COMPARISON OF DATA MINING ALGORITHMS IN THE DIAGNOSIS OF TYPE II DIABETES

Sadri Sa’di¹, Amanj Maleki¹, Ramin Hashemi², Zahra Panbechi¹ and Kamal Chalabi¹

¹Department of Computer Engineering, Shahbestar Branch, Islamic Azad University, Shahbestar, Iran
²Executive Master of Business Administration (EMBA) Student, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

ABSTRACT

Diabetes is one of the most prevalent diseases in the world today with high mortality and morbidity rate, thus one of the biggest health problems in the world. There are many ways to diagnose the disease; one of these methods is data mining algorithms. The use of data mining on medical data has brought about important, valuable, and effective achievements, which can enhance the medical knowledge to make necessary decisions. In this paper, the diagnosis of type II diabetes is done through data mining algorithms. The dataset used for the diagnosis of type II diabetes includes 768 samples from diabetic patients taken from Pima Indians Dataset. In this article, Naive Bayes, RBF Network, and J48 are the data mining algorithms used to diagnose type II diabetes. The so-called algorithms perform diagnosis using Weka. Finally, the algorithms were compared to determine which one was more accurate in diagnosis of type II diabetes. The results revealed that Naive Bayes, having accuracy rate of 76.95%, enjoyed the highest accuracy for diagnosis of type II diabetes.

KEYWORDS

Data Mining, Diabetes type II, Naive Bayes, RBF Network, J48

1. INTRODUCTION

Diabetes is one of the most prevalent metabolic diseases. Distinguishing features of diabetes are overeating and over drinking. Diabetes is caused by reduced insulin production in the body or loss of effectiveness of insulin in the metabolism of sugar. Diabetic patients are divided into two groups of type I and type II. Type I diabetes usually appears before age of 30 due to the lack or deficiency of insulin; patients with type 1 diabetes are usually thin and need insulin injections. Type II diabetes usually occurs over 40 years of age [1]. The causes of type II diabetes are factors such as overweight and obesity, lack of physical activity, poor diet, and family history. In patients with type II diabetes who are often obese, diet and exercise are the main treatments. Type II diabetes develops slowly and shows long-term symptoms. Usually, in diagnosing type II diabetes, the symptoms are type I diabetes-like; however, since it is developed slowly, it does not usually attract the patient’s attention. Generally when the amount of glucose in the blood is high, the environment in which different body organs perform their duties is defected; at this time, the body makes symptoms appear to show that the body is not an appropriate condition. These symptoms are experienced by diabetics. If this condition, caused by various factors, is left untreated, it may cause problems such as heart attacks, strokes, blindness, kidney failure, and blood vessels disease which may lead to amputation, nerve damage, and sexual impotence [1, 2, 3].

DOI:10.5121/ijcsa.2015.5501
In many cases, the diagnosis is generally based on patient test results and physicians’ experience. Thus, diagnosis is a complex task requiring high skills and experience. Early diagnosis and medical care of patients can greatly reduce the problems of patients [4, 5]. So far, various methods have been proposed to solve these problems. One of these methods is data mining. Naive Bayes, RBF Network, and J48 are data mining algorithms that are used in this article. In recent years, machine learning methods such as data mining are widely used in anticipation especially in medicine. Data mining is used to enhance the accuracy of diagnosis [3]. Here, the purpose of diagnosis of type II diabetes, by data mining algorithms, is to achieve a high accuracy rate and objective medical diagnosis of a dataset.

