Learning Dynamic Bayesian Network Discriminatively for Human Activity Recognition

Xiaoyang Wang and Qiang Ji
Dept. of ECSE, Rensselaer Polytechnic Institute, USA
$\{$*wangx16, jiq*$\}$*@rpi.edu*

Abstract

The purpose of this paper is to develop an approach to learn dynamic Bayesian network (DBN) discriminatively for human activity recognition. DBN is a generative model widely used for modeling temporal events in human activity recognition. The parameters of the DBN models are usually learned through maximizing likelihood or expected likelihood. However, activity is often recognized through identifying the activity class with the highest posterior probability. Hence, there is discrepancy between the learning and classification criteria. In this paper, we focus on developing a discriminative parameter learning approach for hybrid DBNs that has a consistent criterion during training and testing. Our approach is applicable to parameter learning with both complete data and incomplete data, and empirical studies show the proposed discriminative learning approach outperforms the maximum likelihood or EM algorithm in activity recognition tasks.

1 Introduction

The dynamic Bayesian networks (DBNs) [16, 7] received increasing attention in human action and activity recognition during recent years. DBNs gain their advantage through explicitly modeling both the spatial and temporal dependencies among different entities in human activity. In the training process, most of these approaches learn one representative DBN model for each activity through maximum likelihood (ML) estimation, or expectation maximization (EM) if with incomplete data. During testing stage, the activity is recognized through picking the model with the highest likelihood given the image observations, or equivalently, with the highest posterior probability given prior on activities. In this manner, we can see a clear discrepancy between the learning objective and the classification criterion. The parameters learned by ML or EM algorithm, though capturing the data dependency well, may not maximize the classification accuracy.

The solution to reduce the discrepancy between the training and testing objective is to learn the generative DBN models discriminatively. The learning procedure maximizes the conditional likelihood instead of the joint likelihood and ensure a consistent criterion during learning and testing. Previous researches have shown that, for classification tasks, discriminative learning often works better than generative learning even for a generative model. A. Ng and M. Jordan [10] have provided a theoretical and empirical comparison of generative learning and discriminative learning for Bayesian network classifiers.

One difficulty associated with discriminative learning is that, the conditional likelihood function, unlike the general likelihood function, is not decomposable over the structure. Thus, no analytical solution is available to determine the parameters. Researchers [3, 2, 12, 15, 14, 4] tried to solve this problem with numeric optimization. These discriminative learning approaches mainly focus on discrete and static Bayesian networks. However, in human activity recognition, we often involve continuous attributes and dynamic Bayesian networks, so these approaches are not directly applicable. In recent literature, on one hand, although the inference and generative learning of the hybrid DBN [8, 9] have been widely investigated, there are no discriminative approaches, to the best of our knowledge, proposed in the literature for learning the parameters of hybrid DBNs with both discrete and continuous variables. On the other hand, though there are approaches [5, 11] proposed for discriminatively learning HMMs (which can be viewed as a simple case of the DBNs) in speech recognition domain, its generalization to more complex DBN has not been investigated before. In this paper, considering the general requirements for modeling and recognizing activities, we propose a discriminative parameter learning approach for general hybrid DBN under a gradient-based framework.

In summary, we focus on developing algorithm to learn hybrid DBN models discriminatively for human activity recognition. In the learning process, through

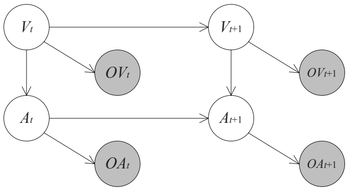

Figure 1. Activity DBN model

maximizing the conditional likelihood, we can reduce the discrepancy between the learning and testing criteria. Under a gradient-search framework, our approach can handle DBN model with hybrid variables, and work well for the case of incomplete data.

2 Activity Modeling with DBNs

We introduce a DBN model for human activity recognition. The model consists of two levels: the feature level and the state level. While the feature level encodes the observations from the images, the state level abstracts the basic states of the activity.

The features used for activity recognition include the kinematic and appearance features. The kinematic features OV are evaluated based on the estimated bounding box position and moving velocity. The appearance features OA are selected from a feature pool consisting of the HOG (histogram of oriented gradient) and HOF (histogram of optical flow) feature.

The DBN structure of our activity model is shown in Fig. 1. The clear nodes represent two physical states: motion state (V), and appearance state (A). The shaded nodes denote observations. Besides the nodes, there are two types of links in the model: intra-slice links and inter-slice links. The intra-slice links from V to A denotes the possible causality between the subject's motion and appearance. The inter-slice links show the temporal evolution and capture the dynamic relationships between states at different times.

