A Study on the Use of Statistical Tests for Experimentation with Neural Networks

Julián Luengo, Salvador García, and Francisco Herrera

University of Granada, Department of Computer Science and Artificial Intelligence,
E.T.S.I. Informática, 18071 Granada, Spain
julianlm@ugr.es, {salvagl,herrera}@decsai.ugr.es

Abstract. In this work, we get focused on the use of statistical techniques for behavior analysis of Artificial Neural Networks in the task of classification. A study of the non-parametric tests use is presented, using some well-known models of neural networks. The results show the need of using non-parametric statistic, because the Artificial Neural Networks used do not verify the hypothesis required for classical parametric tests.

1 Introduction

Nowadays, statistical validation of published results is an important task [3]. Due the increasing number of real-world applications and frameworks for Machine Learning (ML), developing and modifying new algorithms is relatively easy. However, every development made must be exposed in front of existing work. The question then is, how could we compare and rank them? and much more important, is our comparison right made? Usually, we can't demonstrate which algorithm is better by theoretically, and we only counting on empirical results to achieve this goal.

In a typical paper of ML, and Artificial Neural Networks (ANNs) by extension, a new algorithm or improvement has been proposed, and there exists an implicit hypothesis that such an enhancement yields an improved performance over the existing algorithm(s). A number of data sets is selected for testing, the algorithms are run over them and the quality of the resulting models is evaluated by means of an appropriate measure (commonly, the classification accuracy). The final step, and the topic we want to show, is the use of statistical tests which really suits the initial conditions.

In fact, a low proportion of publications uses statistical techniques to comparing the obtained results. However, their presence is growing notoriously, and most of reviews claim for their use. When we found statistical studies, they are based on the mean and variance, using parametrical tests (ANOVA, t-test,...)[1,2,4,12].

In this work, we will focus on the use of statistical techniques for the analysis of ANNs in classifications tasks, studying the use of parametric and non-parametric statistical tests [8,11]. Indeed, we analyze the required conditions which allow the use of parametric tests, and we will show results obtained using non-parametric tests.

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 72–79, 2007.

To achieve the proposed goals, we will use some well-known models of ANNs applied to classification of data sets [6,10]. In addition, we will show the results in which the need of non-parametric statistical is left patent, since used ANNs don't verify the initial hypothesis which allow the use of parametric tests.

The remain of the paper is organized as follows. In Section 2, we describe the ANN models used in the study. Section 3 explores the needed conditions in order to correctly apply and analyze the parametric tests decision. A presentation of the non-parametric tests and their experimental use is given in Section 4. Finally, in Section 5 we reach our conclusion.

2 Preliminaries: Artificial Neural Networks, Data Sets and Experimentation Framework

In this section, we will briefly describe the algorithms used and the data set chosen. We also show the details of the experimentation we have done. We have used the next models of ANNs:

- Multi-Layer Perceptron (MLP) with Backpropagation[6]: This class of networks consists of multiple layers of computational units with directed connections to the neurons of the subsequent layer, in a feed-forward way and which weights are adjusted with backpropagation. As an activation function the units of these networks apply a sigmoid function . We have used two configurations for MLP Backpropagation model: *MLP Backpropagation 1x5* has 1 hidden layer with 5 perceptrons. *MLP Backpropagation 1x25* has 1 hidden layer with 25 perceptrons.
- Radial Basis Function Network (RBFN)[6]: Radial basis functions have been applied in the area of neural networks as a replacement for the sigmoidal function. RBF networks have 2 layers of processing: In the first, input is mapped onto each RBF in the 'hidden' layer. The number of neurons is fixed at 50 neurons.
- RBFN Decremental[10]: In the classical approach (see above), the number of hidden units is fixed a priori. The authors have proposed an algorithm that adds hidden units to the network based on the novelty of the new data, and augments it with a pruning strategy (which removes hidden neurons with little contribution to the output). The configuration used has 20 initial neurons, *alpha* value of 0.3 and *percent* value of 0.1.

We have selected a set of data sets taken from the UCI repository. Altogether, we have used 7 data sets to make the study. In Table 1, we summarize the properties of these data sets.

With this data, two kinds of validations have been carried out. In 10-fold cross validation each data set have been partitioned in ten folds, and iteratively 9 of those are taken to train the ANN, so the last fold is taken for testing the learning of the network. With Hold out partition at 50% the considered data set

Table 1. Data Sets used for experimentation

Data set	# Instances	# Attributes	# Classes
breast	682	10	2
cleveland	303	13	5
crx	689	16	2
glass	214	9	7
iris	150	4	3
pima	768	8	2
wisconsin	699	10	2

is divided into two parts with same number of instances. The network uses one part to train, and the complementary for test. For each type, we have repeated the experiments 5 times for 10fcv, and 25 times for hold out partitions. In this way, 50 runs and their respective validations have been carried out, and the tests results are summarized in Table 2.

