MINING WITH RARE CASES

Gary M. Weiss
Department of Computer and Information Science
Fordham University

Abstract: Rare cases are often the most interesting cases. For example, in medical diagnosis one is typically interested in identifying relatively rare diseases, such as cancer, rather than more frequently occurring ones, such as the common cold. In this chapter we discuss the role of rare cases in data mining. Specific problems associated with mining rare cases are discussed, followed by a description of methods for addressing these problems.

Key words: Rare cases, small disjuncts, inductive bias, sampling.

1. INTRODUCTION

Rare cases are often of special interest. This is especially true in the context of data mining, where one often wants to uncover subtle patterns that may be hidden in massive amounts of data. Examples of mining rare cases include learning word pronunciations (Van den Bosch, Weijters, Van den Herik & Daelmans, 1997), detecting oil spills from satellite images (Kubat, Holte & Matwin, 1998), predicting telecommunication equipment failures (Weiss & Hirsh, 1998) and finding associations between infrequently purchased supermarket items (Liu, Hsu & Ma, 1999). Rare cases warrant special attention because they pose significant problems for data mining algorithms.

We begin by discussing what is meant by a rare case. Informally, a *case* corresponds to a region in the instance space that is meaningful with respect to the domain under study and a *rare case* is a case that covers a small region of the instance space and covers relatively few training examples. As a concrete example, with respect to the class *bird*, *non-flying bird* is a rare case since very few birds (e.g., ostriches) do not fly. Figure 1 shows rare

cases and common cases for unlabeled data (Figure 1a) and for labeled data (Figure 1b). In each situation the regions associated with each case are outlined. Unfortunately, except for artificial domains, the borders for rare and common cases are not known and can only be approximated.

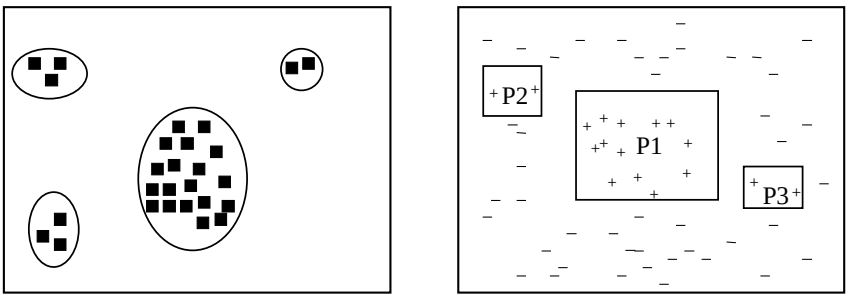

Figure -1. Rare and common cases in unlabeled (a) and labeled (b) data

One important data mining task associated with unsupervised learning is *clustering*, which involves the grouping of entities into categories. Based on the data in Figure 1a, a clustering algorithm might identify four clusters. In this situation we could say that the algorithm has identified one common case and three rare cases. The three rare cases will be more difficult to detect and generalize from because they contain fewer data points. A second important unsupervised learning task is *association rule mining*, which looks for associations between items (Agarwal, Imielinski & Swami, 1993). Groupings of items that co-occur frequently, such as *milk* and *cookies*, will be considered common cases, while other associations may be extremely rare. For example, *mop* and *broom* will be a rare association (i.e., case) in the context of supermarket sales, not because the items are unlikely to be purchased together, but because neither item is frequently purchased in a supermarket (Liu, et al., 1999).

Figure 1b shows a *classification problem* with two classes: a positive class *P* and a negative class *N*. The positive class contains one common case, *P1*, and two rare cases, *P2* and *P3*. For classification tasks the rare cases may manifest themselves as *small disjuncts*. Small disjuncts are those disjuncts in the *learned* classifier that cover few training examples (Holte, Acker & Porter, 1989). If a decision tree learner were to form a leaf node to cover case *P2*, the disjunct (i.e., leaf node) will be a small disjunct because it covers only two training examples. Because rare cases are not easily identified, most research focuses on their learned counterparts—small disjuncts.