Data mining is a self-knowledge discovery and a process for the analysis of large databases, providing unknown, hidden, meaningful, and useful patterns automatically obtained from large-scale databases [6]. Data mining algorithms such as neural networks [6], support vector machines [3], decision tree [7] and ... have been used successfully in various medical fields. These algorithms have been able to be a decent solution in different diseases’ diagnostic systems such as diabetes [3], heart disease [6], breast cancer [9], etc. Also, use of data mining algorithms makes new information and relationships embedded in large and complex datasets evident through inference and learning new patterns and relations [7, 8]. The use of large amounts of patients’ data to diagnose disease, using data mining algorithms, improves the accuracy of these algorithms [7, 8]. The data used in this article includes 768 samples from diabetic patients obtained from Pima Indians Diabetes Data Set. And the data are used for data mining algorithms, and diagnosis is done by Weka through data mining algorithms.

Significant research has been performed in the field of diabetes diagnosis by data mining algorithms. Farahmandian et al. [3] carried out the diagnosis of diabetes on 768 samples using Pima Indians Dataset. The data mining method used is algorithms such as SVM, KNN, Naive Bayes, ID3, C4.5, C5.0, and CART, 80% of data is used for education and 20% for testing. Comparing algorithms used, SVM algorithm which is more accurate than other algorithms enjoyed a accuracy rate of 81.77%.

In another study [10], diabetes was predicted using 768 samples of Pima Indians Dataset. The data mining method used is to predict multi-layer perceptron neural networks (MLP). MLP has an input layer with eight Pima Indians Dataset features and an output layer that predicts, including the ninth property which is zero (Normal) and one (Sick). 60% of data is for education, 20% for testing data, and 20% for the application set. The accuracy obtained for education is 97.61% in MLP.

In [11], using Pima Indians Dataset, containing 768 samples, the researchers set to diagnose diabetes. The applied data mining techniques are algorithms such as KNN, Amalgam KNN, K-means, EM, and ANFIS. Comparing these algorithms, amalgam KNN was more accurate than others.

In the second part, the data mining algorithms and datasets used for the diagnosis of type II diabetes are presented. In the third part, the results of Naive Bayes, RBF, and J48 algorithms for the diagnosis of type II diabetes are discussed, and then the results of three algorithms are compared to decide which method is more accurate in diagnosis. In the fourth part, conclusion and future works are presented.
2. THE DATA MINING ALGORITHMS FOR DIAGNOSIS OF TYPE II DIABETES

Advances in data storage led to large databases. Considering valuable information in these databases, attempts began to extract information. Data mining is a new scientific field which has formed the research in the field of statistics, machine learning, computer science, especially database management, [8]. In data mining, an effort is put to make interactions between statistics and computer science. Making the connection, due to the assumptions, difficult transformations of theoretical concepts to computer algorithms, and many computing algorithms, is difficult [8, 12, 13]. Interactions in data mining are provided as models such as regression, time series, clustering, classifications, graphs, etc. The data are samples of a population; the sample’s largeness is due to bulky data, issues such as how to analyse data at the right time and the ability to extract patterns and carrier models on data. Data mining is to search hidden patterns among data [8, 12, 13, 14]. After processing the data, results are stored on a database or new warehouses. The structure of a typical data mining system has shown that its input consists of large data and the output consists of the required data, eventually leading to knowledge discovery from databases [12, 13, 14].

Data mining techniques are used because of their high use in medical fields [3]. Due to the high cost of health services both for patients and the government, data mining can be used to save on the treatment costs using data analysis [6]. Using data mining, the best treatment and the most economical methods can be identified to reduce the cost of treatment. Using data mining, chronic and dangerous diseases can be identified; the result of which is to better manage patients and assign higher priority to these diseases. And in emergencies and situations that the needed time and resources to deal with all patients are limited, these patients are the first priority [6, 7, 8]. And also, drug side effects and drug contradictions of some treatments can be identified. To treat other patients, the results of these patterns can be used [6]. Accordingly, type II diabetes including 768 samples is diagnosed through data mining algorithms such as Naive Bayes, RBF Network, and J48 that will be explained throughout the paper.

Pima Indians Diabetes Dataset [15] consists of nine features for each sample. Applying these features, type II diabetes is diagnosed via data mining algorithms (Table 1). The first eight features are inputs, and feature IX is the output. Feature IX values are zero and one; zero indicating the Normal, and one, the Sick. For data classification, feature IX is classified into two classes of Normal and Sick.

