International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9820-9825
© Research India Publications. http://www.ripublication.com

Comparison and Analysis of Linear Regression & Artificial Neural Network

**Ki-Young Lee[1], Kyu-Ho Kim[1,*], Jeong-Jin Kang[2], Sung-Jai Choi[3],
Yong-Soon Im[4], Young-Dae Lee[5], Yun-Sik Lim[6]**

[1]*Department of Medical IT, Eulji University, Seongnam 13135, Korea.*
[2]*Department of Information and Communication, Dong Seoul University, Seongnam 13117, Korea.*
[3]*Departmentof Electronic Engineering,Gachon University, Seongnam 13120, Korea.*
[4]*Department of Computer Information & Communication, Kookje University, Pyeongtaek 17731, Korea.*
[5]*The International Promotion Agency of Culture Technology (IPACT), Seoul 05719, Korea.*
[6]*Department of Electrical Engineering, Yeoju Institute of Technology, Yeoju 12652, Korea.*

[1,*]*Corresponding Author: Kyu-Ho Kim*

Abstract

Recently, Artificial intelligence is gaining traction after learning of data on computers after focusing on learning data on computers through big data analysis. In this thesis, companies were expected to use accounting data to generate continuous investments in industries in the future industries using accounting data. In other words, after studying survey data on the status of school companies in terms of school establishment type, school class, income statement, region, and industry, we studied which algorithm is more efficient. The learning process of this paper is comparing and analyzing the results of accuracy through the algorithm of linear regression and deep learning neural network through supervised learning based on correct answer label. In addition, we investigated the problem of Neural Net with respect to the number of hidden layers. In conclusion, we present the conclusion as to which algorithm is more efficient if the numeric data is a single attribute value.

Keywords: BigData, Supervised Learning, Linear Regression, Neural Network

INTRODUCTION

AI technology has emerged as a key technology to change the future of the future. The first boom of the inference and seeking center of the 1960s, the second boom with the expert system of the 1980s, and the third boom with recent big data and deep running. There are two major types of AI research. The first is the research method (Supervised learning) of how to effectively express human knowledge and perform logical inference based on it. The second is a study (Unsupervised learning) of how computers learn their own knowledge from many data and perform their assigned tasks [1]. The following figure1 illustrates the distinction between artificial intelligence and machine running and dimneoning. Firstly, the beginning was artificial intelligence, followed by data - based machine learning, and deep - run based on artificial neural networks, which have recently begun to attract attention.

Artificial Intelligence (AI) refers to 'CI (Computer Intelligence)', a concept that emerged in 1956 as a human being with human sense and thinking ability. On the other hand, machine learning basically uses algorithms to analyze data, learn through analysis, and make judgments or predictions based on the learning contents. Finally, Deep Learning is a method of 'learning' the computer itself through a large amount of data and algorithms, and then performing a parallel execution method. If existing artificial intelligence is to code specific instructions for decision-making directly into software, then Deep-Learning is an artificial neural network that was originally created by machine learning researchers, Characteristics and the structure of neurons are applied to artificial intelligence. That is, deep learning is artificial intelligence developed in artificial neural network. Data is learned by using information input / output layer similar to brain neurons [2]. The following are the forecasts for future AI areas in our lives. Experts in industry, academia and research believe that the effects of AI are very broad and include computer and IT, financial accounting, medical and biotechnology, legal services, manufacturing, chemicals, online customer service, transportation, toys and games, music and entertainment, Media, publishing, editing, and so on [3][4][5][6].

This paper aims to compare the artificial neural networks of linear learning and deep learning of machine learning to predict the results for future industries. And we investigated the solution of the overfitting problem through artificial neural network.

