Appears in *Proceedings of the 12th International Conference on Machine Learning*,
Morgan Kaufmann, 1995, 558-565.

Learning with Rare Cases and Small Disjuncts

Gary M. Weiss
Rutgers University/AT&T Bell Labs
200 Laurel Avenue
Middletown, NJ 07748
gary.m.weiss@att.com

Abstract

Systems that learn from examples often create a disjunctive concept definition. Small disjuncts are those disjuncts which cover only a few training examples. The problem with small disjuncts is that they are more error prone than large disjuncts. This paper investigates the reasons *why* small disjuncts are more error prone than large disjuncts. It shows that when there are rare cases within a domain, then factors such as attribute noise, missing attributes, class noise and training set size can result in small disjuncts being more error prone than large disjuncts and in rare cases being more error prone than common cases. This paper also assesses the impact that these error prone small disjuncts and rare cases have on inductive learning (i.e., on error rate). One key conclusion is that when low levels of attribute noise are applied only to the training set (the ability to learn the correct concept is being evaluated), rare cases within a domain are *primarily* responsible for making learning difficult.

1. INTRODUCTION

Many systems that learn from examples create a disjunctive concept definition. The coverage, or size, of each disjunct is defined as the number of training cases that it correctly classifies. Small disjuncts are those disjuncts which cover only a few training cases. The problem with small disjuncts is that they often have a much higher error rate than large disjuncts and are therefore considered *error prone* (Holte, Acker & Porter, 1989). Furthermore, although small disjuncts may individually cover only a few examples, collectively they can cover a significant percentage (e.g., 20%) of the total examples. Thus, they cannot be disregarded if a high level of predictive accuracy is to be achieved.

Small disjuncts, however, are not inherently more error prone than large disjuncts. This paper will show that when there are rare cases within a domain, factors such as attribute noise, missing attributes, class noise and training set size can cause small disjuncts to be error prone. This paper will also show that the rare cases within a domain are themselves error prone (i.e., they have a higher error rate than the common cases). Rare cases are defined as those cases which occur relatively infrequently within a domain. Finally, this paper will assess the impact that error prone rare cases and small disjuncts have on learning (i.e., on error rate).

It is important to understand the relationship between rare cases and small disjuncts. Rare cases exist in the underlying population from which training and test cases are chosen, while small disjuncts are a consequence of learning. Rare cases tend to *cause* small disjuncts to be formed during learning. Thus, it is more appropriate to talk about the problem with rare cases than the problem with small disjuncts, since the former exists independent of the inductive learning system being used. This paper examines the effect of small disjuncts on learning, as well as the effect of rare cases on learning, so that the results in this paper can be related to previous work in the field (previous work has focused exclusively on small disjuncts).

2. BACKGROUND

Several papers have investigated the problem of small disjuncts (see Holte, et al., 1989; Quinlan, 1991; Ali & Pazzani, 1992; Danyluk & Provost, 1993; Weiss, 1994), however none have provided a comprehensive explanation of *why* small disjuncts are error prone. Rather, previous work has concentrated on determining the effect of small disjuncts on inductive learning. One exception, however, is the work of Holte and colleagues (1989), which showed that bias can cause small disjuncts to be error prone. Also of interest, Danyluk and Provost (1993) asserted that, in the domain they were studying, learning from noisy data was hard due to a difficulty in distinguishing between

noise and true exceptions, especially since errors in measurement and classification often occur systematically rather than randomly. Weiss (1994) investigated a variant of this assertion where the errors were generated by random class noise (i.e., no systematic errors were introduced).

This paper extends the work presented in Weiss (1994). It utilizes artificial domains over which greater experimental control can be exerted and investigates the effect that systematic and random attribute noise have on learning with small disjuncts.

3. WHY ARE SMALL DISJUNCTS SO ERROR PRONE?

Although it is well known that small disjuncts are more error prone than large disjuncts, the only current explanation for this behavior is the effect of bias (Holte, et al., 1989). This section will explain how attribute noise, missing attributes, class noise and training set size can interact with rare cases within a domain to cause error prone small disjuncts and error prone rare cases. Bias will not be discussed since it has already been studied in detail (Holte, et al., 1989). However, it should be noted that all of the factors studied in this paper are associated with a domain, while bias is a property of a specific inductive learner.

