ELSEVIER

Neurocomputing 48 (2002) 155–173

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Multiple comparison procedures applied to model selection

Joaquín Pizarro, Elisa Guerrero, Pedro L. Galindo*

Departamento Lenguajes y Sistemas Informáticos e Inteligencia Artificial Grupo "Sistemas Inteligentes de Computación", Universidad de Cádiz, Spain

Received 31 October 2000; accepted 6 June 2001

Abstract

This paper presents a new approach to model selection based on hypothesis testing. We first describe a procedure to generate different scores for any candidate model from a single sample of training data and then discuss how to apply multiple comparison procedures (MCP) to model selection. MCP statistical tests allow us to compare three or more groups of data while controlling the probability of making at least one Type I error. The complete procedure is illustrated on several model selection tasks, including the determination of the number of hidden units for feed-forward neural networks and the number of kernels for RBF networks.

Keywords: Model selection; Multiple comparison procedures; Generalization; Network size; Problem complexity

1. Introduction

Many model selection algorithms have been proposed in the literature [35]. The existing procedures can roughly be categorized as analytical or resampling based. Analytical approaches require certain assumptions of the underlying statistical model. Resampling based methods involve much more computation, but they remove the risk of making faulty statements due to unsatisfied assumptions [10]. With the computer power currently available, this does not seem to be an obstacle.

Standard methods of model selection include classical hypothesis testing [35], maximum likelihood [2], Bayes method [29], cross-validation [31], Akaike's information criterion [1] and many more. Probably the most widely accepted procedure

* Corresponding author. Tel.: 34-956-016434; fax: 34-956-016437.
E-mail address: pedro.galindo@uas.es (P.L. Galindo).

0925-2312/02/$ - see front matter
PII: S0925-2312(01)00653-1

is the use of an information criterion based on choosing the model with the maximized log-likelihood function minus a penalty. However, there is little agreement about what the form of the penalty function should be. Although, there is active debate within the research community regarding the best method for comparison, statistical model selection is a reasonable approach [21].

We consider the general problem of determining which of a set of competing models is better. A statistical approach to model selection should try to find out which model is better on average. One way to define "on average" is to consider the performance of these algorithms averaged over all the training sets that might be drawn from the underlying distribution [25]. In a real situation, the underlying distribution is unknown, and we only have a finite size sample to work with.

The simplest approach to estimate the error for each model is to divide available data into a training set and a disjoint test set (hold-out method). However, the relative performance may be dependent on the choice of training and test sets. One way to improve this estimate is to repeatedly partition the data into disjoint training and test sets and to take the mean of the test set errors for these different experiments. The goal of our strategy will be the correct design of these batteries of experiments to take into account the sources of variation that should be controlled and to analyze the errors for each model to determine if differences among models exist.

In the following sections, we first describe the design of a randomized data collecting procedure, taking into account the different sources of variation that could exist. After collecting the data, our goal will be to make inferences about k population means. To get around this problem, we need tests that compare groups of data. These are the analysis of variance tests (parametric/nonparametric, independent/repeated measures). Although, these tests allow us to reject the null hypothesis that the groups' means are all equal, they do not determine where the significant differences lie [14]. To accomplish this, a naïve approach is to test each possible pair of groups by a paired t-test. However, multiple t tests are not appropriate because the probability of a Type I error increases with the number of comparisons made. Statistical methods to compare three or more means while controlling the probability of making at least one Type I error are called *multiple comparisons procedures* [15]. We briefly discuss these methods, including Fisher's LSD, Tukey's HSD, Bonferroni, Sidak, Scheffé, Dunnett and Hsu's RSMCB procedures and comment their potential advantages.

Finally, we apply these techniques to the determination of the optimal complexity of a model on various artificial and real problems (both, classification and regression tasks are considered) and show examples where the appropriate model complexity is known in advance, observing that it performs well in these situations.

2. Design of the experiment

To design and evaluate statistical tests, the first step is to identify the sources of variation that must be controlled by each test. A source of variation is anything that could cause an observation to have different numerical value from other

observations. Dietterich [7] studied different statistical tests for comparing supervised learning algorithms and the sources of variation that a good statistical test should control:

Random variation in the selection of the test data: On any particular randomly-drawn test data set, one model may outperform another, even though on the whole population the two models could perform identically.

Random variation in the selection of the training data: On any particular randomly drawn training data set, one model may outperform another, even though on average, the two models have the same accuracy. Even small changes to the training set, such as adding or deleting a few data points, may cause large changes in the estimated parameters of models.

Internal randomness in the estimation of model parameters: If the estimation of parameters is analytical and its determination is unique, this source may be omitted because there is no internal randomness. However, in an iterative approach the results depend critically on the starting state. Most of the iterative procedures suffer from internal randomness due to the initialization of the parameter set. This parameter set depends on the model complexity, so it is different in value and number for each model.

Ideally, the population is considered to have an infinite number of samples. However, in real situations, the amount of data available is only a subset (sometimes only a few data) of the overall population. A fundamental assumption is that this collection of known cases is representative of the entire population. For a finite set of data, these sources of variation should be controlled as follows:

- The learning algorithms should be executed multiple times over different training and test sets to control the variation due to the choice of training and test data sets.
- If any model is trained and tested on a given training and test data set, any other model should be trained and tested with the same set. This ensures that all models are compared under the same conditions. It also helps to control the variations due to the choice of training and test data sets, and allows us to apply statistical pairwise tests.
- Each unstable algorithm should be executed several times, taking different starting states for each training data set to reduce the variance due to internal randomness.

