Summary report

Biliang Wang
bwang135@uottawa.ca
Nov. 3, 2019

Part A

In this experiment, the seismic bumps dataset was rebalanced using three different approaches:oversampling (random oversampling) the minority class, under-sampling (repeated edited nearest neighbors) the majority class and balanced sampling (SMOTE-ENN). Four algorithms were trained against the four datasets using ten-fold cross validation, which were decision tree, k nearest neighbors, naive bayes and rule-based. The results can be summarized below:

Table 1: Model performance after applying sampling methods

		Decision Tree	Nearest Neighbors	Naive Bayes	Rule-based
Original Dataset	precision	0.094	0.181	0.088	0.040
	recall	0.153	0.112	0.900	0.041
	accuracy	0.848	0.908	0.383	0.873
Oversampling Dataset	precision	0.899	0.827	0.594	0.497
	recall	1.000	1.000	0.917	0.496
	accuracy	0.944	0.895	0.645	0.497
Under-sampling Dataset	precision	0.267	0.583	0.118	0.066
	recall	0.376	0.329	0.871	0.071
	accuracy	0.865	0.926	0.460	0.843
Balanced sampling Dataset	precision	0.890	0.916	0.698	0.533
	recall	0.953	0.994	0.923	0.551
	accuracy	0.910	0.947	0.740	0.491

Clearly, compared with original dataset, performances of model were better after the sampling method, though performances of models trained against under-sampling dataset only improved marginally.

Models trained against oversampling and balanced sampling datasets both achieved better result than the original and under-sampling datasets. Simply to avoid over-fitting, we chose balanced sampling dataset for further experiments.

The accuracies of the four algorithms against each one of the ten folds when trained against the balanced sampling dataset can been seen below:

Table 2: accuracies of models against one fold

Fold	Decision Tree	Nearest Neighbors	Naive Bayes	Rule-based
1	0.8308	0.8905	0.8905	0.4826
2	0.8358	0.9104	0.9104	0.5174
3	0.8980	0.9403	0.9403	0.4925
4	0.9303	0.9751	0.9751	0.5124
5	0.9426	0.9626	0.9626	0.5262
6	0.9327	0.9426	0.9426	0.4663
7	0.9551	0.9800	0.9800	0.5037
8	0.9501	0.9476	0.9476	0.5486
9	0.9102	0.9375	0.9375	0.4950
10	0.9450	0.9825	0.9825	0.5100
avg	0.8879	0.9365	0.7169	0.4963
stdev	0.0807	0.0650	0.0557	0.0194

The following table demonstrates the calculation of a pair t-test on the result above.

Table 3: Pair t-test of different models

Fold	DT-KNN	DT-NB	DT-RB	KNN-NB	KNN-RB	NB-RB
1	-0.0597	0.0746	0.3483	0.1343	0.4080	0.2736
2	-0.0746	0.1169	0.3184	0.1915	0.3930	0.2015
3	-0.0423	0.1045	0.4055	0.1468	0.4478	0.3010
4	0.0448	0.0547	0.4179	0.0995	0.4627	0.3632
5	-0.0200	0.2244	0.4165	0.2444	0.4364	0.1920
6	-0.0100	0.1970	0.4663	0.2070	0.4763	0.2693
7	-0.0249	0.2195	0.4514	0.2444	0.4763	0.2319
8	0.0025	0.2444	0.4015	0.2419	0.3990	0.1571
9	-0.0273	0.2302	0.4152	0.2575	0.4425	0.1850
10	-0.0375	0.2675	0.4350	0.3050	0.4725	0.1675
avg	-0.0486	0.1711	0.3916	0.2197	0.4402	0.2206
stdev	0.0157	0.1364	0.0613	0.1207	0.0456	0.0750
p-value	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000

According to the p-value of the four algorithms, there is a statistically significant difference in the accuracies obtained by the four algorithms at the $\alpha = 0.05$ level.

We then reduced the feature size of the balanced dataset to 10 and 8 by applying tree-based and L1-based feature selection techniques.

The following table shows accuracies of four models trained against dataset after feature selection.

Table 4: accuracies of models trained against feature-selected dataset

	Decision Tree			Nearest Neighbors			Naive Bayes			Rule-based		
F	No	Tree	L1	No	Tree	L1	No	Tree	L1	No	Tree	L1
1	0.83	0.85	0.83	0.89	0.89	0.82	0.89	0.67	0.60	0.48	0.48	0.46
2	0.84	0.84	0.82	0.91	0.87	0.83	0.91	0.67	0.66	0.52	0.52	0.51
3	0.90	0.91	0.90	0.94	0.92	0.89	0.94	0.75	0.76	0.49	0.46	0.52
4	0.93	0.97	0.96	0.98	0.97	0.97	0.98	0.80	0.81	0.51	0.49	0.51
5	0.94	0.93	0.94	0.96	0.90	0.91	0.96	0.81	0.77	0.53	0.48	0.48
6	0.93	0.93	0.91	0.94	0.95	0.91	0.94	0.75	0.79	0.47	0.52	0.49
7	0.96	0.92	0.93	0.98	0.94	0.95	0.98	0.76	0.75	0.50	0.51	0.53
8	0.95	0.91	0.90	0.95	0.95	0.91	0.95	0.79	0.74	0.55	0.44	0.53
9	0.91	0.92	0.93	0.94	0.93	0.92	0.94	0.73	0.74	0.50	0.50	0.48
10	0.95	0.95	0.93	0.98	0.97	0.94	0.98	0.77	0.73	0.51	0.54	0.49
avg	0.89	0.90	0.88	0.94	0.93	0.88	0.94	0.72	0.67	0.50	0.51	0.48
diff	0.0%	0.8%	-1.0%	0.0%	-0.1%	-6.1%	0.0%	**-23.6%-28.9%**		0.0%	2.3%	-4.3%

The feature selection techniques do not lead to an overall improvement in accuracies in this dataset, since the differences between accuracies are less than 10%(most of them are less than 5%). On the contrary, the accuracies of Naive Bayes model decreased more than 20%.

Part B

The following table demonstrates the accuracies of the four algorithms against the four datasets.

Table 5: accuracies of four models

Dataset	Decision Tree	Nearest Neighbors	Naive Bayes	Rule-based
Seismic	0.8479(3)	0.9083(1)	0.3827(4)	0.8765(2)
Labor	0.8772(3)	0.9474(1)	0.9298(2)	0.5439(4)
Iris	0.9533(1)	0.9533(1)	0.9533(1)	0.2800(4)
Voting	0.9494(1)	0.9218(3)	0.9448(2)	0.4943(4)
avg rank	2.0	1.5	2.3	3.5

Based on the Friedman test, $\overline{R} = \frac{k+1}{2} = 2.5$, $n\sum_j (R_j - \overline{R})^2 = 9.16$ and $\frac{1}{n(k-1)}\sum_{ij} (R_i j - \overline{R})^2 = 2$, so the Friedman statistic is 4.58. The critical value for $k = 4$ and $n = 4$ at the $\alpha = 0.05$ level is 7.8, so we accept the null hypothesis that all algorithms perform equally, which means that the average ranks as a whole don't display significant different.

Further analysis is carried out on a pairwise level. The critical difference in Nemenyi test is as follows:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}$$

where q_α depends on the significance level α as well as k: for $\alpha = 0.05$ and $k = 4$ it is 2.569, leading to a critical difference of 2.345. No significant difference between four algorithms. The Nemenyi diagram is as follows: