Scientific Journal of Informatics
Vol. 7, No. 2, November 2020

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Diagnosis of Lung Disease Using Learning Vector Quantization 3 (LVQ3)

Dwi Marisa Midyanti[1], Syamsul Bahri[2], Rahmi Hidayati[3]

[1]Computer System Department, Faculty of Mathematics and Natural Sciences,
Universitas Tanjungpura Pontianak, Indonesia
Email: [1]dwi.marisa@siskom.untan.ac.id, [2]syamsul.bahri@siskom.untan.ac.id,
[3]rahmihidayati@siskom.untan.ac.id

Abstract

Lung disease is one of the diseases with the highest number of patients in Indonesia. Lung disease is a disease with many types and symptoms that are almost the same as each other. This study uses an artificial neural network Learning Vector Quantization 3 (LVQ3), to diagnose lung disease. The data used in this study were 113 medical records, with seven types of lung disease, and 27 symptoms of the disease. From the experimental results, the best LVQ3 parameters from this study are using m = 0.15, and the learning rate = 0.15. LVQ3 produces the best accuracy value for training data at 87.5% of 80 data, and accuracy for test data 88% of 33 data.

Keywords: Lung disease diagnosis, Neural Network, LVQ3

1. INTRODUCTION

Lung disease is a disease that has been widely studied by researchers in Indonesia. It's because lung disease has a large number of patients in Indonesia. Besides, lung disease has many types of diseases with symptoms that are almost the same as each other. In 2016, [1] used the Forward Chaining method to diagnose lung disease. The system's probability of accuracy was 84.21%, using five types of lung disease with 27 symptoms inputted. Sumiati [2] uses the Certainty Factor method to diagnose lung disease using seven types of lung disease. This study produces applications that can help diagnose lung disease that provides information about lung disease, solutions, and percentages of trust based on the results of the Certainty Factor method. Certainty Factor is also used by [3] to diagnose lung disease with an accuracy rate of 85.18% for 27 test data. Forward Chaining method and Certainty Factor method are methods widely used in expert systems with output in classification.

Artificial Neural Networks can also solve classification problems. One method in artificial neural networks is Learning Vector Quantization (LVQ), which Teuvo Kohonen introduced in 1989.LVQ is a method for supervised competitive training. The competitive layer will automatically learn to classify the given vector input. If some input vectors have very close distances, they will be grouped in the same class [4]. According to [5], Kohenen improvised the LVQ algorithm, namely LVQ2, LVQ2.1, and LVQ3. LVQ3 has been used by [6] to identify blood images.

In research [6], LVQ3 can recognize all classes when using 90% of training data. Based on the above research, this study aims to apply the LVQ3 method to diagnose lung disease and determine the accuracy produced by LVQ3. Another objective is to determine m, the LVQ3 parameter's effect on changes in the error value in the LVQ3 network.

2. METHODS

Teuvo Kohonen introduced LVQ3 in 1990. According to Kohonen in [5], LVQ is a pattern classification method in which each ouput unit represents a particular class or category. The weight vector for an output unit is often referred to as a reference (or codebook) vector for the class that unit represents. During training, the output units are positioned to approximate the decision surfaces of the theoretical Bayes Classifier.

The LVQ architecture in this study can see in Figure 1.

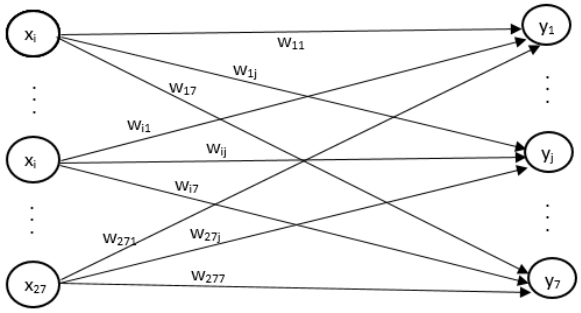

Figure 1. LVQ3 Architecture

where:
x = input
y = output
w = LVQ weight

LVQ3 allows the two closest vectors to learn as long as the input vector satisfies the window condition in equation (1) [5].

$$\min \left[\frac{d_{c1}}{d_{c2}}, \frac{d_{c2}}{d_{c1}}\right] > (1 - \varepsilon)(1 + \varepsilon) \quad (1)$$

where typical value of $\varepsilon = 0.2$ are indicated. If one of the two closest vectors, y_{c1}, belong to the same class as the input vector x, and the other vector y_{c2} belongs to the same class, the weight updates are for LVQ2.1. However, LVQ3 extends the training algorithm to provide for training if x, y_{c1}, and y_{c2} belong to the same class. In this case, the weight update are in equation 2 [5].

