A Study on the Use of Non-Parametric Tests for Experimentation with Cluster Analysis

Deepika Singh
M. Tech (Research Scholar), Galgotia College of Engineering and Technology, Greater Noida, INDIA

ABSTRACT
In this paper, we focus on the experimental analysis on the performance in cluster analysis with the use of non-parametric tests on the clustering task. Particularly, we have studied whether the sample of results from multiple trials obtained by conventional clustering algorithms checks the necessary conditions for being analyzed through parametrical tests. The study is conducted by considering the possibilities on clustering experiments. The study obtained state that a parametric statistical analysis could not be appropriate specially when we deal with multiple-problem results. In multiple-problem analysis, we propose the use of non-parametric statistical tests given that they are less restrictive than parametric ones and they can be used over small size samples of results. These conditions are problem-dependent and indefinite, which justifies the need of using non-parametric statistics in the experimental analysis.

Keywords: Statistical analysis of experiments, Evolutionary algorithms, Parametric tests, Non-parametric tests, Clustering.

I. INTRODUCTION

The experimental analysis on the performance of a new method is a crucial and necessary task to carry out in a research on Data Mining, Computational Intelligence techniques. Deciding when an algorithm is better than other one may not be a trivial task. There has been a growing interest for the experimental analysis in the field of evolutionary algorithms. It is noticeable due to the existence of numerous papers which analyze and propose different types of problems, such as the basis for experimental comparisons of algorithms, proposals of different methodologies in comparison or proposals of use of different statistical techniques in algorithms comparison. In this paper we are studying the differences faced by the conventional analysis i.e. parametric tests and non-parametric tests. The experimental method used is clustering which is one of the efficient techniques of data mining.

1.1 STATISTICAL ANALYSIS
Testing (usually called ‘hypothesis testing’) play a major role in statistical investigation. In statistical testing, we are concerned with examining the truth or otherwise, of hypothesis (assumptions, claims, guesses, etc.) about some feature(s) of one or more populations. Almost all large and small sample tests such as t, F and χ2 are based on the assumptions that the parent population (from which the sample is drawn) has a specific distribution, such as normal distribution. The distributions are usually defined through some parameters. Non-parametric tests do not require such assumption. Hence nonparametric tests are also known as distribution free tests. The term nonparametric refers to the fact that there are no parameters involved in the traditional sense of the term parameter used generally. Nonparametric test statistics utilize some simple aspects of sample data such as the signs of measurements, order relationships or category frequencies. Therefore, stretching or compressing the scale does not alter them. As a consequence, the null distribute on of the nonparametric test statistic can be determined without regard to the shape of the parent population distribution. The inferences drawn from tests based on the parametric tests such t, F and χ2 may be seriously affected when the parent population distributions is not normal. These effects could be more if when sample size is small. Thus when there is doubt about the distribution of the parent population, a nonparametric method should be used. In many situations particularly in social and behavioural sciences observations are difficult or impossible to take on numerical scales. Nonparametric tests are well suited under such situations.

First step in statistical testing is formulation of a hypothesis. A hypothesis is a statement about the population. Its plausibility is evaluated on the basis of information obtained by sampling from the population. A test generally involves two hypotheses. An assertion about the population in favor of the ‘existing’ situations is taken as null hypothesis and denoted as H0. The negation of the null hypothesis is known as alternative hypothesis and...
denoted as H1. H1 plays a decisive role in classifying a test as one-sided or two sided. We first develop a statistic T (say) on the basis of the sample observations. The statistic T decides whether to reject or accept the null hypothesis. Usually T follows some distribution. Based on this distribution the range of T is divided into two groups; the critical region and the region of acceptance. If the sample point falls in critical region, we reject null hypothesis. The size of the critical region depends on the importance we wish to incur which ultimately gives the significance level of the test. It is denoted by α. It represent the probability of rejecting the null hypothesis when it is true, also known as Type I error. Type II error is the probability of rejecting H0 when H1 is true and denoted by β. Commonly used significance levels are 5% and 1% (α= .05 and .01). Finally, a conclusion is drawn on the basis of the value of T falling or not falling in the critical region.

1.2 CONDITIONS FOR THE SAFE USE OF NONPARAMETRIC TESTS

In order to distinguish a nonparametric test from a parametric one, we must check the type of data used by the test. A nonparametric test is that which uses nominal or ordinal data. This fact does not force it to be used only for these types of data. It is possible to transform the data from real values to ranking based data. In such way, a non-parametric test can be applied over classical data of parametric test when they do not verify the required conditions imposed by the test.

As a general rule, a non-parametric test is less restrictive than a parametric one, although it is less robust than a parametric when data are well conditioned. Parametric tests have been commonly used in the analysis of experiments. For example, a common way to test whether the difference between two algorithms' results is non-random is to compute a paired t-test, which checks whether the average difference in their performance over the data sets is significantly different from zero. When comparing a set of multiple algorithms, the common statistical method for testing the differences between more than two related sample means is the repeated-measures ANOVA (or within-subjects ANOVA). The "related samples" are again the performances of the algorithms measured across the same problems. The null-hypothesis being tested is that all classifiers perform the same and the observed differences are merely random. Unfortunately, Parametric tests are based on assumptions which are most probably violated when analyzing the performance of computational intelligence and data mining algorithms. These assumptions are:

**Independence**: In statistics, two events are independent when the fact that one occurs does not modify the probability of the other one occurring.

- When we compare two optimization algorithms they are usually independent.
- When we compare two machine learning methods, it depends on the partition.

- The independency is not truly verified in 10-fcv (a portion of samples is used either for training and testing in different partitions.
- Hold out partitions can be safely take as independent, since training and test partitions do not overlap.

**Normality**: An observation is normal when its behaviour follows a normal or Gauss distribution with a certain value of average μ and variance σ. A normality test applied over a sample can indicate the presence or absence of this condition in observed data.

- Kolmogorov–Smirnov: it compares the accumulated distribution of observed data with the accumulated distribution expected for a Gaussian distribution, obtaining the p-value based on both discrepancies. Therefore, it is a quality of fit procedure that can be used to test the hypothesis of normality in the population distribution. However, this method performs poorly because it possess very low power.
- Shapiro–Wilk: it analyzes the observed data for computing the level of symmetry and kurtosis (shape of the curve) in order to compute the difference with respect to a Gaussian distribution afterwards, obtaining the p-value from the sum of the squares of these discrepancies. The power of this test has been shown to be excellent. However, the performance of this test is adversely affected in the common situation where there is tied data.
- D’Agostino–Pearson: it first computes the skewness and kurtosis to quantify how far from Gaussian the distribution is in terms of asymmetry and shape. It then calculates how far each of these values differs from the values expected with a Gaussian distribution, and computed a single p-value form the sum of these discrepancies. The performance of this test is not as good as that of Shapiro–Wilk’s procedure, but it is not affected by tied data.

### II. ANALYSIS OF CLUSTER ANALYSIS ALGORITHM

Since 80’s the clustering algorithm began to process data, the research of clustering algorithms has not stopped. Facing the different application requirements, the researchers suggest many kind of clustering algorithm, so it can be seen in clustering algorithm has many applications. The clustering algorithm can be defined by a simple to describe “through clustering operation, data object is divided into subsets in order to as far as possible similarity in the same subset and as far as dissimilar in the different subset.”

#### 2.1 Concepts of Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and
are “dissimilar” to the objects belonging to other clusters. We can show this with a simple graphical example:

![Figure 1: A Simple Graphical Example](image)

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Cluster analysis originates “like attracts like “the simple idea, so the clustering or classification is according to the data characteristic. So a high quality clustering algorithm must satisfy the two conditions: the similarity of data or the object in same subset is the strongest; the similarity of data or object in different subset is the weak. The quality of clustering usually depends on similarity measuring method and the realization of the way used by the clustering algorithm, but also depends on the algorithm it can find all or part of the hidden pattern.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

### 2.2 Analyses of Clustering Algorithms

Now summarizing the existing clustering methods, we have boiled it down to the following categories:

1. **Partitioning Methods**, a given database that contains n data object or tuple generates the number for the K cluster partition based on clustering algorithm. The standard of Division (or known similarity function) usually called Euclidean distance, for the categorical data attributes you can use the Jaccard coefficient. All K-means and Fuzzy C-means is the most famous two in this method, these clustering methods are very applicable for finding globular clusters mall in medium database. But in order to clustering for large-scale data set, and clustering complex shape, the method needs to be further expanded.

2. **Hierarchical Methods**, the method is decompose a given data object set for many levels, so it will build a clustering tree. According to the hierarchical decomposition is based on bottom-up or top-down principle, it can be further divided into condensed and division. The former to each individual as a separate class and it process with the data similarity, then a sufficiently large data gradually merged into larger categories; the latter conversely, the entire set as a category, and then gradually divided into small different types. In order to make up for no traceable deficiency of decomposition or polymerization, hierarchical clustering method often combine some other methods, such as circular positioning. The typical hierarchical clustering methods such as BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

3. **Density-based Methods**, it see cluster as a high density area what is splitting a space by low density region. The basic idea is following: as long as the adjacent region of the dot density (number of data points) beyond a certain threshold, will continue to clustering, till the field must contain at least a certain number of points. The algorithm is based on the sample of neighborhood conditions, the whole sample space is split by low density interval, and it does not need to know the cluster number in advance, it complete the clustering task through one scan. The difference of Density-based Methods and other methods of a fundamental is that it is based on the density not on the variety of distance, so it can overcome the situation that the algorithm based the distance can only be found in “dough “ cluster. There are some typical algorithms such as OPTICS (Ordering Points to Identify the Clustering Structure)

4. **grid-based method**, this method first is divided data space into a grid unit, and then mapped the data set of samples into the grid cell; each cell density will be calculated. According to given density threshold to judge whether each grid unit is high density unit, so several adjacent dense grid unit formed clusters. The main advantage of this approach is fast processing speed, which independent of the number of data objects, but is with each dimension unit number of the quantization space. At present the common grid clustering algorithm includes Wave Cluster

5. **model-based method**, the method assumes a model for every cluster, then looks for the model data sets it can well meet the model. Such as through the construction of density function that reflecting the spatial distribution to achieve clustering, its theory is that the data is generated according to the underlying probability distribution , this clustering method can optimize the given data and certain
mathematical model more adaptability, such as RSDE algorithm.

(6) Spectral-based Clustering. The method convert the data set into an undirected connected graph, the sample of data set is Vertex data of diagram, the edge weights can reflect the degree of similarity between the graph vertices. Then the clustering problem transfers into the graph partition problem. The best division effect is the weighted summation of maximum spanning between internal vertex of a sub graph, and minimize the weight of between the vertices of a graph, such as GRC (Graph-based Relaxed Clustering ) . So we can see there are many clustering algorithms, every method has its characteristics, and every kind of clustering algorithms can be seen its own useful in many application.

III. NONPARAMETRIC TESTS

In the discussion of the tests for comparisons of two methods over multiple cases of problems we will make two points. We shall warn against the widely used t-test as usually conceptually inappropriate and statistically unsafe. Another, even more rarely used test is the sign test which is weaker than the Wilcoxon test but also has its distinct merits. The other message will be that the described statistics measure differences between the methods from different aspects, so the selection of the test should be based not only on statistical appropriateness but also on what we intend to measure.

3.1 The Sign Test

A popular way to compare the overall performances of algorithms is to count the number of cases on which an algorithm is the overall winner. When multiple algorithms are compared, pairwise comparisons are sometimes organized in a matrix.

Some authors also use these counts in inferential statistics, with a form of binomial test that is known as the sign test. If the two algorithms compared are, as assumed under the null-hypothesis, equivalent, each should win on approximately N/2 out of N problems. The number of wins is distributed according to the binomial distribution; the critical number of wins can be found in next Table. For a greater number of cases, the number of wins is under the null-hypothesis distributed according to N(N/2,SQRT(N)/2), which allows for the use of z-test: if the number of wins is at least N/2+1.96*SQRT(N)/2 (or, for a quick rule of a thumb, N/2+SQRT(N)), the algorithm is significantly better with p < 0.05. Since tied matches support the null-hypothesis we should not discount them but split them evenly between the two methods; if there is an odd number of them, one is ignored.

\[
R^+ = \sum_{d_i > 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i)
\]

\[
R^- = \sum_{d_i < 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i)
\]

