The Problem with Noise and Small Disjuncts

Gary M. Weiss
AT&T Bell Labs/Rutgers University
gweiss@paul.rutgers.edu

Abstract

Systems that learn from examples often create a disjunctive concept definition. The disjuncts in the concept definition which cover only a few training examples are referred to as small disjuncts. The problem with small disjuncts is that they are more error prone than large disjuncts, but may be necessary to achieve a high level of predictive accuracy [Holte, Acker, and Porter, 1989].

This paper extends previous work done on the problem of small disjuncts by taking noise into account. It investigates the assertion that it is hard to learn from noisy data because it is difficult to distinguish between noise and true exceptions. In the process of evaluating this assertion, insights are gained into the mechanisms by which noise affects learning. Two domains are investigated. The experimental results in this paper suggest that for both Shapiro’s chess endgame domain [Shapiro, 1987] and for the Wisconsin breast cancer domain [Wolberg, 1990], the assertion is true, at least for low levels (5-10%) of class noise.
1. Introduction

It is important to understand the effect that noise has on inductive learning, because noise is generally present in real world domains and may be the limiting factor in how well a learning system performs if a large training set is available. This paper has two goals: to demonstrate the effect that noise has on inductive learning, and to provide an understanding of the mechanisms by which noise affects inductive learning. With regard to the latter goal, this paper will investigate the assertion that learning with noise is difficult because of small disjuncts in the concept definition to be learned.

One method of improving the predictive performance of an inductive learning system is to use an overfitting avoidance strategy (e.g., pruning). Since these strategies largely operate by eliminating small disjuncts, understanding the role of small disjuncts in learning from noisy data will lead to a better understanding of how overfitting avoidance strategies work and when they are most effective.

2. The Problem

This section will define the problem with small disjuncts, as described in [Holte, Acker, and Porter, 1989], and then define the problem with noise and small disjuncts.

2.1 The Problem with Small Disjuncts

Systems that learn from examples often create a disjunctive concept definition, where each term in the disjunct can be thought of as defining a subconcept of the concept to be learned. Some of these subconcepts may describe only a few cases, whereas others may describe many cases. The subconcepts which individually cover only a small fraction of the overall cases are called small disjuncts.

As defined in [Holte et al., 1989], the coverage of a disjunct is defined as the number of training examples it correctly classifies. That paper showed that, although small disjuncts individually cover only a small fraction of the examples, collectively they can cover a significant percentage (e.g., 20%) of the total examples. The problem with small disjuncts is that they are more error prone than large disjuncts, but the effect of eliminating all small disjuncts is "difficult to predict".

This paper defines coverage in terms of the number of test cases that are correctly classified. Since the number of test cases exceeds the number of training cases for the experiments performed in this paper, this definition yields a more granular measure\(^1\). However, since this definition of coverage is defined over the test set, it has the disadvantage of not being able to be employed in the learning strategy (which isn’t a

---

\(^1\) Both [Holte et al., 1989] and this paper use the KPa7KR chess end game domain, with 200 training and 3000 test cases. Since there are so many more test cases than training cases, the cases which fall into the band of disjuncts of size 1-3 in Holte et al roughly correspond to those that fall into the band of size 15-45 in this paper.
concern in this paper).

Section 4 will empirically test whether the observations made by Holte, *et al.*, hold true for the domains used in this paper.

### 2.2 The Problem with Noise and Small Disjuncts

The initial motivation for this paper came from [Danyluk and Provost, 1993]. That paper described the NYNEX MAX system, which is used for diagnosing problems in the local loop of the telephone system. MAX did not perform well when trained on noisy data, and the reason given was that learning from noisy data is difficult because it is difficult to distinguish between noise and true exceptions, especially since errors in measurement and classification often occur systematically rather than randomly.

This paper will investigate only part of this reason (using a domain other than MAX). The following assertion will be investigated in this paper:

**Assertion:** It is hard to learn from noisy data *because* it is difficult to distinguish between noise and true exceptions.

For this assertion to be true for a given domain, two conditions must hold. The first is that small disjuncts must collectively cover a significant percentage of the examples, since if they do not then we could completely eliminate them from the learned definition and have only a minimal decrease in predictive accuracy. The second condition is that the negative impact of noise on small disjuncts must outweigh the impact of noise on large disjuncts— if not, the small disjuncts cannot be blamed for making learning difficult in the presence of noise. These two conditions will be evaluated in sections 4 and 5.

It is important to note that the assertion refers to true exceptions rather than small disjuncts, and that these two terms are not identical. True exceptions are rare or uncommon cases which are positive instances of the concept to be learned, and tend to cause small disjuncts to be formed during learning. These small disjuncts will cover the true exceptions, but may also cover other cases which should not be covered. The experiments in this paper deal with small disjuncts rather than true exceptions because the latter cannot easily be identified. One final condition which must hold for the above assertion to be true is that the true exceptions must be necessary to achieve good predictive accuracy. Since there is no easy way to evaluate this for the domains used in this paper, we must be satisfied with checking that the small disjuncts cover a significant percentage of the examples.

### 3. Description of Experiments

This section will describe the inductive learning program used, the problem domains, and the experimental methodology.

#### 3.1 C4.5

C4.5, a descendant of ID3, is a program for inducing decision trees from a set of preclassified training examples [Quinlan, 1993]. C4.5 is used for all of the experiments in this paper, and was modified by the author to collect statistics relating to disjunct size and to optionally disable the default pruning strategy (pruning would obscure the small
disjuncts in the underlying concept definition).

For the majority of experiments, C4.5 was run in one of the following two configurations:

- with its pruning strategy enabled and its default parameter settings
- with its pruning strategy disabled and with the -m1 option to disable the default stopping criterion\(^2\).

In the second configuration, C4.5 will build a decision tree that is guaranteed to correctly classify all training cases (assuming consistent cases). For a few experiments, pruning is disabled and larger values are given to the -m option, to simulate a more aggressive overfitting avoidance strategy.

### 3.2 The Problem Domains

#### 3.2.1 KPa7KR Chess Endgame

The first problem domain is the chess endgame King+Rook versus King+Pawn on a7 (abbreviated KPa7KR) described by Shapiro in [Shapiro, 1987]. There are 3196 total examples, each of which represents a board position. Each example has 36 attributes, and belongs to the class "won" or "nowin" (the class distribution is approximately equal).

#### 3.2.2 Wisconsin Breast Cancer Database

The Wisconsin breast cancer database was obtained from the University of Wisconsin hospitals, Madison, from Dr. William Wolberg [Wolberg, 1990]. There are 699 total examples. Each example has nine attributes (sample number is omitted), and belongs to the class "benign" or "malignant" (about 2/3 of the examples are benign).

In the interest of space, only the most important experimental results for this domain will be displayed or discussed, unless they differ significantly from those of the chess domain. Unless otherwise noted, all results are for the chess domain.

### 3.3 Experimental Methodology

For all experiments, the examples are randomly split into disjoint training and test sets. For each experiment, seven independent runs are performed (with different randomly selected training and test sets), and the results are averaged together. All measurements (e.g., error rates) are based on the test sets.

Varying levels of randomly generated class noise are used in the experiments. The examples are considered to be initially noise-free. For the purposes of this paper, n\% noise means that with probability n/100, a value is randomly selected from the remaining alternatives\(^3\). This means that when 50\% class noise is applied to a domain with binary valued classes, there is no information provided by the class.

---

2. The default value for the -m option is 2, which stops a node from being split if the resulting nodes will have 2 or fewer outcomes.

3. This definition is consistent with the one used in [Langley and Kibler, 1991], but is very different from the one in [Quinlan, 1986], in which the same value can be selected again. For classes with two values, this makes a factor of 2 difference.
In all cases, the resulting overall noise level may vary slightly from the specified level, since the noise is applied to each case independently. For cases where noise is applied to both the test and training sets, this is done before the cases are randomly split (no attempt is made to preserve the noise levels between these two sets). There is also no artificial attempt to equalize noise levels between classes, nor maintain the class proportions between the test and training sets. However, since all results are based on seven independent runs, these factors are not expected to significantly influence the results.

The experimental parameters and their values are shown in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise:</td>
<td>0%, 5%, 10%, 20%, 30%, 50%</td>
</tr>
<tr>
<td>level</td>
<td>class</td>
</tr>
<tr>
<td>type</td>
<td>test and training, test only, training only</td>
</tr>
<tr>
<td>applied to</td>
<td></td>
</tr>
<tr>
<td>training set size</td>
<td>200 cases</td>
</tr>
<tr>
<td>pruning strategy</td>
<td>none, C4.5 default, -m20, -m30</td>
</tr>
</tbody>
</table>

Table 1. Experimental Parameters

4. The Problem with Small Disjuncts

This section will empirically explore the problem with small disjuncts which was discussed in section 2.1. The reason this section is included, even though the emphasis of this paper is on the effect of noise on small disjuncts, is to verify that small disjuncts collectively cover a significant percentage of examples in these two domains, and are more error prone than large disjuncts.

Figures 1 and 2 shows the results of running C4.5 with its pruning strategy disabled. It plots the coverage against cumulative total cases, cumulative errors, and the cumulative error rate, each expressed as a percentage (note that for all graphs in this paper, coverage is plotted using a logarithmic scale). For example, by looking at the point where the curves corresponding to errors and error rate intersect (at coverage 35) in figure 1, we see that for the cases covered by the disjuncts with size 1-35, the error rate (for these cases) is 50%, and they account for 50% of the total errors but for only 5% of the total cases (this shows the problem with small disjuncts).
The plot of cumulative total cases (matched and errors) shows that the majority of the cases are concentrated toward the high coverages for both domains, but that small disjuncts nonetheless cover a significant percentage of the cases. The cumulative errors are concentrated much more toward the small disjuncts (as previously stated, for the chess domain at coverage 35 we have seen 50% of the total errors but only 5% of the total cases). Finally, the error rate is high for small disjuncts, and, for the most part, drops as the disjunct size increases. These results agree with those from [Holte et al., 1989].

Figures 1 and 2 show that small disjuncts cover a significant percentage of the cases. This is more evident for the breast cancer domain, but even for the chess domain, disjuncts of size \( \leq 100 \) cover about 20% of the total cases (100 may not sound small, but it corresponds to a disjunct of size 7 if the definition of coverage from Holte et al., were used).

5. The Problem with Noise and Small Disjuncts

This section shows what happens when noise is introduced into the data. Section 5.1 shows what happens when class noise is applied to the test and training sets. Section 5.2 argues that there are two basic effects of noise, and demonstrates these effects by running experiments where noise is applied to only the test or the training set.
5.1 Class Noise Applied to All Examples

In this set of experiments, varying levels of class noise are applied to both the training and test examples. The overfitting avoidance strategy which is used is specified.

Figures 3 and 4 show the effect of noise on error rate and error factor (for the chess domain), respectively, with pruning disabled. The error factor is a function of coverage, 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 13 means that 13 times more errors had been seen than expected (if coverage had no effect on error rate).

![Figure 3. Effect of Noise on Error Rate](image1)

![Figure 4. Effect of Noise on Error Factor](image2)

Figure 3 shows that the addition of 5% class noise causes the error rate for small disjuncts to increase, but from that point on it decreases as more noise is added. Figure 4 is even more insightful, since the error factor normalizes for the different overall error rates. This figure shows that as the amount of noise increases, the error factor for small disjuncts decreases. This suggests that as the noise level increases, a greater percentage of the errors may come from the larger disjuncts.