RELATED RESEARCH

Regression Model

In general, a linear relationship is a line showing a constant increase or decrease ratio, indicating a trend line of data.In other words, the value of the dependant variable is changed to

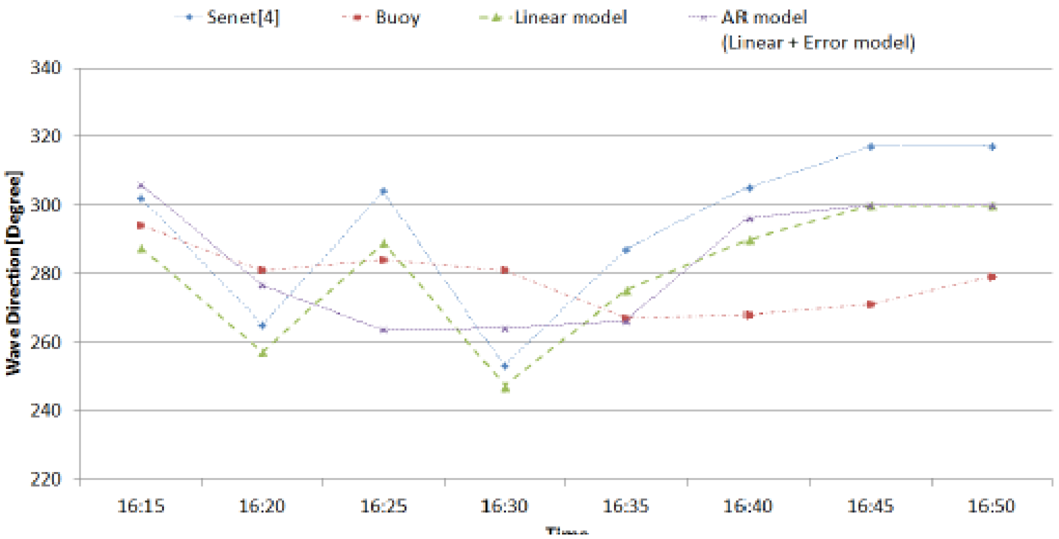

Figure1. Predicting Wave Information Through Linear Regression in Paper on "Wave Information Estimation and Revision Using Linear Regression Model(2016)"

a constant pattern, and the relationship between these variables is referred to as linear regression in the linear function of the primary function, and the statistical modelling method is the most commonly used method [7]. Multiple linear regression analysis assumes a linear relationship between multiple independent variables (X1, X2, ..., Xn) and dependent variables (Y), which calculates the effects of each independent variable (β) using the following expressions [8]. $Y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta n\ xn + \varepsilon$ In multiple linear regression analysis, many input variables are susceptible to noise data or contain unnecessary information, thereby reducing the predictive power of regression analysis [9].

In this paper, They compare and analyze the performance of the wave information prediction algorithm in order to improve the stability of the navigation and the efficiency through the maritime traffic control due to frequent occurrence of marine accidents. In addition, by introducing the linear regression model, Algorithm. First, this paper is a process of finding waves. It is related to this paper in that it predicts the predicted dangerous waves by applying linear regression algorithms[10]. Figure 1 shows the result of the linear regression.

Neural Network

Neurons as a mathematical model are interconnected to form a network, which is called a neural network, as humans are connected to neurons, which are basic structural tissues of the brain. The basic function of neurons is acceptance of information, processing of computation, output of information, and the formation of neurons by connecting many neurons together. Artificial neural network theory is a form in which the output is performed by a predetermined nonlinear function on multiple inputs. All neurons in the neural network model are divided into an input layer, a hidden layer, and an output layer depending on the function, and each layer is functionally connected. The input layer connects the external input mode and is transmitted in units of hidden layers according to the input unit. Here, the hidden layer is the inner processing unit layer of the neural network and the output layer is used to generate the output mode[8]. In Figure 3, the artificial neural network is connected by a group of nodes, which are similar to the network of vast neurons in the brain. In the figure 2 above, each circular node represents an artificial neuron, and an arrow represents the input from one neuron output to another neuron.

