Discriminative versus Generative Parameter and Structure Learning of Bayesian Network Classifiers

Franz Pernkopf PERNKOPF@TUGRAZ.AT
Department of Electrical Engineering, Graz University of Technology, Inffeldgasse 16c, A-8010 Graz, Austria

Jeff Bilmes BILMES@EE.WASHINGTON.EDU
Department of Electrical Engineering, University of Washington, Box 352500, Seattle, WA 98195 USA

Abstract

In this paper, we compare both discriminative and generative parameter learning on *both* discriminatively *and* generatively structured Bayesian network classifiers. We use either maximum likelihood (ML) or conditional maximum likelihood (CL) to optimize network parameters. For structure learning, we use either conditional mutual information (CMI), the explaining away residual (EAR), or the classification rate (CR) as objective functions. Experiments with the naive Bayes classifier (NB), the tree augmented naive Bayes classifier (TAN), and the *Bayesian multinet* have been performed on 25 data sets from the UCI repository (Merz et al., 1997) and from (Kohavi & John, 1997). Our empirical study suggests that discriminative structures learnt using CR produces the most accurate classifiers on almost half the data sets. This approach is feasible, however, only for rather small problems since it is computationally expensive. Discriminative parameter learning produces on average a better classifier than ML parameter learning.

1. Introduction

There are two paradigms for learning statistical classifiers: Generative and discriminative methods (Bahl et al., 1986),(Jebara, 2001). Generative classifiers learn a model of the joint probability of the features and the corresponding class label and perform predictions (classification) by using Bayes rule to compute the posterior probability of the class variable. The standard approach to learn a generative classifier is maximum likelihood (ML) estimation, possibly augmented with a (Bayesian) smoothing prior. Discriminative classifiers directly model the class posterior probability, or use a discriminant function. Hence, the classification problem is solved by optimizing that which is most important for classification accuracy. There are several compelling reasons for using discriminative rather than generative classifiers, one of which is that the classification problem should be solved most simply and directly, and never via a more general problem such as the intermediate step of estimating the joint distribution. This is indeed the goal of the support-vector approach to classification (Vapnik, 1998). Here, however, we attempt to form a generative model that does not include the complexity necessary to actually *generate* samples from the true distribution. Rather, we desire generative models that include complexity enough only so that they discriminate well.
Indeed, Friedman et al. (Friedman et al., 1997) observed that there can be a discrepancy between a Bayesian network learned by optimizing likelihood and the predictive accuracy of using this network as classifier since the entire data likelihood is optimized rather than only the class conditional likelihood. A sufficient (but not necessary) condition for optimal classification is for the conditional likelihood (CL) to be exact. Hence, the network structure and parameters which maximize the CL are of interest, since that criterion is equivalent to minimizing the KL-divergence (Cover & Thomas, 1991) between the true and the approximate conditional distribution (Bilmes, 2000). Unfortunately, the CL function does not decompose and there is no closed-form solution for determining its parameters.

In current approaches, either the structure or the parameters are learned in a discriminative manner by maximizing CL. Greiner and Zhou (Greiner & Zhou, 2002) introduced an optimization approach by computing the maximum CL parameters using a conjugate gradient method after the structure of the network has been established. In (Grossman & Domingos, 2004) the CL function is used to learn the structure of the network, where the parameters are determined by ML estimation. They use hill climbing search with the CL function as a scoring measure, where at each iteration one edge is added to the structure which complies with the restrictions of the network topology (e.g. tree

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005.

augmented naive Bayes (TAN)) and the definitions of a Bayesian network. The classification rate (CR) has also been used as an objective function for discriminative structure learning (Keogh & Pazzani, 1999),(Pernkopf, 2005). Bilmes (Bilmes, 2000),(Bilmes, 1999) introduced the *explaining away residual* (EAR) for discriminative structure learning of dynamic Bayesian networks for enhancing the performance in speech recognition. An empirical and theoretical comparison of discriminative and generative classifiers (logistic regression and the naive Bayes (NB) classifier) is given in (Ng & M., 2002). It is shown that for small sample sizes the generative NB classifier can outperform the discriminatively trained model. Therefore, hybrid models have been proposed (Raina et al., 2004) to obtain the best of both worlds.

In this paper, we empirically compare both discriminative and generative parameter training on *both* discriminative *and* generatively structured Bayesian network classifiers (see Figure 1). As far as we know, our work is the first evaluation of discriminative parameter training on discriminatively structured networks. In our work, parameter training has been performed either by ML estimation or by optimizing the CL. For structure learning of the TAN and the Bayesian multinet we use the following scoring functions: The conditional mutual information (CMI) is utilized as a generative approach, and the CR or an approximation of the EAR measure as a discriminative method. We focus on

		Structure learning	
		Generative (CMI)	Discriminative (EAR, CR)
Parameter learning	Generative (ML)	✓	✓
	Discriminative (CL)	✓	✓

Figure 1. Our strategies for Bayesian network classifi er learning

three network topologies: naive Bayes (NB), the TAN classifier, and Bayesian multinets.

The paper is organized as follows: In Section 2 we introduce Bayesian networks. Then we briefly present the different network topologies and the approaches for generative and discriminative structure and parameter learning. In Section 3 we report classification results on 25 data sets from the UCI repository (Merz et al., 1997) and from (Kohavi & John, 1997) using all combinations of generative/discriminative structure/parameter learning. Additionally, we give the number of parameters which have to be trained for the particular network structure and the number of classifications it takes to establish the structure using the CR scoring function for the TAN structure. Conclusions are presented in Section 4.

