Analysis of Data for Diabetics Patient

Korobi Saha Koli  
Sajjad Waheed

Department of Information and Communication Technology,  
Mawlana Bhashani Science and Technology University,  
Santosh-1902, Tangail, Bangladesh

URL: http://dx.doi.org/10.19044/esj.2017.v13n15p216

Abstract

Diabetes, a disease responsible for different kinds of diseases such as heart attack, kidney disease, blindness and renal failure etc. The most common disorder is the endocrine (hormone) system, occurs when blood sugar levels in the body consistently stay above normal. There are two types of diabetic; one is body's inability to make insulin and another is body not responding to the effects of insulin. In our developing country Bangladesh, Diabetes is a costly disease whose risk is increasing at alarming rate. This paper evaluates the selected classification algorithms for the classification of some Diabetes patient datasets. Classification algorithms considered here are Naive Bayes classification (NBC), Bagging algorithm, KStar algorithm, Logistic algorithm and Hoeffding tree. These algorithms are evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity. Collected datasets of diabetes affected people are firstly preprocessed then some investigation based on mentioned algorithm has been executed successfully. From the investigation result it is found that, KStar algorithm is the best as it gives high accuracy with the low error. Here it is said that, some parameters are responsible for diabetes.

Keywords: Diabetics in Bangladesh, Type-2 Diabetics, Data Pre-processing, Box Plot, Histogram, Classification Algorithms

Introduction

Diabetes is not only a disease but also responsible for distinctive types of diseases. Various types of complicated diseases including kidney diseases, heart problem, and blindness can be occurred due to diabetes. It also increases the blood sugar level of the affected people. The micro and macro coronary heart diseases have broken out when the glucose level of human body is raised more than 200mgI/dL (Abdullah A. Aljumah et al.,
This situation happens when the diabetes is uncontrolled. The prevalence of type-2 diabetes is increasing at an alarming rate in developed and developing countries all over the world (Diabetes Group, 2002; M. Pradhan and R.K. Sahu, 2011). It is one of the vital reasons for death and numerous health problems. As a result, diabetes affected patients are becoming burden of their families and the health care system. Large numbers of studies have proved that, this disease (type 2) can be smoothly prevented by changing life style and food habits (S. Chaudhuri et al., 1989; N. Landwehr et al., 2005). Numerous peoples in developed countries have suffered from obesity. The eating habits of the western countries are very peculiar. Most of the people eat junk foods that are full of carbohydrates and saturated fats. These foods are not contained fiber. For this reason, they are easily affected by several types of diseases specially diabetes (V.P. Kumar and L.Velide, 2014; R. Bagdi and P. Patil, 2012, K. Ahmed et al., 2015).

Diabetes is one of the most alarming diseases in our country, Bangladesh. According to “International Diabetes Federation”, 7.1 million diabetes affected people / patients have been found and 129312 adults fall to death due to diabetes. Two types of diabetes are: Type-1 and Type-2. For Type-1 diabetes, it is a metabolic disorder where high level of blood sugar occurs. Commonly diseases like heart attacks, strokes, blindness and kidney failure occurred due to type 2 diabetes. The dramatic increase in type 2 diabetes is therefore inextricably linked to the global obesity epidemic (L. W. Yun et al., 2008; M. Pradhan and R.K. Sahu, 2011, K. Ahmed et al., 2014).

The diagnosis of diabetes is a vital and tedious task. Data mining technique is very popular way to detect diabetes or any types of diseases. It is a popular way to find out hidden and potential information from the databases. Data mining contains some methods like classification, clustering, association rule that are effective to analysis the data. Recently, many organizations are also utilizing this technique to make various decisions. So, data mining contributes the methodology to examine the practical information of data for decision making (H.W. Ian, 2005; K. Ottenbacher, 2001; N. Otsu, 1979; Sung-Hyuk Cha and Charles Tappert, 2009; N. Patton et al., 2006; J. Manyika, 2012; A. Thusoo, 2009, K. Ahmed et al., 2013).

The main goal of our research is to develop a system that can be used to which attribute are responsible for Diabetes and comparing Classification algorithms performance based on the diabetes patient data.