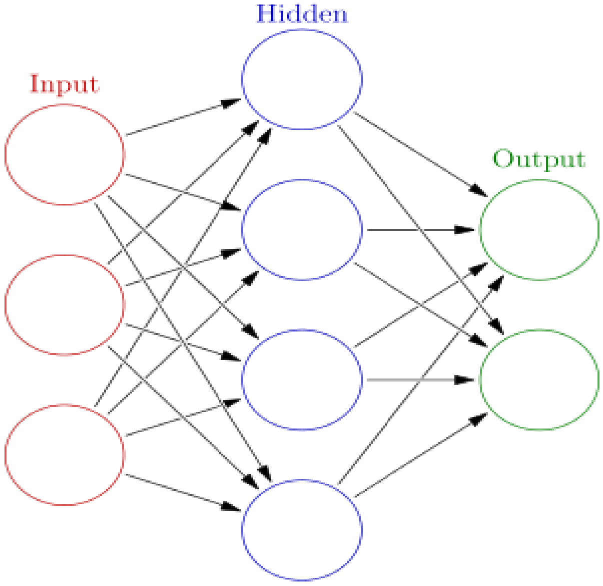

Figure 2. Artificial Neural Network that Outputs Result Value through Hidden Layer

Performance of artificial neural networks, It is known that the number of hidden layers, the number of nodes in the hidden layer, the number of learning iterations, learning rate, and momentum affect the optimization [11]. Back Propagation Algorithm is the most widely used method for learning artificial neural networks because it continuously adjusts the difference between the target result and the result calculated by the model This is a method that learns to minimize the error value and verifies the predicted object using the learning result [12]. According to the article "Electricity Price Prediction Based on Supervised Learning and Neural Network Algorithms", the actual values are predicted through the neural network as the meteorological variables and the electricity prices until the past several points.

Neural networks have been widely used for time series prediction because the structure is flexible to train time series data. The number of nodes in the output layer is one, and the predicted output value of the electricity sales unit price is obtained from neural network learning [13]. According to the "forecasting model of agricultural products price by optimizing the hidden layer of artificial neural network" which is related to the neural network experimental method used in this paper, the number of hidden layers is 17 and the momentum value is 0.5, It is confirmed that the learning rate is predicted to be close to the actual price when the value is 0.4 [8]. Therefore, in this paper, it is necessary to clarify the number of hidden layers in order to obtain accurate values.

Possibility of overfitting according to the number of HiddenLayer layers

Most of the artificial neural network models consist of a single input / output layer and a few hidden layers. The input layer is a set of input element nodes required for learning. These nodes are connected to each node of the hidden layer in Fully Connected form, but the nodes in the same layer are not interconnected. Since there is no clear rule for determining the optimal number of nodes for each floor, the number of nodes is determined by a heuristic method or a trial and error method. There is no hidden layer or it is composed of one or more layers, and there is no clear decision rule like the determination of the number of nodes in the hidden layer. According to a study by Tam & Kiang (1992), a comparison between one with and without a hidden layer showed excellent results with one artificial neural network [11], According to the research by Collins et al. (1988), Dutta &Shekhar (1988) and Salchenberger et al. (1992), the predictive rate of artificial neural networks did not improve.

Therefore, in most previous studies, artificial neural networks with a single hidden layer are used to enhance the performance of artificial neural networks and reduce learning time [14]. In the network structure including the hidden layer, if the number of nodes of the hidden layer is increased, the classification ability is improved and more patterns can be recognized, but it is likely to result in excessive fitting. Therefore, it is necessary to determine the appropriate number of cloaking. There is no requirement that the number of hidden nodes in the neural network should be several. However, since the model is constructed using the appropriate level of nodes, if there are too many nodes, explanatory power will be exploited due to overfitting problem, but in many cases, the value becomes meaningless in prediction. In addition, the number of nodes, connection function, and activation function should be controlled in a single layer. Above-the-shelf hindrances impair the speed of artificial neural networks [15].

