Performance Evaluation Of Linear Regression And Neural Networks On Forecasting Numerical Data Sets

[1] B. A Ganesh, Asst professor in Department of computer science and engineering,

Vignan's Institute of Engineering for Women, Visakhapatnam.

[2] Rama Janaki Ramireddy, Asst professor in Department of CSE,

Vignan's Institute of Engineering for Women, Visakhapatnam.

[3] N. K. Santosh ,Asst professor in Department of computer science and engineering,

Vignan's Institute of Engineering for Women, Visakhapatnam.

[4] S. Ram Prasad Reddy, Assoc. professor in Department of CSE

Vignan's Institute of Engineering for Women, Visakhapatnam.

IJERT

Abstract

Data mining is a subject which deals with knowledge discovery from a huge set of repositories. It is the confluence of different subjects like statistics, machine learning, artificial intelligence, etc. There are different functionalities which are provided by data mining. Prediction is one of the functionality of data mining which can be done by using regression techniques and neural networks. This work is a study on the performance evaluation of the regression and neural networks models on baby weight forecasting. The study was conducted between linear regression model and neural network. Neural network was constructed by using the back propagation algorithm with one input node and one output node with one hidden layer. The comparative analysis study was done on different data sets based on the correlation of the data.

***KEYWORDS**: LINEAR REGRESSION, NEURAL NETWORKS, PERFORMANCE EVALUATION*

1. INTRODUCTION

Linear regression is an approach to modeling the relationship between scalar dependent variable y and one or more explanatory variables denoted x. The case of one explanatory variable is called simple linear regression while for more than one explanatory variable; it is termed as multiple linear regression. The linear regression model can be represented in the form of

$$Y = a + bx + e$$

Where the x is an independent variable and the "residual" e is a random variable with mean zero. The coefficients a and b are determined by the condition that the sum of the square residuals is as small as possible.

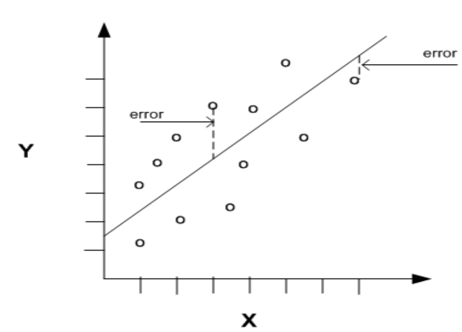

Figure 1

A **neural network** is a collection of input and output units. The connections between the units have associated weights. During the learning stage, the weights were adjusted so that the correct class label can be predicted. Hence it is also referred to as "connectionist learning". The algorithm that is used to train the network in this system is back propagation algorithm.

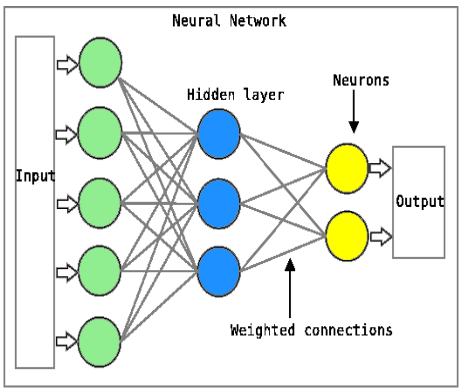

Figure 2: Simple Neural Network Model

2. Related work

The literature so far, comprises of papers briging out the comparision of these two models by considering the specific applications such as comparison on defect color data on CRT color display [1]; comparison of rainfall data [2]; comparison of trafic accidents data [3]. The comparisons were carried out by considering the statastical measures like MSE(mean square error), R^2 values and correlation coffiecients. This research was conducted on general numerical data by considering the correlation among the attribute values.

3.1 Datasets and methods:

3.1.1 Baby weight data set

For forecasting the baby weight at different stages , Gestation period (in weeks) and weight (in ounce) are considered as attributes. Gestation is the period during which an embryo develops (about 266 days in humans). The average human gestation length is calculated as 40 weeks. Childbirth occurring anywhere between 37-42 weeks is considered to be normal, whereas a baby born before 37 weeks is called pre-term and the baby born after 42 weeks is called post-term.

