CSI_5155_Homework_2

Student Name: Lingfeng Zhang
Student Number: 300134245
In this homework, I used scikit-learn to analyze data.

- dataset preprocessing
  - Converting original dataset to pandas.DataFrame format to analyze.
  - Encoding categorical features into numbers by using LabelEncoder
  - For dataset Congressional+Voting+Records and Labor+Relations, missing values occur. Missing value imputation is applied in this dataset. SimpleInputer in scikit-learn is a simple method to impute missing values. Missing values are inputted by the most frequent category.
- Classifier usage and hyper-parameters
  - Decision Tree: sklearn.tree.DecisionTreeClassifier()
  - Rule-based: There is no rule-based classifier in scikit-learn, so Orange.classification.rules.CN2Learner() is applied in this case. The numpy data format must be transferred to Orange.data.Table format in Orange3. However, Orange is so heavy that my Mac run it extremely slow, so I use sklearn.dummy.DummyClassifier() to replace it and generate results. There are two version codes in the attachment.
  - Naive Bayes: sklearn.naive_bayes.GaussianNB()
  - K-Nearest-Neighbors: sklearn.neighbors.KNeighborsClassifier(n_neighbors=3), choosing 3 nearest neighbors.
- Sampling methods
  - Over Sampling: over sampling from minority. Using scikit-learn method -SMOTE.
  - Under Sampling: under sampling from majority. Using scikit-learn method -ClusterCentroids.
  - Balanced Sampling: combing over sampling and under sampling. Using scikit-learn method -SMOTEENN.
- Feature selection methods
  - L-based feature selection
  - Tree-based feature selection

- decision tree sampling

  In decision tree classifier, applying sampling method can improve the overall performance of imbalanced data. Precision and recall increase significantly after using sampling methods, and they are more worthy to analyze than accuracy in imbalanced data. After under-sampling, average accuracy decreases a little, this is possibly because the data size(majority data size) is shrunk and the probability of the model predicting majority corrected is slightly low. Combing over-sampling and under-sampling perform better than only over-sampling or only under-sampling. So, I chosen balanced-sampling method as the best result to analyze the influence of feature selection later.

	without-sampling	over-sampling	under-sampling	balanced-sampling
fold-1	0.85	0.71	0.71	0.78
fold-2	0.64	0.83	0.65	0.79
fold-3	0.76	0.86	0.76	0.95
fold-4	0.86	0.98	0.85	0.97
fold-5	0.87	0.95	0.79	0.97
fold-6	0.89	0.94	0.94	0.96
fold-7	0.86	0.94	0.94	0.95
fold-8	0.91	0.95	0.79	0.97
fold-9	0.86	0.96	0.91	0.97
fold-10	0.92	0.97	0.88	0.98
average	0.84	0.91	0.82	0.93
precision	0.10	0.89	0.82	0.91
recall	0.15	0.94	0.83	0.95

- decision tree feature selection

  In decision tree model, after feature selection, the dimension of features is reduced. For classifier with higher accuracy,precision and recall, applying feature selection does not overall improve the performance of the classifier.

	without-feature-selection	ExtraTree	LinearSVC
accuracy	0.93	0.93	0.91
precision	0.91	0.93	0.90
recall	0.95	0.95	0.94

- Rule-based sampling

	without-sampling	over-sampling	under-sampling	balanced-sampling
accuracy	0.88	0.5	0.46	0.5
precision	0.08	0.5	0.47	0.56
recall	0.07	0.51	0.48	0.55

- Rule-based feature selection

	without-feature-selection	ExtraTree	LinearSVC
accuracy	0.5	0.51	0.51
precision	0.56	0.55	0.55
recall	0.55	0.53	0.55

- Naive-Bayes sampling

	without-sampling	over-sampling	under-sampling	balanced-sampling
accuracy	0.89	0.59	0.62	0.71
precision	0.18	0.57	0.58	0.77
recall	0.2	0.71	0.86	0.65

- Naive-Bayes feature selection

  In Naive Bayes model, the accuracy of after using balanced sampling could not perform as excellent as other models, so after using feature selection, the overall performance is improved. This is possibly because feature selection can enhance the influence of important features and eliminate unnecessary features.

	without-feature-selection	ExtraTree	LinearSVC
accuracy	0.71	0.65	0.80
precision	0.77	0.86	0.86
recall	0.65	0.42	0.75

