Comparison of Algorithms in Data classification

Jiachen Zhang

Abstract: Classification of binary and multi-class datasets to draw meaningful decisions is the key in today's scientific world. Machine learning algorithms are known to effectively classify complex datasets. This report attempts to study and compare the performance of classification algorithms, including "logistic regression, support vector machine, random forest, linear discriminant analysis, and k-nearest neighbors, and naive Bayes" to Blood Transfusion Service Center data sets. "Random forest" algorithm is found to give the best classification accuracy.

Keywords: Machine learning, Classification

1. Reading the Data

Read the data. The 'Made Donation in March 2007' is the variable to be predicted, so it is transformed to factor.

```r
data_all <- read.table("D:/data/transfusion.data", sep=",", header = TRUE)
colnames(data_all) <- c("Months_since_Last_Donation", "Number_of_Donations",
                        "Total_Volume_Donated_(c.c.)", "Months_since_First_Donation",
                        "Made_Donation_in_March_2007")
data_all$‘Made_Donation_in_March_2007‘ <- factor(data_all$‘Made_Donation_in_March_2007‘)
```

2. Data Exploration

2.1 Summary Statistics

Variables are seperated into two categories, one is 'Made Donation in March 2007', which is a categorical variable, and the other category include the other variabls, which are continuous variables.

The frequency table of 'Made Donation in March 2007' is as follows.

Table 1: Frequency of persons who made donation

Made donation	Frequency
0	570
1	178

In the sample, there are 570 persons who did not made donation in March 2007, which is nearly four times as many as persons who made donation.

Then caculate the summary statistics of other continuous variables. The summary statistics table also shows that there is no NA in the data.

Table 2: Summary statistics of the data

	n	mean	sd	median	min	max	range
Months since Last Donation	748	9.5	8.1	7	0	74	74
Number of Donations	748	5.5	5.8	4	1	50	49
Total Volume Donated (c.c.)	748	1378.7	1459.8	1000	250	12500	12250
Months since First Donation	748	34.3	24.4	28	2	98	96

2.2 Boxplot

Draw boxplot of each variable to show the relationship between the predicted variable and other variables.

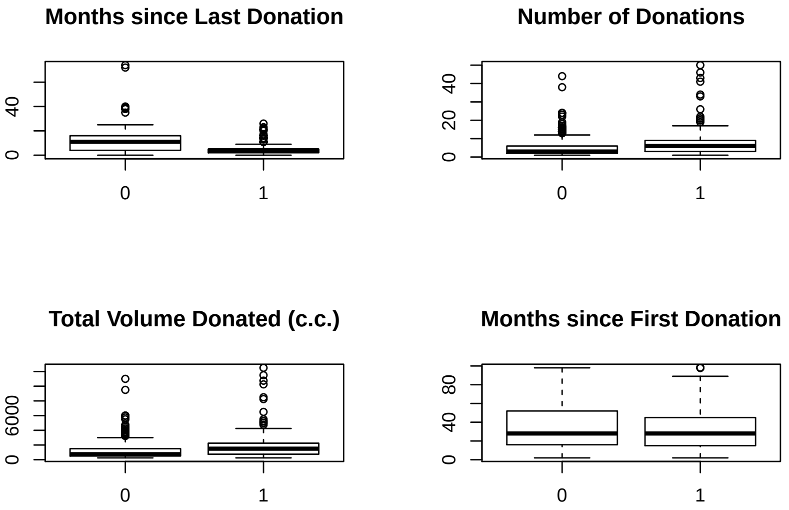

Figure 1: Boxplot

From the above four boxplot, we can see the difference between different 'Donation' values of the median of 'Months since Last Donation', 'Number of Donations', 'Total Volume Donated (c.c.)' is more obvious. Persons who made donation in March 2007 tend to donate more times, and have more total volumn donated, and less months elapsed since last donation for them.

2.3 Correlation Plot

The Correlation plot between the continuous variables is as follows.

