A Novel Algorithm for Scaling up the Accuracy of Decision Trees

Ali Mirza Mahmood[*1], K.Mrutunjaya Rao[2], Kiran Kumar Reddi[3]

1. Assistant Professor, Department of Computer Science, DMSSVH College of Engineering, Machilipatnam, India.
2. Professor, Department of Computer Science & Engineering, Vagdevi College of Engineering, Warangal, India.
3. Assistant Professor, Department of Computer Science, Krishna University, Machilipatnam, India.

Abstract:

Classification is one of the most efficient and widely used data mining technique. In classification, Decision trees can handle high dimensional data, and their representation is intuitive and generally easy to assimilate by humans. The area under the receiver operating characteristic curve, AUC is one of the recently used measures for calculating the performance of a classifier.In this paper, we presented two novel decision tree algorithms namely C4.45 and C4.55, aimed to improve the AUC value over the C4.5, which is a state-of-the-art decision tree algorithm. The empirical experiments conducted on 42 benchmark datasets have strongly indicated that C4.45 and C4.55 has significantly outperformed C4.5 on the AUC value.

Keywords: Decision Trees. Information gain. Area under Curve .Gain ratio . Laplace Correction .Confidence Factor.

1. Introduction

Classification is one of the major data mining technique, which is used for classifying data items into one of the known classes. In classification, a training algorithm is used for training the classifier by a set of data items, and then a test algorithm is used for fine-tuning the classifier, for classifying the newly coming data.

The main applications of classification include medical diagnosis, astronomy, molecular biology, bio-informatics, detecting faults in industrial applications, classifying financial market trends, loan approvals, image and pattern matching.

There are different ways (such as Decision trees, Navie bayes, NBTree, SVM,NN, Decision tables) to built classifiers[1-5]. Decision tree classifier is one of the popular technique, because it does not require any domain knowledge or parameter setting and therefore it is appropriate for Knowledge Exploration. The Receiver Operating Characteristics (ROC) curve has been shown as one of the measures for the quality of ranking [6], which shows the trade off between sensitivity and false positive rate for comparing two classification models.

The reminder of this paper is organized as follows. In section 2, we discussed the related work. In section 3, we presented our new algorithm. In section 4 we describe the experimental settings and results in detail. Finally, in section 5 we draw conclusions and outline our main direction for future work.

2. Related work

In the last decade, classification has attracted much attention from researchers.[7-11].In classification, Decision trees have gone through many new improvements [12],[13**]** in recent years.C4.5 [1] is still the Bench marking decision tree algorithm so, we conducted the comparison of our new algorithms with it. In C4.5, the information gain of each attribute is calculated and then gain ratio is applied for only those attributes, which has information gain value above the average. The splitting criteria is the sensitive issue in building decision tree[14].There have been numerous comparisons of the different classification algorithms[15].No single method has been found to be superior over all others for all data sets. Issues such as accuracy, training time, robustness, interpretability and scalability must be considered [16].The performance of the classification algorithm is usually examined by evaluating the accuracy of the classification [17].In classification accuracy, only the percentage of correctly classified instances are calculated. In ROC [19] the performance of the classifier across the entire range and error costs are caluculated.So, Area under the receiver operating characteristic curve, AUC is one of the recently used measures for calculating the performance of the classifier.

3 New decision tree algorithms for scaling up the accuracy

3.1. C4.45:

C4.5 has been observed to produce poor performance of class probability estimation and Ranking [19].The size of the tree produced by C4.5 is also huge. Our motivation is to further scale up the AUC value of C4.5 and to reduce the size of the tree .Our new algorithms has succeeded in doing so.

The new algorithm of C4.45 and C4.55 can be described as,

Algorithm C4.45 or C4.55 (**D**)
Input: a instances set **D**
Output: a C4.45 or C4.55 tree

1. If the number of instances is under 5, create a leaf node for the tree
2. Otherwise
3. For each attribute, calculate its measure score
4. Select the attribute *A* with the highest score for the test attribute
5. If the highest score is zero, create a leaf node for the tree
6. Otherwise
7. Partition *D* according to the test attribute *A*
8. For each possible value of *A*, create a child node for the tree
9. For each child node, recursively call the algorithm
10. Return a C4.45 or C4.55 tree

In the above algorithm we implements the following techniques to produce C4.45,
1. We increased the least number of instances per leaf node from 2 to 5, to make the size of tree moderate.
2. Pruning the tree is stopped.
3. We used the laplace correction to smooth the count of the leaves.

