International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 1, February 2018, pp. 333~343
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i1.pp333-343

An Influence of Measurement Scale of Predictor Variable on Logistic Regression Modeling and Learning Vector Quntization Modeling for Object Classification

Waego Hadi Nugroho[1], Samingun Handoyo[2], Yusnita Julyarni Akri[3]
[1,2]Departement of Statistics, Faculty of Sciences, Universitas Brawijaya Malang, Indonesia
[3]Department of Midwifery Educators, Tribhuwana Tunggadewi University, Malang, Indonesia

Article Info

Article history:

Received Sep 20, 2017
Revised Jan 2, 2018
Accepted Jan 16, 2018

Keyword:

Logistic Regression
LVQ
Model performance
Object classification
Predictor variable scale

ABSTRACT

Much real world decision making is based on binary categories of information that agree or disagree, accept or reject, succeed or fail and so on. Information of this category is the output of a classification method that is the domain of statistical field studies (eg Logistic Regression method) and machine learning (eg Learning Vector Quantization (LVQ)). The input argument of a classification method has a very crucial role to the resulting output condition. This paper investigated the influence of various types of input data measurement (interval, ratio, and nominal) to the performance of logistic regression method and LVQ in classifying an object. Logistic regression modeling is done in several stages until a model that meets the suitability model test is obtained. Modeling on LVQ was tested on several codebook sizes and selected the most optimal LVQ model. The best model of each method compared to its performance on object classification based on Hit Ratio indicator. In logistic regression model obtained 2 models that meet the model suitability test is a model with predictive variables scaled interval and nominal, while in LVQ modeling obtained 3 pieces of the most optimal model with a different codebook. In the data with interval-scale predictor variable, the performance of both methods is the same. The performance of both models is just as bad when the data have the predictor variables of the nominal scale. In the data with predictor variable has ratio scale, the LVQ method able to produce moderate enough performance, while on logistic regression modeling is not obtained the model that meet model suitability test. Thus if the input dataset has interval or ratio-scale predictor variables than it is preferable to use the LVQ method for modeling the object classification.

Corresponding Author:

Waego Hadi Nugroho,
Departement of Statistics,
Faculty of Sciences,
Universitas Brawijaya,
Jln. Veteran Malang 65145, East Java Indonesia.
Email: whn@ub.ac.id

1. INTRODUCTION

The input data for the classification or categorization of an object can be either an image or an attribute. Before being used as input, the image must go through a preprocessing stage called feature extraction. There are two processes at the phase of features extraction; Texture feature extraction and shape feature extraction [1], [2]. The method of categorizing or classifying an object based on the features or attributes attached to the object is a study in the field of statistics and machine learning. Applied from the method is very wide in various aspects of life that come from the field of exact, engineering, and social. The

most crucial thing is the information resulting from the classification used as the basis for decision-making or policy. Although the attributes of the object are very wide in scope, they are either tangible or intangible, observable or unobservable. However, in order to apply a method of classification, it is necessary to characterize the objects arranged in a particular data structure. The most commonly used data structure as an input argument of a classification method is an input-output pair.

In the statistical field, the data in the input-output pair format should be considered as a causality where the set of observed values in the input attribute will determine the observed value of the output attribute [3]. The attributes of an object that has a special property that has a unique or single value on the object are known by the term variable. When an object is observed on the basis of 5 variables, it will get 5 observation values associated with the object. These types of observational values have 4 types of measurement scales: nominal, ordinal, interval, and ratio. This type of measurement scale in statistics will greatly influence the selection of the most suitable analytical methods. Suppose that if the output variables (variables affected by the input variables) are nominal or ordinal, then the statistical modeling for the classification of objects is logistic regression [4], [5]. On the other hand, the machine learning method does not require a causality between the input-output variables and also does not concern the type of measurement scale in the output variable [6], [7].

Hand and Henley [8] have reviewed the methods used in object classification. They concluded that the classification methods which are easy to understand (such as regression, nearest neighbour and tree-based methods) are much more appealing, both to users and to clients, than are methods which are essentially black boxes (such as Artificial Neural Network). They also permit more ready explanations of the sort of reasons why the methods have reached their decisions. Meanwhile Dreiseitl and Ohno-Machado [9] sampled 72 papers comparing both logistic regression and neural network models on medical data sets. They analyzed these papers with respect to several criteria, such as the size of data sets, model parameter, selection scheme, and performance measure used in reporting model results. They said that where performance was compared statistically, there was a 5:2 ratio of cases in which it was not significantly better to use neural networks.