Table 2. Results for ANNs used

Using 10-fold cross validation								
Method	MLP Backprop.-1x25		MLP Backprop.-1x5		RBFN Decremental		RBFN	
Dataset	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.
Breast	0.96	0.01	0.96	0.01	0.83	0.06	0.86	0.04
Cleveland	0.51	0.07	0.49	0.10	0.35	0.09	0.35	0.10
Crx	0.85	0.05	0.82	0.09	0.45	0.02	0.45	0.02
Glass	0.50	0.10	0.46	0.14	0.29	0.12	0.37	0.13
Iris	0.74	0.10	0.75	0.13	0.90	0.09	0.86	0.09
Pima	0.74	0.05	0.70	0.09	0.68	0.05	0.62	0.12
Wisconsin	0.97	0.02	0.96	0.05	0.84	0.09	0.86	0.07
Using validation by hold out partition								
Method	MLP Backprop.-1x25		MLP Backprop.-1x5		RBFN Decremental		RBFN	
Dataset	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.	Mean	St. Desv.
Breast	0.97	0.01	0.97	0.01	0.82	0.07	0.83	0.05
Cleveland	0.53	0.04	0.48	0.08	0.33	0.08	0.37	0.11
Crx	0.84	0.05	0.82	0.08	0.47	0.05	0.47	0.03
Glass	0.48	0.05	0.49	0.08	0.29	0.08	0.33	0.09
Iris	0.78	0.06	0.79	0.07	0.89	0.07	0.84	0.06
Pima	0.71	0.02	0.69	0.06	0.65	0.05	0.62	0.10
Wisconsin	0.97	0.01	0.97	0.00	0.84	0.09	0.86	0.07

3 Study on the Basic Conditions for Parametric Tests Using Artificial Neural Networks

In this section we will analyze the needed conditions which allow parametric test usage, and their fulfillment referred to the data sets and algorithms used.

In [8], the distinction between parametric tests and non-parametric tests is based upon measure level used over analyzed data. In such way, a parametric test use data in a real values contained in an interval. Although we dispose of that kind of values, a parametric test cannot be always used. It is possible that some initial suppositions are not fulfilled, resulting in loss of accuracy and credibility.

Therefore, in meanings of using parametric test, the fulfillment of these initial conditions is required[8,11]:

- Independency: Two events are independent if the occurrence of the first does not affect to the probability of the occurrence of the second.
- Normality: A observation is normal when its behavior follows a normal distribution with mean μ and variance σ. We can apply a normality test over the sample to verify whether if this condition is accomplished. We will use the Kolmogorov-Smirnov test. It compares the observed data accumulated distribution versus expected accumulated distribution from a Gaussian distribution, obtaining a p value based on the lack of similarity between them.
- Heteroscedasticity: This property indicates that a violation of the equality of variances exists. Levene's test is used to verify if k samples show this homogeneity. When sampled data does not verify normality condition, it is safer using Levene's test than Bartlett's one[11], which it is another test to check the same property.

As Demšar points out in [3], *independency* is not truly verified in 10-fold cross validation (since a portion of data set could be used either for training and testing in different partitions). Hold out partition can be safely taken as independent, since training and test partitions does not overlap.

We have applied the Kolmogorov-Smirnov test of *normality* with error probability $p = 0.05$ (we have used SPSS) for both 10fcv and hold out partitions. Table 3 shows the results, where the symbol '*' points out that normality is not verified. The value in parenthesis is the p value of confidence needed to reject hypothesis of normality.

Table 3. Results for Kolmogorov-Smirnov test

10-fold cross validation							
	Breast	cleveland	crx	glass	iris	pima	wisconsin
MLP backpropagation-1x25	* (.00)	* (.02)	* (.04)	(.20)	* (.00)	* (.00)	* (.00)
MLP backpropagation-1x5	* (.00)	(.20)	* (.00)	(.20)	* (.00)	* (.00)	* (.00)
RBFN Decremental	* (.00)	(.05)	* (.00)	(.08)	* (.00)	(.20)	* (.00)
RBFN	* (.00)	* (.04)	* (.00)	(.20)	* (.00)	* (.00)	* (.00)
Hold out partition							
	Breast	cleveland	crx	glass	iris	pima	wisconsin
MLP backpropagation-1x25	(.06)	(.20)	* (.00)	(.05)	* (.01)	(.20)	* (.00)
MLP backpropagation-1x5	* (.00)	* (.00)	* (.00)	* (.00)	(.20)	* (.00)	(.20)
RBFN Decremental	* (.02)	(.20)	* (.00)	(.20)	* (.00)	(.14)	* (.00)
RBFN	* (.01)	* (.01)	* (.00)	* (.01)	* (.00)	* (.00)	* (.01)

Referred to *heteroscedasticity* study, Table 4 shows results of Levene test for 10-fcv and hold out partitions, where the symbol '*' points out the variances of the distributions of the algorithms for a given data set which are not homogeneous.

