Optimization Learning Vector Quantization Using Genetic Algorithm for Detection of Diabetics

Inggih Permana[1], Nesdi Evrilyan Rozanda[2], Fadhilah Syafria[3], Febi Nur Salisah[4]
[1,2,4]Department of Information Systems, Faculty of Science and Technology, Universitas Sultan Syarif Kasim (UIN SUSKA) Riau, Pekanbaru-Riau, 28293
[3]Department of Informatics Engineering, Faculty of Science and Technology, Universitas Sultan Syarif Kasim (UIN SUSKA) Riau, Pekanbaru-Riau, 28293

Article Info

Article history:

Received Feb 11, 2018
Revised May 30, 2018
Accepted Aug 21, 2018

Keywords:

Diabetics
GA
LVQ
Weight vectors

ABSTRACT

This study proposed the method to improve the result of Learning Vector Quantization (LVQ) by optimizing the weight vectors using a genetic algorithm (GA) to detect the diabetics. Initial value of individuals for GA is taken from weight vectors which come from the last m iterations of LVQ training result. The result of experiment showed that there is a significant increase in sensitivity level, however there is a significant decrease in specificity level. It means the proposed method success in improving the LVQ ability to recognized the diabetics, but it lowers the ability of LVQ to recognize the people unaffected by diabetes.

Corresponding Author:

Inggih Permana,
Department of Information Systems, Faculty of Science and Technology,
Universitas Sultan Syarif Kasim (UIN SUSKA) Riau,
Pekanbaru-Riau, 28293.
Email: inggihpermana@uin-suska.ac.id

1. INTRODUCTION

Diabetes Mellitus (DM) is a disease that occurs when the pancreas can not to secretion enough insulin [1]. DM may increase the risk of vessel damage, blindness, kidney disease, cardiac disease, nerve damage, stroke, birth defects [2]. DM is one of the major health problems that occur in Indonesia. The prevalence of DM patients (diabetics) in Indonesia is increasing every year. Among the 1980s, the prevalence of diabetics in people with age over 15 years is 1,5% to 2,3% [3]. In 2001, in urban areas, the prevalence diabetics of people aged between 25-64 years old is 5.7% [4]. In 2013, the prevalence of diabetics in urban, rural and whole Indonesia are 6,8%, 7%, and 6,9% [3]. From twelve million diabetics in Indonesia, there was 69,6% undiagnosed since the disease were affected. This means that most diabetics in Indonesia realized the disease when it is severe. The reason is that the DM appears over many years so it was not realized by the sufferer.

Computer technology for the early detection of diabetics is a solution to resolve the problems. This has been done by previous researchers by developing the variety of algorithms, such as: (1) changing the algorithms artificial immune recognition system by adding fuzzy k-nearest neighbor [5]; (2) combining the algorithm between centripetal sped up particle swarm optimization and multi-layer perceptron algorithm [6]; (3) by using naive bayes and decision tree methods [7].

This research used learning vector quantization (LVQ) methods for early diabetics detect. LVQ is the pattern classification method where the entire output unit represents the certain classes [8]. LVQ were the fastest and easiest applied and intuitive [9]. LVQ is chosen in this research because LVQ has been the success to applied in many areas of research, like in: (1) identification hand-writing [10]; (2) real-time monitoring [11]; and (3) array analysis from a sensor [12].

However, LVQ has a lack; the training result of weight vectors is depended on the parameters initialization that used in LVQ. A weight vectors is important in LVQ because this vector will determined the data classification to the certain classes. In this research, weight vector will references to determine someone who diagnose as the diabetics or not. The weakness of LVQ can be resolved by optimized the weight vectors use a genetic algorithm (GA). GA is a searching algorithm that inspired from the natural evolution theory. The result of the previous research show that GA can optimize the ability of classified algorithm [13-17].

LVQ optimization using GA has been done in previous studies. The optimization is done by finding the initial value of LVQ weight vector using GA [18-19]. Although the initial value of the weight vector affects the LVQ result, it can not guarantee that LVQ training produces the optimal representative vector. The method offered in this study do the opposite, LVQ training done first, then GA is used to optimize the vector representative of the LVQ training results. In addition, there are also studies that optimize LVQ by finding directly weight vectors using GA [20]. In such techniques, the LVQ training process is not done. LVQ is only used in the classification process only. The method offered in this study did not eliminate the LVQ training process since the initial individuals used on GA derived from the LVQ training.

