A Quantitative Study of Small Disjuncts: Experiments and Results*

Gary M. Weiss and Haym Hirsh

Department of Computer Science
Rutgers University
New Brunswick, New Jersey 08903
{gweiss, hirsh}@cs.rutgers.edu

July 14, 2000

Abstract

Systems that learn from examples often express the learned concept in the form of a disjunctive description. Disjuncts that correctly classify few training examples are known as small disjuncts and are interesting to machine learning researchers because they have a much higher error rate than large disjuncts. Previous research has investigated this phenomenon by performing *ad hoc* analyses of a small number of datasets. In this paper we present a quantitative measure for evaluating the effect of small disjuncts on learning and use it to analyze 30 benchmark datasets. We investigate the relationship between small disjuncts and pruning, training set size and noise, and come up with several interesting results.

1 INTRODUCTION

Systems that learn from examples often express the learned concept as a disjunction. The size of a disjunct is defined as the number of training examples that it correctly classifies (Holte, Acker, and Porter 1989). A number of empirical studies have demonstrated that learned concepts include disjuncts that span a large range of disjunct sizes and that the small disjuncts—those disjuncts that correctly classify only a few training examples—collectively cover a significant percentage of the test examples (Holte, Acker, and Porter 1989; Ali and Pazzani 1992; Danyluk and Provost 1993; Ting 1994; Van den Bosch et al. 1997; Weiss and Hirsh 1998). It has also been shown that small disjuncts often correspond to rare cases within the domain under study (Weiss 1995) and cannot be totally eliminated if high predictive accuracy is to be achieved (Holte et al. 1989). Previous studies have shown that small disjuncts have much higher error rates than large disjuncts and contribute a disproportionate number of the total errors. This phenomenon is known as "the problem with small disjuncts".

There are two reasons for studying the problem with small disjuncts. The first is that small disjuncts can help us answer important machine learning questions, such as: how does the amount of available training data affect learning, how does pruning work and when is it most effective, and how does noise affect the ability to learn a concept? Thus, we use small disjuncts as a lens through which to examine important issues in machine learning. The second reason for studying small disjuncts is to learn to build machine learning programs that "address" the problem with small disjuncts. These learners will improve the accuracy of the small disjuncts without significantly decreasing the accuracy of the large disjuncts, so that the overall accuracy of the learned concept is improved. Several researchers have attempted to build such learners. One approach involves employing a maximum specificity bias for

* This is an expanded version of the paper "A Quantitative Study of Small Disjuncts" that was presented at the Seventeenth National Conference on Artificial Intelligence (AAAI-2000) in Austin Texas.

learning small disjuncts, while continuing to use the more common maximum generality bias for the large disjuncts (Holte et al. 1989; Ting 1994). Unfortunately, these efforts have produced, at best, only marginal improvements. A better understanding of small disjuncts and their role in learning may be required before further advances are possible.

In this paper we use small disjuncts to gain a better understanding of machine learning. In the process of doing this, we address a major limitation with previous research—that very few datasets were analyzed: Holte et al. (1989) analyzed two datasets, Ali and Pazzani (1992) one dataset, Danyluk and Provost (1993) one dataset, and Weiss and Hirsh (1998) two datasets. Because so few datasets were analyzed, only relatively weak qualitative conclusions were possible. By analyzing thirty datasets, we are able to draw some quantitative conclusions, as well as form more definitive qualitative conclusions than previously possible.

For those readers who would like more information on small disjuncts, a brief survey of the research on this topic is provided at: http://www.cs.rutgers.edu/~gweiss/small_disjuncts.html.

2 DESCRIPTION OF EXPERIMENTS

The results presented in this paper are based on 30 datasets, of which 19 were collected from the UCI repository (Blake and Merz 1998) and 11 from researchers at AT&T (Cohen 1995; Cohen and Singer 1999). Numerous experiments were run on these datasets to assess the impact of small disjuncts on learning, especially as factors such as training set size, pruning strategy, and noise level are varied. The majority of experiments use C4.5, a program for inducing decision trees (Quinlan 1993). C4.5 was modified by the authors to collect information related to disjunct size. During the training phase the modified software assigns each disjunct/leaf a value based on the number of training examples it correctly classifies. The number of correctly and incorrectly classified examples associated with each disjunct is then tracked during the testing phase, so that at the end the distribution of correctly/incorrectly classified test examples by disjunct size is known. For example, the software might record the fact that disjuncts of size 3 collectively classify 5 test examples correctly and 3 incorrectly. Some experiments were repeated with RIPPER, a program for inducing rule sets (Cohen 1995), in order to assess the generality of our results.

Since pruning eliminates many small disjuncts, consistent with what has been done previously, pruning is disabled for C4.5 and RIPPER for most experiments (as is seen later, however, the same trends are seen even when pruning is not disabled). C4.5 is also run with the –m1 option, to ensure that nodes continue to be split until they only contain examples of a single class, and RIPPER is configured to produce unordered rules so that it does not produce a single default rule to cover the majority class. All experiments employ 10-fold cross validation and the results are therefore based on averages of the test set calculated over 10 runs. Unless specified otherwise, all results are based on C4.5 without pruning.

3 AN EXAMPLE: THE VOTE DATASET

In order to illustrate the problem with small disjuncts and introduce a way of measuring this problem, we examine the concept learned by C4.5 from the Vote dataset. Figure 1 shows how the correctly and incorrectly classified test examples are distributed across the disjuncts in this concept. Each bin in the figure spans 10 sizes of disjuncts. The leftmost bin shows that those disjuncts that classify 0-9 training examples correctly cover 9.5 test examples, of which 7.1 are classified correctly and 2.4 classified incorrectly. The fractional values occur because the results are averaged over 10 cross-validated runs. Disjuncts of size 0 occur because when C4.5 splits a node using a feature *f*, the split uses all possible feature values, whether or not the value occurs within the training examples at that node.

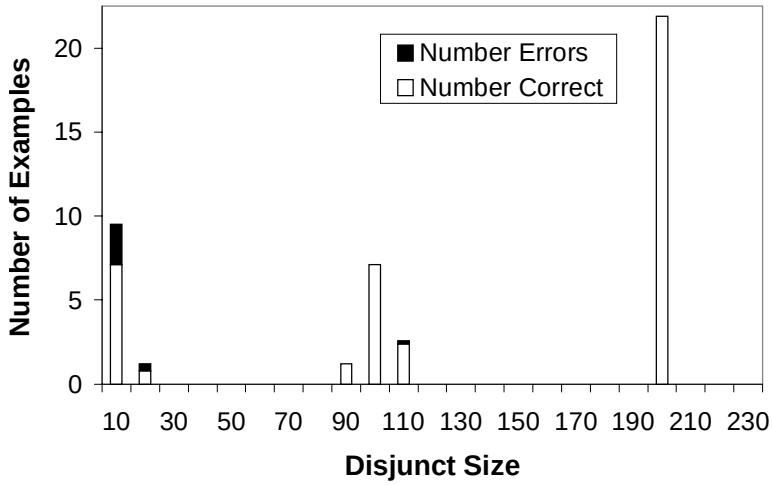


Figure 1: Distribution of Examples for Vote Dataset

Figure 1 clearly shows that the errors are concentrated toward the smaller disjuncts. Analysis at a finer level of granularity shows that the errors are skewed even more toward the small disjuncts—75% of the errors in the leftmost bin come from disjuncts of size 0 and 1. Those readers interested in seeing the distribution of correctly and incorrectly classified examples using a bin size of 1 should refer to Appendix F, Figures F1.1.3 and F1.1.4. One may also be interested in the distribution of disjuncts, as opposed to the distribution of examples. As it turns out, of the 50 disjuncts that make up the learned concept, 45 of them are associated with the leftmost bin (i.e. have a disjunct size less than 10). The actual distribution of disjuncts is shown in Appendix F, Figure F1.1.2.

The data may also be described using a new measure, *mean disjunct size*. This measure is computed over a set of examples as follows: each example is assigned a value equal to the size of the disjunct that classifies it, and then the mean of these values is calculated. For the concept shown in Figure 1, the mean disjunct size over all test examples is 124—one can also view this as the center of mass of the bins in the figure. The mean disjunct size for the incorrectly (correctly) classified test examples is 10 (133). Since 10 << 133, the errors are heavily concentrated toward the smaller disjuncts.

In order to better show the degree to which errors are concentrated toward the small disjuncts, we plot, for each disjunct size *n*, the percentage of test errors versus percentage of correctly classified test examples covered by disjuncts of size *n* or less. Figure 2 shows this plot for the concept induced from the Vote dataset. It shows, for example, that disjuncts with size 0-4 contribute 5.1% of the correctly classified test examples but 73% of the total test errors. Since the curve in Figure 2 is above the line Y=X, the errors are concentrated toward the smaller disjuncts.

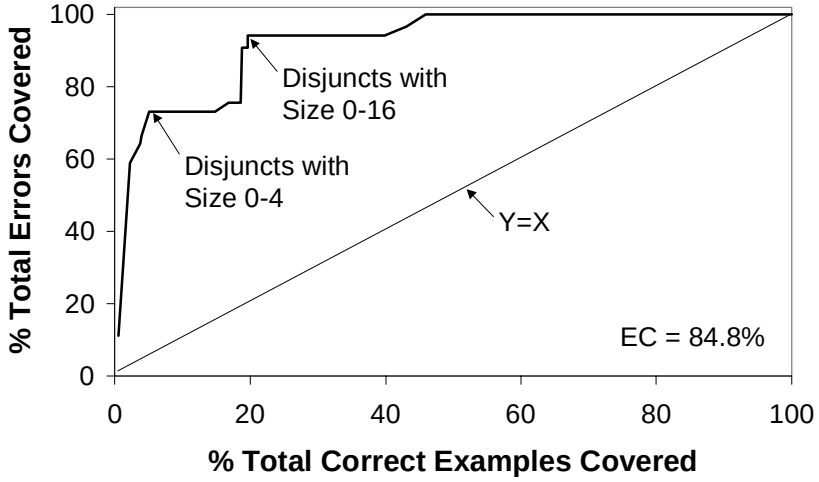


Figure 2: Error Concentration Curve for the Vote Dataset

To make it easy to compare the degree to which errors are concentrated in the small disjuncts for different concepts, we introduce a measurement called error concentration. *Error Concentration* (EC) is defined as the percentage of the total area *above* the line Y=X in Figure 2 that falls under the EC curve. EC may take on values between 100% and –100%, but is expected to be positive—a negative value indicates that the errors are concentrated more toward the larger disjuncts than the smaller disjuncts. The EC value for the concept in Figure 2 is 84.8%, indicating that the errors are highly concentrated toward the small disjuncts.

4 RESULTS

In this section we present the EC values for 30 datasets and demonstrate that, although they exhibit the problem with small disjuncts to varying degrees, there is some structure to this problem. We then present results that demonstrate how small disjuncts are affected by pruning, training set size, and noise. Due to space limitations, only a few key results are presented in this section. More detailed results are presented in the Appendix.

4.1 Error Concentration for 30 Datasets

C4.5 was applied to 30 datasets and the results, ordered by EC, are summarized in Table 1. We also list the percentage of test errors contributed by the smallest disjuncts that cover 10% of the correctly classified test examples. Note that, although there is a wide range of EC values and many concepts have high EC values, none of the concepts have a negative EC.

Table 1: Error Concentration Results for 30 Datasets

EC Rank	Dataset	Dataset Size	Error Rate	Largest Disjunct	Number Leaves	% Errors at 10% Correct	Error Conc.
1	kr-vs-kp	3196	0.3	669	47	75.0	87.4
2	hypothyroid	3771	0.5	2697	38	85.2	85.2
3	vote	435	6.9	197	48	73.0	84.8
4	splice-junction	3175	5.8	287	265	76.5	81.8
5	ticket2	556	5.8	319	28	76.1	75.8
6	ticket1	556	2.2	366	18	54.8	75.2
7	ticket3	556	3.6	339	25	60.5	74.4
8	soybean-large	682	9.1	56	175	53.8	74.2
9	breast-wisc	699	5.0	332	31	47.3	66.2
10	ocr	2688	2.2	1186	71	52.1	55.8
11	hepatitis	155	22.1	49	23	30.1	50.8
12	horse-colic	300	16.3	75	40	31.5	50.4
13	crx	690	19.0	58	227	32.4	50.2
14	bridges	101	15.8	33	32	15.0	45.2
15	heart-hungarian	293	24.5	69	38	31.7	45.0
16	market1	3180	23.6	181	718	29.7	44.0
17	adult	21280	16.3	1441	8434	28.7	42.4
18	weather	5597	33.2	151	816	25.6	41.6
19	network2	3826	23.9	618	382	31.2	38.4
20	promoters	106	24.3	20	31	32.8	37.6
21	network1	3577	24.1	528	362	26.1	35.8
22	german	1000	31.7	56	475	17.8	35.6
23	coding	20000	25.5	195	8385	22.5	29.4
24	move	3028	23.5	35	2687	17.0	28.4
25	sonar	208	28.4	50	18	15.9	22.6
26	bands	538	29.0	50	586	65.2	17.8
27	liver	345	34.5	44	35	13.7	12.0
28	blackjack	15000	27.8	1989	45	18.6	10.8
29	labor	57	20.7	19	16	33.7	10.2
30	market2	11000	46.3	264	3335	10.3	4.0

While dataset size is not correlated with error concentration, error rate clearly is—concepts with low error rates (<10%) tend to have high EC values. Based on the error rate (ER) and EC values, the entries in Table 1 seem to fit naturally into the following three categories.

1. **High-EC/Low-ER:** includes datasets 1-10
2. **Medium-EC/High-ER:** includes datasets 11-22
3. **Low-EC/High-ER:** includes datasets 23-30

Note that there are no learned concepts with very high EC and high ER, or with low EC and low ER. Of particular interest is that fact that for those datasets in the High-EC/Low-ER group, the largest disjunct in the concept classifies a significant portion of the total training examples, whereas this is not true for the datasets in the Low-EC/High-ER group.

A table similar to Table 1, but expanded to include the results for C4.5 with pruning, appears in Appendix A, Table A1.1. The main results for the pruning case can be summarized by comparing the averages over the 30 datasets: for C4.5 without pruning, the average EC is 47.1% whereas for C4.5 with pruning, the average is 33.5%. Thus, even with pruning, the small disjuncts still account for many of the overall errors.

4.2 Comparison with Results from RIPPER

Some learning methods, such as neural networks, do not have a notion of a disjunct, while others, such as nearest neighbor methods, do not form disjunctive concepts, but generate something very similar, since clusters of examples can be viewed as disjuncts (Van den Bosch et al. 1997). C4.5 is used for most experiments in this paper because it is well known and forms disjunctive concepts. In order to support the generality of any conclusions we draw from the results using C4.5, we compare the EC values for C4.5 with those of RIPPER, a rule learner that also generates disjunctive concepts. The comparison is presented in Figure 3, where each point represents the EC values for a single dataset. Since the results are clustered around the line Y=X, both learners tend to produce concepts with similar EC values, and hence tend to suffer from the problem with small disjuncts to similar degrees. The agreement is especially close for the most interesting cases, where the EC values are large—the same 10 datasets generate the largest 10 EC values for both learners.

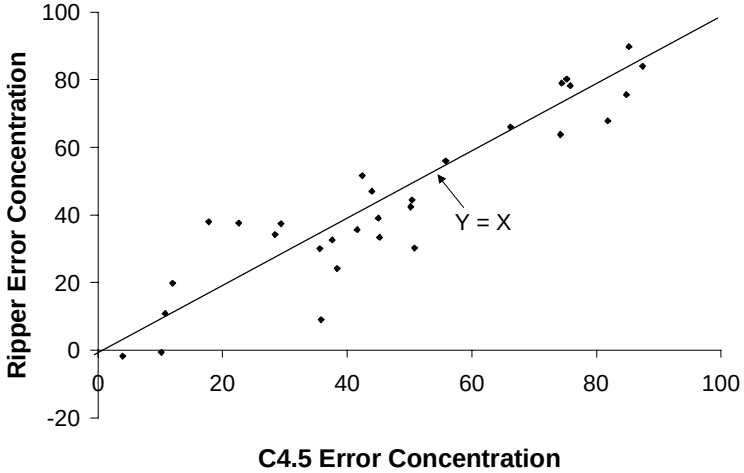


Figure 3: Comparison of C4.5 and RIPPER EC Values

The agreement shown in Figure 3 supports our belief that there is a fundamental property of the underlying datasets that is responsible for the EC values. We believe this property is the relative frequency of rare and general cases in the "true", but unknown, concept to be learned. We recognize, however, that a concept that has many rare cases when expressed as a disjunctive concept may not have them when expressed in a different form. We believe this does not significantly decrease the generality of our results given the number of learners that form disjunction-like concepts.

Additional information about the concepts generated by RIPPER is contained within the Appendix.

Detailed experimental results for RIPPER, similar to the results presented in Table 1 for C4.5, appear in Appendix A, Table A1.2. A comparison of the C4.5 and RIPPER EC values when pruning is used appears in Appendix B, Figure B2. The main difference with pruning is that then C4.5 tends to produce much higher EC values than RIPPER, perhaps indicating that RIPPER's pruning strategy tends to remove more small disjuncts from the learned concept. For completeness, the error rates for C4.5 and RIPPER are compared without and with pruning, in Appendix B, Figures B3 and B4, respectively. The results indicate that overall, RIPPER outperforms C4.5 when pruning is used, but when pruning is not used C4.5 outperforms RIPPER.

4.3 The Effect of Pruning

Pruning is not used for most of our experiments because it partially obscures the effects of small disjuncts. Nonetheless, small disjuncts provide an opportunity for better understanding how pruning works. Figure 4 displays the same information as Figure 1, except that the results are generated using C4.5 with pruning. Pruning causes the overall error rate to decrease to 5.3% from 6.9%.

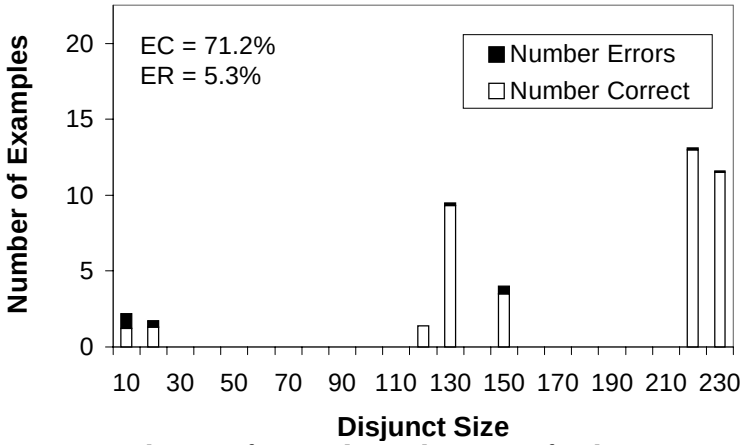


Figure 4: Distribution of Examples with Pruning for the Vote Dataset

Comparing Figure 4 with Figure 1 shows that with pruning the errors are less concentrated toward the small disjuncts (the decrease in EC from 84.8% to 71.2% confirms this). It is also apparent that with pruning far fewer examples are classified by disjuncts with size less than 30. This is because the distribution of disjuncts has changed—whereas before there were 45 disjuncts of size less than 10, after pruning there are only 7 (see Appendix F, Figure F1.2.2). Thus pruning eliminates most small disjuncts and many of the "emancipated" examples (i.e., those examples that would have been classified by the eliminated disjuncts) are then classified by the larger disjuncts. Overall, pruning causes the EC to decrease for 23 of the 30 datasets—and the decrease is often large. Looking at this another way, pruning causes the mean disjunct size associated with both the correct and incorrectly classified examples to increase, but the latter increases more than the former. Even after pruning the problem with small disjuncts is still quite evident—after pruning the average EC for the first 10 datasets is 50.6%.

Figure 5 plots the absolute improvement in error rate due to pruning against EC rank. The first 10 datasets, which are in the low-ER/high-EC group, show a moderate improvement in error rate. The datasets in the high-ER/medium-EC group, which starts with the Hepatitis dataset, show more improvement, but have more room for improvement due to their higher error rate. The datasets in the high-ER/low-EC group, which start with the Coding dataset, show a net *increase* in error rate. These results suggest that pruning helps when the problem with small disjuncts is quite severe, but may actually increase the error rate in other cases.

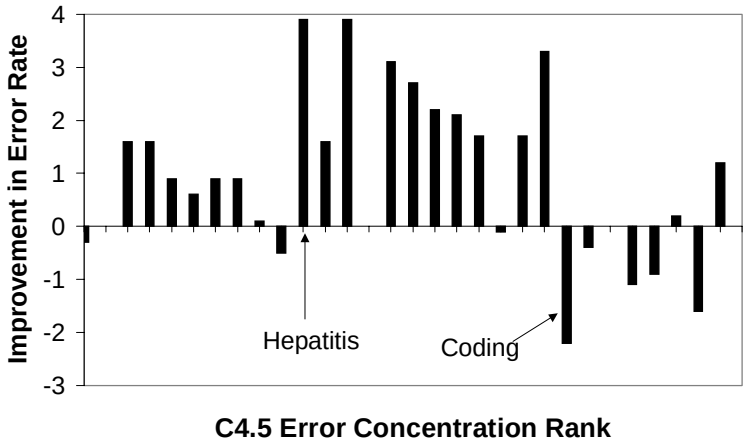


Figure 5: Improvement in Error Rate versus EC Rank

Pruning is the most widespread strategy for addressing the problem with small disjuncts. As was shown earlier, pruning eliminates many small disjuncts. The emancipated examples are then classified using other disjuncts. While this tends to cause the error rate of these other disjuncts to increase, the overall error rate of the concept tends to decrease. Pruning reduces C4.5's average error rate on the 30 datasets from 18.4% to 17.5%, while reducing the EC from 84.8% to 71.2%. It is useful to compare this average 0.9% error rate reduction to an "idealized" strategy where the error rate for the small disjuncts is equal to the error rate of the other (i.e., medium and large) disjuncts. While we do not expect such a strategy to be achievable, it provides a way of gauging the effectiveness of pruning at addressing the problem of small disjuncts.

Table 2 compares the error rates (averaged over the 30 datasets) resulting from various strategies. The idealized strategy is applied using two scenarios, where the smallest disjuncts covering 10% (20%) of the training examples are assigned an error rate equal to the error rate of the disjuncts covering the remaining 90% (80%) of the examples.

Table 2: Comparison of Pruning to Idealized Strategy

Strategy	No Pruning	Default Pruning	Idealized (10%)	Idealized (20%)
Average Error Rate	18.4%	17.5%	15.2%	13.5%

Table 2 shows that the idealized strategy, even when only applied to 10% of the examples, significantly outperforms C4.5's pruning strategy. These results provide a motivation for finding strategies that better address the problem with small disjuncts. The detailed results for each of the 30 datasets appear in Appendix C, Table C3.

For many real-world problems, such as identifying those customers likely to buy a product, one is more interested in finding individual classification rules that are extremely precise (i.e., have low error rate) than in finding the concept with the best *overall* accuracy. Given that previous results indicate that pruning tends to decrease the precision of the larger, more precise disjuncts (compare the results in Figures 1 and 4), this suggests that pruning may be counterproductive in many cases. To investigate this further, we allow each concept to grow, by starting with the largest disjunct and progressively adding smaller disjuncts. We then calculate the resulting error rate (on the test set) for each concept, with and without pruning, at the point at which it covers 10%, 20%, … ,100% of the total training examples. Because we expect the larger disjuncts to have lower error rates, we expect the error rate of the concept to increase as it is grown to cover more examples.

Table 3 shows the error rates, with and without pruning for the points at which the coverage of the training set is 10%, 30%, 50%, 70% and 100%. Table 3 also displays the difference in error rates (actually the increase in error rate with pruning). Because many of the differences are positive, we see that pruning often leads to poorer performance. An expanded version of Table 3, which shows the results at each 10% increment, appears in Appendix C, Table C4.

Table 3: Effect of Pruning when Concept Built from Largest Disjuncts

Dataset	% Error Rate at 10% covered			% Error Rate at 30% covered			% Error Rate at 50% covered			% Error Rate at 70% covered			% Error Rate at 100% covered		
	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ
kr-vs-kp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.6	0.3	0.3
hypothyroid	0.1	0.3	-0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.5	0.5	0.0
vote	3.1	0.0	3.1	1.0	0.0	1.0	0.9	0.0	0.9	2.3	0.7	1.6	5.3	6.9	-1.6
splice-junction	0.3	0.9	-0.6	0.2	0.3	-0.1	0.3	0.2	0.1	2.4	0.6	1.8	4.2	5.8	-1.6
ticket2	0.3	0.0	0.3	2.7	0.8	1.9	2.5	0.7	1.8	2.5	1.0	1.5	4.9	5.8	-0.9
ticket1	0.1	2.1	-1.9	0.3	0.6	-0.3	0.4	0.4	0.0	0.3	0.3	0.0	1.6	2.2	-0.5
ticket3	2.1	2.0	0.1	1.7	1.2	0.5	1.4	0.7	0.6	1.5	0.5	1.0	2.7	3.6	-0.9
soybean-large	1.5	0.0	1.5	5.4	1.0	4.4	5.3	1.6	3.7	4.7	1.3	3.5	8.2	9.1	-0.9
breast-wisc	1.5	1.1	0.4	1.0	1.0	0.0	0.6	0.6	0.0	1.0	1.4	-0.4	4.9	5.0	-0.1
ocr	1.5	1.8	-0.3	1.9	0.8	1.1	1.3	0.6	0.7	1.9	1.0	0.9	2.7	2.2	0.5
hepatitis	5.4	6.7	-1.3	15.0	2.2	12.9	15.0	9.1	5.9	12.8	12.1	0.6	18.2	22.1	-3.9
horse-colic	20.2	1.8	18.4	14.6	4.6	10.0	11.7	5.3	6.3	10.7	10.6	0.1	14.7	16.3	-1.7
crx	7.0	7.3	-0.3	7.9	6.5	1.4	6.3	7.3	-0.9	7.8	9.3	-1.6	15.1	19.0	-3.9
bridges	10.0	0.0	10.0	17.5	0.0	17.5	16.8	2.0	14.9	14.9	9.4	5.4	15.8	15.8	0.0
heart-hungarian	15.4	6.2	9.2	18.4	11.4	7.0	15.6	10.9	4.7	16.0	16.4	-0.4	21.4	24.5	-3.1
market1	16.6	2.2	14.4	12.2	7.8	4.4	12.7	12.1	0.6	14.5	15.9	-1.4	20.9	23.6	-2.6
adult	3.9	0.5	3.4	3.6	4.9	-1.3	8.9	8.1	0.8	8.3	10.6	-2.3	14.1	16.3	-2.2
weather	5.4	8.6	-3.2	10.6	14.0	-3.4	16.4	19.4	-3.1	22.7	24.6	-1.9	31.1	33.2	-2.1
network2	10.8	9.1	1.7	12.5	10.7	1.8	12.7	14.7	-2.0	15.1	17.2	-2.1	22.2	23.9	-1.8
promoters	10.2	19.3	-9.1	10.9	10.4	0.4	14.1	15.7	-1.6	19.6	16.8	2.8	24.4	24.3	0.1
network1	15.3	7.4	7.9	13.1	11.8	1.3	13.2	15.5	-2.3	16.7	17.3	-0.6	22.4	24.1	-1.7
german	10.0	4.9	5.1	11.1	12.5	-1.4	17.4	19.1	-1.8	20.4	25.7	-5.3	28.4	31.7	-3.3
coding	19.8	8.5	11.3	18.7	14.3	4.4	21.1	17.9	3.2	23.6	20.6	3.1	27.7	25.5	2.2
move	24.6	9.0	15.6	19.2	12.1	7.1	21.0	15.5	5.6	22.6	18.7	3.8	23.9	23.5	0.3
sonar	27.6	27.6	0.0	23.7	23.7	0.0	19.2	19.2	0.0	24.4	24.3	0.1	28.4	28.4	0.0
bands	13.1	0.0	13.1	34.3	16.3	18.0	34.1	25.0	9.1	33.8	26.6	7.2	30.1	29.0	1.1
liver	27.5	36.2	-8.8	32.4	28.1	4.3	28.0	30.1	-2.2	30.7	31.8	-1.2	35.4	34.5	0.9
blackjack	25.3	26.1	-0.8	25.1	25.8	-0.8	24.8	26.7	-1.9	26.1	24.4	1.7	27.6	27.8	-0.2
labor	25.0	25.0	0.0	17.5	24.8	-7.3	23.6	20.3	3.2	24.4	17.5	6.9	22.3	20.7	1.6
market2	44.1	45.5	-1.4	43.1	44.3	-1.2	42.5	44.2	-1.7	43.3	45.3	-2.0	45.1	46.3	-1.2
Average	**11.6**	**8.7**	**2.9**	**12.5**	**9.7**	**2.8**	**12.9**	**11.4**	**1.5**	**14.2**	**13.4**	**0.8**	**17.5**	**18.4**	**-0.9**

Table 3 shows that when we look at the error rates for each concept, averaged over all 30 datasets (i.e., the last row in the table), pruning results in a *higher* overall error rate in all cases, except when all disjuncts are included in the performance evaluation. For example, if we only consider the largest disjuncts that cover 50% of the total training examples, then C4.5 with pruning generates concepts with an average error rate of 12.9%, whereas C4.5 without pruning generates concepts with an average error rate of 11.4%. Looking at the individual results in this situation, pruning does worse for 17 of the datasets, better for 9 of the datasets, and the same for 4 of the datasets. However, the magnitude of the differences is much greater in the cases where pruning performs worse (see the scatter plot in Appendix C, Figure C2). These averaged results for the 30 datasets are summarized in Figure 6.

