Analyzing Diabetic Data using Classification

Nazim Razali¹, Aida Mustapha¹, Syed Zulkarnain Syed Idrus², Mohd Helmy Abd Wahab³, Siti Aida Fatimah Madon¹

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
² Universiti Malaysia Perlis, Jalan Wang Ulu Arau, Sg. Chuchuh, 01000 Kangar, Perlis
³ Faculty Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400 Batu Pahat, Johor, Malaysia
E-mail: aidam@uthm.edu.my

Abstract. In modern day, Diabetes has become one of prominent disease that affecting people all around the world. Over 425 million people worldwide with a majority of them are adults living in low and middle income countries had been affected by diabetes. According to the Diabetes Atlas from International Diabetes Federation in 2017, the number of people affected by diabetes disease are expected to increase to over 629 million by 2045. Along with the increasing of the computer science and information technology in researches, its has continued advent into many field including the medical and healthcare field which the cases of this disease and the symptoms are well recorded and documented. This paper aims to use several data mining techniques such as Naive Bayes, Sequential Minimal Optimization (SMO), RepTree and Simple Logistic Regression for classifying whether positive or negative result of diabetes diagnostic. Then, the results will be measured using confusion matrix in term of accuracy, precision and recall as evaluation metric for measuring the classification performance.

1. Introduction

According to Internationnal Diabetes Federation Diabetes Atlas [1], there are about 425 million people around 20 to 79 years were suffering from diabetes and it is expected to rise to 629 million by year 2045. Unfortunately, this disease is getting worse where diabetes are increasing in every third class country where people living in low and middle-income. Poorly, every 1 in 2 from 212 million of people do not realize that they are suffering from diabetes which already claimed 4 million life.

Diabetes is the condition where a person’s blood sugar levels are above normal from 4.4-6.1 mmol/L. Diabetes is actually referred to Diabetes Mellitus (DM) and people who suffered with DM are called “diabetics”. Frequent urination, increased thirst and increased hunger are the usual symptoms of high blood sugar. If diabetes is left untreated, it can cause many complications. A patient can also suffer from diabetic ketoacidosis, nonketotic hyperosmolar coma or even worse, death. Apart from that, diabetic patients are also suffering from other fatal diseases such as heart disease, stroke, and chronic kidney failure. Foot ulcers and eyes damages are common as the body fails to function normally. Once sugar level in the blood increased, it called pre-diabetes because the sugar level is not so high than the normal value.

Diabetes is due to low production of insulin in human body or in the cells of the body by the pancreas. Insuline in important to ensure glucose reach each cell of the body. The International
Diabetes Federation [2] categorizes diabetes into three types: Type 1 Diabetes, Type 2 Diabetes, and Gestational Diabetes. In Type-1 Diabetes, it manifests as an auto-immune disease occurring at a very young age of below 20 years. This type of diabetes is referred as “insulin-dependent diabetes mellitus” (IDDM) or “juvenile diabetes”. The person who suffers from this type of diabetes will suffer it throughout their life and had to depend on insulin injection. Type 2 diabetes is in the state when the various organs of the body become insulin resistant and this increases the demand for insulin and at this point, the pancreas does not produce the required amount of insulin. This type is referred as “non-insulin-dependent diabetes mellitus” (NIDDM) or “adult-onset diabetes”.

While the main reason of Type 1 diabetes is largely unknown, the main reason for Type 2 diabetes is obesity and can be controlled by exercise and proper diet. If the blood sugar level is not decreasing after with the control of exercises and diet, medicine will be prescribed to control the blood sugar level. Whereas, gestational diabetes tends to occur in pregnant women who do not have a history of diabetes but suddenly develop high blood sugar and the insulin produced is not sufficient. Pregnancies among older women are of higher risk as it will develop into Type 2 diabetes. One can avoid the complications associated with them if the disease is detected at an early state [3].

According to World Health Organization (WHO), 415 million people worldwide was estimated to have diabetes among adults aged 20 to 79 years in 220 countries with Type 2 in 2015 [4]. Through data mining, patient histories and medical data can be used to detect the disease based on symptoms at an early stage, hence starting early treatment. This can also help in avoiding further complications [5]. Considering the importance of early medical diagnosis of this disease, this paper aims to analyze diabetic data through a classification task in data mining.

Section 2 reviews all works related to the diabetes cancer. Section 3 presents the classification methodology used to perform the data mining task along with the dataset and the evaluation metrics. Section 4 presents the results and finally Section 5 concludes with some direction for future work.

2. Related Work
Application of data mining in diabetes diagnostic has been widely used in researches. [6] has used data mining techniques to cluster type of diabetes using simple K-means algorithm and classify it using Naive Bayes, Random Tree, C4.5 Decision Tree and Simple Logistic Regression based on 14 attributes of 650 diabetic patients data. The results show that C4.5 Decision Tree successfully to outperform other 3 classifier. However, the work of [7] shows opposite result if compared to [6].