<table>
<thead>
<tr>
<th>No. of Feature</th>
<th>Feature</th>
<th>Descriptions and Feature values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of times Pregnant</td>
<td>Numerical values</td>
</tr>
<tr>
<td>2</td>
<td>Plasma Glucose Concentration</td>
<td>Numerical values</td>
</tr>
<tr>
<td>3</td>
<td>Diastolic Blood Pressure</td>
<td>Numerical values in (mm Hg)</td>
</tr>
<tr>
<td>4</td>
<td>Triceps Skin Fold Thickness</td>
<td>Numerical values in mm</td>
</tr>
<tr>
<td>5</td>
<td>2-Hour Serum Insulin</td>
<td>Numerical values in (mu U/ml)</td>
</tr>
<tr>
<td>6</td>
<td>Body Mass Index (BMI)</td>
<td>Numerical values in (weight in kg/(height in m)^2)</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes Pedigree Function (DPF)</td>
<td>Numerical value</td>
</tr>
<tr>
<td>8</td>
<td>Age</td>
<td>Numerical values</td>
</tr>
<tr>
<td>9</td>
<td>Diagnosis of type 2 diabetes disease</td>
<td>sick=1 Normal=0</td>
</tr>
</tbody>
</table>
2.1. Naive Bayes

This algorithm provides a predictive model in relation to the likelihood of certain outcomes. Naive Bayes algorithm measures patterns or relationships among data by counting the number of observations. The algorithm then creates a model that reflects the patterns and their relationships. After creating this model, it can be used as a predictive model. This algorithm will help to create models for classification and prediction of several objectives. This algorithm is used to solve problems, identify and classify data between different data classifications. Naive Bayes is a learning probability algorithm adopted from Bayesian theory and a classification based on the probabilities. The Naive Bayes algorithm makes possible the model creation in the quickest time that will be available for issues such as classification by two or more classes [3, 9].

2.2. Radial Basis Function (RBF) Network

RBF neural networks enjoy a strong mathematical base, based on regularization hypothesis for solving complex problems. RBF advantage over other neural networks is that these networks can be trained in a shorter time to obtain the answer. RBF has three input, output, and hidden layers. The hidden layer has the Gaussian transfer function. Gaussian function is used with a certain width and center to stimulate the neurons in the hidden layer [5, 6]. Formula (1) shows the non-linear Gaussian function to stimulate neurons in the hidden layer.

\[
G_i(x) = \exp \left[ -\frac{\|x - x_i\|^2}{2\sigma_i^2} \right]
\]

According to formula (1), x is the network input vector; \(x_i\) is the vector function center of i-th radial basis; \(\|x - x_i\|\) is the distance between the input vector and the function center of i-th radial basis; \(\sigma_i\) is the width of the Gaussian function in the hidden neuron layer. The width consists of a multiple value of the average distance between the centers of radial basis function. The linear function used to stimulate neurons in the output layer is calculated using formula (2) [5, 6].

\[
F(x) = \sum_{i=1}^{n} w_{ij} G_i(\|x - x_i\|) + b
\]

According to formula (2), \(w_{ij}\) is the neurons’ link weight of the hidden and output layers. Gaussian function neurons of the hidden layer and the output layer neurons use a linear function, defined by a matrix forming RBF weight. To find the optimal weights in RBF, formula (3) is used [5, 6].

\[
W = (G^T G + \lambda G_0)^{-1} G^T d
\]

According to formula (3), G is the Gaussian function in formula (1); d is the target matrix; W is the weight vector, and \(\lambda\) is the parameter setting [5, 6].

