Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052059 doi:10.1088/1757-899X/546/5/052059

Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection

Nadisa Karina Putri[1], Zuherman Rustam[1*], Devvi Sarwinda[1]

[1]Department of Mathematics, Universitas Indonesia, Depok Indonesia

*Corresponding author: rustam@ui.ac.id

Abstract. Diabetes mellitus or commonly referred as diabetes is a metabolic disorder caused by high blood sugar level and the pancreas does not produce insulin effectively. Diabetes can lead to relentless disease such as blindness, kidney failure, and heart attacks. Early detection is needed in order for the patients to prevent the disease being more severe. According to the non-normality and huge dataset in medical data, some researchers use classification methods to predict symptoms or diagnose patients. In this study, Learning Vector Quantization (LVQ) is used to classify the diabetes dataset with Chi-Square for feature selection. The result of the experiment shows that the best accuracy is achieved at 80% and 90% of the data training and the performance measurement, which are precision, recall, and f1 score are the highest when the model contains all the features in the dataset.

1. Introduction

Diabetes is a metabolic disease that is becoming a serious problem around the world, caused by high blood sugar level as the pancreas does not effectively produce insulin, a hormone that controls levels of glucose in the blood [1]. Human cells need glucose for their energy, but unfortunately the cells cannot convert glucose into energy directly. Therefore, insulin is needed to help the cell absorbing the glucose from the blood. The glucose that has been converted into energy will be used for human activity or stored as fat.

According to the International Diabetes Federation, over 425 million people are currently affected worldwide, with severe development such as causing blindness, kidney failure, and heart attacks [2]. Although it is a difficult and incurable disease, it can be controlled, preventing it from developing. It is important for sufferers to be aware of the disease as early as possible, often by using several mathematic methods to predict it.

In medical research, the data presented are usually huge. Data Mining methods has an ability to visualize the data in medical fields which can later be used for various kinds of predictions. One of data mining techniques that is useful to categorize and predict symptoms in medical data is classification. The classification method can categorize the information in medical diagnostic field which is usually vague and incomplete, furthermore it will help much in decision making [3].

There are several classification methods that has been studied by researchers such as Support Vector Machine, Decision Tree, and Neural Network [4]. Neural network performs based on the work of human mind. It has an ability to extract pattern and detect a complex trend. LVQ is one of the neural network method that works by applying 'winner take all' strategy. In LVQ, the winning vector is vector that has the smallest Euclidean distance.

Research using data mining techniques that classify diabetes use feature selection, or applied Learning Vector Quantization method for prediction, some being: Sahan et al (2005) adopted Attribute Weight Artificial Immune System (AWAIS) in their work as collected by using 10-fold cross validation. It earned 75.87% accuracy for the Pima Indians Diabetes datasets [5]. D. Enachescu et al (2005) examined breast cancer prediction using LVQ and found that the accuracy for this prediction ranged from 77% to 100%. The result varies because it depends on the parameters used in the LVQ [6]. M. Sinecen et al (2009) proposed some Artificial Neural Networks (ANN) method for prostate cancer diagnosis. The result shown that between single hidden layer ANN, two hidden layer ANN, LVQ ANN, and RBF ANN, the FF2 ANN received the highest accuracy of 85.8% [7]. Kumari A, et al (2013) proposed a Radial based Kernel Support Vector Machine for diabetes data classification. They use the data from UCI Repository which is trained and tested using the SVM classifier. The model received 65.8% accuracy for the training data and 78.2% accuracy for those tested [8]. Saravananathan et al (2016) analyzed diabetic data using several classification methods so that they could compare which has the highest accuracy. The methods presented by this work are J48, k-Nearest Neighbor (k-NN), SVM, and Regression Tree CART. The proposed work used WEKA software to calculate the execution time and error rate. The result found that J48 method performed the best accuracy of 67.16% compared with others [3]. Sisodia et al (2018) used Naïve Bayes classification for early detection of diabetes disease. The highest accuracy from this method is at 76.7% [9]. Z. Rustam et al (2018) adopted SVM and Fuzzy C-Means for intrusion detection system. The research found that FCM performed a higher accuracy at 95.09% than the SVM method at 94.43% [10].

