A HYBRID GENERATIVE-DISCRIMINATIVE LEARNING ALGORITHM FOR BAYESIAN NETWORK STRUCTURE

XIAO-BO JIN, XIN-WEN HOU, CHENG-LIN LIU

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing 100080, P.R. China
E-MAIL: {xbjin,xwhou, liucl}@nlpr.ia.ac.cn

Abstract

The discriminative learning of Bayesian networks benefits the classification accuracy as compared to generative learning. Previous approaches mostly learn either the structure or the parameters in a discriminative manner based on the scoring+ search paradigm. Many works have focused on structure learning by optimizing a discriminative scoring function but the resulted structure is still generative in the sense that the class variable is not conditioned on attribute variables. On the other hand, searching Markov Blanket in a constrained space can generate a hybrid generative-discriminative structure. In this paper, we propose a new Hybrid Generative-Discriminative (HGD) algorithm for learning Bayesian network structure. The algorithm searches the neighboring structures by optimizing a cross-validated classification rate (CR) criterion to give a really discriminative structure. We select the initial structure and design neighborhood operators appropriately such that the learning procedure is computationally feasible. Our empirical study on a large suite of bench-mark datasets shows that the proposed HGD+ CR algorithm yields better classification results than BN classifiers with only discriminative scores.

Keywords: Classification; Bayesian Network; Generative Learning; Discriminative Learning; Generative Structure; Discriminative Structure; Markov Blanket

1. Introduction

Bayesian network (BN) is a directed graphical model that represents the probabilistic relationship between attributes. The Naive Bayes (NB) classifier [1], a simplistic form of BN that assumes class-conditionally independence between attributes, has led to a wide use because of its simple implementation and acceptable classification accuracy. Due to the inherent dependence between attributes in practical classification problems, to learn a more sophisticated network structure by relaxing the conditional independence assumption is expected to improve the classification performance. A general paradigm, called scoring+search, is to search a network structure topology by maximizing a defined score function. The score should consider both the classification capability and the parsimony of the structure, while the optimization is usually performed by local search (hill-climbing), which moves to neighboring structures or parameter configurations iteratively to find a structure or configuration of maximum score. Another paradigm, called constraint-based learning [2], performs conditional independence test to explore the dependence and independence relationships between the attributes.

The scoring+search paradigm has been more widely pursued. The concrete methods vary in the initial structure topology, the form of score function, the definition of structure neighborhood, the combination of structure learning and parameter learning, etc. Very often, the NB is taken as a seed for learning augmented structures by adding arcs between attribute nodes for accounting for their dependence. The resulted structure is called augmented naive Bayes networks (ABNs). The k-dependence ABN [3] is a BN admitting each attribute to have at most k attributes except the class node as its parents. The special structure of Bayesian network augmented naive Bayesian classifier (BAN) [4] is an N-dependence ABN (N is the number of attributes). The tree-augmented naive Bayes (TAN) of Fried-man et al. [4] is a 1-dependence BN classifier, a special case of k-dependence ABN. A variant of TAN, called forest-augmented naive Bayesian network (FAN) [5], builds a forest depending on the relationship between the attributes. The learning of the above structures takes the maximum likelihood (ML) as the objective score function. On the other hand, Langseth and Nielsen proposed a hierarchical naive Bayes model (HNB) [6] which adds latent variables between the class node and the attributes by optimizing the cross-validation classification accuracy. The (or cross-validation accuracy) on the training data is a discriminative learning criterion while the ML is a generative learning criterion. Another discriminative criterion pertinent to classification accuracy is the conditional likelihood (or conditional log-likelihood, CLL) [7, 8], which measures the a posteriori probability of class variable conditioned on the attributes.

Whereas the above discussions concern generative/discriminative learning of BN structures, the learning of BN parameters can also be dichotomized into generative
and discriminative. Again, the ML criterion has been popularly used for parameter estimation in previous works. Yet in recent years, learning parameters by optimizing a discriminative criterion is shown to yield higher classification performance. For example, Greiner and Zhou [9] maximize the CLL criterion by the conjugate gradient method when giving the network structure. Pernkopf and Bilmes [10] investigated into the effects of various combinations of generative/discriminative learning of structures and parameters. They observed that the discriminative structure learning is sufficient to produce good classifiers even though the parameters are estimated by generative learning. Further, the combination of discriminative structure learning and generative parameter learning was shown to perform best out of all the generative classifiers they tried.

In structure learning of BNs, previous methods mostly use the NB structure as the initial state and the neighborhood structures are obtained by adding augmenting arcs between attributes. This restricted neighborhood makes the final learned structure remains an ANB, which is still generative in the sense that the class variable is not conditioned on attribute variables. In this paper, we propose a new hybrid generative-discriminative (HGD) algorithm for learning the structure of BN classifier. The algorithm first learns a generative ANB structure (specifically, the TAN) as the initial state. The initial structure is then discriminatively modified using two simple operations: reversing the arcs and deleting the arcs. The first operation can generate a hybrid generative-discriminative structure which has the class node conditioned on some attribute variables. And by the second one, we can obtain a reduced structure which contains the MB of the class node. On one hand, the proposed algorithm extends the search space outside ANB, on the other hand, it restricts the search space to the skeleton (undirected graph) of ANB structure. For the scoring function for evaluating the structures, we use the cross-validation (holdout method) CR to avoid over-fitting. At last, the generation of MB also makes feature selection be an integral part of our learning method. Previous method HBayes-TAN[11] split the attribute set into two parts learned separately by the logistic regression and the generative method (TAN), unlike our method learned in a single framework.

The rest of this paper is organized as follows. Section 2 reviews the structure learning and parameter learning method of BNs. Section 3 introduces our proposed HGD learning algorithm. The empirical evaluation results with comparison to state-of-the-art classifiers are given in Section 4. Concluding remarks are given in Section 5.

2. Bayesian Network Classifier

BNs [12] are graphical models that encode the joint probability distribution of a set of random variables $X = \{X_1, X_2, X_N\}$ as the nodes connected by directed edges in a graph. It is defined as the pair $\theta = \{G, \theta\}$, $\theta$ is its structure and $\theta$ is the parameter vector attached to the structure $G$. The joint distribution on $B$ can be factored by $P_B(X) = \prod_{i=1}^{N} P_B(X_i | Pa(X_i))$. Given a collection of i.i.d. observed data $D = \{x^1, x^2, \cdots x^M\}$ where $x^i = \{x^i_1, x^i_2, \cdots x^i_N, c^i\}$ and $c^i$ is the class label of the $i$’th example, the optimal prediction for the test example $x$ is the class which maximize $P(c | x)$ [13] to make the classification accuracy as high as possible.

2.1. Structure Learning

2.1.1 Generative Learning

The structure learning of BN is dichotomized as: generative learning and discriminative learning [10]. The standard approach for generative learning is to find a structure which maximizes the likelihood. Friedman et al. introduced TAN by extending Chow and Liu’s result [14] from the tree structure to the TAN structure where the class variable has no parent and each attribute has the class variable and at most one other attribute as parent. The algorithm generates an optimal 1-dependence tree structure by maximizing the sum of conditional mutual information (CMI) between attributes conditioned the class variables which make the likelihood maximum, where

$$LL(B | D) = M \cdot \sum_{Pa(X_j) / C} I_\theta (X_i; Pa(X_i) / C | C) + C_0$$

(1)

where $C_0$ is a constant term.

2.1.2 Discriminative Learning

Discriminative structure learning optimizes the discriminative objective function such as CLL, classification rate (CR) [15], explaining away residual (EAR) [16] instead of ML. CR essentially measures the classification accuracy on the training data when given the network structure. CLL is directly relevant for the classification purposes, where $CLL(B | D) = \sum_{i=1}^{M} \log P_B(c^i | x^i)$. Note that for CR criterion a classifier is trained and tested both on the training data $D$. EAR is regard as an approximation to the expected log posteriori and defined as $I(X_i; X_j)$ where $X_i, X_j$ are the attributes. Both of CR and CLL need to travel all the training data twice for each structure estimation. One is to learn structure parameters; the other is to evaluate the structure. The CR (CLL) criterion is computationally expensive evaluated on each new structure. But they directly measure the classification performance of the network.
structure and can be expected to give a better result. When the locally optimal solution is reached during the hill climbing search, the learning is terminated.

2.1.2 Parameters Learning

As contrast to the structure learning, the parameter learning may be divided into: generative learning and discriminative learning. Generative learning maximizing the log likelihood \( P_D(x_i | Pa(x_i)) \). In this paper, we use a Dirichlet prior to smooth the ML estimation \[ (\cdot) \]

\[
P^*_D(x_i | Pa(x_i)) = \frac{M \cdot P_D(x_i, Pa(x_i))}{M \cdot P_D(Pa(x_i)) + N_0}
\]

where \( N_0 \) is empirically set to 5 as in Friedman et al.’s experiments. \( P_D(\cdot) \) is the ML estimation of the probability on dataset \( D \). When the objective function of the BN classifier is CLL, a form of the discriminative learning, there is no known closed form for the optimal parameters estimates. ELR algorithm [9] found the locally optimal estimates by a numeric method such as conjugate gradient with line search.

3. Hybrid Generative-Discriminative Structure Learning (HGD)

Although the discriminative learning generally performs better than the generative learning, the resulted ABN structure in previous methods is still generative in the sense that the class variable is not conditioned on attribute variables. In the following we will introduce generative /discriminative BN structure and HGD algorithm.

3.1. Generative and Discriminative BN structure

There are two paradigms for the statistical learning of the classifiers: generative and discriminative [12]. The generative learning models the distribution of the inputs as well as outputs explicitly or implicitly by learning the prior class probability and the class-conditional probability. While discriminative learning directly models the posterior class probability, or use a discriminant function. In general, the discriminative learning can obtain a better classification performance than the generative learning. Based on the above definition, BN structures may fall into two types: A generative structure and discriminative structure. The generative structure as shown in Figure 1a) has the following probability decomposition regarding the probability:

\[
P_B(C | X) = P_B(C) \prod_{i=1}^{N} P_B(X_i | Pa(X_i))
\]

(3)

ANB,HNB,TAN belong to such structures. The posterior probability of the discriminative structure as shown in Figure 1b) (general containing generative substructure) may be decomposed as:

\[
P_B(C | X) = P_B(C | Pa(C)) \prod_{Y \in Children(C)} P_B(X | Pa(X))
\]

(4)

Figure 1. Generative and Discriminative Structure:

a) A generative structure (ANB);
b) A generative-discriminative structure(DAG)

Note that \( Pa(X_i) \) and \( Pa(X) \) in the above formulas contain the class node. The first item in (4) directly models the posterior probability dependent on some attributes of the input and it corresponds to a discriminative substructure. The latter production items estimate the class-conditional probability on other attributes, which produce a generative substructure. The set of nodes comprising \( X_i \)’s parents, \( X_i \)’s children and the parents of \( X_i \)’s children in \( G \) is called \( X_i \)’s Markov blanket (MB). We can think of the node \( X_i \)’s MB as being the minimal set of nodes that isolates \( X_i \) from the rest of the graph \( G \). All nodes except the class node \( C \) in (4) comprise MB of the class node \( C \). So MB can be contained in a generative-discriminative structure.

3.2. HGD algorithm

The maximal weighted spanning tree (MST) in TAN maximizes \( LL(B_T | D) \) [4] in which the dependence relationships between the attributes are kept as much as possible. The expected generative-discriminative structure can be generated from such an informative structure. So we

\[
P_B(C | X) = P_B(C) \prod_{i=1}^{N} P_B(X_i | Pa(X_i))
\]

(3)
first build a generative structure by TAN algorithm.

Next we run the discriminative learning by the hill-climbing algorithm as many previous works [17, 7]. At each point we consider its all neighborhood DAGs and estimate their discriminative scores. The neighborhood graphs include all DAGs by inverting or deleting any edge on current graph. We have no consideration about the addition for shrinkage of the graph structure. During discriminative learning of the network structure, we can find a generative-discriminative structure which defines the MB of the class node. The posteriori probability can be estimated by (4).

Two discriminative functions are adopted for the estimation of the network structures: CR and CLL. For the evaluation, the training data is split into two parts: one is used for learning the parameters attached to the structure; the other is used for estimating the CR and CLL measures. It is different from the commonly CR scoring criterion which is determined through a classifier trained and tested both on the training data. Holdout CR (CLL) avoids the over-fitting in discriminatively-trained BN. In general, the higher the posteriori probability of a example on its true class is, the bigger the probability it is correctly classified. The results of two measures in the sense of classification is similar shown in our experiments. Parameters in each network structure are set to their maximum likelihood values as (2). ML estimation is extremely fast compared with the discriminative parameters learning [10].

We see from the algorithm paradigm that the initial TAN structure and the holdout CR (CLL) make the classification performance of HGD be no worse than that of TAN. It is expected that the BN classifier with a generative-discriminative structure performs better than a generative structure. The deletion operation generate a network structure containing less nodes. However, the evaluation of the CR in the search space is quite computationally expensive. In the neighborhood of each graph, there is $O(N)$ graph structures to be evaluated. Although its search space complexity is $O(2^{2N-1})$ smaller than the full search space $O(2^{(N+1)N/2})$, it still is a NP-hard problem. When the number of the examples and the nodes is high, HGD will be computationally infeasible. In our future research, we will try to the other random search strategy such as simulated annealing, tabu search or genetic search and simpler discriminative measures to accelerate the searching of the optimal graph.

4. Experiments

In this section, we will demonstrate the effectiveness of HGD algorithm. We compare our algorithm (HGD+CR and HGD+CLL) with other five state-of-the-art algorithms on 25 benchmark datasets from Friedman et al.[4]. These include 23 datasets from UCI repository [18] and two extra datasets designed by Kohavi and John for feature selection experiments. We use the same settings as in [4] to choose 5-fold cross-valid or hold out testing for the experiments. These datasets has the same preprocessing steps: the continuous variables have been discretized using the procedure A proposed by Fayyad and Irani [19] and the instances with missing values are eliminated.

NB, TAN, HGD+CR and HGD+CLL are developed on Torch Library [20], a C++ platform for the machine learning research. The other three algorithms run on Weka toolkit [21] and all parameters are set by default. Figure 2 displays the accuracies of the seven algorithm using abbreviations:

- **NB**: the naive Bayes classifier
- **TAN**: the tree-augmented naive Bayes Classifier
- **HGD+CR** and **HGD+CLL**: two hybrid learning classifier, the former using the CR score and the latter CLL score.
- **C4.5**: the decision-tree induction method developed by Quinlan[22]. In Weka library is a variant of C4.5 called J4.8 algorithm.
- **K2**: K2 is a hill climbing algorithm by Greg Cooper and Ed Herskowitz [23]. It first assumes that a node has no parents and then adds incrementally that node among the predecessors in the ordering as a parent which increases the probability of the resultant structure by the largest amount. It stops adding parents to the node when the addition of no additional single parent can increase the probability.
- **AODE**: the aggregating one-dependence estimators [24] is a extended NB to weaken the attribute independence assumption by averaging all of a constrained class of classifiers.

Hold-out estimation in HGD+CR and HGD+CLL splits the training data into two parts: 90% is used for the parameter learning and the rest 10% for structure learning by maximizing CR or CLL. From Figure 2, we can see the average accuracy 84.04% of NB is slightly better than Pernkopf and Bilmes’s [10] result 82.94%, but the difference between TAN and NB about the average accuracy is consistent with Pernkopf and Bilmes’s, i.e., 84.04% - 84.23% = 84.23% - 82.94%. It is important that HGD+CR outperforms other algorithms significantly on most datasets. For examples, it achieves a 5%-6% difference higher than TAN on some datasets and 4% on average accuracy. On another hand, HGD+CR and HGD+CLL are near each other on the classification performance. NB, K2,C4.5 show similar accuracies on average. C4.5 performs better than HGD + CR on three datasets. From the average accuracy, we see TAN and AODE have a improvement than NB, K2,C4.5 as shown in...
[4, 8, 24]. To determine the difference between two algorithms is significant or not, we gather p-value by paired t-test [25] into Figure 3. Each item in the table has a value that indicates the classifier in the row is better than the other ones in column and the positive sign indicates a improvement. We can see the item corresponding to both of the row and the column of the HGD + CR, which show a important significance than other classifiers (p < 0.05 is denoted by boldface).

![Figure 2: Experimental results: classification accuracy (×100)](image)

![Figure 3: Each value in the grid indicates whether the classifier in the row improves the one in the column or not](image)

Our experimental settings are the same as [10]. From the results of Pernkopf and Bilmes, the TAN with discriminative criterion (TAN+CR+ML) can improve 1.3% on average than the TAN with the generative criterion (TAN+CM1+ML)(ML is the criterion for parameter learning). In our work, using both discriminative structure and discriminative criterion improves 4.1% on average than TAN, which is higher than [11] (1.1%) on average but with more time complexity. So, both discriminative structure and discriminative criterion are effective, and our results show that the discriminative structure is even more effective than discriminative criterion.

5. Conclusions and future work

We propose a new hybrid generative-discriminative(HGD) algorithm for structure learning of BN classifiers. The algorithm falls in the scoring+search paradigm, searching Markov blanket in the restricted space with appropriate neighborhood operations of moderate complexity, and using the cross-validation classification rate (CR) as objective for avoiding over-fitting. In experiments on a large suite of benchmark datasets from the UCI Repository, the proposed learning algorithm is demonstrated to outperform the previous methods NB, TAN, K2, and AODE and the discriminative structure is even more effective than discriminative criterion. In our future work, we intend to explore even better heuristic approximations in terms of the tradeoff between the time complexity and the classification performance, refine the search space and neighborhood operations, and elaborate the theoretical explanations for the excellent performance of the approach.

5. Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under grant no.6012130.

References

[6]. Langseth, H., Nielsen, T.D.: Classification using

[7]. Grossman, D., Domingos, P.: Learning bayesian network classifiers by maximizing conditional

discriminative learning of bayesian network classifier
via boosted augmented naive bayes. In: ICML ’05:
Proceedings of the 22nd international conference on
Machine learning, New York, NY, USA, ACM Press
(2005) 369–376

[9]. Greiner, R., Zhou,W.: Structural extension to logistic
regression: discriminative parameter learning of belief
net classifiers. In: Eighteenth national conference on
Artificial intelligence, Menlo Park, CA, USA, American Association for Artificial Intelligence (2002)
167–173

[10]. Pernkopf, F., Bilmes, J.: Discriminative versus
generative parameter and structure learning of
Bayesian network classifiers. In: ICML ’05:
Proceedings of the 22nd international conference on
Machine learning, New York, NY, USA, ACMPress
(2005) 657–664

bayesian classifier. In: In Proceedings of the 19thInter
national FLAIRS Conference. (2006)

[12]. Bishop, C.M.: Pattern Recognition and Machine

Classification. Wiley Interscience (2000)

distributions with dependence trees. IEEE

classifiers: A comparison of distribution based and
classification-based approaches. Proceedings of the
Seventh International Workshop on Artificial
Intelligence and Statistics (1999) 225–230

[16]. Bilmes, J.: Dynamic bayesian multinet. In: UAI ’00:
Proceedings of the 16th Conference on Uncertainty in
Artificial Intelligence, San Francisco, CA, USA,
Morgan Kaufmann Publishers Inc. (2000) 38–45

[17]. Heckerman, D.: A tutorial on learning with Bayesian
networks. Technical report,Microsoft Research,
Redmond, Washington (1996)

[18]. Blake, C., Merz, C.: UCI repository of machine
learning databases, University of California, Irvine,
Dept. of Information and Computer Sciences.
http://www.ics.uci.edu/mllearn/MLRepository.html
(1998)

of continuous-valued attributes for classification

modular machine learning software library. Technical

[21]. Witten, I.H., Frank, E.: Data Mining: Practical machine
learning tools and techniques. 2nd edn. Morgan
Kaufmann, San Francisco (2005)

Morgan Kaufmann Publishers Inc., San Francisco,CA,
USA (1993)

[23]. Cooper, G.F., Herskovits, E.: A bayesian method for the
induction of probabilistic networks from data. Machine
Learning 9(4) (1992) 309–347

bayes: Aggregating one-dependence estimators.

273–324.