MAXIMUM MARGIN STRUCTURE LEARNING OF BAYESIAN NETWORK CLASSIFIERS

Franz Pernkopf, Michael Wohlmayr

Signal Processing and
Speech Communication Laboratory
Graz University of Technology, Graz, Austria

Manfred Mücke

Research Lab Computational
Technologies and Applications
University of Vienna, Wien, Austria

ABSTRACT

Recently, the margin criterion has been successfully used for parameter optimization in graphical models. We introduce maximum margin based *structure* learning for Bayesian network classifiers and demonstrate its advantages in terms of classification performance compared to traditionally used discriminative structure learning methods. In particular, we provide empirical results for generative structure learning and two discriminative structure learning approaches on handwritten digit recognition tasks. We show that maximum margin structure learning outperforms other structure learning methods. Furthermore, we present classification results achieved with different bitwidth for representing the parameters of the classifiers.

Index Terms— Bayesian network classifiers, discriminative learning, margin learning, custom-precision

1. INTRODUCTION

There are two fundamental approaches for learning probabilistic classifiers: generative and discriminative learning [1]. Generative learning optimizes the joint probability distribution of the features and the class labels using maximum likelihood (ML) estimation. The class label is usually predicted using the maximum a-posteriori estimate of the class posteriors obtained by applying Bayes rule. Discriminative learning uses an objective function, e.g. classification rate (CR), conditional likelihood (CL), or margin, that optimize the model for the classification task. Discriminative learning may lead to better classification performance, particularly when the class conditional distributions poorly approximate the true distribution [1]. Unfortunately, discriminative scores are usually not decomposable, while generative scores, e.g. log likelihood, are decomposeable, i.e. they can be written as sum of terms where each term depends on the variable and its conditioning variables (parents).

Learning the graph structure of a Bayesian network classifier is a challenging task. Recently, approaches for finding the optimal generative Bayesian network structure have been proposed. These methods are based on dynamic programming [2], branch-and-bound techniques [3], or search over various variable orderings [4]. More methods and a comprehensive overview can be found in [5] and references therein. Discriminative structure learning[1] is not less difficult because of the non-decomposability of the scores. Discriminative structure learning methods – relevant for learning Bayesian network classifiers – are usually approximate methods based on local search heuristics. In [6], a greedy hill climbing heuristic is used to learn a classifier structure using the CR as score. Particularly, at each iteration one edge is added to the structure which complies with the restrictions of the network topology and the acyclicity constraints of a Bayesian network. In a similar algorithm, the CL has been applied for discriminative structure learning [7]. Recently, we introduced a computationally efficient order-based greedy search heuristic for finding discriminative structures [8]. This algorithm finds structures with similar performance as greedy hill climbing at lower computational costs. This ordering heuristic first establishes an ordering of the N features according to classification based information measures. Given the resulting ordering, the algorithm efficiently discovers high-performing discriminative network structure with $\mathcal{O}\left(N^{k+1}\right)$ score evaluations where k indicates the tree-width[2] of the learned sub-graph over the attributes. Our order-based structure learning is based on the observations in [9] and shows similarities to the K2 heuristic [10]. However, we proposed to use a discriminative scoring metric (i.e. CR) and suggest approaches for establishing the variable ordering based on conditional mutual information [11]. Further generative and discriminative *parameter* learning methods for Bayesian network classifiers are summarized in [8, 12].

In this paper, we apply greedy hill climbing and order-based heuristics for learning discriminative classifier structures. In contrast to previous work, we replace the CR score by the maximum margin (MM) criterion. One of the most successful discriminative classifiers, namely the support vector machine (SVM), finds a decision boundary which maximizes the margin between samples of distinct classes resulting in good generalization properties [13] of the classifier. Recently, the margin criterion has been applied to learn the parameters of probabilistic models. Taskar et al. [14] observed that undirected graphical models can be efficiently trained to maximize the margin. More recently, Guo et al. [15] introduced the maximization of the margin for parameter learning based on convex relaxation to Bayesian networks. We proposed to use a conjugate gradient algorithm for maximum margin optimization of the parameters and show its advantages with respect to computational requirements [12]. Since then, different margin-based training algorithms have been proposed for HMMs in [16, 17] and references therein.

To the best of our knowledge, this is the first work using the margin score for *structure* learning. We empirically evaluate our margin-based discriminative structure learning heuristics on two handwritten digit recognition tasks. We use naive Bayes (NB) as well as generatively and discriminatively optimized tree augmented naive

This work was supported by the Austrian Science Fund (Project number P22488-N23) and (Project number S10604-N13).

[1]Discriminative scoring functions (e.g. classification rate, conditional likelihood, or margin) are used for structure learning.

[2]The tree-width of a graph is defined as the size of the largest clique (i.e. number of variables) of the moralized and triangulated directed graph minus one. Since there can be multiple triangulated graphs, the tree-width is defined by the triangulation where the largest clique has the fewest number of variables. More details are given in [1] and references therein.

Bayes (TAN) [18] structures. Maximum margin structure learning outperforms recent generative and discriminative structure learning results [8]. Additionally, we perform classification experiments with variable bitwidth of the classifier parameters. While learning is performed in double-precision floating-point arithmetic guaranteeing both wide range and high accuracy, we are interested in investigating if the same effort is required for classification. Our experiments show that the presented classifier is extremely robust against truncation and therefore allows for use of a fixed-point number format while achieving a classification performance comparable to classification using double-precision floating-point arithmetic. This finding can be used to achieve a high-performance implementation of our classifier in custom or reconfigurable hardware or on fixed-point DSPs.

The paper is organized as follows: In Section 2, we introduce our notation, ML parameter learning as well as NB and TAN structures. Section 3 introduces different structure learning heuristics. In Section 4, we introduce the MM criterion for discriminative structure learning. In Section 5, we present experimental results on handwritten digit recognition. Section 6 concludes the paper and gives a perspective on future work.

2. BAYESIAN NETWORK CLASSIFIERS

A Bayesian network [19] $\mathcal{B} = \langle \mathcal{G}, \Theta \rangle$ is a directed acyclic graph $\mathcal{G} = (\mathbf{Z}, \mathbf{E})$ consisting of a set of nodes $\mathbf{Z}$ and a set of directed edges $\mathbf{E}$ connecting the nodes. This graph represents factorization properties of the distribution of a set of random variables $\mathbf{Z} = \{Z_1, \ldots, Z_{N+1}\}$. The variables in $\mathbf{Z}$ have values denoted by lower case letters $\mathbf{z} = \{z_1, z_2, \ldots, z_{N+1}\}$. We use boldface capital letters, e.g. $\mathbf{Z}$, to denote a set of random variables and correspondingly boldface lower case letters denote a set of instantiations (values). Without loss of generality, in Bayesian network classifiers the random variable Z_1 represents the class variable $C \in \{1, \ldots, |C|\}$, where $|C|$ represents the number of classes and $\mathbf{X}_{1:N} = \{X_1, \ldots, X_N\} = \{Z_2, \ldots, Z_{N+1}\}$ denotes the set of random variables representing the N attributes of the classifier. In a Bayesian network each node is independent of its non-descendants given its parents. The set of parameters which quantify the network are represented by Θ. Each random variable Z_j is represented as a local conditional probability distribution given its parents Z_{Π_j}, i.e. $P_\Theta\left(Z_j | Z_{\Pi_j}\right)$. The training data consists of M independent and identically distributed samples $\mathcal{S} = \{\mathbf{z}^m\}_{m=1}^M = \{(c^m, \mathbf{x}_{1:N}^m)\}_{m=1}^M$ where $M = |\mathcal{S}|$. The joint probability distribution of a sample $\mathbf{z}^m$ is determined as

$$P_\Theta\left(\mathbf{Z} = \mathbf{z}^m\right) = \prod_{j=1}^{N+1} P_\Theta\left(Z_j = z_j^m | Z_{\Pi_j} = z_{\Pi_j}^m\right).$$

In this work, we restrict ourselves to NB and TAN structures. The NB network assumes that all the attributes are conditionally independent given the class label. As reported in [18], the performance of the NB classifier is surprisingly good even if the conditional independence assumption between attributes is unrealistic or even wrong for most of the data. Friedman et al. [18] introduced the TAN classifier which is based on structural augmentations of the NB network. In order to relax the conditional independence properties of NB, each attribute may have at most one other attribute as an additional parent. This means that the tree-width of the attribute induced sub-graph is unity, i.e. we have to learn a 1-tree over the attributes. A TAN classifier example is shown in Figure 1. In [8], we noticed that 2-trees over the features often do not improve classification performance significantly without regularization. Therefore, we limit the experiments to NB and TAN structures.

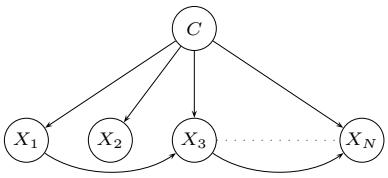

Fig. 1. An example of a TAN classifier structure.

3. STRUCTURE LEARNING HEURISTICS

This section provides three structure learning heuristics. Note that the parameters during structure learning are optimized generatively using maximum likelihood estimation [19].

3.1. Generative Structure Learning

The conditional mutual information (CMI) [11] between the attributes given the class variable is computed as:

$$I\left(X_i; X_j | C\right) = E_{P(X_i, X_j, C)} \log \frac{P\left(X_i, X_j | C\right)}{P\left(X_i | C\right) P\left(X_j | C\right)}.$$

This measures the information between X_i and X_j in the context of C. Friedman et al. [18] gives an algorithm for constructing a TAN network using this measure. First, the pairwise CMI $I\left(X_i; X_j | C\right) \quad \forall \quad 1 \leq i \leq N$ and $i < j \leq N$ is computed. Then, an undirected 1-tree is built using the maximal weighted spanning tree algorithm [19] where each edge connecting X_i and X_j is weighted by $I\left(X_i; X_j | C\right)$. The undirected 1-tree is transformed to a directed tree. Therefore, a root variable is selected and all edges are directed away from this root. Finally, the class node C and the edges from C to all attributes $X_1, \ldots, X_N$ are added.

3.2. Greedy Discriminative Structure Learning

A Bayesian network is initialized to NB and at each iteration we add the edge that, while maintaining a partial 1-tree, gives the largest improvement of the scoring function. Basically, two scoring functions have been considered: the CR [6] and the CL [7]. Structure learning is terminated if there is no edge which further improves the score. Thus, if we might obtain a partial 1-tree (forest) over the attributes. This approach is computationally expensive since each time an edge is added, the scores for all $\mathcal{O}\left(N^2\right)$ edges need to be re-evaluated due to the discriminative non-decomposable scoring functions we employ. Overall, for learning a 1-tree structure, $\mathcal{O}\left(N^3\right)$ score evaluations are necessary. In our experiments, we consider either the CR or the margin (defined in the next section) as scoring objective. Both are discriminative learning criteria. The CR or margin computation can be accelerated by techniques presented in [8].

3.3. Order-based Discriminative Structure Learning

In [8] an order-based greedy algorithm (OMI) has been introduced which is able to find a discriminative TAN structure with only $\mathcal{O}\left(N^2\right)$ score evaluations compared to the greedy algorithm above ($\mathcal{O}\left(N^3\right)$). The classification results of the order-based greedy algorithm are not statistical significantly different compared to the greedy algorithm. The order-based algorithm consists of 2 steps:

Step 1: Establish an ordering: First a total ordering $\prec$ of the variables $\mathbf{X}_{1:N}$ according to the CMI is established. Therefore, the feature that is most informative about C is selected first. The next node in the order is the node that is most informative about C conditioned on the first node. More specifically, this algorithm forms an ordered sequence of nodes $\mathbf{X}_{\prec}^{1:N} = \left\{X_{\prec}^{1}, X_{\prec}^{2}, \ldots, X_{\prec}^{N}\right\}$ according to

$$X_{\prec}^{j} \leftarrow \arg \max_{X \in \mathbf{X}_{1:N} \setminus \mathbf{X}_{\prec}^{1:j-1}} \left[I\left(C; X | \mathbf{X}_{\prec}^{1:j-1}\right)\right],$$

where $j \in \{1, \ldots, N\}$.

Step 2: Selecting parents with respect to a given order to form a 1-tree: Once the variables are ordered $\mathbf{X}_{\prec}^{1:N}$, the parent $X_{\Pi_j} \in \mathbf{X}_{\Pi_j} = \mathbf{X}_{\prec}^{1:j-1}$ for each $X_{\prec}^{j}$ ($j \in \{3, \ldots, N\}$) is selected. In case of a small size of $\mathbf{X}_{\Pi_j}$ (i.e. N) and of 1-trees a computational costly scoring function to find X_{Π_j} can be used. Again, we suggest to use both, either the CR or the margin, for learning a discriminative structure. We connect a parent to $X_{\prec}^{j}$ only when the scoring objective is improved, and otherwise leave $X_{\prec}^{j}$ parentless (except C). This might result in a partial 1-tree (forest) over the attributes.

4. DISCRIMINATIVE MAXIMUM MARGIN (MM) SCORE

The multi-class margin [15] of sample m can be expressed as

$$d_{\Theta}^{m} = \min_{c \neq c^m} \frac{P_{\Theta}\left(c^m | \mathbf{x}_{1:N}^{m}\right)}{P_{\Theta}\left(c | \mathbf{x}_{1:N}^{m}\right)} = \frac{P_{\Theta}\left(c^m, \mathbf{x}_{1:N}^{m}\right)}{\max_{c \neq c^m} P_{\Theta}\left(c, \mathbf{x}_{1:N}^{m}\right)}.$$

For the sake of brevity, we only notate instantiations of the random variables. If $d_{\Theta}^{m} > 1$, then sample m is correctly classified and vice versa. The magnitude of d_{Θ}^{m} is related to the confidence of the classifier about its decision. Taking the logarithm, we obtain

$$\log d_{\Theta}^{m} = \log P_{\Theta}\left(c^m, \mathbf{x}_{1:N}^{m}\right) - \max_{c \neq c^m} \left(\log P_{\Theta}\left(c, \mathbf{x}_{1:N}^{m}\right)\right).$$

Usually, the maximum margin approach maximizes the margin of the sample with the smallest margin for a separable classification problem [20], i.e. the objective function is written as $M\left(\mathcal{B}|\mathcal{S}\right) = \min_{m=1,\ldots,M} \log d_{\Theta}^{m}$. For the non-separable problem, we aim to relax this by introducing a soft margin, i.e. we focus on samples with $\log d_{\Theta}^{m}$ close to zero. For this purpose, we consider the *hinge* loss function

$$M\left(\mathcal{B}|\mathcal{S}\right) = \sum_{m=1}^{M} \min\left[1, \lambda \log d_{\Theta}^{m}\right],$$

where the scaling parameter $\lambda > 0$ controls the margin with respect to the loss function and is set by cross-validation. Maximizing this function with respect to the parameters Θ implicitly increases the log-margin, whereas the emphasis is on samples with $\lambda \log d_{\Theta}^{m} < 1$, i.e. samples with a large positive margin are considered as constant factor during on the optimization. We use $M\left(\mathcal{B}|\mathcal{S}\right)$ as score for discriminative structure learning, i.e. the CR criterion in the discriminative structure learning heuristics (see Section 3.3 and 3.2) is replaced by $M\left(\mathcal{B}|\mathcal{S}\right)$.

5. EXPERIMENTS

We present results for handwritten digit recognition. In the following, we provide details about the MNIST and the USPS data sets:
MNIST Data: The MNIST data [21] contains 60000 samples for training and 10000 digits for testing. We down-sample the gray-level images by a factor of two which results in a resolution of 14×14 pixels, i.e. 196 features.
USPS Data: This data set contains 11000 uniformly distributed handwritten digit images from zip codes of mail envelopes. The data set is split into 8000 images for training and 3000 for testing. Each digit is represented as a 16×16 grayscale image, where each pixel is considered as feature.

For structure learning we use the algorithms introduced in Section 3. In particular, we apply the following approaches for learning TAN structures:

- TAN-CMI: Generative TAN structure learning using conditional mutual information [18].
- TAN-CR: Discriminative TAN structure learning maximizing the CR using the naive greedy heuristic [6, 8].
- TAN-MM: Discriminative TAN structure learning maximizing the margin using the naive greedy heuristic (this paper).
- TAN-OMI-CR: Discriminative TAN structure learning maximizing the CR using the order-based heuristic [8].
- TAN-OMI-MM: Discriminative TAN structure learning maximizing the margin using the order-based heuristic (this paper).

The CR and MM scores are determined by 5-fold cross-validation on the training data. Zero probabilities in the conditional probability tables are replaced with small values ($\varepsilon = 10^{-5}$). Furthermore, we used the same data set partitioning for various learning algorithms.

Classifier Structure	MNIST	USPS
NB	83.73±0.37	87.10±0.61
TAN-CMI	91.28±0.28	91.90 ±0.50
TAN-OMI-CR	92.28±0.27	92.40±0.48
TAN-OMI-MM	**92.71**±0.26	**95.47**±0.37
TAN-CR	**92.63**±0.26	92.70±0.47
TAN-MM	**93.15**±0.25	**95.40**±0.37

Table 1. Classification results in [%] for MNIST and USPS data with standard deviation. Best structure learning results are emphasized using bold font.

Table 1 shows the classification rates for MNIST and USPS for various learning methods. The classification rate is improving for more complex structures using ML parameter learning. Discriminatively optimized structures, i.e. TAN-OMI-CR, TAN-OMI-MM, TAN-CR, and TAN-MM significantly outperform generatively learned, i.e. TAN-CMI, and NB structures. Furthermore, margin optimized discriminatively learned structures, i.e. TAN-OMI-MM and TAN-MM, are significantly better compared to TAN-OMI-CR and TAN-CR, respectively. As already reported in [8], the greedy hill climbing heuristic is only marginally better than the ordering-based (OMI) heuristic. The difference in classification performance is insignificant. However, the OMI heuristic uses only $\mathcal{O}\left(N^2\right)$ score evaluations compared to the greedy algorithm ($\mathcal{O}\left(N^3\right)$) for learning TAN structures.

Furthermore, we consider the implementation in hardware, where custom-precision arithmetic can be implemented. Therefore,

knowledge about the optimal bitwidth of the classifier parameters while maintaining the classification performance is essential. Any unsigned fixed-point variable of bitwidth $w = 64$ covers a range of $[0 \dots 2^{w-b} - 1]$, where b is the position of the binary point. Given the range of our parameters in the logarithmic domain, we can approximate double-precision floating-point calculations by a fixed-point number format with $b = 12$ integer and $f = 52$ fractional bits, totalling 12+52=64 bits ($Q12.52$). In our classification experiments (see Figure 2), we reduce the total width of the format by truncating the n least significant bits where n is varied between zero ($Q12.52$) and 63 ($Q1.0$). Results on both data sets show that compared to double-precision floating point (see Table 1) a bitwidth of 14 for fixed point variables (i.e. 2 fractional bits) is sufficient to obtain good classification performance. This results in reduced storage and computational requirements when implemented in custom or reconfigurable hardware.

