An experimental assessment of decision trees and decision tree ensembles

Student number: 100203952. Blackboard ID: aeb17nnu

1 Introduction

The aims of this project are to evaluate the accuracy of decision trees and ensembles using a set of classification experiments, with the aim of finding if different versions of the ID3 decision tree improves performance for different data and data types.

1.1 Hypotheses

- I expect that decision trees implementing randomness will produce less accurate results than those that do not use randomness.
- I expect that the use of averaging ensemble classifiers will produce more accurate results than the use of independent classifiers.
- I expect that the built in Weka ensembles will produce more accurate data than the TreeEnsemble classifier due to the use of randomness in TreeEnsemble likely reducing accuracy.
- For the FacesUCR case study, I expect that rotation forest will give a more accurate result than all other ensembles and classifiers due to it being proved to be more performant [1].

2 Data Description

For the given experiments, 3 provided data sets were used to complete analysis. A group of discrete data-sets, a group of continuous data-sets, and the FacesUCR case study. The first 2 groups of data-sets were acquired from the UCI Archive and the FacesUCR data-set was taken from timeseriesclassification.com.

A description of data-sets 1 and 2 can be seen in Table 1 and Table 2 respectively. FacesUCR is a case study called FacesAll which is a rotationally aligned version of another case study called FaceAll with a different train/test split. FaceAll is used to map an outline of a persons face onto a one dimensional series. A description of this data is summarised in Table 3.

3 Classifier Description

The classifier ID3Coursework is an implementation of the Weka classifier ID3 which is a decision tree that builds a tree using information gain calculations for the splitting criterion. ID3 only works for nominal value data-sets.

3.1 Design Choices

ID3Coursework is more flexible than ID3 by allowing for a choice of 4 different split measures. These include information gain, Gini, chi squared and chi squared with the yates correction. These split measures are represented as classes that implement the AttributeSplitMeasure class. This is an improvement over base ID3 because the use of different split criteria are better for certain problems.

Another feature that ID3Coursework has that is not present in base ID3 is the ability to use numeric data-sets. This is done by checking whether an attribute is numeric and then if it is, the classifier creates a random value between the maximum and minimum value for that attribute and then performing a binary split on the data based on whether a value falls below of above the split value.

The custom tree ensemble TreeEnsemble uses a collection of default 50 copies of the ID3Coursework classifier where each classifier uses a random split measure to combine results from each measure to hopefully give a more accurate result. The uses a proportion of the attributes from the data randomised for each classifier. This proportion is defaulted to 50%, but this and the number of 50 classifiers can be changed when constructing the TreeEnsemble to allow for tuning.

3.2 Other Classifiers Used

- Weka ID3, the ID3 decision tree based classifier used as the basis for ID3Coursework, but is limited to information gain splitting measure and only assessing nominal attributes.
- J48, another decision tree based classifier that can assess numeric values.
- Random forest, an ensemble that uses RandomTree as a base classifier and the Gini index as a splitting measure. Uses bagging and sub sampling on each node.
- Rotation forest, which uses C4.5 as a base. This ensemble is similar to random forest but combines resampling and random subspace sampling.

4 Results

The results of the following experiments can be found in the appendix.

4.1 Experiment 1

This experiment was done to test the accuracy of ID3, J48 and ID3Coursework using the various splitting measures. The accuracy was checked using the UCI discrete and continuous data. The results can be seen in Tables 4 and 5 respectively. These results show that Id3Coursework produced more accurate results than base ID3 with similar results for every splitting measure, and J48 was more accurate than both.

4.2 Experiment 2

Experiment 2 is similar to the first however the comparison is between tuning the ID3Coursework classifier or the use of the TreeEnsemble with 50% or 100% proportion of attributes. Tables 4 and 5 show that tuning ID3Coursework has little effect on its accuracy measure, and Table 6 shows that TreeEnsemble does not calculate a significantly different accuracy when using a proportion of 50% or 100% but does produce better results than the independent ID3Coursework classifier.

4.3 Experiment 3

Experiment 3 was carried out to assess the accuracy of TreeEnsemble compared to built in Weka ensembles. These results are shown in Table 7 where you can see that TreeEnsemble has a similar accuracy to RandomForest, Bagging and LogitBoost, is significantly better than AdaBoost, but RotationForest is more accurate.

4.4 Experiment 4

Experiment 4 is an analysis of the case study data-set FacesUCR to determine which classifier or ensemble is most accurate, shown in Table 8. Similar to previous experiments we see that ensembles produce better results than independent classifiers, TreeEnsemble has a similar performance to RandomForest, and RotationForest has the highest accuracy.

5 Conclusions

It is apparent from testing that the tuning of my ID3Coursework classifier does not drastically change its accuracy, performs better than base ID3, but worse than J48. Results show that ensembles are much more performant than independent classifiers. TreeEnsemble is relatively performant having a similar effectiveness to that of RandomForest. However, RotationForest proved to be much more accurate than the alternatives, confirming my hypotheses.

