Appendix to the article
A Comparison of Pruning Criteria for Probability Trees
in *Machine Learning*, 2009.

Daan Fierens Jan Ramon Hendrik Blockeel Maurice Bruynooghe

http://www.cs.kuleuven.be/∼dtai/pruning-ml09/
daan.fierens@cs.kuleuven.be

This appendix contains detailed results for three groups of experiments:

- The main experiments in which we compare the six pruning criteria on 26 datasets. See Section 1.
- Experiments with alternative splitting criteria (in the main experiments we always used information gain as the splitting criterion). See Section 2.
- The experiments on manipulated data in which we assess the influence of certain characteristics of the data (number of classes, class skew, number of examples). See Sections 3 to 5.

Contents

1 Main Experiments

1.1 Results for all Performance Measures

In the paper we report the results of experiments on 26 dataset for the performance measures AUC, RMSE and tree size. In this section, we give the results for some additional performance measures: conditional log-likelihood, classification accuracy and running time.

1.1.1 CLL

CLL denotes the negative normalized conditional log-likelihood:

$$\frac{-log_2(\prod_{i=1}^{N} p_{\mathcal{T}}(c_i|e_i))}{N} = \frac{-\sum_{i=1}^{N} log_2(p_{\mathcal{T}}(c_i|e_i))}{N}$$

where c_i denotes the true class label of example e_i, $p_{\mathcal{T}}(c_i|e_i)$ denotes the probability of this class label as predicted by the tree $\mathcal{T}$, and N denotes the total number of examples. Note that because CLL is the *negative* normalized conditional log-likelihood, lower CLL is better.

Table 1: CLL for the six pruning criteria (lower is better).

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	0.579	0.662 ∘	0.651 ∘	0.602	0.599	0.725 ∘
australian credit	0.502	0.501	0.513	0.523	0.549	0.647 ∘
balance scale	0.922	0.962	0.956	0.954	0.898	0.789 •
breast wisconsin	0.226	0.261	0.214	0.223	0.239	0.225
chess kr-kp	0.046	0.068 ∘	0.045	0.047	0.048	0.036
diabetes	0.766	0.769	0.779	0.764	0.820	0.916 ∘
german credit	0.806	0.806	0.858	0.841	0.967 ∘	1.051 ∘
heart cleveland	0.745	0.774	0.768	0.748	0.785	0.829
ionosphere	0.427	0.404	0.438	0.450	0.427	0.448
mushroom	0.002	0.004	0.002	0.002	0.003	0.002
pen digits	0.388	0.595 ∘	0.565 ∘	0.411 ∘	0.390	0.411 ∘
primary tumor	3.266	3.649 ∘	3.533 ∘	3.207	3.492 ∘	3.778 ∘
segment	0.293	0.385 ∘	0.360 ∘	0.307	0.304	0.318 ∘
soybean	1.453	1.901 ∘	1.892 ∘	1.510	1.562 ∘	1.736 ∘
splice	0.282	0.282	0.276	0.288	0.305 ∘	0.337 ∘
thyroid	0.068	0.084 ∘	0.077	0.068	0.073	0.091 ∘
vehicle	0.941	1.071 ∘	0.980	0.952	1.070 ∘	1.161 ∘
voting	0.214	0.222	0.212	0.211	0.222	0.217
vowel	1.684	2.232 ∘	2.152 ∘	1.828 ∘	1.676	1.693
yeast	1.759	1.778	1.779	1.716	1.878 ∘	2.426 ∘
biodegradability	0.880	0.886	0.892	0.891	0.907	0.925
cora	1.303	1.435 ∘	1.398 ∘	1.337	1.368	1.600 ∘
diterpenes	2.010	2.261 ∘	2.217 ∘	1.989	2.013	2.221 ∘
gene	1.692	1.950 ∘	1.870 ∘	1.704	1.789 ∘	1.878 ∘
hiv	0.184	0.197 ∘	0.188 ∘	0.189 ∘	0.200 ∘	0.208 ∘
mutagenesis	0.854	0.833	0.856	0.827	0.882	0.821
average	0.857	0.96	0.941	0.869	0.903	0.98
wins/ties/losses		4/0/22	4/1/21	7/1/18	2/0/24	4/1/21
significant w/t/l		0/13/13	0/16/10	0/23/3	0/18/8	1/9/16
average rank	1.87	4.46	3.60	2.63	3.85	4.60

1.1.2 Classification Accuracy

Table 2: Accuracy (%) for the six pruning criteria (higher is better).

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	86.7	81.2 ∘	81.5 ∘	84.4	87.3	86.4
australian credit	84.9	85.6	85.0	85.1	85.2	81.3
balance scale	76.1	71.2 ∘	74.7	74.6	78.2	77.2
breast wisconsin	95.0	94.6	95.2	94.9	95.2	94.4
chess kr-kp	99.3	98.5 ∘	99.4	99.3	99.4	99.5
diabetes	73.3	73.2	73.2	72.7	73.8	69.3
german credit	71.0	69.7	70.8	70.5	70.5	68.2
heart cleveland	77.8	73.8	77.1	76.5	77.8	75.1
ionosphere	89.9	91.0	89.8	89.2	90.7	89.8
mushroom	100.0	100.0	100.0	100.0	100.0	100.0
pen digits	95.4	88.4 ∘	89.0 ∘	93.5 ∘	96.1 •	96.1 •
primary tumor	37.4	22.8 ∘	22.5 ∘	37.7	37.8	34.5
segment	94.9	92.6 ∘	93.1 ∘	93.9 ∘	95.8 •	95.9
soybean	77.0	51.1 ∘	53.0 ∘	71.3 ∘	80.1	78.5
splice	94.3	94.6	94.5	94.0	93.8	92.7 ∘
thyroid	98.8	98.5	98.7	98.8	98.8	98.2 ∘
vehicle	71.0	63.0 ∘	68.5	69.0	72.2	71.4
voting	95.0	95.1	95.0	95.1	95.4	94.6
vowel	71.8	40.4 ∘	43.5 ∘	57.2 ∘	81.5 •	82.5 •
yeast	56.8	54.7	54.6	57.0	56.2	52.1 ∘
biodegradability	68.3	67.0	72.5	67.7	73.1	73.4
cora	77.9	74.4 ∘	75.4 ∘	77.1	75.7	71.9 ∘
diterpenes	59.1	43.5 ∘	44.2 ∘	53.6 ∘	60.5	60.1
gene	63.3	55.3 ∘	56.5 ∘	62.9	62.4	61.1
hiv	96.5	96.5	96.6	96.4 ∘	96.6	95.9 ∘
mutagenesis	66.7	65.8	67.5	65.5	69.8	71.5
average	79.9	74.7	75.8	78.4	80.9	79.7
w/t/l		4/0/22	8/1/17	6/1/19	19/0/7	10/1/15
signif w/t/l		0/14/12	0/17/9	0/20/6	3/23/0	2/19/5
avg rank	2.87	4.81	3.87	3.67	2.00	3.79

1.1.3 Running Time

We measured the time in seconds to build a tree on the entire dataset on a Intel Pentium 4 machine with 1.70GHz CPU and 1GB RAM. For NoPruning we show absolute times in seconds, for the other criteria we show relative times with respect to NoPruning. We show the results for attribute-value ('non-relational') datasets and relational datasets separately, and we list the datasets in ascending order of running time (for NoPruning).

Note that, as explained in the paper, the running time is almost the same for all pruning criteria, except Rand.

Table 3: Running times for the six postpruning criteria on the attribute-value datasets.

	absolute time for NoPruning	relative time for MDL	relative time for BIC	relative time for Chi	relative time for EBP	relative time for Rand
voting	0.3s	1.01	1.01	1.02	1.01	3.03
breast wisconsin	0.7s	1.01	1.01	1.01	1.01	4.17
heart cleveland	1.1s	1.01	1.01	1.01	1.01	2.27
ionosphere	1.3s	1.01	1.01	1.01	1.02	5.03
chess kr-kp	1.4s	1.01	1.01	1.01	1.01	9.64
mushroom	1.6s	1.02	1.02	1.02	1.01	12.48
thyroid	1.9s	1.01	1.01	1.02	1.00	5.26
annealing	2.7s	1.01	1.01	1.02	1.01	2.86
australian credit	3.0s	1.01	1.01	1.02	1.02	2.38
balance scale	3.3s	1.04	1.04	1.04	1.04	1.25
soybean	3.6s	1.03	1.03	1.04	1.02	1.92
segment	5.0s	1.00	1.00	1.00	1.00	6.67
primary tumor	5.9s	1.02	1.02	1.02	1.01	1.49
german credit	6.4s	1.01	1.01	1.01	1.01	2.38
vowel	6.6s	1.03	1.03	1.03	1.02	2.94
diabetes	7.7s	1.02	1.02	1.02	1.02	1.64
splice	9.0s	1.06	1.06	1.07	1.07	7.14
vehicle	9.0s	1.02	1.01	1.03	1.03	2.78
pen digits	23.5s	1.01	1.01	1.01	1.01	6.25
yeast	32.7s	1.02	1.02	1.02	1.03	1.54
average		1.02	1.02	1.02	1.02	4.16

Table 4: Running times for the six postpruning criteria on the relational datasets.

	absolute time for NoPruning	relative time for MDL	relative time for BIC	relative time for Chi	relative time for EBP	relative time for Rand
biodegradability	7.9s	1.02	1.02	1.02	1.02	2.08
mutagenesis	10.4s	1.04	1.04	1.04	1.03	1.37
gene	21.6s	1.04	1.04	1.06	1.02	1.81
cora	38.5s	1.03	1.03	1.04	1.06	1.23
diterpenes	52.8s	1.02	1.02	1.02	1.03	2.56
hiv	1185.2s	1.00	1.00	1.01	1.00	1.43
average		1.03	1.03	1.03	1.03	1.75

1.2 Pairwise Comparison of all Pruning Criteria

In the paper we only report wins/ties/losses of each pruning criterion versus RAND. Here we report the wins/ties/losses between all pairs of pruning criteria. As in the paper, we consider both 'plain' wins/ties/losses (without significance tests), and significant wins/ties/losses.

1.2.1 Pairwise 'Plain' Wins/Ties/Losses

The entries in the tables below should be interpreted as follows. The entry "0/0/26" in Table 5, in the row for MDL and the column for BIC, indicates that MDL wins 0 times, ties 0 times and loses 26 times versus BIC. Note that all the wins/ties/losses in the last column (the column for RAND) are the wins/ties/losses that are reported in the paper (i.e., the wins/ties/losses of each pruning criterion versus RAND).

Table 5: Plain wins/ties/losses for the performance measure AUC.

	BIC	CHI	EBP	NOPRUNING	RAND
MDL	0/0/26	0/0/26	1/0/25	3/0/23	1/0/25
BIC		11/1/14	11/0/15	7/1/18	8/1/17
CHI			14/0/12	11/1/14	7/1/18
EBP				12/0/14	5/0/21
NOPRUNING					12/1/13

Table 6: Plain wins/ties/losses for the performance measure RMSE.

	BIC	CHI	EBP	NOPRUNING	RAND
MDL	4/0/22	4/0/22	7/0/19	15/0/11	4/0/22
BIC		10/1/15	12/0/14	17/1/8	8/1/17
CHI			14/0/12	18/1/7	4/1/21
EBP				21/0/5	9/0/17
NOPRUNING					4/1/21

Table 7: Plain wins/ties/losses for the performance measure CLL.

	BIC	CHI	EBP	NOPRUNING	RAND
MDL	5/0/21	5/0/21	9/0/17	14/0/12	4/0/22
BIC		8/1/17	13/0/13	18/1/7	4/1/21
CHI			19/0/7	19/1/6	7/1/18
EBP				20/0/6	2/0/24
NOPRUNING					4/1/21

Table 8: Plain wins/ties/losses for the performance measure accuracy.

	BIC	Chi	EBP	NoPruning	Rand
MDL	7/0/19	6/0/20	3/0/23	11/0/15	4/0/22
BIC		11/1/14	4/0/22	12/1/13	8/1/17
Chi			6/0/20	13/1/12	6/1/19
EBP				20/0/6	19/0/7
NoPruning					10/1/15

Table 9: Plain wins/ties/losses for tree size.

	BIC	Chi	EBP	NoPruning	Rand
MDL	26/0/0	26/0/0	26/0/0	26/0/0	26/0/0
BIC		14/2/10	23/0/3	25/1/0	15/1/10
Chi			23/0/3	25/1/0	14/1/11
EBP				26/0/0	3/0/23
NoPruning					0/1/25

1.2.2 Pairwise Significant Wins/Ties/Losses

Table 10: Significant wins/ties/losses for the performance measure AUC.

	BIC	Chi	EBP	NoPruning	Rand
MDL	0/15/11	0/11/15	0/12/14	1/11/14	0/12/14
BIC		1/15/10	1/14/11	2/13/11	0/14/12
Chi			1/22/3	4/16/6	0/22/4
EBP				2/21/3	0/25/1
NoPruning					2/19/5

Table 11: Significant wins/ties/losses for the performance measure RMSE.

	BIC	Chi	EBP	NoPruning	Rand
MDL	0/18/8	0/13/13	3/12/11	9/11/6	0/12/14
BIC		0/18/8	3/15/8	12/10/4	0/18/8
Chi			5/18/3	14/9/3	2/18/6
EBP				17/9/0	1/17/8
NoPruning					1/9/16

Table 12: Significant wins/ties/losses for the performance measure CLL.

	BIC	Chi	EBP	NoPruning	Rand
MDL	0/18/8	0/12/14	2/15/9	7/12/7	0/13/13
BIC		0/16/10	5/15/6	10/12/4	0/16/10
Chi			6/18/2	14/9/3	0/23/3
EBP				14/11/1	0/18/8
NoPruning					1/9/16

Table 13: Significant wins/ties/losses for the performance measure accuracy.

	BIC	Chi	EBP	NoPruning	Rand
MDL	0/22/4	0/18/8	1/13/12	3/12/11	0/14/12
BIC		1/19/6	0/17/9	4/14/8	0/17/9
Chi			0/19/7	6/15/5	0/20/6
EBP				8/18/0	3/23/0
NoPruning					2/19/5

Table 14: Significant wins/ties/losses for tree size.

	BIC	Chi	EBP	NoPruning	Rand
MDL	20/6/0	24/2/0	25/1/0	25/1/0	22/4/0
BIC		14/6/6	19/7/0	25/1/0	15/3/8
Chi			20/5/1	25/1/0	10/11/5
EBP				25/1/0	1/4/21
NoPruning					0/1/25

2 Experiments with Other Splitting Criteria

An important parameter of the learning algorithm for probability trees is the splitting criterion. For all the experiments reported in the paper or in the other sections of this appendix, we used information gain as the splitting criterion. However, we also performed all the main experiments (on 26 datasets, with 6 pruning criteria and 6 performance measures) with gain ratio and gini index as the splitting criteria. We report the results below.

2.1 Influence of the Splitting Criterion: Summary

We now first give some summaries of the results with the three splitting criteria. As earlier, for each performance measure and each pruning criterion, we report the average results, the plain wins/ties/losses, the significant wins/ties/losses, and the ranks. These summaries show that the differences between the results for the various splitting criteria are small, and that the conclusions formulated in the paper hold for all of the splitting criteria.

Table 15: Influence of the splitting criterion on the performance measure AUC.

Splitting criterion		RAND	MDL	BIC	CHI	EBP	NOPRUNING
information gain	average	88.8	84.9	86.6	88.2	88.1	88.7
gain ratio	average	88.9	80.7	84.3	87.7	87.4	88.7
gini index	average	88.4	83.7	85.8	87.7	87.9	88.7
information gain	w/t/l		1/0/25	8/1/17	7/1/18	5/0/21	12/1/13
gain ratio	w/t/l		0/0/26	5/1/20	6/1/19	5/0/21	14/1/11
gini index	w/t/l		0/0/26	8/1/17	11/1/14	5/0/21	12/1/13
information gain	signif w/t/l		0/12/14	0/14/12	0/22/4	0/25/1	2/19/5
gain ratio	signif w/t/l		0/8/18	0/14/12	0/21/5	0/23/3	2/21/3
gini index	signif w/t/l		0/10/16	0/14/12	0/22/4	0/23/3	3/21/2
information gain	avg rank	2.33	5.65	3.56	3.17	3.50	2.79
gain ratio	avg rank	2.21	5.90	4.08	3.13	3.35	2.33
gini index	avg rank	2.44	5.85	3.63	2.83	3.42	2.83

Table 16: Influence of the splitting criterion on the performance measure RMSE.

Splitting		RAND	MDL	BIC	CHI	EBP	NOPRUNING
information gain	average	0.387	0.411	0.403	0.392	0.394	0.415
gain ratio	average	0.386	0.420	0.408	0.389	0.392	0.414
gini index	average	0.387	0.412	0.404	0.392	0.393	0.415
information gain	w/t/l		4/0/22	8/1/17	4/1/21	9/0/17	4/1/21
gain ratio	w/t/l		3/0/23	5/1/20	7/1/18	9/0/17	4/1/21
gini index	w/t/l		4/0/22	6/1/19	5/1/20	8/0/18	5/1/20
information gain	signif w/t/l		0/12/14	0/18/8	2/18/6	1/17/8	1/9/16
gain ratio	signif w/t/l		0/12/14	0/17/9	2/19/5	2/18/6	1/9/16
gini index	signif w/t/l		0/12/14	0/17/9	1/19/6	1/17/8	1/9/16
information gain	avg rank	2.17	4.69	3.29	3.13	3.12	4.60
gain ratio	avg rank	2.13	4.87	3.58	2.79	3.04	4.60
gini index	avg rank	2.13	4.69	3.25	3.17	3.19	4.56

Table 17: Influence of the splitting criterion on the performance measure CLL.

Splitting		RAND	MDL	BIC	CHI	EBP	NOPRUNING
information gain	average	0.857	0.960	0.941	0.869	0.903	0.980
gain ratio	average	0.854	1.010	0.974	0.861	0.896	0.977
gini index	average	0.861	0.976	0.955	0.879	0.903	0.983
information gain	w/t/l		4/0/22	4/1/21	7/1/18	2/0/24	4/1/21
gain ratio	w/t/l		2/0/24	4/1/21	9/1/16	4/0/22	4/1/21
gini index	w/t/l		3/0/23	4/1/21	6/1/19	5/0/21	4/1/21
information gain	signif w/t/l		0/13/13	0/16/10	0/23/3	0/18/8	1/9/16
gain ratio	signif w/t/l		0/9/17	0/17/9	1/21/4	0/19/7	1/9/16
gini index	signif w/t/l		0/12/14	0/15/11	0/23/3	0/17/9	2/8/16
information gain	avg rank	1.87	4.46	3.60	2.63	3.85	4.60
gain ratio	avg rank	1.94	4.98	3.58	2.56	3.65	4.29
gini index	avg rank	1.90	4.46	3.56	2.83	3.62	4.63

Table 18: Influence of the splitting criterion on the performance measure accuracy.

Splitting		RAND	MDL	BIC	CHI	EBP	NOPRUNING
information gain	average	79.9	74.7	75.8	78.4	80.9	79.7
gain ratio	average	80.3	73.8	75.1	78.7	80.9	79.6
gini index	average	80.1	75.0	76.2	78.7	80.8	79.6
information gain	w/t/l		4/0/22	8/1/17	6/1/19	19/0/7	10/1/15
gain ratio	w/t/l		4/0/22	7/1/18	6/1/19	15/0/11	10/1/15
gini index	w/t/l		5/0/21	8/1/17	7/1/18	19/0/7	9/1/16
information gain	signif w/t/l		0/14/12	0/17/9	0/20/6	3/23/0	2/19/5
gain ratio	signif w/t/l		0/14/12	0/18/8	0/21/5	2/24/0	2/18/6
gini index	signif w/t/l		0/16/10	0/18/8	0/20/6	2/24/0	2/19/5
information gain	avg rank	2.87	4.81	3.87	3.67	2.00	3.79
gain ratio	avg rank	2.69	5.02	3.69	3.52	2.25	3.83
gini index	avg rank	2.90	4.83	3.73	3.67	2.04	3.83

Table 19: Influence of the splitting criterion on tree size.

Splitting		RAND	MDL	BIC	CHI	EBP	NOPRUNING
information gain	average	34.9	7.6	14.3	29.6	61.1	193.7
gain ratio	average	50.4	6.9	14.4	41.6	65.3	210.5
gini index	average	32.8	7.4	14.2	31.0	62.0	196.2
information gain	w/t/l		26/0/0	15/1/10	14/1/11	3/0/23	0/1/25
gain ratio	w/t/l		26/0/0	19/1/6	17/1/8	4/0/22	0/1/25
gini index	w/t/l		26/0/0	15/1/10	12/1/13	2/0/24	0/1/25
information gain	signif w/t/l		22/4/0	15/3/8	10/11/5	1/4/21	0/1/25
gain ratio	signif w/t/l		25/1/0	15/9/2	11/13/2	1/8/17	0/1/25
gini index	signif w/t/l		26/0/0	15/3/8	8/11/7	1/5/20	0/1/25
information gain	avg rank	3.29	1.00	2.96	3.15	4.69	5.94
gain ratio	avg rank	3.60	1.02	2.69	3.21	4.54	5.94
gini index	avg rank	3.17	1.00	2.94	3.25	4.69	5.94

2.2 Results for the Splitting Criterion Gain Ratio

In the previous section, we have shown summaries of the results obtained with gain ratio as the splitting criterion. In this section we report these results in detail.

2.2.1 AUC

Table 20: AUC for the six pruning criteria with gain ratio as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	88.6	81.6 ∘	84.0 ∘	86.5	85.2	90.9 •
australian credit	91.1	90.3	91.0	91.0	91.3	90.1
balance scale	88.7	82.9 ∘	87.4	88.3	89.5	92.6 •
breast wisconsin	97.4	96.0	97.9	97.7	96.9	98.0
chess kr-kp	99.9	99.7	99.9	99.9	99.8	100.0
diabetes	77.8	74.3	77.7	77.6	77.7	74.6
german credit	71.3	66.8 ∘	71.1	71.3	70.7	71.4
heart cleveland	81.0	75.6 ∘	81.8	80.7	81.5	81.5
ionosphere	93.3	91.1	94.3	94.6	94.1	95.0
mushroom	100.0	99.9	100.0	100.0	100.0	100.0
pen digits	99.4	98.2 ∘	98.3 ∘	99.4	99.4	99.4
primary tumor	73.0	50.3 ∘	52.7 ∘	69.1 ∘	71.3	69.9
segment	99.4	98.6 ∘	98.8 ∘	99.3	99.4	99.5
soybean	96.9	87.8 ∘	87.8 ∘	95.6 ∘	96.1 ∘	96.0 ∘
splice	98.5	97.9 ∘	98.3	98.6	98.6	98.4
thyroid	99.6	99.2	99.3	99.6	99.4	99.6
vehicle	88.9	84.6 ∘	86.3 ∘	88.4	88.4	87.6
voting	97.7	96.8	97.7	97.7	97.0	98.0
vowel	92.9	83.6 ∘	84.5 ∘	91.5 ∘	92.2	92.2
yeast	79.7	71.7 ∘	72.9 ∘	79.5	78.8	75.5 ∘
biodegradability	77.6	54.0 ∘	74.1	73.8	77.4	79.0
cora	91.8	86.7 ∘	87.9 ∘	92.4	91.4	91.1
diterpenes	85.9	67.6 ∘	68.8 ∘	81.8 ∘	85.3	85.2
gene	87.0	51.7 ∘	55.4 ∘	87.0	86.1 ∘	85.6 ∘
hiv	75.5	54.7 ∘	67.8 ∘	74.8	52.4 ∘	76.3
mutagenesis	78.8	57.1 ∘	75.2	63.7 ∘	72.0	80.1
average	88.9	80.7	84.3	87.7	87.4	88.7
w/t/l		0/0/26	5/1/20	6/1/19	5/0/21	14/1/11
signif w/t/l		0/8/18	0/14/12	0/21/5	0/23/3	2/21/3
avg rank	2.21	5.90	4.08	3.13	3.35	2.33

2.2.2 RMSE

Table 21: RMSE for the six pruning criteria with gain ratio as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	0.352	0.374	0.364	0.358	0.351	0.408 ○
australian credit	0.328	0.326	0.328	0.330	0.340	0.365 ○
balance scale	0.443	0.471 ○	0.450	0.444	0.430	0.419 ●
breast wisconsin	0.199	0.206	0.195	0.200	0.195	0.201
chess kr-kp	0.080	0.106 ○	0.077	0.079	0.076	0.068
diabetes	0.423	0.426	0.424	0.424	0.436	0.464 ○
german credit	0.439	0.438	0.436	0.439	0.464 ○	0.479 ○
heart cleveland	0.422	0.438	0.422	0.424	0.434	0.443
ionosphere	0.235	0.253	0.232	0.236	0.229	0.234
mushroom	0.011	0.022	0.011	0.011	0.014	0.011
pen digits	0.277	0.341 ○	0.333 ○	0.289 ○	0.274 ●	0.292 ○
primary tumor	0.857	0.900 ○	0.894 ○	0.857	0.876 ○	0.911 ○
segment	0.236	0.266 ○	0.258 ○	0.248 ○	0.238	0.252 ○
soybean	0.558	0.674 ○	0.674 ○	0.567	0.564	0.609 ○
splice	0.221	0.228	0.224	0.223	0.229	0.253 ○
thyroid	0.115	0.123	0.115	0.115	0.115	0.145 ○
vehicle	0.481	0.521 ○	0.502 ○	0.486	0.512 ○	0.536 ○
voting	0.197	0.195	0.202	0.201	0.201	0.207
vowel	0.645	0.754 ○	0.747 ○	0.667 ○	0.645	0.650
yeast	0.655	0.668	0.662	0.641 ●	0.670 ○	0.779 ○
biodegradability	0.437	0.490 ○	0.452	0.447	0.451	0.450
cora	0.507	0.560 ○	0.538	0.506	0.511	0.599 ○
diterpenes	0.695	0.727 ○	0.723 ○	0.678 ●	0.686 ●	0.733 ○
gene	0.616	0.755 ○	0.733 ○	0.615	0.632 ○	0.656 ○
hiv	0.174	0.183 ○	0.178 ○	0.176 ○	0.184 ○	0.189 ○
mutagenesis	0.424	0.470 ○	0.433	0.46 ○	0.441	0.422
average	0.386	0.420	0.408	0.389	0.392	0.414
w/t/l		3/0/23	5/1/20	7/1/18	9/0/17	4/1/21
signif w/t/l		0/12/14	0/17/9	2/19/5	2/18/6	1/9/16
avg rank	2.13	4.87	3.58	2.79	3.04	4.60

2.2.3 CLL

Table 22: CLL for the six pruning criteria with gain ratio as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	0.589	0.666 ∘	0.632	0.601	0.606	0.732 ∘
australian credit	0.518	0.515	0.519	0.53	0.549	0.630 ∘
balance scale	0.878	0.997 ∘	0.903	0.872	0.863	0.755 •
breast wisconsin	0.222	0.253	0.214	0.224	0.227	0.220
chess kr-kp	0.045	0.068 ∘	0.043	0.045	0.046	0.035
diabetes	0.781	0.782	0.784	0.782	0.823	0.928 ∘
german credit	0.831	0.812	0.820	0.836	0.935 ∘	1.019 ∘
heart cleveland	0.809	0.843	0.814	0.818	0.856	0.894
ionosphere	0.335	0.365	0.330	0.340	0.313	0.337
mushroom	0.002	0.007	0.002	0.002	0.003	0.002
pen digits	0.389	0.619 ∘	0.593 ∘	0.407 ∘	0.388	0.418 ∘
primary tumor	3.295	3.728 ∘	3.658 ∘	3.241	3.542 ∘	3.824 ∘
segment	0.293	0.391 ∘	0.371 ∘	0.308	0.303	0.321 ∘
soybean	1.346	1.868 ∘	1.868 ∘	1.384	1.435 ∘	1.635 ∘
splice	0.271	0.293 ∘	0.283	0.270	0.284	0.325 ∘
thyroid	0.074	0.085 ∘	0.079	0.072	0.076	0.096 ∘
vehicle	1.010	1.162 ∘	1.088	1.010	1.130 ∘	1.209 ∘
voting	0.227	0.235	0.234	0.234	0.242	0.234
vowel	1.741	2.327 ∘	2.284 ∘	1.848 ∘	1.747	1.761
yeast	1.800	1.892	1.864	1.734 •	1.902 ∘	2.432 ∘
biodegradability	0.826	0.975 ∘	0.885	0.868	0.884	0.918
cora	1.263	1.506 ∘	1.428 ∘	1.247	1.307	1.587 ∘
diterpenes	2.018	2.298 ∘	2.274 ∘	1.988	2.003	2.215 ∘
gene	1.673	2.463 ∘	2.348 ∘	1.657	1.787 ∘	1.901 ∘
hiv	0.186	0.214 ∘	0.199 ∘	0.191 ∘	0.217 ∘	0.212 ∘
mutagenesis	0.774	0.905 ∘	0.801	0.878 ∘	0.823	0.770
average	0.854	1.010	0.974	0.861	0.896	0.977
w/t/l		2/0/24	4/1/21	9/1/16	4/0/22	4/1/21
signif w/t/l		0/9/17	0/17/9	1/21/4	0/19/7	1/9/16
avg rank	1.94	4.98	3.58	2.56	3.65	4.29

2.2.4 Classification Accuracy

Table 23: Accuracy for the six pruning criteria with gain ratio as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	86.1	83.1	83.8	84.3	87.9	86.5
australian credit	84.5	85.9	84.8	84.7	84.5	81.1 ∘
balance scale	77.1	74.0	75.6	76.1	78.4	78.0
breast wisconsin	95.0	94.9	95.4	95.0	95.6	94.4
chess kr-kp	99.3	98.5 ∘	99.4	99.4	99.4	99.5
diabetes	72.7	73.1	72.6	72.5	73.3	68.4
german credit	70.8	69.8	71.2	71.2	70.8	68.2
heart cleveland	74.6	73.3	75.0	74.1	74.9	72.4
ionosphere	93.8	92.9	93.7	93.3	93.8	93.1
mushroom	100.0	99.9	100.0	100.0	100.0	100.0
pen digits	95.4	88.2 ∘	88.8 ∘	93.8 ∘	96.1 •	96.1 •
primary tumor	37.3	24.3 ∘	25.3 ∘	33.6	36.3	33.1
segment	95.3	92.3 ∘	92.6 ∘	93.6 ∘	95.8	95.8
soybean	81.2	54.4 ∘	54.4 ∘	75.8 ∘	82.7	80.4
splice	94.4	94.3	94.3	94.1	94.0	92.4 ∘
thyroid	98.7	98.5	98.7	98.6	98.7	98.1 ∘
vehicle	69.9	62.1 ∘	65.0	67.4	71.1	70.2
voting	94.6	95.3	94.5	94.7	94.9	93.8
vowel	71.4	36.5 ∘	38.2 ∘	56.8 ∘	80.9 •	82.2 •
yeast	55.9	46.7 ∘	47.9 ∘	55.1	56.3	52.2 ∘
biodegradability	72.9	59.4 ∘	70.3	72.0	71.8	71.7
cora	78.3	70.9 ∘	73.8	78.2	77.1	72.6 ∘
diterpenes	59.8	42.7 ∘	43.4 ∘	53.7 ∘	60.8	60.4
gene	64.7	45.2 ∘	48.1 ∘	65.0	63.3	61.9 ∘
hiv	94.2	96.4	96.5	96.4	96.5	95.7
mutagenesis	70.8	65.6	69.8	67.0	69.6	72.2
average	80.3	73.8	75.1	78.7	80.9	79.6
w/t/l		4/0/22	7/1/18	6/1/19	15/0/11	10/1/15
signif w/t/l		0/14/12	0/18/8	0/21/5	2/24/0	2/18/6
avg rank	2.69	5.02	3.69	3.52	2.25	3.83

2.2.5 Tree Size

Table 24: Tree size for the six pruning criteria with gain ratio as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	21.8	6.5 •	7.6 •	11.8 •	20.3	78.1 ◦
australian credit	7.8	2.5 •	6.8	10.5	32.5 ◦	101.5 ◦
balance scale	20.5	8.9 •	14.4 •	15.2 •	52.3 ◦	122.8 ◦
breast wisconsin	7.9	3.6 •	11.8 ◦	9.5	10.3	29.5 ◦
chess kr-kp	28.6	19.4 •	29.0	27.3	28.8	42.8 ◦
diabetes	14.5	4.0 •	12.8	12.6	49.4 ◦	192.6 ◦
german credit	24.4	1.5 •	19.3	32.5	108.5 ◦	200.3 ◦
heart cleveland	8.2	2.2 •	13.3 ◦	7.8	21.4 ◦	53.9 ◦
ionosphere	7.8	4.3 •	10.6	11.9 ◦	10.2	22.0 ◦
mushroom	11.8	10.7	11.8	11.8	11.6	11.8
pen digits	164.6	36.5 •	40.1 •	133.9 •	223.5 ◦	301.3 ◦
primary tumor	18.6	0.1 •	0.5 •	4.6 •	81.0 ◦	144.0 ◦
segment	42.7	12.4 •	13.6 •	33.5 •	61.3 ◦	91.7 ◦
soybean	31.3	7.3 •	7.3 •	17.9 •	62.3 ◦	110.1 ◦
splice	34.2	13.8 •	21.2 •	43.6 ◦	78.1 ◦	127.4 ◦
thyroid	9.5	3.3 •	4.4 •	10.5	10.5	63.3 ◦
vehicle	41.7	10.7 •	14.9 •	28.8 •	117.4 ◦	199.3 ◦
voting	4.1	1.9 •	5.5	5.6	6.9	24.1 ◦
vowel	95.8	7.9 •	9.0 •	43.3 •	158.2 ◦	176.6 ◦
yeast	50.6	5.2 •	6.0 •	25.4 •	139.3 ◦	432.1 ◦
biodegradability	24.0	0.6 •	24.9	22.6	40.2 ◦	78.6 ◦
cora	30.7	6.5 •	8.2 •	32.3	83.5 ◦	177.5 ◦
diterpenes	52.9	2.3 •	2.6 •	15.4 •	64.2 ◦	114.6 ◦
gene	79.1	1.1 •	4.3 •	73.1	199.2 ◦	287.5 ◦
hiv	469.9	4.0 •	66.7 •	437.1	19.8 •	2275.2 ◦
mutagenesis	8.1	0.9 •	7.3	3.7 •	7.3	13.4 ◦
average	50.4	6.9	14.4	41.6	65.3	210.5
w/t/l		26/0/0	19/1/6	17/1/8	4/0/22	0/1/25
signif w/t/l		25/1/0	15/9/2	11/13/2	1/8/17	0/1/25
avg rank	3.60	1.02	2.69	3.21	4.54	5.94

2.3 Results for the Splitting Criterion Gini Index

In this section we report in detail the results obtained with gini index as the splitting criterion.

2.3.1 AUC

Table 25: AUC for the six pruning criteria with gini index as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	88.8	82.9 ∘	83.3 ∘	87.3	85.7 ∘	90.7
australian credit	90.9	90.5	91.5	91.0	91.1	89.8
balance scale	87.6	83.3 ∘	85.3 ∘	85.6 ∘	87.1	91.6 •
breast wisconsin	97.1	96.1	97.6	97.1	97.0	97.8
chess kr-kp	99.8	99.6	99.8	99.8	99.7	99.9
diabetes	78.3	76.6	78.4	77.8	77.5	76.1
german credit	72.0	68.2 ∘	71.6	71.2	70.9	70.6
heart cleveland	83.6	80.3	84.1	83.0	83.5	83.4
ionosphere	92.2	91.9	93.5	93.1	93.3	93.6
mushroom	100.0	99.9	100.0	100.0	100.0	100.0
pen digits	99.4	98.5 ∘	98.6 ∘	99.3	99.3	99.4
primary tumor	72.9	50.0 ∘	50.2 ∘	71.6	74.1	71.5
segment	99.4	98.7 ∘	98.9 ∘	99.4	99.4	99.5
soybean	96.3	87.0 ∘	88.2 ∘	95.5 ∘	95.3 ∘	95.2 ∘
splice	98.5	98.1 ∘	98.3	98.5	98.4	98.2
thyroid	99.6	99.1	99.2	99.6	99.3	99.6
vehicle	88.9	85.0 ∘	87.6	88.9	88.9	88.3
voting	97.9	97.3	97.9	98.0	97.6	98.1
vowel	92.7	87.0 ∘	88.2 ∘	92.3	91.9	92.1
yeast	79.3	76.9 ∘	77.0 ∘	80.0	79.0	75.7 ∘
biodegradability	73.5	67.5	75.6	73.4	77.2	78.7
cora	91.6	88.6 ∘	89.0 ∘	92.1	91.5	90.4
diterpenes	85.3	68.0 ∘	69.7 ∘	82.1 ∘	84.8	84.7
gene	86.4	79.2 ∘	81.6 ∘	87.2	86.2	85.6
hiv	74.6	69.1 ∘	72.9 ∘	74.8	64.5 ∘	76.2 •
mutagenesis	71.4	57.2 ∘	73.9	60.8 ∘	72.1	79.0 •
average	88.4	83.7	85.8	87.7	87.9	88.7
w/t/l		0/0/26	8/1/17	11/1/14	5/0/21	12/1/13
signif w/t/l		0/10/16	0/14/12	0/22/4	0/23/3	3/21/2
avg rank	2.44	5.85	3.63	2.83	3.42	2.83

2.3.2 RMSE

Table 26: RMSE for the six pruning criteria with gini index as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	0.346	0.368 ∘	0.364	0.357	0.351	0.406 ∘
australian credit	0.324	0.324	0.325	0.327	0.339	0.371 ∘
balance scale	0.447	0.468	0.456	0.456	0.437	0.428
breast wisconsin	0.205	0.217	0.196	0.208	0.199	0.201
chess kr-kp	0.083	0.113 ∘	0.081	0.083	0.082	0.071
diabetes	0.422	0.423	0.423	0.421	0.436	0.458 ∘
german credit	0.435	0.436	0.445	0.440	0.477 ∘	0.490 ∘
heart cleveland	0.400	0.416	0.404	0.407	0.419	0.431
ionosphere	0.288	0.279	0.276	0.288	0.274	0.281
mushroom	0.009	0.021	0.009	0.009	0.012	0.009
pen digits	0.273	0.328 ∘	0.320 ∘	0.284 ∘	0.269 •	0.286 ∘
primary tumor	0.844	0.900 ∘	0.900 ∘	0.841	0.855 ∘	0.905 ∘
segment	0.233	0.262 ∘	0.254 ∘	0.245 ∘	0.234	0.248 ∘
soybean	0.577	0.696 ∘	0.688 ∘	0.59 ∘	0.584	0.622 ∘
splice	0.223	0.222	0.222	0.226	0.235 ∘	0.256 ∘
thyroid	0.111	0.121 ∘	0.113	0.113	0.112	0.141 ∘
vehicle	0.476	0.519 ∘	0.493 ∘	0.484	0.503 ∘	0.526 ∘
voting	0.192	0.190	0.198	0.196	0.193	0.207
vowel	0.625	0.723 ∘	0.710 ∘	0.650 ∘	0.622	0.630
yeast	0.646	0.646	0.646	0.643	0.661 ∘	0.773 ∘
biodegradability	0.454	0.463	0.457	0.458	0.458	0.457
cora	0.506	0.530 ∘	0.524 ∘	0.514	0.522 ∘	0.593 ∘
diterpenes	0.696	0.733 ∘	0.725 ∘	0.685 •	0.691	0.734 ∘
gene	0.623	0.666 ∘	0.655 ∘	0.624	0.636 ∘	0.655 ∘
hiv	0.172	0.177 ∘	0.174	0.175 ∘	0.177 ∘	0.186 ∘
mutagenesis	0.449	0.476 ∘	0.441	0.471 ∘	0.446	0.426 •
average	0.387	0.412	0.404	0.392	0.393	0.415
w/t/l		4/0/22	6/1/19	5/1/20	8/0/18	5/1/20
signif w/t/l		0/12/14	0/17/9	1/19/6	1/17/8	1/9/16
avg rank	2.13	4.69	3.25	3.17	3.19	4.56

2.3.3 CLL

Table 27: CLL for the six pruning criteria with gini index as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	0.580	0.659 ∘	0.650 ∘	0.606	0.609	0.726 ∘
australian credit	0.511	0.509	0.516	0.530	0.552	0.654 ∘
balance scale	0.913	0.990	0.962	0.955	0.916	0.795 •
breast wisconsin	0.242	0.270	0.225	0.247	0.236	0.228
chess kr-kp	0.053	0.079 ∘	0.048	0.052	0.055	0.040
diabetes	0.771	0.775	0.777	0.770	0.826 ∘	0.913 ∘
german credit	0.810	0.806	0.860	0.858	0.994 ∘	1.088 ∘
heart cleveland	0.738	0.775	0.759	0.762	0.805	0.850
ionosphere	0.436	0.412	0.427	0.442	0.415	0.440
mushroom	0.002	0.008	0.002	0.002	0.003	0.002
pen digits	0.387	0.580 ∘	0.550 ∘	0.406 ∘	0.386	0.412 ∘
primary tumor	3.260	3.738 ∘	3.734 ∘	3.251	3.405 ∘	3.759 ∘
segment	0.291	0.381 ∘	0.359 ∘	0.307	0.298	0.314 ∘
soybean	1.441	1.988 ∘	1.939 ∘	1.495	1.538 ∘	1.706 ∘
splice	0.277	0.281	0.280	0.281	0.300 ∘	0.337 ∘
thyroid	0.071	0.084 ∘	0.076	0.070	0.074	0.093 ∘
vehicle	0.995	1.155 ∘	1.050 ∘	1.006	1.106 ∘	1.177 ∘
voting	0.211	0.221	0.225	0.221	0.222	0.230
vowel	1.684	2.168 ∘	2.088 ∘	1.786 ∘	1.675	1.693
yeast	1.788	1.821	1.820	1.764	1.872 ∘	2.404 ∘
biodegradability	0.879	0.895	0.940	0.922	0.951	0.999
cora	1.291	1.435 ∘	1.406 ∘	1.294	1.356	1.590 ∘
diterpenes	2.022	2.318 ∘	2.263 ∘	1.999	2.024	2.218 ∘
gene	1.709	1.907 ∘	1.856 ∘	1.712	1.809 ∘	1.890 ∘
hiv	0.185	0.196 ∘	0.188 ∘	0.189 ∘	0.199 ∘	0.208 ∘
mutagenesis	0.849	0.925 ∘	0.826	0.914	0.843	0.781 •
average	0.861	0.976	0.955	0.879	0.903	0.983
w/t/l		3/0/23	4/1/21	6/1/19	5/0/21	4/1/21
signif w/t/l		0/12/14	0/15/11	0/23/3	0/17/9	2/8/16
avg rank	1.90	4.46	3.56	2.83	3.62	4.63

2.3.4 Classification Accuracy

Table 28: Accuracy for the six pruning criteria with gini index as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	86.4	83.0	83.6	85.0	87.4	86.2
australian credit	84.8	85.9	85.0	84.4	85.1	80.9
balance scale	77.8	75.5	77.2	76.8	78.3	77.4
breast wisconsin	95.1	94.5	95.3	94.8	95.4	94.7
chess kr-kp	99.3	98.2 ∘	99.4	99.3	99.3	99.5
diabetes	72.6	73.2	73.1	72.7	73.8	69.4
german credit	71.3	69.7	70.8	71.7	70.2	68.3
heart cleveland	77.9	73.6	77.2	75.8	77.0	74.2
ionosphere	89.2	90.8	90.6	89.3	91.0	90.0
mushroom	100.0	99.9	100.0	100.0	100.0	100.0
pen digits	95.3	89.3 ∘	89.8 ∘	93.8 ∘	96.1 •	96.0 •
primary tumor	37.4	24.4 ∘	24.4 ∘	36.3	38.1	34.7
segment	95.2	92.3 ∘	92.8 ∘	94.0 ∘	96.0	96.0
soybean	79.4	51.0 ∘	52.5 ∘	73.4 ∘	80.3	78.8
splice	94.4	94.6	94.5	94.1	93.7	92.3 ∘
thyroid	98.8	98.5	98.8	98.7	98.7	98.2 ∘
vehicle	70.1	61.6 ∘	67.3	68.7	71.9	71.2
voting	95.0	95.4	94.8	95.1	95.1	93.9
vowel	71.8	45.0 ∘	47.2 ∘	60.8 ∘	80.8 •	82.3 •
yeast	56.7	55.3	55.1	56.7	56.8	52.5 ∘
biodegradability	69.1	68.0	71.5	68.3	73.0	73.3
cora	78.2	74.1 ∘	75.0 ∘	77.8	75.8	71.8 ∘
diterpenes	58.7	41.3 ∘	42.8 ∘	53.6 ∘	59.7	59.4
gene	63.5	55.1 ∘	57.2 ∘	63.6	62.0	60.7
hiv	96.6	96.5	96.6	96.5 ∘	96.6	95.9 ∘
mutagenesis	68.0	63.8	68.1	65.8	69.8	71.5
average	80.1	75.0	76.2	78.7	80.8	79.6
w/t/l		5/0/21	8/1/17	7/1/18	19/0/7	9/1/16
signif w/t/l		0/16/10	0/18/8	0/20/6	2/24/0	2/19/5
avg rank	2.90	4.83	3.73	3.67	2.04	3.83

2.3.5 Tree Size

Table 29: Tree size for the six pruning criteria with gini index as the splitting criterion.

	RAND	MDL	BIC	CHI	EBP	NOPRUNING
annealing	17.4	4.9 •	5.5 •	13.3	18.5	77.5 ○
australian credit	5.0	2.2 •	10.9 ○	10.9 ○	27.7 ○	92.0 ○
balance scale	24.5	8.2 •	13.9 •	14.2 •	52.1 ○	127.6 ○
breast wisconsin	7.3	3.8 •	11.0 ○	7.5	11.2 ○	27.4 ○
chess kr-kp	30.5	19.2 •	31.6	29.3	30.7	48.0 ○
diabetes	10.0	3.8 •	13.3	8.6	48.6 ○	180.4 ○
german credit	11.0	2.0 •	27.0 ○	25.6 ○	106.2 ○	172.6 ○
heart cleveland	5.9	3.3 •	13.3 ○	6.6	21.3 ○	49.7 ○
ionosphere	6.9	2.4 •	15.0 ○	10.1	14.7 ○	22.4 ○
mushroom	10.9	8.1 •	10.9	10.9	10.7	10.9
pen digits	142.0	32.1 •	35.8 •	111.5 •	211.9 ○	283.2 ○
primary tumor	12.4	0.0 •	0.0 •	4.9 •	76.2 ○	138.7 ○
segment	36.3	11.7 •	13.1 •	29.9 •	56.8 ○	86.4 ○
soybean	30.3	6.4 •	6.9 •	18.8 •	69.0 ○	111.8 ○
splice	33.3	13.9 •	20.7 •	40.8 ○	76.0 ○	121.4 ○
thyroid	7.5	3.1 •	4.2 •	10.0 ○	10.2	61.9 ○
vehicle	25.6	7.6 •	13.7 •	24.9	102.0 ○	176.8 ○
voting	3.4	2.1 •	6.8 ○	5.3 ○	6.9 ○	21.7 ○
vowel	72.4	8.9 •	10.6 •	37.2 •	138.5 ○	161.7 ○
yeast	30.0	5.0 •	5.3 •	30.9	118.2 ○	425.2 ○
biodegradability	4.9	1.9 •	17.3 ○	10.0	34.0 ○	63.2 ○
cora	19.0	4.9 •	5.8 •	31.6 ○	92.9 ○	168.4 ○
diterpenes	41.5	2.0 •	2.7 •	13.9 •	54.2 ○	108.9 ○
gene	27.7	7.7 •	9.4 •	54.8 ○	167.8 ○	239.3 ○
hiv	233.1	26.6 •	59.9 •	241.5	49.2 •	2110.6 ○
mutagenesis	3.8	1.0 •	5.4 ○	1.7 •	6.1	14.1 ○
average	32.8	7.4	14.2	31.0	62.0	196.2
w/t/l		26/0/0	15/1/10	12/1/13	2/0/24	0/1/25
signif w/t/l		26/0/0	15/3/8	8/11/7	1/5/20	0/1/25
avg rank	3.17	1.00	2.94	3.25	4.69	5.94

3 Influence of the Number of Classes

In the paper we discuss some experiments in which we assessed the influence of certain characteristics of the data (number of classes and class skew) on the performance of some pruning criteria. In the paper we only report results partially (i.e., for the relevant pruning criteria, the main performance measures and on one dataset only). In the following sections we give the full results. In addition, we also report the influence of the number of examples (by means of learning curves, see Section 5).

In this section we consider the experiments in which we varied the number of classes in multi-class datasets. We performed this experiments on all datasets that have ten or more classes: 'cora', 'diterpenes', 'gene', 'pen digits', 'soybean', 'vowel' and 'yeast'.[1]

3.1 'Cora' Dataset

We performed an experiment on the 'Cora' dataset in which we varied the number of classes from 10 (as in the original dataset) to 3.

3.1.1 Experimental Setup

For an experiment with C classes, we considered only examples having one of the C most frequent classes, and we computed the class distribution in the original dataset taking into account only these examples. We then sampled examples from the original dataset (without replacement) according to this class distribution. For all values of C we sampled the same total number of examples in order to eliminate any influence of this factor on the results. Concretely, we always sampled 657 examples.[2] For each value of C we performed 10 runs of 5-fold cross-validation with new samples in every run.

[1] We did not use 'primary tumor' (21 classes) since it does not have a sufficient number of examples for this experiment.

[2] The reason for using 657 examples is that this is the number of examples having one of the three most frequent classes in the original dataset. Hence for $C = 3$, 657 is the highest number of distinct examples that we can possibly use (so for $C = 3$ all examples having one of the three classes are selected and no sampling is needed).

3.1.2 AUC

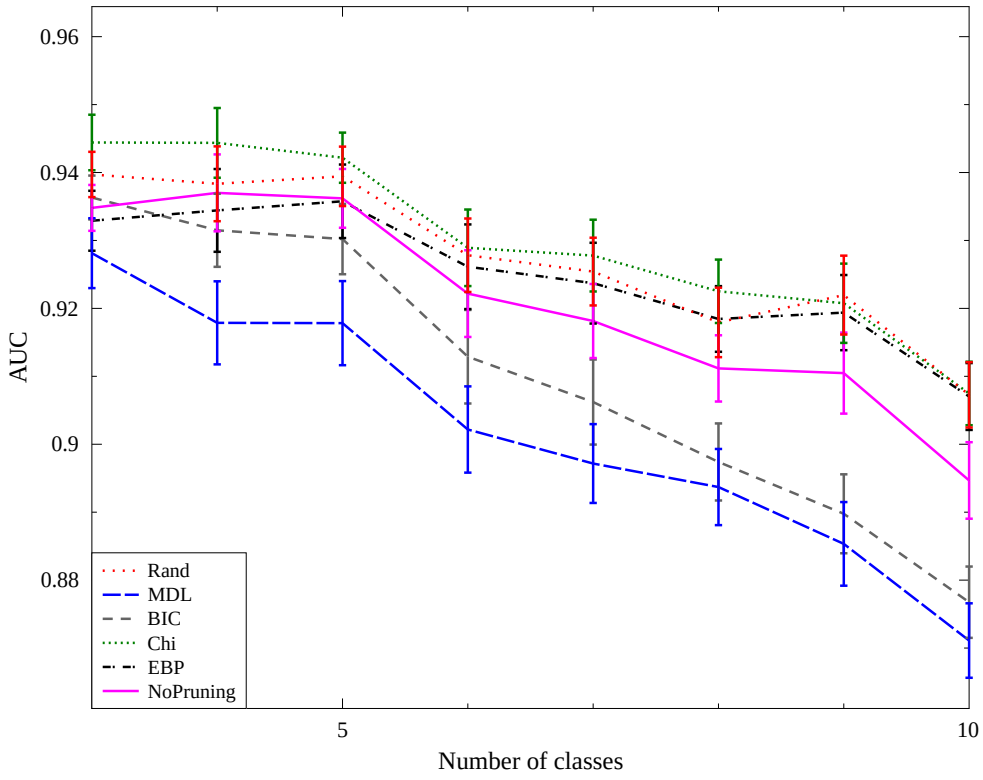

Figure 1: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.1.3 RMSE

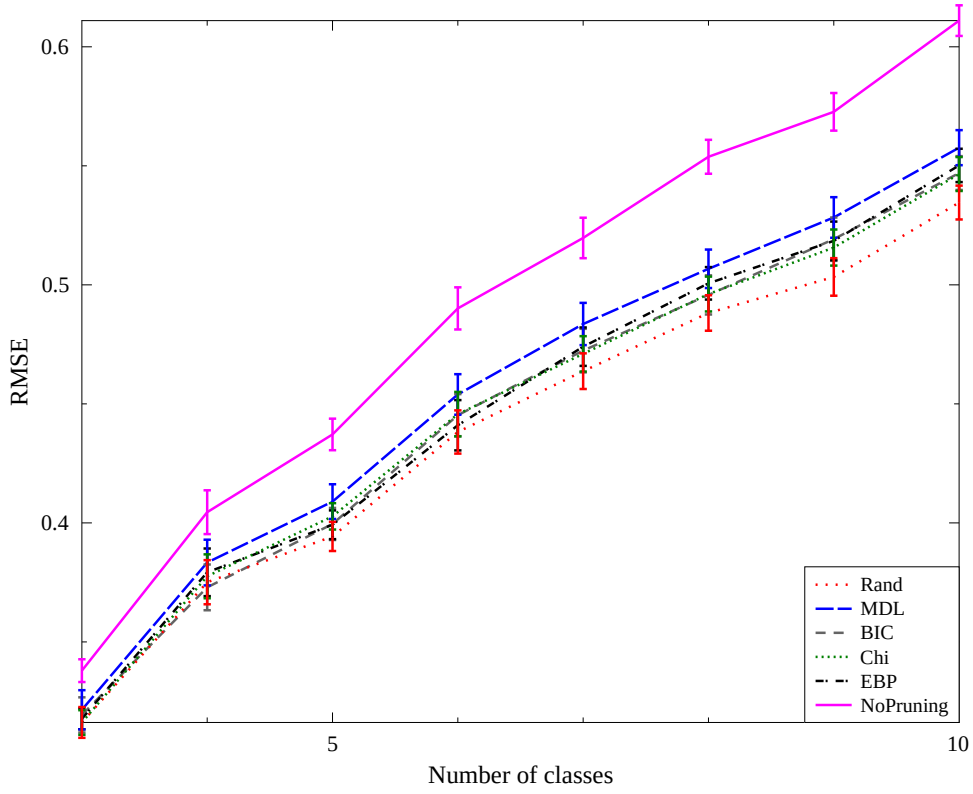

Figure 2: Influence of the number of classes on the performance measure RMSE (lower is better).

3.1.4 CLL

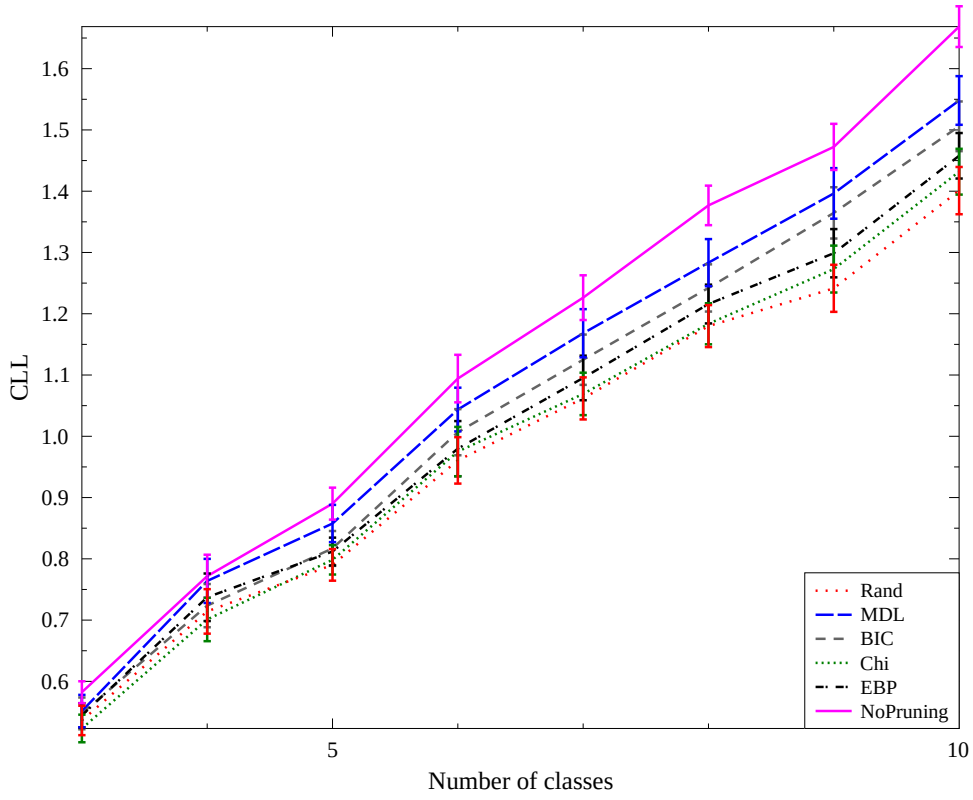

Figure 3: Influence of the number of classes on the performance measure CLL (lower is better).

3.1.5 Calibration Error

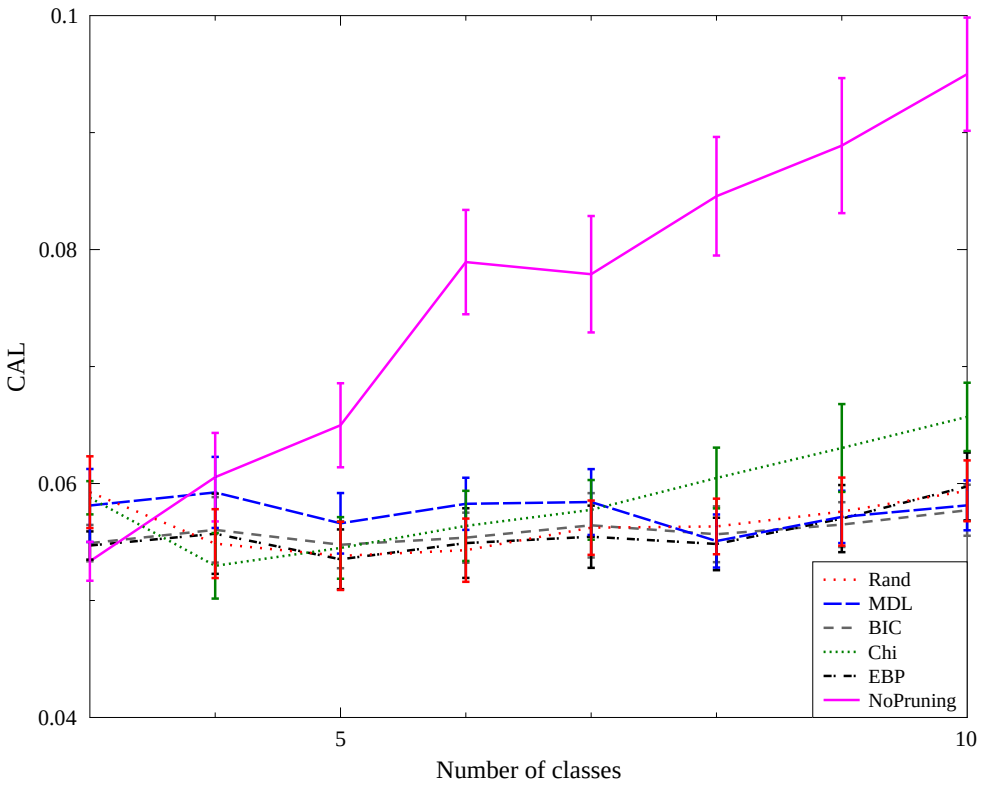

Figure 4: Influence of the number of classes on the performance measure CAL (lower is better).

3.1.6 Classification Accuracy

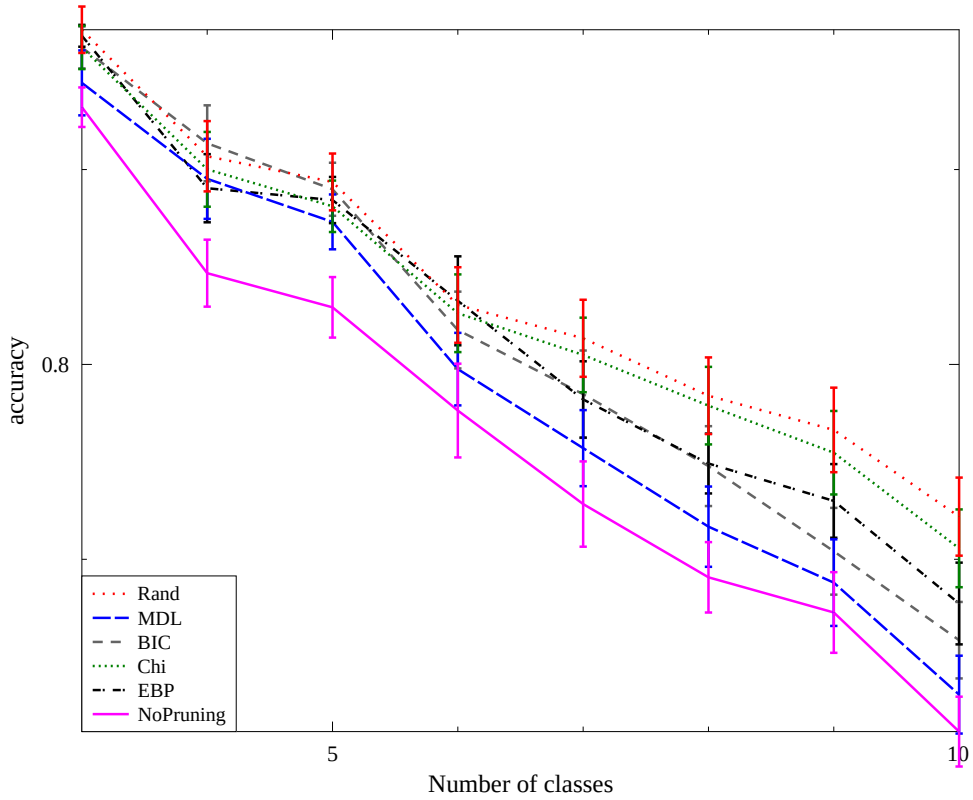

Figure 5: Influence of the number of classes on the performance measure accuracy (higher is better).

3.1.7 Tree Size

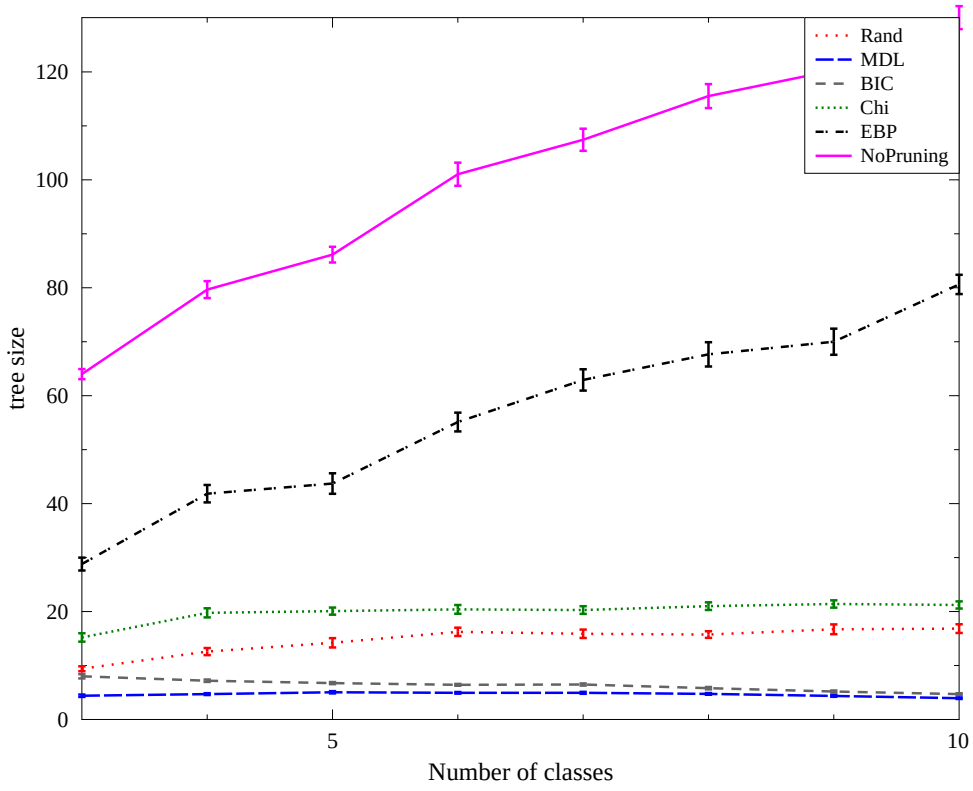

Figure 6: Influence of the number of classes on tree size (lower is better).

3.2 'Diterpenes' Dataset

In the experiment on the 'diterpenes' dataset we varied the number of classes from 23 (as in the original dataset) to 3. We always used 1157 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.2.1 AUC

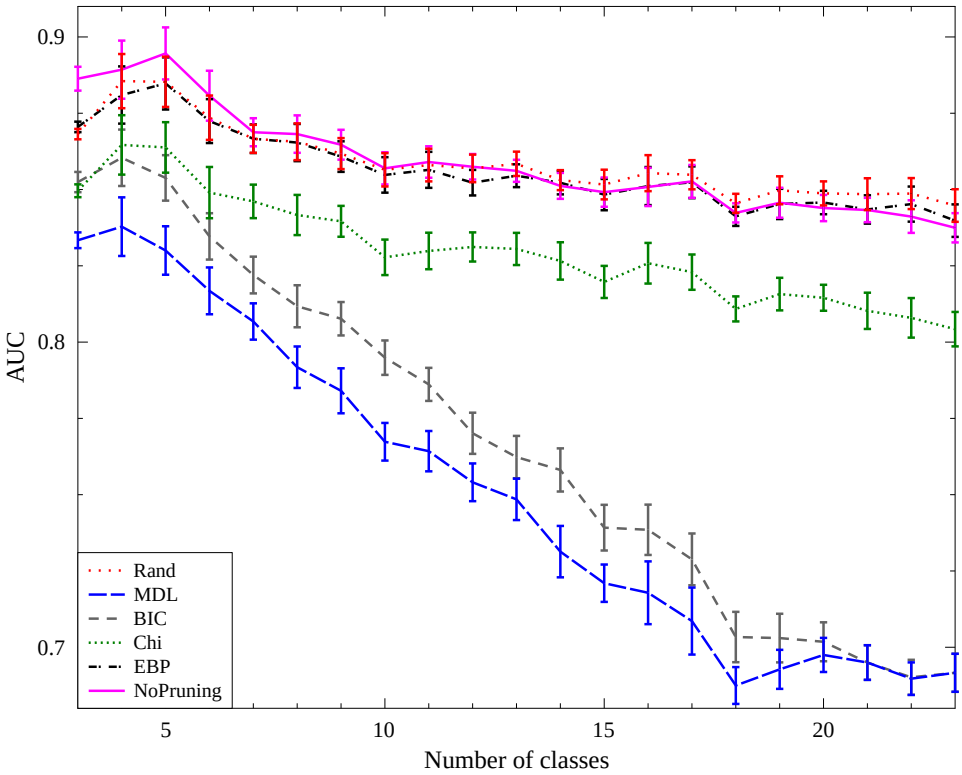

Figure 7: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.2.2 RMSE

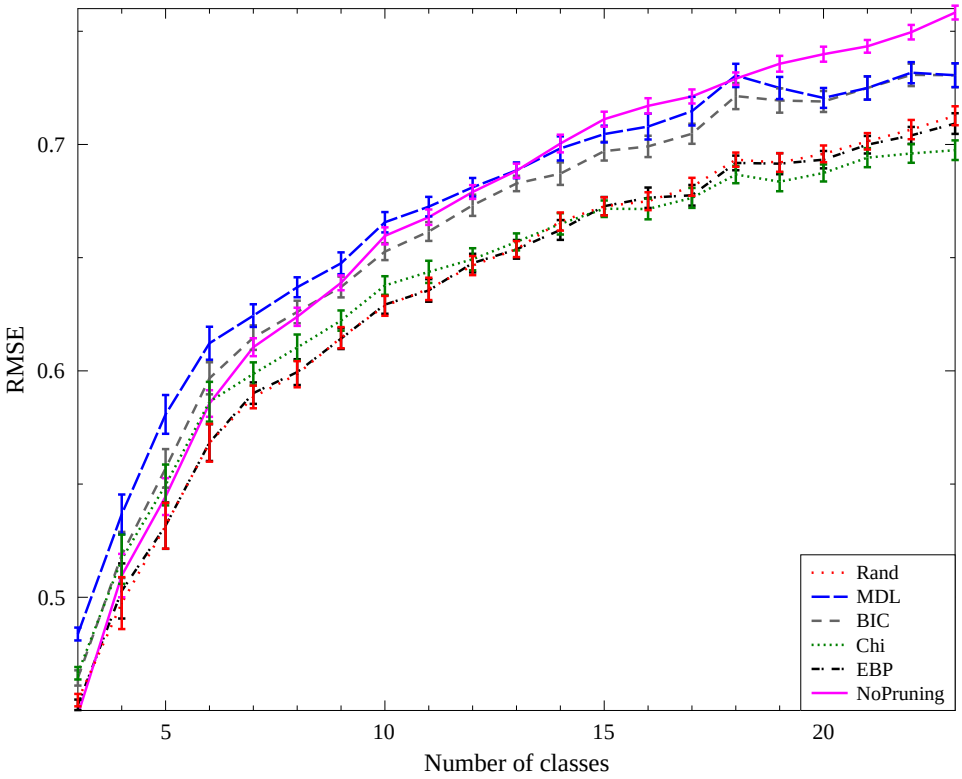

Figure 8: Influence of the number of classes on the performance measure RMSE (lower is better).

3.2.3 CLL

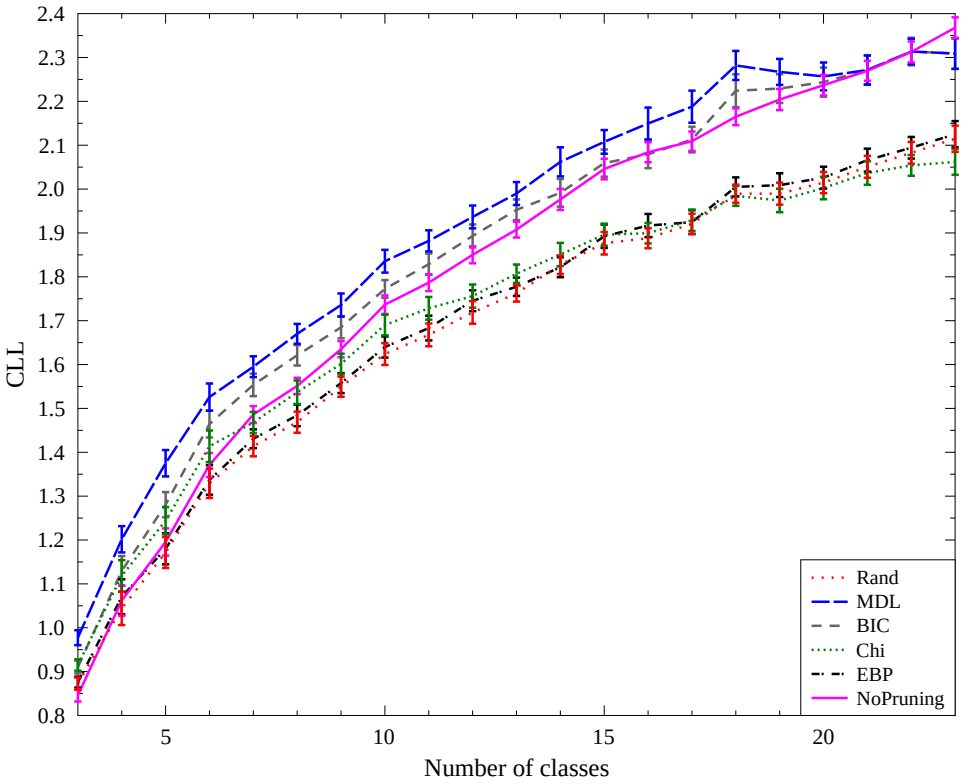

Figure 9: Influence of the number of classes on the performance measure CLL (lower is better).

3.2.4 Calibration Error

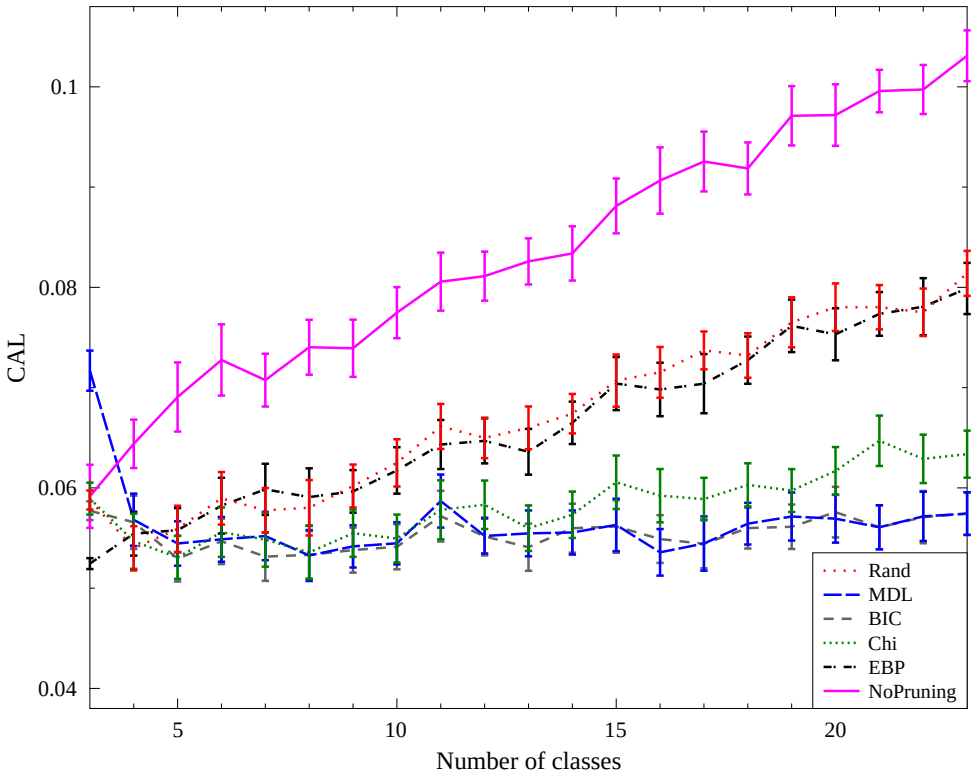

Figure 10: Influence of the number of classes on the performance measure CAL (lower is better).

3.2.5 Classification Accuracy

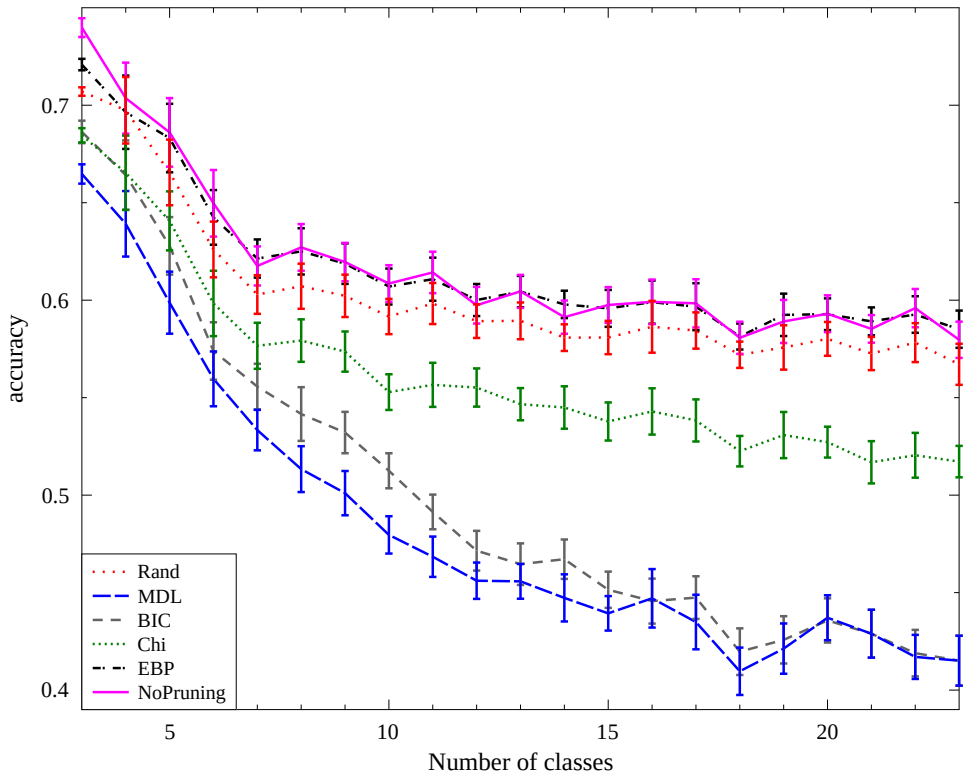

Figure 11: Influence of the number of classes on the performance measure accuracy (higher is better).

3.2.6 Tree Size

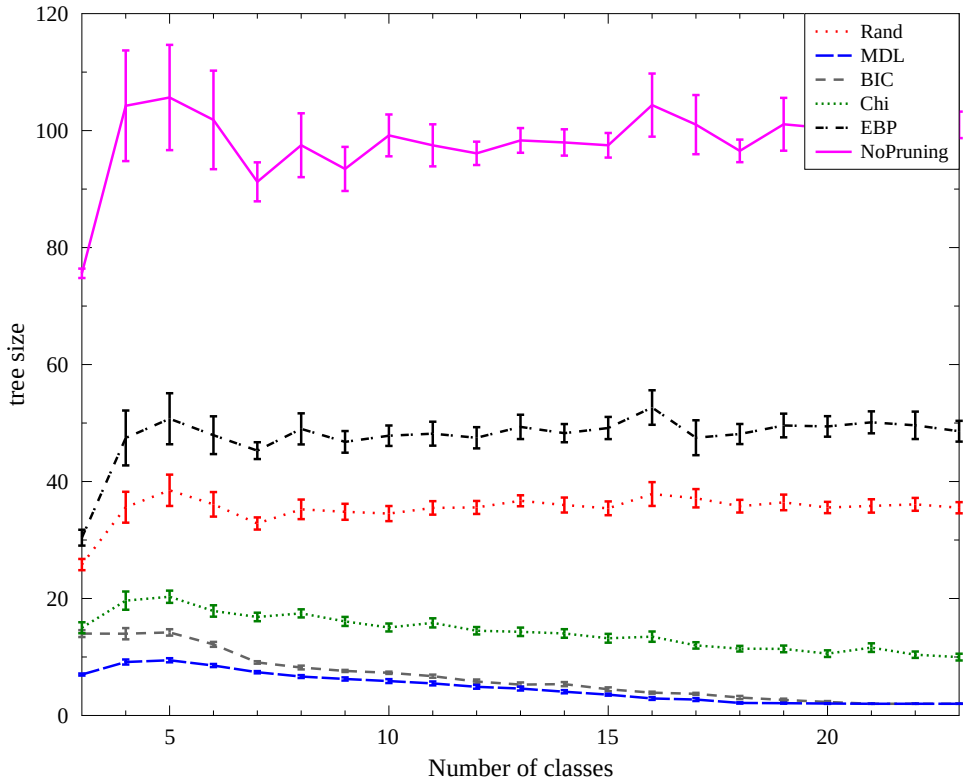

Figure 12: Influence of the number of classes on tree size (lower is better).

3.3 'Gene' Dataset

In the experiment on the 'gene' dataset we varied the number of classes from 10 (as in the original dataset) to 3. We always used 901 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.3.1 AUC

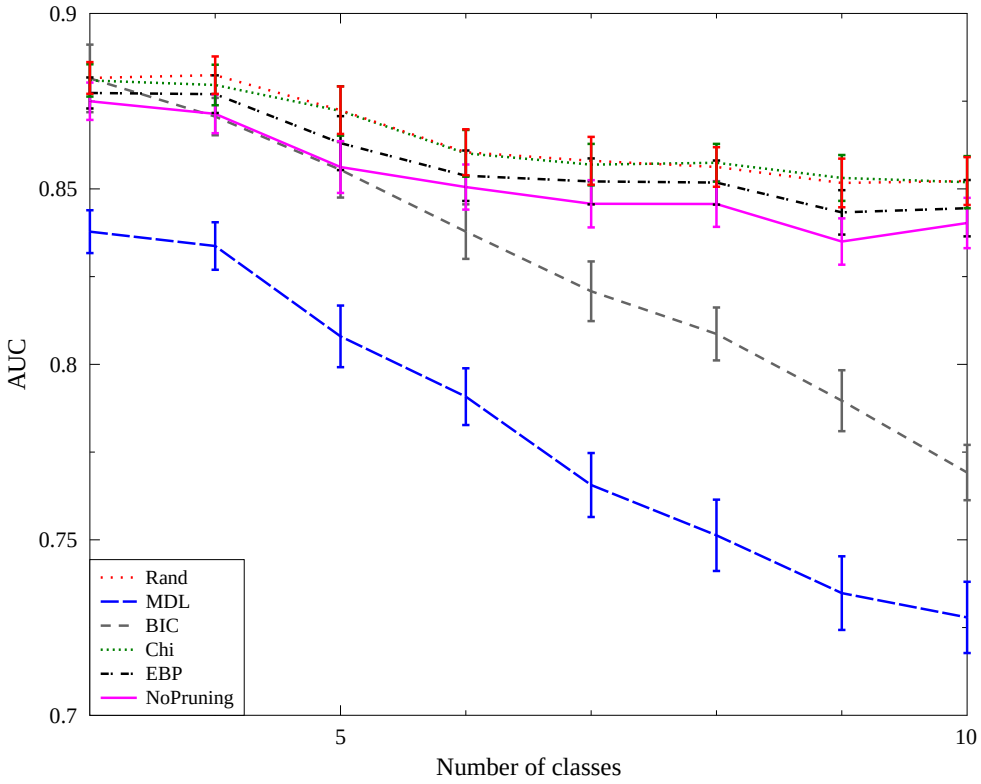

Figure 13: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.3.2 RMSE

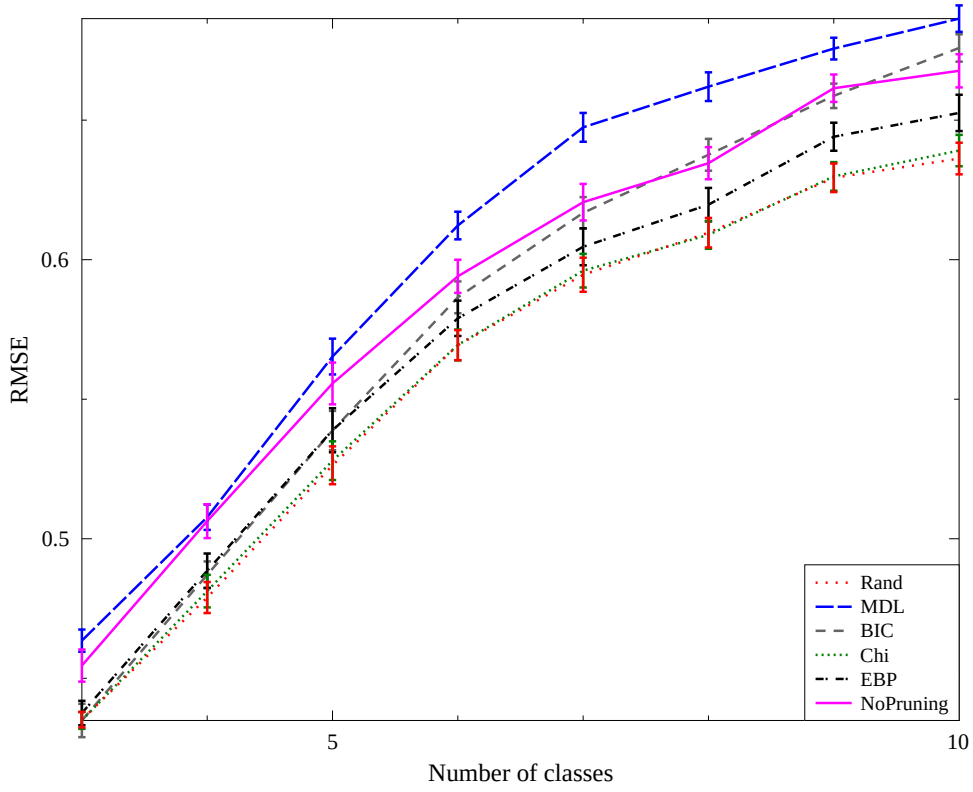

Figure 14: Influence of the number of classes on the performance measure RMSE (lower is better).

3.3.3 CLL

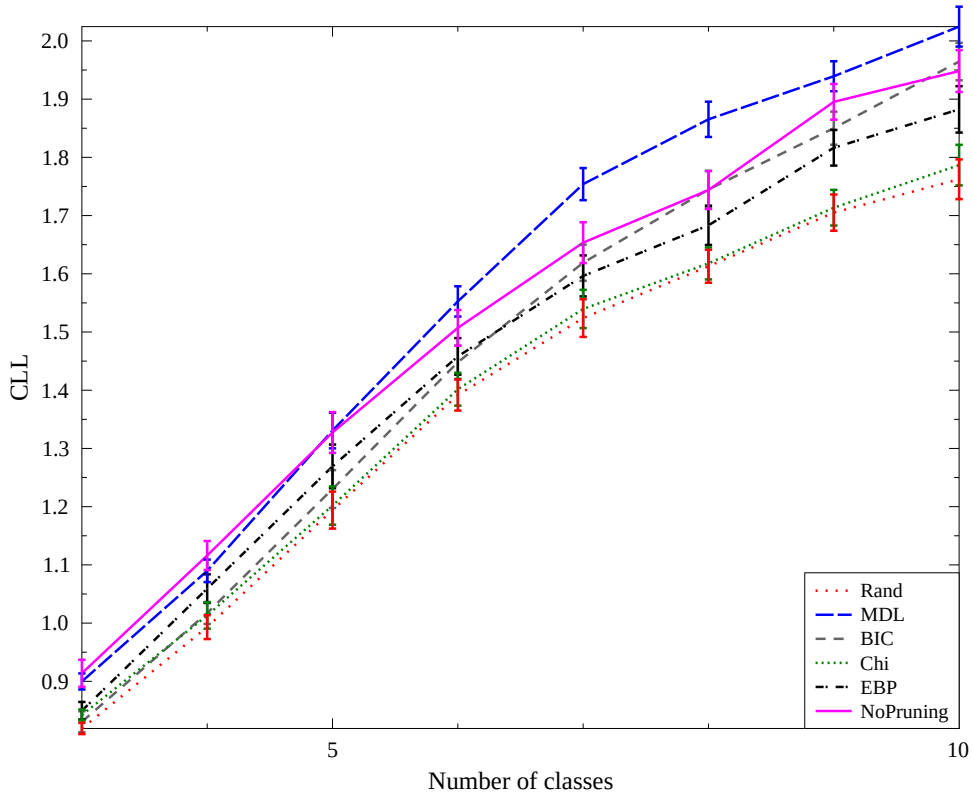

Figure 15: Influence of the number of classes on the performance measure CLL (lower is better).

3.3.4 Calibration Error

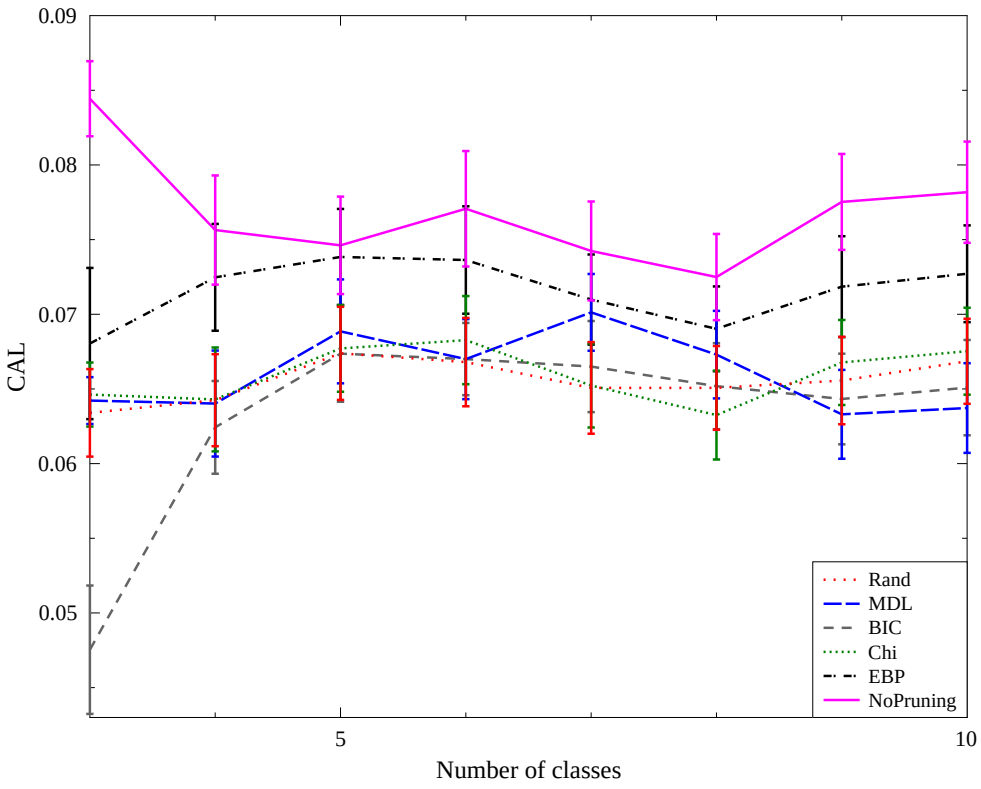

Figure 16: Influence of the number of classes on the performance measure CAL (lower is better).

3.3.5 Classification Accuracy

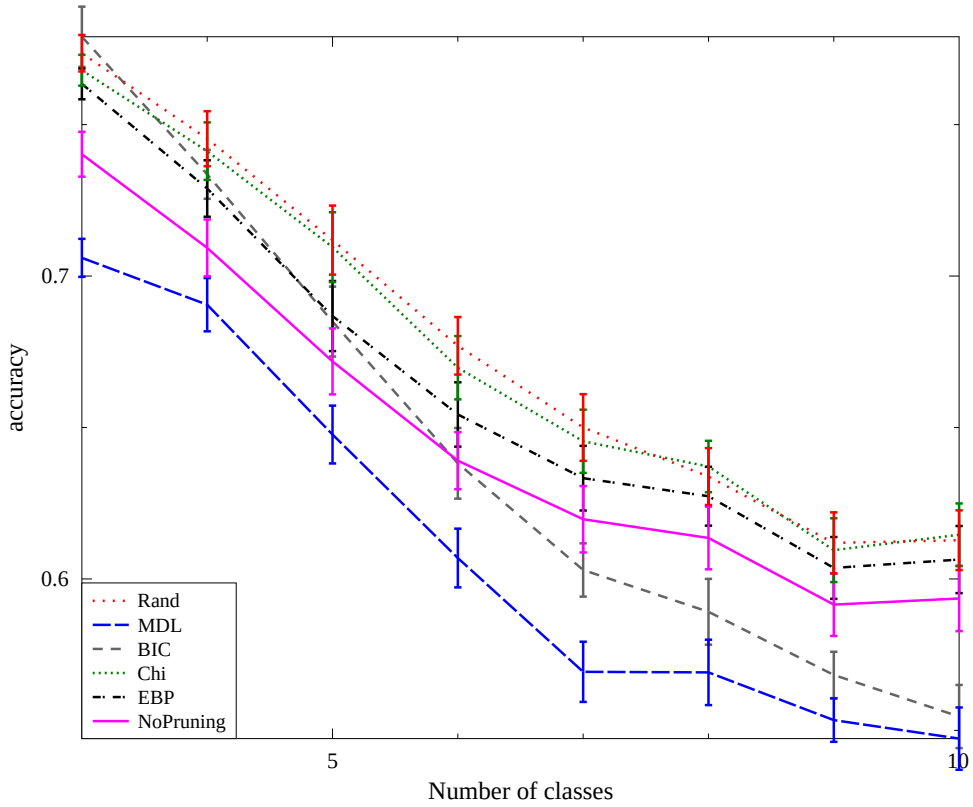

Figure 17: Influence of the number of classes on the performance measure accuracy (higher is better).

3.3.6 Tree Size

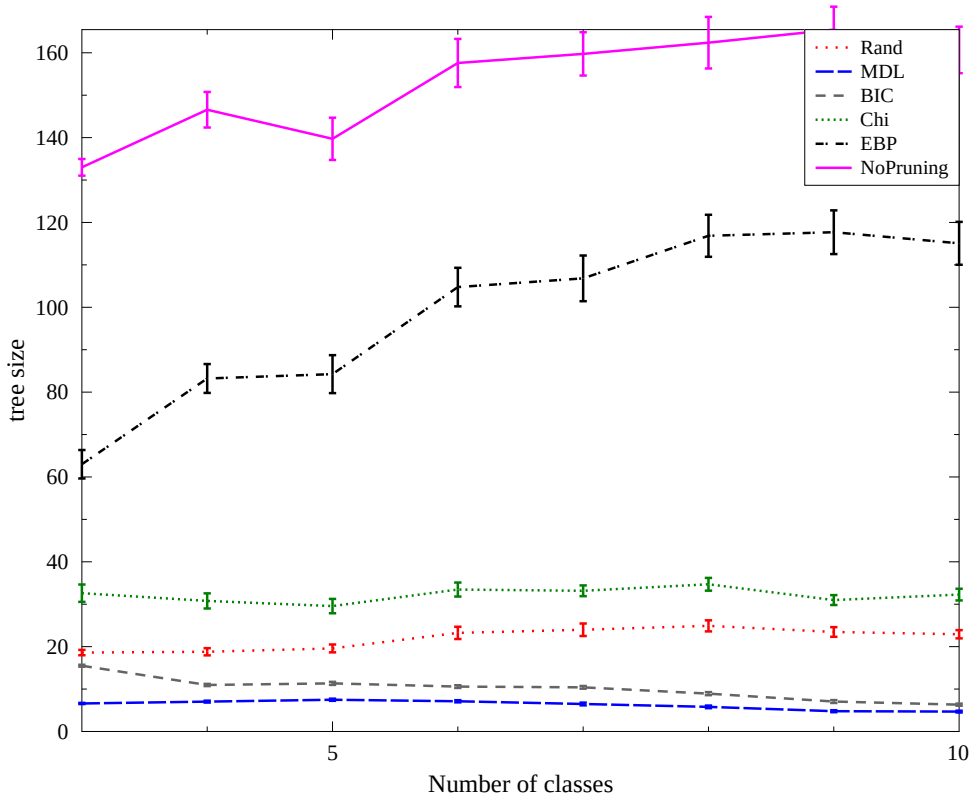

Figure 18: Influence of the number of classes on tree size (lower is better).

3.4 ‘Pen Digits’ Dataset

In the experiment on the ‘pen digits’ dataset we varied the number of classes from 10 (as in the original dataset) to 3. We always used 2340 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.4.1 AUC

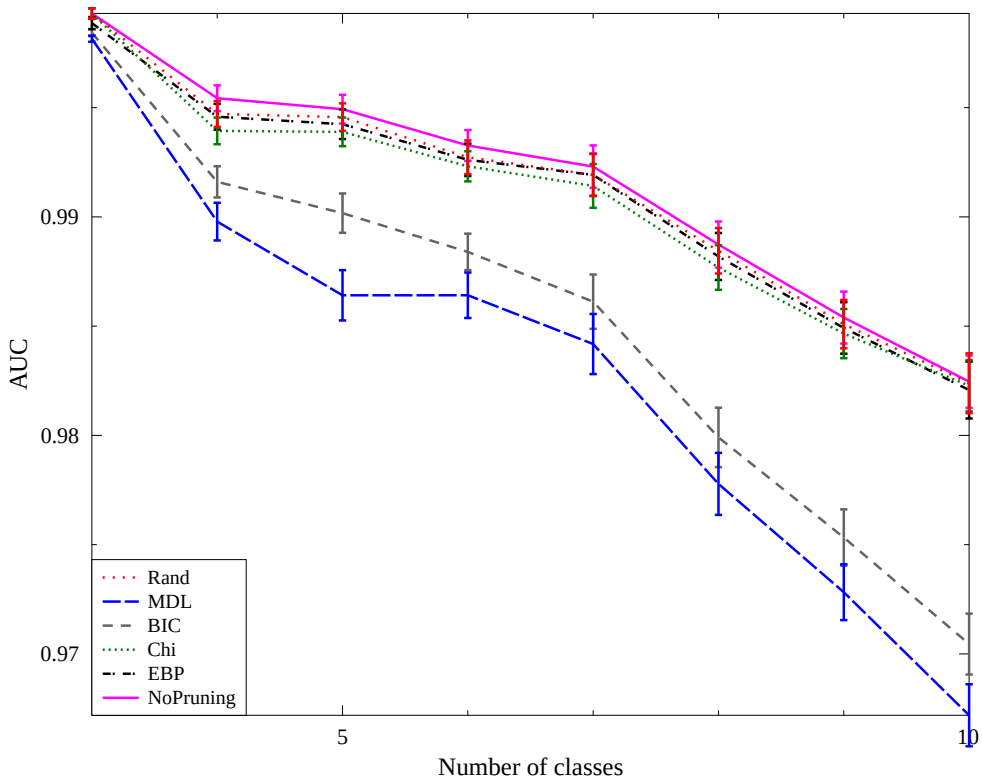

Figure 19: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.4.2 RMSE

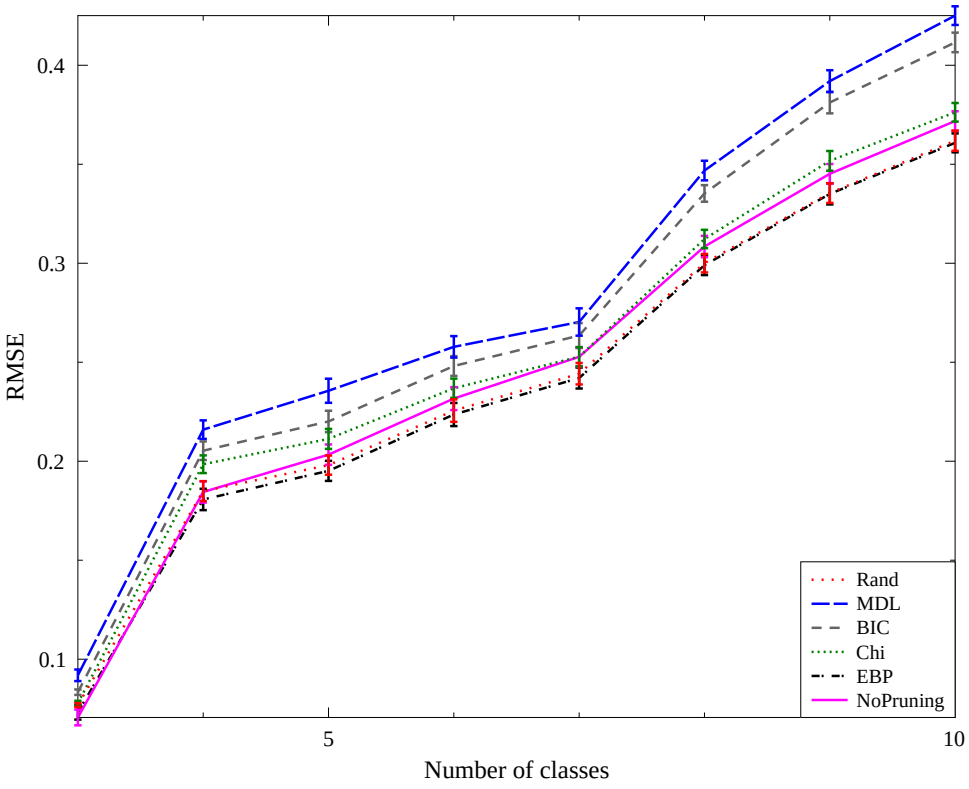

Figure 20: Influence of the number of classes on the performance measure RMSE (lower is better).

3.4.3 CLL

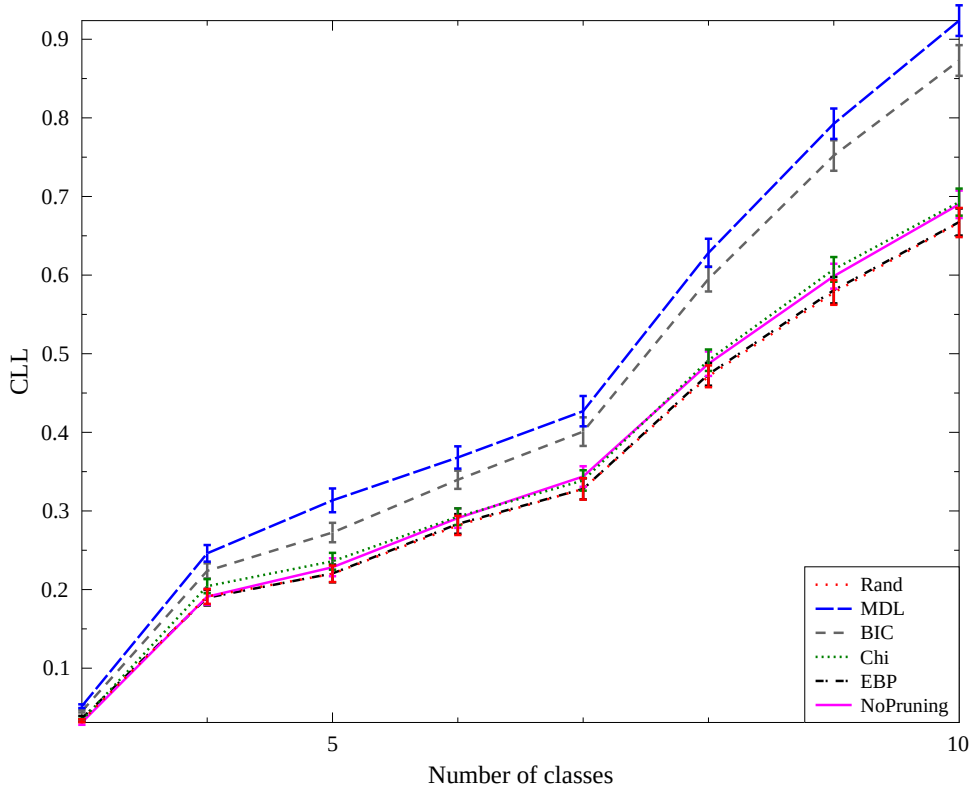

Figure 21: Influence of the number of classes on the performance measure CLL (lower is better).

3.4.4 Calibration Error

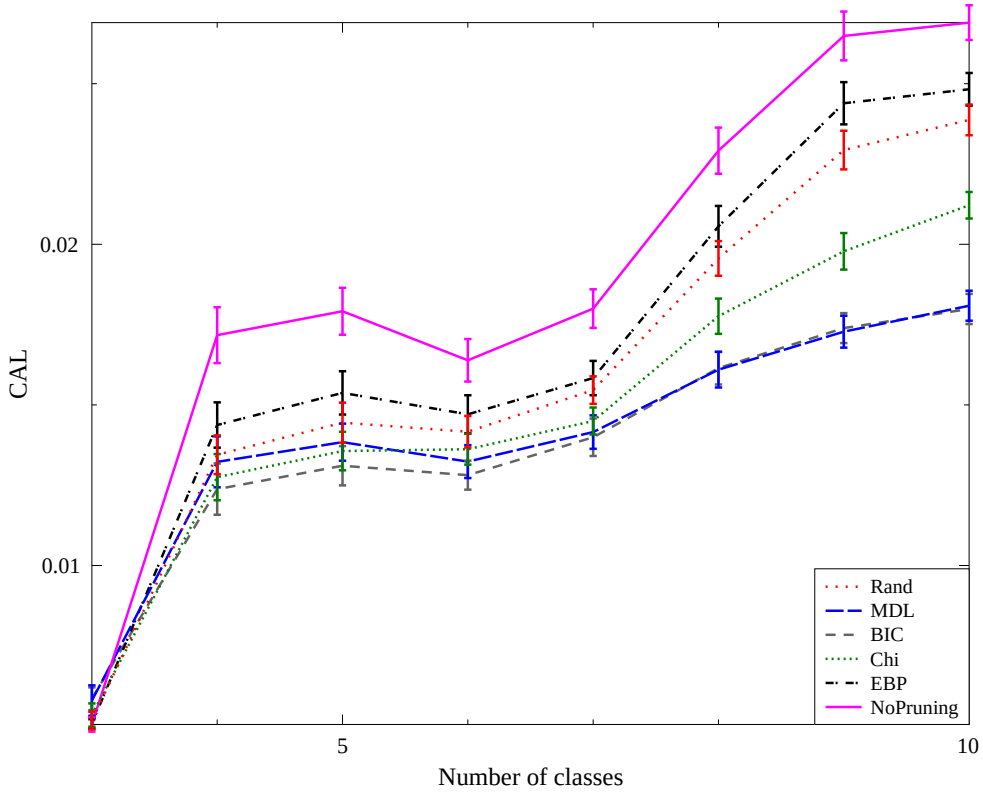

Figure 22: Influence of the number of classes on the performance measure CAL (lower is better).

3.4.5 Classification Accuracy

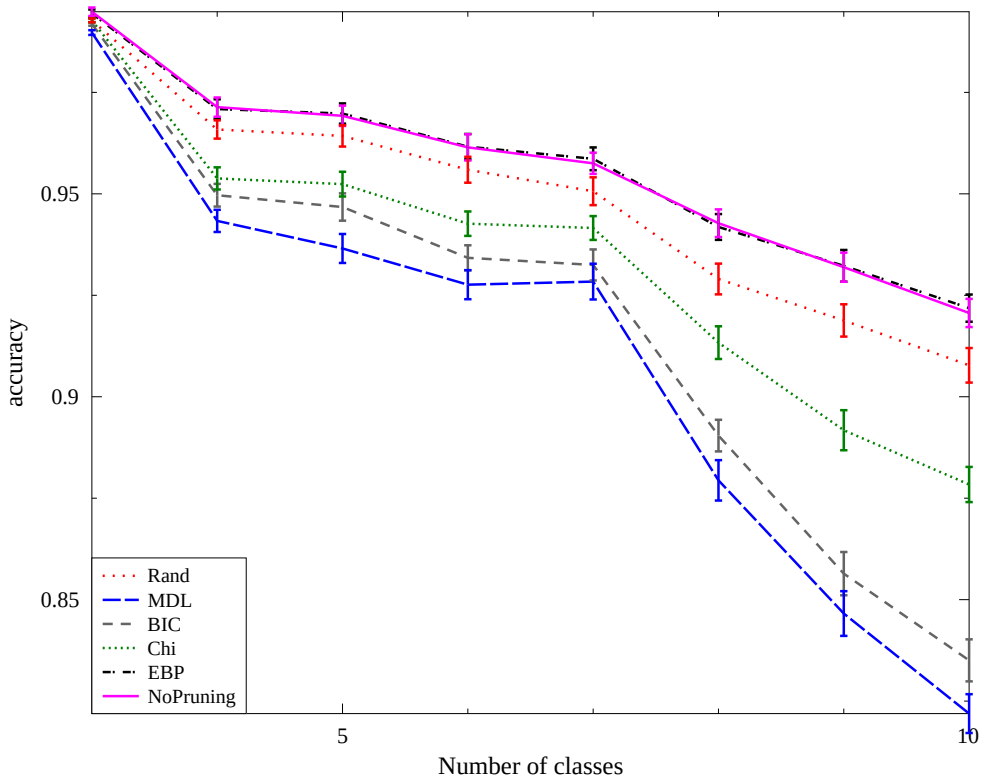

Figure 23: Influence of the number of classes on the performance measure accuracy (higher is better).

3.4.6 Tree Size

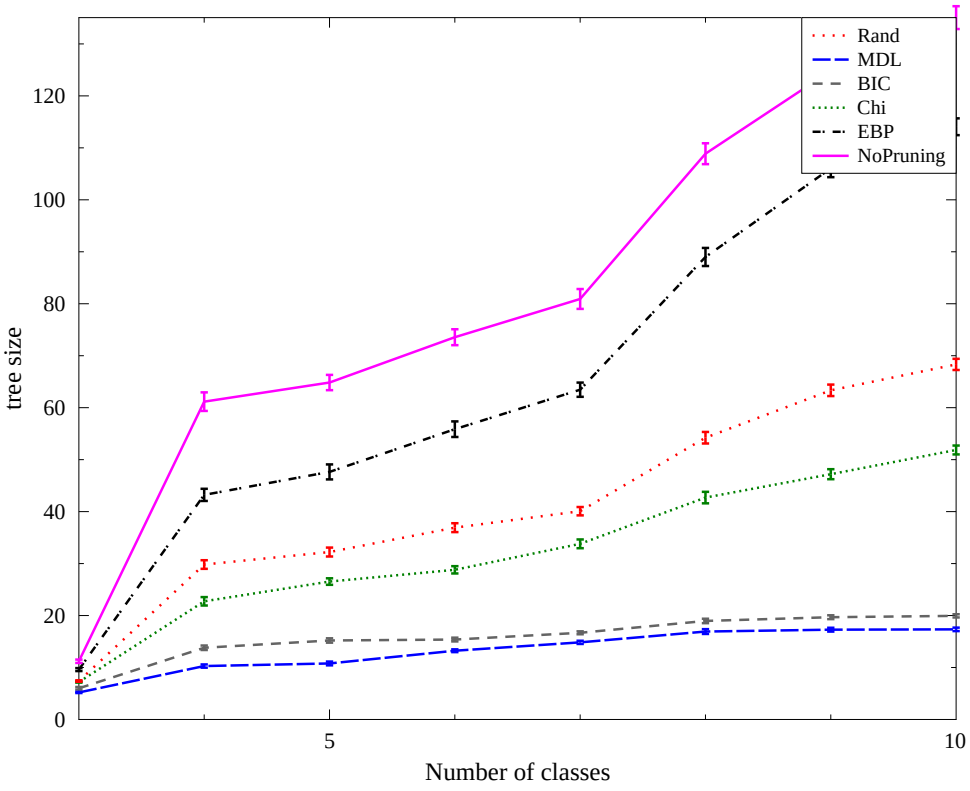

Figure 24: Influence of the number of classes on tree size (lower is better).

3.5 'Soybean' Dataset

In the experiment on the 'soybean' dataset we varied the number of classes from 15 (as in the original dataset) to 3. We always used 274 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.5.1 AUC

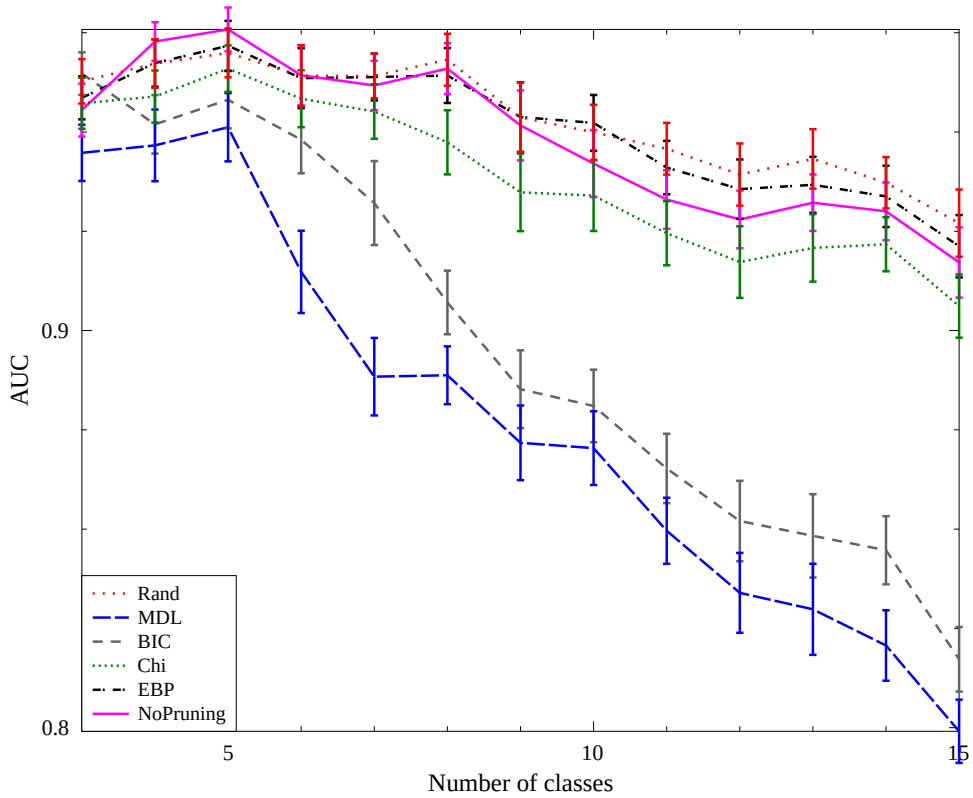

Figure 25: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.5.2 RMSE

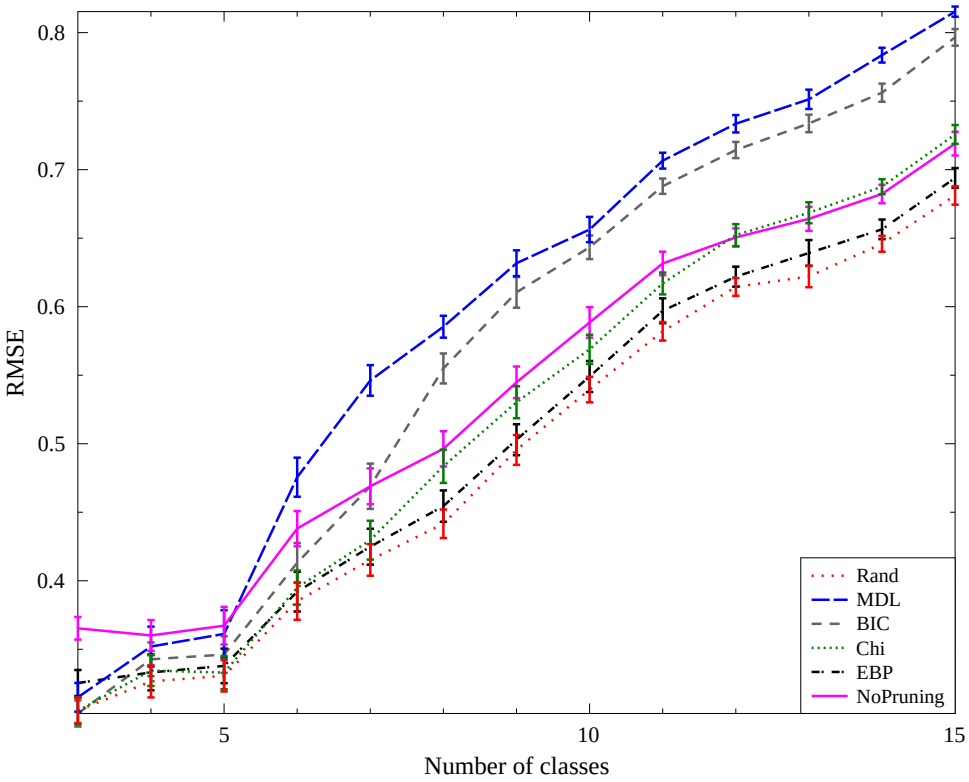

Figure 26: Influence of the number of classes on the performance measure RMSE (lower is better).

3.5.3 CLL

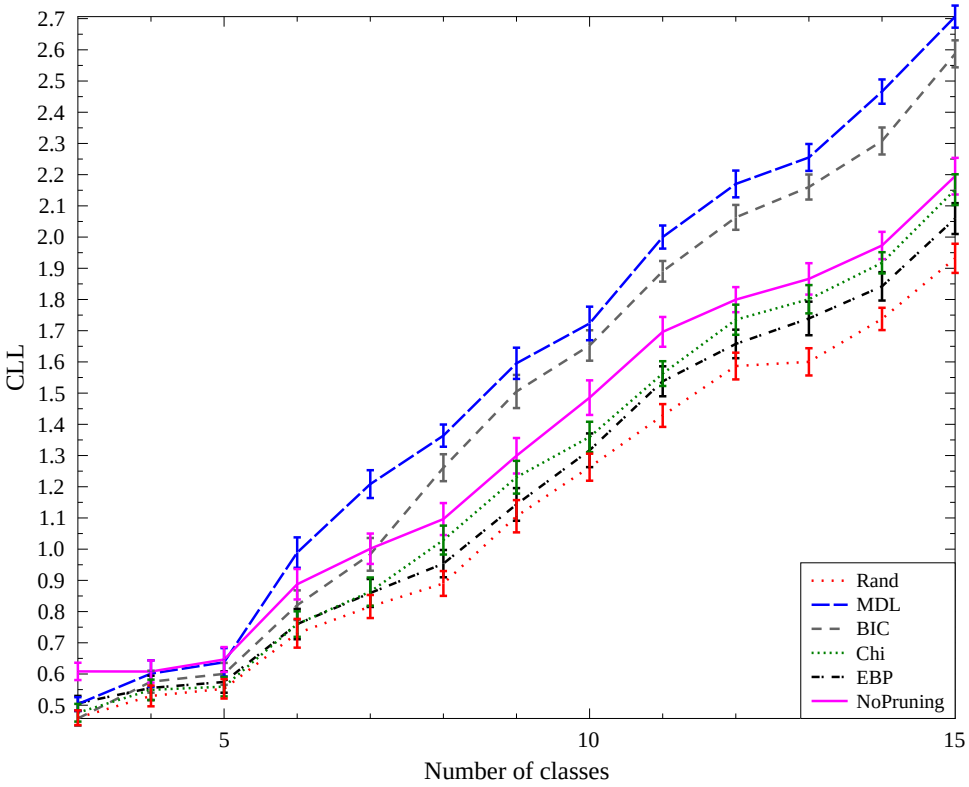

Figure 27: Influence of the number of classes on the performance measure CLL (lower is better).

3.5.4 Calibration Error

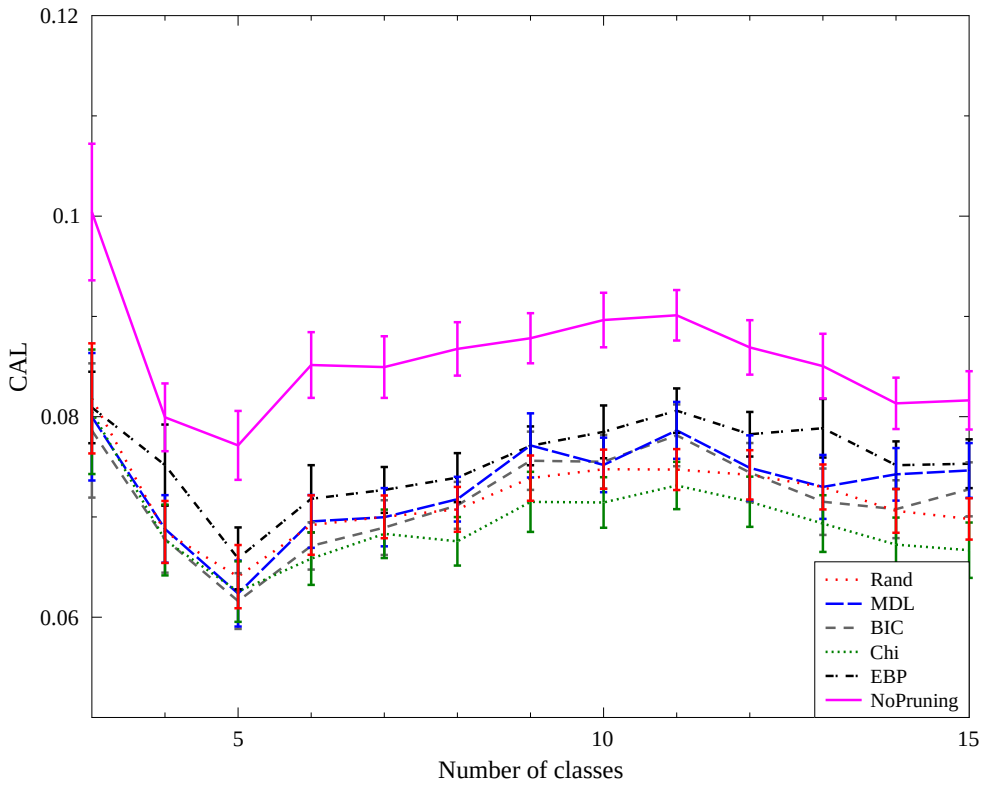

Figure 28: Influence of the number of classes on the performance measure CAL (lower is better).

3.5.5 Classification Accuracy

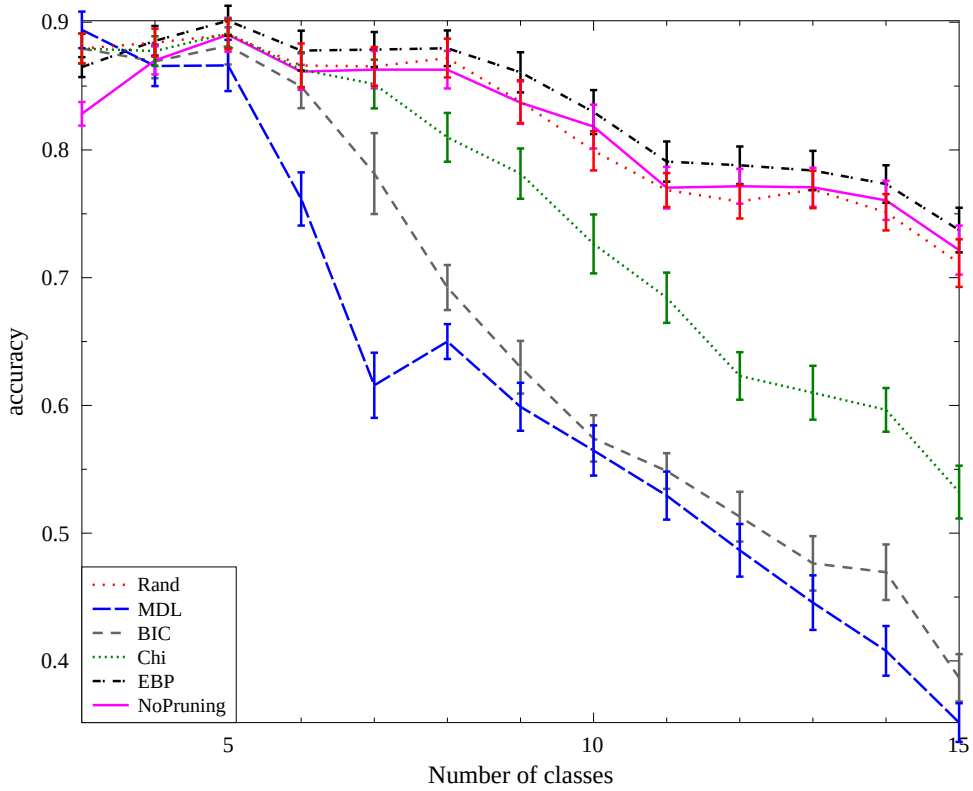

Figure 29: Influence of the number of classes on the performance measure accuracy (higher is better).

3.5.6 Tree Size

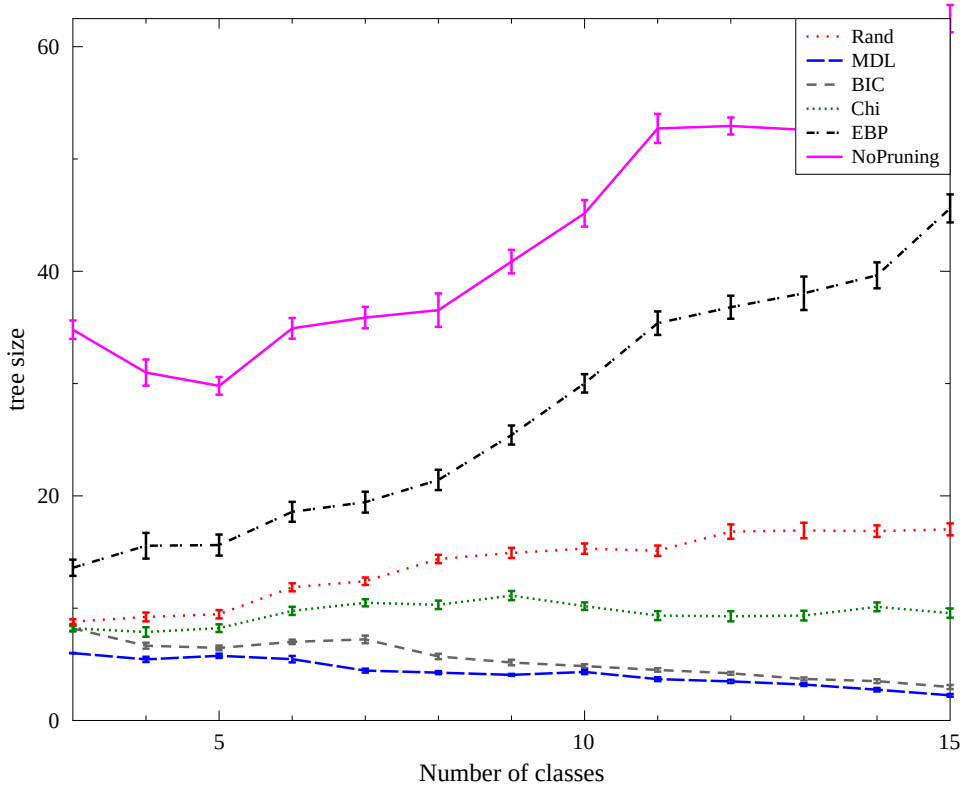

Figure 30: Influence of the number of classes on tree size (lower is better).

3.6 'Vowel' Dataset

In the experiment on the 'vowel' dataset we varied the number of classes from 11 (as in the original dataset) to 3. We always used 270 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.6.1 AUC

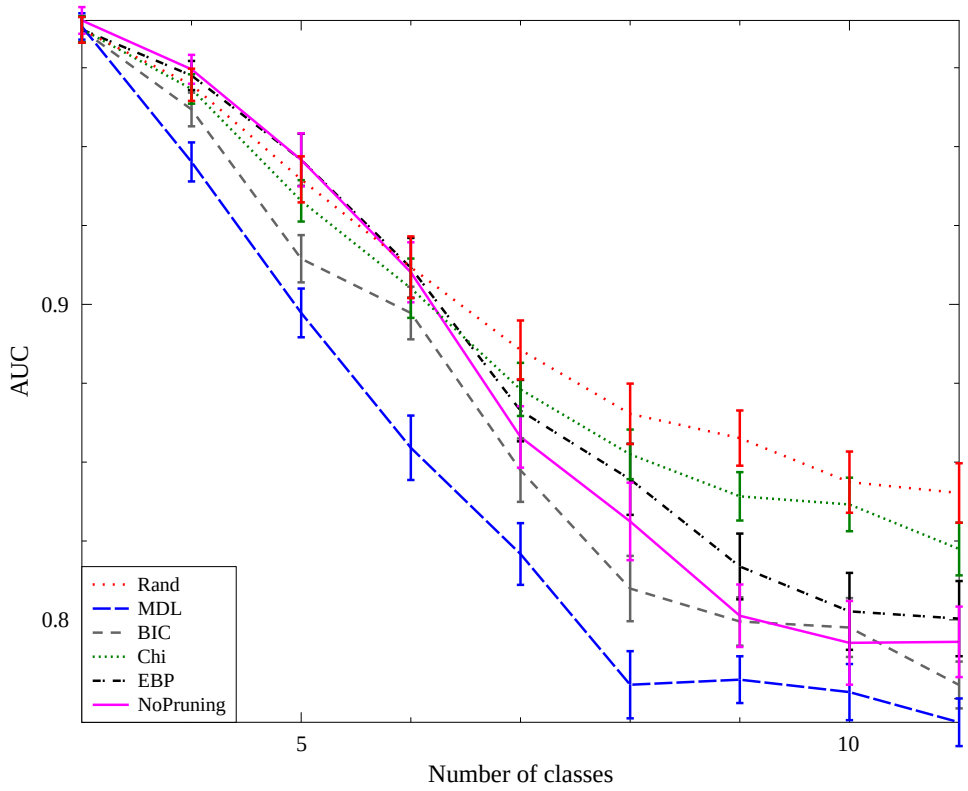

Figure 31: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.6.2 RMSE

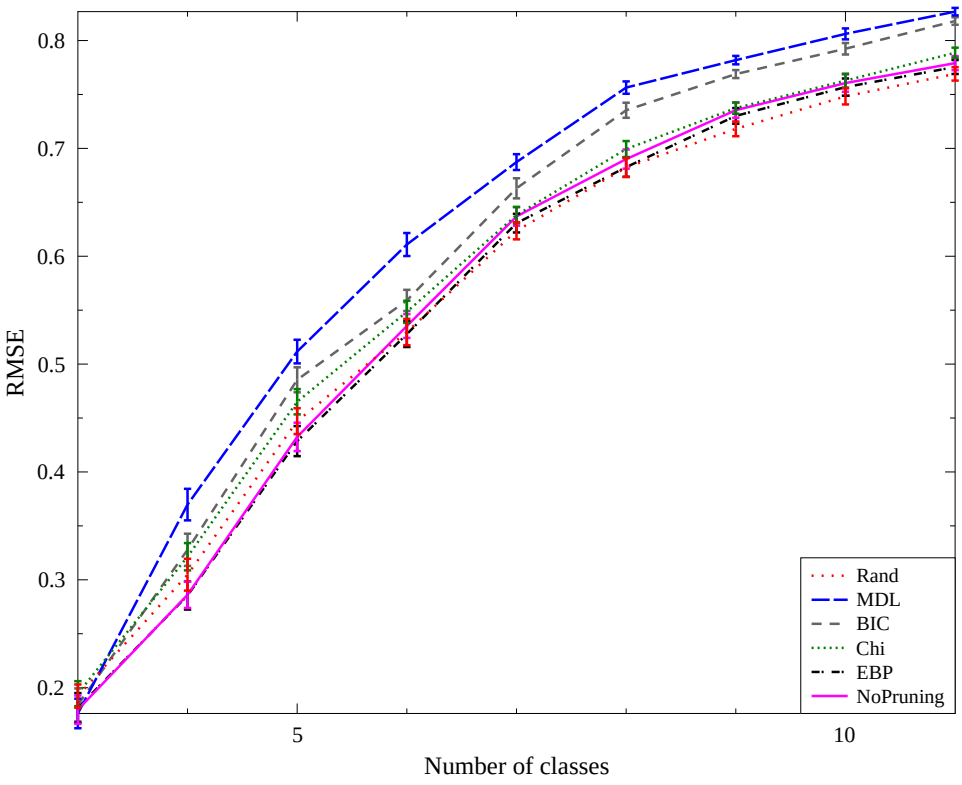

Figure 32: Influence of the number of classes on the performance measure RMSE (lower is better).

3.6.3 CLL

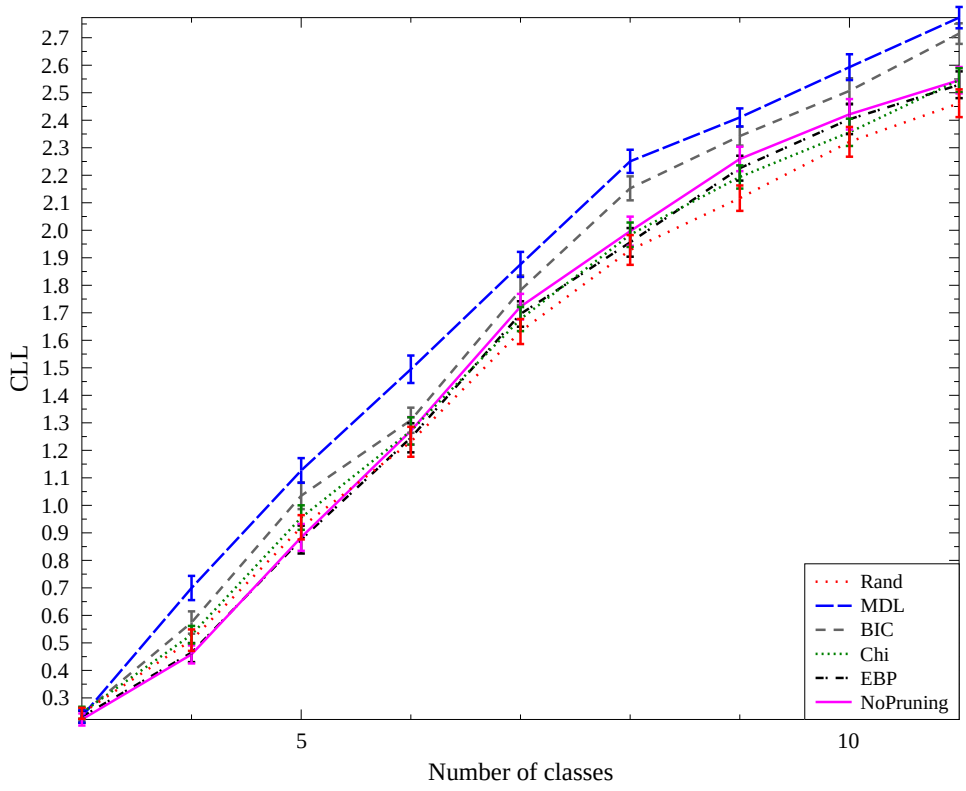

Figure 33: Influence of the number of classes on the performance measure CLL (lower is better).

3.6.4 Calibration Error

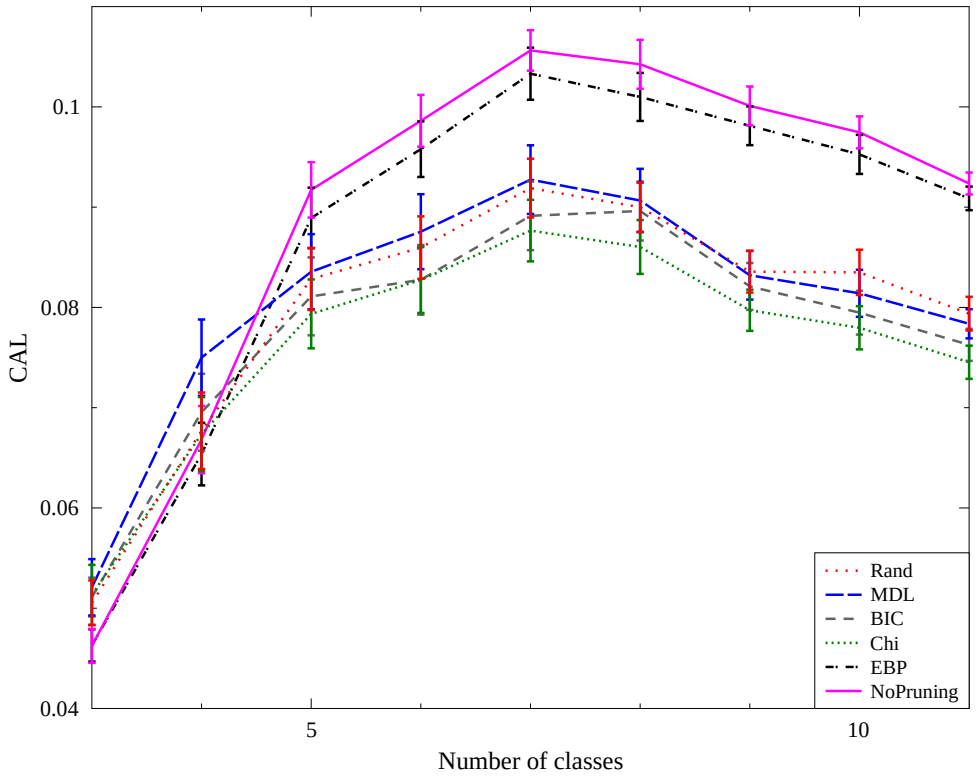

Figure 34: Influence of the number of classes on the performance measure CAL (lower is better).

3.6.5 Classification Accuracy

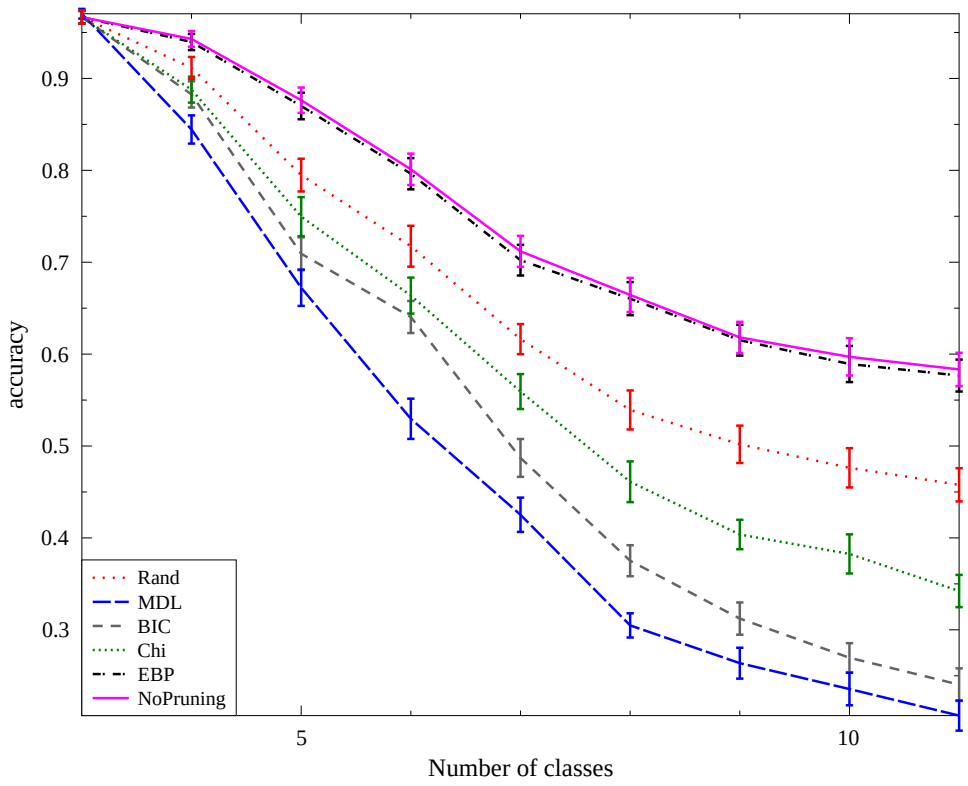

Figure 35: Influence of the number of classes on the performance measure accuracy (higher is better).

3.6.6 Tree Size

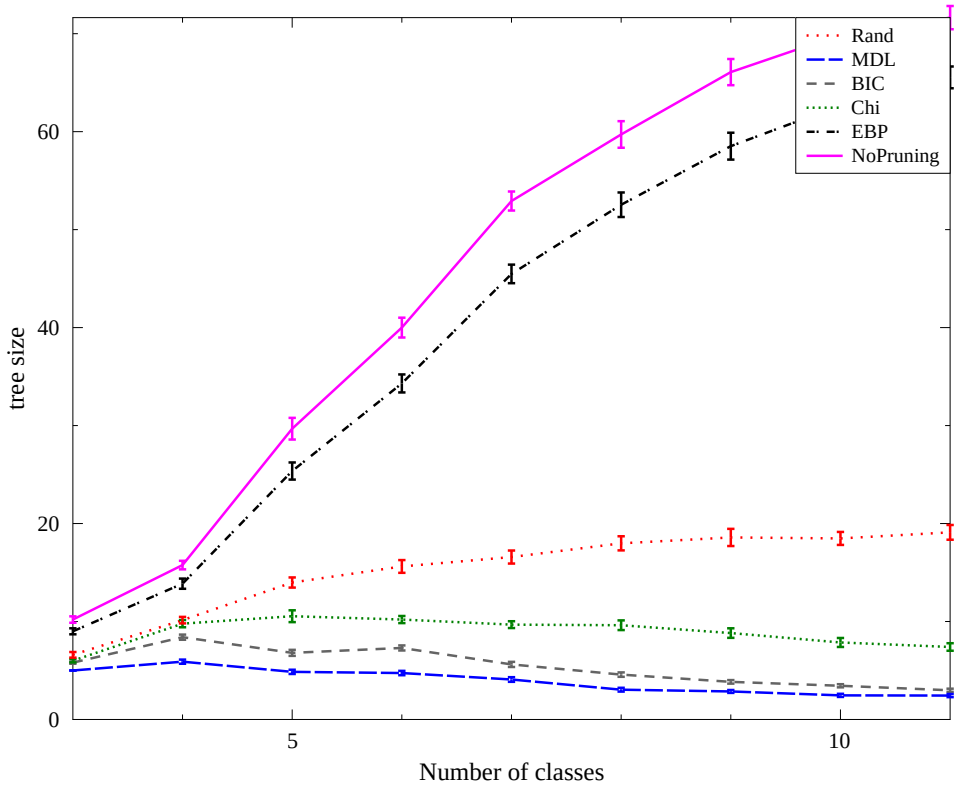

Figure 36: Influence of the number of classes on tree size (lower is better).

3.7 'Yeast' Dataset

In the experiment on the 'yeast' dataset we varied the number of classes from 10 (as in the original dataset) to 3. We always used 1136 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 3.1.1).

3.7.1 AUC

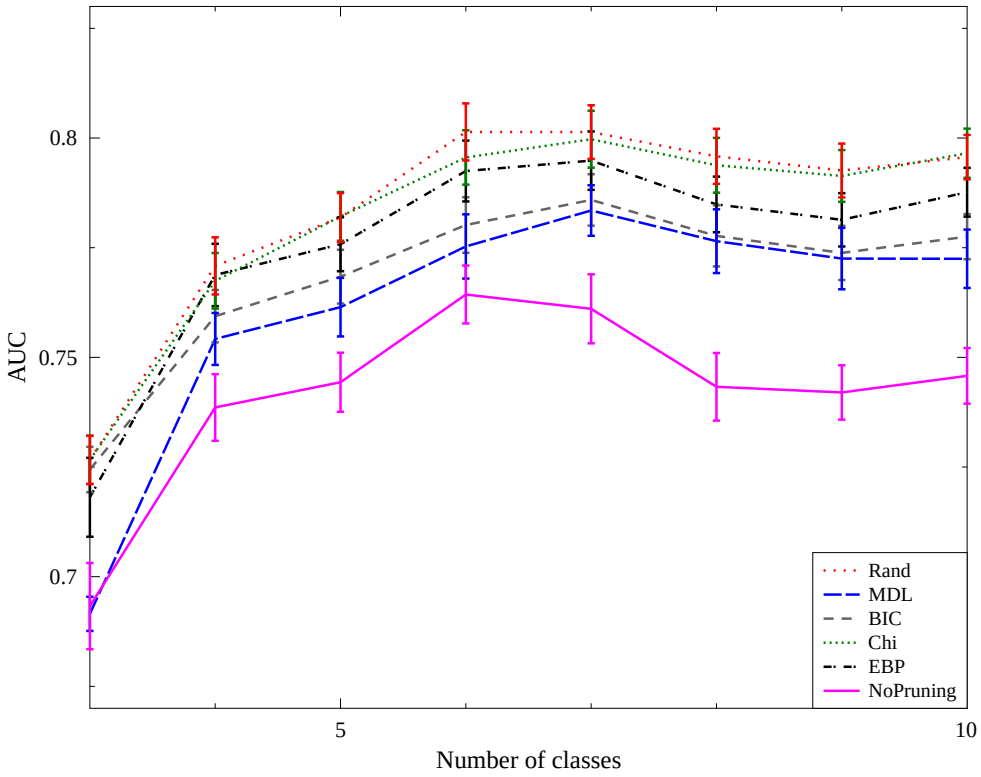

Figure 37: Influence of the number of classes on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

3.7.2 RMSE

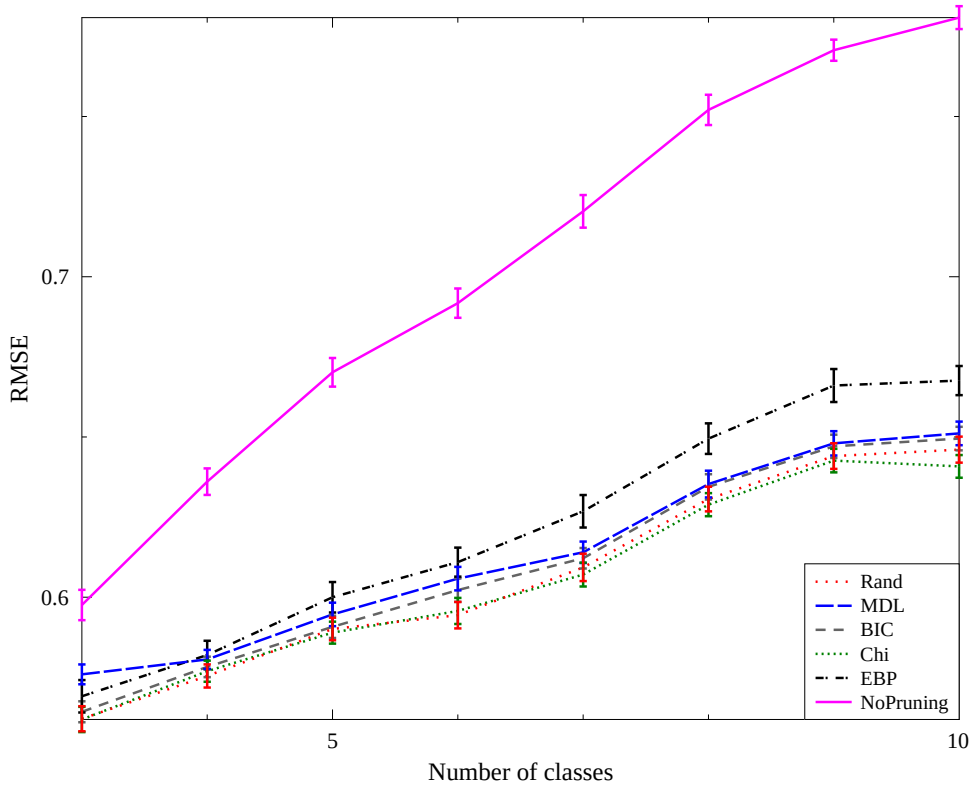

Figure 38: Influence of the number of classes on the performance measure RMSE (lower is better).

3.7.3 CLL

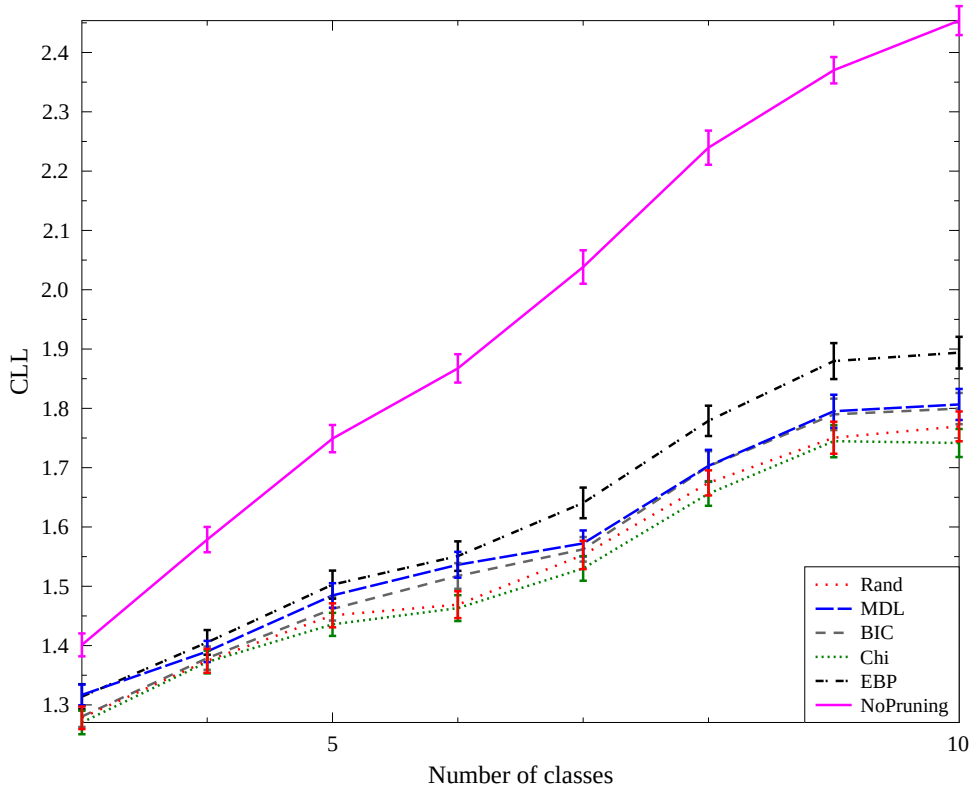

Figure 39: Influence of the number of classes on the performance measure CLL (lower is better).

3.7.4 Calibration Error

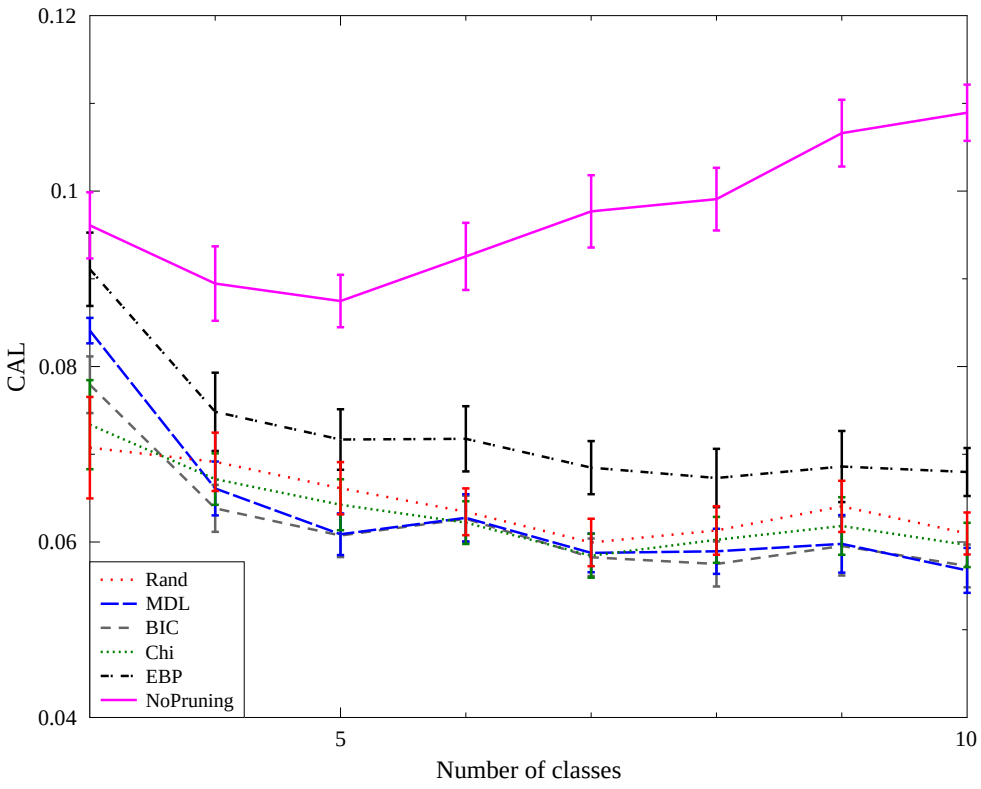

Figure 40: Influence of the number of classes on the performance measure CAL (lower is better).

3.7.5 Classification Accuracy

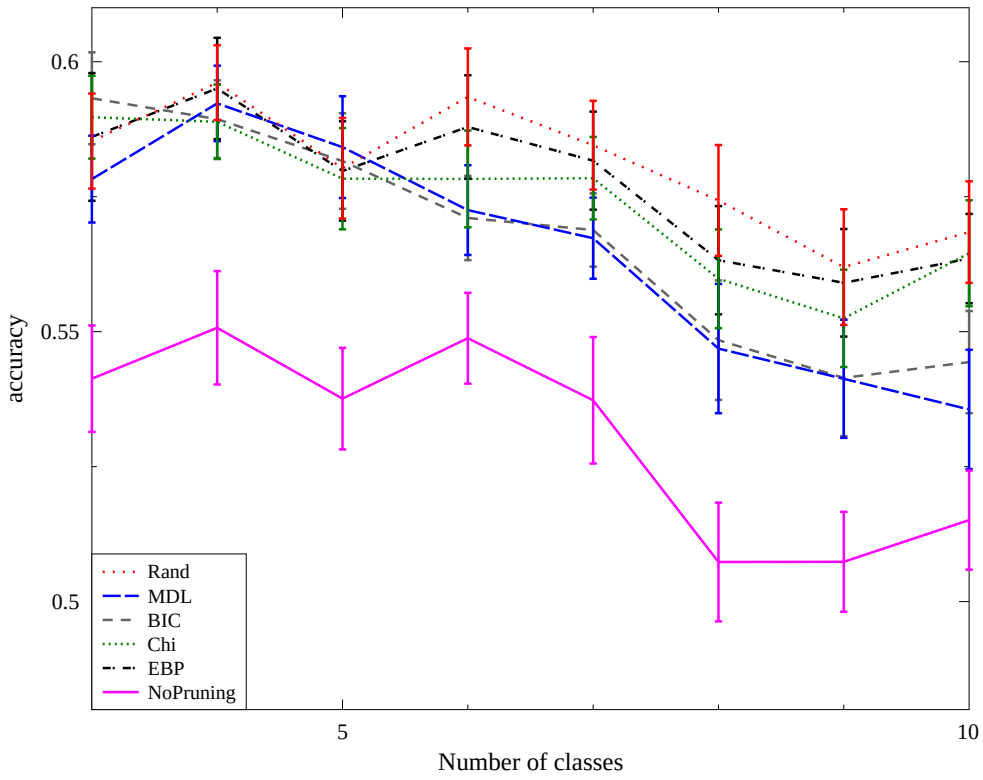

Figure 41: Influence of the number of classes on the performance measure accuracy (higher is better).

3.7.6 Tree Size

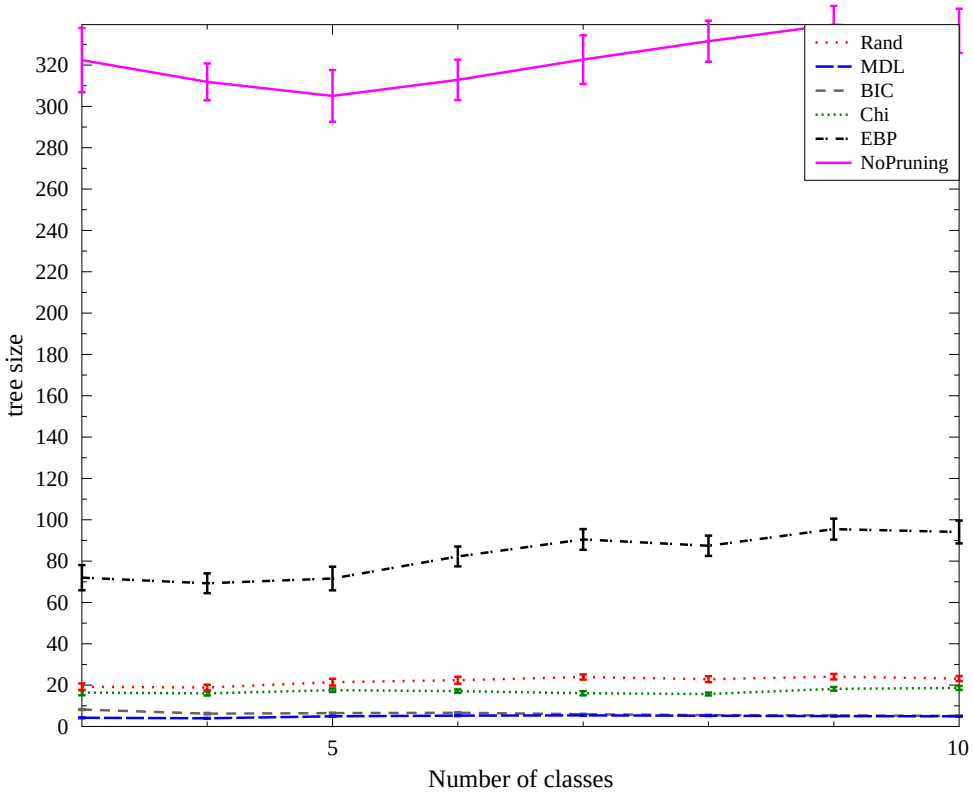

Figure 42: Influence of the number of classes on tree size (lower is better).

4 Influence of the Class Skew

We now consider the experiments in which we varied the skew of the class distribution of two-class datasets. We performed this experiment on all two-class datasets that have a sufficient (≥ 500) number of examples for this experiment: 'australian credit', 'breast', 'chess', 'diabetes', 'german credit' and 'hiv'.[3]

4.1 'Australian Credit' Dataset

We performed an experiment on the 'australian credit' dataset in which we varied the percentage of examples having the minority class from 50% (balanced) to 2.5% (strongly skewed).

4.1.1 Experimental Setup

For each minority class percentage that we consider, we use the same total number of examples in order to eliminate any influence of this factor on our results. Concretely, we always use 366 examples. Given the required number of examples from the majority and minority class, these examples were sampled randomly (without replacement) from all the examples in the original dataset.[4] For each minority class percentage we performed 10 runs of 5-fold cross-validation with new samples in every run.

[3] We also performed the experiments on the 'mushroom' dataset but we leave out these results because the differences in performance between the different algorithms are very small (because 'mushroom' is a very "easy" dataset).

[4] The reason why we use 366 examples in total is that this is the largest number of examples that can be used such that the required number of examples of each class can always be obtained by sampling without replacement (i.e.,using a larger total number of examples than 366 is only possible if we sample with replacement and use examples multiple times; we chose not to do this).

4.1.2 AUC

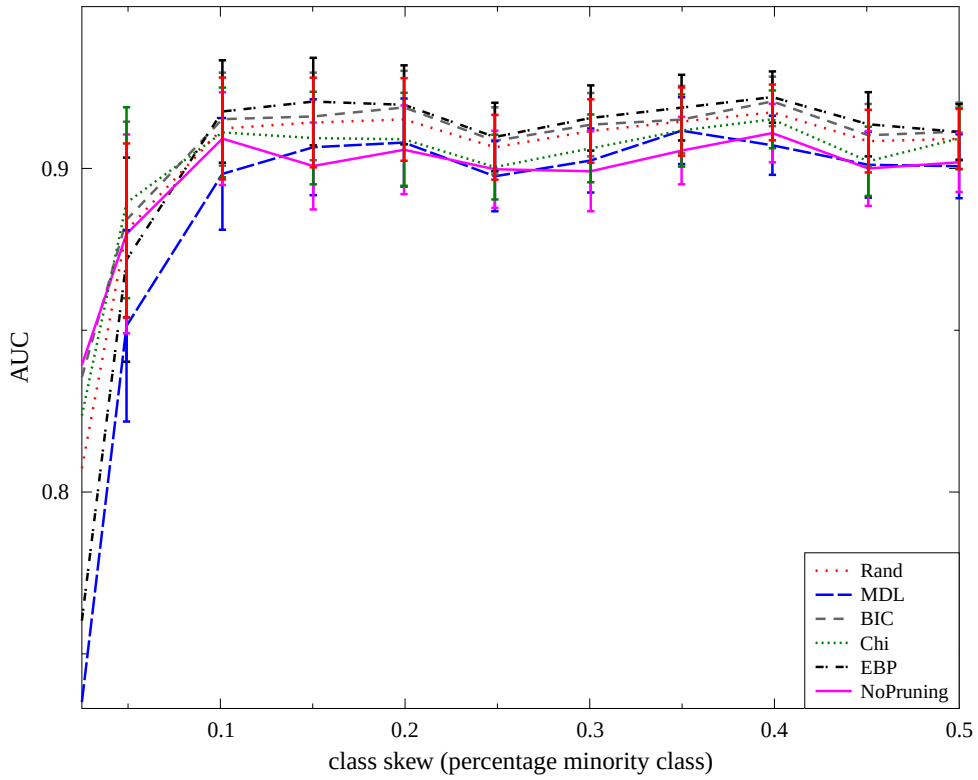

Figure 43: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.1.3 RMSE

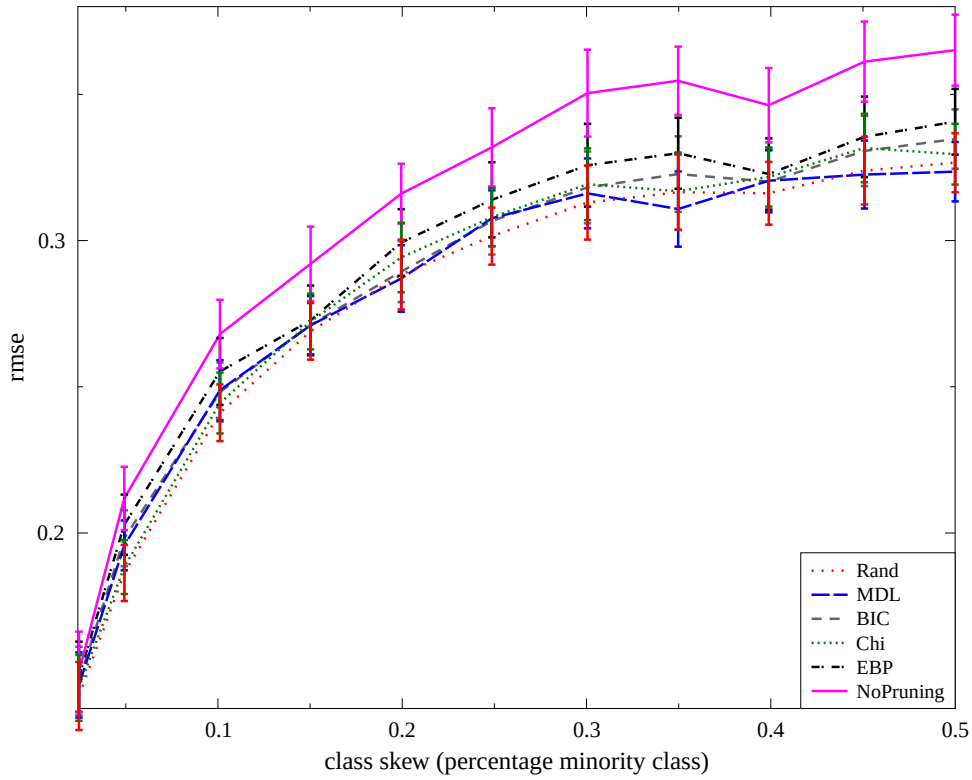

Figure 44: Influence of class skew on the performance measure RMSE (lower is better).

4.1.4 CLL

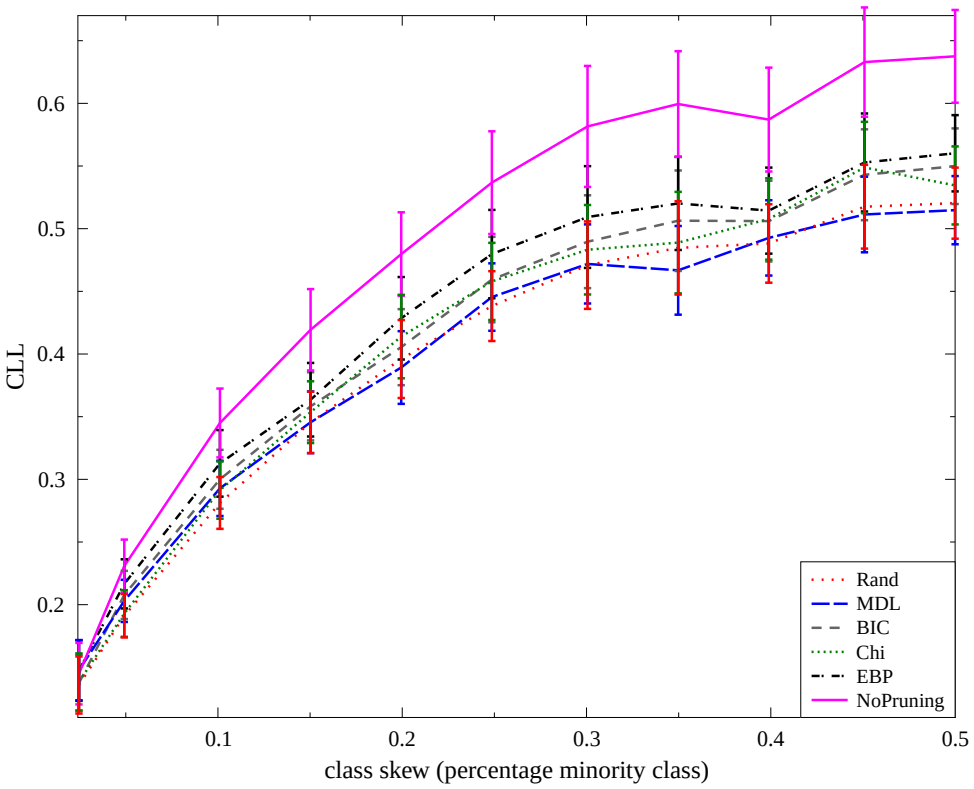

Figure 45: Influence of class skew on the performance measure CLL (lower is better).

4.1.5 Calibration Error

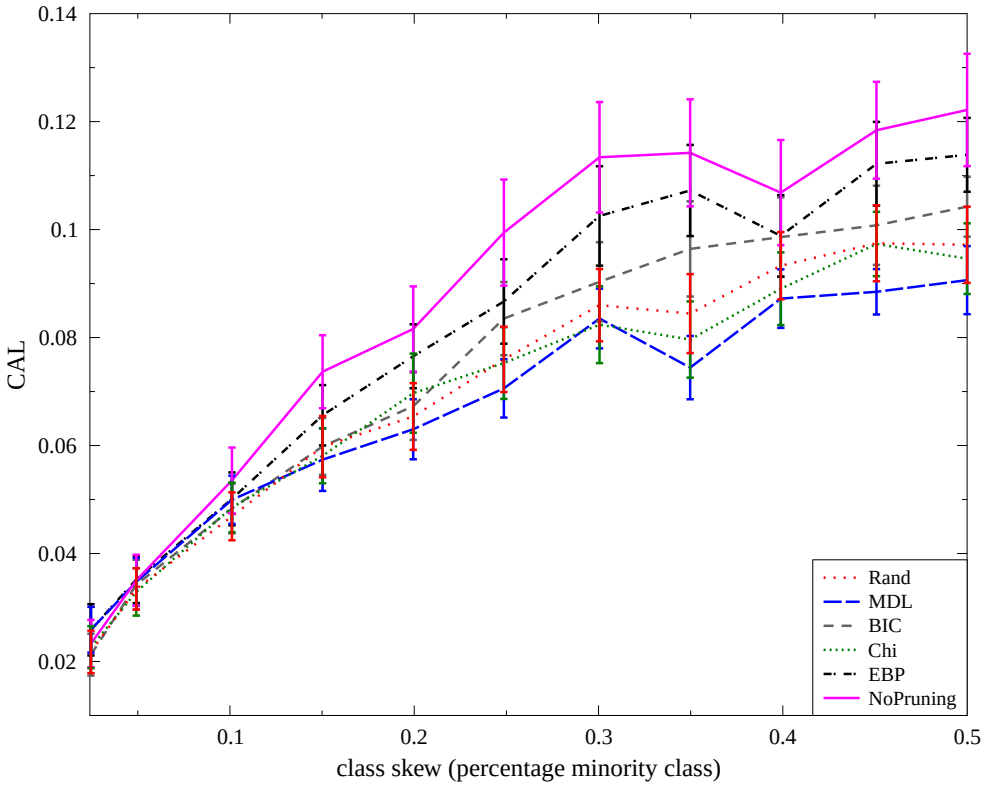

Figure 46: Influence of class skew on the performance measure CAL (lower is better).

4.1.6 Classification Accuracy

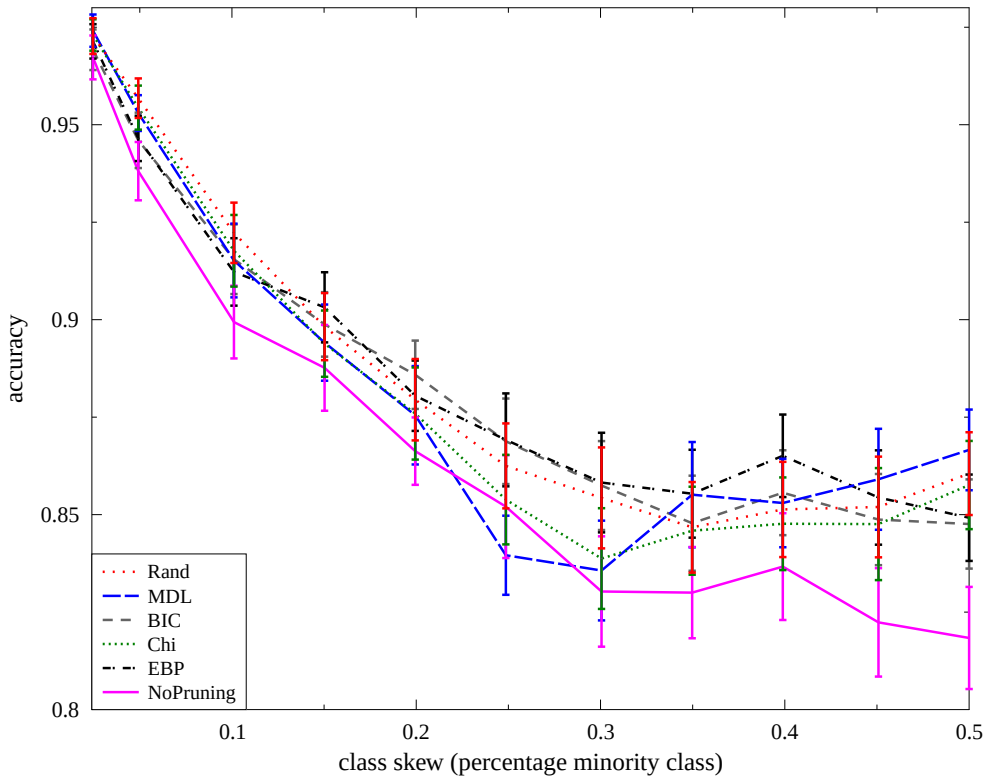

Figure 47: Influence of class skew on the performance measure accuracy (higher is better).

4.1.7 Tree Size

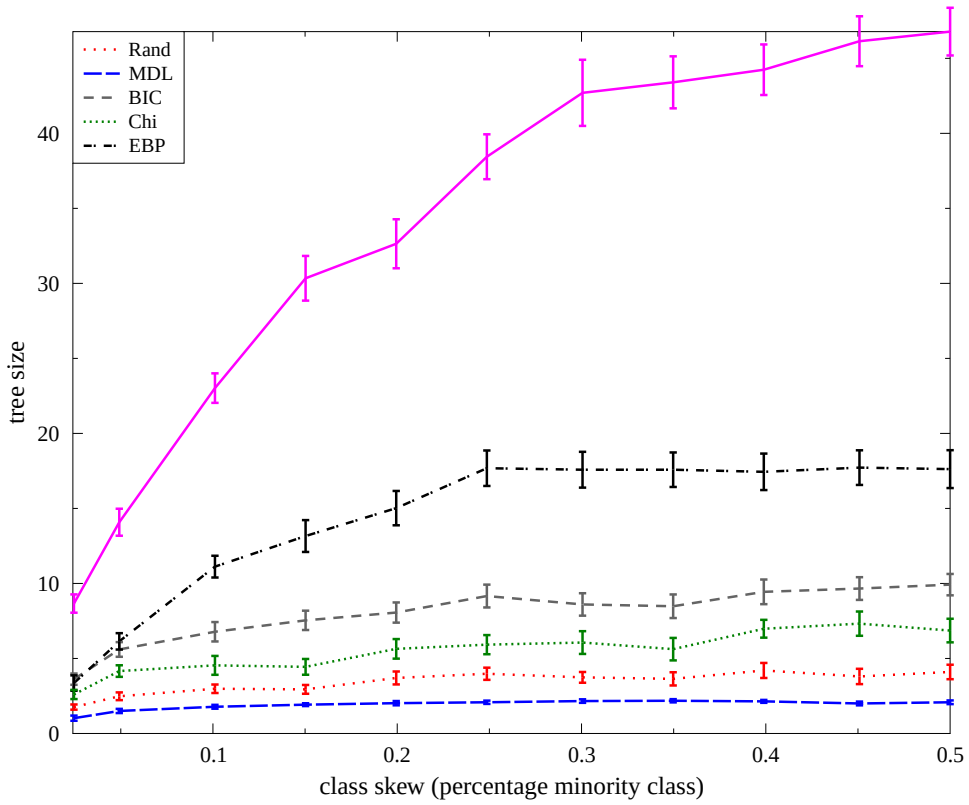

Figure 48: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

4.2 'Breast' Dataset

In the experiment on the 'breast' dataset we varied the minority class percentage from 50% (balanced) to 2.5% (strongly skewed). We always used 455 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 4.1.1).

4.2.1 AUC

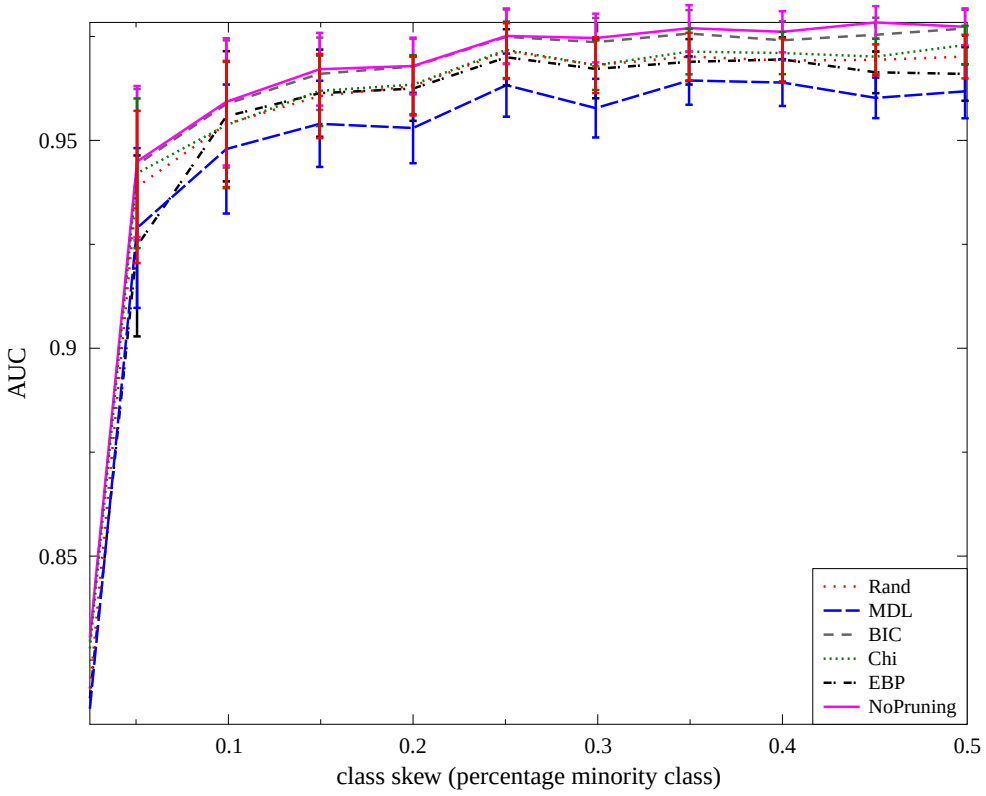

Figure 49: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.2.2 RMSE

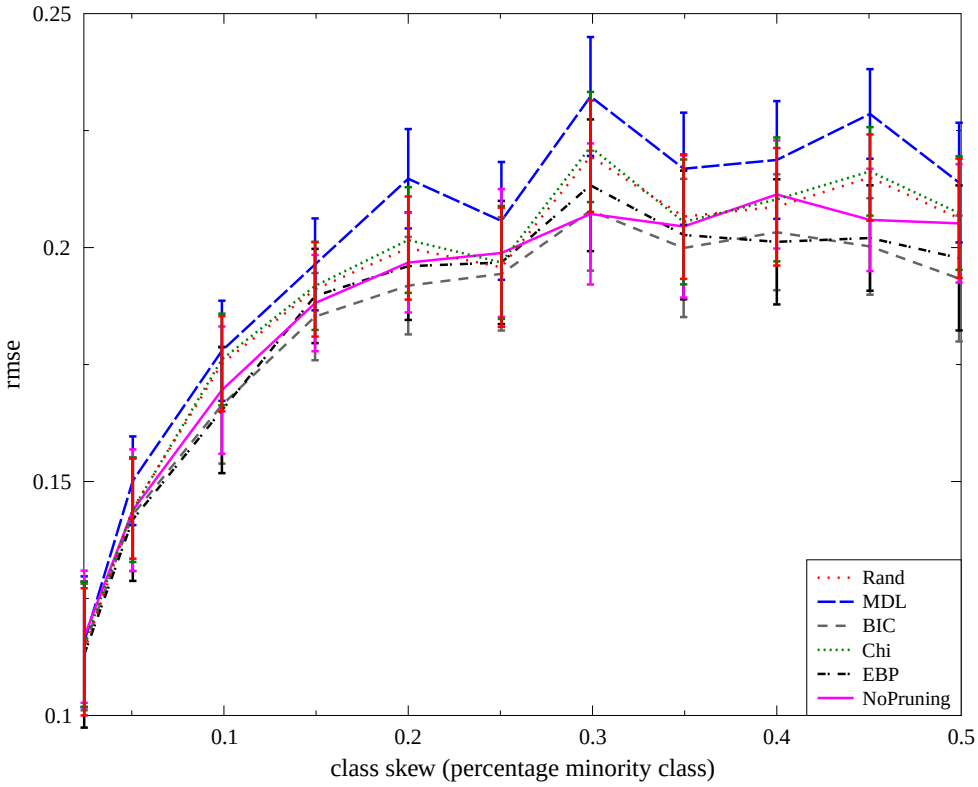

Figure 50: Influence of class skew on the performance measure RMSE (lower is better).

4.2.3 CLL

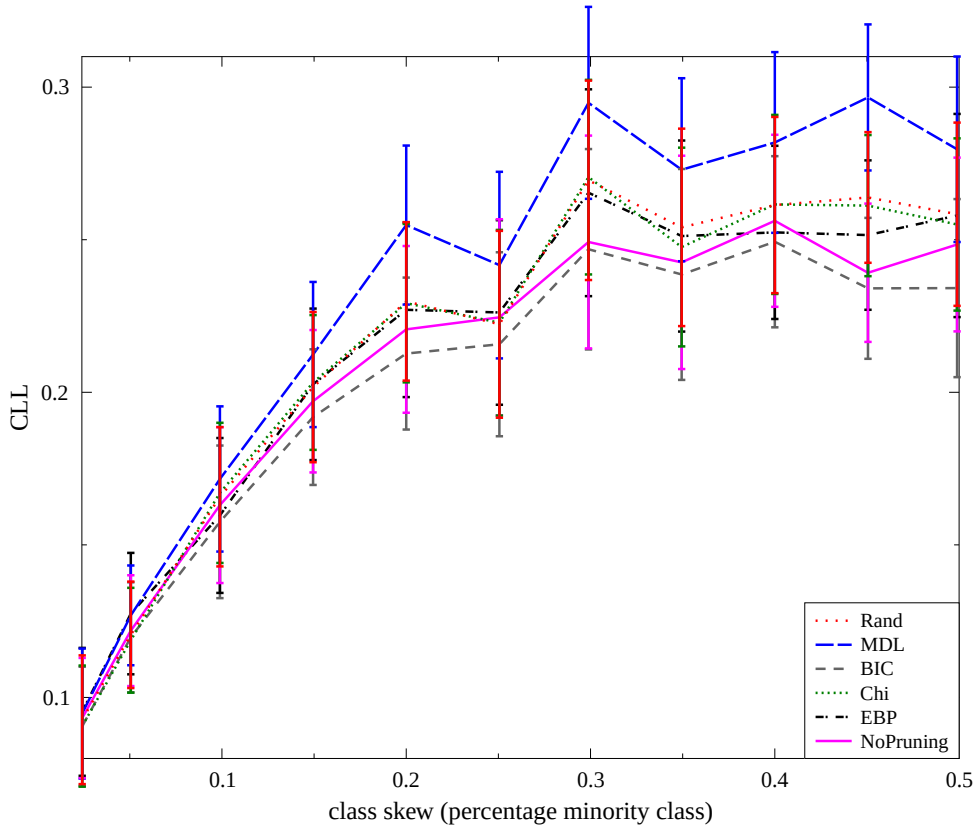

Figure 51: Influence of class skew on the performance measure CLL (lower is better).

4.2.4 Calibration Error

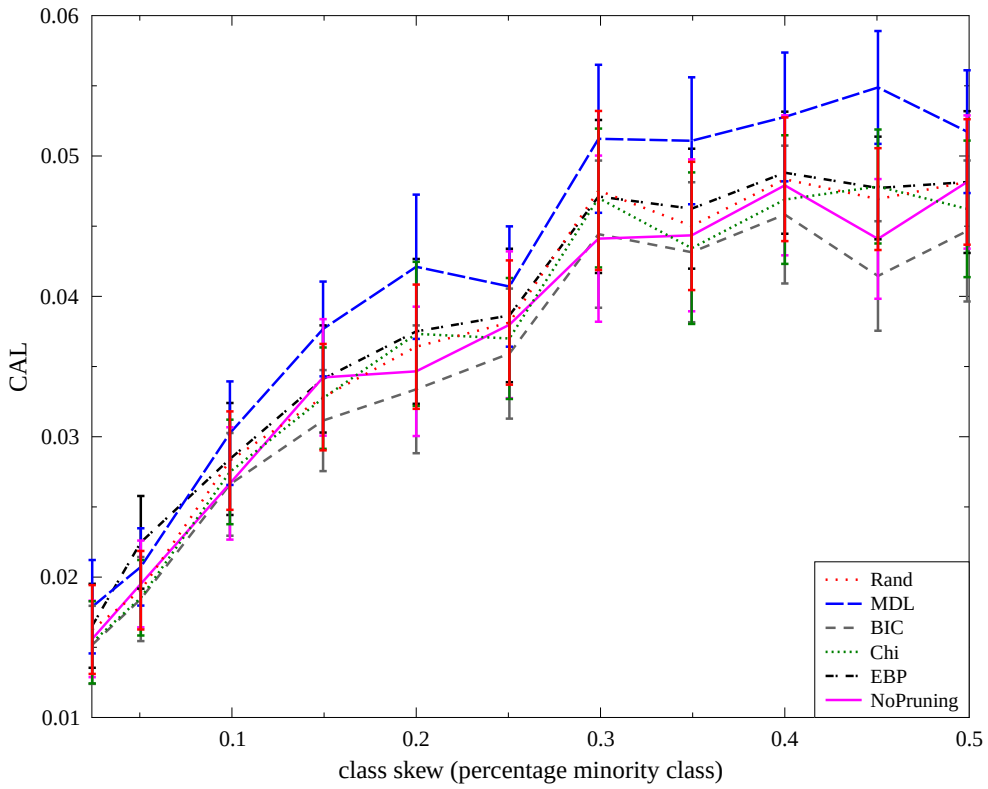

Figure 52: Influence of class skew on the performance measure CAL (lower is better).

4.2.5 Classification Accuracy

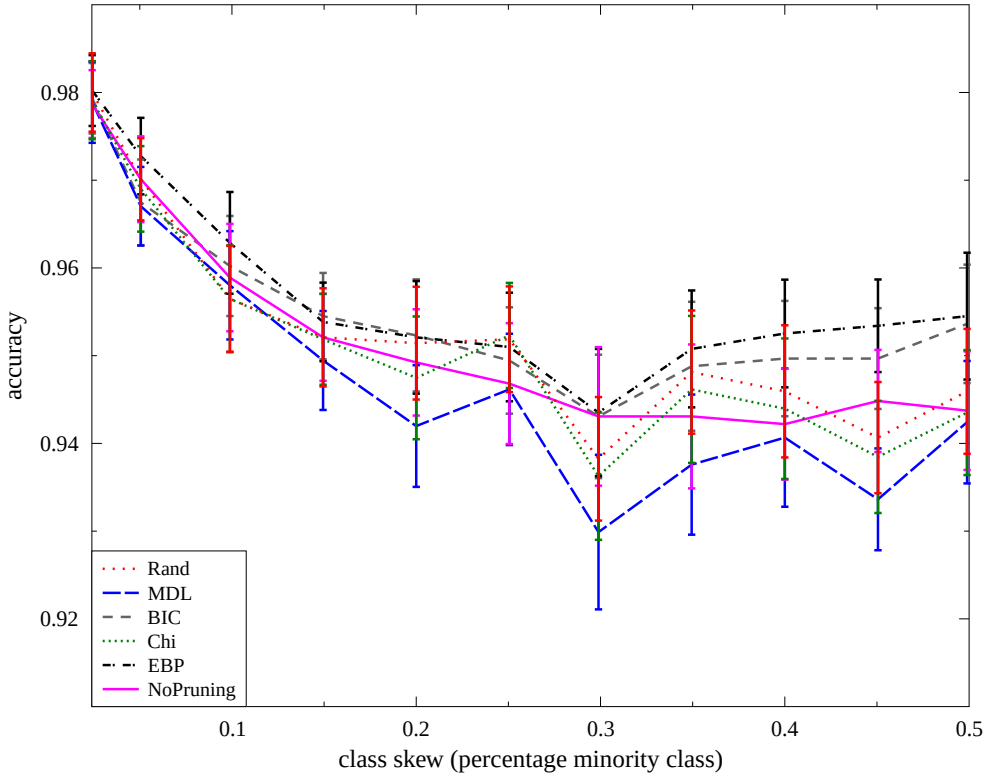

Figure 53: Influence of class skew on the performance measure accuracy (higher is better).

4.2.6 Tree Size

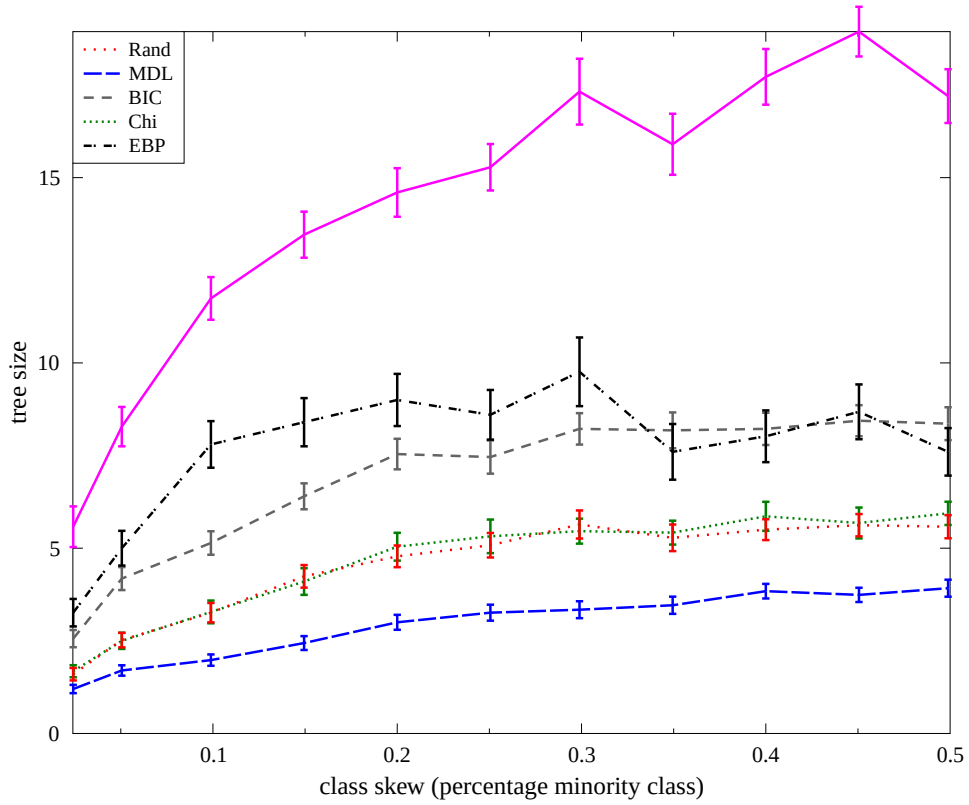

Figure 54: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

4.3 'Chess' Dataset

In the experiment on the 'chess' dataset we varied the minority class percentage from 50% (balanced) to 2.5% (strongly skewed). We always used 1711 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 4.1.1).

4.3.1 AUC

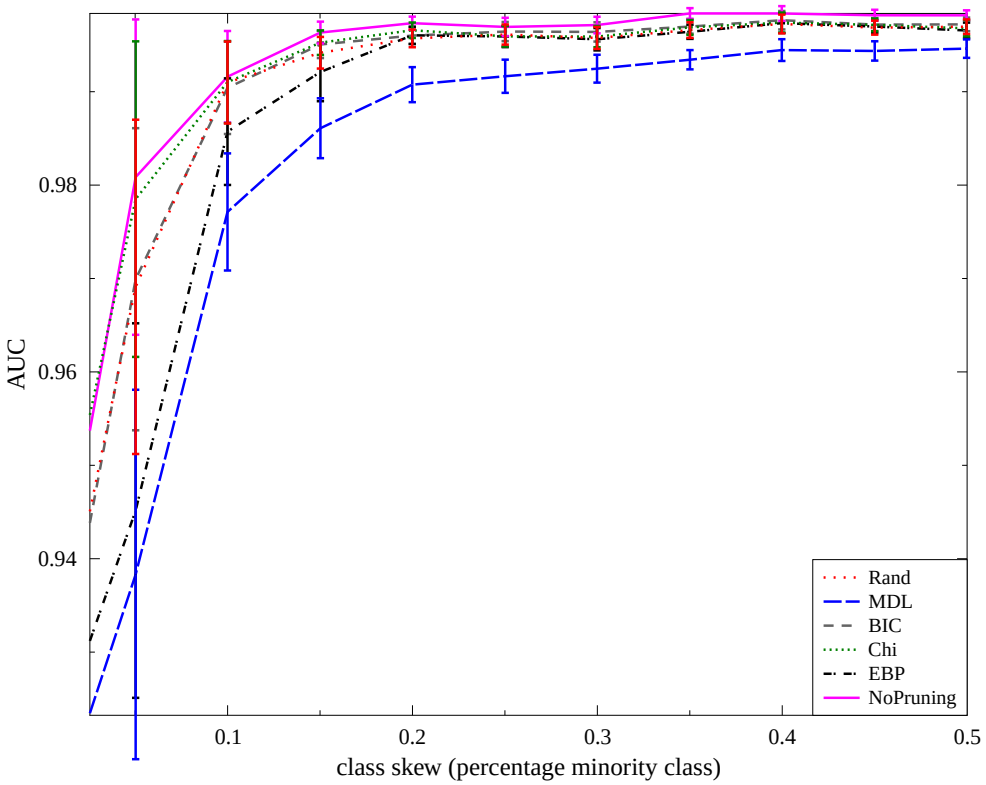

Figure 55: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.3.2 RMSE

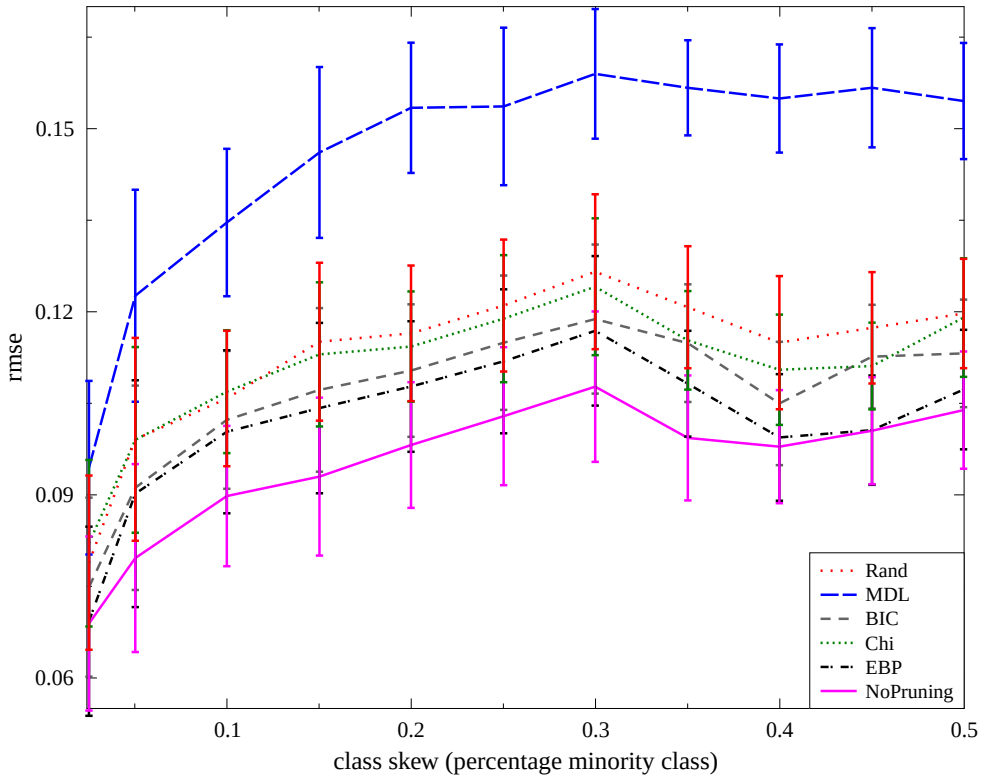

Figure 56: Influence of class skew on the performance measure RMSE (lower is better).

4.3.3 CLL

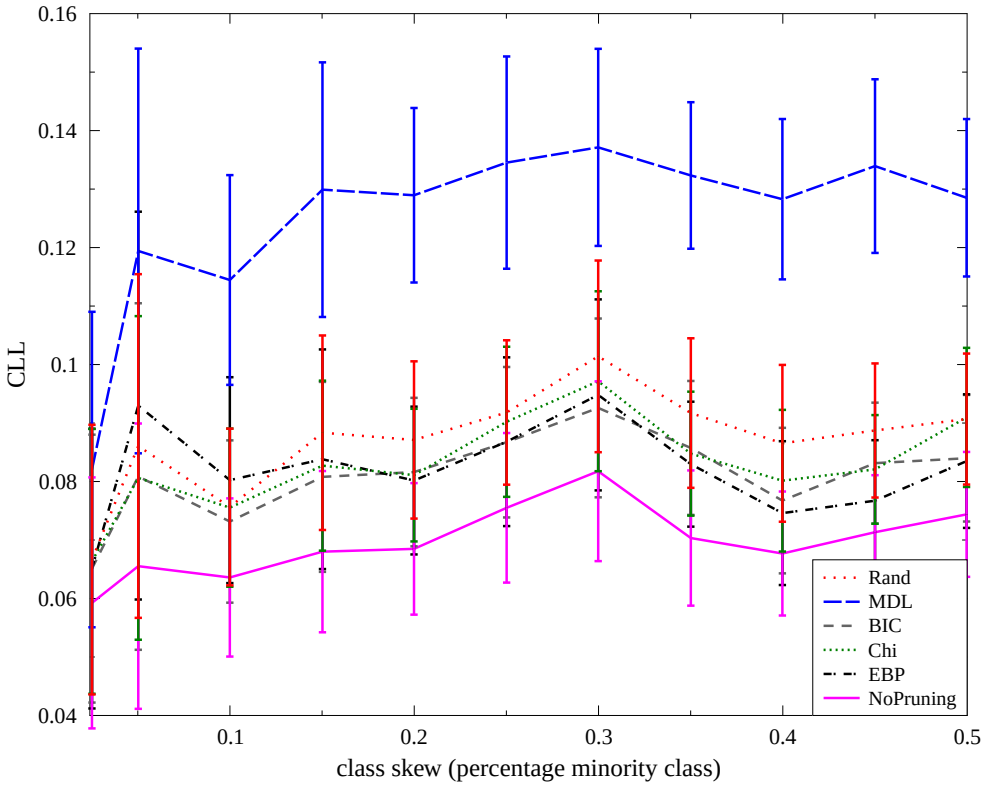

Figure 57: Influence of class skew on the performance measure CLL (lower is better).

4.3.4 Calibration Error

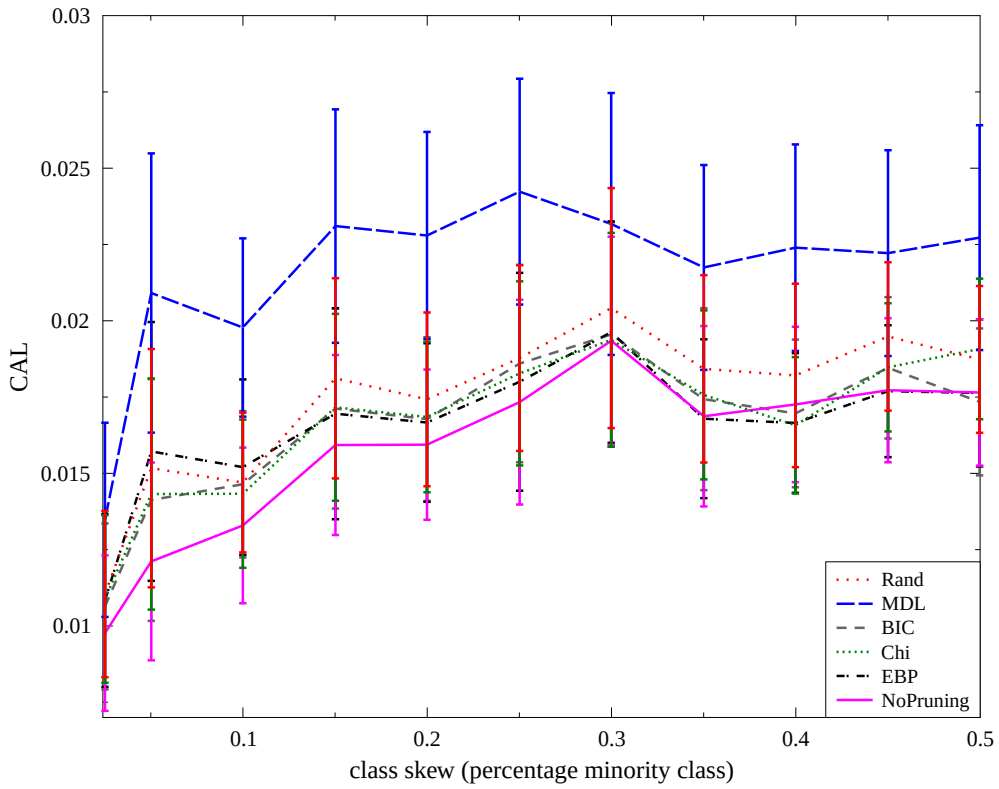

Figure 58: Influence of class skew on the performance measure CAL (lower is better).

4.3.5 Classification Accuracy

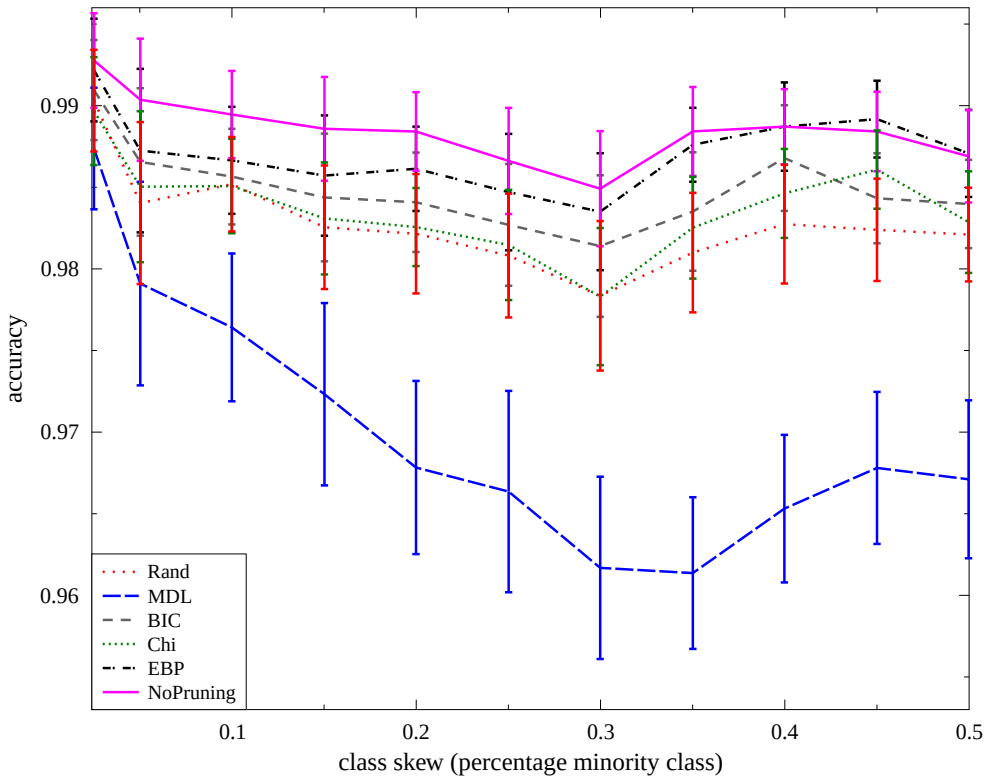

Figure 59: Influence of class skew on the performance measure accuracy (higher is better).

4.3.6 Tree Size

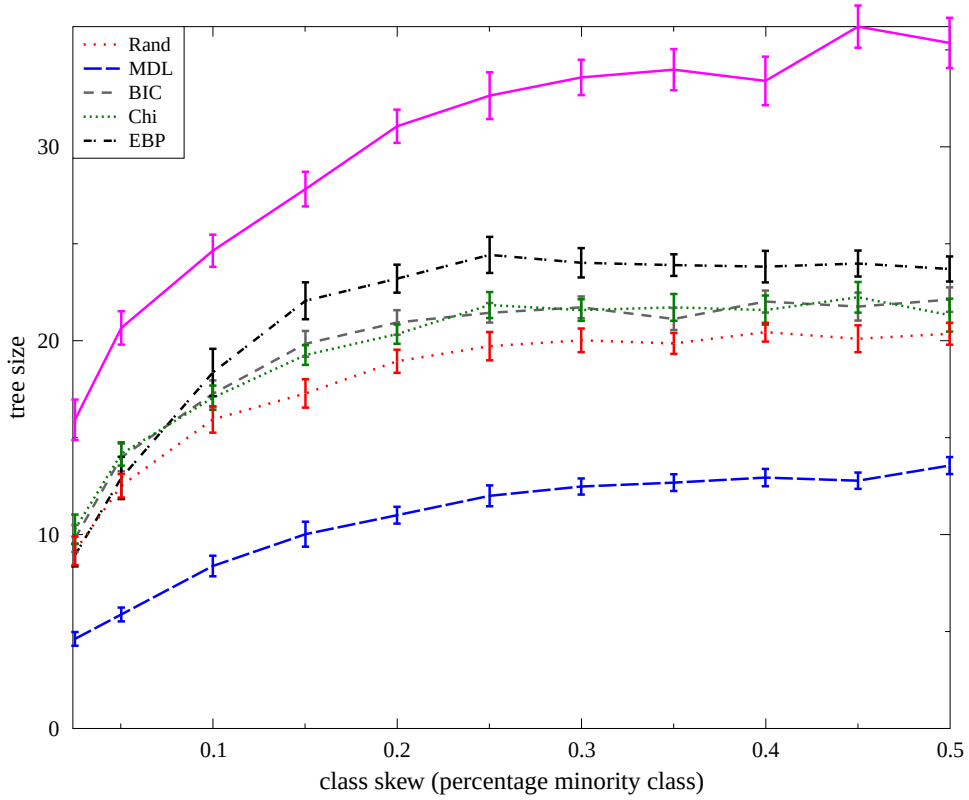

Figure 60: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

4.4 'Diabetes' Dataset

In the experiment on the 'diabetes' dataset we varied the minority class percentage from 50% (balanced) to 2.5% (strongly skewed). We always used 512 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 4.1.1).

4.4.1 AUC

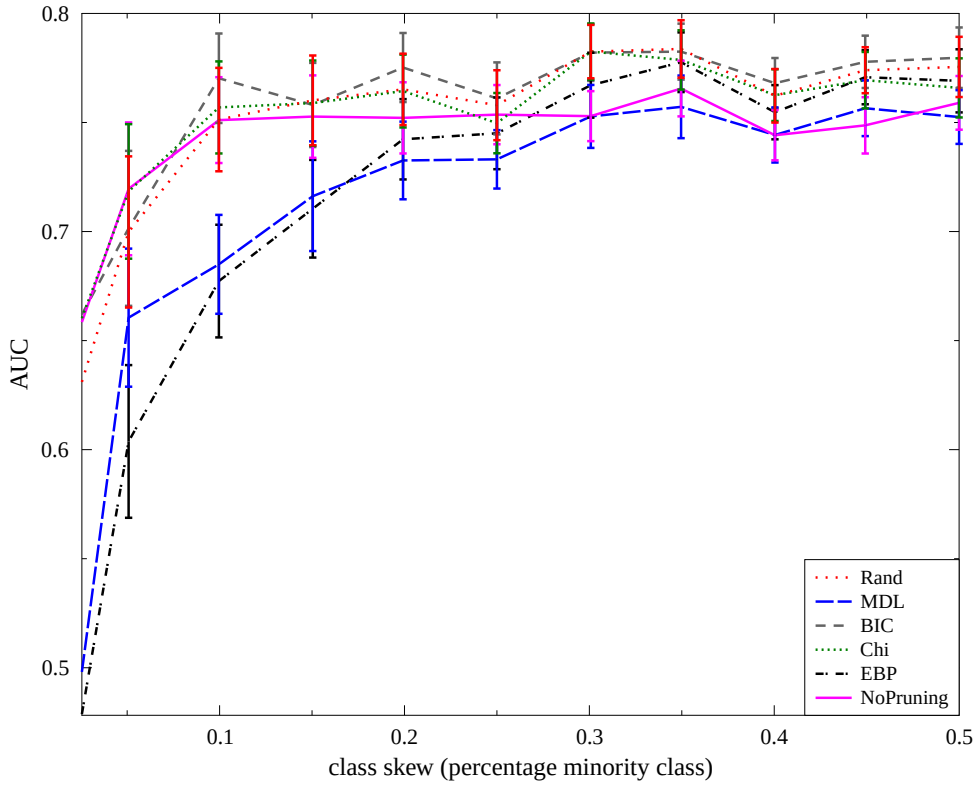

Figure 61: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.4.2 RMSE

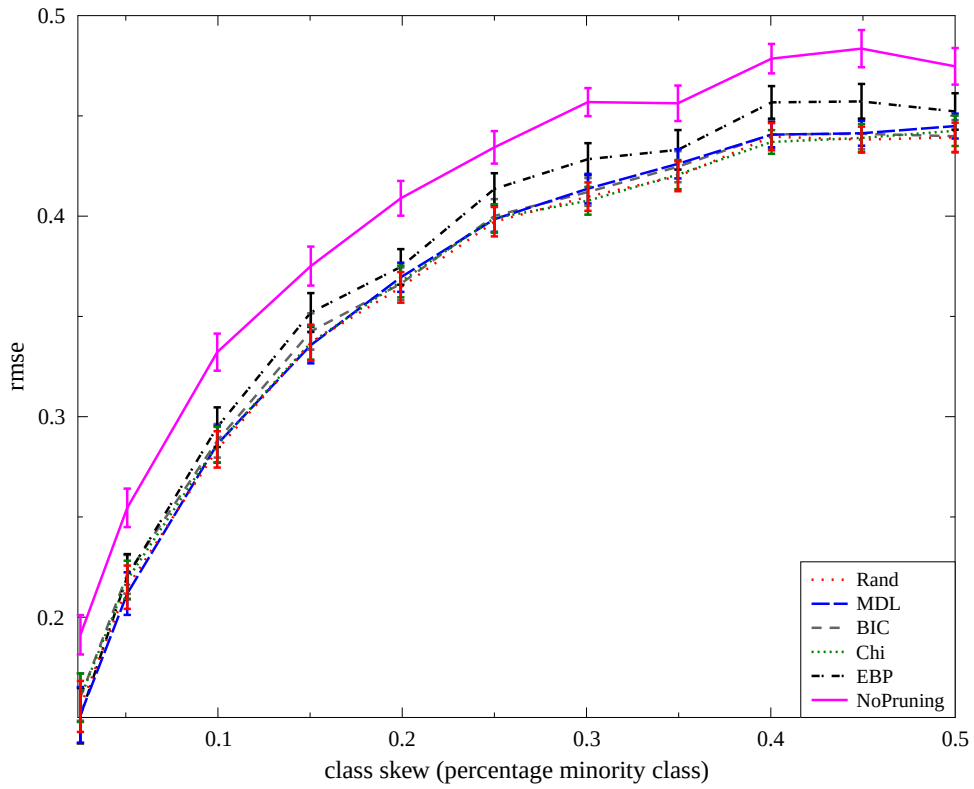

Figure 62: Influence of class skew on the performance measure RMSE (lower is better).

4.4.3 CLL

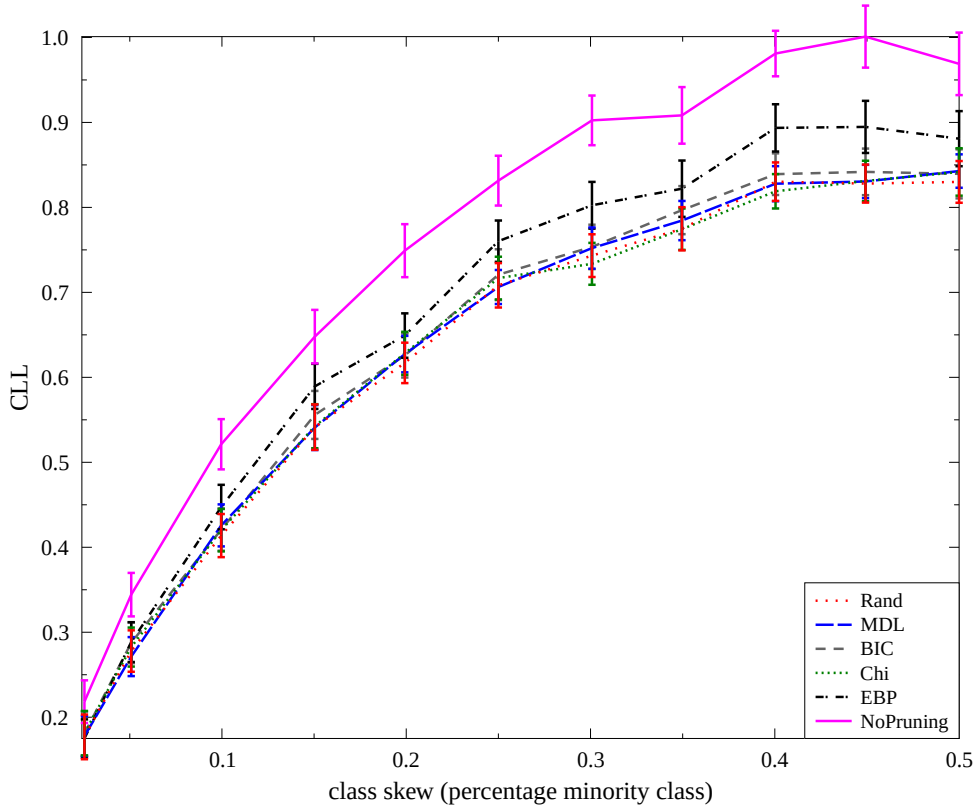

Figure 63: Influence of class skew on the performance measure CLL (lower is better).

4.4.4 Calibration Error

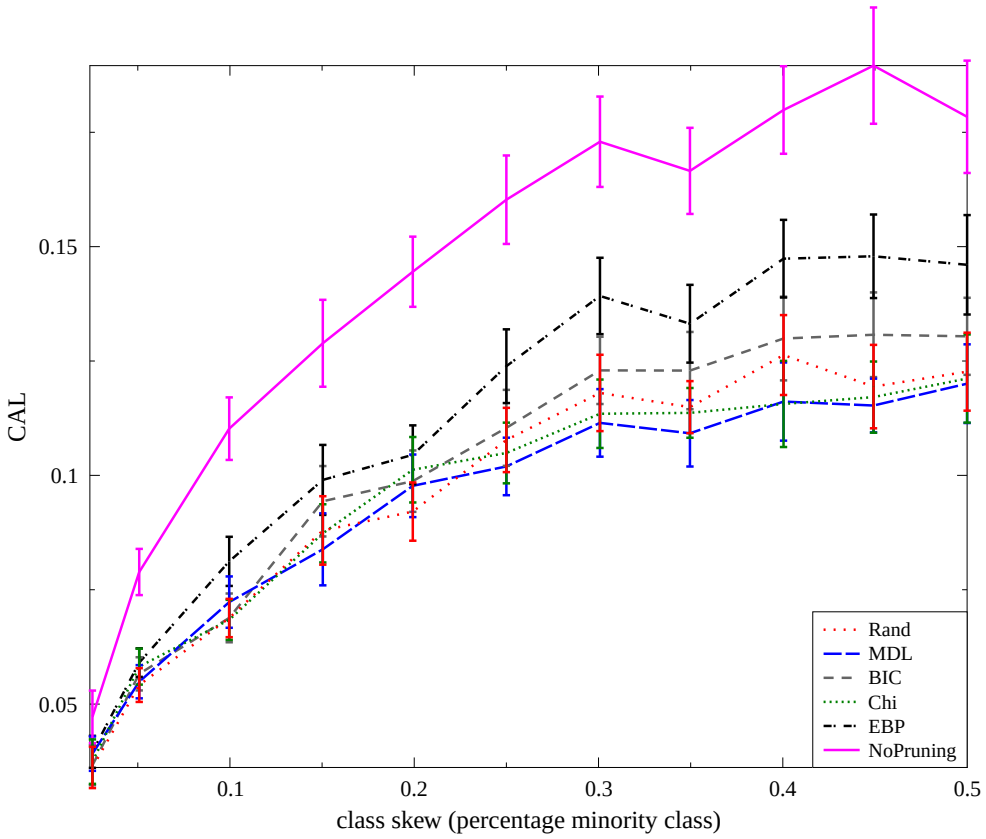

Figure 64: Influence of class skew on the performance measure CAL (lower is better).

4.4.5 Classification Accuracy

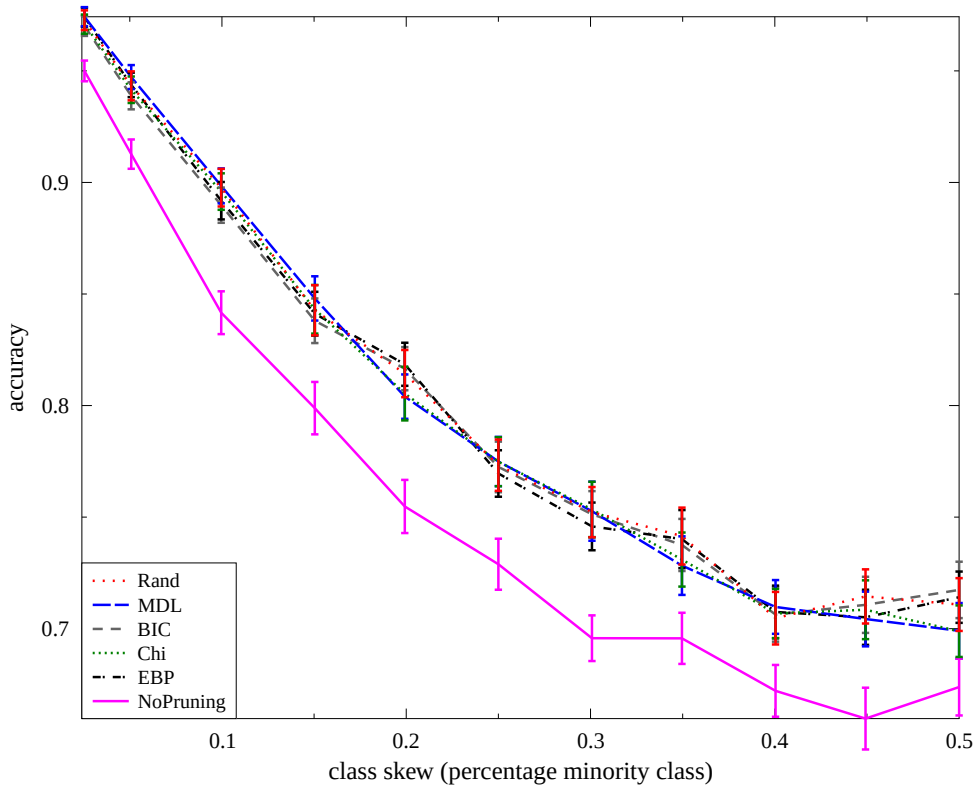

Figure 65: Influence of class skew on the performance measure accuracy (higher is better).

4.4.6 Tree Size

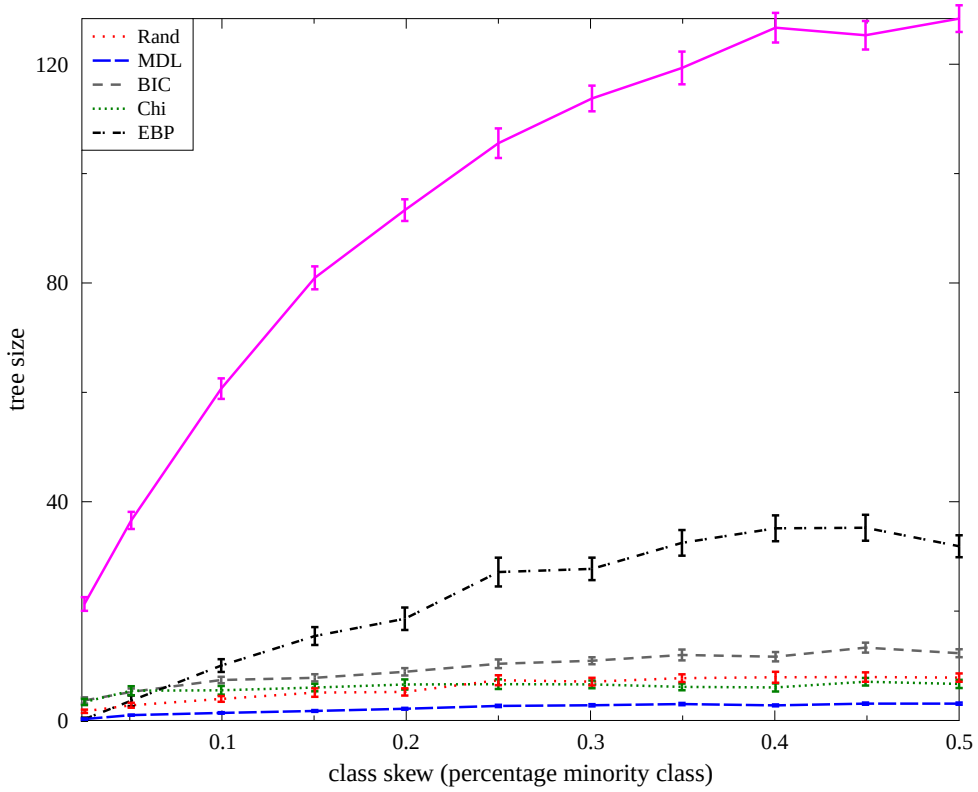

Figure 66: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

4.5 'German Credit' Dataset'

In the experiment on the 'german credit' dataset we varied the minority class percentage from 50% (balanced) to 2.5% (strongly skewed). We always used 600 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 4.1.1).

4.5.1 AUC

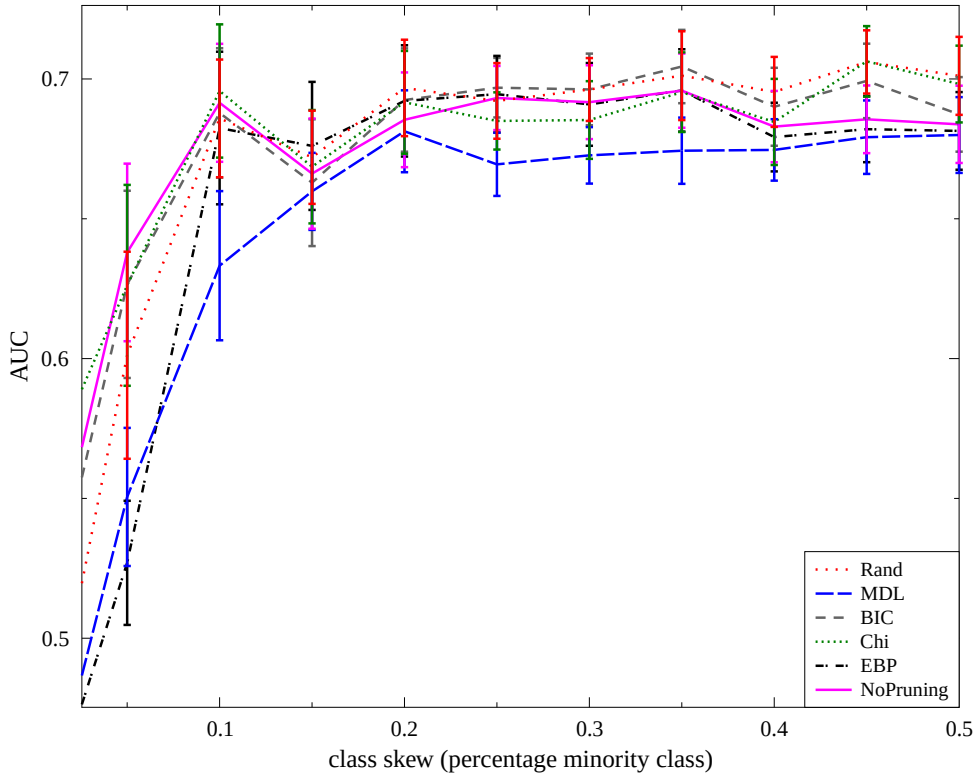

Figure 67: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.5.2 RMSE

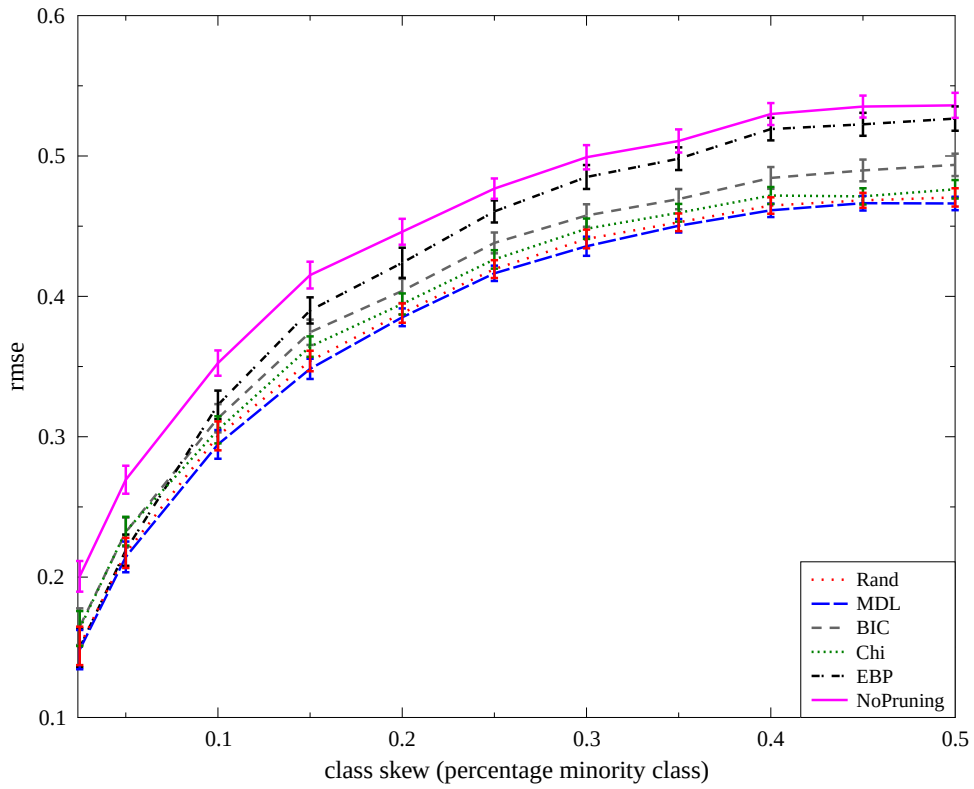

Figure 68: Influence of class skew on the performance measure RMSE (lower is better).

4.5.3 CLL

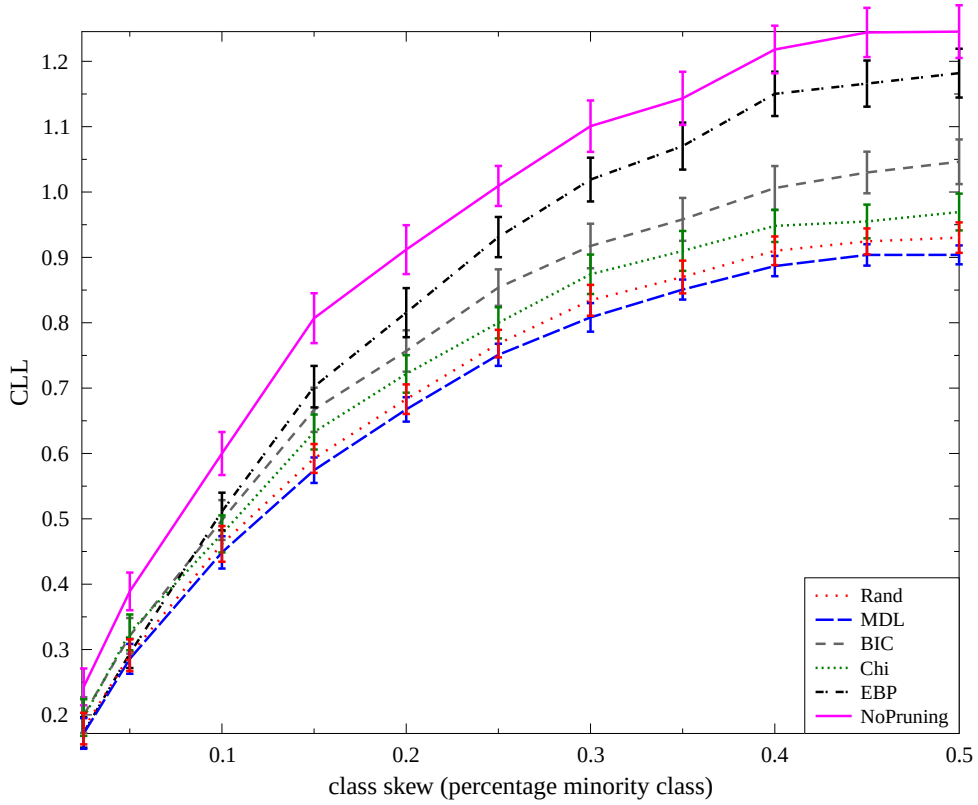

Figure 69: Influence of class skew on the performance measure CLL (lower is better).

4.5.4 Calibration Error

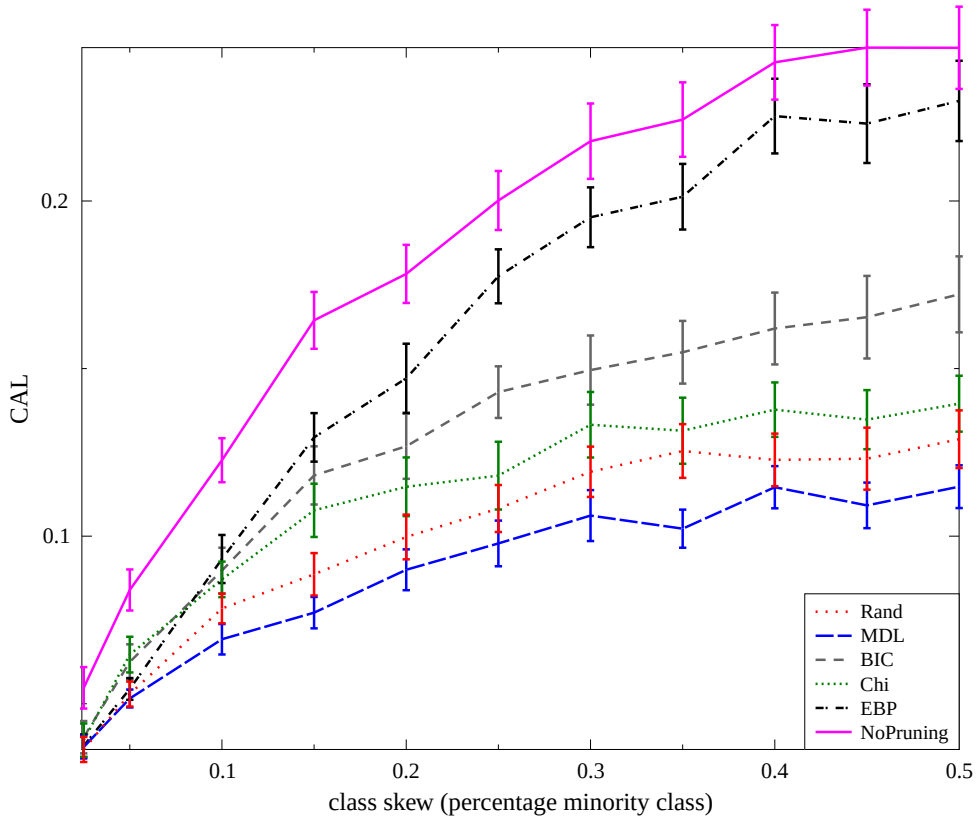

Figure 70: Influence of class skew on the performance measure CAL (lower is better).

4.5.5 Classification Accuracy

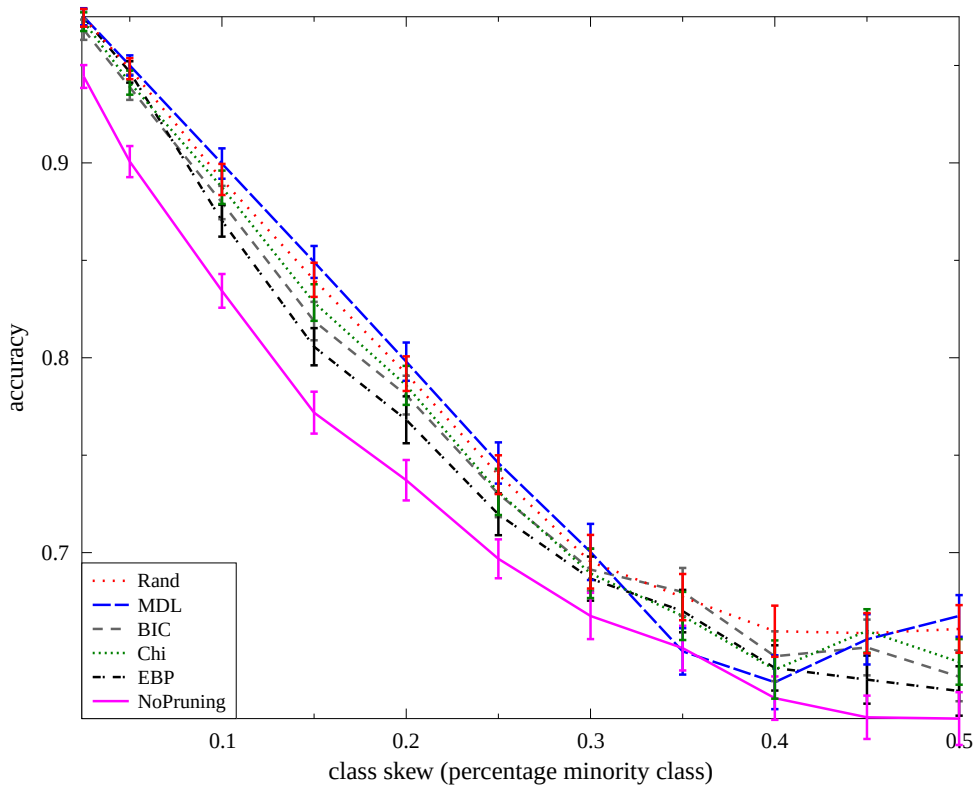

Figure 71: Influence of class skew on the performance measure accuracy (higher is better).