2.3. J48

J48 algorithm is a decision tree that uses the concept of entropy with a training dataset. The decision tree is a method to display a series of rules, leading to a class or value. Decision trees used to predict batch variables, are called classification trees since they locate the samples in
categories or classes. In J48 algorithm, every feature of the data is used to make a decision by splitting into smaller subsets. J48 is a pruned or a non-pruned tree from C4.5 tree. J48 uses a statistical value called the Information Gain to determine how much a property can separate the training data according to their classification. The information interest of a feature is the amount of entropy reduction that can be achieved by separating data through this feature. J48 algorithm predicts the output objective using a classification tree structure [3, 9, 16].

3. RESULTS AND DISCUSSION

In this section, the results of data mining algorithms such as Naive Bayes, RBF Network, and J48 are presented and compared to decide which is more accurate in the diagnosis of type II diabetes. The results of the above algorithms are obtained. 768 samples were used; 70% of the data, i.e. 538 samples using Weka tool were selected for training and 30% i.e. 230 samples for testing data. The results of this study were obtained from the testing data of 230 samples. Accordingly, the results of Naive Bayes, RBF Network, and J48 are based on Tables 2, 3, and 4. Each of these tables manifests the details of the accuracies obtained from two classes of type II diabetes. The two classes are obtained from feature IX of Table 1 since feature IX specifies the Sick or Normal individuals. And these two classes, using the first eight features of Table 1, diagnose type II diabetes.

<table>
<thead>
<tr>
<th>Class</th>
<th>ROC Area</th>
<th>F-Measure</th>
<th>Recall</th>
<th>Precision</th>
<th>TP Rate</th>
<th>FP Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>sick</td>
<td>0.845</td>
<td>0.624</td>
<td>0.611</td>
<td>0.638</td>
<td>0.158</td>
<td>0.611</td>
</tr>
<tr>
<td>Normal</td>
<td>0.845</td>
<td>0.834</td>
<td>0.842</td>
<td>0.826</td>
<td>0.389</td>
<td>0.842</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.845</td>
<td>0.768</td>
<td>0.77</td>
<td>0.768</td>
<td>0.317</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>ROC Area</th>
<th>F-Measure</th>
<th>Recall</th>
<th>Precision</th>
<th>TP Rate</th>
<th>FP Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>sick</td>
<td>0.764</td>
<td>0.556</td>
<td>0.514</td>
<td>0.607</td>
<td>0.152</td>
<td>0.514</td>
</tr>
<tr>
<td>Normal</td>
<td>0.764</td>
<td>0.82</td>
<td>0.848</td>
<td>0.793</td>
<td>0.486</td>
<td>0.848</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.764</td>
<td>0.737</td>
<td>0.743</td>
<td>0.737</td>
<td>0.381</td>
<td>0.743</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>ROC Area</th>
<th>F-Measure</th>
<th>Recall</th>
<th>Precision</th>
<th>TP Rate</th>
<th>FP Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>sick</td>
<td>0.743</td>
<td>0.667</td>
<td>0.75</td>
<td>0.6</td>
<td>0.228</td>
<td>0.75</td>
</tr>
<tr>
<td>Normal</td>
<td>0.743</td>
<td>0.819</td>
<td>0.772</td>
<td>0.871</td>
<td>0.25</td>
<td>0.772</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.743</td>
<td>0.771</td>
<td>0.765</td>
<td>0.771</td>
<td>0.243</td>
<td>0.765</td>
</tr>
</tbody>
</table>
In Tables 2, 3, and 4, true positives (TP) are equivalent to the number of samples that are true detected positives. False positives (FP) are equal to the number of samples that are false detected positives, and ROC determines the receiver operating characteristic [3].

Precision, recall, and F-measure criteria have been achieved from formulas 4, 5, and 6, respectively [3]. In formula 5, false negatives (FN) are equal to the number of samples that are false detected negatives [3].

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (4)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (5)
\]

\[
\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)
\]

According to Tables 2, 3, and 4, the highest accuracy of TP Rate at Sick class was 0.75, belonging to J48 algorithm. The highest accuracy of TP Rate at Normal class is equal to 0.848, belonging to RBF Network algorithm. Figure 1 compares the TP rate accuracy in both Sick and Normal classes for the algorithms used.