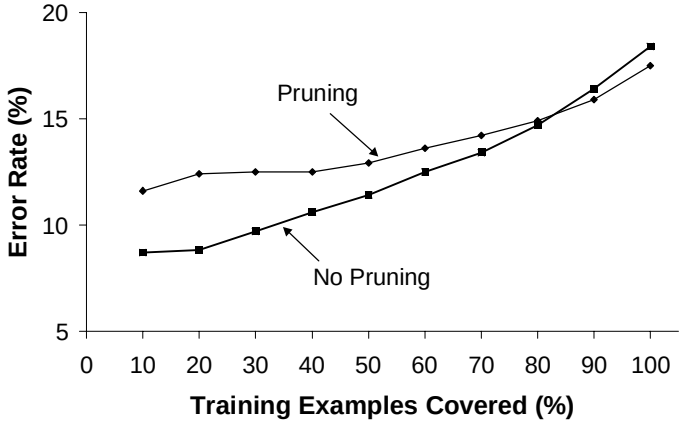


Figure 6: Averaged Error Rate Based on Concept Built from Largest Disjuncts

Figure 6 clearly demonstrates that under most circumstances pruning does not produce the best results.[1] While it produces marginally better results when predictive accuracy is the evaluation metric, it produces much poorer results when one can be very selective about the classification "rules" that are used. These results confirm the hypothesis that when pruning eliminates some small disjuncts, the emancipated examples wind up increasing the error rate of the larger disjuncts. The overall error rate is reduced only because the error rate of the emancipated examples is lower than their original error rate. Pruning redistributes the errors such that the errors are more uniformly distributed than before. This is exactly what we do not want to happen when we have the opportunity to conditionally classify an example. The fact that pruning actually hurts more than it helps for most situations in Table 3, and that the break-even point is all the way at 80%, is quite compelling.

4.4 The Effect of Training Set Size

Small disjuncts provide an opportunity to better understand how training set size affects learning. We again apply C4.5 to the Vote dataset, except that this time a different 10% (not 90%) of the dataset is used for training for each of the 10 cross-validation runs. Thus, the training set size is 1/9 the size it was previously. As before, each run employs a different 10% of the data for testing. The resulting distribution of examples is shown in Figure 7.

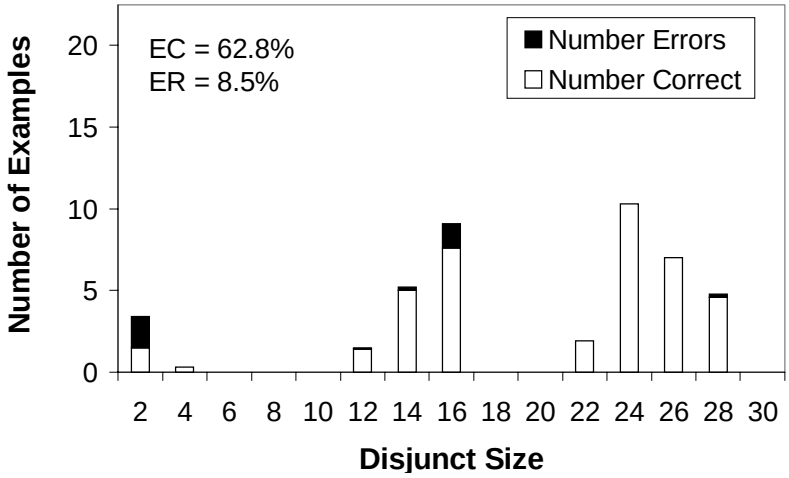


Figure 7: Distribution of Examples (10% Training Data)

Comparing the distribution of errors between Figures 1 and 7 shows that errors are less concentrated toward the smaller disjuncts in Figure 7. This is consistent with the fact that the EC decreases from 84.8% to 62.8% and the mean disjunct size over all examples decreases from 124 to 19, while the mean disjunct size of the errors decreases only slightly from 10.0 to 8.9. Figures similar to Figure 7, also for individual datasets, are presented in Appendix F, Figures F1.3, F2.3, and F3.3. The results for all 30 datasets are provided in Appendix D, Table D1. Those results demonstrate a similar phenomenon—for 27 of the 30 datasets the EC decreases as the training set size decreases.

These results suggest that the definition of small disjuncts should factor in training set size. To investigate this further, the error rates of disjuncts with specific sizes (0, 1, 2, etc.) were compared as the training set size was varied. Because disjuncts of a specific size for most concepts cover very few examples, statistically valid comparison were possible for only 4 of the 30 datasets (Coding, Move, Adult, and Market2); with the other datasets the number of examples covered by disjuncts of a given size is too small. The results for the Coding dataset are shown in Figure 8. Results for the remaining three datasets appear in Appendix D, Figures D2 - D4.

[1] Figure 6 corrects a minor error that is present in the shortened AAAI-2000 version of this paper. In the figure in the AAAI paper, the x-axis was mistakenly labeled as measuring recall instead of the percentage of training examples covered.

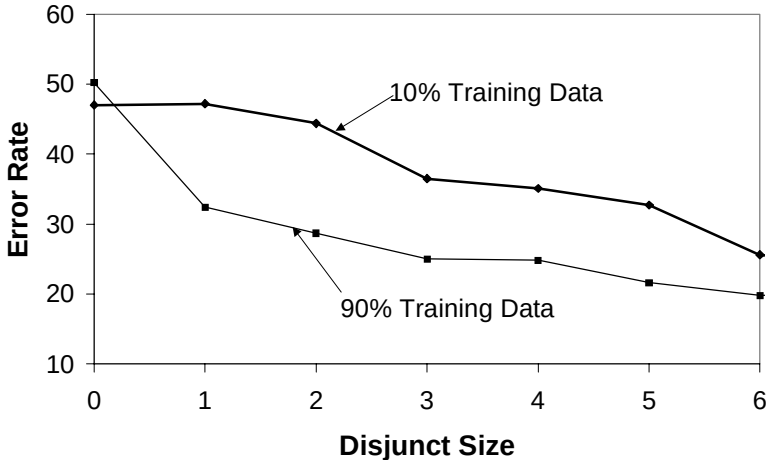


Figure 8: Effect of Training Size on Disjunct Error Rate

Figure 8 shows that the error rates for the smallest disjuncts decrease significantly when the training set size is increased. These results further suggest that the definition of small disjuncts should take training set size into account.

4.5 The Effect of Noise

Rare cases cause small disjuncts to be formed in learned concepts. The inability to distinguish between these rare cases (i.e., true exceptions) and noise may be largely responsible for the difficulty in learning in the presence of noise. This conjecture was investigated using synthetic datasets (Weiss 1995) and two real-world datasets (Weiss and Hirsh 1998). We extend this previous work by analyzing 27 datasets (technical difficulties prevented us from handling 3 of the datasets).

All experiments involved applying either random class noise or random attribute noise to the data. A total of 3 scenarios were used:

1. Random class noise is applied to the training set (the test set is untouched)
2. Random attribute noise is applied to the training set (the test set is untouched)
3. Random attribute noise is applied to both the training and test sets

Random class noise is never applied to the test set, since that would make no sense (since we evaluate the results using the class value associated with the test set examples). The scenario where random class noise is applied only to the training set allows us to evaluate the ability of the learner to learn the correct concept in the presence of attribute noise. The scenario where attribute noise is applied to both the training and test set corresponds to the real-world situation where errors in measurement affect all examples. When we say n% random class noise is applied to a dataset, we mean that for n% of the examples the class value is replaced by a randomly selected valid class value (possibly the same value as the original value). Given this definition, all information is lost only when 100% class noise is applied to the dataset. Attribute noise is defined similarly, except that if the attribute is numerical, then a random value is generated within the range defined by the minimum and maximum values. It should be pointed out that comparing results with attribute noise across datasets is problematic, since the datasets contain differing number of attributes, and hence the effect of attribute noise is not expected to be equal.

We begin by examining the effect that noise has on the error rate, the error concentration, and the number of leaves in the induced decision tree. So that we can easily see any general trends, we initially focus on the results averaged over the 27 datasets. Figure 9 shows the results for error rate, Figure 10 for error concentration, and Figure 11 for the number of leaves. The values for each data point can be found by referring to Appendix E, Table E1.1, which also shows how noise affects the mean disjunct statistics. In all cases, measurements are taken at the following levels of noise: 0%, 3%, 5%, 10%, 20%, 30%, 40%, and 50%. In the Figures, the curves are labeled to identify which of the 3 types of noise is used: class noise (Class), attribute noise applied to the training set (AttrTrain) or

attribute noise applied to both the training and test sets (AttrBoth). If the label has the suffix "-Prune" then C4.5's default pruning strategy was used; otherwise pruning was disabled.

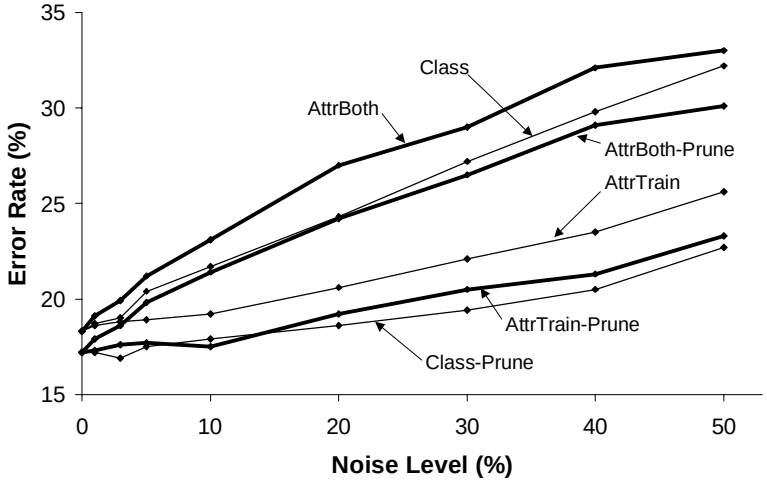


Figure 9: Effect of Noise on Error Rate (averaged over 27 datasets)

Figure 9 shows that, as expected, with two very minor exceptions, the error rate increases with increasing levels of noise.[2] Note that pruning improves the performance of the learned concepts when there is noise present—more so than when there is no noise. As expected, the error rate is higher when attribute noise is applied to both the training and test sets than when it is applied to just the training set. One interesting result is that pruning is much more able to correct for class noise than attribute noise.

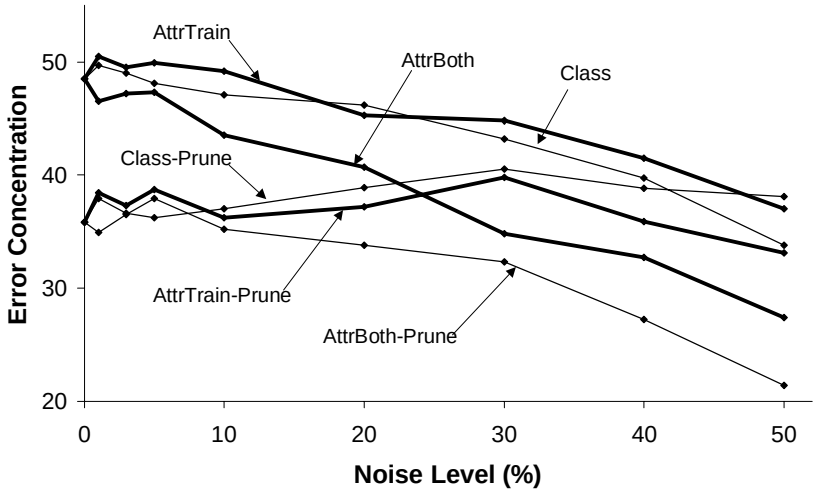


Figure 10: Effect of Noise on Error Concentration (averaged over 27 datasets)

Figure 10 shows that for four of the six scenarios the EC decreases relatively consistently. This means that as the error rate increases, a greater percentage of the errors come from the larger disjuncts. This is not surprising, since at 100% noise the EC must approach 0 (at that point there is no information in the data). The results show, however, that when there is noise only in the training set—either class noise or attribute noise—with pruning the EC remains relatively constant.

[2] The error rate decreases slightly, from 17.2% to 16.9%, for the case where the class noise goes from 1% to 3% and pruning is used (8 of the 27 datasets show an increase in error rate, 13 show a decrease and 6 show no change). The error rate decreases from 17.7% to 17.5% for the case where attribute noise applied to the training set increases from 5% to 10% (16 of the datasets show an increase, 9 show a decrease, and 2 show no difference--the decreases tended to be larger than then increases). The decreases, especially for the class noise case, may be due to the fact that the noise causes more aggressive pruning, which ultimately benefits the learned concept.

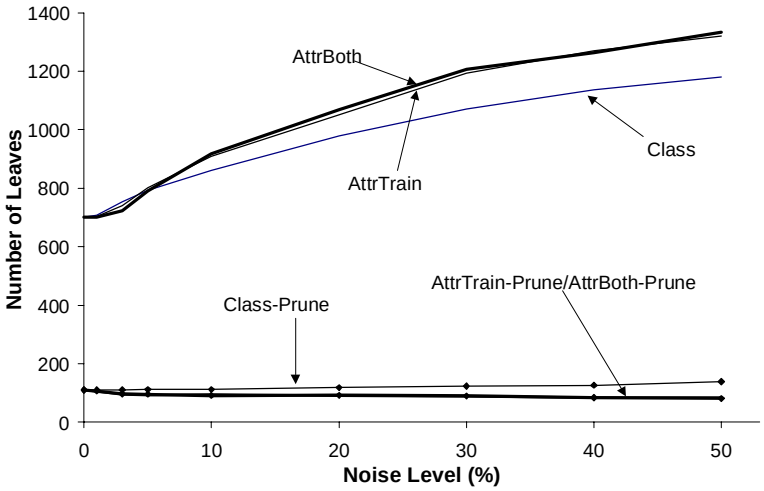


Figure 11: Effect of Noise on Number of Leaves (averaged over 27 datasets)

Figure 11 shows that the number of leaves in the tree increases as the noise level increases when there is no pruning, but that pruning dramatically slows down this increase. Note that we get essentially identical results whether there is attribute noise in the training set or in the training and test set. This is expected, since noise in the test set cannot affect the construction of the decision tree. It is worth noting that attribute noise causes the size of the tree to grow faster than class noise. In addition, class noise does not affect the complexity of all induced concepts equally. For low-ER/high-EC group, 10% class noise causes the mean disjunct size of these concepts to shrink, on average, to one-ninth the original size; for the datasets in the high-ER/low-EC group, the same level of noise causes almost no change in the mean disjunct size—the average drops by less than 1%.

The detailed results for class noise (Appendix E, Table E2.1) indicate that there is a subtle trend for datasets with higher EC values to experience a greater increase in error rate from class noise. What is much more apparent, however, is that many concepts with low EC values are *extremely* tolerant of noise, whereas none of the concepts with high EC's are. For example, two of the low-EC datasets, blackjack and labor, are so tolerant of noise that when 50% random class noise is added to the training set (i.e., the class value is replaced with a randomly selected valid value 50% of the time), the error rate on the test set increases by less than 1%. The other effect is that as the amount of class noise is increased, the EC tends to decrease. Thus, as noise is added, across almost all of the concepts a greater percentage of the errors come from the larger disjuncts. This helps explain why we find a low-ER/high-EC group of concepts and a high-ER/medium-EC group of concepts: adding noise to concepts in the former increases their error rate and decreases their error concentration, making them look more like concepts in the latter group.

5 DISCUSSION

Many of the results in this paper can be explained by understanding the role of small disjuncts in learning. We begin with the understanding that learning algorithms tend to form large disjuncts to cover general cases and small disjuncts to cover rare cases (although the bias of the learner is also a factor). Concepts with many rare cases are harder to learn than those with few, since general cases can be more accurately sampled with less training data. The results in Table 1 support this, since concepts with low error rates tend to have some very general cases. For example, the first 10 entries in Table 1, which fall into the "High-EC/ low-ER" group, include a single disjunct that, on average, classifies 43% of the correctly classified training examples. This at least partially explains why the datasets with low error rate have a high error concentration—they contain very general cases that can be learned quite well. The Vote dataset demonstrates this quite clearly since the largest disjunct learned in each of its 10 cross-validated runs never covers any test errors.

Pruning operates by removing some of the more error-prone small disjuncts. This will cause some

of the rare cases to be mistakenly classified along with more general cases, since the pruning strategy may not be able to distinguish between rare cases and noisy data. The emancipated examples are then distributed throughout the other disjuncts in the concept, which tends to spread out the errors and reduce the error concentration.

The results of our experiments which vary training set size (Appendix D, Table D1) show that for 27 of 30 datasets, the error concentration increases as the training set size increases. The reason this occurs is that as the training set size increases, the rare cases are more likely to be sampled, which will allow them to be represented in the learned concept. With small training set sizes, the rare cases are likely to be "missed" and they will wind up being classified along with the general cases, which will cause the EC to move closer toward 0. An important question is what will happen if the training set size grows without bound. Based on our results it appears that the average disjunct size will grow, even though new small disjuncts may be introduced due to more thorough sampling of the data. The error rate will also continue to improve, until in reaches a plateau. Figure 8 and Figures D1 through D4 in Appendix D show that as the training set size increases, the error rate of a disjunct of fixed size tends to *decrease*. The key question is whether, at the point at which the plateau is reached and additional data results in no improvement in error rate, the small disjuncts will have a higher error rate than the large disjuncts—and if so, why? Could it be that the smaller disjuncts, which correspond to the (relatively) rare cases in the concept to be learned, are inherently more error prone? Or perhaps there is noise in the data that prevents the rare cases from being learned but is not sufficient to prevent the more general cases from being learned.

Almost all strategies for addressing the problem with small disjuncts treat small and large disjuncts differently. Consequently, if we hope to address this problem, we need a way to effectively distinguish between the two. The definition that a small disjunct is a disjunct that correctly classifies few training examples (Holte, et al. 1989) is not particularly helpful in this context. What is needed is a method for determining a good threshold *t*, such that disjuncts with size less than *t* have a much higher error rate than those with size greater than *t*. Based on our results we suggest that the threshold *t* should be based on the relationship between disjunct size and error rate, since error rate is not related to disjunct size in a simple way, and more specifically, using error concentration. Based on the EC curve in Figure 2, for example, it seems reasonable to conclude that the threshold for the Vote dataset should be 4, 16, or a value in between. For datasets such as Market2 or Labor, where the EC is very low, we may choose not to distinguish small disjuncts from large disjuncts at all.

6 CONCLUSION

This paper provides insight into the role of small disjuncts in learning. By measuring error concentration on concepts induced from 30 datasets, we demonstrate that the problem with small disjuncts occurs to varying degrees, but is quite severe for many of these concepts. We show that even after pruning the problem is still evident, and, by using RIPPER, showed that our results are not an artifact of C4.5.

Although the focus of the paper was on measuring and understanding the impact of small disjuncts on learning, we feel our results could lead to improved learning algorithms. First, error concentration can help identify the threshold for categorizing a disjunct as small, and hence can be used to improve the effectiveness of variable bias system in addressing the problem with small disjuncts. The EC value could also be used to control the pruning strategy of a learning algorithm, since low EC values seem to indicate that pruning may actually decrease predictive accuracy. A high EC value is also a clear indication that one is likely to be able to trade-off reduced recall for greatly improved precision.

Acknowledgments

We would like to thank William Cohen for supplying the AT&T datasets, and for detailed information about RIPPER.

REFERENCES

Ali, K. M. and Pazzani, M. J. 1992. Reducing the Small Disjuncts Problem by Learning Probabilistic Concept Descriptions, in T. Petsche editor, *Computational Learning Theory and Natural Learning Systems*, Volume 3.

Blake, C. L. and Merz, C. J. 1998. UCI Repository of ML Databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Dept. of Computer Science.

Cohen, W. 1995. Fast Effective Rule Induction. In Proceedings of the Twelfth International Conference on Machine Learning, 115-123.

Cohen, W. and Singer, Y. 1999. A Simple, Fast, and Effective Rule Learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 335-342. Menlo Park, Calif.: AAAI Press.

Danyluk, A. P. and Provost, F. J. 1993. Small Disjuncts in Action: Learning to Diagnose Errors in the Local Loop of the Telephone Network. In *Proceedings of the Tenth International Conference on Machine Learning*, 81-88.

Holte, R., C., Acker, L. E., and Porter, B. W. 1989. Concept Learning and the problem of small disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 813-818. San Mateo, CA: Morgan Kaufmann.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Ting, K. M. 1994. The Problem of Small Disjuncts: its Remedy in Decision Trees. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, 91-97.

Van den Bosch, A., Weijters, A., Van den Herik, H. J. and Daelemans, W. 1997. When Small Disjuncts Abound, Try Lazy Learning: A Case Study. In *Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning*, 109-118.

Weiss, G. M. 1995. Learning with Rare Cases and Small Disjuncts. In *Proceedings of the Twelfth International Conference on Machine Learning*, 558-565.

Weiss, G. M. and Hirsh, H. 1998. The Problem with Noise and Small Disjuncts. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 574-578.

Appendices

The figures and tables on the following pages contain our detailed data relating to small disjuncts. For easy reference, some of the figures and tables that appeared in the main body of the paper are reproduced here. The appendices are organized as follows:

A. Basic Small Disjuncts Statistics for 30 Datasets **16**
A1: Error Concentration Tables 16
A2: Mean Coverage Statistics 19

B. Comparison of C4.5 and RIPPER **20**

C. Effect of Pruning **22**

D. Effect of Training Set Size on Small Disjuncts **25**

E. Effect of Noise on Small Disjuncts **27**
E1: Summary Results Averaged over all Datasets 27
E2: Effect of Noise on Error Rate and Error Concentration 28
E3: Effect of Noise on Disjunct Sizes 32

F. Detailed Analysis of Selected Datasets **36**
F1: Vote Dataset 37
F2: Move Dataset 40
F3: Adult Dataset 43

Appendix A
Basic Small Disjunct Statistics for 30 Datasets

A1. Error Concentration Tables

The error concentration tables include the error concentration values for all 30 datasets, as well as a few other descriptive variables. A table of results is provided for each of the two learners (C4.5 and RIPPER), and each table contains results for the two different pruning configurations (no pruning, default pruning strategy). Table A1.1 provides the results for C4.5 and Table A1.2 the results for RIPPER.

Each table contains 30 rows (one for each dataset) and 12 fields, each of which is described below. The values in fields 4-10 are calculated as the averages over the 10 10-fold cross-validation runs.

1. **EC Rank:** a value between 1 and 30, where 1 indicates the highest EC value and 30 the lowest EC value. To allow easy comparison between tables, the EC rank is computed only using C4.5 without pruning, and hence is the same for all tables.
2. **Dataset:** the name of the dataset
3. **Dataset Size:** the total number of instances in the dataset
4. **Prune:** The value "no" indicates pruning was disabled and "yes" means that the default pruning strategy was used
5. **Error Rate:** the error rate of the learner on the test set
6. **Largest Disjunct:** the size of the largest disjunct in the concept
7. **Number of Leaves/Rules:** the number of leaves/rules in the C4.5 decision tree/RIPPER ruleset
8. **% Errors at 10% correct:** the percentage of the total test errors that are contributed by the smallest disjuncts that cover the first 10% of the correct examples
9. **% Errors at 20% correct:** defined similarly to the previous field
10. **% Correct at 50% errors:** the percentage of the total correctly classified test examples that are contributed by the smallest disjuncts that cover the first 50% of the test errors
11. **Cov. at EF < 2:** The Error Factor is defined as the cumulative % of total errors covered divided by the cumulative % of total cases that are covered. This field displays the first point at which the Error Factor drops to below 2 for good (i.e., which is equivalent to the point at which the error rate drops below twice the overall error rate).
12. **Error Concentration (EC):** a measure of the degree to which errors are concentrated toward the small disjuncts. It is described in Section 3 of the body of this report (and in Figure 1). The EC may range from –100 to +100.
    - A value of +100 indicates that all of the errors are found in the smallest disjunct(s), before even a single correctly classified example is found.
    - A value of 0 indicates that the errors are distributed evenly throughout the disjuncts, and are not concentrated toward the small or large disjuncts.
    - A value of –100 indicates that all of the errors are found in the largest disjunct(s), without even a single correctly classified example found in these disjuncts.

