# Bayesian Networks for Phonetic Classification Using Time-Scale Features

*Franz Pernkopf and Tuan Van Pham*

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
pernkopf@tugraz.at, v.t.pham@tugraz.at

10.21437/Interspeech.2006-234

## Abstract

We present a phonetic classification approach based on Bayesian networks using time-scale features which are extracted from the discrete Wavelet transform. We apply Bayesian networks using discriminative and generative parameter and/or structure learning for classifying the speech frames into silence, voiced, unvoiced, mixed sounds, and two more categories, voiced closure and release of plosives. Gender dependent/independent experiments have been performed on the TIMIT database. The experiments show that (i) our time-scale features mostly outperform standard MFCC features, (ii) discriminative learning of Bayesian networks is superior to the generative approach.

**Index Terms**: phonetic classification, time-scale features, wavelet transform, Bayesian networks, discriminative learning.

## 1. Introduction

One of the most critical tasks in speech processing and speech applications is automatic phonetic classification. In speech recognition, a phonetic classifier is needed to enhance the endpoint detection performance in order to increase the word recognition rate. The accurate discrimination between phonetic classes will improve the output quality of data-driven speech synthesizers by adjusting non-uniform scaling factors of each phonetic class based on time-scaling modification algorithms. Voice activity detection which is employed by most speech applications is a direct application of phonetic classification.

The speech classification is done by training a model to learn differences of statistical distributions of the acoustic features between different phonetic classes [1]. The acoustic features can be derived in the time domain, e.g., zero crossing rate, energy level, and autocorrelation coefficients [2]. Frequently used features in the frequency domain are cepstrum pitch detection [3] and mel frequency cepstral coefficients (MFCC) [4]. These features are based on the short-time Fourier transform (STFT) which shows a shortcoming of the rigid time-frequency (TF) plane. The extracted features from the discrete Wavelet transform (DWT) which overcome the shortcomings of the STFT by a flexible resolution of the TF plane can improve the classification rate significantly [5].

Different classification approaches, e.g., Gaussian mixture model and multi-layer perceptron have been studied in [6] for phonetic classification. To the best of our knowledge, discriminatively trained Bayesian networks, which achieves promising results in other classification domains [7], have not been used for phonetic classification so far. 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. The usual approach for learning a generative model is maximum likelihood (ML) estimation. Discriminative classifiers directly model the class posterior probability. Maximizing the conditional likelihood (CL) of the class given the attributes results in optimizing the ability to correctly predict the class. Unfortunately, the CL for Bayesian networks is not decomposable, i.e., there is no closed-form solution. Recently, some approaches have been suggested to learn the structure and/or parameters discriminatively by maximizing the class conditional likelihood (CL) or the classification rate (CR). An excellent overview is provided in [7].

In this paper, we apply both discriminative parameter learning by optimizing the CL and generative parameter training (ML estimation) on *both* discriminatively *and* generatively structured Bayesian networks. The naive Bayes (NB) and the tree augmented naive Bayes (TAN) classifiers are used. Time-scale features based on the DWT are employed to improve the phonetic classification. The features are derived by applying a wavelet decomposition at the $4^{th}$ scale on every windowed speech frame of 16ms length and 8ms overlap. Based on these features, the speech frames are classified into four classes: voiced (V), unvoiced (U), silence (S), and mixed (M) sounds, or into six classes which include two additional classes of voiced closure (VC) and release of plosives (R).

The paper is organized as follows: Section 2 introduces Bayesian network classifiers, the NB and TAN structures, and generative/discriminative structure learning. Feature extraction based on DWT is presented in Section 3. Experiments on the TIMIT database and discussions are presented in Section 4. Section 5 concludes and gives perspectives for future research.

