Empirical Performance Assessment of Nonlinear Model Selection Techniques

Elisa Guerrero Vázquez, Joaquín Pizarro Junquera, Andrés Yáñez Escolano, and Pedro L. Riaño Galindo

Grupo Sistemas Inteligentes de Computación
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Cádiz
11510 Puerto Real, Spain
{elisa.guerrero,joaquin.pizarro,pedro.galindo,
andres.yaniez }@uca.es
http://www2.uca.es/grup-invest/sic/

Abstract. Estimating Prediction Risk is important for providing a way of computing the expected error for predictions made by a model, but it is also an important tool for model selection. This paper addresses an empirical comparison of model selection techniques based on the Prediction Risk estimation, with particular reference to the structure of nonlinear regularized neural networks. To measure the performance of the different model selection criteria a large-scale small-samples simulation is conducted for feedforward neural networks.

1 Introduction

The choice of a suitable model is very important to balance the complexity of the model with its fit to the data. This is especially critical when the number of data samples available is not very large and/or is corrupted by noise. Model selection algorithms attempt to solve this problem by selecting candidate functions from different function sets with varying complexity, and specifying a fitness criterion, which measures in some way the lack of fit. Then, the class of functions that will likely optimize the fitness criterion is selected from that pool of candidates.

In regression models, when the fitness criterion is the sum of the squared differences between future observations and models forecasts, it is called Prediction Risk. While estimating Prediction Risk is important for providing a way of estimating the expected error for predictions made by a model, it is also an important tool for model selection [11].

Despite the huge amount of network theory and the importance of neural networks in applied work, there is still little published work about the assessment on which model selection method works best for nonlinear learning systems. The aim of this paper is to present a comparative study of different model selection techniques based on the Minimum Prediction Risk principle in regularized neural networks.

F.J. Garijo, J.C. Riquelme, and M. Toro (Eds.): IBERAMIA 2002, LNAI 2527, pp. 450–459, 2002.

Section 2 studies the Generalized Prediction Error for nonlinear systems introduced by Moody [7] which is based upon the notion of the effective number of parameters. Since it cannot be directly calculated, algebraic or resampling estimates are reviewed taking into account regularization terms in order to control the appearance of several local minima when training with nonlinear neural networks.

Results varying the number of hidden units, the training set size and the function complexity are presented in the Simulation results section. Conclusions follow up.

2 Model Selection Techniques

The appearance of several local minima in nonlinear systems suggests the use of regularization techniques, such as weight decay, in order to reduce the variability of the fit, at the cost of bias, since the fitted curve will be smoother than the true curve [9]. Regularization adds a penalty Ω to the error function ε to give:

$$\hat{\varepsilon} = \varepsilon + \lambda\Omega \tag{1}$$

where the decay constant λ controls the extent to which the penalty term Ω influences the form of the solution.

In particular, weight decay consists of the sum of the squares of the adaptive parameters in the network where the sum runs over all weights and biases:

$$\Omega = \frac{1}{2}\sum_i w_i^2 \tag{2}$$

It has been found empirically that a regularizer of this form can lead to significant improvements in network generalization.[1]

Prediction Risk measures how well a model predicts the response value of a future observation. It can be estimated either by using resampling methods or algebraically, by using the asymptotic properties of the model.

Algebraic estimates are based on the idea that the resubstitution error ε_{Res} is a biased estimate of the Prediction Risk ε_{PR}, thus the following equality can be stated:

$$\varepsilon_{PR} = \varepsilon_{Res} + \text{Penalty\_Term} \tag{3}$$

where the penalty-term represents a term which grows with the number of free parameters in the model. Thus, if the model is too simple it will give a large value for the criterion because the residual training error is large, while a model which is too complex will have a large value for the criterion because the complexity term is large. The minimum value for the criterion represents a trade-off between bias and variance.