Table A1.1: Error Concentration Table for C4.5

EC Rank	Dataset	Dataset Size	Prune	Error Rate	Largest Disjunct	Number Leaves	% Errors at 10% Correct	% Errors at 20% Correct	% Correct at 50% Errors	Cov. at EF < 2	Error Conc.
1	kr-vs-kp	3196	no	0.3	669	47	75.0	87.5	1.1	527	87.4
			yes	0.6	669	29	35.4	62.5	15.6	529	65.8
2	hypothyroid	3771	no	0.5	2697	38	85.2	90.7	0.8	2705	85.2
			yes	0.5	2732	15	90.7	90.7	0.7	2737	81.8
3	vote	435	no	6.9	197	48	73.0	94.2	1.9	194	84.8
			yes	5.3	221	10	68.7	74.7	2.9	218	71.2
4	splice-junction	3175	no	5.8	287	265	76.5	90.6	4.0	82	81.8
			yes	4.2	479	55	41.6	45.1	25.9	452	56.6
5	ticket2	556	no	5.8	319	28	76.1	83.0	2.7	354	75.8
			yes	4.9	442	9	48.1	55.0	12.8	445	47.4
6	ticket1	556	no	2.2	366	18	54.8	90.5	4.4	362	75.2
			yes	1.6	410	5	46.7	94.4	10.3	410	73.0
7	ticket3	556	no	3.6	339	25	60.5	84.5	4.6	333	74.4
			yes	2.7	431	6	37.0	49.7	20.9	432	31.0
8	soybean-large	682	no	9.1	56	175	53.8	90.6	9.3	18	74.2
			yes	8.2	61	62	48.0	57.3	14.4	18	39.4
9	breast-wisc	699	no	5.0	332	31	47.3	63.5	10.7	323	66.2
			yes	4.9	345	14	49.6	78.0	10.0	341	68.8
10	ocr	2688	no	2.2	1186	71	52.1	65.4	8.9	141	55.8
			yes	2.7	1350	37	40.4	46.4	34.3	99	34.8
11	hepatitis	155	no	22.1	49	23	30.1	58.0	17.2	19	50.8
			yes	18.2	89	9	24.2	46.3	26.3	82	16.8
12	horse-colic	300	no	16.3	75	40	31.5	52.1	18.2	31	50.4
			yes	14.7	137	6.3	35.8	50.4	19.3	68	27.2
13	crx	690	no	19.0	58	227	32.4	61.7	14.3	7	50.2
			yes	15.1	267	23	45.2	62.5	11.5	190	51.6
14	bridges	101	no	15.8	33	32	15.0	37.2	23.2	11	45.2
			yes	15.8	67	2	14.9	28.9	50.1	58	6.4
15	heart-hungarian	293	no	24.5	69	38	31.7	45.9	21.9	45	45.0
			yes	21.4	132	10	19.9	37.7	31.8	52	19.8
16	market1	3180	no	23.6	181	718	29.7	48.4	21.1	11	44.0
			yes	20.9	830	135	28.4	44.6	23.6	75	33.6
17	adult	21280	no	16.3	1441	8434	28.7	47.2	21.8	10	42.4
			yes	14.1	5018	419	36.6	53.2	17.6	561	42.4
18	weather	5597	no	33.2	151	816	25.6	47.1	22.4	8	41.6
			yes	31.1	573	496	26.2	46.0	22.2	13	44.2
19	network2	3826	no	23.9	618	382	31.2	46.9	24.2	12	38.4
			yes	22.2	1685	151	30.8	48.2	21.2	39	36.2
20	promoters	106	no	24.3	20	31	32.8	48.7	20.6	2	37.6
			yes	24.4	26	16	17.2	31.1	37.0	15	12.8
21	network1	3577	no	24.1	528	362	26.1	44.2	24.1	11	35.8
			yes	22.4	1470	142	24.4	43.4	27.2	25	31.8
22	german	1000	no	31.7	56	475	17.8	37.5	29.4	2	35.6
			yes	28.4	313	92	29.6	46.8	21.9	13	40.4
23	coding	20000	no	25.5	195	8385	22.5	36.4	30.9	1	29.4
			yes	27.7	415	2077	17.2	31.6	34.9	1	21.6
24	move	3028	no	23.5	35	2687	17.0	33.7	30.8	0	28.4
			yes	23.9	216	366	14.4	24.4	42.9	2	9.4
25	sonar	208	no	28.4	50	18	15.9	30.1	32.9	2	22.6
			yes	28.4	50	15	15.1	28.0	34.6	2	20.2
26	bands	538	no	29.0	50	586	65.2	65.2	54.1	0	17.8
			yes	30.1	279	3	0.8	4.7	58.3	29	-18.4
27	liver	345	no	34.5	44	35	13.7	27.2	40.3	1	12.0
			yes	35.4	59	22	17.6	31.8	34.8	14	16.2
28	blackjack	15000	no	27.8	1989	45	18.6	31.7	39.3	65	10.8
			yes	27.6	3053	22	16.9	29.7	44.7	154	9.2
29	labor	57	no	20.7	19	16	33.7	39.6	49.1	3	10.2
			yes	22.3	24	4	14.3	18.4	40.5	14	8.2
30	market2	11000	no	46.3	264	3335	10.3	21.6	45.5	0	4.0
			yes	45.1	426	856	12.2	23.9	44.7	0	6.0

Table A1.2: Error Concentration Table for RIPPER

EC Rank	Dataset	Dataset Size	Prune	Error Rate	Largest Disjunct	Number Leaves	% Errors at 10% Correct	% Errors at 20% Correct	% Correct at 50% Errors	Cov. at EF < 2	Error Conc.
1	kr-vs-kp	3196	no	0.8	669	43	92.9	92.9	2.2	49.8	84.0
			yes	0.8	669	28	56.8	92.6	5.4	49.1	74.6
2	hypothyroid	3771	no	1.2	2696	25	96.0	96.0	0.1	76.3	89.8
			yes	0.9	2732	14	97.2	97.2	0.6	84.3	93.0
3	vote	435	no	6.0	197	27	75.8	75.8	3.0	57.6	75.6
			yes	4.1	221	8	62.5	68.8	2.8	54.9	64.8
4	splice-junction	3175	no	6.1	422	106	62.3	76.1	7.9	46.5	67.8
			yes	5.8	552	46	46.9	75.4	10.7	49.3	69.0
5	ticket2	556	no	6.8	261	32	71.0	91.0	3.2	55.9	78.2
			yes	4.5	405	9	73.3	74.6	7.8	73.2	57.4
6	ticket1	556	no	3.5	367	18	69.4	95.2	1.6	100.0	80.2
			yes	1.6	410	7	41.5	95.0	11.9	92.6	74.0
7	ticket3	556	no	4.5	333	28	61.4	81.5	5.6	77.0	79.0
			yes	4.0	412	8	71.3	71.3	9.0	82.3	51.6
8	soybean-large	682	no	11.3	61	65	69.3	69.3	4.8	40.4	63.8
			yes	9.8	66	36	17.8	26.6	47.4	2.6	12.8
9	breast-wisc	699	no	5.3	355	25	68.0	68.0	3.6	57.8	66.0
			yes	4.4	370	10	14.4	39.2	31.4	28.7	12.4
10	ocr	2688	no	2.6	804	29	50.5	62.2	10.0	41.1	56.0
			yes	2.7	854	26	29.4	32.6	24.5	15.4	30.6
11	hepatitis	155	no	20.3	60	19	19.3	47.7	20.8	40.7	30.2
			yes	22.3	93	5	25.5	28.3	57.2	2.6	-0.4
12	horse-colic	300	no	22.0	73	27	20.7	47.2	23.9	35.8	44.4
			yes	15.7	141	6	13.8	20.5	36.6	10.7	8.6
13	crx	690	no	17.0	120	31	32.5	50.3	19.7	38.4	42.4
			yes	15.1	272	6	16.4	31.9	39.1	11.2	10.8
14	bridges	101	no	14.5	39	14	41.7	41.7	35.5	56.2	33.4
			yes	18.3	71	4	19.1	22.2	55.0	30.0	-2.4
15	hungarian-heart	293	no	23.9	67	28	25.8	44.9	24.8	34.7	39.0
			yes	18.8	138	7	17.9	29.3	42.6	25.6	7.2
16	market1	3180	no	25.0	243	46	32.2	57.8	16.9	36.7	47.0
			yes	21.3	998	18	19.0	34.5	43.4	5.2	11.4
17	adult	21280	no	19.7	1488	104	36.9	56.5	15.0	39.0	51.6
			yes	15.2	9293	31	9.8	29.5	67.9	0.7	-14.6
18	weather	5597	no	30.2	201	142	23.8	42.1	24.8	27.2	35.6
			yes	26.9	1148	34	18.8	31.2	35.4	8.8	19.8
19	network2	3826	no	23.1	77	23	25.6	45.9	22.9	30.3	24.2
			yes	22.6	1861	15	15.3	34.4	39.5	7.1	9.0
20	promoters	106	no	19.8	24	15	20.0	50.6	20.0	54.3	32.6
			yes	11.9	32	9	0.0	0.0	54.1	2.2	-32.4
21	network1	3577	no	23.4	79	26	18.9	29.7	46.0	13.0	9.0
			yes	23.3	1765	14	16.0	34.4	42.0	12.9	9.0
22	german	1000	no	30.8	99	34	12.1	31.2	35.0	0.9	30.0
			yes	29.4	390	8	14.7	32.5	32.4	0.4	12.8
23	coding	20000	no	28.2	206	773	22.6	37.6	29.2	20.1	37.4
			yes	28.3	894	158	12.7	21.7	46.5	0.0	5.2
24	move	3028	no	32.1	45	79	25.9	44.5	25.6	31.1	34.2
			yes	24.1	320	43	10.9	19.5	63.1	0.2	-9.4
25	sonar	208	no	31.0	47	15	32.6	41.2	23.9	41.3	37.6
			yes	29.7	59	8	23.1	27.8	25.4	44.2	28.2
26	bands	538	no	21.9	62	41	25.6	36.9	29.2	28.8	38.0
			yes	26.0	118	14	22.1	39.5	24.0	32.5	21.8
27	liver	345	no	34.0	28	32	28.2	37.4	32.0	36.7	19.8
			yes	32.1	69	9	13.6	33.2	34.7	12.7	14.6
28	blackjack	15000	no	30.2	1427	193	12.3	24.2	42.3	0.6	10.8
			yes	28.1	4893	15	16.8	22.1	45.3	3.3	4.0
29	labor	57	no	24.5	21	10	0.0	55.6	18.3	4.7	-0.6
			yes	18.2	25	6	0.0	3.6	70.9	16.4	-22.8
30	market2	11000	no	48.8	55	12	10.4	21.1	49.8	0.6	-1.8
			yes	40.9	2457	8	7.7	17.7	50.2	0.1	-1.6

A2. Mean Coverage Statistics

The error concentration describes the degree to which errors are concentrated toward the smaller disjuncts. Another measure that we use to describe a concept is *mean coverage*. The mean coverage is computed by labeling each test example with the disjunct size of the node that it is classified by, and then taking the average of all of these values—that is, the mean coverage is a weighted average where the weights are the disjunct sizes. The mean correct coverage/mean error coverage are defined similarly, except that the averages are computed only using the correctly/incorrectly classified test examples. The mean ratio is defined as mean correct coverage divided by mean error coverage. A mean ratio greater than 1 indicates that the correctly classified examples tend to be classified by larger disjuncts than the incorrectly classified examples.

Table A.2: Mean Coverage Statistics for C4.5

Dataset	Dataset Size	Unpruned C4.5				Pruned C4.5			
		Mean Coverage	Mean Correct Coverage	Mean Error Coverage	Mean Ratio	Mean Coverage	Mean Correct Coverage	Mean Error Coverage	Mean Ratio
kr-vs-kp	3196	401.8	403.0	28.2	14.3	409.8	411.4	125.1	3.3
hypothyroid	3771	2181.7	2190.1	330.6	6.6	2238	2247.1	335.9	6.7
vote	435	124.4	132.9	10.0	13.3	169.7	175.5	66.0	2.7
splice-junction	3175	95.8	100.9	13.4	7.5	277.9	286.5	81.8	3.5
ticket2	556	224.9	236.5	36.0	6.6	400.7	410.9	200.9	2.0
ticket1	556	277.6	282.6	51.3	5.5	342.5	346.9	78.6	4.4
ticket3	556	241.9	249.2	46.0	5.4	379.0	383.5	214.7	1.8
soybean-large	682	19.9	21.5	4.1	5.2	29.0	30.0	17.7	1.7
breast-wisc	699	195.3	203.3	44.3	4.6	228.6	237.1	61.5	3.9
ocr	2688	625.4	635.3	192.0	3.3	794.7	804.1	451.6	1.8
hepatitis	155	21.6	25.2	8.7	2.9	70.0	73.7	53.1	1.4
horse-colic	300	33.7	37.5	14.2	2.6	97.7	101.6	75.0	1.4
crx	690	15.5	17.6	6.9	2.6	184.5	199.3	101.2	2.0
bridges	101	14.2	16.3	3.2	5.1	60.3	61.4	54.3	1.1
heart-hungarian	293	37.1	42.8	19.6	2.2	91.9	96.9	73.5	1.3
market1	3180	42.2	50.1	16.8	3.0	379.3	416.9	237.4	1.8
adult	21280	182.6	212.6	28.5	7.5	2065.1	2244.8	967.4	2.3
weather	5597	24.4	30.6	12.0	2.6	110.0	144.2	34.2	4.2
network2	3826	163.7	192.9	70.8	2.7	948.2	1061.8	549.1	1.9
promoters	106	6.5	7.4	3.8	1.9	13.0	14.3	9.0	1.6
network1	3577	149.1	173.7	71.5	2.4	782.6	871.9	472.5	1.8
german	1000	9.5	11.8	4.4	2.7	132.7	160.6	62.3	2.6
coding	20000	8.2	9.9	3.3	3.0	62.4	68.9	45.4	1.5
move	3028	6.2	6.9	3.8	1.8	57.6	59.0	53.1	1.1
sonar	208	29.1	31.2	23.7	1.3	29.1	31.2	23.7	1.3
bands	538	6.0	8.2	0.4	20.5	249.8	239.2	274.4	0.9
liver	345	22.3	23.2	20.7	1.1	30.9	32.9	27.3	1.2
blackjack	15000	909.6	937.8	836.2	1.1	1711.8	1752.3	1605.7	1.1
labor	57	11.5	11.4	11.6	1.0	18.9	19.2	17.7	1.1
market2	11000	35.0	35.5	34.3	1.0	84.7	87.6	81.2	1.1
Averages	**3553**	**203.9**	**211.3**	**65.01**	**4.7**	**415.0**	**435.7**	**215.0**	**2.1**

Appendix B
Comparison of C4.5 and RIPPER

This appendix compares the performance of C4.5 and RIPPER by comparing the error concentrations (Figures B1 and B2) and error rates (Figures B3 and B4) of the concept induced by these learners for each of the 30 datasets, with and without the use of pruning.

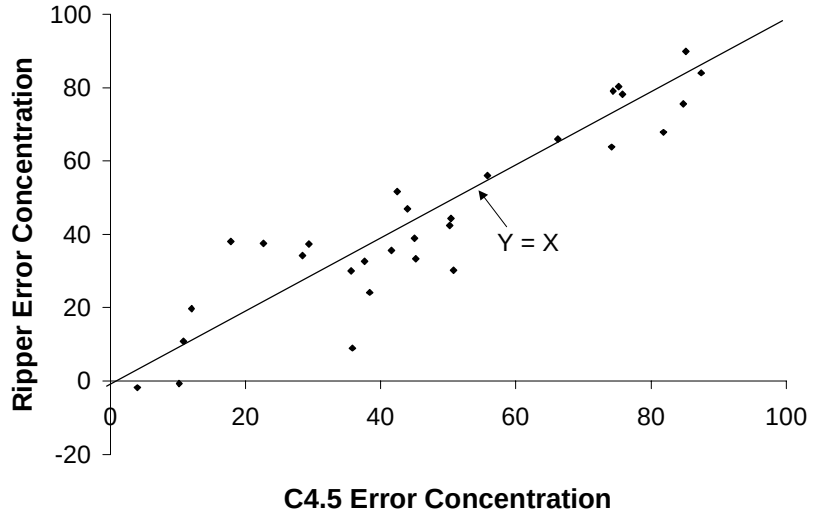


Figure B1: Comparison of Error Concentrations without Pruning

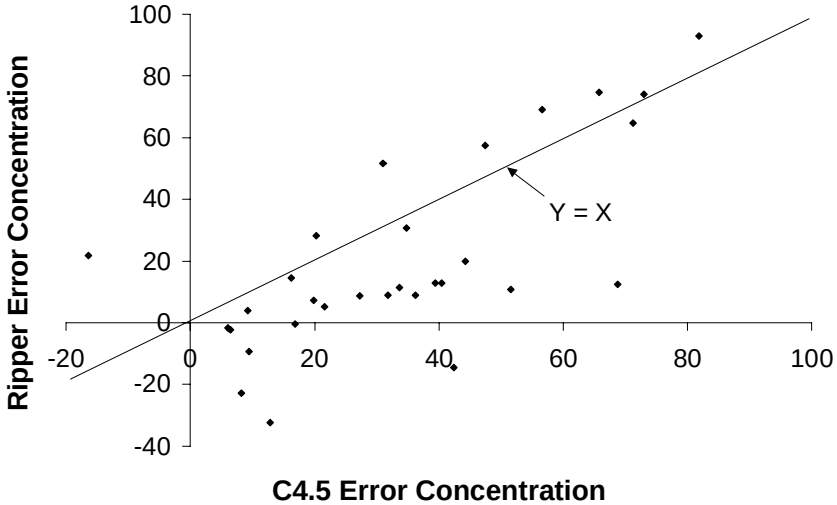


Figure B2: Comparison of Error Concentrations with Pruning

C4.5 tends to have a higher error concentration, with or without pruning. This suggests that C4.5 has a more specific bias than RIPPER. For the case without pruning, this difference might be influenced by the fact that C4.5 will try to perfectly fit the training data, but RIPPER will not, due to some of its heuristics that effectively pre-prune the rules. With pruning RIPPER's EC decreases much more than C4.5's, indicating that it may have a more aggressive pruning strategy.

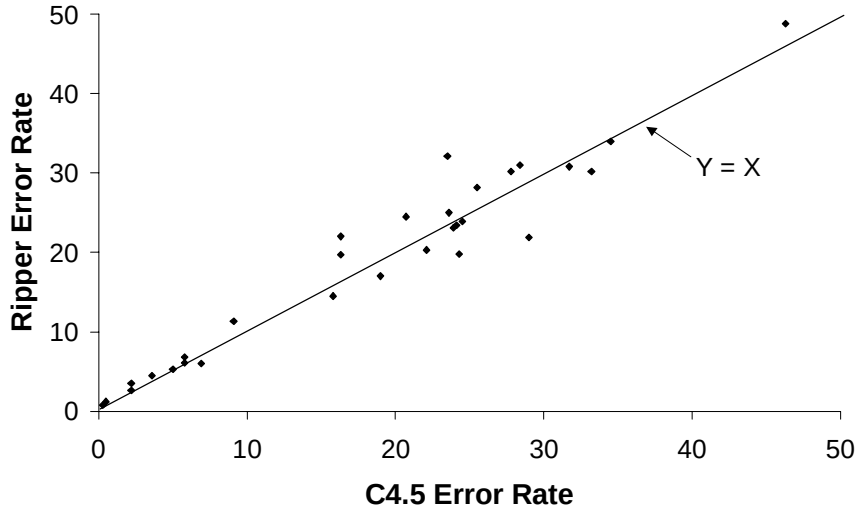


Figure B3: Comparison of C4.5 and RIPPER Error Rates w/o Pruning

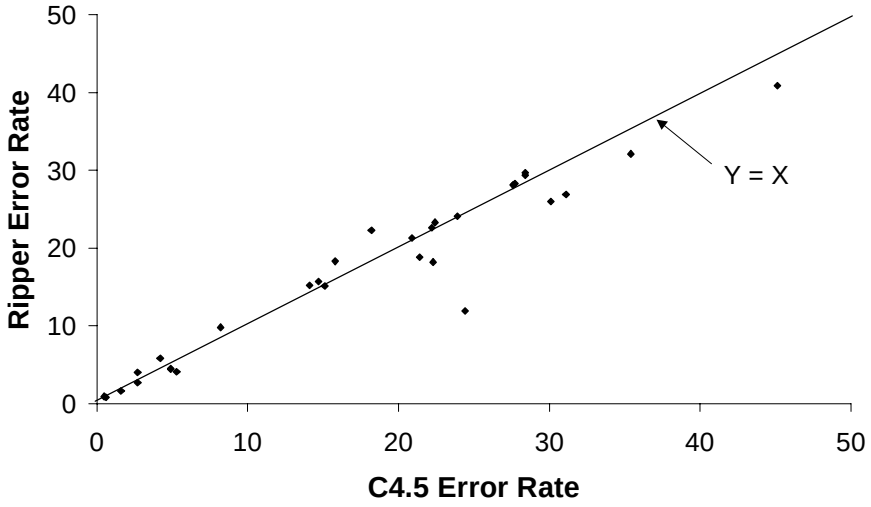


Figure B4: Comparison of C4.5 and RIPPER Error Rates with Pruning

C4.5 seems to perform slightly better when there is no pruning. With pruning, RIPPER does significantly better than C4.5 on a number of datasets. Note that with pruning RIPPER outperforms C4.5 most when the error rate is above 20%-- this might indicate that RIPPER's pruning strategy is more suitable for domains with higher error rates or a lot of noise.

Appendix C:
The Effect of Pruning

The results in this appendix show the impact of pruning on small disjuncts. Tables C1 and C2 present results that have been averaged over all 30 datasets. Table C3 shows more detailed results than what was presented in Section 4.3, Table 2 (see Section 4.3 for a description of the idealized pruning strategy).

Table C1: Comparison of Averaged Summary Statistics

Pruning Strategy	Error Rate	Largest Disjunct	Number Leaves	% Errors at 10% Correct	% Errors at 20% Correct	% Correct at 50% Errors	Error Conc
No Pruning	18.4	412	914	39.5	56.7	21.0	47.1
With Pruning	17.5	742	170	31.6	46.2	26.4	33.5

Table C2: Comparison of Mean Statistics

Pruning Strategy	Mean Coverage	Mean Correct Coverage	Mean Error Coverage	Mean Ratio
No Pruning	203.9	211.3	65.0	4.7
With Pruning	415.0	435.7	215.0	2.1

Table C3: Comparison of Error Rate with Pruning vs. Idealized Strategy

Dataset	Unpruned ER (%)	Pruned ER (%)	Idealized (smallest 10%)			Idealized (smallest 20%)		
			Ideal ER (%)	Absolute % decrease	Relative % decrease	Ideal ER (%)	Absolute % decrease	Relative % decrease
kr-vs-kp	0.3	0.6	0.1	0.5	86.1	0.0	0.6	92.2
hypothyroid	0.5	0.5	0.1	0.4	83.5	0.1	0.4	88.3
vote	6.9	5.3	2.2	3.1	59.0	0.5	4.8	89.9
splice-junction	5.8	4.2	1.6	2.6	62.3	0.7	3.5	82.9
ticket2	5.8	4.9	1.6	3.3	67.2	1.3	3.6	73.6
ticket1	2.2	1.6	1.1	0.5	30.2	0.3	1.3	83.3
ticket3	3.6	2.7	1.6	1.1	40.3	0.7	2.0	73.4
soybean-large	9.1	8.2	4.9	3.3	40.4	1.2	7.0	85.8
breast-wisc	5.0	4.9	3.0	1.9	39.0	2.3	2.6	52.1
ocr	2.2	2.7	1.2	1.5	56.2	1.0	1.7	64.3
hepatitis	22.1	18.2	18.1	0.1	0.8	13.0	5.2	28.8
horse-colic	16.3	14.7	12.9	1.8	12.2	10.4	4.3	29.0
crx	19.0	15.1	15.0	0.1	0.8	10.1	5.0	33.1
bridges	15.8	15.8	15.1	0.7	4.7	12.8	3.0	18.7
heart-hungarian	24.5	21.4	19.8	1.6	7.7	18.0	3.4	15.9
market1	23.6	20.9	19.4	1.5	7.0	16.6	4.3	20.5
adult	16.3	14.1	13.4	0.7	5.2	11.4	2.7	19.2
weather	33.2	31.1	29.1	2.0	6.4	24.7	6.4	20.5
network2	23.9	22.2	19.4	2.8	12.8	17.2	5.0	22.3
promoters	24.3	24.4	19.3	5.1	20.8	17.1	7.3	30.0
network1	24.1	22.4	20.7	1.7	7.7	18.1	4.3	19.1
german	31.7	28.4	29.8	-1.4	-4.8	26.6	1.8	6.3
coding	25.5	27.7	22.8	4.9	17.8	21.4	6.3	22.8
move	23.5	23.9	22.1	1.8	7.6	20.3	3.6	15.1
sonar	28.4	28.4	27.0	1.4	4.8	25.7	2.7	9.4
bands	29.0	30.1	13.6	16.5	54.7	15.1	15.0	49.9
liver	34.5	35.4	33.6	1.8	5.2	32.4	3.0	8.5
blackjack	27.8	27.6	25.8	1.8	6.4	24.7	2.9	10.4
labor	20.7	22.3	16.1	6.2	27.7	16.5	5.8	26.2
market2	46.3	45.1	46.2	-1.1	-2.5	45.8	-0.7	-1.5
Averages	**18.4**	**17.5**	**15.2**	**2.3**	**25.6**	**13.5**	**4.0**	**39.7**

Table C4: Effect of Pruning on Larger Disjuncts

This table is an expanded version of Table 3 that appears in Section 4.3. See the description that precedes Table 3 for a description of this table.

Dataset	% Error Rate at 10% covered			% Error Rate at 20% covered			% Error Rate at 30% covered			% Error Rate at 40% covered			% Error Rate at 50% covered			% Error Rate at 60% covered			% Error Rate at 70% covered			% Error Rate at 80% covered			% Error Rate at 90% covered			% Error Rate at 100% covered		
	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ	prune	none	Δ
kr-vs-kp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.3	0.0	0.3	0.5	0.1	0.4	0.6	0.3	0.3
hypothyroid	0.1	0.3	-0.2	0.2	0.1	0.0	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.5	0.5	0.0
vote	3.1	0.0	3.1	1.2	0.0	1.2	1.0	0.0	1.0	1.2	0.0	1.2	0.9	0.0	0.9	1.6	0.8	0.8	2.3	0.7	1.6	2.3	1.9	0.4	2.3	1.8	0.5	5.3	6.9	-1.6
splice-junction	0.3	0.9	-0.6	0.3	0.5	-0.2	0.2	0.3	-0.1	0.2	0.2	-0.1	0.3	0.2	0.1	1.2	0.7	0.5	2.4	0.6	1.8	2.8	0.9	2.0	2.7	2.2	0.5	4.2	5.8	-1.6
ticket2	0.3	0.0	0.3	1.8	0.3	1.5	2.7	0.8	1.9	2.5	0.9	1.6	2.5	0.7	1.8	2.5	0.6	1.9	2.5	1.0	1.5	2.4	0.9	1.5	2.4	2.2	0.2	4.9	5.8	-0.9
ticket1	0.1	2.1	-1.9	0.2	1.3	-1.1	0.3	0.6	-0.3	0.4	0.5	0.0	0.4	0.4	0.0	0.3	0.3	0.0	0.3	0.3	0.0	0.2	0.5	-0.2	1.0	1.2	-0.2	1.6	2.2	-0.5
ticket3	2.1	2.0	0.1	1.9	1.4	0.5	1.7	1.2	0.5	1.4	0.9	0.4	1.4	0.7	0.6	1.5	0.6	0.9	1.5	0.5	1.0	1.4	0.7	0.7	1.8	1.4	0.4	2.7	3.6	-0.9
soybean-large	1.5	0.0	1.5	3.8	0.0	3.8	5.4	1.0	4.4	6.1	1.2	4.9	5.3	1.6	3.7	5.0	1.5	3.5	4.7	1.3	3.5	4.2	2.8	1.4	4.4	6.2	-1.8	8.2	9.1	-0.9
breast-wisc	1.5	1.1	0.4	0.7	1.4	-0.7	1.0	1.0	0.0	0.7	0.7	0.0	0.6	0.6	0.0	0.7	1.4	-0.7	1.0	1.4	-0.4	1.9	2.2	-0.2	3.3	3.3	0.0	4.9	5.0	-0.1
ocr	1.5	1.8	-0.3	2.2	1.1	1.1	1.9	0.8	1.1	1.5	0.6	0.9	1.3	0.6	0.7	1.6	1.0	0.6	1.9	1.0	0.9	1.8	1.0	0.8	1.7	1.3	0.4	2.7	2.2	0.5
hepatitis	5.4	6.7	-1.3	10.8	3.3	7.5	15.0	2.2	12.9	14.4	7.6	6.8	15.0	9.1	5.9	13.6	10.9	2.8	12.8	12.1	0.6	14.6	15.5	-1.0	16.3	19.9	-3.7	18.2	22.1	-3.9
horse-colic	20.2	1.8	18.4	17.9	3.3	14.6	14.6	4.6	10.0	11.4	5.8	5.5	11.7	5.3	6.3	10.6	7.2	3.3	10.7	10.6	0.1	10.7	11.6	-0.9	11.0	14.4	-3.4	14.7	16.3	-1.7
crx	7.0	7.3	-0.3	8.0	7.0	1.0	7.9	6.5	1.4	7.3	5.6	1.6	6.3	7.3	-0.9	6.9	8.1	-1.2	7.8	9.3	-1.6	8.2	12.3	-4.1	11.4	16.3	-4.9	15.1	19.0	-3.9
bridges	10.0	0.0	10.0	20.0	0.0	20.0	17.5	0.0	17.5	15.0	0.7	14.4	16.8	2.0	14.9	16.1	8.4	7.6	14.9	9.4	5.4	14.1	12.8	1.3	14.1	14.6	-0.4	15.8	15.8	0.0
heart-hungarian	15.4	6.2	9.2	18.3	6.8	11.5	18.4	11.4	7.0	15.9	10.2	5.6	15.6	10.9	4.7	15.2	12.5	2.7	16.0	16.4	-0.4	17.5	19.0	-1.5	20.2	21.4	-1.2	21.4	24.5	-3.1
market1	16.6	2.2	14.4	13.3	4.9	8.4	12.2	7.8	4.4	12.0	9.9	2.1	12.7	12.1	0.6	13.2	14.4	-1.3	14.5	15.9	-1.4	16.1	18.1	-2.0	18.4	20.8	-2.4	20.9	23.6	-2.6
adult	3.9	0.5	3.4	3.4	3.3	0.1	3.6	4.9	-1.3	8.8	7.2	1.5	8.9	8.1	0.8	8.0	9.5	-1.5	8.3	10.6	-2.3	9.2	12.0	-2.8	11.3	14.1	-2.8	14.1	16.3	-2.2
weather	5.4	8.6	-3.2	8.4	10.2	-1.8	10.6	14.0	-3.4	13.5	16.4	-3.0	16.4	19.4	-3.1	19.6	22.2	-2.7	22.7	24.6	-1.9	25.6	27.5	-1.9	28.6	30.9	-2.3	31.1	33.2	-2.1
network2	10.8	9.1	1.7	12.0	7.6	4.4	12.5	10.7	1.8	12.9	12.9	0.1	12.7	14.7	-2.0	14.0	15.7	-1.7	15.1	17.2	-2.1	17.2	18.1	-0.9	19.0	20.9	-1.9	22.2	23.9	-1.8
promoters	10.2	19.3	-9.1	10.9	9.4	1.5	10.9	10.4	0.4	13.7	11.0	2.8	14.1	15.7	-1.6	19.0	15.6	3.3	19.6	16.8	2.8	22.6	20.1	2.5	23.7	22.6	1.1	24.4	24.3	0.1
network1	15.3	7.4	7.9	13.1	8.7	4.4	13.1	11.8	1.3	13.4	14.3	-0.9	13.2	15.5	-2.3	15.0	16.0	-1.0	16.7	17.3	-0.6	18.2	19.4	-1.2	20.2	21.4	-1.2	22.4	24.1	-1.7
german	10.0	4.9	5.1	11.4	8.8	2.6	11.1	12.5	-1.4	11.9	16.0	-4.1	17.4	19.1	-1.8	18.9	24.1	-5.2	20.4	25.7	-5.3	22.5	27.6	-5.1	25.9	30.2	-4.3	28.4	31.7	-3.3
coding	19.8	8.5	11.3	16.6	12.0	4.6	18.7	14.3	4.4	19.6	16.2	3.4	21.1	17.9	3.2	22.7	19.2	3.5	23.6	20.6	3.1	25.1	21.9	3.3	26.3	23.1	3.2	27.7	25.5	2.2
move	24.6	9.0	15.6	22.4	10.0	12.4	19.2	12.1	7.1	20.5	13.2	7.3	21.0	15.5	5.6	21.8	17.5	4.3	22.6	18.7	3.8	22.9	20.8	2.1	23.0	22.6	0.4	23.9	23.5	0.3
sonar	27.6	27.6	0.0	25.5	25.5	0.0	23.7	23.7	0.0	21.6	21.6	0.0	19.2	19.2	0.0	21.7	21.1	0.6	24.4	24.3	0.1	26.6	26.5	0.1	27.2	27.2	0.0	28.4	28.4	0.0
bands	13.1	0.0	13.1	26.3	10.1	16.2	34.3	16.3	18.0	34.2	22.4	11.8	34.1	25.0	9.1	34.0	25.8	8.2	33.8	26.6	7.2	33.8	27.4	6.4	33.1	28.2	4.9	30.1	29.0	1.1
liver	27.5	36.2	-8.8	30.0	34.3	-4.3	32.4	28.1	4.3	30.1	27.5	2.6	28.0	30.1	-2.2	29.8	31.0	-1.2	30.7	31.8	-1.2	32.3	32.6	-0.4	34.0	33.6	0.5	35.4	34.5	0.9
blackjack	25.3	26.1	-0.8	25.1	23.5	1.6	25.1	25.8	-0.8	24.7	27.6	-2.9	24.8	26.7	-1.9	26.6	23.9	2.7	26.1	24.4	1.7	25.2	24.8	0.4	26.0	26.1	-0.1	27.6	27.8	-0.2
labor	25.0	25.0	0.0	25.0	25.0	0.0	17.5	24.8	-7.3	18.6	22.1	-3.6	23.6	20.3	3.2	24.3	20.6	3.6	24.4	17.5	6.9	24.4	15.6	8.8	21.6	16.6	5.0	22.3	20.7	1.6
market2	44.1	45.5	-1.4	42.3	45.3	-3.0	43.1	44.3	-1.2	42.8	44.3	-1.5	42.5	44.2	-1.7	42.7	44.5	-1.8	43.3	45.3	-2.0	44.0	45.9	-1.9	44.6	46.2	-1.7	45.1	46.3	-1.2
Average	11.6	8.7	**2.9**	12.4	8.8	**3.6**	12.5	9.7	**2.8**	12.5	10.6	**2.0**	12.9	11.4	**1.5**	13.6	12.5	**1.1**	14.2	13.4	**0.8**	14.9	14.7	**0.3**	15.9	16.4	**-0.5**	17.5	18.4	**-0.9**

Figures C1 – C4 are scatter plots that display the information presented in Table C4 for 4 of the 10 coverage values (20%, 50%, 70% and 100%). The concepts represented as a point in each scatter plot are built by starting with the largest disjunct and then adding disjuncts until the specified percentage of examples are covered. Each scatter plot contains 30 points since there are 30 datasets. Note that in most cases the error rate without pruning is lower than the error rate with pruning. Figure C5 is a copy of Figure 6 that appears in Section 4.3.

