Comparative Study of Decision Tree and Neural Network for the Analysis of Diabetes Data

Gauri D. Kalyankar  
Dept. of CSE,  
Rajarambapu Institute of Technology, Islampur  
gaurikalyankar93@gmail.com

Shivananda R. Poojara  
Dept. of CSE,  
Rajarambapu Institute of Technology, Islampur  
shivu.poojara@gmail.com

Nagaraj V. Dharwadkar  
Dept. of CSE,  
Rajarambapu Institute of Technology, Islampur  
dharwadkarn@gmail.com

Manoj A. Patil  
Dept. of IT,  
Rajarambapu Institute of Technology, Islampur  
manupatil85@gmail.com

Abstract—Now days Diabetic Mellitus (DM) has become a very big health problem in worldwide as the number of diabetic patients are increasing every day. The Diabetes Mellitus is one of the critical deceases and has long term complications associated with it. Patients who suffer from Diabetes may suffer from different health problems such as heart stroke, kidney failure, high blood pressure etc. Hence it is necessary to use efficient techniques to process and analyze such diabetes data for the diagnosis of Diabetic Mellitus with better accuracy. Different data mining, machine learning and statistics can be used to analyze medical dataset efficiently. In this paper two machine learning algorithms Decision Tree and Neural Network are used to analyze Pima Indian Diabetes Dataset. This paper aims to compare the performance accuracy of both the classifier against Pima Indian Diabetes Dataset.

Keywords—Diabetes Mellitus, Machine Learning Algorithms, Classifiers, Decision Tree, Neural Network

I. INTRODUCTION

Now days from healthcare industries large amount of data is generating. This data contains lots of useful information that can be used to take efficient medical decisions. Hence it is necessary to store and process such data to extract knowledge from it and by using that take efficient decisions. There are various techniques such as data mining, machine learning, statistical analysis etc can be used to analyze medical data. By using these recent techniques early diagnosis of various diseases can be done so that patients can be provided with better cure and care [1].

In current situations the numbers of people suffering from Diabetes are increasing per day. Diabetic Mellitus (DM) belongs to the family of Non Communicable Diseases (NCD). It is one of the dangerous diseases as it has long term impacts on patient’s body. Diabetic Mellitus (DM) is categorized into three types. First is Type 1 DM which is called as Insulin - Dependent Diabetes Mellitus (IDDM). Type 1 DM is caused when patient’s body is unable to produce insulin and requires injecting insulin externally to the patient. The second type of DM is Type 2 DM which is called as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of diabetes is caused when patient’s body cell are not able to use insulin properly. Third type of DM is gestational diabetes which is caused in pregnant women due to the development of high blood sugar without the knowledge of

Pre-diagnosis of Diabetes. This type of DM is very dangerous and can lead to Type 2 DM [1]. As numbers of patients of this serious disease are increasing day by day, the size of diabetic data set is also increasing. With the utilization of machine learning algorithm, it can be possible to analyze the diabetes data and accordingly provide early diagnosis and better treatment [1] [4].

Machine learning is the set of different methods that can be used to find patterns from the dataset and then use those patterns to predict future conditions or to make efficient decisions under some conditions. Machine learning introduces different algorithms that can be used to make understanding of the current situations to machine and then based upon that machines can take decisions. Machine learning can make decisions independently at its own [6] [7].

Machine learning is categorized into two types, supervised learning and unsupervised learning.

1. Supervised Learning: In supervised learning the input and output both is known. This type of machine learning is known as supervised learning because it learns from the training data set and creates knowledge base and applies it on the testing data set to make appropriate decisions.

2. Unsupervised Learning: In Unsupervised learning we are aware about only input. Hence the very important task of unsupervised learning algorithm is to build up class labels automatically. Unsupervised learning algorithm finds relationship between the data and then same characterized data are grouped together which is known as cluster. Cluster analysis is the second name of unsupervised learning [6] [7].

In this work two machine learning algorithms, Decision Tree (C4.5/J.48) and Neural Network (Multilayer Perception) are used for the analysis of diabetes data.

II. PRIOR WORK

The review of prior work gives information about various works that has been carried out by various researchers for the analysis of diabetes data sets. These researchers have used different techniques and developed systems for the analysis of health care data. These techniques include data
mining, machine learning, Hadoop, statistic techniques or combination of these techniques

Eswari et al. (2015) proposed the system which is based on Hadoop and MapReduce for the analysis of diabetic data. By this analysis the proposed system is able to predict diabetic type’s prevalent and complications associated with it [1].

Aiswarya Iyer et al. (2015), used classification method to find out patterns from the diabetes datasets. They implemented two algorithms, Naïve Byse and Decision Tree by using Weka. Authors have done comparison of performance of these two classifiers against Pima Diabetes Datasets. Authors showed experimental results of each classifier [3].

Gaganjot Kaur and Amit Chhabra (2014), proposed the approach for the prediction of diabetes. Authors improved J48 for increasing the rate of accuracy of data mining procedures. Authors used MATLAB with WEKA as an API for creating the J48 classifiers. The experimentation is done on the Pima Indian Diabetes Datasets. They have shown the experimental results which conclude the efficient improvement over the existing J48 algorithm [9].

Sujata Joshi and S. R. Priyanka Shetty (2015), implemented different classification techniques namely J48, Naive Bye, CART, KNN, Random Forest, Bayesian network, REP, Conjunctive rule learning etc for the analysis of diabetes data sets. This experimentation is done on the Pima Indian Diabetes Datasets. From the experimental results authors conclude that J48 is the best classifier for the diabetes data analysis [10].

Olaiya Folorunsho (2013), measured the performance of two classifiers, Artificial Neural Network and Decision Tree on medical data-sets. On the basis of different performance measurement parameters performance measurement is done. For the experimentation there was data which contains 200 records is used. Experimental results showed that performance of Decision Tree is better than the performance of Artificial Neural Network [11].

Nipiyoti Sarma et al. (2014), compared the measurement of performance and accuracy of Naive Bye and Decision Tree classifier for the diagnosis of Diabetes. For the comparison, authors have done survey of various papers in this area and by considering the results of different researchers and mentioned the accuracy parameters for the performance measurement. The surveying and comparison is done upon Pima Indian Diabetes Dataset [12].

In prior research most of the researchers are concluded that Decision Tree and Neural Network give better performance for the analysis of healthcare data. Hence for this work we have chosen these two classifiers for the analysis of diabetes data.

III. PROPOSED METHOD

A. System Architecture

The system architecture contains different tasks like applying Decision Tree and Neural Network on diabetes dataset, analysis of results of each classifier and comparing the accuracy of both classifiers. Figure1 describes overall architecture of proposed work.

B. Diabetes Dataset

To perform the experimentation Pima Indians Diabetes Dataset has been used. This dataset is located at UCI Machine Learning Repository [8]. The dataset contains 8 attributes and 1 class variable. This data set contains total 768 diabetic and non-diabetic women records whose age is above 21 years. All attributes in data set are numeric. For this work the dataset is considered as complete data set having no missing value in it. Following table describes the attributes in data set [8].
Table 1: Attribute Names

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Attribute Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of times pregnant</td>
</tr>
<tr>
<td>2</td>
<td>Plasma glucose concentration a 2 hours in an oral glucose tolerance test</td>
</tr>
<tr>
<td>3</td>
<td>Diastolic blood pressure (mm Hg)</td>
</tr>
<tr>
<td>4</td>
<td>Triceps skin fold thickness (mm)</td>
</tr>
<tr>
<td>5</td>
<td>2-Hour serum insulin (mu U/ml)</td>
</tr>
<tr>
<td>6</td>
<td>Body mass index (weight is in kg and height in m)</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes pedigree function</td>
</tr>
<tr>
<td>8</td>
<td>Age (years)</td>
</tr>
<tr>
<td>9</td>
<td>Class variable (0 or 1)</td>
</tr>
</tbody>
</table>

**C. J48 Decision Tree**

J48 is java based open source implementation of C4.5 Decision Tree algorithm which enables to make decision tree. C4.5 is improved from the ID3 algorithm and developed by Ross Quinlan. C4.5 Decision Tree is used for the classification. C4.5 can process both continuous and discrete values, missing values and can do pruning of tree after the construction of tree [13].

C4.5 is used to solve the classification problems. C4.5 is based on information gain ratio which can be calculated by using entropy. The information gain ratio is used to select test attribute at each node of tree and this measure is called selection criteria of an attribute. The entropy, information gain and gain ratio can be calculated using below equations and recursively computed on sub tree [5].

The number of classes is denoted by c, and p(S, j) is the proportion of instances in S, that are assigned to jth class. Therefore, the entropy of attribute S is calculated as:

\[
\text{Entropy}(S) = \sum_{j=0}^{c} p(S, j) \times \log p(S, j)
\]

Accordingly, the information gain by a training data set T is defined as:

\[
\text{Gain}(S, T) = \text{Entropy}(S) - \sum_{v \in \text{Values}} \frac{\text{TS,v}}{T} \times \log \frac{\text{TS,v}}{T}
\]

Where Values (TS) is the set of values of S in T, TS is the Subset of T induced by S, and (TS, v), is the subset of T in which attribute S has a value of v.