$$y_c(t+1) = y_c(t) + \beta(t)[x(t) - y_c(t)] \quad (2)$$

for both y_{c1} and y_{c1}. The learning rate $\beta(t)$ is a multiple of the learning rate $\alpha(t)$ that is used if y_{c1} and y_{c1} belong to different classes. The appropriate multiplier is typically between 0.1 and 0.5, with smaller value corresponding to a narrower windows . The formula shown in equation (3) [5].

$$\beta(t) = m\,\alpha(t) \text{ for } 0.1 < m < 0.5 \quad (3)$$

The number of iterations = 100. The minimum error used is 0.05. Errors calculated using Mean Square Error (MSE) as shown in equation (4) [7].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Target\ Output - Actual\ Output)^2 \quad (4)$$

The number of iterations and MSE used to stop the loop of LVQ then accuracy calculated by the confusion matrix

3. RESULT AND DISCUSSION

3.1. Data Collection

The data used in this study are the results of medical records from 113 patients with seven types of lung disease, namely Pneumonia, Tuberculosis, Bronchitis, Chronic Obstructive Pulmonary Disease (COPD), Asthma, Pneumothorax, and Pleural Effusion. There are 27 symptoms obtained from medical records:

G1 : Cough with phlegm
G2 : Coughing up phlegm with blood
G3 : Occasional Cough
G4 : Difficulty breathing
G5 : Shortness of breath
G6 : Shortness of breath, accompanied by a wheezing
G7 : Nausea / vomit
G8 : Chest pain
G9 : Body weakness
G10 : Difficulty Swallowing
G11 : Headache
G12 : High fever
G13 : Fever (> 1 week)
G14 : Fever in the afternoon and evening
G15 : Fever temperature up-down
G16 : Sneezing in the morning
G17 : Night sweats
G18 : Weight loss
G19 : Decreased appetite
G20 : Active smokers
G21 : Dust Allergy
G22 : Fur Allergy

G23 : Feeling of being unwell
G24 : Had Bronchitis
G25 : Had Emphysema
G26 : Excessive mucus production
G27 : Feel sick after doing physical activity

From 113 data, 70% used for training data, and 30% used for test data.

3.2. Testing parameters m LVQ3

This test do to shows the effect of the value of m on the results of training data using learning levels (α) 0.1-0.5 and the value of $\varepsilon = 0.2$ The m values observed in this study were 0.15, 0.25, 0.35, and 0.45. Figure 2 is a comparison between LVQ3 m parameters.

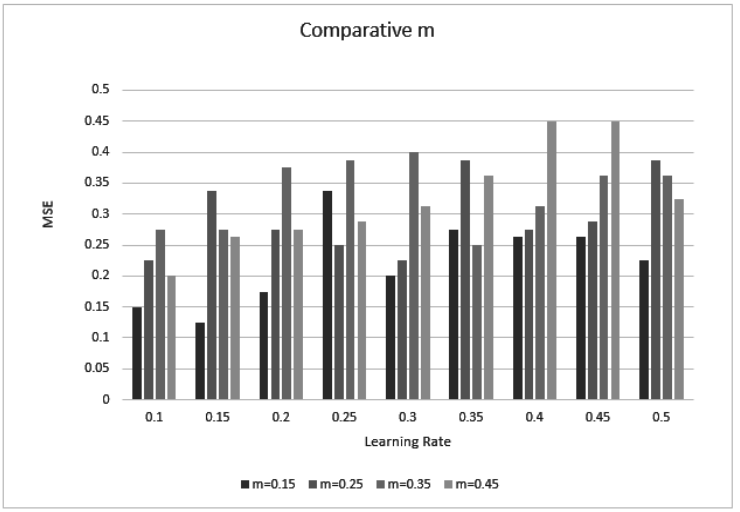

Figure 2. Comparative m

In Figure 2, it can see that the error value using the m parameter moves up-down. The minimum error in the training data obtained using m = 1.5 and α = 0.15. MSE values using m = 1.5 and α = 0.15 can be seen in Figure 2. The minimum MSE value is 0.125, which means the accuracy of the training data is 87.5%. LVQ3 can recognize 70 training data from 80 training data used.

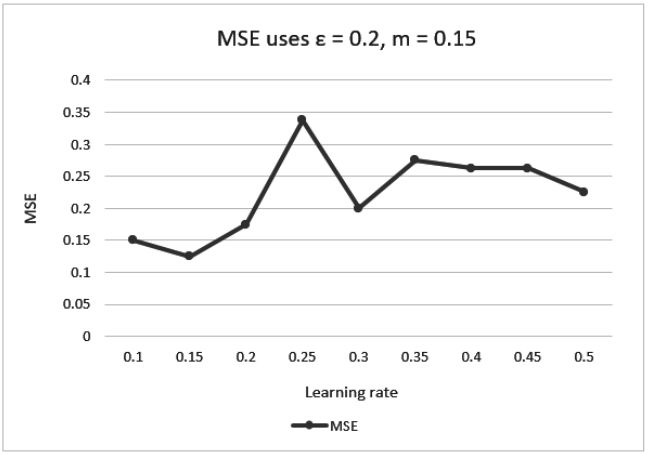

Figure 3. MSE uses ε = 0.2, and m =0.15

3.2. Testing parameters ε LVQ3

The test carried out to see the effect of the ε parameters on the MSE of training data using α = 0.1-0.5 and m = 0.15. LVQ3 has a typical value of 0.2. Therefore, we observed the effect of ε with a range from 0.1 to 0.4. The results can see in Figure 4.

