Selecting an Appropriate Statistical Test for Comparing Multiple Experiments in Evolutionary Machine Learning

José Otero
Dept. de Informática
Universidad de Oviedo
jotero@lsi.uniovi.es

Luciano Sánchez
Dept. de Informática
Universidad de Oviedo
luciano@uniovi.es

Jesús Alcalá-Fdez
Dept. de Informática
Universidad de Jaén
jalcala@decsai.ugr.es

Abstract

It is widely admitted that any experimentation in Genetic Machine Learning must be accompanied by a suitable statistical test. When making pairwise comparisons, a de facto standard is testing that the expectation of the differences between the empirical errors of the algorithms being compared is not null. A t-test is used if the gaussian assumption holds, or a non parametric test otherwise.

But, when multiple comparisons are needed, there is not such an agreement in the methodology that one must follow. In this work we survey the most relevant bibliography in this area and also outline some conclusions on our own, about the power and Type I error estimations of the best approaches. Our recommendations will be based on an extensive empirical analysis, where synthetic data (and, therefore, with known theoretical properties) was used, so that all possible conCEgurations of the null hypothesis are accounted.

1 Introduction

There are many papers in Evolutionary Machine Learning literature showing comparisons between several algorithms. Often, the authors compare their own method with the most relevant ones. As stated in [8], there are different factors that make the use of a statistical test mandatory in this kind of competition: the error metric employed, the election of training and test sets and the nature of the algorithm when this is not deterministic. A typical experimentation in Evolutionary Machine Learning is driven by questions similar to the following ones:

- Is the proposed algorithm the one with the least medium error?
  - If so, are the differences with the other algorithms statistically signiCEcant? If they are, the proposed algorithm is the best among the analyzed.
  - If the differences with the best of the other algorithms are not statistically signiCEcant, the proposed algorithm has equivalent error to those algorithms.
- If the proposed algorithm is not the one with the least medium error,
  - Are the differences with the best algorithms statistically signiCEcant? If not, the proposed algorithm is equivalent (in error) to the best one.
  - Otherwise, we can conclude that the proposed algorithm is not the best one.

This sequence of steps involves a composite assert, comprising the simultaneous comparison of an algorithm with every other one. This leads to some kind of mean comparison test and a Multiple Comparison Procedure (MCP) [11, 13].

1.1 MCP test in evolutionary Machine Learning

In particular, the study here presented addresses those experiments that consist on solving a problem (chosen from the usual catalog in the literature of this area [2]) using an implementation of an algorithm, with the goal of determining which is the one with the lowest error. In the following, the *experimental design* is given by the set of problems

to solve, the measurements performed, the implementation details and the context of experiment realization [3]. It is remarked that there is not consensus in the Evolutionary Machine Learning community about which are the statistical test and the experimental designs to use in the papers of this area [8, 15, 19, 23] and some argue that frequently the data lacks of one or more of the properties needed to apply a given test [4, 19]. Even more, when several algorithms are compared [16], this implies the use of specific statistical test [11, 13]. Additionally, some researchers argue that it is impossible to extract conclusions about the performance of a given algorithm from the data sets used normally [12].

Apart from these considerations, it is widely acknowledged that a suitable test must be included in every experimentation involving more than two algorithms. However, the use of MCP tests in Evolutionary Machine Learning is not so widespread as it is in other research areas (e.g. analysis of clinical trials.) Nevertheless, in order to make a bibliographic study we can make use of the methodologies used in these areas [5], because the experiments are similar. Therefore, we have conducted a review of the most relevant literature in statistical tests, which we will try to apply to Evolutionary Machine Learning experiments. We have payed special attention to the verification of the parametric conditions of the data, in order to use the correct test in each situation, as proposed in [19, 20] or, alternatively, employ a bootstrap technique to empirically obtain the distribution of the statistics [21, 22].

1.2 Summary

The exposition is organized in three parts. In Section 2, the taxonomy of the most relevant MCP s is done. In Section 3, the application of this kind of test to Evolutionary Machine Learning experiments is explained. Finally, in Section 4 an empirical study of a selection of this tests is done, using *multifold cross validation experimental setup* in a synthetic problem with known solution. Some conclusions about the power and Type I error estimations follow this point. The experiments follow the technique proposed in [9] where all configurations of truth (or Null Hypothesis) are tested. This is important because of the behavior of a test may depend on how many of the algorithms tested *actually* have equivalent or different mean test error.