As explained above, we also need a method to obtain different measures of error for each learning algorithm. In order to get different data sets for training and testing the most common procedures are the following:

- *Systematic selection of subsets from the original data set*: K-fold cross-validation related techniques ($2\times$CV, $10\times$CV, leave-one-out, stratified CV, ..., etc.) might be included in this approach.
- *Repeated permutations* [19]: New data sets are created by permuting available data randomly. This has the same effect as sampling N data randomly without

replacement. After generation of data, holdout may be applied. It is also known as random subsampling.
- *Bootstrapping* [6,9]: New data sets are created by sampling N data randomly with replacement, so the resulting data set has the same size as the original, but some data have been left out (these data are used for testing purposes) and others are duplicated.
- Combinations of the above.

Depending on the strategy selected, we may find nonindependent training sets (in all cases, at different levels), nonindependent test sets (resampling cases), very small test sets (leave-one-out), or a limited number of data sets that may be generated (systematic selection of subsets). As a consequence, any design devised to work with a finite set of data will violate the fundamental assumption of statistical methodology, sampling independence. What is more, statistical designs cannot be viewed as rigorously correct, but only approximate.

In the design of the experiment we use repeated permutations followed by two-fold cross-validation, because it gives a trade-off between large test sets and completely disjoint training sets, at least, on pairs of consecutive samples. We recommend at least 30 error measures per model, in order to guarantee that the error samples will provide a good estimate of the distribution of errors. For a given training and test set, each unstable algorithm is trained for 10 times. We focus our study on the model behavior on average, so the mean of these errors is considered to be the actual error of the model. Extreme error values (the minimum and maximum error estimates) are excluded in the computation of the error mean to reduce the influence of the appearance of local minima in the learning process. The strategy is summarized as follows:

```
for iteration=1 to 30 (at least)
    Random selection of training and test sets
    for model=1 to H (H=number of models)
        for r=1 to 10 (to avoid internal randomness)
            Train model
            Error(r) = Compute Error Measure
        end
        GlobalError(iteration,model)=Average(Error) (excluding extreme cases)
    end
end
```

3. Data analysis

As a first step, we may consider the use of a paired sample t-test to assess whether the means of two populations are not different. However, if we are interested in testing whether the means of more than two populations are equal, the appropriate inferential statistic will depend on the underlying distributions [4]. Both, the parametric test (ANOVA) and the nonparametric test (Kruskal–Wallis), are adequate for testing the differences between more than two samples. They look

at how much variation or spread there is in each sub-group. The more within-group variation that there is in each sub-group, the more difficult it will be to positively say that there is a difference among the groups. However, if the populations from which data to be analyzed violate some assumptions, the results of the analysis may be incorrect or misleading. ANOVA test may be used if the following assumptions/requirements are met:

- *Normal distribution*: The ANOVA test functions fairly well with deviations from normality if the sample sizes are nearly equal [3]. This assumption has been tested using the method of Kolmogorov–Smirnov and we have nearly always found that the distribution of results follows a Gaussian curve.
- *Homoscedasticity—Homogeneous variances*: The most common method employed to test for homogeneity of variances is Bartlett's test [23]. This test is powerful when the sample populations are normal, but it is badly affected by nonnormal populations. ANOVA's are pretty reliable even if the equal variance assumption is violated, if the sample sizes are all equal. In our design the number of error measures is the same in all the models. In the experiments, lower complex models exhibit greater variance due to underfitting, while more complex models exhibit nearly equal variances between them. We expect robustness against these inequalities.
- *Independence of observations*: This assumption is in practice difficult to test. We must think about the experimental design. As the sources of variation have been taken into account, we assume random and independent data samples. Strictly speaking, the independence of the samples is not verified in our design, given that different results have been obtained from splitting randomly the available data, which are finite sized. However, by considering pairwise comparisons, the violation of this assumption might be considered secondary.

Fortunately, the analysis of variance is robust with respect to the assumption of the underlying population's normality, operating well even with considerably heterogeneity of variances, as long as all groups have the same size. Anyhow, if the data are highly skewed or if the variances of the different populations are very unequal, then we can, either transform the data to change the scale of the values or use a non-parametric version of the analysis of variance, called Kruskal–Wallis test. The Kruskal–Wallis test provides a nonparametric alternative to the ANOVA test for comparing more than two populations based on independent random samples by using rank sums to calculate an H-test statistic that possesses an approximate χ^2 sampling distribution. This test is 95% as powerful as a single-factor ANOVA test, and much better when the assumptions of the ANOVA test are not true.

Repeated measures designs [18], often referred to as within-subjects designs, offer greater statistical power relative to sample size. They test for significant differences among the means of two or more groups when the observations come from matched units. With such designs you actually get more power because you can factor between-subject differences out of the error term, thereby resulting in larger F values. When the proper assumptions are not met, nonparametric

Friedman's analysis of variance by ranks may be used. In this case, no assumptions are made about the population.

We should be very careful when applying non-parametric tests [13], because they are less sensitive to the detection of differences when the assumptions are satisfied. Therefore, if a parametric test is appropriate, it should be used because it provides a better chance of finding significances when they exist. Only when the parametric test is not appropriate, should a nonparametric test be used.

4. Multiple comparison procedures

When comparing more than two means, analysis of variance tells you whether the means are significantly different from each other, but it does not tell you which means differ from each other [17].

The first idea that comes to mind is to test each possible difference by a paired t-test. The problem with multiple individual comparisons is that when we compute several tests we increase our chances of obtaining a significant result by chance alone. We should bear in mind that each comparison is typically done with the level of significance set at a probability of 0.05 which means that on 5% of occasions we will reject the null hypothesis when in fact it is true. This means that the level of significance for the experiment soon rises to unreasonable levels. For example, if a one-way analysis of variance is computed on five groups and indicates that there is a significant difference among the groups, then there will be a total of $n(n-1)/2 = 10$ pairwise comparisons that can be made. If a simple t-test comparison is made on each of these ten possible comparisons and an 0.05 level of significance is used for each, then the experimentwise level of significance is $1-(1-0.05)^{10} = 1-0.5987 = 0.4013$. In other words, there is a 40% chance of making a Type I error. Statistical methods to compare three or more means while controlling the probability of making at least one Type I error are called MCP.