Attribute noise will be considered first. Small disjuncts (by definition) individually classify fewer training cases correctly than large disjuncts; therefore they are expected to classify fewer test cases correctly. When attribute noise is introduced, it corrupts some of the original noise-free cases, making them look like other cases. During training, this means that even low levels of attribute noise can cause common cases to overwhelm rare cases. For example, if a case with binary attribute vector *11111* occurs 100 times more frequently than one with *01111*, then 2% attribute noise will cause the former case to "obscure" or "overwhelm" the latter and, if they have different classes, the wrong subconcept will be learned. Furthermore, if during learning some generalization occurs, then a common case can overwhelm a rare case even when the attribute vectors do not match exactly. Consequently, rare cases will have a higher error rate than common cases.

During testing, a case will be misclassified for one of two reasons. The first reason is that noise corrupted the test case to look like another, for which a different classification was learned. In this situation, the resulting misclassifications will tend to be distributed evenly throughout the disjuncts, independent of the coverage of the disjunct (this assumption is based on the noise model being used; see Section 5). The second reason is that the wrong classification was learned for the test case during training. Referring to the above example, it is because *01111* is overwhelmed by *11111*. However, all we know about *11111* is that it needs to be more common than *01111*. Thus, these misclassifications will still be spread out with respect to disjunct size—at least to some degree. Thus, although each disjunct will tend to have roughly the same number of misclassified cases, small disjuncts will have fewer correctly classified cases and hence be more error prone than large disjuncts.

Missing attributes will also cause rare cases and small disjuncts to be error prone. With an attribute missing, some previously distinct cases will now appear identical. Consequently, the classification that is learned will be determined by which cases occur most frequently. Hence, rare cases will be more error prone than common cases. Small disjuncts will also be error prone because the misclassifications will tend to be distributed independent of disjunct size.

Class noise has a fundamentally different effect than does attribute noise on rare cases and small disjuncts. During training, class noise does not cause one case to look like another (i.e., attribute vectors to overlap). However, if the level of class noise is sufficiently high, then the wrong class may be learned—with this being much more likely for rare cases and those disjuncts which cover few cases. For example, with 40% class noise it is more likely that the wrong class will be learned for a disjunct that covers 10 cases than one that covers 1000 cases—in the former case there is a greater statistical chance that more than 50% of the training cases are corrupted. A similar argument holds for rare cases. The effect of class noise on the test set is independent of disjunct size or how rare a case is—n% class noise tends to increase the error rate by n%. So, class noise during testing will result in errors which will be distributed evenly throughout the disjuncts. Since small disjuncts classify fewer test cases correctly than large disjuncts, this will result in small disjuncts being more error prone. However, unlike the effect of attribute noise, this effect may only be significant at high levels of noise.

Training set size also has an impact on learning. Rare cases will have a higher error rate than common cases since they are less likely to be found in the training set. The small disjuncts will tend to be more error prone because they cover fewer correct cases than large disjuncts.

4. THE PROBLEM DOMAINS

To allow the impact of rare cases on learning to be easily assessed, artificial domains are used. This makes it possible to construct domains with and without rare cases, and in which these rare cases are easily identified. Another advantage of using artificial domains is that they make it possible to start off "clean"—with no noise, missing attributes or inconsistent cases within a domain, and in which 100% predictive accuracy is possible.

Each domain has five binary attributes and a binary class. The class is determined by a parity function which computes the class based on whether there is even or odd parity or by a voting function which computes the class based on whether the attributes contain more 0's or 1's.

Each domain is defined by selecting one of these two functions to compute the class and then selecting a distribution of cases. Since one purpose of this paper is to examine how rare cases affect learning, two distributions were designed—a *uniform* distribution which contains no rare cases and a *skewed* distribution which contains a mixture of rare and common cases. The name of each domain corresponds to the name of the function used to compute the class: prefixed by *eq* if the uniform distribution is used and no prefix if the skewed distribution is used.

For the uniform distribution, each of the 32 (2^5) distinct cases will occur with the same frequency. The skewed distribution was designed so that: 1) it contains common cases, rare cases and cases in between these two extremes and 2) the rare cases collectively cover a significant percentage of the overall cases. This type of distribution has been seen in existing domains, including the KPa7KR chess endgame domain (Holte, et al., 1989), the NYNEX MAX domain (Danyluk & Provost, 1993) and the Wisconsin breast cancer domain (Weiss, 1994).

