Int. J. Appl. Math. Comput. Sci., 2012, Vol. 22, No. 4, 867–881
DOI: 10.2478/v10006-012-0064-z

NONPARAMETRIC STATISTICAL ANALYSIS FOR MULTIPLE COMPARISON OF MACHINE LEARNING REGRESSION ALGORITHMS

BOGDAN TRAWIŃSKI [*], MAGDALENA SMĘTEK [*], ZBIGNIEW TELEC [*], TADEUSZ LASOTA [**]

[*] Institute of Informatics
Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: {bogdan.trawinski,magdalena.smetek,zbigniew.telec}@pwr.wroc.pl

[**]Department of Spatial Management
Wrocław University of Environmental and Life Sciences, ul. Norwida 25/27, 50-375 Wrocław, Poland
e-mail: tadeusz.lasota@up.wroc.pl

In the paper we present some guidelines for the application of nonparametric statistical tests and post-hoc procedures devised to perform multiple comparisons of machine learning algorithms. We emphasize that it is necessary to distinguish between pairwise and multiple comparison tests. We show that the pairwise Wilcoxon test, when employed to multiple comparisons, will lead to overoptimistic conclusions. We carry out intensive normality examination employing ten different tests showing that the output of machine learning algorithms for regression problems does not satisfy normality requirements. We conduct experiments on nonparametric statistical tests and post-hoc procedures designed for multiple $1 \times N$ and $N \times N$ comparisons with six different neural regression algorithms over 29 benchmark regression data sets. Our investigation proves the usefulness and strength of multiple comparison statistical procedures to analyse and select machine learning algorithms.

Keywords: machine learning, nonparametric statistical tests, statistical regression, neural networks, multiple comparison tests.

1. Introduction

The field of machine learning has been intensively developed recently. Many novel approaches and techniques to solve classification and regression problems have been proposed and the issue of assessing and comparing with competitive and classical methods has arisen. Demšar (2006) surveyed over 120 papers presented in a series of conferences in the field. He stated that only about half of the papers contained some statistical procedure either for determining an optimal method or for comparing the performances among themselves, and pairwise t-tests were the prevailing method for assessing statistical significance of differences. His work initiated a series of studies aimed to establish systematic procedures for comparing the performance of a number of classifiers over multiple data sets (Demšar, 2006; Derrac *et al.*, 2011; Garcia and Herrera, 2008; Garcia *et al.*, 2009; 2010; Luengo *et al.*, 2009). Their authors argue that the commonly used paired tests, i.e., the parametric t-test and its nonparametric alternative Wilcoxon signed rank tests, are not adequate when conducting multiple comparisons due to the so-called multiplicity effect (Salzberg, 1997). They recommend to employ rank-based nonparametric Friedman or Iman and Davenport tests followed by proper post-hoc procedures for identifying pairs of algorithms which differ significantly.

The most frequently used statistical tests to determine significant differences between two machine learning algorithms are the t-test and the Wilcoxon signed-ranks test (Wilcoxon, 1945). However, the former is a parametric one and requires that the necessary conditions for a safe usage of parametric tests should be fulfilled, i.e., independence, normality, heteroscedasticity (Sheskin, 2011; Zar, 2009). It is not the case in majority of experiments in a machine learning (García and Herrera, 2008; Luengo *et al.*, 2009). Thus, the nonparametric Wilcoxon matched pairs test, which is less powerful than the t-test, should be employed. But, when the researcher wants to confront a newly developed technique with a number of known algorithms or choose the best one out of a set of several algorithms, the pairwise comparisons are not proper. In such situations he/she loses control over the so-called familywise error rate due to an accumulated error coming from the combination of pairwise comparisons. Therefore, he/she should perform tests adequate to multiple comparisons together with a set of post-hoc procedures to compare a control algorithm with other algorithms ($1 \times N$ comparisons) or to perform all possible pairwise comparisons ($N \times N$ comparisons).

First of all, the Friedman test (Friedman, 1937) or its more powerful derivative, the Iman and Davenport test (Iman and Davenport, 1980), should be performed. A usable characteristic of these tests is that they rank the algorithms from the best performing one to the poorest one. However, both tests can only inform the researcher about the presence of differences among all samples of results compared. Two more alternatives can be also applied, the Friedman aligned ranks (Hodges and Lehmann, 1962) and the Quade (Quade, 1979) test, which differ in the way of computing the rankings and may lead to better results depending on the features of the experimental study considered. After the null hypotheses have been rejected, you may proceed with the post-hoc procedures in order to find the particular pairs of algorithms which produce differences. The latter comprise Bonferroni–Dunn's, Holm's, Hochberg's, Hommel's, Holland's, Rom's, Finner and Li's procedures in the case of $1 \times N$ comparisons, and Nemenyi's, Shaffer's, and Bergmann–Hommel's procedures in the case of $N \times N$ comparisons.

The Bonferroni–Dunn scheme (Dunn, 1961) leads to the statement that the performance of two algorithms is significantly different if the corresponding average of rankings is at least as great as its critical difference. More powerful is Holm's routine (Holm, 1979), which checks sequentially hypotheses ordered according to their p-values from the lowest to the highest. All hypotheses for which p-value is less than the significance level α divided by the number of algorithms minus the number of a successive step are rejected. All hypotheses with greater p-values are supported. Holland's and Finner's procedures (Holland and Copenhaver, 1987; Finner, 1993) also adjust the value of α in a step-down manner as Holm's method does. Hochberg's procedure (Hochberg, 1988) operates in the opposite direction to the former, comparing the largest p-value with α, the next largest with $\alpha/2$, and so forth until it encounters a hypothesis it can reject. Rom (1990) devised a modification to Hochberg's step-up procedure to increase its power. In turn, Li (2008) proposed a two-step rejection procedure.

When all possible pairwise comparisons need to be performed, the easiest is Nemenyi's procedure (Nemenyi, 1963). It assumes that the value of the significance level α is adjusted in a single step by dividing it merely by the number of comparisons performed. It is a very simple way but has little power. Shaffer's static routine (Shaffer, 1986), in turn, follows Holm's step down method. At a given stage, it rejects a hypothesis if the p-value is less than α divided by the maximum number of hypotheses which can be true given that all previous hypotheses are false. Bergmann–Hommel's scheme is characterized by the best performance, but it is also most sophisticated and therefore difficult to understand and computationally expensive. It consists in finding all the possible exhaustive sets of hypotheses for a certain comparison and all elementary hypotheses which cannot be rejected. The details of the procedure are described by Bergmann and Hommel (1988) as well as García and Herrera (2008), and the rapid algorithm for conducting this test is presented by Hommel (1994).

All the above mentioned procedures were described in detail by Demšar (2006), Derrac *et al.* (2011), García and Herrera (2008), Garcia *et al.* (2009; 2010) as well as Luengo *et al.* (2009), and used to a series of experiments on neural network algorithms (Luengo *et al.*, 2009), genetics-based machine learning algorithms (García *et al.*, 2009), decision trees and other classification algorithms (García and Herrera, 2008; García *et al.*, 2010), evolutionary and swarm intelligence algorithms (Derrac *et al.*, 2011). These experiments were conducted using machine learning algorithms and benchmark data sets devoted to classification problems and function approximation. In this paper we focus on regression algorithms and employ benchmark data sets for regression problems in statistical tests.

So far, the authors of the present paper have investigated several methods to construct regression models: evolutionary fuzzy systems, neural networks, decision trees, and statistical algorithms using MATLAB, KEEL, and WEKA data mining systems (Graczyk *et al.*, 2009; Krzystanek *et al.*, 2009; Król *et al.*, 2008; Lasota *et al.*, 2010), on the basis of a real-world case of real estate appraisals.

In our investigations we have employed nonparametric Friedman and Wilcoxon paired comparison tests (Lasota *et al.*, 2010; Smętek and Trawiński, 2011) as well as post-hoc procedures devoted to multiple comparisons (Lasota *et al.*, 2011; Lughofer *et al.*, 2011). The Wilcoxon sign rank test was applied also by Zaman and Hirose (2011) to pairwise comparisons in their research into ensemble methods with small training sets.