4.1. Description

In general, multiple comparisons of several groups should be performed only as a follow-up analysis to the appropriate analysis of variance F-test, i.e., only after we have determined that sufficient evidence exists of differences among the means [30]. In this section, we will describe some of the methods that adjust for the multiplicity of tests. In all multiple comparison testing, equal sample sizes are desirable for maximum power and robustness, and the experiment has been designed keeping this in mind. So, all procedures are presented for analysis with equal n.

Let $\bar{y}_i$ and n_i be the mean and sample size of group i and $\bar{y}_j$ and n_j be the mean and sample size of group j, respectively. Two groups will be considered significantly different if their corresponding means are bigger than the 'critical value'.

As shown below, all the tests define 'critical values' based on the square-root of the estimated variance of $\bar{y}_i - \bar{y}_j$, that will be noted as $\hat{\sigma}_{ij}$:

$$\hat{\sigma}_{ij} = S_{\mathrm{VNE}} \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = S_{\mathrm{VNE}} \sqrt{\frac{2}{n}},$$

where S_{VNE} is the within-sample variation with $(n-k)$ degrees of freedom, being k the number of models considered, and n the number of samples for each model:

$$S_{\mathrm{VNE}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{y}_i)^2}{n \cdot k - k}.$$

A large number of multiple comparison procedures have been developed. Among the most commonly used methods are the following:

Fisher's least significant difference (*LSD*) [5]: If the overall F-ratio (which tests the hypothesis that all group means are equal) is statistically significant, we can safely conclude that not all the treatment means are identical and then, and only then, we compare all possible combinations of the group means, taking two at a time, while controlling the level of significance. Two groups are not significantly different if:

$$|\bar{y}_i - \bar{y}_j| \geqslant t(\alpha, n-k)\hat{\sigma}_{ij},$$

where $t(\alpha, n-k)$ is the α-level critical value from a two-tailed Student's t distribution with $(n-k)$ degrees of freedom. The method is undesirable if the number of groups is large, for, in fixing a significance level, we are controlling the individual probability of false rejection for each pair, rather than the overall probability of some false rejection.

Tukey's honestly significant differences (*HSD*) [33]: It is based on a Studentized range distribution (q statistic) which is similar to the Student distribution but taking into account the number of treatments being considered. Two groups are not significantly different if:

$$|\bar{y}_i - \bar{y}_j| \geqslant q(\alpha, k, n-k)\hat{\sigma}_{ij},$$

where $q(\alpha, k, n-k)$ is the α-level critical value of a studentized range distribution of k independent normal random variables with $(n-k)$ degrees of freedom [24].

Bonferroni correction: The Bonferroni approach is a follow-up analysis to the ANOVA method [30] and is based on the following result. If c comparisons are to be made, each with confidence coefficient $(1-\alpha/c)$, then the overall probability of making one or more Type I errors is at most α. Two groups are not significantly different if:

$$|\bar{y}_i - \bar{y}_j| \leqslant t(\alpha/c, n-k)\hat{\sigma}_{ij},$$

where $t(\alpha/c, n-k)$ is the α/c-level critical value from a two-tailed Student's t distribution with $(n-k)$ degrees of freedom.

Sidak test: The Sidak test [17] is a variant on the Bonferroni approach, using a t-test for pairwise multiple comparisons, where the α significance level for multiple

comparisons is adjusted to tighter bounds than for the Bonferroni test:

$$|\bar{y}_i - \bar{y}_j| \leqslant t(1-(1-\alpha)^c, n-k)\hat{\sigma}_{ij}.$$

Scheffé test [27,28]: It assumes all possible pairs and all possible combinations of means are to be tested. It works by first requiring the overall F-test of the null hypothesis be rejected. Two groups are not significantly different if:

$$|\bar{y}_i - \bar{y}_j| \leqslant \sqrt{(k-1)F(\alpha, k-1, n-k)}\hat{\sigma}_{ij},$$

where $F(\alpha, k-1, n-k)$ is the α-level critical value of an F distribution with $(k-1)$ numerator degrees of freedom and $(n-k)$ denominator degrees of freedom.

Dunnett test [8]: It is a t-statistic that is used when the researcher wishes to compare each treatment group mean with the mean of the control group, and for this purpose has better power than alternative tests. Any group is significantly different from the control one if:

$$|\bar{y}_i - \bar{y}_{\text{Control}}| \geqslant w_2\hat{\sigma}_{ij},$$

where $w_2 = t^{\rho}_{k-1,n-k,\alpha}$ is the percentile of the maximum of a two-tailed multivariate t distribution with common correlation ρ and $(n-k)$ degrees of freedom at the α-level.

Ranking, selection and multiple comparisons with the best treatment (RSMCB). Hsu [16] developed a method in which each sample mean of a treatment is compared with the best of the other treatments, allowing some of them to be eliminated as worse than best, and allowing one treatment to be identified as best if all others are eliminated. Any group is significantly different from the best one if:

$$|\bar{y}_i - \bar{y}_{\text{Best}}| \geqslant w_1\hat{\sigma}_{ij}.$$

The critical coefficient w_1 is the same as that for Dunnett's, but one-tailed confidence bound.