2 Multiple Comparisons Procedures, bibliographic survey.

There are some books devoted to this topic [11, 13] even in conjunction with resampling [22]. In these books can be found most of the MCP s commented here. There are also some review articles as [6, 7, 17, 18] focused in MCP s. In the Evolutionary Machine Learning area, the most remarkable cite is [14] and outside of this area but with application (it belongs to area of clinical trials) is [5].

The researcher willing to compare his own method with the most relevant in the state of the art, can employ a series of simultaneous Hypothesis Testing, if r is the number of algorithms:

$$H_{0i} : \mu_1 = \mu_i$$
$$H_{1i} : \mu_1 \neq \mu_i$$
$$i \in 2..r$$

where the algorithm with index equal to 1 is compared with the following $r - 1$ algorithms.

The important thing is that the researcher wishes to assess (with a low probability of making a mistake) something like "algorithm 1 has different mean error than algorithm 2, but it is equivalent to 3, ...etc", that is, an affirmation composed of *simultaneous* individual assertions about the result of each test. In this situation, when simultaneous hypothesis testing is done, the multiplicative effect of Type I error appears: if Type I error (reject null hypothesis when it is true, in our case, find differences when there are not such) probability is α, the probability of no mistake in the n tests are $(1 - \alpha)^n$, then the probability of *at least* one mistake is $\alpha_t = 1 - (1 - \alpha)^n$. This effect, well known in Statistics [11, 13] it s usually ignored in Evolutionary Machine Learning literature [19]. There are two classic alternatives to overcame this effect, originally proposed by Fisher, the so called "two-step methods", the "one-step methods" along with a more recent third class called "multi-step methods" [11] or "sequential methods".

Alternatively, the methods can be classified involving the nature of the testing being performed,

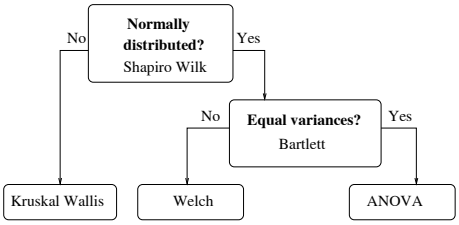

Figure 1: Graphic representation of statistical test sequence to be applied depending on the parametric conditions satisfied.

all with all, all with the best, all against a control,...[13].

2.1 Two-step methods.

Two-step methods try to *protect* the MCP against Type I error performing first an F test (analysis of variance or the non-parametric counterpart). In this first stage, the null hypothesis is that there are no statistically significant differences in the results. The alternative hypothesis is that at least there are two algorithms that have different mean error. If there are r experiments in total:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_r$$
$$H_1 : \mu_i \neq \mu_j \text{ for some } i, j \in 1..r$$

It is necessary to verify the parametric conditions of the data in order to decide which is the sequence of test to be applied, as shown in Figure 1. In this work, independence between samples is assumed.

If the null hypothesis is rejected (there are statistically significant differences), then the proposed algorithm is compared with the others, using the same confidence level as in the previous test. For each rejected hypothesis, it is assumed that a difference has been found. If the null hypothesis is not rejected, then no additional test is performed and the conclusion is that there are no differences between the algorithms being compared. Fisher LSD (Least Significant Difference) test belongs to this kind of procedure [4].

It is important that the protection against Type I error occurs only under the null hypothesis, that is, if there are no differences at all, the probability to erroneously reject the null hypothesis equals the significance level of the first test. But if there are some differences, the probability of erroneously reject at least one true null hypothesis will exceed that level. Note that this means to find differences that doesn t exist. This is called to protect the test *in the weak sense* [11, 13].

2.2 One-step methods.

One-step methods account for multiplicity of the Type I error and adjust the significance level of each individual test (or the corresponding p-values) with the aim that the MCP Type I error does not exceed the desired value. If the significance level of the MCP is α_t, then for each individual test the level can be fixed at $\alpha = 1 - (1 - \alpha_t)^{1/n}$ (being n the number of comparisons), known as Dunn-Sidák adjustment. If $(1 - \alpha)^n$ is approximated by $1 - n\alpha$ then $\alpha = \alpha_t/n$, known as Bonferroni adjustment. This adjustments guard the MCP against Type I error *in strong sense*, that is, in all possible configurations of true and false null hypothesis [11, 13].