Figures 3 and 4 do not indicate how the distribution of small disjuncts changes as noise is added. One possibility is that as noise is added, the number of small disjuncts increases. This would make it possible for the percentage of errors contributed by small disjuncts to increase as noise is added, even though the error rate for these disjuncts doesn’t increase. Figures 5 through 7 address this.
Figure 5 shows that as noise is added to the chess domain, the number of cases covered by small disjuncts rises dramatically. Figure 6 shows when low to moderate levels of noise (5-20%) are added, the percentage of errors contributed by very small disjuncts increases, but this is not true for disjuncts with coverage >30 nor for higher levels of noise. Figure 7 shows a somewhat similar effect for the breast cancer domain. Thus, based on these two domains, one cannot generally conclude that as more noise is added, a greater percentage of the errors will come from the small disjuncts.

Figure 5. Effect of Noise on Distribution of Cases

Figure 6. Effect of Noise on Distribution of Errors (Chess Domain)

Figure 7. Effect of Noise on Distribution of Errors (Breast Cancer Domain)
As explained in section 2.2, for small disjuncts to be the cause of learning with noise being difficult, small disjuncts must account for a majority of the errors. In section 4 we saw that this is true for both domains when noise is not present. Figures 4 and 6 demonstrate that for the chess domain this still holds true for low to moderate levels of noise (5-20% noise), but not at higher levels. At these higher levels, disjuncts with coverage < 30 still cover about half of the errors, but they also now cover almost half of the total cases (figure 5). Thus, it appears that at low to moderate noise levels the inability to distinguish between true exceptions and noise may be responsible for making learning difficult, but at higher rates the effect of noise on the more prevalent large disjuncts dominates. For this domain, the crossover point occurs at around 20% class noise.

As figure 7 shows, small disjuncts play an even larger role in the breast cancer domain than in the chess domain (because the breast cancer domain is mainly populated by small disjuncts). However, the effect of noise on the error factor for this domain (not shown), also indicates that at high levels of noise the small disjuncts do not contribute a disproportionate percentage of the errors.

Figure 8 shows that pruning can be used to improve the error rate given class noise. Note that with n% random class noise applied to the test set, the optimal error rate is n%—you cannot learn to predict random noise. This optimal error rate provides a good basis against which to judge performance.

Note that pruning has little effect when there is no noise or 50% noise. For 10% or 20% noise pruning is very effective, but at 30% it has little impact. This is consistent with the results from figure 6, since one would expect pruning to be most effective when small disjuncts contribute most of the errors.

5.2 Noise Applied to Training or Test Sets

In the previous section, random noise was applied to both the training and test sets, since this best reflects what happens in the real world. Frequently, however, when the effects of noise are studied, noise is applied to the training set but not to the test set [Quinlan, 1986]. In this case, what is being studied is the ability to learn the "correct" concept (i.e., the noise free definition) when noise is present.
Noise can be thought of as having two distinct, albeit interacting, effects:

1. Noise applied to the training set perturbs the concept definition that is learned (this results in additional errors even if no noise is applied to the test set).

2. Noise applied to the test set causes additional errors (even if the correct concept definition is learned).

Applying noise to the training and test sets separately allows these two effects to be measured separately, which will lead to a clearer understanding of the problem of noise and small disjuncts. The first effect of noise is to introduce many small disjuncts (which are not true exceptions). This will have some effect on predictive accuracy. The second effect is clearly independent of disjunct size, which makes it less interesting for the purposes of this paper. However, the second effect is important since it causes the effect of noise on large disjuncts to dominate the effect of noise on small disjuncts, once some threshold level of noise is reached (assuming large disjuncts cover the majority of the examples).

Figures 9 and 10 show the error rates when noise is applied to different parts of the data.