With this model, we construct one DBN model for each activity and perform activity recognition by finding the model with the highest likelihood, or finding the model with the highest posterior probability given prior on activity classes. Both of them can be evaluated by the forward propagation of dynamic junction tree.

3 DBN Parameter Learning

In this section, we introduce a discriminative learning approach to learn the DBN models for all activity models together, which can reduce the discrepancy between the training and testing objective.

3.1 Discriminative Learning Formulation

The goal of classification is to predict the class label c given the evidence E. Under the Bayesian decision framework, the optimal prediction for data E is the class that maximizes $P(c|E)$. In activity recognition, as we can evaluate the likelihood $P(E|c)$ for each activity c, we can compute the posterior probability based on Bayesian theorem.

$$P(c|E) = \frac{P(E|c)P(c)}{\sum_{c'} P(E|c')P(c')} \tag{1}$$

As DBN usually captures the joint distribution of sequence of variables, it is typically learned by maximizing the log likelihood of all training sequences:

$$\hat{\Theta} = \arg\max_{\Theta} P(\mathcal{E}|\Theta) = \arg\max_{\Theta} \sum_{n} \log P(E^n|\Theta)$$

Here $\mathcal{E} = (E^1, \ldots, E^N)$ denotes N training sequences.

For activity recognition, with generative learning, we learn the parameters $\hat{\Theta}_c$ of each activity model c independently through maximize its likelihood:

$$\hat{\Theta}_c = \arg\max_{\Theta} P(\mathcal{E}_c|\Theta) = \arg\max_{\Theta} \sum_{E\in\mathcal{E}_c} \log P(E|\Theta)$$

where $\mathcal{E}_c$ denotes the training sequences for activity c.

In this way, we can ensure to obtain a representative model for each activity, but it can not guarantee the best performance in classification, since the objective function of the maximum likelihood learning is not consistent with our prediction criterion $P(c|E)$. Hence, a better objective function for learning DBNs for activity recognition would be the conditional log likelihood:

$$CLL(C|\mathcal{E}) = \sum_{n=1}^{N} \log P(c^n|E^n) = \sum_{c} \sum_{E\in\mathcal{E}_c} \log P(c|E)$$

where $C = (c^1, c^2, \ldots, c^N)$ denote the activity labels of all N training sequences.

Maximizing the conditional likelihood is not trivial since the CLL objective is non-convex in general. However, we can optimize it locally through gradient search. The limited memory BFGS with Armijo line search [1] is employed to perform the optimization.

A key step for the optimization is to evaluate the gradient of the CLL. In general, for sample (E, c), the gradient of CLL with respect to model parameter Θ is

$$\begin{aligned}\frac{\partial \log P(c|E)}{\partial\Theta} &= \frac{\partial \log[P(E|c)P(c)]}{\partial\Theta} - \frac{\partial \log P(E)}{\partial\Theta} \\ &= \frac{\partial \log P(E|c)}{\partial\Theta} - \frac{\partial \log P(E)}{\partial\Theta}\end{aligned} \tag{2}$$

Please note that the second term of Eqn. 2 can be evaluated as the expectation of the first term, as in Eqn. 3.

$$\frac{\partial \log P(E)}{\partial\Theta} = \sum_{c} P(c|E) \frac{\partial \log P(E|c)}{\partial\Theta} \tag{3}$$

Now we consider this gradient with respect to the parameter $\Theta_{c'}$ of a specific activity model c'. With Eqn. 3 and the fact

$$\frac{\partial \log P(E|c)}{\partial \Theta_{c'}} = 0 \quad \text{if } c' \neq c$$

we can get

$$\frac{\partial \log P(c|E)}{\partial \Theta_{c'}} = \begin{cases} (1 - P(c|E))\frac{\partial \log P(E|c)}{\partial \Theta_c} & \text{if } c' = c \\ -P(c'|E)\frac{\partial \log P(E|c')}{\partial \Theta_{c'}} & \text{if } c' \neq c \end{cases}$$

As $P(c|E)$ can be evaluated with Eqn. 1, we mainly focus on computing $\partial \log P(E|c)/\partial \Theta_c$ when evaluating the derivative of the CLL. Please note that $\partial \log P(E|c)/\partial \Theta_c$ is just the derivative of the log likelihood of DBN model c with respect to Θ_c.

3.2 Discriminative Learning for Hybrid DBN

For learning the parameter of hybrid DBN with discrete parent and continuous Gaussian child nodes (DP-CC), we have the following parameters.

