A Naive Bayes Data Mining Approach for Classification of Cancer Dataset

Nidhi Dangi¹, Mr. Prashant Ahlawat²

¹Department of CSE, Gurgaon Institute of Technology & Management, Gurgaon, Haryana
²Asst. Professor, Department of CSE, Gurgaon Institute of Technology & Management, Gurgaon, Haryana

Abstract
Decision tree based techniques are being continuously evolved for classification of medical datasets. Different algorithms based on soft computing and hard computing has been developed to apply on medical datasets. Hybrid techniques find their way out to search for the solution of cancer databases. Some are inspired by nature and some are inspired by biological phenomenon. Biological inspired methods are successfully developed and applied for different medical problems. In current research, an intelligent technique Naive based decision tree have been applied and evaluated successfully to classify lung cancer based datasets. A comparison has been made with other techniques to check the effectiveness of the proposed method. Simulation results shows that proposed technique achieved appropriate accuracy to classify cancer data sets which is more as compared to other techniques. TP rate, ROC and Precision is highest for proposed method amongst other method. Hence, proposed algorithm is optimum to classify cancer related datasets.

Introduction
Data mining is the process of digging data for discovering latent patterns which can be translated into valuable information. Data mining usage witnessed unprecedented growth in the last few years, the usefulness of data mining techniques has been realized in Healthcare domain. This realization is in the wake of explosion of complex medical data. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain.

Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on. The data mining techniques that have been applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain.

Association
Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction, which is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer’s buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers and therefore they can put beers and crisps next to each other to save time for customer and increase sales.

Classification
Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into
two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

**Decision Trees**

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. We use the following decision tree to determine whether or not to play tennis.

**Human and Artificial Neurones**

Much is still unknown about how the brain trains itself to process information, so theories abound. In the human brain, a typical neuron collects signals from others through a host of fine structures called dendrites. The neuron sends out spikes of electrical activity through a long, thin stand known as an axon, which splits into thousands of branches. At the end of each branch, a structure called a synapse converts the activity from the axon into electrical effects that inhibit or excite activity in the connected neurones. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

**Components of a neuron**

**The synapse**

**Human Neurones To Artificial Neurones**

We conduct these neural networks by first trying to deduce the essential features of neurones and their interconnections. We then typically program a computer to simulate these features. However because our knowledge of neurones is incomplete and our computing power is limited, our models are necessarily gross idealisations of real networks of neurones.

**The neuron model**
Problem Formulation

Proposed Method

1. Feed Input Data, set all the parameters of NBT like number of neurons, error limit, learning rate etc
2. Train the NBT Network. Attributes are mapped to each node.
3. Supervised Learning through attribute selection feature filtering at each node. Check the error with actual and expected output
4. Evaluate the Testing Parameters. Check the max error rate
5. Record the classification Accuracy, Computation time

Stop

Challenges

1) Tedious and time consuming Mathematical Formulation
2) Memory consumption
3) In previous researches, researchers have used methods which are dependent on each other. Thus, error in any one of the feature result in malfunction of the complete classification system.

Advantages of Proposed Method

- Fast Convergence
- Linearize the non linear data
- Intelligent Decision Making
Objectives

1. To develop a NBT algorithm for the classification of cancer data sets.
2. To evaluate the developed algorithm on cancer data sets.
3. Compare the proposed method with other algorithms.

Other Algorithms

• Bayesian Networks
  A Bayesian network (also referred to as Bayesian belief network, belief network, probabilistic network, or causal network) consists of a qualitative part, encoding existence of probabilistic influences among a domain's variables in a directed graph, and a quantitative part, encoding the joint probability distribution over these variables. Each node of the graph represents a random variable and each arc represents a direct dependence between two variables. The directed graph is a representation of a factorization of the joint probability distribution. As there can be many graphs that are capable of encoding the same joint probability distribution.

• Support Vector Machine
  SVMs are inspired by the Structural Risk Minimization principle from statistical learning theory. In their basic form, SVMs attempt to perform classification by constructing hyper planes in a multidimensional space that separates the cases of different class labels. It backs both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models like linear, polynomial. In the last years, SVMs have been widely investigated and used in a lot of different fields and for various classification tasks, due to their good performances. Learning algorithms such as neural network & SVMs, both trained with different parameters and input features, showed that SVMs produce the most robust results.

• Multilayer Perceptron Tree
  A supervised multilayer perceptron tree (SMLPT) is a trained feed forward neural network model that maps sets of input data onto a set of appropriate outputs. A SMLPT consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. SMLPT utilizes a supervised learning technique called back propagation for training the network through attribute selection feature. SMLPT is a modification of the standard multilayer perceptron and can distinguish data that are not linearly separable.

Results and Discussion

Computation Time For Different Decision Tree Algorithms

1) Naïve Bayes Tree
   Time taken to build model: 0.02 seconds

2) Multilayer Perceptron Tree
   Time taken to build model: 8.16 seconds

3) Support Vector Machine Tree
   Time taken to build model: 0.48 seconds

Simulation Results

Comparative analysis is done for checking the effectiveness of the proposed method. As observed in Table 1, we can see that proposed method NBT is having high classification rate with appropriate accuracy as compared to other decision tree algorithms and less error rate. It achieves high value of testing parameters (TP-True Positive, FP-False Positive, ROC-Region of Curve).

<table>
<thead>
<tr>
<th>Technique</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.44</td>
<td>0.2376</td>
<td>0.4702</td>
<td>57.5927</td>
<td>104.0271</td>
<td>78.125</td>
</tr>
<tr>
<td>Method</td>
<td>Recall</td>
<td>TP Rate</td>
<td>FP Rate</td>
<td>Precision</td>
<td>F-Measure</td>
<td>ROC Area</td>
</tr>
<tr>
<td>------------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>Decision Tree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi Layer Perceptron</td>
<td>0.12</td>
<td>0.3247</td>
<td>0.5302</td>
<td>78.7205</td>
<td>117.2976</td>
<td>65.625</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.12</td>
<td>0.3438</td>
<td>0.5863</td>
<td>83.3333</td>
<td>129.7132</td>
<td>65.625</td>
</tr>
</tbody>
</table>

Table 2: Accuracy by Class with Naïve Bayes Tree (NBT)

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cancerous</td>
<td>0.556</td>
<td>0.556</td>
<td>0.13</td>
<td>0.625</td>
<td>0.588</td>
<td>0.773</td>
</tr>
<tr>
<td>Cancerous</td>
<td>0.87</td>
<td>0.444</td>
<td>0.833</td>
<td>0.851</td>
<td>0.773</td>
<td></td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.781</td>
<td>0.356</td>
<td>0.775</td>
<td>0.777</td>
<td>0.773</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Accuracy by Class with Multilayer Perceptron Tree (MLPT)

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cancerous</td>
<td>0.851</td>
<td>0.624</td>
<td>0.763</td>
<td>0.805</td>
<td>0.691</td>
<td></td>
</tr>
<tr>
<td>Cancerous</td>
<td>0.376</td>
<td>0.149</td>
<td>0.516</td>
<td>0.435</td>
<td>0.691</td>
<td></td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.71</td>
<td>0.483</td>
<td>0.69</td>
<td>0.695</td>
<td>0.691</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Accuracy by Class with Support Vector Machine Tree (SVMT)

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>F-Measure</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cancerous</td>
<td>0.866</td>
<td>0.706</td>
<td>0.744</td>
<td>0.8</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>Cancerous</td>
<td>0.294</td>
<td>0.149</td>
<td>0.481</td>
<td>0.365</td>
<td>0.637</td>
<td></td>
</tr>
</tbody>
</table>
From Table 2., we can see that True positive rate ,precision of solution, F measure and ROC value is highest for proposed technique NBT as compared to other methods as shown in table 2,3,4.Hence,proposed method NBT is effective in testing parameters

Conclusion

The proposed NBT based approach evolved as optimal approach to classify the cancer datasets with a remarkable accuracy of 78.125% and fast computation time of 0.02 seconds as compared to MLP technique and other classification methods. With such high accuracy in proposed method, it will be easy to identify the cancer and non cancer patients from different attributes of people for large data chunks where other decision tree algorithms fail to achieve high accuracy.

References