EXPERIMENT

Experimental Data & Experimental Environment

Experimental data is data of school companies in 2016, and attributes include school class, income statement (sales, operating profit, net profit), regional distribution, and industry distribution. There are 431 data tuples. The attribute of industry type distribution used in the experiment is composed of nominal data, which is not easy to analyze. Therefore, we changed to the numeric data type. The following Table 1 is an example of a part of the experimental original data. Table 2 below shows the data values modified by the Numeric Data format.

Table 1. Experimental Data

NO.	Type of establishment	Level of School	Total	Profit	Short Profit	Area	Industry Distribution
1	Private	University	1654620000	95381000	100961000	Non Capital	Service
3	Private	University	778078175	257721941	0	Capital	Manufacturing

Table 2. Parts of Numeric Data that Changed Attribute Values for Experiment

Total	Profit	Short Profit	Target
1654620000	95381000	100961000	10
778078175	257721941	0	1
2898053	-5344293	-2136160	1
326074511	-302045681	-302055502	2
626569372	17490193	-53611374	3
332141000	-104449428	-788727	4

We used the Rattle library of R-Programming to support Regression Model and Neural Net. R-Programming is a big data open source program that supports statistics, data mining, and so on. The Rattle library is used because the GUI has the advantage of being able to combine with the package to use only certain functions and to display it in graph form.

3.2 Regression Linear and Neural Net Algorithm Analysis

Figure 3 shows the result using the regression linear model. The solid line represents the actual input data value and the dotted line represents the value in the predicted form. The target is captured as net gain and you can see how close it is to the actual value. The R-square value is 0.6111, which means that the variance of the causal variables is about 60%. In statistics, the coefficient of determination is a measure of the extent to which an estimated linear model is appropriate for a given data set. It refers to the percentage of the variable that can be explained by the applied model among the variation of the response variable. The usual sign of the coefficient of determination is.

It is generally interpreted as the explanatory power of the model, but it should be interpreted because it increases as the explanatory variable enters the model. To solve this problem, adjustment decision coefficients have been proposed. The value of the coefficient of determination is between 0 and 1, and the higher the correlation between the dependent variable and the independent variable, the closer to 1. In other words, a regression model with a decision coefficient close to zero has a low usefulness, whereas a larger decision coefficient has a high usefulness of a regression model [16]. The R-sqaure value is known to have a significant value when it is 0.6 or more.

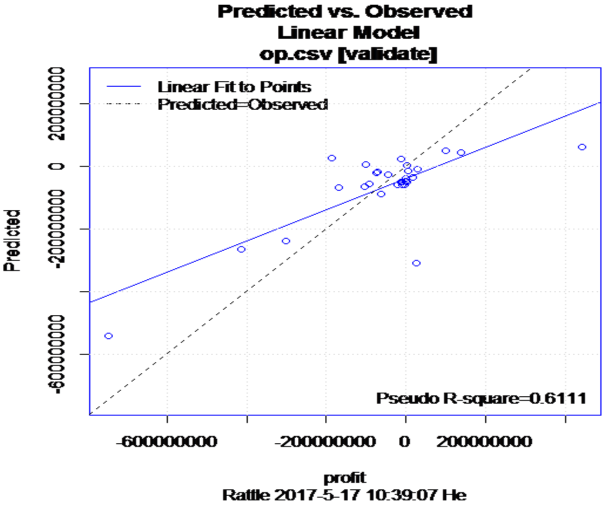

Figure 3. Linear Regression

Figure 4 shows the results using the Neural Net Model. Similarly, the dotted line indicates the predicted result value, and the solid line generally indicates that the more the number of hidden layers, the more optimized it is for the specific training data. However, as shown in Figure 4 below, it can be seen that the R-sqaure decreases as the number of hidden layers increases. This can be seen as a problem caused by insufficient amount of training data.