- Discrete parents - continuous child (DP-CC)
  The local conditional probability distribution is parameterized with conditional Gaussian $p(x_{t,i}|\pi_{t,i} = j) \sim N(\mu_{ij}, \Sigma_{ij})$, so we focus on learning the parameters (μ_{ij}, Σ_{ij}). To ensure the covariance matrix be positive semidefinite, we reparameterize Σ_{ij} as $A_{ij} = \Sigma_{ij}^{-\frac{1}{2}}$.

In the case of complete data, computing the derivative $\partial \log P(E|c)/\partial \Theta_c$ is relatively easy since $P(E|c)$ is completely decomposable based on the DBN structure. We compute this gradient as [1].

$$\frac{\partial \log P(E|c)}{\partial \mu_{ij}} = \sum_t \left[A_{ij}^2 (x_{t,i} - \mu_{ij})\right]$$

$$\frac{\partial \log P(E|c)}{\partial A_{ij}} = \sum_t \left[A_{ij}^{-1} - A_{ij}(x_{t,i} - \mu_{ij})(x_{t,i} - \mu_{ij})^T\right]$$

3.3 Incomplete Data

When the training data is incomplete, $P(E|c)$ is not decomposable, so evaluating the derivative of CLL becomes difficult. One natural choice is the EM algorithm, with the objective function substituted by the CLL. However, in this case, in M step of EM, there is no analytical solution for estimating the parameter and we still need to go through the optimization procedure for the "completed" case. To avoid this double-looped optimization procedure, an efficient way is needed to directly compute the gradient of CLL with incomplete data. We resort this to the existing exact inference algorithms in the hybrid model. In the following parts, we show the derivative $\partial \log P(E|c)/\partial \theta$ for the DP-CC case different from the solved discrete parent and discrete child nodes (DP-DC) case [2].

In DP-CC case, the derivatives $\partial \log P(E|c)/\partial \mu_{ij}$ and $\partial \log P(E|c)/\partial A_{ij}$ can be computed as follows

$$\frac{\partial \log P(E|c)}{\partial \mu_{ij}} = -\sum_t P(\pi_{t,i} = j|E) A_{ij}^2 \mu_{ij} + A_{ij}^2 \sum_t E_{p(x_{t,i}, \pi_{t,i}=j|E)}\{x_{t,i}\}$$

$$\frac{\partial \log P(E|c)}{\partial A_{ij}} = -\sum_t P(\pi_{t,i} = j|E) A_{ij}^{-1} - A_{ij} \sum_t E_{p(x_{t,i}, \pi_{t,i}=j|E)}\{(x_{t,i} - \mu_{ij})(x_{t,i} - \mu_{ij})^T\}$$

here

$$E_{p(x_{t,i}, \pi_{t,i}=j|E)}\{x_{t,i}\} = P(\pi_{t,i} = j|E) E_{p(x_{t,i}|\pi_{t,i}=j,E)}\{x_{t,i}\}$$

$$E_{p(x_{t,i}, \pi_{t,i}=j|E)}\{(x_{t,i} - \mu_{ij})(x_{t,i} - \mu_{ij})^T\} = P(\pi_{t,i} = j|E) E_{p(x_{t,i}|\pi_{t,i}=j,E)}\{(x_{t,i} - \mu_{ij})(x_{t,i} - \mu_{ij})^T\}$$

where

$$E_{p(x_{t,i}|\pi_{t,i}=j,E)}\{(x_{t,i} - \mu_{ij})(x_{t,i} - \mu_{ij})^T\} = cov[x_{t,i}] + (E[x_{t,i}] - \mu_{ij})(E[x_{t,i}] - \mu_{ij})^T$$

Since $P(\pi_{t,i} = j|E), E[x_{t,i}], cov[x_{t,i}]$ [2] can be obtained through the inference in the hybrid dynamic Bayesian network [9], we can compute the derivative $\partial \log P(E|c)/\partial \mu_{ij}$ and $\partial \log P(E|c)/\partial \Sigma_{ij}$. Further, based on discussions in section 3.1, we can obtain the gradient of the CLL with respect to μ_{ij} and Σ_{ij}.

4 Experiments

We apply the discriminative learning algorithm on the KTH human activity dataset [13]. KTH dataset has 6 human activities: walking, jogging, running, boxing, hand waving and hand clapping. Each activity is performed by 25 subjects (performers) in four different environments. For each subject, there are 24 videos in total with 3 to 4 activity sequences in each video.

