ScienceDirect

Information Sciences

Volume 180, Issue 10, 15 May 2010, Pages 2044-2064

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]

Show more

https://doi.org/10.1016/j.ins.2009.12.010

Get rights and content

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.

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 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, 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.

Section snippets

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,…

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 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 …

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,…

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…

…

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…

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 …

…

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…

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.…

References (54)

A hybrid decision tree/genetic algorithm method for data mining
Information Sciences (2004)

L.J. Eshelman
The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination

Z. Lei *et al.*
Designing of classifiers based on immune principles and fuzzy rules
Information Sciences (2008)

F. Martínez-Estudillo *et al.*
Evolutionary product-unit neural networks classifiers
Neurocomputing (2008)

I. Partalas *et al.*
Greedy regression ensemble selection: theory and an application to water quality prediction
Information Sciences (2008)

V. Rivas *et al.*
Evolving RBF neural networks for time-series forecasting with EvRBF
Information Sciences (2004)

D. Shilane *et al.*
A general framework for statistical performance comparison of evolutionary computation algorithms
Information Sciences (2008)

C.-J. Tsai *et al.*
A discretization algorithm based on class-attribute contingency coefficient
Information Sciences (2008)

S. Tsumoto
Contingency matrix theory: statistical dependence in a contingency table
Information Sciences (2009)

A. Ulaş *et al.*
Incremental construction of classifier and discriminant ensembles
Information Sciences (2009)

View more references

Cited by (1625)

A label noise filtering method for regression based on adaptive threshold and noise score

2023, Expert Systems with Applications

Show abstract

Application of asymmetric proximal support vector regression based on multitask learning in the stock market

2023, Expert Systems with Applications

Show abstract

A reconstructed feasible solution-based safe feature elimination rule for expediting multi-task lasso

2023, Information Sciences

Show abstract

X-MODE: Extended Multi-operator Differential Evolution algorithm

2023, Mathematics and Computers in Simulation

Show abstract

CKD.Net: A novel deep learning hybrid model for effective, real-time, automated screening tool towards prediction of multi stages of CKD along with eGFR and creatinine

2023, Expert Systems with Applications

Show abstract

Pneumothorax prediction using a foraging and hunting based ant colony optimizer assisted support vector machine

2023, Computers in Biology and Medicine

Show abstract

View all citing articles on Scopus

Recommended articles (6)

Research article

Grey wolf optimizer with cellular topological structure

Expert Systems with Applications, Volume 107, 2018, pp. 89-114

Show abstract

Research article

Boosted hunting-based fruit fly optimization and advances in real-world problems

Expert Systems with Applications, Volume 159, 2020, Article 113502

Show abstract

Research article

Social mimic optimization algorithm and engineering applications

Expert Systems with Applications, Volume 134, 2019, pp. 178-191

Show abstract

Research article

A Novel Approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics

Information Sciences, Volume 417, 2017, pp. 186-215

Show abstract

Research article

Multilevel threshold image segmentation with diffusion association slime mould algorithm and Renyi's entropy for chronic obstructive pulmonary disease

Computers in Biology and Medicine, Volume 134, 2021, Article 104427

Show abstract

Research article

Chaotic simulated annealing multi-verse optimization enhanced kernel extreme learning machine for medical diagnosis

Computers in Biology and Medicine, Volume 144, 2022, Article 105356

Show abstract

View full text