Performance Comparison of Naïve Bayes and J48 Classification Algorithms

Anshul Goyal and Rajni Mehta, Assistant Professor, CSE

JCDM COE, Kurukshetra University, Kurukshetra
JCDM COE, Kurukshetra University, Kurukshetra
goyal.anshul90@gmail.com, er.rajnimehta@gmail.com

Abstract
Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. This paper has been carried out to make a performance evaluation of Naïve Bayes and j48 classification algorithm. Naïve Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper sets out to make comparative evaluation of classifiers NAÏVE BAYES AND J48 in the context of financial institute dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool.

The experiments results shown in this paper are about classification accuracy and cost analysis. The results in the paper on this dataset also show that the efficiency and accuracy of j48 and Naïve bayes is good.

Keywords: Naïve Bayes, j48, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, ROC Curve.

Data Mining Advancement in technology has brought great growth in the volume of the data available on the internet, digital libraries, news sources and company-wide intranets. This makes a huge number of databases and information repositories available, but it is impossible to manually organize, analyze and retrieve information from this data. This generates an eminent need of methods that can help users to efficiently navigate, summarize, and organize the data so that it can further be used for applications ranging from market analysis, fraud detection and customer retention etc. Therefore, the techniques that perform data analysis and may uncover important data patterns are needed. One of these techniques is known as data mining.

Data mining refers to extracting or mining knowledge from large amounts of data [1][2]. Data Mining is performed on Database-oriented data sets and applications, Relational databases, data warehouses, transactional databases and advanced data sets and advanced applications such as Object-Relational database Temporal Databases, Sequence Databases, and Time-Series Databases, Spatial Databases and Spatiotemporal Databases Text Databases and Multimedia Databases, Heterogeneous Databases and Legacy Databases, Data Streams and the World Wide Web[1].

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, once the software is made that can learn how to classify the data items into groups. For example, we can apply classification in application that given all past records of employees who left the company, predict which current employees are probably to leave in the future. In this case, we divide the employee’s records into two groups that are leave and stay. And then it can be asked from the data mining software to classify the employees into each group [3]. Given a collection of records. Each record contains a set of attributes; one of the attributes is the class. Find a model for class attribute as a function of the values of other attributes. Goal: previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification—A Two-Step Process
Model construction: describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.[1]

The set of tuples used for model construction: training set. The model is represented as classification rules, decision trees, or mathematical formula.

Model usage: for classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur.

Classification Process (1):
Model Construction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Prof</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Assoc</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Assoc</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

If rank = "professor" OR years = 6 THEN tenured = "yes"
### J48

**Classification by Decision Tree Induction**

**Decision tree**
1. A flow-chart-like tree structure
   - Internal node denotes a test on an attribute
   - Branch represents an outcome of the test
   - Leaf nodes represent class labels or class distribution

2. Decision tree generation consists of two phases
   - **Tree construction**
     - At start, all the training examples are at the root
     - Partition examples recursively based on selected attributes
   - **Tree pruning**
     - Identify and remove branches that reflect noise or outliers

3. Use of decision tree: Classifying an unknown sample
   - Test the attribute values of the sample against the decision tree [4][5].

**Algorithm for Decision Tree Induction**
1. Basic algorithm (a greedy algorithm)
   - Tree is constructed in a top-down recursive divide-and-conquer manner
     - At start, all the training examples are at the root
     - Attributes are categorical (if continuous-valued, they are discredited in advance)

Examples are partitioned recursively based on selected attributes.

Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

2. Conditions for stopping partitioning
   - All samples for a given node belong to the same class
   - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
   - There are no samples left [4][5].

**Extracting Classification Rules from Trees**
1. Represent the knowledge in the form of IF-THEN rules
2. One rule is created for each path from the root to a leaf
3. Each attribute-value pair along a path forms a conjunction
4. The leaf node holds the class prediction
5. Rules are easier for humans to understand [6].

**Example**
- If age = “<=30” AND student = “no” THEN buys computer = “no”
- If age = “<=30” AND student = “yes” THEN buys computer = “yes”
- If age = “31…40” THEN buys computer = “yes”
- If age = “>40” AND credit rating = “excellent” THEN buys computer = “yes”
- If age = “>40” AND credit rating = “fair” THEN buys computer = “no”

**Naïve Bayes Classifier (I)**
A simplified assumption: attributes are conditionally independent:
- Greatly reduces the computation cost, only count the class distribution [10].

\[
P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}
\]

**Naïve Bayesian Classifier (II)**
Given a training set, we can compute the probabilities

**Bayes theorem:**
1. \( P(C|X) = P(X|C) \cdot P(C) / P(X) \).
2. \( P(X) \) is constant for all classes.
3. \( P(C) = \text{relative freq of class C samples} \) such that \( p \) is maximum=c Such that \( P(X|C) \cdot P(C) \) is maximum
4. Problem: computing \( P(X|C) \) is unfeasible! [9][10].

**WEKA Tool**
In Weka datasets should be formatted to the ARFF format.
- The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file, the Preprocess panel has facilities for importing data from a database, and for preprocessing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.
Performance Investigation and Results

Experiments are performed on the Bank data set by using Classification algorithm using WEKA tool. The first step is to find the total no. of instances of the given data using both, Naïve bayes and j48 classification algorithm. In the next step, experiment calculates the Classification accuracy and cost analysis.

Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

<table>
<thead>
<tr>
<th></th>
<th>Classified As</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a=Male</td>
</tr>
<tr>
<td>a</td>
<td>87</td>
</tr>
<tr>
<td>b</td>
<td>67</td>
</tr>
</tbody>
</table>

The above shown Table is about Confusion matrix that calculates that actual and predicted classification i.e. Total no. of true positives for class a is 87 and Total no. of False positive for class a is 67. Total no. of true positives for class b is 78 and Total no. of false positive for class b is 68.

Calculations of true positive rate have been done putting the following formula on the confusion matrix.

True Positive Rate = \( \frac{TP}{TP + FN} \)

False positive Rate = \( \frac{FP}{FP + TN} \)

True Positive Rate for Class a = \( \frac{87}{87 + 67} = 0.565 \)
False positive Rate for Class a = \( \frac{68}{68 + 78} = 0.466 \)
True Positive Rate for Class b = \( \frac{78}{78 + 68} = 0.534 \)
False Positive Rate for Class b = \( \frac{67}{67 + 87} = 0.435 \)

Precision = \( \frac{TP}{TP + FP} \)

Precision for Class a = \( \frac{87}{87 + 68} = 0.561 \)
Precision for Class b = \( \frac{78}{78 + 67} = 0.538 \)

F-measure = \( 2 \times \frac{Precision \times Recall}{Precision + Recall} \)

F-measure for Class a = \( 2 \times \frac{0.561 \times 0.565}{0.561 + 0.565} = 0.563 \)
F-measure for Class b = \( 2 \times \frac{0.538 \times 0.534}{0.538 + 0.534} = 0.536 \)

Graphs using j48

Cost analysis of j48 for Class Male

Tables and graphs using j48

Bank Data set have total no. of 300 instances. Gender class has been chosen randomly from bank dataset. When J48 algorithm is applied on the dataset the confusion matrix is generated for class Gender having two possible values i.e. Male and Female.
Tables and graphs using Naïve Bayes

Tables and graph for class Gender

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>Classified As</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>93</td>
<td>61</td>
<td>Male</td>
</tr>
<tr>
<td>b</td>
<td>82</td>
<td>64</td>
<td>Female</td>
</tr>
</tbody>
</table>

The above shown Table is about Confusion matrix that calculates that actual and predicted classification i.e. Total no. of true positives for class a is 93 and Total no. of False positive for class a is 61. Total no. of true positives for class b is 64 and Total no. of false positive for class b is 8.

True Positive Rate=Diagonal element/Sum of relevant row
False positive Rate=Diagonal element/Sum of relevant row

True Positive Rate for class a=93/93+61=.604
False positive Rate for class a=82/82+64=.562
True Positive Rate for class b=64/64+82=.438
False positive Rate for class b=61/61+93=.396

Precision= Diagonal element/Sum of relevant Column
Precision for Class a=93/93+82=.531
Precision for Class b=64/64+61=.512

F-measure=2* Precision* Recall/ Precision+ Recall
F-measure for Class a =2*.531*.604/.531+.604=.565
F-measure for Class b =2*.512*.438/.512+.438=.472

Graphs using Naïve Bayes

The above shown tables and graphs represents Confusion matrix, True positive rate , False Positive Rate, Precision, Recall, ROC Area, F-Measure on the basis of class Gender in which there are two categories Male and Female. There are two graphs for each class .Graph 1st is about sample size and True positive Rate. Graph 2nd is for sample size and cost benefit analysis. These graphs are drawn on the basis of confusion matrix and cost matrix. In this we are using Naïve Bayes classification algorithm.

Conclusion

Both the algorithms are performed on the given Bank data set and their results are Presented in this section

Performance evolution on the basis of Gender

<table>
<thead>
<tr>
<th></th>
<th>Classification Accuracy</th>
<th>Cost Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Naïve Bayes</td>
<td>J48</td>
</tr>
<tr>
<td>Male</td>
<td>48.33%</td>
<td>52.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>J48</td>
</tr>
<tr>
<td>Female</td>
<td>51%</td>
<td>52%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>J48</td>
</tr>
<tr>
<td></td>
<td>147</td>
<td>144</td>
</tr>
</tbody>
</table>

Classification accuracy and cost analysis of class Gender

The proposed method uses bank dataset. The experiments have been performed using the weka tool. Bank data set have been taken from UCI repository having 300 instances and 9 attributes. J48 is a simple classifier technique to make a...
decision tree, efficient result has been taken from bank dataset using weka tool in the experiment. Naive Bayesian classifier also showing good results. The experiments results shown in the study are about classification accuracy and cost analysis. J48 gives more classification accuracy for class gender in bank dataset having two values Male and Female. The result in the study on these datasets also shows that the efficiency and accuracy of j48 and Naive bayes is good. Classification technique of data mining is useful in every domain of our life e.g. universities, Crime, Medical etc.

Future work
Classification is an important technique of data mining and has applications in various fields. In the present study few issues like high dimensionality, Scalability and accuracy are focused but there are still many issues that can be taken into consideration for further research which are as follows:

1. Different algorithms which are not included in WEKA can be tested. Also, experiments with various feature selection techniques should be compared.
2. Classification technique of data mining is useful in every domain of our life e.g. University domain category wise, Medical domain, crime domain, Auto Price, Zoo etc.
3. Confusion matrix given by weka tool can also be removed by changing in implementation of algorithms that are already used in weka tool.

References