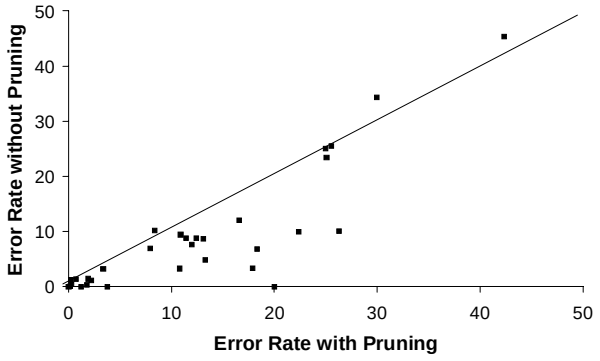


Figure C1: Effect of Pruning when 20% Covered

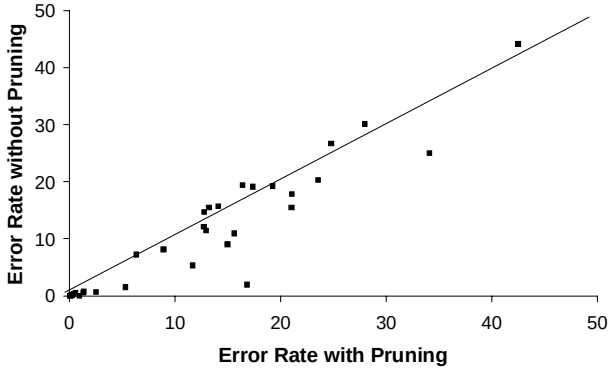


Figure C2: Effect of Pruning when 50% Covered

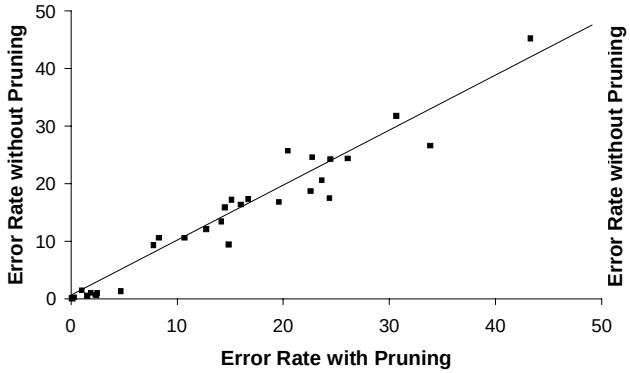


Figure C3: Effect of Pruning when 70% Covered

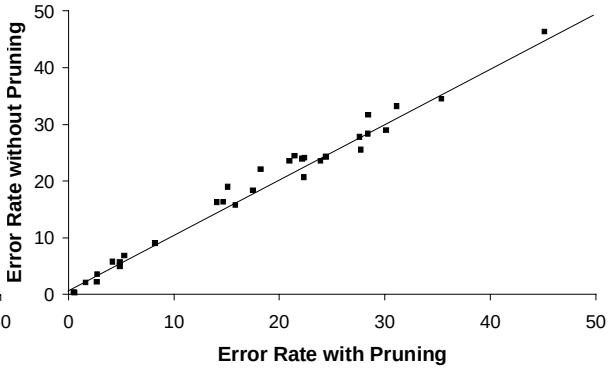


Figure C4: Effect of Pruning when 100% Covered

Figures C1-C4: Effect of Pruning when Concept Built from Largest Disjuncts

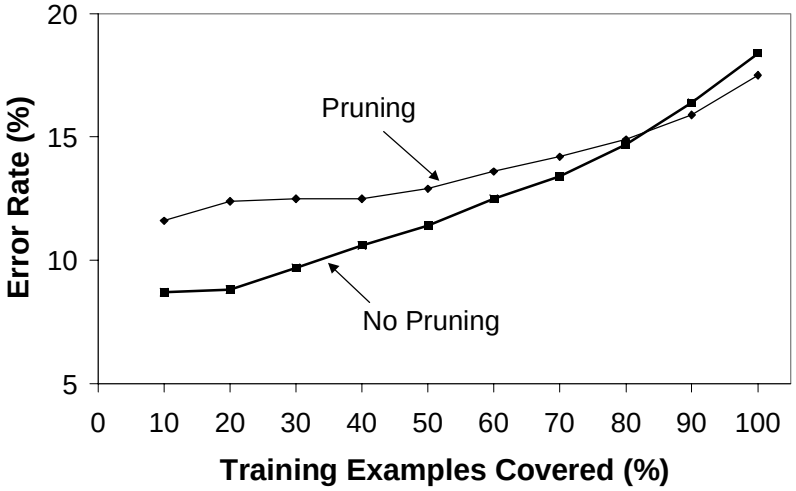


Figure C5: Averaged Error Rate based on Concept Built from Largest Disjuncts

Appendix D
Effect of Varying Training Set Size

This appendix presents data related to varying the training set size. Table D1 shows how the training set size impacts the distribution of the errors, by providing the Error Concentration and mean disjunct size values for different training set sizes (the error rates are also listed). It also shows, in the last column, the amount by which the Error Concentration decreases and the error rate increases, as the training set size is reduced by a factor of 9.

Note: The Mean Stats field is of the form X (Y/Z), where Y is the mean disjunct size of the correctly classified test examples, Z is the mean disjunct size of the incorrectly classified test examples, and X is the ratio of these two values (Y divided by Z). Mean disjunct size is defined in Section 3 of this paper.

Table D1: Effect of Training Set Size on Small Disjuncts

Domain	Dataset Size	90% Training Data			50% Training Data			10% Training Data			90% => 10%	
		Error Rate	Error Conc	Mean Stats	Error Rate	Error Conc	Mean Stats	Error Rate	Error Conc	Mean Stats	EC decrease	Err Rate increase
kr-vs-kp	3196	0.3	87.4	14.3 (403/28)	0.7	88.4	11.9 (224/19)	3.9	74.2	4.8 (47/10)	13.2	3.6
hypothyroid	3771	0.5	85.2	3.9 (1290/331)	0.6	83.8	7.9 (1224/156)	1.3	91.0	13.6 (232/17)	-5.8	0.8
vote	435	6.9	84.8	13.3 (133/10)	6.7	76.2	5.6 (78/14)	9.0	62.6	2.2 (20/9)	22.2	2.1
splice-junction	3175	5.8	81.8	7.5 (101/13)	6.3	80.6	10.3 (98/10)	8.5	76.0	6.9 (31/5)	5.8	2.7
ticket2	556	5.8	75.8	6.6 (237/36)	5.7	78.8	6.4 (143/22)	7.0	36.4	1.9 (42/22)	39.4	1.2
ticket1	556	2.2	75.2	5.5 (283/51)	3.2	85.2	7.6 (170/22)	2.9	47.6	2.8 (39/14)	27.6	0.7
ticket3	556	3.6	74.4	5.4 (249/46)	4.1	51.2	2.9 (170/58)	9.5	67.2	4.1 (43/10)	7.2	5.9
soybean-large	682	9.1	74.2	5.2 (22/4)	13.8	66.0	3.4 (12/3)	31.9	48.4	2.5 (4/1)	25.8	22.8
breast-wisc	699	5.0	66.2	4.6 (203/44)	5.4	65.0	4.4 (130/29)	9.2	36.6	1.7 (32/18)	29.6	4.2
ocr	2688	2.2	55.8	3.3 (635/192)	2.9	50.2	2.4 (388/163)	8.9	50.6	2.4 (110/46)	5.2	6.7
hepatitis	155	22.1	50.8	2.9 (25/9)	22.5	52.6	2.8 (24/9)	22.2	31.8	1.5 (9/6)	19.0	0.1
horse-colic	300	16.3	50.4	2.6 (38/14)	18.7	53.4	2.7 (27/10)	23.3	45.2	1.7 (10/6)	5.2	7.0
crx	690	19.0	50.2	2.6 (18/7)	19.1	42.6	2.0 (13/6)	20.6	46.0	2.7 (13/5)	4.2	1.6
bridges	101	15.8	45.2	5.1 (16/3)	14.6	27.0	2.3 (12/5)	16.8	10.0	1.4 (6/4)	35.2	1.0
heart-hungarian	293	24.5	45.0	2.2 (43/20)	22.1	41.6	2.2 (32/14)	23.7	21.6	1.4 (10/7)	23.4	-0.8
market1	3180	23.6	44.0	3.0 (50/17)	23.9	42.2	2.5 (37/15)	26.9	32.2	1.8 (22/12)	11.8	3.3
adult	21280	16.3	42.4	7.5 (213/29)	17.2	45.2	10.3 (205/20)	18.6	48.6	9.2 (80/9)	-6.2	2.3
weather	5597	33.2	41.6	2.6 (31/12)	32.7	38.0	2.7 (34/13)	34.0	34.0	2.0 (25/12)	7.6	0.8
network2	3826	23.9	38.4	2.7 (193/71)	24.9	34.2	2.3 (103/44)	27.8	35.4	1.9 (61/32)	3.0	3.9
promoters	106	24.3	37.6	1.9 (7/4)	22.4	20.6	1.9 (7.0/4)	36.0	10.8	1.4 (3/2)	26.8	11.7
network1	3577	24.1	35.8	2.4 (174/72)	25.1	35.4	2.5 (111/44)	28.6	31.4	1.9 (47/25)	4.4	4.5
german	1000	31.7	35.6	2.7 (12/4)	33.3	33.4	3.6 (17/5)	34.3	24.8	1.8 (10/5)	10.8	2.6
coding	20000	25.5	29.4	3.0 (10/3)	30.6	28.0	2.4 (7/3)	38.4	21.4	1.6 (4/3)	8.0	12.9
move	3028	23.5	28.4	1.8 (7/4)	25.9	26.8	1.8 (6/3)	33.7	15.8	1.4 (4/3)	12.6	10.2
sonar	208	28.4	22.6	1.3 (31/24)	27.3	29.2	1.2 (23/18)	40.4	2.8	1.1 (8/8)	19.8	12.0
bands	538	29.0	17.8	20.5 (8/0.4)	30.7	15.2	9.6 (5/1)	36.8	10.0	1.6 (2/1)	7.8	7.8
liver	345	34.5	12.0	1.1 (23/21)	36.4	5.4	1.0 (20/19)	40.5	3.0	1.1 (15/14)	9.0	6.0
blackjack	15000	27.8	10.8	1.1 (938/836)	27.9	9.4	1.1 (722/661)	29.4	10.0	1.1 (209/189)	0.8	1.6
labor	57	20.7	10.2	1.0 (11/12)	17.0	4.4	1.0 (8/8)	30.3	11.4	1.2 (3/2)	-1.2	9.6
market2	11000	46.3	4.0	1.0 (36/34)	45.7	2.8	1.0 (33/32)	47.3	3.2	1.0 (19/19)	0.8	1.0
Average		18.4	47.1		18.9	43.8		23.4	34.7		12.4	5.0

Table D1 shows several trends. As expected, the error rates tend to increase as the training set size decreases. The Error Concentration is also shown to decrease in all but 3 of the 30 cases. As the training set size is reduced, the mean size of the correctly and incorrectly classified test cases become smaller, although the ratio of the two does not show nearly as clear a pattern.

Figures D1- D4 show the error rates of disjuncts of size 0-6, for 4 datasets, as the training set size is varied by a factor of 9. Note that with the exception of the Market2 dataset, the error rates clearly tend to be higher when there is less training data. The lack of this effect for the Market2 dataset is likely due to its EC being very close to 0, in which case there is little difference between the error rate of small and large disjuncts.

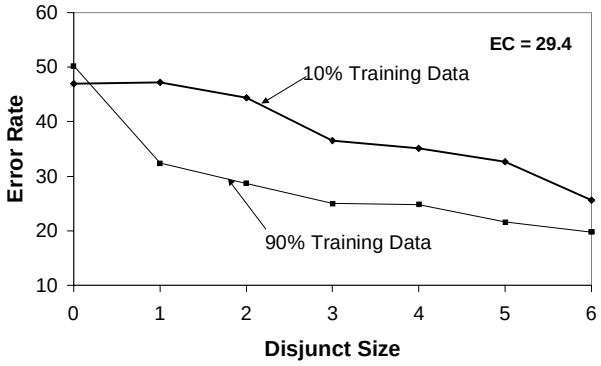


Figure D1: Coding Domain

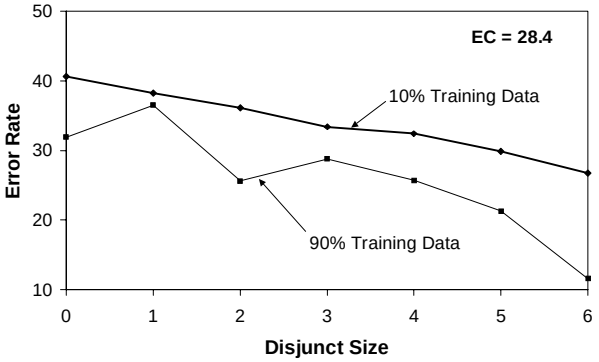


Figure D2: Move Domain

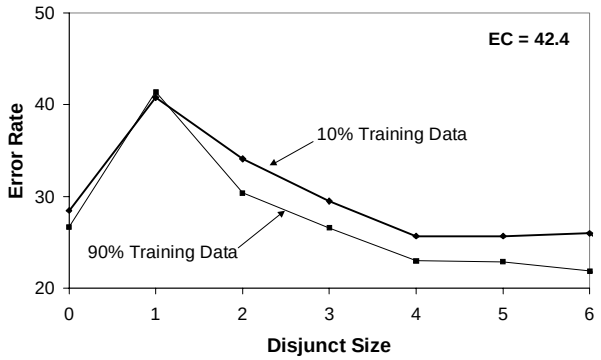


Figure D3: Adult Domain

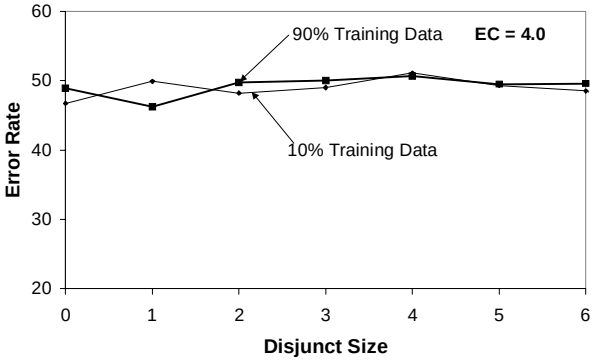


Figure D4: Market2 Domain

Appendix E:
Effect of Noise

E1. Summary Results Averaged over all Datasets

The experimental results that involve noise are summarized here—all results are based on the average over all 27 datasets that noise was applied to. Appendix E2 and E3 contain the results for each of the individual 27 datasets.

Table E1.1: Impact of Noise on Average of 27 Datasets

Type of Noise	Pruning	Measure	Noise Level								
			0%	1%	3%	5%	10%	20%	30%	40%	50%
Class	**No**	Error Rate	18.3	18.7	19.0	20.4	21.7	24.3	27.2	29.8	32.2
		EC	48.5	49.7	49.0	48.1	47.1	46.2	43.2	39.7	33.8
		Number Leaves	702.1	706.8	753.4	794.0	860.7	979.3	1070.1	1136.8	1179.7
		Disjunct Size (All)	202.0	122.3	91.3	81.3	76.2	70.3	72.6	86.6	89.4
		Disjunct Size (Errors)	64.1	52.7	48.0	49.0	48.9	48.4	53.1	65.1	71.2
		Disjuct Size (Correct)	209.7	129.3	97.3	86.8	82.8	76.9	79.2	94.2	96.1
		Disjunct Ratio	3.3	2.5	2.0	1.8	1.7	1.6	1.5	1.4	1.3
Class	**Yes**	Error Rate	17.2	17.2	16.9	17.5	17.9	18.6	19.4	20.5	22.7
		EC	35.8	37.9	36.6	36.2	37.0	38.9	40.5	38.8	38.1
		Number Leaves	110.3	111.0	110.9	112.3	112.3	118.7	123.5	126.2	139.4
		Disjunct Size (All)	428.3	429.4	427.6	425.7	420.0	118.7	123.5	331.8	139.4
		Disjunct Size (Errors)	219.7	212.0	217.9	212.2	204.9	176.3	168.1	165.1	170.3
		Disjuct Size (Correct)	450.6	453.9	451.2	450.1	445.9	382.1	370.3	353.2	346.2
		Disjunct Ratio	2.1	2.1	2.1	2.1	2.2	2.2	2.2	2.1	2.0
Attribute (Training set)	**No**	Error Rate	18.3	18.6	18.8	18.9	19.2	20.6	22.1	23.5	25.6
		EC	48.5	50.5	49.5	49.9	49.2	45.3	44.8	41.5	37.0
		Number Leaves	702.1	702.4	741.0	801.2	909.8	1052.5	1192.5	1270.2	1321.7
		Disjunct Size (All)	202.0	200.4	199.4	203.2	196.9	172.7	160.5	155.1	144.9
		Disjunct Size (Errors)	64.1	69.6	71.6	80.2	81.0	86.0	93.3	97.9	105.6
		Disjuct Size (Correct)	209.7	208.5	208.5	211.6	206.5	180.9	168.2	161.3	149.9
		Disjunct Ratio	3.3	3.0	2.9	2.6	2.5	2.1	1.8	1.6	1.4
Attribute (Training set)	**Yes**	Error Rate	17.2	17.3	17.6	17.7	17.5	19.2	20.5	21.3	23.3
		EC	35.8	38.4	37.3	38.7	36.2	37.2	39.8	35.9	33.1
		Number Leaves	110.3	106.4	96.3	96.0	92.0	92.2	90.0	84.6	80.9
		Disjunct Size (All)	428.3	418.7	411.9	401.4	382.1	385.3	378.3	562.0	584.7
		Disjunct Size (Errors)	219.7	212.1	212.9	210.0	203.0	229.5	233.3	456.7	500.9
		Disjuct Size (Correct)	450.6	441.1	433.9	421.7	402.0	403.9	398.7	576.3	597.1
		Disjunct Ratio	2.1	2.1	2.0	2.0	2.0	1.8	1.7	1.3	1.2
Attribute (Both)	**No**	Error Rate	18.3	19.1	19.9	21.2	23.1	27.0	29.0	32.1	33.0
		EC	48.5	46.5	47.2	47.3	43.5	40.7	34.8	32.7	27.4
		Number Leaves	702.1	700.6	723.6	792.0	917.4	1069.5	1207.2	1262.8	1334.0
		Disjunct Size (All)	202.0	181.9	172.4	183.2	174.4	136.2	114.7	112.9	132.1
		Disjunct Size (Errors)	64.1	59.8	55.6	69.8	73.8	65.4	74.7	80.8	107.6
		Disjuct Size (Correct)	209.7	190.3	180.9	194.9	186.3	148.1	122.4	120.1	138.2
		Disjunct Ratio	3.3	3.2	3.3	2.8	2.5	2.3	1.6	1.5	1.3
Attribute (Both)	**Yes**	Error Rate	17.2	17.9	18.6	19.8	21.4	24.2	26.5	29.1	30.1
		EC	35.8	34.9	36.5	37.9	35.2	33.8	32.3	27.2	21.4
		Number Leaves	110.3	105.9	95.6	93.5	95.6	91.0	89.5	84.0	83.4
		Disjunct Size (All)	428.3	407.6	385.7	398.6	353.8	314.2	315.1	495.7	617.1
		Disjunct Size (Errors)	219.7	209.0	196.5	210.1	194.8	174.9	207.3	418.0	536.9
		Disjuct Size (Correct)	450.6	430.8	409.0	424.6	377.3	339.7	337.9	514.5	638.4
		Disjunct Ratio	2.1	2.1	2.1	2.0	1.9	1.9	1.6	1.2	1.2

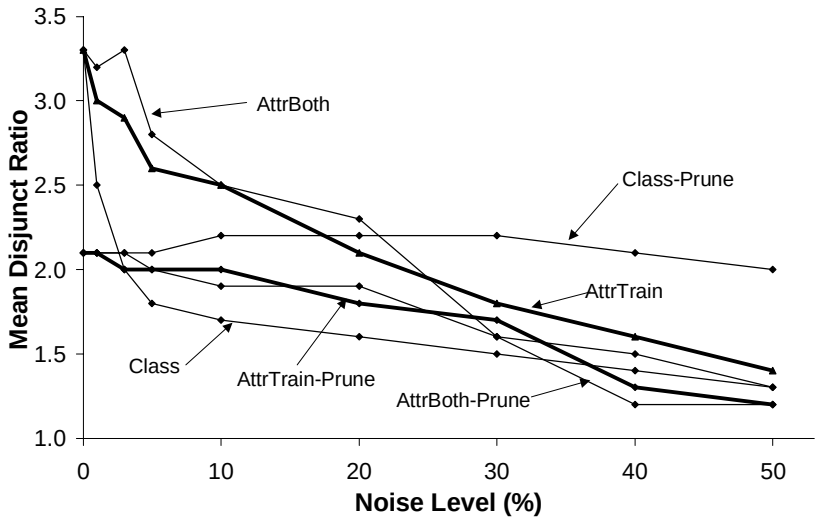


Figure E1.1: Effect of Noise on Averaged Mean Disjunct Ratio

Figure E1.1 shows the impact of different types of noise on the mean disjunct ratio. The mean disjunct ratio is defined as the means disjunct size of the correctly classified test examples divided by the mean disjunct size of the incorrectly classified test examples. The mean disjunct ratio is similar to the EC in that both represent the distribution of errors by disjunct size. A Figure similar to the one above, but displaying EC on the y-axis, appears in the body of this paper in Figure 10.

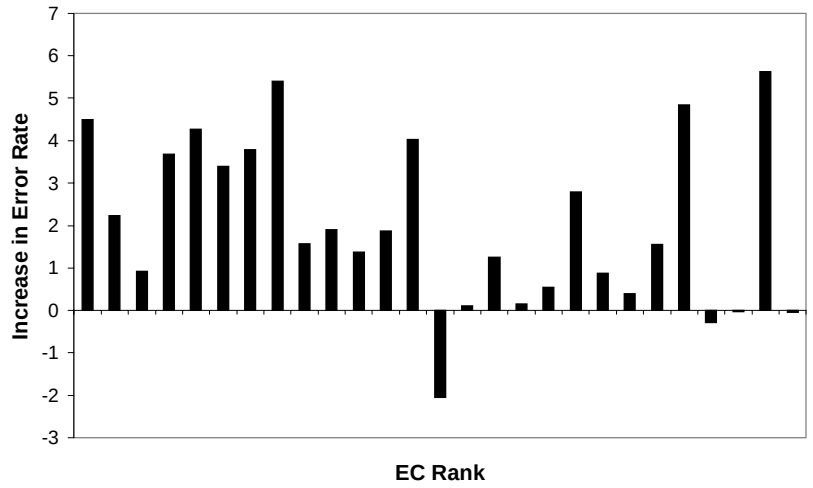


Figure E1.2: Sensitivity to 5% Class Noise

Figure E1.2 shows that those datasets with high EC tend to be more susceptible to class noise than those with low EC.

