Assignment 4 Part B: T Test

Group 21: Robert Prust, Naomi Bassett, Alan Vey, Octavian Tuchila

Introduction

In this section we are performing t-tests on the classification error of our algorithms in an attempt to assess whether they are statistically different to each other for each emotion. Our null hypothesis is that they are the same given some reasonable significance level alpha.

Results

DT = Decision Tree
ANN = Artificial Neural Network
CBR = Case Based Reasoning

'+' indicates a t-test performed on the classification error of two algorithms

Table entries are in the format (h, p) where h gives information about the null hypothesis (h = 0 accept, h = 1 reject) and p is the p-value of the t-test

Clean Data	DT + ANN	DT + CBR	ANN + CBR
Anger	(0, .49)	(1, .00)	(1, .00)
Disgust	(0, .79)	(1, .00)	(1, .00)
Fear	(0, .54)	(1, .00)	(1, .00)
Happiness	(1, .00)	(1, .00)	(1, .00)
Sadness	(0, .45)	(1, .00)	(1, .00)
Surprise	(0, .30)	(1, .00)	(1, .00)

Noisy Data	DT + ANN	DT + CBR	ANN + CBR
Anger	(0, .05)	(0, .23)	(1, .02)
Disgust	(0, .65)	(1, .00)	(1, .00)
Fear	(0, .37)	(1, .01)	(1, .00)
Happiness	(0, .05)	(1, .00)	(1, .00)
Sadness	(0, .40)	(1, .01)	(1, .00)
Surprise	(0, .63)	(1, .00)	(1, .00)

Question 1: T-Test Performance Comparisons

The results can be interpreted as follows, always cross referenced with the actual data to determine which of the two are better if a statistical difference exists. It is also important to note that p-values close to the alpha value could indicate incorrect acceptance or rejection of the null hypothesis:

- If an emotion (row) has all ones we should be able to identify the best algorithm for classification by examining the actual data.
- If an emotion (row) has all zeros then all algorithms identify it in an equal manner. However if there are large differences in the p-value we can establish which is probabilistically better.
- Otherwise we have to have a much closer look at the p-values in an attempt to predict which algorithm is best with the highest probability.

Clean Data

- Anger: We notice that there is a statistical difference between DT and CBR and between ANN and CBR however not between DT and ANN with fairly high confidence. By observing the data we notice the classification rate for CBR is significantly lower than for the other two. We can infer that CBR is definitely worse at identifying anger and DT and ANN are statistically equally good.
- Disgust: We notice that there is a statistical difference between DT and CBR and between ANN and CBR however not between DT and ANN with high confidence. By observing the data we notice the classification rate for CBR is significantly lower than for the other two. We can infer that CBR is definitely worse at identifying anger and DT and ANN are statistically very similar.
- Fear: We notice that there is a statistical difference between DT and CBR and between ANN and CBR however not between DT and ANN with fairly high confidence. By observing the data we notice the classification rate for CBR is significantly lower than for the other two. We can infer that CBR is definitely worse at identifying anger and DT and ANN are statistically equally good.
- Happiness: We notice a statistical difference between all algorithms and by examining the data see that ANNs seem to outperform the other algorithms when it comes to classification error.
- Sadness: We notice that there is a statistical difference between DT and CBR and between ANN and CBR however not between DT and ANN with fairly high confidence. By observing the data we notice the classification rate for CBR is significantly lower than for the other two. We can infer that CBR is definitely worse at identifying anger and DT and ANN are statistically equally good.
- Surprise: We notice that there is a statistical difference between DT and CBR and between ANN and CBR however not between DT and ANN with fairly high confidence.

By observing the data we notice the classification rate for CBR is significantly lower than for the other two. We can infer that CBR is definitely worse at identifying anger and DT and ANN are statistically equally good.

We cannot find a statistical difference in the classification error between Decision Trees and Artificial Neural Networks for any emotion other than happiness. In all cases we can say that Case Based Reasoning performs the worst though after cross-referencing with the data. For happiness we can say that with that Artificial Neural Networks are probabilistically the best at classifying it.

Noisy Data

- Anger: We notice that the null hypothesis is accepted for DT and ANN and for DT and CBR but not for ANN and CBR. However the p-value for ANN and CBR is close to our significance level suggesting uncertainty in this result. The same goes for DT and ANN. After cross referencing with the data this leads us to believe that ANN is better at classifying this emotion and DT and CBR are similarly bad however we cannot make this statement with any degree of certainty.

- Disgust: DT and ANN show no statistical difference and the other two comparisons show a significant difference. Our p-values strongly indicate both of these statements to be valid and thus we conclude after cross-referencing with our data that DT and ANN are equally good at classifying this emotion and CBR is the worst.

- Fear: DT and ANN show no statistical difference and the other two comparisons show a significant difference. Our p-values strongly indicate both of these statements to be valid and thus we conclude after cross-referencing with our data that DT and ANN are equally good at classifying this emotion and CBR is the worst.