In this study, a learning vector quantization method is used to classify the diabetes dataset achieved from Kaggle online database [11]. The chi-square feature selection was employed and compared the accuracy for the features and selected all those that were relevant.

2. Methods

2.1. Data

The dataset used in this paper is retrieved from Kaggle's Diabetes Dataset [11]. The dataset contains 8 features with 1 class and 2001 instances. The features are described in Table 1.

Table 1. Description of The Features in Dataset

No.	Feature Name	Feature Abbreviation
1.	Pregnant	preg
2.	Plasma Glucose Concentration	gluc
3.	Blood Pressure	dbp
4.	Skin Thickness	sft
5.	Insulin	ins
6.	Body Mass Index (BMI)	bmi
7.	Diabetes Pedigree Function	dbf
8.	Age	age
9.	Diabetic or Non-Diabetic	cl

2.2. Chi-square Feature Selection

Chi-square feature selection is one of the filter method as a part of supervised feature selection [12]. In order to find the best features or the features order from the most prominent one, calculate the χ^2 score for each feature X, and begin by building Table 2. As seen in Table 2, the set of training set has 2 classes, which are 1 and 0. After that, calculate the expected value (E), for each P, Q, R, S using Equation 1.

$$E_P = (P + R)\frac{P + Q}{C} \tag{1}$$

Consequently, use Equation 4 to calculate the χ^2 score. Equation 4 is obtained from Equation 3 and Equation 2.

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k} \tag{2}$$

$$\chi^2 = \frac{(P - E_P)^2}{E_P} + \frac{\left(Q - E_Q\right)^2}{E_Q} + \frac{(R - E_R)^2}{E_R} + \frac{(S - E_S)^2}{E_S} \tag{3}$$

$$\chi^2 = \frac{C(PS - RQ)^2}{(P + R)(Q + S)(P + S)(R + Q)} \tag{4}$$

After calculating the χ^2 score, the features chosen are those with the maximum χ^2 score.

Table 2. Chi Square Feature Selection Table

	Class 1	Class 0	Total
X exist	P	Q	P+Q=B
X not exist	R	S	R+S=C-B
Total	P+R=A	Q+S=C-A	C

2.3. Learning Vector Quantization (LVQ)

LVQ is a derivative form of the artificial neural network which uses supervised learning and nearest neighbor pattern classifier [13]. It adopted competitive learning and has a similar architecture with the Kohonen Self Organizing Map (SOM) founded by Prof. Teuvo Kohonen in 1982. The basic concept of this method is to get as near as possible to the distribution of input vector in order to minimize the error of classification. This can be done by calculating the Euclidean distance between the input vector and weight vector [14]. The smallest Euclidean distance will be called as the winning vector, where the winning vector will be updated and continued until the termination condition [15]. For more details on the process, the algorithm of LVQ training can be described as follows:

Step 1: Initialized the initial weight vector with learning rate α

Step 2: For each input x, calculate the Euclidean distance and choose the winning vector with the minimum Euclidean distance using Equation 5.

$$D(j) = \sqrt{\sum_{i=1}^{n}\left(\mathbf{x_i} - \mathbf{w_{ij}}\right)^2} \tag{5}$$

with x as the input vector, w as the weight vector (winner), and n as the number of attributes [8].

Step 3: Update the weight vector using Equation 6 if the class in the in the neuron j is equal to the target

$$\mathbf{w_j}(new) = \mathbf{w_j}(old) + \alpha\left[x - \mathbf{w_j}(old)\right] \tag{6}$$

Step 4: Update the weight vector using Equation 7 if the class in the in the neuron j is not equal to the target

$$\mathbf{w_j}(new) = \mathbf{w_j}(old) - \alpha\left[x - \mathbf{w_j}(old)\right] \tag{7}$$

Step 5: Perform steps 3 and 4 for each input vector in the training

Step 6: Reduce α

Step 7: Until specified number of epoch is reached, repeat step 2 to 6
Step 8: Test for stopping condition

2.4. Statistical Measures

In this paper, we evaluate our proposed method with several measurements that can be described in Table 3 and Table 4.