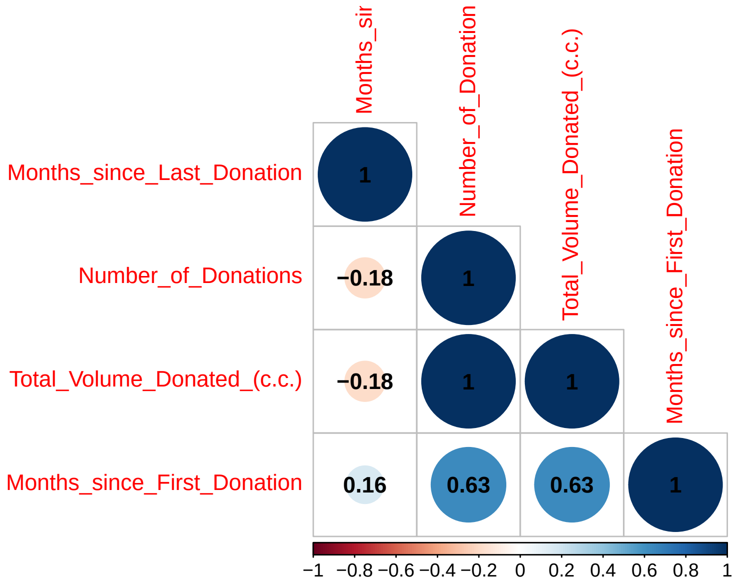

Figure 2: Correlation plot

Since Total Volume Donated (c.c.) have the very high correlation with other variables, so we are dropping the variable.

```
data <- data_all[,-3]
```

3.Model Building

Step1: Training and Testing Data Separation

Separate the data into training data and testing data. Here 20% of data are taken as testing data.

```
set.seed(1234)
library(caret)
train_index = createDataPartition(data_all$`Made_Donation_in_March_2007`, times = 1, p = 4/5, list = FA
don_train <- data[train_index[,1],]
don_test <- data[-train_index[,1],]
```

Step2: Evaluation Parameters

Some evaluation parameters in data mining are accuracy, precision, recall, and F measure. Where TP -True Positive, TN - True Negative, FP - False Positive and FN - False Negative. Here positive represents the variable 'made donation in March 2007' equals to 1, because we generally assume it is more important to recognize the ones who made donation, and the positive often represents the minority group of 0 and 1. The four parameters are defined as follows.

- Accuracy is defined as the number of accurately classified instance divided by a total number of instance in the dataset as in (1).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \ldots (1)$$

- Precision is the average probability of relevant retrieval as described in (2).

$$Precision = \frac{TP}{TP + FP} \ldots (2)$$

- The recall is defined as the average probability of complete retrieval as defined in (3).

$$Recall = \frac{TP}{TP + FN} \ldots (3)$$

- F- Measure is the calculated by using both precision and recall as shown in (4).

$$FMeasure = \frac{2 * (\text{ Precision } * \text{ Recall })}{\text{Precision*Recall}} \ldots (4)$$

Step3: Classification Algorithms

Then we use seven classification algorithms to predict the variable "Made Donation in March 2007". The algorithms are logistic regression, support vector machine, random forest, linear discriminant analysis, k-nearest neighbors, naive Bayes. The formula is as follows.

```
# create a formu_donla
formu_don <- Made_Donation_in_March_2007 ~
  Months_since_Last_Donation + Number_of_Donations +
  Months_since_First_Donation
```

In the process of every algorithm, the following steps are followed:

(1)Fit the model with the train set;

(2)Predict with the test set;

(3)Compute accuracy, precision, recall, F1;

(4)Compute confusing matrix.

Due to limited space, only the code of classification is not showed.

4. Comparison of Classification Algorithms

Finally, compare confusing matrix, accuracy, precision, recall, F1, ROC curves of the seven algorithms.

4.1 Confusion Matrix

Table 3: Confusion matrix

	Reference					
Prediction	logistic regression		support vector machine		random forest	
	0	1	0	1	0	1
0	111	32	112	31	101	24
1	3	3	2	4	13	11

	Reference							
Prediction	linear discriminant analysis		k-nearest neighbors		naive Bayes		decision tree	
	0	1	0	1	0	1	0	1
0	111	33	110	27	109	32	107	22
1	3	2	4	8	5	3	7	13

From the above confusing matrix, we can see that for all seven algorithms, the prediction of 0 class is better than 1 class. Random forest and decision tree perform better on prediction of 1 class.

4.2 Draw ROC Curves

Draw ROC curves of the seven algorithms.