Wilcoxon’s test performs individual comparisons between two algorithms (pairwise comparisons). The p-value in a pairwise comparison is independent from another one. If we try to extract a conclusion involving more than one pairwise comparison in a Wilcoxon’s analysis, we will obtain an accumulated error coming from the combination of pairwise comparisons. In statistical terms, we are losing the control on the Family Wise Error Rate (FWER), defined as the probability of making one or more false discoveries among all the hypotheses when performing multiple pairwise tests. The true statistical significance for combining pairwise comparisons is given by:

\[
p = P(\text{Reject } H_0 | H_0 \text{ true})
\]

\[
= 1 - P(\text{Accept } H_0 | H_0 \text{ true})
\]

\[
= 1 - P(\text{Accept } A_k = A_i, i = 1, \ldots, k - 1 | H_0 \text{ true})
\]

\[
= 1 - \prod_{i=1}^{k-1} P(\text{Accept } A_k = A_i | H_0 \text{ true})
\]

\[
= 1 - \prod_{i=1}^{k-1} (1 - P(\text{Reject } A_k = A_i | H_0 \text{ true}))
\]

\[
= 1 - \prod_{i=1}^{k-1} (1 - p H_i)
\]

So, a pairwise comparison test, such as Wilcoxon’s test, should not be used to conduct various comparisons involving a set of algorithms, because the FWER is not controlled. The expression defined above
computes the true significance obtained after performing several comparisons, hence the level of significance cannot be set before performing the comparisons and the statistical significance cannot be known a priori.

In order to carry out a comparison which involves more than two methods, under the assumption of being significant at a certain level of significance, established previously to the statistical analysis, the multiple comparisons tests should be used. In this part, we describe the most used test for performing multiple test comparisons together with a set of post-hoc procedures to compare a control method with other methods (1 x n comparisons). We refer to the Friedman Test and derivatives

### 3.3 Multiple Sign Test

The multiple sign test, allows us to compare all of the other algorithms with a control labeled algorithm. It carries out the following steps:

1. Represent by \( x_{i, CONTROL} \) and \( x_{i,J} \), the performances of the control and the jth classifier in the ith data set.
2. Compute the signed differences \( d_{ij} = x_{i,J} - x_{i,CONTROL} \). In other words, pair each performance with the control and, in each data set, subtract the control performance from the jth classifier.
3. Let \( r_j \) equal the number of differences, \( d_{ij} \), that have the less frequently occurring sign (either positive or negative) within a pairing of an algorithm with the control.
4. Let \( M_1 \) be the median response of a sample of results of the control method and \( M_j \) be the median response of a sample of results of the jth algorithm

### 3.4 Contrast Estimation based on Medians

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

It carries out the following steps:

1. For every pair of \( k \) algorithms in the experiment, we compute the difference between the performances of the two algorithms in each of the \( n \) data sets. In other words, we compute the differences
   
   \[
   D_{i(u,v)} = X_{i,CONTROL} - X_{i,J},
   \]

   where \( i = 1; \ldots ; n; u = 1; \ldots ; k; \) and \( v = 1; \ldots ; k \). We form performance pairs only for those in which \( u < v \).
2. We find the median of each set of differences and call it \( Z_{uv} \). We call \( Z_{uv} \) the unadjusted estimator of \( M_u - M_v \). Since \( Z_{uv} = Z_{av} \), we have only to calculate \( Z_{av} \) for the case where \( u < v \). There are \( k(k - 1)/2 \) of these medians. Also note that \( Z_{uu} = 0 \).
3. We compute the mean of each set of unadjusted medians having the same first subscript and call the result \( m_u \); that is, we compute
   
   \[
   m_u = \frac{\sum_{j=1}^{k} Z_{uj}}{k}, \quad u = 1, \ldots, k
   \]
4. The estimator of \( M_u - M_v \) is \( m_u - m_v \), where \( u \) and \( v \) range from 1 through \( k \). For example, the difference between \( M_1 \) and \( M_2 \) is \( m_1 - m_2 \).

### 3.5 The Friedman Test

Friedman’s test is used for answering this question: In a set of \( k \) samples (where \( k \geq 2 \)), do at least two of the samples represent populations with different median values? It is a non-parametric procedure employed in a hypothesis testing situation involving a design with two or more samples. It is the analogous of the repeated measures ANOVA in non-parametrical statistical procedures; therefore, it is a multiple comparison test that aims to detect significant differences between the behavior of two or more algorithms.