[7] has proposed 4 different classification algorithms such as C4.5 Decision Tree or also known as J48 in WEKA software, PART Decision Tree, Multi Layer Perceptron (MLP) and Naive Bayes to be applied on 768 instances of dataset from National Institute of Diabetes and Digestive and Kidney Diseases extracted from Kaggle database [8] for diabetes diagnostic. As the results, show that Naive Bayes successfully outperform 3 classifier in classifying whether the patients is positive tested or negative tested for diabetes.

As well as work of [9] showed that Naive Bayes is a prominent classifier in diabetes classification or prediction. [9] the same dataset as in work of [7] to be executed using 3 classification algorithm such as Naive Bayes, Support Vector Machine and Decision Tree. The result shows that Naive Bayes provide the optimum result among all classification algorithms employed.

Finally, [10] has employed 4 classifier such as k-Nearest Neighbor, Support Vector Machine (SVM), Multilayer Perceptron Neural Network, and Naive Bayes for diabetes diagnostic. The benchmark dataset similar to the work of [7] and [9] has been used in this research. Thus, SVM show the highest classification accuracy outperformed the others with 78.83%.
3. Methodology
In data mining, classification deals with identifying the problem by prediction of target classes or categories accurately in the dataset. The paper explores the aspect of Naive Bayes, Sequential Minimal Optimization (SMO), REPTree and Simple Logistic Regression as classification algorithms in data mining for classifying the diabetes diagnostic dataset. The main objective of this research work is to evaluate the performance of classification methods for diabetes dataset based on the numerical input and imbalance dataset constraints.

The model proposed in this paper followed the methodology from previous work of [6]. There are two important stages and the classification methodology that shown in Figure 1 for implementation. The experiment are performed by using 10-folds cross validation for training and testing data.

- Stage-1: Data pre-processing. To combat imbalance dataset and generate new instances for undersample target class. Randomise the newly generated datasets.
- Stage-2: Applying the classification algorithms to dataset and evaluate the performance of classification methods for in term of accuracy, precision and recall.

![Figure 1. Classification Methodology [11]](image)

3.1. Dataset
The dataset used in the classification experiment is the Pima Indians Diabetes Database of National Institute [12] sourced from the Kaggle Database [8]. The data has records of 768 total instances. However, the datasets are imbalance in target class where 500 instances of target class for “Negative” or “0” and 268 instances of target class for “Positive” or “1”. Thus, the SMOTE oversampling method [13] have been applied to combat the imbalance dataset which generated 1036 instances consisting 500 instances of target class for “Negative” or “0” and 536 instances of target class for “Positive” or “1”. Then, the dataset were randomise to shuffle the order of newly generated synthetic target class for “Positive” or “1” in the dataset. Table 1 shows a brief description of the attribute that is being considered.
Table 1. Experimental results

| Attribute                  | Relabeled |
|----------------------------|-----------|
| 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     |

3.2. Algorithms

The classification algorithms used in the experiment include the Naive Bayes, Sequential Minimal Optimization (SMO), Reduced Error Pruning Tree (REPT), and Simple Logistic Regression. The results are then analyzed based on accuracy, precision, and recall metrics.

- **Naive Bayes** is the most popular and simplest probabilistic algorithm in classification. It has ability to handle missing value and imbalance data. It defines all the attributes are independence or no depency between all the attributes except the attributes that become the target class. Thus, it neglect the effect of correlation of other attributes and solely dependence to the target class.

- **Sequential Minimal Optimization (SMO)** proposed by John Platt in 1998 to provide solution for training a support vector machine (SVM) classifier [14]. Training a SVM requires a lot computational power and time consuming to handle quadratic programming (QP) optimization problem. Thus, SMO enhanced SVM by breaks QP as smallest as possible and solved it analytically.

- **RepTree** or also known as Reduced Error Pruning Tree (REPT) is fast decision tree learning that uses information gain to build the decision or regression tree, and prunes the tree using reduced error pruning. It uses the regression tree logic and creates multiple trees in different iterations, then selects the best from all generated trees [15]. Missing values are treated but it sorts values for numerical attributes only once.

- **Simple Logistic Regression** is a simple but commonly used algorithm in data mining for binary or two-class classification when there are only two possible classes such as Yes/No, Positive/Negative or Survived/Death. This algorithm is easy to implement and can be used as the baseline for any binary classification problem. Logistic regression estimates the relationship between one dependent binary variable and independent variables.

3.3. Evaluation Metrics

Naive Bayes, SMO, REPTree and Simple Logistics Regression algorithms were used for this research. The evaluation are performed by using 10-folds cross validation. Accuracy, precision and recall have been used as evaluation metric since it have been used widely in standard data mining fieldwork like in work of [6]. This evaluation metric were derived from a confusion matrix as shown in Table 2, where the class 1 and 2 can be positive and negative. The definition can be defined as follows.

