Prediction of Type-2 Diabetes Foot Ulcer - A Comparative Study with Classification Algorithm

R. Christy Pushpaleela and R. Padmajavalli

1Bharathiar University,
Coimbatore.
Department of Computer Science,
Women’s Christian College, Chennai.
leela_viswa@yahoo.com

2Bharathiar University,
Coimbatore.
Bhaktavatsalam Memorial College for Women,
Chennai.
padamahari2002@yahoo.com

Abstract

This research paper focuses upon the study of foot ulcer-diabetic data and implement various classification algorithms on different kinds of medical datasets of foot ulcer-Diabetic to evaluate its comparative performance. Performance analysis records the most frequently used algorithms on respective medical datasets and most efficient classification algorithm to analyze the foot ulcer-Diabetic disease based on the comparative study between C4.5 and SVM on high dimensional datasets. The scope of research work is aimed at predicting diabetics for foot ulcer patients by comparing various classification algorithms.

Keywords: Classification, prediction, clustering, diabetics, KNN, decision tree, SVM, C4.5.
1. Data Mining

Data mining is the process of discovering actionable information from huge sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Patterns and trends can be collected and defined as a data mining model. Diabetes is a group of metabolic illness due to high blood sugar levels. It may also cause many complication like as heart problem, kidney issues, foot ulcers. There are three types of diabetics issue namely.

Type 1 - Affects adult and children's- Insulin dependent diabetes.

Type 2 - Affects only Adult -Non Insulin dependent diabetes and Gestational Diabetes - This will occur when pregnant women without a previous history of diabetes develop a high blood glucose level.

Data mining involves pulling out of useful information from the huge volume of data. Data mining technique has been applied in various fields like banking, insurance, medicine etc.

Data mining process followed the below process to get the final model for predication.

- Problem Definition.
- Preparation.
- Data Exploring.
- Model Building for predication.
- Exploring & Validating models.
- Deploying & Updating models.

There are two forms of data analysis that can be used for extracting models.

Classification and Prediction.

2. Primary Objective

The Primary Objectives of the present work is as follows:

- To compile the aggregated real time data set from reputed Hospitals.
• To Pre-process the clustered real time patient data for foot ulcer-diabetic prediction consuming both manual techniques and Preprocessing tools like Weka.
• To apply the classification algorithm for pre-processed foot data to identify the optimum Predictor.
• To classify the efficient algorithm based on the performance analysis report employing best accuracy rate.

Related work

Auria, Laura and Moro, R. A [6] introduced a statistical technique, SVM which provide a higher accuracy of company classification into solvent and insolvent. SVM's are more accurate and convenient even when input data is non-monotonous and non-linearly separable. SVM's do not deliver parametric score but can offer support for recognizing different financial ratios.

Dr K S OzaMr V S Kumbhar [8] presented a comparison of different data mining techniques like k-nearest neighbor (KNN), Bayesian network, Decision trees, support vector machines based on their accuracy, evaluation time and rate of error.

NongyaoNai-aruna and RungruttikarnMoungmaia [5] introduced a web application by using a use of disease classifiers and a real data set. Various algorithms were used to classify the risk of diabetes mellitus. Decision Tree, Regression, Naive Baye and Random Forest are used to classify the risk of diabetes. Based on the various random selection found out that Random Forest algorithm is effective in increasing the accuracy value and predicting the diabetes.

VrushaliBalpande and RakhiWajgi [4]presented a research work for prediction of diabetes using data mining technique. Data mining has been applied in various fields like medicine, marketing, banking, etc. In medicine, predictive data mining is used to diagnose the disease at the earlier stages itself and helps the physicians in treatment planning procedure. Existing Data mining method which is efficient in predicting diseases is used for prediction of Diabetes.

Kawsar Ahmed, TasnubaJesmin [1] presented a research work for 20 classification algorithms are compared by measuring accuracies, speed and robustness using WEKA tool to classify diabetes patient data.

J.S.Raikwal and KanakSaxena [3] proposed methodology to predict the future disease of the patients using data patterns. SVM and KNN algorithm is used to classify data -For example medical patients data to find hidden patterns for targets, like predicting the future diseases of the patients.

Vincenzo Lagani a, Franco Chiarugi [13] implemented a Data-mining analyses of the DCCT/EDIC data allow the identification of true predictive models for diabetes-related impediments.

3. Data Pre-Processing Steps

• Step 1. Obtained real time medical data sets are pre-processed manually.
• Step 2: Excel data is converted into .CVS format further to .Arff format for effective processing.
• Step 3: Weka tool is launched alongside explorer window where in "Preprocess tab" is selected.
• Step 4: Arff file format is opened to choose the attributes filed in the data set (e.g. number of instances, attributes etc).
• Step 4: Respective data set is entered and implemented by selecting Visualization button.

