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. This homework attempts to study and compare the performance of classification algorithms, including *logistic regression, support vector machine, random forest, linear discriminant analysis, k-nearest neighbors, naive Bayes and decision tree* with Blood Transfusion Service Center dataset. The performances of algorithms change when variables used to predict are removed or kept. Overall, k-nearest neighbors, random forest and decision tree perform better than other algorithms.

Keywords: classification; evaluation parameters; comparison

(The main body is about five pages)

1 Read the Data

Read the data. The 'Made Donation in March 2007' is the variable to be predicted, so it is transformed to factor class. Note that there is no NA in the data.

2 Data Exploration

2.1 Summary Statistics

Variables are seperated into two categories. One category includes 'Made Donation in March 2007'. It is a categorical variable, and is the variable to be predicted. The other category includes the other four variables. They continuous variables, and are used to predicted 'Made Donation in March 2007'.

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 by group. Here Group 0 includes persons who did not made donation in March, and Group 1 made donation.

Mean of 'Months since Last Donation' for Group 0 is 10.8, while for Group 1 is 5.5; mean of 'Number of Donations' for Group 0 is 4.8, while for Group 1 is 7.8; mean of 'Total Volume Donated (c.c.)' for Group 0 is 1200.4, while for Group 1 is 1949.4. These three variables seem to vary between the different groups.Mean of 'Months since First Donation' for Group 0 is 34.8, and for Group 1 is 32.7, they are quite close.

Other statistics, such as median, standard deviation, for the two groups can also be compared in the similar way. The conclusion from the summary statistics table is consistent with the following boxplot.

Table 2: Summary statistics of Group 0 and 1

	vars	n	mean	sd	median
Group 0					
Months since Last Donation	1	570	10.8	8.4	11
Number of Donations	2	570	4.8	4.8	3
Total Volume Donated (c.c.)	3	570	1200.4	1186.7	750
Months since First Donation	4	570	34.8	24.6	28
Group 1					
Months since Last Donation1	1	178	5.5	5.2	4
Number of Donations1	2	178	7.8	8.0	6
Total Volume Donated (c.c.)1	3	178	1949.4	2009.2	1500
Months since First Donation1	4	178	32.7	23.6	28

2.2 Boxplot

Draw boxplots of the four continuous variable grouped by 'Made Donation in March 2007' respectively, to show the difference of sample between Group 0 and Group 1.

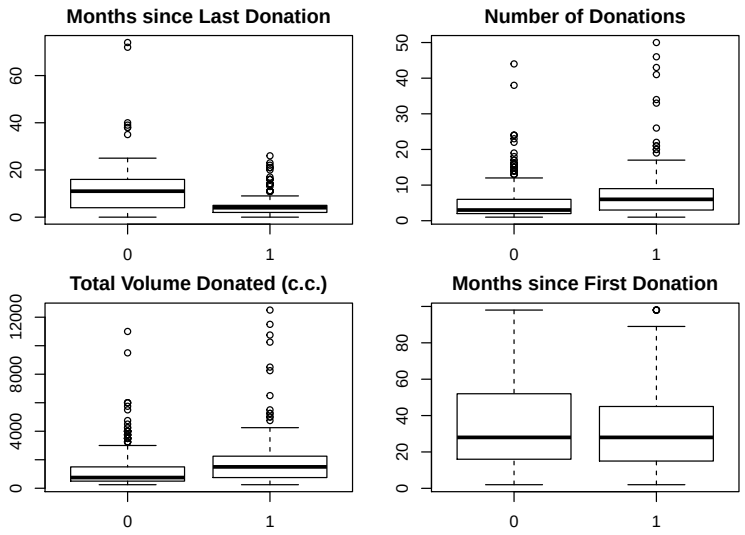

Figure 1: Boxplot

From the above four boxplot, we can see the differences between two groups of the median of 'Months since Last Donation', 'Number of Donations', 'Total Volume Donated (c.c.)' are more obvious. The time inverval between the last donation and the March 2007 donation tends to be shorter for persons who made donation in March 2007, and they tend to donate more times, and have more total volumn donated.

However, the quantiles of 'Months since First Donation' look close from the boxplot, which is consistent with the conclusion from the summary statistics table.

2.3 Correlation Plot

The correlation plot between the four continuous variables is as follows.

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

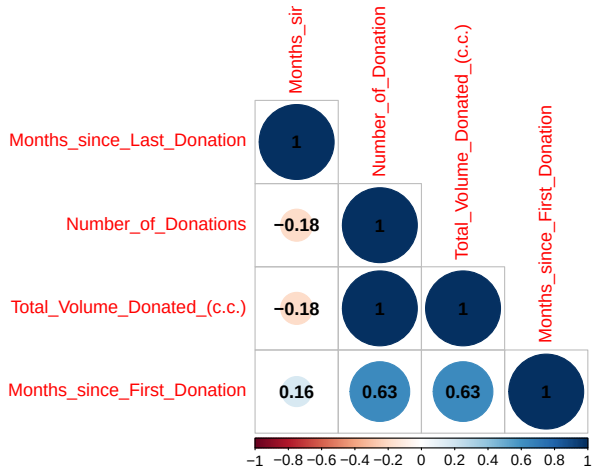

Figure 2: Correlation plot

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

3 Model Building Preparations

3.1 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(8); library(caret)
train_index <- createDataPartition(data_all$`Made_Donation_in_March_2007`, times = 1, p = 4/5, list = FALSE)
```

3.2 Evaluation Parameters

Some evaluation parameters in data mining are accuracy, precision, recall, and F measure. Here TP - True Positive, TN - True Negative, FP - False Positive and FN - False Negative. '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 evaluation 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.$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$
- Precision is the average probability of relevant retrieval.$Precision = \frac{TP}{TP+FP}$
- The recall is defined as the average probability of complete retrieval.$Recall = \frac{TP}{TP+FN}$
- F- Measure is the calculated by using both precision and recall.$F\ Measure = \frac{2*(\text{ Precision } * \text{ Recall })}{\text{Precision*Recall}}$

4 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 and decision tree.*

4.1 Use 'Months since Last Donation', 'Number of Donations', 'Months since First Donation' to Classify

4.1.1 Formula and Steps of Classification Algorithms

The formula is defined as follows.

```
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 confusing matrix, accuracy, precision, recall, F1 and auc.

4.1.2 Comparison of Classification Algorithms

Finally, compare confusing matrix and evaluation parameters (accuracy, precision, recall, F1, auc) of the seven algorithms, to decide which algorithm has the best performance.

4.1.2.1 Confusion Matrix

Table 3: Confusion matrix

	Reference													
Prediction	logistic		svm		rf		lda		knn		naiveBayes		deci tree	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	110	33	110	32	102	28	110	34	108	26	109	32	108	27
1	4	2	4	3	12	7	4	1	6	9	5	3	6	8

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

4.1.2.2 Compare by Evaluation Parameters

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

Table 4: Comparison of classification algorithms (use three variables to classify)

	Accuracy	Precision	Recall	F1	AUC
logistic regression	0.75168	0.33333	0.05714	0.09756	0.55128
support vector machine	0.75839	0.42857	0.08571	0.14286	0.60161
random forest	0.73154	0.36842	0.20000	0.25926	0.57652
linear discriminant analysis	0.74497	0.20000	0.02857	0.05000	0.48194
k-nearest neighbors	0.78523	0.60000	0.25714	0.36000	0.70299
naive Bayes	0.75168	0.37500	0.08571	0.13953	0.57402
decision tree	0.77852	0.57143	0.22857	0.32653	0.68571

The value of TP is low compared with TN, which can also been intui consistent with the confusing matrix. Because of the low TP, the precision and recall of the seven algorithms is not high overall.Except for linear discriminant analysis, the AUC of all algorithms are above 0.5. The auc of k-nearest neighbors is the highest, which is 0.703.

Among the seven algorithms, k-nearest neighbors has the highest accuracy(78.523%), precision(60.000%), recall(25.714%), F1(0.36000), AUC(0.70299). Thus, we can conclude that k-nearest neighbors is the best algorithm in this case.

However, it is worth mentioning that k-nearest neighbors is not always the best algorithm with this dataset. When the random seed is changed, the evaluation parameters of all algorithms change, and the rank of the algorithms will also be slightly different. Support vector machine, random forest, k-nearest neighbors, naive Bayes, and decision tree are all likely to be the best algorithm, depending on the concrete partition of train and test dataset.

4.2 Use Two Variables to Classify

Before running the code of classification with two variables, my prediction is as follows:

- When 'Months since First Donation' is dropped, the evaluation parameters of most algorithms may increase, because 'Months since First Donation' seems to be an interfere factor;
- When 'Months since Last Donation' or 'Number of Donations' is dropped, the evaluation parameters of most algorithms may decrease, because the value of these two variables vary obviously between two groups.

Overall, the result shown below is consistent with the prediction. 'Months since First Donation' and 'Months since Last Donation' are removed respectively, and the evaluation parameters (accuracy, precision, recall, F1 and auc) are calculated. Because the result of removing 'Number of Donations' is similar to removing 'Months since Last Donation', it is not shown in this homework due to the space.

4.2.1 Remove 'Months since First Donation' (irrelevant variable)

Table 5: Comparison of classification algorithms (remove 'Months since First Donation'(irrelevant variable))

	Accuracy	Precision	Recall	F1	AUC
logistic regression	0.76510	0.50000	0.05714	0.10256	0.63621
support vector machine	0.76510	0.50000	0.02857	0.05405	0.63435
random forest	0.77181	0.52381	0.31429	0.39286	0.66815
linear discriminant analysis	0.76510	0.50000	0.05714	0.10256	0.63621
k-nearest neighbors	0.75168	0.33333	0.05714	0.09756	0.55128
naive Bayes	0.75168	0.37500	0.08571	0.13953	0.57402
decision tree	0.77181	0.55556	0.14286	0.22727	0.67063

The evaluation parameters of logistic regression (all parameters), support vector machine (except for recall), random forest (all parameters), linear discriminant analysis (all parameters) increase, while k-nearest neighbors (all parameters) decrease and decision tree (all parameters) derease, naive Bayes (all parameters) stay the same. In this case, random forest and decision tree become the best algorithms.

Besides, it shows that dropping a seemingly 'useless' variable make the result of classification more precise most of the time, but it also depends on the specific algorithm.Note that the change depends on the partition of train and test dataset.

(It is easier to compare the parameters by the combination of the result tables in appendix.)

Table 6: Comparison of classification algorithms (remove 'Months since Last Donation'(relevant variable))

	Accuracy	Precision	Recall	F1	AUC
logistic regression	0.75168	0.25000	0.02857	0.05128	0.50776
support vector machine	0.75168	0.25000	0.02857	0.05128	0.50776
random forest	0.75168	0.45000	0.25714	0.32727	0.62422
linear discriminant analysis	0.75168	0.25000	0.02857	0.05128	0.50776
k-nearest neighbors	0.74497	0.36364	0.11429	0.17391	0.56950
naive Bayes	0.73154	0.22222	0.05714	0.09091	0.49325
decision tree	0.77852	0.66667	0.11429	0.19512	0.72494

4.2.2 Remove 'Months since Last Donation' (relevant variable)

Note the equal value of some parameters in the above table is just a coincidence, because of the small sample size of the test dataset. Most parameters stay the same or decrease. Only some parameters of k-nearest neighbors and decision tree have a slight increase. In this case, decision tree become the best algorithm. Besides, this indicates that when a 'useful' variable is removed, the result of classification algorithms will be less precise most of the time.

Again, note that the change depends on the partition of train and test dataset.

(It is easier to compare the parameters by the combination of the result tables in appendix.)

5 Conclusions and Shortcomings

The necessity of extracting the valuable information from raw data has arisen in many fields of life like medical area, business areas etc. In this homework, the comparison analysis of classifiers for the prediction of blood transfusion is performed. Status of donation is predicted using seven different classifiers. Experimental result shows that different classifiers behave differently on the same dataset.

From the analysis, we observed that out of seven classifiers, k-nearest neighbors performed best when using three variables to classify, and decision tree and random forest are better when using two variables. Besides, the comparison between the result using three and two variables indicates that we can increase the preciseness by dropping the irrelevant variable and only keep the highly correlated variables.

However, due to the small sample size of the dataset, there might be some coincidence, and the result of comparison may change when the algorithms are performed on other datasets. Also, the small number of variables also leads to the overall bad performance of the classification algorithms.

6 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

7 Appendix

In order to compare the results, the combination of the three result tables is shown below.

Table 7: Comparison of classification algorithms

	Accuracy	Precision	Recall	F1	AUC
Use three variables to classify					
logistic regression	0.75168	0.33333	0.05714	0.09756	0.55128
support vector machine	0.75839	0.42857	0.08571	0.14286	0.60161
random forest	0.73154	0.36842	0.20000	0.25926	0.57652
linear discriminant analysis	0.74497	0.20000	0.02857	0.05000	0.48194
k-nearest neighbors	0.78523	0.60000	0.25714	0.36000	0.70299
naive Bayes	0.75168	0.37500	0.08571	0.13953	0.57402
decision tree	0.77852	0.57143	0.22857	0.32653	0.68571
Remove 'Months since First Donation'(irrelevant variable)					
logistic regression	0.76510	0.50000	0.05714	0.10256	0.63621
support vector machine	0.76510	0.50000	0.02857	0.05405	0.63435
random forest	0.77181	0.52381	0.31429	0.39286	0.66815
linear discriminant analysis	0.76510	0.50000	0.05714	0.10256	0.63621
k-nearest neighbors	0.75168	0.33333	0.05714	0.09756	0.55128
naive Bayes	0.75168	0.37500	0.08571	0.13953	0.57402
decision tree	0.77181	0.55556	0.14286	0.22727	0.67063
Remove 'Months since Last Donation'(relevant variable)					
logistic regression	0.75168	0.25000	0.02857	0.05128	0.50776
support vector machine	0.75168	0.25000	0.02857	0.05128	0.50776
random forest	0.75168	0.45000	0.25714	0.32727	0.62422
linear discriminant analysis	0.75168	0.25000	0.02857	0.05128	0.50776
k-nearest neighbors	0.74497	0.36364	0.11429	0.17391	0.56950
naive Bayes	0.73154	0.22222	0.05714	0.09091	0.49325
decision tree	0.77852	0.66667	0.11429	0.19512	0.72494