IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 9, 2019, accepted November 25, 2019, date of publication November 28, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956477

Positive Sequential Data Modeling Using Continuous Hidden Markov Models Based on Inverted Dirichlet Mixtures

RU WANG[1] **AND WENTAO FAN**[1,2], **(Member, IEEE)**
[1]Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[2]Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China

Corresponding author: Wentao Fan (fwt@hqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61876068, in part by the Natural Science Foundation of Fujian Province under Grant 2018J01094, and in part by the Promotion Program for Young and Middle-Aged Teacher in Science and Technology Research of Huaqiao University under Grant ZQN-PY510.

ABSTRACT The hidden Markov model (HMM) has long been one of the most commonly used probability graph models for modeling sequential or time series data. It has been widely used in many fields ranging from speech recognition, face recognition, anomaly detection, to gene function prediction. In this paper, we theoretically propose a variant of the continuous HMM for modeling positive sequential data which are naturally generated in many real-life applications. In contrast with conventional HMMs which often use Gaussian distributions or Gaussian mixture models as the emission probability density, we adopt the inverted Dirichlet mixture model as the emission density to build the HMM. The consideration of inverted Dirichlet mixture model in our case is motivated by its superior modeling capability over Gaussian mixture models for modeling positive data according to several recent studies. In addition, we develop a convergence-guaranteed approach to learning the proposed inverted Dirichlet-based HMM through variational Bayes inference. The effectiveness of the proposed HMM is validated through both synthetic data sets and a real-world application regarding anomaly network intrusion detection. Based on the experimental results, the proposed inverted Dirichlet-based HMM is able to achieve the detection accuracy rates that are about 4%∼9% higher than those ones obtained by the compared approaches.

INDEX TERMS Hidden Markov models, inverted Dirichlet distribution, variational Bayes, mixture models, intrusion detection.

I. INTRODUCTION

The hidden Markov model (HMM) [1], [2] is one of the most commonly used probability graphical models for modeling sequential or time series data, such as video, audio, text, etc. It has been widely used in many fields ranging from handwritten word recognition, speech recognition, speech synthesis, face recognition, anomaly detection, to gene function prediction [3]–[8]. The HMM contains a set of hidden states that are assumed to form a Markov chain, and each of these states is associated with a probability distribution that controls the emission of the observed data.

In continuous HMMs, a common choice for the emission density is the Gaussian distribution (or mixture of Gaussian distributions) [2], [9]. However, by taking the nature of the data into account, several works have shown that other distributions may become effective alternatives to Gaussian for modeling data with non-Gaussian structure [10]–[12]. Examples of these research works include inverted Dirichlet-based models [13]–[16], have demonstrated better performance than Gaussian-based models in modeling semi-bounded data (i.e. positive vectors), which naturally appear in several pattern recognition and computer vision applications [13]. The inverted Dirichlet (ID) distribution, developed by Tiao and Cuttman [17], is a multivariate generalization of the inverted Beta distribution. For the Gaussian distribution, the domain of its probability density function is not bounded,

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

i.e. $(-\infty, +\infty)$. However, for the ID distribution, the domain of its probability density function is semi-bounded, i.e. $[0, +\infty)$. This tighter bounded domain makes the ID distribution more suitable than Gaussian for modeling positive data. Moreover, the ID distribution is a multivariate generalization of the inverted Beta distribution. In contrast with Gaussian distribution which only allows symmetric modes, the ID distribution permits both multiple symmetric and asymmetric modes, which results in more flexibility and better modeling capability. Moreover, the ID distribution is also closely related to the multivariate Student's-t distribution as discussed in [17]. Therefore, motivated by its advantages over Gaussian mixture models, we consider the ID mixture model as the emission density, to propose the ID-based continuous HMMs for modeling positive sequential data sequences.