ATTRIBUTE NAME	**MEASURE**
GESTATION PERIOD	IN WEEKS
WEIGHT	OUNCE

3.2 Temperature data set

In the data set of temperature, Celsius and Fahrenheit are the attributes and following is the conversion formula from Celsius to Fahrenheit

$$°C \times 9/5 + 32 = °F$$

ATTRIBUTE NAME	**MEASURE**
TEMPARATURE	CELSIUS AND FARENHEIT

3.3 Input and Output variable choice:

In case of neural network, we have considered one input and one output, where gestation period in weeks and temperature in Celsius are given as input in different cases and the output variables are weight and temperature in Fahrenheit. Similarly in the case of linear regression we have taken gestation period and temperature in Celsius are the independent variables and weight and temperature in Fahrenheit as dependent variables.

3.4 Back-Propagation Learning:

The back-propagation algorithm has emerged as the workhorse for the design of a special class of layered feed forward networks known as multilayer perceptrons (MLP) [4]. A multilayer perceptron has an input layer of source nodes and an output layer of neurons (i.e., computation nodes); these two layers connect the network to the outside world. In addition to these two layers, the multilayer perceptron usually has one or more layers of hidden neurons, which are so called because these neurons are not directly accessible. The hidden neurons extract important features contained in the input data.The training of an MLP was accomplished by using a BACK PROPAGATION (BP) algorithm.

Back propagation learns by iteratively processing a set of training sample, comparing the network's prediction for each sample with the actual known value. For each training sample, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual value.

IJERT

These modifications are made in the "backwards" direction, that is, from the output layer, through each hidden layer down to the first layer.

3.4.1 Initialize the weights:

The weights in the network are initialized to small random number (e.g., ranging from -1.0 to 1.0 or -0.5 to 0.5). Each unit has a bias associated with it. The biases are similarly initialized to small random numbers.

3.4.2 Propagate the inputs forward

In this step, the net input and output of each unit in the hidden and output layers are computed.[4] First the training sample is fed to the input layer of the network. Note that for unit j in the input layer, its output is equal to its input, that is, $O_j = I_j$ for input unit j. The net input to each unit in the hidden and output layers is computed as a linear combination of its inputs. The inputs to the unit are, in fact, the outputs of the units connected to it in the previous layer. To compute the net input to the unit, each input connected to the unit is multiplied by its corresponding weight, and this is summed. Given a unit j in a hidden or output layer, the net input, I_j, to unit j is

$$\mathbf{I_j=\Sigma_i W_{ij} O_i+\theta_j,}$$

Where w_{ij} is the weight of the connection from unit i in the previous layer to unit j; O_i is the output of unit i from the previous layer; and θ_j is the bias of the unit. The bias acts as a threshold in that is serves to vary the activity of the unit.

Each unit in the hidden and output layers takes its net input and then applies an activation function to it[4]. The function symbolizes the activation of the neuron represented by the unit. The sigmoid function is used. Given the net input I_j to unit j, the O_j, the output of unit j, is computed as

$$\mathbf{O_j} = \frac{1}{1+e^{-Ij}}$$

This function is also referred to as a squashing function, since it maps a large input domain onto the smaller range of 0 to 1. The logistic function is non-linear and differentiable, allowing the model classification problems that are linearly inseparable.

3.4.3 Back propagate the error

The error is propagated backwards by updating the weights and biases to reflect the error of the network's prediction [4]. For a unit j in the output layer, the error Err_j is computed by

$$\mathbf{Err_j = O_j\,(1\text{-}O_j)\,(T_j\text{-}O_j)}$$

Where O_j is the actual output of unit j, and T_j is the true output and O_j $(1\text{-}O_j)$ is the derivative of the logistic function.

The error of a hidden layer unit j is

$$\mathbf{Err_j=O_j(1\text{-}O_j)\Sigma_k Err_k W_{jk},}$$

Where w_{jk} is the weight of the connection from unit j to a unit k in the next higher layer, and Err_k is the error of unit k.

The weights and biases are updated to reflect the propagated errors. Weights are updated by the following equations, where Δw_{ij} is the change in weight w_{ij}:

$$\mathbf{\Delta w_{ij} = (l)\ Err_j O_i}$$

$$\mathbf{w_{ij} = w_{ij}+\Delta w_{ij}}$$

The variable l is the learning rate, a constant typically having a value between 0.0 and 1.0. Back propagation learns using a method of gradient descent to search for a set of weights that can model the given classification problem so as to minimize the mean squared distance between the network's class prediction and the actual class label of the samples. The learning rate helps to avoid getting stuck at a local minimum at a local minimum in decision space and encourages finding the global minimum. If the learning rate is too small, then learning will occur at a very slow pace. If the learning rate is thumb is to set the learning rate to 1/t, where t is the number of iterations through the training set so far.

