Discriminant Model Selection for Belief Net Structures

Yuhong Guo and Russ Greiner
Department of Computing Science
University of Alberta
Edmonton, Alberta T6G 2E8
{ yuhong, greiner }@cs.ualberta.ca

Abstract

Bayesian belief nets (BNs) are often used for classification tasks, typically to return the most likely class label for a specified instance. Many BN-learners, however, attempt to find the BN that maximizes a different objective function — viz., likelihood, rather than classification accuracy — typically by first using some model selection criterion to identify an appropriate graphical structure, then finding good parameters for that structure. This paper considers a number of possible criteria for selecting the best structure, both generative (AIC, MDL, CH) and discriminative: Conditional AIC (CAIC), Conditional MDL (CMDL), resubstitution Classification Error (CE) and Bias^2+Variance (BV). We empirically compare these criteria against a variety of different "correct BN structures", both real-world and artificial, over a range of complexities. We also comparing (1) using the *entire* training sample first to learn the best parameters and then to evaluate the models, versus (2) use only a partition for parameter estimation and another partition for evaluation. Our results show that the discriminant model selection criteria, especially BV, tend to work better than the generative ones, in terms of identifying the optimal structure, and that it is better to partition the training sample. We find similar behaviour whether we are learning the parameters that maximize likelihood or the ones that maximize *conditional* likelihood.

1 Introduction

While belief networks (BNs, a.k.a. Bayesian networks, graphical models) are generative models, capable of modeling a joint probability distribution over a set of variables, they are typically used *discriminatively* for some classification task — *e.g.*, to predict the probability of some disease, given some evidence about the patient. This has motivated the growing body of work on learning an effective BN-classifier from a datasample.

Each BN includes a graph that represents the direct dependencies among the variables. In general, learning an effective BN-classifier requires first finding a good structure (a.k.a. model), then determining appropriate parameters for this model. The first step requires searching through a space of models, seeking the element that optimizes some *model selection criteria*. This paper investigates a number of criteria, towards determining which work best in practice — *i.e.*, which will minimize classifier error on unseen data.

This is not a trivial challenge. While one can typically improve classification performance *on the training data* by increasing the complexity of the model, this usually increases the number of parameters that must be estimated, which typically increases parameter variance, leading to inferior generalization error — *i.e.*, worse performance on unseen data. "Model selection criteria" attempt to operationalize this balance between complexity and goodness of training data, by providing a single number for each network structure. A good model selection criterion is especially important when have limited training data, which is the standard case.

Earlier work [VG00a] evaluated several standard *generative* criteria, where the goal is a structure that produces the best fit to the underlying distribution (using likelihood, Equation 3). Our current paper considers several of these, including Akaike's Information Criterion (AIC) [Boz87], Minimum Description length (MDL) [Ris89], and the CH measure [CH92].

As noted above, our overall goal is different, as we are seeking a structure that leads to good *discriminant* performance, *i.e.*, which has the best classification performance on the unknown testing data. We therefore consider several alternative *discriminative* critera, including Conditional AIC (CAIC), Conditional MDL (CMDL), resubstitution Classification Error (CE), and Bias^2+Variance (BV).

There are one remaining issue to consider: Each learner is given a corpus of training data, which it can use to first find the appropriate parameters for each structure, and then to evaluate the various structures. The learner could use the entire training sample for *both tasks*, or it could first partition the training sample into two subsamples, and then use the first for model selection, and the second for parameter instantiation. We also explore this "undivided sample" versus "divided sample" below.

We therefore compare the three classical generative model selection criteria with the four discriminant model selection criteria, using different sampling, over a range of training sizes. Our preliminary experimental results suggest that, while discriminant model selection criteria (especially BV), performs better than generative model selection criteria in most cases, this is not universal. We hypothesize the performance of criteria is related to complexity of the Markov blanket around the query variable, within the true generative model (which we identify with the true discriminative model [GGS97]). We therefore systematically explore the effectiveness of various model selection criteria across generative models with a wide range of Markov blanket complexities. Our experimental results confirm that this complexity term does influence the performance of different model selection criteria. Finally, we explore model selection both when the parameter learning algorithm attempts to optimize the simple likelihood of the data (which is the standard approach), and also *conditional*/ likelihood.

As a final preliminary comment, we note that there are many reasons to select some single model, many of which relate more to prior assumptions and constraints, than to performance. In this paper, however, we are *only* concerned with eventual classification performance, as measured by Equation 1.

The rest of this section discusses related work. Section 2 provides the framework for this paper, describing belief networks, model selection criteria and parameter learning. Section 3 presents our experimental setup and results.

The webpage [Guo04] contains additional information about the experiments reported here, as well as other related results.

1.1 Related Work

There is a large of literature on the general model selection problem, proposing a variety of schemes including MDL [Ris89], BIC [Sch78] and AIC [Boz87]; our analysis includes each of these schemes

There is also a considerable literature on structure learning and generative model selection for belief networks in particular; see [Hec98] for a detailed overview on this subject. MDL is used frequently to evaluate candidate structures [LB94, Suz78, FGG97]. [FY96] examined the sample complexity of the MDL-based belief network learning. [VG00a] provides a comprehensive comparision between MDL, AIC and Cross-Validation (CV) when learning belief network structures *generatively*. While we borrow some of the techniques from these projects, recall our goal is learning the structure that is best for a *discriminant* classification task.

As noted earlier, belief nets are often used for this classification task. This dates back (at least) to NaïveBayes classifiers [DH73], and has continued with various approaches that include feature selection [LS94], and alternative structures [FGG97, CG99, CG01]. [KMST99] compared several model selection criteria (unsupervised/supervised marginal likelihood, supervised prequential likelihood, cross validation) on a subset of Bayesian networks regarded as "pruned Naive Bayes"; their result suggested that the supervised prequential likelihood performs best. [GD03] presented an algorithm for discriminative learning belief networks that used the log conditional likelihood of the class variable given the evidence (Equation 2) as the model selection criterion.

Our work differs from those others, as it is the first to provide a comprehensive comparison between classical generative model selection criteria and discriminant criteria on the task of learning good structures for BN-classifier. We also propose several new discriminant model selection criteria, some motivated variants of generative criteria (CAIC and CMDL), and one (BV) motivated by the classification task in general [Rip96].

2 Framework

2.1 Belief Network Classifiers

We assume there is a stationary underlying distribution $P(\cdot)$ over n (discrete) random variables $\mathcal{V} = \{V_1, \ldots, V_n\}$, which we encode as a "(Bayesian) belief net" (BN) — a directed acyclic graph $B = \langle \mathcal{V}, A, \Theta \rangle$, whose nodes $\mathcal{V}$ represent variables, and whose arcs A represent dependencies; see Figure 1. Each node $D_i \in \mathcal{V}$ also includes a conditional-probability-table (CPtable) $\theta_i \in \Theta$ that specifies how D_i's values depend (stochastically) on the values of its immediate parents. In particular, given a node $D \in \mathcal{V}$ with immediate parents $\mathbf{F} \subset \mathcal{V}$, the parameter $\theta_{d|\mathbf{f}}$ represents the network's term for $P(D = d \,|\, \mathbf{F} = \mathbf{f})$ [Pea88].

The user interacts with the belief net by asking *queries*, each of the form "What is $P(C = c \,|\, \mathbf{E} = \mathbf{e})$?" — *e.g.*, What is $P($`Cancer = true`$\,|\,$`Gender=female, Smoke=true`$)$? — where $C \in \mathcal{V}$ is a single "query variable", $\mathbf{E} \subset \mathcal{V}$ is the subset of "evidence variables", and c (resp., $\mathbf{e}$) is a le-

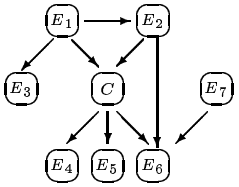

Figure 1: Example of a Belief Net Structure

gal assignment to C (resp., $\mathbf{E}$).[1]

Given any unlabeled instance $\{E_1 = \mathbf{e}_1, \ldots, E_k = \mathbf{e}_k\}$ (which we will write as $\mathbf{E} = \mathbf{e}$), the belief net B will produce a distribution over the values of the query variable; perhaps $B(\texttt{Cancer = true} \,|\, \mathbf{E} = \mathbf{e}) = 0.3$ and $B(\texttt{Cancer = false} \,|\, \mathbf{E} = \mathbf{e}) = 0.7$. In general, the associated H_B classifier system will then return the value $H_B(\mathbf{e}) = \text{argmax}_c\{B(C = c|\mathbf{E} = \mathbf{e})\}$ with the largest posterior probability — here return $H_B(\mathbf{E} = \mathbf{e}) =$ false as $B(\texttt{Cancer = false} \,|\, \mathbf{E} = \mathbf{e}) > B(\texttt{Cancer = true} \,|\, \mathbf{E} = \mathbf{e})$.

A good belief net classifier is one that produces the appropriate answers to these unlabeled queries. We will use "classification error" (aka "0/1" loss) to evaluate the resulting B-based classifier H_B

$$\text{err}(B) = \sum_{\langle \mathbf{e}, c\rangle : H_B(\mathbf{e}) \neq c} P(\mathbf{e}, c) \qquad (1)$$

Our goal is a belief net B^* that minimizes this score, with respect to the true distribution $P(\cdot)$. While we do not know this distribution *a priori*, we can use a sample drawn from this distribution, to help determine which belief net is optimal. We will use a training set S of $m = |S|$ complete instances, where the ith instance is represented as $\langle c^i, \mathbf{e}^i_1, \ldots, \mathbf{e}^i_n\rangle$. This paper focuses on the task of learning the BN-structure $G = \langle \mathcal{V}, A\rangle$ that allows optimal classification performance on unseen examples.

