Handling Imbalanced Data Sets in Insurance Risk Modeling

Edwin P. D. Pednault, Barry K. Rosen, and Chidanand Apte

IBM T. J. Watson Research Center
P.O. Box 218
Yorktown Heights, New York 10598
pednault@us.ibm.com, rosen@watson.ibm.com, apte@us.ibm.com

Abstract

As owners of cars, homes, and other property, consumers buy property and casualty insurance to protect themselves against the unexpected: i.e., accidents, fire, theft, etc. Such events occur very rarely at the level of individual policyholders. Data sets constructed for the purpose of insurance risk modeling are therefore highly imbalanced. In any given time period, most policyholders file no claims, a small percentage file one claim, and an even smaller percentage file two or more claims. This paper presents some of the tree-based learning techniques we have developed to model insurance risks. Two important aspects of our approach that distinguish it from other tree-based methods are that it incorporates a split-selection criterion tailored to the specific statistical characteristics of insurance data, and it uses constraints on the statistical accuracies of model parameter estimates to guide the construction of splits in order to overcome selection biases that arise because of the imbalance that is present in the data.

Introduction

The property and casualty (P&C) insurance business deals with the insuring of tangible assets; e.g., cars, boats, homes, etc. The insuring company evaluates the risk of the asset being insured taking into account characteristics of the asset as well as the owner of the asset. Based on the level of risk, the company charges a certain fixed, regular premium to the insured. Actuarial analysis of policy and claims data plays a major role in the analysis, identification, and pricing of P&C risks.

Actuaries develop insurance risk models by segmenting large populations of policies into predictively accurate risk groups, each with its own distinct risk characteristics. A well-known segment is male drivers under age 25 who drive sports cars. Examples of risk characteristics include mean claim frequency, mean claim severity amount, pure premium (i.e., frequency times severity), and loss ratio (i.e., pure premium over premium charged). Pure premium is perhaps the most important risk characteristic because it represents the minimum amount that policyholders in a risk group must be charged in order to cover the claims generated by that risk group. Actual premiums charged are ultimately determined based on the pure premiums of each risk group, as well as on the cost structure of the insurance company, its marketing strategy, competitive factors, etc.

It is common for large insurance firms to have millions of policies in each of their operating territories, with many years of accumulated claim data. The actuarial departments of insurance companies make use of this data to develop risk models for the markets served by their companies. The availability of large quantities of insurance data represents both an opportunity and a challenge for machine learning.

One of these challenges is that individual policyholders tend to file claims very infrequently. For example, for automobile insurance, only about 2% of policyholders will file claims in any given quarter. Data sets constructed for the purpose of insurance risk modeling are therefore highly imbalanced with respect to the proportion of data records of policyholders that filed claims in a given time period versus those that filed no claims.

In addition to the sparsity of claim records, insurance data poses several other challenges from the point of view of developing suitable learning algorithms. One challenge is that specialized, domain-specific equations must be used to estimate claim frequency and claim severity. Another challenge is that some types of claims can take several years to settle, most notably bodily injury claims. Consequently, the claim amount can be a missing value. A suitable learning algorithm would therefore have to compensate for such "partially labeled" training examples. A third challenge is that insurance actuaries demand tight confidence bounds on the risk parameters that are obtained; i.e., the risk groups identified by the learning algorithm must be *actuarially credible*.

The above challenges have motivated our own research (Pednault 1999; Hosking, Pednault, and Sudan 1997) and have lead to the development of the IBM ProbE[TM] (*Prob*abilistic *E*stimation) predictive modeling kernel (Apte *et al*. 1999). This C++ kernel embodies several innovations that address the challenges posed by insurance data. A tree-based learning algorithm is employed to identify risk groups. However, unlike other tree-based methods, such as CHAID (Biggs, de Ville, and Suen 1991; Kass 1980), CART (Breiman *et al*. 1984), C4.5 (Quinlan 1993), SPRINT (Shafer, Agrawal, and Mehta 1996), and QUEST (Loh and Shih 1997), a domain-specific optimization criterion is used to identify suitable splits during tree building.

The optimization criterion has the effect of assigning different weights to claim records than to nonclaim records. This weight imbalance seems to help compensate for the imbalance in the proportion of claim versus nonclaim records; however, we have found that the compensating effect of the optimization criterion does not in and of itself solve the data set imbalance problem.

