Prediction of Diabetes using a Classification model

Dr.Sujni Paul, Dr.C.Beulah Christalin Latha
Assistant Professor, Assistant Professor [S.G]
Al Dar University College, Karunya University, Dubai, Coimbatore, UAE, India
Email: sujni@aldar.ac.ae

Abstract

This research paper helps in predicting diabetes using a data mining technique classification by building a model. In medical diagnosis knowledge discovery is a very crucial part. Diabetes mellitus is a very extensively growing chronic disease and this is being a great challenge worldwide. Today it is common to different age groups ranging from children to adults. With diabetes patients, doubling every year especially in the UAE and India there is a need to curb this epidemic and help those who are affected by this disease so that they can live a peaceful life. Continuous monitoring of health indicators ensures prompt medical attention and reduction in fatalities. The primary challenge to continuously monitor diabetes is that glucose level measurement requires invasive methods. Data mining is growing in relevance to solving such real world disease problems through its tools. This paper proposes to use the UCI repository Pima Diabetes dataset and generates a classification model in order to predict diabetes using recursive partitioning algorithm. The results indicate that the efficiency of the algorithm can be improved by selection of appropriate features and the appropriate training set for the model.

Key terms:
Data Mining, Knowledge Discovery, Diabetes Prediction, Classification Model

1. Introduction

Data mining is the process of extracting hidden knowledge from large volumes of data. This knowledge is then presented in such a way that humans can easily understand it. Prediction of diseases by analysis of voluminous historical data is one of the most significant applications of data mining. Medical data mining is the process of finding useful patterns that would be helpful in medical diagnosis. The predictability of diabetes will be more effective since earlier detection of disease will be helpful for the patients to take care of themselves. Classification is a supervised learning machine learning technique that helps in construction of models that can be used for prediction. In this paper, we propose a classifier that will detect diabetes with better performance. Diabetes happens when a human body fails to produce insulin, which is required to maintain the rate of glucose. Diabetes can be controlled by taking insulin injections, regular exercise and healthy diet. However, the means for a complete cure of the disease is rare, especially when it is detected at a later stage. Diabetes leads to many other types of diseases such as blindness, blood pressure, cholesterol, heart disease, etc. This paper presents a classification model for diabetic prediction.

1.1 Significance of the Study
The significance of this study is to detect an efficient model that can predict the risk of diabetes with improved accuracy. As diabetes is a very threatening disease, which in turn leads to other complications, early prediction of this disease will help the patients to keep their sugar levels intact by taking healthy diet with required drugs. It helps to maintain the sugar level under control.

2. Review of Related Literature and Studies

In order to support our research paper, the below mentioned literature were reviewed.

The research paper entitled “Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree” referencing “Sudajai Lowanichchai, Saisunee Jabjone, Tidanut Puthasimma, 2012“ talks about diabetes analysis in elders. The result showed that the RandomTree model has the highest accuracy in the classification is 99.60 percent when compared with the medical diagnosis that the error MAE is 0.004 and RMSE is 0.0447. The NBTree model has lowest accuracy in the classification is 70.60 percent when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454.

The research paper “Using Bayes Network for Prediction of Type-2 Diabetes”, Yang Guo, Guohua Bai 2010, Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden concluded the following. The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The dataset used was the Pima Indian diabetes dataset. Preprocessing was used to improve the quality of data. Classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally, Weka tool was used to do simulation, and the accuracy of the resulting model was 72.3%.

Gyorgy J. Simon, Pedro J.Caraballo, et al. “Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus, 2015 proposed the method of distributional association rule mining to identify sets of risk factors and the corresponding patient subpopulations that are significantly increased risk of progressing to diabetes. To discover sets of risk factor, here uses bottom up summarization algorithm, which produces most suitable summary that describes subpopulations at high risk of diabetes. The Subpopulation identified by this summary covered highest risk of patients, had low overlap and was at very high risk. This method is used for when the patient having high risk.

J.Tuomilehto, “Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance”, in proceedings of International Journal of Medical Research, vol. 344, no. 18, pp. 1343-1350, 2001 proposed the data mining relationship for efficient classification. They applied data mining techniques to classify diabetes clinical data and predict whether the patient was affected with diabetes or not. They presented a system that gave training data on that data feature relevance analysis is done then comparison of classification algorithm, selecting classifier then improved classification algorithm is applied and then found out the evaluation that compared with training data. They applied C4.5 Algorithm, which gave classification rate of 91%.