Biases are updated by the following equations below, where $\Delta\theta_j$ is the change in bias θ_j:

$$\mathbf{\Delta\theta_j = (l)\ Err_j}$$

$$\mathbf{\theta_j = \theta_j+\Delta\theta_j}$$

IJERT

Weights and biases are updated after the presentation of each sample, referred to as case updating. Alternatively, the weight and bias increments could be accumulated in variables, so that the weights and biases are updated after all of the samples in the training set have been presented.This latter strategy called epoch updating, where a single iteration through the training set is an epoch. In theory, the mathematical derivation of back propagation employs epoch updating,

3.4.4 Terminating condition

Training stops when

- all Δw_{ij} in the previous epoch were so small as to be below some specified threshold, or
- the percentage of samples misclassified in the previous epoch is below some threshold, or
- A prespecified number of epochs have expired.

3.5 Linear regression

In linear regression, data are modeled using a straight line. Linear regression is the simplest form of regression. Bivariate linear regression models a random variable, Y (called response variable), as linear function of another random variable, X(called a predictor variable),that is,

$$Y = \alpha + \beta X,$$

Where the variance of Y is assumed to be constant, α and β are regression coefficients specifying the Y-intercept and slope of the line, respectively. These coefficients can be solved for by the method of least squares, which minimizes the error between the actual data and the estimate of the line.

The least squares estimator a, b for α, β respectively in the method of least squares are:

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \frac{\Sigma y - b\Sigma x}{n}$$

where x, y values represent the attribute values.

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between -1 and +1.

- If value is 1 then all the data points lie perfectly along a straight line with positive slope.
- If value is -1 then also all the data points lie perfectly along a straight line but with a negative slope.
- A value that is close to either +1 or -1 signifies clustering of the data points around a straight line.
- If the value of correlation is 0, it indicates the presence of nonlinearity.

Corelation coffiecient:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

here X,Y are the corresponding attributes

4.Research approach and objective:

By considering the hypothetical statement " The correlation among the independent and dependent attributes influences the technique that adopted for predicting the future values" and proved the above statement with the help of temparature data set. Since the correlation coefficient for celsius and farenheit is exactly one the linear regression is performing better than neural networks but not all attributes and data sets will have correlation coefficient equals to exactly one and it ranges from -1 to 1.Hence a question raises regarding the performance of technique adopted for prediction of unknown values different data sets. forecasting the baby weight in order to evaluate the performance.

Since in the acual data set the gestation period and baby weight are correlated but not having the correlation coefficient value exactly one so,comparative analysis is made on two techniques i.e linear regression and neural networks for

In this paper work we conducted two experiments .the first experiment was to prove that correlation influences technique adopted for forecasting and the second experiment was conducted to evaluate the performance of the adopted techniques.The results were verified and tabulated.

IJERT

5. Experimental Results

Experiment 1: comparing the linear regression output with neural network output for temparature data set whose r=1

Table 1: Celsius Vs Farenheit Temparature

Celsius	Farenheit	Linear regression Output	Neural network output	Difference b/w Linear regression output and Actual output	Difference between Neural network and actual output.
101	213.8	213.8	208.65	1.53E-05	5.15
102	215.6	215.6	209.89	1.53E-05	5.71
103	217.4	217.4	211.1	3.05E-05	6.3
104	219.2	219.2	212.3	3.05E-05	6.9
105	221	221	213.48	3.05E-05	7.52
106	222.8	222.8	214.64	3.05E-05	8.16
107	224.6	224.6	215.78	1.53E-05	8.82
108	226.4	226.4	216.91	0	9.49
109	228.2	228.2	218.01	1.53E-05	10.19
110	230	230	219.1	1.53E-05	10.9
111	231.8	231.8	220.17	1.53E-05	11.63
112	233.6	233.6	221.22	3.05E-05	12.38
113	235.4	235.4	222.25	1.53E-05	13.15
114	237.2	237.2	223.27	1.53E-05	13.93
115	239	239	224.26	1.53E-05	14.74

IJERT

The T**able 1** shows that when the correlation cofficient is one, the difference between Linear regression and its actual out put value is 0. Hence it is shown that linear regression is more accurately predicts the output value

5.1 Result analysis for temparature data set:

From the plotted figures 1(a) and 1(b) , it is clear that for the data set whose correlation coefficient is exactly one, both the techniques are working good but comparatively linear regression is performing better than neural networks.