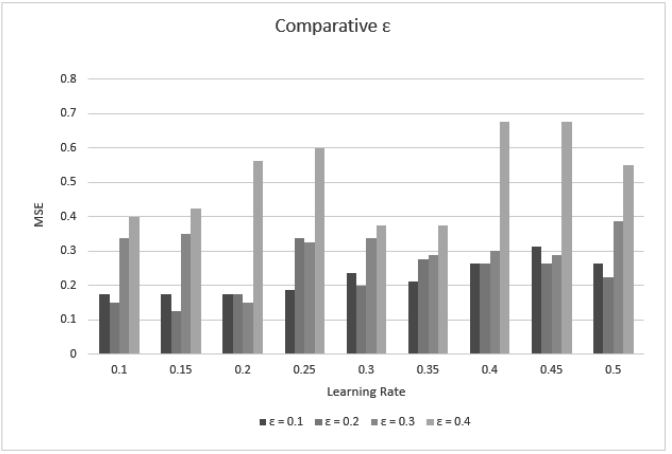

Figure 4. Comparative ε

Figure 4 shows that the MSE train data using ε = 0.1 and ε = 0.2 do not differ much, but the MSE minimum is obtained from ε = 0.2 with α = 0.15. used to calculate output so that it can produce MSE values from test data.

The results of the accuracy of the test data using the Confusion matrix can see in Table 1.

Table 1. Accuracy of test data using a confusion matrix

Actual	Prediction							Recall (%)
	1	2	3	4	5	6	7	
1	7	1	1	0	0	0	0	78
2	0	5	0	0	0	0	0	100
3	0	0	2	0	0	0	0	100
4	0	0	1	5	0	0	0	83
5	0	0	0	0	4	1	0	80
6	0	0	0	0	0	2	0	100
7	0	0	0	0	0	0	4	100
Precision (%)	100	83	50	100	100	67	100	88

Table 1 shows that LVQ3 can correctly identify Pneumonia, COPD, Asthma, and Pleural Effusion from the test data, but errors occur when detection TB, Bronchitis, and Peneumotoraks. It's can be caused by the similarity of symptoms from one disease to another and can create by the lack of training data on several types of diseases. From 33 test data, LVQ3 can recognize 29 test data so that 88% accuracy obtained.

4. CONCLUSION

The LVQ3 method can diagnose lung disease with an accuracy of 87.5% training data and 88% test data accuracy. From 80 training data, LVQ3 can recognize 70 data accurately. For 33 test data, LVQ3 can identify 29 data precisely. The best parameter of LVQ3 in this study is using m = 0.15, and $\alpha = 0.15$.

5. REFERENCES

[1] Rahmawati, E., & Wibawanto, H. (2016). Sistem Pakar Diagnosis Penyakit Paru-Paru Menggunakan Metode Forward Chaining. *Jurnal Teknik Elektro*, *8*(2), 64-49.

[2] Sumiati, Badriyah, R., & Ariyani, A. (2017). Sistem Pakar Untuk Diagnosa Penyakit Paru - Paru Menggunakan Metod5e Certainty Factor Di Puskesmas Citangkil. *Jurnal ProTekInfo*, *4*, 34-42.

[3] Iqbal, M., Setyaningsih, F., & Bahri, S. (2019). Implementasi Metode Certainty Factor Dalam Sistem Pakar Diagnosis Penyakit Paru-paru Berbasis Android. *Coding : Jurnal Komputer dan Aplikasi, 7*(3), 155-164.

[4] Kusumadewi, S. (2004). *Membangun Jaringan Syaraf Tiruan Menggunakan MATLAB & EXCEL LINK.* Yogyakarta: Graha Ilmu.

[5] Fausett, L. V. (1994). *Fundamentals of neural networks: Architectures, algorithms, and applications*. Englewood Cliffs, NJ: Prentice-Hall.

[6] Putra, F., & Syafria, F. (2018). Penerapan Learning Vector Quantization 3 (LVQ3) untuk Mengidentifikasi Citra Darah Acute Lymphoblastic Leukemia (ALL) dan Acute Myeloid Leukemia (AML). *CoreIT*, *4*(1), 27-33.

[7] Tukur, U., & Shamsuddin, S. (2015). Radial Basis Function Network Learning with Modified Backpropagation Algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *13*(2), 369 – 378.