Data Mining and Model Simplicity: A Case Study in Diagnosis

Gregory M. Provan
Rockwell Science Center
1049 Camino dos Rios
Thousand Oaks, CA 91360.
provan@risc.rockwell.com

Moninder Singh
Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389
msingh@gradient.cis.upenn.edu

Abstract

We describe the results of performing data mining on a challenging medical diagnosis domain, acute abdominal pain. This domain is well known to be difficult, yielding little more than 60% predictive accuracy for most human and machine diagnosticians. Moreover, many researchers argue that one of the simplest approaches, the naive Bayesian classifier, is optimal. By comparing the performance of the naive Bayesian classifier to its more general cousin, the Bayesian network classifier, and to selective Bayesian classifiers with just 10% of the total attributes, we show that the simplest models perform at least as well as the more complex models. We argue that simple models like the selective naive Bayesian classifier will perform as well as more complicated models for similarly complex domains with relatively small data sets, thereby calling into question the extra expense necessary to induce more complex models.

Introduction

In any data mining task, one key question that needs to be determined is the type of model that one attempts to learn from the database. One rule of thumb is to try the simplest model first, and see how well the model fits the data, incrementally increasing model complexity to try to obtain better fit of model to data. During this process, attributes that are determined to be irrelevant to the particular task (e.g., classification) may be deleted from the database, other attributes included, etc.

In this paper we present results on applying data mining techniques to a medical diagnosis domain, the diagnosis of acute abdominal pain. The diagnosis of acute abdominal pain is well known to be a difficult task both for physician and machine. Depending on the assumptions used in reporting statistics the accuracy rates vary, but most machine diagnostic systems achieve accuracy of little more than 60% (Todd & Stamper 1994). Moreover, there has been great debate in the literature about whether the naive Bayesian classifier is optimal for this domain (Todd & Stamper 1994). The naive Bayesian model makes the strong assumption that the attributes are conditionally independent given the class variable, yet has been shown to perform remarkably well in this domain, and possibly better than any other approach (Todd & Stamper 1994).

One paradoxical question we attempt to address is that fact that no approach outperformed the naive Bayesian classifier on this domain. In particular, we examine two questions pertaining to model simplicity in this domain: (a) does modeling attribute dependence given the class variable improve performance?; and (b) how many attributes facilitate accurate diagnosis? In addressing hypothesis (a), we compare the performance of the naive Bayesian classifier with that of the Bayes network classifier, an extension of the naive Bayesian classifier that models attribute non-independence given the class variable. The Bayesian network classifier (Singh & Provan 1995) has been shown to outperform the naive Bayesian classifier on several UC-Irvine domains, so it may prove better than the naive Bayesian classifier on this domain. In addressing hypothesis (b), we compare the performance of classifiers using all attributes with those using attributes selected by an attribute-selection algorithm that has produced small models whose accuracy rivals that of much larger models that contain all attributes (Provan & Singh 1996; Singh & Provan 1995). Clearly, since data collection can be a costly process, having the smallest model possible is an advantage if the small model has accuracy comparable to that of the full model.

The following sections of the paper describe in turn the Bayesian network representation and performance measure, the application domain, our experimental studies, and a summary of our contributions.

Bayesian Network Representation

This section summarizes the Bayesian network representation and the performance measure used in this paper.

Representation: A Bayesian network consists of a qualitative network structure $\mathcal{G}$ and a quantitative probability distribution θ over the network structure. The qualitative network structure $\mathcal{G}(N, V)$ consists of a directed acyclic graph (DAG) of nodes N and arcs V, where $V \subseteq N \times N$. Each node i corresponds to a discrete random variable A_i with finite domain Ω_{A_i}.

Arcs in the network represent the dependence relationships among the variables $A = \{A_1, A_2, ..., A_n\}$. An arc into node i from node j may represent probabilistic dependence of A_i on A_j, and is precisely specified using the notion of parents of a node. The parents of A_i, $pa(A_i)$, are the direct predecessors of A_i in $\mathcal{G}$. The absence of an arc from node i to j indicates that variable A_j is conditionally independent of variable A_i given $pa(A_j)$.

