

# Analysis of Performance Evaluation Metrics to Combat the Model Selection Problem

Ranganath Kothamasu, Samuel.H.Huang  
Intelligent CAM Systems Laboratory  
University of Cincinnati, Cincinnati, OH 45221

**William.H.VerDuin**  
VerTech LLC, Chagrin Falls, OH 44022

## ABSTRACT

Classification of data has been researched in different arenas of machine learning and statistical approximation techniques. Various techniques ranging from neural networks, clustering, discriminant functions to regression models have been employed for this purpose. Since no technique can produce the best model in all applications, it is essential to evaluate the individual models. This is often referred to as the model selection problem. Akin to the function approximation scenario, several approaches have been proposed and in this paper we evaluate their efficiency. However, they are not without drawbacks as can be seen from our experiments. Our focus in this study is to evaluate the performance metrics with respect to their efficiency in optimizing accuracy with complexity. The application of Akaike Information Criteria (AIC) to classification problems has also been explored in this paper.

**Keywords:** *AIC, multiple comparison procedures, performance evaluation*

## 1. INTRODUCTION

Computational Intelligence and statistical algorithms are increasingly applied to solve problems with no or difficult analytical solutions (first principle models). Their applications range from process optimization and control, image processing, medical diagnosis, gene mapping to natural language processing etc. The reason for their vast application domain is primarily their adaptability and universal applicability.

These algorithms, especially those belonging to the soft computing arena, are highly adaptable and hence this often leads to the notion of easy usability. This might be true in ideal situations, although in practice any modeling using soft computing or statistical algorithms is usually a laborious process. The modeling procedure is a

cycle that comprises of three broad stages – *data preprocessing, learning and validation*. The complexity of using these methodologies lies in the absence of sound governing principles (within these stages) that would theoretically point to the next modeling step. For instance, *feature extraction* belonging to the preprocessing stage is indispensable in most applications, yet there are no universally applicable features nor is there a definitive methodology to identify the ideal set of features for any domain (from the data alone). Such complexities result in modeling being an intensively iterative process.

Additional complexity is induced by the fact that these algorithms yield highly incomparable solutions – in form, complexity etc. Since a multitude of these algorithms can be applied to any given problem it becomes imperative to select the “best” among them. This problem is referred to as the *model selection* problem and essentially it aims at selecting the model that has achieved a good compromise between its *complexity* and *precision*. It is also defined in terms of the compromise between the modeling *bias* and *variance* (Duda et al, 2001).

Any supervised learning task is a derivative of the generalized learning problem which consists of estimating the function  $f$  defined as  $y = f(x, w)$ , where  $y$  are the required outputs and  $x$  are the inputs. The learning problem is defined as approximating  $f$  in the sense of minimizing the risk functional defined below (Vapnik, 1999).

$R(w) = \int L(y, f(x, w))p(x, y) dx dy$ , where  $p(x, y)$  is the joint probability distribution and  $L$  is a loss function and  $w$  is the set of estimated parameters. According to the above notation, if  $y$  is continuous it is a function approximation task and if it is categorical the problem is a classification task. The loss function “L” varies with inductive principle and at times with the learning task. Model selection based only on the empirical risk is not a desirable approach as it does not take into consideration the intrinsic dimensionality of the model and its generalization ability.

Section 2 gives a summary of various evaluation methods that assist in model selection. Section 3 examines the problem in the function approximation arena and section 4 in the classification arena.

## 2. MODEL EVALUATION

As mentioned previously, model evaluation is done with the goal of selecting the best available model for the given dataset. Several criteria, tests and methods have been proposed in the literature. Some of them are summarized below.

Traditionally criteria like SSE, MSE, MAP,  $R^2$ ,  $R^2_{(adj)}$ , PRESS are used to validate a model. The usual approach is to split the available data into *learning* and *validation* sets (Cherkassky and Mulier, 1998). The algorithms are supplied with the learning sets to create the model and are later validated with the above-mentioned criteria on the validation dataset. The definition of the above parameters is given in table 1.