Performance improvement of logistic regression model on microarray data with the Bayesian approach to gene selection and classification using the logistic regression model. The method can effectively identify important genes consistent with the known biological findings while the accuracy of the classification is also high [10]. In addition, the performance comparison between ANN and logistic regression in various fields are done by Felicisimo, et al [11] did Mapping landslide susceptibility, M. Shafiee, et al [12] did Forecasting Stock Returns in Iran Stock Exchange, and Kamley S, et al [13] did Forecasting of Share Market. From various studies, ANN method used is back propagation or support vector machine. Both methods are widely used because it has a topology and learning methods that are easy to understand. On the other hand, there is also ANN method known as Learning Vector Quantization (LVQ) which is still rarely found applied, because this method has a competitive layer that works using the principle of the self-organizing map [14], [15] so it has a structure and learning method that is difficult to understood. LVQ is a classification method in which each unit of output represents a class that can be used for grouping where the number of target groups or classes is pre-determined.

Based on the above exposure, this paper will examine the implementation of logistic regression and LVQ network for object classification in three datasets with input variables (predictors) with different measurement scales, respectively are intervals, ratios and nominal for data 1, data 2, and data 3. In the logistic regression modeling is done parameter estimation stage, testing the model parameters, then test the goodness of fit, finally obtained a suitable model. In LVQ network modeling the 4 codebooks are tested ie 2, 10, 30, and 50. The best models produced by both logistic regression and LVQ networks will be evaluated for classification using Hit Ratio [16], ie the proportion of sample observations that can be classified by the classification model appropriately. Implementation of both methods using software R.

2. LITERATURE REVIEW

2.1. Binary Logistic Regression Analysis

Binary logistic regression is a logistic regression with response variables that are categorical values of binary or dichotomous. The variable response of Bernoulli's Y distributes with the following probability functions [3]:

$$f(y_i) = \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i};\ y_i = 0,1 \quad (1)$$

The logistic regression model :

$$\pi(x_i) = \frac{e^{(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})}}{1 + e^{(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})}} \quad \begin{matrix} i=1, \ 2, \ \ldots, n \\ j=1, \ 2, \ \ldots, p \end{matrix} \tag{2}$$

Where: n=Number of observations
p=Number of predictor variables
β_0=Intercept
β_j=Logistic regression coefficient from the j th jth predictor variable
x_{ji}=The value of the j-th predictor variable on the i-th observation.
While the logit form is:

$$g(x_i) = \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right) \tag{3}$$

In logistic regression, the conditional distribution pattern of the response variable is
Y=π (x_i) + ε, which has 2 types of error:
Y=1 then ε=1-π (x) with probability π (x)
Y=0 then ε=-π (x) with a chance of 1-π (x)

So the error distribution has the mean equal to zero, variance {π (x) (1-π (x))} and follows the Binomial distribution.

2.1.1. Estimation of Model Parameters

The method to estimate the logistic regression model parameters is Maximum Likelihood Estimation (MLE). The model parameter is estimated from the $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ vector, the value $\boldsymbol{\beta}^T$ is obtained by maximizing the likelihood function (L (β)) through derivation of its parameters. The likelihood function is a joint probability function of the variables x_i and y_i. The probability distribution function for each (x_i,y_i), is

$$f(x_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \ , \ i = 1,2, \ldots, n \tag{4}$$

where, $\pi(x_i) = \frac{exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})}{\left[1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})\right]}$

According to Hosmer and Lemeshow [4], if inter-observations are assumed to be independent, the likelihood function is the multiplication of each probability distribution in Equation (4).

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \\ \ell(\boldsymbol{\beta}) &= \ell n \left[L(\boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{n} \left[y_i \, \ell n \left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) + \ell n\big(1 - \pi(x_i)\big)\right] \\ &= \sum_{i=1}^{n} \left[y_i\big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\big) + \ell n\big(1 + exp\big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\big)\big)^{-1}\right] \\ &= \sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right) - \sum_{i=1}^{n} \ell n \left(1 + exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right)\right) \end{aligned}$$

Max ℓn likelihood is obtained by derivating $\ell(\boldsymbol{\beta})$ to β and equating it with zero.

$$\begin{aligned} \frac{\partial \, \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} y_j x_{ji} - \sum_{i=1}^{n} x_{ji} \left[\frac{exp\big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\big)}{1 + exp\big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\big)}\right] \\ 0 &= \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} x_{ji} \ \pi(x_i) \\ 0 &= \sum_{i=1}^{n} x_{ji} \left(y_i - \pi(x_i)\right) \end{aligned} \tag{5}$$

Since Equation (5) is non-linear, the solution of this equation becomes difficult to resolve analytically, requiring an iterative solution such as the Newton-Raphson method [5].