2. RESEARCH METHOD

2.1. Optimization of weight vector of LVQ using GA

The optimize of LVQ comprise four stages. First initialize the parameter values of LVQ and GA. The second stage is LVQ training. The third stage is the establishments initial individual for GA. The last stage there is optimization using GA. The illustration of the process can be shown in equationure 1.

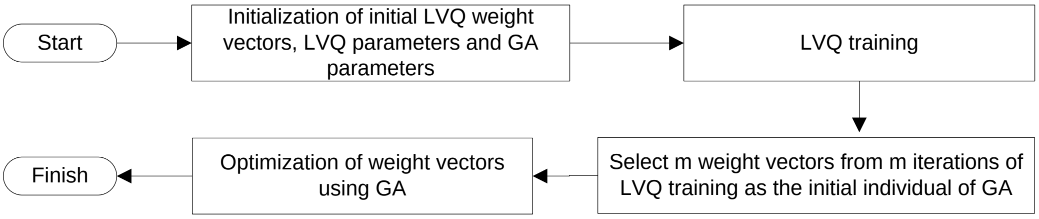

Start
Initialization of initial LVQ weight vectors, LVQ parameters and GA parameters
LVQ training
Select m weight vectors from m iterations of LVQ training as the initial individual of GA
Optimization of weight vectors using GA
Finish

Figure 1. The LVQ optimization process uses GA

The parameters that need to be initialized on LVQ are the initial value of the weight vectors, the number of iterations, the learning rate, the decrement learning rate. The parameters that need to be initialized on GA are the number of generations, the number of individuals, crossover probability (Pc) and mutation probability (Pm).

The initial individuals for GA in this research were taken from m last iteration of LVQ training. The chromosome from the individuals of GA in this research was represented in the real form. For more details show Figure 2.

This study uses stochastic universal sampling (SUS) as a selection method. SUS is a method that has a bias is 0 and has the complexity is O (N) [21]. Basically, SUS is a roullete wheel with N pointers. N is the number of individuals selected. The first individual is a random value between 1 and 1 / N. The next individual is 1 / N from the previous individual. Figure 3 is an illustration of how SUS works. In the picture is selected five individuals so that there are five pointers and the distance between individuals is 0.2. Based on the selection result, I1, I1, I2, I4 and I5 were selected.

The crossover technique that use in this research is line-crossover, and the mutation is done by adding a small random value. Formula of line-crossover can be seen in Equation 5.

$$CGen_i = PGen_{1i} + \lambda(PGen_{2i} - PGen_{1i}) \tag{1}$$

In the Equation 1, CGeni is i-th gen of a child, PGen1i is i-th gen of a first parent, Pgen2i is i-th gen of a second parent, λ is a random value between 0 until 1.