4.2. Discussion

Although, there is no correct procedure to use, most researchers believe that procedures like Fisher's protected LSD procedure should not be used since they do not control the overall confidence level nor the experimentwise error rate. The remaining procedures discussed in this section keep the experimentwise error rate at the specified significance level, but they might be less powerful for testing all pairwise comparisons. Bonferroni, Sidak and Scheffe's tests are most conservative methods. LSD is the least conservative, but, as we mentioned above, it is not recommended. Tukey's HSD test is somewhat in-between, and it is used frequently when a comparison between all pairs of means is needed. It is preferred when the number of groups is large as it is very conservative. Bonferroni test gives shorter confidence intervals than other methods if c is small. Scheffé's test gives shorter intervals than Bonferroni's method if n is large. Dunnett's test is used when you want to compare the mean of a control to the other group means, rather than

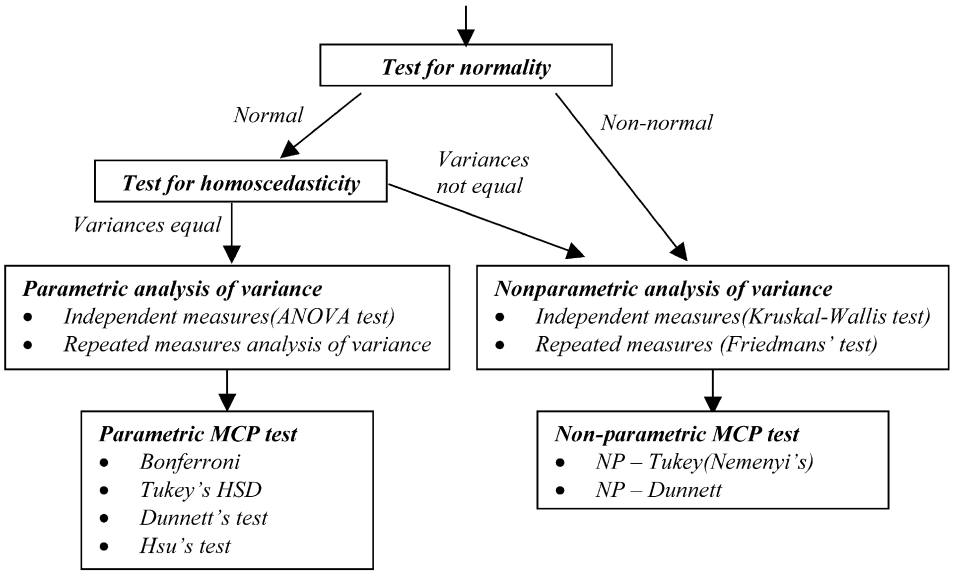

Fig. 1. Schematic illustration of the complete procedure for significance testing.

comparing all means to each other. Hsu's is similar to Dunnett's test, except that it is considered known prior to the experiment which treatment is the best.

The choice of a multiple comparison test should be also governed by a logical analysis of the seriousness of making an error. If falsely rejecting the null hypothesis would have serious consequences, then we should select a more conservative method. In this case, the level of significance for each comparison should be set very low ($\alpha = 0.001$). If the experiment is exploratory, then more powerful tests may be considered and a moderate level of significance ($\alpha = 0.1$) might be selected.

Finally, let us consider the situation where the assumptions of normality are not met, and the nonparametric Kruskal–Wallis test is applied. In this case, we will also have the need of a nonparametric multiple comparison test. This may be done using rank sums instead of means, resulting in tests analogous to Tukey (Nemenyi's method) or Dunnett (Steel method) testing. These techniques are discussed at length in [34].

Fig. 1 shows an schematic illustration of the recommended procedure for testing. In our experiments, we do not test for homoscedasticity, given that all the groups have the same size.

5. Experimental results

In order to illustrate our strategy we conducted a range of experiments on both simulated and real data sets. Unless stated otherwise, the original data were not

preprocessed. To compute the error measure (the ratio of misclassified patterns for classification problems and the mean squared error, MSE, for regression problems), two-fold crossvalidation is used.

We consider three different algorithms for classification to which we refer as KNN (K-nearest neighbor), MLP (multilayer perceptron) and RBF (Radial basis function network).

KNN implements the most basic instance-based method [12]: The K-nearest neighbor algorithm with Euclidean distance, where ties have been solved randomly. We have not considered the use of a reject option in instances where there is not a clear 'winner'.

MLP represents a multilayer perceptron having two layers of weights with full connectivity between adjacent layers. One linear output unit, M 'tanh' hidden units and no direct input–output connections [20]. In the experiments, the weights of the MLP network were randomly initialized and 200 iterations were performed using the Levenberg–Marquardt algorithm [2].

RBF represents a radial basis function network having one hidden layer for which the combination function is the Euclidean distance between the input vector and the weight vector, and the activation function is the exponential [2]. The placement of the kernel functions has been accomplished using the k-means algorithm. The width of the basis functions has been set to

$$\sigma = \frac{\|\max(\vec{x}_i - \vec{x}_j)\|}{\sqrt{2n}},$$

where n is the number of kernels and $\|\cdot\|$ denotes the Euclidean norm. The second layer of the network is a linear mapping from the RBF activations to the output nodes. Output weights are computed via matrix-pseudoinversion [26]. Only for RBF algorithm, we normalize each component to have zero mean and unit variance based on training test statistics.

For regression we have considered two different algorithms to which we refer as POL (polynomial fitting) and NLLSQ (nonlinear least squares fitting). POL finds the coefficients of a polynomial of degree N that fits the data in a least-squares sense. NLLSQ finds the coefficients to best fit a given nonlinear function to the data in the least-squares sense.

In most of the cases, the populations follow a normal distribution (Kolmogorov–Smirnoff test is applied), so ANOVA test and parametric multiple comparison test are used. In this situation, analysis of variance F-value was always significant, and it was not necessary to use a repeated measures design. Occasionally, there are experiments where one or two populations do not fit normal distributions, and Kruskal–Wallis and nonparametric Tukey tests are applied.