Because rare events by definition occur rarely, it is all too easy to construct splits in which one branch corresponds to a small subset of the data in which the rare event does not occur. At first glance, it might appear that progress is being made. However, the smaller the subset of data one considers, the higher is the probability that the rare event will not occur simply by random chance. ProbE ameliorates the small sample size problem by using constraints on the statistical accuracies of model parameter estimates to guide the construction of splits. The constraints eliminate small splinter groups from consideration because splinter groups yield inaccurate parameter estimates for rare events. The constraints further ensure that the resulting risk groups are actuarially credible.

Top Down Identification of Risk Groups

The traditional method used by actuaries to construct risk models involves first segmenting the overall population of policyholders into a collection of risk groups based on a set of factors, such as age, gender, driving distance to place of employment, etc. The risk parameters of each group are then estimated from historical policy and claims data. Ideally, the resulting risk groups should be homogeneous with respect to risk; i.e., further subdividing the risk groups by introducing additional factors should yield substantially the same risk parameters. Actuaries typically employ a combination of intuition, guesswork, and trial-and-error hypothesis testing to identify suitable factors. The human effort involved is often quite high and good risk models can take several years to develop and refine.

ProbE replaces manual exploration of potential risk factors with automated search. Risk groups are identified in a top-down fashion by a method similar to those employed in classification and regression tree algorithms (Biggs, de Ville, and Suen 1991; Breiman *et al*. 1984; Kass 1980; Quinlan 1993; Shafer, Agrawal, and Mehta 1996; Loh and Shih 1997). Starting with an overall population of policyholders, ProbE recursively divides policyholders into risk groups by identifying a sequence of factors that produce the greatest increase in homogeneity within the subgroups that are produced. The process is continued until each of the resulting risk groups is either declared to be homogeneous (via tree pruning) or is too small to be further subdivided from the point of view of actuarial credibility.

One of the key differences between ProbE and other classification and regression tree algorithms is that splitting factors are selected based on statistical models of insurance

risks. In the case of ProbE, a joint Poisson/log-normal model is used to enable the simultaneous modeling of frequency and severity, and hence pure premium.

This choice of statistical model was strongly influenced by the fundamental nature of the claims process. For property and casualty insurance, the claims process consists of claims being filed by policyholders at varying points in time and for varying amounts. In the normal course of events, wherein claims are not the result of natural disasters or other widespread catastrophes, loss events that result in claims (i.e., accidents, fire, theft, etc.) tend to be randomly distributed in time with no significant pattern to the occurrence of those events from the point of view of insurance risk. Policyholders can also file multiple claims for the same type of loss over the life of a policy. These properties are the defining characteristics of Poisson random processes.

In addition to modeling the distribution of claims over time, the amounts of those claims must also be modeled. Log-normal distributions were selected for this purpose based on an examination of actual historical automobile claims data. The claim amounts were found to have a highly skewed distribution. Most claims were small in value relative to the maximum amounts covered by the policies, but a significant proportion of large claims were also present. When the claim amounts were log transformed, the skewness virtually disappeared and the resulting distribution was found to be highly Gaussian in shape. These properties are the defining characteristics of log-normal distributions.

The risk modeling algorithms in ProbE are designed to model frequency and severity simultaneously using a joint Poisson/log-normal model. The risk groups that are identified are therefore homogeneous with respect to pure premium (i.e., the product of frequency and severity). Moreover, the joint Poisson/log-normal model explicitly takes into account the peculiarities of insurance data discussed in the introduction.

The Joint Poisson/Log-Normal Model

The optimization criterion used to identify splitting factors is based on the principles of maximum likelihood estimation. Specifically, the negative log-likelihood of each data record is calculated assuming a joint Poisson/log-normal statistical model, and these negative log likelihoods are then summed to yield the numerical criterion that is to be optimized. Minimizing this negative log-likelihood criterion causes splitting factors to be selected that maximize the likelihood of the observed data given the joint Poisson/log-normal models of each of the resulting risk groups.

Historical data for each policy is divided into distinct time intervals for the purpose of data mining, with one data record constructed per policy per time interval.