---

4. For example, assume an arbitrary disjunct has an error rate of n%, with no noise applied to the test set. If we now add 100% class noise, and we have two classes, then the other class will now always be selected. The resulting error rate will be 100-n%, which is independent of the disjunct size. A similar analysis can be done for levels of class noise other than 100%.
Figure 9 empirically demonstrates that there are two effects, and that individually each is significantly smaller than the combined effect. Figure 10 shows that pruning dramatically decreases the error rate when noise is present in only the training set (the optimal error rate only applies when noise is present in the test set). In fact, for 0%, 5%, and 10% noise, the error rate is virtually constant. In contrast, under the same conditions, the experimental results for the breast cancer domain (not shown) show almost a 3% increase in error rate between the 0% and 5% noise levels—probably because more true exceptions are being pruned, since they are more prevalent in the breast cancer domain.

Figures 11 and 12 show the effect of pruning on each of the two effects of noise. Figure 11 shows that pruning has no effect when noise is present in only the test set (this is expected, since this effect of noise is independent of disjunct size). Figure 12 shows that pruning is very effective in combating noise on the training set, and as more noise is added, the relative merit of more aggressive pruning strategies increases.

![Figure 11. Effect of Pruning when Noise applied only to Test Set](image1)

![Figure 12. Effect of Pruning when Noise applied only to Training Set](image2)

In summary, there are two competing effects of noise: the effect of noise on training and the effect of noise on testing. For the former, small disjuncts are mainly affected; for the latter, all disjuncts are affected to the same degree. So, when these two factors are combined (noise is present in both the test and training sets), the effect on small disjuncts dominates only when the training component dominates—at low noise levels. If noise is only present in the training set, then small disjuncts play an even more central role—they are more likely to contribute the majority (and a disproportionate percentage) of the
total errors and the assertion that true exceptions in the concept definition make learning hard is more likely to hold.

This explains why pruning helps most when noise is either applied to only the training set, or to both sets but the level of noise is low— in these situations small disjuncts contribute most of the errors.

6. Next Steps and Future Research

One logical next step is to provide a better theoretical basis for the phenomena observed in this paper. This can be done by creating an artificial domain with two nearly identical concepts, which only differ in that one has no small disjuncts. The effects of noise on these two concepts can then be analyzed.

The experiments in this paper used C4.5 with training set sizes of 200 and with varying levels of random class noise. Future research should determine how learning with small disjuncts is affected by random and systematic attribute noise, systematic class noise, and training set size. In addition, the effect of using alternative overfitting avoidance strategies should be investigated, since some of the results in this paper depend on C4.5’s default pruning strategy ([Holte, et al., 1989] suggest using strategies which test both significance and error-rate).

Some further research has already been undertaken. [Weiss, 1995] examines the effect of systematic and random attribute noise as well as training set size on learning with small disjuncts, using two artificial domains.

7. Conclusion

This paper investigated the assertion that true exceptions make learning from noisy data difficult. For both the KPa7KR chess end-game domain and the Wisconsin breast cancer domain, the experimental results in this paper suggest that this assertion is true for low levels of class noise (where the effect of noise on the test set does not dominate). The assertion is even more likely to hold when the class noise is limited to the training set—the most realistic scenario, since it is the purpose of learning to predict the class of the test cases. However, these results need to be kept in perspective— the assertion was evaluated by examining the behavior of the small disjuncts, not by comparing the results of learning on a domain with true exceptions and then with the true exceptions removed (see [Weiss, 1995] for this).

This paper also showed some trends and effects which are likely to exist independent of a particular domain. The most important was that noise affects both training and testing, and that once a certain level of noise is reached, the effect of noise on testing will overwhelm the "problem with noise and small disjuncts".

8. Acknowledgements

Thanks to Andrea Danyluk, Foster Provost, Rob Holte and especially Haym Hirsh for many interesting discussions and helpful comments on this paper.
References