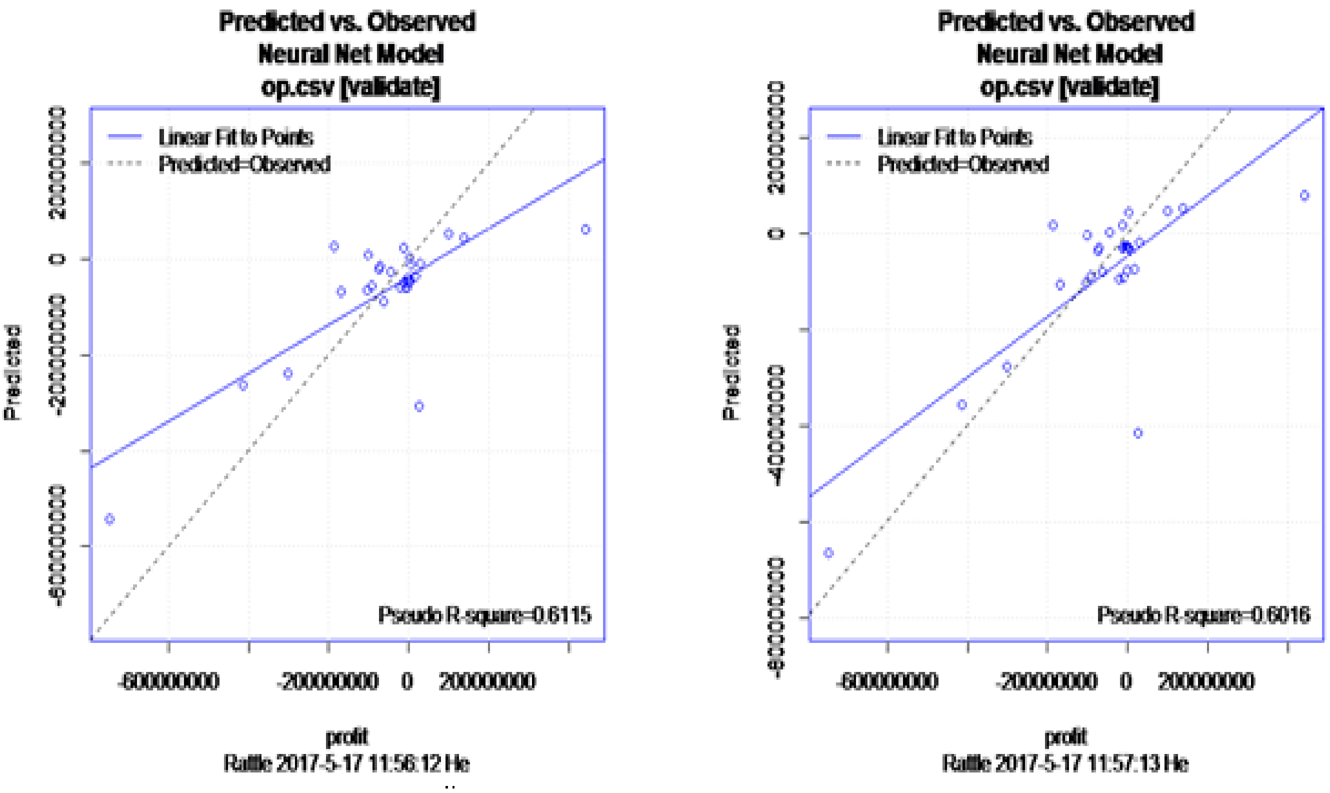

Figure 4. HiddenLayer=1, HiddenLayer=5 Neural Net

In other words, you need to increase the amount of training data to get the right results. Another solution to this is through dropout. Dropout refers to learning a hidden neural network by omitting some neurons in the hidden layer. In this way, when the learning of the omitted network is finished, the learning is performed repeatedly while omitting the other neurons at random [17]. In this paper, a comparison and analysis of Linear Regression and Neural Net Model was not conducted. In summary, linear regression analysis showed that the numerical value was constant even though the number of data was not large. However, when applying Neural Net Model, it is necessary to consider the overfitting problem according to the number of data.