The main goal of the study presented in this paper is to review the methodology and investigate the potential of multiple comparison statistical procedures to analyse and select machine learning regression algorithms and provide the machine learning community, especially non-statisticians, with guidelines for using advanced nonparametric tests in the case there is a need to evaluate a newly devised algorithm by comparing it with a number of classic and benchmark techniques. In the comparison with the recent works we focus on the assessment of regression algorithms using multiple data sets. Evaluating the performance of multiple regressors, in terms of their accuracy over multiple data sets, is safe because the researcher does not need to assume anything about the sampling scheme and consider the variance of the results obtained.

It is only required that the output ensures reliable estimates of the algorithms' accuracy on individual data sets. Moreover, such an approach provides the independence of the measurements made. The second goal was to survey and carry out the tests for the normality of the result produced by regression algorithms over multiple data sets. This is an important issue because researchers often face a dilemma about whether the output of their experiments has a Gaussian distribution and they are legitimized in applying a parametric test, such as the t-test, or it is safer to use a nonparametric one. Another question may be if the number of measurements performed is sufficient to assume that the values obtained come from a normal distribution. The tests for normality may help with a proper decision. In our case study we applied 6 regression neural algorithms to 29 benchmark data sets and performed a statistical analysis of the results obtained using nonparametric tests and post-hoc procedures designed especially for multiple comparisons. Our preliminary work was presented at the KES2010 conference (Graczyk *et al.*, 2010).

The paper is organized as follows. Section 2 describes the statistical approach to analyze the results of experiments including testing normality, nonparametric tests for pairwise and multiple comparisons and post-hoc procedures adequate for multiple comparisons. Section 3 describes the design of experiments we conducted comprising regression neural algorithms and benchmark data sets we employed. Section 4 presents the results we obtained as well as a thorough analysis of the statistical significance of the output. Finally, in Section 5, we point out the conclusions of the paper.

2. Tests for normality, paired and multiple comparisons

2.1. Testing normality. Very often data sets which are used in validation and verification experiments are not normal in their nature. The non-normality may occur because of their inherent random structure or the presence of outliers. Tens of normality tests have been proposed by statisticians and their performance depends greatly on the distribution type and sample size. A comprehensive review of tests for normality presented by Thode (2002), describing their design, theory, and application. In our case study we applied ten different tests chosen from among the most popular tests of normality.

Pearson's chi-square goodness of fit test checks if an observed frequency distribution differs from the normal distribution (Plackett, 1983). However, it is characterized by less power compared with other tests. The Lilliefors corrected Kolmogorov–Smirnov (K–S) test compares the cumulative distribution of data with the expected cumulative normal distribution (Lilliefors, 1967). This test is different from the K–S test because unknown parameters of the population are estimated, while the statistic remains the same. The Anderson–Darling test is based on the Cumulative Distribution Function (CDF) approach and performs well for small sample sizes (Anderson and Darling, 1954). It is a modification of the Kolmogorov–Smirnov test and gives more weight to the tails than the K–S test. It is one of the most powerful statistical tools for detecting most departures from normality.

The Shapiro–Wilk test calculates the W statistic defined as the ratio of the square linear combination of the ordered sample to the usual sum of squares of deviations from the mean (Shapiro and Wilk, 1965). It is the most preferred test of normality because of its good power properties compared with a wide range of alternative tests. The test statistic of the Shapiro–Francia test is the squared correlation between the ordered sample values and the expected ordered quantiles of the standard normal distribution (Royston, 1993) and is assessed to perform well.

Other types of normality test are called moment tests. They are derived from the recognition that the departure from normality may be detected based on sample moments, namely, skewness and kurtosis. Skewness is the third moment of a distribution and describes its asymmetry. A symmetric distribution has zero skewness, an asymmetric distribution with the largest tail to the right has positive skewness, and a distribution with a longer left tail has negative skewness. D'Agostino test for skewness in normally distributed data has a null hypothesis that data are symmetrical, i.e., skewness is equal to zero (D'Agostino, 1970). This test is useful for significant skewness in normally distributed data. Kurtosis is the fourth moment of distribution and measures both peakedness and tail heaviness of a distribution relative to that of the normal distribution. The Anscombe–Glynn test of kurtosis for normal samples has a null hypothesis that data have kurtosis equal to 3 and is detecting a significant difference of kurtosis in normally distributed data (Anscombe and Glynn, 1983). The two most widely known goodness-of-fit measures of departure from normality are the tests proposed by Jarque and Bera (1987) as well as D'Agostino *et al.* (1990). The former is based on sample kurtosis and skewness, whereas the latter on transformations of these moments. Their statistics have approximately a chi-square distribution with 2 degrees of freedom when the population is normally distributed.

Not long ago, Szekély and Rizzo (2005) proposed the Energy test (E-test), a rotation invariant and consistent goodness-of-fit test for multivariate and univariate distributions based on the Euclidean distance between statistical observations. It is practical to be applied for arbitrary dimension and sample size. The inventors proved it is a powerful competitor to existing tests, and is very sensitive against heavy tailed alternatives.

Several comparative analyses of different normality tests have been published recently. Results of Keskin's study showed that the Shaphiro–Wilk test was the most powerful from among the Kolmogorov–Simirnov, chi-square, Shaphiro–Wilk, D'Agostino–Pearson, skewness and kurtosis normality tests (Keskin, 2006). Yazici and Yolacan (2007) studied 12 different normality tests and compared their powers. They concluded that the best results of the simulation study were achieved for the Kuiper, Vasicek, and Jarque–Bera tests. They also stated that the Shapiro–Wilk statistic provides a superior omnibus indicator of non-normality, although from the practical point of view it should be used for sample sizes between 20 and 50. Romão *et al.* (2010) compared the performance of 33 normality tests, for various sample sizes, considering several significance levels and for a number of symmetric, asymmetric and modified normal distributions. One of their conclusions was that when the nature of the non-normality is unknown, the Shapiro–Chen, Shapiro–Wilk as well as Barrio–Cuesta-Albertos–Matrán–Rodríguez-Rodríguez quantile correlation (BCMR) tests should be applied. Tanveer-ul-Islam (2011) showed the overall superiority of the Anderson–Darling test to the Jarque–Bera one, which belongs to the most popular and widely used tests in the field of economics, as well as to the Shapiro–Francia, D'Agostino–Pearson and Lilliefors tests.

2.2. Wilcoxon's test for pairwise comparisons. The Wilcoxon signed-ranks test (Wilcoxon, 1945) is a nonparametric counterpart of the paired t-test, which ranks the differences in performances of two algorithms over each data set. It omits the signs, and compares the ranks for the positive and the negative differences. The differences are ranked based on their absolute values; in case of ties average ranks are computed. Let d_i be the difference between the performance scores of the two algorithms on the i-th out of n data sets. Let R^+ be the sum of ranks for the data sets on which the second algorithm outperformed the first, and R^- the sum of ranks for the opposite,

$$R^+ = \sum_{d_i>0} \operatorname{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \operatorname{rank}(d_i)$$

$$R^- = \sum_{d_i<0} \operatorname{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \operatorname{rank}(d_i) \tag{1}$$

Ranks of d_i= 0 are divided in half and added to the sums (see Eqn. (1)). If T denotes the smaller sum, i.e., T =min(R^+, R^-), the statistic

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}. \tag{2}$$

for a larger number of data sets, for example, greater than 25, will be approximately normally distributed. For n up to 25, exact critical values for T can be found in tables published in numerous statistical textbooks.

If the requirements for the paired t-test are met, then the Wilcoxon test has less statistical power than the t-test. However, it is safer because it does not assume normal distributions and is more sensible than the t-test when the number of observations is small, e.g., less than 30.

Wilcoxon's test is fit to be used for pairwise comparisons between two algorithms. If we wanted to draw a conclusion comprising more than one pairwise comparison, we would obtain an accumulated error resulting from the combination of pairwise comparisons. Thus, we lose control on the so-called Family Wise Error Rate (FWER), which is defined as the probability of drawing false conclusions when conducting multiple pairwise tests. In consequence, it is not recommended to use pairwise comparison tests, such as Wilcoxon's test, to perform comparisons involving a number of algorithms, since the FWER is not controlled. In order to conduct comparisons which comprise more than two algorithms, tests adequate for multiple comparisons should be used.