4.1 Discriminative vs. Generative Learning

We first focus on comparing the generative learning approach with the discriminative learning approach with different training size. The generative learning approach we used is the EM algorithm. For discriminative learning, as our approach can only guarantee a local

[1] A_{ij}^{-1} denotes the pseudo inverse of A_{ij}

[2] For simplicity, we denote $E_{p(x_{t,i},\pi_{t,i}=j|E)}[x_{t,i}]$ as $E[x_{t,i}]$, $E_{p(x_{t,i},\pi_{t,i}=j|E)}\{(x_{t,i} - E[x_{t,i}])(x_{t,i} - E[x_{t,i}])^T\}$ as $cov[x_{t,i}]$

optimum of the conditional log likelihood, one critical issue is the initialization of the model parameters. In all our experiments, we use the result of the generative learning as the initialization for discriminative learning.

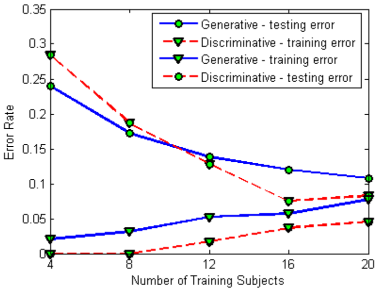

Figure 2. Discriminative learning vs. generative learning on KTH dataset

We first compare the training error of these two approaches based on the results in Fig. 2. Since adding more training subjects will introduce more variations to training data, both approaches have slightly higher training errors when subject number is larger. But it is obvious the discriminatively learned DBN performs consistently better than generatively learned DBN.

We also compare the testing error of the discriminatively learned model with generatively learned model. When the number of training sequences is large, the discriminatively learned models perform obviously better than the generatively learned models. More specifically, given sequences of 16 and 20 subjects for training, the error rates of discriminative learning are 4.5% and 2.5% lower than the generatively learning respectively. However, when the number of training subjects become smaller, discriminative learning suffers more from overfitting than generative learning. We can see that the classification error of the discriminatively learned model is 4.4% higher than generatively learned model given the sequences from 4 subjects for training.

4.2 Comparison with Other Approaches

We compare our approach with the state-of-art approaches on KTH dataset. As in the works of Yuan et al. [17] and Laptev et al. [6], we use the same training, validation and testing split of data with models trained on sequences of 16 subjects. We compare our results in Table 1. While the performance of our generatively learned DBN is about 4% worse than the state-of-art approaches, the discriminative learning DBN can achieve comparable results to the state-of-art approaches.

Table 1. Comparison with previous work.

	Recognition rate
Our method - Generative	88.0%
Our method - Discriminative	92.5%
Yuan et al. [17]	93.3%
Laptev et al. [6]	91.8%

5 Conclusion

In this paper, we propose a discriminative parameter learning method for hybrid dynamic Bayesian network in human activity recognition. Compared to the generative learning approaches, our approach has a more consistent objective in the training stage with the classification criterion, which can guarantee a better classification performance on the training set. Moreover, based on our experiments on the real data from KTH activity dataset, we demonstrate the advantage of discriminative learning over generative learning.

References

[1] M. Avriel. *Nonlinear Programming: Analysis and Methods*. Dover Publishing, 2003.

[2] R. Greiner, X. Su, S. Shen, and W. Zhou. Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Machine Learning*, 59:297–322, 2005.

[3] R. Greiner and W. Zhou. Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *AAAI*, pages 167–173, 2002.

[4] Y. Jing, V. Pavlovic, and J. M. Rehg. Efficient discriminative learning of Bayesian network classifier via Boosted Augmented Naive Bayes. In *ICML*, 2005.

[5] B. Juang, W. Chou, and C. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech and Ausio Processing*, 1997.

[6] I. Laptev, M. Marszalek, and C. Schmid. Learning realistic human actions from movies. *CVPR*, 2008.

[7] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. *CVPR*, 2007.

[8] S. Monti and G. Cooper. Learning hybrid bayesian networks from data. *Learning in graphical models*, 1999.

[9] K. Murphy. Inference and learning in hybrid bayesian networks. *Technical report, UC Berkeley*, 1998.

[10] A. Ng and M. Jordan. On discriminative versus generative classifiers: a comparison of logistic regression and naive bayes. *NIPS*, 2002.

[11] Y. Normandin. Hidden markov models, maximum mutual information estimation, and the speech recognition problem. *PhD. dissertation, McGill University*, 1991.

[12] T. Roos, H. Wettig, P. Grunwald, and P. Myllymaki. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, 2005.

[13] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[14] J. Su, H. Zhang, C. Ling, and S. Matwin. Discriminative parameter learning for bayesian networks. *ICML*, 2008.

[15] H. Wettig, P. Grunwald, T. Roos, P. Myllymaki, and H. Tirri. When discriminative learning of bayesian network parameters is easy. *IJCAI*, 2003.

[16] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Regh. A scalable approach to activity recognition based on object use. *ICCV*, 2007.

[17] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. *CVPR*, 2009.