The null hypothesis for Friedman’s test is \( H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \); the median of the population \( i \) represents the median of the population \( j \), \( i \neq j \), \( 1 \leq i \leq k \), \( 1 \leq j \leq k \). The alternative hypothesis is \( H_1 : \) Not \( H_0 \), so it is non-directional.

Next, we describe the tests computations. It computes the ranking of the observed results for algorithm \( (r_j \), for the algorithm \( j \) with \( k \) algorithms) for each function, assigning to the best of them the ranking 1, and to the worst the ranking \( k \). Under the null hypothesis, formed from supposing that the results of the algorithms are equivalent and, therefore, their rankings are also similar, the Friedman’s statistic

\[
\chi^2_F = \frac{12N}{k(k + 1)} \left[ \sum_{j=1}^{k} R_j^2 - \frac{k(k + 1)^2}{4} \right]
\]

distributed according to \( \chi^2 \) with \( k - 1 \) degrees of freedom, being \( R_j = \frac{1}{N} \sum_{i=1}^{n} r_{ij}^2 \), and \( N \) the number of cases of the problem considered. The critical values for the Friedman’s statistic coincide with the established in the \( \chi^2 \) distribution when \( N > 10 \) and \( k > 5 \).

### 3.6 The Iman and Davenport Test

Iman and Davenport proposed a derivation from the Friedman’s statistic given that this last metric produces a conservative undesirably effect. The proposed statistic is

\[
F_F = \frac{(N - 1) \chi^2_F}{N(k - 1) - \chi^2_F}
\]

and it is distributed according to a \( F \) distribution with \( k - 1 \) and \((k - 1)(N - 1)\) degrees of freedom.
Computation of the p-values given a $\chi^2$ or $F$ statistic can be done by using the algorithms in Abramowitz, M.: Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables. Dover, New York (1974). Also, most of the statistical software packages include it.

The rejection of the null hypothesis in both tests described above does not involve the detection of the existing differences among the algorithms compared. They only inform us about the presence of differences among all samples of results compared. In order to conducting pairwise comparisons within the framework of multiple comparisons, we can proceed with a post-hoc procedure. In this case, a control algorithm (maybe a proposal to be compared) is usually chosen. Then, the post-hoc procedures proceed to compare the control algorithm with the remain $k - 1$ algorithms. Next, we describe three post-hoc procedures:

### 3.7 Friedman Aligned Ranks Test

The Friedman test is based on $n$ sets of ranks, one set for each data set in our case; and the performances of the algorithms analyzed are ranked separately for each data set. Such a ranking scheme allows for intra-set comparisons only, since inter-set comparisons are not meaningful. When the number of algorithms for comparison is small, this may pose a disadvantage. In such cases, comparability among data sets is desirable and we can employ the method of aligned. In this technique, a value of location is computed as the average performance achieved by all algorithms in each data set. Then, it calculates the difference between the performance obtained by an algorithm and the value of location. This step is repeated for algorithms and data sets. The resulting differences, called aligned observations, which keep their identities with respect to the data set and the combination of algorithms to which they belong, are then ranked from 1 to $kn$ relative to each other. Then, the ranking scheme is the same as that employed by a multiple comparison procedure which employs independent samples; such as the Kruskal–Wallis test. The ranks assigned to the aligned observations are called aligned ranks.

The Friedman Aligned Ranks test statistic can be written as

$$T = \frac{(k - 1)\left[\sum_{j=1}^{k} \hat{R}_j^2 - (kn^2/4)(kn + 1)^2\right]}{\left\{k(nk + 1)(2kn + 1)/6\right\} - (1/k)\sum_{j=1}^{n} \hat{R}_j^2}$$

where $R_i$ is equal to the rank total of the $i$th data set and $R_j$ is the rank total of the $j$th algorithm.

The test statistic $T$ is compared for significance with a chi-square distribution for $k - 1$ degrees of freedom. Critical values can be found at. Furthermore, the p-value could be computed through normal approximations (Abramowitz, M.: Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables. Dover, New York (1974)). If the null hypothesis is rejected, we can proceed with a post hoc test.

### 3.8 Quade Test

The Friedman test considers all data sets to be equal in terms of importance. An alternative to this could take into account the fact that some data sets are more difficult or the differences registered on the run of various algorithms over them are larger. The rankings computed on each data set could be scaled depending on the differences observed in the algorithms’ performances. The Quade test conducts a weighted ranking analysis of the sample of results. The procedure starts finding the ranks $r_i$ in the same way as the Friedman test does. The next step requires the original values of performance of the classifiers $x_i$. Ranks are assigned to the data sets themselves according to the size of the sample range in each data set. The sample range within data set $i$ is the difference between the largest and the smallest observations within that data set:

$$\text{Range in data set } i = \max_j\{x_{ij}\} - \min_j\{x_{ij}\}$$

Obviously, there are $n$ sample ranges, one for each data set. Assign rank 1 to the data set with the smallest range, rank 2 to the second smallest, and so on to the data set with the largest range, which gets rank $n$. Use average ranks in case of ties. Let $Q_1, Q_2, \ldots , Q_n$ be the ranks assigned to data sets 1, 2, ..., $n$, respectively. Finally, the data set rank $Q_i$ is multiplied by the difference between the rank within data set $i$, $r_i$, and the average rank within data sets, $(k + 1) / 2$, to get the product $S_{ij}$, where

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

is a statistic that represents the relative size of each observation within the data set, adjusted to reflect the relative significance of the data set in which it appears. For convenience and to establish a relationship with the Friedman test, we will also use rankings without average adjusting:

$$\mathcal{W}_{ij} = Q_i \left[ r_i^j \right]$$

Let $S_i$ denote the sum for each classifier, Next we must to calculate the terms:

$$A_2 = n(n + 1)(2n + 1)(k)(k + 1)(k - 1)/72$$

$$B = \frac{1}{n} \sum_{j=1}^{k} S_{ij}^2$$

The test statistic is

$$T_3 = \frac{(n - 1)B}{A_2 - B}$$
which is distributed according to the F-distribution with \( k - 1 \) and \( (k - 1)(n - 1) \) degrees of freedom. Moreover, the p-value could be computed through normal approximations (Abramowitz, M.: Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables. Dover, New York (1974)). If \( A_2 = B \), consider the point to be in the critical region of the statistical distribution and calculate the p-value as \((1/k!)^{n-1}\). If the null hypothesis is rejected, we can proceed with a post hoc test.

IV. PROBLEM STUDIED

The study obtained state that a parametric statistical analysis could not be appropriate specially when we deal with multiple-problem results. In multiple-problem analysis, we propose the use of nonparametric statistical tests given that they are less restrictive than parametric ones and they can be used over small size samples of results. As in the cases of any data mining technique i.e. cluster analysis.

V. CONSIDERATIONS AND RECOMMENDATIONS ON THE USE OF NONPARAMETRIC TESTS

This section notes some considerations and recommendations concerning the nonparametric tests presented in this tutorial. Their characteristics as well as suggestions on some of their aspects and details of the multiple comparisons tests are presented. With this aim, some general considerations and recommendations are given first (Section 5.1). Then, some advanced guidelines for multiple comparisons with a control method (Section 5.2) and multiple comparisons among all methods (Section 5.3) are provided.