2.3. Friedman's and Iman–Davenport's tests for multiple comparisons. In Figs. 1 and 2, multiple comparison nonparametric tests as well as post-hoc procedures for 1 × N and $N \times N$ comparisons, which were considered for classification tasks and function approximation by García *et al.* (2010) and Derrac *et al.* (2011), are summarized. They were applied in the study reported in the present paper for regression problems. It illustrates that first the Friedman test and/or its two alternatives, the Friedman aligned ranks and the Quade test, should be conducted in order to detect whether statistically significant differences occur among the examined algorithms. Moreover, these tests rank the algorithms from the best performing one to the poorest one. If statistical significance is revealed, then the researcher may proceed to accomplish post-hoc procedures to point out which pair of algorithms differ significantly.

The Friedman test (Friedman, 1937) is a nonparametric counterpart of the parametric two-way analysis of variance. The goal of this test is to determine whether there are significant differences among the

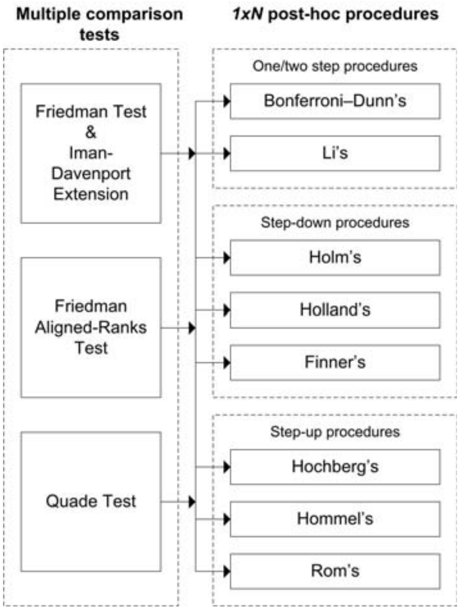

Fig. 1. Nonparametric tests and post-hoc procedures for $1 \times N$ comparisons.

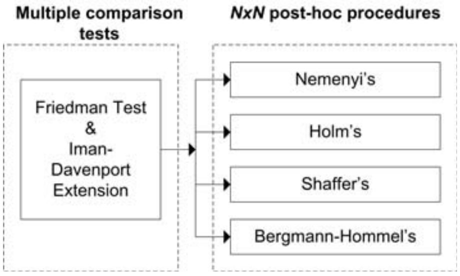
Multiple comparison tests
NxN post-hoc procedures
Friedman Test & Iman-Davenport Extension
Nemenyi's
Holm's
Shaffer's
Bergmann-Hommel's

Fig. 2. Nonparametric tests and post-hoc procedures for $N \times N$ comparisons.

algorithms considered over given sets of data. The test determines the ranks of the algorithms for each individual data set, i.e., the best performing algorithm receives the rank of 1, the second best rank 2, etc.; in the case of ties average ranks are assigned. Let r_i^j be the rank of the j-th of k algorithms on the i-th of n data sets. The Friedman test compares the average ranks of algorithms, $R_j = 1/n \sum_i r_i^j$. The null hypothesis states that all the algorithms perform equivalently and therefore their ranks R_j should be equal. Under this hypothesis the Friedman statistic

$$\chi_F^2 = \frac{12n}{k(k+1)} \left[\sum_j R_j^2 + \frac{k(k+1)^2}{4} \right] \tag{3}$$

is χ_F^2 distributed with $k-1$ degrees of freedom, when n and k are large enough, i.e., $n > 10$ and $k > 5$ (García *et al.*, 2010).

Iman and Davenport (1980) proved that Friedman's χ_F^2 is too conservative and devised a better statistic

$$F_F = \frac{(n-1)\chi_F^2}{n(k-1)\chi_F^2}, \tag{4}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom.

Due to a limited space we do not present in this paper an analysis for aligned Friedman's and Quade's tests, but the results obtained were similar to the output produced with the Friedman's one. The formulas for Friedman aligned ranks and Quade test statistics, the examples of computing the ranks for the Friedman, Friedman aligned, and Quade tests as well as p-values are shown by Derrac *et al.* (2011).

2.4. Post-hoc procedures for $1 \times N$ and $N \times N$ comparisons. The Friedman, Iman–Davenport, Friedman aligned, and Quade tests can only detect significant differences over the whole multiple comparison, although they are not in a position to establish interrelations between the algorithms under consideration. If the null hypothesis of equivalence of rankings is rejected by these tests, the researcher may proceed with post-hoc procedures. In Tables 1 and 2 a set of post-hoc procedures is presented for $1 \times N$ and $N \times N$ comparisons, respectively. For each procedure a brief outline of its scheme and the formula for computation of the Adjusted P-Value (APV) are given. The notation used in Tables 1 and 2 is as follows:

- indexes i and j apply to a given comparison or hypothesis in the family of hypotheses. Index i always concerns the hypothesis whose APV is being determined and index j refers to another hypothesis in the family;
- p_j is the p-value calculated for the j-th hypothesis;
- k is the number of predictors being compared.

3. Experimental setup

3.1. Tools and algorithms used in experiments. All experiments were conducted using *KEEL (Knowledge Extraction based on Evolutionary Learning)*, a tool for creating, learning, optimizing and evaluating various models ranging from soft computing ones to support vector machines, decision trees for regression, and linear regression. KEEL contains several dozen of algorithms for data pre-processing, designing and conducting the experiments, data post-processing and evaluating and visualizing the results obtained, which have been bound into one flexible and user friendly system. KEEL has been developed in the Java environment by a group of Spanish research centres and is available for free

Table 1. Post-hoc procedures for 1 × N comparisons.

Procedure	Description	APV_i
Bonf (Dunn, 1961)	Calculates the adjusted value of a in a single step by dividing it by the number of comparisons, i.e., $(k-1)$.	$\min\{v; 1\}$, where $v = (k-1)p_i$.
Li (Li, 2008)	It is a two-step rejection procedure: *Step 1*: reject all H_i for $p_{k-1} \leq \alpha$. Otherwise, accept the hypothesis associated with p_{k-1} and go to Step 2. *Step 2*: reject any remaining H_i with $p_i \leq (1-p_{k-1})/(1-\alpha)\alpha$.	$p_i/(p_i + 1 - p_{k-1})$.
Holm (Holm, 1979)	Rejects H_1 to H_{i-1} if i is the smallest integer such that $p_i > \alpha/(k-i)$.	$\min\{v; 1\}$, where $v = \max\{(k-1)p_j : 1 \leq j \leq i\}$.
Holl (Holland and Copenhaver, 1987)	Rejects H_1 to H_{i-1} if i is the smallest integer so that $p_i > 1-(1-\alpha)^{k-i}$.	$\min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{k-j} : 1 \leq j \leq i\}$.
Finn (Finner, 1993)	Rejects H_1 to H_{i-1} if i is the smallest integer so that $p_i > 1-(1-\alpha)^{(k-1)/i}$.	$\min\{v; 1\}$, where $v = \max\{1-(1-p_j)^{(k-1)/j} : 1 \leq j \leq i\}$
Hoch (Hochberg, 1988)	This step-up procedure works in the opposite direction that step-down ones do, comparing the largest p-value with α, the next largest with $\alpha/2$, the next with $\alpha/3$, and so forth until it encounters a hypothesis it can reject. All hypotheses with smaller p-values are then rejected as well.	$\max\{(k-j)p_j : k-1 \geq j \geq i\}$.
Rom (Rom, 1990)	It modifies Hochberg's procedure to increase its power. It operates almost in the same way as the Hochberg one, except that the α values are calculated through the expression $a_{k-1} = \frac{1}{i}\left[\sum_{j=1}^{i-1}\alpha^j - \sum_{j=1}^{i-2}\binom{i}{k}\alpha_{k-1-j}^{i-j}\right]$ (5) where $a_{k-1} = \alpha$ and $a_{k-2} = \alpha/2$.	$\{\max\{(r_{k-j})p_j : k-1 \geq j \geq i\}$, where r_{k-j} can be obtained from Eqn. (5), $r = \{$1, 2, 3, 3.814, 4.755, 5.705, 6.655 $\ldots\}$.
Homm (Hommel, 1988)	First, the largest j for which $p_{n-j+k} > k\alpha/j$ for all $k = 1, \ldots, j$ should be found. If no such j exists, all hypotheses can be rejected, otherwise all hypotheses for which $p_i \leq \alpha/j$ are rejected.	The algorithm to compute APV_i for Hommel's procedure can be found in the work of García *et al*. (2010).

Table 2. Post-hoc procedures for $N \times N$ comparisons.