Real-time-Data from Leading Hospital

455 data of patient (355 diabetes patients & 100 non diabetes patients) is collected from leading hospital. There are male and female patients whose age between 15 to 76 years old and sample data set listed below.

<table>
<thead>
<tr>
<th>Patient Id</th>
<th>Age</th>
<th>SEX</th>
<th>BMI</th>
<th>BPH</th>
<th>BPL</th>
<th>SugarP</th>
<th>SugarPo</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR01</td>
<td>53</td>
<td>F</td>
<td>31.71</td>
<td>139</td>
<td>73</td>
<td>236</td>
<td>412</td>
</tr>
<tr>
<td>STAR02</td>
<td>60</td>
<td>M</td>
<td>28.32</td>
<td>123</td>
<td>77</td>
<td>172</td>
<td>260</td>
</tr>
<tr>
<td>STAR03</td>
<td>62</td>
<td>M</td>
<td>31.71</td>
<td>130</td>
<td>80</td>
<td>187</td>
<td>268</td>
</tr>
<tr>
<td>STAR04</td>
<td>54</td>
<td>F</td>
<td>28.32</td>
<td>154</td>
<td>84</td>
<td>162</td>
<td>353</td>
</tr>
<tr>
<td>STAR05</td>
<td>67</td>
<td>M</td>
<td>28.33</td>
<td>140</td>
<td>89</td>
<td>143</td>
<td>250</td>
</tr>
<tr>
<td>STAR06</td>
<td>40</td>
<td>M</td>
<td>28.44</td>
<td>128</td>
<td>67</td>
<td>130</td>
<td>189</td>
</tr>
<tr>
<td>STAR07</td>
<td>55</td>
<td>F</td>
<td>32.55</td>
<td>130</td>
<td>90</td>
<td>156</td>
<td>240</td>
</tr>
<tr>
<td>STAR08</td>
<td>62</td>
<td>F</td>
<td>31.71</td>
<td>108</td>
<td>62</td>
<td>283</td>
<td>345</td>
</tr>
<tr>
<td>STAR09</td>
<td>61</td>
<td>F</td>
<td>28.32</td>
<td>125</td>
<td>78</td>
<td>112</td>
<td>197</td>
</tr>
<tr>
<td>STAR10</td>
<td>56</td>
<td>M</td>
<td>31.71</td>
<td>130</td>
<td>80</td>
<td>128</td>
<td>192</td>
</tr>
<tr>
<td>STAR11</td>
<td>55</td>
<td>F</td>
<td>28.32</td>
<td>105</td>
<td>71</td>
<td>137</td>
<td>209</td>
</tr>
<tr>
<td>STAR12</td>
<td>68</td>
<td>F</td>
<td>28.32</td>
<td>120</td>
<td>70</td>
<td>171</td>
<td>224</td>
</tr>
<tr>
<td>STAR13</td>
<td>43</td>
<td>M</td>
<td>28.44</td>
<td>125</td>
<td>84</td>
<td>98</td>
<td>196</td>
</tr>
<tr>
<td>STAR14</td>
<td>35</td>
<td>F</td>
<td>32.55</td>
<td>110</td>
<td>70</td>
<td>168</td>
<td>269</td>
</tr>
<tr>
<td>STAR15</td>
<td>45</td>
<td>M</td>
<td>31.71</td>
<td>103</td>
<td>78</td>
<td>123</td>
<td>231</td>
</tr>
</tbody>
</table>

Figure 2: Weka Tools – Pre-Process Flow
4. Classification Procedures

There are several classifications methods propose by researchers. Some of the methods are

- Support Vector Machine.
- Decision tree.
- Bayesian classification.
- C4.5.
- K-NN.

**SVM Support Vector Machine**

A support vector machine is a Classification method.

It is mostly used in classification and regression problems. SVM will support data mining, text mining, and pattern recognition. SVM provides the user can easy analysis real data or artificial data. SVM delivers good and optimal solutions [3][6][11][15].

**J48(C4.5)**

It is an algorithm used to generate a decision tree and is a successor of earlier ID3 Algorithm. C4.5 algorithm is a greedy algorithm developed by Ross Quinlan, used for the induction of decision trees. It can be used for classification of the data and so referred to as statistical classifier [15][25].

**Decision Tree**

Decision Tree will build a predictive model which is mapped to a tree organization. Means form of left and right child, root node compensations. Decision is a knowledge representation structure consisting of nodes & branches organized in the form of Tree such that, every internal non-leaf node is labeled with values of the attributes [15].

**Naïve Bayes**

It is a Standard probabilistic Naïve Bayes classifier. Naive Bayes classifiers are the group of simple probabilistic classifiers based on applying Bayes' theorem. Naïve Bayes classifiers are
highly scalable. Naive Bayes model is easy to build for very large data sets [17] [11].

