AdaBoost

Tobias Dreher, Dmitriy Zharkov

November 2012

Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering
Norwegian University of Science and Technology

Contents

1 Results

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	87.8	-	87.4
DTC	1	A	100	-	91.3
NBC	5	-	65.5	30.9	86.3
NBC	10	-	59.7	38.8	87.7
NBC	20	-	58.1	49.2	85.5
DTC	5	A	86.4	13.7	91.3
DTC	10	1	32.9	15.3	35.4
DTC	10	2	57.4	5.3	73.1
DTC	10	A	84.5	9.1	91.6
DTC	20	A	84.6	13.1	93.8
NBC + DTC	5	2	65.5 / 57.6	30.9 / 8.4	87.9
NBC + DTC	10	2	59.7 / 56.5	38.8 / 17.6	89.1
NBC + DTC	20	2	58.1 / 57.3	49.2 / 19.5	90.1

Table 1.1: pendigits.tra

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	90.2	-	90.8
DTC	1	A	100	-	98.6
NBC	5	-	75.7	27.8	91.4
NBC	10	-	73.5	35.4	90.5
NBC	20	-	73.1	39.7	90.1
DTC	5	A	97.1	3.6	98.6
DTC	10	1	70.9	0.0	71.3
DTC	10	2	79.8	18.4	83.0
DTC	10	A	97.4	3.4	98.6
DTC	20	A	97.1	3.6	98.9
NBC + DTC	5	2	75.1 / 70.4	27.8 / 26.5	88.8
NBC + DTC	10	2	73.5 / 75.4	35.4 / 25.6	90.3
NBC + DTC	20	2	73.1 / 76.4	39.7 / 42.7	88.7

Table 1.2: nursery.txt

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	80.2	-	79.4
DTC	1	A	100	-	83
NBC	5	-	66.7	15.9	78.1
NBC	10	-	61.7	26	82
NBC	20	-	59.7	36.2	81.5
DTC	5	A	84.4	17.5	83
DTC	10	1	50.4	1.9	50.3
DTC	10	2	73.6	9.1	80.8
DTC	10	A	77.3	27.8	83.9
DTC	20	A	75.2	39.3	84.8
NBC+DTC	5	2	66.7 / 73.6	15.9 / 2.2	82.6
NBC+DTC	10	2	61.7 / 72.3	26 / 13.2	83
NBC+DTC	20	2	59.7 / 72.8	36.2 / 15.5	84.9

Table 1.3: sat.trn

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	86.9	-	89.6
DTC	1	A	99.7	-	94.6
NBC	5	-	79.7	8.8	89.2
NBC	10	-	78.3	11.1	92.4
NBC	20	-	76.9	16.3	91.3
DTC	5	A	89.2	12.9	94.6
DTC	10	1	51.9	1.1	55.4
DTC	10	2	75.1	22.6	88.7
DTC	10	A	89.2	12	94.2
DTC	20	A	89.1	13.9	94.8
NBC+DTC	5	2	79.7 / 76.8	8.8 / 14.7	91.6
NBC+DTC	10	2	78.3 / 76.2	11.1 / 16.7	92
NBC+DTC	20	2	76.9 / 70.5	16.3 / 35.9	93.1

Table 1.4: segment.dat

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	72.9	-	64.1
DTC	1	A	100	-	73.7
NBC	5	-	50.2	26.3	65.7
NBC	10	-	46.5	29	68.2
NBC	20	-	44.3	31.3	69.2
DTC	5	A	82.9	21.1	73.7
DTC	10	1	27.7	10.6	34.8
DTC	10	2	45.6	3.9	50
DTC	10	A	82.2	24.6	75.8
DTC	20	A	80	27.2	77.3
NBC+DTC	5	2	50.2 / 42.7	26.3 / 5.2	70.7
NBC+DTC	10	2	46.5 / 41.8	29 / 6.8	70.7
NBC+DTC	20	2	44.3 / 42.8	31.3 / 10.2	72.7

Table 1.5: vowel-context.data

Classifier	No	Depth	Train Acc avg (%)	Std Deviation	Test Accuracy (%)
NBC	1	-	89.3	-	89.7
DTC	1	A	98.8	-	94.5
NBC	5	-	74.6	18.3	90.2
NBC	10	-	67.7	35.1	92.7
NBC	20	-	63.6	66.2	91.9
DTC	5	A	97.7	2	95.1
DTC	10	1	89.9	0	89.1
DTC	10	2	91.9	4.6	92.7
DTC	10	A	97.3	2.5	95.2
DTC	20	A	97.3	2.6	95.3
NBC+DTC	5	2	74.6 / 90.9	18.3 / 1.5	90.4
NBC+DTC	10	2	67.7 / 91.2	35.1 / 4.9	91.8
NBC+DTC	20	2	63.6 / 92.9	66.2 / 6	93

Table 1.6: page-blocks.txt

2 Evaluation

The first interesting fact is that even thought NBC is a very weak classifier, the results are quite satisfying even without boosting. All the data sets contain multi class problems, for example the data set *vowel-context* (table 1.5) has 11 different labels, so the apriori probability for a class would be 9,1%. One single NBC still classifies 64% of the test instances correctly. However, the improvement with increasing number of NBCs is not very big. In some cases like *pendigits* (table 1.1) there is even a decrease of the test accuracy from 10 to 20 NBCs. The reason for that might be that the weight of a hypothesis is not computed on the basis of the majority of instances but on their weights. In case when only few instances could be classified by a hypothesis the hypothesis still can have a big weight when these instances also have big weights.