Table 1 shows how the skewed distribution is formed by dividing the 32 distinct cases unequally into five bands and then duplicating the cases in each band by the specified amount. There are a total of 96 cases. The training and test sets are formed by randomly selecting cases (with replacement) from the resulting distribution. Each case in band 5 is rare (selected only 1/96 of the time), but collectively these cases cover 1/6 of the total cases. To ensure that comparisons between the uniform and skewed distributions are meaningful, distinct cases were assigned to each band so that the class distributions are identical for each band (i.e., 50-50). Had this not been done, the experiments could be biased since a band which contains cases with an uneven class distribution (e.g., 80-20) would be easier to learn from.

Table 1: Distribution of Cases

Band	# distinct cases	Dup. factor
1	2	16
2	2	8
3	4	4
4	8	2
5	16	1

5. THE EXPERIMENTS

Experiments were run to determine the behavior of small disjuncts and rare cases. For each experiment, five independent runs were performed and the results averaged together. For the experiments focusing on small disjuncts, the cases for each run were randomly drawn from the distribution described in Table 1. These cases were then divided into a training and a test set, which were then fed into C4.5, a program for inducing decision trees from a set of preclassified training examples (Quinlan, 1993). C4.5 was modified by the author to to collect statistics relating to disjunct size and to disable the default pruning strategy, since pruning might obscure the small disjuncts in the underlying concept definition. The same method was used for the experiments focusing on the rare cases, except that five test sets were used instead of one—with each test set containing cases from only one of the five bands described in Table 1. This makes it possible to compare the error rates of rare and common cases.

Although the cases from the problem domains were initially noise-free, most of the experiments added some form of noise. Noise was applied to either the class label or to all of the attributes, in either a random or systematic fashion. N% random noise signifies that, with probability N/100, a value is randomly selected from the *remaining* alternatives (Quinlan, 1986). A very simple model of systematic noise is used in which noise only corrupts values in one direction. For example, 5% systematic attribute noise means that each attribute value of 0 is corrupted to a 1 with probability .05, but a 1 is never corrupted to a 0.[1]

Noise can be applied to the training set, test set or both sets. For the majority of experiments, either attribute noise was applied to both sets or class noise was applied to the training set (it does not make sense to apply class noise to the test set, since it is the class label that the learner is trying to predict). Frequently, however, when the effects of noise are studied, noise is applied to the training set but not to the test set (Quinlan, 1986). What is being studied in this case is the ability to learn the correct concept when noise is present. Table 2 summarizes the experimental parameters.

Table 2: Experimental Parameters

Parameter	Values
Domain	parity, eqparity, voting, eqvote
Noise:	
level	varies (0%-40%)
type	attribute, class
model	random, systematic
applied to	training set, test set, both
Training set size	1000 (varies in Section 6.4)
Test set size	2000
Pruning strategy	no pruning

1. This does *not* imply that the noise we are attempting to model has a random component. Imagine a domain with a binary class which depends on whether the measured voltage is > 5 volts, and where the voltmeter consistently (systematically) adds 1 volt to the true voltage. Each case will be misclassified only if the true voltage is between 4-5 volts. The systematic noise causes the value to be corrupted in one direction only, based on the statistical likelihood of the voltage being within the 4-5 volt range.

6. RESULTS AND DISCUSSION

This section presents the results of the experiments described in Section 5. These results show how learning from a domain which contains rare cases is affected by the following factors: random and systematic attribute noise (Section 6.1), missing attributes (Section 6.2), class noise (Section 6.3), and training set size (Section 6.4). Within each subsection, results are presented which demonstrate that the rare cases and small disjuncts are error prone, followed by results which demonstrate their impact on learning (i.e., on the overall error rate).

6.1 ATTRIBUTE NOISE

Sections 6.1.1 and 6.1.2 present the experiments in which random and systematic attribute noise (respectively) are applied to both the training and test sets. Section 6.1.3 presents the experiments where attribute noise is applied to either the training or the test set.

6.1.1 Random Attribute Noise

Experiments were run with varying levels of random attribute noise applied to both the training and test sets. Due to the large training set size which was used, all distinct cases were trained on. This was done to isolate the impact that noise has on learning from the impact that training with an incomplete training set has on learning.

