CO395 Machine Learning
CBC Assignment #4B
T-Test

Group 11
Baisheng Song(bs2111) Jiayun Ding(jd1611)
Yiming Lin(yl8411) Yan Liu(ybl11)

December 1, 2014

1 Results

1.1 Clean dataset

	DT vs. ANN	DT vs. CBR	ANN vs. CBR
Emotion 1	3.4297	7.2103	2.4577
Emotion 2	2.2527	4.9130	2.9371
Emotion 3	0.7478	4.2695	2.2951
Emotion 4	4.7052	5.1989	0.5519
Emotion 5	1.1476	4.7703	3.5431
Emotion 6	2.6257	3.3366	1.6148

Table 1: t-values for every emotion and learning algorithm on the *clean* dataset

	DT vs. ANN	DT vs. CBR	ANN vs. CBR
Emotion 1	different	different	similar
Emotion 2	similar	different	different
Emotion 3	similar	different	similar
Emotion 4	different	different	similar
Emotion 5	similar	different	different
Emotion 6	similar	different	similar

Table 2: h-values for every emotion and learning algorithm on the *clean* dataset

1.2 Noisy dataset

	DT vs. ANN	DT vs. CBR	ANN vs. CBR
Emotion 1	6.1220	2.8199	-2.3353
Emotion 2	3.7721	3.9875	0.7858
Emotion 3	2.8850	3.3686	0.0036
Emotion 4	3.9887	4.0040	-0.4604
Emotion 5	-0.3930	1.6809	1.6629
Emotion 6	3.6246	5.4845	1.6254

Table 3: t-values for every emotion and learning algorithm on the *noisy* dataset

	DT vs. ANN	DT vs. CBR	ANN vs. CBR
Emotion 1	different	different	similar
Emotion 2	different	different	similar
Emotion 3	different	different	similar
Emotion 4	different	different	similar
Emotion 5	similar	similar	similar
Emotion 6	different	different	similar

Table 4: h-values for every emotion and learning algorithm on the *noisy* dataset

2 Questions

2.1 Performance

Based on the error rates we calculated from three different learning methods, for the clean dataset, we can gain a conclusion that the CBR performs better than ANN, and ANN performs better than the DT. But from the t-values we get for the clean dataset, we get the conclusion that although CBR performs differently to DT, DT Performs similarly as the ANN at most emotions and ANN performs similarly as CBR at most emotions, this is because for the clean dataset, any method is trained well without errors, therefore predictions will be given at a high CR level.

For noisy dataset, all three algorithms perform similarly, and it is difficult to assess whether any of them has performed significantly better. But ANN and CBR do better than DT. For the case of ANN vs. CBR, according to the t-test, the distributions of errors for both algorithms across all the emotions are statistically similar.

It would not be a general case, in some situations, for example for a linear program, DT can do the job quite well while ANN may have a higher error due to its complex structure. And also for a simple program, CBR may not be able to use its advantage of dividing complex problem into simple problems, therefore, ANN may perform better than it.

2.2 Adjustment of significance level

If applying an ordinary t-test in this situation, the significance value would apply to each comparison, so the chance of incorrectly finding a significant difference would increase with the number of comparisons. As we were performing multiple statistical tests, this increases the chance that a significant result is produced simply by chance. We adjusted our significance level to counter this problem by using Bonferroni Correction. The Bonferroni Correction works by simply dividing significance level (0.05 in our case) by the number of tests (3 in our case) to produce a new significance level 0.05/3 (0.01667) for an individual test.

2.3 Type of T-test

We used the MATLAB built-in function `ttest`, paired t-test. Since the results are not independent between three different algorithms, each of the ten folds is trained using the same training examples and targets and used to predict corresponding classifications on the same test examples. Therefore, it makes sense to pair these predictions of the same data and compare the differences and check whether they are similar or not using T-test.