2.1.2. Parameter Significance Testing

a. Simultaneous Testing

The simultaneous test is performed to examine the role of each predictor variable in the model simultaneously. Statistical hypotheses and test statistics are as follows [3]:

Hypothesis: H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus H_1: at least one $\beta_j \neq 0$;

$$G = -2\ln\left(\frac{L_0(\boldsymbol{\beta})}{L_p(\boldsymbol{\beta})}\right) = -2\left[\ln L_0(\boldsymbol{\beta}) - \ln L_p(\boldsymbol{\beta})\right] \quad (6)$$

where: L_0: $loglikelihood$ with β_j; j=1,2,…,p

L_p: $loglikelihood$ without β_j; j=1,2,…,p

The G value is compared with the statistic $\mathcal{X}^2_{(db,0.05)}$ with degrees of freedom corresponding to the estimated number of parameters. H_0 will be accepted if p-value is greater than the probability of doing type I error of α.

b. Partial Testing

Partial testing is used to test the effect of each parameter β_j on the model individually or separately with regard to other parameters. The partial test results will show whether a predictor variable is eligible to enter the model or not. If hypothesi test yields β_j significant, then an β_j enter in model. Hypothesis: H_0: β_j=0 versus H_1: $\beta_j \neq 0$.

Wald statistic test,

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

The test statistic used is the Wald test, reject H_0 if the value of $|W_j| > Z_{(1-\frac{\alpha}{2})}$ or p-value <0.05, so it can be concluded there is influence between predictor variables with response variables [4]. In addition to following the normal distribution, the Wald statistical test squared $(W)^2$ will follow the chi-square $\mathcal{X}^2$ with the degree of freedom one [3].

2.1.3. Goodness of Fit Test

Model fit test on logistic regression, can use test statistic called goodness of fit test. This test statistic is used to find out how big the effectiveness of the model formed in explaining the response variable, so the model can represent the actual condition represented by the data used in the analysis. The hypothesis and test statistic are as follows [3]:

Hypothesis: H_0: The model appropriate vs H_1: The model is not appropriate.

The test statistic used is χ^2 $Pearson$ ie $\chi^2 = \sum_{i=1}^{n} e_i^2$. H_0 is rejected if $\chi^2 > \chi^2_{(\alpha,(n-p-1))}$.

2.2. Artificial Neural Network with Competitive Layer

Artificial neural networks (ANN) with competitive layers have three layers: the input layer, the hidden layer, and the output layer. In this case the competitive layer lies in the hidden layer. Neurons in networks with competitive layers compete for active rights. One of the competitive layer network model is LVQ. There are two learning methods in ANN namely supervised learning and unsupervised learning. In Supervised learning, every pattern given as input for ANN, has been known output. The difference between the ANN output and the desired output (target) is called an error. This error quantity is used to correct ANN weight so that ANN can produce output as close as possible to known target pattern. ANN learning algorithm using this method is Hebbian, Perceptron, Adaline, Boltzman, Hopfield, Backpropagation, and LVQ.

One of the aims of ANN modeling is for classification. According Fausett [6], the basis of classification on ANN is to use the optimal weight of the learning process. The weights are the amounts or values that exist on the connection between neurons that transfer data from one layer to another, which serves to regulate the network so as to produce the desired output. In addition to having advantages that do not have to meet the classical assumptions and varian homogenity of errors, ANN also has a weakness that takes longer training time to perform calculations in the formation of models [13].

LVQ is a classification method in which each output unit presents a class with a specified target class. LVQ uses a supervised competitive learning algorithm version of the Kohonen Self-Organizing Map (SOM) algorithm. LVQ network architecture according to Kaski and Kohonen [15] can be seen in Figure 1.

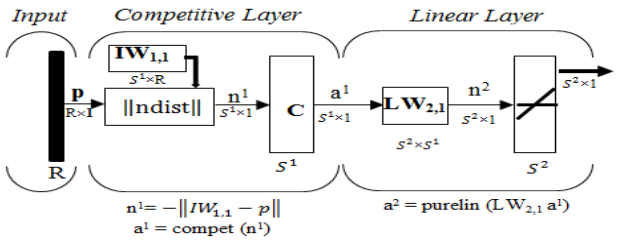

Figure 1. The architecture LVQ network

where: R=number of elements in the input vector
S^1=the number of competitive neurons
S^2=number of linear neurons

Based on Figure 1., the LVQ network consists of three layers: an input layer, a competitive layer, and an output layer. Learning in the competitive layer aims to classify each input vector to a hidden layer. While learning on the output layer aims at transforming the subclass on the competitive layer into the classification of the target class that has been established. Learning the competitive layer as a subclass and learning from the output layer as the target class. According to Putri (2012), there are two factors that influence the learning process in LVQ namely initial initialization and training rate.