When 5% attribute noise was applied to the parity domain, the error rate increased as cases became increasingly rare. More specifically, the error rates for bands 1-5 were: 15.53%, 16.67%, 20.30%, 22.86% and 36.01%, respectively. This finding clearly demonstrates that rare cases are more error prone than common cases. Figure 1 shows that, for the same experiment, the parity domain also has error prone small disjuncts: the cumulative error rate decreases with increasing coverage and at low coverages a greater percentage of errors than cases are seen. Note that the curves are cumulative with respect to coverage, because at coverage *n* the various quantities are measured on the cases which are covered by disjuncts in the *range* 0-*n*.[2] When the same experiments were run for the eqparity domain, the curves representing the cumulative cases and cumulative errors overlapped almost exactly and the error rate remained constant.

Since a major concern of this paper is the error prone nature of small disjuncts, it would be useful to know at what coverages errors are concentrated. A new statistic, called *error factor*, can concisely present this information. Error factor is a function of disjunct size and is defined as the cumulative percentage of the total errors divided by the cumulative percentage of the total cases. For example, an error factor of 2 at coverage 100 means that, between coverage 0-100, twice as many errors were seen than expected if coverage had no effect on error rate. Figure 2 shows the error factor for the domains used in this paper (note that the error factor is undefined until some cases are covered). Figure 2 demonstrates that small disjuncts are more error prone than large disjuncts. Namely, when there is a skewed distribution of cases (i.e., parity and voting), small disjuncts are more error prone than large disjuncts (error factor > 1), but when the distribution is uniform, this is not true (or much less true). Thus, random attribute noise causes small disjuncts to be more error prone than large disjuncts when there are rare cases in a domain.

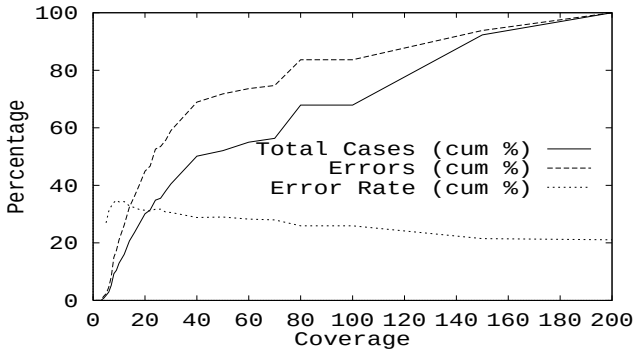

Figure 1: Effect of 5% Attribute Noise on Parity

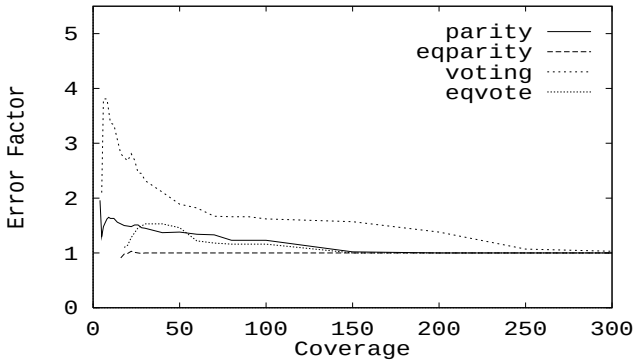

Figure 2: Error Factor with 5% Attribute Noise

The impact that attribute noise has on learning can be assessed by comparing the error rates for the skewed and uniform distributions. However, to fully understand this comparison, it is useful to first examine what happens as increasing levels of noise are applied to a domain—in this case, the parity domain. Figure 3 shows us that the error factor *decreases* as the noise level increases. This is because, at high noise levels, a greater percentage of the errors are contributed by the large disjuncts. This is not very surprising since they collectively cover more cases than the small disjuncts. This is clearly demonstrated by Figure 4, which shows how the errors are distributed by coverage.

2. Since the *cumulative* error rate is being plotted (as opposed to the error rate at each specific coverage band), the differences in error rate appear smaller than they really are. For example, in Figure 1 the error rate of the cases covered by disjuncts with size 100-200 is 11.8%.

