Deep learning application using neural network classification for cyberspace dataset with backpropagation algorithm and log-linear models

Baiq Siska Febriani Astuti[a,1,*], Tuti Purwaningsih [b,2]

[a] Graduate Student of Statistics Department, Institut Teknologi Sepuluh Nopember
[b] Lecturer of Statistics Department, Universitas Islam Indonesia
[1] baiqsiskafebriani@gmail.com *; [2] tuti.purwaningsih@uii.ac.id
* corresponding author

ABSTRACT

This study aims to classify bloggers in the Kohkiloye and Boyer Ahmad Province in Iran where causes of users tend cyberspace on there. The database was got from UCI Machine Learning Repository. There are 100th object and 6th variables. All of the variables were Professional Bloggers, Political and Social Space (LPSS), Local Media Turnover (LMT), Political Caprice, Topics, and Degree. This study has using Artificial Neural Network with backpropagation algorithm and Log-linear models for classify Bloggers (Cyber Space). We classify blogger to two groups: professional bloggers and seasonal (temporary) bloggers. The result of this study is Neural network with backpropagation algorithm has been shown to be useful tool for prediction, especially for this case. From this study, we can see on the result that miss-classification with backpropagation algorithm less than using Log-Linear Models.

Keywords:
Artificial Neural Network
Backpropagation Algorithm
Log-Linear Models
Classification.

I. Introduction

Today is the era of technology like internet which is everyone have internet on their smartphone, notebook, etc. Internet is probably the most valuable asset for billions people in around the world. It plays key roles in our social lives, governance, education, economy and became inspirable part of freedom speech. All of people in the world have access internet especially in Iran where is include into Persian cyberspace.

According to the research, conducted by opponent initiative, internet usage in Iran is increasing at sharp rate and it reaches about 48% annually, thus transforming Iran into a leader compared to other middle east countries. Nowadays approximately 37 million citizen of Iran have access to the internet and while internet penetration rate is growing. Iran's internet culture is rapidly developing [5].

Cyberspace is a part of internet which is like Blog. Blog is a discussion or informational websites published on the world wide web consisting of discrete, often informal diary-style text entries [7]. It has entries about specific issues to informing the others opinion about given issues. There is much topics became headline in the blog suitable what users want to exposed. Topics sometimes include brief about historical, healthy, political, government, and another social issues.

Blog has positive and negative impact. The causes of bloggers tendency to main of topics in that blog. Recognizing the causes of bloggers and the main of parameters of their approach are among major issues which the macro planning of the countries are determined based on these modern technologies and their users and provides vital data for planners and governments[6].

Due to the regional context of the province, population, social and political structure, the users of cyberspace in Kohkiloye and Boyer Ahmad in Iran are recognized to each other and their behavior in cyberspace will be gathered in a valuable and useful database by cognitive and research methods and through collecting data and analyzing information databases, social networks, blogs, websites, and virtual communities which are used by them[6].

It will help us to survey the effect of the users tends to cyberspace, use neural network to distinguish categorize into two groups which is professional and seasonal blogger. So that, the parameters has

been collected from survey like LPSS, LMT, topic, caprice, and degree which thus explains about characteristic of user cyberspace. Finally, from thus the researcher want to classify about bloggers users to two groups using Neural Network with backpropagation algorithm and Log-Linear models using program R version 3.4.1

II. Artificial neural network

A. Neural Networks

Trentin and Freno said that Neural networks are flexible, nonparametric modeling tools. They can perform any complex function mapping with desired accuracy. An ANN is typically composed of several layers of many computing elements called nodes. Each node receives an input signal from other nodes or external inputs and after processing the signals locally through a transfer function, it outputs a transformed signal to other nodes or final result [1]. A single-input neuron is shown in Figure 1. The scalar input p is multiplied by the scalar weight w to wp form, one of the terms that is sent to the summer. The other input, 1, is multiplied by a bias b and then passed to the summer. The summer output often referred to as the net input, goes into a transfer function, which produces the scalar neuron output. (Some authors use the term "activation function" rather than transfer function and "offset" rather than bias) [2].
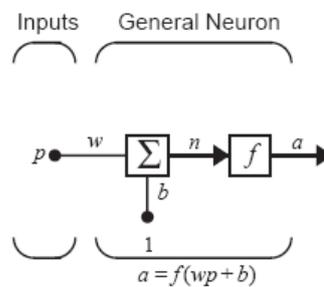
Fig. 1. Single-input Neuron