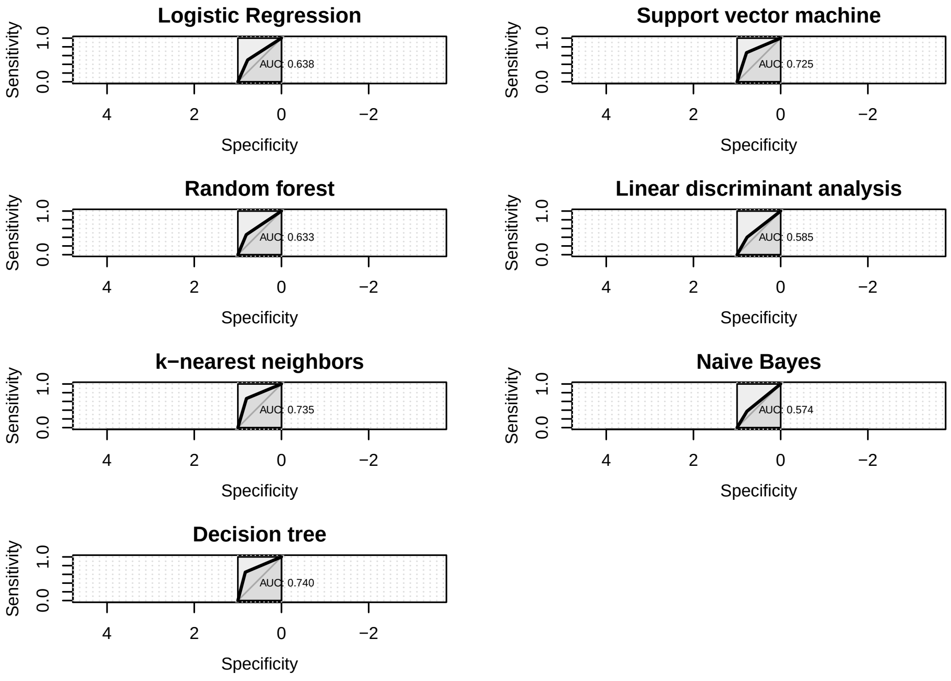

Figure 3: ROC curve of seven algorithms

We can see the AUC of seven algorithms are all above 0.5, the auc of decision tree is the highest, which is 0.740, and the auc of linear discriminant analysis, naive Bayes are blow 0.6.

4.3 Compare by Evaluation Parameters

The accuracy, precision, recall and F1 of the seven algorithms are shown in the table below.

Table 5: Comparison of classification algorithms

	Accuracy	Precision	Recall	F1
logistic regression	0.76510	0.50000	0.08571	0.14634
support vector machine	0.77852	0.66667	0.11429	0.19512
random forest	0.75168	0.45833	0.31429	0.37288
linear discriminant analysis	0.75839	0.40000	0.05714	0.10000
k-nearest neighbors	0.79195	0.66667	0.22857	0.34043
naive Bayes	0.75168	0.37500	0.08571	0.13953
decision tree	0.80537	0.65000	0.37143	0.47273

According to the confusing matrix, the TP is low compared with TN. So the precision and recall of the seven algorithms is not high.

Among the seven algorithms, Decision tree has the highest accuracy, recall, F1, AUC, and the precision is close to the precision of support vector machine and k-nearest neighbors. Thus, we can conclude that in this case, decision tree is the best algorithm.

Reference

[1] Ul Hassan, C. A., Khan, M. S. and Shah, M. A. (2018) ‘Comparison of Machine Learning Algorithms in Data classification’, 2018 24th International Conference on Automation and Computing (ICAC), Automation and Computing (ICAC), 2018 24th International Conference on, pp. 1–6. doi: 10.23919/IConAC.2018.8748995.

[2] Aasim, O. (2019, October 11). Machine Learning Project 17 — Compare Classification Algorithms. Medium. https://towardsdatascience.com/machine-learning-project-17-compare-classification-algorithms-87cb50e1cb60

[3] Blog, G. (2020, June 26). Best way to learn kNN Algorithm using R Programming. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/

[4] Rungta, K. (2021, February 26). Decision Tree in R | Classification Tree & Code in R with Example. Guru99. https://www.guru99.com/r-decision-trees.html

[5] S. (2020, September 20). Blood Donation Analysis. Kaggle. https://www.kaggle.com/shivan118/blood-donation-analysis