An Approach to Imbalanced Data Sets Based on Changing Rule Strength

Jerzy W. Grzymala-Busse
Xinqun Zheng
Department of Electrical Engineering
and Computer Science
University of Kansas
Lawrence, KS 66045
Jerzy@eecs.ukans.edu
zheng@eecs.ukans.edu

Linda K. Goodwin
Department of Information Services
and the School of Nursing
Duke University
Durham, NC 27710
goodw010@mc.duke.edu

Witold J. Grzymala-Busse
RS Systems, Inc.
Lawrence, KS 66047
witek@argus.rs-systems.com

Abstract

This paper describes experiments with a challenging data set describing preterm births. The data set, collected at the Duke University Medical Center, was large and, at the same time, many attribute values were missing. However, the main problem was that only 20.7% of the total number of cases represented the important preterm birth class. Thus the data set was imbalanced. For comparison, we include results of experiments on another imbalanced data set, the well-known breast cancer data set. Our approach to dealing with this imbalanced data set was to induce a rule set using our standard procedure: the LEM2 algorithm of the LERS rule induction system and then increase the rule strength for all rules describing preterm births by multiplying all such rule strengths by the same number called a strength multiplier. The rules strength for any rule describing the majority class, fullterm birth, remained unchanged. The optimal strength multiplier was determined experimentally using our optimality criterion: the maximum of the sum of sensitivity and specificity.

Introduction

Approximately one of every ten infants is born preterm (premature). Preterm birth is the leading cause of death in infants, and those who survive frequently suffer from lifelong handicaps and require health care that costs about one million dollars in the first year of life. (Creasy and Herron, 1981) developed a manual preterm risk scoring tool that was widely used for nearly a decade, but later evaluated as ineffective for accurate identification of most preterm births (Creasy, 1993). A decade of manual preterm risk scoring tools yielded only 17–38% positive predictive values (McLean and Walters, 1993), thus data-driven decision support tools are needed to improve diagnosis in this complex domain.

This paper describes a series of experiments with preterm birth data provided by the Duke University Medical Center. Duke's data set includes a sample of 19,970 women that is ethnically diverse and includes 1,229 variables. Duke's data subset was partitioned into two parts: training (14,977 cases) and testing (4,993 cases). The prenatal data set collected at the Duke University Medical Center is associated with many technical challenges. First of all, it is large.

Secondly, the data set contains many missing attribute values. For example, an average attribute of the training data set is not specified for 32.7% of cases. Even worse, for the two mutually disjoint subsets of the main set (1,229 attributes), identified by experts from the Duke University Medical Center as important, the first set containing 52 attributes and the second set containing 54 attributes, named Duke-1 and Duke-2, respectively, have even more missing attribute values. The Duke-1 data set contains laboratory test results. The Duke-2 data set represents the most essential remaining attributes that, according to experts, should be used in diagnosis of preterm birth. Duke-1 has 64.8% missing attribute values, Duke-2 has 36.1% missing attribute values.

There are many approaches to handle missing attribute values in data mining (Grzymala-Busse 1991; Grzymala-Busse *et al*. 1999b; Michalski *et al*. 1986; Quinlan 1993). So far we experimented with the closest fit algorithm for missing attribute values, based on replacing a missing attribute value by an existing value of the same attribute in another case that resembles as much as possible the case with the missing attribute values (Grzymala-Busse *et al*. 1999b).

Furthermore, both data sets, Duke-1 and Duke-2, are imbalanced because only 3,103 training cases are preterm, all remaining 11,874 cases are fullterm. Similarly, in the testing data set, there are only 1,023 preterm cases while the number of fullterm cases is 3,970. Since both data sets, Duke-1 and Duke-2, yield similar results, for brevity we will present results only for Duke-1.

The data sets are further complicated by numerical attributes. Usually data with numerical attributes are consistent, i.e., for any two cases with the same vectors of attribute values, the outcome is the same. This is not the case with Duke's data set. Even with all 1,229 attributes the training data set is inconsistent. Thus, discretization, the process of converting numerical attributes into symbolic attributes, is a difficult problem for this preterm-birth data. Our solution is based on preserving the existing rate of conflicting cases (i.e., keeping the same inconsistency level). Following this approach, the numerical attribute values of the training data set were sorted for every attribute. Every value *v* was replaced by the interval [*v*, *w*), where *w* was the next bigger values than *v* in the sorted list. This approach to discretization is very cautious since, in the training data set, we put only one attribute value in each interval. For testing data sets, values were replaced by the corresponding intervals taken from the training data set. It is possible that a few values come into the same interval. This method was selected to keep the same inconsistency level of the data.

