Knowledge Discovery in Telecommunication Services Data Using Bayesian Network Models

Kazuo J. Ezawa
Room 7E-523

Steven W. Norton
Room 7B-511A

AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974
kje@ulysses.att.com norton@ulysses.att.com

Abstract

Fraud and uncollectible debt are multi-billion dollar problems in the telecommunications industry. Because it is difficult to know which accounts will go bad, we are faced with the difficult knowledge-discovery task of characterizing a rare binary outcome using large amounts of noisy, high-dimensional data. Binary characterizations may be of interest but will not be especially useful in this domain. Instead, proposing an action requires an estimate of the probability that a customer or a call is uncollectible. This paper addresses the discovery of predictive knowledge bearing on fraud and uncollectible debt using a supervised machine learning method that constructs Bayesian network models. The new method is able to predict rare event outcomes and cope with the quirks and copious amounts of input data. The Bayesian network models it produces serve as an input module to a normative decision-support system and suggest ways to reinforce or redirect existing efforts in the problem area. We compare the performance of several conditionally independent models with the conditionally dependent models discovered by the new learning system using real-world datasets of 4-6 million records and 600-800 million bytes.

I. Introduction

Every year, the telecommunications industry incurs several billion dollars in uncollectible debt. Despite the fact that collectible revenues are considerably higher (more than 100 billion dollars annually), controlling uncollectibles is an important problem in the industry.

If an oracle unerringly identified customers who would not pay their bills or phone calls for which we could not collect, policy making would be simple. Instead, we can never really be certain about a customer or a call. That is not to say though that we must be entirely uninformed. To support policy making that reduces the level of uncollectible debt, we need only provide an estimate of this probability of uncollectible debt. In fact, unqualified black and white assessments will not be particularly useful. Instead, a probability model could and should be devised as input to a normative decision-support system [Ezawa, 1993]. That way a variety of actions might be considered, ranging from inaction to call disconnect in the most extreme cases. The question to be addressed then is how to feasibly develop a useful probability model.

The datasets employed here contain a mixture of call-detail and customer-summary information. While large by research standards (4-6 million records and 600-800 million bytes) they are *small* by telecommunications industry standards. The interesting outcomes (*i.e.* the non-paying customers) are rare, comprising just 1 or 2% of the population. Compounding the difficulty is the reality of unequal misclassification costs. Non-paying customers that initially slip through undetected will be identified within a couple of billing cycles anyway. As bad as that may be, the greater potential problem is incorrectly classifying valuable paying customers. In today's highly competitive telecommunications market, dissatisfied customers have a range of options to choose from; the corresponding revenue might well be lost forever. And finally, the data are described by more than thirty variables, some discrete and some continuous. Many of the discrete variables have large unordered outcome sets. The continuous variables are not normally distributed. And last but not least, missing values are all too common.

Some learning methods simply cannot hope to process this much data in a timely manner because they must process it many times over before converging to a solution [Baldi and Chauvin, 1991]. Efficient decision tree learners that use recursive partitioning [Quinlan, 1993] often have difficulty with discrete variables that have large unordered outcome sets, such as telephone exchange or city name. In addition, their pruning mechanisms are easily thwarted by widely disparate class proportions, all too often returning a single node tied to the majority class instead of a meaningful tree structure. Even though we enriched our datasets by selecting a subpopulation more likely to be uncollectible, now 9 to 12% instead of 1 to 2%, these systems still have difficulties characterizing the minority class. Lastly, decision tree learners offer little support for problem domains with unequal misclassification costs. (To our knowledge, none has considered misclassification costs that vary from example to example the way they do in this problem.) At the moment, appropriate treatment of unequal

misclassification costs is an open research area [Catlett, 1995, Pazzani *et al*, 1994]. All of this is merely to suggest the kinds of difficulties this data poses to learning systems in general, whether they are regression systems, nearest-neighbor systems, neural networks, *etc.*

This paper presents the Advanced Pattern Recognition and Identification (APRI) system, a Bayesian supervised machine-learning system. Comparisons between APRI and standard methods such as discriminant analysis and recursive tree builders can be found elsewhere [Ezawa and Schuermann, 1995]. Instead, the large call-detail datasets mentioned above will be used here to compare the performance of several conditionally-independent probabilistic models to the performance of conditionally-dependent models constructed by APRI.