In order to learn HMMs, most of the approaches in the literature adopt the EM algorithm [18], which in the case of HMMs is known as the Baum-Welch (B-W) algorithm, to find maximum likelihood solutions for HMMs [19], [20]. Unfortunately, the EM algorithm contains several limitations such as its suboptimal generalization performance, dependency on initialization and over-fitting. To tackle these problems, several research works [10], [21], [22] have adopted variational Bayes inference to learn the HMMs [23], [24], which is an effective inference algorithm based on deterministic approximation. In the variational Bayes inference, the overfitting associated with maximum likelihood can be avoided by the incorporation of prior knowledge (or belief) in a principled way and then marginalizing (i.e. integrating) over the model parameters instead of making point estimates of their values. Furthermore, in contrast with other well-known inference approaches such as Markov chain Monte Carlo (MCMC), variational Bayes is more computationally efficient and can assess the convergence of the learning algorithm in a systematical way. Motivated from its merits as mentioned above, we aim to develop an effective and efficient variational Bayes inference approach to learning the proposed ID-based continuous HMMs.

The contributions of this work can be summarized as follows. First, we theoretically propose a variant of the HMM based on ID mixture models. To the best of our knowledge, this is the first work to consider the ID mixture model as the emission density of the continuous HMM model for modeling positive sequential data. Second, we derive complete inference procedures for learning the proposed ID-based HMM model using variational Bayes. Third, the proposed ID-based HMM model is validated through both simulated data sets and is applied to a real-world problem of network intrusion detection.

The rest of the paper is organized as follows: Section 2 describes the ID-based HMM model. In Section 3, an effective approach based on variational Bayes is developed for learning the ID-based HMM model. Section 4 reports the experimental results using both simulated data sets and a real-world application. Conclusion is provided in Section 5.

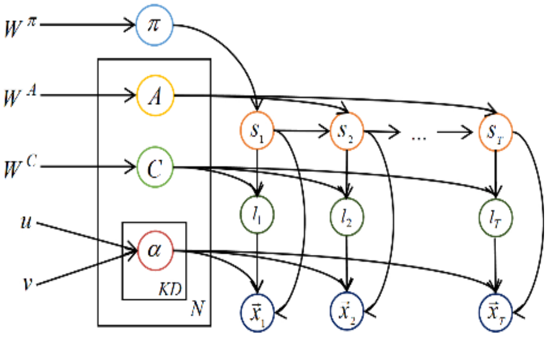

FIGURE 1. The proposed ID-based HMM model.

II. THE INVERTED DIRICHLET-BASED HMMS

Given a D-dimensional random vector $\vec{x} = (x_1, \ldots, x_D)$ which follows an inverted Dirichlet (ID) distribution with parameter $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{D+1})$, its probability density function (pdf) is given by [17]:

$$\mathrm{ID}(\vec{x}|\vec{\alpha}) = \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1} \left(1 + \sum_{d=1}^{D} x_d\right)^{-\sum_{d=1}^{D+1} \alpha_d} \quad (1)$$

where $\Gamma(\cdot)$ is the gamma function which is defined by $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \mathrm{d}t$, $x_d > 0$ for $d = 1, \ldots, D$, and $\alpha_d > 0$ for $d = 1, \ldots, D+1$.

Given a HMM model with N states, if the hidden emission density of each state follows a mixture of K ID distributions, then the ID-based HMM model can be modeled by a set of parameters $\Phi = \{\vec{\pi}, A, C, \vec{\alpha}\}$, where $\vec{\pi} = \{\pi_i\}_i^N$ represents the initial-state probability vector, $A = \{a_{ij}\}_{i,j}^{N,N}$ denotes the state transition matrix, $C = \{c_{ik}\}_{i,k}^{N,K}$ denotes the mixture coefficient matrix with c_{ik} represents the mixing proportion of the kth mixture component under the ith state, and $\vec{\alpha} = \{\vec{\alpha}_{ik}\}_{i,k}^{N,K}$ denotes the parameter of the kth ID component under the i state.