K.C. Tan, E.J. Teoh, Q. Yua, K.C. Goh, “A hybrid evolutionary algorithm for attribute selection in data mining”, 2008, talks about short filtering method which removes undesirable features before classification begins while the wrapper method applies classification algorithm to select optimal features. Wrapper method gives higher classification accuracy. The only drawback of the wrapper approach would be a longer runtime because the ML algorithm has to run iteratively in the search for the attribute subsets.
3. Conceptual Framework of the Study

3.1 Data Mining

Data mining, an interdisciplinary subfield of computer science, is the process of discovering patterns from large data sets. It involves artificial intelligence and machine learning techniques to discover hidden patterns or knowledge in voluminous data. The overall goal of the data mining process is to extract knowledge from a data set and visualize it in a form that can further be used for better purpose. Data mining analyzes data from different viewpoints, categorizes them, and summarizes them into valuable information. It helps us to uncover hidden patterns from the data. These patterns are used for constructing models that are helpful for prediction. Data mining is a widely used concept in today’s world in many fields such as finance, medicine, commerce and customer relationships. Various data mining techniques can be used for prediction purposes. Some of the techniques are classification, association rule mining, clustering and decision trees.

3.2 Classification

Classification is a supervised learning technique that classifies samples into different groups. A classification algorithm can be trained with a sample dataset known as the training set and a classification model can be constructed. During the training process, a classification algorithm detects the relationships between the values of the predictors and the values of the class attribute. The identified relationships are then summarized into a model, which can be applied to predict a dataset whose class attribute values are unknown. The model can then be used to predict data in test dataset. In a classification problem, the learning process knows the target groups or class labels prior. The simplest type of classification is called as binary classification. In binary classification, the target attribute or the class attribute consists of only two labels or two possible values. Classification model predicts categorical class labels and prediction models predict continuous valued functions. As an example, classification models can be built to categorize loan applications as either safe or risky or a prediction model, which will predict the expenditures.

3.3 Programming Environment

R is a language that is used for statistical computation and graphical support. R is an integrated suite of software that contains algorithms for data manipulation, statistical calculation and data visualization. It has a conducive environment for data analysis, statistical investigation, data interpretation and visualization. In our paper, we use the R programming environment to build the classification model and use it to predict data with better performance accuracy.

3.4 Problem Statement

Diabetes mellitus is a common disease that affects a vast majority of the people in many parts of the world. Diabetes affects people usually after the age of 20. According to
WHO statistics, the global prevalence of diabetes among adults above 18 years of age has risen to 8.5% in 2014. Diabetes prevalence has been increasing more in middle and low-income countries. It becomes a cause for other illnesses also like blindness, kidney failure, cholesterol and heart diseases. The deaths due to diabetes and high blood glucose are on the rise. Prediction of diabetes at an early stage would help the patients to maintain the sugar level under control. As data mining techniques prove to be good in predictive analyses, a data mining approach is used to predict the risk of diabetes in the proposed approach. The performance of the algorithm is also measured and improved using feature selection and selection of training set.

4. Methodology

4.1 Research Design

The sample dataset is selected and divided into training and test dataset. Feature selection is an important problem in knowledge discovery. The main aim is to find a feature subset that produces higher classification accuracy. After selection of features, the classification algorithm is applied to build the classification model. Then the model is applied to the test set for predicting the diabetes risk. The performance metrics are measured and evaluated. The proposed work is shown in Fig. 1.

![Fig. 1 Proposed Framework](image)

Feature selection is a data-preprocessing step. This will select the subset of features from whole feature set based on statistical score and will remove redundant features that do not contribute to performance. The types of approaches for feature selection are filter, wrapper and embedded methods.

4.2 Sample and Sampling design

The Pima Indian Diabetes dataset that is available in the UCI repository is chosen as the sample for the experimental setup. This dataset consists of diabetic and non-diabetic records. It consists of eight attributes and a class attributes. There are 768 total instances available in the data set. All the patients in the dataset are females above 21 years of age and they are Pima Indians. The attributes or features of the dataset are shown in Table 1.