2.4 Classification error vs. F1 measure

Each fold has a different number of positive and negative examples. The F1 measure is only based on the true positives (TP), false positives (FP) and false negatives (FN), it will not be identically distributed since an algorithm will perform better in folds with more positives. This means the sample error for each fold is no longer an independent and identically distribution. To calculate the variance of the sample errors, the central limit theorem is used under the assumption that the samples are independent and identically distribution, this means that the variance used in the t-test calculation will not be representative, and folds with more positive examples will skew the results. However, in addition to TP, FP and FN, classification error also takes true negatives (TN) into account, so the result will not be skewed on folds with more examples. Thus, the error rate is a more representative measure than F1 measures and hence better suitable for the t-test.

2.5 Trade-off

As the number of folds decreases, the size of training and validation set for each fold increases, the accuracy of the confusion matrix will increase. If we assume that the confusion matrix follows some probability distribution, then increasing the number of examples for each fold is like having more samples from this probability distribution. If we have more samples from a given probability distribution, then we can compute a better approximation of its probability density function. So having more test examples per fold allows us to generate a better approximation of the performance of the algorithm for the unseen data.

The disadvantage of increasing number of examples in each fold is that we are decreasing the number of times we can compute the error rate when using t-test.As the number of examples for the t-test decreases, the degree of freedom decreases, and the threshold for statistical significance for the same significance level increases. In this case, we are decreasing the confidence that it will return meaningful results. For instance, if we used the t-test to access only five examples drawn from Normal Distribution with the exact same mean and variance, the t-test might often fail and claim that they come from distributions with two different means.

If the number of folds increases, the size of training and validation set for each fold will decrease, and the classification error for each fold will decrease. Therefore, the difference between the classification errors for different algorithms will decrease. The t-values will be smaller, so it will be less confidence to decide the hypothesis.

2.6 Adding new emotions

The lazy learning algorithm, Case Based Reasoning is more suitable for incorporating the new classes. It is possible to create new cluster and feed examples to the existing CBR using retain method.

The eager learning algorithms, Decision Trees and Artificial Neural Networks would require a complete re-training from scratch using a dataset including the examples of the new emotions. Especially for ANN, it will require the most computational effort due to the large number of parameters. The requirements to correctly predict the new emotion could reconstruct any number of parameters for the network meaning the entire optimization need to be redone.

3 Error Rates

3.1 Error Rates on the Clean Dataset

3.1.1 Decision Trees

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.1000	0.0891	0.1200	0.1287	0.1200	0.1000	0.1188	0.0800	0.1287	0.0900
Emotion 2	0.1200	0.1683	0.1100	0.1089	0.1000	0.0800	0.0990	0.0600	0.1188	0.1300
Emotion 3	0.0600	0.0693	0.0600	0.0792	0.0400	0.0600	0.0792	0.0700	0.0594	0.0600
Emotion 4	0.0900	0.0693	0.0300	0.0594	0.1000	0.0500	0.0792	0.0500	0.0396	0.0600
Emotion 5	0.1100	0.1089	0.0900	0.0693	0.0700	0.1400	0.1386	0.1400	0.1089	0.0900
Emotion 6	0.0400	0.0891	0.0700	0.0495	0.0700	0.0300	0.0594	0.1200	0.0396	0.1100

Table 5: Error rates for each fold and each emotion in the Decision Trees algorithm on the *clean* dataset

3.1.2 Artificial Neural Networks

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.0800	0.0891	0.0800	0.1287	0.0600	0.0800	0.0792	0.0400	0.0396	0.0500
Emotion 2	0.0900	0.0990	0.0800	0.0990	0.0900	0.0800	0.1188	0.0600	0.0792	0.0400
Emotion 3	0.0500	0.0891	0.0300	0.0990	0.0300	0.0500	0.0495	0.0500	0.0594	0.0700
Emotion 4	0.0300	0.0396	0.0000	0.0198	0.0300	0.0400	0.0297	0.0200	0.0297	0.0100
Emotion 5	0.1100	0.0891	0.1300	0.0792	0.0700	0.1400	0.0792	0.0900	0.0891	0.0500
Emotion 6	0.0400	0.0495	0.0400	0.0297	0.0200	0.0500	0.0396	0.0400	0.0396	0.0600

Table 6: Error rates for each fold and each emotion in the Artificial Neural Networks algorithm on the *clean* dataset

