Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification

F Maulidina¹, Z Rustam¹, S Hartini¹, V V P Wibowo¹, I Wirasati¹ and W Sadewo²

¹ Departement of Mathematics, University of Indonesia, Depok 16424, Indonesia
² Department of Neurosurgery, Faculty of Medicine, University of Indonesia, Dr. Cipto Mangunkusumo National General Hospital, Indonesia

E-mail: faisa.maulidina@ui.ac.id

Abstract. Diabetes is a disease that occurs when the blood glucose level is higher than normal and also leads to health problems. Early and accurate diagnosis needs to be carried out on individuals affected by this disease. Furthermore, excellent treatment needs to be provided to prevent worse situations. Some studies have used several machine learning methods to diagnose diabetes. Furthermore, in this study, the Backward Elimination and Support Vector Machine (SVM) algorithm was used to classify the PIMA Indians diabetes dataset. It consisted of 268 diabetic and 500 non-diabetic patients with eight attributes. Backward Elimination is a feature selection method used to remove irrelevant features based on the linear regression model. Using this method, the right features for the model was expected. This method has some advantages which include increasing training time, decreasing complexity and improving performance and accuracy. Therefore, the performance of SVM improved. Based on the experiments, it was discovered that by combining feature selection algorithm (backward elimination) and SVM, the highest accuracy obtained was 85.71% using 90% data training. Therefore, it was concluded that Backward Elimination combined with SVM algorithm is an excellent method to classify diabetes by using the PIMA Indians diabetes dataset.

1. Introduction

Diabetes is a disease that occurs when the blood glucose level is higher than normal. This is due to the inability of the body to release insulin, a hormone produced by the pancreas, effectively [1]. Each individual has different levels of glucose, which increases after every meal and returns back to normal within two hours. This disease causes health problems, including heart and kidney disease, loss of vision, stroke and lower limb amputation [2].

The World Health Organization (WHO) stated that about 422 million worldwide have diabetes and 1.6 million deaths are attributed to this disease each year [3]. Furthermore, the number of cases has been increasing each year. Despite its huge impact, there is currently no cure, but it can be prevented and managed. Therefore, it is important to classify the number of individuals with this disease.

In the medical field, classification is used to classify a patient’s illness, which makes it easier for doctors to give an accurate therapeutic solution. Therefore, several methods in data mining such as Decision Tree, Neural Network and Support Vector Machine have been applied as classifiers [4].
Support Vector Machine uses a technique called Kernel to transform the given data. It is used to discover the optimal boundary between the possible outputs and captures many complex relationships between the data points.

Diabetes classification by using the Pima Indians diabetes dataset has been carried out with various methods, including SVM which obtained an accuracy of 82.2% [5]. N K Putri, Z Rustam and D Sarwinda used Learning Vector Quantization and Chi-Square feature selection to classify diabetes and obtained an accuracy of 100% using 80% and 90% training data [6]. The other method that has also been used with this dataset is K-Nearest Neighbour (KNN) and it obtained an accuracy of 91% [7]. One disadvantage of the KNN method is determining the optimal k parameters. High k values reduces the effect of noise on classification, but makes boundaries between each class less distinct, while k values that are too low causes lesser comparison sample which leads to reduced accuracy [8]. T Nadira and Z Rustam used SVM and feature selection to classify cancer and the accuracy was 99.9999% and 96.4286% on lung and breast cancer dataset respectively [9]. Arfiani, Z Rustam, J Pandelaki and A Siahaan used SVM for acute sinusitis classification and obtained an accuracy of 90% [10]. Z Rustam and Rampisela T V also used SVM to classify Schizophrenia data and obtained an accuracy of 90.1% [11].

Therefore, to classify this diabetes dataset [12], Support Vector Machine (SVM) method was used. The backward elimination selection feature was also included in this study to select all the relevant features in other to obtain a higher accuracy.

2. Material and methods

2.1. Material

2.1.1 Dataset. The dataset was obtained from the Kaggle’s Diabetes Dataset, which was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It consisted of 768 instances with eight features and two class. This dataset consisted of 268 diabetic and 500 non-diabetic patients. The target type of this dataset is diabetic or non-diabetic. The features are shown in table 1.

| Feature Name                  | Feature Symbol |
|-------------------------------|----------------|
| Pregnancies                   | preg           |
| Plasma Glucose Concentration  | pgc            |
| Blood Pressure                | bps            |
| Skin Thickness                | sft            |
| Insulin                       | ins            |
| Body Mass Index (BMI)         | bmi            |
| Diabetes Pedigree Function    | dbf            |
| Age                           | age            |
| Diabetic or Non Diabetic      | class          |

2.2. Methods

2.2.1. Backward Elimination. Backward elimination is a feature selection method that has been used in medical studies [13]. It will removes the least important variables first, and leaves only the most important variables in the model. Firstly, all the features are (variables) to be tested in a regression
model, with a significance level of 0.05. When the p-value of a feature is greater than the significance level (P-value > 0.05), it is eliminated.

This step is repeated until all features become significant (P-value < 0.05). Lastly, the model is fitted with the new set of features. This process is seen in table 2.

In addition, this method has some advantages which include increasing training time, decreasing complexity and improving performance and accuracy. Therefore, this method was useful to select the features before evaluating the model.

![Backward elimination process](image)

**Figure 1.** Backward elimination process

2.2.2. Support Vector Machine. The SVM concept is useful in finding the best hyperplane that plays the role of a separator between two classes. Figure 2(a) draws several patterns which are members positive (+1) and negative (−1) class. Patterns that are joined in the positive class are represented by circles, while the negative class are represented by squares. The variety of the dividing lines is shown in figure 2 (a).
Measuring the hyperplane margin and finding the maximum point should be carried out to obtain the best hyperplane separator between two classes. The distance between the hyperplane and the closest data from each class is called a margin. A support vector is the closest subset of the training data sets. A hyperplane is said to be optimum when it is located in the middle of two classes. Therefore, it has the farthest distance to the outer data in both classes. The best hyperplane shows in figure 2(b) which located in the middle of two classes. Support vectors are the points represented by squares and circles inside the black circle. The concept of SVM is to find a hyperplane that separates data sets into two classes \[13\]. The existing data is denoted by \(\mathbf{x}_i \in \mathbb{R}^d\), while its label is respectively denoted by \(y_i \in \{-1, +1\}\) for \(i = 1, 2, \ldots, l\), where \(i\) is the amount of data. Assuming that class -1 and +1 can be separated by a hyperplane without dimensions, which is determined by:

\[
\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (1)
\]

A pattern \(\mathbf{x}_i\) that belongs to class -1 is defined by:

\[
\mathbf{w} \cdot \mathbf{x} + b \leq 1 \quad (2)
\]

while pattern \(\mathbf{x}_i\) which belongs to class +1 is defined by:

\[
\mathbf{w} \cdot \mathbf{x} + b \geq 1 \quad (3)
\]

The maximum margin was found by maximizing the value of the distance between the hyperplane and its closest point, which is \(\frac{1}{\|\mathbf{w}\|}\). From finding the minimum point of the equation (4) and noting the constraints of equation (5), a Quadratic Programming (QP) problem was formulated.

\[
\min \tau(w) = \frac{1}{2}\|w\|^2 \quad (4)
\]

\[
y_i (\mathbf{x}_i, \mathbf{w} + b) = \frac{1}{2}\|w\|^2 \quad (5)
\]

This problem can be solved by several computational techniques, such as Lagrange Multiplier as shown in equation (6):
\[
L(\vec{w}, b, \alpha) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i((\vec{x}, \vec{w} + b) - 1))
\] (6)

\(\alpha_i\) is a Lagrange multiplier, whose value is zero or positive \((\alpha_i \geq 0)\). By minimizing \(L\) with respect to \(w\) and \(b\), and maximizing \(L\) with respect to \(\alpha_i\), the optimal value of equation (6) was calculated. It was modified as the maximization of problems that only contain \(\alpha_i\) by considering the optimal point \(\nabla L = 0\), as shown in equation (7).

Maximize:

\[
\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \bar{x}_i \bar{x}_j
\] (7)

Subject to:

\[
\alpha_i \geq 0 (i = 1, 2, \ldots l) \sum_{i=1}^{l} \alpha_i y_i = 0
\] (8)

From this calculation, \(\alpha_i\) obtained was mostly positive. This positive correlated data with \(\alpha_i\) is called support vector [14].

To measure accuracy, the data was classified using SVM. Firstly, it was split into two sets, namely training and testing. The training data was used to find the optimal hyperplane in SVM. Furthermore, the model with the training process and accuracy was calculated using the Confusion Matrix which is seen in table 2.

**Table 2.** Confusion Matrix

| Actual vs Predicted | Positive | Negative |
|---------------------|----------|----------|
| Positive            | TP       | FP       |
| Negative            | TN       | FN       |

Classification accuracy formula is written such as:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

3. Results and discussion

The diabetes data set consisted of 768 instances, eight features with two class. Firstly, the backward elimination method was used to select the features that were significant for the model. Furthermore, Support Vector Machine was used to evaluate the selected features. From the results, the significant features were obtained based on the p-value of the variables as show in table 3.

**Table 3.** Description of the selected features in diabetes dataset

| Feature Name | Feature Symbol |
|--------------|----------------|
|              |                |
Plasma Glucose Concentration pgc
Insulin ins
Body Mass Index (BMI) bmi
Diabetes Pedigree Function dbf
Age age
Diabetic or Non Diabetic class

This was a training test where all training data used were test data. After feature selection was applied, the dataset was classified using Support Vector Machine method with selected features. The accuracy obtained using 10% until 90% data training is shown in Table 1. From the result, it was seen that the highest accuracy obtained when using 90% data training was 85.71%, while the lowest accuracy value obtained when using 10% training data was 76.44%. For SVM without feature selection, the highest accuracy obtained when using 90% of data training was 84.41%, while the lowest accuracy obtained when using 20% data training was 75.77%.