## 2. Bayesian network classifier

A Bayesian network [8] $\mathcal{B} = \langle \mathcal{G}, \Theta \rangle$ 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 $\mathbf{E} = \{E_{Z_i,Z_j}, E_{Z_i,Z_k}, \ldots\}$ connecting the nodes where $E_{Z_i,Z_j}$ is an edge from $Z_i$ to $Z_j$. This graph represents factorization properties of the distribution of a set of random variables $\mathbf{Z} = \{C, X_1, \ldots, X_N\} = \{Z_1, \ldots, Z_{N+1}\}$, where each variable in $\mathbf{Z}$ has values denoted by lower case letters $\{c, x_1, \ldots, x_N\}$. We use boldface capital letters, e.g. $\mathbf{Z}$, to denote a set of random variables and correspondingly lower case boldface letters denote a set of instantiations (values). The random variable $C \in \{1, \ldots, |C|\}$ represents the classes, $|C|$ is the cardinality of $C$, $\mathbf{X}_{1:N} = \{X_1, \ldots, X_N\}$ denote the set of random variables of the $N$ attributes of the classifier. Each graph node represents a random variable, while the lack of edges specifies independencies. 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 $\Theta$ represents the set of

parameters which quantify the network. Each node $Z_j$ is represented as a local conditional probability distribution given its parents $Z_{\Pi_j}$. The joint probability distribution of the network is determined by the local conditional probability distributions as $P_\Theta(\mathbf{Z}) = \prod_{j=1}^{N+1} P_\Theta\left(Z_j|Z_{\Pi_j}\right)$.

Generative parameter learning, i.e. ML estimation is discussed in [8]. Discriminative parameter learning by optimizing the CL is presented in [9].

### 2.1. NB and TAN structures

The NB network assumes that all the attributes are conditionally independent given the class label. As reported in the literature [10], 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 is illustrated in Figure 1a.

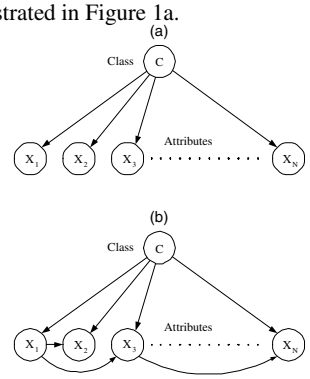


Figure 1: *Bayesian Network: NB (a), TAN (b).*

In order to correct some of the limitations of the NB classifier, Friedman et al. [10] introduced the TAN classifier. A TAN is based on structural augmentations of the NB network, where additional 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 (1-tree). Hence, the maximum number of edges added to relax the independence assumption between the attributes is $N-1$. An example of a TAN network is shown in Figure 1b. A TAN network is typically initialized as a NB network. Additional edges between attributes are determined through structure learning.

### 2.2. Generative structure learning of TAN:

The conditional mutual information (CMI) $I(X_i; X_j|C)$ between the attributes given the class variable is used as score. This measures the information between $X_i$ and $X_j$ in the context of $C$. Friedman et al. [10] give an algorithm for constructing a TAN network using this measure. In the following, we shortly review this algorithm for constructing the classifier structure:

1. Compute the pairwise CMI $I(X_i; X_j|C) \quad \forall \quad 1 \leq i \leq N$ and $i \leq j \leq N$.
2. Build a complete undirected 1-tree using the maximal weighted spanning tree algorithm where each edge connecting $X_i$ and $X_j$ is weighted by $I(X_i; X_j|C)$.
3. Transform the complete undirected 1-tree to a directed tree. Therefore, select a root variable and direct all edges away from this root. Add to this tree the class node $C$ and the edges from $C$ to all attributes $X_1, \ldots, X_N$.

### 2.3. Discriminative structure learning of TAN:

We use an order-based greedy search heuristic for efficient learning of the discriminative structure of a Bayesian network classifier [7]. The best network consistent with a given variable ordering can be found in $\mathcal{O}\left(N^k\right)$ where $k$ is the upper bound of parents per node. This fact is used in our *order mutual information* (OMI) heuristic for learning discriminative structures [7]. Our procedure first looks for an ordering $\prec$ of the variables $\mathbf{X}_{1:N}$ according to the conditional mutual information. If the graph is consistent with the ordering $X_i \prec X_j$ then the parent $X_{\Pi_j} \in \mathbf{X}_{\Pi_j}$ is one of the variables which appear before $X_j$ in the ordering, where $\mathbf{X}_{\Pi_j}$ is the set of possible parents for $X_j$. This constraint ensures that the network stays acyclic. More specifically, our 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

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

where $j \in \{1, \ldots, N\}$. The first node $X_\prec^1$ is the node with the largest information about $C$, i.e. it is most important for $C$.