Table 4. Results for Levene's test

	Breast	cleveland	crx	glass	iris	pima	Wisconsin
Levene 10-fcv	* (.00)	(.15)	* (.00)	(.10)	(.16)	* (.00)	* (.00)
Levene Hold-out	* (.00)	* (.00)	(.13)	* (.01)	(.26)	* (.00)	* (.00)

Finally, we can confirm that no conditions needed for parametric tests are verified:

1. Independency: As we have mentioned before, the use of 10-fold cross validation does not ensure independency of the results for each partition. However, hold out partition does, and [3] shows the most suitable partitions to avoid high Type I error rates.
2. Normality: The most of the Kolmogorov-Smirnov tests have shown that normality is not a common property of the experiments. For this reason, we cannot assume the presence of normality in our experiments.
3. Heteroscedasticity: In a very similar way than normality, heteroscedasticity is not a property we can expect finding in our experiments, due the low proportion of cases which fulfills the test.

An alternative of these are the non-parametric tests [3]. The majority of them are based on the ranking of the algorithms and the data sets used for evaluation.

4 On the Use of Rank-Based Non-parametric Tests: A Short Experimental Study

In this section, we briefly introduce non-parametric tests used and we present an experimental study using the four algorithms.

A non-parametric test is such that uses nominal data, ordinal data or ranked data. However, this does not mean that other data types cannot be used. It could be interesting to transform real data from an interval into ranked data by means of their order, so non-parametric tests can be applied on data which is typically used by parametric tests (when conditions for parametric tests application are not verified). Usually, a non-parametric test is less restrictive than parametric one, but less robust than a parametric test applied over data which verifies all needed conditions.

Next, we show the basis of each non-parametric tests used in this study:

– Friedman test [8], which is a non-parametric test equivalent of the repeated-measures ANOVA. Under the null-hypothesis, it states that all the algorithms are equivalent, so a rejection of this hypothesis implies the existence of differences among the performance of all the algorithms studied. After this, a post-hoc test could be used in order to find whether the control or proposed algorithm presents statistical differences with regards to the remain of methods into the comparison. One of them is the Bonferroni-Dunn test.

Friedman test way of working is described as follows: It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2, and so on. In 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, $R_j = \frac{1}{N}\sum_i r_i^j$. Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be equal, the Friedman statistic:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum jR_j^2 - \frac{k(k+1)^2}{4}\right] \tag{1}$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, when N and k are big enough (as a rule of a thumb, $N > 10$ and $k > 5$).

– Iman and Davenport test [5], which is a non-parametric test, derived from the Friedman test, less conservative than the Friedman statistic:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2} \tag{2}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. Statistical tables for critical values can be found at [8,11].

– Bonferroni-Dunn is a post-hoc test that can be used after Friedman or Iman-Davenport tests when they reject the null hypothesis. It is similar to the Tukey test for ANOVA. This method assumes that the performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference:

$$CD = q_\alpha / \sqrt{\frac{k(k+1)}{6N}} \tag{3}$$

q_α value is the critical value Q' for a multiple non-parametrical comparison with a control (see Table B.16 in [11]).

4.1 Experimental Study: Results and Analysis

In Table 5 we show the result of applying the tests of Friedman and Iman-Davenport, which search for differences in the results. In bold appears the greater value of the compared ones, and if it is the statistical then the null hypothesis is rejected. In our case, both Friedman and Iman-Davenport tests indicate the existence of significant differences between results of 10fcv and hold out validation. Due to these results, a posteriori statistical analysis is needed. In Figures 1 and 2 we show the application of Bonferroni-Dunn test. These graphics represent a bar chart, which height is proportional to the mean rank obtained from each algorithm. If we sum to the lower of those (the best algorithm) the Critical Difference value (CD value), we obtain a horizontal line (denoted as "Threshold"),

Table 5. Results for Friedman and Iman-Davenport tests

Method	Friedman Value	Value of χ^2	Iman-Davenport Value	F_F Value
10fcv	**9.686**	7.815	**5.136**	3.160
Hold out	**8.657**	7.815	**8.657**	3.160