5.1. General considerations

- By using nonparametric statistical procedures, it is possible to analyze any unary performance measure (that is, associated to a single algorithm) with a defined range. This range does not have to be limited; thus, comparisons considering running times, memory requirements, and so on, are feasible.
- Being able to be applied in multi-domain comparisons, nonparametric statistical procedures can compare both deterministic and stochastic algorithms simultaneously, providing that their results are represented as a sample for each pair of algorithm domain.
- For the application of these methods, only a result for each pair of algorithm/domain is required. A known and standardized procedure must be followed to gather them, using average results from several executions when considering stochastic algorithms.
- An appropriate number of algorithms in contrast with an appropriate number of case problems are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of case problems. The previous statement may not be true for the Wilcoxon test. The influence of the number of case problems used is more noticeable in multiple comparison procedures than in Wilcoxon’s test.
- Although Wilcoxon’s test and the post-hoc tests for multiple comparisons are nonparametric statistical tests, they operate in a different way. The main difference lies in the computation of the ranking. Wilcoxon’s test computes a ranking based on differences between case problems independently, whereas the Friedman test and its derivative procedures compute the ranking between algorithms.
- In relation to the sample size (number of case problems when performing Wilcoxon’s or Friedman’s tests in a multi-problem analysis), there are two main aspects to be determined. First, the minimum sample size considered acceptable for each test needs to be stipulated. There is no established agreement about this specification. Statisticians have studied the minimum sample size when a certain power of the statistical test is expected. In our case, the employment of a sample size as large as possible is preferable because the power of the statistical tests (defined as the probability that the test will reject a false null hypothesis) will increase. Moreover, in a multi-problem analysis, the increase of the sample size depends on the availability of new case problems (which should be well known in computational intelligence or data mining field). Second, we have to study how the results are expected to vary if there were a larger sample size available. In all statistical tests used for comparing two or more samples, an increase of the sample size benefits the power of the test. In the following items, we will state that Wilcoxon’s test is less influenced by this factor than Friedman’s test. Finally, as a rule of thumb, the number of case problems in a study should be \( n = a \cdot k \), where \( a \geq 2 \).
- Although there is not a theoretical maximum number of domains to use in a comparison, it can be derived from the central limit theorem that, if this number is too high, the results may be unreliable. If the number of domains grows too much, statistical tests can lose credibility, as they may start highlighting true insignificant hypotheses as significant ones. For the Wilcoxon’s test, a maximum of 30 domains is suggested. For multiple comparisons, a value of \( n \geq 8 \cdot k \) could be too high, obtaining no significant comparisons as a result.
- Taking into account the previous observation and knowing the operations performed by the nonparametric tests, we can deduce that Wilcoxon’s test is influenced by the number of case problems used. On the other hand, both the number of algorithms and case problems are crucial when we refer to multiple comparisons tests (such as Friedman’s test), given that all the critical values depend on the value of \( n \) (see the expressions above). However, the increasing/decreasing of the number of case problems rarely affects the computation of the ranking. In these procedures, the number of functions used is an important factor to be considered when we want to control the FWER.
Another interesting procedure considered in this paper is related to Contrast Estimation based on medians between two samples of results. Contrast Estimation in nonparametric statistics is used for computing the real differences between two algorithms, considering the median measure the most important. Taking into account that the samples of results in computational intelligence experiments rarely fulfill the needed conditions for a safe use of parametric tests, the computation of nonparametric contrast estimation through the use of medians is very useful. For example, one could provide, apart from the average values of accuracies over various problems reported by the methods compared, the contrast estimation between them over multiple problems, which is a safer metric in multi-problem environments.

Finally, we want to remark that the choice of any of the statistical procedures presented in this paper for conducting an experimental analysis should be justified by the researcher. The use of the most powerful procedures does not imply that the results obtained by a given proposal will be better. The choice of a statistical technique is ruled by a trade-off between its power and its complexity when it comes to being used or explained to non-expert readers in statistics.

5.2. Multiple comparisons with a control method

A multiple comparison of various algorithms must be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. Once this test rejects the hypothesis of equivalence of means, the detection of the concrete differences among the algorithms can be done with the application of post-hoc statistical procedures, which are methods used for comparing a control algorithm with two or more algorithms.

An appropriate number of algorithms in contrast with an appropriate number of case problems are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of case problems. In general, p-values are lower on increasing the number of case problems used in multiple comparison procedures (so long as this number does not exceed \( n \geq 8 \cdot k \)); therefore, the differences among the algorithms are more detectable.

As we have suggested, multiple comparisons tests must be used when we want to establish a statistical comparison of the results reported among various algorithms. We focus on cases when a method is compared against a set of algorithms. It could be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. There are three alternatives: the Friedman test with the Iman–Davenport extension, the Friedman Aligned Ranks test, and the Quade test. Once one of these tests rejects the hypothesis of equivalence of medians, the detection of the specific differences among the algorithms can be made with the application of post-hoc statistical procedures, which are methods used for specifically comparing a control algorithm with two or more algorithms.