Procedure	Description	APV_i
Nem (Nemenyi, 1963)	Calculates the adjusted value of α in a single step by dividing it by the number of comparisons accomplished, i.e., $k(k-1)/2$.	$\min\{v; 1\}$, where $v = k(k-1)p_i/2$.
Holm (Holm, 1979)	Step-down method, it rejects H_1 to H_{i-1} if i is the smallest integer such that $p_i > \alpha/(k(k-1)/2 - i + 1)$.	$\min\{v; 1\}$, where $v = \max\{(k(k-1)/2 - j + 1)p_j : 1 \leq j \leq i\}$
Shaf (Shaffer, 1986)	Following Holm's step-down method, at stage j, instead of discarding H_i if $p_i \leq \alpha/(k(k-1)/2 - i + 1)$, discards H_i if $p_i \leq \alpha/t_i$, where t_i is the maximum number of hypotheses which can be true given that any $(i, \ldots, 1)$ hypotheses are false.	$\min\{v; 1\}$, where $v = \max\{t_j p_j : 1 \leq j \leq i\}$.
Berg (Bergmann and Hommel, 1988)	Under the definition: "An index set of hypotheses $I \subseteq \{1, \ldots, m\}$ is called exhaustive if exactly all H_j, $j \in I$, could be true", Bergman and Hommel's procedure rejects all H_j with $j \notin A$, where the acceptance set A, given as $A = \bigcup\{I : I \text{ exhaustive}, \min\{P_i : i \in I\} > \alpha/ \Vert I \Vert\}$, is the index set of null hypotheses which are retained.	$\min\{v; 1\}$, where $v = \max\{\Vert I \Vert \min\{p_j, j \in I\} : I \text{ exhaustive}; i \in I\}$

for non-commercial purposes (Alcalá-Fdez *et al.*, 2009; 2011).

KEEL is designed for different users with different expectations and provides three main functionalities: *Data Management*, which is used for new data set up, data import and export, data edition and visualization, data transformations and partitioning, etc.; *Experiments*, which is used to design and evaluate experiments with the use of selected data and provided parameters; *Education*, which is used to run experiments step-by-step in order to display the learning process. KEEL algorithms employed to carry out the experiments are listed in Table 3, where references to source papers are shown. Details of the algorithms can also be found on the KEEL web site

Table 3. Neural machine learning algorithms used in the study.

Code	KEEL name	Description
MLP	Regr-MLPerceptron Conj-Grad	Multilayer perceptron for modeling (Moller, 1990)
RBF	Regr-RBFN	Radial basis function neural network (Broomhead and Lowe, 1998)
RBI	Regr-Incremental-RBFN	Incremental radial basis function neural network (Plat, 1991)
RBD	Regr-Decremental-RBFN	Decremental radial basis function neural network (Broomhead and Lowe, 1998)
IRP	Regr-iRProp+	Multilayer perceptrons trained with the iRProp+ algorithm (Igel and Hüsken, 2003)
SON	Regr-SONN	Self organizing modular neural network (Smotroff *et al.*, 1991)

at www.keel.es.

3.2. Benchmark data sets used in experiments. Twenty nine benchmark data sets for regression were used in experiments. They were downloaded from four web sites:
http://archive.ics.uci.edu/ml/datasets,
http://www.liaad.up.pt/~ltorgo/Regression/Datasets.html,
http://sci2s.ugr.es/keel/datasets.php,
http://funapp.cs.bilkent.edu.tr/datasets.

In order to reduce the size of data sets, feature and instance selection was accomplished. Since different feature selection methods often produce different results, we combined the output of three methods available in the Statistica Data Miner package (Hill and Lewicki, 2007). We computed the average ranking of feature importance provided by variable screening (VSC) as well as variable importance measures for the gradient boosting machine (VIB) and random forests (VIF) and chose the best features for each data set. In turn, instance selection was carried out by randomly drawing a subset of samples. Moreover, outliers were removed by means of the three sigma method. Then the data were normalized using the min-max approach. Table 4 presents information about the data sets: code, name, number of instances and features, number of the link to site they come from (see the list above). Six machine learning algorithms were run in KEEL individually for 29 data sets using 10-fold cross validation (10cv) and the prediction accuracy was measured with the Root Mean Square Error (RMSE).

Table 4. Data sets used in the experiments.

Code	Name	Inst.	Feat.	Link
01	Abalone	4027	8	1,2
02	Ailerons	7154	8	2,3
03	Delta ailerons	6873	5	2
04	Stock	950	5	2,3
05	Bank8FM	4318	8	2
06	California Housing	7921	8	2,3
07	2Dplanes	6560	7	2
08	House (8L)	7358	8	2
09	House (16H)	5626	8	2
10	Delta Elevators	4691	5	3
11	Elevators	7560	7	2
12	Friedman Example	8217	5	1,2
13	Kinematics	8190	8	2,4
14	Computer Activity (1)	6570	8	2
15	Computer Activity (2)	6953	6	2
16	Boston Housing	461	4	1,2
17	Diabetes	43	2	2,4
18	Machine-CPU	188	6	1,4
19	Wisconsin Breast Cancer	152	6	1,2
20	Pumadyn (puma8NH)	2984	8	2
21	Pumadyn (puma32H)	1245	5	2
22	Baseball	337	6	4
23	Plastic	1650	2	3,4
24	Ele2-4 — Electrical-Length	1056	4	3
25	Ele1-2 — Electrical-Length	495	2	3
26	Weather-Izmir	1461	7	3,4
27	Weather-Ankara	1609	8	3,4
28	Mortgage	1049	6	3,4
29	Concrete Strength	72	5	1

4. Statistical analysis of the results of experiments

RMSE values obtained for 6 neural algorithms over 29 data sets are shown in Table 5. The lowest median and InterQuartile Range (IQR) were obtained with MLP and RBF algorithms, whereas the biggest values were produced by RBD and SON algorithms. In turn, the highest values of skewness and kurtosis were MLP, RBI, and RBD, which indicates their output has the distribution more widely different form the normal one than the results provided by other algorithms. Both skewness and kurtosis are equal to zero for normal distribution in the Statistica package.

4.1. Tests of normality. The results of ten normality tests conducted for the output produced by individual neural algorithms over all 29 data sets are shown in Table 6. In all normality tests the null hypothesis stated that data were sampled from a normal distribution whereas the alternative hypothesis stated the opposite. If the p-value for an individual null hypothesis is less than the significance level α (in our study $\alpha = 0.05$), then this hypothesis is rejected. In Table 6, p-values less than 0.05 leading to the rejection of normality hypotheses were

Table 6. Results of normality tests in terms of p-values.

Normality test	MLP	RBF	RBI	RBD	IRP	SON
Pearson's chi-square	*0.0069*	0.0961	0.0961	*0.0005*	0.2533	0.4962
Lilliefors	*0.0045*	0.0638	*0.0471*	*0.0076*	0.1721	0.4262
Anderson–Darling	*0.0001*	0.0746	*0.0025*	*0.0001*	0.1631	0.3767
Shapiro–Wilk	*0.0000*	*0.0370*	*0.0005*	*0.0001*	0.1200	0.4282
Shapiro–Francia	*0.0000*	*0.0359*	*0.0008*	*0.0002*	0.0842	0.3007
D'Agostino skewness	*0.0019*	*0.1280*	*0.0260*	*0.0177*	0.1888	0.4837
Anscombe-Glynn kurtosis	*0.0000*	*0.1624*	*0.0094*	*0.0087*	0.1376	0.2827
Jarque–Bera	*0.0000*	*0.0350*	*0.0020*	*0.0020*	*0.0490*	0.4070
D'Agostino–Pearson	*0.0000*	*0.0260*	*0.0001*	*0.0001*	*0.0453*	0.3166
E-statistic (Energy)	*0.0000*	0.0781	*0.0060*	*0.0000*	0.1682	0.3764

distinguished in the italic font. It can be easily seen that all tests but one indicated the non-normality of MLP, RBI, and RBD output, whereas the results provided by IRP and RBF, had the normal distribution in the majority of tests. It is only in the case of SON that no null hypothesis could be rejected.

Table 5. Comparison of accuracy in terms of RMSE values for models built over 29 data sets.