Therefore, the information gain ratio of attribute S is defined as:

\[
\text{GainRatio}(S, T) = \frac{\text{Gain}(S, T)}{\text{SplitInfo}(ST)}
\]

Where SplitInfo (S, T) is calculated as:

\[
\text{SplitInfo}(S, T) = \sum_{v \in \text{Values}} \frac{\text{TS,v}}{T} \times \log \frac{\text{TS,v}}{T}
\]

**D. Neural Network**

Neural Network is the classifier which is mostly used as tool for solving various classification problems. A multilayer perception is feed forward neural network model follows back-propagation algorithm technique that is widely utilized for number of classification problems. The working of neural network is same as working of human brain. It is supervised learning model hence need to train first. It learns how to convert input data in to desired outcome hence most of the time it is used for the pattern classification [11] [15].

The multilayer perception is based on error correction learning. The working of error correction learning is as follows. From the system response at neuron j, at iteration t, yj(t) and preferred outcome dj(t) for a given input pattern an instantaneous error \( e_j(t) \) is defined as \( e_j(t) = d_j(t) - y_j(t) \). The theory of gradient learning is used to modify each weight in the network by correcting the current value of the weight with a value that is proportional to the present input and error at the weight i.e.

\[
w_{jk}(t+1) = w_{jk}(t) + \eta \delta_j(t) x_k(t)
\]

\( \eta(t) \) is the learning rate parameter. At the iteration \( t \), the \( w_{jk}(t) \) is the weight which connects the output of neuron \( k \) to the input neuron \( j \). The local error \( d_j(t) \) can be directly calculated from \( e_j(t) \) at the output neuron or can be calculated as a weighted sum of errors at the internal neurons [2].

The sensitivity of a cost function with respect to each weight in current network is calculated by Back-propagation and each weight is updated which is proportional to the calculated sensitivity. The advantage of this process is that it can be employed with local information and can be implemented with very few multiplications performed per weight, which is very effective approach [2].
WEKA is the open source data mining tool used throughout the work for the experimentation. WEKA software contains different machine learning algorithms that can be used for data mining tasks. In 1994 the WEKA tool is developed by WEKA team. WEKA is platform independent software which is easily available at any time. The people who have very less knowledge about the data mining can also use this tool for the experimentation. WEKA has become one of the popular tools for the analysis of data. WEKA contains many powerful functions which can be used to analyze complex data [9] [14].

The following tasks can be performed using WEKA

1. Data pre-processing, missing value imputation and visualization
2. Attribute Selection
3. Classification Tree
4. Prediction
5. Clustering
6. Association rules

IV. RESULTS AND DISCUSSION

A. Input Dataset

Pima Indian Diabetes Dataset (PIDD) is used for experimentation. Pima Indian Diabetes Dataset contains 8 attributes and one binary class variable (1 is for diabetic patient and 0 is for non-diabetic patient). PIDD contains total 768 records of female patients.

We have used 70:30 percentage split for performing training and testing. Then J48 and Multilayer Perception are separately applied on the training dataset which contains 538 records and then knowledgebase is applied on testing dataset. WEKA tool is used for the experimentation.

B. Comparative Study of Decision Tree and Neural Network

For the testing dataset accuracy of classifier is calculated. There are total 230 records are used for testing purpose out of which correctly and incorrectly classified instances by each classifier are recorded. Following tables describes the results of each classifier.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Results (out of 230)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly classified instances</td>
<td>159</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td>71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification</th>
<th>Results (out of 230)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly classified instances</td>
<td>178</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td>52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>69.1304</td>
</tr>
<tr>
<td>Multilayer Perception</td>
<td>77.3930</td>
</tr>
</tbody>
</table>

From the results of both the classifiers, it is clear that the accuracy of Neural Network is higher than the accuracy of Decision Tree classifier and hence the performance of Neural Network is better than the performance of Decision Tree. Neural Network is efficient algorithm for the analysis of diabetes data.
V. CONCLUSION

In health care sector Diabetic Mellitus (DM) has become a big challenge now days. With the help of technology it is necessary to build systems that can be used for early diagnosis of such type of diseases. Machine learning techniques are very useful to process and analyze such medical data sets. In this work we have used two classifiers Decision Tree (J48) and Neural Network (Multilayer Perception) for the analysis of diabetes data sets. The experimentation is done on WEKA tool. The Decision Tree gives 69.1304 % accuracy and Neural Network gives 77.3900% accuracy. The performance of Neural is better than Decision Tree.

In future work we can improve the algorithms for getting more accuracy and also can use different machine learning techniques to improve the accuracy.

REFERENCES