Table 1. Definition of the traditional criteria used in model evaluation

Criteria	Definition
SSE (Sum of Squared Error)	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$
MSE (Mean Squared Error)	$\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
MAP (Mean Absolute Percent error)	$\left(\frac{100}{n}\right) * \sum_{i=1}^n  (y_i - \hat{y}_i) / y_i $
$R^2$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
$R^2_{(adj)}$	$R^2_{(adj)} = 1 - \left( \frac{SSE/n-k}{SST/n-1} \right)$

$y_i$  - Actual Outputs

$\hat{y}_i$  - Predicted Outputs

$n$  -# of Patterns

$k$  -# of parameters

The quality of the model is in inverse proportion to the magnitude of the first three criteria and also to the deviation of the last two criteria from “1”. However it has to be mentioned that these criteria are not appropriate for model selection in all situations. For instance,  $R^2$  should not be used for comparison of modeling algorithms that do not satisfy the criteria that  $\sum e_i = 0$  and  $\sum y_i e_i = 0$  where  $e_i$  are the corresponding residuals. This is especially true in the case of neural networks where situations with  $R^2$  greater than 1 are often encountered. Besides the restrictions mentioned above these criteria do not explicitly take into account the underlying dimensionality of the model (except  $R^2_{(adj)}$ ) and the complexity of data into account. Some criteria specific to neural networks were explored by Ripley (1995).

Statistical tests classified under Multiple Comparison Procedures (MCP) are another category of model evaluation techniques that are often used to compare a set of possible models to the given data. Such tests include McNemar’s test, a test for difference of error proportions, resampled paired t test, k-fold cross validated paired t test and 5x2cv paired t test. (Dietrich, 1997).

The basic concept of these tests is to check for significant difference in error (or its proportions) from the various models developed. Since the usual practice is to check for this difference among the error vectors from the same dataset, care must be taken to compensate for correlation. Secondly, the *multiplicity effect* that arises out of simultaneous pair wise comparisons between the models should also be taken into consideration (because of increased chances of Type I error).

The Hochberg and Tamhane test is appropriate for the function approximation problems and it is based on *studentized maximum modulus* distribution. Dunn (1961) proposed a test based on *studentized t distribution* that can reveal any significant differences between error proportions (well suited for classification problems). An excellent summary of both these tests is given in Feelders and Verkooijen (1996). These two tests are part of the focus in this study.

A third evaluation strategy is to construct a *penalization* form of criteria that enhances the empirical risk with a term that disfavors complex models (Domingos,1999). There are several penalization forms and AIC (Akaike Information Criterion) as defined below is one of them (Ishikawa, 1996).

$$AIC = nl \log(\hat{\sigma}^2) + 2k$$

where,  $k$  is the number of independent estimated parameter,  $l$  is the number of output units and  $\hat{\sigma}^2$  is the maximum likelihood estimate of the mean square error.

### 3. FUNCTION APPROXIMATION

In this section we compare the performance of different evaluation criteria in the domain of function approximation. We use an approach similar to the one proposed by Lawrence et al (1997) where a randomly initialized *teacher network* is used to extract the training and testing data. Networks of varying complexities called *student networks* are then trained on the learning dataset and validated with the testing data. In our case study we chose a neural network consisting of 3 neurons as the teacher network. This network is used to extract the necessary data that consists of standard normal random inputs (2-dimensional) and the respective outputs. The data was split into learning and validation sets comprising of 140 and 60 patterns respectively.

Networks of varying size are trained with the learning data (using the conjugate gradient descent algorithm) for 500 epochs. Various evaluation criteria along with the Hochberg and Tamhane confidence intervals (at 95% level) are given in Tables 2 and 3 respectively.

Table 2. Values of various evaluation criteria for the function approximation problem

Hidden Neurons	MSE	R <sup>2</sup>	R <sup>2</sup> <sub>(adj)</sub>	AIC
2	0.006471215	0.417	0.3256	-285.4319
3	0.000054294	0.7061	0.631	-564.2748
5	0.000040264	0.8028	0.7017	-566.2111
10	0.000220444	0.8472	0.5254	-424.2004
15	0.000046811	0.8769	8.2639	-477.1716
20	0.004963404	0.8229	1.4976	-157.3482

Table 3. Confidence intervals from pairwise Hochberg and Tamhane test

Model						
	1	2	3	4	5	6
1	-	[-2.777 2.790]	[-2.777 2.790]	[-2.777 2.789]	[-2.777 2.790]	[-2.781 2.784]
2	-	-	[-2.780 2.780]	[-2.780 2.779]	[-2.780 2.780]	[-2.787 2.778]
3	-	-	-	[-2.780 2.779]	[-2.780 2.780]	[-2.787 2.777]
4	-	-	-	-	[-2.780 2.780]	[-2.787 2.778]
5	-	-	-	-	-	[-2.787 2.778]

From Table 2 we cannot conclusively select a model because of varying indications from the different criteria, although AIC points to the model with 5 hidden neurons which is the closest to the original model (3 hidden neurons). From Table 3 it is evident that the Hochberg & Tamhane test concludes that all models are equally good.