Set	MLP	RBF	RBI	RBD	IRP	SON
01	0.121	0.125	0.143	0.147	0.124	0.126
02	0.073	0.076	0.079	0.471	0.077	0.096
03	0.085	0.091	0.094	0.169	0.088	0.090
04	0.049	0.045	0.072	0.076	0.081	0.155
05	0.049	0.053	0.092	0.110	0.085	0.152
06	0.125	0.134	0.144	0.235	0.148	0.149
07	0.043	0.044	0.051	0.117	0.066	0.175
08	0.103	0.106	0.121	0.147	0.113	0.141
09	0.084	0.087	0.096	0.098	0.092	0.098
10	0.107	0.107	0.119	0.203	0.111	0.131
11	0.090	0.089	0.105	0.110	0.108	0.103
12	0.036	0.040	0.051	0.095	0.075	0.136
13	0.073	0.087	0.138	0.127	0.145	0.156
14	0.061	0.061	0.075	0.369	0.078	0.078
15	0.064	0.063	0.074	0.547	0.065	0.099
16	0.088	0.093	0.110	0.111	0.095	0.123
17	0.200	0.208	0.174	0.227	0.199	0.189
18	0.089	0.088	0.099	0.131	0.099	0.133
19	0.426	0.265	0.270	0.296	0.268	0.269
20	0.136	0.174	0.318	0.190	0.186	0.204
21	0.046	0.057	0.084	0.102	0.158	0.158
22	0.168	0.169	0.169	0.183	0.158	0.162
23	0.152	0.167	0.177	0.183	0.158	0.174
24	0.019	0.019	0.046	0.195	0.027	0.046
25	0.080	0.089	0.085	0.099	0.084	0.129
26	0.021	0.021	0.046	0.057	0.027	0.061
27	0.020	0.019	0.052	0.058	0.038	0.151
28	0.012	0.014	0.042	0.047	0.019	0.035
29	0.092	0.118	0.122	0.125	0.104	0.255
Skew	2.864	1.045	1.708	1.872	0.879	0.444
Kurt	11.20	1.218	3.503	3.578	1.340	0.812
Med	0.084	0.088	0.096	0.131	0.095	0.136
IQR	0.058	0.065	0.0640	0.093	0.068	0.058

Supplementary to the statistical normality tests are graphical methods which include among others, histograms and Quantile-Quantile (Q-Q) plots. The former are useful devices for exploring the shape of the underlying frequency distribution of a set of continuous data, for screening the outliers, skewness, peakedness, tails, etc. The latter technique, in turn, plots the quantiles of the first data set against those of the second data set. If the two data sets come from a population with the same distribution, the points fall approximately along the 45-degree reference line. The greater the departure from the reference line, the greater the evidence that the two data sets have come from populations with different distributions. The histograms with an ideal Gaussian distribution and normal Q-Q plots for the results provided by individual algorithms are presented in Figs. 3–8.

A general observation is that none of the algorithms produces perfect Gaussian output. Taking into account the scale of the Y-axes of histogram charts, the distribution of points on Q-Q plots, and the values of skewness and kurtosis in Table 5, it can be concluded that MLP, RBI, and RBD accuracies deviate from a Gaussian distribution more than others. However, the graphical approach is not too accurate and requires that the researcher be experienced in such analyses.

In our case study we acquired about 30 measurement points for each algorithm, which was a moderate number, i.e., not too small and not too big. Therefore it was reasonable to conduct normality tests. The normality tests try to find the answer to the question how far a given distribution deviates from the ideal Gaussian one. Because the tests estimate deviations from Gaussian using different methods, they produce different results. This explains our output for IRP and RBF neural algorithms, where some tests indicated the normality of distribution and some did not.

Our general conclusion was that we were justified to employ nonparametric tests discarding parametric ones, which reveal greater power provided the normality requirements are satisfied. However, it is argued, e.g.,

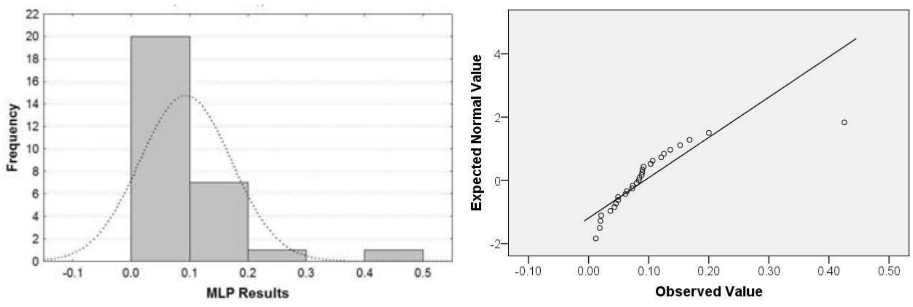

Fig. 3. Histogram (left) and normal Q-Q plot (right) of MLP results.

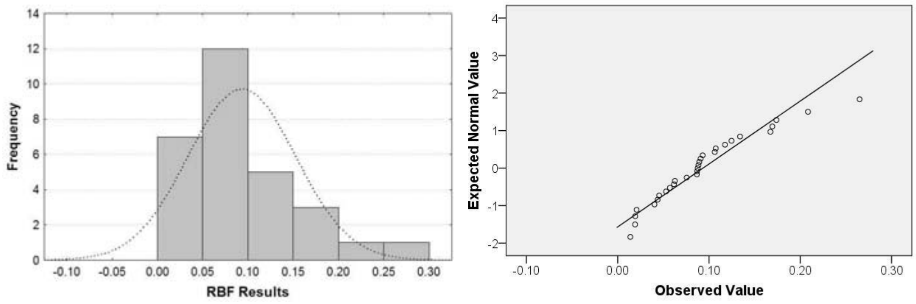

Fig. 4. Histogram (left) and normal Q-Q plot (right) of RBF results.

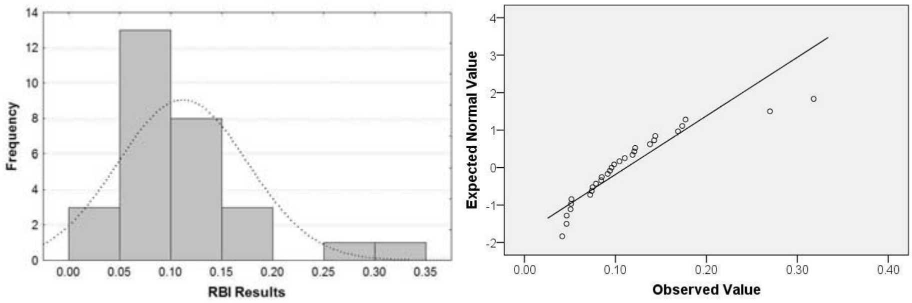

Fig. 5. Histogram (left) and normal Q-Q plot (right) of RBI results.

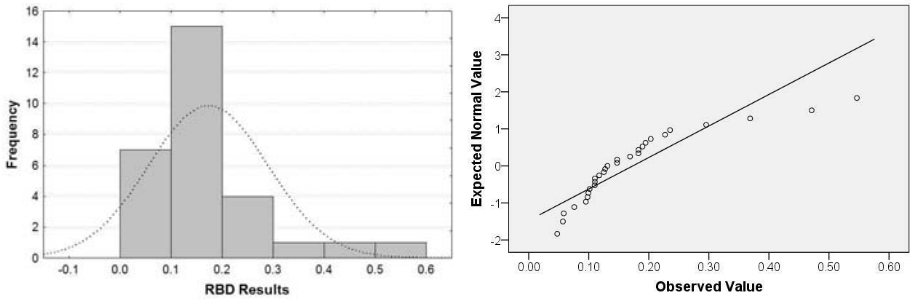

Fig. 6. Histogram (left) and normal Q-Q plot (right) of RBD results.

(Motulsky, 2010; Sheskin, 2011; Zar, 2009) that the decision whether to use a parametric or a nonparametric test is more sophisticated and the researcher should not automatically rely only on normality tests. If data deviate significantly from a Gaussian distribution, maybe the problem can be solved by transforming all the values to their logarithms or reciprocals, rather than using a nonparametric test. Eliminating outliers may also lead to the desired results. Moreover, in the case of small samples, normality tests do not have enough power

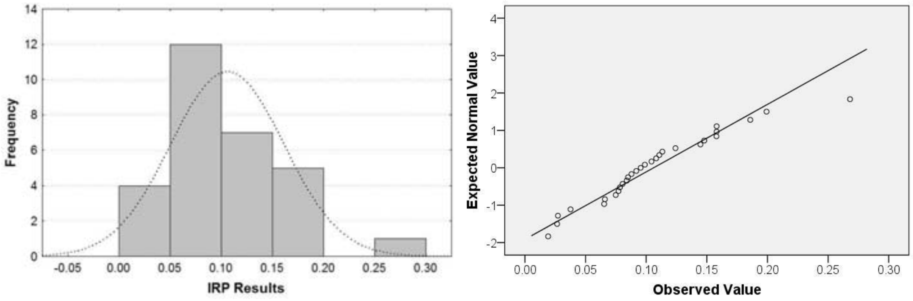

Fig. 7. Histogram (left) and normal Q-Q plot (right) of IRP results.