- K-nearest-neighbors sampling

	without-sampling	over-sampling	under-sampling	balanced-sampling
accuracy	0.91	0.80	0.56	0.96
precision	0.13	0.76	0.56	0.93
recall	0.07	0.89	0.56	0.99

- K-nearest-neighbors feature selection

	without-feature-selection	ExtraTree	LinearSVC
accuracy	0.96	0.83	0.86
precision	0.93	0.81	0.84
recall	0.99	0.90	0.94

- Comparing different models with best results(after balanced sampling)

	Decision-Tree	Rule-based	Naive-Bayes	KNN
accuracy	0.93	0.5	0.71	0.96
precision	0.91	0.56	0.77	0.93
recall	0.95	0.55	0.65	0.99

- Comparing different models with best results(after balanced sampling) with ten-fold

  I used Friedman test to analyze the significant difference of models.

	Decision-Tree	Rule-based	Naive-Bayes	KNN
fold-1	0.78	0.51	0.5	0.83
fold-2	0.79	0.48	0.68	0.94
fold-3	0.95	0.49	0.80	0.93
fold-4	0.97	0.47	0.83	0.98
fold-5	0.97	0.50	0.73	0.99
fold-6	0.96	0.46	0.69	0.98
fold-7	0.95	0.51	0.7	0.97
fold-8	0.97	0.53	0.7	0.98
fold-9	0.97	0.56	0.67	0.97
fold-10	0.98	0.45	0.73	0.98

Ranking these numbers for each fold:

	Decision-Tree	Rule-based	Naive-Bayes	KNN
fold-1	2	3	4	1
fold-2	2	4	3	1
fold-3	1	4	3	2
fold-4	2	4	3	1
fold-5	2	4	3	1
fold-6	2	4	3	1
fold-7	2	4	3	1
fold-8	2	4	3	1
fold-9	1	4	3	2
fold-10	2	4	3	1
average	1.8	3.9	3.1	1.2

average rank $= \frac{k+1}{2} = \frac{4+1}{2} = 2.5$

Friedman statistic $= n \sum_j (R_j - \bar{R})^2 = 45 > critical\ value\ around\ 6.0,\ where\ \alpha = 0.05$

In conclusion, these models have significant difference. So, we have to analyze which pairs are significant different.

- accuracy of different models with different dataset

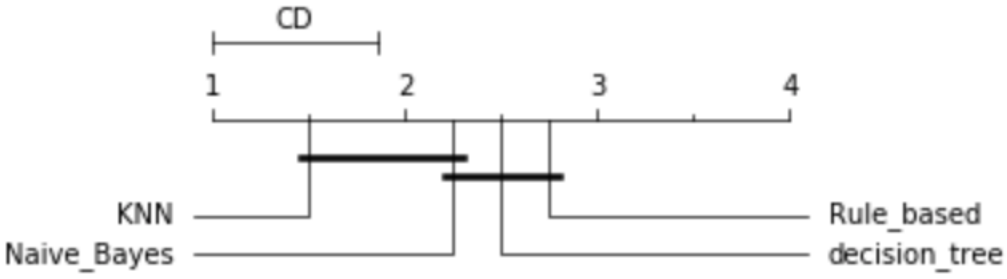
	Decision-Tree	Rule-based	Naive-Bayes	KNN
Seismic Bumps	0.84	0.88	0.89	0.91
Labor+Relations	0.81	0.54	0.82	0.95
Iris	0.95	0.34	0.95	0.97
Congressional+Voting+Records	0.94	0.54	0.92	0.92

Ranking the accuracy of different models for each dataset

	Decision-Tree	Rule-based	Naive-Bayes	KNN
Seismic Bumps	4	3	2	1
Labor+Relations	3	4	2	1
Iris	2	4	3	1
Congressional+Voting+Records	1	4	2	3
average	2.5	2.75	2.25	1.5

	DT-RB	DT-NB	DT-KNN	RB-NB	RB-KNN	NB-KNN
difference	0.25	0.25	1.0	0.5	1.25	0.75

Same way to do Friedman test



We can see rule-based and K-nearest neighbor are most significantly different. In addition, decision tree and K-nearest neighbor are also significantly different.

Conclusion: Different models for different datasets perform differently. Overall, for imbalanced dataset, after using sampling methods can improve the model performance. If the model could not work well after sampling, applying feature selection can improve the model performance in most cases. For testing different models significant difference, we can use Friedman test to analyze whether models are significantly different. If models are significantly different, then we use Nemenyi graph to analyze which pair of models are significantly different.