According to this statement different model selection criteria have appeared in the statistics literature for linear models and unbiased nonlinear models, such as Mallow's CP estimate, the Generalized Cross-Validation (GCV) formula, Akaike's Final Prediction Error (FPE) and Akaike's Information Criteria (AIC) [5], etc. For general nonlinear learning systems which may be biased and may include weight decay or

other regularizers Moody [7] was the first to introduce an estimate of Prediction Risk, the Generalized Prediction Error (GPE), which for a data sample of size *n* can be expressed as:

$$GPE(\lambda) = \mathcal{E}_{\mathrm{Res}} + 2\hat{\sigma}^2 \frac{\hat{p}_{eff}(\lambda)}{n} \tag{4}$$

where $\hat{\sigma}^2$ is an estimate of the noise variance on the data and the regularization parameter λ controls the effective number of parameters peff(λ) of the solution. As suggested in [6] it is not possible to define a single quantity which expresses the effective number of weights in the model. peff(λ) usually differs from the true number of model parameters *p* and depends upon the amount of model bias, model nonlinearity, and our prior model preferences as determined by λ and the form of the regularizer. See [6] for a detailed determination of peff(λ) and $\hat{\sigma}^2$.

The effective number of parameters can then be used in a generalization of the AIC for the case of additive noise, denoted by Murata as NIC (Network Information Criterion) [8]. The underlying idea of NIC is to estimate the deviance for a data set of size *n*, compensating for the fact that the weights were chosen to fit the training set:

$$NIC = n * \log(\mathcal{E}_{\mathrm{Res}}) + 2 * \hat{p}_{eff}(\lambda) \tag{5}$$

Alternatively, data resampling methods, such as k-fold Cross-validation (kCV) or bootstrap estimation make maximally efficient use of available data, but they can be very CPU time consuming for neural networks. A nonlinear refinement of CV is called 10NCV [7].

In both, kCV and kNCV, the dataset is randomly split into k mutually exclusive folds or subsets of approximately equal size. The training process is repeated k times, each time leaving out one of the k subsets to test, but kNCV uses as starting point weights of a network trained on all available data rather than random initial weights for retraining on the k subsets.

We consider that models which minimize GPE, NIC, kCV and kNCV are optimal in the average loss sense. We can use these criteria to select a particular model from a set of possible models.

3 Simulation Results

This paper focuses on feedforward neural networks with a single layer of units with hyperbolic tangent activation functions. Architectures considered are limited to single hidden layer networks because of their proven universal approximation capabilities and to avoid further increasing complexity.

The networks were trained by ordinary least-squares using standard numerical optimisation algorithms for H hidden units ranging from 1 to M. The training algorithm was Levenberg-Marquardt. For a network with H hidden units, the weights for the previously trained network were used to initialise H-1 of the hidden units,

while the weights for the H^{th} hidden unit were generated from a pseudorandom normal distribution. The decay constant λ was fixed to 0.002.

All simulations were performed 1000 times, each time generating a new different data set of size N. Model selection results were averaged to reduce the influence of model variability on network size selection by introducing the possibility of escaping local minima.

We used artificially generated data from the following target functions:

$$y = 1.8*\tanh(3.2*x + 0.8) - 2.5*\tanh(2.1*x + 1.2) - 0.2*\tanh(0.1*x - 0.5) + \xi \quad (6)$$

$$y = -5*x^5 - 1.8*x^4 + 23.27*x^3 + 8.79*x^2 - 15.33*x - 6 + \xi \quad (7)$$

where $x \in [-2,2]$ and ξ is a Gaussian zero mean, i.i.d. sequence which is independent of the input with variance $\sigma=0.5$.

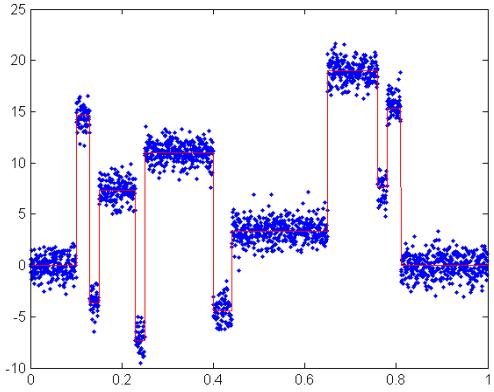

Fig. 1. Low-noise and noise-free block functions from Donojo-Johnstone benchmarks

Alternatively, in order to study a case of higher nonlinearity we considered the low-noise block function from the Donoho-Jonstone benchmarks (fig. 1).These benchmarks have one input, high nonlinearity and random noise can be added to produce an infinite number of data sets. Sarle [10] checked that the MLP easily learned the block function at all noise levels with 11 hidden units and there was overfitting with 12 or more hidden units when training with 2048 samples.