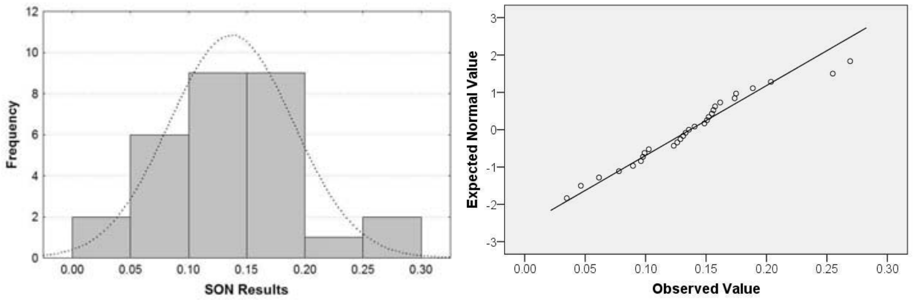

Fig. 8. Histogram (left) and normal Q-Q plot (right) of SON results.

to discriminate between Gaussian and non-Gaussian populations. With large samples, on the other hand, the parametric t-tests and ANOVA are fairly resistant to violations of this requirement and it is not much critical whether data are non-Gaussian or not.

4.2. Nonprametric significance tests.

Statistical analysis of the results of experiments was performed using software available on the web page of the Research Group Soft Computing and Intelligent Information Systems at the University of Granada (`http://sci2s.ugr.es/sicidm`). This open source JAVA program calculates multiple comparison procedures: the Friedman, Iman–Davenport, Bonferroni–Dunn, Holm, Hochberg, Holland, Rom, Finner, Li, Shaffer, and Bergamnn-Hommel tests as well as adjusted p-values. An adjusted p-value can be directly taken as the p-value of a hypothesis belonging to a comparison of multiple algorithms. If the adjusted p-value for an individual null hypothesis is less than the significance level, in our study $\alpha = 0.05$, then this hypothesis is rejected (Wright, 1992). For paired comparisons, nonparametric Wilcoxon signed ranks tests were made using the Statistica package.

The Friedman and Iman–Davenport tests were performed in respect of average ranks, which use χ^2 and F statistics, respectively. The calculated values of these statistics were 88.17 and 43.44, respectively, whereas the critical values at $\alpha = 0.05$ are $\chi^2(5) = 12.83$ and $F(5, 140) = 2.28$, so the null hypotheses were rejected. Thus, we were justified in proceeding to post-hoc procedures. Average rankings of neural algorithms over 29 data sets for produced by the Friedman test are shown in Table 7.

Table 7. Average rank positions of neural algorithms determined during the Friedman test.

1-st	2-nd	3-rd	4-th	5-th	6-th
MLP (1.55)	RBF (2.21)	IRP (3.24)	RBI (3.90)	SON (4.79)	RBD (5.31)

Unadjusted and adjusted p-values for the Bonferroni–Dunn, Holm, Hochberg, Hommel, Holland, Rom, Finner, and Li post-hoc procedures for $1 \times N$ comparisons, where MLP was the control algorithm, are displayed in Table 8. In all tests the Bonferroni–Dunn procedure provided the highest adjusted p-values, whereas the Finner and Li procedures yielded the lowest adjusted p-values. In turn, the Holm, Hochberg, Hommel post-hoc procedures gave the same results for all tests; slightly lower p-values were produced by the Holland and Rom procedures. Thus, our study confirmed some observations made for classification algorithms. The Bonferroni–Dunn procedure is the simplest one but also the least powerful. The Li and Finner procedures seem to be multiple comparison tests with the highest power. As for machine learning algorithms, MLP revealed significantly better performance than other algorithms except for RBF.

In Table 9, p-values for the Wilcoxon test, unadjusted values and adjusted p-values for the Nemenyi, Holm, Shaffer and Bergmann–Hommel tests for $N \times N$

Table 8. Adjusted p-values for 1 × N comparisons of neural algorithms over 29 data sets for the Friedman test (MLP is the control algorithm).

p-values	RBD	SON	RBI	IRP	RBF
pUnadjust	*2.01E-14*	*4.18E-11*	*1.82E-06*	*0.000584*	0.182355
pBonf	*1.00E-13*	*2.09E-10*	*9.09E-06*	*0.002918*	0.911776
pHolm	*1.00E-13*	*1.67E-10*	*5.45E-06*	*0.001167*	0.182355
pHoch	*1.00E-13*	*1.67E-10*	*5.45E-06*	*0.001167*	0.182355
pHomm	*1.00E-13*	*1.67E-10*	*5.45E-06*	*0.001167*	0.182355
pHoll	*1.00E-13*	*1.67E-10*	*5.45E-06*	*0.001167*	0.182355
pRom	*9.53E-14*	*1.59E-10*	*5.45E-06*	*0.001167*	0.182355
pFinn	*1.00E-13*	*1.05E-10*	*3.03E-06*	*0.000729*	0.182355
pLi	*2.45E-14*	*5.11E-11*	*2.22E-06*	*0.000713*	0.182355

Table 9. Adjusted p-values for N × N comparisons of neural algorithms over 29 data sets.

Alg vs Alg	pWilcox	pUnadj	pNeme	pHolm	pShaf	pBerg
MLP vs RBD	*0.000031*	*2.01E-14*	*3.01E-13*	*3.01E-13*	*3.01E-13*	*3.01E-13*
MLP vs SON	*0.000079*	*4.18E-11*	*6.27E-10*	*5.85E-10*	*4.18E-10*	*4.18E-10*
RBF vs RBD	*0.000003*	*2.67E-10*	*4.01E-09*	*3.47E-09*	*2.67E-09*	*2.67E-09*
RBF vs SON	*0.000013*	*1.41E-07*	*2.11E-06*	*1.69E-06*	*1.41E-06*	*8.46E-07*
MLP vs RBI	*0.000247*	*1.82E-06*	*0.000027*	*0.000020*	*0.000018*	*0.000013*
RBD vs IRP	*0.000060*	*0.000025*	*0.000381*	*0.000254*	*0.000254*	*0.000178*
MLP vs IRP	*0.000192*	*0.000584*	*0.008754*	*0.005252*	*0.004085*	*0.003502*
RBF vs RBI	*0.000055*	*0.000584*	*0.008754*	*0.005252*	*0.004085*	*0.003502*
IRP vs SON	*0.000021*	*0.001586*	*0.023797*	*0.011105*	*0.011105*	*0.006346*
RBI vs RBD	*0.000095*	*0.004007*	0.060100	*0.024040*	*0.024040*	*0.016027*
RBF vs IRP	*0.002746*	*0.035240*	0.528603	0.176201	0.140961	0.070480
RBI vs SON	*0.002947*	0.068025	1.000000	0.272099	0.272099	0.136050
MLP vs RBF	*0.001654*	0.182355	1.000000	0.547065	0.547065	0.547065
RBI vs IRP	0.116956	0.182355	1.000000	0.547065	0.547065	0.547065
RBD vs SON	0.537723	0.292436	1.000000	0.547065	0.547065	0.547065

comparisons for all possible 15 pairs of algorithms are placed. The p-values below 0.05 indicate that the respective algorithms differ significantly in prediction errors; they were marked with the italic font. It should be noted that with 15 hypotheses the differences between pairwise and multiple comparisons become apparent. The Wilcoxon test allows rejecting 13 hypotheses whereas the Holm, Shaffer and Bergmann–Hommel ones discard only 10 while Nemenyi's method just 9. MLP revealed significantly better performance than any other algorithm but RBF.

