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,