We assume that among the candidate models there exists model M_c that is closest to the true model in terms of the expected Prediction Risk, $E[PR](M_c)$. Suppose a model selection criterion selects model M_k which has an expected Prediction Risk of $E[PR](M_k)$. Observed efficiency is defined as the ratio that compares the Prediction Risk between the closest candidate model, M_c, and the model selected by some criterion M_k.

$$observed \quad efficiency = \frac{PR(M_c)}{PR(M_k)} \tag{8}$$

Tables from 1 to 9 show observed efficiency for different target functions when the numbers of training examples are 25, 50 and 100.

First column shows the number of hidden units, ranging from 1 to 10 hidden units for hyperbolic tangent target function (6) and for the 5^{th} degree target function (7), and models ranging from 1 to 20 for low-noise block target function. For each of the 1000 realizations the criteria select a model and the observed efficiency of this selection is recorded, where higher observed efficiency denotes better performance.

Next columns show the counts for the different model selection criteria: NIC, 10NCV, 10CV, GPE and the Prediction Risk (PR) computed over a sample size of 2000. These results are one way to measure consistency, and we might therefore expect the consistent model selection criteria to have the highest counts. Last two rows show the mean observed efficiency and the rank for each criterion. The criterion with the highest averaged observed efficiency is given rank 1 (better) while the criterion with the lowest observed efficiency is given rank 4 (lowest of the 4 criteria considered).

Table 1. Simulation results for a data sample size of N=25 and target function (6)

Models	NIC	10NCV	10CV	GPE	PR
1	1	10	7	2	4
2	431	646	653	567	790
3	158	180	155	156	137
4	87	63	85	83	28
5	60	23	24	43	13
6	39	11	8	27	7
7	39	11	10	19	5
8	25	8	12	15	6
9	29	9	11	19	3
10	131	39	35	69	7
Efficiency	0.8080	0.9030	0.9090	0.8480	1.0
Rank	4	2	1	3	

Table 2. Simulation results for N=50 and target function (6)

Models	NIC	10NCV	10CV	GPE	PR
1	0	0	0	0	0
2	529	713	741	631	851
3	164	161	132	149	120
4	86	54	48	65	11
5	55	29	24	45	6
6	47	13	13	31	3
7	22	3	7	13	2
8	15	4	8	10	1
9	22	8	10	17	1
10	60	15	17	39	5
Efficiency	0.8891	0.9521	0.9520	0.9125	1.0
Rank	4	1	2	3	

Table 3. Simulation results for N=100 for target function (6)

Models	NIC	10NCV	10CV	GPE	PR
1	0	0	0	0	0
2	607	702	692	668	873
3	146	165	168	136	106
4	74	66	64	66	12
5	72	33	28	58	4
6	33	14	12	26	2
7	23	2	7	15	0
8	12	5	8	7	0
9	5	1	9	5	0
10	28	12	12	19	3
Efficiency	0.9438	0.9743	0.9716	0.9529	1.0
Rank	4	1	2	3	

Table 4. Simulation results for N=25 and target function (7)

Models	NIC	10NCV	10CV	GPE	PR
1	0	56	60	1	44
2	1	96	77	7	29
3	18	191	149	62	89
4	41	222	185	148	233
5	60	142	151	144	240
6	63	92	125	125	159
7	86	71	85	112	75
8	108	51	67	98	51
9	178	35	55	127	37
10	445	44	46	176	43
Efficiency	0.6820	0.7782	0.7573	0.7370	1.0
Rank	4	1	2	3	

Table 5. Simulation results for N=50 and target function (7)

Mod	NIC	10NCV	10CV	GPE	PR
1	0	0	0	0	1
2	0	4	2	1	3
3	13	75	32	18	16
4	101	264	219	165	239
5	138	248	278	221	351
6	119	161	185	158	199
7	111	103	113	118	103
8	133	53	78	102	46
9	147	53	57	105	20
10	238	39	36	112	22
Efficiency	0.7866	0.8783	0.8558	0.8272	1.0
Rank	4	1	2	3	

Tables 1, 2 and 3 show that for experimental function (6) all methods select models with 2 and 3 hidden units, 10NCV and 10CV perform almost the same, but both are

superior to GPE and NIC. In all the experiments NIC averaged observed efficiency has the last position on the ranking.