Conditional Likelihood: Given a sample S, the empirical "log conditional likelihood" of a belief net B is

$$LCL^{(S)}(B) = \frac{1}{|S|} \sum_{\langle \mathbf{e}, c\rangle \in S} \log(B(c \,|\, \mathbf{e})) \qquad (2)$$

where $B(c \,|\, \mathbf{e})$ represents the conditional probability produced by the belief network B.

[MN89, FGG97] note that maximizing this score will typically produce a classifier that comes close to minimizing the classification error (Equation 1).

[1] This paper focuses on the standard "machine learning" case, where all queries involve the same variable (*e.g.*, all queries ask about `Cancer`), and we assume the distribution of conditioning events matches the underlying distribution, wihch means there is a single distribution from which we can draw instances, which correspond to a set of labeled instances (aka "labeled queries"). See [GGS97] for an alternative position, and the challenges this requires solving.

While this $LCL^{(S)}(B)$ formula closely resembles the (empirical) "log likelihood" function

$$LL^{(S)}(B) = \frac{1}{|S|} \sum_{\langle \mathbf{e}, c\rangle \in S} \log(B(c, \mathbf{e})) \qquad (3)$$

used as part of many *generative* BN-learning algorithms, note [FGG97]

$$LL^{(S)}(B) = \frac{1}{|S|}\left[\sum_{\langle c, \mathbf{e}\rangle \in S} \log(B(c \,|\, \mathbf{e})) + \sum_{\langle c, \mathbf{e}\rangle \in S} \log(B(\mathbf{e}))\right]$$

where only the first term (which resembles our LCL measure) is relevant.

We will measure the complexity of the BN B as the number of free parameters in the network

$$k(B) = \sum_{i=1}^{n} (|V_i| - 1) \prod_{F \in \text{Pa}(V_i)} |F| \qquad (4)$$

where n is the number of variables, $|V|$ is the number of values of any variable V, and $\text{Pa}(V)$ is the set of immediate parents of the node V.

For a belief network structure, given a completely instantiated tuple, a variable C is only dependent on the variables in its Markov Blanket [Pea88], which is defined as the union of X's direct parents, direct children and all direct parents of X's direct children. *E.g.*, in Figure 1, C's Markov Blanket contains all of the E_i's except E_3. We define $k_C(B)$ as the number of parameters in C's Markov blanket, within B, using an obvious analogue to Equation 4.

2.2 Generative Model Selection Criteria

Most of the generative criteria begin with the average empirical log likelihood of the data, Equation 3, as $LL^{(S')}(B)$ on *unseen* data S' is useful as an unbiased estimate of the average generative quality of the distribution B. To avoid overfitting, the MDL and AIC measures add a "regularizing" term that penalizes complex structures, as an embodiment of the trade off between model simplicity and goodness of fit to the training data.

$$MDL^{(S)}(B) = -LL^{(S)}(B) + \frac{k(B)\,\log m}{2m}$$

$$AIC^{(S)}(B) = -LL^{(S)}(B) + \frac{k(B)\,\log e}{m}$$

(Note this MDL is exactly the negative of Bayesian Information Criterion (BIC) [Sch78].)

The third generative model selection criterion is the marginal likelihood — averaged over all possible CPtable values (in the Bayesian framework):

$$CH^{(S)}(B) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + a_{ij})} \prod_{k=1}^{|E_i|} \frac{\Gamma(\alpha_{ijk} + a_{ijk})}{\Gamma(\alpha_{ijk})},$$

where $q_i = \prod_{Y \in \text{Pa}(E_i)} |Y|$ is the number of states of the parents of variable E_i, α_{ijk} are the Dirichlet prior parameters (here set to 1), $\alpha_{ij} = \sum_{k=1}^{|E_i|} \alpha_{ijk}$, and a_{ijk} are the empirical counts — *i.e.*, the number of instances in the datasample S where the ith variable E_i takes its kth value and its parents have their jth value. This measure is equivalent to the BD (Bayesian Dirichlet) metric [HGC95], if we set both the structure prior and each Dirichlet paramter prior to 1.

We also tested the log likelihood criteria (Equation 3) as it is often used in the literature. As it did performs not as well as the other criteria, we decided not to show those result in this paper. They do appear in [Guo04].

2.3 Discriminant Model Selection Criteria

We consider two classes of discriminant Model Selection Criteria; one set are variants of the generative (LL-based) criteria shown above, and the others are based on classification performance.