In the second step of the algorithm, we connect $X_\prec^j$ ($\forall j \in \{2, \ldots, N\}$) to the selected parent $X_\prec^* \in \mathbf{X}_{\Pi_j} = \mathbf{X}_\prec^{1:j-1}$ by maximizing the classification rate

$$X_\prec^* \leftarrow \arg \max_{X \in \mathbf{X}_\prec^{1:j-1}} CR\left(\mathcal{B}_\mathcal{S}|\mathcal{S}\right), \qquad (2)$$

of the current $\mathcal{B}_\mathcal{S}$. ($\mathbf{E} \leftarrow \left\{\mathbf{E} \cup E_{X_\prec^*, X_\prec^j}\right\}$ starting with $\mathbf{E}_{NaiveBayes}$). The classification rate $CR(\mathcal{B}_\mathcal{S}|\mathcal{S}) = \frac{1}{R}\sum_{r=1}^{R} \delta\left(\mathcal{B}_\mathcal{S}\left(x_{1:N}^r\right), c^r\right)$. The expression $\delta\left(\mathcal{B}_\mathcal{S}\left(x_{1:N}^r\right), c^r\right) = 1$ if the Bayesian network classifier $\mathcal{B}_\mathcal{S}\left(x_{1:N}^r\right)$ trained with samples in $\mathcal{S}$ assigns the correct class label $c^r$ to the attribute values $x_{1:N}^r$. The training data consists of $R$ samples $\mathcal{S} = \{\mathbf{z}^r\}_{r=1}^R = \{(c^r, \mathbf{x}_{1:N}^r)\}_{r=1}^R$.

## 3. Time-scale feature extraction

### 3.1. Wavelet-based multiresolution analysis

With the DWT, various positions in the time-frequency plane are analyzed with different time-frequency resolutions which overcomes the limitation of STFT. The advantage of DWT in speech processing is based on the relation between DWT and multiresolution analysis (MRA) which allows the multiscale representation of speech signals in the time-scale domain. A discrete-time signal $x[k]$ can be represented as:

$$x[k] = \sum_m \sum_n \langle \psi_{m,n}, x \rangle \, \tilde{\psi}_{m,n}[k], \qquad (3)$$

where the discrete-time wavelet basis function $\psi_{m,n}[k]$ is constructed from iterated filters, $m, n, k \in \mathbb{Z}$. Based on the MRA, the signal $x[k]$ can be represented as the sum of an approximation plus $L$ details at $L$ decomposition stages:

$$\begin{aligned} x[k] = &\sum_{n=-\infty}^{\infty} X^{(L)}[2n] \cdot g_0^{(L)}[k - 2^L n] + \\ &\sum_{m=1}^{L} \sum_{n=-\infty}^{\infty} X^{(m)}[2n+1] \cdot g_1^{(m)}[k - 2^m n], \end{aligned} \qquad (4)$$

where

$$X^{(L)}[2n] = \left\langle h_0^{(L)}[2^L n - l], x[l] \right\rangle, \quad X^{(m)}[2n+1] = \left\langle h_1^{(m)}[2^m n - l], x[l] \right\rangle, \tag{5}$$

are the approximation coefficients and the detail coefficients, respectively, at the output of the iterated filter bank with $L$ stages. $g_0^{(m)}[k]$ is an equivalent filter obtained through $m$ stages of low-pass synthesis filters $g_0[k]$, preceded by an upsampler by a factor of 2. We call $W_{m,i}(n)$ the sequence of all wavelet coefficients (i.e, the $X^{(L)}[2n]$ and $X^{(m)}[2n+1]$) which are derived by WD at the $m^{th}$ scale of the $i^{th}$ frame, $n$ is the coefficient index, $i \in \mathbb{Z}$.