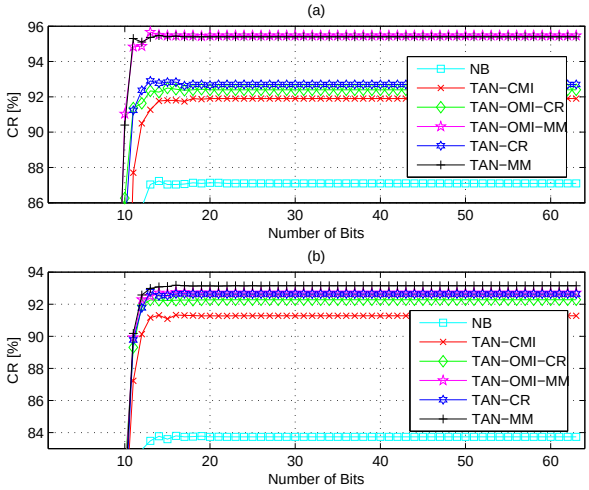

Fig. 2. Classification performance versus number of bits for representing the parameters of the classifiers: (a) USPS data, (b) MNIST data.

6. CONCLUSIONS

We use the maximum margin score for learning discriminative classifier structures. As search heuristic we apply greedy hill climbing and an order-based heuristic. We empirically evaluate our margin-based discriminative structure learning heuristics on the MNIST and USPS handwritten digit recognition tasks. We use naive Bayes as well as generatively and discriminatively optimized tree augmented naive Bayes structures. Maximum margin structure learning outperforms generative and discriminative structure learning results. Additionally, the ordering heuristic performs similar compared to the greedy hill climbing approach at lower computational costs. Custom-precision experiments show that a bitwidth of 14 is sufficient for good classification results. Future work includes the evaluation of our maximum margin structure learning algorithm on further data sets. Furthermore, discriminative parameter learning approaches [8, 12] are investigated for margin-optimized structures.

7. REFERENCES

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] P. Parviainen and M. Koivisto, "Exact structure discovery in Bayesian networks with less space," in *Uncertainty in Artivicial Intelligence (UAI)*, 2009, pp. 436–443.

[3] C.P. de Campos, Z. Zeng, and Q. Ji, "Structure learning of bayesian networks using constraints," in *International Conference on Machine Learning (ICML)*, 2009, pp. 113–120.

[4] M. Teyssier and D. Koller, "Ordering-based search: A simple and effective algorithm for learning Bayesian networks," in *21th Conf. on Uncertainty in AI (UAI)*, 2005, pp. 584 – 590.

[5] T. Jaakkola, D. Sontag, A. Globerson, and M. Meila, "Learning Bayesian network structure using LP relaxations," in *Intern. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 358–365.

[6] E.J. Keogh and M.J. Pazzani, "Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches," in *7th Intern. Workshop on Artificial Intelligence and Statistics*, 1999, pp. 225–230.

[7] D. Grossman and P. Domingos, "Learning Bayesian network classifiers by maximizing conditional likelihood," in *21st Inter. Conf. of Machine Lerning (ICML)*, 2004, pp. 361–368.

[8] F. Pernkopf and J. Bilmes, "Efficient heuristics for discriminative structure learning of Bayesian network classifiers," *Journal of Machine Learning Res.*, vol. 11, pp. 2323–2360, 2010.

[9] W.L. Buntine, "Theory refinement on Bayesian networks," in *7th Conference on Uncertainty in AI (UAI)*, 1991, pp. 52–60.

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

[11] T. Cover and J. Thomas, *Elements of information theory*, John Wiley & Sons, 1991.

[12] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek, "Maximum margin Bayesian network classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence, submitted*, 2011.

[13] V. Vapnik, *Statistical Learning Theory*, Wiley & Sons, 1998.

[14] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[15] Y. Guo, D. Wilkinson, and D. Schuurmans, "Maximum margin Bayesian networks," in *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[16] F. Sha and L. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, pp. 313–316.

[17] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *Intern. Conf. on Machine learning (ICML)*, 2008, pp. 384–391.

[18] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[19] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 1988.

[20] B. Schölkopf and A.J. Smola, *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2001.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.