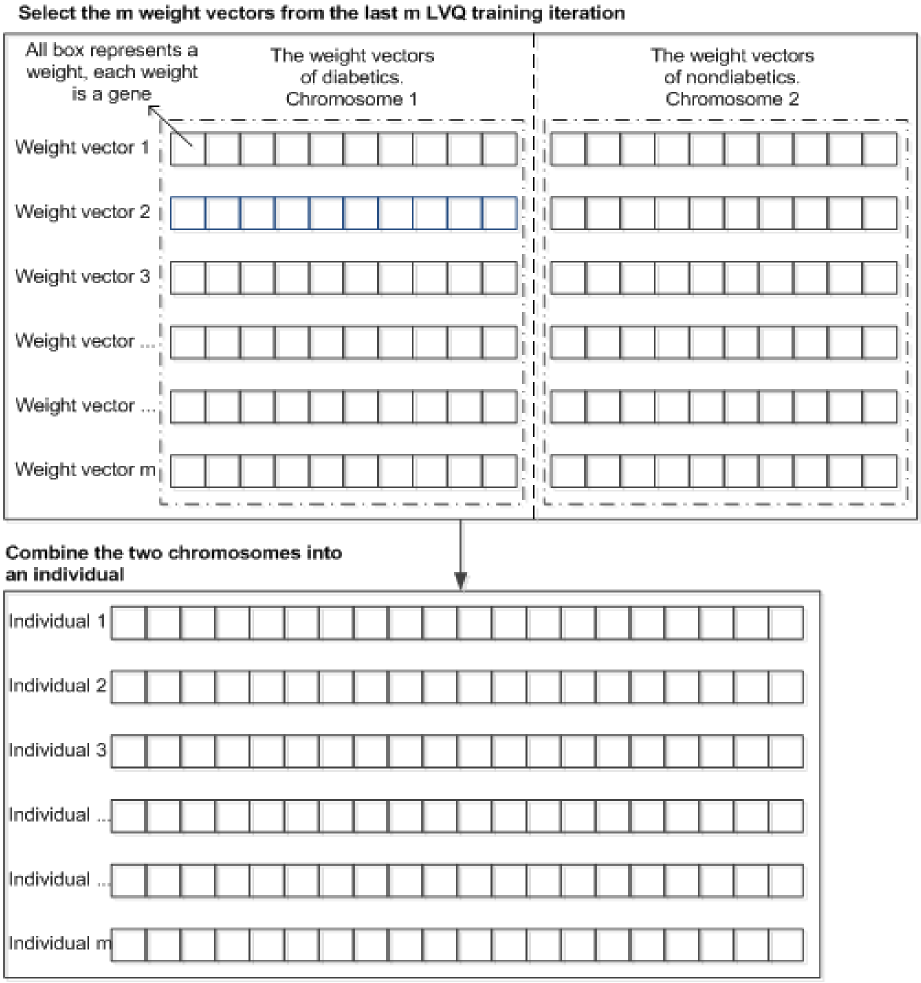

Select the m weight vectors from the last m LVQ training iteration
All box represents a weight, each weight is a gene
The weight vectors of diabetics. Chromosome 1
The weight vectors of nondiabetics. Chromosome 2
Weight vector 1
Weight vector 2
Weight vector 3
Weight vector ...
Weight vector ...
Weight vector m
Combine the two chromosomes into an individual
Individual 1
Individual 2
Individual 3
Individual ...
Individual ...
Individual m

Figure 2. Early individual GA creation

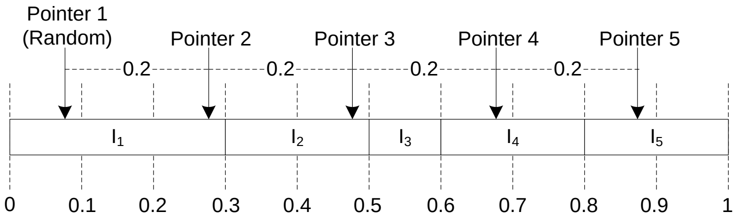


Figure 3. Illustration of how SUS works

2.2. Fitness Function

Fitness values in this research was computed with Equation 2.

$$Fitness = \frac{(1-\text{Accuracy}) + (1-\text{Sensitivity}) + (1-\text{Spesificity})}{3} \tag{2}$$

Accuracy, sensitivity and specificity are got by Equation 3, Equation 4 and Equation 5.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{4}$$

$$Spesificity = \frac{TN}{FP+TN} \tag{5}$$

Where, true positive (TP) is the number of DM patients classified as DM, false negative (FN) is the number of non-DM patients classified as DM, true negative (TN) is the number of non-DM patients classified as non-DM patients, false positive (FP) is the number of DM patients classified as non-DM patients.

2.3. Dataset

Dataset where used in this research is Pima Indians Database. This dataset can download on website repository learning machine of the University of California Irvine (https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). The purpose of this set of data is for classified the Pima Indians people that affected by DM or not according to personal data and heath check. This set of data were stage from 768 data where 268 people affected by DM and 500 people unaffected by DM. There are 7 attribute in this set of data. There is the number of pregnancy, plasma glucose concentrate, diastolic blood pressure, the thickness of triceps skin folds, body weight, DM genealogical history and age.