4. Conclusions

In this paper, we analyze the predicted value of profits by applying Linear Regression and Neural Net Model with industry data by school companies. As a result of comparison, it can be said that Neural Net has a slightly higher accuracy, but the more the HiddenLayer layer is, the lower the result is. This shows that the more the number of HiddenLayers is, the more the OverFitting problem occurs with respect to the weight. To solve this problem, it is effective to increase the number of training data by increasing the number of data of the attribute value, or to reduce the hidden layer layer to solve the OverFitting problem.

REFERENCES

[1] KEIT PD issue report, 16, 03 (2016), 29-51

[2] Evolution of artificial intelligence, another industrial revolution. *LG Business Insight*, (2015).

[3] Jung, Y.G., Choi, J.A., Cha, B.H., Analysis of Radioactive Contamination Normal Level of Numerical Isotope using Clustering Methods. *The Journal of the Institute of Internet, Broadcasting and Communication (JIIBC)*, 14, 6 (2014), 41-46.

[4] Min, H., Heo, J.Y., Document Clustering Scheme for Large-scale Smart Phone Sensing. *The Journal of the Institute of Internet, Broadcasting and Communication (JIIBC)*, 14, 1 (2014), 253-258.

[5] Hong, Y.S., Park, C.K., Cho, S.S., Suck-Joo Hong4*, Intelligence Transportation Safety Information System. *International Journal of Internet, Broadcasting and Communication (IJIBC)*, 6, 2 (2014), 20-24.

[6] Yutaka matsuo, Artificial intelligence and Deep Learning. ISBN 979-11-86008-23-2.

[7] G.A.F Seber and A.J. Lee, Linear Regression Analysis. *John Wiley & Sons Publishers*, USA, 2012.

[8] Bae, K.T., Kim, C.J., An Agricultural Estimate Price Model of Artificial Neural Network by Optimizing

Hidden Layer. *Journal of Intelligent Information Systems*, 12 (2016), 161-169.

[9] Park, H.J., Jang, K.Y., Lee Y.H., Kim, W.J., and Kang P.S., Prediction of Correct Answer Rate and Identification of Significant Factors for CSAT English Test Based on Data Mining Techniques. *KIPS Tr. Software and Data Eng*, 4, 11 (2015), 509-520.

[10] Lim, D.H., Kim, J.S., Lee, B.G., Wave Information Estimation and Revision Using Linear Regression Model. *Journal of Korea Multimedia Society*, 19, 8 (2016), 1377-1385.

[11] Cho, Y.H. and Kim, I.H., Predicting the Performance of Recommender Systems through Social Network Analysis and Artificial Neural Network. *Journal of Intelligent Information Systems*, 16, 4 (2010), 159-172.

[12] Ryoo, E.C., Ahn, H.C. and Kim, J.K., The Audience Behavior-based Emotion Prediction Model for Personalized Service. *Journal of Korea Intelligent Information Systems Society*, 19, 2 (2013), 73-85.

[13] Kim, H.S. and Shin, H.J., Electricity Price Prediction Based on Semi-Supervised Learning and Neural Network Algorithms. *Journal of the Korean Institute of Industrial Engineers*, 39, 1 (2013), 30-45.

[14] Kim, J.B. and Kim, Y.I., The Influence of Weight Adjusting Method and the Number of Hidden Layer's Node on Neural Network's Performance. *Journal of Korean Association of Information Systems*, 9, 1 (2000), 27-44.

[15] Jang, I.D. and Wee, S.M., The Analysis Telecommunication Service Market with Data Mining. *Journal of Korea Information Science Society*, 28, 2 (2001), 1-3.

[16] https://ko.wikipedia.org/wiki/%EA%B2%B0%EC%A0%95%EA%B3%84%EC%88%98

[17] Lee, H.W., Kim, N.R., Lee, J.H., Deep Neural Network Self-training Based on Unsupervised Learning and Dropout. *International Journal of Fuzzy Logic and Intelligent Systems*, 17, 1 (2017), 1-9.