THE THEORY OF BAYESIAN NETWORKS AND ITS APPLICATION IN SPEECH RECOGNITION

Zhiling Long

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, MS 39762
long@isip.msstate.edu

ABSTRACT

Bayesian networks theory is under significant development in recent years. It has been found very powerful in solving various data analysis problems in areas such as expert systems, decision support systems, and pattern recognition. A Bayesian Network (BN), or Belief Network, is a directed graphical model. It efficiently encodes probabilistic relationships among a set of random variables. In this paper, we give an introductory review of the fundamental concepts and methodology underlying the Bayesian networks theory. We also demonstrate the application of Bayesian nets to Automatic Speech Recognition (ASR). Experiment results from state-of-the-art research work indicate that BNs are promising techniques for speech recognition.

1. INTRODUCTION

A Bayesian network is a graphical model that represents probabilistic relationships among a set of random variables. It is a compact and computationally efficient representation of probability distributions. Over the last decade, it has become a popular method for encoding uncertainty in artificial intelligence. Now it is playing a crucial role in modern expert systems, decision support systems, etc. [1]. More and more researchers in such related fields as pattern recognition are also starting to realize the power of this technique because of the outstanding effectivity it has been demonstrating in data analysis problems. In recent years, there has been significant progress in algorithms for learning Bayesian networks directly from data. The technology is still under fast development.

Bayesian networks can easily deal with incomplete data sets; they are good at learning causal relationships between the random variables under study; they help to integrate prior knowledge with data, if used in combination with Bayesian statistical techniques; and lastly, they offer an efficient approach for avoiding the overfitting of data. These advantages distinguish Bayesian networks from other data analysis methods such as rule bases, artificial neural networks, and decision trees [4].

In this paper we will review the basic concepts behind Bayesian networks. We will then describe the fundamental algorithms for learning structures and parameters of these networks from data. Finally, we will discuss how to apply Bayesian networks theory to speech recognition. We will mainly describe the pioneering research by Zweig at University of California at Berkeley, and present some of his speech recognition results with Bayesian nets.

2. CONCEPTS AND METHODOLOGY

2.1. Representation

Bayesian Networks (BNs), or Belief Networks, are directed graphical models, in which nodes represent random variables and directed arcs represent conditional independence assumptions. To determine a Bayesian network, we also need to specify Conditional Probability Distributions (CPDs) at each node. If the random variables are discrete, the probability distributions turn into Conditional Probability Tables (CPTs), which list the probabilities that a child node takes on each of its possible values given various combinations of values of its parents. For a node without any parents, the associated table gives the prior probabilities instead of the conditional ones.

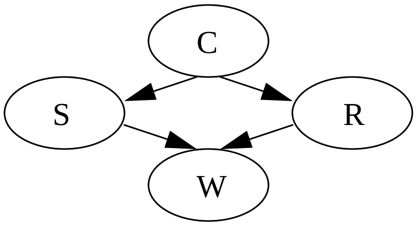

Figure 1: The structure of a simple Bayesian network.

The conditional independence relationships in BNs allow us to represent the joint probability more compactly. The simplest conditional independence relationship encoded in a Bayesian network can be stated as follows: if the parents of a node are given, the node is independent of its ancestors. As an example, in the network shown in Figure 1, we have

$$P(W|C,S,R) = P(W|S,R) \tag{1}$$

because W is independent of C given its parents S and R.

Nodes in Bayesian Nets could also be of continuous values. In this case, we have conditional probabilistic distributions at each node. The most common distribution for such continuous random variables is the Gaussian distribution. Even more complicated, there could be both continuous and discrete nodes existing in a Bayesian net.

A stochastic process, which is a vector of random variables evolving over time, can also be modelled by a Bayesian net. These temporal models are called Dynamic Bayesian Networks (DBNs). DBNs allow the states of the system to be represented as a set of separate variables.

2.2. Inference

The most common task we wish to accomplish using Bayesian networks is probabilistic inference. Bayesian nets can be used for both diagnostic reasoning (from effects to causes) and causal reasoning (from causes to effects). We use Bayes' rule for inference.

$$P(X = x|Y = y) = \frac{P(X = x)P(Y = y|X = x)}{\sum_i P(X = x_i)P(Y = y|X = x_i)} \tag{2}$$

An important issue with inference is the computational efficiency. The conditional independence properties help greatly in enhancing the efficiency. This issue has been discussed in detail in [2][5].