Time-varying risk characteristics are then assumed to remain constant within each time interval; that is, for all intents and purposes their values are assumed to change only from one time interval to the next. The choice of time scale is dictated by the extent to which this assumption is appropriate given the type of insurance being considered and the business practices of the insurer. For convenience, quarterly intervals will be assumed to help make the discussion below more concrete, but it should be noted that monthly or yearly intervals are also possible

Assuming that data is divided into quarterly intervals, most data records will span entire quarters, but some will not. In particular, data records that span less than a full quarter must be created for policies that were initiated or terminated mid-quarter, or that experienced mid-quarter changes in their risk characteristics. In the case of the latter, policy-quarters must be divided into shorter time intervals so that separate data records are created for each change in the risk characteristics of a policy. This subdivision must be performed in order to maintain the assumption that risk characteristics remain constant within the time intervals represented by each data record. In particular, subdivision must occur when claims are filed under a policy in a given quarter because the filing of a claim can itself be an indicator of future risk (i.e., the more claims one files, the more likely one is to file future claims). The actual time period covered by a database record is the *earned exposure* of that record.

For Poisson random processes, the time between claim events follows an exponential distribution. Moreover, no matter at what point one starts observing the process, the time to the next claim event has the same exponential distribution as the time between claim events. It can further be shown that the probability density for the total time T (i.e., the total earned exposure) between $k+l$ claim filings (where k is the number of settled claims and l is the number of open claims) is given by

$$f(T \mid k+l) = \lambda^{k+l} e^{-\lambda T} , \tag{1}$$

where λ is the claim frequency of the risk group. The maximum likelihood estimate used by ProbE for the frequency parameter λ is thus the same one that is typically used by actuaries for estimating frequency:

$$\widehat{\lambda} = \frac{k+l}{T} = \frac{\text{Total Number of Claims}}{\text{Total Earned Exposure}} . \tag{2}$$

In the case of claim amounts, the joint probability density function for the severities $s_1, ..., s_k$ of k settled claims is given by:

$$f(s_1, ..., s_k) = \frac{e^{-\dfrac{\sum_{i=1}^{k}(\log(s_i) - \mu_{\log})^2}{2\sigma_{\log}^2}}}{\left(\sqrt{2\pi}\,\sigma_{\log}\right)^k \prod_{i=1}^{k} s_i} . \tag{3}$$

The estimates of the mean log severity $\mu_{\log}$ and the variance of the log severity $\sigma_{\log}$ are likewise the ones typically used for log-normal distributions:

$$\widehat{\mu}_{\log} = \frac{1}{k} \sum_{i=1}^{k} \log(s_i) . \tag{4}$$

and

$$\widehat{\sigma}_{\log}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\log(s_i) - \widehat{\mu}_{\log}\right)^2 . \tag{5}$$

Only fully settled claims are considered when applying Equations (4) and (5). The severity fields of unsettled claims are often used to record reserve amounts; i.e., the money that insurers hold aside to cover pending claims. Reserve amounts are not actual losses and therefore are not used to develop models for predicting actual losses.

Negative log-likelihoods are calculated for each database record in a risk group based on Equations (1) and (3). The nonconstant terms in the negative log-likelihoods are then summed and used as the criterion for selecting splitting factors in the top-down identification of risk groups. The constant terms do not contribute to the selection of splitting factors and, hence, are omitted to reduce the amount of computation.

With constant terms removed, the negative log-likelihood score for the i'th database record is:

$$\xi_i = \begin{cases} \widehat{\lambda} t_i, & \text{non-claim} \\[2mm] \widehat{\lambda} t_i + \log\left(\dfrac{\widehat{\sigma}_{\log}}{\widehat{\lambda}}\right), & \text{open claim} \\[3mm] \widehat{\lambda} t_i + \log\left(\dfrac{\widehat{\sigma}_{\log}}{\widehat{\lambda}}\right) + \dfrac{(\log(s_i) - \widehat{\mu}_{\log})^2}{2\widehat{\sigma}_{\log}^2}, & \text{settled claim} \end{cases} \tag{6}$$

where t_i is the earned exposure for the i'th record. Note that the Poisson portion of the model contributes an amount $\widehat{\lambda} t_i + \log(1/\widehat{\lambda})$ to the score of each claim record and an amount $\widehat{\lambda} t_i$ to the score of each nonclaim record. The sum of these values equals the negative logarithm of Equation (1). The log-normal portion of the model contributes nothing to the scores of nonclaim records, and an amount $\log(\widehat{\sigma}_{\log}) + (\log(s_i) - \widehat{\mu}_{\log})^2/(2\widehat{\sigma}_{\log}^2)$ to the score of each settled claim record. The sum of these values equals the negative logarithm of Equation (3) with constant terms (i.e., $\sum_{i=1}^{k} \log(\sqrt{2\pi}\, s_i)$) removed. In the case of open claim records, an expected value estimate of the log-normal score is constructed based on the scores of the settled claim records. After dropping constant terms from this expected value estimate, open claim records contribute an amount $\log(\widehat{\sigma}_{\log})$ to the log-normal portions of their scores.

