Searching Most Efficient Neural Network Architecture Using Akaike's Information Criterion (AIC)

Gaurang Panchal
Charotar University of Science and Technology,
Changa, Anand-388 421, INDIA

Amit Ganatra
Charotar University of Science and Technology,
Changa, Anand-388 421, INDIA

Y.P.Kosta
Charotar University of Science and Technology,
Changa, Anand-388 421, INDIA

Devyani Panchal
Charotar University of Science and Technology,
Changa, Anand-388 421, INDIA

ABSTRACT
The problem of model selection is considerably important for acquiring higher levels of generalization capability in supervised learning. Neural networks are commonly used networks in many engineering applications due to its better generalization property. An ensemble neural network algorithm is proposed based on the Akaike information criterion (AIC). Ecologists have long relied on hypothesis testing to include or exclude variables in models, although the conclusions often depend on the approach used. The advent of methods based on information theory, also known as information-theoretic approaches, has changed the way we look at model selection The Akaike information criterion (AIC) has been successfully used in model selection. It is not easy to decide the optimal size of the neural network because of its strong nonlinearity. We discuss problems with well used information and propose a model selection method.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *Connectionism and neural nets.*

General Terms
General terms: Algorithms, Design.

Keywords
Neural Network, Hidden Neurons, Akaike's Information Criterion (AIC), Correct Classification Rate (CRR)

1. INTRODUCTION
Akaike's information criterion, developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971 and proposed in Akaike (1974), is a measure of the goodness of an estimated statistical model. It is grounded in the concept of entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality and can be said to describe the tradeoff between bias and variance in model construction. The AIC is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best.

From the AIC value if top three models are in a tie and the rest are far worse, but one should not assign a value above which a given model is 'rejected'.

The AIC is a basis of comparison and selection among several statistical models. As we all know the goodness of parameters of a model can be calculated by the expected log likelihood, means the larger the expected log likelihood is better explanation. In looking at the relationship between the bias and the number of free parameters of a model [1], it is found that,

(Maximum log likelihood of a model) – (number of free parameters of the model)

It is an asymptotically unbiased estimator of the mean expected log likelihood. AIC estimator of Kullback –Leibler information is

AIC = -2 * (maximum log likelihood of the model) + 2 * (number of free parameters of the model). (1)

2. PROBLEM STATEMENT
The goal is to find most efficient neural network architecture. It is difficult to select number of hidden neurons while designing neural network architecture [1].

3. AKAIKE'S INFORMATION CRITERION
In the general case, the AIC [5] is,

$$AIC = -2 * \ln(likelihood) + 2 * k \quad (2)$$

Where ln is the natural logarithm, k is the number of parameters in the statistical model and RSS is the residual sums of squares (Calculation of RSS value is discussed later in this paper). AIC can also be calculated using residual sums of squares [5] from regression

$$AIC = n * \ln(RSS/n) + 2 * K \quad (3)$$

Where n is the number of data points (observations). AIC requires a bias-adjustment small sample sizes. If ratio of n/K < 40 then uses bias adjustment

$$AIC = -2*\ln(L) + 2*K + (2*K*(K+1))/(n-K-1) \quad (4)$$

For example, consider 3 candidate models for the growth model, their RSS values, and assume n = 100 samples in the data.Table1 shows the Calculation of AIC values for different Models. AIC is calculated using different RSS values and different no of free parameters. Lower AIC is better for model

Table 1. Table captions should be placed above the table

K	RSS	AIC
4	25	100*ln(25/100) + 2*4 + (2*4*(4 + 1))/(100 - 4 -1) = -130.21
3	26	100*ln(26/100) + 2*3 + (2*3*(3 + 1))/(100- 3- 1) = -128.46
3	27	100*ln(27/100) + 2*3 + (2*3*(3 + 1))/(100- 3- 1) = -124.68

4. MODEL SELECTION WITH AIC

The best model is determined by examining their relative distance to the "truth". The first step is to calculate the difference between lowest AIC model and the others as

$$\Delta i = AICi - \min AIC \quad (5)$$

Where Δi is the difference between the AIC of the individual models and min AIC is the minimum AIC value of all models [5]. The smallest value of AIC is –130.21(using equation 5). Thus the Δi is show in table1.

