

Guideline Generation from Data by Induction of Decision Tables Using a Bayesian Network Framework

Subramani Mani, MBBS, MS, Michael J. Pazzani PhD

Dept. of Information and Computer Science, University of California, Irvine CA 92697

Decision tables can be used to represent practice guidelines effectively. In this study we adopt the powerful probabilistic framework of Bayesian Networks (BN) for the induction of decision tables. We discuss the simplest BN model, the Naive Bayes and extend it to the Two-Stage Naive Bayes. We show that reversal of edges in Naive Bayes and Two-stage Naive Bayes results in simple decision table and hierarchical decision table respectively. We induce these graphical models for dementia severity staging using the Clinical Dementia Rating Scale (CDRS) database from the University of California, Irvine, Alzheimer's Disease Research Center. These induced models capture the two-stage methodology clinicians use in computing the global CDR score by first computing the six category scores of memory, orientation, judgment and problem solving, community affairs, home and hobbies and personal care, and then the global CDRS. The induced Two-Stage models also attain a clinically acceptable performance when compared to domain experts and could serve as useful guidelines for dementia severity staging.

INTRODUCTION AND BACKGROUND

The goal of evidence-based medicine is to improve the quality of patient care while keeping it cost-effective. This in turn mandates development of efficient practice guidelines. Traditionally, the task of guideline generation has been performed by a committee of experts by looking at published evidence and then arriving at a consensus. The format of the guidelines also varies from a set of *if...then rules* to flowcharts. Since formal verification methods are seldom employed to check for completeness, ambiguities lie hidden in the guidelines. Usually, the inadequacies are noted at the time of implementation of the guidelines in a hospital or an out-patient clinical practice setting. This leaves room for a subjective interpretation (see [1] for a good review). Researchers have argued for the feasibility, utility and advantages of the format of decision tables for guidelines [2], [1]. The decision table format is expressive, the flow of logic clear and it could easily be implemented as a set of *if...then rules* or *case* statements. An enriched form of the decision table including de-

tailed documentation in the form of background evidence, pointers to literature etc. called the *augmented decision table* has also been proposed [1]. With the advent of Electronic Medical Records (EMR) plus the availability of research data in electronic format, there has been a surge of interest in generating practice guidelines and disease models from data [3], [4]. In this paper we propose a method of induction of decision tables from data using a Bayesian Network framework. In the ensuing section we discuss the key ideas of our graphical models. We then describe the dataset used for this study, the algorithms used, and the experimental set up. Next we present our results and discuss the significance of our induction approach and compare it with related work. Finally, we conclude by summarizing our contributions.

Decision Tables and Bayesian Networks

Bayesian Networks (BN) have emerged in recent years as a powerful formalism to handle uncertainty in complex domains. They offer a compact and expressive graphical structure to represent knowledge, and researchers have developed powerful algorithms for propagating evidence across the network. Qualitatively BNs are directed acyclic graphs, and the direction of arrows could be interpreted as causal influences when such a semantics is present in the domain. These structures are strengthened quantitatively by the incorporation of prior (for nodes without parents) and conditional probabilities. The probabilities can easily be estimated from data or could be assigned by an expert. See [5] for a general discussion of BNs and [6] for a good review of BN model building issues. But the use of BNs in generating disease models or practice guidelines have lagged behind probably because of the computational complexity in the general case [7]. This work explores special classes of BN models representing simple and hierarchical decision tables. We also introduce the Two-Stage Naive Bayes model extending the Naive Bayes structure. Naive Bayes (NB) [8] is the simplest Bayesian network model for any domain. In this BN all the attributes are conditioned on the class node only and there are no other interactions. Even though it assumes that the attributes are independent given the class, it

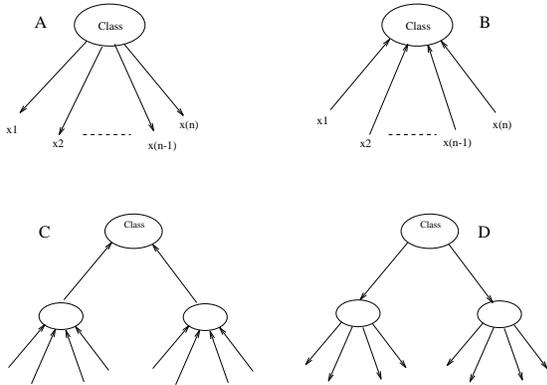


Figure 1: Decision Tables as Graphical Models

is a robust classifier and is known to perform very well on real-world problems. From a comprehensibility point of view it has great limitations. There is no meaningful structure to the model and hence no intermediate concepts can be ascertained. See Figure 1, A.

