Classification of Poverty Levels Using k-Nearest Neighbor and Learning Vector Quantization Methods

Santoso and Mohammad Isa Irawan

***Abstract*—Poverty is the inability of individuals to fulfill the minimum basic needs for a decent life. The problem of poverty is one of the fundamental problems that become the central attention of the local government. One of the government efforts to overcome poverty is using the alleviation programs. Government often faces some difficulties to sort out of the poverty levels in the society. Therefore it is necessary to conduct a study that helps the government to identify the poverty level so that the aid did not miss the targets. In order to tackle this problem, this paper leverages two classification methods: k-nearest neighbor (k-NN) and learning vector quantization (LVQ). The purpose of this study is to compare the accuracy of the value of both methods for classifying poverty levels. The data attributes that are used to characterize poverty among others include: aspects of housing, health, education, economics and income. From the testing results using both methods, the accuracy of k-NN is 93.52%, and the accuracy of LVQ is 75.93%. It can be concluded that the classification of poverty levels using k-NN method gives better performance than using LVQ method.**

***Index Terms*—k-nearest neighbor, learning vector quantization, poverty.**

I. Introduction

POVERTY is one of the fundamental problems that became the central attention of the government in any country. One important aspect to support the poverty reduction strategy is the availability and accuracy of poverty data and the ability to deliver aid to the right targets. Measurement of poverty that can be trusted can be a powerful instrument for policy makers to focus attention on an area with the living conditions of the poor. The poverty data may be used to evaluate government policies on poverty and set targets for the poor with the aim to improve their conditions [1].

One of the government's efforts to reduce poverty is through several programs for poverty countermeasures. In this case, the government is often difficult to sort out the levels of poverty in the society, especially poor households. It follows that the distribution of aid is sometimes not well targeted. To support the successful implementation of the program, especially with regards to poverty reduction, we need a study that could assist the government in identifying and classifying poor households that have traits or characteristics of poverty that is almost the same. By knowing the information regarding poverty criteria of each class, it is expected that the local government policy program can be arranged so that it is more focused on target or targets to be achieved.

Manuscript received February 29, 2016; accepted March 21, 2016.

The authors are with the Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia. Emails: santosos17march@gmail.com, mii@its.ac.id

Classification problems have been widely discussed by researchers in many contexts and disciplines that reflect the benefits and broad appeal as one of the steps in the data analysis. Accuracy and precision in the classification of data are very important. In recent years, classification method has been proven in helping many people's work, such as in medical [2], [3], [4], image classification [5], text classification [6] etc. Some methods of classification are often used, such as rule-based, neural networks, support vector machines, naïve Bayes and nearest neighbor.

With the existence of several methods, problems that often arise are the type of methods that should be chosen. The research that has been done before is on classification of heart diseases using k-nearest neighbor and genetic algorithm (GA) [2]. This study provides results that the use of GA in the k-NN method for classification of heart diseases has a better accuracy rate compared with the k-NN method without GA. In addition, there is also a research on the comparative results of the classification of diabetes mellitus using back propagation neural network and learning vector quantization [4].

Based on the description above, in the present study the authors will examine and compare the performance of the k-nearest neighbor (k-NN) and Learning vector quantization (LVQ) methods for poverty level classification problems. The benefits of this research is to increase the depth of knowledge about the classification technique using k-NN and LVQ methods and can provide a reference method for accuracy comparison in classification problems. So the results of this classification may be considered by the government in identifying and classifying poor households.

II. Related Work

A. Poverty

Definition of poverty used in different countries vary. Poverty is often seen as the inability to pay for minimal living expenses although some experts argue that poverty is also a lack of access to services such as education, health, information, and a lack of public access to development and political participation.

Planning agency of national development defines poverty as a condition in which a person or group of people unable to meet their basic rights to maintain and develop a dignified life. The central bureau of statistics (CBS) defines poverty

as the inability of individuals to meet minimum basic needs for a decent life. A condition which is under the minimum requirement standard value line called the poverty line or the poverty threshold [1].