DTC has proved to be better than NBC. In data sets shown above DTC is about 3-8% better than NBC. This is reasonable because NBC uses the naive assumption that all of the features are conditionally independent of one another. DTC considers the dependencies and is therefore better. At least, when the decision tree is deep enough. As shown in *pendigits* (table 1.1) and *segment* (table 1.4) the performance of DTC is very poor at depth 1. Similar to NBC, there is only single digit percentage increase after using a higher number of hypotheses.

Not only a low depth of the decision tree can lead to unsufficient results but also a high one. In this case the classifier is overspecialized. This leads to very good results during the training phase but not in the tests. Fortunately, the data sets we use don't suffer from this problem. So, even the infinit depth leads to good results during the tests.

Unfortunately, there is no increase in accuracy after using both classifiers. The reason for that is that after NBC is done classifying, the weights of most instances are very small so that DTC only concentrates on "bad" or inconsistent instances. Additionally, the depth of DTC is only 2 which is not as good as infinit depth.

The difference between train and test accuracy is partly huge. The reason for that is that for the average training accuracy all hypotheses are considered equal, that means that their weights aren't taken into account. So, even if a hypothesis is right for very few instances it still counts the same lika a hypothesis that is right for almost all instances.

With increasing number of hypotheses the average train accuracy decreases and the standard deviation increases. This shows that the later hypotheses are specialized for only few instances and are bad for the majority of the data. However, the weight of these hypotheses should be smaller than of the earlier hypothese. Figure 2.1 shows the distribution of the weights.

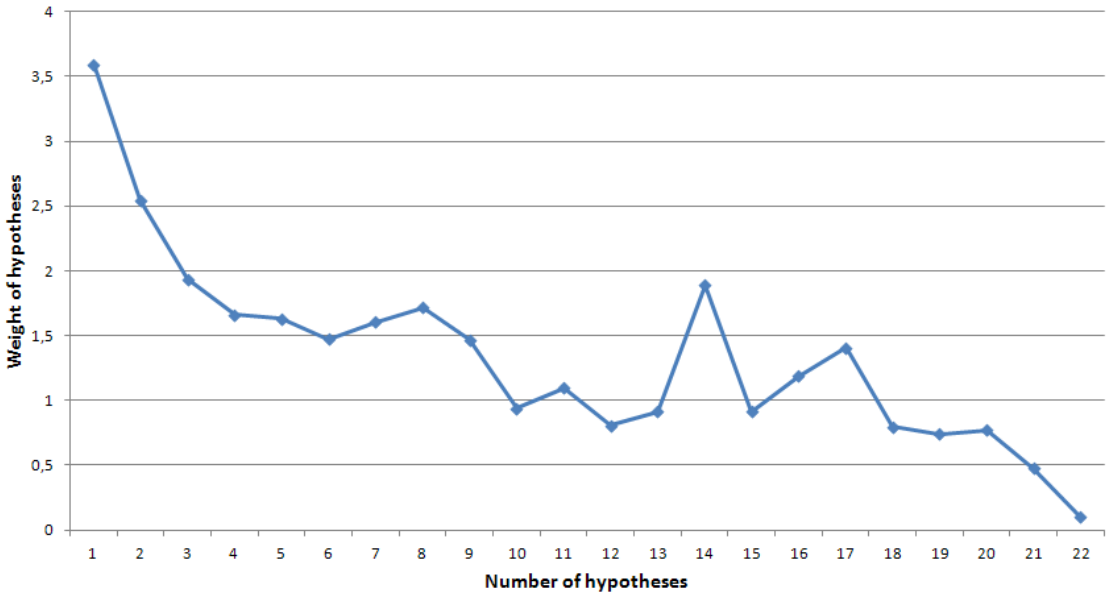

Figure 2.1: Distribution of the hypotheses' weights

AdaBoost has problems with inconsistent data. Early in the project we tried both NBC and DTC with data sets which contained instances with the same values but different labels. After a few rounds of boosting the inconsistend instances gained so much weight that it was impossible for the other instances to "recover". Therefore, after a curtain point, all classifiers concentrated only on the inconsistent data and were not able to produce a proper hypothesis. We think, a detailed preprocessing of the complete data set would improve the results and make AdaBoost more stable.

Also, dividing the real-number-features into bins has a great influence on the results. If there are too many bins then NBC and DTC would show almost 100% accuracy on the training data but would perform poorly on the training set (overspecialization). On the other hand, if there are only a couple of bins then the statements made by the hypotheses are too general.

After all, AdaBoost has shown less improvement than expected. It would be interesting to see if other classifiers like Artificial Neural Networks or Support Vector Machines could improve the performance of AdaBoost. Also, more attetion should be directed to the preprocessing because it can play the crucial role in optimizing the results.