K-NN

K-NN is a type of instance-based learning. KNN is a group of simple algorithm like as Classification and Regression. The main advantages of KNN are below [5] [22].

- Easy implementation.
- Robust.
- Very low cost.

5. Disease Summary

Diabetic foot ulcer [33]

A diabetic foot ulcer is an open sore or wound that occurs in approximately 15 percent of patients with diabetes. Commonly located on the bottom of the foot [3] [2].

Diabetes Foot Ulcer - Symptoms

Diabetic ulcers are most commonly affected

- Poor circulation.
- High blood sugar (hyperglycemia).
- Nerve damage.
- Irritated or wounded feet.

Doctor will identify the scale range between 0 to 2 parameter

- 0: Means ulcer not present but foot at risk
- 1: Means Ulcer present but no infection
- 2: Means ulcer deep exposing joints and tendons

Risk Factors for Diabetic Foot Ulcers Patients [1] [5] [7] [8].

- Poor quality shoes.
- Poor hygiene (not washing regularly or thoroughly).
- Improper trimming of toenails.
- Alcohol consumption.
- Eye disease from diabetes.
- Heart disease.
- Kidney disease.
- Obesity.

Figure 4: Symptoms Ulcer Foot Ulcer [3] [2].
Data Set (Diabetic foot ulcer- Data Set)

<table>
<thead>
<tr>
<th>Sno</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient Id</td>
</tr>
<tr>
<td>2</td>
<td>Name</td>
</tr>
<tr>
<td>3</td>
<td>Age</td>
</tr>
<tr>
<td>4</td>
<td>SEX</td>
</tr>
<tr>
<td>5</td>
<td>BMI</td>
</tr>
<tr>
<td>6</td>
<td>WeightKG</td>
</tr>
<tr>
<td>7</td>
<td>HeightIN</td>
</tr>
<tr>
<td>8</td>
<td>BPH</td>
</tr>
<tr>
<td>9</td>
<td>BPL</td>
</tr>
<tr>
<td>10</td>
<td>SugarPre</td>
</tr>
</tbody>
</table>

Table 2: Data Set attribute

<table>
<thead>
<tr>
<th>S.no</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>SugarPost</td>
</tr>
<tr>
<td>12</td>
<td>HBA1CP</td>
</tr>
<tr>
<td>13</td>
<td>Hemoglobin</td>
</tr>
<tr>
<td>14</td>
<td>UREA</td>
</tr>
<tr>
<td>15</td>
<td>Creatinie</td>
</tr>
<tr>
<td>16</td>
<td>GlycemicIndex</td>
</tr>
<tr>
<td>17</td>
<td>Single/multiple ulcer</td>
</tr>
<tr>
<td>18</td>
<td>Alcohol</td>
</tr>
<tr>
<td>19</td>
<td>Etiology</td>
</tr>
<tr>
<td>20</td>
<td>Culture test</td>
</tr>
</tbody>
</table>

6. Classification Matrix

The basic phenomenon used to classify the diabetic foot ulcer classification using classifier is its performance and accuracy. The performance of a chosen classifier is validated based on error rate and computation time [11][13].

Table 3: Confusion Matrix

<table>
<thead>
<tr>
<th>Predictable</th>
<th>Categorized as a Healthy(0)</th>
<th>Categorized as a not Healthy(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Healthy (0)</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Really not Healthy (1)</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Where

**TP** - True Positive & **FN** – False Negative

**TN** – true Negative & **FP** - False Positive

For measuring accuracy rate the following mathematical model is used [22][18][21][2].

**Accuracy** = **TP+TN / TP+FP+ TN +FN**
7. **Assessment Result**

Table 4: Performance Report for Classification

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>92.22</td>
</tr>
<tr>
<td>C4.5</td>
<td>85.68</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>75.45</td>
</tr>
<tr>
<td>KNN</td>
<td>73.77</td>
</tr>
</tbody>
</table>

Figure 5: Performance Report for Classification

8. **Conclusion and Future Work**

Performance comparison of SVM and C4.5 Classifiers for foot ulcer-Diabetic disease are recorded to obtain effective methodology for prediction. Preprocessing is carried out using Weka for building optimum predictability model. The maximum classification accuracies of the SVM and C4.5 classifiers were found to be 92.22% and 85.68% respectively. Based on the comparative study, it is concluded that SVM Classifier method produces better accuracy ranges than C4.5 and other classification algorithm for foot ulcer disease prediction in terms of Classification accuracy.

Furthermore several available classifications techniques will also be considered to perform risk factor prediction to arrive at concluding the best suitable algorithm for diabetic prediction in foot ulcer.

**References**


[16] Thair N.P., Survey of Classification Techniques in Data Mining.

[17] Thangadurai K., Nandhin N., Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus.