**Proposed Methodology**

**Data Collection**

We collected 300 data of patients, among which 150 are diabetes patients and 150 are non-diabetes patients. We have collected raw data from
different diagnosis centers and government hospitals of Bangladesh by considering some risk factors like age, weight, gender, area, income, profession, marital status etc. Both the patients and non-patients data are added for process where all the participants’ ages about 25 to 75 are exhibited. The whole data collection process is performed based on a handmade questioner which is created through background study on diabetics. About 10 questions are selected for assessment of diabetics of Bangladeshi population.

**Data Pre-processing**

First time, raw data looks like a junk data. Data pre-processing is a vital term of data mining. Data pre-processing handles are not only data duplications but also data missing inconsistency in the dataset. Data reduction, transformation, integration, discretization, cleaning are considered as the major tasks of the data pre-processing. Box Plot analysis is used here to handle noisy data of numerical formatted data (K. Ahmed et al., 2012).

**Classifications**

**Classification via Naïve Bayes Algorithm**

The example of Naïve Bayes algorithm is applied on Diabetes data set and the confusion matrix is generated for class having possible values are shown in Figure 1 and Figure 2 shown the Plot for Testing data set.

```plaintext
=== Confusion Matrix ===
  a  b  <-- classified as
   208  7  |   a = yes
   81  4  |   b = no
```

**Fig 1.** The confusion matrix of Naïve Bayes algorithm.

**Fig 2.** Plot for Testing data set of Naïve Bayes Algorithm

218
Classification via KStar Algorithm

The example of KStar algorithm is applied on Diabetes data set and the confusion matrix is generated for class having possible values are shown in Figure 3 and Figure 4 shown the Plot for Testing data set.

\[
\begin{array}{ccc}
\text{Confusion Matrix} \\
\text{a} & \text{b} & \text{classified as} \\
213 & 2 & a = \text{yes} \\
45 & 40 & b = \text{no}
\end{array}
\]

**Fig 3.** The confusion matrix of KStar algorithm

![Plot for Testing data set for KStar Algorithm.](image)

Classification via Bagging Algorithm

The example of Bagging algorithm is applied on Diabetes data set and the confusion matrix is generated for class having possible values are shown in Figure 5 and Figure 6 shown the Plot for Testing data set.

\[
\begin{array}{ccc}
\text{Confusion Matrix} \\
\text{a} & \text{b} & \text{classified as} \\
211 & 4 & a = \text{yes} \\
75 & 10 & b = \text{no}
\end{array}
\]

**Fig 5.** The confusion matrix of Bagging algorithm.
**Logistic Model tree Algorithm**

The example of Logistic Model Tree algorithm is applied on Diabetes data and the confusion matrix is generated for class having possible values are shown in Fig 7 and Figure 8 shown the Plot for Testing data set.

```
=== Confusion Matrix ===
   a  b  <-- classified as
   206 9 | a = yes
   77 8 | b = no
```

**Fig 7.** The confusion matrix of Logistic model tree algorithm

**Hoeffding Tree Algorithm**

The example of Hoeffding Model Tree algorithm is applied on Diabetes data and the confusion matrix is generated for class having possible values are shown in Fig 9 and Figure 10 shown the Plot for Testing data set.
\[\text{Confusion Matrix}\]
\[
\begin{array}{c|cc}
\text{a} & \text{b} \\
\hline
\text{206} & \text{9} & \text{a = yes} \\
\text{77} & \text{8} & \text{b = no} \\
\end{array}
\]

**Fig 9.** The confusion matrix of Hoeffding model tree algorithm

**Fig 10.** Plot for Testing data set of Hoeffding algorithm.

**Results and Analysis**

Performance of Selected classification algorithms were evaluated with diabetics datasets. Diabetics dataset contains 300 diabetics patient records from Bangladesh. The experimental results under the framework of WEKA (Version 3.8). The experimental results are partitioned into several sub items for easier analysis and evaluation. With Each algorithm, we have observed Accuracy, Precision, Sensitivity, Specificity, Relative Absolute Error (RAE), Root Mean squared error (RMSE) and Mean absolute error (MAE) whose can be defined as follows:

**Accuracy:**

The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

Accuracy = \(\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}\)

**Precision:**

Precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives).