![Figure 1. Comparison of TP Rate criterion in algorithms used for diagnosis of Type II Diabetes](image1)

In Tables 2, 3 and 4, the highest FP Rate accuracy at Sick class is 0.228, belonging to J48 algorithm, and the highest FP Rate accuracy at Normal class is 0.486, belonging to RBF Network algorithm. Figure 2 also compares the accuracy of FP Rate at both Normal and Sick classes for the algorithms used.

![Figure 2. Comparison of FP Rate criterion of the algorithms used for diagnosis of type II Diabetes](image2)
According to Tables 2, 3, and 4, the highest precision accuracy is at Sick class, 0.638, belonging to Naive Bayes algorithm. And the highest precision accuracy at Normal class is 0.871, belonging to J48 algorithm. Figure 3 compares the precision accuracy rate in both Normal and Sick classes for algorithms used.

![Figure 3. Comparison of precision criterion in algorithms used for the diagnosis of type II diabetes](image)

According to Tables 2, 3, and 4, the highest Recall accuracy in Sick class is 0.75, belonging to J48 algorithm. Also, the highest Recall accuracy in Normal class is 0.848, belonging to RBF Network algorithm. Figure 4 compares the Recall accuracy in both Normal and Sick classes for the algorithms used. Recall and TP Rate accuracies are equal in both Normal and Sick classes.

![Figure 4. Comparison of Recall criterion in algorithms used for diagnosis of type II diabetes](image)

Tables 2, 3, and 4 show that the highest F-Measure accuracy in Sick class is 0.667, belonging to J48 algorithm, and the highest F-Measure accuracy in Normal class is equal to 0.834, belonging to Naive Bayes algorithm. Figure 5 compares the F-Measure accuracy in both Normal and Sick classes for the algorithms used.
Also according to Tables 2, 3, and 4, the highest ROC Area accuracy in both Normal and Sick classes is equal to 0.845, belonging to Naive Bayes algorithm. ROC Area criterion for Normal and Sick classes of each algorithm has the same value. Figure 6 compares ROC Area accuracy in both Normal and Sick classes of the algorithms used.

Table 5 shows the error rates obtained from testing 30% of the data, i.e. 230 samples of 768 samples from Naive Bayes, RBF Network, and J48 algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean Absolute Error (MAE)</th>
<th>Root Mean Squared Error (RMSE)</th>
<th>Relative Absolute Error (RAE)</th>
<th>Root Relative Squared Error (RRSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.2677</td>
<td>0.3863</td>
<td>59.5622%</td>
<td>82.7801%</td>
</tr>
<tr>
<td>RBF Network</td>
<td>0.3573</td>
<td>0.4231</td>
<td>79.4914%</td>
<td>90.6753%</td>
</tr>
<tr>
<td>J48</td>
<td>0.3206</td>
<td>0.4239</td>
<td>71.3381%</td>
<td>90.8521%</td>
</tr>
</tbody>
</table>
Table 5 shows that each of the MAE, RMSE, RAE and RRSE criteria is related to the error rate of each algorithm in the diagnosis of type II diabetes. Naive Bayes has the minimum error in MAE, RMSE, RAE and RRSE criteria compared with other algorithms. Figure 7 compares the algorithms in the Mean Absolute Error (MAE) criterion.

Figure 7. Comparison of the MAE criterion in algorithms used for the diagnosis of type II diabetes

Figure 7 manifests that Naive Bayes algorithm has the lowest error rate in MAE criterion compared with J48 and RBF Network algorithms. J48 algorithm has a lower error rate in MAE criterion compared to RBF Network algorithm. Figure 8 shows a comparison of algorithms in the Root Mean Squared Error (RMSE) criterion.