Assume that we have collected a D-dimensional sequence $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ with T observations, the associated latent variables of the complete-data are $S = \{s_1, \ldots, s_T\}$ and $L = \{l_1, \ldots, l_T\}$, where $s_t \in [1, N]$ is the unobserved state sequence, and $l_t \in [1, K]$ is the unobserved variable that indicates from which mixture component that the tth observation is generated. The graphical representation of the proposed ID-based HMM model is shown in Figure 1.

Based on the framework the ID-based HMM model, the probability of the complete data $\{X, S, L\}$ given the set of parameters Φ is defined by

$$p(X, S, L|\Phi) = \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} \mathrm{ID}(\vec{x}_t|\vec{\alpha}_{s_t l_t})\right]. \quad (2)$$

Then, the likelihood of model parameters Φ given the observation sequence X is formulated by

$$p(X|\Phi) = \sum_{S,L} \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} \mathrm{ID}(\vec{x}_t|\vec{\alpha}_{s_t l_t})\right]. \quad (3)$$

Since we are constructing a full Bayesian model, each random variable of the ID-based HMM is associated with a prior. Following [10], [21], [22], the common choice of the priors for parameters $\vec{\pi}$, A and C is the Dirichlet distribution $Dir(\cdot)$

$$p(\vec{\pi}) = \mathrm{Dir}(\pi_1, \ldots, \pi_N|\phi_1^{\pi}, \ldots, \phi_N^{\pi}), \tag{4}$$

$$p(A) = \prod_{i}^{N} \mathrm{Dir}(a_{i1}, \ldots, a_{iN}|\phi_{i1}^{A}, \ldots, \phi_{iN}^{A}), \tag{5}$$

$$p(C) = \prod_{i}^{N} \mathrm{Dir}(c_{i1}, \ldots, c_{ik}|\phi_{i1}^{C}, \ldots, \phi_{ik}^{C}). \tag{6}$$

For the parameter $\vec{\alpha}$ of the ID distributions, Gamma distribution $Gam(\cdot)$ is a reasonable choice for its prior since $\vec{\alpha}$ has to be positive. Thus, we have

$$\begin{aligned} p(\vec{\alpha}) &= \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D} \mathrm{Gam}(\alpha_{ikd}|u_{ikd}, v_{ikd}) \\ &= \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D} \frac{v_{ikd}^{u_{ikd}-1}}{\Gamma(u_{ikd})}\alpha_{ikd}^{u_{ikd}-1}e^{-v_{ikd}\alpha_{ikd}}. \end{aligned} \tag{7}$$

III. MODEL LEARNING THROUGH VARIATIONAL BAYES

The proposed ID-based HMM model is learnt using variational Bayes [23], [24], which is an effective inference algorithm based on deterministic approximation. The main idea of variational Bayes is to find an approximation $q(S, L, \Phi)$ for the posterior distribution $p(S, L, \Phi|X)$, where $\{S, L, \Phi\}$ denotes the set of all latent variables and parameters. In order to have a tractable inference for the $q(S, L, \Phi)$, we adopt the mean-field assumption [24] such that $q(S, L, \Phi)$ can be factorized as

$$q(S, L, \Phi) = q(S)q(L)q(\vec{\pi})q(A)q(C)q(\vec{\alpha}) \tag{8}$$

Based on the framework of variational Bayes, the log marginal probability of X can be written as

$$\begin{aligned} \ln p(X) &= \int q(S, L, \Phi)\ln p(X, S, L, \Phi)dSdLd\Phi \\ &\quad - \int q(S, L, \Phi)\ln p(S, L, \Phi|X)dSdLd\Phi \\ &= F(q) + KL(q||p), \end{aligned} \tag{9}$$