II. The Bayesian Network Approach

Theoretically, the *Bayesian Classifier* [Fukunaga, 1990] provides optimal classification performance. As a practical matter, however, its implementation is infeasible. Recent advances in evidence propagation algorithms [Shachter 1990, Lauritzen 1988, Pearl 1988, Jensen 1990, Ezawa 1994] and computer hardware allow us to approximate the ideal Bayesian classifier using Bayesian network models [Cheeseman 1988, Herskovits 1990, Cooper 1992, Buntine 1993, Langley 1994, Provan 1995].

A. The Bayesian Network

The classification problem can be addressed using the joint probability $\Pr\{\pi, \mathbf{X}\}$ of classes or populations π and the variables $\mathbf{X}$ that describe the data. The classification of an observation is based on the conditional probability $\Pr\{\pi \mid \mathbf{X}\}$. Assessing this probability directly is often infeasible due to computational resource limitations. The conditional probability of the attributes given the classes, $\Pr\{\mathbf{X} \mid \pi\}$, and the unconditional probability of the classes, $\Pr\{\pi\}$, are often assessed instead by analyzing a preclassified training data set. With those probabilities in hand, Bayes' rule then yields the desired conditional probability $\Pr\{\pi \mid \mathbf{X}\}$.

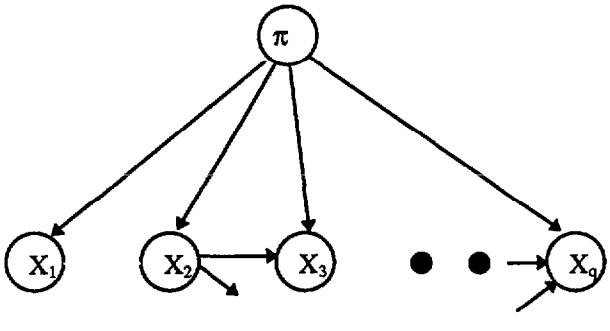

Figure 1: A Bayesian Network Model

Rewriting $\Pr\{\pi, \mathbf{X}\}$ as a product distribution can shed light on the underlying probabilistic structure. Figure 1 depicts a model where $\Pr\{\pi, \mathbf{X}\}$ is further factored to attribute level relationships, in particular

$$\Pr\{\pi\} \times \Pr\{X_1 \mid \pi\} \times \Pr\{X_2 \mid \pi\} \times \Pr\{X_3 \mid X_2, \pi\} \times \ldots$$

In this standard graphical notation, the conditional probability of a variable depends only on its parents. And given its parents, a variable is conditionally independent of the other variables in the network.

B. APRI

The Advanced Pattern Recognition & Identification (APRI) system developed at AT&T Bell Labs is a Bayesian network-based supervised machine learning system that constructs graphical probability models like those described earlier, using the entropy-based concept of mutual information to perform dependency selection. It first selects a set of variables and then a set of dependencies among the selected variables using heuristically identified mutual information thresholds. We settled on this approach to reduce the training time with special emphasis on repeated reading of the training dataset. APRI is extremely efficient in this regard. In fact, it reads the training dataset no more than five times.

APRI constructs graphical probability models using a four-step process. It takes three inputs: a database of training cases and two parameters T_{pf} and T_{ff}, each between zero and one. T_{pf} governs variable selection, and T_{ff} governs selection of variable-to-variable links for the final model.

APRI first scans the database to identify the outcome sets for each variable. For continuous variables it either estimates the kernel density or uses information-based discretization. If the class variable is continuous in the latter case, APRI first defines the class outcomes either by discretization or kernel density estimation.

In the second step, APRI chooses the variables for the final model. It computes the mutual information between the class node and each variable, then ranks the variables accordingly. Without loss of generality, let the indices from 1 to K reflect the ranking, so that $I(\pi; X_1) \geq I(\pi; X_2) \geq I(\pi; X_3)$ and so on. APRI selects the smallest number of variables J out of the entire pool of K variables, such that

$$\sum_{i=1}^{J} I(\pi; X_i) \geq T_{pf} \sum_{i=1}^{K} I(\pi; X_i)$$

In other words, the parameter T_{pf} establishes a mutual information threshold for choosing relevant variables. A value of 1 indicates that all the variables should be incorporated in the model. Lesser values indicate that less informative variables should be excluded. In the final graphical model, the class node becomes the parent of each of the selected variables.