2. Bayesian network classifier

A Bayesian network (Pearl, 1988),(Cowell et al., 1999) $B = \langle G, \Theta \rangle$ is a directed acyclic graph G which represents factorization properties of the distribution of a set of random variables $Z = \{C, X_1, \ldots, X_N\} = \{Z_1, \ldots, Z_{N+1}\}$, where each variable in Z has values denoted by lower case letters $\{c, x_1, \ldots, x_N\}$. The random variable $C \in \{1, \ldots, |C|\}$ represents the classes, $|C|$ is the cardinality of C, $X_{1:N}$ denote the random variables of the N attributes of the classifier. Each graph node depicts a random variable, while the lack of an edges specifies some independences property. Specifically, in a Bayesian network each node is independent of its non-descendants given its parents. These conditional independence relationships reduce both number of parameters and required computation. Symbol Θ represents the set of parameters which quantify the network. Each node Z_j is represented as a local conditional probability distribution given its parents Z_{Π_j}. We use $\theta^j_{i|k}$ to denote a specific conditional probability table entry, the probability that variable Z_j takes on its ith value assignment given that its parents Z_{Π_j} take the kth assignment, i.e. $\theta^j_{i|k} = P\left(z^m_j = i | z^m_{\Pi_j} = k\right) = \prod_{i=1}^{|Z_j|} \prod_k \left(\theta^j_{i|k}\right)^{u^{j,m}_{i|k}}$, where the last equality follows from the fact that $u^{j,m}_{i|k}$ is 1 for $z^m_j = i$ and $z^m_{\Pi_j} = k$, and is 0 elsewhere, i.e. $u^{j,m}_{i|k} = \mathbb{1}_{\left\{z^m_j = i, z^m_{\Pi_j} = k\right\}}$. The training data consists of M samples $\mathcal{S} = \{z^m\}_{m=1}^M = \{(c^m, x^m_{1:N})\}_{m=1}^M$. We assume a complete data set with no missing data values. The joint probability distribution of the network is determined by the local conditional probability distributions as

$$P_\Theta(Z) = \prod_{j=1}^{N+1} P_\Theta\left(Z_j | Z_{\Pi_j}\right) \tag{1}$$

and the probability of a sample m is

$$P_\Theta(Z = z^m) = \prod_{j=1}^{N+1} \theta^j_{i|k} = \prod_{j=1}^{N+1} \prod_{i=1}^{|Z_j|} \prod_k \left(\theta^j_{i|k}\right)^{u^{j,m}_{i|k}}. \tag{2}$$

2.1. NB, TAN, and Bayesian multinet structures

The NB network assumes that all the attributes are conditionally independent given the class label. As reported in the literature (Friedman et al., 1997), the performance of the NB classifier is surprisingly good even if the conditional independence assumption between attributes is unrealistic in most of the data. The structure of the naive Bayes classifier represented as a Bayesian network is illustrated in Figure 2a. In order to correct some of the limitations of the NB classifier, Friedman et al. (Friedman et al., 1997) introduced the TAN classifier. A TAN is based on structural augmentations of the NB network, where additional

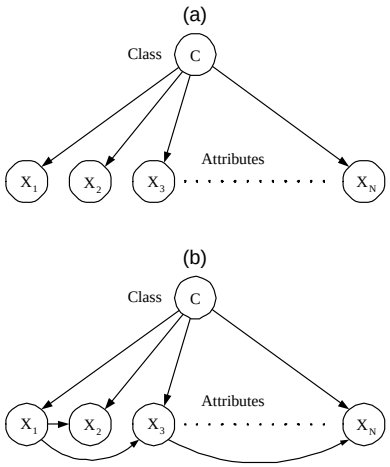

Figure 2. Bayesian Network: (a) NB, (b) TAN.

edges are added between attributes in order to relax some of the most flagrant conditional independence properties of NB. Each attribute may have at most one other attribute as an additional parent which means that the tree-width of the attribute induced sub-graph is unity. Hence, the maximum number of edges added to relax the independence assumption between the attributes is $N-1$. Thus, two attributes might not be conditionally independent given the class label in a TAN. An example of a TAN network is shown in Figure 2b. A TAN network is typically initialized as a NB network. Additional edges between attributes are determined through structure learning.

Bayesian multinets (Geiger & Heckerman, 1996) further generalize the TAN approach. In the TAN network, the dependencies amongst the attributes are the same for all values of the class node C. A Bayesian multinet has different edges for each class. Hence, depending on the class label, we can have a different (1-tree) network structure.

2.2. Structure learning

In this paper, the following approaches are used to learn the structure of the TAN network and the Bayesian multinet.

Generative structure learning: We use the CMI (Cover & Thomas, 1991) between the attributes given the class variable

$$I\left(X_i;X_j|C\right)=\sum_{x_i,x_j,c}P\left(x_i,x_j,c\right)\log\frac{P\left(x_i,x_j|c\right)}{P\left(x_i|c\right)P\left(x_j|c\right)}. \tag{3}$$

This measures the information between X_i and X_j in the context of C. Note, CMI is *not* a discriminative structure learning method, as it can be shown that augmenting the structure according to CMI will produce a guaranteed non-decrease in likelihood but does not necessarily help discrimination (Bilmes, 2000).