If the database records for a risk group contain k settled claims and l open claims, then the sum of the above scores is given by:

$$\xi = \hat{\lambda}\left(\sum_{i=1}^{N} t_i\right) + (k+l)\log\left(\frac{\hat{\sigma}_{\log}}{\hat{\lambda}}\right) \quad (7)$$

$$+ \left(\frac{1}{2\hat{\sigma}_{\log}^2}\right)\sum_{i=1}^{k}(\log(s_i) - \hat{\mu}_{\log})^2 \; .$$

In the above equation, N is the total number of database records for the risk group, the first k of which are assumed for convenience to be settled claim records. Equation (7) is then summed over all risk groups to yield the overall score of the risk model. The top-down procedure described in the previous section identifies risk groups by minimizing the overall score in a stepwise fashion, where each step involves dividing a larger risk group into two smaller risk groups so as to reduce the value of the overall score to the maximum extent possible.

From the point of view of machine learning, the important thing to note about the above equations is that insurance-specific quantities such as earned exposure and claim status enter into both the equations for estimating model parameters and the equations for selecting splitting factors. Earned exposure effectively plays the role of a weighting factor, while claim status plays the role of a correction factor that adjusts for missing data in one of the two data fields to be predicted (i.e., the settled claim amount given that a claim was filed). The equations thus take into account the peculiarities of insurance data discussed in the introduction.

Equation (7) essentially replaces the entropy calculations used in many standard tree-based data mining algorithms. It should be noted that entropy is, in fact, a special case of negative log-likelihood. Its calculation need not be restricted to categorical or Gaussian (least-squares) distributions. The development of the joint Poisson/log-normal model presented above illustrates the general methodology one can employ to customize the splitting criteria of tree-based data mining algorithms to take into account data characteristics that are peculiar to specific applications.

Actuarial Credibility

ProbE's top-down modeling procedure is constrained to produce risk groups that are actuarially credible. In actuarial science, credibility (Klugman, Panjeer and Willmot 1998) has to do with the accuracy of the estimated risk parameters (in this case, frequency, severity, and ultimately pure premium). Accuracy is measured in terms of statistical confidence intervals; that is, how far can the estimated risk parameters deviate from their true values and with what probability. A fully credible estimate is an estimate that has a sufficiently small confidence interval. In particular, estimated parameter values X must be within a certain factor r of their true (i.e. expected) values $E[X]$ with probability at least p:

$$P\left\{\left|\frac{X - E[X]}{E[X]}\right| \le r\right\} \ge p \; . \quad (8)$$

Typical choices of r and p used by actuaries are $r = 0.05$ and $p = 0.9$. In other words, X must be within 5% of $E[X]$ with 90% confidence.

To ensure that actuarially credible risk groups are constructed, ProbE permits a maximum fractional standard error to be imposed on the estimated pure premiums of each risk group. In the process of subdividing larger risk groups into smaller risk groups, ProbE only considers splitting factors that yield smaller risk groups that obey this constraint. Specifically, each risk group must satisfy the following inequality:

$$\frac{\sqrt{Var[X]}}{E[X]} \le r' \; , \quad (9)$$

where X is the pure premium estimate of the risk group, $E[X]$ is the expected value of the pure premium, $Var[X]$ is the variance of the pure premium estimate, and r' is the maximum allowed fractional standard error. If a splitting factor that satisfies Equation (9) cannot be found for a given risk group, that risk group is declared to be too small to be subdivided and no further refinement of the risk group is performed. Actuarial credibility is ensured by the fact that, for any pair of values of p and r in Equation (8), there exists a corresponding value of r' for Equation (9) such that

$$P\left\{\left|\frac{X - E[X]}{E[X]}\right| \le r\right\} \ge p \quad \text{if and only if} \quad \frac{\sqrt{Var[X]}}{E[X]} \le r' \; .$$

In particular, if X is approximately Gaussian and $p = 0.9$, then the corresponding value for r' as a function of r is

$$r' = \frac{r}{1.645} \; . \quad (10)$$

For a 5% maximum error with 90% confidence, the corresponding value of r' would thus be 0.00304 (i.e., 3.04%).