### 3.2. Feature extraction

In this paper, we want to classify six types of phonetic classes which have the same phonetic characteristics as silence, voiced (vowels, semivowels, diphthongs and nasals), unvoiced (unvoiced fricatives), mixed (voiced fricatives and glottal fricatives), and two classes composing plosives: voiced closure and release.

Depending on the phonetic properties of the input speech frames, the power of details increases from $1^{st}$ scale to $4^{th}$ scale for voiced frames and vice versa for unvoiced frames. There is no power change over various scales for mixed and silence frames. Furthermore, we observe from statistical distribution of speech sound that the power of voiced frames is mostly concentrated in the low-frequency subbands in the range 0-4 kHz, and much less in the high-frequency subbands. This is reversely for the unvoiced frames, and relatively equal energy distribution occurs for the mixed sounds. The voiced closure of voiced plosives show a periodic structure and slightly high energy of approximation part (at $4^{th}$ scale). Some voiced consonants which are considered as mixed class have low power as voiced closure but has higher standard deviation at the $1^{st}$ detail part (high-frequency subband). These property can be used as specific representations of the defined classes. A set of the time-scale features is derived as follows:

- ***Power delta*** *(D)* is the power difference between approximation with detail at highest scale and detail at lowest scale:

$$D(i) = \frac{1}{N_{f3}} \sum_{n=1}^{N_{f3}} W_i^2(n) - \frac{1}{N_{f1}} \sum_{n=N_{f1}+1}^{N_f} W_i^2(n). \tag{6}$$

- ***Power ratio*** *($PR_1$)* is the power ratio between approximation and three details at three highest scales:

$$PR_1(i) = \frac{N_{f1} - N_{f4}}{N_{f4}} \frac{\sum_{n=1}^{N_{f4}} W_i^2(n)}{\sum_{n=N_{f4}+1}^{N_{f1}} W_i^2(n)} \tag{7}$$

- ***Power ratio*** *($PR_2$)* is the power ratio between details of two lowest scales and approximation with detail at highest scale:

$$PR_2(i) = \frac{N_{f3}}{N_f - N_{f2}} \frac{\sum_{n=N_{f2}+1}^{N_f} W_i^2(n)}{\sum_{n=1}^{N_{f3}} W_i^2(n)} \tag{8}$$

- ***Power of approximation*** *($PA$)*:

$$PA(i) = \frac{\sum_{n=1}^{N_{f4}} W_i^2(n)}{N_{f4}} \tag{9}$$

- ***Standard deviation of detail at lowest scale*** *($SD$)*:

$$SD(i) = \sqrt{\frac{\sum_{n=N_{f1}+1}^{N_f} (W_i(n) - \overline{W_i(n)})^2}{N_{f1}}} \tag{10}$$

- ***Peak delta*** *($PD$)* is the distance between the peak values of the first and second lobe obtained from the autocorrelation function of each speech frame:

$$PD(i) = R_i(j_1) - R_i(j_2) \tag{11}$$

where $R_i$ is the autocorrelation function that shows the peak values at lag $j_1$ and $j_2$.

- ***Logarithmic short-term energy*** *(LgSE)*:

$$LgSE = 0.5 + \frac{16}{ln(2)} ln\left(1 + \frac{\sum_{n=1}^{N_f} x(n)^2}{32}\right) \tag{12}$$

- ***Zero crossing rate*** *(ZCR)*:

$$ZCR = \sum_{n=1}^{N_f} |sgn[x(n) - sgn(x(n-1))]| \tag{13}$$

where $N_{f1} = \frac{N_f}{2}$, $N_{f2} = \frac{N_f}{4}$, $N_{f3} = \frac{N_f}{8}$, $N_{f4} = \frac{N_f}{16}$ are indeces of the approximation and detail parts in the sequence of the wavelet coefficients, and $N_f$ is number of samples in one speech frame. The evolution of extracted features is shown in Fig. 2.