The third step is akin to the second one, save that it identifies relationships between variables. In particular, it computes the conditional mutual information $I(X_i; X_j \mid \pi)$ between pairs of the J previously identified variables, where $i \neq j$. These candidate links are rank ordered. The highest ranked are then selected until the cumulative value is just T_{ff} times the total conditional mutual information. Directionality of these links is based on the mutual information variable ranking determined in the second step.

In the fourth and final step, APRI computes $\Pr\{\pi\}$ and $\Pr\{X_i \mid C(X_i)\}$ where $C(X_i)$ represents the parents of X_i, including the class node π.

Reading the dataset from secondary storage is a key element missing from analyses that assume data are stored in fast random-access memory [Herskovits 1990, Cooper 1992, Quinlan 1993, Provan 1995]. For problem domains like ours, the luxury of sufficient random-access memory is unlikely to be available in the near term. APRI is quite efficient in this regard, reading the database just four or five times: once in the first step for discrete classes or twice for continuous classes, then once in each of the three remaining steps.

C. Alternative Methods

A number of other authors have developed algorithms that search for graphical models by computing joint probabilities $P\{B_s,D\}$, where B_s is a Bayesian network structure and D a dataset [Herskovits 1990, Cooper 1992, Heckerman 1994, Provan 1995]. This kind of approach was not taken here, because it is impractical for applications with massive datasets.

Creation of the data structures that support these programs could be a problem. In K2 the supporting data structure is the *index tree* [Cooper, 1992, pp. 316-317]. Just two variables with 1,000 or more outcomes (*e.g.*, originating and terminating cities) could yield an index tree with 1,000,000 cells or more. If even a few more variables with small outcome sets are brought together as the parents of the same node, constructing and storing the index tree would be infeasible in typical computing environments. K2's search exacerbates this difficulty by considering all possible links at each step, eliminating no attributes from consideration and making no distinctions based on the size of outcome sets.

Even if these programs succeed in creating and storing their support structures, they face another run-time problem. If a dataset is too large to hold in memory, they must read it at least once for each arc in the final graphical model. For K2, it appears that the dataset would have to be read $O(n(u+1))$ times to create a model, where n is a number of nodes and u the maximum number of parents per node. With 33 variables and allowing 2 parents per node, K2 might need to read the dataset 99 times. If the training data consists of several million records and perhaps hundreds of millions of bytes of data, as in the application described here, reading and re-reading the data will become the limiting factor. APRI reads the dataset just four or five times during model creation, for any n and u.

Still other authors have investigated the possibility of very simple probabilistic models [Langley and Sage, 1994]. The naive Bayesian classifier estimates probabilities using a fully independent model incorporating all the given variables. That is:

$$\begin{aligned}\Pr\{\pi, \mathbf{X}\} &= \Pr\{\mathbf{X}|\pi\} \times \Pr\{\pi\} \\ &= \Pr\{X_1|\pi\} \times \Pr\{X_2|\pi\} \times \ldots \times \Pr\{X_n|\pi\} \times \Pr\{\pi\}\end{aligned}$$

Recognizing that certain variables might be irrelevant or even damaging to the classification, Langley and Sage also implement a selective Bayesian classifier that assumes independence but uses a forward search to develop a limited variable set and uses an error metric as a stopping criterion. In their experiments, the selective classifier is never significantly worse than the full independent naive classifier, and is often significantly better.

Conditionally independent models are easier to create and require much less storage than conditionally dependent models. Of course, if the variables are in fact conditionally dependent in important ways, the accuracy of independent Bayesian classifiers will suffer. Conditionally dependent models with many interrelated attributes tend to require more time to create and more space to store. And of course, additional dependencies require additional data if the learned model is to be robust.