E2. Effect of Noise on Error Rate and Error Concentration

This part of the Appendix contains the error rate and error concentration that results from applying noise to each of the 27 datasets. There are a total of 6 tables:

- Tables E2.1/E2.2: Class noise without/with pruning
- Tables E2.3/E2.4: Attribute noise applied to the training set without/with pruning
- Tables E2.5/E2.6: Attribute noise applied to the training and test sets, without/with pruning

Table E2.1: Effect of Class Noise (No Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% Noise		50% Noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.3	87.4	1.3	93.8	2.9	94.0	4.8	91.6	8.1	91.6	12.6	86.0	19.4	75.2	23.1	69.2	27.1	59.0
hypothyroid	0.5	85.2	1.0	86.4	1.7	91.4	2.7	88.4	6.0	91.2	13.6	80.4	19.6	80.6	28.9	75.0	39.2	62.0
vote	6.9	84.8	7.6	70.4	6.4	83.4	7.8	80.4	9.0	63.4	13.1	70.8	19.1	75.0	23.9	67.0	23.7	63.4
splice-junction	5.8	81.8	6.6	84.0	7.6	82.2	9.5	79.6	11.6	78.6	16.7	71.8	21.7	65.0	25.2	58.4	31.3	51.4
ticket2	5.8	75.8	6.1	73.8	6.3	65.8	10.1	83.4	10.6	78.0	14.9	76.0	18.3	73.8	22.3	64.4	27.5	61.6
ticket1	2.2	75.2	3.1	82.6	5.0	81.4	5.6	72.0	7.9	76.2	14.1	82.0	16.9	79.2	21.6	72.4	26.3	61.8
ticket3	3.6	74.4	4.5	62.2	5.4	70.6	7.4	78.8	9.0	74.6	12.8	80.8	18.6	71.2	26.3	74.8	26.4	60.8
soybean-large	9.1	74.2	9.5	74.8	11.1	70.4	14.5	67.2	19.9	62.8	27.4	61.2	35.0	46.8	43.5	44	55.9	31.2
breast-wisc	5.0	66.2	5.2	76.8	6.9	81.2	6.6	68.8	7.9	72.0	8.6	76.6	11.7	72.2	11.9	68.2	11.6	66.2
hepatitis	22.1	50.8	24.7	52.2	18.8	49.8	24.0	52.8	24.1	47.4	30.3	48.8	30.7	44.8	32.3	49.6	35.6	26.0
horse-colic	16.3	50.4	17.3	53.2	19.7	45.2	17.7	42.4	21.0	50.0	19.3	24.8	24.7	43.8	23.7	32.6	29.3	29.0
crx	19.0	50.2	18.8	47.4	18.6	49.8	20.9	49.2	22.5	52.0	25.4	50.0	28.0	40.6	30.9	41.2	32.3	29.0
bridges	15.8	45.2	15.8	46.2	20.8	55.0	19.8	35.8	18.8	45.2	23.9	65.8	28.6	58.0	32.6	44.2	22.6	39.4
hungar-heart	24.5	45.0	24.2	48.2	22.8	36.6	22.5	40.2	23.8	41.2	21.1	35.2	25.8	36.6	25.9	32.4	22.5	13.2
market1	23.6	44.0	25.0	45.6	23.7	39.8	23.7	39.2	26.2	37.0	25.2	33.8	28.1	32.8	27.4	33.4	28.8	29.6
adult	16.3	42.4	16.5	37.6	16.9	33.8	17.6	31.0	18.5	28.0	20.2	25.6	22.3	24.0	24.6	24.4	26.9	22.2
weather	33.2	41.6	32.6	41.2	33.3	41.6	33.4	38.8	34.0	36.8	36.6	36.8	38.5	32.4	38.6	28.2	41.6	26.0
network2	23.9	38.4	24.6	40.0	24.2	37.8	24.4	39.2	25.1	39.8	25.0	37.2	26.3	34.0	25.6	32.2	28.3	31.8
promoters	24.3	37.6	20.5	45.8	25.4	39.0	27.1	33.2	40.4	60.4	37.6	52.0	29.5	53.4	44.1	38.2	44.2	45.6
network1	24.1	35.8	25.3	36.8	24.5	37.0	25.0	37.6	25.5	36.8	26.6	37.2	28.2	35.0	28.2	29.8	26.6	24.4
german	31.7	35.6	31.5	34.0	31.8	36.2	32.1	32.6	34.9	27.8	35.2	28.8	35.7	20.2	39.2	20.8	41.7	23.0
move	23.5	28.4	24.1	32.4	24.7	28.6	25.1	29.8	25.3	28.6	28.2	27.4	31.4	27.2	32.2	20.6	36.8	17.8
sonar	28.4	22.6	28.8	31.0	25.0	19.0	33.2	25.0	31.8	16.0	28.3	30.0	36.6	26.6	32.2	13.4	43.8	35.0
liver	34.5	12.0	36.2	12.0	32.5	15.0	34.2	17.2	33.4	4.4	36.8	5.6	36.2	-5.0	42.3	2.8	42.6	-4.6
blackjack	27.8	10.8	27.8	10.0	27.7	11.6	27.8	9.8	27.9	10.2	28.0	10.2	28.1	11.0	28.3	6.6	28.4	9.4
labor	20.7	10.2	20.7	19.0	22.7	24.8	26.3	32.8	15.3	19.2	27.7	10.4	29.0	9.0	21.0	21.4	21.3	-3.2
market2	46.3	4.0	46.3	3.4	46.2	2.6	46.3	2.2	46.7	3.2	47.5	2.4	47.3	3.0	48.4	6.0	47.5	1.4
Averages	**18.3**	**48.5**	**18.7**	**49.7**	**19.0**	**49.0**	**20.4**	**48.1**	**21.7**	**47.1**	**24.3**	**46.2**	**27.2**	**43.2**	**29.8**	**39.7**	**32.2**	**33.8**

Table E2.2: Effect of Class Noise (with Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% Noise		50% Noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.6	65.8	0.7	68.4	0.7	64.8	0.7	63.6	0.7	60.2	1.0	73.0	1.6	73.0	2.3	73.2	4.4	81.0
hypothyroid	0.5	81.8	0.4	82.4	0.5	80.6	0.5	87.0	0.5	90.4	0.9	88.8	0.9	91.0	1.5	91.6	5.3	92.2
vote	5.3	71.2	5.3	68.6	5.3	68.2	5.3	71.6	5.3	67.6	5.3	74.6	6.2	73.4	7.4	64.8	8.1	67.0
splice-junction	4.2	56.6	4.1	57.4	4.7	61.6	4.0	56.6	4.4	60.2	5.0	68.8	5.5	69.6	7.0	76.4	12.5	76.2
ticket2	4.9	47.4	4.9	39.2	4.9	47.0	5.0	55.6	5.4	41.4	5.9	41.8	6.8	68.4	8.5	68.2	12.9	67.4
ticket1	1.6	73.0	1.6	80.0	2.5	63.2	2.4	80.2	2.2	85.0	2.7	78.0	3.8	75.2	4.5	60.6	11.9	80.6
ticket3	2.7	31.0	3.2	41.0	2.3	38.0	3.1	52.8	3.3	31.6	3.6	30.8	4.9	67.6	7.5	60.0	8.6	59.2
soybean-large	8.2	39.4	8.5	44.8	8.6	37.0	9.2	40.6	11.4	57.4	10.5	42.2	11.1	57.8	13.8	51.8	20.4	63.4
breast-wisc	4.9	68.8	4.9	64.8	5.7	65.8	5.6	47.6	5.7	50.8	4.9	64.8	7.4	68.2	7.2	54.2	8.2	60.6
hepatitis	18.2	16.8	18.2	23.8	17.5	12.2	21.4	18.4	19.4	42.6	22.5	21.4	21.3	28.0	25.2	44.6	27.8	45.0
horse-colic	14.7	27.2	14.7	28.2	15.0	39.2	14.7	20.8	16.3	27.8	15.7	20.6	17.3	44.8	15.7	36.8	20.3	16.8
crx	15.1	51.6	15.4	52.8	13.5	42.0	14.4	48.6	15.2	51.2	14.5	51.0	14.4	51.4	17.5	46.6	19.7	38.4
bridges	15.8	6.4	15.8	5.8	15.8	4.2	14.8	0.0	14.8	0.8	15.8	2.4	18.8	3.6	15.8	8.2	15.8	6.4
hungar-heart	21.4	19.8	21.1	21.0	20.7	23.8	20.8	10.2	20.4	18.6	20.1	12.8	22.8	26.8	22.1	22.6	21.1	25.6
market1	20.9	33.6	21.6	37.2	20.8	39.0	21.4	39.0	22.5	36.4	22.5	33.4	24.2	36.6	24.3	29.0	25.9	31.0
adult	14.1	42.4	14.0	46.4	14.0	46.6	14.1	45.8	14.2	49.2	14.2	46.6	14.6	50.2	14.6	46.4	15.4	47.4
weather	31.1	44.2	31.1	42.6	30.8	43.0	31.3	42.4	33.1	42.0	35.3	39.4	37.2	36.0	38.5	32.4	40.8	29.0
network2	22.2	36.2	22.7	36.0	21.9	32.4	21.9	34.6	22.9	38.8	23.6	36.0	24.4	35.2	25.0	35.4	27.2	35.4
promoters	24.4	12.8	23.6	35.6	26.4	17.2	27.1	18.4	29.3	16.4	25.2	54.8	23.5	35.8	39.4	16.6	36.5	15.4
network1	22.4	31.8	22.9	31.6	22.8	31.4	23.3	35.2	24.1	34.2	25.0	36.8	26.0	36.4	26.6	33.8	25.4	28.0
german	28.4	40.4	28.5	39.8	28.5	37.8	26.9	36.4	29.0	40.2	29.5	39.4	29.2	33.4	29.0	35.8	34.0	36.8
move	23.9	9.4	24.3	8.8	24.2	9.8	24.6	9.8	24.3	6.8	26.8	8.0	27.4	10.8	28.8	8.2	32.1	9.4
sonar	28.4	20.2	27.9	27.4	25.4	21.0	32.8	26.2	31.8	19.0	28.8	32.4	38.0	33.0	33.7	17.8	41.9	31.6
liver	35.4	16.2	35.1	10.4	33.0	20.4	32.8	15.8	33.6	6.8	38.8	12.8	36.2	4.0	40.3	1.6	43.2	-1.2
blackjack	27.6	9.2	27.7	8.8	27.6	8.4	27.6	9.4	27.9	7.4	27.8	6.4	27.8	3.6	28.3	6.2	28.2	6.6
labor	22.3	8.2	22.3	13.0	19.0	25.6	20.7	6.0	19.0	10.6	31.0	25.8	27.3	-25.8	22.7	18.4	19.3	-24.8
market2	45.1	6.0	45.1	7.0	45.4	8.6	45.8	5.4	45.7	6.4	46.9	7.2	46.1	5.4	46.5	6.2	45.9	4.0
Averages	**17.2**	**35.8**	**17.2**	**37.9**	**16.9**	**36.6**	**17.5**	**36.2**	**17.9**	**37.0**	**18.6**	**38.9**	**19.4**	**40.5**	**20.5**	**38.8**	**22.7**	**38.1**

Table E2.3: Effect of Attribute Noise on Training Set (No Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% Noise		50% noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.3	87.4	0.8	96.6	2.0	92.2	2.7	92.4	6.7	87.4	10.9	80.2	17.0	71.4	23.4	57.2	27.1	49.2
hypothyroid	0.5	85.2	0.5	83.8	0.8	87.8	1.1	93.4	1.3	92.0	2.1	93.6	3.0	92.8	3.2	89.4	4.2	89.8
vote	6.9	84.8	6.7	86.2	6.2	80.2	7.1	79.6	7.6	88.2	7.1	77.6	8.7	79.0	7.6	67.2	8.0	76.8
splice-junction	5.8	81.8	6.9	84.0	6.9	79.8	8.9	81.2	9.9	77.2	14.1	72.4	16.6	68.4	21.5	65.0	24.1	59.4
ticket2	5.8	75.8	6.8	80.6	7.2	83.6	7.0	89.2	7.9	80.6	8.8	74.2	8.6	52.2	10.4	49.6	13.5	32.8
ticket1	2.2	75.2	2.9	77.0	2.3	83.8	3.8	84.8	3.1	83.4	6.5	59.8	10.6	78.8	12.1	65.4	18.4	52.6
ticket3	3.6	74.4	4.1	77.2	4.3	77.6	4.7	76.8	6.6	78.4	6.8	70.2	8.3	70.8	10.1	73.8	10.3	71.2
soybean-large	9.1	74.2	8.8	72.6	10.7	72.8	10.7	70.4	11.0	69.2	18.3	60.8	22.1	60.0	31.5	50.4	40.5	53.4
breast-wisc	5.0	66.2	6.3	67.2	6.2	81.4	5.4	68.6	5.0	68.2	5.6	65.2	4.7	65.2	5.2	62.8	7.0	70.8
hepatitis	22.1	50.8	21.4	58.0	24.7	57.6	24.0	56.0	18.1	52.2	24.0	53.4	19.4	41.6	20.1	43.4	25.2	45.6
horse-colic	16.3	50.4	17.3	48.0	15.3	45.0	14.0	47.0	18.7	49.8	16.3	33.6	19.3	25.4	19.3	41.6	20.3	30.2
crx	19.0	50.2	18.8	51.2	18.7	48.2	19.1	48.6	20.0	52.6	23.9	49.6	24.5	45.0	26.4	34.2	28.0	42.6
bridges	15.8	45.2	14.8	42.6	21.8	51.4	19.7	51.4	17.7	46.4	15.8	54.6	21.8	35.2	19.7	50.6	21.6	19.0
hungar-heart	24.5	45.0	23.8	41.2	21.8	35.4	20.7	45.8	19.4	32.4	21.8	40.8	20.1	37.6	22.5	42.4	23.8	45.6
market1	23.6	44.0	23.5	44.0	23.5	44.0	23.5	44.0	23.5	44.0	23.5	44.0	23.5	44.0	23.5	44.0	23.5	44.0
adult	16.3	42.4	16.7	41.8	16.7	41.6	17.3	41.4	18.3	41.2	19.7	42.6	21.0	42.0	22.0	44.2	22.9	44.4
weather	33.2	41.6	32.3	40.4	32.0	40.4	32.0	39.2	33.0	40.2	31.1	35.0	31.5	31.6	31.3	28.8	30.5	25.4
network2	23.9	38.4	23.8	36.8	23.8	36.4	23.2	33.8	23.8	34.8	22.5	33.2	23.2	33.2	23.8	32.8	24.9	34.8
promoters	24.3	37.6	21.6	47.4	24.5	33.0	29.1	50.2	27.2	53.0	30.1	45.4	32.8	54.8	29.5	27.4	36.6	15.4
network1	24.1	35.8	24.3	38.0	24.5	36.6	24.0	36.8	23.9	36.6	24.6	33.4	26.4	36.4	24.5	27.8	26.0	26.2
german	31.7	35.6	33.4	35.0	31.6	36.2	33.4	38.2	31.7	31.2	31.8	31.8	33.8	33.0	37.6	35.4	35.6	25.0
move	23.5	28.4	24.9	30.2	25.3	33.0	26.2	29.4	27.7	29.0	32.5	25.6	34.0	18.2	38.4	17.4	40.3	14.6
sonar	28.4	22.6	27.4	28.4	30.2	20.2	32.6	32.6	31.2	42.2	29.3	21.6	31.3	34.6	31.3	31.4	39.5	16.4
liver	34.5	12.0	36.8	17.6	36.2	14.2	33.3	3.4	39.4	3.6	39.4	4.0	36.8	-0.2	39.7	2.2	40.9	1.6
blackjack	27.8	10.8	27.9	11.4	27.9	9.4	27.8	7.4	28.1	4.0	28.0	4.0	28.4	0.4	28.8	-2.6	30.1	-5.8
labor	20.7	10.2	24.0	23.6	17.3	11.6	15.7	1.2	16.0	6.0	17.0	12.0	26.7	52.0	27.7	32.6	24.0	14.8
market2	46.3	4.0	45.6	3.8	44.7	4.4	44.0	5.4	42.8	4.0	43.5	5.2	43.8	6.6	44.9	6.2	45.1	3.8
Averages	**18.3**	**48.5**	**18.6**	**50.5**	**18.8**	**49.5**	**18.9**	**49.9**	**19.2**	**49.2**	**20.6**	**45.3**	**22.1**	**44.8**	**23.5**	**41.5**	**25.6**	**37.0**

Table E2.4: Effect of Attribute Noise on Training Set (with Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% noise		50% Noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.6	65.8	0.7	72.6	0.8	62.8	1.1	69.6	3.1	62.8	5.1	60.4	10.4	78.0	15.1	66.4	23.3	61.2
hypothyroid	0.5	81.8	0.5	85.4	0.7	84.8	0.7	83.6	1.1	87.6	1.5	87.0	2.4	88.8	2.8	85.4	3.5	59.8
vote	5.3	71.2	6.0	72.6	5.3	64.6	5.3	69.8	4.6	66.4	6.7	76.8	6.9	67.0	6.9	62.2	6.7	78.8
splice-junction	4.2	56.6	4.4	59.8	4.3	62.0	4.7	66.4	5.5	67.4	8.4	72.0	9.9	72.0	11.3	72.4	14.3	76.2
ticket2	4.9	47.4	5.6	48.8	5.4	53.0	6.1	51.4	7.0	53.0	8.6	27.6	8.4	23.4	10.1	16.2	9.9	9.6
ticket1	1.6	73.0	1.8	73.6	2.5	86.6	2.9	52.0	3.4	54.8	5.9	70.2	9.9	81.2	11.3	75.2	13.0	55.4
ticket3	2.7	31.0	3.4	54.2	3.6	57.0	4.1	60.6	5.8	66.4	7.2	52.8	9.2	66.8	9.4	66.4	9.7	67.8
soybean-large	8.2	39.4	9.2	52.6	10.7	51.0	10.4	57.2	11.3	47.8	17.6	50.0	20.1	52.6	29.6	51.0	38.2	52.0
breast-wisc	4.9	68.8	5.4	66.0	5.4	47.0	5.3	46.8	4.9	47.6	4.7	56.8	5.6	65.4	5.0	65.8	7.0	72.0
hepatitis	18.2	16.8	19.5	30.8	22.6	28.4	23.3	53.6	19.4	46.4	21.4	42.2	22.6	43.0	17.5	21.4	20.8	35.0
horse-colic	14.7	27.2	15.3	28.8	14.7	32.4	14.7	15.6	17.0	38.6	16.3	32.4	19.7	26.4	19.3	37.0	19.0	22.2
crx	15.1	51.6	15.8	53.0	14.2	44.8	14.9	49.4	14.4	37.4	15.5	48.0	16.4	54.0	17.8	43.2	20.1	54.8
bridges	15.8	6.4	15.8	-0.8	16.7	8.6	14.8	0.4	14.8	0.0	14.8	2.4	14.8	0.2	14.8	0.0	14.8	0.0
hungar-heart	21.4	19.8	21.5	27.4	23.5	30.4	19.7	35.0	19.1	29.0	21.4	35.4	20.7	37.0	22.5	39.4	22.1	39.8
market1	20.9	33.6	21.0	33.8	21.0	33.8	21.0	33.8	21.0	33.8	21.0	33.8	21.0	33.8	21.0	33.8	21.0	33.8
adult	14.1	42.4	14.2	40.0	14.3	39.4	14.6	40.6	15.1	35.0	16.3	33.0	17.4	38.8	17.6	9.8	17.8	1.2
weather	31.1	44.2	30.8	42.4	30.2	42.0	31.0	42.0	31.6	40.8	30.8	35.6	31.0	33.0	30.8	28.8	30.5	26.4
network2	22.2	36.2	22.2	32.8	22.8	37.4	21.6	32.8	23.1	38.8	22.7	34.2	23.3	36.2	24.0	33.8	24.9	36.8
promoters	24.4	12.8	19.7	22.8	23.7	18.0	28.2	44.6	19.7	11.8	30.1	18.4	34.6	38.2	29.5	36.6	37.5	30.8
network1	22.4	31.8	22.5	32.0	22.9	33.8	22.1	32.4	24.3	38.4	24.8	36.0	26.7	38.0	24.2	29.8	25.8	29.4
german	28.4	40.4	29.8	40.4	27.8	38.6	27.5	39.2	27.1	34.0	29.1	40.2	28.6	36.4	30.1	41.0	31.4	25.2
move	23.9	9.4	24.8	6.0	23.6	6.2	25.6	9.0	26.1	5.8	30.2	3.4	33.4	8.2	36.1	5.2	36.3	5.6
sonar	28.4	20.2	26.9	25.6	29.3	16.6	31.7	28.6	29.8	38.6	29.3	22.4	32.2	38.0	31.3	31.6	39.0	15.8
liver	35.4	16.2	35.1	18.0	36.2	11.8	35.4	7.2	38.2	5.2	39.1	5.0	36.5	0.6	38.6	0.2	40.9	1.6
blackjack	27.6	9.2	27.8	10.2	27.8	7.8	27.7	8.4	28.2	3.2	28.2	2.4	28.5	2.2	29.4	-2.0	30.3	-2.2
labor	22.3	8.2	22.3	3.0	21.0	2.2	20.7	10.0	14.0	-19.2	19.0	19.2	23.0	11.8	27.7	11.6	27.7	-0.2
market2	45.1	6.0	44.5	3.8	43.3	4.8	42.9	3.6	43.3	6.0	42.4	5.6	41.6	4.8	42.4	6.0	42.6	4.6
Averages	**17.2**	**35.8**	**17.3**	**38.4**	**17.6**	**37.3**	**17.7**	**38.7**	**17.5**	**36.2**	**19.2**	**37.2**	**20.5**	**39.8**	**21.3**	**35.9**	**23.3**	**33.1**

Table E2.5: Effect of Attribute Noise on Training and Test Sets (No Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% Noise		50% noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.3	87.4	2.2	66.0	6.1	61.0	9.9	64.8	17.7	59.6	28.2	48.2	35.9	38.6	37.8	29.0	42.9	16.8
hypothyroid	0.5	85.2	1.2	79.4	1.6	89.8	2.1	90.2	3.8	88.0	6.4	89.0	8.1	83.2	9.7	79.0	13.0	78.6
vote	6.9	84.8	5.7	77.0	6.9	83.4	8.5	70.8	6.9	71.2	11.5	48.8	11.9	61.2	19.3	50.8	25.5	39.4
splice-junction	5.8	81.8	6.6	76.2	9.5	72.8	10.7	76.4	15.2	64.0	19.8	60.2	23.6	50.6	29.0	41.2	31.5	39.4
ticket2	5.8	75.8	7.9	77.2	10.1	76.4	10.4	73.0	11.1	68.0	16.5	65.6	20.3	60.2	18.1	54.8	18.5	52.0
ticket1	2.2	75.2	3.6	81.0	7.4	65.2	8.8	78.0	12.4	72.2	21.0	72.2	17.6	59.4	26.8	53.6	27.5	55.2
ticket3	3.6	74.4	4.5	73.6	7.4	66.2	9.7	69.2	12.2	82.0	12.1	64.6	20.7	65.2	21.2	59.2	18.2	47.2
soybean-large	9.1	74.2	10.7	70.0	18.2	72.6	23.4	60.6	31.2	62.0	49.5	53.2	59.9	45.4	70.2	43.8	74.5	32.8
breast-wisc	5.0	66.2	6.0	65.4	6.6	76.6	8.6	72.0	6.4	62.6	6.9	54.0	9.0	61.4	10.7	56.2	10.2	52.0
hepatitis	22.1	50.8	18.2	53.8	21.2	45.8	19.4	31.4	18.2	50.4	29.8	50.8	19.9	54.0	27.2	47.8	24.0	45.6
horse-colic	16.3	50.4	16.7	42.0	13.3	36.6	19.7	50.6	25.7	28.6	24.7	18.8	28.0	29.4	34.0	28.6	30.3	24.0
crx	19.0	50.2	19.9	51.4	21.0	43.4	20.6	56.0	23.9	38.4	25.8	41.2	33.0	21.4	37.4	20.6	40.4	19.4
bridges	15.8	45.2	19.8	39.2	20.8	47.4	23.6	50.0	15.8	26.6	22.6	65.2	27.5	21.2	26.6	40.2	26.4	26.0
hungar-heart	24.5	45.0	21.5	46.6	21.5	46.2	22.0	38.6	25.1	37.2	22.4	41.4	23.1	36.4	29.6	33.2	28.2	7.4
market1	23.6	44.0	23.5	41.8	24.6	37.4	26.1	39.2	27.0	30.6	30.8	29.0	32.7	27.2	34.1	26.4	36.4	20.4
adult	16.3	42.4	16.8	40.8	18.0	45.4	18.9	43.0	20.6	41.2	24.3	40.0	25.9	29.6	27.6	26.8	28.9	25.0
weather	33.2	41.6	32.6	40.4	32.8	38.0	32.1	39.2	32.6	32.4	34.5	33.2	34.0	24.2	36.3	24.6	36.0	17.0
network2	23.9	38.4	24.5	38.6	25.7	39.2	25.9	36.4	26.8	38.4	27.2	37.2	27.6	31.8	28.0	33.2	28.3	31.2
promoters	24.3	37.6	29.0	33.8	24.3	47.4	15.1	38.8	28.9	56.0	31.0	48.8	50.9	16.2	33.9	34.2	40.6	-1.2
network1	24.1	35.8	25.8	36.2	27.4	38.4	27.1	35.6	27.7	36.6	28.8	34.0	28.8	30.0	30.1	27.2	29.7	24.0
german	31.7	35.6	32.2	31.0	31.4	30.2	32.4	36.0	35.5	38.2	37.7	25.8	38.6	28.2	38.0	26.0	40.8	24.4
move	23.5	28.4	24.8	31.6	25.6	28.0	27.9	28.8	31.8	23.6	38.9	17.4	43.7	13.6	45.5	10.6	46.3	4.4
sonar	28.4	22.6	33.6	34.4	30.9	38.8	27.0	39.4	32.6	18.8	35.1	17.2	33.2	20.6	42.8	17.4	41.8	-6.8
liver	34.5	12.0	35.1	3.8	33.3	7.4	40.0	7.4	41.1	6.2	40.0	6.0	40.0	2.4	44.3	2.0	41.2	1.8
blackjack	27.8	10.8	28.0	10.0	28.4	7.8	28.9	13.0	29.8	6.0	31.0	5.6	32.4	1.6	33.2	-1.6	34.0	-0.4
labor	20.7	10.2	20.7	9.2	19.0	30.0	30.0	34.4	19.3	30.6	26.7	26.4	10.7	24.0	27.7	16.4	29.0	61.8
market2	46.3	4.0	45.4	5.2	44.7	3.8	44.8	5.2	45.1	4.6	45.8	4.2	45.2	1.6	46.7	1.8	46.7	1.2
Averages	**18.3**	**48.5**	**19.1**	**46.5**	**19.9**	**47.2**	**21.2**	**47.3**	**23.1**	**43.5**	**27.0**	**40.7**	**29.0**	**34.8**	**32.1**	**32.7**	**33.0**	**27.4**

Table E2.6: Effect of Attribute Noise on Training and Test Sets (with Pruning)

Dataset	0% Noise		1% Noise		3% Noise		5% Noise		10% Noise		20% Noise		30% Noise		40% noise		50% Noise	
	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC	ER	EC
kr-vs-kp	0.6	65.8	2.0	39.6	4.2	27.2	7.3	42.2	11.5	23.4	20.7	31.2	29.9	35.8	34.4	28.0	40.2	22.2
hypothyroid	0.5	81.8	1.0	81.2	1.3	85.6	1.7	87.4	2.7	84.4	4.4	83.4	5.4	73.4	6.6	72.8	7.9	34.2
vote	5.3	71.2	5.5	67.2	6.7	57.6	6.0	37.6	6.2	42.6	10.1	46.8	11.0	33.2	17.9	32.4	23.9	36.0
splice-junction	4.2	56.6	4.8	52.2	5.9	51.0	8.1	59.0	9.4	44.0	14.3	43.4	17.8	39.6	23.2	43.2	23.0	32.0
ticket2	4.9	47.4	7.5	47.0	7.5	45.8	10.4	54.0	9.7	52.2	11.7	42.8	11.7	1.6	12.4	17.8	13.3	36.0
ticket1	1.6	73.0	2.9	77.6	5.4	61.6	7.7	67.8	10.2	61.6	15.8	67.6	15.3	60.2	20.7	58.4	24.3	60.8
ticket3	2.7	31.0	3.2	37.0	5.6	66.4	8.8	72.2	8.8	61.2	10.1	60.0	13.1	57.8	14.6	36.0	12.2	29.6
soybean-large	8.2	39.4	10.4	38.2	15.4	49.6	20.5	55.2	28.1	43.4	44.2	50.6	54.5	49.0	67.1	38.0	73.8	48.4
breast-wisc	4.9	68.8	6.0	49.0	5.4	68.8	7.0	55.6	5.7	48.0	7.6	54.8	9.5	59.6	10.6	55.0	10.4	50.0
hepatitis	18.2	16.8	18.1	37.0	21.9	30.6	21.2	26.2	14.3	31.4	27.9	31.2	21.9	40.8	27.3	51.0	23.3	37.2
horse-colic	14.7	27.2	14.7	26.2	12.0	29.8	17.3	39.8	22.0	27.8	24.7	30.6	25.7	27.6	29.7	19.2	31.0	23.0
crx	15.1	51.6	15.2	52.2	16.1	43.0	15.5	47.8	19.3	41.2	21.0	31.8	29.1	32.4	28.1	24.6	33.2	22.0
bridges	15.8	6.4	14.8	0.0	14.8	0.0	14.8	0.0	15.8	0.6	14.8	0.0	14.8	0.0	14.8	0.0	14.8	0.0
hungar-heart	21.4	19.8	22.8	34.6	20.4	35.4	24.8	33.4	25.5	34.6	23.5	34.0	24.1	37.8	29.6	33.6	28.9	12.6
market1	20.9	33.6	21.8	39.4	22.2	35.2	24.0	35.8	24.6	37.4	26.3	31.4	28.0	33.0	29.3	31.0	32.5	31.0
adult	14.1	42.4	14.5	38.0	15.1	39.8	16.1	36.2	17.2	34.8	19.5	42.0	20.0	42.6	21.8	2.8	22.5	10.8
weather	31.1	44.2	31.0	43.0	31.5	41.4	31.1	40.2	32.1	36.0	33.8	34.4	34.2	24.8	36.0	24.6	35.6	16.4
network2	22.2	36.2	22.2	32.2	23.6	35.4	24.3	38.6	25.7	36.8	26.6	36.6	26.8	30.8	28.2	36.0	28.1	31.4
promoters	24.4	12.8	26.3	29.4	29.3	41.4	18.0	31.2	27.1	41.6	23.4	34.4	50.7	26.6	30.4	18.6	39.5	-5.4
network1	22.4	31.8	23.7	34.8	25.3	37.2	25.1	37.0	27.3	41.6	28.1	34.2	28.4	29.2	29.6	31.6	29.9	27.2
german	28.4	40.4	26.7	35.2	28.9	38.0	29.8	36.6	28.9	36.2	29.2	35.6	32.1	37.4	31.8	30.0	35.5	21.0
move	23.9	9.4	25.1	13.6	27.0	5.6	25.5	9.0	32.1	13.0	37.2	12.4	41.5	11.0	42.9	10.0	43.5	8.0
sonar	28.4	20.2	32.7	31.4	29.9	36.0	26.5	37.6	32.1	16.2	34.1	12.6	35.1	29.0	42.3	16.0	41.8	-10.8
liver	35.4	16.2	36.0	7.0	33.9	11.4	38.0	11.2	41.1	7.0	39.7	4.8	40.3	1.8	44.3	2.0	41.5	4.6
blackjack	27.6	9.2	28.1	10.2	28.3	8.4	28.9	11.2	29.7	6.4	30.8	3.2	32.3	5.4	33.0	2.8	34.0	3.4
labor	22.3	8.2	22.3	-17.2	19.0	-1.2	31.3	13.6	26.3	39.0	29.7	14.0	17.7	44.6	33.3	14.2	22.3	-6.0
market2	45.1	6.0	44.3	7.0	44.5	5.4	43.7	7.4	43.8	7.4	44.4	8.0	44.4	5.8	45.5	5.8	45.2	1.0
Averages	**17.2**	**35.8**	**17.9**	**34.9**	**18.6**	**36.5**	**19.8**	**37.9**	**21.4**	**35.2**	**24.2**	**33.8**	**26.5**	**32.3**	**29.1**	**27.2**	**30.1**	**21.4**

E3. Effect of Noise on Disjunct Sizes

This appendix is similar to Appendix E2, except that it meaures the effect that noise has on disjunct size and the number of leaves in the decision tree induced by C4.5. Ideally these results would have been presented along with those in Appendix E2, but that much information could not be presented on a single page. The tables are presented in the same order as in Appendix E2:

- Tables E3.1/E3.2: Class noise without/with pruning
- Tables E3.3/E3.4: Attribute noise applied to the training set without/with pruning
- Tables E3.5/E3.6: Attribute noise applied to the training and test sets, without/with pruning

For each Table, the results are given for the following levels of noise: 1%, 5%, 10%, 20%, 30%, and 50%. Due to space considerations, we do not present the results for 3% and 40% noise, as we had done in the tables in Appendix E2. We present the results for 0% noise below, rather than in each table, to save space (the type of noise does not matter since the noise level is 0%).