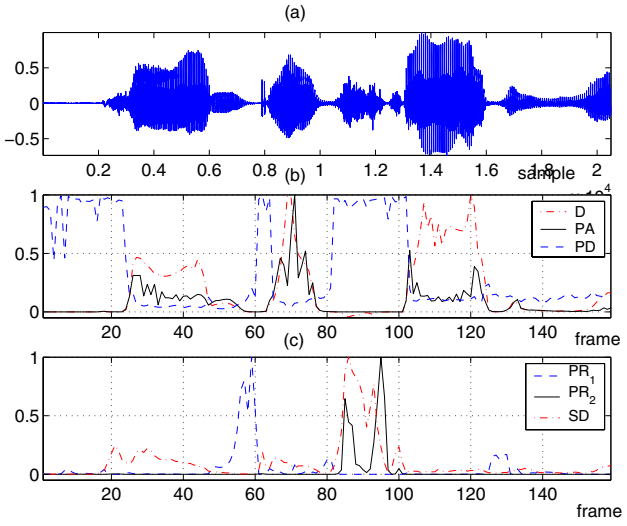


Figure 2: Speech signal (a), evolution of features (b), (c).

## 4. Experiments and Evaluations

We apply both discriminative parameter learning by optimizing the CL [9] and generative parameter training (ML estimation) [8] on both discriminatively (TAN-OMI) and generatively (TAN-CMI) structured Bayesian networks. We use the NB and the TAN classifier topology. Experiments have been performed on data from the

TIMIT speech corpus using the dialect speaking region 4 which consists of 16 male and 16 female speakers, 320 utterances, and 121629 frames in total. All speech sounds are sampled at 16 kHz. The distributions of four phonetic classes V/U/S/M and six phonetic classes V/U/S/M/VC/R are 23.08%, 60,37%, 13.54%, 3.01% and 20.9%, 54.66%, 12.26%, 2.74%, 6.08%, 3.36%, respectively. We perform classification experiments on data of male speakers (Ma), female speakers (Fe), and both (Ma+Fe) genders. The data have been split into 2 mutually exclusive subsets of $\mathcal{D} \in \{\mathcal{S}_1, \mathcal{S}_2\}$ where the size of the training data $\mathcal{S}_1$ is 70% and of the test data $\mathcal{S}_2$ is 30% of $\mathcal{D}$. Throughout the experiments, we use exactly the same data partitioning. Additionally, to our 8 time-scale features (TSF) (see Section 3), we perform experiments using baseline features, i.e. , 12 MFCC + Log-Energy. The attributes in the data sets are continuous-valued. Since the classifiers are constructed for multinomial attributes, the features have been discretized using the algorithm in [11] where the codebook is produced using only the training data. Zero probabilities in the conditional probability tables of the Bayesian networks are replaced with a small epsilon $\varepsilon = 0.00001$. Discriminative parameter learning is currently implemented in a naive way. We either perform 15 iterations of the gradient descent algorithm or prematurely terminate the parameter optimization in case of convergence.

Table 1 and Table 2 present the classification performance for all different generative/discriminative classifiers for 4 and 6 phonetic classes, respectively.

Table 1: *Classification results* [%] *for 4 classes.*

| Classifier Struct. Learn. Param. Learn. | | NB - ML | NB - CL | TAN CMI ML | TAN CMI CL | TAN OMI ML | TAN OMI CL |
|---|---|---|---|---|---|---|---|
| Data set | Features | | | | | | |
| Ma+Fe | TSF | 88.36 | 88.53 | 90.67 | 90.69 | 90.92 | **90.96** |
| Ma | TSF | 89.68 | 89.80 | 91.11 | 91.12 | 91.25 | **91.27** |
| Fe | TSF | 87.85 | 87.92 | 89.55 | 89.57 | 90.35 | **90.38** |
| Ma+Fe | MFCC | 88.58 | 88.76 | 90.61 | 90.62 | 90.64 | 90.62 |
| Ma | MFCC | 89.01 | 89.25 | 90.86 | 90.88 | 91.24 | **91.27** |
| Fe | MFCC | 88.59 | 88.65 | 89.85 | 89.84 | 89.92 | 89.92 |