Data sourced from CBS poverty is often the basis for the implementation of the poverty reduction program by the government. To the best of our knowledge, CBS issued two types of data on poverty: the macro poverty data and micro poverty data. Both of these data have criteria, measurement, and coverage of different poverty. Macro poverty is calculated using the basic needs approach that covers the basic needs of food and non-food. The second approach is a micro poverty estimation used in non-monetary approach. Differences that occur in addition to the methods and approaches are also scope. Macro poverty only covers the poor, while the micro poverty besides the poor also includes near-poor population [7].

B. Classification

Classification is a process to find a model that describes or distinguishes the concept or class of data, in order to be able to predict the class of an object whose class is unknown [8].

In the classification, given the record number of the so-called training set, which consists of several attributes, the attribute can be either continuous or categorical, one attribute indicates the class to record. Evaluating the performance of a model built by classification algorithms can be done by counting the number of records in the test correctly predicted (accuracy) or false (error rate) by the model. Accuracy is defined as follows:

$$Accuracy = \frac{Number\ \ of\ \ correct\ \ predictions}{Total\ \ predictions}$$

C. Literature Review of Classification

The research that has been done before, among others classification of heart diseases using k-nearest neighbor and genetic algorithm [2]. This study provides results that the application of GA into the k-NN method for classification of heart diseases has a better accuracy rate compared with k-NN method without GA. In addition, there is also a research on the comparative results of the classification of diabetes mellitus using back propagation neural network and learning vector quantization [4]. This research shows the results of data classification using back-propagation of diabetes provides a higher degree of accuracy or accurate in reading patterns compared to data classification using LVQ network. The other research is on comparison of the three methods for classification, support vector machine (SVM), k-nearest neighbor (k-NN), backpropagation applied to image retrieval [5]. Generally all three methods have good accuracy and fast computational time. The result is as follows. The k-NN method has the best results among the three, and also k-NN method does not need training process like SVM and backpropagation.

III. k-NEAREST NEIGHBOR ALGORITHM

The k-nearest neighbor (k-NN) method was first introduced by Fix and Hodges in 1951 and 1952 [6], [9] and later developed by Cover and Hart in 1967 [10]. k-Nearest Neighbor (k-NN) is a method to classify the object based on the distance learning data closest to the object. The working principle of k-NN is to find the shortest distance between the data to be evaluated with the k neighbor in the closest training data [11].

In the learning phase, the algorithm simply stores the vectors of features and classification of learning data. In the classification phase, the same features are calculated for the test data. The distance of this new vector of all learning data vector is calculated, and k closest data are taken. The best value of k highly depends on the data. In general, a high value of k will reduce the effect of noises on the classification, but makes the boundaries between each classification becomes more blurred. Good value of k can be selected by using parameter optimization, for example by using cross-validation. Special case where the predicted classifications are based on learning closest data (in other words, $k = 1$) is called the k-nearest neighbor algorithm. The purpose of the k-NN algorithm is to classify new objects based on attributes and training samples, where the results of the new test samples were classified by the majority of the categories k nearest neighbors. In the process of classification, this algorithm does not use any model to be matched and only based on memory. The k-NN classification algorithm uses adjacency as the predictive value of the test sample new ones. According to [11], the ratings for the k nearest neighbor based on the similarity is calculated using Euclidean distance which a defined as follows:

$$D(X, Y) = \sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2} \tag{1}$$

The k-nearest neighbor algorithm can be written as follows [3]:

1) Let k be number of nearest neighbors and D be the set of training samples Y_i
2) For each test sample X_i do compute using Euclidean distance for every sample Y_i of D:
   a) Select the k close set training samples to test sample X_i
   b) Classify the sample X_i based on majority class among its nearest neighbors.

Some of the advantages of k-NN are a) it is very simple to implement and easy to justify the outcome. Although k-NN has these advantages, it has some disadvantages such as: a) high computational cost since it needs to compute the distance of each test instance to all training samples, b) requires large memory proportional to the size of training set, c) Low accuracy rate in multi-dimensional data sets with irrelevant features, d) there is no rule of thumb to determine value of parameter k.