Now consider a network with several layers. Each layer has its own weight matrix w, its own bias vector b, a net input vector n, and an output vector. We need to introduce some additional notation to distinguish between these layers. We will use superscripts to identify the layers. Specifically, we append the number of the layer as a superscript to the names for each of these variables. Thus, the weight matrix for the first layer is written as w^1, and the weight matrix for the second layer is written as w^2. This notation is used in the three-layer network shown in Figure 2 [1].
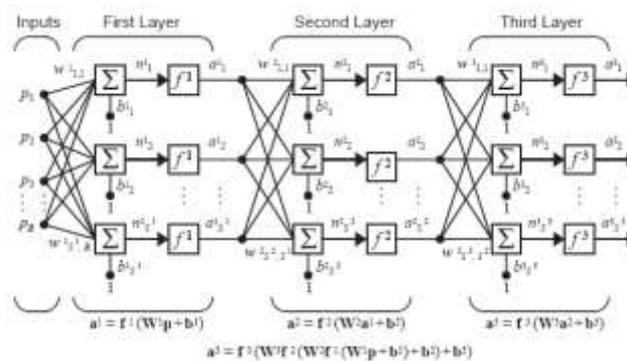
Fig. 2. Three Layer Network

B. Backpropagation Algorithm

The backpropagation algorithm looks for the minimum of the error function in weight space using the method of gradient descent. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. Since this method requires computation of the gradient of the error function at each iteration step, we must guarantee the continuity and differentiability of the error function. Obviously, we have to use a kind of activation function other than the step function used in perceptrons, because the composite function produced by interconnected

perceptrons is discontinuous, and therefore the error function too. One of the more popular activation functions for backpropagation networks is the sigmoid, a real $s\_c=R\to(0,1)$ defined by the expreesion[3].

$$s_c = \frac{1}{1+e^{-cx}}$$ (1)

The constant c can be selected arbitrarily and its reciprocal 1/c is called the temperature parameter in stochastic neural networks. The shape of the sigmoid changes according to the value of c, as can be seen in Figure 3. The graph shows the shape of the sigmoid for c=1, c= 2 and c=3. Higher values of c bring the shape of the sigmoid closer to that of the step function and in the limit c→1 the sigmoid converges to a step function at the origin. In order to simplify all expressions derived in this chapter we set c=1, but after going through this material the reader should be able to generalize all the expressions for a variable c. In the following we call the sigmoid s_1 (x) just s(x) [3].
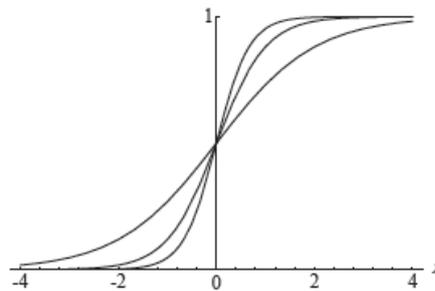
Fig. 3. three sigmoid (for c=1, c=2, and c=3) [3]

The derivative of the sigmoid with respect to x as follow as

$$\frac{d}{dx}s(x) = \frac{e^{-x}}{(1+e^{-x})^2} = s(x)\big(1-s(x)\big)$$ (2)

An alternative to the sigmoid is the symmetrical sigmoid S(x) defined as

$$(x) = 2s(x) - 1 = \frac{1-e^{-x}}{1+e^{-x}}$$ (3)

C. Log-Linear Models

Log-linear models play a considerable role in statistics and machine learning, special classes are often known through different names depending on the application domains and on various details: exponential families, maximum entropy models conditional random fields, binomial and multinomial logistic regression. A conditional log-linear model which we could a conditional exponential family, as model form [4]:

$$p(x|K,\alpha) = \frac{1}{z(K,\alpha)} b(K,\alpha)exp\big(\alpha^T \emptyset(K,x)\big)$$ (4)