A Simple Decision Table (**SDT**) can be viewed as a Naive Bayes structure with the direction of edges reversed i.e. the class or outcome node is conditionally dependent on all the attributes. Hence we also refer to it as Reverse Naive Bayes. See Figure 1, B. This framework is infeasible if the number of attributes is very large—there will be 2^n rows in the SDT where n is the number of attributes. Here we make the assumption that the attributes of the SDT have been identified by an expert to be included in the model or guideline and n is reasonably small. Hence we do not focus on feature selection or reducing space complexity even though the Hierarchical Decision Table and the Two-Stage Naive Bayes which are described below address this issue to some extent.

The Hierarchical Decision Table (**HDT**) is composed of SDTs linked hierarchically so that intermediate concepts can be represented. This type of model, apart from reducing complexity makes concept representation expressive. This is well suited for guideline representation if the domain semantics support such a structured layout of attributes. See Figure 1, C. If the intermediate concepts are directly available, that part of the model alone is required for decision making. They could also be derived from the primitive attributes of the first layer.

The Two-Stage NB is an interesting model capable of representing intermediate concepts and is computationally very efficient. Each node has a maximum of one parent and the edges are directed away

from the class (outcome) node to the intermediate concepts which in turn have edges directed to the others. See Figure 1, D. Reversing the direction of the edges in this model yields the HDT model. Even though it is an extension of the NB model, it is more expressive as it can represent intermediate concepts.

There are many advantages of approaching Decision Tables from a BN perspective. First, the model generation fits into a probabilistic and graph theoretic framework which has a solid mathematical foundation. Second, there are efficient algorithms to propagate evidence across these types of models to compute classification or outcome. Third, the framework offers a mechanism to improve the classification accuracy of the DT while retaining their expressive power. Fourth, extending the basic model to construct HDT and Two-Stage NB can be handled easily within the framework, for estimating the probabilities from data and for evidence propagation.

METHODS

Sample Description

We illustrate these models using the Clinical Dementia Rating Scale (**CDRS**) [9] dataset obtained from the UCI Alzheimer’s Disease Research Center (**ADRC**). The staging of dementia severity using the CDRS is done by clinicians in two steps. First, the clinician uses CDRS criteria to rate the patient’s level of impairment for each of the six categories of memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Second, he or she uses another set of CDRS criteria to assign a global level of impairment (global CDRS score) based on the six CDRS category scores. This complex clinical task calls for extending SDT and NB, to HDT and the Two-Stage NB models described earlier. The sample consisted of the initial visits of 678 subjects seen at the University of California, Irvine Alzheimer’s clinic between 1988 and 1996. Subjects ranged from normal aging to severely demented and were carefully assessed and diagnosed using a standard, four hour battery of tests which included the CERAD protocol for diagnosing Alzheimer’s Disease. In order to eliminate inter-observer variability in applying the Alzheimer’s Disease Cooperative Study Unit (**ADCSU**) criteria [10] for scoring the CDRS global and category scores, the ADRC clinicians spent two years developing and validating a computerized scoring algorithm of the CDRS category and global scores [11] and these derived scores

served as the gold standard for generating and evaluating our models.

Component Architecture

The various components incorporated in our architecture included the Train Test Partitioner, Discretizer, CPB Module, CAP module, HUGIN and the Classification module. Figure 2 shows how the different components of the datamining environment are integrated. We now describe the important components.

CPB Module This module implemented in C, computes the prior and conditional probabilities taking a dataset and any arbitrary Bayesian Network structure as input. The output is a Bayesian Network incorporating the probabilities and formatted in the NET language, which can be read and compiled by HUGIN.

CAP Module Also written in C, this module calls the HUGIN API functions. It facilitates compilation of the network, propagation of evidence provided as test set instances, and outputs the posterior probability vector of the outcome (class) node. Automated batch processing of the test set is made feasible using this component.

HUGIN The HUGIN [12] module is a commercial package for constructing BNs.

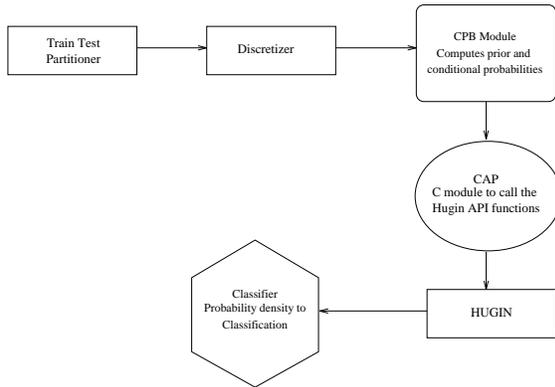


Figure 2: Graphical Model Building Components

Experimental Methods

We generated NB and SDT models (see Figure 1, A and B) for the first stage i.e. computing the six category scores of memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. For example, six attributes measuring activities of daily living—eating, dressing, grooming, medication taking, bathing and elimination were used to compute the category score of *personal care*. We also generated these simple models to compute the global CDR score from the gold standard category scores. We then modeled the complete Two-Stage methodology clinicians employ in computing global CDRS.