Though AIC was close to the original model (3 neurons), it cannot be concluded that it did in fact select a model that best fits the data and it is also not valid to assume that a 3 or close to 3 hidden neuron network is a good fit for the finite data. (This validates the theory that for finite data, the best fit is not necessarily a model identical to the true parametric form [Cherkassky & Mulier,1998]. Though, it is interesting that the best fit here is in fact of a higher complexity than the true parametric form.)

To confirm that the 5 hidden neuron model is in fact superior, a generalization test was performed where noisy inputs were presented to the networks. The (additive) noisy inputs were generated as  $I(i) = I(i) + wgn(i, dBW)$  where “ $I$ ” is the original input value and “ $wgn$ ” is white gaussian noise with power specified by “ $dBW$ ”. The result from the test is given in Table 4. It can be seen from the table and Figure 1 that the network with 5 neurons has the best overall predictions at various noise levels.

Table 4. MSE values of the models at different noise levels

Neurons \ Noise (dBW)	2	3	5	10	15	20
1	0.0112	0.07	0.0456	0.049	0.0611	0.0367
5	0.0365	0.1191	0.0673	0.095	0.1177	0.0429
10	0.0871	0.2113	0.099	0.2325	0.2257	0.1023
15	0.1864	0.2735	0.1332	0.3805	0.3768	0.1684
20	0.2669	0.4267	0.2189	0.4318	0.8173	0.197
25	0.2106	0.3823	0.1502	0.4577	0.7028	0.1944
30	0.3304	0.4181	0.1871	0.4428	0.8066	0.2133
35	0.3803	0.4053	0.1985	0.4856	0.8117	0.2375
40	0.4897	0.4747	0.2371	0.6171	0.8084	0.3408
45	0.613	0.3692	0.184	0.5839	0.8462	0.2364
50	0.549	0.4485	0.2114	0.5048	0.9228	0.2672
AvgMSE	0.2873	0.3272	0.1575	0.3892	0.5906	0.1852



A real world problem in the form of *EColi* dataset (available by anonymous ftp from <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>) was selected for further analysis of these evaluation strategies. Table 7 details the composition of this dataset (8 inputs and 8 classes). As before, the data was normalized to facilitate computation of AIC and this was done using the technique specified by Mirkin (1996) and shown below.

$$v_{normalized} = \frac{v - P_v}{\sqrt{1 - \sum_v P_v^2}}$$

The AIC values and the confidence intervals on the error proportions are given in Tables 8 and 9 respectively. From Table 8, it is evident that the AIC values indicate an inferior classification capability (positive AIC values) and that the network with 2 hidden neurons is the best of the lot. From Table 9, it can be seen that none of the classifiers have identical classification capabilities (no closed interval containing '0') and that their performance is in the order 3,6,5,4,2,1.

Table7. Classes in the Ecoli dataset

Class ID	Class Name	Number of patterns
CP	Cytoplasm	143
IM	inner membrane without signal sequence	77
PP	Periplasm	52
IMU	inner membrane, un-cleavable signal sequence	35
OM	outer membrane	20
OML	outer membrane lipoprotein	5
IML	inner membrane lipoprotein	2
IMS	Inner membrane, cleavable signal sequence	2

Table 8. AIC values of networks developed for classifying the EColi dataset

# of Hidden Neurons	AIC
2	146.5111
3	170.166
5	199.6422
10	299.781
15	406.0171
20	472.2449

Table9. Confidence Intervals for difference in error

Model	2	3	4	5	6
1	0.0037718 0.01603	0.19189 0.20415	0.12258 0.13484	0.16219 0.17445	0.17209 0.18435
2		0.18199 0.19425	0.11268 0.12494	0.15229 0.16455	0.16219 0.17445
3			-0.075436 -0.063178	-0.035832 -0.023574	-0.025931 -0.013673
4				0.033475 0.045733	0.043376 0.055634
5					0.0037718 0.01603

The test of generalization accomplished by inducing additive white gaussian noise yielded the error proportions shown in Table 10. It is evident from the average error values and box plot in Figure 3 that network 3 is in fact the best and the performance of the networks is in the order 3,4,6,5,2,1 which is close to what is concluded from the above test. This is also in total contradiction to that indicated by AIC.