Discriminative structure learning: Bilmes (Bilmes, 2000) introduced an objective function $I\left(X_i;X_j|C\right)-I\left(X_i;X_j\right)$ which is called *explaining away residual (EAR)*. This measure prefers edges which are mutually informative conditioned on the class variable but simultaneously are not mutually informative unconditionally. The EAR measure is in fact an approximation to the expected log posterior, and so improving EAR is likely to decrease the KL-divergence (Cover & Thomas, 1991) between the true posterior and the resultant approximate posterior. The EAR measure is originally defined in terms of vectors of random variables. In our work, we use a scalar approximation of the EAR. Moreover, to simplify further, our approximate EAR-based structure learning procedure sequentially adds edges with an EAR value larger than zero to form the TAN network or Bayesian multinet, starting with the edge which corresponds to the largest EAR value. Additionally, we consider additions only that correspond to the tree-width unity constraint of the TAN and the Bayesian multinet structure.

We also evaluate a second discriminative measure, namely the CR (Keogh & Pazzani, 1999; Pernkopf, 2005)

$$CR=\frac{1}{|\mathcal{S}|}\sum_{m\in\mathcal{S}}^{|\mathcal{S}|}\delta\left(B_{\mathcal{S}}\left(x_{1:N}^m\right),c^m\right), \tag{4}$$

where $|\mathcal{S}|$ is the size of the training data $\mathcal{S}$. The expression $\delta\left(B_{\mathcal{S}}\left(x_{1:N}^m\right),c^m\right)=1$ if the Bayesian network classifier $B_{\mathcal{S}}\left(x_{1:N}^m\right)$ trained with samples $\mathcal{S}$ assigns the correct class label c^m to the attribute values $x_{1:N}^m$. Note that the CR scoring measure is determined from a classifier trained and tested on the same data $\mathcal{S}$. Exactly the same portion of data is used for learning the structure of the classifiers either with CMI, EAR, or CR (see Section 3).

In the case of the TAN structure, the network is initialized to NB and with each iteration we add the edge which gives the largest improvement of the classification rate. The CR approach is the most computationally expensive, as a complete re-classification of the training set is needed on each new evaluation of an edge. The CR, however, is the discriminative criterion that has the fewest number of approximations, so we expect it to do well. The greedy hill climbing search is terminated when there is no edge which further improves the score.

2.3. Parameter learning

Generative parameter learning: The parameters of the generative model are learned by maximizing the log likelihood of the data which leads to the ML estimation of $\theta_{i|k}^j$.

The log likelihood function of a fixed structure of B is

$$LL\left(B|\mathcal{S}\right)=\sum_{m=1}^{M}\log P_{\Theta}\left(Z=z^{m}\right)= \sum_{m=1}^{M}\sum_{j=1}^{N+1}\log P_{\Theta}\left(z_{j}^{m}|z_{\Pi_{j}}^{m}\right)= \sum_{m=1}^{M}\sum_{j=1}^{N+1}\sum_{i=1}^{|Z_{j}|}\sum_{k}u_{i|k}^{j,m}\log\left(\theta_{i|k}^{j}\right). \tag{5}$$

It is easy to show that the ML estimate of the parameters is

$$\theta_{i|k}^{j}=\frac{\sum_{m=1}^{M}u_{i|k}^{j,m}}{\sum_{m=1}^{M}\sum_{l=1}^{|Z_{j}|}u_{l|k}^{j,m}}, \tag{6}$$

using Lagrange multipliers to constrain the parameters to a valid normalized probability distribution.

Discriminative parameter learning: As mentioned above, for classification purposes, having a good approximation to the posterior probability is sufficient. Hence, we want to learn parameters so that CL is maximized. Unfortunately, CL does not decompose as does ML. Consequently, there is no closed-form solution and we have to resort to iterative optimization techniques. In our experiments, discriminative parameter learning is performed after the structure of the network has been determined. The objective function for the conditional log likelihood is

$$CLL\left(B|\mathcal{S}\right)=\log\prod_{m=1}^{M}P_{\Theta}\left(C=c^{m}|X_{1:N}=x_{1:N}^{m}\right)= \sum_{m=1}^{M}\log\frac{P_{\Theta}\left(C=c^{m},X_{1:N}=x_{1:N}^{m}\right)}{\sum_{c=1}^{|C|}P_{\Theta}\left(C=c,X_{1:N}=x_{1:N}^{m}\right)}=\sum_{m=1}^{M}\log\frac{P_{\Theta}\left(Z\right)}{\sum_{Z_{1}=1}^{|C|}P_{\Theta}\left(Z\right)}. \tag{7}$$

Similar to (Greiner & Zhou, 2002) we use a conjugate gradient descent algorithm with line-search (Press et al., 1992). Therefore, the derivative of the objective function is