The quantitative parameter set θ consists of the conditional probability distributions $P(A_i|\, pa(A_i))$ necessary to define the joint distribution $P(A_1,A_2,..., A_n)$. We can write the unique joint distribution specified by $\mathcal{G}$ as

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i|pa(A_i)). \qquad (1)$$

The naive Bayesian classifier assumes that the attributes are conditionally independent given the class variable C. A naive Bayesian classifier is a Bayesian network whose structure is restricted to having arcs only from the class node to the feature nodes, i.e. the only parent node in Equation 1 is the node for C. The joint distribution is thus given by

$$P(C, A_1, A_2, \ldots, A_n) = P(C) \prod_{i=1}^{n} P(A_i|C). \qquad (2)$$

Performance: The performance of Bayesian networks is measured by conducting inference on the networks using belief updating method, such as Lauritzen-Spiegelhalter's (1988) clique-tree inference algorithm. Inference in naive Bayesian networks is linear in the number of attributes, and is done using Bayes' rule and the assumption of feature independence within each class. Inference is more complicated for Bayesian networks: it is NP-hard (Cooper 1990). Pearl (1988), among others, reviews Bayesian network inference; details are beyond the scope of this paper.

Application Domain

This section discusses the domain of acute abdominal pain, focusing on the models used for the diagnosis.

Diagnosis of Acute Abdominal Pain

The diagnosis of acute abdominal pain is considered to be a classic Bayesian problem, as findings are probabilistically (rather than deterministically) related to underlying diseases, and prior information can make a significant difference to a successful diagnosis.

The most serious common cause of acute abdominal pain is appendicitis, and in many cases a clear diagnosis of appendicitis is difficult, since other diseases such as Non-Specific Abdominal Pain (*NSAP*) can present similar signs and symptoms (findings). The tradeoff is between the possibility of an unnecessary appendectomy and a perforated appendix, which increases mortality rates five-fold. The high incidence of acute abdominal pain, coupled with the poor diagnosis accuracy, make any improvements in diagnostic accuracy significant.

The Use of (Naive) Bayesian Classifiers

A full model for this domain typically has three variable types: observable, intermediate (latent) and disease. Observable variables correspond to findings that can be observed directly, such as nausea, vomiting and fever. Disease variables correspond to diseases that are the underlying causes for a case of acute abdominal pain, such as appendicitis or NSAP. Latent variables correspond to physiological states that are neither directly observable nor are underlying diseases, but are clinically relevant (as determined by the domain expert) to determining a diagnosis. Examples include peritonitis and inflammation. Such models typically do not make strong assumptions about conditional independence of latent or observable variables given the disease variable. Models with such a structure are described in (Provan 1994; Todd & Stamper 1994).

A naive model typically ignores the class of latent variables, or if it includes any latent variables it assumes that they are independent of any observable variables given the disease variable. This latter assumption does not correspond to known physiological principles; in addition, including latent variables should improve diagnostic performance, since more information is being used. Hence, it appears that a full model should outperform a naive model.[1]

However, neither the empirical nor the theoretical evidence fully supports this hypothesis. The empirical evidence provides inconclusive evidence about the effect on diagnostic accuracy of capturing dependencies in Bayesian models. Following de Dombal et al.'s publication of a successful naive Bayesian model for the diagnosis of acute abdominal pain (de Dombal *et al.* 1972), many researchers have studied empirically the effect of independence assumptions on diagnostic accuracy. Some studies have demonstrated the influence on diagnostic accuracy of capturing dependencies. For example, Seroussi (1986) reported a 4% increase in diagnostic accuracy (from 63.7% to 67.7%) by accounting for pairwise interactions using a Lancaster model; other researchers (Fryback 1978; Norusis & Jacquez 1975) have shown that capturing conditional dependencies may improve diagnostic accuracy. In contrast, other studies have shown no statistically significant difference between the two approaches (Todd & Stamper 1993), and some have

[1] We ignore the cost of data collection in model construction.

even found independence Bayesian classifiers to be optimal (Todd & Stamper 1994; Edwards & Davies 1984; de Dombal 1991). Fryback (1978) has studied the sensitivity of diagnostic accuracy to conditional independence assumptions in a Bayesian model for medical diagnosis. He showed empirically that large models with many inappropriate independence assumptions can be less accurate than smaller models which do not have to make such inappropriate independence assumptions. Fryback suggests that model size should be increased incrementally in cases where conditional independence assumptions are not all known, rather than starting from a large model. The most detailed comparison of several different approaches (Todd & Stamper 1994) has shown the naive classifier to outperform classifiers based on a neural network, decision tree, Bayesian network (hand-crafted network structure with induced parameters), and nearest neighbor model, among others.