Existing research indicates that rare cases and small disjuncts pose difficulties for data mining. Experiments using artificial domains show that rare cases have a much higher misclassification rate than common cases (Weiss, 1995; Japkowicz, 2001), a problem we refer to as the problem with rare cases. A large number of studies demonstrate a similar problem with small disjuncts. These studies show that small disjuncts consistently have a much higher error rate than large disjuncts (Ali & Pazzani, 1995; Weiss, 1995; Holte, et al., 1989; Ting, 1994; Weiss & Hirsh, 2000). Most of these studies also show that small disjuncts collectively cover a substantial fraction of all examples and cannot simply be eliminated—doing so will substantially degrade the performance of a classifier. The most thorough empirical study of small disjuncts showed that, in the classifiers induced from thirty real-world data sets, most errors are contributed by the smaller disjuncts (Weiss & Hirsh, 2000).

One important question to consider is whether the rarity of a case should be determined with respect to some absolute threshold number of training examples ("absolute rarity") or with respect to the relative frequency of occurrence in the underlying distribution of data ("relative rarity"). If we use absolute rarity, then if a rare case covers only three examples from a training set, then it should be considered rare. However, if additional training data are obtained so that the training set increases by factor of 100, so that this case now covers 300 examples, then absolute rarity says this case is no longer a rare case. However, if the case covers only 1% of the training data in both situations, then relative rarity would say it is rare in both situations. From a practical perspective we are concerned with both absolute and relative rarity since, as we shall see, both forms of rarity pose problems for virtually all data mining systems.

This chapter focuses on rare cases. In the remainder of this chapter we discuss problems associated with mining rare cases and techniques to address these problems. Rare *classes* pose similar problems to those posed by rare cases and for this reason we comment on the connection between the two at the end of this chapter. A comprehensive discussion of rare classes and, more generally, class imbalance, is provided in Chapter **XX**.

2. WHY RARE CASES ARE PROBLEMATIC

Rare cases pose difficulties for data mining systems for a variety of reasons. The most obvious and fundamental problem is the associated lack of data—rare cases tend to cover only a few training examples (i.e., absolute rarity). This lack of data makes it difficult to detect rare cases and, even if the rare case is detected, makes generalization difficult since it is hard to

identify regularities from only a few data points. To see this, consider the classification task shown in Figure 2, which focuses on the rare case, *P3*, from Figure 1b. Figure 2a reproduces the region from Figure 1b surrounding *P3*. Figure 2b shows what happens when the training data is augmented with only positive examples while Figure 2c shows the result of adding examples from the underlying distribution.

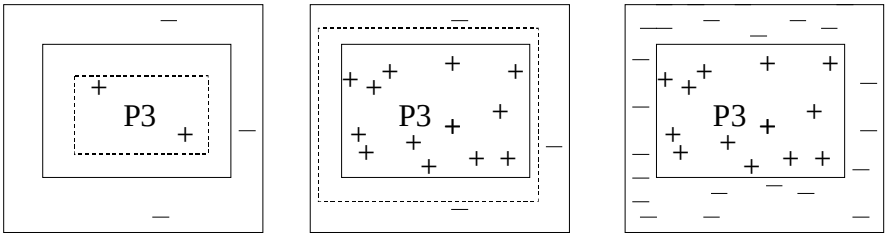

Figure -2. The problem with absolute rarity

The learned decision boundaries are displayed in Figure 2a and Figure 2b using dashed lines. The learned boundary in Figure 2a is far off from the "true" boundary and excludes a substantial portion of *P3*. The inclusion of additional positive examples in Figure 2b addresses the problem with absolute rarity and causes all of *P3* to be covered/learned—although some examples not belonging to *P3* will be mistakenly assigned a positive label. Figure 2c, which includes additional positive and negative examples, corrects this last problem (the learned decision boundary nearly overlaps the true boundary and hence is not shown). Figures 2b and 2c demonstrate that additional data can address the problem with absolute rarity. Of course, in practice it is not always possible to obtain additional training data.

Another problem associated with mining rare cases is reflected by the phrase: *like a needle in a haystack*. The difficulty is not so much due to the needle being small—or there being only one needle—but by the fact that the needle is obscured by a huge number of strands of hay. Similarly, in data mining, rare cases may be obscured by common cases (relative rarity). This is especially a problem when data mining algorithms rely on greedy search heuristics that examine one variable at a time, since rare cases may depend on the conjunction of many conditions and any single condition in isolation may not provide much guidance. As a specific example of the problem with relative rarity, consider the association rule mining problem described earlier, where we want to be able to detect the association between *mop* and *broom*. Because this association occurs rarely, this association can only be found if the minimum support (*minsup*) threshold, the number of times the

association is found in the data, is set very low. However, setting this threshold low would cause a combinatorial explosion because frequently occurring items will be associated with one another in an enormous number of ways (most of which will be random and/or meaningless). This is called the *rare item problem (*Liu, et al., 1999).

The metrics used during data mining and to evaluate the results of data mining can also make it difficult to mine rare cases. For example, because a common case covers more examples than a rare case, classification accuracy will cause classifier induction programs to focus their attention more on common cases than on rare cases. As a consequence, rare cases may be totally ignored. Furthermore, consider the manner in which decision trees are induced. Most decision trees are grown in a top-down manner, where test conditions are repeatedly evaluated and the best one selected. The metrics (e.g., information gain) used to select the best test generally prefer tests that result in a balanced tree where purity is increased for most of the examples, over a test that yields high purity for a relatively small subset of the data but low purity for the rest (Riddle, Segal & Etzioni, 1994). Thus, rare cases, which correspond to high purity branches covering few examples will often not be included in the decision tree. The problem is even easier to understand for association rule mining, since rules that do not cover at least *minsup* examples will never be considered.

The bias of a data mining system is critical to its performance. This extra-evidentiary bias makes it possible to generalize from specific examples. Unfortunately, the bias used by most data mining systems impacts their ability to mine rare cases. This is because many data mining systems, especially those used to induce classifiers, employ a maximum-generality bias (Holte, et al., 1989). This means that when a disjunct that covers some set of training examples is formed, only the most general set of conditions that satisfy those examples are selected. This can be contrasted with a maximum-specificity bias, which would add all possible, shared, conditions. The maximum-generality bias will work well for common cases/large disjuncts but does not work well for rare cases/small disjuncts. This leads to the problem with small disjuncts described earlier. Attempts to address the problem of small disjuncts by carefully selecting the bias of the learner are described in Section 3.2.

Noisy data may also make it difficult to mine rare cases, since, given a sufficiently high level of background noise, a learner may not be able to distinguish between true exceptions (i.e., rare cases) and noise-induced ones (Weiss, 1995). To see this, consider the rare case, *P3*, in Figure 1b. Because *P3* contains so few training examples, if attribute noise causes even a few negative examples to appear within *P3*, this would prevent *P3* from being learned correctly. However, common cases such as *P1* are not nearly as

susceptible to noise. Unfortunately, there is not much that can be done to minimize the impact of noise on rare cases. Pruning and other overfitting avoidance techniques—as well as inductive biases that foster generalization—can minimize the overall impact of noise, but, because these methods tend to remove both the rare cases and "noise-generated" ones, they do so at the expense of the rare cases.

3. TECHNIQUES FOR HANDLING RARE CASES

A number of techniques are available to address the issues with rare cases described in the previous section. We describe only the most popular techniques.

3.1 Obtain Additional Training Data

Obtaining additional training data is the most direct way of addressing the problems associated with mining rare cases. However, if one obtains additional training data from the original distribution, then most of the new data will be associated with the common cases. Nonetheless, because some of the data will be associated with the rare cases, this approach may help with the problem of "absolute rarity". However, this approach does not address the problem of relative rarity at all, since the same proportion of the training data will cover common cases. Only by *selectively* obtaining additional training data for the rare cases can one address the issues with relative rarity (such a sampling scheme would also be quite efficient at dealing with absolute rarity). Japkowicz (2001) applied this non-uniform sampling approach to artificial domains and demonstrated that it can be very beneficial. Unfortunately, since one can only identify rare cases for artificial domains, this approach generally cannot be implemented and has not been used in practice.[1] However, based on the assumption that small disjuncts are the manifestation of the rare cases in the learned classifier, this approach can be approximated by preferentially sampling examples that fall into the small disjuncts of some initial classifier. This approach warrants additional research.

[1] Because rare *classes* are trivial to identify, it is straightforward to increase the proportion of rare classes in the training data. Thus this approach is routinely used to address the problem with relative rarity for rare classes.

3.2 Use a More Appropriate Inductive Bias

Rare cases tend to cause error-prone small disjuncts to be formed in a classifier induced from labeled data (Weiss, 1995). As discussed earlier, the error prone nature of small disjuncts is at least partly due to the bias used by most learners. Simple strategies that eliminate all small disjuncts or use statistical significance testing to prevent small disjuncts from being formed perform poorly (Holte et al., 1989). A number of studies have investigated more sophisticated approaches for adjusting the bias of a learner in order to minimize the problem with small disjuncts.

Holte et al. (1989) modified CN2 so that its maximum generality bias is used only for large disjuncts. A maximum specificity bias was then used for small disjuncts. This was shown to improve the performance of the small disjuncts but degrade the performance of the large disjuncts, yielding poorer overall performance. This occurred because the "emancipated" examples—those that would previously have been classified by small disjuncts—were then misclassified at an even higher rate by the large disjuncts. Going on the assumption that this change in bias was too extreme, a selective specificity bias was then evaluated. This yielded further improvements, but not enough to improve overall classification accuracy.

This approach was subsequently refined to ensure that the more specific bias used to induce the small disjuncts does not affect—and therefore cannot degrade—the performance of the large disjuncts. This was accomplished by using different learners for examples that fall into large disjuncts and examples that fall into small disjuncts (Ting, 1994). While the results of this study are encouraging and show that this hybrid approach can improve the accuracy of the small disjuncts, the results were not conclusive. Carvalho & Freitas (2002a, 2002b) use essentially the same approach, except that the set of training examples falling into each individual small disjunct are used to generate a separate classifier.

A final study advocates the use of instance-based learning for domains with many rare cases/small disjuncts, because of the highly specific bias associated with this learning method (Van Den Bosch, 1997). The authors of this study were mainly interested in learning word pronunciations, which, by their very nature, have "pockets of exceptions" (i.e., rare cases) that cause many small disjuncts to be formed during learning. Results are not provided to demonstrate that instance-based learning outperforms others learning methods in this situation. Instead the authors argue that instance-based learning methods should be used because they store all examples in memory, while other approaches ignore examples when they fall below some utility threshold (e.g., due to pruning).

In summary, several attempts have been made to perform better on rare cases by using a highly specific bias for the induced small disjuncts. These methods have shown only mixed success. We view this approach to addressing rarity to be promising and worthy of future investigation.

3.3 Using More Appropriate Metrics

Data mining can better handle rare cases by using evaluation metrics that, unlike accuracy, do not discount the importance of rare cases. These metrics can then better guide the data mining process and better evaluate the results of data mining. Precision and recall are metrics from the information retrieval community that have been used to mine rare cases. Given a classification rule *R* that predicts target class *C*, the recall of *R* is the percentage of examples belonging to *C* that are correctly identified while the precision of *R* is the percentage of times the rule is correct. Rare cases can be given more prominence by increasing the importance of precision over recall. Timeweaver (Weiss, 1999), a genetic-algorithm based classification system, searches for rare cases by carefully altering the relative importance of precision versus recall. This ensures that a diverse population of classification rules is developed, which leads to rules that perform well with respect to precision, recall, or both. Thus, precise rules that cover rare cases will be generated.

Two-phase rule induction is another approach that utilizes precision and recall. This approach is motivated by the observation that it is very difficult to optimize precision and recall simultaneously—and trying to do so will miss rare cases. PNrule (Joshi, Agarwal & Kumar, 2001) uses two-phase rule induction to focus on each measure separately. In the first phase, if high precision rules cannot be found then lower precision rules are accepted, as long as they have relatively high recall. So, the first phase focuses on recall. In the second phase precision is optimized. This is accomplished by learning to identify false positives within the rules from phase 1. Returning to the needle and haystack analogy, this approach identifies regions likely to contain needles in the first phase and then learns to discard the hay strands within these regions in the second phase. Two-phase rule induction deals with rare cases because the first phase is sensitive to the problem of small disjuncts while the second phase allows the false positives to all be grouped together, making it easier to identify the false positives. Experimental results indicate that PNrule performs competitively with other disjunctive learners on easy problems and is able to maintain its high performance as more complex concepts with many rare cases are introduced—something the other learners cannot do.

3.4 Employ Non-Greedy Search Techniques

Most data mining algorithms are greedy in the sense that they make locally optimal decisions without regard to what may be best globally. This is done to ensure that the data mining algorithms are tractable. However, because rare cases may depend on the conjunction of many conditions and any single condition in isolation may not provide much guidance, such greedy methods are often ineffective when dealing with rare cases. Thus, one approach for handling rare cases is to use more powerful, global, search methods. Genetic algorithms, which operate on a population of candidate solutions rather than a single solution, fit this description and cope well with attribute interactions (Goldberg, 1989). For this reason genetic algorithms are being increasingly used for data mining (Freitas, 2002) and several systems have used genetic algorithms to handle rare cases. In particular, Timeweaver (Weiss, 1999) uses a genetic algorithm to predict very rare events and Carvalho and Freitas (2002a, 2002b) use a genetic algorithm to "discover small disjunct rules".

More conventional learning methods can also be adapted to better handle rare cases. For example, Brute (Riddle, et al., 1994) is a rule-learning algorithm that performs an exhaustive depth-bounded search for accurate conjunctive rules. The goal is to find accurate rules, even if they cover relatively few training examples. Brute performs quite well when compared to other algorithms, although the lengths of the rules needs to be limited to make the algorithm tractable. Brute is capable of locating "nuggets" of information that other algorithms may not be able to find.

Association-rule mining systems generally employ an exhaustive search algorithm (Agarwal et al., 1993). However, while these algorithms are in theory capable of finding rare associations, they become intractable if the minimum level of support, *minsup*, is set small enough to find rare associations. Thus, such algorithms are heuristically inadequate for finding rare associations and suffer from the rare item problem described earlier. This problem has been addressed by modifying the standard Apriori algorithm so that it can handle multiple minimum levels of support (Liu et al., 1999). Using this approach, the user specifies a different minimum support for each item, based on the frequency of the item in the distribution. The minimum support for an association rule is then the lowest *minsup* value amongst the items in the rule. Empirical results indicate that these enhancements permit the modified algorithm to find meaningful associations involving rare items, without producing a huge number of meaningless rules involving common items.

3.5 Utilize Knowledge/Human Interaction

Knowledge, while generally useful when data mining, is especially useful when rare cases are present. Knowledge can take many forms. For example, an expert's domain knowledge, including knowledge of how variables interact, can be used to generate sophisticated features capable of identifying rare cases (most experts naturally tend to identify features that are useful for predicting rare, but important, cases). Knowledge can also be applied interactively during data mining to help identify rare cases. For example, in association rule mining a human expert may indicate which preliminary results are interesting and warrant further mining and which are uninteresting and should not be pursued. At the end of the data mining process the expert can also help distinguish between meaningful rare cases and spurious correlations.

3.6 Employ Boosting

Boosting algorithms, such as AdaBoost, are iterative algorithms that place different weights on the training distribution each iteration (Schapire, 1999). Following each iteration boosting increases the weights associated with the incorrectly classified examples and decreases the weights associated with the correctly classified examples. This forces the learner to focus more on the incorrectly classified examples in the next iteration. Because rare cases are difficult to predict, it is reasonable to believe that boosting will improve their classification performance. A recent study showed that boosting can help with rarity if the base learner can effectively trade-off precision and recall (Joshi et al., 2002). An algorithm, RareBoost (Joshi, Kumar & Agarwal, 2001), has been developed that modifies the standard boosting weight-update mechanism to improve the performance of rare classes and rare cases.

3.7 Place Rare Cases Into Separate Classes

Rare cases complicate classification tasks because different rare cases may have little in common between them, making it difficult to assign the same class value to all of them. One possible solution is to reformulate the original problem so that the rare cases are viewed as separate classes. The general approach is to 1) separate each class into subclasses using clustering (an unsupervised learning technique) and then 2) learn after re-labeling the training examples with the new classes (Japkowicz, 2002). Because multiple clustering experiments were used in step 1, step 2 involves learning multiple models, which are subsequently combined using voting. The performance

results from this study are promising, but not conclusive, and additional research is needed.

4. CONCLUSION

This chapter describes various problems associated with mining rare cases and methods for addressing these problems. While a significant amount of research on rare cases is available, much of this work is still in its infancy. That is, there are no well-established, proven, methods for generally handling rare cases. We expect research on this topic to continue, and accelerate, as increasingly more difficult data mining tasks are tackled.

This chapter covers rare cases. Rare classes, which result from highly skewed class distributions, share many of the problems associated with rare cases. Furthermore, rare cases and rare classes are connected. First, while rare cases can occur within both rare classes and common classes, we expect rare cases to be more of an issue for rare classes (e.g., rare classes will never have any very common cases). A study by Weiss & Provost (2003) confirms this connection by showing that rare classes tend to have smaller disjuncts than common classes (small disjuncts are assumed to indicate the presence of rare cases). Other research shows that rare cases and rare classes can also be viewed from a common perspective. Japkowicz (2001) views rare classes as a consequence of between-class imbalance and rare cases as a consequence of within-class imbalances. Thus, both forms of rarity are a type of data imbalance. Recent work further demonstrates the similarity between rare cases and rare classes by showing that they introduce the same set of problems and that these problems can be addressed using the same set of techniques (Weiss, 2004). More intriguing still, some research indicates that rare classes per se are not a problem, but rather it is the rare cases within the rare classes that are the fundamental problem (Japkowicz, 2001).

REFERENCES

Agarwal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data; 1993.

Ali, K., Pazzani, M. HYDRA-MM: learning multiple descriptions to improve classification accuracy. International Journal of Artificial Intelligence Tools 1995; 4.

Carvalho, D. R., Freitas, A. A. A genetic algorithm for discovering small-disjunct rules in data mining. Applied Soft Computing 2002*;* 2(2):75-88.

Carvalho, D. R., Freitas, A. A. New results for a hybrid decision tree/genetic algorithm for data mining. Proceedings of the Fourth International Conference on Recent Advances in Soft Computing; 2002.

Freitas, A. A. Evolutionary computation. In Handbook of Data Mining and Knowledge Discovery; Oxford University Press, 2002.

Goldberg, D. E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, 1989.

Holte, R. C., Acker, L. E., Porter, B. W. Concept learning and the problem of small disjuncts. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence; 1989.

Japkowicz, N. Concept learning in the presence of between-class and within-class imbalances. Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence, Springer-Verlag; 2001.

Japkowicz, N. Supervised learning with unsupervised output separation. International Conference on Artificial Intelligence and Soft Computing; 2002.

Japkowicz, N., Stephen, S. The class imbalance problem: a systematic study. Intelligent Data Analysis 2002; 6(5):429-450.

Joshi, M. V., Agarwal, R. C., Kumar, V. Mining needles in a haystack: classifying rare classes via two-phase rule induction. SIGMOD ’01 Conference on Management of Data; 2001.

Joshi, M.V., Agarwal, R. C., Kumar, V. Predicting rare classes: can boosting make any weak learner strong? Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

Joshi, M. V., Kumar, V., Agarwal, R. C. Evaluating boosting algorithms to classify rare cases: comparison and improvements. First IEEE International Conference on Data Mining; 2001.

Kubat, M., Holte, R. C., Matwin, S. Machine learning for the detection of oil spills in satellite radar images. Machine Learning 1998; 30(2):195-215.

Liu, B., Hsu, W., Ma, Y. Mining association rules with multiple minimum supports. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999.

Riddle, P., Segal, R., Etzioni, O. Representation design and brute-force induction in a Boeing manufacturing design. Applied Artificial Intelligence 1994; 8:125-147.

Schapire, R. E. A brief introduction to boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.

Ting, K. M. The problem of small disjuncts: its remedy in decision trees. Proceeding of the Tenth Canadian Conference on Artificial Intelligence; 1994.

Van den Bosch, A., Weijters, T., Van den Herik, H. J., Daelemans, W. When small disjuncts abound, try lazy learning: A case study. Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning; 1997.

Weiss, G. M. Learning with rare cases and small disjuncts. Proceedings of the Twelfth International Conference on Machine Learning; Morgan Kaufmann, 1995.

Weiss, G. M., Hirsh, H. Learning to predict rare events in event sequences. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; 1998.

Weiss, G. M. Timeweaver: a genetic algorithm for identifying predictive patterns in sequences of events. Proceedings of the Genetic and Evolutionary Computation Conference; Morgan Kaufmann, 1999.

Weiss, G. M., Hirsh, H. A quantitative study of small disjuncts. Proceedings of the Seventeenth National Conference on Artificial Intelligence; AAAI Press, 2000.

Weiss, G. M. Mining with Rarity—Problems and Solutions: A Unifying Framework. SIGKDD Explorations 2004: 6(1):7-19.