Repeated Measures Multiple Comparison Procedures Applied to Model Selection in Neural Networks

Elisa Guerrero Vázquez[1], Andrés Yañez Escolano[1], Pedro Galindo Riaño[1], Joaquín Pizarro Junquera[1],

[1] Universidad de Cádiz, Departamento de Lenguajes y Sistemas Informáticos, Grupo de Investigación "*Sistemas Inteligentes de Computación*"
C.A.S.E.M. 11510 - Puerto Real (Cádiz), Spain
{elisa.guerrero, andres.yaniez, pedro.galindo, joaquin.pizarro@uca.es}

Abstract.

One of the main research concern in neural networks is to find the appropriate network size in order to minimize the trade-off between overfitting and poor approximation. In this paper the choice among different competing models that fit to the same data set is faced when statistical methods for model comparison are applied. The study has been conducted to find a range of models that can work all the same as the cost of complexity varies. If they do not, then the generalization error estimates should be about the same among the set of models. If they do, then the estimates should be different and our job would consist on analyzing pairwise differences between the least generalization error estimate and each one of the range, in order to bound the set of models which might result in an equal performance. This method is illustrated applied to polynomial regression and RBF neural networks.

1 Introduction

In the model selection problem we must balance the complexity of a statistical model with its goodness of fit to the training data. This problem arises repeatedly in statistical estimation, machine learning, and scientific inquiry in general. Instances of model selection problem include choosing the best number of hidden nodes in a neural network, determining the right amount of pruning to be performed on a decision tree, and choosing the degree of a polynomial fit to a set of points. In each of these cases, the goal is not to minimize the error on the training data, but to minimize the resulting generalization error [8]. The model selection problem is coarsely prefigured by Occam's Razor: given two hypotheses that fit the data equally well, prefer the simpler one. For example, in the case of neural networks, oversized networks learn more rapidly since they easily monitor local minima, but exhibit poor generalization performance because of their tendency to fit data noise with the true response [11].

In a previous work [10] we proposed a resampling based multiple comparison technique where a randomized data collecting procedure was used to obtain

independent error measurements. However repeated measures experimental designs offer greater statistical power relative to sample size because the procedure takes into account the fact that we have two o more observations on the same subject, and uses this information to provide a more precise estimate of experimental error [9].

This paper focuses on the use of repeated measures tests, specifically repeated measures Anova and Friedman tests, that allow to reject the hypothesis null that the models are all equal, but they do not pinpoint where the significant differences lie.

Multiple comparison procedures (MCP) are used to find out which pair or pairs of models are significantly different from each other [6].

The outline of this paper is as follows. In section 2 we briefly introduce parametric and nonparametric statistical tests, including the multiple comparison procedures that could be applied for a repeated measures design. In section 3 our strategy using statistical tests for model selection is described and illustrated in the application of RBF networks and polynomial regression.

2 Repeated measures tests and post-hoc tests

Our objective is to consider a set of several models simultaneously, compare them and come to a decision on which to retain. Repeated measures tests allow to reject the hypothesis null that the models are all equal [2].

The application of the within-subjects ANOVA test implies the following assumptions to be satisfied: the scores in all conditions are normally distributed and each subject is sampled independently from each other subject, and, sphericity [5]. When these assumptions are violated, a non-parametric test should be used instead [3]. In our experiments we use Friedman test that is based on the median of the sample data [7].

If repeated measures Anova F test is significant and only pairwise comparisons should be made Girden [5] recommended the use of separated error terms for each comparison in order to protect the overall familywise probability of the type I error.

The approach for multiple comparisons proposed here consists on comparing only groups between which we expect differences to arise rather than comparing all pairs of treatment levels. The less tests we perform the less we have to correct the significance level, and the more power is retained. This approach is applied to the step-down sequentially rejective Bonferroni procedure [1] that is more powerful and less conservative that the singlestep Bonferroni correction.

Nemenyi test is a nonparametric multiple comparison method [12] that we will use when the repeated measures Anova assumptions are not met. This test uses rank sums instead of means.

3 Experimental Results

Our experiments can be outlined as follows:

1. Take the whole data set and create at least 30 new sets by bootstrapping [4].
2. Apply models with a degree of complexity ranging from *m* to *n* (in our experiments *m* = 1 and *n* = 25) and generate as many error measures per model as number of sets have been created in the previous step.
3. Calculate the error mean per model and select the model with minimum error mean.
4. Test repeated measures Anova assumptions
5. Apply statistical tests to check if we might reject the null hypothesis that all groups are equal:
   5.1. Repeated measures Anova test, if its assumptions are met.
   5.2. Friedman test if assumptions are not met.
6. If the null hypothesis is not rejected select the simplest model.
7. If the null hypothesis is rejected, apply a Multiple Comparison Procedure
   7.1. Sequentially rejective Bonferroni if repeated measures Anova was applied.
   7.2. Nemenyi test if Friedman test was applied.
8. Select the less complex model from the subset of models that are not significantly different from the model with minimum error mean.