2.4. Experimental Setup

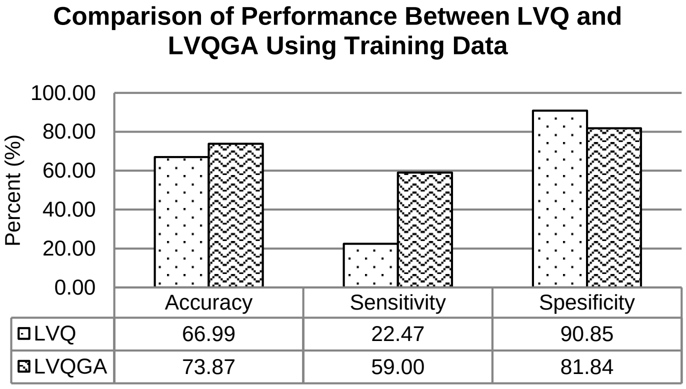
The evaluation was performed using 10-fold cross-validation. To apply the evaluation, then the data set is divided into ten groups. Nine groups will be used as training data while another will be used as test data. For each combination of parameters ten experiments were performed so that all groups were once test data. After all experiments were calculated the average accuracy, sensitivity and specificity.

3. RESULTS AND ANALYSIS

Figure 4 is the performance comparison between LVQ and LVQGA using training data. The figure shows GA increases accuracy of LVQ by 10.27% ((73.87-66.99)/66.99). Sensitivity level of LVQ increased 162.57% ((59.00-22.47)/22.47) but specificity level of LVQ decreased 9.92% ((90.85-81.84)/90.85). Based on this comparison it can be concluded, in the training data, GA increases LVQ accuracy through high increase of sensitivity level even though specificity level decrease.

Figure 5 is the performance comparison between LVQ and LVQGA using testing data. GA decreases accuracy of LVQ by 1.17% ((71.03-70.20)/71.03). However, there is a significant increase of sensitivity level of LVQ by 85.86% ((54.94-29.56)/29.56). Specificity level of LVQ decreased 16.04% ((93.43-78.44)/93.43). Based on this comparison it can be concluded, in the testing data, GA decrease LVQ accuracy through a high decrease of specificity level even though sensitivity level increase.

This study use sensitivity to measure capability of algorithms for recognize diabetics whereas specificity to measure capability of algorithms for recognize non-diabetics. Therefore, base on comparisons of performance between LVQ and LVQGA it can be concluded that GA improve LVQ capability in recognizing diabetics, but lower LVQ capability in recognizing non-diabetics.



	Accuracy	Sensitivity	Spesificity
□LVQ	66.99	22.47	90.85
▨LVQGA	73.87	59.00	81.84

Figure 4. Comparisons of performance between LVQ and LVQGA in training data

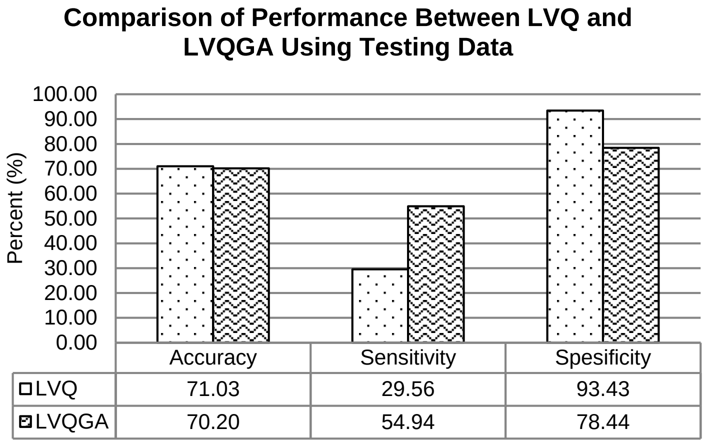


Figure 5. Comparisons of performance between LVQ and LVQGA in testing data

4. CONCLUSION

In the training data, the algorithm offered improves the accuracy of LVQ algorithm in the detection of diabetics. But in the testing data, LVQGA algorithm decreases LVQ accuracy. This means there has been over-fitting on the training data train process so that the model generated by LVQGA is too exclusive to the training data.

Whether on training data or on testing data, GA increases LVQ sensitivity level, but there is also a significant decrease in the level of specificity. This means the method offered increases LVQ's ability to recognize people affected by diabetes, but LVQ's ability to recognize people who are not affected by diabetes.

REFERENCES

[1] Alberti KGMM, Zimmet PF. “Definition, Diagnosis and Classification of Diabetes Mellitus and Its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus”. Provisional Report of a WHO Consultation. *Diabetic medicine*. 1998; 15(7): 539-553.

[2] Forbes JM, Cooper ME. “Mechanisms of Diabetic Complications”. *Physiological Reviews*. 2013; 93(1): 137-188.