Figure 8. Comparison of RMSE criterion in algorithms used for the diagnosis of type II diabetes

Figure 8 demonstrates that Naive Bayes has the lowest error in RMSE criterion compared to J48 and RBF Network algorithms. RBF Network algorithm has a lower error in the RMSE criterion compared to J48 algorithm. It can be said that the error rate of RMSE criterion in J48 and RBF Network algorithms are equal because they both are 0.4231 and 0.4239, respectively, without a significant difference.
Figure 9 shows comparisons of algorithms in Relative Absolute Error (ARE) criterion.

![Relative Absolute Error (RAE)](image)

Figure 9. Comparison of RAE criterion in algorithms used for the diagnosis of type II diabetes

Figure 9 shows that Naive Bayes algorithm has the minimum error in RAE criterion compared to J48 and RBF Network algorithms. And J48 algorithm has a lower rate of error in RAE criterion compared to RBF Network algorithm.

Figure 10 compares the algorithms in Root Relative Squared Error (RRSE) criterion.

![Root Relative Squared Error (RRSE)](image)

Figure 10. Comparison of RRSE criterion in algorithms used for the diagnosis of type II diabetes

Figure 10 states that Naive Bayes algorithm has the minimum error in RRSE criterion compared to J48 and RBF Network algorithms, and Network RBF algorithm has a lower rate of error in RRSE criterion compared to J48 algorithm. However, the error rate of RRSE in RBF Network and J48 algorithms are equal since both are 90.6753% and 90.8521%, respectively with no significant difference.

Table 6 shows the accuracy rate obtained from testing 30% of the data, i.e. 230 samples of 768 samples from Naive Bayes, RBF Network and J48 algorithms.
Table 6. Evaluation of testing set on 230 samples

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correctly Classified Instances (%)</th>
<th>Incorrectly Classified Instances (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>76.9565 (177 data)</td>
<td>23.0435 (53 data)</td>
</tr>
<tr>
<td>RBF Network</td>
<td>74.3478 (171 data)</td>
<td>25.6522 (59 data)</td>
</tr>
<tr>
<td>J48</td>
<td>76.5217 (176 data)</td>
<td>23.4783 (54 data)</td>
</tr>
</tbody>
</table>

According to Table 6, Naive Bayes algorithm had the most accuracy for the diagnosis of type II diabetes which is equal to 76.95%. This means that Naive Bayes algorithm, evaluating the testing dataset of 230 samples, were able to correctly classify 177 samples. J48 algorithm with accuracy rate of 76.52% was able to correctly classify 176 samples; J48 algorithm accuracy is almost equal to the Naive Bayes algorithm. In fact, Naive Bayes algorithm could correctly classify one more data than J48 algorithm. The least accuracy belongs to RBF Network algorithm, 74.34%. This means that the RBF was able to correctly classify 171 samples through evaluation of the dataset of 230 samples. Figure 11 compares data classification accuracy by the algorithms for the diagnosis of type II diabetes.

4. CONCLUSIONS AND THE FUTURE WORKS

One of the most dangerous diseases is diabetes causing big problems such as heart attacks, strokes, blindness, kidney failure, and blood vessel disease which may lead to amputation, nerve damage, and sexual impotence. Data mining is an important technique for diagnosis of diseases; hence, data mining algorithms such as Naive Bayes, RBF Network, and J48 were used to diagnose type II diabetes. There were 768 samples for diagnosis of the disease; 230 samples were selected for testing. Naive Bayes algorithm with accuracy rate of 76.95% had the highest accuracy compared to J48 and RBF Network for diagnosis. That is why Naive Bayes algorithm had the lowest error rate in MAE, RMSE, RAE, and RRSE criteria compared with J48 and RBF...
Network algorithms. It is recommended to diagnose diabetes with other methods such as Neuro-Fuzzy networks and compare them with the algorithms used in this paper, to determine the better method to diagnose the disease.

REFERENCES