In order to illustrate our strategy we conducted a range of experiments on simulated data sets. Thirty data sets Z=(X,Y) for several sample sizes were simulated according to the following experimental functions:

$$y = sin(x+2)^2 + x, \quad x \in (-2,+2) \tag{1}$$

$$y = -4.9389x^5 - 1.7917x^4 + 23.2778x^3 + 8.7917x^2 - 15.3380x - 6 + x \ , x \in (-1,3) \tag{2}$$

where x is gaussian noise.

3.1 Polynomial Fitting

We considered the problem of finding the degree N of a polynomial P(x) that better fits a set of data in a least squared sense. Polynomials with degrees ranging from 1 to 25 are used. The only aspect of the polynomials that remains to be specified is the degree.

Tables 1 and 2 show the results according to the experimental function (1) for each of the following sample sizes n = 25, 100, 500 and for each of the following N (0,0.5), N(0,1), N(0,3) and N(0,5) gaussian noise. In this case, all the experiments conducted showed that the least generalization error mean was reached with a polynomial with degree 5.

Tables 3, 4 and 5 show the results according to the experimental function (2) for each of the following sample sizes n = 250, 500, 1000 and for each of the following N (0,0.5), N(0,1), N(0,3) and N(0,5) gaussian noise.

Table 1. Polynomial Degrees (Models) with minimum test error mean, and percentage of times that occurred in our experiments.

		noise's variance			
	models (%)	0.5	1	3	5
data set size	25	**5 (93.33%)** 6 (6.67%)	4 (6.67%) **5 (90%)** 7 (3.33%)	4 (16.67%) **5 (80%)** 6 (3.33%)	4 (30%) **5 (70%)**
	50	**5 (86.67%)** 6 (10%) 7 (3.33%)	**5 (83.33%)** 6 (13.33%) 7 (3.33%)	**5 (93.33%)** 7 (3.33%) 8 (3.33%)	**5 (90%)** 6 (10%)
	100	**5 (83.33%)** 6 (3.33%) 7 (10%) 8 (3.33%)	**5 (83.33%)** 6 (13.33%) 8 (3.33%)	**5 (33.33%)** 6 (20%) 7 (3.33%) 8 (3.33%)	**5 (86.67%)** 6 (13.33%)
	500	**5 (86.67%)** 6 (6.67%)) 7 (3.33%) 10 (3.33%)	**5 (90%)** 6 (6.66%) 7 (3.33%)	**5 (96.67%)** 6 (3.33%)	**5 (80%)** 6 (6.66%) 7 (3.33%) 8 (3.33%) 9 (3.33%) 10 (3.33%)

Table 2. Selected models that are not significant different from the model with minimum test error mean (confidence level = 0.05).

		noise's variance			
	models (%)	0.5	1	3	5
data set size	25	2 (3.33%) 3 (13.33%) **4 (73.33%)** 5 (10%)	1 (10%) 3 (40%) **4 (50%)**	1 (13.33%) 2 (6.67%) **3 (60%)** 4 (20%)	1 (26.67%) 2 (23.33%) **3 (43.33%)** 4 (6.67%)
	50	4 (3.33%) **5 (96.67%)**	4 (16.67%) **5 (83.33%)**	4 (40%) **5 (60%)**	**4 (83.33%)** 5 (16.67%)
	100	**5 (100%)**	**5 (100%)**	**5 (100%)**	4 (3.33%) **5 (96.67%)**
	500	**5 (100%)**	**5 (100%)**	**5 (100%)**	**5 (100%)**

Table 3. Models with minimum generalization error mean.

		noise's variance				
	models (%)	0.1	0.25	0.5	0.75	1
data set size	250	14	14	13	12	11
	500	15	14	14	13	11
	1000	17	14	14	14	12

Table 4. Models with minimum test error mean and percentage of times that occurred in our experiments.

data set size	models (%)	noise's variance: 0.1	0.25	0.5	0.75	1
	250	12 (40%) 13 (6.67%) **14 (60%)** 15 (13.33%) 16 (10%)	11 (16.67%) 12 (20%) 13 (16.67%) **14 (36.67%)** 15 (6.67%) 16 (3.33%)	**11 (50%)** 12 (26.67%) 13 (16.67%) 14 (6.67%)	8 (3.33%) 9 (6.67%) 10 (6.67%) **11 (60%)** 12 (6.67%) 13 (10%) 14 (3.33%) 16 (3.33%)	8 (6.67%) 9 (16.67%) 10 (26.67%) **11 (33.33%)** 12 (13.33%) 15 (3.33%)
	500	14 (33.33%) **15 (36.67%)** 16 (13.33%) 17 (10%) 18 (3.33%) 20 (3.33%)	12 (3.33%) 13 (3.33%) **14 (63.33%)** 15 (10%) 16 (16.67%) 18 (3.33%)	11 (20%) **12 (36.67%)** 13 (13.33%) 14 (20%) 15 (3.33%) 16 (6.67%)	**11 (56.67%)** 12 (20%) 13 (6.67%) 14 (10%) 15 (3.33%) 16 (3.33%)	9 (3.33%) 10 (3.33%) **11 (53.33%)** 12 (30%) 13 (3.33%) 14 (3.33%) 15 (3.33%)
	1000	14 (16.67%) **15 (46%)** 16 (10%) 17 (16.67%) 18 (3.33%) 20 (3.33%) 21 (3.33%)	**14 (73.33%)** 15 (16.67%) 17 (3.33%) 18 (3.33%) 19 (3.33%)	11 (3.33%) 12 (13.33%) 13 (10%) **14 (63%)** 15 (6.67%) 16 (3.33%)	11 (10%) 12 (40%) 13 (10%) **14 (33.33%)** 15 (10%) 17 (3.33%) 20 (3.33%)	**11 (53.33%)** 12 (30%) 13 (6.67%) 14 (6.67%) 18 (3.33%)

Table 5. Selected models that are not significant different from the model with minimum test error mean (confidence level = 0.05).

data set size	models (%)	noise's variance: 0.1	0.25	0.5	0.75	1
	250	9 (6.67%) 10 (3.33%) 11 (36.67%) **12 (46.67%)** 13 (6.67%)	9 (6.67%) 10 (3.33%) **11 (90%)**	8 (46.67%) 9 (33.33%) **10 (60%)**	**8 (96.67%)** 9 (3.33%)	1 (10%) 3 (10%) 4 (3.33%) **8 (73.33%)** 9 (3.33%)
	500	**12 (53.33%)** 13 (46.67%)	**11 (100%)**	9 (20%) **10 (46.67%)** 11 (33.33%)	8 (40%) **9 (53.33%)** 10 (6.67%)	**8 (76.67%)** 9 (20%) 11 (3.33%)
	1000	**14 (100%)**	11 (3.33%) **12 (76.67%)** 13 (20%)	**11 (100%)**	9 (6.67%) 10 (16.67%) **11 (76.67%)**	8 (10%) **9 (60%)** 10 (26.67) 11 (3.33%)

3.2 Radial Basis Function Neural Networks

RBF neural network having one hidden layer for which the combination function is the Euclidean distance between the input vector and the weight vector. We use the exp activation function, so the activation of the unit is a Gaussian "bump" as a function of the inputs. The placement of the kernel functions has been accomplished using the k-means algorithm. The width of the basis functions has been set to

$$s = \frac{\left\| max(x_i - x_j) \right\|}{\sqrt{2 \cdot n}} \quad (3)$$

where n is the number of kernels.

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.

Tables 6, 7 and 8 show the results according to the experimental function (1) for each of the following sample sizes n = 50, 100, 500 and for each of the following N (0,0.5), N(0,3) and N(0,5) gaussian noise.

Table 6. Models with minimum generalization error mean.

		noise's variance		
	models	0.5	3	5
data set size	50	11	11	10
	100	13	12	11
	500	14	14	13

Table 7. Models with minimum test error mean and percentage of times that occurred in our experiments.

		noise's variance		
	models (%)	0.5	3	5
data set size	50	8 (16.67%) **12 (66.67%)** 13 (16.67%)	9 (20.33%) **10 (43.33%)** 11 (16.67%) 13 (16.67%)	**10 (86.67%)** 11 (13.33%)
	100	12 (30%) **13 (53.33%)** 15 (16.67%)	**10 (40%)** 11 (26.67%) 12 (16.67%) 13 (16.67%)	10 (20%) **11 (60%)** 12 (20%)
	500	14 (16.67%) **15 (83.33%)**	12 (36.67%) **13 (43.33%)** 15 (20%)	11 (23.33%) **12 (60%)** 13 (16.67%)

Table 8. Selected models that are not significant different from the model with minimum test error mean (confidence level = 0.05).

		noise's variance		
	models (%)	0.5	3	5
data set size	50	6 (16.67%) 8 (20%) **9 (63.33%)**	7 (36.37%) **8 (63.33%)**	**7 (80%)** 8 (20%)
	100	**9 (83.33%)** 10 (16.67%)	**8 (100%)**	7 (23.33%) **8 (76.67%)**
	500	**10 (73.33%)** 11 (26.67%)	**9 (76.67%)** 10 (23.33%)	**9 (100%)**

Tables 9, 10 and 11 show the results according to the experimental function (2) for each of the following sample sizes n = 250, 500, 1000 and for each of the following N (0,0.25), N(0,0.5) and N(0,1) gaussian noise.

