging to the family. Although it could seem that the selection of data sets determined by the difference of accuracy between two classifiers may not influence on the overall comparison, the graphic shows us that it occurs. Furthermore, the lines drawn follow a parallel behavior, indicating us the relation and magnitude of power among the four procedures. In Figure 5(b) we illustrate the average APV for all the comparisons of classifiers. We can notice that the conservatism of the Nemenyi test is obvious with respect to the rest of procedures. The benefit of using a more advanced testing procedure is similar with respect to the following less-powerful procedure, except for Holm's procedure.

Finally, our recommendation on the usage of a certain procedure depends on the results obtained in this paper and in our experience in understanding and implementing them:

- We do not recommend the use of Nemenyi's test, because it is a very conservative procedure and many of the obvious differences may not be detected.

- When we use a considerable number of data sets with regards to number of classifiers, we could proceed with the Holm procedure.



(a) Total number of hypotheses rejected

(b) Average APV in all comparisons

Figure 5: All comparisons

- However, conducting the Shaffer static procedure means a not very significant increase of the difficulty with respect to the Holm procedure. Moreover, the benefit of using information about logically related hypothesis is noticeable, thus we strongly encourage the use of this procedure.

- Bergmann-Hommel's procedure is the best performing one, but it is also the most difficult to understand and computationally expensive. We recommend its usage when the situation requires so (i.e., the differences among the classifiers compared are not very significant), given that the results it obtains are as valid as using other testing procedure.

5. Conclusions

The present paper is an extension of Demšar (2006). Demšar does not deal in depth with some topics related to multiple comparisons involving all the algorithms and computations of adjusted p-values.

In this paper, we describe other advanced testing procedures for conducting all pairwise comparisons in a multiple comparisons analysis: Shaffer's static and Bergmann-Hommel's procedures. The advantage that they obtain is produced due to the incorporation of more information about the hypotheses to be tested: in $n \times n$ comparisons, a logical relationship among them exists. As a general rule, the Bergmann-Hommel procedure is the most powerful one but it requires intensive computation in comparisons involving numerous classifiers. The second one, Shaffer's procedure, can be used instead of Bergmann-Hommel's in these cases. Moreover, we present the methods for obtaining the adjusted p-values, which are valid p-values associated to each comparison useful to be compared with any level of significance without restrictions and they also provide more information. We have illustrated them with a case study and we have checked that the new described methods are more powerful than the classical ones, Nemenyi's and Holm's procedures.

Acknowledgments

This research has been supported by the project TIN2005-08386-C05-01. S. García holds a FPU scholarship from Spanish Ministry of Education and Science. The present paper was submitted as a regular paper in the JMLR journal. After the review process, the action editor Dale Schuurmans encourages us to submit the paper to the special topic on Multiple Simultaneous Hypothesis Testing. We are very grateful to the anonymous reviewers and both action editors who managed this paper for their valuable suggestions and comments to improve its quality.

Appendix A. Source Code of the Procedures

The source code, written in JAVA, that implements all the procedures described in this paper, is available at `http://sci2s.ugr.es/keel/multipleTest.zip`. The program allows the input of data in CSV format and obtains as output a LaTeX document.

References

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, and F. Herrera. KEEL: A software tool to assess evolutionary algorithms to data mining problems. *Soft Computing. doi: 10.1007/s00500-008-0323-y*, 2008. In press.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Anaylisis and Machine Intelligence*, 29(1):173–180, 2007.

G. Bergmann and G. Hommel. Improvements of general multiple test procedures for redundant systems of hypotheses. In P. Bauer, G. Hommel, and E. Sonnemann, editors, *Multiple Hypotheses Testing*, pages 100–115. Springer, Berlin, 1988.

P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

S. Esmeir and S. Markovitch. Anytime learning of decision trees. *Journal of Machine Learning Research*, 8:891–933, 2007.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029. Morgan-Kaufmann, 1993.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

N. García-Pedrajas and C. Fyfe. Immune network based ensembles. *Neurocomputing*, 70(7-9): 1155–1166, 2007.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75: 800–802, 1988.

Y. Hochberg and D. Rom. Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference*, 48:141–152, 1995.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure. *Biometrika*, 75:383–386, 1988.

G. Hommel and G. Bernhard. A rapid algorithm and a computer program for multiple test procedures using procedures using logical structures of hypotheses. *Computer Methods and Programs in Biomedicine*, 43:213–216, 1994.

G. Hommel and G. Bernhard. Bonferroni procedures for logically related hypotheses. *Journal of Statistical Planning and Inference*, 82:119–128, 1999.

R. L. Iman and J. M. Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics*, pages 571–595, 1980.

C. Marrocco, R. P. W. Duin, and F. Tortorella. Maximizing the area under the ROC curve by pairwise feature combination. *Pattern Recognition*, 41:1961–1974, 2008.

G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Mathematical Statistics, 2004.

J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado. Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning Research*, 6:783–816, 2005.

P. B. Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, 1963.

M. Núñez, R. Fidalgo, and R. Morales. Learning in environments with unknown dynamics: Towards more robust concept learners. *Journal of Machine Learning Research*, 8:2595–2628, 2007.

A. B. Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8: 761–773, 2007.

J. R. Quinlan. *Programs for Machine Learning*. Morgan Kauffman, 1993.

D. M. Rom. A sequentially rejective test procedure based on a modified bonferroni inequality. *Biometrika*, 77:663–665, 1990.

J.P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831, 1986.

J.P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, 2003.

R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73: 751–754, 1986.

P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons, 2004.

S. P. Wright. Adjusted p-values for simultaneous inference. *Biometrics*, 48:1005–1013, 1992.

Y. Yang, G. Webb, K. Korb, and K. M. Ting. Classifying under computational resource constraints: anytime classification using probabilistic estimators. *Machine Learning*, 69:35–53, 2007a.

Y. Yang, G. I. Webb, J. Cerquides, K. B. Korb, J. Boughton, and K. M. Ting. To select or to weigh: A comparative study of linear combination schemes for superparent-one-dependence estimators. *IEEE Transcations on Knowledge and Data Engineering*, 19(12):1652–1665, 2007b.

J. H. Zar. *Biostatistical Analysis*. Prentice Hall, 1999.