In the machine learning community, several researchers have found induced classifiers to be fairly robust to independence assumptions (Langley, Iba, & Thompson 1992; Singh & Provan 1995; Domingos & Pazzani 1996). ¿From a theoretical perspective, there is little research into why this might be the case. For example, Hilden (1984) has outlined a class of probability distributions for which modeling conditional dependencies does not improve diagnostic accuracy. Domingos and Pazzani (1996) have shown that modeling conditional independencies does not improve diagnostic accuracy (i.e., classification tasks), but does improve probability estimation. However, much more work remains to be done.

Experimental Studies

Acute Abdominal Pain Database

The abdominal pain data used for this study consists of 1270 cases, each with 169 attributes. The class variable, final diagnosis, has 19 possible values, and the variables have a number of values ranging from 2 to 32 values. This data was collected and pre-screened by Todd and Stamper, as described in (Todd & Stamper 1993). The resulting database addresses acute abdominal pain of gynaecological origin, based on case-notes for patients of reproductive age admitted to hospital, with no recent history of abdominal or back pain. In compiling the database, the first 202 cases were used in the design of the database itself; thus, they cannot be used for the purpose of testing any model. Moreover, out of the 1270 cases, the diagnosis of only 895 cases was definitely known (definite diagnoses); the remaining 375 cases were assigned the best possible diagnosis, as a presumed diagnosis. Finally, 97 patients occur more than once in the database.

An additional 53 variables representing pathophysiological states and refinements of the final diagnosis were recorded. However, these variables were not used by us since their values are ordinarily no more observable than the final diagnosis. The final diagnosis is used as a measure of diagnostic performance.

Experimental Design

Our experiments address two hypotheses:

1. Does a Bayesian network classifier have better accuracy than a naive Bayesian classifier?
2. Can attribute selection produce networks with comparable accuracy (even through they are a fraction of the size of the full networks)?

Four Bayesian networks were induced from the data, using only the first 169 attributes. The networks were pairs of naive and Bayesian network classifier, with each pair consisting of networks containing all attributes and attributes selected based on an information criterion. For inducing networks with all attributes, we run the algorithm in question. For inducing networks with *selected* attributes, we first select the attributes, and then run the induction algorithm using data for only the selected attributes. We ran a set of preliminary experiments to determine the selection algorithm that produced final networks with the highest predictive accuracy. In comparing three information-based algorithms (as described in (Singh & Provan 1996)) and a belief-network wrapper approach (as described in (Singh & Provan 1995; Provan & Singh 1996)), we decided on using an information-based approach we call CDC to select attributes.

To define the CDC metric, we need to introduce some notation. Let C be the class variable and A represent the attribute under consideration. Let Δ be the subset of attributes already selected. Let k be the number of classes and let m be the number of possible values of A. Moreover, let $\Delta = \{A_1, \ldots, A_q\}$ and let s be the cardinality of the cross product of the sets of values of these variables, i.e., $s = \times_{i=1}^{q} |A_i|$. Given that the set Δ is instantiated to its l^{th} unique instantiation, let $p_{i/l}$ represent the probability that the class variable is instantiated to its i^{th} value and let $p_{j/l}$ be the probability that A is instantiated to its j^{th} value. Similarly, $p_{ij/l}$ is the probability that the class variable takes on its i^{th} value and the attribute A takes on its j^{th} value, given that Δ is instantiated to its l^{th} unique instantiation. The CDC metric is a conditional extension of the complement of Mantaras's (1991) distance metric (d_N), i.e. $1 - d_N$. Following Singh and Provan (1996), we can then define the CDC metric as follows:

$$CDC(A, \Delta) = \frac{\sum_{l=1}^{s} p_l \left(H_{C_l} + H_{A_l} - H_{Cell_l}\right)}{\sum_{l=1}^{s} p_l H_{Cell_l}},$$

where $H_{Cell_l} = -\sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij/l} \log p_{ij/l}$, $H_{A_l} = -\sum_{j=1}^{m} p_{j/l} \log p_{j/l}$, and $H_{C_l} = -\sum_{i=1}^{k} p_{i/l} \log p_{i/l}$.