Table 2. Calculation of Δi

K	RSS	AICc	Δ i
4	25	-130.21	0
3	26	-114.15	1.75
3	27	-98.73	5.53

To quantify the plausibility of each model as being the best approximating, we need an Estimate of the likelihood of our model given our data

$$L(Model \mid data)$$

Interestingly, this proportional ( ) to the exponent of

(-0.5*Δi) so that

$$L(Model \mid data) \alpha \exp(-0.5 * \Delta i)$$

The right hand side of above is known as the relative likelihood of the model, given the data. A better means of interpreting the data is to normalize the relative likelihood [5] values as

$$\sum i = \exp(-0.5 * \Delta i) / \sum_{i}^{R} \exp(-0.5 * \Delta i) \quad (6)$$

Table 3. Exponent of Delta

K	RSS	AICc	Δ i	exp(-0.5*Δ i)
4	25	-130.21	0	1
3	26	-128.46	1.75	0.4166
3	27	-124.68	5.52	0.0631
				Sum = 1.4798

The sum of the relative likelihoods is 1.4798, so we obtain the Akaike weights for each by dividing the relative likelihood by 1.4798.

Table 4. Akaike's Weights for different Models

K	RSS	AICc	Δ i	Wi
4	25	-130.21	0	0.6758
3	26	-128.46	1.75	0.2816
3	27	-124.68	5.52	0.0427

Where Wi are known as Akaike weights for model I and the denominator is simply the sum of the relative likelihoods for all candidate models. For example, using the earlier values from the 3 growth models:

For the above example, the first model is (0.6758/0.2816) = 2.4 times more likely to be the best explanation for growth compared to second Model only and (0.6758/0.0427) = 15.8 times more likely than third model only.

As a general rule of thumb, the confidence set of candidate models (analogous to a confidence interval for a mean estimate) include models with Akaike weights that are within 10% of the highest, which is comparable with the minimum cut-off point (i.e., 8 or 1/8) suggested by Royall (1997) as a general rule-of-thumb for evaluating strength of evidence.

For the above example, this would include any candidate model with a value greater than (0.6758*0.10) = 0.0676. Thus, we would probably exclude the third model only from the model confidence set because its weight, 0.0427<0.0676

5. RESIDUAL SUM OF SQUARED

We want to find a model which has an equation of the form

$$y = \beta 0 - \beta 1 \quad (7)$$

We want to find RSS of the red line. First we need the equation of this line. We know that the equation of a line can always be determined by two points on the line.

In statistics, the residual sum of squares (RSS) is the sum of squares of residuals. It is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data.

In general: total sum of squares = explained sum of squares + residual sum of squares.

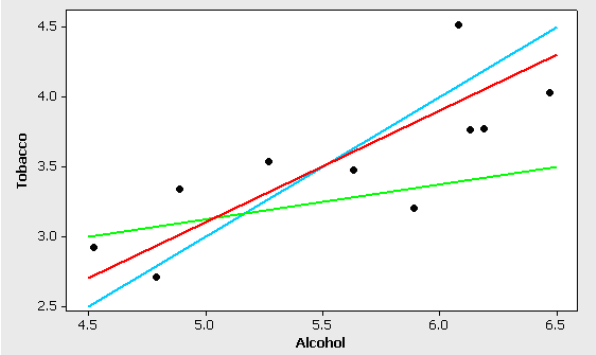

Figure 1. Selecting Red Line for RSS Calculation

To find the equation of this line we need to find their intercept and their slope.
Slope:

$$\beta 1 = (y2 - y1)/(x2 - x1) \quad (8)$$

is equal to (4.3-2.7) / (6.5-4.5) = 0.8

Intercept:

$$\beta 0 = y1 - \beta 1 - x1 \quad (9)$$

is equal to 2.7-0.8•4.5 = -0.9

With this we get the equation

$$y = -0.9 + 0.8x \quad (10)$$

Let's consider observation #4 with values x4=4.89 and y4=3.34 Now using the x value in our equation we get Y4 = -0.9 + 0.8*4.89 = 3.01. This is called the fitted value.

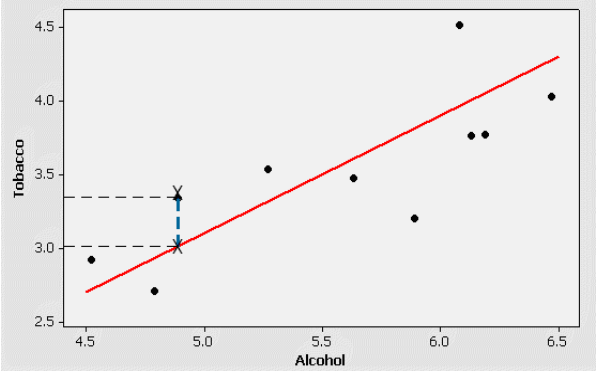

Figure 2. Residual (Errors) at Point Three

We have two y values. y4 = 3.34 is actual observation and Y4 = 3.01 is what our line predicted we should observe for an x value of x4=4.89.The difference between them is

$$\varepsilon 4 = y4 - Y4 = 3.34 - 3.01 = 0.33$$

ε4 is called the residual (or error).