Table 3. Confusion Matrix

Actual Vs. Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 4. Terminology of Statistical Measurement [5]

Name	Formula	Function
Accuracy (A)	$A = \frac{TP + TN}{TP + TN + FP + FN}$	Measurement of the algorithm in prediction
Precision (P)	$P = \frac{TP}{TP + FP}$	Measure classifier correctness
Recall (R)	$R = \frac{TP}{TP + FN}$	Measure classifier sensitivity or completeness
F1 Score (F1)	$F1 = 2 \times \frac{P \times R}{P + R}$	Measure the weighted average of the precision and recall

3. Experimental Results

The diabetes dataset contains 8 features with 1 class as described in Section 2. There are 2001 instances from the dataset. Before getting into the classification process, we perform chi-square feature selection in order to determine the features that are significant for the model. Later on, we evaluate the selected features with learning vector quantization classifier. From experimental result, we get the feature order based on the maximum value or the χ^2 score, where it can be seen in Table 5.

After applying the feature selection, the dataset is then classified using the Learning Vector Quantization based on the feature selected. The accuracy from each number of features are shown in Table 6. As illustrated, the accuracy is very high for the 80% and 90% of data training. From Table 6, we also evaluate our model with accuracy precision, recall, and F-1 score for 80% percentage of data training. In Figure 1, it shows the comparison of the model performance for 1 feature, 2 features, 3 features, 4 features, and all features without selection.

Based on the model performance shown in Figure 1, it can be observed that the model performance increase as the number of features increases. The highest performance is reached by 8 features, which means that model without feature selection. From this output, we can conclude that all the features in the dataset are significant for the prediction. With the chi-square feature selection, we may acknowledge the features order from the most significant to the least significant. Based on the χ^2 score, the features order is: Pregnant, Blood Pressure, Age, Skin Thickness, Glucose, Insulin, Body Mass Index (BMI), and Diabetes Pedigree Function.

Table 5. Chi Square Feature Selection

Number of Features	Selected Features
1	‘preg’
2	'preg', 'dbp'
3	'preg', 'dbp', 'age'
4	'preg', 'dbp', 'age', 'sft'
5	'preg', 'dbp', 'age', 'sft', 'gluc'
6	'preg', 'dbp', 'age', 'sft', 'gluc', 'ins'
7	'preg’, 'dbp', 'age', ‘sft', 'gluc', 'ins', 'bmi'
8 (without feature selection)	'preg’, 'dbp', 'age', ‘sft', 'gluc', 'ins', 'bmi', 'dbf'

Table 6. Accuracy based on % Data Training

% Data Training	Accuracy of *i* Features (%)							
	1	2	3	4	5	6	7	8
10	70.71	70.15	70.65	71.98	70.09	70.82	69.93	70.54
20	79.61	78.92	78.99	78.55	79.17	79.30	78.49	78.99
30	84.70	84.70	84.56	84.63	84.63	84.70	84.70	84.63
40	99.33	99.33	99.33	99.33	99.33	99.33	99.33	99.33
50	99.20	99.20	99.20	99.20	99.20	99.20	99.20	99.20
60	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00
70	98.66	98.66	98.66	98.66	98.66	98.66	98.66	98.66
80	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
90	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