Setting data as input on LVQ network must be in input-output pair format. In this case the output data will serve as a target in the learning process. Suppose in the format as follows:

$$\{s^{(q)}: t^{(q)}\},\ q=1, 2, ..., Q \tag{7}$$

Where: $s^{(q)}$=vector/input matrix
$t^{(q)}$=the output vector

LVQ consists of a competitive layer that includes a competitive subnet and a linear output layer. In a competitive layer, each neuron is assigned to a class. Different neurons in the competitive layer, it is possible to have the same class. Each class is then paired with one of the neurons in the output layer. Thus the number of neurons in the competitive layer, at least as much as the number of neurons in the linear output layer [14]. The relationship between the input vector and one of the weight vectors is measured by the Euclid distance. A subnet is used to find the smallest element in the input data.

$$n^{(1)} = \begin{bmatrix} \|x - \boldsymbol{W}_1^{(1)}\| \\ \|x - \boldsymbol{W}_2^{(1)}\| \\ \vdots \\ \|x - \boldsymbol{W}_Q^{(1)}\| \end{bmatrix} \tag{8}$$

An element given value 1 indicates that the input vector belongs to the intended class, and an element is assigned a value of 0 if the input vector is not included in the desired class. This can be represented by a subnet as a vector with the following vector functions:

$\mathbf{a}^{(1)}$=compet ($n^{(1)}$) (9)

Linear output layer of LVQ network, used to combine subclasses into a single class. This is done using the weight matrix $\boldsymbol{W}^{(2)}$, ie the weight matrix having elements:

$$w_{ij} = \begin{cases} 1, & if\ neuron\ i\ included\ in\ a\ subclass\ j \\ 0, & if\ neuron\ i\ excluded\ in\ adubclass\ j \end{cases} \tag{10}$$

In addition, the weight matrix $\boldsymbol{W}^{(1)}$, on the competitive layer must be trained using the Kohonen SOM rule as follows:

At each iteration, each training vector is entered into the network as input x and the Euclid distance from the input vector to each prototype vector (weighted matrix column) is calculated. Neuron j* wins the

competition if Euclid's distance between x and j* prototype vector is the smallest. The activation value $a^{(1)}$ is multiplied by n $\boldsymbol{W}^{(2)}$ in its right position to obtain input n $^{(2)}$. The output $a^{(2)}=n^{(2)}$, as long as the transfer function in the output neuron is an identity function. The $a^{(2)}$ also has only one value in the element k*, indicating that the input vector belongs to the class k*. Kohonen rules are used to fix the weights on the hidden layer. If x is correctly classified, the weight vector $w_{.j*}^{(1)}$ is the winner so that the hidden neuron is moved closer to x.

$$\Delta w_{.j*}^{(1)} = \alpha\left(x - w_{.j*}^{(1)}\right) \text{ if } \left(a_{k^*}^2 = t_{k^*} = 1\right) \quad (11)$$

But if x is classified incorrectly, it is obvious that the wrong hidden neurons win the competition. In this case, the weight is moved away from the x.

$$\Delta w_{.j*}^{(1)} = -\alpha\left(x - w_{.j*}^{(1)}\right) \text{ if } \left(a_{k^*}^2 = t_{k^*} \neq 1\right) \quad (12)$$

After the training, the final weights (w) will be used for the next simulation, test or classification [15].

2.3. Accuracy of Classification

According to Dianiati (2013), prior to classification, the data is divided into two datasets. The first part is the training dataset used to form the optimal model of artificial neural networks, while the second part is the testing dataset to test the optimal model obtained from the training dataset. Hair, *et al.*[5] explains the principle of sharing the most popular proportion of training-testing datasets is 50-50, but most researchers also use the 60-40 or 75-25 division principle, since there is no standard rule about dividing the dataset. The precision in classification can be determined by calculating the value of Hit Ratio, ie the proportion of observational samples that can be classified by the classification function [16]. The value of Hit Ratio can be calculated using the following formula:

$$Hit\ Ratio = \frac{number\ of\ objek\ classified\ accurately}{total\ sampel}\ X\ 100\% \quad (13)$$

The Hit Ratio value calculated according to the Equation (13) shows performance of classification function. A biger Hit Ratio value indicates a better classification method.

3. RESEARCH METHOD

3.1. Data Characteristics

The data used are secondary data obtained from three previous studies. The three datasets have different response and predictor variables. The explanation of the data used is shown in Table 1.