In this kind of test, it is possible to use just the rankings obtained when establishing a classification between the algorithms, and even employ them to measure their performance differences. However, this cannot be used to conclude that a given proposal outperform the rest, unless the null hypothesis is rejected.

Although, by definition, post-hoc statistical procedures can be applied in an independent way from the rejection of the null hypothesis, it is advisable to check this rejection firstly. Holm’s procedure can always be considered better than Bonferroni–Dunn’s procedure, because it appropriately controls the FWER and it is more powerful than Bonferroni–Dunn’s procedure. We strongly recommend the use of Holm’s method in a rigorous comparison. Nevertheless, the results offered by the Bonferroni–Dunn test are suitable to be visualized in graphical representations.

Hochberg’s procedure is more powerful than Holm’s procedure. The differences between it and Holm’s procedure are in practice rather small. We recommend the use of this test together with Holm’s method.

An alternative to directly performing a comparison between a control algorithm and a set of algorithms is the Multiple Sign test. It has been described in this paper, and an example of its use has been provided. We have shown that this procedure is rapid and easy to apply, but it has low power with respect to more advanced techniques. We recommend its use when the differences reported by the control method with respect to the rest of algorithms are very clear for a certain performance metric.

Apart from the well-known Friedman test, we can use two alternatives which differ in the ranking computation. Both the Friedman Aligned Rank test and the Quade test can be used under the same circumstances as the Friedman test. The differences in power between Friedman Aligned Ranks test and the Quade test are unknown, but we encourage the use of these tests when the number of algorithms to be compared is low.

As we have described, the Quade test adds to the ranking computation of Friedman’s test a weight factor computed through the maximum and minimum differences in a problem. This implies that those algorithms that obtain further positive results in diverse problems could benefit from this test. The use of this test should be regulated, because it is very sensitive to the choice of problems. If a researcher decided to include a subgroup of an already studied group of problems where in most of them the proposal obtained good results, this test would report excessive significant differences. On the other hand, for specific problems in which we are interested in quantifying the real differences obtained between algorithms, the use of this test can be justified. We recommend the use of this procedure under justified circumstances and with special caution.

In relation to the post-hoc procedures shown, the differences of power between the methods are rather small,
with some exceptions. The Bonferroni–Dunn test should not be used inspite of its simplicity, because it is a very conservative test and many differences may not be detected. Five procedures (those of Holm, Hochberg, Hommel, Holland, and Rom) have a similar power. Although the Hommel and Rom procedures are the two most powerful procedures, they also are the most difficult to be applied and to be understood. A good alternative is to use the Finner test, which is easy to comprehend and offers better results than the remaining tests, except the Li test in some cases.

• The Li test is even simpler than the Finner, Holm, or Hochberg tests. This test needs to check only two steps and to know the greatest unadjusted p-value in the comparison, which is easy to obtain. The author declares that the power of his test is highly influenced by the p-value of the last hypothesis of the family and, when it is lower than 0.5, the test will be more powerful than the rest of post-hoc methods. However, we recommend that it be used with care and only when the differences between the control algorithm and the rest seem to be high in the performance measure analyzed.

5.3. Multiple comparisons among all methods
• When comparing all algorithms among themselves, 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.
• 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 is computationally expensive. We recommend its use when the situation requires it (that is, when the differences among the algorithms compared are not very significant), given that the results it obtains are as valid as using other testing procedures.

CONCLUSIONS

In this work, we have studied a complete set of non-parametric statistical procedures. The wide set of methods considered, ranging from basic techniques such as the Sign test or Contrast Estimation, to more advanced approaches such as the Friedman Aligned and Quade tests, include tools which can help practitioners in many situations in which the results of an experimental study need to be contrasted.

This study has been extended with a list of considerations, in which we discuss some important issues concerning the behavior and applicability of these tests (and emphasize the use of the most appropriate test depending on the circumstances and type of comparison).

Finally, we encourage the use of nonparametric tests whenever there exists a necessity of analyzing results obtained by clustering algorithms for continuous optimization problems in multi-problem analysis, due to the fact that the initial conditions that guarantee the reliability of the parametric tests are not satisfied. The techniques presented here can help to cover these necessities, providing the research community with reliable and effective tools for incorporating a statistical analysis into the experimental methodologies.

REFERENCES