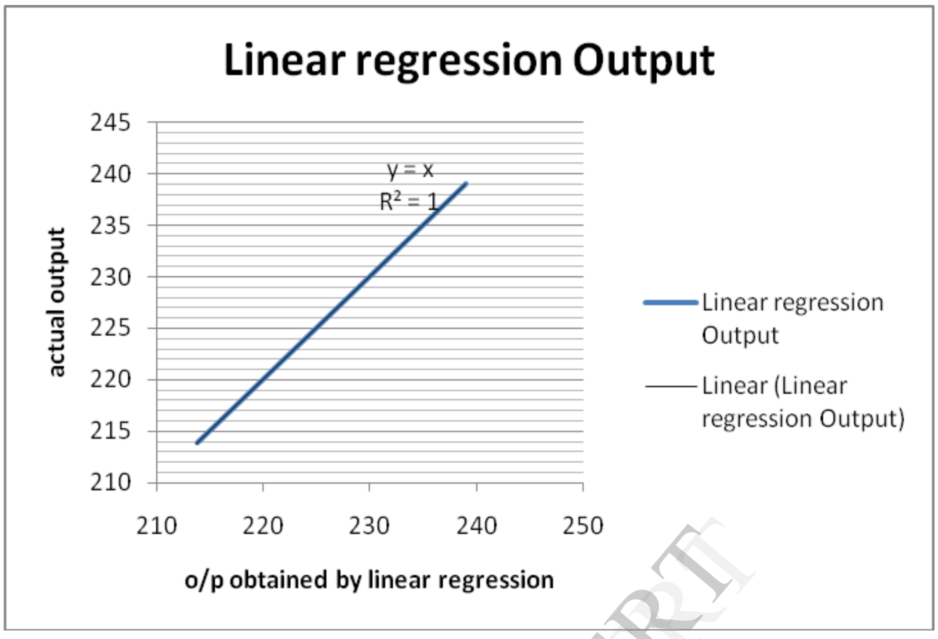

Figure 1(a)

IJERT

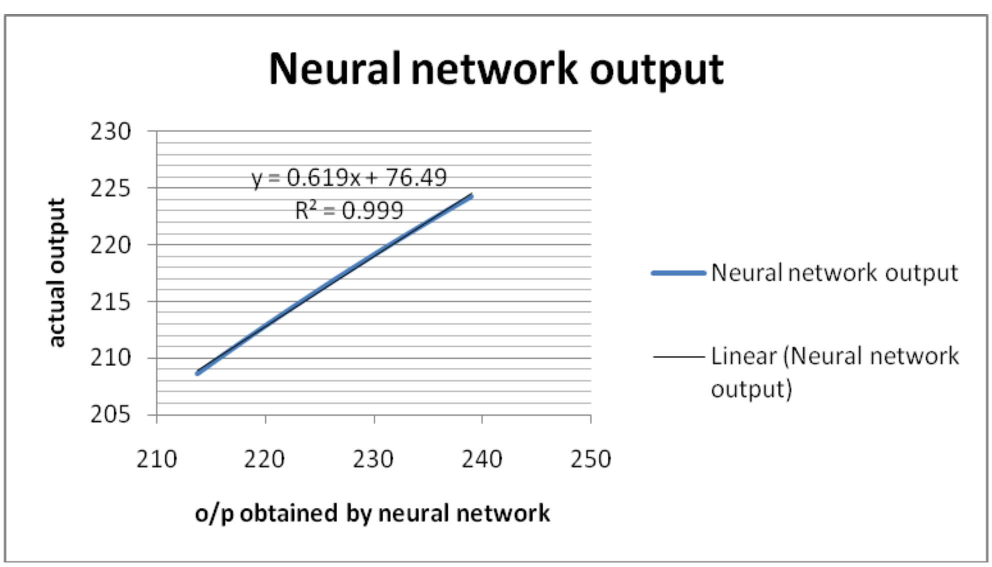

Figure 1(b)

Experiment 2: comparing the linear regression output with neural network output for baby weight whose r≠1

When the correlation coefficient is not equal to one between the attribute values, the comparision is as shown in the table 2.