Where
- x is a variable in a set V.
- K is the conditional variable
- α is parameter vector i $\mathbb{R}^d$
- $\emptyset$ is future function $(K,x) \to \mathbb{R}^d$; note that we sometimes write $(x;K)$ or $(K;x)$ instead of $(K;x)$ to stress the fact that K is a condition.
- b is a nonnegative function $(K,x) \to \mathbb{R}^+$; we will call it the background function of the model
- $z(K,\alpha)$, called the partition function, is a normalization factor:
    $$z(K,\alpha) = \sum_x b(K,x)exp\big(\alpha^T \emptyset(K,x)\big)$$ (5)
    When the context is unambiguous, we will sometimes leave the condition K as well as the parameter vector a implicit, and also simply write Z instead of $z(K,\alpha)$, thus we will write[4]:
    $$p(x) = \frac{1}{z}b(x)exp\big(\alpha^T \emptyset(x)\big)$$ (6)
    Or

$$p(x) \propto \frac{1}{z} b(x) exp\left(\alpha^T \emptyset(x)\right) \tag{7}$$

III. Design of Study

The dataset has been collected at website http://archive.ics.uci.edu/ml/datasets/BLOGGER. Three are 6th variables which is Political Social Space (LPSS), Local Media Turnover (LMT), Political Caprice, Topics, and Degree. The number of objects in this study are 100th. The characteristic of the dataset shown on table 1.

Table 1. characteristic of the datasets

Variables	Specification	
Bloggers	0 (No / Seasonal Bloggers) 1 (Yes/ Professional Bloggers)	Output
Political Social Space (LPSS)	0 (No) 1 (Yes)	
Local Media Turnover (LMT)	0 (No) 1 (Yes)	
Political Caprice	0 (Left) 1 (Middle) 2 (Right)	
Topics	0 (Impression) 1 (Political) 2 (Tourism) 3 (News) 4 (Scientific)	Inputs
Degree	0 (Low) 1 (Medium) 2 (High)	

In this study, researcher using Artificial Neural Network to solve problem of classification bloggers which methods is Neural Network using backpropagation algorithm and Log-Linear Models. The purpose of this study is firstly what is the appropriate neural network architecture for this particular data set, secondly what is the best method to solved in this case.

IV. Results and Discussion

The database was got from http://archive.ics.uci.edu/ml/datasets/BLOGGER. There are 100th object and 6th variables. All of the variables were bloggers (cyberspace), Political and Social Space (LPSS), Local Media Turnover (LMT), Political Caprice, Topics, and Degree. This study has using Artificial Neural Network with backpropagation algorithm and Log-linear for classify Professional Bloggers. We classify blogger to two groups: professional bloggers and seasonal (temporary) bloggers. In this case, we assume that the Factors of bloggers are Political and Social Space (LPSS), Local Media Turnover (LMT), Political Caprice, Topics, and Degree.

A. Classification using Neural Network with Backpropagation Algorithm

It is necessary to know how many neurons and hidden layer are required to achieve the optimal results. So, in the first experiment the numbers of neurons in hidden layer are investigated and the result show in fig 4. In this case, we use perceptron which has two hidden layer. The both of hidden layer has four and two activations, show in figure 4. That perception has five input which is lpss, lmt, topic, caprice and degree which one output bloggers (cyberspace) case.

Figure 4 was classification result using backpropagation algorithm, optimum error for this case is 2.588766 for 13579 steps. The mean of value 13579 steps which have optimum solution for this case as large as 13579.

Error: 2.588776   Steps: 13579
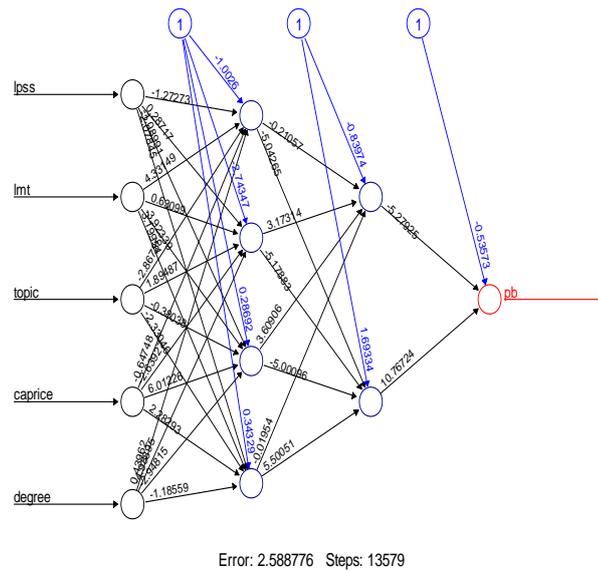
Fig. 4. Perceptron with backpropagation algorithm