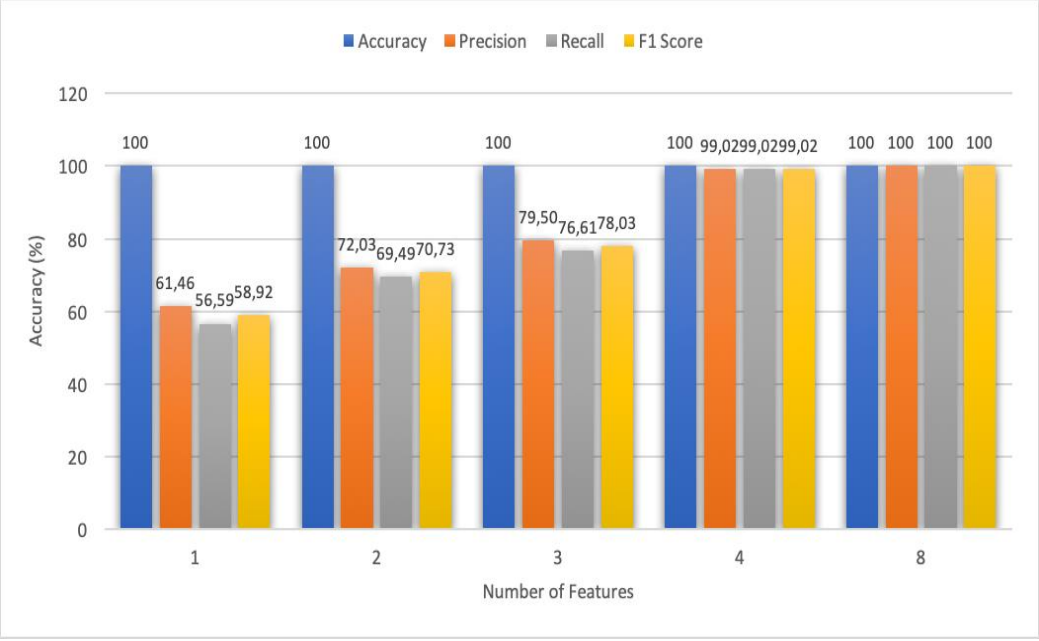

Figure 1. Model Performance

4. Conclusion

In this paper, we have developed Chi-Square feature selection and Learning Vector Quantization for the diabetes dataset. Chi-square is applied to recognize the feature's significance for the prediction of diabetes. It can also help to sort the features from the most prominent to the least. To evaluate the model, the performance indicators for this data are accuracy, sensitivity, recall or precision, and f1 score.

From the experimental result, we can conclude that the highest accuracy for diabetes dataset using chi-square feature selection and LVQ can be obtained at 80% and 90% data training. Because the accuracy is 100% for all features, the model presented is good for the dataset. By this result, the user of this method can choose how many features they want to use to get the desired accuracy.

Although LVQ has given good performance for diabetes data classification, LVQ still has both advantages and disadvantages. The advantages include summarizing large dataset into small size vector for classification, the model produced can be updated and adjusted gradually, and the algorithm and formula is relatively easy to understand. On the other hand, the disadvantages are the model needs to calculate distance for each attribute and the accuracy and performance is influenced by some parameters used including learning rate and epoch rate, and initial weight vector.

For future research, it is recommended to apply this method for another dataset or to predict another disease. It is considered to use the dataset with a lot of features so later on it can be tested if the feature selection is significant to increase the accuracy for the model. By doing this research, it is hoped that the result is better than the traditional methods and will be helpful across the medical fields.

Acknowledgement

This research was financially supported by University of Indonesia, with PIT. 9 2019 research grant scheme (ID number NKB-0039/UN2.R3.1/HKP.05.00/2019).