2.3 Sequential methods.

The power of a statistical test (one minus Type II error probability, failing to reject a null hypothesis when it is false) decreases as α decreases too. Because of this, one-step methods tend to be conservative [11, 13]. Sequential methods try to overcame this problem by sequentially adjust the level of the individual test. The idea behind this is as follows, the p-values are ordered (from greater to smaller or vice versa, depends on the method), then the Bonferroni or Dunn-Sidák adjustment is applied and a new MCP is constructed, possibly extracting one of the individual test. Because the new MCP has less hypothesis, the effect of p-value adjustment is smaller. To this class of test belong Holm method, Simes-Hochberg method or Hommel method. These tests protect against Type I error in the strong sense and their power does not decrease as much as one step method does.

Finally, in Evolutionary Machine Learning experiments, not only it is important to control the Type I error probability, it is important to control the *number* of true null hypothesis incorrectly rejected, controlling the False Discovery Rate (FDR). In some works after the seminal work [1], techniques to control the number of incorrectly rejected

hypothesis are proposed [11, 13].

3 MCP s in Evolutionary Machine Learning experiments.

In the context of this work, it must be explained how MCP s are employed to answer the questions pointed in section 1.
We need to obtain a sample of mean error from each algorithm tested, solving a known problem. Given the size of standard datasets (e.g. [2]) this implies the use of cross-validation or resampling experimental designs. In this work we consider only cross-validation, in order to minimize the overlapping of train sets between samples [8, 19].
The researcher must decide then how to protect the MCP against Type I error. This leads to perform a previous F type test or a p-value adjustment.

If the researcher performs a previous F test, then it is necessary to carefully test if the parametric conditions of a given test are fulfilled. In Figure 1 are shown the corresponding F test to each set of parametric conditions, lack of independence is not considered here. If the null hypothesis of the F test is rejected, then there are statistically significant differences in the mean samples. In this case, a test of means comparison is performed between the algorithm proposed and each of the competitors. The significance level of these tests is the same as the F test. Here it is again necessary to verify the parametric conditions of the test applied, as stated in [20]. In Figure 2 it is shown the sequence of test to be applied for each set of parametric conditions. The result of the mean comparison tests performed is a set of p-values, which size is the number of algorithms being compared minus one. If the corresponding p-value of a given comparison is less than the significance level chosen, then it is assumed that a statistical significant difference exist between the corresponding algorithms. In this point, the answers to the questions pointed in Section 1 can be found:

- If the proposed algorithm has the least mean error:
  - If the null hypothesis of the F test is rejected and also the null hypothesis of the test (means comparison between this algorithm and the best of the others,) then the proposed algorithm is the best.
  - If the null hypothesis of the F test is rejected and not the null hypothesis of the test, then the proposed algorithm is equivalent to the best.
  - If the null hypothesis of the F test is not rejected, all the algorithms are equivalent.
- If the proposed algorithm has not the least mean error:
  - If the null hypothesis of the F test is rejected and also the null hypothesis of the test, then the proposed algorithm is not the best.
  - If the null hypothesis of the F test is rejected and not the null hypothesis of the test, then the proposed algorithm is equivalent to the best.
  - If the null hypothesis of the F test is not rejected, all the algorithms are equivalent.

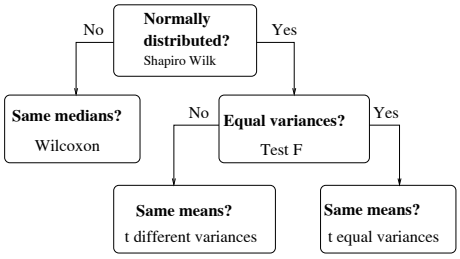

Figure 2: Graph of statistical tests used in combination with k-fold cross validation experimental design.

If the researcher does not perform the previous F test and performs p-value adjustment, he must pick one of the methods commented. After the application of the chosen method, a set of p-values is obtained, and each null hypothesis of the mean comparison test can be rejected or not, with a signification level that is different than the desired α level of the MCP and guarantees that the type I error of the MCP is under that value.
Finally, if the aim of the researcher is to control the number of true null hypothesis incorrectly rejected, he must pick one of the methods that control the FDR.

In these two later situations, the questions stated before can be also answered, the procedure is similar to the two-step method, without the previous F test.

4 Empirical analysis.

We are not aware of works where there is a guide describing the systematic use of MCP s for comparing Evolutionary Machine Learning algorithms. The most remarkable cite advocating the use of these tests is [14], but it is not focused in the topics of this work. In particular, we want to analyze the dependence between the type I error and the power vs. the the number of partitions, for different signi-Ecation levels. The results of the experiments will clarify if there are differences that make us to prefer classical MCP s or those that adjust the p-values.