IV. LEARNING VECTOR QUANTIZATION ALGORITHM

Learning vector quantization network was first introduced by Kohonen Tuevo. LVQ is a network of artificial nerves that make learning in supervised competitive layer. A competitive layer will automatically learn to classify input vector are grouped into classes that have been defined through a network

that has been trained. The classes were obtained as a result of competitive layer depends only on the distance between the input vectors. If the two input vectors closer together, it would put both the competitive layer input vectors into the same class.

LVQ network is a network classifying the pattern so that each unit of output states of a class or category. The weight vector for the output unit is often called the reference vector for the class declared by the unit. During the training output unit searched his position by adjusting the weight through unsupervised training [12].

The following is the algorithm of learning vector quantization (LVQ):

1) Set:
   The initial weight input variable j to go to classes (clusters to-i , W_ij, $i = 1, 2, ..., k$; and $j = 1, 2, ..., m$
   Enter the epoch: Max epoch
   Parameter learning rate: α
   Reduction learning rate: Dec α
   Allowed minimum learning rate: Min α
2) Enter:
   Data input: x_{ij} with $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$
   Target class: T_k with $k = 1, 2, ..., n$
3) Set initial conditions: $epoch = 0$
   Do it if: $(epoh \leq MaxEpoh)$ and $(\alpha \geq Min\ \alpha)$
   $epoch = epoch + 1$
   Work for $i = 1$ to n
   Determine j such that $|x_i - W_j|$ minimum with $j = 1, 2, ..., k$
   Fix W_j with the following provisions:
   If $C_j = T$ then
   $W_j(new) = W_j(old) + \alpha(x - W_j)$
   If $C_j \neq T$ then
   $W_j(new) = W_j(old) - \alpha(x - W_j)$
4) Reduction of the learning rate value

V. Results and Discussions

The data used in this research are collected from the office of the central bureau of statistics. Source of data used is the data targeted households Documenting Social Protection Program in 2011 by taking a sample of the data as much as 216 households. Determination of targeted households approach based on the characteristics of poor households consisting of four 14 criteria/poverty indicators that is:

Of the 14 attributes of the data set, there is data that need to be converted into numerical form that can be used as input to the training and testing process. So we need further data transformation and normalization of data. While the output data of the target classes that is classifies into three grade categories.

The purpose of this study was to find a comparison of the level of accuracy of the method k-NN and LVQ. In the next section, we will discuss the accuracy of the classification results by using both methods.

A. Implementation of k-NN method

The classification method k-nearest neighbor is divided into two processes, namely the processes of training and testing.

TABLE I
Criteria for Poverty

Attribute	Description
x_1	Broad of floor
x_2	Floor type
x_3	Wall type
x_4	Source of illumination
x_5	Cooking fuel
x_6	Sources of drinking water
x_7	Type of toilet/WC
x_8	Ownership of assets
x_9	Income
x_{10}	Education
x_{11}	Jobs
x_{12}	Treatment capabilities
x_{13}	Consumption
x_{14}	Capability to buy

The training process k-NN is using sample data that consists of variables and target class is taken from the number of classification classes as input. While in the testing process, k-NN is using the distance calculation value for the attributes of each test data against all the attributes in the training data with Euclidean distance formula. Furthermore, we generate a number of value k nearest neighbor, where the results of the new test data is classified based on the majority of the class category of the k nearest neighbor.

The accuracy of the test results using k-NN method in terms of two parameters: k nearest neighbor and the amount of training data. k-NN test method is done by determining the value of k and the amount of training data is used. The test result accuracy of classification in terms of the value of k and the amount of training data is presented in Table II as follows:

TABLE II
The accuracy of k-NN method with 216 testing data reviewed from k parameters.

Parameter	Accuracy (%)	Elapsed time (s)
$k = 3$	87.037	6.21
$k = 4$	93.52	4.09
$k = 5$	86.57	3.93
$k = 6$	89.81	3.95
$k = 7$	83.8	3.91
$k = 8$	87.5	3.85
$k = 9$	80.56	3.90
$k = 10$	84.72	3.89

Table II shows the results of trials using k-NN method. The trial was conducted using the method k parameter value, where k is changed from 3 to 10. Based on the value of k that is used, the highest accuracy results seen in the value of $k = 4$ is 93.52%. Next trial using training data as much as 162 data and the remaining 54 the data used as a test data. The accuracy of the results of the trials in terms of the parameter $k = 3$ to 10 for 54 testing the data can be viewed in Table III below:

Furthermore, the results of the trial using k-NN method is presented in a graphical form in Fig. 1 below:

Figure 1 shows the level of accuracy of k-NN method using the amount of training data respectively of 216 and 162 data set. The graph above shows the highest accuracy with parameter $k = 4$ the number of training data as much as 216

TABLE III
THE ACCURACY OF k-NN METHOD WITH 54 TESTING DATA REVIEWED FROM k PARAMETERS.

Parameter	Accuracy (%)	Elapsed time (s)
$k = 3$	79.63	4.36
$k = 4$	81.48	4.09
$k = 5$	81.48	4.08
$k = 6$	83.33	4.06
$k = 7$	83.33	3.93
$k = 8$	81.48	3.69
$k = 9$	77.78	3.91
$k = 10$	81.48	3.86

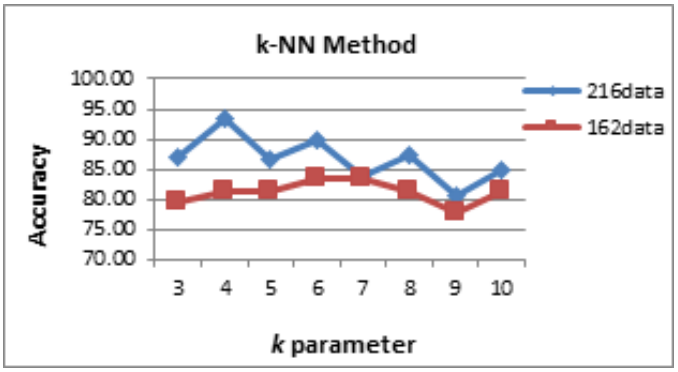

Fig. 1. Graph accuracy LVQ with 216 training dataset reviewed from the k parameters.

data. It can also be seen that the amount of training data 216 of data, the accuracy values decreased at the parameters $k = 5$, $k = 7$ and $k = 9$.

B. Implementation of LVQ method

Learning vector quantization (LVQ) is a network of single layer consisting of two layers of input and output. Input layer consists of 14 units of input taken from the variable criteria of poverty, while the output unit consists of three units of output which are taken from the number of grade classifications. LVQ network architecture in this research are presented in Fig. 2.

Descriptions of Fig. 2 are as follows:

- x_i is a vector of training as much as (x_1 until x_{14})
- T is the target for as many as three targets training vectors are t_1, t_2 and t_3 determined based on two existing classes
- w_j is the weight vector for the j-th output unit is $(w_{1j}, w_{2j}, ..., w_{14j})$
- C_j is a category / class of computational results by unit of j-th output, consists of three classes, namely C_1, C_2 and C_3
- $\|x - W_j\|$ is the Euclidean distance between the input vector and the weight vector for the j-th output unit.

The accuracy of the test results using LVQ in terms of learning rate parameters, the number of iterations and the amount of training data. LVQ trials are done by changing the value of learning rates. In this experiment, we use the following learning rates 0.01, 0.05 and 0.1 with the number of iterations used from 50 to 500 iterations. The first test is done by using the amount of training data and testing as many as 216 data. The accuracy of classification results can be seen in Table IV below:

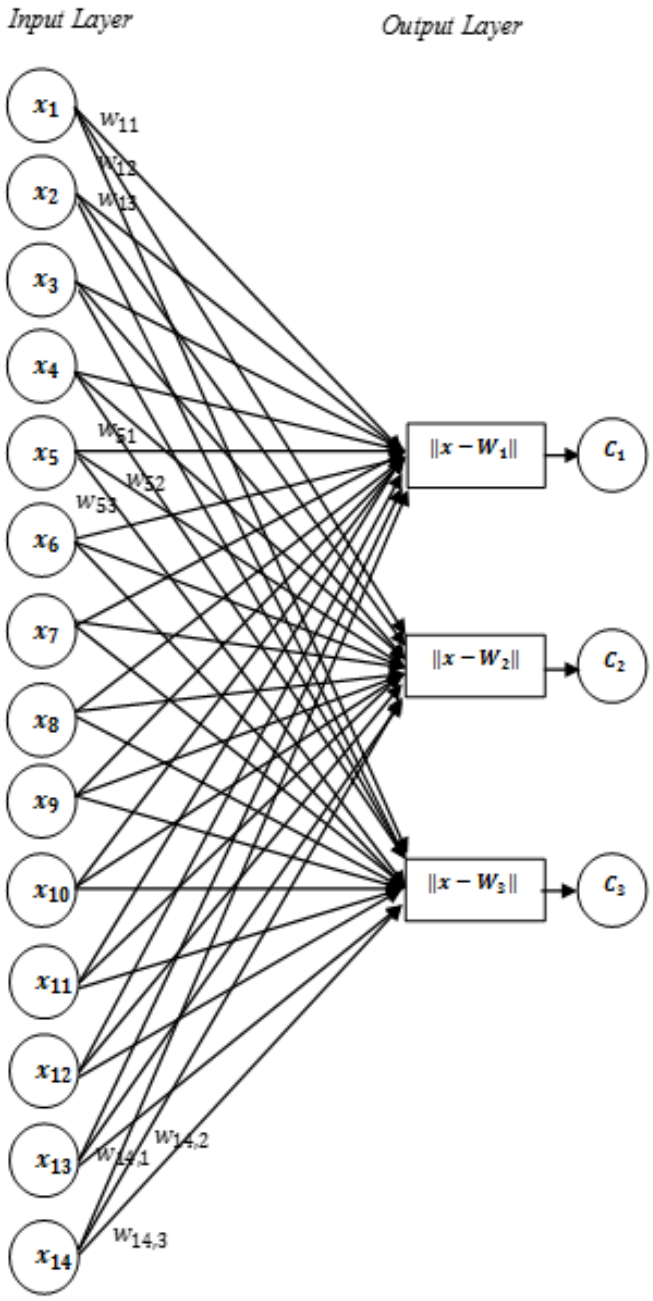

Fig. 2. LVQ network architecture issues poverty level.

Iteration Table IV shows the results of experiments using LVQ. The test is done by changing the value of learning rate as follows 0.01, 0.05 and 0.1 as well as the number of iterations between 50 and 500 iterations. Accuracy results in Table IV are obtained from trials by running a program with as much data as 216 trained data by using learning rate = 0.1, 0.05 and 0.01. Here is a graph of accuracy using LVQ for the amount of training data as much as 216 data presented in Fig. 3 below:

Figure 3 shows the level of accuracy of LVQ using the amount of training data as much as 216 data. The graph above shows the accuracy of the learning rate 0.01, 0.1, 0.05 and between 50 to 500 iterations. In Fig. 3 shows that the accuracy value decreased for the 0.01 when iterating over 200 iterations.

TABLE IV
THE ACCURACY RESULTS OF LVQ METHOD WITH 216 TESTING DATA TO BE REVIEWED FROM THE AMOUNT OF ITERATIONS AND LEARNING RATE.

iteration	Learning rate (α)		
	0.1	0.05	0.01
50	72.68	77.78	75
100	75	75.46	75.93
200	74.07	73.15	77.78
300	77.78	77.31	75.93
400	73.61	74.07	76.39
500	74.54	74.54	75