Results with 0% Noise

Dataset	0% Noise (no pruning)				0% Noise (with pruning)			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	47.0	401.8	28.2	403.0	29.0	409.8	125.1	411.4
hypothyroid	38.0	2181.7	330.6	2190.1	14.5	2238.0	335.9	2247.1
vote	48.0	124.4	10.0	132.9	10.0	169.7	66.0	175.5
splice-junction	265.0	95.8	13.4	100.9	54.7	277.9	81.8	286.5
ticket2	28.0	224.9	36.0	236.5	9.1	400.7	200.9	410.9
ticket1	18.0	277.6	51.3	282.6	5.0	342.5	78.6	346.9
ticket3	25.0	241.9	46.0	249.2	6.2	379.0	214.7	383.5
soybean-large	175.0	19.9	4.1	21.5	62.0	29.0	17.7	30.0
breast-wisc	31.0	195.3	44.3	203.3	13.9	228.6	61.5	237.1
hepatitis	23.0	21.6	8.7	25.2	9.0	70.0	53.1	73.7
horse-colic	40.0	33.7	14.2	37.5	6.3	97.7	75.0	101.6
crx	227.0	15.5	6.9	17.6	22.5	184.5	101.2	199.3
bridges	32.0	14.2	3.2	16.3	2.2	60.3	54.3	61.4
hungar-heart	38.0	37.1	19.6	42.8	9.8	91.9	73.5	96.9
market1	718.0	42.2	16.8	50.1	135.0	379.3	237.4	416.9
adult	8434.0	182.6	28.5	212.6	419.0	2065.1	967.4	2244.8
weather	816.0	24.4	12.0	30.6	496.1	110.0	34.2	144.2
network2	382.0	163.7	70.8	192.9	150.9	948.2	549.1	1061.8
promoters	31.0	6.5	3.8	7.4	16.3	13.0	9.0	14.3
network1	362.0	149.1	71.5	173.7	142.0	782.6	472.5	871.9
german	475.0	9.5	4.4	11.8	91.5	132.7	62.3	160.6
move	2687.0	6.2	3.8	6.9	365.5	57.6	53.1	59.0
sonar	586.0	6.0	0.4	8.2	2.7	249.8	274.4	239.2
liver	35.0	22.3	20.7	23.2	22.0	30.9	27.3	32.9
blackjack	45.0	909.6	836.2	937.8	22.2	1711.8	1605.7	1752.3
labor	16.0	11.5	11.6	11.4	4.2	18.9	17.7	19.2
market2	3335.0	35.0	34.3	35.5	856.3	84.7	81.2	87.6
Averages	**702.1**	**202.0**	**64.1**	**209.7**	**110.3**	**428.3**	**219.7**	**450.6**

Table E3.1: Effect of Class Noise (No Pruning)

Dataset	1% Noise				5% Noise				10% Noise				20% Noise				30% Noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	116.7	203.7	9.9	206.2	317.9	61.3	2.7	64.2	518.4	27.0	1.5	29.3	832.2	13.3	1.3	15.1	1072.8	8.1	1.5	9.6	1393.8	4.8	1.7	5.9
hypothyroid	81.6	506.9	46.6	511.4	227.9	132.5	14.0	135.9	396.7	73.9	6.0	78.2	636.1	47.7	18.4	52.3	919.3	35.1	10.8	41.1	1346.6	22.9	12.5	29.7
vote	55.6	96.2	14.5	102.9	71.6	50.5	5.8	54.3	86.0	35.1	8.3	37.7	146.6	15.5	2.4	17.5	166.4	10.9	2.0	13.0	197.6	7.1	2.3	8.6
splice-junction	307.6	61.9	6.3	65.9	441.1	32.6	4.6	35.6	569.5	21.5	3.0	24.0	781.9	11.7	2.6	13.5	978.4	8.0	2.2	9.7	1218.7	4.8	2.2	6.0
ticket2	32.4	166.8	31.3	175.6	56.0	49.7	11.0	54.1	82.9	23.9	4.3	26.2	103.5	18.5	4.2	21.0	135.6	12.5	2.6	14.8	161.7	7.6	2.6	9.5
ticket1	24.4	191.6	15.3	197.2	39.8	91.1	11.8	95.8	62.7	33.5	8.2	35.7	102.6	18.4	4.0	20.7	125.0	12.6	2.9	14.6	166.5	7.1	2.5	8.7
ticket3	28.0	204.1	46.0	211.5	54.5	56.5	9.7	60.2	69.1	38.9	8.8	41.8	97.3	18.1	3.5	20.3	129.8	12.9	2.7	15.2	161.4	8.9	2.8	11.0
soybean-large	192.0	18.9	4.0	20.4	305.8	11.5	2.7	12.9	399.9	9.0	2.5	10.6	537.2	4.3	1.2	5.5	659.8	2.8	1.3	3.6	786.8	1.8	1.5	2.3
breast-wisc	32.8	194.4	19.2	203.9	42.6	126.3	18.9	133.9	46.3	92.3	17.7	98.7	65.0	63.7	11.0	68.7	79.9	54.4	9.1	60.5	83.2	51.1	16.2	55.6
hepatitis	23.6	21.5	8.1	25.9	30.1	13.2	4.4	16.0	27.9	15.7	5.6	18.8	35.0	10.8	4.8	13.4	36.4	7.6	4.4	9.0	45.6	5.6	4.0	6.5
horse-colic	41.7	35.5	13.5	40.1	48.5	20.1	11.1	22.1	53.2	20.3	8.2	23.5	65.3	14.4	9.9	15.5	66.7	12.6	8.0	14.1	85.0	9.9	6.6	11.2
crx	223.7	15.9	6.5	18.1	256.9	10.5	4.8	12.0	269.6	10.0	4.3	11.7	293.8	6.8	3.4	8.0	344.9	5.3	2.9	6.3	386.8	4.0	2.8	4.6
bridges	32.6	14.3	3.2	16.4	39.2	10.1	5.6	11.2	36.4	12.1	4.4	13.8	41.9	6.9	3.5	7.9	50.8	5.1	2.1	6.3	55.0	4.9	2.2	5.7
hungar-heart	38.3	38.5	20.0	44.4	39.1	44.6	23.0	50.8	39.9	36.3	17.1	42.3	42.6	25.5	15.1	28.3	44.4	27.9	15.1	32.4	45.8	43.4	36.4	45.5
market1	702.2	46.5	18.4	55.9	730.0	41.6	18.9	48.7	719.6	58.8	27.5	69.8	746.3	48.5	27.2	55.7	645.6	60.7	37.1	69.9	526.3	76.3	48.8	87.4
adult	8984.1	62.2	22.6	70.0	10372.3	38.6	20.4	42.4	11406.5	31.3	20.9	33.6	13016.4	25.2	17.5	27.2	14373.7	23.6	17.5	25.3	15843.7	22.5	16.0	24.9
weather	820.5	26.4	12.3	33.2	849.6	22.2	11.0	27.8	886.3	17.7	10.1	21.6	952.4	13.4	8.3	16.3	993.3	12.1	7.9	14.8	1051.7	11.2	8.2	13.4
network2	382.5	181.5	76.6	215.7	396.4	139.1	53.0	167.0	385.2	250.2	106.0	298.4	355.7	333.3	152.8	393.6	348.9	305.4	157.0	358.5	325.8	280.8	159.4	328.5
promoters	31.9	7.0	4.1	7.8	34.0	6.9	4.7	7.7	35.2	4.4	2.2	5.9	42.7	4.9	2.2	6.5	42.1	4.7	2.1	5.8	46.3	3.4	2.2	4.3
network1	371.6	127.1	56.4	151.0	384.4	118.4	53.4	140.0	380.3	175.9	77.8	209.4	391.0	171.7	76.4	206.3	344.5	208.2	101.3	250.1	236.0	455.0	291.5	514.2
german	472.9	8.8	4.2	10.9	495.9	8.3	3.9	10.3	520.8	6.4	3.7	7.9	539.8	5.5	3.1	6.8	561.1	4.7	3.4	5.5	598.8	3.8	2.9	4.4
move	2694.9	6.0	3.3	6.8	2791.5	5.7	3.5	6.5	2889.1	5.0	3.1	5.6	3200.9	4.2	2.9	4.7	3389.7	3.7	2.4	4.3	3629.9	2.8	2.2	3.1
sonar	17.4	28.0	20.6	31.0	18.1	27.6	22.6	30.0	19.9	24.2	20.3	26.0	21.2	21.2	16.5	23.1	22.6	18.4	14.7	20.5	25.2	14.0	11.0	16.4
liver	36.2	26.6	23.4	28.4	37.8	20.5	17.9	21.9	28.8	29.2	28.8	29.4	27.8	30.6	30.3	30.7	17.9	53.6	58.0	51.0	10.6	77.1	83.4	72.4
blackjack	43.0	970.0	895.2	998.7	45.2	1008.7	939.9	1035.1	39.9	954.6	876.4	984.7	37.9	898.0	824.5	926.5	37.7	985.4	902.4	1017.8	31.3	1205.8	1123.6	1238.3
labor	14.5	11.8	11.3	12.0	16.3	12.2	10.3	12.9	16.1	9.2	6.0	9.8	17.6	6.8	6.0	7.1	18.8	9.2	8.7	9.4	21.6	7.0	7.2	7.0
market2	3279.6	30.7	30.1	31.1	3296.2	34.6	34.2	34.9	3252.5	40.1	37.6	42.2	3311.0	59.3	54.6	63.6	3285.8	54.6	52.8	56.1	3371.4	69.0	69.5	68.6
Averages	**706.8**	**122.3**	**52.7**	**129.3**	**794.0**	**81.3**	**49.0**	**86.8**	**860.7**	**76.2**	**48.9**	**82.8**	**979.3**	**70.3**	**48.4**	**76.9**	**1070.1**	**72.6**	**53.1**	**79.2**	**1179.7**	**89.4**	**71.2**	**96.1**

Table E3.2: Effect of Class Noise (with Pruning)

Dataset	1% Noise				5% Noise				10% Noise				20% Noise				30% Noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	28.1	407.1	112.4	409.1	27.6	399.7	130.4	401.5	29.5	383.7	134.3	385.5	29.2	347.5	94.5	350.1	34.7	330.8	54.6	335.3	54.6	281.8	32.2	293.3
hypothyroid	14.1	2219.1	373.2	2227.0	13.7	2152.1	330.6	2160.8	14.7	2062.6	315.7	2071.0	17.8	1872.4	164.9	1887.0	17.8	1700.8	154.1	1715.3	86.4	1192.1	30.6	1256.8
vote	9.4	171.8	77.9	177.0	7.8	171.7	77.4	176.9	11.0	163.6	68.3	168.9	10.4	149.7	61.7	154.6	7.6	145.0	59.6	150.7	13.4	125.4	44.1	132.5
splice-junction	53.8	276.4	82.0	284.7	58.0	269.0	75.7	277.1	61.0	262.7	68.8	271.7	77.8	243.5	50.7	253.7	93.7	232.9	48.3	243.6	232.3	132.3	13.3	149.3
ticket2	9.1	399.1	200.2	409.2	9.8	391.6	204.2	401.5	8.7	370.2	171.5	381.5	10.7	346.7	159.2	358.6	13.4	305.6	105.8	320.2	29.1	170.0	30.7	190.8
ticket1	5.0	340.5	78.1	344.8	5.8	332.2	54.8	338.8	6.1	323.5	55.8	329.4	6.7	303.5	88.7	309.5	8.4	295.5	77.8	304.1	24.0	152.6	19.9	170.5
ticket3	6.9	377.1	178.6	383.7	7.7	371.2	209.1	376.3	8.1	355.1	173.7	361.1	8.2	350.0	210.5	355.2	11.4	321.4	112.5	332.1	17.1	215.2	69.7	228.9
soybean-large	61.2	29.0	16.5	30.2	61.4	29.0	17.8	30.1	63.2	27.8	12.3	29.8	66.6	25.0	16.2	26.0	65.7	22.2	9.4	23.8	89.7	15.6	6.6	17.9
breast-wisc	14.4	233.0	72.4	241.2	11.8	242.8	108.7	250.7	11.7	232.3	108.4	239.8	14.1	216.4	59.8	224.5	17.2	204.0	46.9	216.6	18.8	188.0	61.5	199.2
hepatitis	9.3	64.7	48.5	68.3	8.8	53.8	39.8	57.6	8.1	60.5	35.0	66.6	9.0	55.6	39.0	60.4	11.3	39.9	26.9	43.4	14.6	20.0	9.9	23.8
horse-colic	6.5	96.5	74.4	100.4	6.4	96.6	79.2	99.6	9.5	84.6	62.3	88.9	8.3	82.1	66.5	85.0	9.8	66.5	42.8	71.5	11.4	47.5	37.5	50.0
crx	21.8	177.5	95.0	192.5	25.3	171.4	94.3	184.4	27.6	180.9	93.0	196.7	28.1	168.0	86.1	181.8	28.7	156.1	80.6	168.7	63.9	106.2	54.4	118.9
bridges	2.2	60.8	54.7	61.9	1.7	69.0	70.6	68.7	1.9	68.7	65.7	69.3	2.7	60.8	56.3	61.6	4.0	56.1	45.5	58.5	5.3	48.5	42.9	49.6
hungar-heart	9.7	98.1	79.7	103.0	9.7	95.6	80.3	99.6	8.7	94.1	74.0	99.2	8.0	96.6	85.5	99.4	9.7	70.9	53.6	76.0	11.8	71.6	55.3	76.0
market1	146.1	371.9	213.8	415.5	171.3	332.8	178.7	374.9	157.4	415.9	245.7	465.2	160.4	363.0	223.9	403.3	166.0	382.1	225.6	432.2	164.1	438.7	279.6	494.4
adult	404.6	2279.5	968.0	2492.6	376.2	2449.5	1091.9	2672.4	358.0	2395.2	943.6	2635.1	370.6	1785.8	751.0	1956.8	448.9	1755.4	625.1	1948.5	472.9	2164.2	1067.0	2363.5
weather	505.2	107.2	36.2	139.2	530.3	96.6	30.4	126.7	551.9	91.1	29.4	121.7	627.1	29.8	14.1	38.4	651.3	26.6	14.0	34.1	733.9	15.2	10.4	18.6
network2	148.3	956.3	561.5	1072.5	147.4	909.9	520.1	1019.2	153.5	856.9	471.0	971.4	159.4	696.5	379.6	794.2	154.7	582.8	306.6	672.1	161.9	460.5	251.0	538.8
promoters	16.3	13.2	8.9	14.6	15.7	12.9	9.1	14.3	15.7	13.7	9.6	15.3	17.2	13.2	6.7	15.4	15.1	13.5	8.5	15.1	19.9	9.1	7.9	9.7
network1	143.1	801.7	476.8	898.2	154.1	761.5	437.9	859.5	146.1	789.9	455.2	896.4	167.6	544.8	297.6	627.1	144.8	435.6	229.3	508.2	103.6	646.0	420.1	722.8
german	89.5	134.2	64.9	161.9	88.2	132.3	66.4	156.6	96.2	113.8	51.7	139.2	101.6	113.1	54.3	137.8	98.9	92.9	43.9	113.1	138.7	68.7	30.4	88.4
move	349.6	59.9	55.5	61.4	364.2	59.8	54.6	61.5	342.6	60.3	55.9	61.7	364.7	55.2	50.8	56.8	328.3	52.1	45.2	54.6	374.7	48.0	41.8	51.0
sonar	14.9	28.1	21.4	30.7	14.6	27.9	23.2	30.1	16.4	24.5	20.5	26.4	17.2	21.6	16.8	23.5	18.9	18.8	14.8	21.2	22.2	14.4	11.5	16.5
liver	23.6	33.3	29.9	35.1	25.9	23.5	21.1	24.6	18.4	34.4	33.9	34.7	16.6	40.8	38.7	42.1	11.7	60.7	63.0	59.4	5.6	86.2	88.9	84.2
blackjack	21.2	1763.6	1656.5	1804.7	23.1	1740.8	1627.9	1783.8	21.2	1761.6	1671.0	1796.6	21.0	1667.2	1587.7	1697.8	18.7	1958.6	1933.0	1968.4	16.0	1818.8	1727.8	1854.5
labor	4.1	20.1	17.8	20.8	3.6	19.6	18.7	19.8	3.5	19.7	19.1	19.8	3.3	20.8	19.4	21.4	3.9	19.4	22.9	18.0	3.8	18.3	24.5	16.8
market2	877.7	72.9	68.9	76.1	861.1	81.2	75.6	86.0	880.9	92.7	86.9	97.6	879.8	86.7	79.1	93.4	940.8	91.2	88.1	93.8	875.2	129.9	128.6	131.0
Averages	**111.0**	**429.4**	**212.0**	**453.9**	**112.3**	**425.7**	**212.2**	**450.1**	**112.3**	**420.0**	**204.9**	**445.9**	**118.7**	**361.3**	**176.3**	**382.1**	**123.5**	**349.5**	**168.1**	**370.3**	**139.4**	**321.7**	**170.3**	**346.2**

Table E3.3: Effect of Attribute Noise on Training Set (No Pruning)

Dataset	1% Noise				5% Noise				10% Noise				20% Noise				30% Noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	107.9	249.1	4.4	251.1	330.2	81.2	4.4	83.3	496.0	51.7	3.9	55.2	669.1	26.4	3.9	29.2	751.7	16.2	4.0	18.8	835.5	9.5	4.8	11.2
hypothyroid	52.3	2140.3	293.3	2149.7	119.8	1972.7	72.7	1994.6	174.0	1855.9	117.3	1878.7	244.1	1375.1	51.5	1403.8	319.0	1073.3	31.9	1105.8	406.0	635.6	30.5	662.2
vote	51.0	122.5	10.2	130.6	58.0	109.2	16.5	116.4	66.4	90.8	7.1	97.7	77.6	64.5	11.2	68.6	93.6	54.9	9.7	59.2	124.6	26.2	4.2	28.1
splice-junction	303.7	71.1	7.2	75.8	409.6	50.0	5.1	54.4	508.9	38.9	4.8	42.6	661.6	19.7	3.4	22.3	761.8	14.3	3.2	16.5	918.4	8.3	2.4	10.1
ticket2	37.7	196.6	28.6	208.9	49.2	160.5	14.0	171.5	53.0	179.6	32.1	192.3	59.0	124.1	36.2	132.6	61.6	97.8	45.4	102.8	65.0	51.8	30.6	55.1
ticket1	22.9	262.4	41.2	268.9	35.2	262.0	28.0	271.2	43.6	229.1	27.8	235.4	60.4	155.0	33.8	163.4	69.2	62.4	14.0	68.2	77.0	28.5	12.3	32.2
ticket3	30.0	235.6	40.3	244.0	37.5	211.4	31.6	220.2	47.6	180.5	33.3	191.0	50.4	161.7	38.1	170.8	57.3	99.4	27.8	105.9	59.1	65.0	18.8	70.3
soybean-large	203.3	20.0	5.4	21.4	308.5	18.1	4.2	19.7	428.9	14.7	3.7	16.1	593.1	9.6	3.5	11.0	677.3	6.6	2.6	7.7	763.1	3.5	1.8	4.7
breast-wisc	31.3	200.4	45.1	210.8	29.0	203.5	44.5	212.6	28.0	204.1	63.0	211.6	27.4	170.5	49.9	177.7	32.3	153.5	49.4	158.6	30.0	116.8	31.7	123.2
hepatitis	22.5	24.0	9.7	27.8	21.3	30.2	12.6	35.7	19.8	38.5	15.7	43.6	16.9	38.3	18.8	44.5	16.1	28.8	15.4	32.0	12.7	38.5	24.7	43.2
horse-colic	40.9	30.0	13.6	33.5	42.9	31.9	15.8	34.6	36.1	41.0	24.4	44.9	28.0	52.4	33.2	56.1	26.7	57.3	38.9	61.7	19.7	43.8	31.4	47.0
crx	236.2	15.5	6.6	17.6	242.8	10.9	4.7	12.4	254.7	13.9	4.9	16.1	326.4	9.2	4.5	10.7	329.4	7.5	4.2	8.6	391.6	4.9	2.8	5.7
bridges	33.7	14.3	4.8	16.0	37.8	11.2	2.9	13.3	38.0	12.4	4.3	14.1	37.8	11.1	3.9	12.5	38.1	11.0	5.2	12.6	41.1	5.7	5.0	5.9
hungar-heart	38.5	40.7	23.6	46.0	35.1	45.3	22.5	51.2	33.3	48.1	33.3	51.6	24.4	56.1	38.5	61.0	23.6	48.9	29.1	53.8	17.6	50.6	30.3	57.0
market1	719.9	42.1	16.6	49.9	719.9	42.1	16.6	49.9	719.9	42.1	16.6	49.9	719.9	42.1	16.6	49.9	719.9	42.1	16.6	49.9	719.9	42.1	16.6	49.9
adult	9234.9	146.2	22.7	170.9	11828.8	88.4	16.5	103.4	14417.9	56.2	14.2	65.5	17445.5	42.8	7.6	51.5	20536.2	23.7	5.5	28.5	23360.8	12.2	3.1	14.9
weather	807.8	26.2	12.3	32.8	748.2	46.6	20.3	59.0	681.8	65.5	27.8	84.1	554.6	88.4	46.4	107.4	496.6	82.7	50.3	97.5	351.1	111.8	84.5	123.8
network2	421.0	193.8	101.3	222.8	479.5	177.9	96.0	202.7	354.9	340.3	177.9	391.1	250.8	338.1	198.1	378.7	225.4	337.5	210.9	375.6	166.9	340.7	223.1	379.6
promoters	30.1	7.8	4.7	8.6	33.4	6.4	3.1	7.7	36.1	6.0	2.9	7.2	34.9	5.8	3.2	6.9	40.0	5.0	2.0	6.5	44.5	4.1	3.4	4.5
network1	382.4	157.5	72.1	184.9	402.7	191.5	101.6	219.9	310.6	402.7	214.7	461.8	233.4	263.0	160.6	296.5	196.3	322.6	185.2	371.9	132.9	317.0	229.1	347.8
german	467.6	10.9	5.1	13.8	444.0	11.3	5.0	14.4	448.9	9.2	5.4	11.0	429.1	9.7	5.1	11.9	440.9	7.2	3.8	8.9	462.3	5.4	3.8	6.2
move	2745.9	5.8	3.4	6.6	3022.6	5.1	3.1	5.9	3276.5	4.5	2.9	5.2	3456.0	3.7	2.6	4.2	3525.8	3.2	2.3	3.7	3538.1	2.3	1.9	2.6
sonar	17.5	29.1	22.6	31.5	18.4	25.0	19.0	28.0	19.3	23.7	15.9	27.2	19.9	22.3	17.4	24.3	21.1	24.0	18.1	26.6	22.8	17.5	15.6	18.7
liver	35.6	20.9	18.9	22.1	32.8	24.3	24.5	24.2	22.0	38.7	38.1	39.1	15.1	86.8	85.3	87.8	8.5	98.5	101.8	96.5	1.7	150.5	150.5	150.5
blackjack	46.2	1009.7	927.4	1041.5	35.0	1469.0	1389.6	1499.5	32.0	1145.3	1114.7	1157.3	25.6	1314.5	1287.3	1325.1	21.1	1462.7	1466.6	1461.2	13.8	1663.6	1740.9	1630.4
labor	14.3	12.5	11.9	12.7	16.1	13.6	13.4	13.6	15.0	11.4	11.4	11.4	15.9	8.6	6.0	9.2	15.6	9.6	6.1	10.8	18.4	6.1	4.4	6.7
market2	2830.7	127.1	126.1	128.0	2095.3	187.0	178.2	194.0	2002.0	171.6	168.6	173.9	2340.9	162.1	154.9	167.6	2692.6	181.7	168.4	192.0	3090.2	150.1	142.1	156.7
Averages	**702.4**	**200.4**	**69.6**	**208.5**	**801.2**	**203.2**	**80.2**	**211.6**	**909.8**	**196.9**	**81.0**	**206.5**	**1052.5**	**172.7**	**86.0**	**180.9**	**1192.5**	**160.5**	**93.3**	**168.2**	**1321.7**	**144.9**	**105.6**	**149.9**

Table E3.4: Effect of Attribute Noise on Training Set (with Pruning)

