Classification Algorithms in Data Mining

Deciding on the classification algorithms
Project Description

Objective

- Data pre-processing and transform the dataset to perform data mining.
- To apply different algorithms to our dataset and predicting the result.
- To evaluate the result obtained between different classifiers on given data set.

Our project is based on the paper “Top 10 algorithms in data mining”. We will be implementing 5 of them with our data set.
What is classification algorithm?

Classification is a method to find that the given data set belongs to which class according to constraints.

Algorithm tries the possible combination of attributes on input data set to predict output data values.

Classification is case of supervised learning.

The following are the algorithms that we will be implementing:
- Decision Trees
  - C5
  - CART
- Naive Bayes
- K-nearest neighbors
- Support Vector Machine
Decision Tree

Notable decision tree algorithm include:
- C.5
- CART

- Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item’s target value.
- Tree models where the target variable can take a finite set of values are called classification trees.
Notable Decision Tree types

- **C5.0**: It constructs classifier on basis of decision tree, to generate data and can use discrete, continuous data.

- It is supervised learning, because data set is labelled with classes.

- To classify data, different species of data is tested using training data.

- C5.0 is more popular because of its simple to read, and can be interpreted by anyone.
Notable Decision Tree types

- **CART**: It stands for Classification & Regression Trees, as it outputs either classification or regression trees using learning trees.

- It predicts continuous or numerous value.

- It uses supervised learning technique to classify or regression trees.

- It is quite fast and affordable too.
### Comparing C5.0 & CART

<table>
<thead>
<tr>
<th><strong>C5.0</strong></th>
<th><strong>CART</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>It uses information gain to classify.</td>
<td>It uses gini impurity for regression.</td>
</tr>
<tr>
<td>In case of over-fitting it uses single – pass pruning process.</td>
<td>It uses cost-complexity pruning.</td>
</tr>
<tr>
<td>It can have 2 or more branches</td>
<td>It has exactly 2 branches.</td>
</tr>
</tbody>
</table>
Once we have the dataset we will have to prepare the data for data mining. We will remove unwanted columns, perform normalization on selected columns and also data binning based on the type of attribute.

Our Dataset

```
> creditCardData <- read.csv(file.choose(), header = TRUE)
> head(creditCardData)
  ID LIMIT_BAL SEX EDUCATION MARRIAGE AGE PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4
1 1 20000 2 2 1 24 2 2 -1 -1 -2 -2 3913 3102 689 0
2 2 12000 2 2 2 26 -1 0 0 0 0 2 2682 1725 2682 3272
3 3 9000 2 2 2 34 0 0 0 0 0 0 29239 14027 13559 14221
4 4 5000 2 2 1 37 0 0 0 0 0 0 46900 48233 49291 28314
5 5 5000 1 2 1 57 -1 0 -1 0 0 0 8617 5670 33835 20940
6 6 5000 1 1 2 37 0 0 0 0 0 0 64400 57069 57608 19394
BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6 default.payment.next.month
1 1 0 0 0 0 0 0 0 0 0 1
2 3455 3261 0 1000 1000 1000 0 2000 1
3 24948 13549 1518 1500 1000 1000 1000 5000 0
4 28939 29847 2000 2019 1200 1100 1069 1000 0
5 19146 19131 2000 36681 10000 9000 689 679 0
6 19619 20024 2500 1815 657 1000 1000 800 0
```

```
> head(creditCardData)
  LIMIT_BAL SEX EDUCATION MARRIAGE AGE PAY_DUE_MONTHS_TOTAL BALANCE_PAY_TOTAL default.payment.next.month
1 0.01010101 2 2 1 30 -2 0.3946213 1
2 0.11111111 2 2 2 30 3 0.3953670 1
3 0.08080808 2 2 2 40 0 0.4069408 0
4 0.04040404 2 2 1 40 0 0.4264339 0
5 0.04040404 1 2 1 60 -2 0.4009969 0
6 0.04040404 1 1 2 40 0 0.4275235 0
```
Implementing Decision tree algorithm C5 for our Dataset

Evaluation on training data (9900 cases):

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Size</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
<td>1917(19.4%)</td>
</tr>
</tbody>
</table>

(a) (b) <-classified as

<table>
<thead>
<tr>
<th>Size</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>7385</td>
<td>282</td>
</tr>
<tr>
<td>1635</td>
<td>598</td>
</tr>
</tbody>
</table>