This was done by generating HDT and the Two-Stage NB models. We compared these results with the NB model. For ascertaining the global CDRS using these models, we did not use the gold-standard category scores in the test instances.

The datasets were randomly partitioned into a two-third training and one-third test sets. The training set was used to learn the probabilities and the test set was used for validation. This was repeated thirty times and the results of the test sets are reported. For each test instance, we got a probability density vector as output showing the posterior probability of each class state after evidence propagation. This was further processed by taking the class with the *maximum posterior probability* as the classification. Ties were resolved randomly.

RESULTS

Table 1 gives the classification accuracy of the various models. For the orientation category score both the NB and SDT models obtained an accuracy of 100%. For the category scores of memory, community affairs, home and hobbies, and personal care the accuracy for the NB model varied between 81.82% and 95.67% and the for the SDT between 53.91% and 83.48%. For the category judgment and problem solving, NB accuracy was 74.78% and that of SDT 65.80%. Except for the category of orientation, the classification accuracy of SDT was lower by 8.98% to 30.44% when compared to NB. On the other hand for the computation of global CDR score from the gold standard category scores the accuracy of the NB model was 81.86% and that of SDT 67.84%. Computation of the global CDR score using the category scores and the raw attributes (features used to compute the category scores) resulted in an accuracy of 72.12% for the NB model. The performance of the HDT model was comparable with an accuracy of 72.44%. The best performance was that of the Two-Stage NB model with an accuracy of 74.78%.

DISCUSSION

A comparison of the classification accuracies of the SDT and NB models for the category scores shows that except in the case of orientation category score, the NB models outperformed the SDT. The difference in accuracy was the least for the subcategory personal care which had only four classes while the others had five. In the case of orientation category, both the models achieved 100% accuracy. This category had only two features and four classes and hence both the models could easily learn the target concept.

Table 1: Mean Classification Accuracy (with standard deviation within braces) of the various models for CDRS category and global scores

CDR Scale	Classification Accuracy in %			
	NB	SDT	HDT	TS NB
Memory	83.67(2.5)	70.80(4.4)	–	–
Orientation	100.00	100.00	–	–
JPS	74.78(2.5)	65.80(3.8)	–	–
Community Affairs	84.35(2.4)	53.91(3.9)	–	–
Home & Hobbies	81.82(2.3)	56.65(2.2)	–	–
Personal Care	95.67(1.4)	83.48(2.5)	–	–
Global CDRS	72.12(2.8)	–	72.44(3.0)	74.78(2.1)

HDT: Hierarchical Decision Table, SDT: Simple Decision Table, NB: Naive Bayes, TS NB: Two-Stage NB, JPS: Judgment & Problem Solving

From a guideline perspective, decision tables are very expressive. They could be used as paper based or online charts. The online charts could further be hyperlinked to take the format of augmented decision tables [1]. The decision tables could also be executed as algorithms [13]. And since the decision tables represent all possible combinations of the feature space as the condition clause (each row being a unique instantiation), the format prevents ambiguities in the guidelines. There are various ways to reduce the dimensionality of the tables. It could be done by an expert or by automated methods. Since the focus of the paper is not on refinement of the tables, we will not go into the details here. The reader is referred to [1] for the relevant discussion and other pointers to literature.

Why Naive Bayes has higher accuracy than SDT?

Even though Naive Bayes has serious limitations (for example, it cannot learn non-linearly separable binary concepts), it is known to be a high performance classifier on most real-world datasets. Structurally, it is a very simple network. See Figure 1, label A. All the attributes are conditioned only on the class node. Since it is reasonable to expect that much care goes into getting the class labels right (any classifier will be thrown off the mark if there is significant labeling error), the classifier gains strength by conditioning *only* on the class node. The probability table for any attribute X_i given the class C , turns out to be reliable as the whole training sample goes into this computation. The NB equation follows.

$$P(C|X) = P(C) \frac{P(X|C)}{P(X)}$$

where

$$P(X) = \prod_{i=1}^n P(X_i); \quad P(X|C) = \prod_{i=1}^n P(X_i|C)$$

and n denotes the number of attributes of the feature vector X .