5.1. Classification experiments

In the first experiment, we use the two-dimensional artificial data (two-class problem) shown in Fig. 2. The training data set consists of 270 points per class

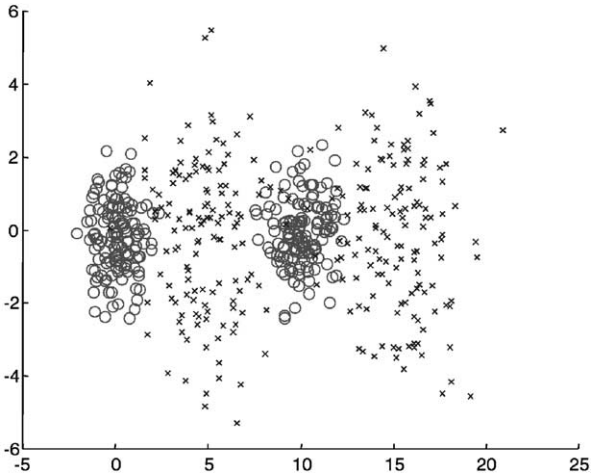

Fig. 2. Sample data distribution for an artificial classification experiment.

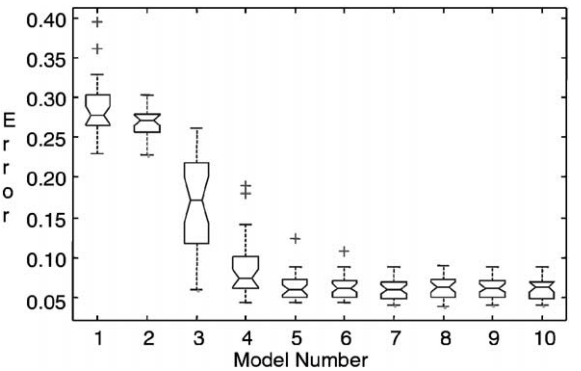

Fig. 3. Box and whisker plot for errors (models 1 through 10).

and was artificially generated from the following bivariate normal distributions:

Class 1: $\mu_1 = (0,0)$, $\Sigma_1 = I$, $\mu_2 = (10,0)$, $\Sigma_2 = I$,
Class 2: $\mu_1 = (5,0)$, $\Sigma_1 = 2I$, $\mu_2 = (15,0)$, $\Sigma_2 = 2I$.

MLP is used as learning algorithm in order to train 10 different models having from 1 through 10 hidden units, respectively. A sample of 30 error measures per model is collected as described in Section 2. Fig. 3 shows a box and whisker plot for errors obtained from each model. The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.

The assumptions to the proper application of the ANOVA test are satisfied, and in this case, the differences among error means are significantly different at

Table 1
MCP Bonferroni's test results (critical value = 0.0199)

Hidden units	Error mean	Error std	Models not significantly different
7 (control)	0.05956	0.0134	7 9 10 8 6 5
9	0.06040	0.0314	7 9 10 8 6 5
10	0.06086	0.0119	7 9 10 8 6 5
8	0.06091	0.0157	7 9 10 8 6 5
6	0.06114	0.0137	7 9 10 8 6 5
5	0.06286	0.0251	7 9 10 8 6 5
4	0.08573	0.0452	4
3	0.16706	0.0727	3
2	0.26877	0.0260	1 2
1	0.2869	0.0154	1 2

Table 2
MCP tests critical values for the estimated variance of differences[a]

Test α-value	Fisher's LSD	Tukey's HSD	Bonferroni correction	Sidak test	Scheffé test	Dunnett test	Hsu's RSMCB
0.1	0.0116	0.0206	0.0183	0.0216	0.0271	0.0170	0.0149
0.05	0.0139	0.0222	0.0199	0.0231	0.0292	0.0189	0.0170
0.01	0.0183	0.0257	0.0234	0.0263	0.0332	0.0227	0.0211

[a] $\hat{\sigma}_{ij} = 0.0273$ and different values of significance ($\alpha = 0.1, 0.05, 0.01$)

the confidence level of $\alpha = 0.05$. Now, we might wish to know which models differ significantly from each other, thus, the application of multiple comparison procedures should be carried out.

We first select the model with the lowest error mean as the control treatment (the a priori best model is, in this case, model 7). After carrying out a multiple comparison procedure, we select the simplest model that is not significantly different from the control model, following Occam's razor criterion [2], so as to obtain better generalization ability.

Table 1 shows the models, which are denoted by the number of hidden units they have, the corresponding error means, standard deviations and Bonferroni test results. Models from 5 through 10 are considered not significantly different from the control model. Two models are not significantly different if the difference between its means is less than 0.01991. Thus, model 5 should be selected, since it is the simplest model not significantly different from that with lowest mean error.

In Table 2, we show critical values for different MCP tests. Let us note that Fisher's LSD is the most powerful, followed by Hsu's RSMCB. However, LSD procedure should not be used since it does not control the experimentwise error rate. For this reason, in successive experiments, results will be reported on Hsu's test.

Now, we will show some examples illustrating our strategy for the determination of model complexity using real data sets. The aim of the experiments will be to

determine the optimal value of K (in the KNN algorithm) or the number of hidden neurons (in the BP and RBF algorithms). Databases, coming from the UCI [22] (Diabetes, Heart, Cancer, Vehicle and Iris) and ELENA Project [32] (Clouds) repositories will be used. A detailed description of each database may be found in the repositories themselves.

