Comparative Experiments for Classification of Diabetes Mellitus

Siti Aida Fatimah Madon¹, Aida Mustapha¹, Mohd Zainuri Saringat¹, Mohd Helmy Abd Wahab², Syed Zulkarnain Syed Idrus³

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
² Faculty Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400 Batu Pahat, Johor, Malaysia
³ Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Jalan Wang Ulu Arau, Sg. Chuchuh, 01000 Kangar, Perlis

E-mail: aidam@uthm.edu.my

Abstract. Diabetes is an affecting people disease nowadays. Over 246 million people worldwide with a majority of them being women had been affected by diabetes. According to the World Health Organization (WHO), by 2025, this number is expected to increase to over 380 million. The disease has been ranked as the fifth deadliest disease in United States with no imminent cure in sight. Along with the increasing of the information technology and its continued advent into the medical and healthcare sector, the cases of this disease and the symptoms are well documented. This paper aims at finding the best performance by four classification algorithms, which are Naive Bayes, Simple Logistics, REPTree, and Sequential Minimal Optimization (SMO). Model testing on 200 tuples with 9 attributes in the diabetes dataset revealed that Naive Bayes achieved highest accuracy of 85%. To improve the overall accuracy, it is necessary to use more data set with larger number of attributes and use a better feature selection method in future works.

1. Introduction

Diabetes or its medical term “diabetes mellitus” is a medical condition where the level of blood sugar is above normal between 4.4-6.1 mmol/L. Diabetes mellitus is non-insulin dependent, whereby a diabetic patient has low production of insulin in their body. This is essentially the task of the pancreas, which is supposed to produce sufficient insulin to the body. But when various organs in the body begin to resist insulin, body system will demand for more production of insulin while the pancreas cannot cater. This abnormal condition comes with frequent urination, increased thirst, and increased hunger. Diabetes must be treated because it may lead to ketoacidosis, nonketotic hyperosmolar coma or even worse, death. Apart from that, patients suffering diabetics will also face long term complications such as heart disease, stroke, chronic kidney failure, foot ulcers, and eyes damaging.

Considering the importance of early medical diagnosis of diabetes, this paper is set to perform a comparative experiment on data mining techniques for prediction of diabetes. Early detection of diabetes can also help in avoiding further complications [1]. To date, various data mining algorithms have been reported in the literature for diabetes classification such as Decision Trees [2, 3, 4], Neural Networks [1], Hierarchical Clustering [5], Random Forest [6], and Naive Bayes [7].
The remaining of this paper proceeds as follows. Section 2 presents the materials and methods of performing the comparative experiments including the algorithm and dataset along with the evaluation metrics. Section 3 presents the results and finally Section 4 concludes with some direction for future work.

2. Materials and Methods

The paper performs comparative experiments on four classification algorithms in predicting positive cases of diabetes using Naive Bayes, REPTree, Simple Logistic, and Sequential minimal optimization (SMO). The experiment is based on the well-known Knowledge Discovery in Databases as shown in Figure 1 [8].

![Figure 1. KDD Methodology [8]](image)

The following subsections deliberates on the dataset, algorithms, and evaluation metrics used for this experiment.

2.1. Dataset

The classification experiment in predicting positive cases of diabetes is applied to the Pima Indians Diabetes Database of National Institute obtained from the University of California at Irvine (UCI) machine learning repository [9]. The data which is used in this project has records of 200 total instances. The datasets do not have any missing values and it does not contain any inconsistent value. Table 1 shows a brief description of the attribute that is being considered.

| Description                          | Feature Name |
|--------------------------------------|--------------|
| Number of times pregnant             | Preg         |
| Plasma glucose concentration         | Plas         |
| Diastolic blood pressure (mm Hg)     | Pres         |
| Triceps skin fold thickness (mm)     | Skin         |
| 2-Hour serum insulin                 | Insu         |
| Body mass index (kg/m)               | Mass         |
| Diabetes pedigree function           | Pedi         |
| Age (years)                          | Age          |
| Class variable (0 or 1)              | Class        |

Figure 2 shows the data distribution for type to determine if the patient is affected by the diabetes. The blue bar graph is the class not affected by diabetes (Negative) and red bar graph is
the class affected by diabetes (Positive). Only 68 patients affected by diabetes and 132 survived from the disease.