Moreover, we compared the behaviour of four methods: the Wilcoxon test, Nemenyi's, Schaffer's and Bergmann–Hommel's procedures for $N \times N$ comparisons depending on the decreasing number of data sets. In Figs. 9 and 10 the number of rejected null hypotheses out of 15 possible pairs of algorithms over a different number of data sets is shown. The decreasing number of data sets was obtained by stepwise elimination of the data sets providing the maximal average prediction error (Fig. 9) and the minimal average RMSE for all algorithms (Fig. 10). In both charts it can be observed that employing only the pairwise Wilcoxon test would lead to overoptimistic conclusions because the number of null hypotheses rejected by this test was from 2 to 7 times greater than the one obtained when following Shaffer's or Bergmann–Hommel's schemes. In a majority of instances, the Shaffer and Bergmann–Hommel procedures turned out to be more powerful than that by Nemeneyi. For multiple comparison procedures the percentage of discarded null hypotheses was larger for a bigger number of benchmark data sets left.

5. Conclusions

In the paper we studied the application of nonparametric statistical tests and post-hoc procedures devised to perform multiple comparisons of regression algorithms over benchmark regression data sets. We conducted experiments on statistical procedures designed especially for multiple 1 × N and N × N comparisons with six neural regression algorithms implemented in the KEEL data mining system.

We carried out intensive normality investigation employing ten different tests chosen from among the most popular tests of normality. The results were unequivocal only for half of the neural algorithms employed. This means that several tests should be conducted and

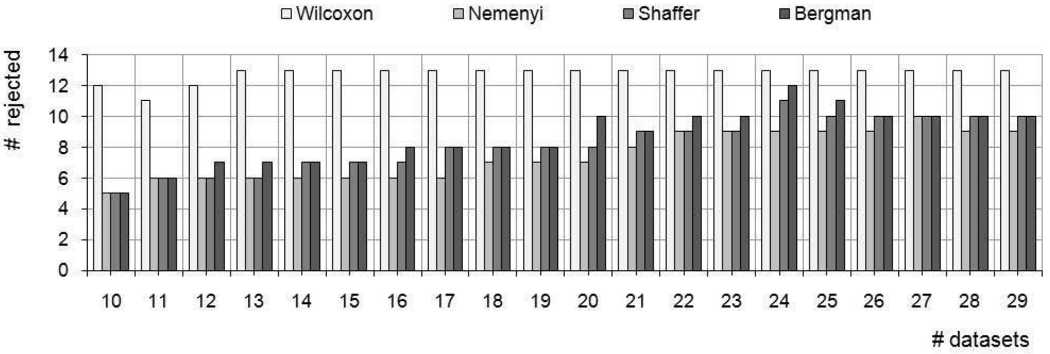

Fig. 9. Number of rejected null hypotheses for the Wilcoxon, Nemenyi and Shaffer tests over a different number of data sets (stepwise eliminating data sets providing the maximal accuracy error).

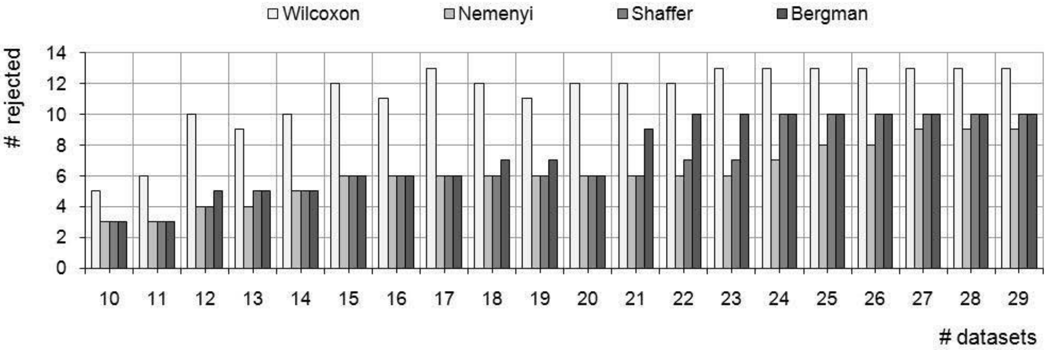

Fig. 10. Number of rejected null hypotheses for the Wilcoxon, Nemenyi and Shaffer tests over a different number of data sets (stepwise eliminating data sets providing the minimal accuracy error).

supplementary analysis is needed using graphical methods such us histograms and quantile-quantile plots.

It is necessary to distinguish between pairwise and multiple comparison tests. The former procedures should be used to examine two algorithms and the latter are valid when contrasting more than two techniques.

Nonparametric methods should be used when the sample sizes are small, especially when the number of instances is less than 30. In turn, for large samples, e.g., for the number of instances greater than 100, due to the central limit theorem, parametric methods are more appropriate because they have more statistical power and therefore are more sensitive.

The nonparametric Wilcoxon test, when employed to multiple comparisons, would lead to overoptimistic conclusions. In our analysis we omitted the nonparametric Sign test, which is less statistically powerful than the Wilcoxon matched pairs test. However, the latter assumes it is possible to rank order the magnitude of differences in matched observations in a meaningful manner, otherwise the Sign test might be more applicable.

Among $1 \times N$ procedures, the Bonferroni–Dunn one is the simplest but is also the least powerful. Holm's procedure has also little power, because it is based on the Bonferonni inequality. Both Rom's and Hommel's procedures are more powerful than Hochberg's procedure due to the fact that sharp inequalities (or equalities) are used in both; however, the power improvement is negligible compared to their complexities. Finner's and Li's procedures yielded the lowest adjusted p-values and thus they seem to have the highest power.

In turn, among $N \times N$ procedures, that of Bergmann and Hommel is the most powerful but it requires intensive computations in comparisons comprising a bigger number of predictors. Thus, Shaffer's static routine or Holm's step down method is recommended. For multiple comparisons, the more data sets used in tests, the larger the number of null hypotheses rejected. Our investigation proved the usefulness and strength of multiple comparison statistical procedures to analyse and select machine learning algorithms.

For the authors proposing novel classification and regression algorithms (Baruque *et al.*, 2011; Czarnowski and Jędrzejowicz, 2011; Jackowski and Woźniak, 2010; Kajdanowicz and Kazienko, 2011; Troć and Unold, 2010), especially $1 \times N$ procedures are recommended because they are simpler and require less measurement points than $N \times N$ procedures to provide meaningful output.

It should be noted that nonparametric tests of significance are based on asymptotic theory, which assumes large samples, therefore their output could not be meaningful if the sample sizes were too small. Thus, more thorough analysis of statistical power and efficiency of individual tests including sample size estimation is also necessary.

Acknowledgment

This work was partially funded by the Polish National Science Centre under the grant no. N N516 483840.

References

Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L. and Herrera, F. (2011). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, *Journal of Multiple-Valued Logic and Soft Computing* **17**(2–3): 255–287.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J. and Herrera, F. (2009). KEEL: A software tool to assess evolutionary algorithms to data mining problems, *Soft Computing* **13**(3): 307–318.

Anderson, T. and Darling, D. (1954). A test of goodness-of-fit, *Journal of the American Statistical Association* **49**(268): 765–769.

Anscombe, F. and Glynn, W. (1983). Distribution of the kurtosis statistic b2 for normal samples, *Biometrika* **70**(1): 227–234.

Baruque, B., Porras, S. and Corchado, E. (2011). Hybrid classification ensemble using topology-preserving clustering, *New Generation Computing* **29**(3): 329–344.

Bergmann, G. and Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses, *in* P. Bauer, G. Hommel and E. Sonnemann (Eds.), *Multiple Hypotheses Testing*, Springer-Verlag, Berlin, pp. 100–115.

Broomhead, D. and Lowe, D. (1998). Multivariable functional interpolation and adaptive networks, *Complex Systems* **11**: 321–355.

Czarnowski, I. and Jędrzejowicz, P. (2011). Application of agent-based simulated annealing and tabu search procedures to solving the data reduction problem, *International Journal of Applied Mathematics and Computer Science* **21**(1): 57–68, DOI: 10.2478/v10006-011-0004-3.

D'Agostino, R. (1970). Transformation to normality of the null distribution of g1, *Biometrika* **57**(3): 679–681.

D'Agostino, R., Belanger, A. and D'Agostino Jr., R. (1990). A suggestion for using powerful and informative tests of normality, *The American Statistician* **44**(4): 316–321.

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

Derrac, J., García, S., Molina, D. and Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, *Swarm and Evolutionary Computation* **1**: 3–18.

Dunn, O. (1961). Multiple comparisons among means, *Journal of the American Statistical Association* **56**(238): 52–64.