The results are given in Tables 3–5. The first row shows the model with the lowest mean of the error measures. In the following rows, models whose means are not significantly different with the lowest are shown. A set of parametric multiple comparison test are used to compare the results and when, necessary, nonparametric Tukey test is used. In all the cases, ANOVA test (and Kruskal–Wallis test when necessary) is significant. In the last row, the model chosen according to Hsu test (or nonparametric Tukey, when needed) and the corresponding error mean is shown. Model N implies N hidden units (MLP and RBF algorithms) or $K = N$ (KNN algorithm). KNN has been trained with values of K between 1 and 15, MLP networks trained with a number of hidden units between 1 and 15 and RBF networks between 1 and 20 kernels.

Table 3
Model selection strategy applied to classification tasks. Algorithm K-NN. ($1 \leqslant K \leqslant 15$)

	Diabetes	Heart	Clouds	Cancer	Vehicle	Iris
Control group (lowest error mean)	11	15	15	6	3	8
Fisher	7–15	5–7, 10–15	8–15	3–15	1–4	3–10
Tukey	7–15	3–15	5–15	3–15	1–6	3–12
Bonferroni	6–15	3–15	5–15	3–15	1–8	3–12
Scheffé	5–15	3–15	5–15	3–15	1–9	1–14
Sidak	6–15	3–15	5–15	3–15	1–7,9	3–12
Dunnett	7–15	3,5–15	7–15	3–15	1–6	3–10,12
Hsu	7–15	3,5–15	7–15	3–15	1–5	3–10,12
Nonparametric Tukey						3–15
Model/Error mean	7/24.40	3/35.23	7/11.85	3/3.54	1/36.56	3/4.17

Table 4
Model selection strategy applied to classification tasks. Algorithm MLP. ($1 \leqslant N \leqslant 15$)

	Diabetes	Heart	Clouds	Cancer	Vehicle	Iris
Control group (lowest error mean)	1	1	15	2	10	3
Fisher	1–2	1–2	11–15	1–3	5–15	2–15
Tukey	1–3	1–2	7–15	1–3	4–15	2–15
Bonferroni	1–3	1–2	7–15	1–3	4–15	2–15
Scheffé	1–5	1–3	7–15	1–4	3–15	2–15
Sidak	1–3	1–2	7–15	1–3	4–15	2–15
Dunnett	1–3	1–2	9–15	1–3	5–15	2–15
Hsu	1–2	1–2	9–15	1–3	5–15	2–15
Nonparametric Tukey			7–15			
Model/Error mean	1/24.21	1/20.31	7/11.85	1/4.52	5/24.55	2/5.14

Table 5
Model selection strategy applied to classification tasks. Algorithm RBF. ($1 \leqslant N \leqslant 20$)

	Diabetes	Heart	Clouds	Cancer	Vehicle	Iris
Control group (lowest error mean)	14	17	20	18	20	11
Fisher	11–20	12–20	18–20	3,4,13–20	15–20	6–20
Tukey	7–20	10–20	15–20	3–7, 9–20	14–20	5–20
Bonferroni	7–20	9–20	14–20	3–7, 9–20	14–20	5–20
Scheffé	6–20	8–20	10–20	3–20	13–20	5–20
Sidak	7–20	10–20	14–20	3–7, 9–20	14–20	5–20
Dunnett	7–20	10–20	17–20	3–5, 11–20	14–20	6–20
Hsu	7–20	10–20	17–20	3–5, 12–20	14–20	6–20
Nonparametric Tukey						
Model/Error mean	7/25.41	10/18.76	17/17.92	3/4.00	14/34.79	6/5.33

We see from Tables 3–5 that Fisher's LSD gives the shortest intervals around the control due to an uncontrolled experimentwise error rate, while Scheffé's test obtains the widest ones due to a well-known limited power of the test. The remaining multiple comparison tests (parametric and nonparametric) gave very similar results. If ANOVA F-test is not significant, which was not the case in any experiment, we might conclude that either the problem has a low-complexity or the sample size is not large enough with respect to the complexity of the problem. In this case statistical tests may be inconclusive.

On the other hand, if the error means seems to be decreasing with the complexity of models and the most complex ones are selected as the best group, more complex models should be analyzed in the experiment. Let us observe, for instance, the results obtained by RBF algorithm on the *clouds* database. Models not significantly different from the control model (model 20), following Hsu criterion, are indeed sorted in descending order of error means, from 17 through 20, and model 17 is selected as the best model. We recommend to repeat the experiment with more complex models (from 21 through 30) and determine if they are not significantly different from the others, thus ensuring that 17 is the best one.

5.2. Regression experiments

Let us consider now the problem of finding the degree N of a polynomial $P(x)$ that better fits a set of data in a least-squared sense. The experimental polynomial is $P(x) = 0.4x^3 - 0.5x^2 - 0.25x + \varepsilon$, where the values $x \in [-1\ 3]$, and ε is zero mean, unit variance Gaussian noise. Fig. 4 shows the set of 160 data points that will be used in the experiment.

Polynomials with degrees ranging from 1 to 10 are used. The only aspect of the polynomials that remains to be specified is the degree (M). A sample of 30 MSE errors for each polynomial has been generated. As ANOVA test assumptions are not satisfied (models 8–10 do not follow normal distribution), Kruskal–Wallis test is used instead. This test is significant and nonparametric Tukey test is applied to determine whether the observed differences in the sample means imply that

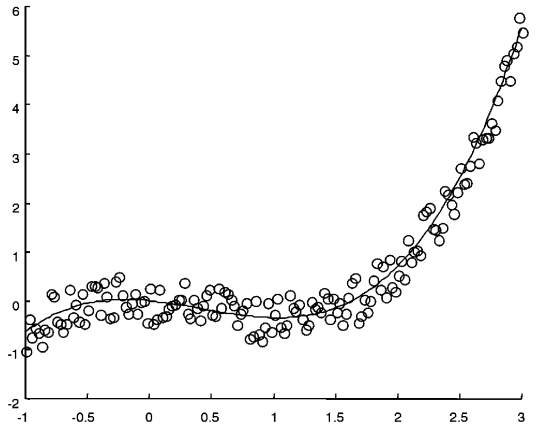

Fig. 4. 160 data points from a curve-fitting problem, with the true curve, a third-order polynomial.