Now consider a situation where the class C is conditioned on π attributes ($\pi_1 \dots \pi_n$ parents). Note that for the SDT, π is X . In this situation we have to compute $2^{(n+1)}$ probabilities (assuming that the parent set P and the class C are binary), as opposed to just $4n$ values in total, for NB. And for computing $2^{(n+1)}$ probabilities, the sample (training) would have to be partitioned into 2^n groups. It is not difficult to see how unreliable the probabilities can turn out to be if the sample size of the partitions are not sufficiently large.

Guideline Model

SDTs are good guideline models since they are expressive, and we propose that SDTs incorporated with class labels generated using the corresponding NB model can serve as effective guidelines. Such models will acquire the classification accuracy of NB while retaining the flexibility of the decision table format.

For the computation of the global CDR score using the category scores and the features used to compute the categories, the NB model and the HDT model gave comparable accuracy, the HDT 0.32% more than NB. But the Two-Stage NB model performed best with an accuracy of 74.78%, a 2.34% increase over HDT. The HDT and the Two-Stage NB follow the two stage approach clinicians use in computing the global CDR score. The only difference between the two models is in the directionality of the edges in the framework of the Bayesian Network. The HDT is very expressive and could be used as a paper-based or online

guideline for computing the global CDR score in two stages. The HDT can be labeled with Two-Stage NB generated class labels to increase the accuracy to about 75%. This compares very favorably with the performance of human experts. The inter-rater agreement described in literature using much smaller sample sizes is about 80% [14].

Related Work

Langley [15] and Kohavi [16] have described algorithms to induce decision tables from data. Their approach is a combination of feature subset selection and computing the probabilities from data for the table. We have extended their work most importantly in two ways. First, we propose a Bayesian Network framework for the problem and show that SDT can be looked at as reverse NB. Second, we develop the HDT and Two-Stage NB models using this approach.

CONCLUSION

We have used a Bayesian Network framework for the induction of simple and hierarchical decision tables. We show that these models could be considered as Naive Bayes and Two-Stage NB with the direction of edges reversed. By using the NB and Two-Stage NB models in class label assignment of SDT and HDT, we increase their classification accuracy while retaining their expressive power. By developing this approach and illustrating it using a real medical dataset to generate clinically useful guidelines, we have shown the efficacy of this methodology. We propose that such decision tables are effective as guidelines by generating models for clinical dementia rating score computation with a performance in accuracy approaching that of domain experts.

Acknowledgements

The authors are grateful to Dr. W.R. Shankle for making available the CDRS dataset. We also thank him for explaining to us in detail the methodology clinicians use in staging dementia severity.

References

1. Richard N. Shiffman. Representation of Clinical Practice Guidelines in Conventional and Augmented Decision Tables. *JAMIA*, 4(5):382–393, 1997.
2. R.R. Holland. Decision Tables: Their use for the Presentation of Clinical Algorithms. *JAMA*, 233:455–7, 1975.
3. GF Cooper, CF Aliferis, R Ambrosino, et al. An evaluation of machine learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.
4. KC Abston, TA Pryor, PJ Haug, and JL Anderson. Inducing practice guidelines from a hospital database. In *JAMIA Supplement*, pages 168–172, 1997.
5. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 1991.
6. David J. Spiegelhalter, Philip A. Dawid, Steffen L. Lauritzen, and Robert G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.
7. GF Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.
8. RO Duda and PE Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.
9. JC Morris. The clinical dementia rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412–4, Nov 1993.
10. JC Morris, LA Coben, and EH et al. Rubin. Clinical Dementia Rating. In M Bergener and SI Finkel, editors, *Treating Alzheimer's and other dementias: clinical applications of recent research advances*, pages 338–346. Springer, New York, 1995.
11. Subramani Mani, William R. Shankle, Malcolm B. Dick, and Michael J. Pazzani. Two-Stage Machine Learning Model for Guideline Development. (In Press), 1998.
12. S.K. Andersen, K.G. Olesen, F.V. Jensen, and F. Jensen. HUGIN—A shell for building bayesian belief universes for expert systems. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1080–1085, San Francisco, CA, 1989. Morgan Kaufmann.
13. J Vanthienen and G Wets. From decision tables to expert system shells. *Data and Knowledge Engineering*, 13:265–282, 1994.
14. WJ Burke, JP Miller, EH Rubin, JC Morris, LA Coben, J Duchek, IG Wittels, and L Berg. Reliability of the washington university clinical dementia rating. *Archives of Neurology*, 45(1):31–2, 1988.
15. Pat Langley. Induction of condensed determinations. In *Proceedings of the Second International Conference of Knowledge Discovery and Data Mining*, pages 327–330. AAAI Press, 1996.
16. Ron Kohavi. The Power of Decision Tables. European Conference on Machine Learning (ECML), 1995.