2.3. Learning Bayesian Nets From Data

As we discussed earlier, the two key elements of a Bayesian network are the graph topology (structure) and the parameters of each CPD. It is possible to learn both of these from data. There are four cases of learning a Bayesian network from training data: structure known, data fully observable; structure known, data partially observable; structure unknown, data fully observable; structure unknown, data partially observable.

2.3.1. Known Structure With Full Observability

We use Maximum Likelihood Estimation (MLE) in this case. The goal is to find the values of the parameters of each CPD which maximizes the likelihood of the training data (N independent cases). The normalized log-likelihood of the training set D is an averaged summation over all nodes:

$$L = \frac{1}{N}\sum_{i=1}^{m}\sum_{l=1}^{S}\log P(X_i|P_a(X_i),D_l) \tag{3}$$

We see that the log-likelihood scoring function decomposes according to the structure of the graph, hence we can maximize the contribution to the log-likelihood at each node independently (assuming the parameters for each node are independent of those for all other nodes). For Gaussian nodes, we may compute the sample mean and variance, and use linear regression to estimate the weight matrix. For other kinds of distributions, more complex procedures are necessary.

2.3.2. Known Structure With Partial Observability

When some of the nodes are hidden, we may use the Expectation Maximization (EM) algorithm to find a locally optimal maximum likelihood estimate of the parameters. In the first step, we compute the expected values for all nodes using an inference algorithm, and then treat these expected values as though they were

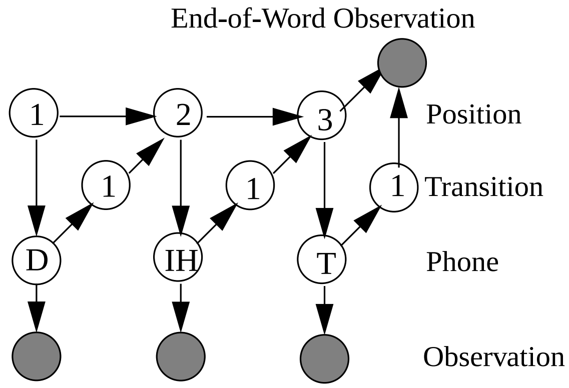

Figure 2: A DBN representation of a simple HMM.

observed (distributions) so that we may maximize the parameters. Then we recompute the expected values, redo the maximization again. This iterative procedure is guaranteed to converge to a local maximum of the likelihood surface. From this procedure, we see that when nodes are hidden, inference becomes a subroutine which is called by the learning procedure. Hence fast inference algorithms are crucial.

2.3.3. Unknown Structure With Full Observability

This case is complicated since we do not know exactly what the structure is like. The strategy here is to search through the model space. We need to select an appropriate model according to a scoring function, and we optimize this function over the space of models.

2.3.4. Unknown Structure With Partial Observability

This is the hardest case of all, where the structure is unknown and there are hidden variables and/or missing data. We can use an iterative method, which alternates between evaluating the expected score of a model with an inference engine, and changing the model structure, until a local maximum is reached. This is also known as the Structural EM (SEM) algorithm.

3. APPLICATION IN ASR

Dynamic Bayesian networks (DBNs) are used to model stochastic processes. It is capable of modeling arbitrary sets of variables with arbitrary conditional independence assumptions. This property enables the construction of explicit models of speech generation and perception, hence makes DBNs suitable for speech recognition.

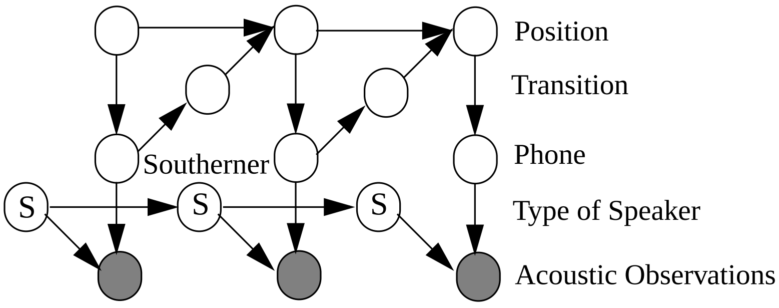

Figure 3: A DBN structured to model speaker type.

3.1. Model Composition With DBNs

In order to apply DBNs to ASR, it is necessary to develop a technique for combining soberer phonetic models into whole word and multiple-word models [3]. For model composition, we need to specify legal submodel sequences first. Stochastic Finite-State Automata (SFSAs) have been used to describe a probability distribution over possible submodel sequences. Bayesian networks are used to specify the behavior of each submodel. The model composition allows for parameter tying between multiple occurrences of the same phone model.

