Associative Based Classification Algorithm For Diabetes Disease Prediction

N. Gnana Deepika, Y.Surekha, G.Lalitha Kumari

1Mtech Student, Dept. of CSE, PVPSIT, India.
2Assistant Professor, PVPSIT, Vijayawada, AP, India.
3Sr. Assistant Professor, PVPSIT, Vijayawada, AP, India

Abstract: Now a day's data mining plays a vital role in prediction of diseases in medical domain. Data mining is the process of exploring hidden patterns from large amounts of disease data to find unknown relationships and patterns to the medical experts. These rules can be used for better clinical decision making and suggestive medicine. Diabetes is the most common disease in all age groups, this disease is predicted sign artificial intelligence model in the past decade. Also, early detection of diabetes is an essential role in diabetes because it can unearth hidden knowledge from a vast amount of medical data. This research provides an optimal approach to filter data and to discover hidden patterns using improved decision tree model. Experimental results proved that proposed model has high true positive rate and precision compared to traditional classification models.

Keywords: Decision Tree, Disease detection, ensemble model, classification, UCI repository.

1. INTRODUCTION

Decision tree, Naïve bayes, neural network and SVM Techniques are used to classify the medical instances based on training data. Even though each technique has its own strengths in finding medical diseases, there are also limitations in several detection techniques, usually takes time to build the model, takes time to load the large volume of data and high false positive rate. Different types of classification mechanisms have been implemented with a series of multi stage classifier to overcome the limitation on high false positive rate. In each stage, binary classifier is used to reduce the medical features by partitioning the data into single class and normal type of diseases.

To define the information measure correctly, the well known concept in information theory, i.e. entropy is used to find the most common information of the random group of samples. Entropy measure is used to compute the impurity level in a group of samples. The information gain and gain ratio measures can be computed as follows:

\[ Entropy = \sum_{i=0}^{n} p_i \log( p_i ) \]

The purpose of data mining is to help the decision maker to discover potential patterns from large volumes of data for decision making. Data mining is a relatively a new framework for intrusion detection and prevention mechanism. Several data mining models that include classification, association, clustering and rule mining techniques have been introduced in the literature for medical disease detection.

Construction of decision trees starts by selecting the attribute that is best allowed to increase the difference of the classes by splitting the attribute into all of its possible values.

Discretization of continuous attributes simply not only broadens the scope of a given range of data mining algorithms able to analyze data in discrete form, but might also dramatically amplify the speed at which these tasks can be carried out. A discrete feature, also known as qualitative features, including sex and level of education, is only able to be limited among a number of values. Continuous features might be ranked if you want and admit to meaningful arithmetic operations.
There are several benefits associated with using discrete values over continuous ones:

- Discretization will reduce the number of continuous features' values, which shows smaller demands on system's storage.
- Discrete features are in close proximity to a knowledge-level representation than continuous ones.
- Data can also be reduced and simplified by discretization. For both users and experts, discrete features are easier to comprehend, use, and explain.
- Discretization makes learning more accurate and faster.
- Existing system suffering does not give correct discretization if the continuous attribute value is empty or noisy.
- Existing approach does not handle data with binary and ratio interval valued attributes.
- Existing entropy-based discretization gives duplicate values on decision tree algorithms.
- Existing rule mining algorithm gives more true negative and false negative error rate.

The rest of the paper is summarized as follows. The work related to the different defect prediction models and feature selection models in software defects are discussed in Section II. In section III, we proposed a novel ensemble learning model for defect prediction. In Section IV, experimental results are evaluated on distinct software defects datasets and finally, Section V describes about conclusion and future scope.

2. RELATED WORK

Support vector machine is an optimization technique for solving a variety of approaches such as classification, learning and outlier problems. The basic support vector machine (SVM) solves the two class problems, in which the data are partitioned by a hyper-plane using support vectors. If the support vector machine fails to separate two classes, then it solves this problem using a kernel function. The support vectors holds a subset of network data to define the boundary between the two attack classes i.e. anomaly and normal. Various kernel functions can be used in the SVM model such as linear, polynomial, Gaussian, regression etc.

In order to overcome the deficiencies of distance-based methods, [1] proposed that each data point of the given data set should really be assigned a degree of outlier. With their view, for example other recent studies, a data point's measure of anomaly should be measured relative to its neighbors; hence they refer to it just like the "local outlier factor" of the data point. Tang et al. [2-4] argued that any outlier doesn't always have to remain of lower density and lower density is a necessary condition to remain an outlier. They modified LOF to search for the "connectivity-based outlier factor" (COF) which they argued, is so much more effective each time a cluster and a neighboring outlier have similar neighborhood densities. Local density is widely measured in terms of k nearest neighbors; LOF and COF both exploit the properties associated with k nearest neighbors of causing given object in the data set. However, it is possible that any outlier lies in a location between objects given by a sparse as well as a dense cluster.

All these models discuss both the advantages and the disadvantages of artificial neural network techniques when applied to the medical domain. Among the advantages are their ability to model dependencies among attributes by minimizing assumptions about the underlying attributes prior to training. It goes on to credit such algorithms with the ability to reduce the number of false positives and with having greater classification power than regression algorithms. The major disadvantage that were discussed include being a 'black box' which produces result using an underlying model that is very difficult for a human to interpret, that they are often more difficult to use in the field due to the high computational cost of training, and that they are prone to overfitting the training data.

In this way objects ought to be in relation to those from neighboring cells to examine if they're outliers. In statistics, regression analysis makes use of to approximate the relationship between attributes. Linear regression and logistic regression are two common models. An outlier in regression analysis is undoubtedly an observation whose value is removed from the prediction. To detect such outliers, the residuals of the observations are computed depending on a trained model. PCA is a feature extraction and feature selection approach, its main aim find the relevant feature from the large search space. Main drawbacks identified with the optimization process are: Linear relationships between the large number of network features and it doesn't handle dynamic data sets.
III. Proposed Model

Decision tree works supports with both nominal and numerical features. It can be robust towards the noise and inconsistent values. Decision tree follows the top down approach and categorizes the entire trained dataset by partitioning them from top most node to the class node. Each node represents the test attribute of the instance and the nodes related to one of the distinct possible values for that feature attribute. A decision tree can easily transform the given set of instances into meaningful patterns from the top node to the attack class node level by level. Creating decision trees require a predefined training dataset to learn interesting patterns in the data. ID3 is one of the most widely used decision tree, which use the greedy and the recursive top-down approach of decision tree structure. Information gain is commonly used attribute selection measure for each node of the ID3 decision tree model.

To mine decision tree shared by two datasets, we need two input datasets $D_1$ and $D_2$. $D_1$ and $D_2$ are assumed to share an identical set of attributes. For the case that they contain different sets of attributes, the user will need to determine equivalence between attributes of $D_1$ and attributes of $D_2$, and then map the attributes of $D_1$ and $D_2$ to an identical set of attributes using the equivalence relation and eliminate those attributes of $Di$ that have no\ equivalent attributes in $Dj$. A shared decision tree is a decision tree, that can be used to accurately classify data in dataset $D_1$ and accurately classify data in dataset $D_2$. A high quality shared decision tree is a decision tree that has high data distribution similarity, and has high shared tree accuracy in both datasets $D_1$ and $D_2$. Data distribution similarity (DS) captures cross-dataset distribution similarity of a tree (DST).

Algorithm: Improved Discretization method.

**Input:**
- $N$: number of examples
- $A_i$: continuous attributes
- $C_j$: class values in training set
- $\text{Global Threshold value}$

**Output:** Interval borders in $A_i$

**Procedure:**

1. For each continuous attribute $A_i$ in training dataset do
2. Do normalize the attribute within 0-1 range
3. Sorting the values of continuous attribute $A_i$ in ascending order
4. For each class $C_j$ in training dataset do
5. Find the minimum (MinValue) using StdDev attribute value of $A_i$ for $C_j$
6. Find the maximum (Max) attribute value of $A_i$ for $C_j$
7. end for
8. Find the cut points in the continuous attribute values based on the Min and Max values of each class $C_j$

**Best Cut point range measure:**

9. Find the conditional probability P($C/A$) on each cut point and select the cut point with maximum probability value

**Stoping criteria:**

10. If the cut point using the maximum probability value is exist and satisfies the global threshold value then it can be taken as an interval border.
12. end for

In this proposed work, medical dataset is consider to find the best decision rules using improved discretizing approach. In this framework diabetes source dataset is prepared and saved as positive and negative. This dataset is given to data conversion process to find the unrealized datasets. After verification of unrealization numerical discretization is applied using proposed approach. In this approach each attribute in the dataset check whether the attribute is continuous or not. If the attribute is continuous then copy attribute values along with class labels to value_class_list variable. After copying sort the value_class_list in value ascending order. For each pair of attributes applying improved chisquare for attribute selection measure identification. For each class after attributes selection find the minimum value using Standard deviation for each class. Similarly find the maximum attribute value for each class.
Disease type clustering:
Step 1: Initialization k;
Step 2: Assign each tuple to its nearest cluster;
Step 3: For each cluster C in the initial partition do
Select the instance with the highest hub rank among the cluster C;
Set the instance as Center .
End for
Step 4: For each instance in the cluster C assign with the Randomization probability
\[ P(x) = \frac{\text{dist}(x,y)}{\sum (x_i - y_j)^2} \]
Step 5: Group each new instance with the similar p (x) With the following objective function
\[ \max z^* = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{dis}(\text{hub}, C) \]
s.t
\[ \sum_{i=1}^{n} d(x_i) > 0 \]
\[ \sum_{i=1}^{n} x_i > 0 \]
Step 6: Update centers
\[ c_{up}(x_i, y_j) = d(x_i, c_k) + \sum_{j=1}^{m} \sum_{k=1}^{r} d(x_i, c_k).d(y_j, c_k) \]
Step 7: Merge two clusters with highest probability Values in the cluster end for
Step 8: if \[ \sum_{k=1}^{2} d(c_{k-1}, c_k) < \lambda \] then
Stop
Else
Return to step 3.