<table>
<thead>
<tr>
<th>Attribute_Id</th>
<th>Attribute Name</th>
<th>Attribute_Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes &amp; Non diabetes Data</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Feature Subset Selection based on Efficient Coding</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Building a classification model</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Prediction</td>
<td></td>
</tr>
</tbody>
</table>

2017, ADRJS, All Rights Reserved.
A1 | Pregnant Times | Number of times pregnant
A2 | Plasma Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test
A3 | Diastolic BP | Diastolic Blood Pressure (mm Hg)
A4 | Skin Thickness | Triceps Skin Fold Thickness (mm)
A5 | Serum Insulin | 2-Hour Serum Insulin (U/ml)
A6 | BMI | Body mass index
A7 | Pedigree | Diabetes Pedigree Function
A8 | Age | Age in years
A9 | Class Variable | Zero or One

Table 1 Attributes of the Pima Diabetes Dataset

The dataset is classified using recursive partitioning algorithm and a model has been built. 70% of the records were chosen to be the training set and the remaining 30% are taken as the test set. The performance of the algorithm has been evaluated using accuracy, sensitivity, specificity and precision.

5. Findings of the Study

5.1 Preliminary Results

A preliminary analysis on the results reveals the following insights in the data. The dataset consists of female patients whose ages range from 21-81. The diabetes risks with respect to age are shown in Fig. 2.

![Diabetes risk based on Age Levels](image)

Fig. 2 Diabetes Risk based on Age Groups

The diabetes risk can also be measured as a factor of plasma glucose levels. The results are shown in Fig. 3.
The serum insulin levels also have an impact on diabetes. The diabetes risk with respect to the serum insulin levels is shown in Fig. 4.

5.2 Construction of the Classification Model

One of the most important machine learning approaches that is widely used in classification and prediction of data is classification. A supervised learning can be used to classify data under known labels and to predict data based on a classification model that is built using classification algorithms. This paper builds a classification model using recursive partitioning algorithm to predict the diabetes risk in the sample data set. The recursive partitioning algorithm builds regression or classification model and the result is obtained in the form of binary trees.
Performance Measures

The performance of the model can be evaluated using various performance metrics. This paper measures the performance of the algorithm using three performance metrics namely, accuracy, sensitivity and specificity. These metrics are calculated from the confusion matrix. The confusion matrix is a table that is used to predict the performance of a classification model on a sample set of data. It is used for summarizing the results of a classifier. It is a matrix that shows the number of True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN). The format of the confusion matrix is shown in Table 2.

<table>
<thead>
<tr>
<th>Actual vs Predicted</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Table 2 Confusion Matrix

The formulae for calculating the performance metrics are shown in equations (1) to (4). Accuracy is a statistical measure that calculates how well a binary classification test identifies or excludes a condition correctly. Sensitivity is also known as recall or true positive rate. Sensitivity measures the proportion of positives that are correctly identified as positives. Another performance measure, the specificity that is also known as the true negative rate measures the proportion of negatives that are correctly identified as negatives. Specificity is also known as precision.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{3}
\]

The rpart algorithm generates rules in the form of binary trees. A sample binary tree
model that has been generated by the rpart algorithm is shown in Fig. 5.

Fig. 5 Sample classification tree

The dataset is divided into training and test sets and the results are evaluated. Attribute or feature subset selection has also been applied in order to increase the accuracy of results. The attribute subset selection focuses on identifying an attribute subset that improves the classification accuracy. The attributes that produced the highest accuracy are shown in Table 3.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1, A2, A3, A6</td>
<td>79.08%</td>
<td>90.56%</td>
<td>56.26%</td>
</tr>
<tr>
<td>A1, A2, A3, A4, A6</td>
<td>79.08%</td>
<td>90.56%</td>
<td>56.26%</td>
</tr>
<tr>
<td>A1, A2, A3, A4, A5, A6</td>
<td>79.08%</td>
<td>90.56%</td>
<td>56.26%</td>
</tr>
<tr>
<td>A2, A4, A5, A6, A8</td>
<td>79.08%</td>
<td>87.42%</td>
<td>62.5%</td>
</tr>
<tr>
<td>A1, A2, A3, A7</td>
<td>78.66%</td>
<td>88.05%</td>
<td>59.99%</td>
</tr>
<tr>
<td>A1, A2, A3, A4, A5, A6, A8</td>
<td>78.66%</td>
<td>87.42%</td>
<td>61.25%</td>
</tr>
<tr>
<td>A1, A2, A3, A4, A5, A8</td>
<td>78.24%</td>
<td>84.28%</td>
<td>66.25%</td>
</tr>
</tbody>
</table>

Table 3 Performance Measures with Highest Value

The feature subset selection has also been tried by eliminating the attributes one by one from the dataset. The performance measures are tabulated in Table 4.

<table>
<thead>
<tr>
<th>Removed Attribute</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Attribute Set</td>
<td>77.82%</td>
<td>86.16%</td>
<td>61.25%</td>
</tr>
<tr>
<td>A8</td>
<td>74.48%</td>
<td>76.73%</td>
<td>70.00%</td>
</tr>
<tr>
<td>A7, A8</td>
<td>79.08%</td>
<td>90.56%</td>
<td>56.26%</td>
</tr>
<tr>
<td>A6, A7, A8</td>
<td>77.41%</td>
<td>88.05%</td>
<td>56.25%</td>
</tr>
<tr>
<td>A5, A6, A7, A8</td>
<td>76.14%</td>
<td>87.42%</td>
<td>53.73%</td>
</tr>
<tr>
<td>A4, A5, A6, A7, A8</td>
<td>74.06%</td>
<td>88.68%</td>
<td>44.99%</td>
</tr>
<tr>
<td>A3, A4, A5, A6, A7, A8</td>
<td>76.15%</td>
<td>94.96%</td>
<td>38.75%</td>
</tr>
<tr>
<td>A2, A3, A4, A5, A6, A7, A8</td>
<td>70.29%</td>
<td>86.79%</td>
<td>37.50%</td>
</tr>
</tbody>
</table>

Table 4 Performance Measures with attribute selection

The results showed highest accuracy when the attributes pedigree and age are removed from the attribute set. Not only the selection of attributes, the selection of the training and test data sets also have an impact on the performance of the algorithm. The algorithm produces greater accuracy when the training set data is increased above 85%. The accuracy of the algorithm when the ratio between training set and test data set is varied is shown in Fig. 6. The attribute sets are renamed as follows: Set 1 : {A1, A2, A3, A6}, Set 2 : {A1, A2, A3, A4, A6}, Set 3: {A1, A2, A3, A4, A5, A6}, Set 4: {A2, A4, A5, A6, A8}, Set 5: {A1, A2, A3, A7}, Set 6 : {A1, A2, A3, A4, A5, A6, A8} and Set 7: {A1, A2, A3, A5, A8}.
The accuracy of the model increases for most of the attribute sets when the training set size is increased. The variation of sensitivity is shown in Fig. 7.

The variation of specificity is shown in Fig. 8.
Fig. 8 Specificity of the classification model for varying training-test set ratio

The ROC curve in a classification model portrays the performance of a classifier system when its discrimination threshold is varied. It is a curve of the sensitivity as a function of the fallout. The ROC curve drawn with true positive rate and false positive rate for the selected attributes is shown in Fig. 9.

Fig. 9 ROC curve of the classification Model

The ROC curve shows a comparison of the selected models. The model with the largest area is considered as the best model. The results show that the model constructed with pedigree attribute removed is the best model. The recall versus precision curve is shown in Fig. 10. This curve shows the trade-off between precision and recall and an appropriate balance between the two.
The curves show that some features have higher precision and some have greater recall. Set 1, Set 3, and Set 7 attributes produce higher precision and the other sets produce higher recall.

6. Conclusions

This paper presents an approach of building a classification model using recursive-partitioning algorithm and implements that model on a dataset for classifying diabetes patients’ data. The model has been trained to classify the diabetes patients from non-diabetes persons and it is used to predict the risk of diabetes on another dataset. The performance of the model has been evaluated using the performance measures such as accuracy, sensitivity and specificity. The performance of the algorithm has been improved by feature subset selection and by varying the size of the training dataset. The models are compared using ROC curve and recall-precision curves.

List of References

Sudajai Lowanichchai, Saisunee Jabjone, Tidanut Puthasimma, (2012)“Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree”.

Yang Guo , Guohua Bai , Yan Hu School of computing Blekinge (2010) Institute of Technology Karlskrona, Sweden, “Using Bayes Network for Prediction of Type-2 Diabetes”.