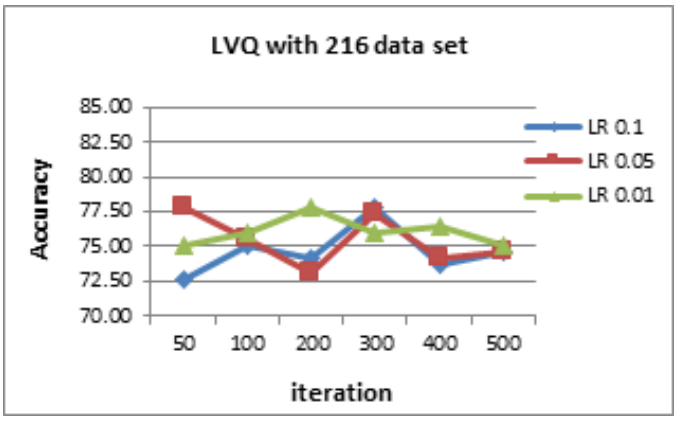

Fig. 3. Graph accuracy LVQ with 216 training dataset reviewed from the number of iterations.

As for the learning rate of 0.1 and 0.05, accuracy value is not stable in any number of iterations used. Next trial with LVQ is performed using training data as much as 162 data and the rest of the data as much as 54 data are being used as a test data. Results of the accuracy of the classification with 162 training data can be viewed in Table V below:

TABLE V
THE ACCURACY RESULTS OF LVQ METHOD WITH 54 TESTING DATA TO BE REVIEWED FROM THE AMOUNT OF ITERATIONS AND LEARNING RATE.

iteration	Learning rate (α)		
	0.1	0.05	0.01
50	81.48	81.48	79.63
100	83.33	79.63	79.63
200	81.48	75.93	83.33
300	83.33	83.33	85.18
400	83.33	83.33	85.18
500	81.48	83.33	83.33

Table V shows the accuracy of the classification using LVQ with 54 test data. The test is done by changing the value of learning rate as follows 0.01, 0.05 and 0.1 as well as the number of iterations between 50 and 500 iterations. Accuracy results in Table V are derived from the test program by running with as much data as 162 trained data by using learning rate = 0.1, 0.05 and 0.01. Here also presented graphs using LVQ accuracy value for amount of training data as much as 162 the data in Fig. 4 below:

Figure 4 shows the level of accuracy of LVQ using the amount of training data as much as 162 data. The graph above shows the accuracy of the learning rate at iteration 0.01, 0.1, 0.05 and between 50 to 500 iterations. Figure 4 shows that

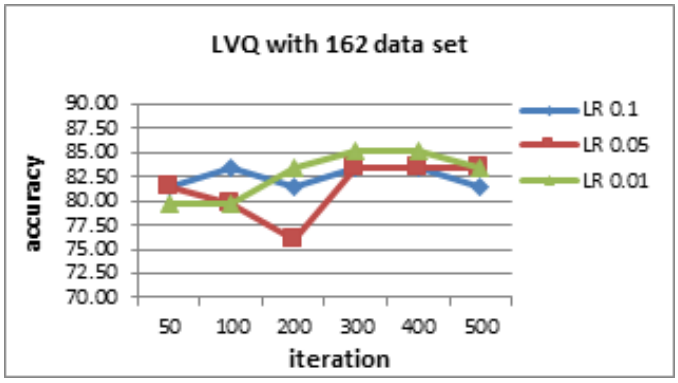

Fig. 4. Graph accuracy of LVQ with 162 training dataset reviewed from the number of iterations.

accuracy value has a constant increase for learning rate 0.01 when the number of iterations over 100 iterations. When the number of iterations is more than 300 iterations, the value of accuracy is getting smaller. When compared with other learning rate parameter, the value of $\alpha = 0.01$ generates fairly good accuracy.

C. Comparison of k-NN and LVQ methods

In a comparison of two methods, different parameters are used. This looks at the k-NN method used k parameters, whereas the parameters used in LVQ is learning rate and the number of iterations. From the test results, comparison of these two methods are based on the classification result by the number of used training data and to further take the average value of the highest accuracy on any parameter that is determined to obtain the results of the comparison of the two methods.

