Appears in *Proceedings of the 15th International Conference on Machine Learning*,
Morgan Kaufmann, 1998, 574-578.

The Problem with Noise and Small Disjuncts

Gary M. Weiss* and Haym Hirsh
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903
gmweiss@att.com, hirsh@cs.rutgers.edu

Abstract

Many systems that learn from examples express the learned concept as a disjunction. Those disjuncts that cover only a few examples are referred to as small disjuncts. The problem with small disjuncts is that they have a much higher error rate than large disjuncts but are necessary to achieve a high level of predictive accuracy. This paper investigates the effect of noise on small disjuncts. In particular, we show that when noise is added to two real-world domains, a significant, and disproportionate number of the total errors are contributed by the small disjuncts; thus, in the presence of noise, it is the small disjuncts that are primarily responsible for the poor predictive accuracy of the learned concept.

1 INTRODUCTION

Systems that learn from examples often express the learned concept as a disjunction. The coverage, or size, of each disjunct is defined as the number of training examples that it correctly classifies (Holte, Acker & Porter, 1989). Small disjuncts are those disjuncts that cover only a few training examples. Although small disjuncts may individually cover only a small fraction of the training examples, collectively they can cover a significant percentage of the training examples. The *problem with small disjuncts* is that they have a higher error rate than large disjuncts but cannot be eliminated without greatly reducing the predictive accuracy of the learned concept.

Early work on small disjuncts investigated a variety of issues, including ways of improving predictive accuracy by eliminating some small disjuncts (Holte, et al., 1989; Quinlan, 1991). Danyluk and Provost (1993) highlighted the role of small disjuncts in learning from noisy data when they speculated that in the telecommunication domain they were studying, learning from noisy data was hard due to a difficulty distinguishing between systematic noise and "true" exceptional cases in the training data. True exceptions and small disjuncts, although similar entities which are sometimes used interchangeably, differ in one important way—true exceptions are defined relative to the "true" (i.e., correct) concept whereas small disjuncts are defined relative to a learned concept. Weiss (1995) investigated the interaction of noise on true exceptions by using artificial datasets and demonstrated that this interaction results in error prone small disjuncts in the learned concept. In this paper we focus on small disjuncts rather than "true exceptions" because for the real world domains we use, the "correct" concept definition is not known, and hence it is not possible to measure the true exceptions.

This paper extends previous work by examining the effect of noise on small disjuncts using real-world datasets and assessing the impact of this effect on the overall learning process. In particular, we show that when noise is added to these datasets, then the concept learned from this data exhibits the *problem with noise and small disjuncts*; that is, the small disjuncts contribute a disproportionate, and significant, number of the total errors (relative to the number of examples they cover) but still cannot be eliminated without adversely affecting the accuracy of the learned concept. Thus, we show that the small disjuncts are primarily responsible for learning being difficult in the presence of noise.

2 DESCRIPTION OF EXPERIMENTS

This section describes the learning program, problem domains and experimental methodology we used to conduct our experiments.

2.1 THE LEARNER

All of the experiments described in this paper use C4.5, a program for inducing decision trees from preclassified training examples (Quinlan, 1993). C4.5 was chosen because it is a popular tool for learning disjunctive

*Also AT&T Labs, Middletown, NJ 07748

concepts and because we were able to modify it, without too much difficulty, to collect statistics relating to disjunct size. For the majority of experiments, C4.5 was run in one of the following two configurations:

— with its default parameters and pruning strategy, and

— with its default parameters but without any pruning and with the –m1 option to disable the default stopping criterion.

The –m option stops a node from being split during the tree-building process if the resulting node covers fewer than the specified number of examples (1 in this case). Thus, in the second configuration, C4.5 will build a decision tree that correctly classifies all training examples if the examples are consistent.

2.2 THE PROBLEM DOMAINS

This paper uses the KPa7KR chess endgame (Shapiro, 1987) and Wisconsin breast cancer (Wolberg, 1990) datasets, which were obtained from the UCI repository of machine learning databases (Murz & Murphy, 1998). These datasets were selected because C4.5 was able to attain high levels of predictive accuracy on them; we wanted to come as close to learning the correct target concept as possible prior to the introduction of artificial noise. The KPa7KR dataset contains 3196 examples with 36 attributes, where each example represents a board position and has the class value "won" or "nowin". The Wisconsin breast cancer dataset contains 699 examples with nine attributes, with each example having the value "benign" or "malignant". The class distribution is approximately equal for the chess endgame domain and is 2:1 in favor of the benign class for the breast cancer domain. The results for the breast cancer domain closely parallel those for the chess domain and therefore in most cases we only display the results for the chess domain (all results are for the chess domain unless noted otherwise).