where $KL(q||p)$ denotes the KL divergence between the distribution p and the approximating distribution q. $F(q)$ is known as the negative free energy, since the KL divergence is nonnegative, $F(q)$ can be also considered as the lower bound of $\ln p(X)$ and is given by

$$\begin{aligned} F(q) &= \int q(S, L, \Phi)\ln\frac{p(X, S, L, \Phi)}{q(S, L, \Phi)}dSdLd\Phi \\ &= \int q(S)q(L)q(\vec{\pi})q(A)q(C)q(\vec{\alpha}) \\ &\quad \times\Bigg[\ln \pi_{s_1} + \sum_{t=1}^{T-1}\ln a_{s_t s_{t+1}} + \sum_{t=1}^{T}\ln c_{s_t l_t} \\ &\quad + \sum_{t=1}^{T}\ln \mathrm{ID}(\vec{x}_t|\vec{\alpha}_{s_t l_t}) + \ln p(\vec{\pi}) + \ln p(C) \\ &\quad + \ln p(\vec{\alpha}) - \ln q(S) + \ln p(A) - \ln q(L) - \ln q(\vec{\pi}) \\ &\quad - \ln q(A) - \ln q(C) - \ln q(\vec{\alpha})\Bigg]dSdLd\Phi \\ &= F(q(\vec{\pi})) + F(q(A)) + F(q(C)) + F(q(\vec{\alpha})) \\ &\quad + \mathrm{Constant}. \end{aligned} \tag{10}$$

In variational Bayes learning framework, the variational posteriors $q(\pi)$, $q(A)$, $q(C)$, $q(\vec{\alpha})$ can be updated by maximizing the free energy $F(q)$ with respect to each variational posterior, holding the others fixed.

A. OPTIMIZATION OF Q(A), Q($\vec{\pi}$) AND Q(C)

In this subsection, we provide detailed steps for optimizing $q(\pi)$, $q(A)$ and $q(C)$. First, according to Eq. (10), we can have a concrete expression about $F(q(A))$ as

$$\begin{aligned} F(q(A)) &= \int q(A)\sum_{S} q(S)\sum_{t=1}^{T-1}\ln a_{s_t s_{t+1}}dA \\ &\quad + \int q(A)\ln p(A)dA - \int q(A)\ln q(A)dA \\ &= \sum_{S} q(S)\sum_{t=1}^{T-1}\langle \ln a_{s_t s_{t+1}}\rangle_{q(A)} \\ &\quad + \langle \ln p(A)\rangle_{q(A)} - \langle \ln q(A)\rangle_{q(A)}, \end{aligned} \tag{11}$$

where the expectations $\langle \ln p(A)\rangle_{q(A)}$ and $\langle \ln q(A)\rangle_{q(A)}$ are given by

$$\begin{aligned} \langle \ln p(A)\rangle_{q(A)} &= \sum_{i=1}^{N}\{\ln\Gamma(\sum_{j=1}^{N}\phi_{ij}^{A}) + \sum_{j=1}^{N}[(\phi_{ij}^{A}-1)\langle \ln a_{ij}\rangle_{q(A)} \\ &\quad - \ln\Gamma(\phi_{ij}^{A})]\}, \end{aligned} \tag{12}$$

and

$$\begin{aligned} \langle \ln q(A)\rangle_{q(A)} &= \sum_{i=1}^{N}\{\ln\Gamma(\sum_{j=1}^{N}W_{ij}^{A}) + \sum_{j=1}^{N}[(W_{ij}^{A}-1)\langle \ln a_{ij}\rangle_{q(A)} \\ &\quad - \ln\Gamma(W_{ij}^{A})]\}. \end{aligned} \tag{13}$$