Table 1. Predictor and Response Variable of the Data

Dataset	Data sources	Predictor variables	Predictor scale	Categorical respon variable		record
Diet on toddlers	Sartika [17]	X_1=Carbohydrates X_2=Vegetables X_3=Side dish X_4=Fruit X_5=Milk	Interval	Bad diet Diet is enough	0 1	108
The granting of credit for seaweed business	Sunadji [18]	X_1=Experience X_2=Duration of education X_3=Employment Intensity X_4=Age X_5=Seaweed cleanliness level X_6==Seaweed Water Content	Rasio	Do not accept credit Accepting credit	0 1	138
Factors that influence the incidence of low birth weight infants (LBW)	Pandin [19]	X_1=Mother Age X_2=Parity X_3=Birth distance X_4=Anemia X_5=Nutrition Status X_6=Education	Nominal	Case of LBW Not a case of LBW	1 2	96

3.2. Research Stages

In this study, the first step is to divide each data into 2 parts, 70% for training and 30% for testing dataset, then continued logistic regression analysis and LVQ analysis. Binary logistic regression analysis was performed using software R. The steps in binary logistic regression analysis were

a. Check for the presence or absence of multicollinearity between independent variables X
b. Parameter estimation
c. Testing parameters simultaneously
d. Partial parameter test
e. The establishment of a logistic regression model
f. Check the suitability of the model
g. Apply binary logistic regression model obtained from training data to testing data
h. Forming a table of precision classification of training and testing models.

While the analysis of LVQ is done using **package class** on software R. The steps in LVQ analysis are:

a. Forming an input matrix on training data and forming a vector or classification factor for training data
b. Initialization weights of the LVQ network
c. Determine the learning rate (α)
d. Renew weight on the competitive layer to obtain optimum weight
e. Form an optimal architecture
f. Re-classified the testing dataset based on the best architecture of the LVQ method formed from the training dataset.

After the completion of the LVQ analysis done then calculated indicator of classification accuracy that is Hit Ratio and next to compare value of Hit Ratio from both methods.

4. RESULTS AND ANALYSIS

4.1. Logistic Regression Modeling

The process of data analysis begins with multicolinearity testing among the predictor variables on each dataset. In the three datasets used in this study, there is no multicollinearity that is indicated by the VIF value greater than 10. The modeling process in logistic regression of the three datasets can be continued by estimating the model parameters of each dataset. Below is the parameter model estimator obtained by the maximum likelihood method, presented in Table 2.

Table 2. The Parameter Estimation Results on All Datasets

Predictors	Values of estimates parameter		
	Dataset 1	Dataset 2	Dataset 3
X_1	5.760	0.234	0.619
X_2	2.184	0.002	-0.120
X_3	2.279	0.035	-0.191
X_4	4.415	-0.089	-1.382
X_5	4.955	0.370	1.092
X_6		-2.976	2.332
Constant	-100.753	45.289	2.153

The full model formed for dataset 1, dataset 2, and dataset 3 respectively are

$$g(x) = -100.753 + 5.76\,X_1 + 2.184\,X_2 + 2.279X_3 + 4.415\,X_4 + 4.955\,X_5\,.$$
$$g(x) = 45.289 + 0.234\,X_1 + 0.002\,X_2 + 0.035\,X_3 - 0.089\,X_4 + 0.370\,X_5 - 2.976\,X_6$$
$$g(x) = 2.153 - 0.619\,X_{11} - 0.120\,X_{21} - 0.191\,X_{31} - 1.382\,X_{41} + 1.092\,X_{51} - 2.332\,X_{61}.$$

Tests on parameter estimators are simultaneously performed to determine the effect of predictor variables contained in the logistic regression model to the response variable as a whole or together. This test is based on the ratio test statistic (G) likelihood ratio test. Here is the hypothesis used:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus H_1: $at\ least\ \beta_j \neq 0$ *j=1, 2, ..., p*

The test results against parameter estimators simultaneously on the datasets 1, 2, and 3 as shown in Figure 3.

Table 3. Simultaneously Parameter Testing of All Datasets

Dataset	-2 log likelihood		Likelihood Ratio Test	p-value
	Model intersep	Model parameter		
1	-43.801	-9.750076	68.103	<0.000
2	-61.508	-7.105	108.805	0.000
3	-46.374	-37.854	17.039	0.009

Based on the testing of parameters simultaneously in Table 2. for dataset 1 it can be seen that the p-value of the likelihood ratio test is <0.000. The value is less than the level of significant α is 0.05, so it was decided to reject H_0 which means that carbohydrates, vegetables, side dishes, fruits, and milk together significantly affect the diet status of children under five. The result of statistical test on dataset 2 can be seen that p-value of likelihood ratio test is 0.000. The value is less than the significant level of α is 0.05, so it is decided to reject H_0 which means that experience, duration of education, labor outpour, age, level of seaweed cleanliness, and seaweed contents significantly affect farming credit to farmers seaweed. While the test results in data 3 can be seen that the p-value of the likelihood ratio test is 0.009. The value is less than α is 0.05, so it is decided to reject H_0 which means that maternal age, parity, gestational distance, anemia, nutritional status, and education together have a significant effect on the incidence of low birth weight babies.

To determine the predictor variables that significantly influence the response variables, it is necessary to test the significance of the parameters in each predictor variables using Wald (W_j) test statistic. The statistical hypothesis tested is H_0: $\beta_j=0$ versus H_1: $\beta_j \neq 0$. Wald test statistic is Chi-square distributed with one degree of freedom. Based on p-value on Wald test statistic, for the first data, it was found that X_2 (vegetable) and X_3 (side dish) variables did not significantly affect the classification of infant diet. In the second data, only predictor X_5 (level of seaweed cleanliness) has a significant effect on the determination of credit for seaweed farmers. As for the third data, it is known that the X_1 (maternal age), X_2 (parity), X_3 (birth distance), and X_5 (nutritional status) did not significantly affect the classification of low birth weight babies. The logistic regression model that is formed based on significant predictor variables are as sown in Figure 4:

Table 4. The Logistic Regression Model of All Datasets and Goodness of Fit Test

Dataset	The final model of logistic regression	*Chi-Square*	df	*p-value*
1	$g(x) = -61.479 + 5.138\,X_1 + 3.815\,X_4 + 2.893\,X_5$	2.897	8	0.941
2	$g(x) = -16.015 + 0.178X_5$	7.305	1	0.007
3	$g(x) = -2.429 + 1.442\,X_{41} + 2.264\,X_{61}$	0.518	2	0.772

Testing the suitability (goodness) model used to determine whether the resulting model is appropriate (feasible). The statistical hypothesis used in this test is: H_0: $\widehat{f_A} = f_A$ (observation frequency=expected frequency) versus H_1: $\widehat{f_A} \neq f_A$ (observation frequency ≠ expected frequency). Based on Table 3. p-value for the 1st and 3rd data has a value of more than 0.05 so the decision is to receive H_0 and it can be concluded that the binary logistic regression model generated is good (appropriate), the model has been sufficient to explain the first data (classification of infant diet), and the 3rd data (classification of low birth weight infants). For the 2nd data because p-value has a value less than 0.05 it can be concluded that the binary logistic regression model generated in the 2nd data has not been suitable or not enough to explain the data. Based on these results, statistically, logistic regression models of the 1st and 3rd data can be used for object classification and should be able to produce fairly good classification accuracy.

4.2. Learning Vector Quantization (LVQ) Optimum Model

To obtain an optimal LVQ network model, LVQ network modeling process is performed on a various number of neurons in a hidden layer called codebook. The amount of codebook used is 2, 10, 30, and 50. The size of the codebook will determine the weight matrix dimension to be calculated to obtain optimal LVQ network. Based on the dimension of this weight matrix, then the weights are randomly initialized. The optimal weight will be obtained through the training process by utilizing the training data input. After all the connecting weights between nodes in different layers of the LVQ network have been obtained, the network output can be obtained by inputting the input data into the network. If used as an input argument is training data then obtained the output of training data. Similarly, if used as an input argument is testing data then obtained the output of testing data. The value of Hit Ratio can be calculated from the output obtained from the network, either output of training or testing data. The initialized weights of the four codebooks for the first data are shown in Table 5.

Table 5. The initial weights between input and hidden layer of the first dataset

Size Codebook	Neuron Hidden	$W_1^{(1)}$	$W_2^{(1)}$	$W_3^{(1)}$	$W_4^{(1)}$	$W_5^{(1)}$
2	1	4.73	3.58	5.01	4.18	5.43
	2	6.70	4.76	4.33	5.00	5.24
10	1	5.34	2.62	4.42	3.36	3.78
	2	4.58	3.71	3.30	3.30	5.00
	...	...	...	...	...	...
	10	6.70	4.40	4.83	3.34	7.50
30	1	5.34	2.62	4.42	3.36	3.78
	2	4.73	3.04	2.80	2.34	3.68
	...	...	...	...	...	...
	30	6.10	4.14	5.49	5.28	5.00
50	1	4.73	3.58	5.01	4.18	5.43
	2	5.60	4.50	2.84	3.30	4.32
	...	...	...	...	...	...
	50	6.18	4.91	5.19	4.00	5.43

As explained in the previous session that data 1 has 5 input variables considered to have an effect on the response variable. In codebook=2, it must be initialized a 5x2 matrix whose elements are the connecting weights between the input layer and the hidden layer, and a 2x1-sized vector whose elements are the connecting weights of the hidden layer to the output layer, whereas in the codebook=50, The dimensions of the weighted matrices and vectors to be initialized are 5x50 and 50x1 which are the connecting weights between the input layer and the hidden layer, and the connecting weights of the hidden layer to the output layer. After the training process, finally we get the optimal weight which some of the elements are presented in Table 6 as follows:

Table 6. The Final Weights between Input and Hidden Layer of the First Dataset

Size Codebook	Neuron Hidden	$W_1^{(1)}$	$W_2^{(1)}$	$W_3^{(1)}$	$W_4^{(1)}$	$W_5^{(1)}$
2	1	4.889	3.827	4.562	3.949	4.954
	2	6.226	4.735	5.110	4.902	6.293
10	1	5.141	3.114	4.345	3.810	3.920
	2	4.912	3.760	3.036	3.047	4.457
	...	...	...	...	...	...
	10	5.939	4.835	4.285	3.892	7.170
30	1	5.340	2.620	4.420	3.360	3.780
	2	4.730	3.040	2.800	2.340	3.680
	...	...	...	...	...	...
	30	5.756	4.469	4.899	5.509	5.255
50	1	4.947	3.848	5.303	4.085	5.130
	2	5.600	4.500	2.840	3.300	4.320
	...	...	...	...	...	...
	50	6.306	5.136	5.327	4.076	5.739

Table 7. Hit Ratio for All Dataset of Various Codebook Size

Data	Codebook size	Hit Ratio	
		Training Data	Testing Data
1	2	84.21	84.4
	10	92.1	81.25
	30	93.4	78.12
	50	96.05	78.12
2	2	67.01	75.61
	10	80.41	65.85
	30	83.5	70.73
	50	91.75	73.17
3	2	61.2	37.93
	10	74.62	44.82
	30	76.11	55.17
	50	76.12	44.82

The final weight of LVQ that has been obtained from LVQ network learning process using training data is then used as network weight. Thus the LVQ Network already has the connecting weights of each node between the layers, so that the LVQ network is ready to be used as a model for object classification. Table 8

is an indicator of LVQ network performance that is Hit Ratio value from three datasets and various codebook.

Based on the value of Hit Ratio in Table 7 we get the best LVQ network model for dataset 1 and dataset 2 is the model with codebook=2, whereas, in dataset 3, the best model is codebook=30. The models are said to be the best models because they have the greatest Hit Ratio value in the data testing. If Table 7 is observed further, it can be said that the increase in the number of codebooks is also followed by the increase of Hit Ratio value in training data, but the incidence does not apply to the value of Hit ratio in data testing. This phenomenon is known as overfitting.

4.3. The Comparison of Accuracy Classification between Logistic Regression and LVQ

The performance of the two methods of classifying objects is compared by the value of Hit Ratio calculated on both training and testing datasets of all the best models. The results of the classification accuracy of all datasets are presented in Table 8 below:

Table 8. The Acuracy of the Best Model from Both Logistic Regression and LVQ

Data	Training		Testing	
	Logistic Regression	LVQ	Logistic Regression	LVQ
	90.8	84.21	84.4	84.4
	-	67.01	-	75.61
	65.7	76.11	55.17	55.17

Based on the percentage of accuracy of classification results in Table 8. it can be seen that in the second dataset (granting of seaweed farming), logistic regression method can not be calculated classification accuracy because the model obtained from the data does not meet the model fit test. This is supported by partial parameter test result only obtained one predictor variable (level of seaweed cleanliness) which have significant effect to response variable (credit approval decision to seaweed farmer). Keep in mind that the second dataset has predictor variables that are all scalable ratios. Different accuracy results are found in the LVQ model that is in the second dataset still obtained the value of Hit Ratio, both in the training and testing dataset of moderate enough size of 67% and 75% respectively.

Modeling on the first dataset is a case that ideally demonstrates that logistic regression model performs equally well compared to LVQ model based on Hit ratio on dataset testing=84%. Having studied more deeply to this logistic regression model, it turns out in the process of diagnostic examination of the error obtained the result that the error of this model is able to meet all the assumptions in the regression modeling. The assumptions are error independenly each others, error has a constant varian, and error has normal distribution. The most difficult thing to be met with regression analysis is the normality assumption of the distribution of error.

In the third dataset obtained the performance of both models are just as bad that is shown by the value of Hit ratio on dataset testing=55%. It should be noted carefully that the predictor variables are all nominal scale measurements. In the logistic regression model, only two categories of predictors (ie, anemia and low educated mothers) had a significant effect on low birth weight (LBW) infants. This implies that logistic regression modeling and LVQ network on predictive variables of nominal scale require more factors that influence the response variable. Although the LVQ model on the training dataset has Hit Ratio=76%, the model also remains unable to classify the test dataset satisfactorily.

5. CONCLUSION

The measurement scale of the predictor variable is very influential on the modeling and performance of the classification model, both logistic regression model, and LVQ model. In the interval-scale predictor variable, the best model of both methods produces an equally high accuracy of Hit Ratio=84%. In the nominal-scale predictor variable, the classification accuracy of the best model in both methods is similarly low: Hit Ratio=55%. While on Ratio-scale predictor variable, logistic regression modeling did not produce the best model, But the resulting LVQ model has an accuracy for fairly moderate object classification, ie Hit Ratio=75%.

REFERENCES

[1] N. Suciati, W.A. Pratomo and D. Purwitasari, *"Batik Motif Classification using Color-Texture-Based Feature Extraction and Backpropagation Neural Network"*, Proceeding IIAI 3rd International Conference on Advanced Applied Informatics, 2014, pp. 517–521.

[2] A.A. Kasim, R. Wardoyo and A. Harjoko, "Batik Classification with Artificial Neural Network Based on Texture-Shape Feature of Main Ornament", *I.J. Intelligent Systems and Applications*, vol. 6, pp. 55-65, 2017.

[3] Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York, 2002.

[4] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Second Edition, Willey-Inter science, Canada, 2000.

[5] J.F. Hair, et al., *Multivariate Data Analysis*, Seventh Edition, Prentice Hall International, New Jersey, 2010.

[6] I. Fausett, *Fundamentals of Neural Network, Architecture, Algoritms and Aplications*, Printice-Hall, London, 1994.

[7] R. Hecht-Nielsen, "Neurocomputing application", *Neural Computer*, pp.445-453, Springer-Verlag, Berlin, 1998.

[8] D.J. Hand, and W.E. Henley, "Statistical Classification methods in consumer credit scoring: a Review", *J.R. Statistics Society*, v ol.160, part.3, pp.521-542, 1997.

[9] Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", *Journal of Biomedical Informatics*, vol. 35, pp. 352–359, 2003.

[10] X.B. Zhoua, *et al*., "Cancer classification and prediction using logistic regression with Bayesian gene selection", *Journal of Biomedical Informatics*, vol. 37, pp. 249–259, 2004.

[11] A. Felicisimo, *et al.*, "Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods*", Landslides,* Springer-Verlag, 2012.

[12] M. Shafiee, *et al*., "Forecasting Stock Returns Using Support Vector Machine and Decision Tree: A Case Study in Iran Stock Exchange", *International Journal of Economy, Management and Social Sciences*, vol/issue: 2(9), pp. 746-751, 2013.

[13] S. Kamley, S. Jaloree and R.S. Thakur, "Performance Forecasting of Share Market using Machine Learning Techniques: A Review", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 6, No. 6, pp. 3196-3204, 2016.

[14] T. Kohonen, "Self-Organizing map", *Springer Series In Information Sciences*, Springer, Berlin, 1995.

[15] S. Kaski and T. Kohonen, "Exploratory data analysis by the selt-organizing map: Structures of welfare and poverty in the world", *Neural Networks in Financial Engineering*, World Scientific, Singapore, 1996.

[16] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Fifth Edition, Prentice Hall International Inc., New Jersey, 2007.

[17] M. Sartika, "Relationship of Diet and Intake of Nutrition in Short Children in Working Area of Kapuas Hilir District Health Center Kapuan Indonesia", Thesis, Universitas Brawijaya Malang, 2012.

[18] Sunadji, "Seaweed Cultivation Development Model in Kupang District, NTT Province, Indonesia (Policy Simulation with Household Economic Approach)", Thesis, Universitas Brawijaya Malang, 2012.

[19] P.K. Pandin, "Factors Influencing Low Birth Weight (LBW) Incidence in Jumpandang Baru Makassar ,Indonesia", Thesis, Universitas Brawijaya Malang, 2015.

BIOGRAPHIES OF AUTHORS

Waego Hadi Nugroho is a Professor and a researcher at Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya Malang. He is also a member of the Indonesian Mathematical Society. He received the B.E. degree in Agriculture from Universitas Brawijaya, Malang, Indonesia in 1976 and the Ph.D degree in Biometrics from University of Adelaide, Adelaide, Australia, in 1985. His research interests include Statistical Modeling, Sampling Tehnique and survey, and Design of Experiments.

Samingun Handoyo is a lecturer and a researcher at Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang. He is also a member of the Indonesian Mathematical Society. He received the B.S. degree in Mathematics from Universitas Brawijaya, Malang, Indonesia in 1997 and the M.Cs degree in Computer Science from Universitas Gadjah Mada, Yogyakarta, Indonesia in 2010. His research interests include statistical computing, Neural Network, Fuzzy System, and Partial Least Square path modeling using R Software.

Yusnita Julyarni Akri is a lecturer and researcher at Department of Midwifery Educators at Tribhuwana Tunggadewi University, Malang. She is also a member of the Indonesian Doctor Society. She received her Bachelor Degree in Medical Sciences from Wijaya Kusuma University, Surabaya, Indonesia in 2003, her M.D. from Wijaya Kusuma University, Surabaya, Indonesia in 2005 and Masters Degree in Health Sciences in 2013 from Sebelas Maret University, Surakarta, Indonesia. Her research interest includes health sciences, midwifery practice, community health, and traditional health practice.