Table 6
Polynomial fitting simulation results (160 data points, nonparametric Tukey test)

Polynomial degree	MSE mean	MSE Std	Polynomial degrees not significantly different
3	7.17314	0.6844	3 4 5 6 7 8 9 10
4	7.18805	0.7006	3 4 5 6 7 8 9 10
5	7.27101	0.7549	3 4 5 6 7 8 9 10
6	7.43246	0.7894	3 4 5 6 7 8 9 10
7	7.71308	1.2842	3 4 5 6 7 8 9 10
8	7.97803	1.8819	3 4 5 6 7 8 9 10
9	8.19752	1.8853	3 4 5 6 7 8 9 10
10	9.02928	4.2481	3 4 5 6 7 8 9 10
2	30.55481	4.3711	2
1	83.92423	13.8307	1

differences exist among the accuracy of the competing polynomials. The overall confidence level is fixed to 0.05.

Table 6 shows polynomial degrees, their corresponding MSE errors mean and standard deviations and degrees of the set of polynomials not significantly different from that of the first column. Two polynomials are not significantly different if the difference between its means is less than the critical value computed in this case through a nonparametric Tukey test as 4.47. Polynomials from degrees 3 to 10 form a not significantly different MSE group and model 3 is selected.

Let us consider now a nonlinear least squares regression problem. The synthetic dataset "add10", coming from the UCI repository [22], uses a function suggested by Friedman [11]. The true function is

$$f(x_1,\ldots,x_{10}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon,$$

Table 7
Nonlinear least squares regression results ('add10' data set)

Model order	MSE mean	MSE Std	Models not significantly different
3	4838.75	70.6805	3 4 5 6 7 8 9
4	4839.83	70.7937	3 4 5 6 7 8 9
5	4840.08	70.4030	3 4 5 6 7 8 9
6	4841.32	69.9378	3 4 5 6 7 8 9
7	4843.28	69.8620	3 4 5 6 7 8 9
8	4844.07	69.9044	3 4 5 6 7 8 9
9	4848.75	97.8153	3 4 5 6 7 8 9
2	19282.31	231.4177	2
1	114593.74	1219.9864	1

where ε is zero mean, unit variance Gaussian noise. The inputs $x_1,\ldots,x_{10}$ are sampled independently from a uniform (0,1) distribution. Let us assume we know that the function has the form:

$$f(x_1,\ldots,x_{10}) = \beta_1 \sin(\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2 + \sum_{i=3}^{N} \beta_i x_{i+1}, \quad N < 10,$$

but we do not know how many input parameters are necessary to fit the data (the true model needs the five first parameters to fit the data and the others are unnecessary). To answer this question nine models are defined:

Model 1: $f(x_1,\ldots,x_{10}) = \beta_1 \sin(\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2$

Model 2: $f(x_1,\ldots,x_{10}) = \beta_1 \sin(\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2 + \beta_3 x_4$

Model 3: $f(x_1,\ldots,x_{10}) = \beta_1 \sin(\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2 + \beta_3 x_4 + \beta_4 x_5$

$\vdots$

Model 9: $f(x_1,\ldots,x_{10}) = \beta_1 \sin(\pi x_1 x_2) + \beta_2 (x_3 - 0.5)^2 + \sum_{i=3}^{9} \beta_i x_{i+1}.$

Table 7 shows the results when the whole data set is used. All the populations follow normal distributions and ANOVA test is significant. All multiple comparison tests select the same groups of models and model 3 is selected.

6. Conclusions

We have assumed that the goal is to find a model having the best generalization performance. In doing this, we have been concerned primarily with the choice of

a subset of models not significantly different from the best rather than with the choice of a single model.

An alternative method has been proposed to model selection, where no distribution assumptions about the data are needed. Our goal has been to determine that, in a finite set of models, it is possible to find a subset, whose differences among error means are not significant with respect to the smallest.

In the design of the experiment for comparing several models, we have taken into consideration all the sources of variations that any statistical test should control. This goal is not achieved completely due to the finite size of available data. At least, we guarantee that different models are evaluated under the same circumstances. A calculation should be done for the number of observations that are needed in order to achieve the objectives of the experiment. If too few observations are taken, the experiment may be inconclusive. If too many are taken, then time, energy, and money may be needlessly expended. We recommend at least 30 error measures per model, in order to guarantee that the error samples will provide a good estimate of the distribution of errors.

After collecting data from a completely randomized design, error means are analyzed. It is well known that a battery of resampled t-tests should never be employed. The more tests we do, the more chance we have of falsely rejecting a null hypothesis and accepting a difference where one does not exist (Type I error). Hence, results obtained using these tests cannot be trusted.

Multiple comparison procedures (parametric and nonparametric) are statistical methods to compare three or more means while controlling the probability of making at least one Type I error. These tests are used only after a significant difference has been demonstrated. When this strategy is applied to a finite set of models, it is possible to find a subset of them whose differences among their error means are not large enough to indicate differences among the corresponding models. A wide range of multiple comparison procedures is commonly present in the literature. Fisher's LSD, Tukey's HSD, Bonferroni, Sidak, Scheffé, Dunnett and Hsu's RSMCB procedures have been discussed. The various procedures trade-off power for control of the experimentwise error rate in different ways. As a conclusion, we can say that there is no "correct" procedure to use.

The complete procedure has been shown to be useful in several model selection problems such as the determination of the optimal degree in polynomial fitting, the determination of the optimal number of hidden units in feedforward networks, the determination of the optimal number of kernels in radial basis function networks and the determination of the optimal K in the K-nearest neighbor algorithm.