[3] “InfoDATIN”. *Pusat Data dan Informasi Kementerian Kesehetan Republik Indonesia*. 2014.

[4] “Laporan Survei Kesehatan Rumah Tangga 2001”. *Badan Penelitian dan Pembangunan Kesehatan, Kementrian Kesehatan Republik Indonesia*. 2002.

[5] Chikh MA, Saidi M, Settouti N. “Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-Nearest Neighbor”. *Journal of Medical Systems*. 2012; 36(5): 2721-2729.

[6] Beheshti Z, Shamsuddin SMH, Beheshti E, Yuhaniz SS. “Enhancement of Artificial Neural Network Learning Using Centripetal Accelerated Particle Swarm Optimization for Medical Diseases Diagnosis”. *Soft Computing*. 2014; 18(11): 2253-2270.

[7] Iyer A, Jeyalatha S, Sumbaly R. 2015. “Diagnosis of Diabetes Using Classification Mining Techniques”. *International Journal of Data Mining & Knowledge Management Process*. 2015; 5(1): 1-14.

[8] Fausett L. “Fundamentals of Neural Networks: Architectures, Algorithms and Applications”. New Jersey: Prentice-Hall. 1994: 187.

[9] Hammer B. “Two or Three Thinks that We Do Not Know About Learning Vector Quantization but We Should Consider”. MIWOCI Workshop-2013. 2013: 6-12.

[10] Ho TK. “Recognition of Handwritten Digits by Combining Independent Learning Vector Quantizations”. *Proceedings of The Second International Conference on Document Analysis and Recognition*. Tsukuba Science City. 1993: 818-821.

[11] Mouy X, Bahoura M, Simard Y. “Automatic Recognition of Fin and Blue Whale Calls for Real-Time Monitoring in The St. Lawrence”. *The Journal of the Acoustical Society of America*. 2009; 126(6): 2918-2928.

[12] Ciosek P, Wróblewski W. “The Analysis of Sensor Array Data with Various Pattern Recognition Techniques”. *Sensors and Actuators B: Chemical*. 2006; 114(1): 85-93.

[13] Karegowda AG, Manjunath AS, Jayaram MA. “Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes”. *International Journal on Soft Computing*. 2011; 2(2): 15-23.

[14] Melin P, Herrera V, Romero D, Valdez F, Castillo O. “Genetic Optimization of Neural Networks for Person Recognition Based on the Iris”. *TELKOMNIKA, Telecommunication Computing Electronics and Control*. 2012; 10(2): 309-320.

[15] Liu S, Tai H, Ding Q, Li D, Xu L, Wei Y. “A Hybrid Approach of Support Vector Regression with Genetic Algorithm Optimization for Aquaculture Water Quality Prediction”. *Mathematical and Computer Modelling*. 2013; 58(3): 458-465.

[16] Singh S, Gill J. “Temporal Weather Prediction using Back Propagation based Genetic Algorithm Technique”. *International Journal of Intelligent Systems and Applications*. 2014; 6 (12): 55-61.

[17] WeiKoh J, Tan TS, EnChuah Z, Soh SS, Arif M, Leong K. “Genetic Algorithm Optimized Back Propagation Neural Network for Knee Osteoarthritis Classification”. *Research Journal of Applied Sciences, Engineering and Technology*. 2014; 8(16): 1787-1793.

[18] Sen O, Zhengxiang S, Jianhua W, Degui C. “Application of LVQ Neural Networks Combined with Genetic Algorithm in Power Quality Signals Classification”. *Proceedings of The International Conference on Power System Technology*. Kunming. 2002: 491-495.

[19] Wang JM, Wen YQ. “Application of Genetic LVQ Neural Network in Credit Analysis of Power Customer”. *Proceedings of Fourth International Conference on Natural Computation*. 2008: 305-309.

[20] Chen N, Ribeiro B, Vieira AS, Duarte J, Neves JC. “Hybrid Genetic Algorithm and Learning Vector Quantization Modeling for CostSensitive Bankruptcy Prediction”. *Proceedings of The Second International Conference on Machine Learning and Computing. Bangalore*. 2010: 213-217.

[21] Baker JE. “Reducing Bias and Inefficiency in The Selection Algorithm”. *Proceedings of The Second International Conference on Genetic Algorithms*. 1987: 14-21.