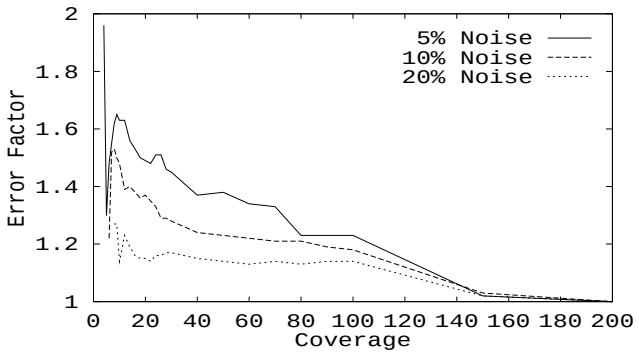

Figure 3: Effect of Attribute Noise on Error Factor

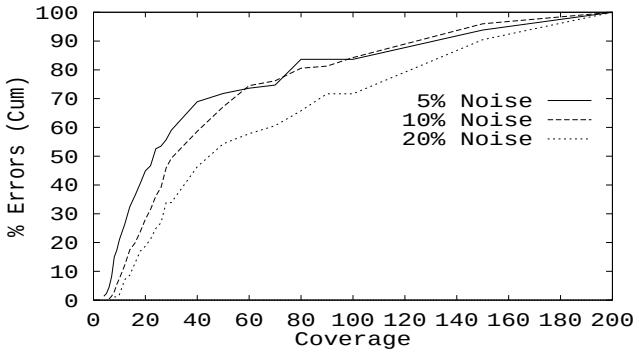

Figure 4: Effect of Attribute Noise on Error Distribution

Table 3, which compares the results of learning with and without rare cases, demonstrates the impact that rare cases and attribute noise have on learning. Note that as the noise level increases, rare cases become less responsible for learning with noise being difficult (Figure 4 shows why this is so). In fact, at high noise levels, the skewed distribution performs better than the uniform distribution, probably because it is still able to perform well on the common cases. Although the impact of small disjuncts is greatest at the 2% noise level, one is still not able to conclude that at this level rare cases/small disjuncts are *responsible* for making learning difficult (when random attribute noise is applied to both training and test sets).

Table 3: Impact of Random Attribute Noise

Noise level	Error rate (%) parity/eqparity	Delta	Error rate (%) voting/eqvote	Delta
2%	10.03/8.77	13.4%	4.07/3.58	12.8%
3%	14.78/13.35	10.2%	6.00/5.73	4.6%
5%	21.08/20.06	5.0%	9.08/8.66	4.7%
10%	32.34/35.76	-10.0%	15.53/15.44	0.6%
20%	41.09/48.98	-17.5%	26.61/27.27	-2.4%

6.1.2 Systematic Noise

Danyluk and Provost (1993) stated that it was the combination of small disjuncts and systematic noise which made learning difficult in the NYNEX MAX domain. Based on this assertion, the experiments involving random attribute noise presented in Section 6.1.1 were repeated using systematic attribute noise. The results were very similar to those for random attribute noise and, in the interest of space, are not all presented here. Table 4 demonstrates the impact that systematic attribute noise has on learning with rare cases. While there are some differences between these results and the ones for random attribute noise, it still cannot be said that rare cases/small disjuncts are responsible for making learning difficult.

Table 4: Impact of Systematic Attribute Noise

Noise level	Error rate (%) parity/eqparity	Delta	Error rate (%) voting/eqvote	Delta
2%	5.13/4.72	8.3%	2.68/2.04	27.1%
3%	7.40/7.17	3.2%	3.36/2.87	15.7%
5%	12.33/11.38	8.0%	6.04/4.36	16.2%
10%	19.42/20.37	-4.8%	8.96/8.58	4.3%
20%	28.33/35.78	-23.0%	15.32/19.83	-25.7%

6.1.3 The Two Effects of Attribute Noise

Noise can be thought of as having two distinct, albeit interacting, effects (Weiss, 1994). Noise applied to the training set prevents the correct concept definition from being learned, while noise applied to the test set causes the correct concept definition, had it been learned, to misclassify some test cases.[3] Applying noise to the training and test sets separately allows these two effects to be measured separately, leading to a clearer understanding of how noise affects learning, especially with rare cases.

In the previous experiments, noise was applied to both the training and test sets. This section demonstrates what occurs when noise is applied to only one of these sets. As mentioned earlier, when noise is applied to only the training set, then the ability to learn the correct concept definition is being evaluated. Although the situation where noise is applied only to the test set is not studied as frequently, it arises whenever steps are taken to clean up the training data.