Precision = \(\frac{\text{TP}}{\text{TP} + \text{FP}}\)

**Specificity**

Specificity is the True negative rate that is the proportion of negative tuples that are
correctly identified.
Specificity= TN/(TN+FP)
All measures can be calculated based on four values, namely True Positive, False Positive, False Negative, and True Negative
These values are described below.
□ True Positive (TP) is a number of correctly classified that an instances positive.
□ False Positive (FP) is a number of incorrectly classified that an instance is positive.
□ False Negative (FN) is a number of incorrectly classified that an instance is negative.
□ True Negative (TN) is a number of correctly classified that an instance is negative.

Mean absolute error
Mean absolute error, MAE, is the average of the difference between predicted and actual value in all test cases; it is the average prediction error.
The formula for calculating MAE is given in equation shown below:

\[
\frac{1}{n} \sum_{i=1}^{n} |a_i - c_i|
\]

Assuming that the actual output is a, expected output is c.

Root Mean-Squared Error
RMSE is frequently used measure of differences between values predicted by a model or estimator and the values actually observed from the thing being modeled or estimated. It is just the square root of the mean square error as shown in equation given below:

\[
\sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - c_i)^2}
\]

Assuming that the actual output is a, expected output is c.

Table-1, Table-2, Table-3,Table-4 show the performance of the different classifiers algorithms for diabetics data set. Fig. 11 and Fig. 12 show the barchart of performance level of different algorithms. Fig. 13 Fig. 14 and Fig. 15 exhibit the decision tree on trained diabetics data set.

<table>
<thead>
<tr>
<th>Algorithm name</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaiveBayes</td>
<td>0.707</td>
<td>0.719</td>
<td>0.976</td>
<td>0.047</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.713</td>
<td>0.727</td>
<td>0.958</td>
<td>0.094</td>
</tr>
<tr>
<td>KStar</td>
<td>0.843</td>
<td>0.826</td>
<td>0.991</td>
<td>0.471</td>
</tr>
</tbody>
</table>
According to Table 1 there are accuracy, precision, sensitivity, specificity for different algorithm & here we can see the highest accuracy, precision, sensitivity, specificity belongs to KStar algorithm.

Table 2. Performance of Error Result for using training set of Diabetics Data Set

<table>
<thead>
<tr>
<th>Algorithm name</th>
<th>MAE</th>
<th>RMSE</th>
<th>RAE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaiveBayes</td>
<td>0.382</td>
<td>0.437</td>
<td>94.00</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.371</td>
<td>0.432</td>
<td>91.35</td>
</tr>
<tr>
<td>KStar</td>
<td>0.26</td>
<td>0.332</td>
<td>63.99</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.361</td>
<td>0.415</td>
<td>88.87</td>
</tr>
<tr>
<td>Hoeffding</td>
<td>0.382</td>
<td>0.437</td>
<td>94.00</td>
</tr>
</tbody>
</table>

Table 2 is showing the MAE, RMSE, RAE for different algorithm & here we can see the minimum error rate value providing by KStar algorithm.

Fig 11. Comparisons of all algorithms for training set
The decision tree J48 in fig 13 showing that, if one (age between 34-50) suffering from other diseases and if there is any existence of diabetics in his/her family history, he/she may suffer from diabetics and if one’s weight become high he/she can also suffer from this diabetics.
The decision tree J48 in fig 14 showing that, if one’s age more than 34 and don’t do any physical exercise and weight almost 53 or more, then he/she can suffer by diabetics.

The decision tree J48 in fig 15 showing that, if one is addicted by tobacco then he/she may be affected by diabetics.
Conclusion

In this research work, comparison among numerous classification algorithm such as Naïve Bayes classification (NBC), Bagging, KStar, Logistic and Hoeffding tree algorithm based on their effectiveness has been analyzed. Popular algorithms were considered for evaluating their classification performance on classifying diabetic affected patient dataset. Based on the above classification and investigation results, it can be clearly visualize that highest accuracy, precision, sensitivity, specificity belong to the KStar algorithm. It is observed that, KStar algorithm is the best as it minimizes error rate by providing excellent rate of accuracy on expected result. Moreover, it can also be decided that the mostly affected persons are found between the age limit of 27-59 years according to decision tree. In addition, those who are over weight, do not exercise regularly, tobacco addicted and ancestor can be affected by diabetics.

References:


