Detection of Type 2 Diabetes Mellitus Disease with Data Mining Approach Using Support Vector Machine

Bayu Adhi Tama[1], Afriyan Firdaus[2], Rodiyatul FS[3]
[1,2]*Faculty of Computer Science, University of Sriwijaya*
[1]*bayu@unsri.ac.id,* [2]*afriyan_firdaus@yahoo.com,* [3]*ebecedeh@yahoo.co.id*

Abstract

Diabetes is a chronic disease and a major problem of morbidity and mortality in developing countries. The International Diabetes Federation (IDF) estimates that 285 million people around the world have diabetes. This total is expected to rise to 438 million within 20 years. Type 2 diabetes (TTD) is the most common type of diabetes and accounts for 90-95% of all diabetes. Detection of TTD from various factors or symptoms became an issue which was not free from false presumptions accompanied by unpredictable effects. According to this context, data mining could be used as an alternative way, help us in knowledge discovery from data. This paper utilize support vector machine (SVM) in the data mining process to acquire information from historical data of patient medical records. It offers a decision-making support through early detection of TTD for physicians and others.

Keywords: data mining, medical record, support vector machine, type 2 diabetes mellitus.

1. Introduction

Diabetes is an illness which occurs as a result of problems with the production and supply of insulin in the body [1]. People with diabetes have high level of glucose or "high blood sugar" called *hyperglycaemia*. This leads to serious long-term complications such as eye disease, kidney disease, nerve disease, disease of the circulatory system, and amputation thas is not the result of an accident.

Diabetes also imposes a large economic impact on the national healthcare system. Healthcare expenditures on diabetes will account for 11.6% of the total healthcare expenditure in the world in 2010. About 95% of the countries covered in this report will spend 5% or more, and about 80% of the countries will spend between 5% and 13% of their total healthcare dollars on diabetes [2].

TTD is the most common type of diabetes and accounts for 90-95% of all diabetes patients and most common in people older than 45 who are overwieght. However, as a consequence of increased obesity among the young, it is becoming more common in children and young adults [1].

Diabates has no obvious clinical symptoms and not been easy to know, so that many diabetes patient unable to obtain the right diagnosis and the treatment. Therefore, it is important to take the early detection, prevent and treat diabetes disease, especially for TTD.

Recent studies by the National Institute of Diabetes and Digestive and Kidney Diseases (DCCT) in UK have shown that effective control of blood sugar level is beneficial in preventing and delaying the progression of complications of diabetes [3]. Data mining techniques could be used as an alternative way in discovering knowledge from the patient medical records and SVMs has shown remarkable success in the area of employing Computer Aided Diagnostic systems (CAD) as a "second opinion" to improve diagnostic decisions [4]. SVMs also have demonstrated highly competitive performance in numerous real-world application such medical diagnosis, SVMs as one of the most popular, state-of-the-art data mining tools for knowledge discovery and data mining [9].

Several studies have been conducted regarding TTD detection. Rule extraction from SVMs has been conducted by Barakat [5], an experts system based on pricipal component analysis (PCA) and adaptive neuro-fuzzy inference systems, Polat and Gunes reported in [6], In [7] Yu et al combined Quantum Particle Swarm Optimization (QPSO) and Weighted Least Square (WLS)-SVM to diagnose type 2 diabates, whereas Huang et al used complementary of three classification techniques such as Naive Bayes, C4.5, and IB1 can be found in [8].

This paper is organized as follow: a brief explained of support vector machine and medical data used is provided in section 2. The detailed information is given for each subsections. Section 3 gives result and discusion, and finally, in section 4 we conclude the paper with summarization of the result by emphasizing this study and also mentioning for future research.

2. Material and Method

2.1. Datasets Used

This datasets was taken from one of public hospital in Palembang, South Sumatera. All patients of this database are men and women at least 10 years old. The variable takes the value "TRUE" and "FALSE", where "TRUE" means a positive test for TTD and "FALSE" means a negative test for TTD.

There are 185 cases, where 71,9% (133) cases in class "TRUE" and 17,8% (53) cases in class "FALSE". There are ten clinical attributes: (1) Gender (male, female), (2) Body mass (thin, medium, overweight), (3) Blood pressure (< 140/90, ≥ 140/90), (4) Hyperlipidemia (true, false), (5) Fasting blood sugar (FBS) (< 126 mg/dl, ≥ 126 mg/dl) , (6) Instant blood sugar (< 200 mg/dl, ≥ 200 mg/dl), (7) Case history (true, false), (8) Diabetes Gest history (true, false), (8) Habitual Smoker (true, false), (9) Plasmainsulin (high, low), and (10) Age (children, adult, old).

2.2. Support Vector Machine

Support vector machine (SVMs) are supervised learning methods that generate input-output mapping functions from a set of labeled training datasets. The mapping function can be either a classifiaction function or a regression function [9]. According to Vapnik [10], SVMs has strategy to find the best hyperplane on input space called the structural minimization principle from statistical learning theory.

Given the training datasets of the form $\{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}$ where c_i is either 1 ("yes") or 0 ("no"), an SVM finds the optimal separating hyperplane with the largest margin. Equation (1) and (2) represents the separating hyperplanes in the case of separable datasets.

$$w.x_i+b \geq +1, \text{ for } c_i = +1 \quad (1)$$

$$w.x_i+b \leq -1, \text{ for } c_i = -1 \quad (2)$$

The problem is to minimize $|w|$ subject to constraint (1). This is called constrainted quadratic programming (QP) optimization problem represented by:

$$\text{minimize } (1/2)\ \|w\|^2$$

$$\text{subject to } c_i(w.x_i - b) \geq 1 \quad (3)$$

The optimal value for the SVM can be shown to be:

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i,j} \alpha_i \alpha_j c_i c_j x_i^T x_j \quad (4)$$

where the α constitutes a dual representation for the weight vector in terms of the training set:

$$w = \sum_i \alpha_i c_i x_i \quad (5)$$

So that support vector α_i has a positive value.

2.2. Performance Evaluation

2.2.1 Performance Metrics

Error can be measured in different ways. Error refers to difference between the output of the systems and the desired response. In classification problems, performance measurement can be represented with confusion matrix as Figure 1 [9].

2.2.2 Estimation Model

Estimating the model can be used to estimate its future prediction accuracy. The simple method is holdout, which partitions the data into two mutually exclusive subsets called training set and test set (aka holdout set) [9].

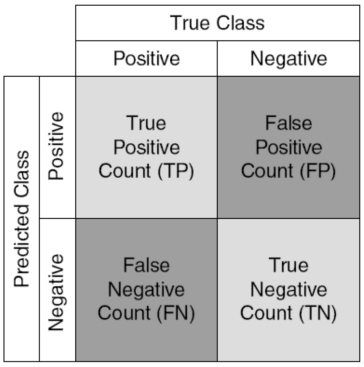

Figure 1. Simple confusion matrix

In order to minimize the bias associated with training and holdout data, one can use methodology called *k*-fold cross validation. In *k*-fold cross validation, the complete datasets is split into *k* subsets with equal size and then the model is trained and tested *k* times. The cross validation accuracy (CVA) is defined as [9]:

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i \quad (6)$$

3. Result and Discussion

The experiment compares performance of SVM with 4 other classical methods such as k-Nearest Neighbour classifier, neural network, C4.5 Decision Tree, and instance based learner (IB1). This SVM uses polynomial as kernel method with C = 250007 and E = 1. The classifier is validated using *k*-fold cross validation with *k* = 10.

The following graph (Figure 2) shows performance of each classifiers in percent. Here, SVM performs the best among other methods with accuracy 97,3%. Detail accuracy of SVM can be shown as Table 1. *k*-Nearest Neighbour with *k* = 10

has the same worst performance with C4.5 Decision Tree which has accuracy 91,89%. Surprisingly, IB1 has also the same worst performance with accuracy 91, 89%. This result opposes to Huang et al [8] which stated that IB1 performed the best among the classical methods (e.g. C4.5 Decision Tree).