Table 9. Models with minimum generalization error mean.

		noise's variance		
	models (%)	0.25	0.5	1
data set size	250	14	13	12
	500	14	12	12
	1000	17	15	13

Table 10. Models with minimum test error mean and percentage of times that occurred in our experiments.

		noise's variance		
	models (%)	0.25	0.5	1
data set size	250	12 (6.67%) 13 (33.33%) **14 (40%)** 15 (20%)	11 (33.33%) 12 (6.67%) **13 (40%)** 15 (20%)	10 (16.33%) 11 (23.33%) **12 (30.67%)** 13 (16.33%) 14 (13.33%)
	500	13 (26.67%) **14 (40%)** 15 (33.33%)	**12 (43.33%)** 14 (36.67%) 15 (20%)	**11 (33.33%)** 12 (26.67%) 13 (23.33%) 15 (16.67%)
	1000	**14 (40%)** 15 (20%) 16 (13.33%) 17 (13.33%) 18 (13.33%)	12 (13.33%) 13 (10%) **14 (63.33%)** 15 (10%) 16 (3.3%)	11 (20%) 12 (3,33%) **13 (60%)** 14 (16.67%)

Table 11. Selected models that are not significant different from the model with minimum test error mean (confidence level = 0.05).

		noise's variance		
	models (%)	0.25	0.5	1
data set size	250	10 (13.33%) 11 (20%) **12 (46.67%)** 13 (20%)	9 (40%) **10 (60%)**	5 (20%) **7 (46.67%)** 9 (23.33%) 10 (10%)
	500	**12 (46.67%)** 13 (53.33%)	**10 (36.67%)** 11 (63.33%)	**9 (76.67%)** 10 (23.33%)
	1000	**13 (100%)**	**11 (100%)**	9 (20%) **10 (76.67%)** 11 (3.33%)

In all the experiments, when the data set was large enough, the parametric tests were applied. Otherwise the repeated measures Anova assumptions were not satisfied and then, nonparametric test were used instead.

A systematic underfitting is observed in the method proposed. This underfitting is produced by: the use of bootstrapping techniques in the simulated data sets, and the strategy followed by our method. However, we can guarantee with high statistical reliability that the performance of the selected model is as good as the least test error mean.

4 Conclusions

In this work we have presented a model selection criterion that consists on finding the group of models that are not significant different from the model with the minimum test error mean, in order to select the model with less complexity (Occam's Razor).

The experimental results show that this criterion produces underfitting when the data set size is small but it works very well when the data set size is large enough, thus our method improves the widely used criterion of selecting the model with the least test error mean. To avoid underfitting problems another selection criterion could be applied at the cost of a larger complexity.

5 References

1. Chen, T., Seneta, E.: A stepwise rejective test procedure with strong control of familywise error rate. University of Sidney, School of Mathematics and Statistics, Research Report 99-9, March (1999)
2. Dean A., Voss, D.: Design and Analysis of Experiments. Springer-Verlag New York (1999)
3. Don Lehmkuhl, L: Nonparametric Statistics: Methods for Analyzing Data Not Meeting Assumptions Required for the Application of Parametric Tests, Journal of Prosthetics and Orthotics, 3 (8) 105-113 (1996)
4. Efron, B., Tibshirani, R.: Introduction to the Bootstrap, Chapman & Hall, (1993)
5. Girden, E. R.: Anova Repeated Measures, Sage Publications (1992)
6. Hochberg, Y., Tamhane A. C.: Multiple Comparison Procedures, Wiley (1987)
7. Hollander, M., Wolfe, D. A.: Nonparametric Statistical Methods, Wiley (1999)
8. Kearns, M., Mansour, Y.: An experimental and theorical comparison of model selection methods. Machine Learning, 27(1), (1997)
9. Minke, A.: Conducting Repeated Measures Analyses: Experimental Design Considerations, Annual Meeting of the Southwest Educational Research Association, Austin, (1997)
10. Pizarro, J., Guerrero, E., Galindo, P.: A statistical model selection strategy applied to neural networks. Proceedings of the European Symposium on Artificial Neural Networks Vol 1, pp. 55-60, Bruges (2000)
11. Vila, J.P., Wagner, V., Neveu, P.:Bayesian nonlinear model selection and neural networks: a conjugate prior approach. IEEE Transactions on neural networks, vol 11,2, march (2000)
12. Zar, J. H.: Biostatistical Analysis, Prentice Hall (1996)