Improved Decision tree measure:
Modified Information or entropy is given as
\[ \text{ModInfo}(D) = -S_i \sum_{i=1}^{m} \log \sqrt{S_i} , \text{m different classes} \]
\[ \text{ModInfo}(D) = -S_i \sum_{i=1}^{2} \log \sqrt{S_i} = -S_i \log \sqrt{S_1} + S_2 \log \sqrt{S_2} \]
Where \( S_i \) indicates set of samples which belongs to target class ‘anamoly’, \( S_2 \) indicates set of samples which belongs to target class ‘normal’ Information or Entropy to each attribute is calculated using
\[ \text{Info}_A(D) = \sum_{i=1}^{y} |D_i|/|D| \times \text{ModInfo}(D_i) \]
The term \( D_i / |D| \) acts as the weight of the jth partition. \( \text{ModInfo}(D) \) is the expected information required to classify a tuple from D based on the partitioning by a improved c45 algorithm.

Datasets
In our experiment , we investigated four different classification models along with preprocessing techniques.
Sample dataset:
@relation pima_diabetes
@attribute 'preg' real
@attribute 'plas' real
@attribute 'pres' real
@attribute 'skin' real
@attribute 'insu' real
@attribute 'mass' real
@attribute 'pedi' real
@attribute 'age' real
@attribute 'class' { tested_negative, tested_positive}
@data
6,148,72,35,0,33.6,0,627,50,tested_positive
1,85,66,29,0,26.6,0,351,31,tested_negative
8,183,64,0,23.3,0,672,32,tested_positive
1,89,66,23,94,28.1,0,167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0,201,30,tested_negative
3,78,50,32,88,31,0,248,26,tested_positive
10,115,0,0,0,35.3,0,134,29,tested_negative
Sample Decision tree Patterns:

\[
\begin{align*}
\text{pres} & \leq 122.0 \rightarrow \text{insu} \leq 846.0 \\
\text{preg} & \leq 17.0 \rightarrow \text{mass} \leq 67.1 \\
\text{pres} & > 0.0 \rightarrow \text{mass} \leq 67.1 \\
\text{skin} & \leq 99.0 \land \text{pedi} \leq 2.42 \rightarrow \text{mass} \leq 67.1 \\
\text{plas} & > 0.0 \land \text{age} \leq 81.0 \rightarrow \text{pedi} \leq 2.42 \\
\text{pres} & > 0.0 \land \text{insu} \leq 846.0 \rightarrow \text{plas} > 0.0 \\
\text{pedi} & \leq 2.42 \rightarrow \text{insu} \leq 846.0 \\
\text{pedi} & \leq 2.42 \land \text{plas} \geq 0.0 \rightarrow \text{pres} \leq 122.0 \\
\text{preg} & \leq 17.0 \land \text{age} \leq 81.0 \rightarrow \text{class} ! = \text{tested_positive} \\
\text{pedi} & \leq 2.42 \land \text{pres} \geq 0.0 \rightarrow \text{mass} \leq 67.1 \\
\text{skin} & \leq 99.0 \land \text{age} \leq 81.0 \rightarrow \text{mass} \leq 67.1 \\
\text{skin} & \leq 99.0 \land \text{class} ! = \text{tested_positive} \rightarrow \text{mass} \leq 67.1 \\
\text{pedi} & \leq 2.42 \rightarrow \text{pres} \geq 0.0 \\
\text{age} & \leq 81.0 \land \text{pedi} \leq 2.42 \rightarrow \text{mass} \leq 67.1 \\
\end{align*}
\]

Number of Iterations :49

F-Measure : 0.96983
Recall : 0.93767
TP rate : 0.98886
FP rate : 0.01114
Classification Accuracy 0.94836

Performance Analysis:

Table 1: Accuracy Performance of the proposed and traditional models

<table>
<thead>
<tr>
<th>DataSize</th>
<th>Algorithm</th>
<th>TruePositive</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>C4.5</td>
<td>0.87</td>
<td>85.89</td>
</tr>
<tr>
<td>500</td>
<td>SVM</td>
<td>0.867</td>
<td>89.56</td>
</tr>
<tr>
<td>500</td>
<td>CART</td>
<td>0.921</td>
<td>91.78</td>
</tr>
<tr>
<td>500</td>
<td>Proposed</td>
<td>0.954</td>
<td>95.81</td>
</tr>
</tbody>
</table>

V. Conclusion

In this paper, we have proposed an efficient scalable improved decision tree construction algorithm which results in high processing speed and small scale. Because of this reason, it is most suitable for large datasets. Our proposed algorithm has many advantages, but the important thing is that it requires only one pass over the training dataset for the entire construction of decision tree. So it significantly reduces the IO cost.

Diabetes is the most common disease in all age groups, this disease is predicted sign artificial intelligence model in the past decade. Also, early detection of diabetes is an essential role in diabetes because it can unearth hidden knowledge from a vast amount of medical data. This research provides an optimal approach to filter data and to discover hidden patterns using improved decision tree model. Experimental results proved that proposed model has high true positive rate and precision compared to traditional classification models.

References