![Figure 2. Class Distribution](image)

2.2. Algorithms

The present work intends to create a mining model based on four classification algorithms in order to provide a simpler solution. The results have been analyzed using statistical methods.

- **Naive Bayes**: The Naive Bayes Algorithm is a probabilistic algorithm that is sequential in nature, following steps of execution, classification, estimation and prediction. Naive Bayes overcomes various limitations including omission of complex iterative estimations of the parameter and can be applied on a large dataset in real time. The equation for Naive Bayes is shown in Equation 1, where \( p(C_k|x) \) is the posterior probability for a feature vector \( x = (x_1, \ldots, x_n) \) and \( k \) possible outcomes, \( p(C_k) \) is the prior probability, \( p(x|C_k) \) is the likelihood probability, and \( p(x) \) is the evidence.

\[
p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{1}
\]

- **Sequential Minimal Optimization (SMO)**: SMO is a sequential minimal optimization algorithm for training a support vector classifier [10]. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. SMO works by breaking down the dual form of the SVM optimization problem as shown in Equation 2 into many smaller optimization problems that are more easily solvable,

\[
f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} < x^{(i)}, x > + b \tag{2}
\]

where:
- \( m \) is the number of training examples
- \( x^{(i)} \) is the \( i \)-th training example feature vector
- \( < x^{(i)}, x^{(j)} > \) represents the inner product of the \( x^{(i)} \) and \( x^{(j)} \) feature vectors
- \( y^{(i)} \) is the class for the \( i \)-th training example
- \( \alpha_i \) is the Lagrange multiplier associated with the \( i \)-th training example
- \( C \) is a regularization parameter (larger values introduce less regularization)
• **RepTree:** RepTree uses the regression tree logic and creates multiple trees in different iterations based on C4.5 algorithm. After that it selects best one from all generated trees. This leads to a fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance, hence the name Reduced Error Pruning Tree (“REPT”). REPTree builds a decision/regression tree using the information gain as the splitting criterion (refer Equation 3), and prunes it using reduced error pruning. It only sorts values for numeric attributes once.

\[
IG(T, a) = H(T) - H(T|a) \tag{3}
\]

This equation represents the information gain of \( T \) given \( a \), which is defined as the difference between the unconditional Shannon entropy of \( T \) and the expected entropy of \( T \) conditioned on \( a \), where the expectation value is taken with respect to the induced distribution on the values of \( a \).

• **Simple Logistic Regression:** Logistic Regression is one of the most simple and commonly used data mining algorithms for binary classification. It is easy to implement and can be used as the baseline for any two-class classification problem. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables. It is similar to a linear regression, but the curve is constructed using the natural logarithm of the outcomes of the target variable, rather than the probability. The equation for Logistic Regression is shown in Equation 4,

\[
p = \frac{1}{e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)}} \tag{4}
\]

where \( b_i \) defines the steepness of the curve.

2.3. Model Evaluation

The evaluation metrics used in measuring the performance for Naive Bayes, Simple Logistics, REPTree, and SMO algorithms are accuracy, precision, and recall, where TP is positive tuples, TN is negative tuples, FP is incorrectly classified positive tuples, and FN is incorrectly classified negative tuples (Refer to Equation 5 to Equation 7. Model testing was then performed by means of 10-fold cross-validation.

\[
\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{5}
\]

\[
\text{Recall} = TP/(TP + FN) \tag{6}
\]

\[
\text{Precision} = TP/(TP + FP) \tag{7}
\]

3. Results and Discussion

The purpose of the experiments is to compare the performance of four classification algorithms, which are Naive Bayes, Simple Logistics, REPTree, and Sequential Minimal Optimization (SMO) on diabetes data based on the numerical input constraints. Table 2 shows the results of various measures across all algorithms and Table 3 shows the error report. Next, Table 4 shows the comparative performance of accuracy, which shows the accuracy result of rightly classified (true positives) and wrongly classified instances (true negatives). The accuracy of Naive Bayes method found to be 85.0%, Simple Logistics is 73.5%, REPTree is 72.5%, and SMO is 75.5%. These results are then illustrated in Figure 3. Based on the diabetes dataset consisting of 200 tuples with 9 attributes, it was found that Naive Bayes algorithm performed the best when compared to other classifiers.
Table 2. Detailed Results