Comparisons were made to discover the better method between SVM with and without backward elimination. It was concluded that a higher accuracy was obtained when SVM with backward elimination was used, which was 85.71%. The comparison between SVM with and without backward elimination is seen in the Table below.

Table 4. Comparison of the accuracy between SVM with and without feature selection

| Data Training | Accuracy with feature selection | Accuracy without feature selection |
|---------------|---------------------------------|-----------------------------------|
| 10%           | 76.44%                          | 76.01%                            |
| 20%           | 77.23%                          | 75.77%                            |
| 30%           | 78.06%                          | 76.39%                            |
| 40%           | 76.78%                          | 76.13%                            |
| 50%           | 77.86%                          | 75.78%                            |
| 60%           | 78.57%                          | 76.62%                            |
| 70%           | 77.48%                          | 76.62%                            |
| 80%           | 81.16%                          | 79.22%                            |
| 90%           | 85.71%                          | 84.41%                            |

4. Conclusion
In this study, Backward Elimination for feature selection and SVM method for classifying the model was used to obtain the best accuracy value. Backward Elimination was used to select the significant features for the model, based on the p-value from each feature. The performance indicator to evaluate the model for this dataset was accuracy.

Backward Elimination method with the SVM algorithm had a higher accuracy value compared to using only the SVM method. The accuracy for the SVM algorithm was 84.41%, while the Backward Elimination accuracy value with the SVM Algorithm was 85.71%, when using 90% data training for both methods.

Although SVM has shown a good performance to classify diabetes, it still has some advantages and disadvantages. SVM performs better when there is a clear margin of separation between classes and it is relatively efficient in memory. But during the test phase, it may be very slow and problematic to overcome. Furthermore, when the dataset becomes noisy, i.e. target classes are overlapping, SVM does not perform well.

For upcoming studies, the use of another method for this data set is recommended for a better performance. Also, this method may be applied to predict another disease. In addition, using a dataset with many features should be considered to ascertain whether feature selection is important to improve
the accuracy during the test. From this proposed study, it is expected that this result is better compared to other methods and will benefit the medical field.

**Acknowledgements**

This research is fully supported financially by The Indonesian Ministry of Research, Technology, with a KEMENRISTEK/BRIM PDUPT 2021 research grant scheme.

**References**

[1] National Institute of Diabetes and Digestive and Kidney Diseases(NIDDK). *Diabetes*, retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[2] U.S. Department of Health and Human Services (DHHS). *Diabetes detection initiative: Finding the undiagnosed*, retrieved from http://www.ndep.nih.gov/ddi/about/index.htm.

[3] World Health Organization. *Diabetes*, retrieved from https://www.who.int/health-topics/diabetes#tab=tab_1

[4] Tang J, Alelyani S and Liu H 2014 Feature selection for classification: A review *Data Classification: Algorithms and Applications* CRC Press 37-64

[5] S Karatsiolis and C N Schizas 2012 Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset *IEEE 12th Int. Conf. Bioinform. & Biomed. Eng. (BIBE)* Larnaca 139-144

[6] Putri, NK, Rustam Z and Sarwinda D 2019, Learning Vector Quantization for Diabetes Data Classification with Chi-Square Feature Selection *IOP Conf. Ser.: Mater. Sci. and Eng.* 546 052059

[7] Nurhayati and A N 2014 Implementation of Naive Bayes and K-nearest neighbor algorithm for diagnosis of diabetes mellitus *Proc. of the 13th Int. Conference on Applied Computer and Applied Computation Science* 117-120

[8] Vinoth R et al 2014 A Hybrid Text Classification Approach Using KNN and SVM *International Journal of Advance Foundation and Research in Computer (IJAFRC)* 1 3

[9] Nadira T and Rustam Z 2018 Classification of cancer data using support vector machines with features selection method based on global artificial bee colony *Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Science, AIP Conf. Proc.* 2023

[10] Arfiani, Rustam Z, Pandelaki J and Siahaan A 2019 Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification *IOP Conf. Ser.: Mater. Sci. and Eng.* 546 052011

[11] Rampisela T V and Rustam Z 2018 Classification of Schizophrenia Data Using Support Vector Machine (SVM) *J. Phys.: Conf. Ser.* 1108 012044

[12] Kaggle 2016 Diabetes Dataset, retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database

[13] Steyerberg Ewout 2009 *Clinical prediction models* (New York: Springer International Publishing)

[14] Christianini Nello and John S Taylor 2000 *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge: Cambridge University Press)