Table 1. Detail accuracy of SVM

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
TRUE	0.99	0.08	0.971	0.992	0.981
FALSE	0.92	0.01	0.98	0.923	0.95
Avg	0.97	0.06	0.973	0.973	0.973

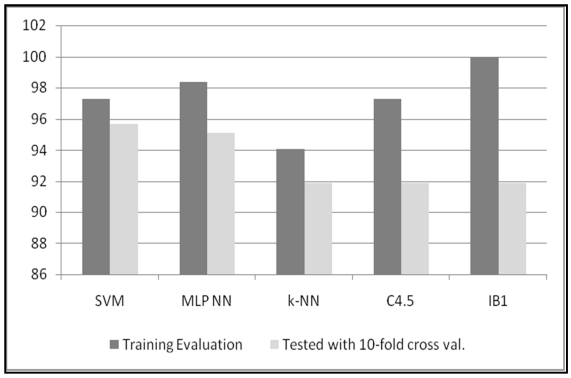

Figure 2. Performance classifiers

Confusion matrix (Table 2) shows that SVM performs most correct classification with class label TRUE is correctly classified as positive test for TTD (131 cases) and only 2 cases are misclassified. However, class label FALSE is also correctly classified as negative test for TTD (46 cases) and only 6 cases are misclassified. It can be concluded that SVM correctly classifies 177 cases of 185 cases.

Table 2. SVM confusion matrix

	TRUE	FALSE
TRUE	**131**	2
FALSE	6	**46**

Rules are extracted using decision tree (C4.5) after trained with SVM [5] as follow:

IF plasmainsulin = high AND
fasting blood sugar ≥ 126 THEN
Class = TRUE (116 cases)

IF plasmainsulin = high AND
fasting blood sugar < 126 AND
family history = true AND
instant blood sugar ≥ 200 THEN
Class = TRUE

IF plasmainsulin = low AND
fasting blood sugar < 126 THEN
Class = FALSE

4. Conclusion

This paper collects and analyzes medical patient record of type 2 diabetes mellitus (TTD) with knowledge discovery techniques to extract the information from TTD patient in one of public hospital in Palembang, South Sumatera.

The experiment has successfully performed with data mining technique. Support vector machine (SVM) as part of data mining technique achieves better performance than other classical methods such as neural network, C4.5, k-nearest neighbour, and IB1. Rules are extracted using decision tree can be used by physician to diagnose TTD disease.

Finally, this experiment is being optimised and later it will focus on increasing the datasets in order to maximized result.

5. References

[1] International Diabetes Federation (IDF), *What is diabetes?*, World Health Organisation, accessed January 2010, http://www.idf.org

[2] Zang Ping, et al. *Economic Impact of* Diabetes, International Diabetes Federation, accessed January 2010, http://www.diabetesatlas.org/sites/default/files/Economic%20impact%20of%20Diabetes.pdf

[3] National Diabetes Information Clearinghouse (NDIC), *The Diabetes Control and Complications Trial and Follow-up Study*, accessed January 2010, http://diabetes.niddk.nih.gov/dm/pubs/control

[4] N. Lavrac, E. Keravnou, and B. Zupan, "Intelligent Data Analysis in Medicine," in *Encyclopedia of Computer Science and Technology*, vol.42, New York: Dekker, 2000

[5] Barakat, et al. "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," *IEEE Transactions on Information Technology in BioMedicine*, 2009.

[6] Polat, Kemal and Salih Gunes. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Expert System with Applications*, pp. 702-710, Elseivier, 2007.

[7] Yue, et al. "An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLS-SVM," *International Symposium on Intelligent Information Technology Application Workshops*, IEEE Computer Society, 2008

[8] Huang, et al. "Feature selection and classification model construction on type 2 diabetic patients' data",

Journal of Artificial Intelligence in Medicine, pp 251-262, Elsevier, 2008.

[9] Olson, David L and Dursun Dulen. *Advanced Data Mining Techniques*, Springer Verlag, Berlin, 2008.

[10] Vapnik, V. *Statistical Learning Theory*, Wiley, 1998.