Tables 5 and 6 show the results of the experiments where noise was applied to the training set *or* the test set. Note that the results from the experiments for random and systematic noise should not be directly compared since, with the same noise level, systematic noise corrupts only half as many values (see Section 5). However, both sets of results indicate that rare cases make it *much* more difficult to learn the correct concept definition (i.e., learn when noise is applied only to the training set) at all but the highest levels of noise. In fact, in this situation rare cases are *primarily responsible* for learning being difficult. Tables 5 and 6 also show that when attribute noise was applied to the test set, the impact of learning with rare

3. Note, however, that if noise in the training set prevents the correct concept definition from being learned, then noise in the test set can actually *improve* the predictive accuracy.

cases was not very significant—the error rates are high regardless of which distribution was used. We can now explain why rare cases only make it slightly harder to learn when noise is applied to both the training and test sets. When rare cases exist within a domain, the presense of noise in the training set has a big effect, but the presense of noise in the test set has very little effect—but the test component contributes the majority of the errors.

Table 5: Two Effects of Random Attribute Noise

Noise applied to	Noise level	Error rate (%) parity/eqparity	Delta	Error rate (%) voting/eqvote	Delta
training set	5%	6.33/0	>100%	1.80/0	>100%
	10%	12.02/6.37	61.4%	3.81/0	>100%
	20%	26.66/30.69	-14%	7.05/1.87	>100%
test set	5%	21.29/20.62	3.2%	9.00/8.64	4.1%
	10%	33.51/33.82	-0.9%	16.33/15.98	2.2%
	20%	46.17/45.55	1.4%	27.05/26.38	2.5%

Table 6: Two Effects of Systematic Attribute Noise

Noise applied to	Noise level	Error rate (%) parity/eqparity	Delta	Error rate (%) voting/eqvote	Delta
training set	5%	0.98/0	>100%	0.79/0	>100%
	10%	3.54/0	>100%	4.21/0	>100%
	20%	10.19/0.68	>100%	7.98/9.60	-18.4%
test set	5%	11.71/11.48	2.0%	5.59/4.66	18.1%
	10%	18.98/20.24	-6.4%	10.82/9.26	15.5%
	20%	29.12/33.44	-13.7%	19.98/17.93	10.8%

6.2 MISSING ATTRIBUTES

Learning is affected when the available set of attributes is insufficient to correctly classify each case. Experiments were run with the first attribute removed from all of the cases in the two domains, but the class was still computed using all five attributes. Figure 5 shows that the missing attribute causes error prone small disjuncts, while Table 7 shows that it also causes error prone rare cases. Although no experiments were run to compare the effect of missing attributes on the skewed and uniform distributions, the results are expected to be consistent with those for learning with attribute noise (see Table 3).

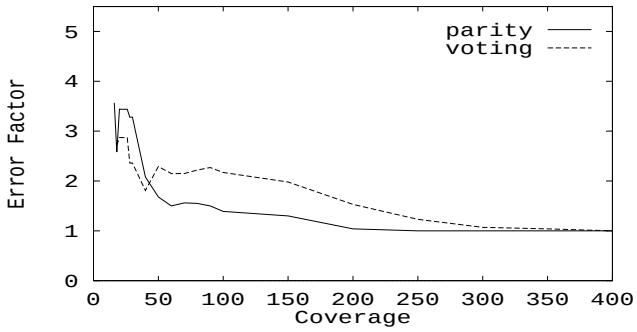

Figure 5: Effect of Missing Attribute on Error Factor

Table 7: Effect of Missing Attribute on Test Bands

Domain	Band 1	Band 2	Band 3	Band 4	Band 5
Parity	0%	0%	0%	30%	85%
Voting	0%	0%	0%	25%	25%

6.3 CLASS NOISE

Experiments were run to determine how class noise affects learning with rare cases. Figure 6 and Table 8 show what happened when random class noise was applied to the training set of the voting domain. The results (most of which are not shown here due to space limitations) demonstrate that class noise is much less of a factor than attribute noise in causing error prone rare cases and error prone small disjuncts. The results demonstrate that class noise only becomes a factor when it exceeds 30%.[4] This assumes that class noise is only applied to the training set. When class noise was applied to both the training and test sets, the resulting error rate was always within 3% of the level of class noise. However, as mentioned earlier, it really only makes sense to apply class noise to the training set, since the *purpose* of learning is to to predict the class of the test cases. Class noise has a very different impact on learning with rare cases than does attribute noise. When the noise level gets high enough for there to be a significant impact on learning (at the 40% noise level), the skewed distribution generally outperforms the uniform distribution, most likely due to the common cases. Thus, rare cases do not make learning particularly susceptible to class noise, although they do make learning susceptible to attribute noise (as was shown in Section 6.1).

