A Detailed Tables of Results

Dataset	(#F	#C	#I)	BTs		SMT				MaxSAT					
				D	%A	Enc		max	min	avg	Enc		max	min	avg	c	%w
ann-thyroid	(21	3	200)	3	100	(213	504)	2.78	0.04	0.13	(566	2780)	0.49	0.06	0.11	3	20
appendicitis	(7	2	106)	3	91	(157	428)	0.44	0.02	0.10	(475	2348)	0.11	0.03	0.04	7	80
biodegradation	(41	2	200)	3	87	(352	914)	234.08	0.33	14.70	(1331	5990)	5.28	1.02	2.31	29	69
divorce	(54	2	150)	3	100	(132	186)	0.39	0.04	0.05	(107	602)	0.02	0.01	0.01	8	100
ecoli	(7	5	200)	3	85	(445	1499)	8.68	0.16	1.39	(2189	11 404)	1.30	0.33	0.84	6	63
glass2	(9	2	162)	4	88	(217	699)	2.39	0.03	0.35	(927	4550)	0.63	0.13	0.22	8	58
ionosphere	(34	2	200)	3	93	(267	650)	22.65	0.28	2.20	(886	3932)	1.04	0.28	0.62	25	82
pendigits	(16	10	110)	3	99	(1470	4499)	125.78	6.48	29.45	(8037	60 198)	5.57	2.75	3.99	17	100
promoters	(58	2	106)	3	100	(104	129)	0.08	0.03	0.03	(4	304)	0.00	0.00	0.00	1	100
segmentation	(19	7	200)	3	95	(592	1175)	22.98	0.45	3.55	(1420	10 187)	1.62	0.15	0.67	16	100
shuttle	(9	7	200)	3	100	(509	1358)	4.52	0.20	0.58	(1498	10 081)	0.47	0.10	0.30	6	77
sonar	(60	2	200)	3	86	(295	631)	12.50	0.31	2.01	(847	3652)	1.03	0.46	0.69	33	96
spambase	(57	2	200)	4	96	(489	1338)	439.87	1.61	42.78	(2015	9652)	41.87	3.03	10.13	41	81
texture	(40	11	200)	3	98	(2073	4576)	1481.71	12.77	150.86	(8093	82 625)	40.57	9.77	24.68	37	96
threeOf9	(9	2	200)	3	100	(108	153)	0.02	0.01	0.01	(10	392)	0.00	0.00	0.00	1	100
twonorm	(20	2	200)	3	97	(463	1135)	52.86	0.20	4.27	(1750	7844)	1.97	0.96	1.49	20	70
vowel	(13	11	200)	4	93	(2292	5771)	464.07	6.53	60.50	(10 183	102 611)	17.62	6.31	11.94	13	97
wdbc	(30	2	200)	4	97	(269	654)	2.85	0.20	0.66	(894	4060)	0.58	0.29	0.42	22	80
wine-recognition	(13	3	178)	3	97	(241	454)	0.43	0.04	0.11	(491	2468)	0.14	0.05	0.09	9	54
wpbc	(33	2	194)	4	74	(302	784)	33.54	0.34	5.24	(1101	5090)	5.15	0.44	1.69	25	85
zoo	(16	7	59)	4	83	(386	651)	2.07	0.20	0.63	(196	2157)	0.08	0.01	0.02	8	100

Table 1: Detailed performance evaluation of computing AXps for BTs. Columns **#F**, **#C** and **#I** report, respectively, the number of features, number of classes and the number of tested instances, in the dataset. (Note that for each dataset, we randomly pick 200 instances to be tested, and if a dataset has a fewer number of instances, we use all available instances.) Columns **D** and **%A** report, respectively, the maximum tree depth and test accuracy of the trained BT. Sub-columns **max**, **min** and **avg** of column **SMT** (resp., **MaxSAT**) show, respectively, the maximum, minimum and average time in second to find an explanation. Sub-column **Enc** reports the SMT (resp. MaxSAT) encoding size: number of variables and number of asserts (clauses for MaxSAT). Sub-column **c** reports the average number of entailment oracle calls. The percentage of won instances by the MaxSAT approach is given as **%w**.

Dataset	(#F	#C	#I)	BTs		Anchor			MaxSAT				
				D	%A	max	min	avg	max	min	avg	c	%w
ann-thyroid	(21	3	200)	3	100	23.46	0.38	2.78	0.49	0.06	0.11	3	100
appendicitis	(7	2	106)	3	91	3.06	0.89	1.33	0.11	0.03	0.04	7	100
biodegradation	(41	2	200)	3	87	65.99	4.43	8.47	5.28	1.02	2.31	29	100
divorce	(54	2	150)	3	100	24.26	13.49	19.27	0.02	0.01	0.01	8	100
ecoli	(7	5	200)	3	85	2.59	0.72	1.41	1.30	0.33	0.84	6	64
glass2	(9	2	162)	4	88	5.18	1.01	2.08	0.63	0.13	0.22	8	100
ionosphere	(34	2	200)	3	93	25.23	2.86	7.51	1.04	0.28	0.62	25	100
pendigits	(16	10	110)	3	99	51.36	4.05	13.92	5.57	2.75	3.99	17	99
promoters	(58	2	106)	3	100	3.20	1.94	2.42	0.00	0.00	0.00	1	100
segmentation	(19	7	200)	3	95	23.62	1.57	4.86	1.62	0.15	0.67	16	100
shuttle	(9	7	200)	3	100	53.88	0.91	7.85	0.47	0.10	0.30	6	100
sonar	(60	2	200)	3	86	117.02	20.34	26.19	1.03	0.46	0.69	33	100
spambase	(57	2	200)	4	96	27.26	3.02	6.74	41.87	3.03	10.13	41	46
texture	(40	11	200)	3	98	507.89	9.96	51.46	40.57	9.77	24.68	37	52
threeOf9	(9	2	200)	3	100	1.03	0.67	0.95	0.00	0.00	0.00	1	100
twonorm	(20	2	200)	3	97	30.43	6.07	14.88	1.97	0.96	1.49	20	100
vowel	(13	11	200)	4	93	21.42	3.68	11.85	17.62	6.31	11.94	13	42
wdbc	(30	2	200)	4	97	15.62	4.66	8.55	0.58	0.29	0.42	22	100
wine-recognition	(13	3	178)	3	97	3.45	0.55	1.75	0.14	0.05	0.09	9	100
wpbc	(33	2	194)	4	74	128.65	1.06	6.33	5.15	0.44	1.69	25	87
zoo	(16	7	59)	4	83	2.38	0.96	1.54	0.08	0.01	0.02	8	100

Table 2: Detailed performance evaluation of computing AXps and Anchors for BTs. (Note that for Anchor tool all the parameters are kept in their default setting.) Columns **#F**, **#C** and **#I** report, respectively, the number of features, number of classes and the number of tested instances, in the dataset. (Note that for each dataset, we randomly pick 200 instances to be tested, and if a dataset has a fewer number of instances, we use all available instances.) Columns **D** and **%A** report, respectively, the maximum tree depth and test accuracy of the trained BT. Sub-columns **max**, **min** and **avg** of column **Anchor** (resp., **MaxSAT**) show, respectively, the maximum, minimum and average time in second to find an explanation. Sub-column **c** reports the average number of entailment oracle calls. The percentage of won instances by the MaxSAT approach is given as **%w**.