Rule Induction

In our research, the main data mining tool was LERS (Learning from Examples based on Rough Sets), developed at the University of Kansas (Grzymala-Busse, 1992). LERS has proven its applicability having been used for years by NASA Johnson Space Center (Automation and Robotics Division), as a tool to develop expert systems of the type most likely to be used in medical decision-making on board the International Space Station. LERS was also used to enhance facility compliance under Sections 311, 312, and 313 of Title III, the Emergency Planning and Community Right to Know. The project was funded by the U. S. Environmental Protection Agency. System LERS was used in other areas as well, e.g., in the medical field to compare the effects of warming devices for postoperative patients and to assess preterm birth (Woolery and Grzymala-Busse, 1994).

LERS handles inconsistencies using rough set theory. The main advantage of rough set theory, introduced by Z. Pawlak in 1982 (Pawlak 1982, 1991; Pawlak *et al.* 1995), is that it does not need any preliminary or additional information about data (like probability in probability theory, grade of membership in fuzzy set theory, etc.). In rough set theory approach inconsistencies are not removed from consideration. Instead, lower and upper approximations of the concept are computed. On the basis of these approximations, LERS computes two corresponding sets of rules: certain and possible, using algorithm LEM2 (Grzymala-Busse, 1992).

Classification

For classification of unseen cases system LERS uses a modified "bucket brigade algorithm" (Booker *et al.* 1990; Holland *et al.* 1986). In this approach, the decision to which concept an example belongs is made using two factors: *strength* and *support*. They are defined as follows: *Strength factor* is a measure of how well the rule has performed during training. The second factor, *support*, is related to a concept and is defined as the sum of scores of all matching rules from the concept. The concept getting the largest support wins the contest.

In LERS, the strength factor is adjusted to be the *strength* of a rule, i.e., the total number of examples correctly classified by the rule during training. The concept *C* for which support, i.e., the following expression

$$\sum_{\text{matching rules } R \text{ describing } C} \text{Strength factor}(R)$$

is the largest is a winner and the example is classified as being a member of *C*.

If an example is not completely matched by any rule, some classification systems use *partial matching*. System AQ15, during partial matching, uses a probabilistic sum of all measures of fit for rules [(Michalski *et al.* 1986).

In the original bucket brigade algorithm, partial matching is not considered as a viable alternative of complete matching. Bucket brigade algorithm depends on default hierarchy instead (Holland *et al.* 1986).

In LERS partial matching does not rely on the user's input. If complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of an example.

For any partially matching rule *R*, the additional factor, called *Matching factor* (*R*), is computed. Matching factor(*R*) is defined as the ratio of the number of matched attribute-value pairs of a rule *R* with the case to the total number of attribute-value pairs of the rule *R*. In partial matching, the concept *C* for which the following expression is the largest

$$\sum_{\text{partially matching rules } R \text{ describing } C} \text{Matching factor}(R) * \text{Strength factor}(R)$$

is the winner and the example is classified as being a member of *C*.

Table 1

		Duke-TR	Duke-ALL	Breast-Cancer
	Error rate in %	21.19	3.18	24.0
Initial	Sensitivity in %	0.59	85.54	33.33
	Specificity in %	98.97	99.67	93.01
	Average rule strength:			
	Basic class	29.42	23.85	2.46
	Complementary class	104.67	77.53	5.20
	Strength multiplier	5.552	44.0	4.0
	Error rate in %	44.98	2.90	35.0
Optimal	Sensitivity in %	53.40	96.73	61.40
	Specificity in %	61.29	97.19	66.43

Sensitivity and Specificity

In many applications, e.g., medical area, we distinguish between two classes: basic and complementary. The basic class is more important, e.g., in medical area, it is defined as the class of all cases that should be diagnosed as affected by a disease or other medical condition, e.g., preterm birth.

The set of all correctly classified (preterm) cases from the basic class are called true-positives, incorrectly classified basic cases (i.e., classified as fullterm) are called false-negatives, correctly classified complementary (fullterm) cases are called true-negatives, and incorrectly classified complementary (fullterm) cases are called false-positives.