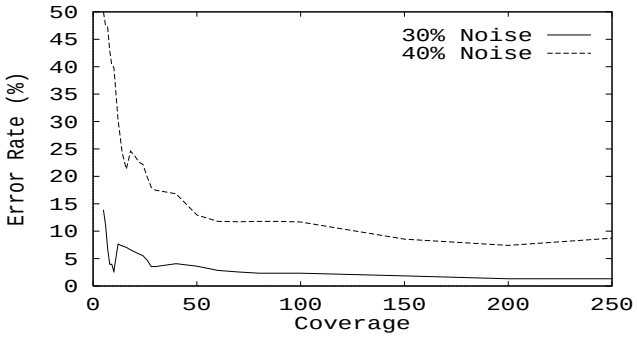

Figure 6: Impact of Class Noise

Table 8: Effect of Class Noise on Test Bands

Noise level	Band 1	Band 2	Band 3	Band 4	Band 5
30%	0%	0%	0%	5%	10%
40%	0%	0%	10%	17.5%	36.25%

4. Had the training set size been smaller, the small disjuncts would have covered fewer cases, in which case class noise may have been a factor at a lower noise level (see the explanation of how class noise affects learning with small disjuncts in Section 3).

6.4 TRAINING SET SIZE

For the experiments described in the previous sections, the training set size was sufficiently large so as to include each of the 32 distinct cases in the training set. The experiments in this section use a variety of training set sizes in order to determine what effect training set size has on learning. The following training set sizes were used: 8, 17, 34, 68, 125, 250, 500, and 1000. Table 9 shows that, for training set size 125, rare cases are more error prone than common cases. This finding also holds true for the other training set sizes.

While the error factor is defined as a function of coverage, Figure 7 only plots the error factor at the point at which 50% of the errors have been seen. This is done so that a separate curve is not needed for every training set size, while still providing a good indication of where the errors are concentrated. The curves do not extend further along the x-axis because the error factor is meaningless when predictive accuracy reaches 100% (i.e., there are no errors). Figure 7 shows that, for the skewed distributions, the larger training set sizes cause small disjuncts to be more error prone than large disjuncts, while the effect is minimal for the uniform distributions.

Table 9: Effect of Training Set Size 125 on Test Bands

Domain	Band 1	Band 2	Band 3	Band 4	Band 5
Parity	0%	0%	0%	12.5%	26.25%
Voting	0%	0%	0%	2.5%	15%

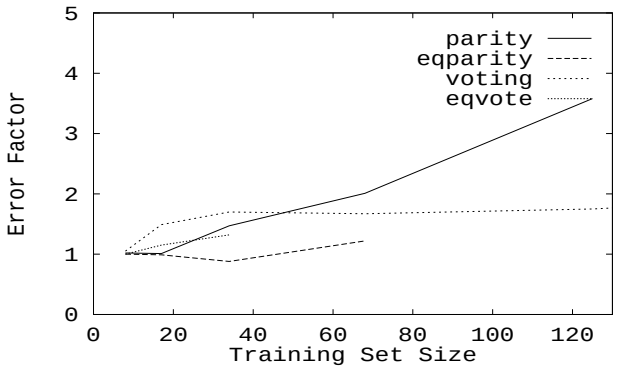

Figure 7: Effect of Training Set Size on Error Factor

Figure 8 shows that the skewed distributions outperformed the uniform distributions when the training set size was small, but that the opposite was true when the training set size was large, with a crossover point in the middle. This occurs because the skewed distribution has an advantage when the training set size is small, since the common cases are still likely to get trained on, but a disadvantage when the training set size is large, because the rare cases may still not get trained on. So, as the training set size grows, rare cases/small disjuncts become increasingly responsible for the errors—and this is exactly what Figure 7 shows.

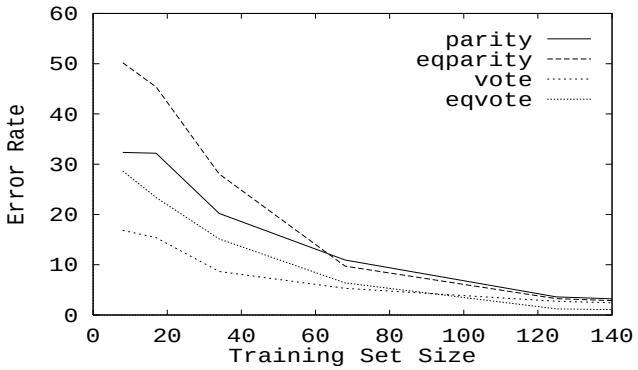

Figure 8: Impact of Training Set Size

7. FUTURE RESEARCH

There are several next steps which suggest themselves. Some "real world" domains should be tested to see if they yield similar results, even though interpreting the results from these domains would be difficult (see Section 4). Some real world domains were used by Weiss (1994), but only the effect of random class noise was examined.