Figure 5 show about capability of every variables influence bloggers (cyberspace) case. Figure 5 explains that bloggers (cyberspace) case have soo much variability, that's why we can say variables of lpss, lmt, topic, caprice and degree are effective for influence bloggers (cyber space) case from this plot at figure 5, we can see the values of bloggers (cyberspace) case is very variability for every variables which is value from -2.5 until 2.5.
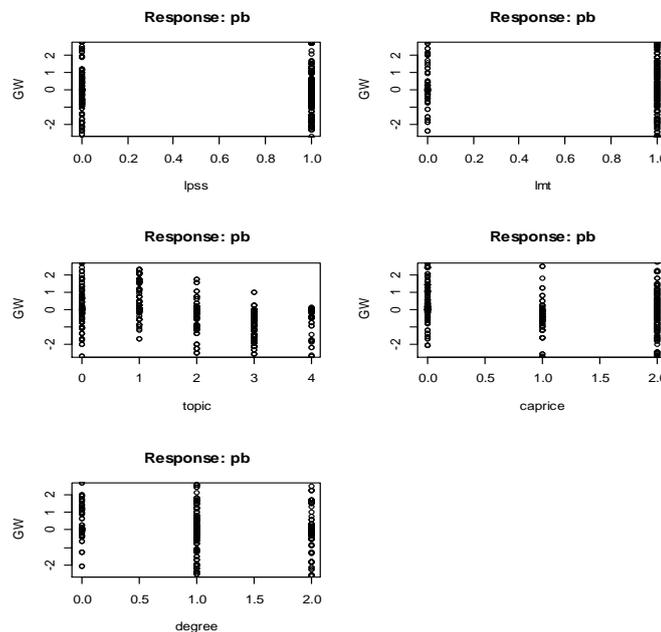
Fig. 5. capability of every variable for influence bloggers (cyberspace) case

B. *Identify the Headings Classification with Log-Linear Models using Neural Networks*

After we calculated the dataset using log-linear with packages "nnet" which are got coefficient for every parameters as given in table 2. From that table, we can see the values of coefficient caprice, topic and lpss are minus. The other way, the values of coefficient degree and lmt are positive. Thus made caprice, topic and lpss are increase so value of classification bloggers must be small. The other way for variables of degree and lmt.

Table 2. Coefficient for every parameters with log-linear using "nnet"

	(Intercept)	Degree	Caprice	topic	lmt	lpss
	0.8764	0.7951	-0.737	-0.231	0.370154	-0.445
	AIC= 118.117			**Residual deviace=106.117**		

It has residual device as large as 106.117 and the value of Akaike information criterion (AIC) is 118.117. The measure of AIC in the model is model has be better if AIC is small and the other way. Figure 6 shown about capability log-linear models predict classify blogger (cyberspace) case. Thus showing that predictions and actual values almost the same.
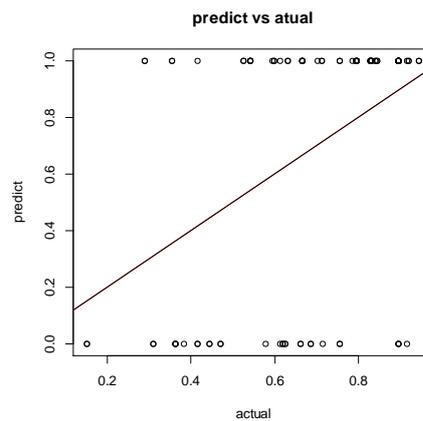
Fig. 6. plot between value of prediction and actual dataset for blogger.

C. Prediction Professional Bloggers

Cross-validation Result on the predictive performance of neural network with backpropagation algorithm are given in table 3. We assume that every of LPSS, LMT, Topic, Caprice, Degree are yes, no, political, middle and medium which result of prediction is Seasonal Bloggers like given in table 3. You can see obvious in table 3.