Dataset	1% Noise				5% Noise				10% Noise				20% Noise				30% Noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	32.2	399.3	89.5	401.4	48.0	375.1	93.6	378.2	73.4	314.0	122.5	320.2	127.8	224.2	90.1	231.4	232.2	117.2	14.3	129.2	388.8	42.9	11.8	52.4
hypothyroid	15.5	2219.4	328.0	2229.0	19.5	2139.2	311.8	2151.8	29.2	2042.2	196.8	2063.5	34.1	1848.9	265.1	1873.6	45.0	1634.9	211.5	1669.7	37.9	1251.0	301.8	1285.1
vote	8.8	172.1	69.7	178.7	10.0	167.9	81.5	172.8	12.8	158.8	88.9	162.1	18.2	139.3	46.6	145.9	23.8	113.2	42.2	118.4	36.8	70.8	19.2	74.5
splice-junction	63.7	275.1	70.7	284.5	70.9	275.7	59.9	286.4	93.1	277.2	56.8	290.2	153.1	250.7	36.9	270.2	177.4	206.7	25.0	226.6	212.2	224.7	11.4	260.2
ticket2	11.6	393.5	199.1	405.0	15.2	350.8	126.9	365.4	15.6	340.0	158.1	353.7	12.8	329.4	230.8	338.8	17.9	292.1	228.5	298.0	10.0	365.0	335.2	368.3
ticket1	6.8	337.5	71.5	342.4	11.5	323.3	139.3	328.7	15.5	308.6	132.7	314.8	27.0	270.4	67.7	283.1	29.4	251.6	45.7	274.2	34.8	185.4	84.5	200.4
ticket3	9.9	374.4	168.5	381.7	11.0	364.9	135.0	374.8	15.0	347.7	107.7	362.4	15.5	340.8	127.0	357.4	19.1	327.2	112.4	348.9	14.6	295.9	124.6	314.3
soybean-large	66.8	29.0	14.7	30.5	84.9	25.5	12.0	27.1	106.1	22.1	12.3	23.3	137.8	17.4	9.1	19.1	161.2	13.2	7.5	14.7	210.4	8.3	5.0	10.4
breast-wisc	14.2	230.1	73.7	239.1	13.7	232.5	98.5	240.0	13.4	233.9	106.9	240.4	15.3	192.7	76.8	198.4	20.7	164.5	60.5	170.6	23.1	122.6	34.4	129.2
hepatitis	9.8	63.3	43.5	68.1	9.3	57.6	29.1	66.3	9.1	63.9	38.7	70.0	10.0	49.7	28.7	55.4	10.4	38.9	21.6	43.9	8.2	49.6	37.2	52.9
horse-colic	7.1	95.2	74.2	99.0	8.2	85.9	73.7	88.0	12.1	76.4	50.1	81.7	10.7	74.1	50.9	78.6	14.9	65.4	44.7	70.5	13.9	49.1	38.1	51.7
crx	20.3	180.8	92.8	197.3	33.1	183.9	98.7	198.8	24.1	165.3	103.5	175.7	26.2	158.4	82.2	172.3	33.9	134.1	68.7	146.9	49.9	105.4	52.4	118.8
bridges	2.6	60.6	60.2	60.7	2.6	64.3	67.4	63.7	1.0	78.3	78.5	78.2	1.9	68.3	69.1	68.2	1.2	74.1	72.6	74.4	1.0	78.3	78.5	78.2
hungar-heart	10.9	91.7	64.0	99.3	14.2	76.7	51.8	82.8	11.8	77.4	58.7	81.8	11.9	77.7	55.7	83.8	12.4	69.1	44.7	75.5	11.7	70.0	45.2	77.0
market1	135.0	382.3	238.8	420.4	135.0	382.3	238.8	420.4	135.0	382.3	238.8	420.4	135.0	382.3	238.8	420.4	135.0	382.3	238.8	420.4	135.0	382.3	238.8	420.4
adult	368.1	1887.9	913.9	2049.1	300.5	1806.6	906.6	1960.9	253.8	1893.2	1087.0	2036.2	281.4	2294.6	1524.1	2444.3	196.9	2098.4	1308.7	2264.8	77.4	6376.2	6319.5	6388.5
weather	501.0	114.0	38.5	147.6	483.7	115.3	42.8	147.9	477.0	107.6	44.3	136.7	417.8	102.9	53.3	125.1	385.8	92.8	54.3	110.1	273.5	138.2	110.4	150.4
network2	169.7	920.7	545.0	1028.0	163.4	720.8	422.9	802.7	182.2	526.8	283.5	599.8	178.2	443.9	267.5	495.7	168.6	372.2	226.9	416.4	129.7	394.5	252.4	441.6
promoters	16.9	13.3	11.4	13.7	15.1	12.1	8.1	13.6	15.7	14.6	12.1	15.2	16.6	12.5	10.8	13.2	19.0	10.1	6.9	11.9	22.6	10.1	8.7	10.9
network1	153.0	744.9	454.4	829.4	163.2	542.5	321.0	605.5	173.1	558.8	303.4	640.7	156.0	392.0	236.0	443.5	138.8	397.7	231.0	458.3	98.6	388.5	274.2	428.1
german	95.4	130.4	62.6	159.2	86.0	124.5	61.3	148.4	76.5	130.6	70.3	153.1	89.9	115.5	54.5	140.6	86.2	115.7	58.7	138.5	54.0	206.4	164.5	225.5
move	366.4	62.8	61.8	63.1	339.8	60.9	56.4	62.4	319.6	65.6	61.6	67.1	301.8	85.6	85.2	85.7	302.7	70.1	65.5	72.4	252.4	133.5	130.5	135.2
sonar	15.2	29.1	23.0	31.3	15.6	25.1	19.6	27.7	16.1	23.8	16.6	26.8	16.8	22.4	17.4	24.4	18.2	24.0	17.7	27.0	19.5	17.7	16.0	18.8
liver	25.4	25.5	22.7	26.9	22.7	34.7	33.2	35.4	14.8	55.0	52.7	56.4	10.9	89.4	87.9	90.4	7.0	100.3	103.5	98.5	1.7	150.5	150.5	150.5
blackjack	25.8	1880.6	1746.4	1932.2	20.3	1923.3	1823.8	1961.4	20.1	1535.5	1500.5	1549.2	19.2	1754.9	1749.9	1756.8	15.2	1954.9	1932.5	1963.9	7.7	3262.8	3321.2	3237.5
labor	4.4	19.0	17.7	19.3	3.4	20.8	19.8	21.1	4.7	18.6	22.2	18.1	3.9	16.8	13.1	17.7	3.7	20.5	18.8	21.0	3.2	21.6	19.3	22.5
market2	716.5	171.6	170.8	172.2	492.0	346.9	337.5	353.9	362.8	498.2	475.5	515.5	260.8	649.4	621.4	670.1	154.7	1072.7	1035.2	1099.5	56.3	1384.6	1338.3	1419.0
Average	**106.4**	**418.7**	**212.1**	**441.1**	**96.0**	**401.4**	**210.0**	**421.7**	**92.0**	**382.1**	**203.0**	**402.0**	**92.2**	**385.3**	**229.5**	**403.9**	**90.0**	**378.3**	**233.3**	**398.7**	**80.9**	**584.7**	**500.9**	**597.1**

Table E3.5: Effect of Attribute Noise on Training and Test Set (no Pruning)

Dataset	1% Noise				5% Noise				10%Noise				20% Noise				30% Noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	125.2	175.7	75.0	178.0	349.4	66.2	21.8	71.1	510.4	31.6	11.8	35.8	681.7	16.9	8.0	20.4	747.1	11.8	7.1	14.4	843.7	6.9	5.8	7.7
hypothyroid	69.6	1916.0	143.4	1936.9	108.7	1838.7	88.8	1875.6	167.8	1649.0	48.6	1712.0	242.5	1154.2	18.2	1232.4	337.3	605.7	21.3	657.3	426.2	344.6	13.1	394.2
vote	47.6	120.6	11.7	127.3	62.0	79.8	16.3	85.7	54.2	85.0	8.1	90.7	74.2	40.2	11.6	43.9	79.4	36.4	12.8	39.6	128.8	12.4	4.8	15.0
splice-junction	301.3	66.0	10.3	69.9	417.1	53.5	5.8	59.2	538.0	26.7	8.1	30.1	671.2	17.3	5.3	20.3	764.5	12.5	5.2	14.7	934.0	8.4	4.6	10.2
ticket2	34.2	216.5	31.3	232.5	55.3	132.6	26.3	145.0	51.1	102.6	20.9	112.9	61.1	54.1	17.1	61.5	68.5	45.0	15.7	52.5	60.9	34.1	16.2	38.2
ticket1	21.8	272.8	33.0	281.7	39.3	187.6	23.8	203.4	48.6	157.2	21.2	176.5	65.8	78.5	12.2	96.2	64.1	58.2	12.3	68.0	68.2	24.4	10.0	29.8
ticket3	26.1	219.5	37.0	228.1	41.2	199.7	40.8	216.8	50.1	166.1	9.1	187.9	48.0	80.3	21.5	88.4	60.8	41.0	16.5	47.4	63.3	31.8	17.4	35.0
soybean-large	207.0	19.9	4.7	21.8	329.5	18.8	5.8	22.7	436.1	11.4	3.5	15.0	557.6	5.1	2.2	8.0	706.0	2.9	1.5	5.0	758.2	1.5	1.1	2.7
breast-wisc	32.4	190.3	40.5	199.8	32.2	173.1	30.6	186.4	25.7	183.0	50.5	192.1	24.6	150.6	55.6	157.6	29.8	105.5	36.1	112.4	27.8	95.8	40.6	102.1
hepatitis	22.1	27.0	12.2	30.3	21.6	36.0	21.4	39.5	19.4	34.9	12.5	39.9	14.6	34.6	17.5	41.9	15.2	35.1	16.6	39.7	13.6	41.4	22.5	47.3
horse-colic	42.9	29.1	14.5	32.0	41.3	36.8	16.8	41.7	37.9	37.0	25.5	41.0	28.3	41.6	33.1	44.4	30.9	30.7	22.7	33.8	27.8	26.8	22.0	28.9
crx	223.5	16.5	6.6	19.0	219.7	19.9	5.7	23.6	257.7	10.6	5.3	12.2	299.9	11.9	6.0	13.9	378.2	6.1	4.4	6.9	417.3	3.6	2.8	4.1
bridges	34.5	10.3	5.8	11.5	31.3	11.5	4.1	13.8	35.3	9.0	5.1	9.8	40.2	8.5	3.0	10.1	34.5	7.4	5.9	8.0	41.1	5.1	4.0	5.6
hungar-heart	36.5	39.1	18.2	44.7	39.6	34.9	22.0	38.6	36.2	45.9	24.4	53.1	33.3	34.3	19.7	38.6	22.8	77.0	46.2	86.2	19.4	40.5	39.5	40.9
market1	738.9	51.1	20.0	60.7	731.4	42.6	17.1	51.6	800.0	33.8	16.1	40.4	877.3	20.7	11.3	24.9	1098.3	12.4	7.1	15.0	1225.1	6.9	5.3	7.8
adult	9182.0	151.7	22.9	177.7	11628.6	118.3	14.1	142.6	14273.9	40.7	7.4	49.4	17644.3	23.0	3.6	29.2	20629.5	10.6	2.9	13.3	23105.3	4.4	1.9	5.4
weather	808.1	25.1	11.6	31.6	729.6	51.8	22.8	65.5	677.0	48.3	25.3	59.4	562.8	69.3	37.0	86.2	469.6	97.1	66.7	112.7	358.6	118.6	98.7	129.7
network2	415.5	183.1	83.8	215.3	420.3	173.9	82.8	205.7	327.0	308.3	153.7	364.8	253.8	332.0	181.9	387.9	197.2	248.6	165.3	280.4	142.0	321.4	215.3	363.2
promoters	32.8	6.5	4.5	7.4	28.0	7.1	4.7	7.6	34.0	7.7	3.8	9.3	39.7	5.6	2.6	6.9	42.7	3.9	2.9	4.9	44.2	3.0	2.7	3.1
network1	393.2	112.6	58.8	131.2	431.7	178.0	87.4	211.8	333.5	201.0	95.0	241.7	222.8	329.9	185.8	388.1	211.3	178.6	109.0	206.7	165.0	155.9	120.3	170.9
german	451.7	12.7	5.7	16.1	451.7	9.4	4.6	11.8	445.6	7.3	3.6	9.4	439.5	5.7	4.2	6.7	427.5	5.9	4.4	6.9	455.1	5.2	3.6	6.3
move	2752.1	5.9	3.3	6.8	3048.5	4.9	2.9	5.6	3253.2	4.0	2.5	4.8	3477.8	2.9	2.1	3.4	3493.9	2.2	1.8	2.6	3466.9	1.8	1.6	1.9
sonar	17.6	27.8	20.7	31.4	18.2	23.8	16.6	26.4	19.0	23.1	20.3	24.4	20.4	20.4	18.0	21.6	21.7	22.4	17.1	25.1	20.3	16.6	17.1	16.3
liver	30.9	23.7	23.3	23.9	25.7	36.1	34.4	37.3	23.9	41.4	38.5	43.4	13.7	54.0	53.7	54.2	9.4	79.0	80.2	78.3	4.9	127.9	126.1	129.1
blackjack	44.3	880.4	810.1	907.8	39.1	1244.8	1113.3	1298.2	29.9	1316.7	1256.9	1342.1	35.2	940.3	904.6	956.3	25.2	1196.8	1178.8	1205.5	16.8	1994.4	1978.4	2002.6
labor	13.7	11.2	10.6	11.4	15.8	10.3	8.0	11.3	13.7	10.9	7.5	11.8	17.1	7.4	4.0	8.6	13.7	12.6	7.3	13.3	19.0	6.0	2.8	7.3
market2	2810.6	99.4	94.4	103.6	2028.3	155.6	146.5	162.9	2271.3	114.6	108.6	119.5	2428.6	136.8	125.7	146.2	2614.5	151.6	148.7	153.9	3164.7	127.0	128.0	126.2
Average	**700.6**	**181.9**	**59.8**	**190.3**	**792.0**	**183.2**	**69.8**	**194.9**	**917.4**	**174.4**	**73.8**	**186.3**	**1069.5**	**136.2**	**65.4**	**148.1**	**1207.2**	**114.7**	**74.7**	**122.4**	**1334.0**	**132.1**	**107.6**	**138.2**

Table E3.6: Effect of Attribute Noise on Training and Test Set (with Pruning)

Dataset	1% Noise				5% Noise				10% Noise				20% Noise				30% noise				50% Noise			
	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr	Lvs	All	Err	Corr
kr-vs-kp	32.8	408.4	258.5	411.5	55.9	352.3	200.6	364.2	61.4	349.6	273.3	359.4	136.5	219.0	149.4	237.1	221.4	107.9	67.0	125.3	391.4	24.9	18.1	29.6
hypothyroid	16.0	2189.4	228.1	2208.3	23.7	2062.7	168.9	2094.4	21.3	1891.2	241.1	1937.6	33.9	1650.9	285.2	1714.2	40.0	1413.2	474.6	1467.2	42.3	1252.5	629.6	1305.6
vote	9.2	170.3	75.2	175.9	9.0	161.6	122.8	164.1	12.8	150.9	97.7	154.4	14.2	129.7	76.6	135.7	18.6	98.8	64.7	103.0	39.2	39.1	22.9	44.1
splice-junction	55.6	279.4	98.1	288.4	85.3	285.1	98.7	301.5	89.5	294.7	156.6	309.1	124.3	267.2	136.7	289.0	159.1	280.2	163.5	305.5	203.5	277.1	184.0	304.9
ticket2	14.4	367.4	147.9	385.3	17.1	311.7	115.1	334.6	20.4	287.8	119.7	305.8	13.1	307.0	154.7	327.1	3.2	408.5	354.0	415.7	17.7	191.3	122.9	201.8
ticket1	5.7	336.9	51.8	345.4	12.8	305.7	79.6	324.7	16.6	279.3	91.1	300.8	26.4	209.9	50.0	239.9	28.1	191.3	53.2	216.2	37.1	118.9	28.2	148.0
ticket3	8.5	370.1	203.2	375.7	14.7	350.2	77.3	376.6	17.4	305.8	72.9	328.3	17.5	309.4	116.6	331.0	18.5	284.3	117.1	309.6	14.2	239.0	165.6	249.2
soybean-large	69.9	28.0	16.7	29.3	92.9	22.5	9.3	25.9	87.7	22.0	13.5	25.3	138.7	13.1	7.3	17.7	164.8	9.2	5.2	14.0	201.0	4.4	3.3	7.8
breast-wisc	12.3	233.5	111.1	241.3	13.0	228.2	99.0	237.9	15.6	193.0	74.3	200.2	16.7	165.6	76.0	172.9	19.4	135.5	55.7	143.9	21.6	102.4	53.2	108.2
hepatitis	8.9	58.2	33.8	63.6	8.7	61.1	41.3	66.5	8.2	56.8	40.0	59.6	8.4	62.7	35.6	73.1	9.8	45.8	30.6	50.1	10.1	51.9	32.2	57.9
horse-colic	6.5	96.3	74.6	100.0	7.8	85.8	55.0	92.3	13.1	77.7	58.7	83.1	12.3	68.6	48.0	75.4	15.3	45.8	34.3	49.8	19.6	37.1	29.6	40.4
crx	33.9	187.8	97.6	204.0	27.3	185.2	105.3	199.8	24.4	170.0	105.8	185.3	23.9	154.8	104.8	168.2	35.5	111.1	79.4	124.2	38.6	112.1	92.8	121.7
bridges	1.0	78.3	78.5	78.2	1.0	78.3	78.5	78.2	1.6	74.5	67.8	75.7	1.0	78.3	78.5	78.2	1.0	78.3	78.5	78.2	1.0	78.3	78.5	78.2
hungar-heart	10.9	83.4	54.0	92.1	11.5	76.6	54.5	84.0	13.2	65.8	46.9	72.3	15.0	48.1	34.9	52.2	10.6	82.4	52.5	91.9	14.2	44.0	41.2	45.2
market1	166.6	283.0	158.7	317.6	133.2	368.4	220.7	415.1	174.2	387.0	220.3	441.4	143.6	426.1	284.9	476.6	169.1	330.7	216.4	375.1	185.8	197.8	130.2	230.3
adult	424.3	1894.4	958.8	2052.7	294.2	1831.0	995.6	1991.0	231.1	1718.1	971.2	1873.3	273.7	1570.6	789.8	1759.1	257.5	1386.2	648.8	1570.3	69.3	7528.9	6879.9	7717.8
weather	500.2	144.7	46.0	189.1	479.9	127.1	47.0	163.2	460.4	92.9	41.2	117.3	420.6	100.0	49.5	125.8	359.3	120.6	84.7	139.2	277.4	129.3	107.1	141.5
network2	159.9	825.9	494.4	920.3	174.6	587.7	310.7	676.3	180.4	482.1	253.4	561.0	181.0	351.0	196.6	407.0	145.4	379.1	249.8	426.4	113.5	334.2	226.9	376.1
promoters	15.1	12.1	8.4	13.5	16.9	14.5	9.3	15.6	19.0	12.6	6.1	15.1	15.4	14.6	10.4	15.9	19.3	9.2	7.9	10.6	19.9	8.2	8.3	8.1
network1	157.2	663.4	375.4	752.9	178.9	406.6	206.9	473.5	170.8	290.9	136.8	348.6	144.2	409.1	233.2	478.0	152.4	195.0	122.5	223.7	126.3	187.7	141.1	207.5
german	87.4	132.8	67.5	156.6	79.5	129.9	67.6	156.4	86.3	104.1	54.7	124.2	88.2	108.9	59.3	129.4	95.3	98.9	55.2	119.5	74.8	199.1	154.6	223.5
move	344.3	58.4	49.2	61.5	331.2	62.9	55.5	65.4	361.0	49.3	42.6	52.5	299.8	67.6	59.8	72.3	294.5	83.5	76.1	88.8	237.0	120.2	111.0	127.3
sonar	14.8	28.0	21.4	31.2	15.0	24.2	17.6	26.6	16.1	23.1	20.6	24.2	16.5	21.7	20.8	22.1	17.9	22.6	16.4	25.9	16.8	17.8	19.0	17.0
liver	19.3	35.4	34.2	36.0	16.3	48.4	46.1	49.9	18.3	45.1	42.1	47.2	9.5	57.8	57.6	57.8	7.3	80.2	81.0	79.7	3.8	132.8	129.7	135.0
blackjack	22.9	1834.1	1701.5	1885.8	22.0	2201.2	2022.5	2273.9	17.9	1770.3	1682.3	1807.5	21.6	1076.3	1050.9	1087.6	17.9	1468.3	1415.2	1493.7	8.6	4062.8	3927.7	4132.4
labor	4.3	18.0	20.1	17.4	3.7	21.6	21.0	21.8	3.7	17.7	13.3	19.3	4.6	17.8	16.8	18.3	4.9	15.1	9.6	16.3	2.2	27.5	26.2	27.9
market2	658.2	188.8	177.5	197.8	399.1	371.8	347.2	390.9	439.1	340.6	316.1	359.7	255.4	578.8	538.9	610.6	129.4	1026.9	984.5	1060.8	63.7	1142.3	1131.9	1150.8
Average	**105.9**	**407.6**	**209.0**	**430.8**	**93.5**	**398.6**	**210.1**	**424.6**	**95.6**	**353.8**	**194.8**	**377.3**	**91.0**	**314.2**	**174.9**	**339.7**	**89.5**	**315.1**	**207.3**	**337.9**	**83.4**	**617.1**	**536.9**	**638.4**

Appendix F
Detailed Analysis of Selected Datasets

This appendix examines the following datasets in detail:

1. Vote dataset
2. Move dataset
3. Adult dataset

For each of these datasets, three sets of figures are presented. Each set of figures shows the distribution of errors by disjuncts size, but under different circumstances. These circumstances are:

- C4.5 without any pruning,
- C4.5's with its default pruning strategy, and
- C4.5 without any pruning but with varying training set sizes

The first two sets of figures each contain the following 6 figures, described below:

1. Distribution of Examples: a plot that shows the number of correctly and incorrectly classified examples by disjuncts size, grouped into "bins" to make the results more readable (the size of the bins can be determined by looking at the labels on the x-axis). This figure provides a higher level view of the information shown in the third and fourth figures. Figure 1 in the body of this paper is this type of figure.

2. Distribution of Disjuncts: a plot that shows the number of disjuncts of a given size. This plot also uses binning to make the results more readable.

3. Distribution of Correct Examples: Shows the distribution of the correct examples only, without any binning (a more detailed view of what is in the first figure).

4. Distribution of Errors: Same as above, but for errors.

5. Error Concentration Curve: Shows how the error concentration value is computed. Figure 2 in the body of this paper is this type of figure.

6. Cumulative Coverage Statistics: This plot shows the cumulative percentage of the total errors, cumulative percentage of the total examples (i.e., correctly and incorrectly classified examples), and the cumulative error rate. For this figure, cumulative means that at x coordinate n, the examples included in the calculation include all examples falling into disjuncts of size 0-n. For example, Figure F1.1.6 shows that for the vote dataset, the disjuncts of size 0-100 cover about 45% of the total examples, but cover over 90% of the total errors (and have an overall error rate of just under 20%).

The third set of figures show the distribution of examples when the training set is varied.

F1 The Vote Dataset

F1.1 Vote Dataset without Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
3	84.8	UCI	435	6.9	197	48	124 (10/133)

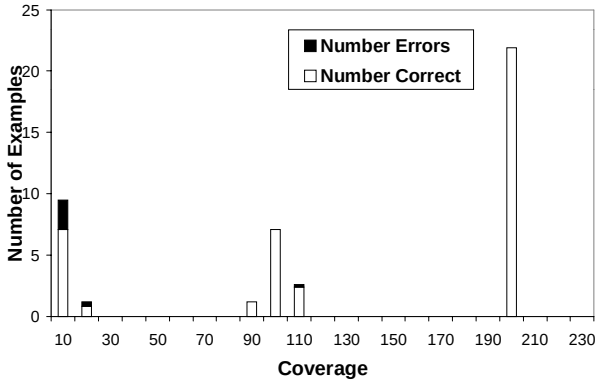


Figure F1.1.1: Distribution of Examples

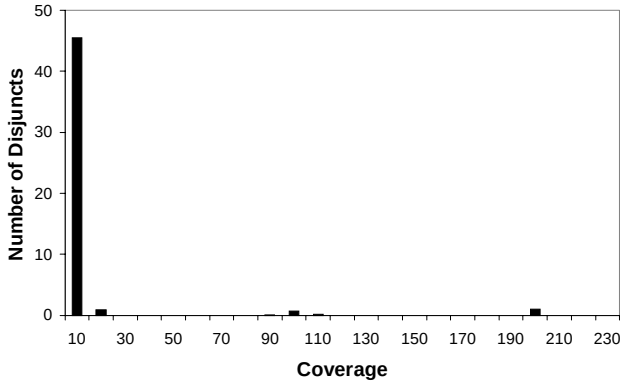


Figure F1.1.2: Distribution of Disjuncts

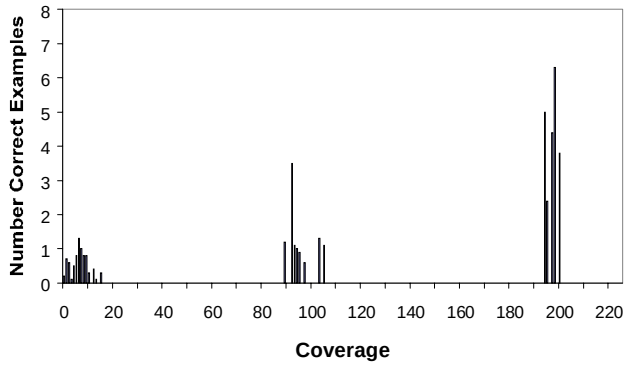


Figure F1.1.3: Distribution of Correct Examples

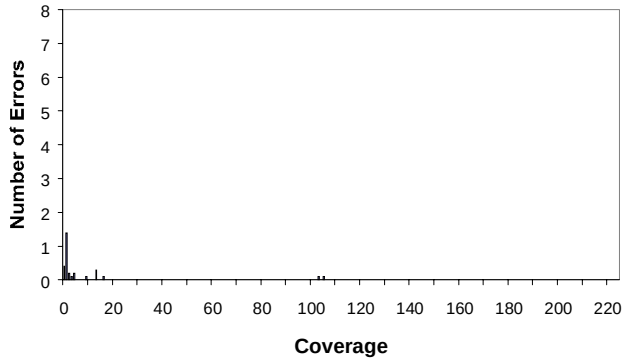


Figure F1.1.4: Distribution of Errors

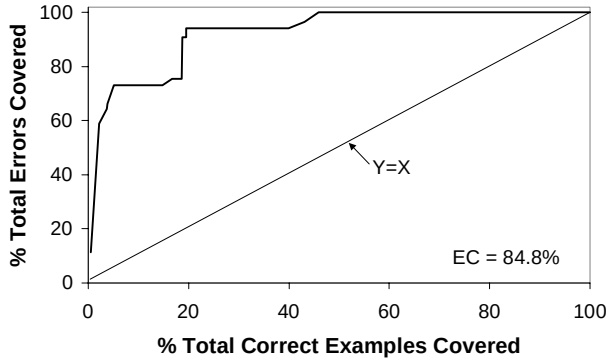


Figure F1.1.5: Error Concentration Curve

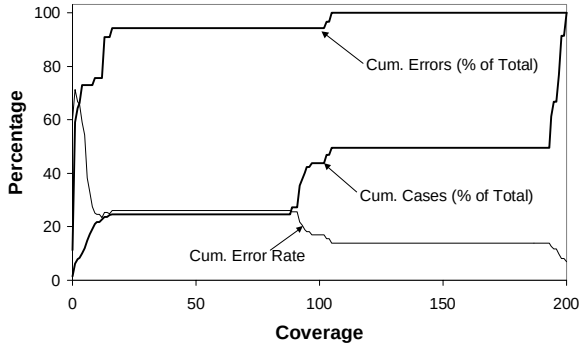


Figure F1.1.6: Cumulative Coverage Statistics

F1.2 Vote Dataset with Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
3	71.2	UCI	435	5.3	220	10	170 (66/176)

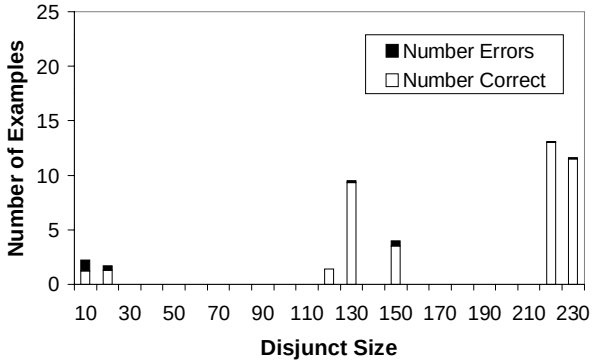


Figure F1.2.1: Distribution of Examples

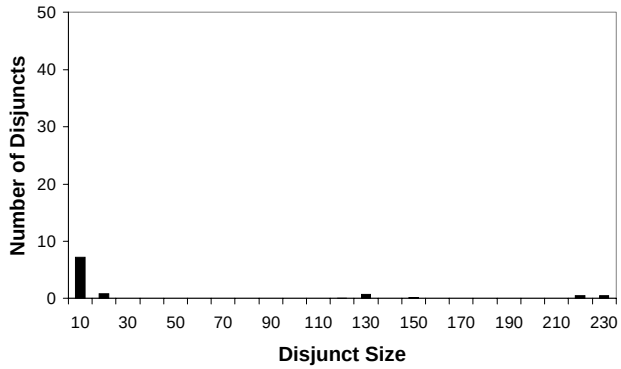


Figure F1.2.2: Distribution of Disjuncts

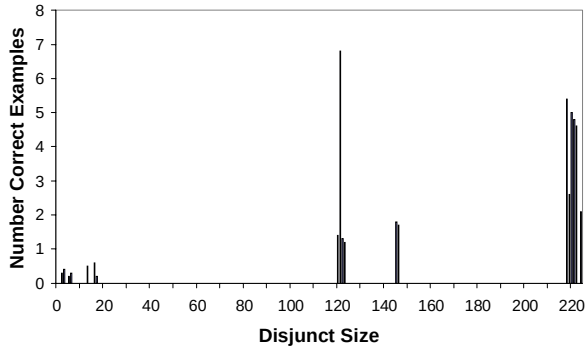


Figure F1.2.3: Distribution of Correct Examples

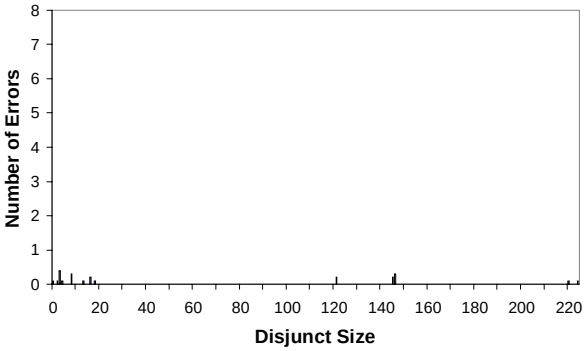


Figure 1.2.4: Distribution of Errors

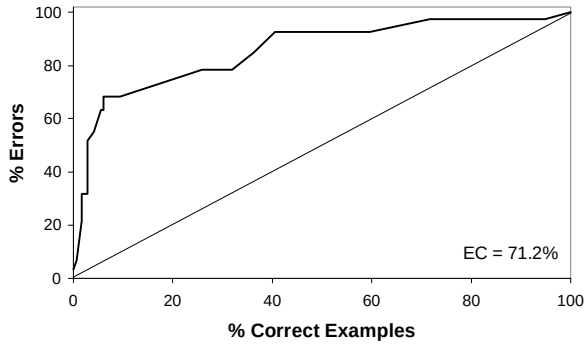
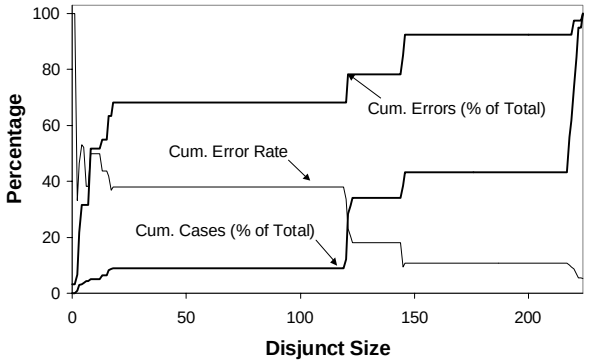