- Happiness: DT and ANN show no statistical difference and the other two comparisons show a significant difference. Our p-values strongly indicate the second of these statements to be valid however we are not as sure about DT and ANN being the same thus we conclude after cross-referencing with our data that DT and ANN are equally good at classifying this emotion with a slight edge to ANN but CBR is the worst.

- Sadness: DT and ANN show no statistical difference and the other two comparisons show a significant difference. Our p-values strongly indicate both of these statements to be valid and thus we conclude after cross-referencing with our data that DT and ANN are equally good at classifying this emotion and CBR is the worst.

- Surprise: DT and ANN show no statistical difference and the other two comparisons show a significant difference. Our p-values strongly indicate both of these statements to be valid and thus we conclude after cross-referencing with our data that DT and ANN are equally good at classifying this emotion and CBR is the worst.

Similarly to the clean data we notice that Artificial Neural Networks and Decision Trees outperform Case based reasoning by far. When it comes to making a distinction between Artificial Neural Networks and Decision Trees we seem to have a slight indication that Artificial Neural Network are slightly better for some emotions. This is probably because Artificial

Neural Networks can train themselves to compensate for erroneous data where Decision Trees become much denser and are dependant on the way all 6 trees results are combined into one.

Conclusion

We notice that for both clean and noisy data Decision Trees and Artificial Neural Networks seem to perform equally well when it comes to classification. The data however suggests that regardless of noise Case Based Reasoning is worst. We believe this is because we have not used weighted regression and we may have a bug in our implementation. In general we see no reason for the deficiencies in the Case Based Reasoning algorithm.

Question 2: Significance Level

We adjusted the significance level alpha using the Bonferroni correction. We chose a value of 5% for our alpha, meaning we are looking to incorrectly find a significant difference no more than 5% of the time. We divided this value by 3, the number of comparisons we are making to ensure this value is not applying to the individual comparisons, but instead all three together i.e. between Decision Trees and Artificial Neural Networks, between Decision Trees and Case Based Reasoning and between Artificial Neural Networks and Case Based Reasoning. This same principal is applied to each of the six emotions.

Question 3: T-Test Type

We are performing the t-test on the same data for each emotion using different learning algorithms and so our samples are dependant. It is the simple case of taking 3 different measurements on the same data. This is why we used a paired t-test.

The paired t-test will take the difference between the classification error for each fold and perform a one sample t-test to compare the mean difference to 0 (assumption that both are normally distributed). This is repeated for each of the six emotions to give us an indication of whether a statistical difference between the algorithms exists.

Question 4: Classification Error vs F1 Measure

In our case, the most significant difference between the classification error and F1 is that the classification error takes into account the number of true negatives, while the F1 measure does not. As a result, whenever there are more than two possible labels which can be given to an example, the classification error is more accurate in describing the performance of the algorithm as it takes into account the entire set of examples as opposed to F1 which only considers the examples relevant to the current label(true positives, false positives, false negatives) and therefore doesn't take into account the size of the entire set.

If one emotion appears very few times(there are very few examples corresponding to that emotion), then the result of the F1 measure is not very reliable, as noisy data can greatly influence its value. The classification rate is more stable however because it computes the accuracy taking into account all the data, and therefore performs better on each of the labels

in particular. As we perform the t-test using a set of emotions, we want to use a measure which better describes the performance of a model for each emotion taking into account the set as a whole, so the classification error is a better choice.

Question 5: N-Fold Cross Validation

If we would use smaller folds then the structures we construct(decision tress, neural networks, CBRs) are less fitted to our data, which may have a negative impact as the approximation can be quite inaccurate; smaller folds could also be effective however as by using them we can avoid over fitting. An advantage of smaller folds is that we would have more t-test results which we can combine in order to compare the three different structures. On the other hand, if we use larger folds we may have data structures which are over fitted so it could be difficult to compare their performance with a t-test as the models themselves are not accurately described by our data structures. Also, we would have fewer results in order to compare any of the models, which may prevent the t-test from giving an accurate description of the data as the t-test does not work well with small data samples. All in all, t-tests are more accurate on larger data samples, but as the size of the sample grows, the accuracy of the t-test does not increase with it in a linear fashion. Therefore, the best results are obtained using a moderately large fold size and as many folds as possible.

Question 6: Adding New Emotion Classes

In terms of engineering effort, the CBR model will adapt well, as it has a lazy evaluation system and will postpone any processing to the next query. The decision tree model will however need to undergo a significant change as all the existing attributes will now have a different gain. Therefore, the trees for all emotions need to be redone in order to compensate for this; also, we would have to construct trees for the new emotions. The neural networks model will need a significant amount of engineering effort, but its structures will not have to be rebuilt completely, such as the decision tree model's structures. The neural networks for the existing emotions will need to be adapted to the new examples while new neural networks will need to be constructed for the new emotions.