Table 3. Prediction classification bloggers (cyberspace) using backpropagation algorithm with R program

LPSS	LMT	TOPIC	CAPRICE	DEGREE	PREDICTION
No	No	Impression	Left	Low	Professional Bloggers (Yes)
Yes	No	Political	Middle	Medium	Seasonal Bloggers (No)
Yes	Yes	Political	Middle	Medium	Professional Bloggers (Yes)
Yes	Yes	Political	Right	Medium	Seasonal Bloggers (No)
Yes	Yes	Political	Right	High	Professional Bloggers (Yes)
Yes	Yes	Tourism	Right	High	Professional Bloggers (Yes)
No	Yes	Tourism	Right	High	Professional Bloggers (Yes)
No	Yes	Tourism	Middle	Low	Seasonal Bloggers (No)
Yes	No	News	Middle	Low	Seasonal Bloggers (No)

Prediction classification bloggers (cyberspace) using Log-Linear Models Neural Network that given at above table 4. From table 4, we can see that LPSS has yes, LMT has no, Topic has political,

Political Caprice has middle and Degree has medium so prediction must be professional bloggers. That is obviously another predictions.

Table 4. Prediction classification bloggers (cyberspace) using Log-Linear Models Neural Network with R program

LPSS	LMT	TOPIC	CAPRICE	DEGREE	PREDICTION
No	No	Impression	left	Low	Professional Bloggers (Yes)
Yes	No	Political	Middle	Medium	Professional Bloggers (Yes)
Yes	Yes	Political	Middle	Medium	Professional Bloggers (Yes)
Yes	Yes	Political	Right	Medium	Seasonal Bloggers (No)
Yes	Yes	Political	Right	High	Professional Bloggers (Yes)
Yes	Yes	Tourism	Right	High	Professional Bloggers (Yes)
No	Yes	Tourism	Right	High	Professional Bloggers (Yes)
No	Yes	Tourism	Middle	Low	Professional Bloggers (Yes)
Yes	No	News	Middle	Low	Seasonal Bloggers (No)

Finally, we can see the different prediction between using backpropagation algorithm and log linear models at above table 3 and table 4. The prediction using backpropagation algorithm and log linear are distinguish on the second observation and the eighth observation which is obvious prediction on above table.

D. Misclassification

In this case, if we use backpropagation for classify which have misclassification as large as 0.1 and if we use log-linear models for classification which have 0.22. Table 5 show result of misclassification for both of methods. Finally, for this case, we can choice backpropagation algorithm to classify bloggers (cyberspace). Causes misclassification backpropagation algorithm less than using Multinomial Log-Linear models.

Table 5. misclassification with backpropagation algorithm and Log-Linear models

	Backpropagation Algorithm		Log-Linear Models	
Group	0	1	0	1
0	25	7	15	17
1	1	67	5	63

V. Conclusion

The researcher using Neural Network with backpropagation algorithm and log-linear models to solve problem classification on this case. There are classify bloggers into two groups which is professional blogger and seasonal bloggers. Neural network with backpropagation algorithm has been shown to be useful tool for prediction, especially for this case. From this study, we can see on the result that misclassification with backpropagation algorithm less than using Log-Linear Models that is 0.1 while misclassification using Log-Linear Models as large as 0.22. And finally, the ANN with backpropagation algorithm has optimum error as large as 2.588766 and optimum solution as large as 13579 which architecture of perceptron has five inputs, two hidden layers, and one output.

References

[1] W. Sibanda and P. Pretorius, "Novel Application of Multi-Layer Perceptrons (MLP) Neural Networks to Model HIV in South Africa using Seroprevalence Data from Antenatal Clinics," 2011.

[2] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural network design*. Martin Hagan, 2014.

[3] J. Feldman and R. Rojas, *Neural Networks: A Systematic Introduction*. Springer Berlin Heidelberg, 1996.

[4] M. Dymetman and C. Xiao, "Log-Linear RNNs: Towards Recurrent Neural Networks with Flexible Prior Knowledge," pp. 1–28, 2016.

[5] Unknown.   Cyberspace   and   Cyber   Warfare   Capabilities   of   Iran. Available:https://securityzap.com/cyberspace-of-iran/ (Accessed at Nov, 1st 2017). (URL)

[6] Gharehchopogh. F.S and Khaze. S.R, "Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study", International Journal of Computer Applications. Volume 47– No.18 (June. 2012) 0975 – 888.

[7] Wikipedia. Blog. Available:https://en.m.wikipedia.org/wiki/Blog (Accessed at Nov, 1st 2017). (URL)

[8] Blogger   Data   Set.   Machine   Learning   Repository Available:http://archive.ics.uci.edu/ml/datasets/BLOGGER.