Sensitivity is the conditional probability of true-positives given basic class, i.e., the ratio of the number of true-positives to the sum of the number of true-positives and false-negatives. Specificity is the conditional probability of true-negatives given complementary class, i.e., the ratio of the number of true-negatives to the sum of the number of true-negatives and false-positives.

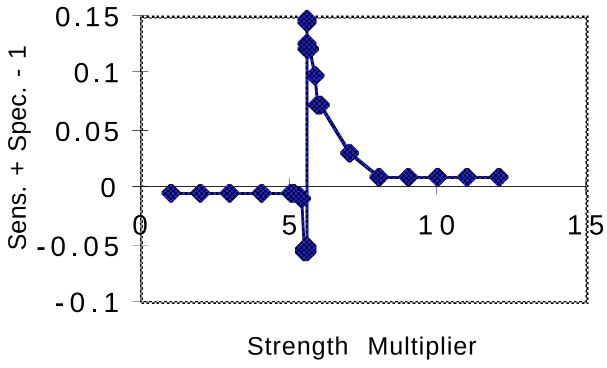

Figure 1. Duke-TR data

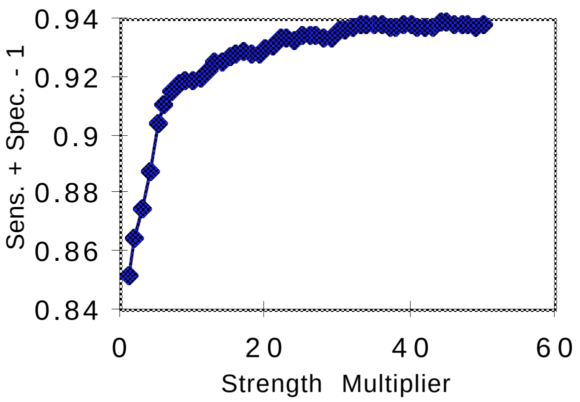

Figure 2. Duke-ALL data

Data Sets

In this paper we present results of experiments on three data sets. The first two data sets were collected at the Duke University Medical Center. The only difference between these two data sets is the approach used to guess missing attribute values. The first data set, Duke-TR, was obtained from the original data set Duke-1 by splitting Duke-1 into training (75%) and testing (25%) data sets, then the training data set was pre-processed using a closest fit approach to missing attribute values.

In the closest fit algorithm for missing attribute values a missing attribute value is replaced by an existing value of the same attribute in another case that resembles as much as possible the case with the missing attribute values. To search for the closest fit case, we need to compare two vectors of attribute values of the given case with missing attribute values and of a searched case. There are many possible variations of the closest fit idea. In this paper we will restrict our attention to the closest fitting cases within