In order to separate the effect of training set size from the effects of noise and missing attributes, the training set size was chosen to be very large relative to the number of distinct cases (except for the experiments described in Section 6.4). Future research should try to determine if the results in this paper hold for smaller training sets.

The noise model described in Section 5 is critical to this paper—a different noise model may have led to different results. Empirical support for this noise model, or for an alternate model, would be useful. Also, the results in this paper failed to demonstrate that when learning with rare cases/small disjuncts, systematic noise is any more problematic than random noise. Future research should continue to investigate the different impact that systematic noise has on learning.

This paper examined how one characteristic of problem domains, the presence of rare cases, affects learning. More emphasis was placed on studying this domain characteristic than on a specific inductive learning system. Perhaps this can be viewed as a profitable future research direction. It may permit a better understanding of the learning task, and allow us to better understand the strengths and weaknesses of the various learning systems which currently exist.

8. CONCLUSION

This paper demonstrated that attribute noise, missing attributes, class noise and training set size can each cause rare cases and small disjuncts to be more error prone than common cases and large disjuncts, respectively. It also provided a theoretical reason for this behavior. This paper demonstrated that when any of these factors, with the

exception of class noise, are present, then rare cases within a domain make inductive learning more difficult. With attribute noise, this difficulty is greatest at low noise levels—at higher noise levels, the impact that noise has on the common cases/large disjuncts dominates. When low levels of attribute noise are applied to both the training and test sets, rare cases make learning only slightly more difficult. However, when low levels of attribute noise are applied only to the training set (the ability to learn the correct noise-free concept is being evaluated), then rare cases within a domain are *primarily* responsible for making learning difficult. High levels of class noise were required to cause rare cases and small disjuncts to be error prone. However, in this situation the rare cases only had a minimal impact on learning.

The effect of training set size on learning depends on whether the domain includes rare and common cases. For large training set sizes, learning from a skewed distribution is more difficult than learning from a uniform distribution, but below a certain training set size, the opposite is true.

This paper also provides insight into the relationship between rare cases and small disjuncts, as well as how each of these should be defined. Rare cases and small disjuncts may need to be defined in relative terms if we want them to be predictors of error prone behavior. That is, in the experiments described in this paper, rare cases and small disjuncts turn out to be error prone only because there are common cases and large disjuncts. If we constructed a domain which contained, in an absolute sense, *only* rare cases and small disjuncts, then the factors described in this paper would not cause these to be error prone. One possible way of defining small disjuncts is in terms of error factor—for example, small disjuncts could be defined as those disjuncts for which the error factor is above some threshold value.

Acknowledgements

I would like to thank Andrea Danyluk and Rob Holte for their comments on an earlier version of this paper, and Foster Provost, Brian Davison, and Rosalie DiSimone-Weiss for comments on the current version. I would especially like to thank Haym Hirsh for valuable early discussions and many helpful comments.

References

Ali, K. M. & Pazzani, M. J. (1992). Reducing the small disjuncts problem by learning probabilistic concept descriptions. Technical Report 92-111, Irvine, CA: University of California at Irvine, Department of Information and Computer Sciences. To appear in T. Petsche (ed.), *Computational Learning Theory and Natural Learning Systems, Volume 3*, Cambridge, Massachusetts. MIT Press.

Danyluk, A. P. & Provost, F. J. (1993). Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network. In *Machine Learning: Proceedings of the Tenth International Conference*, 81-88, San Francisco, CA: Morgan Kaufmann.

Holte, R. C., Acker, L. E. & 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. (1986). The effect of noise on concept learning. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (eds.), *Machine Learning, an Artificial Intelligence Approach, Volume II*, 149-166, Morgan Kaufmann.

Quinlan, J. R. (1991). Technical note: improved estimates for the accuracy of small disjuncts. *Machine Learning, 6*(1).

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

Weiss, G. M. (1994). The problem with noise and small disjuncts. Technical Report ML-TR-38, New Brunswick, NJ: Rutgers University, Department of Computer Science.