Table10. Error proportions when simulated in noisy environment

Neurons \ Noise (dBW)	2	3	5	10	15	20
1	0.6634	0.6436	0.5545	0.6238	0.5941	0.5446
5	0.6634	0.703	0.5248	0.6634	0.703	0.7129
10	0.8218	0.7921	0.6634	0.6634	0.7921	0.7723
15	0.8515	0.8218	0.6238	0.6832	0.6931	0.802
20	0.8218	0.802	0.703	0.703	0.7921	0.7822
25	0.8614	0.8416	0.6931	0.7624	0.7822	0.8119
30	0.8515	0.8515	0.6931	0.6535	0.802	0.7228
35	0.8416	0.7525	0.7129	0.6634	0.802	0.7426
40	0.8515	0.7723	0.6634	0.6832	0.7525	0.7624
45	0.8416	0.8416	0.7129	0.6832	0.7228	0.7525
50	0.8614	0.8416	0.7129	0.703	0.7723	0.7426
Avg Error	0.8119	0.7876	0.6598	0.6805	0.7462	0.7408

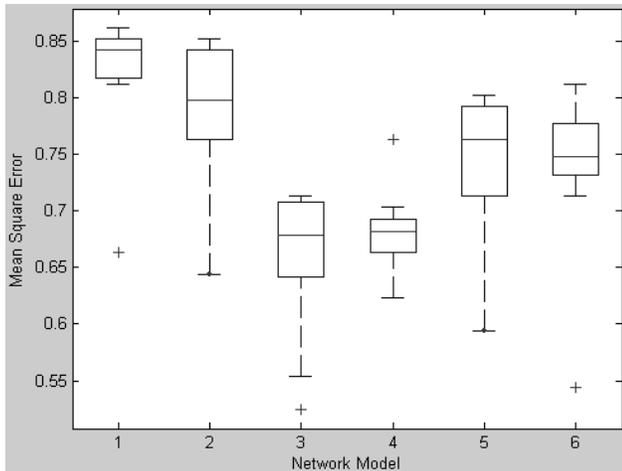


Figure 3. Box and whisker plot of MSE values

## 5. CONCLUSIONS

In this paper the model selection problem is studied in two of the most popular learning tasks, i.e. function approximation and classification. A desirable quality in any evaluation criteria is its capacity to indicate the tradeoff between precision and complexity of the underlying models. It is also particularly desirable to have a universally applicable criterion as the algorithms involved in machine learning and statistical approximation methods yield models that are extremely dissimilar in form and complexity.

Multiple comparison procedures such as the *Hochberg & Tamhane* test for function approximation and *studentized t* test for classification can point to significant differences in the approximation capabilities of the models. However these tests do not take into account the complexity of the models. The Akaike Information Criterion, designed to take into account the complexity as well as the precision of the model, was seen to perform extremely well in the function approximation arena while it falters in the classification domain. Studentized t test yields a better evaluation strategy when compared to AIC for the classification problems. However it has to be noted that this (in fact any statistical test) does not take into account the complexity of the model. A possible solution to this problem is to use *m-fold cross validation* to identify the candidate models. These models can be further filtered by using the comparison techniques.

## 6. REFERENCES

- [1]. Cherkassky. V., Mulier, F. (1998). *Learning From Data: Concepts, theory and methods*, 1st Edition, Wiley-Interscience, 20-31.
- [2]. Diettrich.T.G. (1997). *Approximate statistical tests for comparing supervised classification learning algorithms*, *Neural Computation*, 10, 895-924.
- [3]. Domingos. P. (1999). *The role of Occam's razor in knowledge discovery*, *Data Mining and Knowledge Discovery*, 3(4), 409-425.
- [4]. Duda. R. O., Hart, P. E. and Stork, D. G. (2002). *Pattern Classification*, 2nd Edition, Wiley-Interscience, 466-471.
- [5]. Feelders.A and Verkooijen.W. (1996). *On the Statistical Comparison of Inductive Learning Methods*, Springer-Verlag, 271—279.
- [6]. Ishikawa. M. (1996). Structural learning with forgetting, *Neural Networks*, 9(3), 509-521.
- [7]. Lawrence. S., Giles.L.C., Tsoi.C.A. (1997). Lessons in Neural Network Training: Overfitting May be Harder than Expected, *Proceedings of the fourteenth national conference on Artificial Intelligence*, 540-545.
- [8]. Mirkin.B.(1996). Mathematical classification and clustering, *Kluwer Academic Publishers*, 74-76.
- [9]. Ripley, B. D. (1995). Statistical ideas for selecting network architectures, *Neural Networks: Artificial Intelligence and Industrial Applications*, Springer, 183-190.
- [10]. Vapnik, V. (1999). *The Nature of Statistical Learning Theory*, 2<sup>nd</sup> Edition, Springer, 18-34.