The feature selection algorithm uses a forward selection search, starting with the assumption that Δ

is empty. It then adds incrementally (to Δ) that attribute A (from the available attributes) that maximizes CDC(A, Δ). The algorithm stops adding attributes when there is no single attribute whose addition results in a positive value of the information metric.[2] Complete details are given in (Singh & Provan 1996).

To maintain consistency with Todd and Stamper's careful comparison (Todd & Stamper 1994) of several induction approaches, we adopted their experimental method of using a cross-validation strategy to evaluate the different methods. Since the first 202 cases had been used during the construction of the database itself, they were not used for testing purposes. The remaining 1068 cases were divided into 11 subsets (10 consisting of 101 cases each while the 11th had 58 cases) which were successively used for testing models induced from the remaining 10 sets plus the first 202 cases. Moreover, for each run, we removed from each training set all repeat presentations of any patient. The performance measure we used was the classification accuracy of a model on the test data, where the classification accuracy is the percentage of test cases that were diagnosed correctly. Inference on the Bayesian networks was carried out using the the HUGIN (Anderson *et al.* 1989) system.

Results

Our first set of experiments compared the performance of the four different approaches listed above, averaged over 11 trials. We use the following notation: naive-ALL and CB for the naive Bayesian and Bayesian network classifier using all attributes, respectively; and Naive-CDC and CDC for the naive Bayesian and Bayesian network classifier using selected attributes, respectively.

Table 1 summarizes the network structure for these runs. Note that using attribute selection reduces the network to roughly 10% of the nodes and 2.3% of the edges for the regular Bayesian networks, and roughly 10% of the nodes and 9% of the edges for the naive networks. This is a dramatic reduction, especially considering the fact that the selective networks have comparable performance to their non-selective counterparts.

Table 2 shows the results of variable selection on network structure. This table shows the attributes that were selected on average by our variable selection approach. The attributes are numbered from 1 to 169, as denoted in the original database. The important point to note is that the number of nodes selected comes from a small subset of the full set of nodes.

Table 3 summarizes the predictive accuracies for our experiments. The second row of Table 3 shows that the naive classifier (using all attributes) performed the best over all cases in the database, although not statistically better than the selective naive classifier.

Table 1: "Average" network structure on experiment using all cases. Structure is based on 11 runs, and describes the average number of nodes and edges.

APPROACH	Nodes	Edges
Naive-ALL	170	169
CB	170	589.8
CDC	17.36	19.56
Naive-CDC	17.36	16.36

Table 2: Network structure resulting from variable selection. Structure is based on 11 runs, and describes the nodes selected.

VARIABLE FREQUENCY	VARIABLE SET
always	3, 29, 37, 68, 112, 152, 169
mostly	4, 5, 32, 38
sometimes	26, 45, 167

The third row of Table 3 shows the predictive accuracies when a prior set of nodes was used during network induction. This set of nodes, $\{3, 29, 37, 68, 112, 152, 169\}$, is the set that was always selected by CDC. Using a paired-t test showed that using this prior did not make a statistically significant difference to the predictive accuracies of the induced classifiers over not using this prior (the first row of the table).

The fourth row of Table 3 shows the predictive accuracies when only cases with definite diagnoses were used. Here, the differences over using all cases are statistically significant. Comparing the induction approaches for the definite diagnosis data, there is no statistically significant difference between the naive methods and CDC even though Naive-CDC has the greatest accuracy (around 3% more).

To further study the dependencies in the database, we computed a covariance matrix for the database, and looked at the variables with the highest correlation (conditional on the class variable) in the covariance matrix. Based on this matrix (which we do not reproduce due to its size), and on computing a pairwise correlation measure described in (Domingos & Pazzani 1996), we observe that many variables are relatively highly correlated, yet the naive classifier performed better than any other approach. A second observation is that the variables with the highest variance are always included in the selective networks.

[2] We take the value of the CDC metric to be zero if the denominator is zero.

Table 3: Predictive accuracy for the four induction approaches, based on the different scenarios. The best predictive accuracy in any row is shown in boldface.

SCENARIO	Naive-ALL	CB	CDC	Naive-CDC
all cases	**60.60±5.59**	57.16±5.65	57.88±4.71	58.89±5.18
"network" prior	**61.95±5.18**	57.16±5.65	57.05±5.80	57.09±5.77
definite diagnosis	67.84±5.34	66.91±8.25	69.54±3.16	**70.53±5.22**

Experiments on Synthetic Data

There are two possible explanations for the unexpectedly high accuracy of the naive Bayesian classfier: the classifiers themselves do not require dependencies for this domain,[3] or the data does not allow the more complex classifiers to estimate parameters accurately enough to be able to classify cases accurately.

To disambiguate the effects of classifier and data, we have run a set of experiments using synthetic Bayes networks with controlled dependencies among the attributes given the class variable. Each network has a five-valued class variable and 13 attributes with between 3 and 5 values. We controlled the attribute dependencies given the class variable by defining three different sets of conditional probability distributions, reflecting (given the class variable): (1) attribute independence ($P(A_k|C, A_{k-1}, ..., A_1) = P(A_k|C)$ for $k = 2, ..., 13$), (2) weak attribute dependence, and (3) strong attribute dependence. We simulated 2000 cases from each network, and learned networks from the simulated data.

Table 4 describes the results of inducing networks from the synthetic data, as averaged over 20 runs. This table shows that this classification task is indeed sensitive to attribute dependencies. For the data with no dependencies, the naive and Bayesian network approaches both learned naive networks with statistically indistinguishable predictive accuracies. However, for the data with dependencies the naive and Bayesian network approaches learned networks with significantly different predictive accuracies, differing by almost 20% for the data with strong dependencies, and slightly less for the data with weak dependencies.

Abdominal Pain Data Revisited

Further analysis of our results show that restrictions of the data lead to the naive classifiers outperforming the Bayesian network classifiers. On closer examination of the data (restricting our attention to cases with a "definite diagnosis"), we found that 2 of the 19 classes accounted for almost 67% of the cases, whereas each of the other classes accounted for 7% or less of the cases. For each of the 2 most common classes, since the probability distribution was induced from many cases, the more complex model (CDC) was significantly better than all other methods, correctly classifying about 89% of the cases. Both selective classifiers significantly outperformed both non-selective classfiers on the cases involving these two classes.

On the other hand, on the cases involving the other 17 classes, naive classifiers performed better than the Bayesian networks, with CDC-naive being the best (though not significantly better). This is because the more complex models could not accurately estimate their more complicated distributions from so few cases, leading to poor predictive accuracy.

These results offer some insights into the observed behavior of the various methods on the abdominal data set. In complex domains with many attributes, such as the abdominal pain domain, feature selection may play a very important part in learning good classifiers for diagnosis; this is especially true when the data set is relatively small. In such cases, it is difficult to accurately learn parameters for the larger networks, more so in the case of Bayesian networks which may pick up spurious dependencies.

Moreover, in domains where there are sufficient cases (as for the two main classes in the abdominal pain data set), Bayesian networks should outperform naive Bayesian classifiers since they can easily model attribute dependencies. However, if the number of cases is small, then the simpler method may perform at least as well as the more complex Bayesian networks.

Discussion

One of the key findings of this work is that for this domain, as well as for a large number of other domains described in (Domingos & Pazzani 1996; Singh & Provan 1995; Provan & Singh 1996), simple models (i.e. naive Bayesian classifiers) perform as well as more complex models. A second key finding is that feature selection further increases model simplicity with no performance penalty. In addition, the more complex the model, the better feature selection increases performance (Singh & Provan 1995). We have shown how only 10% of the attributes provide accuracy comparable to using all the attributes. This can lead to significant savings in data collection and inference.

Our experiments with synthetic networks have shown that attribute dependencies given the class vari-

[3] Domingos and Pazzani (1996) suggest that the classification task itself is insensitive to the presence or absence of attribute dependencies given the class variable.

Table 4: Predictive accuracy for induction algorithms for synthetic networks. Dependencies referred to are between attributes, given the class variable. Accuracies are averaged over 20 runs.