Table 2: Gestation period Vs weight

Gestation period	weight	Linear regression output	Neural network output	Difference b/w Linear regression output & Actual output	Difference between Neural network & actual output
21	14	30.44	17.9	16.4473	35.59
22	15	35.20	16.2	20.20884	37.42
26	39	54.25	41.3	15.25499	25.91
32	93.5	82.82	86.7	10.67577	6.8
33	92	87.58	90.62	4.41423	1.38
35	99.1	92.34	94.6	17.01397	19.266
36	106	97.10	98.64	2.002266	0.4711
37	112	101.8	102.7	4.52961	3.68
38	110	106.6	106.8	5.653786	5.4357
39	119	111.3	111	0.960121	0.5666
40	123	116.1	115.1	3.147331	4.1223
41	121	120.9	119.3	2.352699	3.8992
42	132	125.6	123.5	4.223549	2.1054
43	121	130.4	127.7	1.893723	4.5833
44	100	135.2	131.9	14.20116	10.94

IJERT

Table 2 shows The difference between the neural network out put and actual value is less when compared to its counter part.the Network architecture for this neural network is 2-2-1.When the corelation coffiecient is equal to one,the results are in the following manner.

5.2 Result analysis for baby weight data set:

From the plotted figures 2(a) and 2(b) , for this data set whose correlation coefficient is not exactly one, it is observed that neural network is performing better than linear regression.

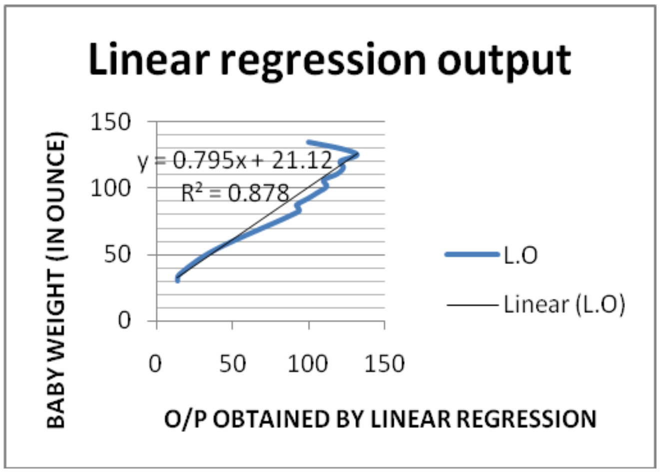

Figure 2(a)

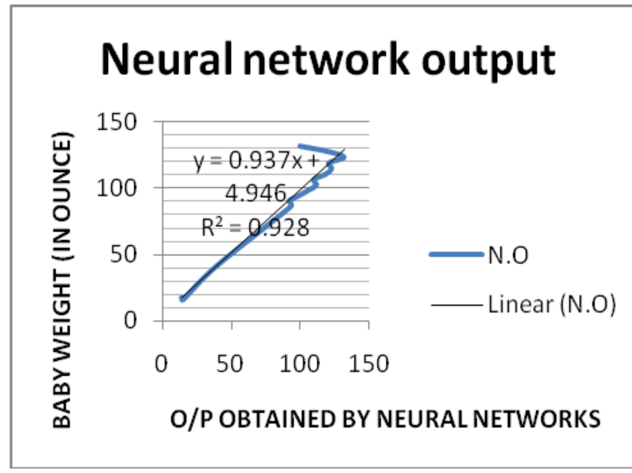

Figure 2(b)

6. Conclusion

From both the experiments we can conclude that Neural network performs well and gives better prediction values in all the cases irrespective of the correlation among the attributes of different data sets whereas linear regression produces good prediction when the correlation value is one.

7. References

[1] Mauridhi Hery PURNOMO,Toshio ASANO,EiJJi SHIMIZU, "A Comparative Study of Neural Network Approach and Linear Regression for Analysis of
Multivariate Data of the Defect Color on the Color CRT Displays",Mem. Fac. Eng., Osaka City Univ., Vol. 38, pp.15-22.

[2] A. El-shafie,M. Mukhlisin, Ali A. Najah and M. R. Taha,"Performance of Artificial NeuralNetwork and regression Techniques for rain fall round off prediction", International Journal of the Physical Sciences Vol. 6(8), pp. 1997-2003.

[3] DR GALAL A ALI and DR CHARLES S.BAKHEIT Comparitive Analasyis of Traffic Accidents in SudanUsing Neural Networks and Statistical Methods"

[4] M.R NarasingaRao, G.R. Sridhar, K.Madhu, A.A.Rao, A Clinical Decesion Support system Using Multilayer Perception Neural Network to Assess Wellbeing in Diabetes", Journal of Association of Physicians India,Volume 57.

IJERT