III. Predicting Uncollectible Debt

Four different models are compared in this section. The dependent models were constructed with APRI, using a 95% cumulative entropy threshold for attribute selection and a 25% or 45% cumulative entropy threshold for dependency selection (T_{pf} and T_{ff} respectively). A fully independent model was also constructed (a naive Bayesian classifier), as was an independent model limited to the attributes selected by APRI (analogous to the selective Bayesian classifier). Because the probability models do not themselves output a binary classification, an uncollectible probability threshold is needed to perform classification. In one set of experiments the uncollectible probability threshold is 50%, a standard value. It was set at a more conservative 70% level in a separate experiment.

The training dataset consists of 4,014,721 records described by 33 attributes. The size of the dataset is 585 million bytes. The attributes are a mixture of call-detail information and customer-summary information. The

baseline probability of a call going bad is 9.90%. Training the conditionally dependent model with APRI took about 4 hours on a SUN Sparc 20 workstation.

For a practical application, we need to train the model on data from a particular period and test it on data from a subsequent period. This means that out-of-sample prediction is not with a typical hold-out sample of the same period but a full set from a subsequent period. Testing models this way is more demanding but also more realistic. The testing dataset used here is indeed from a later period. It contains 5,351,834 records (773 million bytes) and has an 11.88% unconditional probability of a call going bad.

Table 1 shows the results of applying the learned models to this separate dataset taken from a subsequent period. The numbers in square brackets indicate the total number of calls for that cell. Because of the nature of the marketplace, incorrectly classified collectible calls must be minimized at the same time that the number of correctly classified uncollectible calls is maximized. This suggests examining the volume ratio in the fourth column, a quantity that should ideally be minimized. One further point is worth noting. Because of the nature of the business, learned classifiers will never have an impact operationally unless they correctly classify a high enough percentage of uncollectible calls; otherwise it would be better to do nothing. Some classifiers with appealing volume ratios would not meet this requirement.

Table 1: Out-of-Sample Evaluation

Model Type	ICC	CCU	VR
Full Ind. ($\geq$ 50% *)	14.01 % [660,854]	35.06 % [222,874]	2.97 : 1
Full Ind. ($\geq$ 70%)	7.95 % [374,969]	22.94 % [145,799]	2.57 : 1
Limited Ind. ($\geq$ 50%)	4.32 % [203,601]	14.73 % [93,600]	2.18 : 1
Limited Ind. ($\geq$ 70%)	1.48 % [69,752]	5.92 % [37,635]	1.85 : 1
APRI.ff25 ($\geq$ 50%)	6.87% [323,851]	29.53 % [187,678]	1.7:1
APRI.ff25 ($\geq$ 70%)	2.92 % [137,943]	17.56 % [111,606]	1.2:1
APRI.ff45 ($\geq$ 50%)	6.57% [309,784]	31.86 % [202,500]	1.5:1
APRI.ff45 ($\geq$ 70%)	2.85 % [134,305]	21.10% [134,131]	1.0:1

ICC: Incorrect Classification of Collectible Calls
CCU: Correct Classification of Uncollectible Calls
VR: Volume Ratio -- ICC vs. CCU
*: Uncollectible-Call Probability Threshold

APRI's dependent models do a good job at out-of-sample prediction. At the 50% prediction level, *i.e.* when classifying a call as uncollectible requires the predicted probability to be more than 50%, the model identifies about 30% of all uncollectible calls and they missclassify about 6-7% of collectible calls. If we raise the predicted probability threshold to a more conservative 70%, the ratio of false positives to true positives improves even more (VR Column). Being more conservative means that we let a few bad calls slip through, but gain by falsely identifying as uncollectible far fewer calls that are in fact collectible. Neither of the independent models does as well. At any rate, this table suggests that a more conservative threshold may indeed be appropriate for our application.