2.3.1 CMDL and CAIC

The CMDL (conditional MDL) and CAIC (conditional AIC) criteria are discriminative analogues to the generative MDL and AIC criteria, which differ by using log *conditional* likelihood to measure "training error" and by dividing the penalty term by n, as they are considering the log conditional likelihood of only a single class variable.

$$CMDL^{(S)}(B) = -LCL^{(S)}(B) + \frac{k(B)\log m}{2\,m\,n}$$
$$CAIC^{(S)}(B) = -LCL^{(S)}(B) + \frac{k(B)\log e}{m\,n}$$

2.3.2 Empirical Classification Error (CE)

As we use classification error on testing data to measure a BN-classifer's performance, we decided to include its classification error (CE) on *training data* as a discriminant model selection criterion.

$$err^{(S)}(B) = \frac{|\{\langle \mathbf{e}, c\rangle \in S \mid H_B(\mathbf{e}) \neq c\}|}{|S|}$$

2.3.3 Bias^2+Variance

[Rip96] proves that the expected L_2 error of a classifer corresponds to "Bias^2+Variance"

$$BV^{(S)}(B) = \frac{1}{|S|} \sum_{\langle c,\mathbf{e}\rangle \in S} [\,t(c\,|\,\mathbf{e}) - B(c\,|\,\mathbf{e})]^2 + \hat{\sigma}^2[B(c\,|\,\mathbf{e})]$$

where the "true" response $t(c\,|\,\mathbf{e})$ corresponds to the empirical frequence within the training data:

$$t(c\,|\,\mathbf{e}) = \frac{\#_S(C=c, \mathbf{E}=\mathbf{e})}{\#_S(\mathbf{E}=\mathbf{e})}$$

where $\#_S(\mathbf{E}=\mathbf{e})$ is the number of instances in training set S that match this (partial) assignment, and we use the variance estimate provided in [VGH01]: $\hat{\sigma}^2[B(c\,|\,\mathbf{e})] =$

$$\sum_{\theta_{D|\mathbf{f}} \in \Theta} \frac{1}{n_{D|\mathbf{f}}} \left[\begin{array}{l} \sum_{d \in D} \frac{1}{\theta_{d|\mathbf{f}}} [B(d,\mathbf{f},c\,|\,\mathbf{e}) - B(c\,|\,\mathbf{e})\,B(d,\mathbf{f}\,|\,\mathbf{e})]^2 \\ -\,(B(\mathbf{f},c\,|\,\mathbf{e}) - B(c\,|\,\mathbf{e})\,B(\mathbf{f}\,|\,\mathbf{e}))^2 \end{array} \right]$$

which requires summing over the CPtable rows $\theta_{D=d|\mathbf{F}=\mathbf{f}}$, and uses $n_{D|\mathbf{F}=\mathbf{f}} = 1 + |D| + \#_S(\mathbf{F}=\mathbf{f})$ as the "effective sample size" of the conditioning event for this row. (Note this done from a Bayesian perspective, where we first identify each CPtable row with a Dirichlet-distributed random variable, then compute its posterior based on the training sample, and finally use these posterior distributions of the CPtable rows to compute the distribution over the response to the $B_\Theta(c\,|\,\mathbf{e})$, which is afterall just a function of those random variables, Θ.)

2.4 How to Instantiate the Parameters

As mentioned above, a BN includes both a structure *and a set of parameters* for that structure. Given complete training data, the standard parameter learning algorithm, OFE, sets the value of each parameter to its empirical frequency in the datasample, with a Laplacian correction:

$$\theta_{D=d|\mathbf{F}=\mathbf{f}} = \frac{\#_S(D=d, \mathbf{F}=\mathbf{f}) + 1}{\#_S(\mathbf{F}=\mathbf{f}) + |D|}$$

[CH92] prove these generative values optimize the likelihood of the data, Equation 3, for the given structure. (But see Section 3.5.)

The learner has access to a training sample S, to use as it wishes when produces the optimal structure. A simple model selection process will use the "undivided sample" approach: use *all* of S when finding the appropriate instantiation of the parameters. It will then compute a score for this instantiated structure, based again on S. An alternative approach will first partition $S = S_1 \cup S_2$, then use only S_1 when computing the instantiaton, and only S_2 for computing the score. The "5CV" variant will actually use 5 partitions, where 4 are used for setting the parameters and the fifth for evaluating the resulting instantiated structure.

3 Empirical Studies

This section reports on our empirical studies that compare the 7 model selection criteria mentioned above, to help determine when (if ever) to use each, and also to whether to

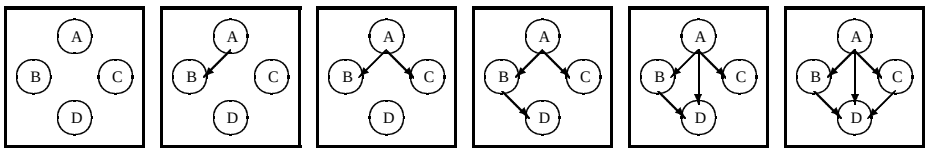

Figure 2: Sequence of structures; (d) is the original structure

partition the training data or not. We therefore asked each of the criteria to identify the appropiate structure, across a range of situations. Section 3.1 first explains how we will run each experiment, and how we will evaluate the results. Section 3.2 presents our first study, on real-world distributions. This data suggests that the complexity of the generative model may determine which criteria works best. The remaining subsections explore this. Section 3.3 (resp., 3.4) considers the performance of the selection criteria on a set of artificial models with a range of complexity, using a single undivided sample (resp., 5CV sample).

Finally, as noted earlier, our goal is to find the structure for an effective classifier. However, the OFE parameter learner is designed to compute the parameters that optimal the *likelihood* of the data, which is *generative*. The ELR algorithm [GZ02], by contrast, attempts to find the parameters that optimize the discrminative *conditional likelihood* score. (This algorithm extends logistic regresion as it applies to arbitrary network structures, while standard logistic regression corresponds to naive bayes.) Section 3.5 therefore considers model selection when ELR is used to find parameters.

3.1 Experimental Setup

In each experiment, we have a specific "true" distribution $P(\cdot)$ — *i.e.*, correct BN-structure and parameters — which we either download, or produce. We generate a number of complete datasamples from this $P(\cdot)$, of various sizes. We produce a set of possible models by modifying the true structure: see below. For each training sample we then run each of the selection criteria (in the appropriate context — divided sample or not). Each criteria produces a single number for each candidate structure. Figure 3 shows this, in the context of the ALARM network (Section 3.2).[2] Each criteria then identifies the structure it considers best, which is the one with the lowest score. Here, for example, CMDL would select the structure labeled "-7", MDL would pick "-9" BV would select "0" and CH, "1". (These numbers correspond to the number of edges added, or deleted, to the initial structure. Hence, the original structure is the one labeled "0".) For each criteria χ, let B^χ be this selected structure, instantiated appropriately. We then compute the error of each B^χ, — $err^{(S')}(B^\chi)$ based on a hold-out sample S' of size $|S'| = 1000$, generated from $P(\cdot)$.

We also detemine which of the structures $B^* = \operatorname{argmin} err^{(S')}(B)$ really was the best — *i.e.*, had the smallest error. (See the "Test Error" line in Figure 3; notice this picks "-2". Observe the structure that is best for a particular sample need not be the original structure!) The score for criterion χ is the ratio $err^{(S')}(B^\chi)/err^{(S')}(B^*)$. For each sample size, we compute the average over 20 repeated trials, each time using different training and testing set. This ratio will be 1 for a perfect criteria; in general, it will be higher.[3]

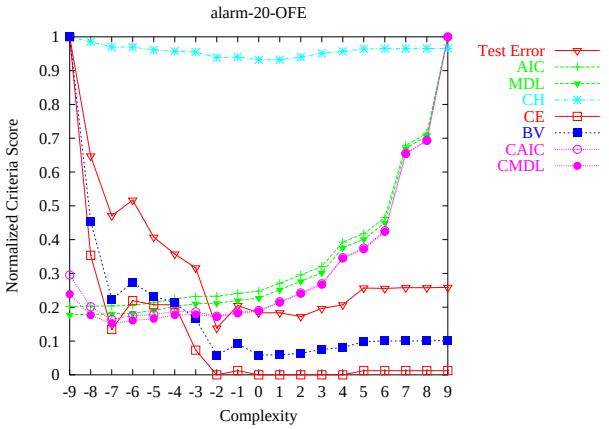

Figure 3: Criteria Score, as function of Structure

Proper model selection is most challenging, and hence more relevant, when given limited training data; this paper therefore considers small training sizes, of 10, 20 and 50.

Generating Sequence of Structures: Given a true BN-structure, G^* we generate a sequence of BN-structure candidates with increasing complexity, as follows:

1. Starting from the original structure, sequentially remove one randomly-selected edge from the Markov blanket (MB) of the class variable, to generate a series of structures whose class variable has decreasing MB size.
2. Starting from the original structure, sequentially add one randomly-selected edge to the Markov blanket of

[2] Each measure is normalized to fit between 0 and 1, with 0 being best.

[3] We considered an alternative evaluation method: simply measure how often each method "correctly" selected the original structure. We prefer our evaluation measure for two reasons: First, the original structure might not be the optimal structure, given this datasample. Second, we wanted to quantify how bad the loss was. (The second reason holds even if we used an evaluation method that dealt with the "optimal structure" B^* rather than "original structure".)

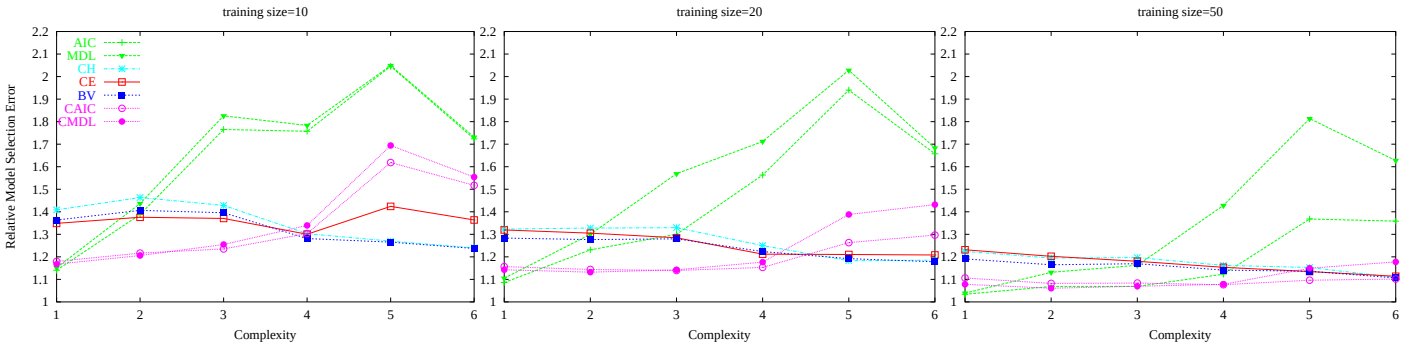

Figure 4: Experiment #2: 7-Variable BN, Undivided Sample — (a) m=10; (b) m=20; (c) m=50

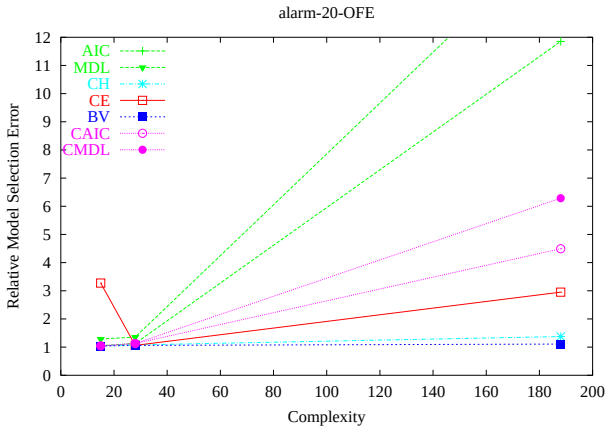

Figure 5: Relative Score

the class variable, to generate a series of structures whose class variable has increasing MB size.

See Figure 2; here (d) is the starting point, and we produce (c), (b) and (a) by deleting existing arcs, and then produce (e) and (f) by adding new arcs.

Note that adding one more arc might not increase the MB size by only one, since adding one arc into the Markov blanket may cause some other edges of the original network to now become part of the Markov blanket of the class variable. Similiarly removing a single arc may reduce the MB size by more than 1.

3.2 Exp#I: Real-World Distribution, Undivided Sample

Our preliminary investigations examined several real belief nets; here we focus ALARM [BSCC89]; see [Guo04] for the others. We choose three different variables to serve as the class variables, which produced three different query forms, whose Markov blankets had a wide range in size: 15, 28, 188. Figure 3 corresponds to one run, with $k_{\text{VentLung}}(\text{ALARM}) = 188$.

As outlined above, we computed the relative score for each criteria, $err^{(S')}(B^{\chi})/err^{(S')}(B^{*})$. Figure 5 is the result when we used a sample of size 20. (Recall every point is the average of 20 runs.) We found that BV performs well throughout, but the other measures appear effective only for simple models, with small MB sizes. Moreover, the discriminative measures (BV, CE, CAIC, CMDL) appear better than AIC and MDL. We found similar performances on other sample sizes, and other networks.

3.3 Exp#II: Artificial Distribution, Undivided Sample

We observed different behaviour of the various selection criteria as we varied the complexity of the Markov blanket around the query variable. To further explore this, we generated a set artificial networks, whose query variables could have arbitrary Markov blanket complexity. We will use the networks here, and in the subsequent sections.

We first randomly generated six groups of belief network structures with varying Markov blanket complexity, where each group includes 30 structures.

Each of these became the gold standard, used to generate datasamples.

We used the experimental apparatus described in the previous section to test the behavior of each criterion, across a spectrum of complexities and a range of sample sizes. The three graphs in Figure 4 show the results for belief networks with seven variables, over samples of $m = 10$, $m = 20$ and $m = 50$, using an undivided training sample. The complexity (on the X axis, from 1 to 6) represents the six group of structures, with increasing generative complexity.

There are several observations from the results, which are consistent across different training sizes. The performance of the criteria BV, CE and CH were much smoother across the generative complexity, as compared to AIC, MDL, CAIC and CMDL. The BV criterion performs consistently well across the whole generative complexity range and different training sizes; it is consistently better than the discriminant CE and the generative CH. CE also performs better than CH on many cases.

(This paper will focus on this $n = 7$ case; we observed similar behaviour in other situations as well; see [Guo04].)

AIC and MDL performs well only when the generative

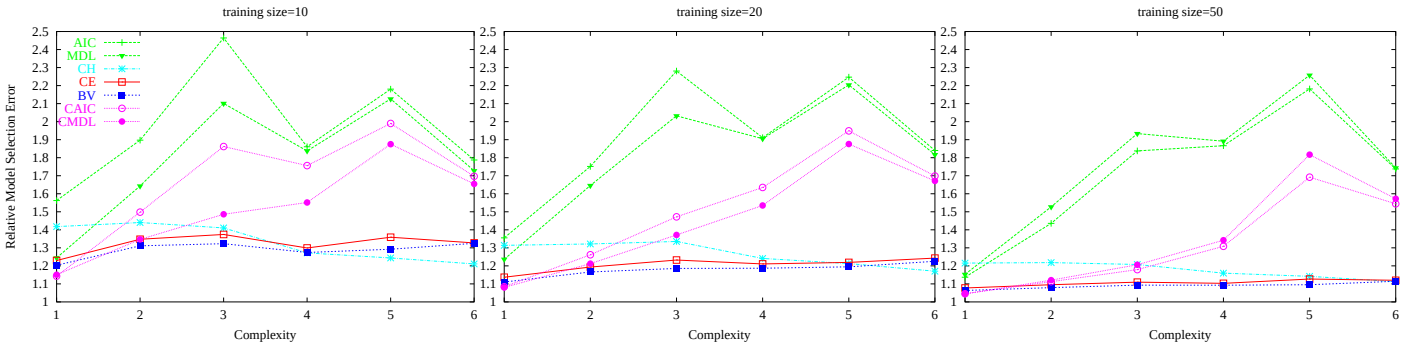

Figure 6: Experiment #III: 7-Variable BN, 5CV Sample — (a) m=10; (b) m=20; (c) m=50

complexity is very small, otherwise, they became the worst among all criteria. CAIC and CMDL are better than AIC and MDL, in that they performs very well on a longer generative complexity range than AIC and MDL. But without exception, CAIC and CMDL become bad when the generative complexity become large. our experiments reveal why: AIC and especially MDL have too strong a preference for simple structures! For smallest training size 10, AIC and MDL *always* picked the simplest structure in the sequence, irrespective of the data (which was sufficient to tell the other measures to prefer other larger structure.) CMDL and CAIC typically avoid complicated structure as well. This is consistent with the [VG00a] observation that these criteria seriously underfit — indeed, for small samples, they almost invariably produced no edges. This suggests the complexity penalty term for AIC and MDL may be too big, and not appropriate for belief network on most cases. Although CAIC and CMDL, with smaller complexity penalty terms, worked better than AIC and MDL, they were not perfect. We also see that increasing the training set size does improve the performance of those complexity penalized criteria, especially CAIC and CMDL. Basically, it is not clear how to pick a godd complexity penalty term for all belief networks, since there are a very large space of structures, and hence a large range of complexities.

In general, however, discriminative criteria seem to work better than generative criteria.

We performed the same experiments on a set of larger belief networks (*e.g.*, with 15 variables) and obtained similar results; see [Guo04]

3.4 Exp#III: Artificial Distribution, 5CV

Here, we used a variant of 5-fold CrossValidation: we first partitioning the training data into 5 partitions $\{S_1, \ldots, S_5\}$ then, for each partition S_i, used the other 4 subsets to set the parameter and used S_i for evaluation. The final score for each structure was the average over these 5 values.

Figure 6 shows the results, on the 7-variable case. We can see that BV is the best for most of the cases across the generative complexity range, and across different training sizes. CE is the almost universally second best, with AIC, MDL, CAIC and CMDL all performing worse. Our explaination is that this 5CV approach makes the data size for evaluation yet smaller, which makes the (conditional) likelihood term in these four criteria become even less accurate. Our experiments also verified this, as we found AIC and MDL pick continue to select the simplest structure in all cases, and CAIC and CMDL pick the simplest structure in most cases.

3.5 Exp#IV: Artificial Distribution, ELR

Finally, we ran the same experiments, but using ELR rather than OFE to instantiate the parameters. Figure 7 shows the result, for the divided-sample case. We see, again, that BV and CE appear to be the best, then CH, followed by the other discriminant measure, and then by the other generative ones.

4 Conclusions

Belief nets are now often used as classifiers. When learning the structure for such a BN-classifier, it is useful to have a criteria for evaluating the different candidate structures. This paper addresses that question.

We proposed a number of novel *discriminative* model selective criteria, some (CAIC, CMDL) analogues of standard generative criteria (commonly used when learning generative models), and one (BV) motivated by the familiar discriminative approach of decomposing error into bias and variance components. We then evaluated these methods, along with the generative ones, across a number of different situations: over queries of different complexities, different sample size, different ways to use the training sample, and also over different ways to instantiate the parameters (generatively vs discriminatively). Our empirical results show that the discriminative model selection criteria generally perform better than the generative criteria, and that BV performs best in most cases. We found that AIC and MDL, the criteria typically often used when learning generative BNs, did poorly as they almost always preferred the simplest structures. While our variants, CAIC and CMDL,

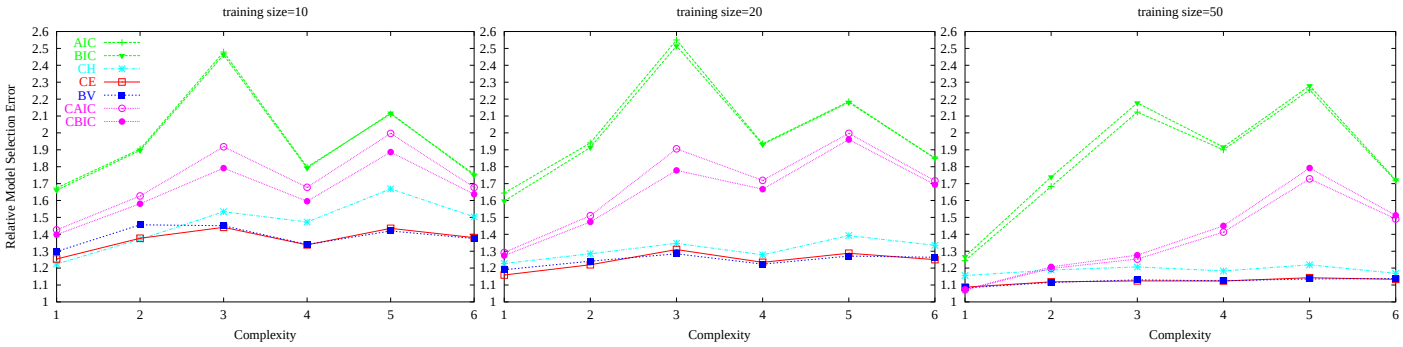

Figure 7: Experiment #IV: 7-Variable BN, 5CV Sample — using ELR — (a) m=10; (b) m=20; (c) m=50

worked better, they were not as stable as BV across the different cases. We attribute this to the observation that it is difficult, if not impossible, to find a single complexity penalty term that applies to every possible query from a belief network. We also found that the cross-validation based approach was more effective than using the entire sample for both instantiating the parameters and evaluation.

In summary, we view our empirical results as confirming two fairly-obvious claims: First, discriminant methods work better when the underlying task is discriminant, and second, cross-validation is a good idea.

See the webpage [Guo04] for additional information.

References

[Boz87] H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrica*, 52, 1987.

[BSCC89] I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *European Conference on Artificial Intelligence in Medicine*, 1989.

[CG99] J. Cheng and R. Greiner. Comparing bayesian network classifiers. In *UAI99*, 1999.

[CG01] J. Cheng and R. Greiner. Learning bayesian belief network classifiers: Algorithms and system. In *Canadian Conference on Artificial Intelligence*, 2001.

[CH92] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning Journal*, 9:309–347, 1992.

[DH73] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[FGG97] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning Journal*, 29, 1997.

[FY96] N. Friedman and Z. Yakhini. On the sample complexity of learning Bayesian networks. In *UAI96*, 1996.

[GD03] D. Grossman and P Domingos. Learning bayesian network classifiers by maximizing conditional likelihood. Technical report, Dept Computer Science & Engineering, U. of Washington, 2003.

[GGS97] R. Greiner, A. Grove, and D. Schuurmans. Learning Bayesian nets that perform well. In *UAI-97*, 1997.

[Guo04] 2004. http://www.cs.ualberta.ca/∼yuhong/DiscriminantModelSelection.

[GZ02] R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminant parameter learning of belief net classifiers. In *AAAI-02*, 2002.

[Hec98] D. Heckerman. A tutorial on learning with Bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*, 1998.

[HGC95] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20(3), 1995.

[KMST99] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of bayesian networks. In *UAI99*, 1999.

[LB94] Wai Lam and Fahiem Bacchus. Learning Bayesian belief networks: An approach based on the MDL principle. *Computation Intelligence*, 10(4):269–293, 1994.

[LS94] P. Langley and S. Sage. Induction of selective bayesian classifiers. In *UAI-94*, 1994.

[MN89] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Rip96] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University, 1996.

[Ris89] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific, 1989.

[Sch78] G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[Suz78] J. Suzuki. Learning bayesian belief networks based on the mdl principle: An efficient algorithm using the branch and bound technique. *Annals of Statistics*, 6:461–464, 1978.

[VG00a] T. Van Allen and R. Greiner. Model selection criteria for learning belief nets: An empirical comparison. In *ICML'00*, 2000.

[VGH01] T. Van Allen, R. Greiner, and P. Hooper. Bayesian error-bars for belief net inference. In *UAI01*, 2001.