If I were to do this research again, I would include other performance improving methods as part of my implementations such as the F1 test and balanced accuracies.

References

[1] A Bagnall, M Flynn, J Large, J Line, A Bostrom, and G Cawley. Is rotation forest the best classifier for problems with continuous features? *arXiv preprint arXiv:1809.06705*, 2018.

A Appendix

Table 1: UCI Discrete Data Meta Data

Relation	Num Attributes	train/test cases	num classes
balance-scale	5	312/312	3
chess-krvk	7	14028/14028	18
chess-krvkp	37	1598/1599	2
connect-4	43	33778/33779	3
contraceptive-method	10	736/737	3
fertility	10	50/50	2
haberman	4	153/153	2
hayes-roth	5	66/66	3
led-display	8	250/250	10
lymphography	19	74/74	4
molecular-promoters	58	53/53	2
molecular-splice	61	1595/1595	3
monks-1	7	278/278	2
monks-2	7	300/301	2
monks-3	7	277/277	2
nursey	9	6480/6480	5
optdigits	65	2810/2810	10
pendigits	17	5496/5496	10
semeion	258	796/796	10
spect-heart	23	133/134	2
tic-tac-toe	10	479/479	2
zoo	17	50/51	7

Table 2: UCI Continuous Data Meta Data

Relation	Num Attributes	train/test cases	num classes
bank	17	2260/2261	2
blood	5	374/374	2
breast-cancer-wisc-diag	31	284/285	2
breast-tissue	10	53/53	6
cardiotocography-10clases	22	1063/1063	10
ecoli	8	168/168	8
glass	10	107/107	6
hill-valley	101	606/606	2
image-segmentation	19	1155/1155	7
ionosphere	34	175/176	2
iris	5	75/75	3
libras	91	180/180	15
musk-2	167	3299/3299	2
oocytes_merluccius_nucleus_4d	42	511/511	2
oocytes_trisopterus_states_5b	33	456/456	3
optical	63	2810/2810	10
ozone	73	1268/1268	2
page-blocks	11	2736/2737	5
parkinsons	23	97/98	2
pendigits	17	5496/5496	10
planning	13	91/91	2
post-operative	9	45/45	3
ringnorm	21	3700/3700	2
seeds	8	105/105	3
spambase	58	2300/2301	2
statlog-image	19	1155/1155	7
statlog-landsat	37	3217/3218	6
statlog-shuttle	10	29000/29000	7
steel-plates	28	970/971	7
synthetic-control	61	300/300	6
twonorm	21	3700/3700	2
vertebral-column-3clases	7	155/155	3
wall-following	25	2728/2728	4
waveform-noise	41	2500/2500	3
wine-quality-white	12	2449/2449	7
yeast	9	742/742	10

Table 3: FacesUCR Meta Data

Relation	Num Attributes	train/test cases	num classes
FacesUCR	131	200/2050	14

Table 4: Results from Experiment 1 with discrete data

Classifier	Accuracy
Id3	0.6698
J48	0.7487
Id3Coursework using IG	0.7114
Id3Coursework using Gini	0.7034
Id3Coursework using Chi Squared	0.7082
Id3Coursework using Chi Squared Yates Correction	0.7091

Table 5: Results from Experiment 1 with continuous data

Classifier	Accuracy
Id3	0.7241
J48	0.8185
Id3Coursework using IG	0.7879
Id3Coursework using Gini	0.7912
Id3Coursework using Chi Squared	0.7852
Id3Coursework using Chi Squared Yates Correction	0.7901

Table 6: Results for Ensembles from Experiment 2 with discrete and continuous data

Classifier	Accuracy Discrete	Accuracy Continuous
Ensemble 50% Attributes	0.7677	0.8497
Ensemble 100% Attributes	0.7201	0.8506

Table 7: Results for Weka Ensembles from Experiment 3 with discrete and continuous data

Classifier	Accuracy Discrete	Accuracy Continuous
Default TreeEnsemble	0.7737	0.8483
RandomForest	0.7814	0.8365
RotationForest	0.8530	0.8607
Bagging	0.7596	0.8386
AdaBoost	0.5875	0.6379
LogitBoost	0.7709	0.8047

Table 8: Results for FacesUCR case study

Classifier	Averaged Accuracy
J48	0.4775
Id3Coursework Info Gain	0.4863
Id3Coursework Gini	0.4960
Id3Coursework Chi Squared	0.4897
Id3Coursework Chi Squared Yates	0.4848
TreeEnsemble 50% Attributes	0.5983
TreeEnsemble 100% Attributes	0.6041
Random Forest	0.5907
Rotation Forest	0.7302
AdaBoost	0.2609