| Algorithms    | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---------------|---------|---------|-----------|--------|-----------|----------|
| Naive Bayes   | 0.803   | 0.309   | 0.835     | 0.803  | 0.809     | 0.806    |
|               | 0.691   | 0.197   | 0.644     | 0.691  | 0.667     | 0.806    |
| Simple Logistics | 0.529  | 0.159   | 0.632     | 0.529  | 0.576     | 0.808    |
|               | 0.735   | 0.365   | 0.727     | 0.735  | 0.729     | 0.808    |
| REP Tree      | 0.811   | 0.441   | 0.781     | 0.811  | 0.796     | 0.712    |
|               | 0.559   | 0.189   | 0.603     | 0.559  | 0.580     | 0.172    |
| SMO           | 0.725   | 0.356   | 0.721     | 0.725  | 0.722     | 0.712    |
|               | 0.871   | 0.471   | 0.782     | 0.871  | 0.824     | 0.700    |
|               | 0.529   | 0.129   | 0.679     | 0.529  | 0.595     | 0.700    |
|               | 0.755   | 0.354   | 0.747     | 0.755  | 0.746     | 0.700    |

Table 3. Error Reports

| Statistics                  | Naive Bayes | Simple Logistics | REP Tree | SMO    |
|-----------------------------|-------------|------------------|----------|--------|
| Kappa statistic             | 0.4856      | 0.3854           | 0.3761   | 0.4233 |
| Mean absolute error         | 0.2735      | 0.3244           | 0.3432   | 0.245  |
| Root mean squared error     | 0.4369      | 0.4079           | 0.4441   | 0.495  |
| Relative absolute error     | 60.8854     | 72.1701          | 76.3586  | 54.511 |
| Root relative squared error | 91.9901     | 86.0966          | 93.7228  | 104.4669|

Table 4. Comparative Results

| Algorithms    | True Positive (TP) | True Negative (TN) |
|---------------|--------------------|--------------------|
| Naive Bayes   | 85.0               | 15.0               |
| Simple Logistic | 73.5             | 26.5               |
| REP Tree      | 72.5               | 27.5               |
| SMO           | 75.5               | 24.5               |

Figure 3. Comparison of Accuracy Performance
4. Conclusion
In this research, the frequently used classification techniques Naïve Bays, Simple Logistic, REP Tree and SMO are analyzed on the medical dataset to find the simpler solution for predicting positive cases of diabetes. The performance evaluation included the accuracy, precision, and recall for the given dataset. The performance of REPTree has the lowest accuracy but within the same range of performance with Simple Logistic and SMO. The results showed that the performances of Naïve Bays is superior to the other techniques, SMO, Simple Logistic and REP Tree for the classification of the diabetes data. To improve the overall accuracy, it is necessary to use more data set with larger number of attributes and use a better feature selection method in future works and may also include hybrid classification models by combine some of the techniques.

References
[1] Kumari S and Singh A 2013 A data mining approach for the diagnosis of diabetes mellitus 2013 7th International Conference on Intelligent Systems and Control (ISCO) (IEEE) pp 373–375
[2] Ahmad A, Mustapha A, Zahadi E D, Masah N and Yahaya N Y 2011 Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus International Conference on Digital Information Processing and Communications (Springer, Berlin, Heidelberg) pp 537–545
[3] Perveen S, Shahbaz M, Guergachi A and Keshavjee K 2016 Procedia Computer Science 82 115–121
[4] Kumar P S and Umatejaswi V 2017 International Journal of Scientific and Research Publications 7 705–709
[5] Ibrahim N H, Mustapha A, Rosli R and Helme H N 2013
[6] Butwall M and Kumar S 2015 International Journal of Computer Applications 120
[7] Das H, Naik B and Behera H 2018 Classification of diabetes mellitus disease (dmd): a data mining (dm) approach Progress in computing, analytics and networking (Springer) pp 539–549
[8] Fayyad U, Piatetsky-Shapiro G and Smyt P 1996 Commun. ACM 39 27–34 ISSN 0001-0782 URL https://doi.org/10.1145/240455.240464
[9] Blake C 1998 http://www. ics. uci. edu/~ mllearn/MLRepository. html
[10] Platt J 1999 AJ, MIT Press, Cambridge, MA