3.2. C4.55:

In the same above algorithm we implemented some other new techniques to produce C4.55,
1. We increased the least number of instances per leaf node from 2 to 5, to make the size of tree moderate.
2. The Confidence factor used for pruning is set from 0.25 to 0.95.
3. We used the binary split when building the tree on nominal attributes.

4 Experimental methods and Results

4.1 Data Sets:

We selected 42 datasets[20] from the UCI repository of machine learning datbases.These are the benchmarked datasets used by almost all the academia and practitioners[9],[19],[4] of classification research field. We downloaded some data sets in format of arff from main website of Weka [21],[22] and for remaining we created the arff format. The description of the data sets with Number of instances, Attributes, Classes, and Missing values have been provided below in the table 1. These data sets represent a wide range of domains and data characteristics.

Table 1. Description of data sets used in the experiments

Sno.	Dataset	Instances	Attrib	Classes	Missing
1	Anneal	898	39	6	Y
2	Anneal.ORIG	898	39	6	Y
3	Audiology	226	70	24	Y
4	Autos	205	26	7	Y
5	Balance-scale	625	5	3	N
6	Breast-cancer	286	10	2	Y
7	Breast-w	699	10	2	Y
8	Colic-h	368	23	2	Y
9	Colic-g	368	28	2	Y
10	Credit	690	16	2	Y
11	Credit-g	1,000	21	2	N
12	Diabetes	768	9	2	N
13	Glass	214	10	7	N
14	Heart-c	303	14	5	Y
15	Heart-h	294	14	5	Y
16	Heart-s	270	14	2	N
17	Hepatitis	155	20	2	Y
18	Hypothyroid	3,772	30	4	Y
19	Ionosphere	351	35	2	N
20	Iris	150	5	3	N
21	Kr-vs-kp	3,196	37	2	N
22	Labour	57	17	2	Y
23	Letter	20,000	17	26	N
24	Liver	345	7	2	N
25	Lymphography	148	19	4	N
26	Mushroom	8,124	23	2	Y
27	Nursery	11,025	9	5	N
28	Primary-tumor	339	18	21	Y
29	Segment	2,310	20	7	N
30	Sick	3,772	30	2	Y
31	Sonar	208	61	2	N
32	Soybean	683	36	19	Y
33	Splice	3,190	62	3	N
34	Tictactoe	958	10	2	N
35	Vehicle	846	19	4	N
36	Vote	435	17	2	Y
37	Vowel	990	14	11	N
38	Waveform	5,000	41	3	N
39	Wdbc	569	31	2	N
40	Wine-red	1,599	11	6	N
41	Wine-white	4,898	12	7	N
42	Zoo	101	18	7	N

We downloaded these datasets from main website of Weka and UCI machine learning repository.

4.2. Results and Discussion:

Our new algorithms are implemented with in the Weka environment. We used 2/3 of the examples of the instances in a dataset for training and 1/3 for testing. The AUC value and the standard deviation of all three algorithms on each data set are obtained via 20 runs of 66 % train/test percentage split. Finally The average of 20 runs is calculated for all the three algorithms.

We conducted a two-tailed t-test [23] with 95% confidence level [18].The comparison results and AUC value of all three algorithms on each data set are shown in Figure 1 and Table 2.The statistically significant upgrade or degradation with a 95% confidence level over C4.5 is indicated by the symbols V and * .Our new algorithms Win ,Tie, and Lose on datasets are represented below the tabular form by w/t/l values as summarized.

Furthermore , it is worth mentioning that, in the 42 datasets we tested, Anneal, Hepatitis, Hypothyroid and Sick are the typical unbalanced datasets, the AUC value of these unbalanced datasets with C4.45 and C4.55 (3 wins out of 4) is higher than the C4.5.

It is also worthwhile to mention that, in the 42 datasets Credit-g, Hypothyroid, Kr-vs-kp, Letter, Mushroom, Nursery, Segment, Sick, Splice, Tic-tac-toe, Vowel, Waveform, Wine-red, Wine-white are 14 large datasets ,the AUC value of these large datasets with C4.45 and C4.55 (12 wins out of 14)is higher than C4.5.It states that C4.45 and C4.55 are accurate for scalable datasets than C4.5.Increasing the accuracy of medical diagnosis from 98% to 99% may cut cost by half because the number of errors is halved[3].

Figure. 1 AUC comparison on 42 datasets

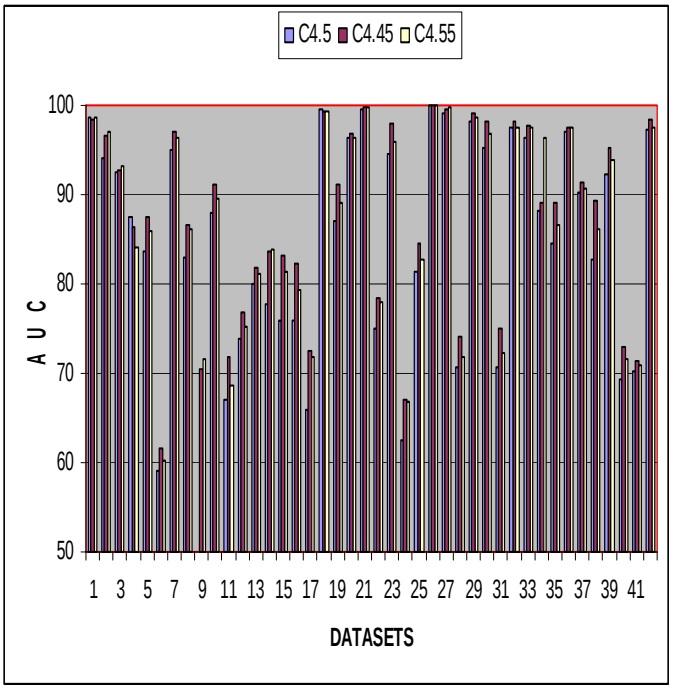

Table 2 Experimental result on AUC value and standard deviation

Dataset	C4.5	C4.45		C4.55	
anneal	98.57±0.12	98.49±0.08	*	98.71±0.07	v
anneal.ORIG	94.00±0.44	96.68±0.07	v	97.05±0.09	v
audiology	92.45±0.16	92.69±0.18	v	93.14±0.15	v
autos	87.42±0.35	86.40±0.34	*	84.18±0.32	*
balance-scale	83.56±0.19	87.58±0.18	v	85.80±0.17	v
breast-cancer	59.10±0.50	61.70±0.43	v	60.19±0.67	v
breast-w	95.02±0.22	97.09±0.10	v	96.30±0.12	v
colic-h	83.02±0.36	86.60±0.23	v	86.17±0.31	v
colic-g	50.00±0.00	70.51±0.41	v	71.69±0.40	v
credit	87.87±0.41	91.15±0.12	v	89.62±0.16	v
credit-g	66.96±0.42	71.88±0.24	v	68.74±0.26	v
diabetes	73.90±0.43	76.90±0.28	v	75.32±0.28	v
Glass	79.96±0.48	81.82±0.44	v	81.17±0.38	v
heart-c	77.79±0.60	83.63±0.35	v	83.95±0.35	v
heart-h	75.95±0.74	83.28±0.47	v	81.40±0.49	v
heart-s	75.85±0.47	82.21±0.34	v	79.27±0.49	v
hepatitis	65.89±0.87	72.48±0.55	v	71.76±0.61	v
hypothyroid	99.47±0.04	99.37±0.03	*	99.41±0.03	*
ionosphere	87.07±0.43	91.16±0.47	v	89.11±0.44	v
iris	96.41±0.21	96.85±0.21	v	96.46±0.23	v
kr-vs-kp	99.65±0.02	99.78±0.01	v	99.74±0.01	v
labor	75.11±0.97	78.48±0.97	v	78.00±1.08	v
letter	94.61±0.02	97.87±0.01	v	95.91±0.02	v
liver	62.60±0.63	67.13±0.53	v	66.89±0.54	v
lymphography	81.39±0.50	84.65±0.46	v	82.72±0.61	v
mushroom	100.0±0.00	100.0±0.00		100.0±0.00	*
nursery	99.15±0.02	99.45±0.01	v	99.75±0.01	v
primary-tumor	70.70±0.19	73.98±0.20	v	71.86±0.21	v
segment	98.28±0.03	99.20±0.02	v	98.66±0.03	v
sick	95.13±0.22	98.22±0.13	v	96.72±0.24	v
sonar	70.64±0.82	74.89±0.80	v	72.19±0.63	v
soybean	97.60±0.10	98.14±0.05	v	97.42±0.07	*
splice	96.42±0.05	97.69±0.05	v	97.42±0.07	v
Tictactoe	88.18±0.24	89.02±0.21	v	96.37±0.12	v
vehicle	84.65±0.20	89.06±0.09	v	86.63±0.13	v
vote	96.98±0.14	97.60±0.12	v	97.47±0.13	v
vowel	90.22±0.12	91.41±0.14	v	90.73±0.10	v
waveform	82.80±0.11	89.37±0.07	v	86.13±0.09	v
wdbc	92.32±0.22	95.18±0.16	v	93.82±0.22	v
wine-red	69.24±0.17	72.99±0.17	v	71.70±0.18	v
wine-white	70.32±0.10	71.42±0.08	v	71.02±0.10	v
zoo	97.34±0.21	98.50±0.16	v	97.47±0.20	v
Average	**84.37**	**87.44**		**86.62**	
	(v/ /*)	**(38/1/3)**		**(38/0/4)**	

v, *: statistically significant upgrade or degradation over C4.5. The mean and *w*/*t*/*l* values are summarized at the bottom of the table

Table 3 Experimental results on training time and standard deviation

Dataset	C4.5	C4.45		C4.55	
anneal	0.0351±0.014	0.0226±0.008	*	0.0977±0.036	v
anneal.ORIG	0.0665±0.022	0.0493±0.022	*	0.3868±0.079	v
audiology	0.0102±0.008	0.0032±0.007	*	0.1071±0.027	v
autos	0.0102±0.008	0.0071±0.008	*	0.1320±0.036	v
balance-scale	0.0055±0.008	0.0055±0.008		0.1727±0.049	v
breast-cancer	0.0016±0.005	0.0008±0.004	*	0.2078±0.079	v
breast-w	0.0063±0.008	0.0063±0.008		0.0525±0.020	v
colic-h	0.0072±0.008	0.0056±0.008	*	0.0625±0.025	v
colic-g	0.0039±0.007	0.0046±0.007	v	0.2469±0.082	v
credit	0.0125±0.006	0.0040±0.007	*	0.2329±0.092	v
credit-g	0.0243±0.008	0.0164±0.004	*	1.0867±0.372	v
diabetes	0.0156±0.005	0.0140±0.005	*	0.1602±0.080	v
Glass	0.0087±0.008	0.0055±0.008	*	0.1008±0.023	v
heart-c	0.0055±0.008	0.0047±0.007	*	0.0657±0.028	v
heart-h	0.0055±0.008	0.0031±0.006	*	0.0516±0.016	v
heart-s	0.0071±0.008	0.0046±0.007	*	0.0649±0.030	v
hepatitis	0.0023±0.006	0.0024±0.006		0.0251±0.010	v
hypothyroid	0.0532±0.016	0.0578±0.027	v	0.0947±0.028	v
ionosphere	0.0405±0.019	0.0352±0.007	*	0.0828±0.015	v
iris	0.0008±0.004	0.0000±0.000	*	0.0102±0.008	v
kr-vs-kp	0.0390±0.021	0.0306±0.020	*	0.2228±0.035	v
labor	0.0000±0.000	0.0008±0.004	v	0.0047±0.007	v
letter	2.9883±0.112	2.1086±0.104	*	13.5812±0.609	v
liver	0.0078±0.008	0.0030±0.006	*	0.1188±0.045	v
lymphography	0.0000±0.000	0.0016±0.005	v	0.0501±0.032	v
mushroom	0.0195±0.007	0.0141±0.005	*	0.1233±0.029	v
nursery	0.0407±0.008	0.0274±0.007	*	1.7336±0.201	v
primary-tumor	0.0078±0.008	0.0038±0.007	*	0.1953±0.053	v
segment	0.1556±0.021	0.1406±0.010	*	0.4016±0.046	v
sick	0.0798±0.037	0.0673±0.024	*	0.1868±0.046	v
sonar	0.0344±0.006	0.0306±0.003	*	0.0727±0.017	v
soybean	0.0164±0.003	0.0180±0.022	v	0.3213±0.060	v
splice	0.1069±0.008	0.0829±0.007	*	0.4771±0.083	v
Tictactoe	0.0046±0.007	0.0024±0.006	*	0.2040±0.038	v
vehicle	0.0499±0.010	0.0406±0.008	*	0.3788±0.127	v
vote	0.0030±0.006	0.0016±0.005	*	0.0204±0.023	v
vowel	0.1110±0.009	0.0993±0.021	*	0.6743±0.132	v
waveform	1.2087±0.046	1.1079±0.051	*	3.2907±0.198	v
wdbc	0.0422±0.021	0.0368±0.008	*	0.0641±0.010	v
wine-red	0.1126±0.023	0.0767±0.005	*	1.4664±0.262	v
wine-white	0.6093±0.062	0.4023±0.033	*	7.5867±0.487	v
zoo	0.0008±0.003	0.0016±0.005	v	0.0360±0.009	v
Average	**0.1419**	**0.1083**		**0.8250**	
	(v/ /*)	**(6/3/33)**		**(42/0/0)**	

v, *: statistically significant upgrade or degradation over C4.**5**. The mean and *w*/*t*/*l* values are summarized at the bottom of the table

The highlights of C4.45 and C4.55 can be summarized as:

1. C4.45 has significantly outperformed on C4.5 on the 42 datasets we tested. Surprisingly C4.45 had registered wins on 38 datsets,ties on 1 dataset and loses on 3 datasets out of 42 datsets. The average AUC value for C4.45(87.44) is better than C4.5(84.37) strongly indicating that C4.45 is better than C4.5.
2. C4.55 has also significantly outperformed on C4.5 on the 42 datasets we tested. C4.55 had registered wins on 38 datsets, ties on 0 datasets and loses on 4 datasets out of 42 datsets. The average AUC value for C4.45(86.62) is better than C4.5(84.37) strongly indicating that C4.45 is better than C4.5.

5 Conclusions and future work

In this paper, we presented two novel decision tree algorithms namely C4.45 and C4.55, aimed to improve the AUC value over the C4.5 decision tree algorithm .The empirical experiments conducted on 42 benchmark datasets has strongly

indicated that C4.45 and C4.55 has significantly outperformed C4.5 on the AUC value.

A direction for the future work is to study the incorporation of more sophisticated methods for calculating measure score.

References

[1] J. R Quinlan,.(1993) C4.5: Programs for MachineLearning.Morgan Kaufmann, Los Altos.
[2] Tzu-Tsung Wong (2009) Alternative prior assumptions for improving the performance of naïve Bayesian classifiers. Data Min Knowl Disc (2009) 18:183–213
[3] Ron Kovi.(2008) Scaling up the accuracy of navie-bayes classifiers: A Decision Tree Hybrid.
[4] Mark Hall, Eibe Frank(2008) Combining Naive Bayes and Decision Tables. Association for the Advancement of Artificial Intelligence
[5] Brijesh Verma . Syed Zahid Hassan(2009) Hybrid ensemble approach for classification. Appl Intell
[6] Rosset S,Perlich C,Zadrozny B(2007) Ranking-based evalution of regression models Knowledge information systems 12(3):331-353.
[7] Bei Hui · Ying Yang · Geoffrey I. Webb(2009) Anytime classification for a pool of instances. Mach Learn (2009) 77: 61–102
[8] Tzu-Tsung Wong(2009) Alternative prior assumptions for improving the performance of naïve Bayesian classifiers.Data Min Knowl Disc (2009) 18:183–213
[9] Xing Zhang · Guoqing Chen · QiangWei(2009) Building a highly-compact and accurate associative classifier. Appl Intell
[10] Weiwei Cheng · Eyke Hüllermeier (2009) Combining instance-based learning and logistic regression for multilabel classification. Mach Learn (2009) 76: 211–225
[11] Lior Rokach(2009) Ensemble-based classifiers. Artif Intell Rev DOI 10.1007/s10462-009-9124-7
[12] Sattar Hashemi · Ying Yang(2009) Flexible decision tree for data stream classificationin the presence of concept change, noise and missing values. Data Min Knowl Disc (2009) 19:95–131
[13] Gary M. Weiss · Ye Tian (2009) Maximizing classifier utility when there are data acquisition and modeling costs. Data Min Knowl Disc (2008) 17:253–282
[14] Drummond, C., & Holte, R. (2000). Exploiting the cost (in)sensitivity of decision tree splitting criteria. Proceedings of the Seventeenth International Conference on Machine Learning (pp. 239–246).
[15] Lim, T.-J., Loh, W.-Y., & Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity,and training time of thirty-three old and new classification algorithms. *Machine Learning, 40:3,* 203–228.
[16] J.Han ,M .Kamber,(2008) "Data Mining: concept and techniques "Second edition,Morgan Kaufmann Publishers.
[17] Margraet H.Dunham,S.Sridhar (2007).Data mining Introductory and advanced topics,Pearson education.
[18] Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning, 45:2,* 171–186.
[19] Liangxiao Jiang · Chaoqun Li · Zhihua Cai(2009). Learning decision tree for ranking. Knowl Inf Syst (2009) 20:123–135
[20] Blake, C., & Merz, C. J.(2000).UCI repository of machinelearning databases. Machine-readable data repository,Department of Information and Computer Science, University of California at Irvine, Irvine, CA. at http://www.ics.uci.edu/mlearn/MLRepository.html.
[21] Ian HW, Eibe F (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco. http://prdownloads.sourceforge.net/weka/datasets-UCI.jar
[22] Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques with java implementations* (2nd ed.). San Mateo: Morgan Kaufmann.
[23] Nadeau ,C;Bengio.y(2003).Inference for the Generalization error.Machine Learning 52(3):239-281.