After that, we obtain parameters' value and the number of iterations that provide the highest accuracy in this research. The next test performed on the same amount of training data with parameters $k = 4$, learning rate of 0.01 and 300 iterations. The results of the training and testing process by using the amount of data as much as 216 data on both methods are show in Table VI.

TABLE VI
COMPARISON OF THE ACCURACY OF CLASSIFICATION RESULTS USING k-NN AND LVQ METHODS.

Method	Parameter(s)	Correct results	Accuracy (%)	Elapsed time (sec)
k-NN	$k = 4$	202	93.52	4.09
LVQ	$epoch = 300$ $\alpha = 0.01$	164	75.93	110.2

Table VI shows the comparison of the results of classification accuracy using k-NN and LVQ. The test is done by using the data as much data as 216 training and test data. The trial results with k-NN method showed that of all the test data, obtained the corresponding results as much as 202 class data so that the value of 93.52% accuracy. While testing with LVQ obtained the appropriate amount of data as much as 164 classes of data, so the value of accuracy obtained at 75.93%. If the terms of performance in the process of running

the program, Table VI shows that the k-NN method is much faster than using LVQ. This is because LVQ takes iteration to obtain final weights during the iteration process. While the k-NN method of distance measurement only done on a dataset so that the time used for running the program is quite short.

VI. Conclusions

Based on the results and discussion, it can be concluded that the accuracy of the classification by using the amount of training data is the same in both methods with the value of each parameter $k = 4$, $\alpha = 0.01$ and 300 iterations values obtained highest accuracy in the k-NN amounted to 93.52%, while highest accuracy on LVQ amounted to 75.93%. In terms of the performance of both methods of classification, k-NN method is faster in the process of running the program when compared to LVQ. From the description above, it can be concluded that the k-NN method is better compared to LVQ in relation to the issues of poverty level classification.

For further research, we can change the type of distances used as well as the parameters that k, learning rate and the number of iterations. In addition, the use of the data type in this research is less suitable to the k-NN method or LVQ thus allowing it to be applied in the case with other types of dataset.

References

[1] CBS, "Analysis and calculation of poverty level in 2008," Central Bureau of Statistics, Jakarta, Tech. Rep., 2008.

[2] M. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using k- nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013, first International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA'13).

[3] A. Karegowda, M. Jayaram, and A. Manjunath, "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology*, vol. 1, no. 3, pp. 147–151, 2012.

[4] A. Nurkhozin, M. Irawan, and I. Mukhlash, "Komparasi hasil klasifikasi penyakit diabetes mellitus menggunakan jaringan syaraf tiruan backpropagation dan learning vector quantization," in *Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA*, 2011, pp. M33–M40.

[5] M. Athoillah, M. Irawan, and E. Imah, "Study comparison of svm-, k-nn-and backpropagation-based classifier for image retrieval," *Jurnal Ilmu Komputer dan Informasi*, vol. 8, no. 1, pp. 17–19, 2015.

[6] E. Fix and J. Hodges Jr., "Discriminatory analysis-nonparametric discrimination: Small sample performance," Report no. 4, USAF School of Aviation Medicine, Texas, Tech. Rep., 1951.

[7] CBS, "Poverty data analysis based on data collection data social protection program in 2011," Central Bureau of Statistics, Jakarta, Tech. Rep., 2012.

[8] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Addison Wesley, Boston, 2006.

[9] E. Fix and J. Hodges Jr., "Discriminatory analysis-nonparametric discrimination: Small sample performance," Report no. 11, USAF School of Aviation Medicine, Texas, Tech. Rep., 1952.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[11] J. Sreemathy and P. Balamurugan, "An efficient text classification using knn and naive bayesian," *International Journal on Computer Science and Engineering*, vol. 4, no. 3, pp. 352–351, 2012.

[12] L. Fausett, *Fundamental of Neural Network: Architectures, Algorithms, and Aplications*. Prentice Hall Inc, New Jersey, 1994.