The degree of the model complexity that is appropriate depends substantially on the sample size. In general, only simple models are stable when the sample size is small. As the sample size increases it become feasible to reliably estimate progressively finer details on the problem by using more complex models. If the size of sample data is not large enough with respect to the complexity of the problem, statistical tests may be inconclusive.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, AC-19(6) (1974) 716–723.
[2] C.M. Bishop, Neural Network for Pattern Recognition, Clarendon Press, Oxford, 1995.
[3] G.E. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters. An Introduction to Design, Data analysis, and Model building, Wiley, New York, 1978.
[4] G.W. Cobb, Introduction to Design and Analysis of Experiments, Springer, New York, 1998.
[5] A. Dean, D. Voss, Design and Analysis of Experiments, Springer Texts in Statistics, Springer, New York, 1999.
[6] A.C. Davidson, D.V. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press, Cambridge, 1997.
[7] T.G. Dietterich, Approximate statistical test for comparing supervised classification learning algorithms, Neural Comput. 7 (10) (1998) 1895–1923.
[8] C.W. Dunnett, New tables for multiple comparison with a control, Biometrics 20 (1964) 482.
[9] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, London, 1993.
[10] A. Feelders, W. Verkooijen, On the statistical comparison of inductive learning methods, Learning from data, Artificial Intelligence and Statistics V, Lecture Notes in Statistics (112), Springer, Berlin, 1996, pp. 271–279.
[11] J. Friedman, Multivariate adaptive regression splines, TR102, November 1988, Laboratory for Computational Statistics, Department of Statistics, Stanford University.
[12] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990.
[13] J.D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, 3rd Edition, Statistics: Textbooks and Monographs, Vol. 131, Marcel Dekker, New York, 1992.
[14] J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Multivariate data analysis with readings, 4th Edition, Prentice-Hall, Englewood Cliffs, NJ, 1995.
[15] Y. Hochberg, A.C. Tamhane, Multiple Comparison Procedures, Wiley, New York, 1987.
[16] J.C. Hsu, Multiple Comparisons: Theory and Methods, Chapman & Hall, London, 1996.
[17] J.D. Jobson, Applied Multivariate Data Analysis, Springer Texts in Statistics, Vol. 1, Springer, New York, 1991.
[18] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 4th Edition, Prentice-Hall, Englewood Cliffs, NJ, 1998.
[19] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, On-line document, http://robotics.stanford.edu/ ~ronnyk.
[20] T. Masters, Advanced Algorithms for Neural Networks, A C++ Source Book, Wiley, New York, 1995.
[21] T. Mitchell, Machine Learning, WCB/McGraw-Hill, New York, 1997.
[22] P.M. Murphy, D.W. Aha, UCI repository of machine learning databases, http://www.ics.uci.edu/ ~mlearn.
[23] J. Newmark, Statistics and Probability in Modern Life, in: W.B. Saunders (Ed.), 1988.
[24] E.S. Pearson, H.O. Hartley, Biometrika Tables for Statisticians, Vol. 1, 3rd Edition, Cambridge University Press, Cambridge, 1966.
[25] J. Pizarro, E. Guerrero, P.L. Galindo, A statistical model selection strategy applied to neural networks, Proceedings of the ninth European Symposium on Artificial Neural Networks, pp. 55–60, Bruges 2000.
[26] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, The Art of Scientific Computing, 2nd Edition, Cambridge University Press, Cambridge, 1992.
[27] H. Scheffé, Analysis of Variance, Wiley, New York, 1959.
[28] H. Scheffé, A method for judging all contrasts in the analysis of variance, Biometrika 40 (1953) 87.
[29] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[30] T. Sincich, Business Statistics by Example, 5th Edition, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[31] M. Stone, Cross-validatory choice and assessment of statistical prediction (with discussion), J. R. Statist. Soc. 1974 (B36) 111–147.
[32] The ELENA Project. Enhanced Learning for Evolutive Neural Architectures. Basic Research ESPRIT project Number 6891. (http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm)
[33] J.W. Tukey, Comparing individual means in the analysis of variance, Biometrics 5 (1949) 99.
[34] J.H. Zar, Biostatistical Analysis, 3rd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1996.
[35] W. Zucchini, An introduction to model selection, J. Math. Psychol. 44 (2000) 1.

Joaquín Pizarro Junquera, graduated from the University of Granada in Computer Science in 1991. Since 1998 he is an Assistant Professor at the Computer Science Department, Cadiz University. He is now a Ph.D. candidate in the department of Computer Science in the area of Pattern Recognition. Since 1998 he has been a research associate at Intelligent Systems Group. His research interests include model selection and neural networks.

Pedro L. Galindo Riaño received his B.Sc. and Ph.D. degrees in Computer Science from the Polytechnic University of Madrid, Spain, in 1989 and 1995, respectively. His Ph.D. was devoted to the application of unsupervised neural networks to speech recognition. He joined the Research Department of ENA Telecommunications company in 1990, where he participated in the R.O.A.R.S. (robust analytical speech recognition system) ESPRIT Project.

In 1992, he joined the University of Cadiz, Spain. At present, Dr. Galindo holds the position of Full Professor. Since 1998 he is the head of the Intelligent Systems Research Group. His main field of research interest is the application of statistical techniques to machine learning. His current activities are focused on the area of model selection.

Elisa Guerrero Vázquez received her B.Sc. degree in Computer Science from the University of Málaga, Spain. In 1996, she joined the Computer Science Department of the University of Cádiz, where she is currently working as a lecturer as well as writing her Ph.D. Thesis in Computer Science. She is a member of the Intelligent Systems Research Group at the University of Cádiz and her research interests include artificial neural networks, pattern recognition, model selection and statistics.