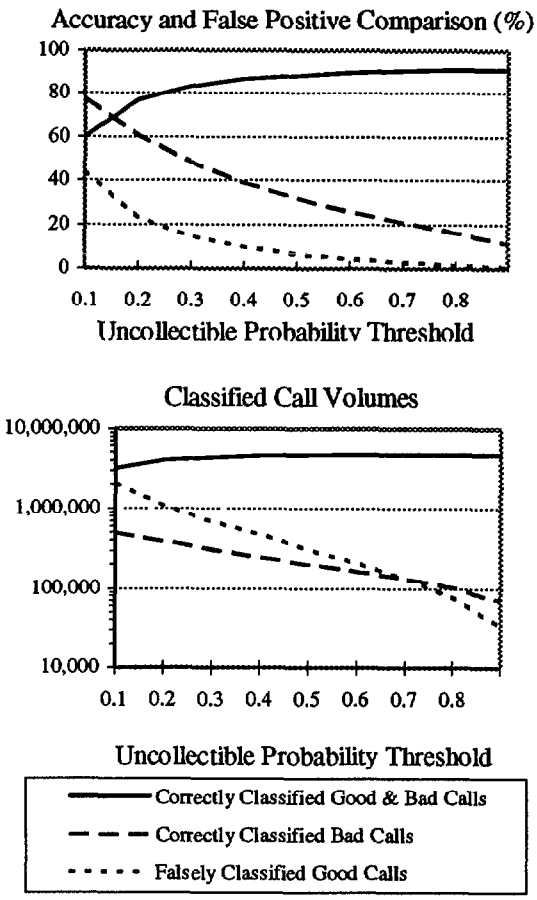

Figure 2: Accuracy & Call Volume *vs.* the Uncollectible Probability Threshold

Figure 2 provides more detail about how changing the uncollectible probability threshold in the range of 10% to 90% affects the proportion of correctly classified calls (collectible & uncollectible), correctly classified uncollectible calls, and falsely classified collectible calls. (It was generated using APRI with a 45% field-to-field

cumulative entropy threshold.) The higher the uncollectible probability threshold, the more conservative we are about classification. Fewer good calls are falsely classified, but at the expense of correctly identifying fewer bad calls.

If one were to focus only on the top half of Figure 2 (Accuracy and False Positive Comparison), one might accept a lower, more aggressive uncollectible probability threshold since throughout the graph, the percentage of correctly classified uncollectible calls (the true positives) is larger than the percentage of incorrectly classified collectible calls (here, the false positives). However, because of the disparity in class proportions, looking at the classification percentages is insufficient. The better thing to do is to examine call volumes as the predicted probability threshold increases.

The bottom half of Figure 2 also shows how the corresponding call volumes change in response to changes in the uncollectible probability threshold. (The vertical axis is in a log scale.) This Figure shows that a more conservative approach is indeed warranted. Only at a threshold of around 70% does the volume of true positives exceed the volume of false positives.

IV. Discussion and Summary

This paper has presented a method for learning Bayesian network models from extremely large datasets. In such cases, the processing bottleneck is likely to be repeatedly reading the training dataset from secondary storage. APRI reads the dataset a constant number of times, not a number of times linear in the size of the final network. It does this by thresholding mutual information to select attributes and dependencies. While this certainly is a coarse heuristic, there is reason to believe it's on the right track. In particular, when APRI's heuristic was used to select the variables in the limited independent model, out-of-sample performance improved over the full independent model in every instance, as measured by the volume ratio.

Overfitting of the learned model to the training data is a potential problem for all inductive learners. In Bayesian network models this problem can manifest itself in probabilities 0 and 1. In APRI the avoidance of probability 1 in the classification process provides protection against overfitting. This feature allows us to separate the impact of model complexity from overfitting. During classification APRI simply skips a variable which causes a classification probability to be one, just as if it was a missing value. Note that the effect of this "pruning" step is not explicitly observable on the network itself, but is implicit in the evaluation of probabilities.

One of the interesting features of predicting uncollectible debt is the requirement of genuine out-of-sample testing datasets from separate time periods. Such testing is essential because of the inevitable lag between model creation and model deployment. Of course, there is the risk that fraud or uncollectible patterns will change in the interim. Seasonal variations could even interfere. Given enough data, these effects might be modeled. In addition, active network policies could also change observed patterns of activity, although there is probably less hope of modeling the effects of untried policies. Despite these potential pitfalls, subsequent-period out-of-sample prediction will remain the real litmus test for this application.