4.1 Experiments

The objective of the empirical analysis is, as mentioned before, determining the best test for its use in Evolutionary Machine Learning experiments. In particular, we will estimate the power and Type I error of some of the tests commented before, in a typical Evolutionary Machine Learning test problem.

In the experimentation described in this section we have compared 10 algorithms, using two of the tests commented before, each one belonging to one of the families of MCP s. Each test was repeated 100 times for each one of the conditions to be compared. In each repetition it is examined if the test correctly distinguishes all the algorithms with different mean error. The fraction of times that this happens is an estimation of the power of the test. Also, in each repetition is examined if the test incorrectly rejects some of the true null hypothesis, i.e. it -Ends inexistent differences. The fraction of times that this happens is an estimation of Type I error probability.
In order to test all possible situations that could happen in a experiment, all possible con-Egurations of truth are tested, following [9]. This means that for $i \in 1 \ldots 9$ *better* algorithms, are generated $j \in 1 \ldots ...9-i$ algorithms *equivalent* to the *control* (the proposed algorithm) and $k = 9-i-j$ *worse* algorithms. The total number of truth con-Egurations for N algorithms being compared is $N\frac{N+1}{2}$.
The problem used in the empirical analysis, de-Ened in [10], consists on a sample of size 500 of a population where there are two classes with equal probability, following a bidimensional normal distribution. The means are $(0, 0)$ and $(2, 0)$. The covariance matrices are I and $4I$, respectively.
The optimal classi-Ger for this problem is quadratic, but we need to obtain a family of classi-Gers with increasing error, from the better ones to the control and equivalents and to the worse. This is achieved by corrupting the labels of the test partition with increasing probability after training with the corresponding partition untouched.
The test to be compared are:

- Holm method [11, 13].
- F test followed by the tests proposed in [20].

Each test is performed at 0.01, 0.05, 0.1 and with 10, 30, 50, 100 partitions.

4.2 Results

The numerical results of power estimation of the compared tests are shown in table 1, where the mean of the results of the 100 repetitions of each combination of signi-Ecance level and number of partitions is shown for the whole set of truth con-Egurations with 10 algorithms being compared (55 con-Egurations in total). As can be seen, the power estimation of both tests are similar, but the power of the modi-Ged LSD test is always higher. In Figure 3 the same data is shown as boxplots, this aids the reader to compare the mean and dispersion of the data or the presence of outliers. As can be seen, Holm s method exhibits higher dispersion of power estimation. Modi-Ged LSD shows less dispersion but outliers appear on right column (0.1 signi-Ecance level). The medians of both tests are comparable.

The numerical results of Type I error estimation of the compared tests are shown in table 2. In this table the mean of the results of the 100 repetitions of each combination of signi-Ecance level and number of partitions is shown to for the whole set of truth con-Egurations. In this case, Holm s method beats modi-Ged LSD. This fact is an experimental evidence of the theoretical analysis found in [11, 13], where the use of LSD alike tests was discouraged, since they do not protect the test *in strong sense*,

Folds	S. Level		
	0.01	0.05	0.10
10	0.477 0.569	0.643 0.746	0.725 0.821
30	0.554 0.623	0.682 0.761	0.742 0.832
50	0.575 0.642	0.690 0.768	0.749 0.831
100	0.565 0.629	0.684 0.764	0.750 0.836

Table 1: Means of power estimations. Each cell shows Holm (left) and modified LSD (right).

Folds	S. Level		
	0.01	0.05	0.10
10	0.001 0.008	0.011 0.038	0.024 0.080
30	0.001 0.007	0.012 0.038	0.026 0.079
50	0.001 0.006	0.010 0.039	0.023 0.078
100	0.002 0.005	0.009 0.037	0.023 0.075

Table 2: Means of Type I error estimations. Each cell shows Holm (left) and modified LSD (right).

as we commented earlier in this paper. Note that if there is no effective protection against Type I error, the researcher can t make assertions about the mean error of the algorithms being compared. Moreover, there is a high probability of mistake when the researcher claims to find differences between any two algorithms.

In table 3, the number of occasions (from all the possible truth configurations with 10 algorithms, that is, 55 configurations) in that the estimated Type I error (from 100 repetitions) exceeds MCP signification level. It is clear that this is much more frequent for LSD alike test than in Holm test. Obviously, if the number of repetitions were infinite, the estimation of Type I error would equal the significance level in the case of Holm test, thus all the values belonging to that column in table 3 would be equal to zero. In a real experimentation, there is no knowledge about the true configuration of the hypothesis (which ones are really true and which ones are false), and the researcher has no information on the behavior of the modified LSD test regarding the significance level of the whole MCP.