Finner, H. (1993). On a monotonicity problem in step-down multiple test procedures, *Journal of the American Statistical Association* **88**(423): 920–923.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**(200): 675–701.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability, *Soft Computing* **10**(13): 959–977.

García, S., Fernández, A. and Luengo, J.and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, *Information Sciences* **180**: 2044–2064.

García, S. and Herrera, F. (2008). An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons, *Journal of Machine Learning Research* **9**: 2677–2694.

Graczyk, M., Lasota, T., Telec, Z. and Trawiński, B. (2010). Nonparametric statistical analysis of machine learning algorithms for regression problems, *in* R. Setchi, I. Jordanov, R.J. Howlett and L.C. Jain (Eds.), *KES 2010*, Lecture Notes in Artificial Intelligence, Vol. 6276, Springer, Heidelberg, pp. 111–120.

Graczyk, M., Lasota, T. and Trawiński, B. (2009). Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA, *in* N.T. Nguyen, R. Kowalczyk and S.-M. Chen (Eds.), *ICCCI 2009*, Lecture Notes in Artificial Intelligence, Vol. 5796, Springer, Heidelberg, pp. 800–812.

Hill, T. and Lewicki, P. (2007). *Statistics: Methods and Applications*, StatSoft, Tulsa.

Hochberg, Y. (1988). A Sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**(4): 800–802.

Hodges, J. and Lehmann, E. (1962). Ranks methods for combination of independent experiments in analysis of variance, *Annals of Mathematical Statistics* **33**: 482–497.

Holland, B. and Copenhaver, M. (1987). An improved sequentially rejective Bonferroni test procedure, *Biometrics* **43**(2): 417–423.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test, *Biometrika* **75**(2): 383–386.

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

Igel, C. and Hüsken, M. (2003). Empirical evaluation of the improved RPROP learning algorithm, *Neurocomputing* **50**: 105–123.

Iman, R. and Davenport, J. (1980). Approximations of the critical region of the Friedman statistic, *Communications in Statistics* **18**: 571–595.

Jackowski, K. and Woźniak, M. (2010). Method of classifier selection using the genetic approach, *Expert Systems* **27**(2): 114–128.

Jarque, C. and Bera, A. (1987). A test for normality of observations and regression residuals, *International Statistical Review* **55**(2): 163–172.

Kajdanowicz, T. and Kazienko, P. (2011). Boosting-based sequential output prediction, *New Generation Computing* **29**(3): 293–307.

Keskin, S. (2006). Comparison of several univariate normality tests regarding type I error rate and power of the test in simulation based small samples, *Journal of Applied Science Research* **2**(5): 296–300.

Król, D., Lasota, T., Trawiński, B. and Trawiński, K. (2008). Investigation of evolutionary optimization methods of TSK fuzzy model for real estate appraisal, *International Journal of Hybrid Intelligent Systems* **5**(3): 111–128.

Krzystanek, M., Lasota, T. and Trawiński, B. (2009). Comparative analysis of evolutionary fuzzy models for premises valuation using KEEL, *in* N.T. Nguyen, R. Kowalczyk and S.-M. Chen (Eds.), *ICCCI 2009*, Lecture Notes in Artificial Intelligence, Vol. 5796, Springer, Heidelberg, pp. 838–849.

Lasota, T., Mazurkiewicz, J., Trawiński, B. and Trawiński, K. (2010). Comparison of data driven models for the validation of residential premises using KEEL, *International Journal of Hybrid Intelligent Systems* **7**(1): 3–16.

Lasota, T., Telec, Z., Trawiński, B. and Trawiński, K. (2011). Investigation of the ets evolving fuzzy systems applied to real estate appraisal, *Journal of Multiple-Valued Logic and Soft Computing* **17**(2–3): 229–253.

Li, J. (2008). A two-step rejection procedure for testing multiple hypotheses, *Journal of Statistical Planning and Inference* **138**(6): 1521–1527.

Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association* **62**(318): 399–402.

Luengo, J., García, S. and Herrera, F. (2009). A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests, *Expert Systems with Applications* **36**: 7798–7808.

Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O. and Lasota, T. (2011). On employing fuzzy modeling algorithms for the valuation of residential premises, *Information Sciences* **181**: 5123–5142.

Moller, F. (1990). A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks* **6**: 525–533.

Motulsky, H. (2010). *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*, 2nd Edn., Oxford University Press, New York, NY.

Nemenyi, P.B. (1963). *Distribution-free Multiple Comparisons,* Ph.D. thesis, Princeton University, Princeton, NJ.

Plackett, R. (1983). Karl Pearson and the chi-squared test, *International Statistical Review* **51**(1): 59–72.

Plat, J. (1991). A resource allocating network for function interpolation, *Neural Computation* **3**(2): 213–225.

Quade, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects, *Journal of the American Statistical Association* **74**: 680–683.

Romão, X., Delgado, R. and Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality, *Journal of Statistical Computation and Simulation* **80**(5): 545–591.

Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**(3): 663–665.

Royston, P. (1993). A pocket-calculator algorithm for the Shapiro–Francia test for non-normality: An application to medicine, *Statistics in Medicine* **12**(2): 181–184.

Salzberg, S. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery* **1**: 317–327.

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

Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality (complete samples), *Biometrika* **52**(3/4): 591–611.

Sheskin, D. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th Edn., Chapman & Hall/CRC, Boca Raton, FL.

Smętek, M. and Trawiński, B. (2011). Investigation of genetic algorithms with self-adaptive crossover, mutation, and selection, *in* E. Corchado, M. Kurzyński and M. Woźniak (Eds.), *HAIS 2011*, Lecture Notes in Artificial Intelligence, Vol. 6678, Springer, Heidelberg, pp. 116–123.

Smotroff, I., Friedman, D. and Connolly, D. (1991). Self organizing modular neural networks, *IEEE International Joint Conference on Neural Networks, IJCNN'91, Seattle, WA, USA*, pp. 187–192.

Székely, G.J. and Rizzo, M. (2005). A new test for multivariate normality, *Journal of Multivariate Analysis* **93**(1): 58–80.

Tanweeer-Ul-Islam (2011). Normality testing—A new direction, *International Journal of Business and Social Science* **2**(3): 115–118.

Thode, H. (2002). *Testig for Normality*, Marcel Dekker, New York, NY.

Troć, M. and Unold, O. (2010). Self-adaptation of parameters in a learning classifier system ensemble machine, *International Journal of Applied Mathematics and Computer Science* **20**(1): 157–174, DOI: 10.2478/v10006-010-0012-8.

Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**: 80–83.

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

Yazici, B. and Yolacan, S. (2007). A comparison of various tests of normality, *Journal of Statistical Computation and Simulation* **77**(2): 175–183.

Zaman, M. and Hirose, H. (2011). Classification performance of bagging and boosting type ensemble methods with small training sets, *New Generation Computing* **29**(3): 277–292.

Zar, J. (2009). *Biostatistical Analysis*, 5th Edn., Prentice Hall, Upper Saddle River, NJ.

Bogdan Trawiński, Ph.D., is an assistant professor at the Faculty of Computer Science and Management of the Wrocław University of Technology, Poland. He received his M.Sc. (1980) and Ph.D. (1986) degrees from the same university. His main research interests are computational intelligence, fuzzy systems, evolving systems, incremental machine learning, ensemble and hybrid methods, and real estate appraisal. He is an author of over 90 scientific publications.

Magdalena Smętek is a 3rd year Ph.D. student at the Faculty of Computer Science and Management, Wrocław University of Technology, Poland. She received her M.Sc. degree from the same university in 2009. Her research interests include computational intelligence, knowledge management, and data mining.

Tadeusz Lasota, Ph.D., is a senior lecturer at the Department of Spatial Management of the Wrocław University of Environmental and Life Sciences, Poland. He received his M.Sc. (1972) and Ph.D. (1978) degrees from the Agricultural University of Wrocław. His main research interests are land and building cadastre, spatial planning and management, management of rural areas, and computational intelligence methods in real estate appraisal.

Zbigniew Telec, Ph.D., is an assistant professor at the Faculty of Computer Science and Management of the Wrocław University of Technology, Poland. He received his M.Sc. (2003) degree from the Wrocław University of Technology and the Ph.D. (2008) degree from the Wrocław University of Economics. His main research interests are ensemble approaches in computational intelligence, fuzzy systems, and evolving systems. He is an author of about 30 scientific publications.

Received: 8 August 2011
Revised: 4 April 2012