Figure 2 illustrates an example of a DBN that is structured for model composition in speech recognition. It is equivalent to a standard HMM.

3.2. Model Structures For ASR

DBNs can be adapted to address the requirements of automatic speech recognition. They can model many of the important factors affecting the speech recognition process, such as articulatory motion, speaking style, noise, etc. [3]. Figure 3 illustrates a DBN structured to model speaker type. More examples, such as DBN structured to model articulatory motion, DBN structured to model speaking rate, perceptually-structured model and combined perceptual-generative model, can be found in Zweig’s dissertation [3].

3.3. Performance

Speech recognition experiments with DBNs on a

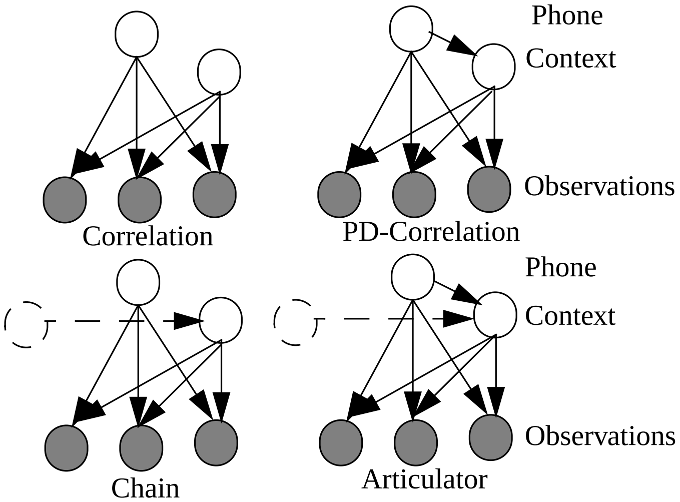

Figure 4: The acoustic models for four of the network topologies tested. The index and transition variables are omitted. The dotted lines indicate conditioning on the previous frame.

large-vocabulary multi-speaker database of isolated words have been described in [3]. Some network structures tested and the corresponding results are shown in Figure 4 and Table 1, respectively.

Here, the "Correlation" network models intra-frame observation correlations in a phone-independent way. The "PD-Correlation" network models the correlations in a phone-dependent way. The "Chain" network models phone-independent temporal correlations. And the "Articulator" network models phone-dependent articulatory target positions and inertial constraints. The Articulator network provided the best performance.

3.4. Advantages Of DBNs-based Speech Recognition

First, arbitrary sets of variables can be associated with each timeslice. This enables a highly expressive representational framework. Second, there are efficient, general-purpose algorithms for doing inference, and no special-purpose algorithms need to be derived for handling extensions to HMMs such as articulator models. Third, sharing variables between submodels leads to a natural way of describing transitional behavior, which is important for modeling coarticulation. Fourth, DBNs are factored representations of a probability distribution, and may have exponentially fewer parameters than unfactored representations such as standard HMMs. Hence these parameters can be estimated more accurately with a fixed amount of data. This is also known as statistical efficiency. Finally, gains in statistical efficiency result in computational efficiency.

Network	Parameters	Error Rate
Baseline HMM	127k	4.8%
Correlation	254k	3.7%
PD-Correlation	254k	4.2%
Chain	254k	3.6%
Articulator	255k	3.4%

Table 1: Word error rates for the four models in Figure 4 using the basic phoneme alphabet.

4. CONCLUSIONS

In this paper, we reviewed the basic concepts behind the theory of Bayesian networks. We also discussed important issues involved in inference and learning a Bayesian network from data. We demonstrated the possibility and the benefits of applying Bayesian networks to speech recognition. Finally we presented experiment results on speech recognition with Bayesian nets from state-of-the-art research work.

5. REFERENCES

[1] N. Friedman and M. Goldszmidt, "Learning Bayesian Networks from Data," *http://www.dsv.su.se/ijcai-99/tutorials/d3.html,* 2000.

[2] K. P. Murphy, "A Brief Introduction to Graphical Models and Bayesian Networks," *http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html,* University of California at Berkeley, 2000.

[3] G. Zweig, "Speech Recognition with Dynamic Bayesian Networks," *Ph.D. Dissertation,* University of California at Berkeley, Berkeley, California, 1998.

[4] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," *Technical Report MSR-TR-95-06,* Microsoft Research, March, 1995.

[5] R. McEliece and S. M. Aji, "The Generalized Distributive Law," *http://www.systems.caltech.edu/EE/Faculty/rjm/papers/GDL.ps,* California Institute of Technology, 1999.