Table 2: *Classification results* [%] *for 6 classes.*

| Classifier Struct. Learn. Param. Learn. | | NB - ML | NB - CL | TAN CMI ML | TAN CMI CL | TAN OMI ML | TAN OMI CL |
|---|---|---|---|---|---|---|---|
| Data set | Features | | | | | | |
| Ma+Fe | TSF | 81.99 | 82.04 | 83.00 | 83.05 | 83.48 | **83.50** |
| Ma | TSF | 82.78 | 82.86 | 83.93 | 83.95 | 84.47 | **84.47** |
| Fe | TSF | 81.41 | 81.51 | 82.24 | 82.27 | 82.76 | 82.75 |
| Ma+Fe | MFCC | 82.08 | 82.16 | 82.89 | 82.91 | 83.39 | 83.40 |
| Ma | MFCC | 82.35 | 82.48 | 83.87 | 83.88 | 84.28 | 84.31 |
| Fe | MFCC | 81.99 | 82.05 | 82.63 | 82.63 | 83.03 | **83.06** |

These tables show that the TAN classifier using discriminative structure and parameter learning (TAN-OMI-CL) outperforms the generative approaches. However, the evaluation of the CR in the OMI algorithm is computationally expensive. Discriminative parameter learning (CL) produces mostly a better classification performance than generative parameter learning (ML).

The proposed time-scale features outperform the baseline MFCC features in most cases. This results from the flexibility of having short basis functions to analyze high-frequency speech components while long ones are applied on low-frequency speech components of the DWT. Additionally, for TSF we have only 8 features compared to 13 MFCC features. This results in a lower complexity of the classifier. The small differences of classification performance between Ma+Fe, Ma, and Fe open an approach for gender independent phonetic classification.

## 5. Conclusion

Bayesian networks are used to classify speech frames into silence, voiced, unvoiced, mixed sounds, and two more categories voiced closure and release of plosives. The classification is based on time-scale features derived from discrete Wavelet transform. Discriminative and generative parameter and/or structure learning approaches are used for learning the Bayesian network model. Gender dependent/independent experiments have been performed on the TIMIT database. Discriminative structure learning of Bayesian networks is superior to the generative approach. Discriminative parameter training improves the classification rate in most cases. Our time-scale features mostly outperform standard MFCC features. Future work includes the investigation of the time-scale features to improve the classifiation rate of the plosives.

## 6. Acknowledgments

We kindly acknowledge the partial support from the MISTRAL Project, financed by the Austrian Research Promotion Agency (www.ffg.at) within the strategic objective FIT-IT under the project contract number 809264/9338.

## 7. References

[1] B. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. on Acous. Speech and Signal Proc.*, vol. 24, no. 3, pp. 201–212, 1976.

[2] D. G. Childers, M. Hahn, and J. N. Larar, "Silence and voiced/unvoiced/mixed excitation classification of speech," *IEEE Trans. on Acoust, Speech, Signal Process.*, vol. 37, no. 11, pp. 1771–1774, 1989.

[3] A.S. Spanias S. Ahmadi, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Trans. on Speech, Audio Proc.*, vol. 7, no. 3, pp. 333–338, 1999.

[4] Z. Xiong and T. Huang, "Boosting speech/non-speech classification using averaged mel-frequency cepstrum," *Proc. IEEE Pacific-Rim Conf. on Multimedia*, 2002.

[5] T. V. Pham and G. Kubin, "DWT-based phonetic groups classification using neural network," *Proc. IEEE Int. Conf. on Acoust, Speech, Signal Process.*, pp. 401–404, 2005.

[6] H.C. Leung, B. Chigier, and J.R. Glass, "A comparative study of signal representations and classification techniques for speech recognition," in *Proc. IEEE Int. Conf. on Acoust, Speech, Signal Process.*, 1993, pp. 657 – 664.

[7] F. Pernkopf and J. Bilmes, "Ordering-based discriminative structure learning for Bayesian network classifiers," in *Technical Report*, 2006.

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

[9] R. Greiner and W. Zhou, "Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers," in *18th Conf. of the AAAI*, 2002, pp. 167–173.

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

[11] U.M. Fayyad and K.B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1027.