When applying the above credibility constraint, the mean and variance of the pure premium estimate are approximated by their empirical estimates, which yields the following approximation:

$$\frac{\sqrt{Var[X]}}{E[X]} \approx \sqrt{\frac{1}{k+l} + \frac{1}{k}\left(e^{\hat{\sigma}_{\log}^2} - 1\right)} \; . \quad (11)$$

Note that this fractional standard error varies as a function of the statistical properties of each risk group. The determination of when a risk group is too small to be subdivided is thus context-dependent. The ability to impose a context-dependent actuarial credibility constraint on the

top-down process by which risk groups are constructed is another important feature of ProbE that distinguishes it from all other tree-based modeling methods, such as CHAID (Biggs, de Ville, and Suen 1991; Kass 1980), CART (Breiman *et al.* 1984), C4.5 (Quinlan 1993), SPRINT (Shafer, Agrawal, and Mehta 1996), and QUEST (Loh and Shih 1997).

Equation (11) can also be used to obtain a rough estimate of the amount of data needed to justify a given number of risk groups. In general, the standard deviation of log severity tends to be close to one; hence, $(\exp(\hat{\sigma}^2_{\log}) - 1) \geq 1$ in most cases. To achieve a 5% maximum error with 90% confidence, a risk group must therefore cover at least 2,164 claim records, or about 108,200 quarterly records given that the average quarterly claim frequency for automobile insurance tends to be about 2%. Multiply 108,200 by the number of risk groups and it becomes quite evident that a very large number of quarterly data records must be considered in order to achieve fully credible results.

Predictive Accuracy

It is important to note that actuarial credibility constraints are used as a stopping criterion in combination with tree pruning using a separate holdout set of training data to estimate generalization error. Traditional pruning methods do not guarantee actuarial credibility. Likewise, actuarial credibility constraints do not prevent overfitting in and of themselves. Pruning and actuarial credibility must be used together to achieve the objectives of insurance risk modeling: i.e., maximize predictive accuracy while ensuring actuarial credibility.

Split Construction

Applying actuarial credibility constraints during split construction is straightforward when splitting on numerical and ordinal (i.e., ordered categorical) data fields. One simply finds the data value that partitions the range of the field into two subsets such that the credibility constraints defined by Equations (9) and (11) are satisfied and the sum of scores given by Equation (7) for the two subsets is minimized.

Nominal (i.e., purely categorical) data fields are somewhat trickier to deal with both because of the combinatorial number of ways of partitioning nominal values, and because of the potential to overfit when dealing with imbalanced data sets. The freedom to group categorical values arbitrarily can exacerbate the small sample size problem discussed in the introduction.

To address these issues, ProbE incorporates a modified version of the bottom-up merging technique used in CHAID (Biggs, de Ville, and Suen 1991; Kass 1980). CHAID performs a stepwise grouping of categories to form larger groups, and eventually a split. The computational complexity is on the order of the square of the number of categories. We have modified CHAID's basic procedure by using credibility constraints to restrict the potential merges that can be performed at each step. In addition, merging continues until a binary split is constructed, whereas CHAID employs a stopping criterion that attempts to identify multiway splits during the merging process. ProbE's method for splitting nominal data fields proceeds as follows:

1) For the nominal data field in question, identify the subsets of the training data that correspond to each categorical value. These subsets will be referred to as segments and segments will be associated with their corresponding categorical values.

2) Merge together all segments for which the number of claim records is less than a given threshold.

3) If two or more segments remain, then repeatedly select and merge pairs of segments so as to minimize the sum of Equation (7) over all resulting segments subject to the condition that, if at least one segment is not actuarially credible, then at least one of the segments in the pair being merged must likewise not be actuarially credible. Continue the merging process until only two segments remain.

4) If there are two remaining segments and one of them is not actuarially credible, then merge the two remaining segments into a single segment.

Step 1 is the same initial step performed by CHAID. Step 2 has no counterpart in CHAID and is introduced in order to stabilize the merging process performed in Step 3. The premerging performed in Step 2 effectively eliminates spurious segments that are too small to be meaningfully compared with other segments. As such, it helps avoid the small sample size problem associated with rare events, such as claim filings. The threshold in Step 2 should be set as small as possible, but large enough to yield meaningful comparisons among pairs of segments in Step 3. The current default threshold used in ProbE is six claim records.