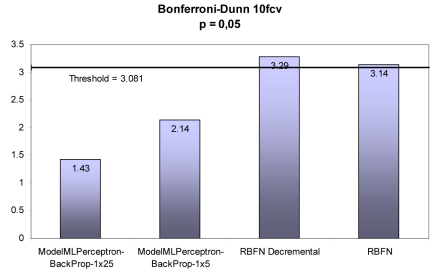

Fig. 1. Bonferroni-Dunn for 10fcv

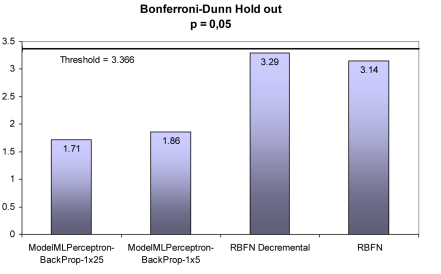

Fig. 2. Bonferroni-Dunn for hold out

and those bars that exceeds this line are algorithms with significantly worse results than the control algorithm (associated with the lowest bar).

As we can see, in Figures 1 and 2, results for 10 fold-cross validation are not the same for both ANN models used: behavior of RBFN based ANNs is different from MLP with Backpropagation-1x25 (the equivalence of means hypothesis is rejected), and MLP with backpropagation-1x5 cannot be distinguished with $p = 0.05$. In the same way, for hold-out, both tests do not find differences between the algorithms with $p = 0.05$.

A further analysis of all results allows us to conclude:

- Observing Figures 1 and 2, the RBFN networks show a behaviour very different from MLP backpropagation. Bonferroni-Dunn test considers the results enough away from themselves in order to detect significant differences among them.
- The greatest differences are found when using a 10 fold cross validation. Using hold out partition and $p = 0.05$ we cannot assume the existence of differences.
- The worst algorithm, RBFN Decremental, is the last in Bofferoni-Dunn as we could expect.
- However, MLP backpropagation-1x25 always computes as the best, Bonferroni-Dunn test does not consider that there exist differences with MLP backpropagation-1x5 in any case, with $p = 0.05$.

5 Conclusions

The present work studies the use of statistical techniques for analysis of ANNs in classification problems, and a further analysis of parametric and non-parametric tests.

The need of using non-parametric tests is pretty clear when analyzing ANNs for classification, since initial conditions required for safe results from parametric tests are not met.

On the use of non-parametric tests, we have shown that Friedman, Iman-Davenport and Bonferroni-Dunn are a good set of tools for testing algorithms.

Indeed, there exist more powerful tests than Bonferroni-Dunn test, i.e. Holm, Hommel and Hochberg test. We can find an example of use of them in [7]. Regarding to the comparison by pairs, the Wilcoxon test may be a good election [9].

Acknowledgement. This work was supported by the project TIN2005-08386-C05-01.

References

1. Alpaydin, E.: Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. Neural Computation 11, 1885–1892 (1999)
2. Castillo-Valdivieso, P.A., Merelo, J.J., Prieto, A., Rojas, I., Romero, G.: Statistical analysis of the parameters of a neuro-genetic algorithm. IEEE Transactions on Neural Networks 13, 1374–1394 (2002)
3. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1–30 (2006)
4. Gao, D., Madden, M., Chambers, D., Lyons, G.: Bayesian ANN classifier for ECG arrhythmia diagnostic system: A comparison study. In: Proceedings of the International Joint Conference on Neural Networks, vol. 4, pp. 2383–2388 (2005)
5. Iman, R. L., Davenport, J. M.: Approximations of the critical region of the Friedman statistic. Communications in Statistics, pp. 571–595 (1980)
6. Rojas, R., Feldman, J.: Neural Networks: A Systematic Introduction. Springer, Heidelberg (1996)
7. Shaffer, J.P.: Multiple Hypothesis testing. Annual Review of Psychology 46, 561–584 (1995)
8. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press, Boca Raton (2000)
9. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics 1, 80–83 (1945)
10. Yingwei, L., Sundararajan, N., Saratchandran, P.: A sequential learning svheme for function approximation using minimal radial basis function neural networks. Neural Computation 9, 361–478 (1997)
11. Zar, J.H.: Biostatistical Analysis. Prentice-Hall, Englewood Cliffs (1999)
12. Zekic-Susac, M., Horvat, J.: Modeling computer and Web attitudes using neural networks. In: Proceedings of the International Conference on Information Technology Interfaces, vol. 4, pp. 2383–2388 (2005)