4.5.6 Tree Size

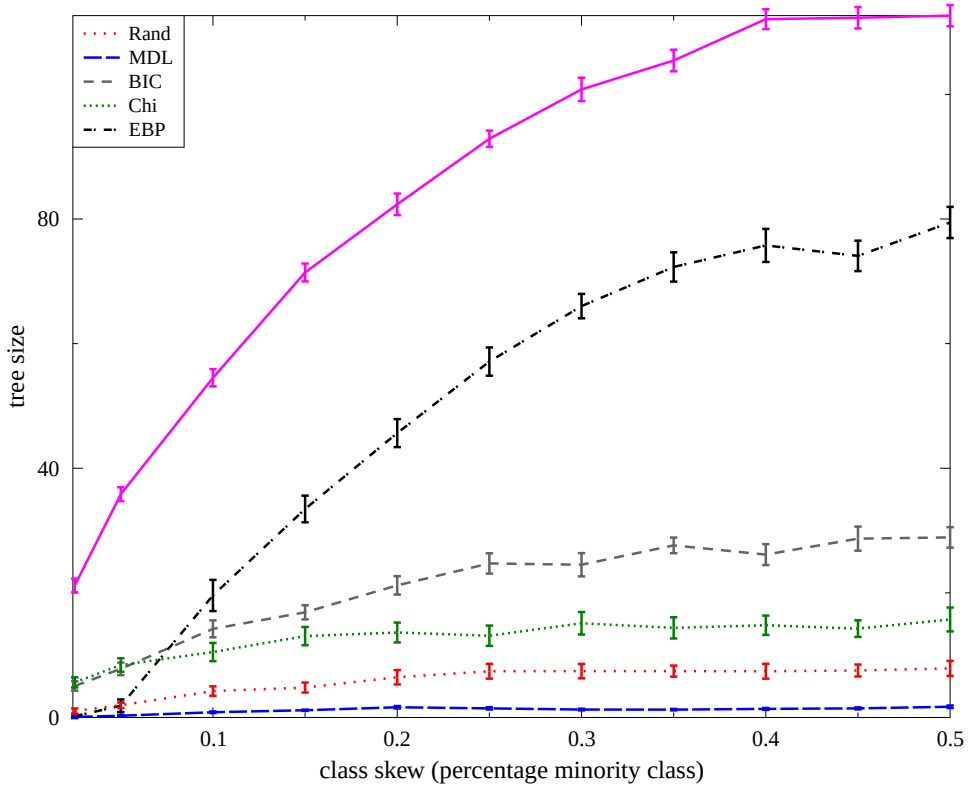

Figure 72: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

4.6 'Hiv' Dataset

4.6.1 Experimental Setup

In the experiment on the 'hiv' dataset we varied the minority class percentage from 3.6% (as in the original dataset) to 50% (balanced). We always used 2972 examples (the way in which we determined this number, and the rest of the experimental setup, is the same as discussed before in Section 4.1.1).

4.6.2 AUC

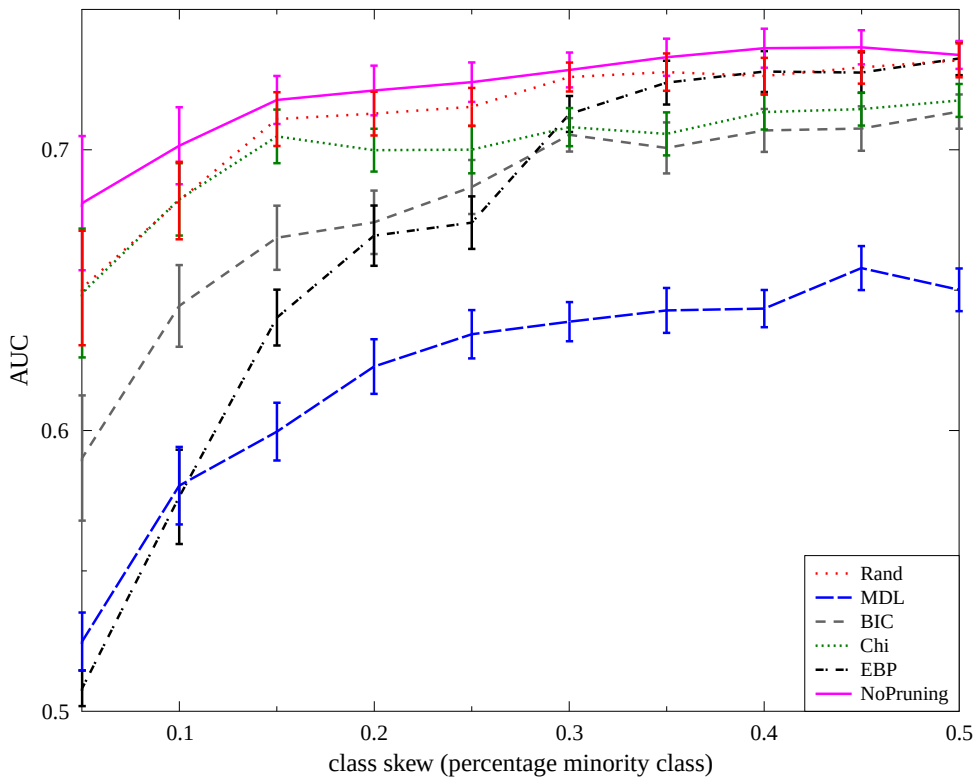

Figure 73: Influence of class skew on the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

4.6.3 RMSE

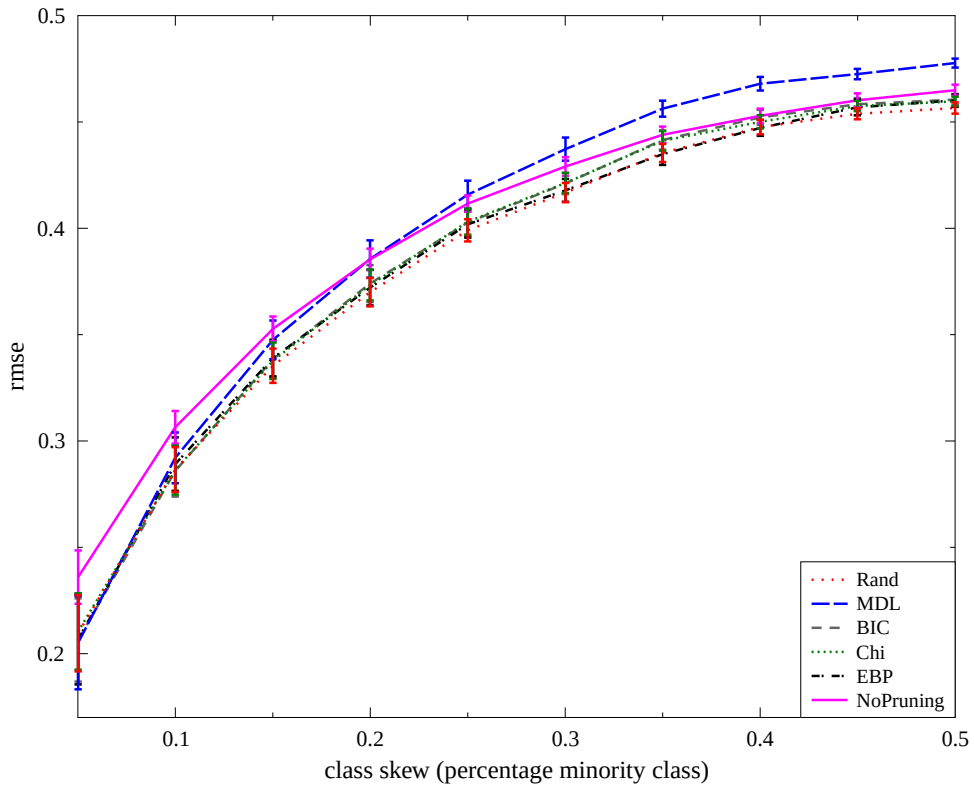

Figure 74: Influence of class skew on the performance measure RMSE (lower is better).

4.6.4 CLL

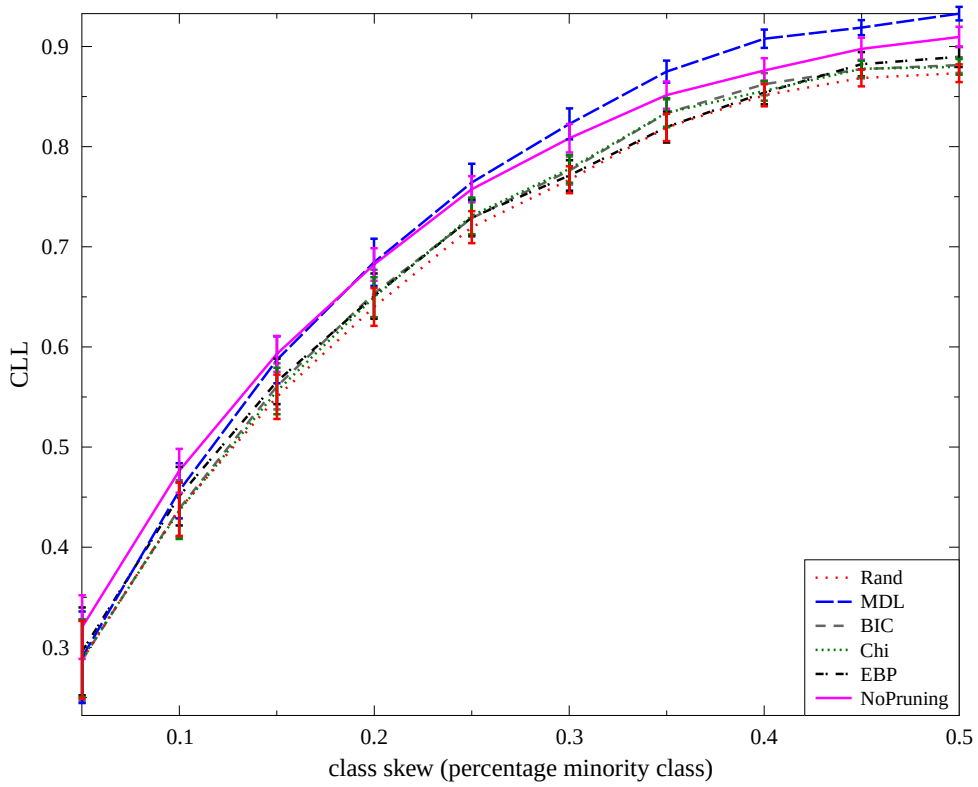

Figure 75: Influence of class skew on the performance measure CLL (lower is better).

4.6.5 Calibration Error

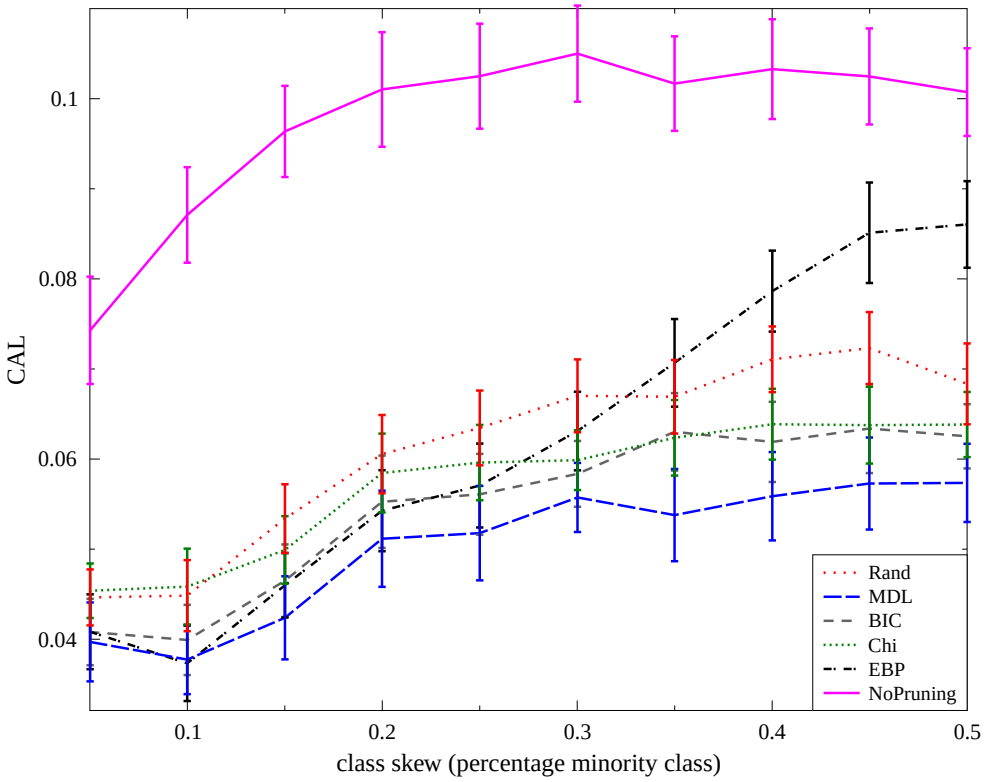

Figure 76: Influence of class skew on the performance measure CAL (lower is better).

4.6.6 Classification Accuracy

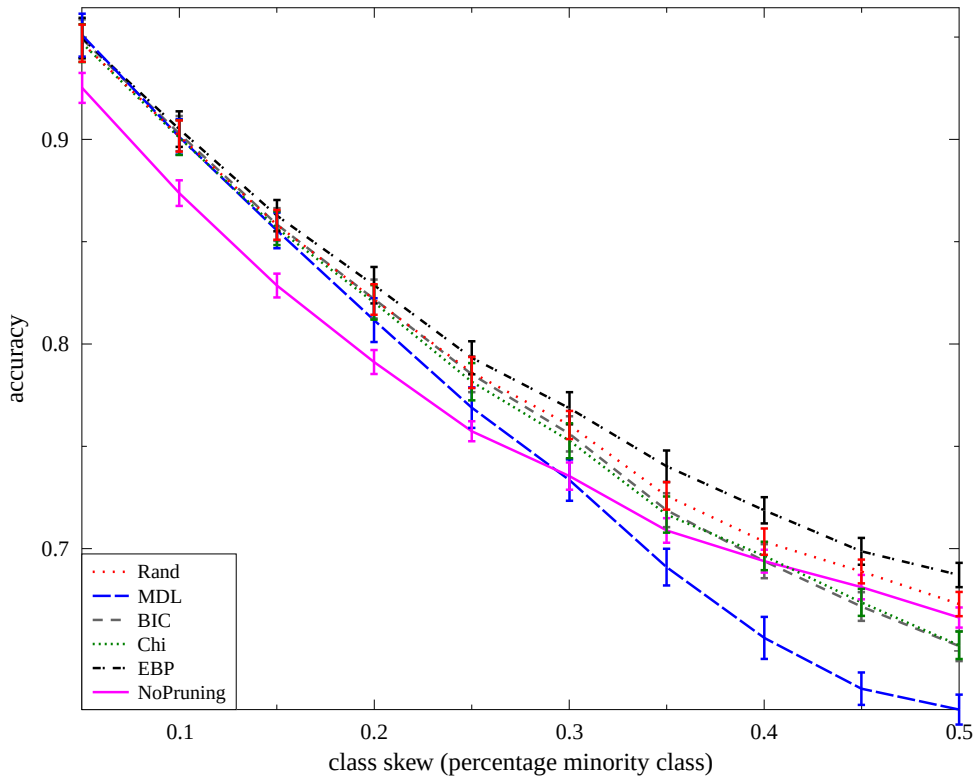

Figure 77: Influence of class skew on the performance measure accuracy (higher is better).

4.6.7 Tree Size

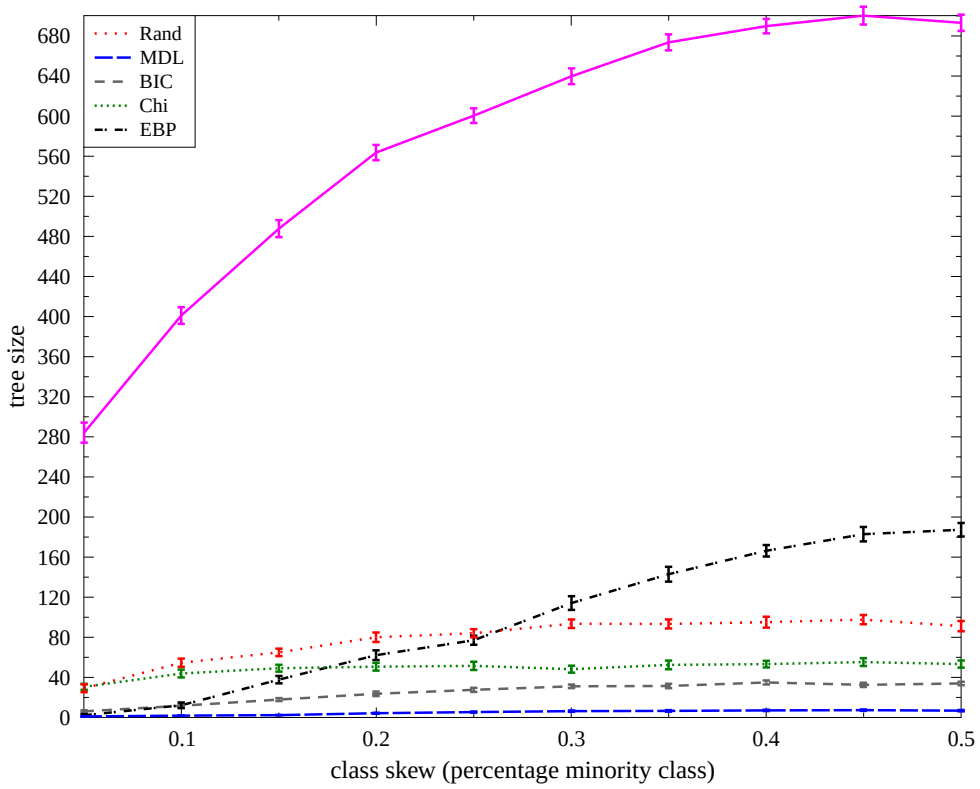

Figure 78: Influence of class skew on tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

5 Influence of the Number of Examples - Learning Curves

To investigate the influence of the number of examples on the performance of the various pruning criteria, we constructed learning curves on the 'hiv' dataset. We choose 'hiv' because it is by far the largest dataset in our study (41768 examples).

5.1 'Hiv' Dataset

We set up our experiments such that all the subsets of the dataset that we use have the same class distribution as the original dataset (i.e., 3.6% positives).

5.1.1 AUC

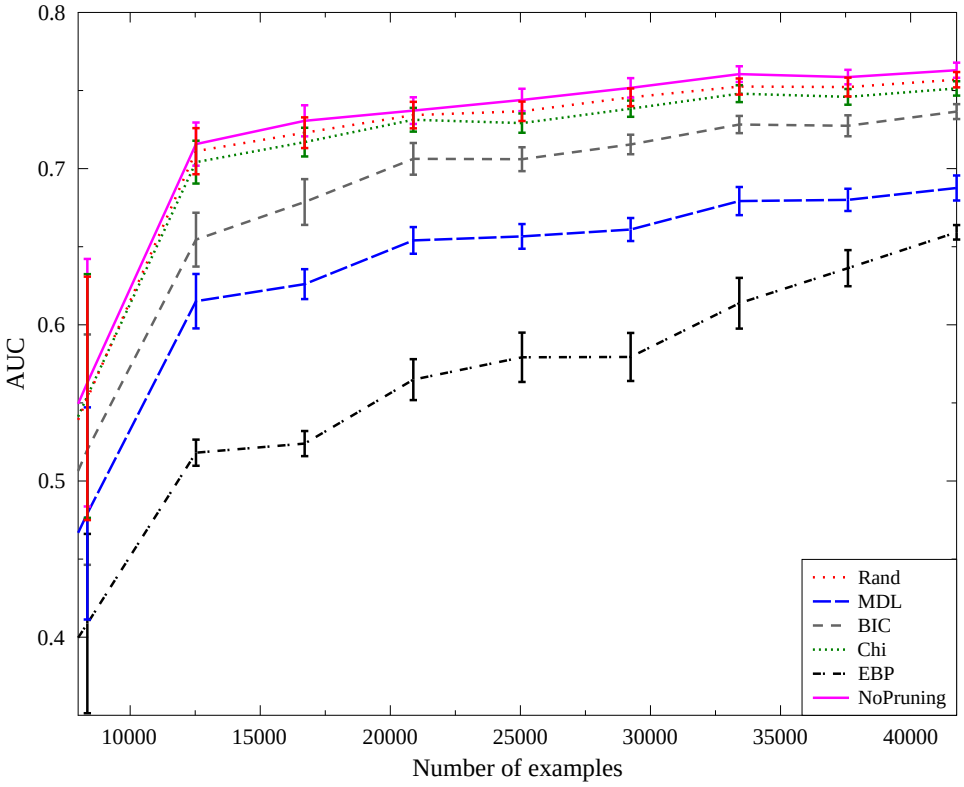

Figure 79: Learning curves for the performance measure AUC (higher is better; bars indicate 95% confidence intervals).

5.1.2 RMSE

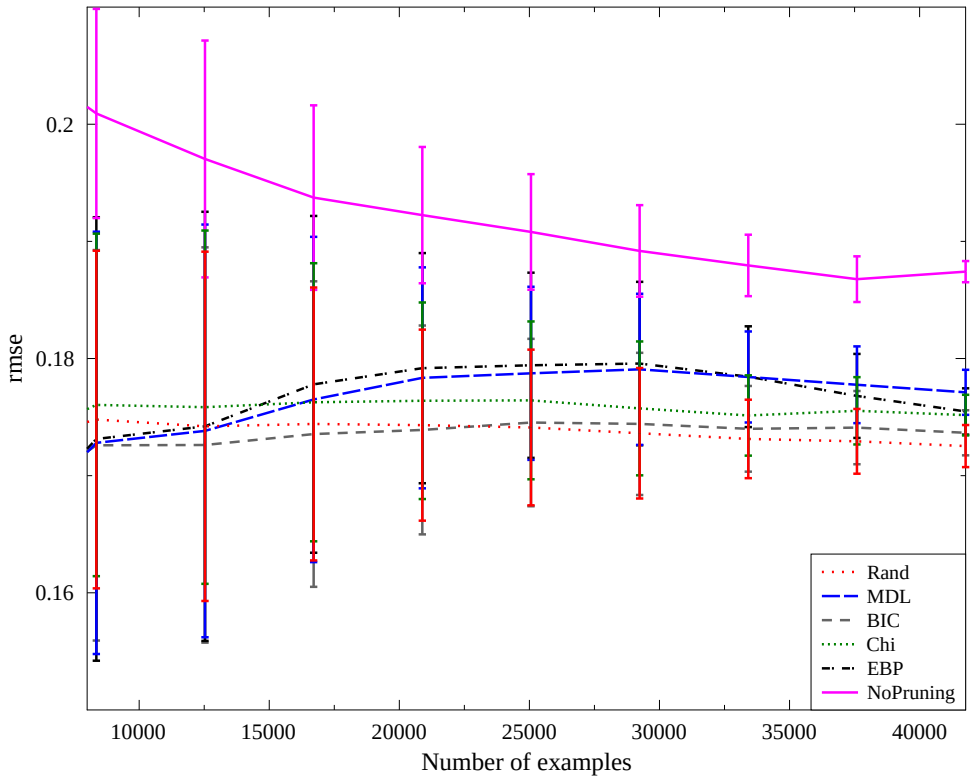

Figure 80: Learning curves for the performance measure RMSE (lower is better).

5.1.3 CLL

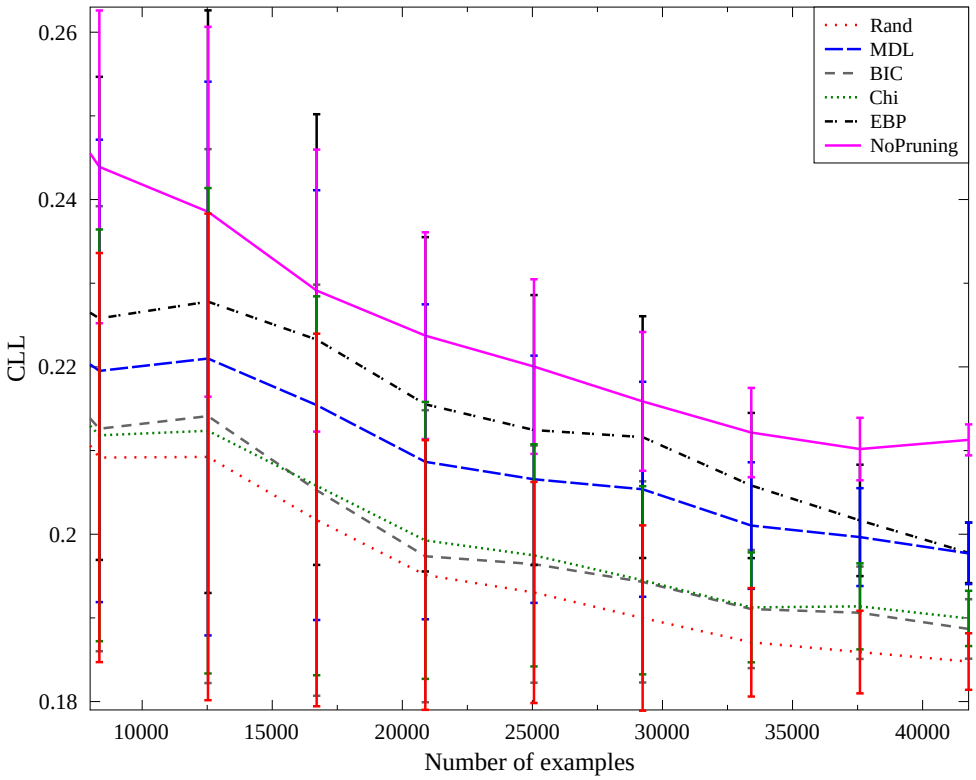

Figure 81: Learning curves for the performance measure CLL (lower is better). The large error bars indicate that CLL is a very unstable measure on this dataset.

5.1.4 Classification Accuracy

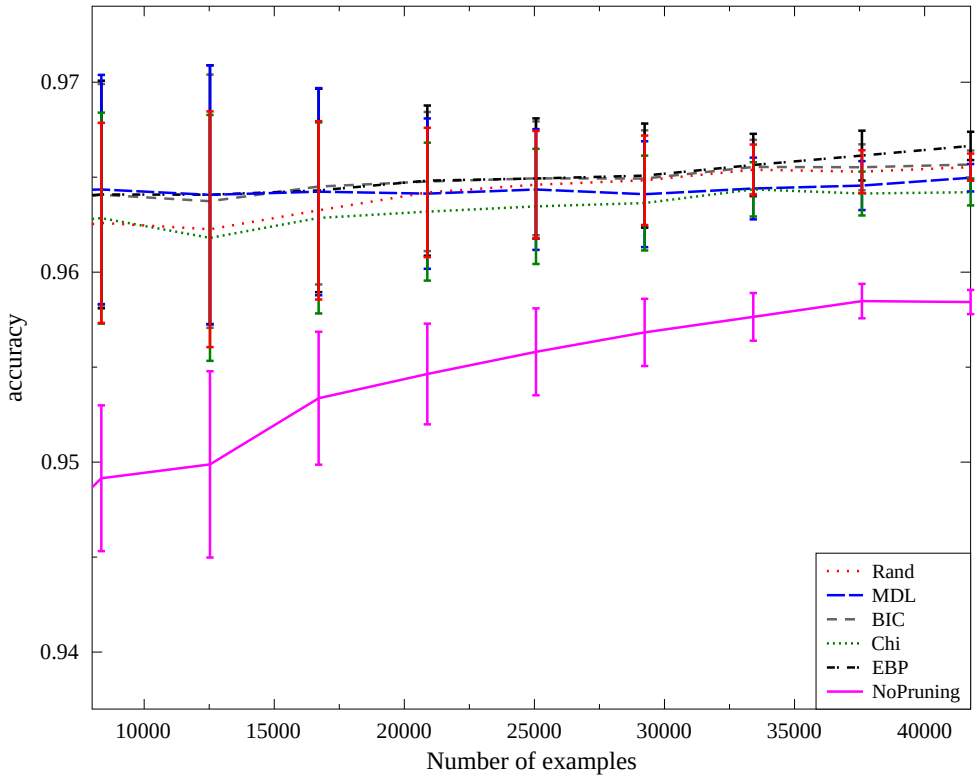

Figure 82: Learning curves for the performance measure accuracy (higher is better).

5.1.5 Tree Size

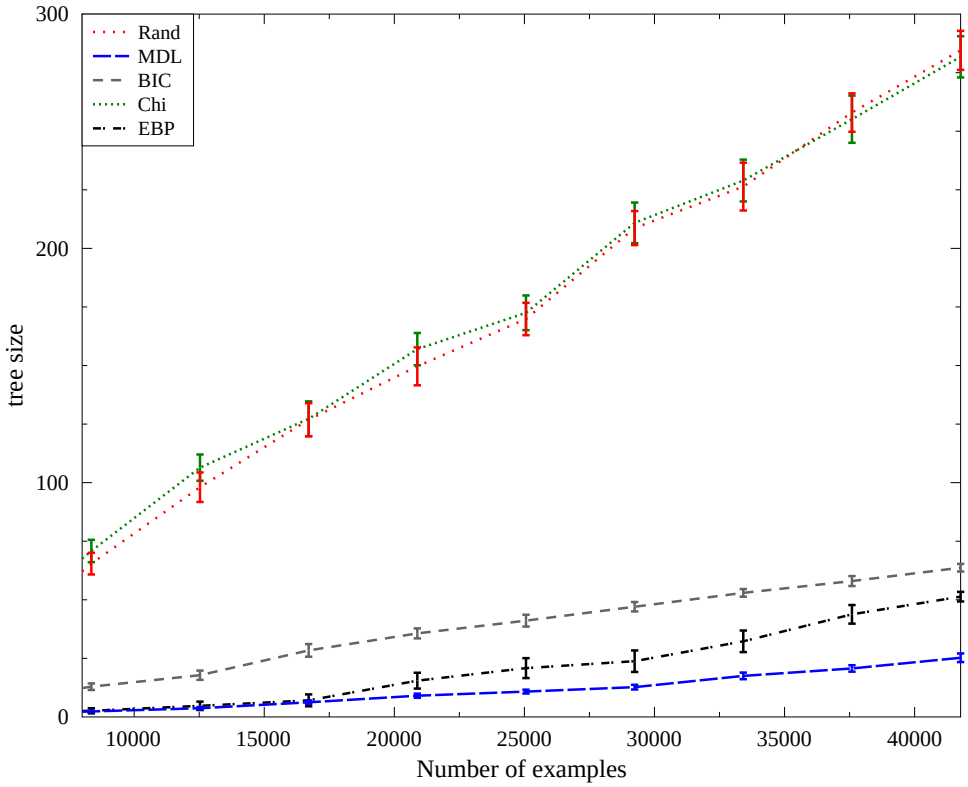

Figure 83: Learning curves in terms of tree size (lower is better). We do not show tree size for NoPruning because it is too high (see next figure).

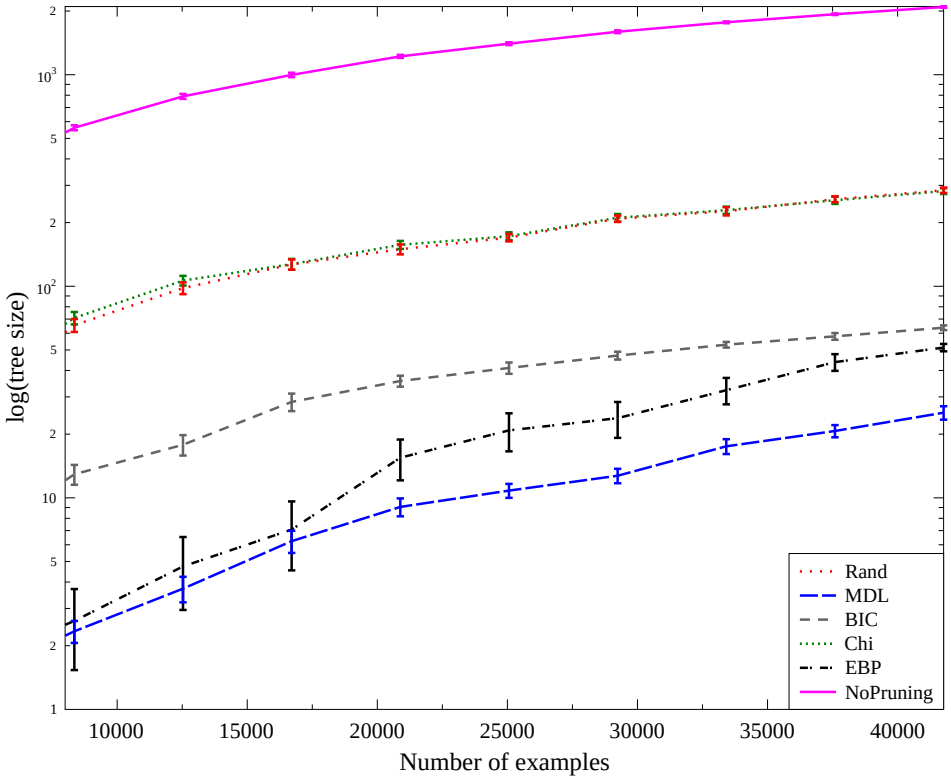

Figure 84: Learning curves in terms of tree size, on a logarithmic axis.