Lastly, it has become popular in the recent literature to invoke a *gold standard network* in the testing and evaluation of learning methods [Chickering, 1995]. At this stage of our research a gold standard is unavailable and probably inappropriate, because an inevitable side-effect of active policies is a change in the patterns of activity related to uncollectible debt. The reason is that the fraud and uncollectible debt problem environment is not static, but dynamic. Whether policies are influenced by APRI or not, today's policies directly impact tomorrow's uncollectible environment. The better the uncollectible debt model, the faster its obsolescence. Success in detecting and treating the uncollectible accounts or calls, in part, changes the character of future uncollectible accounts and calls. The uncollectible problem simply evolves into a different shape and adjusts to the new environment. We will not find a gold standard Bayesian network structure; it changes from moment to moment. Instead, we choose to emphasize the ability to create good models in reasonable amounts of time using very large datasets.

We have demonstrated the performance of APRI, a Bayesian network learning system, on a problem with a rare binary outcome, mixed data types, and extremely large datasets. The need for a probabilistic assessment and the abundance of data limit the set of learning algorithms we can consider. When compared to several conditionally independent probability models, the conditionally dependent models created by APRI do quite well.

Acknowledgments

We are grateful to Til Schuermann and to the referees for their careful readings and constructive comments.

References

Baldi, P., and Chauvin, Y., 1991, "Temporal Evolution of Generalization During Learning in Linear Networks," *Neural Computation*, 3, pp. 589-603.

Buntine, W.L. and Smyth, P., 1993, "Learning from Data: A Probabilistic Framework," Tutorial Program, Ninth Conference on Uncertainty in Artificial Intelligence.

Catlett, J., 1995, "Tailoring Rulesets to Misclassification Costs", Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D., 1988, "AUTOCLASS: A Bayesian Classification System," *Proceedings of the Fifth International Conference on Machine Learning* pp. 54-64, Morgan Kaufmann.

Chickering, D.M., Geiger, D., and Heckerman, D., 1995, "Learning Bayesian Networks: Search Methods and Experimental Results", Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics.

Cooper, G.F. and Herskovits, E., 1992, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, **9**, pp. 309-347.

Ezawa, K.J., 1993, "A Normative Decision Support System", Third International Conference on Artificial Intelligence in Economics and Management.

Ezawa, K.J., 1994, "Value of Evidence on Influence Diagrams", *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 212-220, Morgan Kaufmann.

Ezawa, K.J., and Schuermann, T., 1995, "Fraud/Uncollectible Debt Detection Using a Bayesian Network Based Learning System: A Rare Binary Outcome with Mixed Data Structures," forthcoming in the Eleventh Conference on Uncertainty in Artificial Intelligence.

Fukunaga, K., 1990, *Introduction to Statistical Pattern Recognition*, Academic Press.

Heckerman, D.E., Geiger, D., and Chickering D.M., 1994, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 293- 301.

Herskovits, E.H., and Cooper, G.F., 1990, "Kutato: An entropy-driven system for the construction of probabilistic expert systems from databases", *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 54 - 62.

Jensen, V., Olesen K. G., and Anderson S. K., 1990, "An Algebra of Bayesian Universes for Knowledge-Based Systems," *Networks*, **20**, pp. 637-659.

Langley, P. and Sage, S., 1994, "Induction of Selective Bayesian Classifiers," *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 399-406, Morgan Kaufmann.

Lauritzen, S.L., and Spiegelhalter, D.J., 1988, "Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems," *J. R. Statistics Society, B*, **50**, No. 2, pp. 157-224.

Pazzani, M., C. Merz, P. Murphy, K. ali, T. Hume and C. Brunk, 1994, "Reducing Misclassification Costs", in *Proceedings of the International Conference on Machine Learning*, pp. 217-225, Morgan Kaufmann

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann.

Provan, G.M., and Singh, M., 1995, "Learning Bayesian Networks Using Feature Selection," Preliminary Papers of International Workshop on Artificial Intelligence and Statistics, pp. 450 - 456.

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

Shachter, R. D., 1990, "Evidence Absorption and Propagation through Evidence Reversals", *Uncertainty in Artificial Intelligence*, Vol. 5, pp. 173-190, North-Holland.