2.3 EXPERIMENTAL METHODOLOGY

For each experiment seven independent runs were performed and the results averaged together. For each run, 200 examples were randomly selected and placed into the training set while the remaining examples were placed into the test set. Unless stated otherwise, all measurements are based on the performance of the test set. Varying levels of randomly generated *class* noise are used in the experiments. The examples are considered initially noise-free. A noise level of n% means that with probability n/100 the class value is randomly selected from the *remaining* alternatives. This means that when 50% class noise is applied to a domain with two classes, there is no information provided by the class variable.

For the experiments performed in this paper, coverage is defined in terms of the number of test examples correctly classified, since we felt that this would yield a more fair measure of the true coverage of each disjunct (just as measuring accuracy on the test set yields a more fair measure). However, we do not believe this decision to be critical. For each graph presented in this paper, coverage is displayed on a logarithmic scale, so the behavior of the small disjuncts can be easily identified.

3 THE PROBLEM WITH SMALL DISJUNCTS

Although the focus of this paper is on the problem with noise and small disjuncts, this section will first show that the chess endgame and breast cancer domains exhibit the problem with small disjuncts. Figures 1 and 2 show the results of running C4.5 on the chess endgame and Wisconsin breast cancer domains, respectively, without any artificial noise applied to the datasets. For these figures, and for all figures in this paper with coverage on the x-axis, the value of each curve at coverage *n* is based on the *collective performance* of all the disjuncts with coverage less than or equal to *n*. Thus, the curves labeled "Examples" and "Errors" in Figures 1 and 2 show the percentage of total examples and errors, respectively, covered by these disjuncts (i.e., with size ≤ n) when the learned concept is applied to the test set. The error rate curve shows the error rate of the disjuncts with size ≤ n.

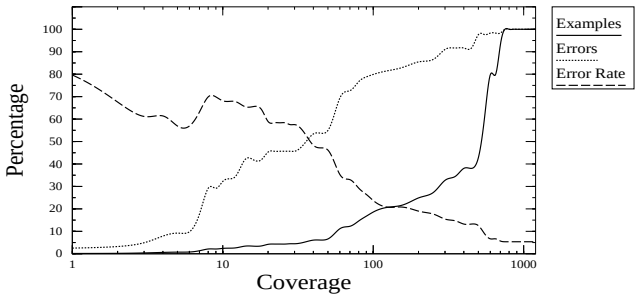

Figure 1: The Effect of Disjunct Size (Chess Domain)

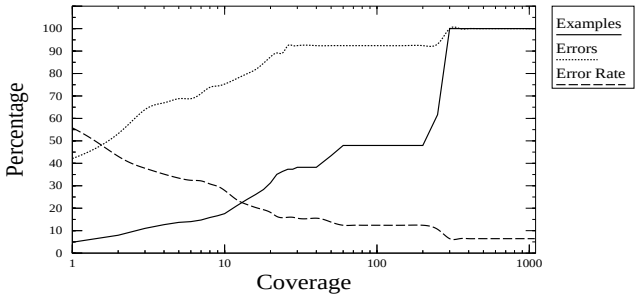

Figure 2: The Effect of Disjunct Size (Cancer Domain)

An example will help clarify the meanings of these curves and demonstrate that small disjuncts are "error prone". In Figure 1, the curves for errors and error rate intersect at coverage 40. The curves tell us that the disjuncts with size ≤ 40 collectively have an error rate of 50% and collectively cover 50% of the total errors, but only cover 5% of the total examples. This clearly demonstrates that small disjuncts are error prone (i.e., they cover a disproportionate number of errors). The error rate for the learner as a whole can be found by

looking at the error rate when 100% of the errors and examples have been covered; we see from this that the overall error rate for the chess endgame domain is 5% and the overall error rate for the breast cancer domain is 6%. The error rate curve also shows that small disjuncts have a higher error rate than large disjuncts, since the error rate decreases (for both domains) as larger disjuncts are included in the error rate calculations.