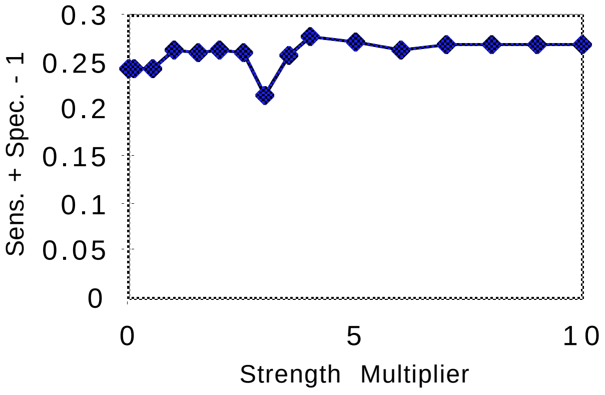

Figure 3. Breast-Cancer data

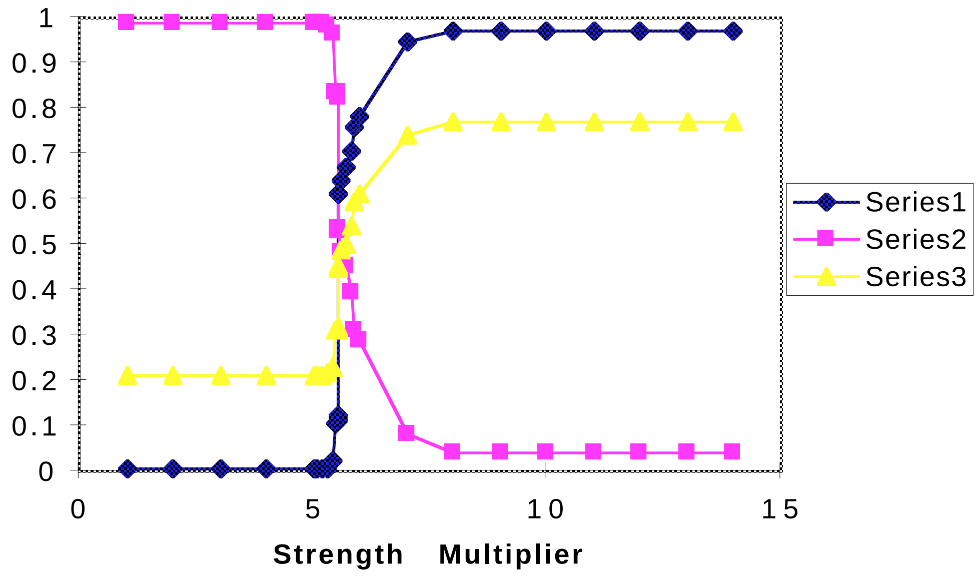

Figure 4. Duke-TR data

the same class. This algorithm is a part of the system OOMIS. During the search, the entire training set within the same class is scanned, for each case a proximity measure is computed, the case for which the proximity measure is the largest is the closest fitting case that is used to determine the missing attribute values.

The testing data set of Duke-TR contained 64.5% of missing attribute values. During matching of testing cases against rules, in the classification process, missing attribute values are ignored for matching.

On the other hand, another data set, Duke-ALL, was obtained from the original data set Duke-1 by—first—preprocessing using a closest fit approach to missing attribute values and—then—by splitting for training (75%) and testing (25%) data sets.

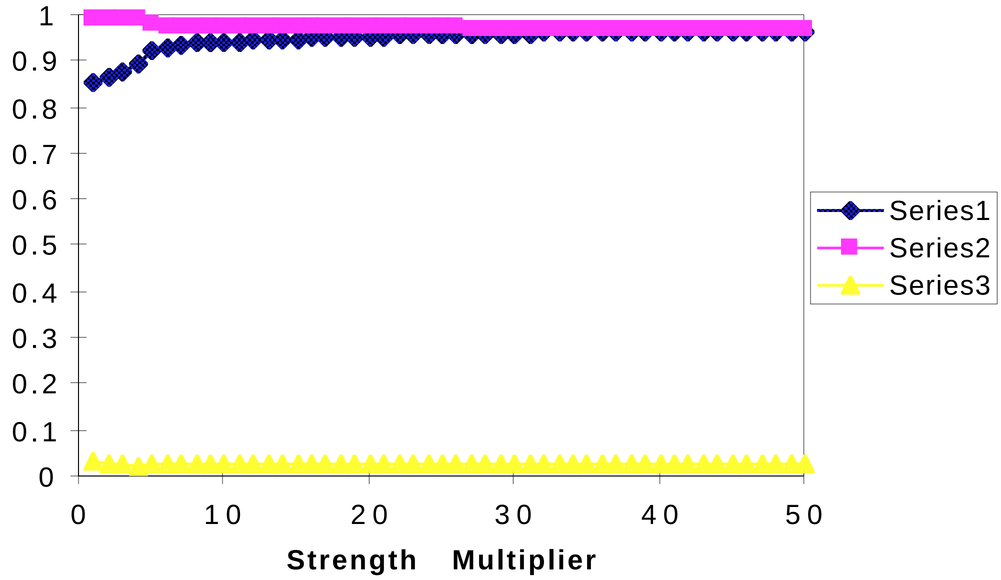

Figure 5. Duke-ALL data

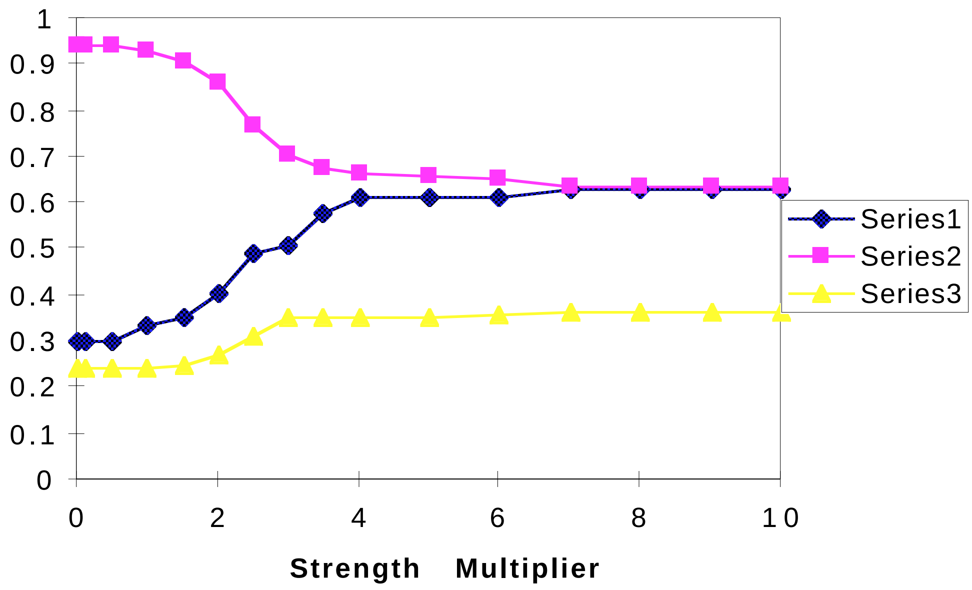

Figure 6. Breast-Cancer data

The quality of the first data set, Duke-TR, is poor, while the quality of the second set, Duke-ALL, is high, therefore we experimented on a third data set, of medium quality, also imbalanced (with 29.7% of cases in the basic class), called Breast-Cancer. This is a well-known data set accessible from the Data Repository of the University of California at Irvine. It contains only 0.35% of missing attribute values.

Results of our experiment are cited in the Table 1 and presented on Figures 1–6. In charts on Figures 4–6, series 1 represents sensitivity, series 2 represents specificity, and series 3 represents the error rate (the ratio of the total number of incorrectly classified cases from both classes to the total number of cases).

Strength Multipliers

In imbalanced data sets with two classes (concepts) one class is represented by the majority of cases while the other class is a minority. Unfortunately, in medical data the smaller class—as a rule—is more important.

In Duke's perinatal training data, only 20.7% of the cases represent the basic class, preterm birth. During rule induction, the average of all rule strengths for the bigger class is also greater than the average of all rule strengths for the more important but smaller basic class. During classification of unseen cases, rules matching a case and voting for the basic class are outvoted by rules voting for the bigger, complementary class. Thus the sensitivity is poor and the resulting classification system would be rejected by diagnosticians.

Therefore it is necessary to increase sensitivity. The simplest idea is to add cases to the basic class in the data set, e.g., by adding duplicates of the available cases. The total number of training cases will increase, hence the total running time of the rule induction system will also increase. However, adding duplicates will not change the knowledge hidden in the original data set, but it may create a balanced data set so that the average rule set strength for both classes will be approximately equal. The same effect may be accomplished by increasing the average rule strength for the basic class. In our research we selected the optimal rule set by multiplying the rule strength for all rules describing the basic class by the same real number called a *strength multiplier* (Grzymala-Busse *et al.*, 1999a).

In general, the sensitivity increases with the increase of the strength multiplier. At the same time, the specificity decreases. It is difficult to estimate what is the optimal value of the strength multiplier. In our experiments the choice of the optimal value of the strength multiplier was based on an analysis presented by Bairagi and Suchindran (1989). Let p be a probability of the correct prediction, i.e., the ratio of all true positives and all false positives to the total number of all cases. Let P be the probability of an actual basic class, i.e., the ratio of all true positives and all false negatives to the total number of all cases. Then

$$p = \text{Sensitivity} * P + (1 - \text{Specificity}) * (1 - P).$$

As Bairagi and Suchindran observed (1989), we would like to see the change in p as large as possible with a change in P, i.e., we would like to maximize

$$\frac{dp}{dP} = \text{Sensitivity} + \text{Specificity} - 1.$$

Thus the optimal value of the strength multiplier is the value that corresponds to the maximal value of Sensitivity + Specificity – 1.

Conclusions

Results of our experiments show that an increase in specificity may be accomplished by changing strength multipliers for rules describing the basic class and by using the LERS classification system.

For poor quality data (Duke-TR), by increasing the strength multiplier (until it reaches an optimal value), a large increase in sensitivity (by 52.81%) is achieved but specificity decreases significantly (by 37.68%), thus the total error rate raises (by 23.79%).

For high quality data (Duke-ALL), under the same circumstances, a significant increase in sensitivity (by 11.19%) and a small decrease in specificity (2.48%) resulted in a decrease of the total error rate (by 0.28%).

And, finally, medium quality data (Breast-Cancer) are characterized by a large increase in sensitivity (by 28.07%) with a decrease in specificity (26.58%), and, at the same time, an increase in the total error rate (by 11%).

For many important applications, e.g., medical area, an increase in sensitivity is crucial, even if it is achieved at the cost of specificity. Thus, the suggested method of increasing the strength multiplier may be successfully applied for machine learning from imbalanced data.

References

Bairagi, R., and Suchindran, C. M. 1989. An estimator of the cutoff point maximizing sum of sensitivity and specificity. *Sankhya*, Series B, *Indian Journal of Statistics* 51: 263–269.

Booker, L. B.; Goldberg, D. E.; and Holland, J. F. 1990. Classifier systems and genetic algorithms. In *Machine Learning. Paradigms and Methods*. Carbonell, J. G. (ed.). Cambridge, MA: The MIT Press, 235–282.

Creasy, R. K., and Herron, M. A. 1981. Prevention of preterm birth. *Seminars in Perinatology* 5: 295–302.

Creasy, R. K. 1993. Preterm birth prevention: where are we? *American Journal of Obstetrics & Gynecology* 168: 1223–1230.

Grzymala-Busse, J. W. 1991. On the unknown attribute values in learning from examples. *Proceedings of the Sixth International Symposium on Methodologies for Intelligent Systems*, ISMIS-91 Charlotte, North Carolina, October 16–19, 1991, 368–377. Lecture Notes in Artificial Intelligence, vol. 542. Berlin, Germany: Springer-Verlag.

Grzymala-Busse, J. W. 1992. LERS—A system for learning from examples based on rough sets. In *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*. Slowinski, R. (ed.). Norwell, MA: Kluwer Academic Publishers, 3–18.

Grzymala-Busse, J. W.; Goodwin, L. K.; and Zhang, X. 1999a. Increasing sensitivity of preterm birth by changing rule strengths. Proceedings of the Eigth Workshop on Intelligent Information Systems (IIS'99), Ustron, Poland, June 14–18, 127–136.

Grzymala-Busse, J. W.; Grzymala-Busse, W. J.; and Goodwin, L. K. 1999b. A closest fit approach to missing attribute values in preterm birth data. *Proceedings of the Seventh International Workshop on Rough Sets, Fuzzy Sets, Data Mining and Granular-Soft Computing* (RSFDGrC'99), Ube, Yamaguchi, Japan, November 8–10, 1999, 405–413. Lecture Notes in Artificial Intelligence, vol. 1711. Berlin, Germany: Springer Verlag.

Holland, J. H.; Holyoak, K. J.; and Nisbett, R. E. 1986. *Induction. Processes of Inference, Learning, and Discovery*. Cambridge, MA: The MIT Press.

McLean, M.; Walters, W. A.; and Smith, R. 1993. Prediction and early diagnosis of preterm labor: a critical review. *Obstetrical & Gynecological Survey* 48: 209–225.

Michalski, R. S.; Mozetic, I.; Hong, J.; and Lavrac, N. 1986. The AQ15 inductive learning system: An overview and experiments. Rep. UIUCDCD-R-86-1260, Department of Computer Science, University of Illinois.

Pawlak, Z.; Grzymala-Busse, J. W.; Slowinski, R.; and Ziarko, W. 1995. Rough Sets. *Communications of the ACM* 38: 89–95.

Pawlak, Z. 1982. Rough sets. *International Journal Computer and Information Sciences* 11: 341–356.

Pawlak, Z. 1991. *Rough Sets. Theoretical Aspects of Reasoning about Data*. Norwell, MA: Kluwer Academic Publishers.

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

Woolery, L. K., and Grzymala-Busse, J. 1994. Machine learning for an expert system to predict preterm birth risk. *Journal of the American Medical Informatics Association* 1: 439–446.