	Cases	CB	CDC	Naive ALL	CDC-naive
"Strong" dependencies	2000	83.65 ± 2.29	82.94 ± 2.10	65.31 ± 2.68	65.7 ± 2.65
"Weak" dependencies	2000	83.06 ± 1.34	82.25 ± 1.78	67.7 ± 1.78	67.7 ± 1.97
No dependencies	2000	87.31 ± 1.00	86.94 ± 1.06	87.31 ± 1.00	86.94 ± 1.06

able do affect classifier performance. Assuming sufficient data with attribute dependencies given the class variable, modeling the dependencies produces a classifier that performs almost 20% better than the naive Bayesian classifier. As a consequence, we argue that the robustness of the naive classifier to attribute correlations is primarily due to the data and the domain under consideration. The robustness of the abdominal pain data to attribute dependencies most likely occurs because the Bayesian network is overfitted to the data.

Our study has shown that data screening can make a big difference to classifier performance. Using the data with just definite diagnoses produces networks with better predictive accuracies than that of the networks induced using all the cases. When all cases are used, the models are both trained and tested on cases whose diagnoses may have been incorrect. This may cause the Bayesian network classifiers to pick up spurious dependencies. In contrast, naive classifiers will not be affected as much: the only thing that changes is the set of probabilities in the network.

Acknowledgements The second author was supported by an IBM Cooperative Fellowship and NSF grant #IRI92-10030. We gratefully acknowledge that HUGIN has been kindly supplied to the second author for his doctoral research.

References

Anderson, S.; Olesen, K.; Jensen, F.; and Jensen, F. 1989. HUGIN - A Shell for building Bayesian Belief Universes for Expert Systems. In *Proceedings IJCAI*, 1080–1085.

Cooper, G. 1990. The computational complexity of probabilistic inference using Belief networks. *Artificial Intelligence* 42:393–405.

de Dombal, F.; Leaper, D.; Staniland, J.; McCann, A.; and Horrocks, J. 1972. Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* 2:9–13.

de Dombal, F. 1991. The diagnosis of acute abdominal pain with computer assistance. *Annals Chir.* 45:273–277.

Domingos, P., and Pazzani, M. 1996. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In *Proc. Machine Learning Conference.*

Edwards, F., and Davies, R. 1984. Use of a Bayesian algorithm in the computer-assisted diagnosis of appendicitis. *Surg. Gynecol. Obstet.* 158:219–222.

Fryback, D. G. 1978. Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research* 11:429–435.

Hilden, J. 1984. Statistical diagnosis based on conditional independence does not require it. *Comput. Biol. Med.* 14:429–435.

Langley, P.; Iba, W.; and Thompson, K. 1992. An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 223–228. AAAI Press.

Lauritzen, S.L., and Spiegelhalter, D.J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)*, 50:157–224.

Lopez de Mantaras, R. 1991. A distance-based attribute selection measure for decision tree induction. *Machine Learning* 6:81–92.

Norusis, M., and Jacquez, J. 1975. Diagnosis I: Symptom nonindependence in mathematical models for diagnosis. *Comput. Biomed. Res.* 8:156–172.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufman.

Provan, G. M., and Singh, M. 1996. Learning Bayesian networks using feature selection. In Fisher, D., and Lenz, H., eds., *Learning from Data: AI and Statistics V, Lecture Notes in Statistics,* 112. Springer Verlag. 291–300.

Provan, G. 1994. Tradeoffs in knowledge-based construction of probabilistic models. *IEEE Trans. on SMC* 11: 287–294.

Seroussi, B. 1986. Computer-aided diagnosis of acute abdominal pain when taking into account interactions. *Method. Inform. Med.* 25:194–198.

Singh, M., and Provan, G. M. 1995. A comparison of induction algorithms for selective and non-selective Bayesian classifiers. In *Proc. 12th Intl. Conference on Machine Learning*, 497–505.

Singh, M., and Provan, G. M. 1996. Efficient learning of selective Bayesian network classifiers. In *Proc. 13th Intl. Conference on Machine Learning.* To appear.

Todd, B. S., and Stamper, R. 1993. The formal design and evaluation of a variety of medical diagnostic programs. Technical Monograph PRG-109, Oxford University Computing Laboratory.

Todd, B. S., and Stamper, R. 1994. The relative accuracy of a variety of medical diagnostic programs. *Methods Inform. Med.* 33:402–416.