Table 6. Simulation results for N=100 and target function (7). NCV, CV and GPE favor models from 4 to 8 hidden units while NIC favors more overfitted models

Models	NIC	10NCV	10CV	GPE	PR
1	0	0	0	0	0
2	0	0	0	0	0
3	1	7	1	1	1
4	152	251	113	184	203
5	207	268	329	270	408
6	191	193	223	199	213
7	126	117	132	117	105
8	109	81	96	91	41
9	113	52	56	76	17
10	101	31	50	62	12
Efficiency	0.9114	0.9491	0.9190	0.9250	1.0
Rank	4	1	3	2	

Table 7. Simulation results for N=25 and low-noise block target function, when the sample size is very small model selection tasks are more difficult, in this case NIC shows a very high variance on the observed efficiency

Models	NIC	10NCV	10CV	GPE	PR
1	0	76	69	0	20
2	2	177	146	2	114
3	0	236	235	4	204
4	12	145	158	31	165
5	26	112	96	44	150
6	73	82	83	120	87
7	79	54	45	126	75
8	103	44	43	146	49
9	96	17	38	121	40
10	61	12	12	99	28
11	88	8	10	50	18
12	75	1	9	59	4
13	48	3	6	17	8
14	33	2	5	19	6
15	31	5	4	20	2
16	26	6	6	17	3
17	26	1	7	15	2
18	27	2	6	14	3
19	37	6	5	16	6
20	157	11	17	80	16
Efficiency	0.7251	0.8046	0.8233	0.7319	1.0
Rank	4	2	1	3	

Tables 4, 5 and 6 show that for experimental function (7) observed efficiency increases as the sample size grows. 10NCV is the most underfitting method for a sample size of 25, while NIC and GPE favor overfitted models.

In contrast to the previous results, we next considered a problem that has a much higher nonlinearity, the low-noise block function. Tables 7, 8 and 9 show that NIC outperforms 10NCV and 10CV when the sample size is 100 while with N=50 all methods perform almost the same. The averaged observed efficiency always grows as the simple size increases.

Table 8. Simulation results for N=50 and low-noise block target function. All criteria show a similar averaged observed efficiency, but 10NCV and 10CV tend to more underfitted models than NIC and GPE

Models	NIC	10NCV	10CV	GPE	PR
1	0	3	2	0	0
2	0	28	18	0	9
3	0	71	89	0	22
4	0	173	138	0	65
5	8	188	187	8	120
6	8	111	97	12	112
7	30	81	71	44	96
8	44	66	85	60	102
9	75	83	70	106	91
10	96	41	64	125	72
11	135	36	52	126	98
12	97	33	47	116	56
13	99	30	14	100	30
14	88	19	10	69	34
15	60	3	13	52	18
16	49	5	4	44	20
17	33	6	5	24	11
18	35	6	10	23	11
19	30	9	7	24	10
20	113	8	17	67	23
Efficiency	0.8269	0.8208	0.8358	0.8322	1.0
Rank	2	1	4	3	

From all the experimental results we can conclude that the performance differences are not great between 10NCV and 10CV, but 10NCV seems to perform better in almost all the sample sizes. 10CV is more computationally demanding than 10NCV. This fact leads us to prefer 10NCV rather than 10CV.