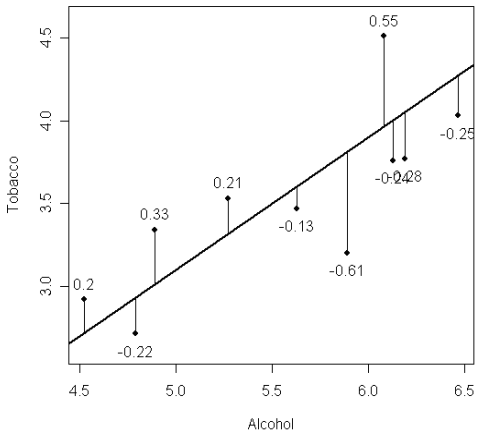

Figure 3. Individual Residual (Error) at Observation Points

Of course we can do the same thing for all the other observations: Find the fitted values, and then find the corresponding residuals ε1 to ε10.

Each of these errors shows how much the actual observed y value differs from what the model predicted. Finally we can combine all these individual errors into one overall error called Residual Sum of Squares (or RSS) [6] as follows:

$$RSS = \sum \varepsilon_i^2 = 1.13 \quad (10)$$

6. EXPERIMENT AND RESULT

We have tested this method for employee retention problem. To find the retention probability of employee of our organization using neural network, first we need to select the Neural Network architecture. We have total 17 inputs (e.g., age, sex, marital status, salary, experiences...) and retention probability is one output of neural network. By using AIC method we are getting following best Neural Network architecture. We assume the single hidden layers.

Table 5. AIC for different NN Architecture

i/p	h/d	o/p	Wgts	RSS	n	K	AIC
17	1	1	18	1	100	18	-416.073
17	2	1	36	1	100	36	-346.231
17	3	1	54	1	100	54	-220.517
17	4	1	72	1	100	72	72.816
17	5	1	90	1	100	90	1539.483
17	6	1	108	1	100	108	**-2860.517**
17	7	1	126	1	100	126	-1393.850
17	8	1	144	1	100	144	-1100.517
17	9	1	162	1	100	162	-974.803
17	10	1	180	1	100	180	-904.961

Where i/p is Input Neurons, h/d is hidden neurons, o/p is output neurons, wgts is total weights is sample data and k is free parameter. For example, inputs are 17 and one output increment the number of hidden neuron by one up to 10 hidden neurons; we get the lowest AIC at [17-6-1].

Table 6. AIC for NN with More number of Hidden Neurons

i/p	h/d	o/p	Wgts	RSS	n	K	AIC
17	2	1	36	1	100	36	-346.231
17	4	1	72	1	100	72	72.816
17	6	1	108	1	100	108	**-2860.517**
17	8	1	144	1	100	144	-1100.517
17	10	1	180	1	100	180	-904.961
17	12	1	216	1	100	216	-829.748
17	14	1	252	1	100	252	-789.929
17	16	1	288	1	100	288	-765.279
17	18	1	324	1	100	324	-748.517
17	20	1	360	1	100	360	-736.379

Similarly we increment the number of hidden neurons by two up to 20.We get the lowest AIC at [17-6-1].Always Low AIC is Best.

7. CONCLUSION

If there is more number of Input Neuron and Output Neurons at that time it is better to use AIC method for Model Selection. If number of inputs are more than it is difficult to select number of hidden neurons. After the Results and experiments, we can conclude that the Akaike's Information Criterion methods are giving best result for selecting Neural Network Architecture.

8. ACKNOWLEDGMENTS

The authors' wishes to thanks all the colleagues for their guidance, encouragement and support in undertaking the research work. Special thanks to the Management for their moral support and continuous encouragement.

9. REFERENCES

[1] S. Raksekaran, G.A. Vijayalakshmi.: Neural Network Fuzzy Logic and Genetic Algorithm Pai. PHI Publication (2005).

[2] Jaiwer Han, Micheline Kamber.: Data Mining, Concepts & Techniques Morgan. Kaufmamn Publication (2005).

[3] Pythia Version 1.02, The Neural Network Designer, http://www.runtime.org

[4] Alyuda NeuroIntelligence 2.2, http://www.aluyada.com.

[5] AIC,Wikipidia,http://www.wikipidia.com

[6] AIC, http://www.modelselection.org

[7] AIC, http://en.wikipedia.org/wiki/Residual_sum_of_squares