References

[1] World Health Organization 2018 *Diabetes Mellitus*, retrived from https://www.who.int/mediacentre/factsheets/fs138/en/ 14 February 2019
[2] International Diabetes Federation 2019 *Diabetes glossary,* retrived from https://idf.org/52-about-diabetes.html 14 February 2019
[3] K. Saravananathan & T. Velmurungan, Analyzing Diabetic Data using Classification Algorithms in Data Mining, *Indian Journal of Science and Technology*, Vol 9(43) (2016)
[4] J. Tang, S. Alelyani, & H. Liu, Feature Selection for Classification: A Review (2014)
[5] S. Sahan, K. Polat, H. Kodaz, & S. Gunes, The Medical Applications of Attribute Weighted Artificial Immune System (AWAIS): Diagnosis of Heart and Diabetes Diseases, *ICARIS 2005,* LNCS 3627, PP. 456-468 (2005)
[6] D. Enaschescu & C. Enaschescu, Learning Vector Quantization for Breat Cancer Prediction, *IEEE Portugese Conference on Artificial Intelligence* (2005)
[7] M. Sinecen, M. Cinar, O. Karal, M. Engin, Y.Z. Atesci, M. Makinaci, & B. Cakmak, Diagnosis of Prostate Cancer Using Artificial Neural Networks, *14th National Biomedical Engineering Meeting 10746387* (2009)
[8] V. Anuja Kumari & R. Chitra, Classification of Diabetes Disease Using Support Vector Machine, *International Journal of Engineering Research and Application (IJERA)*, Vol 3, pp. 1797-1801 (2013)
[9] D. Sisodia & D. Singh Sisodia, Prediction of Diabetes using Classification Algorithms, *International Conference on Computational Intelligence and Data Science (ICCIDS 2018),* 132(2018) 1278-1585 (2018)
[10] Z. Rustam & NPAA. Ariantari, Comparison Betweem Support Vector Machine and Fuzzy Kernel C-Means as Classifiers for Intrusion Detection System Using Chi-Square Feature Selection, *AIP Conference Proceedings 2023 (1), 020214* (2018)
[11] Kaggle 2019 *Diabetes Dataset,* retrived from https://www.kaggle.com/johndasilva/diabetes 17 February 2018

[12] H. Budak & S. Erpolat Tasabat, A Modified T-Score for Feature Selection, *Anadolu University Journal of Science and Technology,* Vol. 17, pp. 845-852 (2016)
[13] V. Badbe, V. Londhe, & G. Shirole, Analysis of Heart Disease by LVQ in Neural Network, *International Journal on Recent and Innovation Trends in Computing and Communication,* Vol. 4, pp. 603-607 (2016)
[14] J. S. Sonawane & D. R. Patil, Prediction of Heart Disease Using Learning Vector Quantization Algorithm, IEEE *Conference on IT in Business, Industry, and Government (CSIBIG)* (2014)
[15] A. Dongoran, S. Rahmadani, M. Zarlis, & Zakarias, Feature Weighting Using Particle Swarm Optimization for Learning Vector Quantization Classifier, *2nd International Conference on Computing and Applied Informatics,* Series 978 (2018)
[16] Z. Rustam & Rika, Face Recognition Using Fuzzy Kernel Learninf Vector Quantization, *IOP Conf. Series: Journal of Physics,* Series 1108 (2018)
[17] Z. Rustam & AS Talita, Fuzzy Kernel C-Means Algoritm for Intrusion Detection Systems, *Journal of Theoretical & Applied Information Technology 81(1),* (2015)
[18] S. Rakshit, S. Manna, S. Biwas, R. Kundu, et al, Prediction of Diabetes Type-II Using A Two-Class Neural Network, *International Conference on Computational Intelligence, Communications, and Business Analytics (CICBA),* pp. 65-71 (2017)
[19] D. B. Manurung, B. Dirgantoro, & C. Setianingsih, Speaker Recognition for Digital Forensic Audio Analysis Using Learning Vector Quantization Method, *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (2018)*
[20] A. K. Dewangan & P. Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques, *International Journal of Engineering and Applied Sciences (IJEAS),* Vol. 2 (2015)
[21] M. Alehegn, R. Joshi, & Dr. P. Mulay, Analysis and Prediction of Diabetes Mellitus Using Machine Learning Algorithm, *International Journal of Pure and Applied Mathematics,* Vol. 118, pp. 871-878 (2018)
[22] Md. Maniruzzaman, N. Kumar, Md. M. Abedin, et al, Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Program, *Computer Methods and Program in Biomedicine,* pp. 23-34 (2017)
[23] N. Barakat, A. P. Bradley, & M. N. Barakat, Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus, *IEEE Engineering in Medical and Biology Society* (2010)
[24] S. Bahassine, A. Madani, M. Al-Sarem, &M. Kissi, Feature Selection Using an Improved Chi-square for Arabic Text Classification, *Journal of King Saud University – Computer and Information Sciences* (2018)
[25] Z. Rustam & T.V. Rampisela, Classification of Schizophrenia Data Using Support Vector Machine (SVM), *Journal of Physics: Conference Series 1108(1), 012038* (2018)