In general, there is not best model selection method. Depending on the particular problem one technique can outperforms another. When N is large, all methods give reasonable efficiency results but crossvalidation-based criteria seem to be slightly better. However, when it comes to the case where N=50 and 100 and high nonlinearity is present, NIC and GPE outperform 10NCV and 10CV. The algebraic estimate of Prediction Risk is also more attractive from the computational perspective. However, it is important to note that the theory of NIC relies on a single well-defined minimum to the fitting function, and it can be unreliable when there are several local minima [8]. Among the different cases presented in this paper GPE shows a more reliable behavior with not great differences between the best technique and GPE.

Table 9. Simulation results for N=100 and low-noise block target function. GPE and NIC show a higher averaged observed efficiency, and favor models from 11 to 16 hidden units, while 10CV and 10NCV models ranging between 9 and 14 hidden units

Models	NIC	10NCV	10CV	GPE	PR
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	12	17	0	1
5	0	30	42	0	5
6	0	75	55	2	8
7	6	60	42	4	33
8	14	66	36	14	25
9	23	90	68	27	45
1	42	92	91	58	96
11	114	112	120	140	97
12	118	72	87	138	99
13	114	84	78	105	97
14	127	79	87	137	135
15	103	66	75	106	102
16	98	44	49	93	77
17	65	30	27	56	68
18	62	16	47	52	27
19	41	29	20	25	27
20	73	43	59	43	58
Efficiency	0.9326	0.8484	0.8581	0.9319	1.0
Rank	1	3	4	2	

Conclusions

The performance of different model selection techniques based on the Prediction Risk estimation in nonlinear regularized neural networks has been studied. We determined relative performance by comparing GPE, NIC, 10NCV and 10CV against each other under different simulated conditions. Which is the best among these competing techniques for model selection is not clear. They can behave quite differently in small sample sizes and directly depend on the nonlinearity of the task.

The similar performance between 10CV and 10NCV lead us to prefer 10NCV since the computational cost is lower. NIC favors overfitted models when low nonlinearity is present while 10NCV favors underfitted models, even in high nonlinearity cases. Although the observed efficiency of GPE is not always the best, it gives reliable results for all the cases and, as well as 10NCV, it provides good estimates of the prediction risk at a lower computational cost.

Acknowledgements. This work has been supported by the Junta de Andalucía (PAI research group TIC-145).

References

1. Bishop C. M.: Neural networks for pattern recognition. Clarendon Press, Oxford (1995)
2. Brake, G., Kok J.N. Vitányi P.M.B.: Model Selection for Neural Networks: Comparing MDL and NIC. In: Proc. European Symposium on Artificial Neural Networks, Brussels, April 20–22 (1994)
3. Larsen J., Hansen L.K.: Generalization performance of regularized neural network models. Proc. IEEE Workshop: Neural Networks for Signal Processing IV, Piscataway, New Jersey (1994) 42–51
4. Lawrence S., Giles C.L., Tsoi A.C.: What Size of Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation. Technical Report. UMIACS-TR-96-22 and CS-TR-3617. Institute of Advanced Computer Studies. University of Mariland. (1996)
5. McQuarrie A., Tsai C.: Regression and Time Series Model Selection. World Scientific Publishing Co. Pte. Ltd. (1998)
6. Moody, J.: The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. NIPS (1992) 847–854
7. Moody, J.: Prediction Risk and Architecture Selection for Neural Networks. In Cherkassky, V., Friedman, J. H., and Wechsler, H., editors, From Statistics to Neural Networks: Theory and Pattern Recognition Applications, NATO ASI Series F. Springer-Verlag (1994)
8. Murata N., Yoshizawa S., Amari S.: Network Information Criterion – Determining the Number of Hidden Units for an Artificial Neural Network Model. IEEE Transactions on Neural Networks (1994) 5, 865–872
9. Ripley B.D. Statistical Ideas for Selecting Network Architectures. Neural Networks: Artificial Intelligence & Industrial Applications, eds. B. Kappend and S. Gielen. Springer, Berlin (1995) 183–190
10. Sarle W.: Donojo-Jonhstone benchmarks: neural nets results (1999) ftp: //ftp.sas.com/ pub/neural/dojo/dojo.html
11. Zapranis A., Refenes A.: Principles of Neural Model Identification, Selection and Adequacy: with applications to financial economics. (Perspectives in neural computing). Springer-Verlag London (1999)