- **Positive (P):** Observation is positive (for example: positive diabetes diagnosed).
- **Negative (N):** Observation is not positive or negative (for example: negative diabetes diagnosed).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

Table 2. Confusion Matrix

|                | Class 1 | Class 2 |
|----------------|---------|---------|
| Predicted Class 1 | TP      | FN      |
| Actual Class 2   | FP      | TN      |

Based on the confusion matrix in Table 2, accuracy can be defined as in Equation 1:

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)
\]

Precision can be defined as in Equation 2 where the total number of correctly classified positive samples are divided by the total number of actual positive samples.

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)
\]

Recall can be defined as in Equation 3 where the total number of correctly classified positive samples divided by the total number of predicted positive samples.

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)
\]

4. Experimental Results

The proposed classification algorithms namely Naive Bayes, SMO, REPTree and Simple Logistic Regression was executed on WEKA tool in two stage which are 1) Pre-processing to combat imbalance dataset and randomise the newly generated synthetic dataset and 2) Applying the classification algorithms before be evaluated the classification performance in term of accuracy, precision and recall. The 10-fold cross validation have been used to evaluate the performance. Overall result in term of accuracy, precision and recall were generated as experimental output of confusion matrix are shown as in Table 3.

Table 3. Confusion Matrix

| Algorithm   | Accuracy | Precision | Recall |
|-------------|----------|-----------|--------|
| Naive Bayes | 0.736    | 0.739     | 0.736  |
| Simple Logistics | 0.757    | 0.758     | 0.758  |
| REPTree     | 0.751    | 0.752     | 0.751  |
| SMO         | 0.740    | 0.741     | 0.740  |

Figure 2 shows the overall result of accuracy in percentage for all proposed classification algorithms. In conclusion, it is observed that Simple Logistic Regression slightly better compare to REPTree (75.10%) and SMO (74.00%) by achieving 75.70% in accuracy while Naive Bayes least accurate by achieving 73.60% accuracy.
5. Conclusions
In conclusion, this research presented an analysis of diabetic data using classification techniques based on Naive Bayes, SMO, REPTree and Simple Logistic Regression. The data undergoing oversampling method called SMOTE to combat imbalance dataset before be randomised to shuffle the newly generated synthetic dataset. The evaluation metric such as accuracy, precision and recall are measured for the given dataset to estimate the performance of each classification techniques. As the result, simple logistic regression have achieved the highest rate of accuracy, precision and recall compare to other three classification algorithms while Naive Bayes is slightly lowest other than three algorithm. It is recommended to use larger number of size of datasets and attributes as well as use better feature selection method in order to improve the classification performance. Besides, combination of classification techniques or hybrid classification such as stacking, boosting and bagging may improve the performance of classification.

Acknowledgements
This research is funded by Universiti Malaysia Perlis.

References
[1] Federation I D 2019 Retrieved from https://www.idf.org/component/attachments/?task=downloadid=1525: Atlas-8e-Global-factsheet
[2] Federation I D 2019 Retrieved from https://www.idf.org/aboutdiabetes/what-is-diabetes.html
[3] for Disease Control C and Prevention 2014 Atlanta, GA, U.S. Department of Health and Human Services, 2014
[4] Ogurtsova K, da Rocha Fernandes J D, Huang Y, Linnenkamp U, Guariguata L, Cho N H, Cavan D, Shaw J and Makaroff L 2017 Diabetes Research and Clinical Practice 128 40–50
[5] Kumari S and Singh A 2013 7th International Conference on Intelligent Systems and Control (ISCO), 2013

[6] Kumar P S and Umatejaswi V 2017 International Journal of Scientific and Research Publications 7

[7] Nidhi, Kumar M and Kakkar L 2018 International Journal for Research in Engineering Application and Management 4 347–351

[8] Kaggle 2019 Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database/version/1

[9] Sisodia D and Sisodia D 2018 Procedia Computer Science 132 1578–1585

[10] Al-Khasawneh A 2018 Handbook of Research on Emerging Perspectives on Healthcare Information Systems and Informatics 281–303

[11] Fayyad U, Piatetsky-Shapiro G and Smyth P 1996 Advances in knowledge discovery and data mining 1–34

[12] Smith J, Everhart J, Dickson W, Knowler W and Johannes R 1988 In Proceedings of the Symposium on Computer Applications and Medical Care, IEEE Computer Society Press 261–265

[13] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 Journal of Artificial Intelligence Research 16 321–357

[14] Platt J 1998 Technical Report MSR-TR-98-14

[15] Kalmeough S 2015 International Journal of Innovative Science, Engineering and Technology 2 438–446