$$\frac{\partial CLL\left(B|\mathcal{S}\right)}{\partial\theta_{i|k}^{j}}=\sum_{m=1}^{M}\left[\frac{\partial\log P_{\Theta}\left(Z\right)}{\partial\theta_{i|k}^{j}}-\frac{\partial\log\sum_{Z_{1}=1}^{|C|}P_{\Theta}\left(Z\right)}{\partial\theta_{i|k}^{j}}\right]= \sum_{m=1}^{M}\left[\frac{\partial P_{\Theta}\left(Z\right)}{\partial\theta_{i|k}^{j}}\frac{1}{P_{\Theta}\left(Z\right)}-\frac{\sum_{Z_{1}=1}^{|C|}\partial P_{\Theta}\left(Z\right)}{\partial\theta_{i|k}^{j}}\frac{1}{\sum_{Z_{1}=1}^{|C|}P_{\Theta}\left(Z\right)}\right]= \sum_{m=1}^{M}\frac{\mathbb{1}_{\left\{z_{j}^{m}=i,z_{\Pi_{j}}=k\right\}}}{P\left(Z_{j}=z_{j}^{m}|Z_{\Pi_{j}}=z_{\Pi_{j}}^{m}\right)}- \sum_{m=1}^{M}\frac{\sum_{Z_{1}=1}^{|C|}\prod_{n=1,n\neq j}^{N+1}P_{\Theta}\left(Z_{n}=z_{n}^{m}|Z_{\Pi_{n}}=z_{\Pi_{n}}^{m}\right)\mathbb{1}_{\left\{z_{j}^{m}=i,z_{\Pi_{j}}^{m}=k\right\}}}{\sum_{Z_{1}=1}^{|C|}\prod_{n=1}^{N+1}P_{\Theta}\left(Z_{n}=z_{n}^{m}|Z_{\Pi_{n}}=z_{\Pi_{n}}^{m}\right)}. \tag{8}$$

The probability $\theta_{i|k}^{j}$ is constrained to $\theta_{i|k}^{j}\geq 0$ and $\sum_{i=1}^{|Z_{j}|}\theta_{i|k}^{j}=1$. We reparameterize the problem to incorporate the constraints of $\theta_{i|k}^{j}$ and use different parameters $\beta_{i|k}^{j}$ as follows

$$\theta_{i|k}^{j}=\frac{\exp^{\beta_{i|k}^{j}}}{\sum_{l=1}^{|Z_{j}|}\exp^{\beta_{l|k}^{j}}}. \tag{9}$$

3. Experiments

Our goal in this paper is to evaluate both discriminative and generative parameter training on *both* discriminative *and* generatively structured Bayesian networks. Additionally, we use different Bayesian network topologies, the NB, TAN, and Bayesian multinet classifier. To perform our evaluations, we compare all of the above techniques on the data sets from the UCI repository (Merz et al., 1997) and from (Kohavi & John, 1997). The characteristics of the data are summarized in Table 1.

We use the same 5-fold cross-validation (CV) and train/test learning schemes as in (Friedman et al., 1997). To make the experimental setup clearer, the data have been split into five mutually exclusive subsets of approximately equal size $\mathcal{D}\in\{\mathcal{S}_{1},\mathcal{S}_{2},\mathcal{S}_{3},\mathcal{S}_{4},\mathcal{S}_{5}\}$. For parameter training (ML, CL) and structure learning (CMI, EAR, CR) of the Bayesian network classifier we use the same data $\mathcal{S}=\mathcal{D}\backslash\mathcal{S}_{c}$ $(1\leq c\leq 5)$. Therefore, we might get different structures for each CV-fold. Throughout the experiments, we use exactly the same data partitioning. Hence, for learn-

Table 1. Data sets used in the experiments.

	DATASET	# FEATURES	# CLASSES	# SAMPLES	
1	AUSTRALIAN	14	2	690	CV-5
2	BREAST	10	2	683	CV-5
3	CHESS	36	2	2130	1066
4	CLEVE	13	2	296	CV-5
5	CORRAL	6	2	128	CV-5
6	CRX	15	2	653	CV-5
7	DIABETES	8	2	768	CV-5
8	FLARE	10	2	1066	CV-5
9	GERMAN	20	2	1000	CV-5
10	GLASS	9	7	214	CV-5
11	GLASS2	9	2	163	CV-5
12	HEART	13	2	270	CV-5
13	HEPATITIS	19	2	80	CV-5
14	IRIS	4	3	150	CV-5
15	LETTER	16	26	15000	5000
16	LYMPHOGRAPHY	18	4	148	CV-5
17	MOFN-3-7-10	10	2	300	1024
18	PIMA	8	2	768	CV-5
19	SHUTTLE-SMALL	9	7	3866	1934
20	VOTE	16	2	435	CV-5
21	SATIMAGE	36	6	4435	2000
22	SEGMENT	19	7	1540	770
23	SOYBEAN-LARGE	35	19	562	CV-5
24	VEHICLE	18	4	846	CV-5
25	WAVEFORM-21	21	3	300	4700

Table 2. Classification results: Empirical accuracy of the classifiers in [%] with standard deviation.

CLASSIFIER STRUCT. LEARN. PARAM. LEARN.	NB - ML	NB - CL	TAN CMI ML	TAN CMI CL	TAN EAR ML	TAN EAR CL	TAN CR ML	TAN CR CL	MULITNET CMI ML	MULTINET CMI CL	MULTINET EAR ML	MULTINET EAR CL
DATA SET												
AUSTRALIAN	86.35 ± 0.99	**86.50** ±**1.05**	82.16 ±1.03	82.16 ±1.03	85.92 ±1.52	85.92 ± 1.52	85.05 ± 1.06	85.05 ± 1.05	81.86 ± 1.44	81.77 ± 1.56	84.47 ± 1.48	84.47 ± 1.48
BREAST	97.39 ± 0.59	97.39 ± 0.59	96.65 ± 0.73	96.65 ± 0.73	97.39 ± 0.59	97.39 ± 0.59	**97.69** ±**0.78**	**97.69** ±**0.78**	96.50 ± 0.44	96.50 ± 0.44	97.39 ± 0.59	97.39 ± 0.59
CHESS	87.14 ± 1.03	88.55 ± 0.98	92.40 ± 0.81	93.43 ± 0.76	94.18 ± 0.72	94.18 ± 0.72	96.06 ± 0.60	**96.25** ±**0.58**	92.50 ± 0.81	92.50 ± 0.81	91.93 ± 0.83	91.93 ± 0.83
CLEVE	**84.12** ±**2.44**	83.44 ±2.60	79.38 ± 2.12	79.72 ± 1.53	81.07 ± 3.51	81.07 ± 3.16	80.06 ± 2.31	81.42 ± 1.61	79.38 ± 2.05	80.40 ± 2.07	80.06 ± 2.32	80.06 ± 2.32
CORRAL	86.66 ± 3.29	88.89 ± 4.31	99.20 ± 0.80	**100.00** ±**0.00**	99.20 ±0.80	**100.00** ±**0.00**	96.80 ±1.96	96.80 ±1.96	98.40 ± 0.98	99.20 ± 0.80	99.20 ± 0.80	99.20 ± 0.80
CRX	85.60 ± 2.25	85.60 ± 1.97	82.40 ± 1.02	82.40 ± 1.02	85.75 ± 1.61	**86.06** ±**1.66**	85.45 ± 1.97	85.45 ± 1.97	82.56 ± 1.27	82.56 ± 1.27	84.22 ± 1.37	84.53 ± 1.25
DIABETES	72.80 ± 1.19	74.75 ± 0.95	70.45 ± 1.41	74.36 ± 1.27	72.80 ± 1.19	74.88 ± 1.00	**75.79** ±**1.38**	**75.79** ±**1.46**	75.39 ± 1.20	75.54 ± 1.78	74.86 ± 1.72	75.51 ± 1.80
FLARE	**83.11** ±**0.51**	**83.11** ±**0.51**	82.93 ±0.37	82.93 ±0.37	**83.11** ±**0.51**	**83.11** ±**0.51**	82.64 ±1.14	82.92 ±1.19	81.71 ±1.17	81.71 ± 1.09	80.58 ± 2.32	81.43 ± 2.15
GERMAN	70.00 ± 0.00	69.90 ± 0.19	69.20 ± 0.49	69.00 ± 0.57	69.70 ± 0.43	69.60 ± 0.43	72.30 ± 1.13	72.20 ± 1.16	72.40 ± 1.89	72.60 ± 1.92	**73.40** ±**1.86**	73.30 ±1.99
GLASS	65.22 ± 1.62	66.70 ± 1.33	65.33 ± 1.41	65.33 ± 1.41	66.72 ± 2.17	68.20 ± 1.65	**71.44** ±**1.27**	70.94 ± 1.26	69.94 ± 1.09	69.94 ± 1.09	70.33 ± 2.72	70.33 ± 2.72
GLASS2	80.38 ± 2.50	81.00 8± 1.73	82.20 ± 2.08	79.70 ± 3.56	80.38 ± 2.50	80.38 ± 1.82	**82.25** ±**2.38**	81.63 ± 2.11	80.32 ± 2.95	80.32 ± 2.95	81.52 ± 2.10	81.52 ± 2.10
HEART	81.85 ± 2.22	82.59 ± 2.79	81.11 ± 1.36	82.59 ± 2.44	81.48 ± 2.49	83.33 ± 2.93	**84.07** ±**2.72**	81.85 ± 2.95	82.96 ± 2.14	83.33 ± 1.94	**84.07** ±**2.24**	**84.07** ±**2.53**
HEPATITIS	87.00 ± 3.96	88.33 ± 4.15	84.33 ± 2.33	85.66 ± 2.96	84.33 ± 3.14	85.67 ± 3.64	**91.67** ±**3.42**	**91.67** ±**3.42**	84.33 ± 4.33	84.33 ± 4.33	85.67 ± 3.64	85.67 ± 3.64
IRIS	93.33 ± 0.00	93.33 ± 0.00	94.00 ± 1.25	94.00 ± 1.25	93.33 ± 0.00	93.33 ± 0.00	93.33 ± 1.06	92.67 ± 0.67	**94.67** ±**1.33**	**94.67** ±**1.33**	93.33 ± 0.00	93.33 ± 0.00
LETTER	74.92 ± 0.61	74.94 ± 0.61	87.46 ± 0.47	87.46 ± 0.47	85.72 ± 0.49	85.72 ± 0.49	- -	- -	**88.26** ±**0.46**	**88.26** ±**0.46**	86.72 ± 0.48	86.72 ± 0.48
LYMPHOGRAPHY	84.77 ± 4.14	84.18 ± 4.16	85.39 ± 2.52	**85.98** ±**2.58**	81.24 ±4.75	81.83 ± 4.54	85.39 ± 3.87	85.39 ± 3.87	**85.98** ±**4.02**	**85.98** ±**4.02**	83.01 ± 4.35	82.42 ± 4.79
MOFN-3-7-10	86.42 ± 1.07	91.50 ± 0.87	91.70 ± 0.86	**94.14** ±**0.73**	91.70 ±0.86	**94.14** ±**0.73**	91.02 ±0.89	93.07 ± 0.79	91.50 ± 0.87	92.68 ± 0.81	91.21 ± 0.88	92.29 ± 0.83
PIMA	71.36 ± 1.84	73.56 ± 1.35	69.66 ± 1.58	71.49 ± 1.86	70.71 ± 1.40	73.05 ± 1.00	**75.65** ±**0.24**	75.39 ± 0.44	75.39 ± 0.68	75.39 ± 0.68	75.13 ± 0.74	75.26 ± 0.77
SHUTTLE-SMALL	98.86 ± 0.24	98.86 ± 0.24	99.48 ± 0.16	99.48 ± 0.16	98.97 ± 0.23	98.97 ± 0.23	99.22 ± 0.20	99.28 ± 0.19	**99.69** ±**0.13**	**99.69** ±**0.13**	99.53 ± 0.16	99.53 ± 0.16
VOTE	91.50 ± 1.00	91.73 ± 0.98	93.58 ± 1.05	93.58 ± 1.05	92.65 ± 0.86	92.88 ± 0.98	**93.82** ±**1.35**	93.59 ± 1.45	93.80 ± 0.58	93.80 ± 0.58	91.48 ± 0.46	91.25 ± 1.03
SATIMAGE	81.75 ± 0.86	81.75 ± 0.86	86.90 ± 0.75	86.90 ± 0.75	81.75 ± 0.86	81.75 ± 0.86	- -	- -	**87.25** ±**0.75**	**87.25** ±**0.75**	85.40 ± 0.79	85.40 ± 0.79
SEGMENT	92.34 ± 0.96	92.47 ± 0.95	94.68 ± 0.81	94.68 ± 0.81	92.60 ± 0.94	92.60 ± 0.94	**95.97** ±**0.71**	**95.97** ±**0.71**	95.06 ± 0.78	95.06 ± 0.78	92.99 ± 0.92	92.99 ± 0.92
SOYBEAN-LARGE	**93.94** ±**0.19**	**93.94** ±**0.19**	89.29 ±1.30	89.29 ±1.30	92.18 ±0.85	92.18 ±0.85	- -	- -	92.89 ± 0.34	92.56 ± 0.52	92.67 ± 0.53	92.67 ± 0.53
VEHICLE	58.78 ± 1.15	58.78 ± 1.15	**68.21** ±**1.23**	**68.21** ±**1.23**	62.00 ±2.19	61.89 ±2.20	67.88 ±0.84	67.64 ± 0.84	67.25 ± 0.93	67.25 ± 0.80	63.46 ± 1.99	63.46 ± 1.99
WAVEFORM-21	78.02 ± 0.60	78.45 ± 0.60	77.74 ± 0.61	77.72 ± 0.61	75.30 ± 0.63	76.00 ± 0.62	78.19 ± 0.60	**78.57** ± **0.60**	77.53 ± 0.61	77.47 ± 0.61	76.83 ± 0.62	76.85 ± 0.62
AVG. PERFORMANCE	82.94	83.60	84.23	84.67	84.01	84.56	**85.53**	85.51	85.10	85.23	84.77	84.86
% OF BEING BEST	12	12	4	16	4	16	**40.91**	27.27	20	20	8	4

Table 3. Comparison of different classifiers using the one-sided paired t-test: Each entry of the table gives the significance of the difference of the classification rate of two classifiers over the data sets. The arrow points to the superior learning algorithm. We use a double arrow if the difference is significant at the level of 0.05.

CLASSIFIER STRUCT. LEARN. PARAM. LEARN.	NB - CL	TAN CMI ML	TAN CMI CL	TAN EAR ML	TAN EAR CL	TAN CR ML	TAN CR CL	MULITNET CMI ML	MULTINET CMI CL	MULTINET EAR ML	MULTINET EAR CL
NB-ML	⇑0.0071	↑0.079	⇑0.041	↑0.080	⇑0.029	⇑0.0007	⇑0.0008	⇑0.014	⇑0.011	⇑0.018	⇑0.015
NB-CL		↑0.16	↑0.098	↑0.17	↑0.078	⇑0.0032	⇑0.0027	⇑0.044	⇑0.035	↑0.057	⇑0.046
TAN-CMI-ML			⇑0.034	←0.18	↑0.16	⇑0.0015	⇑0.0013	⇑0.021	⇑0.01	↑0.11	↑0.097
TAN-CMI-CL				←0.076	←0.19	⇑0.008	⇑0.0062	↑0.094	⇑0.049	↑0.19	↑0.18
TAN-EAR-ML					⇑0.0018	⇑0.0008	⇑0.0005	⇑0.022	⇑0.012	⇑0.027	⇑0.016
TAN-EAR-CL						⇑0.01	⇑0.0081	↑0.11	↑0.078	↑0.16	↑0.13
TAN-CR-ML							←0.19	⇐0.0082	⇐0.025	⇐0.0025	⇐0.0055
TAN-CR-CL								⇐0.012	⇐0.029	⇐0.0045	⇐0.0077
MULTINET-CMI-ML									⇑0.041	←0.11	←0.15
MULTINET-CMI-CL										←0.067	←0.10
MULTINET-EAR-ML											↑0.087

ing and testing the classifiers the same information is available.

The attributes in the data sets are multinomial and continuous-valued. Since the classifiers are constructed for multinomial attributes, the features have been discretized in a manner described in (Fayyad & Irani, 1993) where the codebook is produced using only the training data. Zero probabilities in the conditional probability tables are replaced with a small epsilon $\varepsilon = 0.00001$, implementing a simple but effective form of Dirichlet smoothing.

Table 2 compares the recognition rate of the NB, the TAN, and the Bayesian multinet classifier. We use discriminative and non-discriminative structure learning (EAR or CR vs. CMI) and parameter learning (CL vs. ML) approaches. The best achieved classification accuracy is emphasized by boldface letters. The bottom two lines in the table give the average classification rate (Avg. Performance) over the 25 data sets of the selected parameter and structure learning approach and the percentage over the employed data sets where this technique is best. Note that this row (% of being best) does not sum up to 100% since there can be several classifiers which perform best on a particular data set.

Table 3 presents a summary of the classification results over all 25 data sets from Table 2. We compare all pairs of classifiers using the one-sided paired t-test (Mitchell, 1997). The t-test determines whether the classifiers differ significantly under the assumption that the paired classification differences over the data set are independent and identically normally distributed. In this table each entry gives the significance of the difference in classification rate of two classification approaches. The arrow points to the superior learning algorithm and a double arrow indicates whether the difference is significant at a level of 0.05.

In Figure 3 we show the scatter plot of the classification error comparing the different techniques against the TAN-CR-ML classifier which achieves on average the best performance. Points above the diagonal line in the scatter plot corresponds to data sets where the TAN-CR-ML classifier performs better. The CR score produces in 41% of the cases the best performing network structures. This generally says that discriminative structure learning is sufficient to produce good generative classifiers, even when using maximum likelihood training. Indeed, it has been suggested that a good discriminative structure might possibly obviate discriminative parameter learning (Bilmes, 2000). However, the evaluation of the CR measure is quite computationally expensive — the number of classifications used to learn structure in the TAN-CR case is shown in Figure 4.

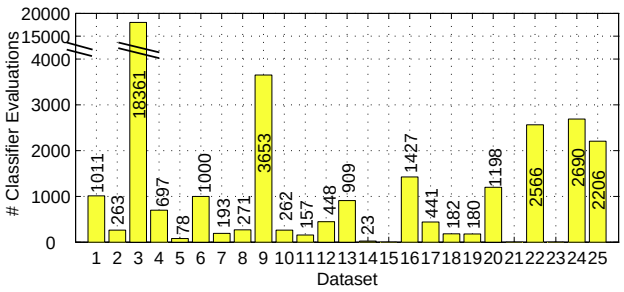

Figure 4. Number of classifier evaluations in TAN-CR structure learning.

Moreover, during structure learning of the TAN-CR, ML parameter training is used. Once the structure is determined we use CL parameter optimization to get the TAN-CR-CL. This might instead be the reason why TAN-CR-ML performs in the majority of the cases better than TAN-CR-CL, 41% versus 27%. Optimizing the structure using CR while learning the parameters using the CL is computationally infeasible. In general, discriminative parameter learning (CL) produces better average classification (Avg. Performance) than ML parameter learning. As noticed in (Bilmes, 2000),(Greiner & Zhou, 2002), discriminative

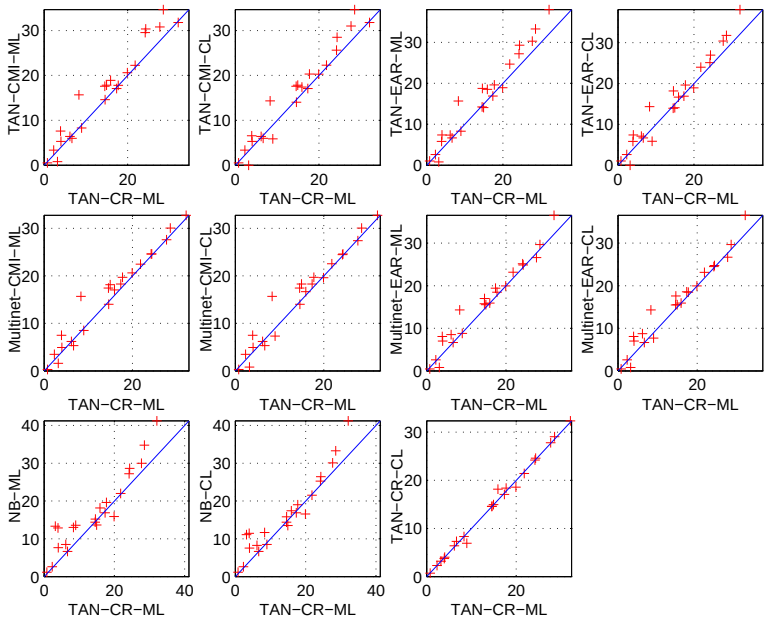

Figure 3. Classifi cation error: TAN-CR-ML versus competing NB, TAN, Bayesian multinet classifi ers. Points above the diagonal line in the scatter plot corresponds to data sets where the TAN-CR-ML classifi er performs better.

structure learning procedures work despite not representing well the generative structure. As mentioned earlier, this is because discriminative models need not represent the joint distribution, rather they need only classify the data well.

Our EAR approximation sometimes produces very sparse networks which produce inferior performance for some data sets. Currently, we add edges in the case of EAR larger than zero. Since this threshold can be inadequate, we intend in future work to use better EAR approximations. Note, the EAR measure can be extended with terms that better approximate the posterior (Bilmes, 1999; Çetin, 2004) and better EAR optimizations have also recently been developed (Narasimhan & Bilmes, 2005).

Lastly, the average number of parameters used for the different classifier structures are summarized in Table 4.

4. Conclusion

Discriminative and generative approaches for parameter and structure learning of Bayesian network classifiers have been compared. For parameter training we compare maximum likelihood estimation and optimizing the conditional likelihood. For structure learning we use either the conditional mutual information, the explaining away residual, or the classification rate as a scoring function. The latter two are discriminative measures. Experiments have been performed with the naive Bayes, the tree augmented naive Bayes, and the Bayesian multinet classifier using 25 data sets. In general, discriminative structure and/or parameter learning produces more accurate classifiers. The best performing network structures are obtained using the classification rate as objective function. Thus, we have empirically found that discriminatively structured (using CR) and maximum likelihood trained classifiers work best out of all the generative classifiers we tried. This approach, however, is the most computationally expensive.

References

Bahl, L., Brown, P., de Souza, P., & Mercer, R. (1986). Maximum Mutual Information estimation of HMM parameters for speech recognition. *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing* (pp. 49–52).

Bilmes, J. (1999). *Natural statistical models for automatic speech recognition*. Doctoral dissertation, U.C. Berkeley.

Bilmes, J. (2000). Dynamic Bayesian multinets. *16th Inter. Conf. of Uncertainty in Artificial Intelligence (UAI)* (pp. 38–45).

Çetin, O. (2004). *Multi-rate modeling, model inference, and estimation for statistical classifiers*. Doctoral dissertation, University of Washington.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. John Wiley & Sons.

Cowell, R., Dawid, A., Lauritzen, S., & Spiegelhalter, D. (1999). *Probabilistic networks and expert systems*. Springer Verlag.

Table 4. Average number of parameters for learned structure of different classifi ers.

CLASSIFIER	NB	TAN			MULITNET	
STRUCT. LEARN.	-	CMI	EAR	CR	CMI	EAR
DATA SET						
AUSTRALIAN	71.0	546.2	154.2	219.0	534.2	162.8
BREAST	37.4	113.0	37.4	51.0	109.6	36.0
CHESS	75.0	149.0	149.0	151.0	142.0	145.0
CLEVE	32.6	89.4	49.0	60.6	83.0	51.2
CORRAL	13.0	23.0	21.4	17.8	22.0	20.4
CRX	74.6	562.6	172.2	235.0	537.4	171.8
DIABETES	18.2	46.2	20.6	32.6	45.0	20.4
FLARE	37.0	169.4	39.4	69.4	171.8	49.4
GERMAN	88.2	581.0	241.8	441.0	551.6	234.0
GLASS	92.8	267.8	126.4	167.0	190.6	121.0
GLASS2	12.6	27.8	16.2	15.8	25.8	15.8
HEART	17.4	31.8	19.4	20.2	31.2	18.4
HEPATITIS	31.4	59.8	54.6	32.2	55.4	52.4
IRIS	21.8	46.4	21.8	32.6	35.6	19.8
LETTER	6265.0	94015.0	94015.0	-	93990.0	93990.0
LYMPHOGRAPHY	130.2	573.4	285.4	203.8	398.0	256.2
MOFN-3-7-10	21.0	39.0	39.0	35.0	38.0	38.0
PIMA	17.0	37.8	18.6	26.6	37.4	18.6
SHUTTLE-SMALL	307.0	2085.0	1483.0	713.0	1204.0	1036.0
VOTE	65.0	185.0	111.4	105.0	184.0	116.0
SATIMAGE	2093.0	21389.0	2093.0	-	21810.0	7102.0
SEGMENT	916.0	8931.0	1875.0	603.8	8438.0	3299.0
SOYBEAN-LARGE	974.0	3539.0	2750.0	-	2329.0	2081.8
VEHICLE	185.4	816.6	371.8	603.8	789.8	369.0
WAVEFORM-21	80.0	221.0	212.0	125.0	213.0	189.0

Fayyad, U., & Irani, K. (1993). Multi-interval discretizaton of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1022–1027).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Geiger, D., & Heckerman, D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, *82*, 45–74.

Greiner, R., & Zhou, W. (2002). Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *18th Conf. of the AAAI* (pp. 167–173).

Grossman, D., & Domingos, P. (2004). Learning bayesian network classifiers by maximizing conditional likelihood. *21st Inter. Conf. of Machine Lerning (ICML)* (pp. 361–368).

Jebara, T. (2001). *Discriminative, generative and imitative learning*. Doctoral dissertation, Media Laboratory, MIT.

Keogh, E., & Pazzani, M. (1999). Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. *Proceedings of 7th International Workshop on Artificial Intelligence and Statistics* (pp. 225–230).

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

Merz, C., Murphy, P., & Aha, D. (1997). UCI repository of machine learning databases. . Department of Information and Computer Science, University of California, Irvine, URL: www.ics.uci.edu/˜mlearn/MLRepository.html.

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Narasimhan, N., & Bilmes, J. (2005). A supermodular-submodular procedure with applications to discriminative structure learning. *21st Inter. Conf. on Uncertainty in Artificial Intelligence (UAI)*.

Ng, A., & M., J. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems 14*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Pernkopf, F. (2005). Bayesian network classifiers versus selective k-NN classifier. *Pattern Recognition*, *38*, 1–10.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical recipes in C*. Cambridge Univ. Press.

Raina, R., Shen, Y., Ng, A., & McCallum, A. (2004). Classification with hybrid generative/discriminative models. *Advances in Neural Information Processing Systems 16*.

Vapnik, V. (1998). *Statistical learning theory*. Wiley & Sons.