Figure F1.2.5: Error Concentration Curve



FigureF1.2.6: Cumulative Coverage Statistics

F1.3 Vote Dataset with Varying Training Set Size

Training Set Size	Error Rate	Number Leaves	Coverage Means	EC
10% (43)	8.0%	6.4	18.8 (8.9/19.8)	62.8%
50% (217)	6.7%	27.2	73.2 (13.8/77.5)	76.2%
90% (391)	6.9	48.4	124.4 (10.0/132.9)	84.8%

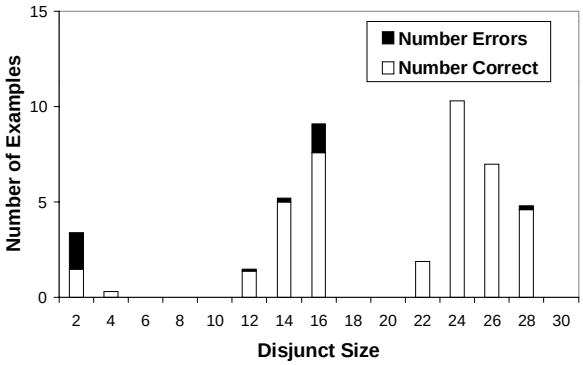


Figure F1.3.1:Distribution with 10% Training Data

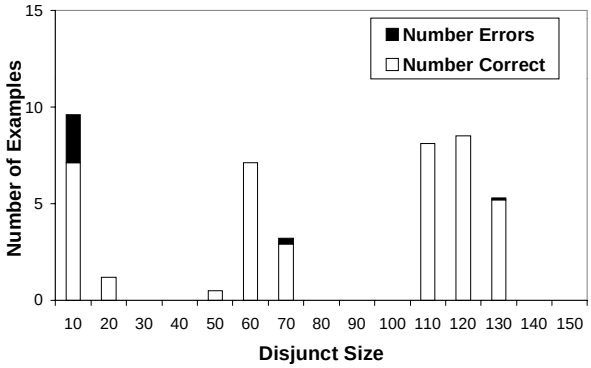


Figure F1.3.2: Distribution with 50% Training Data

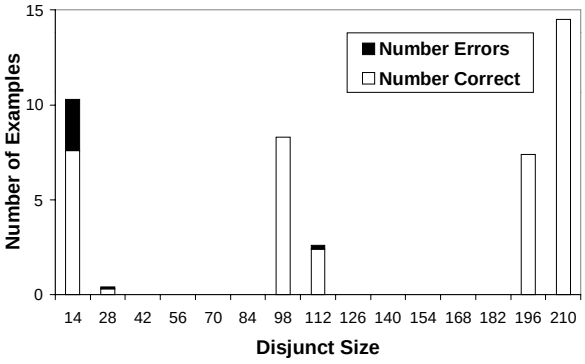


Figure F1.3.3: Distribution with 90% Training Data

Table F1.3.1: Number of Disjuncts in Learned Concept by Coverage Band

Training	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15
10%	4.1	.3	0	0	0	.1	.4	.5	0	0	.1	.4	.3	.2	0
50%	24.9	.3	0	0	.1	.7	.2	0	0	0	.4	.4	.2	0	0
90%	46.1	.3	0	0	0	0	.8	.2	0	0	0	0	0	.3	.7

The EC increases as the training set size increases, because as more data is available for training, the number of errors in the larger disjuncts is dramatically reduced, leaving more of the errors in the relatively smaller disjuncts.

F2. The Move Dataset

F2.1. Move Dataset without Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
24	28.4	ATT	3028	23.5	35	2678	6.2 (3.8/6.9)

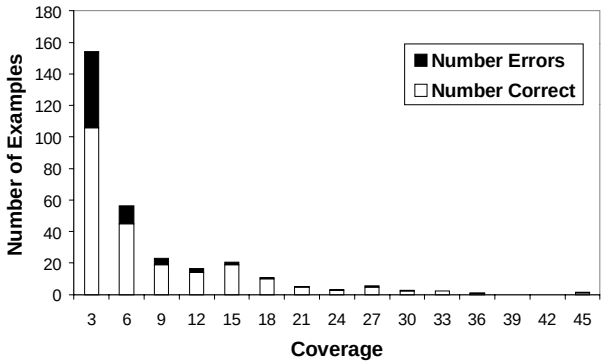


Figure F2.1.1: Distribution of Examples

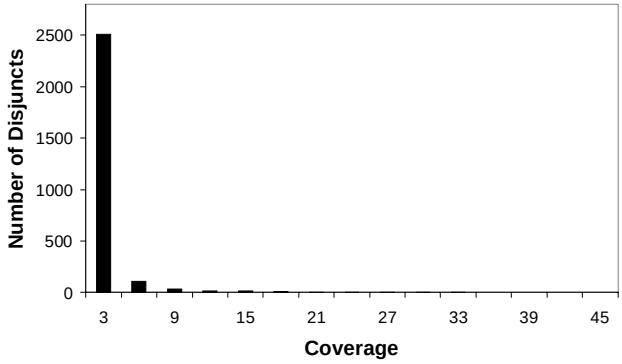


Figure F2.1.2: Distribution of Disjuncts

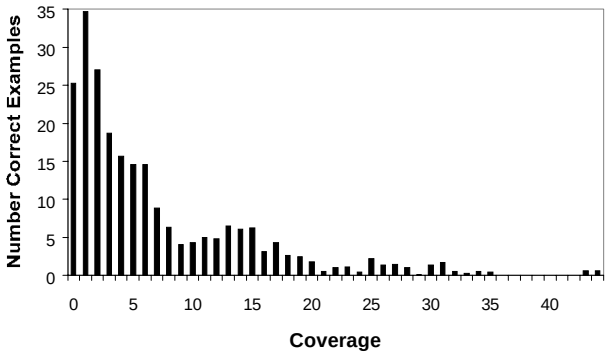


Figure F2.1.3: Distribution of Correct Examples

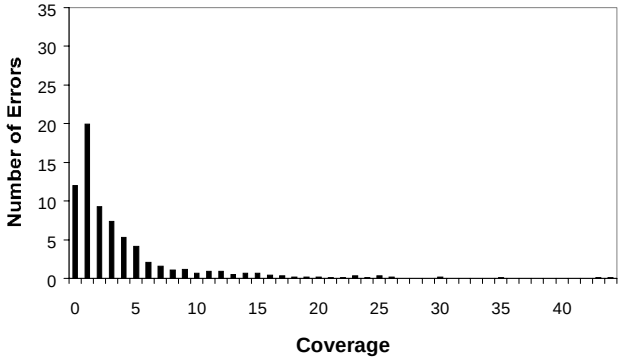


Figure F2.1.4: Distribution of Errors

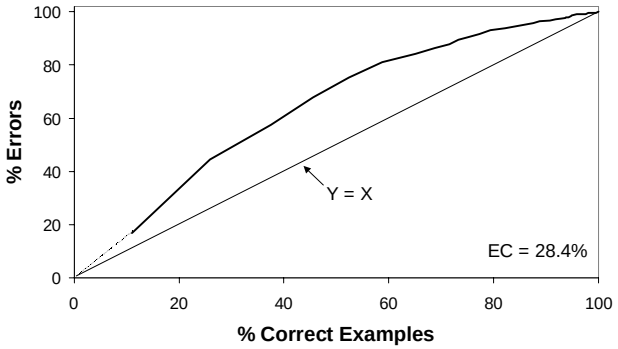


Figure F2.1.5: Error Concentration Curve

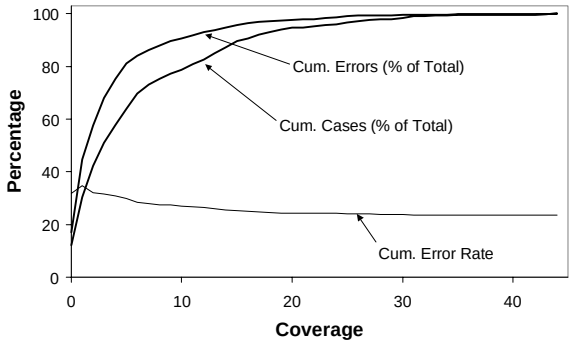


Figure 2.1.6: Cumulative Coverage Statistics

F2.2 Move Dataset with Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
24	9.4	ATT	3028	23.9	216	366	57.6 (53.1/59.0)

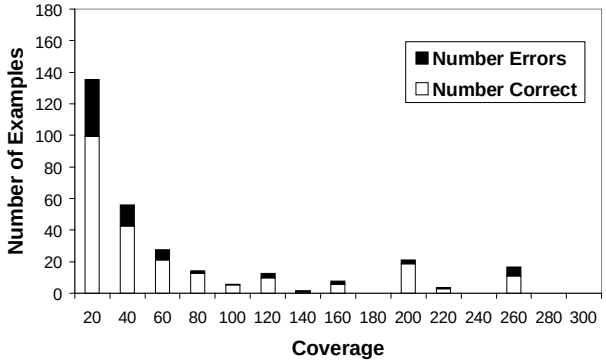


Figure F2.2.1: Distribution of Examples

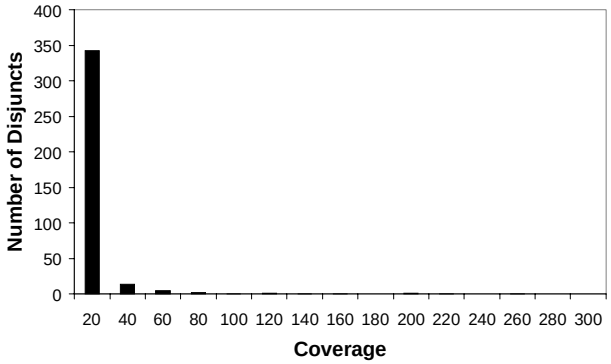


Figure F2.2.2: Distribution of Disjuncts

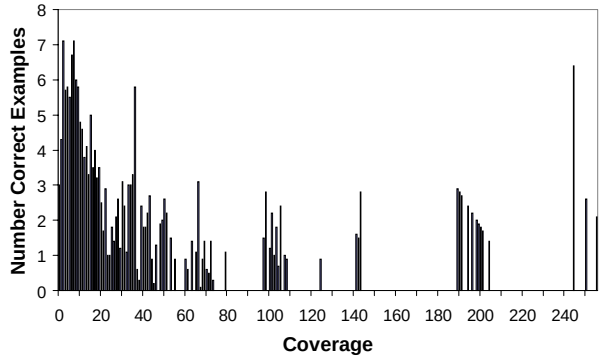


Figure F2.2.3: Distribution of Correct Examples

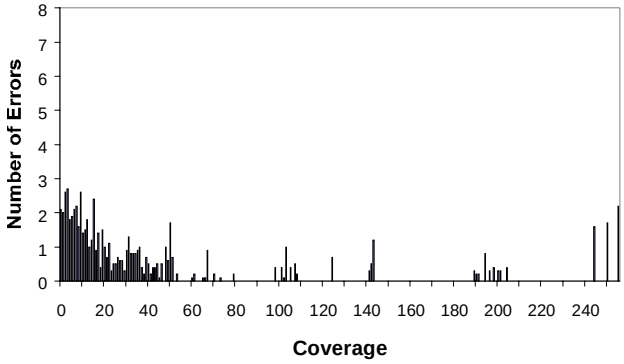


Figure F2.2.4: Distribution of Errors

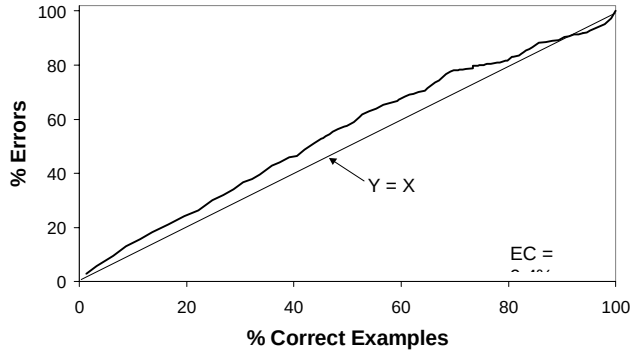


Figure F2.2.5: Error Concentration Curve

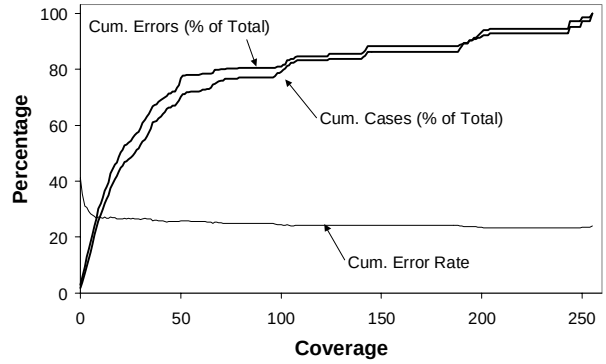


Figure F2.2.6: Cumulative Coverage Statistics

F2.3 Move Dataset with Varying Training Set Size

Training Set Size	Error Rate	Number Leaves	Coverage Means	EC
10% (303)	33.7%	388	3.2 (2.5/3.6)	15.8%
50% (1514)	26.0%	1601	4.9 (3.1/5.6)	26.8%
90% (2726)	23.5%	2687	6.2 (3.8/6.9)	28.4%

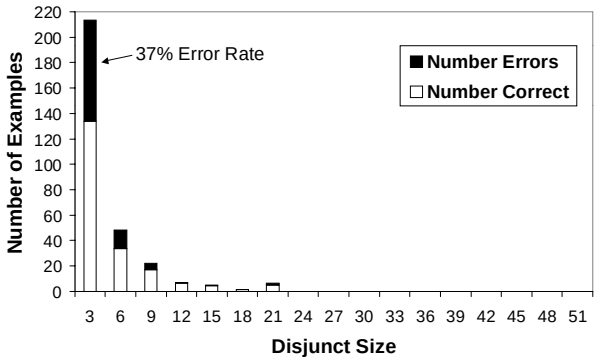


Figure F2.3.1: Distribution with 10% Training Data

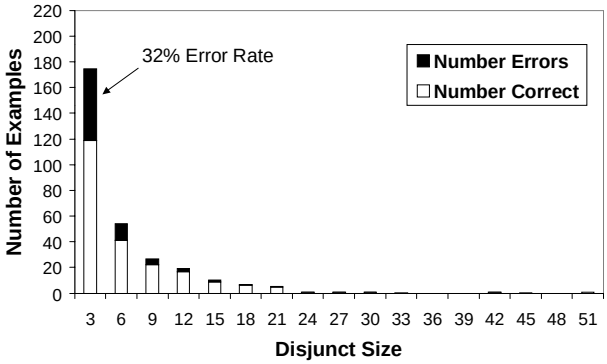


Figure F2.3.2: Distribution with 50% Training Data

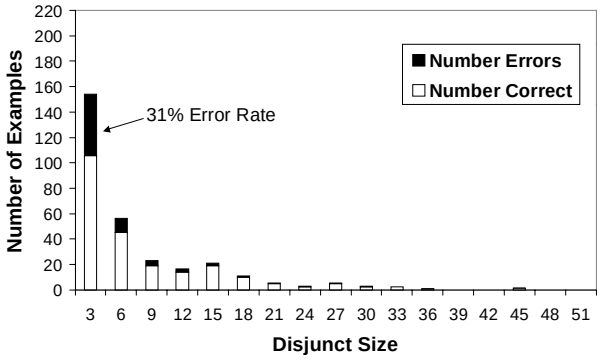


Figure F2.3.3: Distribution with 90% Training Data

Table F2.3.1: Number of Disjuncts in Learned Concept by Coverage Band

Training	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
10%	370.2	12.8	3.4	0.7	0.5	0.1	0.3	0	0	0	0	0	0	0	0	0	0
50%	1497.2	67.3	17.4	9.3	4.7	1.9	1.8	0.3	0.2	0.2	.1	0	0	.1	.2	0	.1
90%	2505.7	105.6	33.7	13.7	12.0	7.2	3.4	1.1	1.7	1	1	.5	0	0	.2	0	0

Note that the EC increases as the training set size increases. This is explained by the fact that most of the errors occur in the first bin (coverage 0-3) and, while the error rate of the first bin decreases as training set size increases, it does not go down as quickly as the overall error rate (a 16% decrease versus a 30% decrease).

F3 The Adult Dataset

F3.1 Adult Dataset without Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
17	42.4	UCI	21280	16.3	1441	8434	182.6 (28.5/212.6)

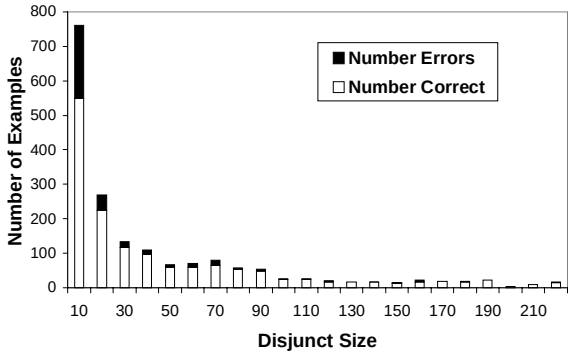


Figure F3.1.1: Distribution of Examples

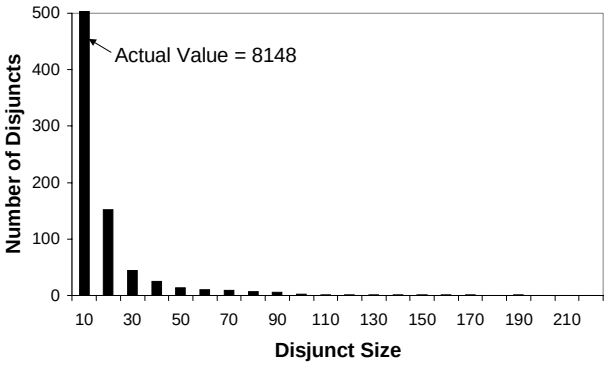


Figure F3.1.2: Distribution of Disjuncts

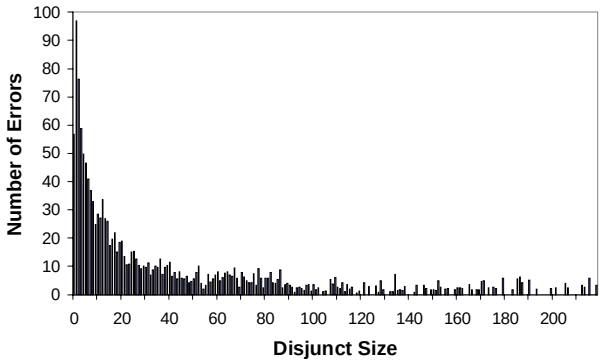


Figure F3.1.3: Distribution of Correct Examples

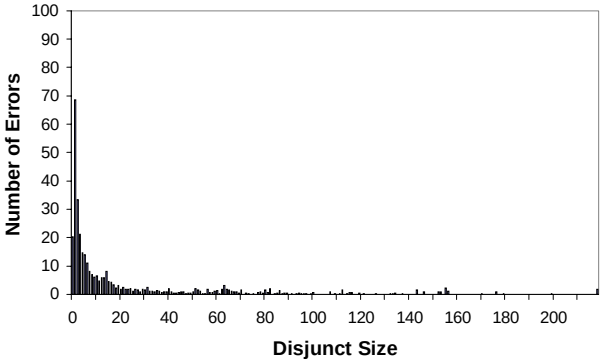


Figure F3.1.4: Distribution of Errors

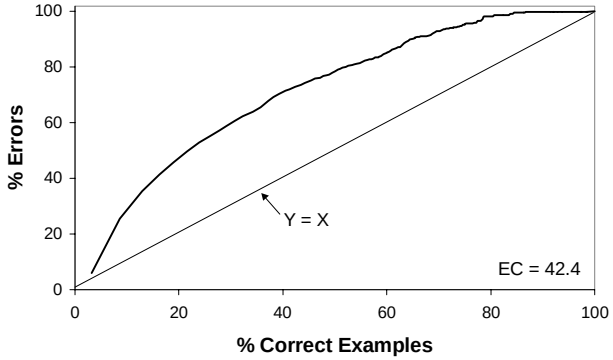


Figure F3.1.5: Error Concentration Curve

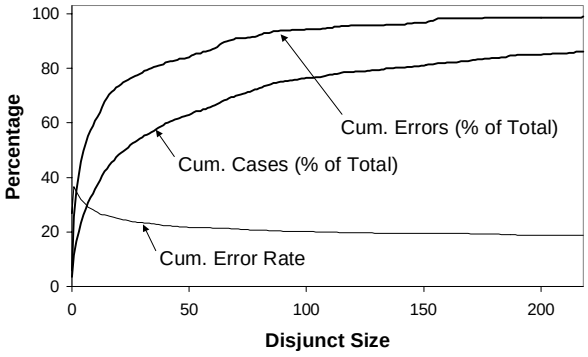


Figure F3.1.6: Cumulative Coverage Statistics

F3.2 Adult Dataset with Pruning

EC Rank	EC	Source	Dataset Size	Error Rate	Largest Disjunct	# Leaves	Mean Cov
17	42.4	UCI	21280	14.1	5017	419	2065 (967/2245)

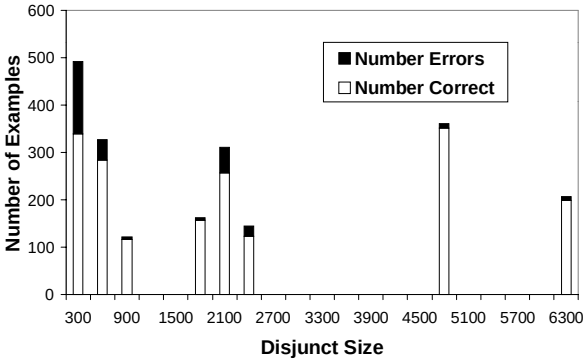


Figure F3.2.1: Distribution of Examples

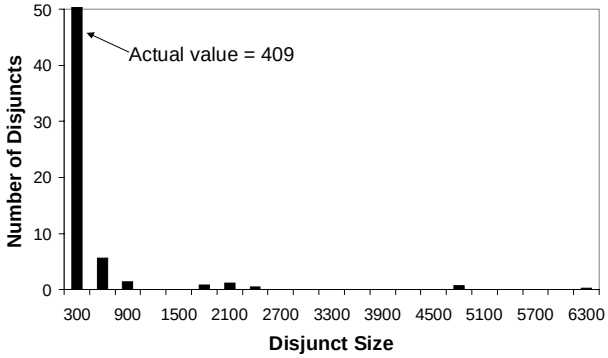


Figure F3.2.2: Distribution of Disjuncts

Intentionally Left Blank
(resolution insufficient for display)

Figure F3.2.3: Distribution of Correct Examples

Intentionally Left Blank
(resolution insufficient for display)

Figure F3.2.4: Distribution of Errors

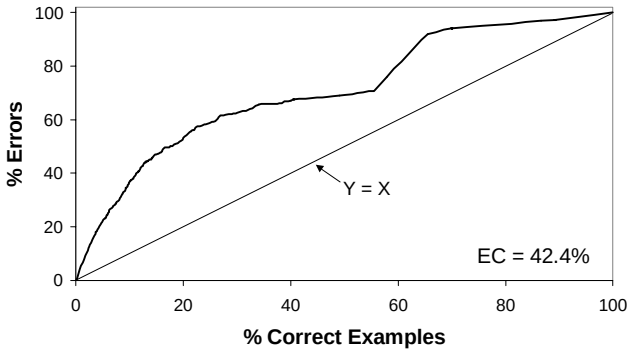


Figure F3.2.5: Error Concentration Curve

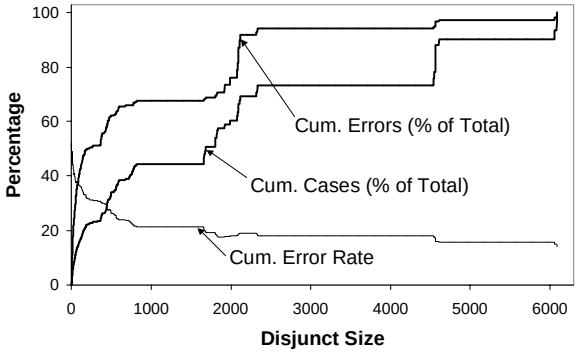


Figure F3.2.6: Cumulative Coverage Statistics

F3.3 Adult Dataset with Varying Training Set Size

Training Set Size	Error Rate	Number Leaves	Coverage Means	EC
10% (2128)	18.6%	1478.9	66.8 (8.7/80.0)	48.6%
50% (10640)	17.2%	5529.5	172.8 (19.9/204.6)	45.2%
90% (19152)	16.3%	8434.4	182.6 (28.5/2120)	42.4%

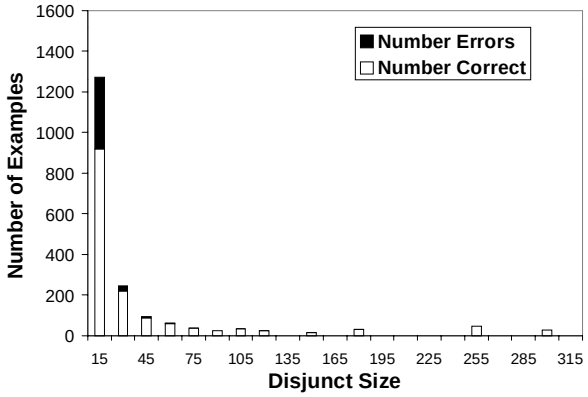


Figure F3.3.1: Distribution with 10% Training Data

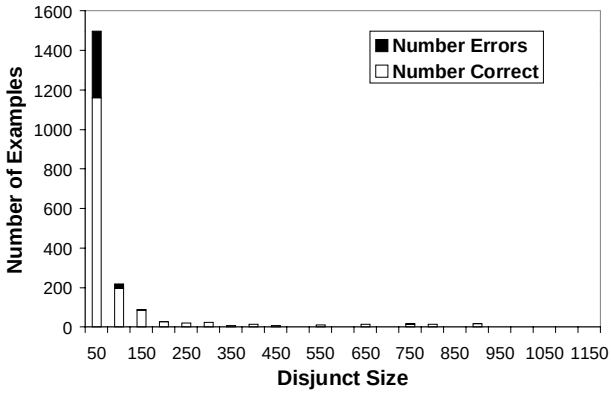


Figure F3.3.2: Distribution with 50% Training Data

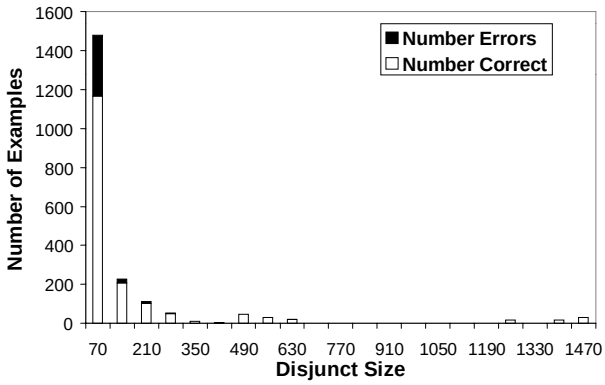


Figure F3.3.3: Distribution with 90% Training Data

Table F3.3.1: Number of Disjuncts in Learned Concept by Coverage Band

Training	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21
10%	1460	12.1	2.7	1.2	0.6	0.3	0.4	0.2	0.0	0.1	0.0	0.2	0.0	0	0.0	0.0	0.2	0.0	0	0.1	0.0
50%	5508	15.1	3.5	0.8	0.4	0.4	0.2	0.2	0.1	0.0	0.1	0.0	0.1	0	0.1	0.1	0.0	0.1	0	0.0	0.0
90%	8403	21.2	5.5	1.8	0.3	0.1	0.9	0.5	0.3	0.0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.1	0	0.1	0.2