Folds	S. Level		
	0.01	0.05	0.10
10	1 9	1 18	1 19
30	1 9	5 17	4 15
50	0 8	1 15	1 17
100	2 4	2 13	2 14

Table 3: Number of times, from all truth configurations with 10 algorithms, in that the estimated Type I error exceeds the signification level. Holm test (left), modified LSD (right).

5 Conclusions and future work.

As exposed in section 1, there is no consensus in which experimental designs and which statistical tests should be applied in Evolutionary Machine Learning experiments. In this work we follow [20], where the parametric conditions must be observed in order to apply the correct statistical test in each situation.

Additionally, when many algorithms are being compared simultaneously, and statements involving all the algorithms are to be made, a MCP is needed [19]. According to our experiments, if a researcher in Evolutionary Machine Learning wants to protect the MCP against Type I error, the classic approach consisting in a previous F test must be abandoned and a one-step or sequential procedure [11, 13] must be employed. If the researcher wishes to control the number of erroneously rejected individual null hypothesis (i.e. false differences), the procedures proposed in [1] must be employed.

Lastly, it is remarked that this survey of the statistical tests suitable to Evolutionary Machine Learning experiments is not complete. At least, the procedures that control the FDR must be analyzed. Finally, given the results found in [21], the development of a MCP that relies in individual bootstrap tests seems to be convenient.

6 Acknowledgment

This work was funded by Spanish M. of Education, under the grant TIN2005-08386-C05-05.

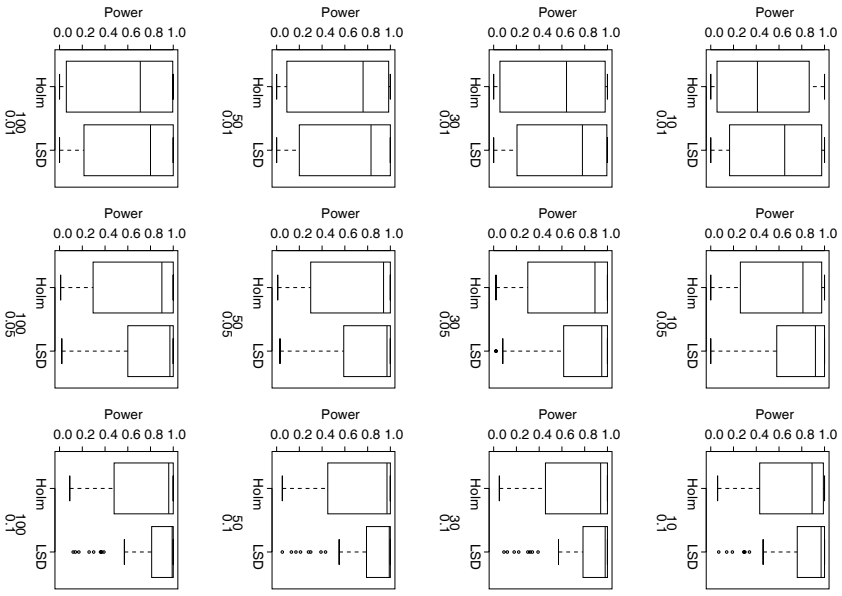

Figure 3: Boxplots of power estimations. Each subpicture shows Holm (left) and modified LSD (right). At the left of each plot is indicated the number of folds and the significance level.