Steps 3 and 4 form the counterpart to the bottom-up merging process employed by CHAID. In our method, however, the process is constrained to always produce segments that satisfy desired statistical constraints on the segments (i.e., actuarial credibility constraints in the case of insurance risk modeling). The credibility constraint in Step 3 greatly reduces the potential to overfit when constructing splits for nominal fields with many categorical values. Note that this algorithm might fail to find a suitable split, in which case ProbE will not split on the nominal field in question.

The fact that statistical constraints are applied as an integral part of the method for constructing splits is a distinguishing feature of our method. Statistical constraints are used to guide the construction process itself. By contrast, other methods, such as CHAID (Biggs, de Ville, and Suen 1991; Kass 1980), CART (Breiman *et al*. 1984), C4.5 (Quinlan 1993), SPRINT (Shafer, Agrawal, and Mehta 1996), and QUEST (Loh and Shih 1997), apply corresponding statistical constraints only after splits have been constructed. A deficiency of this latter approach is that splits may be constructed that violate the statistical constraints, causing them to be eliminated from further consideration, even though it may have been possible to construct alternate splits that actually do satisfy the desired constraints.

Discussion

When actuarial credibility constraints and related thresholds are deactivated, ProbE tends to identify numerous small splinter groups in the training data that contain unrealistically few claim records. These splinter groups yield unreliable estimates of frequency, severity, and, hence, pure premium. In fact, in one run, one such splinter group had the lowest pure premium as estimated on both the training data and the holdout data used for pruning, but had the highest pure premium as estimated on a separate test set. When actuarial constraints and related thresholds are activated so that they then guide the construction of splits, such erratic behavior is effectively eliminated. Moreover, in comparison tests that we performed (Apte *et al*. 1999), ProbE consistently produced superior risk models compared to those constructed using SPRINT (Shafer, Agrawal, and Mehta 1996), the only off-the-shelf classification and regression tree package available to us that could handle the data volumes required for this application. The success we have had with ProbE suggests that using statistical constraints to guide the construction of splits is likely to be a worthwhile approach in other application domains in which imbalanced data sets are common.

Adapting our split construction method to other domains requires only that suitable statistical constraints be used in place of the actuarial credibility constraints and related thresholds discussed in the previous section. Actuarial credibility requires that the minority class (i.e., claim records) be sufficiently well represented in each and every branch of the tree. This type of constraint might not be appropriate for all applications. For example, for classification problems, it is conceivable for classes to separable at least in some regions of the input space. Rather than constraining fractional (i.e., relative) standard errors as is done in actuarial credibility, one might instead impose constraints on absolute standard errors. The effect in the case of classification problems would be to allow splits in which one branch contains no minority classes, but only if the majority class is sufficiently well represented in order to satisfy the absolute standard error constraints. Other types of constraints are also conceivable, and we believe that using constraints to guide split selection is an idea that is ripe for further exploration, particularly when dealing with imbalanced data sets.

References

Apte, C., Grossman, E., Pednault, E. P. D., Rosen, B. K., Tipu, F. A., and White, B. 1999. Probabilistic estimation-based data mining for discovering insurance risks. *IEEE Intelligent Systems* 14(6):49-58.

Biggs, D., de Ville, B., and Suen, E. 1991. A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics* 18(1):49-62.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees.* New York: Chapman & Hall.

Hosking, J. R. M., Pednault, E. P. D., and Sudan, M. 1997. A statistical perspective on data mining. *Future Generation Computer Systems* 13:117-134.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2):119-127.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. 1998. *Loss Models: From Data to Decisions.* New York: John Wiley & Sons.

Loh, W.-Y., and Shih, Y.-S. 1997. Split selection methods for classification trees. *Statistica Sinica* 7:815-840.

Pednault, E. P. D. 1999. Statistical learning theory. In Wilson, R. A., and Keil, F. C. eds. *The MIT Encyclopedia of the Cognitive Sceinces*, 798-800. Cambridge, Mass.: MIT Press.

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

Shafer, J., Agrawal, R., and Mehta, M. 1996. SPRINT: A scalable parallel classifier for data mining. In Proceedings of the 22nd International Conference on Very Large Databases, 544-555. San Mateo, California: Morgan Kaufmann.