3.1.3 Case Based Reasoning

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.0700	0.0594	0.0200	0.0792	0.0300	0.0500	0.0594	0.0400	0.0297	0.0300
Emotion 2	0.0700	0.0594	0.0600	0.0495	0.0600	0.0600	0.0594	0.0900	0.0495	0.0400
Emotion 3	0.0500	0.0693	0.0400	0.0396	0.0200	0.0400	0.0198	0.0200	0.0396	0.0400
Emotion 4	0.0400	0.0297	0.0100	0.0099	0.0300	0.0300	0.0297	0.0200	0.0099	0.0100
Emotion 5	0.0600	0.0198	0.0700	0.0297	0.0300	0.0700	0.0396	0.0600	0.0990	0.0400
Emotion 6	0.0300	0.0594	0.0200	0.0297	0.0100	0.0700	0.0297	0.0100	0.0099	0.0200

Table 7: Error rates for each fold and each emotion in the Case Based Reasoning algorithm on the *clean* dataset

3.2 Error Rates on the Noisy Dataset

3.2.1 Decision Trees

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.1600	0.1400	0.1700	0.1700	0.1584	0.1400	0.1800	0.1600	0.1500	0.1000
Emotion 2	0.1300	0.1000	0.1300	0.1400	0.1089	0.0900	0.1200	0.1500	0.0800	0.1200
Emotion 3	0.1700	0.1600	0.1100	0.1500	0.1980	0.1700	0.2700	0.2500	0.2000	0.1600
Emotion 4	0.1200	0.0800	0.1200	0.1800	0.0594	0.0800	0.1100	0.1200	0.0800	0.1100
Emotion 5	0.0800	0.1000	0.1200	0.1200	0.1287	0.0800	0.0900	0.1400	0.1000	0.1200
Emotion 6	0.1000	0.1000	0.1100	0.1600	0.1188	0.1000	0.1500	0.1800	0.1300	0.1100

Table 8: Error rates for each fold and each emotion in the Decision Trees algorithm on the *noisy* dataset

3.2.2 Artificial Neural Networks

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.1000	0.1100	0.1100	0.1000	0.0990	0.1100	0.0900	0.0600	0.0500	0.0500
Emotion 2	0.1000	0.0900	0.0800	0.1100	0.0693	0.0400	0.0900	0.0800	0.0900	0.0500
Emotion 3	0.1300	0.1700	0.1400	0.0300	0.0693	0.1500	0.1900	0.1000	0.1500	0.0800
Emotion 4	0.0500	0.0700	0.0900	0.0300	0.0693	0.0500	0.0500	0.0400	0.0600	0.0700
Emotion 5	0.1100	0.1200	0.0900	0.1000	0.0990	0.2000	0.0700	0.1300	0.1100	0.1000
Emotion 6	0.0700	0.0800	0.0700	0.0700	0.0693	0.1100	0.0900	0.0700	0.1400	0.0500

Table 9: Error rates for each fold and each emotion in the Artificial Neural Networks algorithm on the *noisy* dataset

3.2.3 Case Based Reasoning

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Emotion 1	0.1700	0.0800	0.1500	0.1200	0.1485	0.1300	0.1000	0.1100	0.0700	0.1000
Emotion 2	0.0800	0.0800	0.0800	0.1400	0.0396	0.0400	0.0700	0.0700	0.0600	0.0500
Emotion 3	0.1700	0.1100	0.1200	0.0700	0.1386	0.1500	0.1700	0.1000	0.0900	0.0900
Emotion 4	0.0400	0.0500	0.0700	0.0700	0.0693	0.0600	0.0700	0.0500	0.0700	0.0600
Emotion 5	0.0900	0.0500	0.1200	0.0700	0.0990	0.1300	0.0500	0.1000	0.0900	0.1000
Emotion 6	0.0900	0.0300	0.0600	0.0500	0.0594	0.0500	0.0800	0.0700	0.1000	0.0600

Table 10: Error rates for each fold and each emotion in the Case Based Reasoning algorithm on the *noisy* dataset