Attribute usage:

100.00% PAY_DUE_MONTHS_TOTAL
6.99% MARRIAGE
3.96% AGE
3.90% EDUCATION
1.48% BALANCE_PAY_TOTAL.BALANCE_PAY_TOTAL
1.26% SEX
0.33% LIMIT_BAL.LIMIT_BAL

Time: 0.2 secs

We will now calculate the Accuracy and Error rate for the C5 as follows:

Accuracy = (True Positive + True Negative)/All

Error rate = 1 – Accuracy or (False Positive + False Negative)/All

Accuracy = (15091+1133) / 20100 = 0.80965 = 80.72%

Error rate = 1 – 0.8072 = 0.1928 = (606+3270)/20100 = 19.28%
Implementing Decision tree algorithm CART for our Dataset

\[
\text{Accuracy} = \frac{(14932+1362)}{20100} = 0.80965 = 81.06\% \\
\text{Error rate} = 1 - 0.8106 = 0.1894 = \frac{(765+3041)}{20100} = 19.035\%
\]
Naive Bayes classifier

In machine learning, it is a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Bayes' theorem provides a way of calculating the posterior probability, $P(c|\, x)$, from $P(c)$, $P(x)$, and $P(x|\, c)$. Naive Bayes classifier assumes that the effect of the value of a predictor ($x$) on a given class ($c$) is independent of the values of other predictors. This assumption is called class conditional independence.

\[
P(c|\, x) = \frac{P(x|\, c)P(c)}{P(x)}
\]

\[
P(c|X) = P(x_1|\, c) \times P(x_2|\, c) \times \cdots \times P(x_n|\, c) \times P(c)
\]
Implementing Naïve Bayes algorithm to our Dataset

Naïve Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
0 1
0.779798 0.220202

We can note from the above output:

- **Accuracy** = 80.3%
- **Specificity** = 96.4%
- **Sensitivity** = 24%
In pattern recognition, the k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.
Implementing K-nearest neighbors to our Dataset

We can see that by changing the k value we will get different accuracy and we can see that accuracy is good when k=11.

Accuracy = \( \frac{15329+118}{20100} = 76.85\% \)

Error rate = \( 1 - 0.7685 = 0.2315 = \frac{4285+368}{20100} = 23.15\% \)
In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. An SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

Application of SVM are as follows:

- SVMs are helpful in text and hypertext categorization.
- Classification of images can also be performed using SVMs.
- Hand-written characters can be recognized using SVM.
- The SVM algorithm has been widely applied in the biological and other sciences.
Implementing Support vector machine to our Dataset

▶ We will now calculate the Accuracy and Error rate for the SVM as follows:

▶ **Accuracy** = \( \frac{15301 + 861}{20100} = 0.8040 = 80.4\% \)

▶ **Error rate** = \( 1 - 0.804 = 0.196 = \frac{3542 + 396}{20100} = 19.6\% \)
In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Receiver operating characteristic or ROC Curve
For a perfect classifier the ROC curve will go straight up the Y axis and then along the X axis. A classifier with no power will sit on the diagonal, whilst most classifiers fall somewhere in between.

From the plotted ROC curve we can see that performance of the SVM model is pretty good.

Also we find the Area Under Curve which is 73%
Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. The goal of cross validation is to define a dataset to "test" the model in the training phase, in order to limit problems like overfitting.

Other important Model Evaluation Techniques

**Bootstrapping**
- In statistics, bootstrapping can refer to any test or metric that relies on random sampling with replacement.
- Bootstrapping allows assigning measures of accuracy. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

**Cross-validation**
- Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.
- The goal of cross validation is to define a dataset to "test" the model in the training phase, in order to limit problems like overfitting.
Outliers

Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains.

In our dataset we can see that how many of them past their due dates and total payment due distribution with the help of boxplots.

Boxplot for Payment due months

Boxplot for Total Payment Due
Conclusion

- For our dataset we got good prediction accuracy using CART and SVM.
- With the help of ROC curve we evaluated how good is our model is.
- Discussed on other important evaluation techniques and also on Outlier Analysis.
- There are few other very good algorithms like Apriori which is more suitable for dataset with categorical attributes.
- Choosing on the algorithm depends on the dataset and in our project we have shown how to use those in R and choose the best among them.
Thank You