Figures 1 and 2 show that most examples are covered by the larger disjuncts, but the smaller disjuncts nonetheless cover a large percentage of the examples. This is more evident for the breast cancer domain, but even for the chess endgame domain disjuncts of size ≤ 100 are much more error prone than the larger disjuncts and cover about 20% of the total examples. These results are consistent with those described by Holte and colleagues (1989). In addition, since the small disjuncts cover too many examples to be simply dropped from the learned concept without significantly impacting the accuracy of the concept, these results also demonstrate that these domains exhibit the problem with small disjuncts.

4 THE PROBLEM WITH NOISE AND SMALL DISJUNCTS

This section will show that for the chess and breast cancer domains, noise results in small disjuncts being mainly responsible for the errors in the learned concept. For these experiments, no pruning is done unless specified and class noise is applied to both the training and test sets.

Figure 3 shows what happens to the error rate as the noise rate is varied (recall that for coverage of *n*, the "collective" error rate is based on all disjuncts with size ≤*n*). The figure 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.

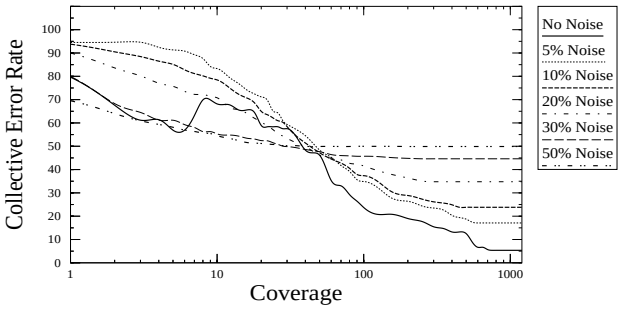

Figure 3: Effect of Noise on Error Rate

To make it easier to see the degree to which errors are concentrated toward the small disjuncts, we will use a statistic called the error factor, first introduced by Weiss (1995). The error factor is defined as:

$$Error\ Factor(cov) \equiv \frac{\%\ cumulative\ errors(cov)}{\%\ cumulative\ examples(cov)}$$

The error factor is a function of coverage and is essentially the "Errors" curve divided by the "Examples" curve. For example, the error factor at coverage 40 in Figure 1 is 10 (50%/5%), which indicates that disjuncts with size ≤ 40 contribute 10 times more errors than expected if coverage had no effect on error rate.

Figure 4, which plots the error factor versus coverage, shows the effect of noise on small disjuncts even more clearly than Figure 3, since error factor is a relative measure which takes into account the different overall error rates resulting from learning with the different levels of class noise. Figure 4 shows that as the amount of noise increases the error factor for small disjuncts decreases. This indicates that as the noise level increases either the percentage of errors contributed by the small disjuncts decreases and/or the percentage of examples covered by the small disjuncts increases.

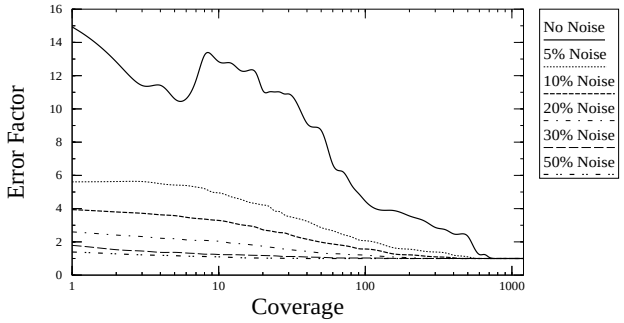

Figure 4: Effect of Noise on Error Factor

Noise added to the training data will undoubtably affect the concept that is learned and will therefore affect the small disjuncts in the learned concept. Figure 5 addresses this by showing how various noise levels affect the number of examples covered by the small disjuncts.

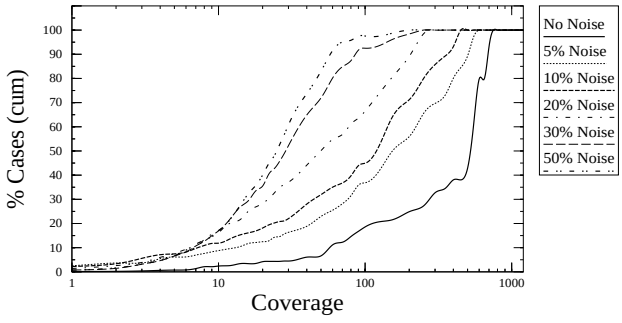

Figure 5: Effect of Noise on Distribution of Cases

Figure 5 shows that as more noise is added to the data, the number of examples covered by small disjuncts increases dramatically. For example, disjuncts of size ≤ 100 cover 3 times as many examples when the noise level increases from no noise to 10% noise. Figure 5 confirms what we and others had suspected—that noisy data will cause a learner to form "erroneous" small disjuncts.

Figure 6 shows how the distribution of errors changes as noise is applied to the domain. It shows that when the noise level is less than 20%, small disjuncts with size ≤ 30 account for an even greater percentage of the total errors than when there was no noise. Thus, we now have an explanation of why the error factor in Figure 4 decreased as additional noise was introduced—it was because the number of examples covered by the small disjuncts increased at a faster rate than the number of errors contributed by these disjuncts. Note that once the noise level reaches 30%, then disjuncts with coverage ≤ 30 no longer cover a disproportionate number of the errors—they cover half of the errors but also cover almost half of the total examples. The breast cancer domain exhibits similar trends.

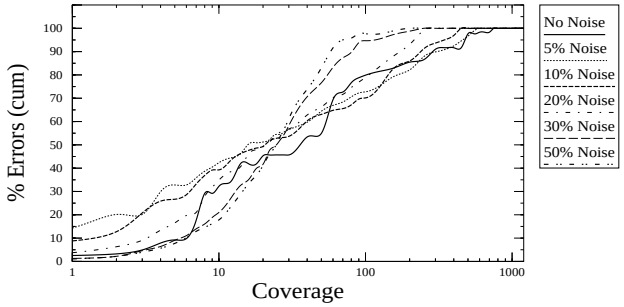

Figure 6: Effect of Noise on Distribution of Errors

We can summarize the results from Figures 3–6 as follows: in the presence of noise, small disjuncts have a higher error rate than large disjuncts and cover a significant number of the total cases and total errors. As a consequence, small disjuncts contribute a disproportionate and very significant number of the errors. All of this holds true until very high levels of noise are applied, at which point the impact of noise on the large disjuncts becomes important relative to the impact of noise on small disjuncts—at which point small disjuncts can no longer be blamed for the poor performance of the learned concept.

Since overfitting avoidance strategies such as pruning are more likely to eliminate small disjuncts than large disjuncts, it is interesting to see how these strategies will affect the error rate and how this can be related to the role of small disjuncts. Figure 7 shows how pruning affects the overall error rate. Since it is not possible to predict random class noise, the optimal error rate will equal the noise rate. This figure shows that the default pruning strategy improves the error rate in the presence of class noise and improves it the most when the noise rate is between 10% and 20%. This is explained by the fact that in this range the small disjuncts have very high error rates (Figure 3) and contribute a very large percentage of the total errors (Figure 6). The strategy which uses C4.5's -m20 option to prevent nodes from being formed when fewer than 20 examples are covered also improves the error rate, except when there is no noise. This strategy also outperforms the default pruning strategy when there are very high levels of noise (e.g., 30%), indicating that in such cases a very aggressive overfitting avoidance strategy is needed to adequately learn the correct concept.

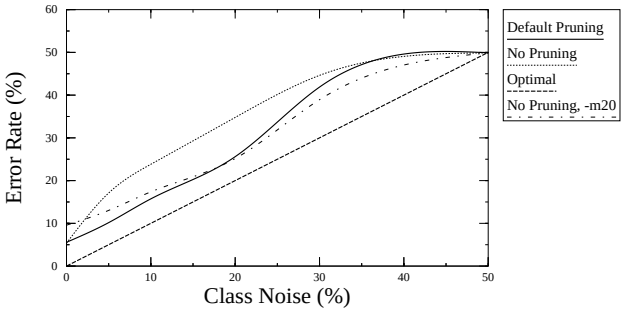

Figure 7: Effect of Pruning on Overall Error Rate

5 UNDERSTANDING THE EFFECT OF NOISE ON SMALL DISJUNCTS

In the experiments described in the previous section, the training and tests sets were generated from the same distribution. While this is the most realistic scenario, when one is trying to *understand* the effect of noise on learning, noise is frequently only applied to either the training or test set.

5.1 THE EFFECT ON TRAINING

Noise applied only to the training set tests the ability to learn the "correct" concept in the presence of noise (Quinlan, 1986). That is, by limiting the noise to the training set, we can evaluate the sensitivity of the learner to noise. We can accomplish this evaluation, even without knowing the "correct" concept, by using the noise-free test data to approximate the correct concept.

As shown earlier, noise in the training set introduces additional "erroneous" small disjuncts into the learned concept. Experiments identical to those described earlier were repeated with the artificial noise restricted to the training set. Graphs corresponding to those shown in Figures 3–6 were generated. The results indicated that under these circumstances small disjuncts have an even more significant impact on learning and, in particular, contribute a greater percentage of the errors than when noise was applied to both the training and test sets.

5.2 THE EFFECT ON TESTING

It is also meaningful to study the effect of noise on the test set. This situation corresponds to the scenario in which the training data is "cleaned up", perhaps by using more costly measurement equipment, in the hope of achieving improved predictive accuracy.[1] Experiments in which the noise was limited to the test set were run and the results showed that, relative to the case where noise

[1] However, if systematic noise is applied to the test set, better predictive accuracy may be obtained by leaving the noise in the training set.

was applied to both the training and test sets, the small disjuncts had much less of a negative impact on learning.

5.3 DISCUSSION

The results described in the previous two subsections can be explained by examining how noise affects small disjuncts. First of all, noise in the training set will influence the concept that is learned but noise in the test set cannot. Since small disjuncts are based on the learned concept, we can conclude that noise in the test set cannot cause small disjuncts to be formed. Futhermore, noise in the test set will tend to affect all disjuncts equally (Weiss, 1995). This explains why the effect of noise on small disjuncts is less dramatic when noise is applied to both the training and test sets than when it is limited to the training set—in the former case noise in the test set reduces the relative difference in error rates between the small and large disjuncts. When noise is applied to only the test set, the effect is greatly diminished, and would disappear completely if the learner were able to learn the correct concept prior to the introduction of artificial noise. For a more in depth description about *how* noise affects small disjuncts, refer to Weiss (1995).

6 CONCLUSION

This paper investigated the effect of noise on small disjuncts and how this effect impacts the overall learning process. For both the KPa7KR chess end-game domain and the Wisconsin breast cancer domain, the experimental results in this paper show that small disjuncts are responsible for learning being difficult. Only at very high levels of class noise do the large disjuncts contribute a relatively large percentage of the total errors. This paper also showed some trends and effects that we feel are likely to hold for learning in general and not just for the two domains used in this paper. In particular, we feel that 1) noise tends to decrease the number of large disjuncts and increase the number of small disjuncts in the learned concept, 2) relatively low levels of noise will increase the percentage of errors contributed by small disjuncts, but this effect will diminish as higher levels of noise are applied, and 3) noise in the test set has an equalizing effect which decreases the impact of the small disjuncts on learning.

We believe these results are important because they provide some insight into *how* noise affects learning and how the effect of noise manifests itself in the learned concept. Given the prevalence of noise in real-world problem domains, such an understanding is critical. This work also provides additional justification for overfitting avoidance strategies and hopefully provides some additional insights into why these strategies work, how they can be improved and the limitations of such strategies.

Acknowledgements

Thanks to Andrea Danyluk, Foster Provost and Rob Holte for helpful comments and interesting discussions on the role of small disjuncts in learning. The authors would also like to thank the members of the Rutgers Machine Learning Research Group for their many constructive comments.

References

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.

Merz, C. J., & Murphy, P. M. (1998). *The UCI Repository of machine learning databases* [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

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), 93-98.

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

Shapiro, A. D. (1987). *Structured Induction in Expert Systems*, Addison-Wesley.

Weiss, G. M. (1995). Learning with rare cases and small disjuncts, In *Machine Learning: Proceedings of the Twelfth International Conference*, 558-565, San Francisco, CA: Morgan Kaufmann.

Wolberg, W. H. & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences*, U.S.A. (Vol 87), 9193-9196.