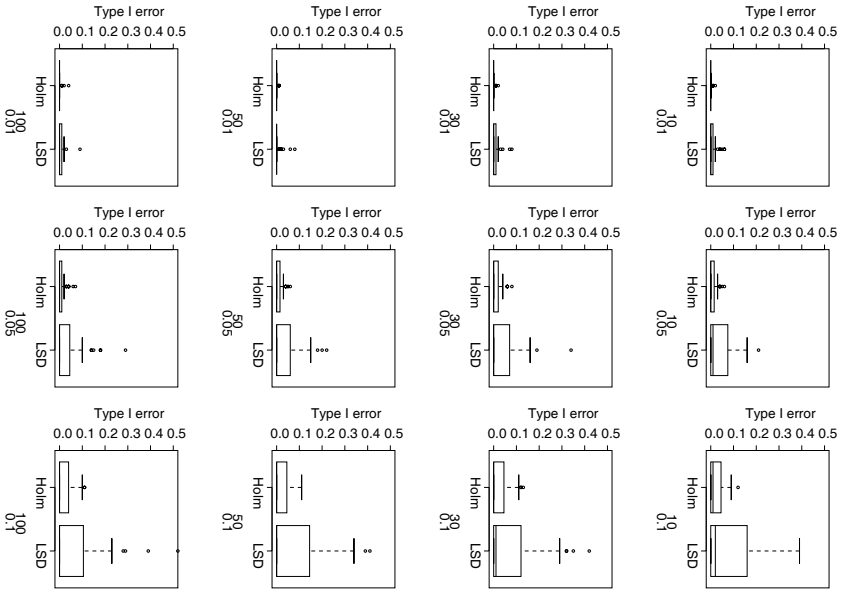

Figure 4: Boxplots of Type I error estimations. Each subpicture shows Holm (left) and modified LSD (right). At the left of each plot is indicated the number of folds and the significance level.

References

[1] Benjamini, Y., Hochberg, T. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. B 57(1): 289-300. 1995

[2] Blake, C.L., Merz, C.J. UCI Repository of machine learning databases. http://www.ics.uci.edu. University of California, Department of Information and Computer Science. 1998

[3] Cochran W.G., Cox G.M. Experimental Designs. Wiley. 1992

[4] Cohen, P.R., Empirical Methods for Artificial Intelligence. MIT Press. 1995

[5] Cook R.J., Farewell V.T. Multiplicity Considerations in the Design and Analysis of Clinical Trials. J. R. Statist. Soc. A. 159, Part 1: 93-110. 1996

[6] Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol 279(1): R1-R8. 2000.

[7] Sandrine Dudoit, Mark J. van der Laan, Katherine S. Pollard. Multiple Testing I, II, III. U.C. Berkeley Division of Biostatistics Working Paper Series Year 2003 Paper 138, 139, 140.

[8] Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10(7): 1895-1923. 1998

[9] Einot, I. and Gabriel, K. R. A Study of the Powers of Several Methods of Multiple Comparison, Journal of the American Statistical Association 70(351): 574-583. 1975

[10] Haykin, S. Neural Networks, A Comprehensive Foundation.Prentice Hall. 1999

[11] Hochberg, Y., Tamhane A. C., Multiple comparison procedures, John Wiley & Sons, Inc. 1987

[12] Holte R. C.: Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning 11(1): 63-90. 1993

[13] Hsu, J.C. (1996). Multiple Comparisons: Theory and Methods, Chapman and Hall,London.

[14] Jensen D., Cohen P. R. Multiple Comparisons in Induction Algorithms. Machine Learning 38(3): 309-338. 2000

[15] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of International Joint Conference on Artificial Intelligence (1995)

[16] Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning 40(3): 203-228. 2000

[17] O Brien PC and Shampo MA: Statistical considerations for performing multiple tests in a single experiment. Mayo Clinic Proceedings. 63:813-815. 1988

[18] Rafter J.A., Abell M. L., James P. Braselton. Multiple Comparison Methods for Means. SIAM Review, Society for Industrial and Applied Mathematics 44(2): 259-278. 2002

[19] Salzberg S. L. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data minig and Knowledge Discovery 1(3): 317-328. 1997

[20] Herrera F., Hervás C., Otero J., Sánchez L.. Un estudio empírico preliminar sobre los tests estadísticos más habituales en el aprendizaje automático. Tendencias de la Minería de Datos en España 403-412. Lecture Notes in Computer Science. ed. Digital @3D.

[21] Sánchez L., Otero J., Alcalá J. Assesing the differences in accuracy between GFSs with bootstrap tests. Actas del congreso Eusocat 2005 1021-1026

[22] Westfall P.H., Young S.S. Resampling-based Multiple Testing. Wiley, New York. 1993

[23] Whitley D., Watson J.P., Howe A., Barbulescu L. Testing, Evaluation and Performance of Optimization and Learning Systems. Keynote Address: Adaptive Computing in Design and Manufacturing. 2002

Actas de las I Jornadas sobre Algoritmos Evolutivos y Metaheurísticas

JAEM'07

Editadas por
Enrique Alba
Francisco Chicano
Francisco Herrera
Francisco Luna
Gabriel Luque
Antonio J. Nebro

Zaragoza, 12 y 13 de Septiembre de 2007