In the above equation, W_{ij}^{A} is defined by

$$W_{ij}^{A} = \sum_{t=1}^{T-1}\omega_{ijt}^{A} + \phi_{ij}^{A}. \tag{14}$$

where ω_{ijt}^{A} is defined by

$$\begin{aligned} \omega_{ijt}^{A} &= \sum_{S} q(S)\delta(s_t = i, s_{t+1} = j) \\ &= q(s_t = i, s_{t+1} = j), \end{aligned} \tag{15}$$

where δ represents the transition probability from state i at time t to state j at time $t+1$. The expected value $\langle \ln a_{ij}\rangle_{q(A)}$ can be obtained by

$$\langle \ln a_{ij}\rangle_{q(A)} = \Psi(W_{ij}^{A}) - \Psi\Bigg(\sum_{j=1}^{N}W_{ij}^{A}\Bigg), \tag{16}$$

where $\Psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x)$ is the digamma function.

Then, $F(q(A))$ in (11) can be rewritten as

$$F(q(A)) = \int q(A) \ln \left[\frac{\prod_{i=1}^{N} \prod_{j=1}^{N} a_{ij}^{(W_{ij}^{A}-1)}}{q(A)} \right] dA. \quad (17)$$

Consequently, the optimized variational posterior $q(A)$ can be obtained based on Gibbs inequality and the maximization of $F(q(A))$ as

$$q(A) = \prod_{i=1}^{N} \mathrm{Dir}(a_{i1}, \ldots, a_{iN} | W_{i1}^{A}, \ldots, W_{iN}^{A}). \quad (18)$$

Similarly, the optimized variational posteriors $q(\pi)$ and $q(C)$ can be obtained by maximizing $F(q)$ with respect to $q(\vec{\pi})$ and $q(C)$, respectively as

$$q(\vec{\pi}) = \mathrm{Dir}(\pi_1, \ldots, \pi_N | W_1^{\pi}, ..., W_N^{\pi}), \quad (19)$$

$$q(C) = \prod_{i=1}^{N} \mathrm{Dir}(c_{i1}, ..., c_{iK} | W_{i1}^{C}, \ldots, W_{iK}^{C}), \quad (20)$$

where the associated hyperparameters are given by

$$W_i^{\pi} = \omega_i^{\pi} + \phi_i^{\pi}, \quad (21)$$

$$\omega_i^{\pi} = \sum_{S} q(S)\delta(s_1 = i) = q(s_1 = i), \quad (22)$$

$$W_{ik}^{C} = \sum_{t=1}^{T} \omega_{ikt}^{C} + \phi_{ik}^{C}, \quad (23)$$

$$\omega_{ikt}^{C} = \sum_{S,L} q(S)q(L)\delta(s_t = i, l_t = k) = q(s_t = i, l_t = k). \quad (24)$$

In our case, the forward-backward algorithm as discussed in [2] is used to calculate ω_{ijt}^{A}, ω_i^{π} and ω_{ikt}^{C}.

B. OPTIMIZATION OF Q($\vec{\alpha}$)

The optimization of $q(\vec{\alpha})$ is carried out by maximizing $F(q)$. Based on (10), $F(q(\vec{\alpha}))$ is given by

$$\begin{aligned} F(q(\vec{\alpha})) &= \int q(\vec{\alpha}) \sum_{S,L} q(S)q(L) \sum_{t=1}^{T} \ln \mathrm{ID}(\vec{x}_t | \vec{\alpha}_{s_t,l_t}) d\vec{\alpha} \\ &\quad + \int q(\vec{\alpha}) \ln p(\vec{\alpha}) d\vec{\alpha} - \int q(\vec{\alpha}) \ln q(\vec{\alpha}) d\vec{\alpha} \\ &= \sum_{S,L} q(S,L) \sum_{t=1}^{T} \langle \ln \mathrm{ID}(\vec{x}_t | \vec{\alpha}_{s_t,l_t}) \rangle_{q(\vec{\alpha})} \\ &\quad + \langle \ln p(\vec{\alpha}) \rangle_{q(\vec{\alpha})} - \langle \ln q(\vec{\alpha}) \rangle_{q(\vec{\alpha})}, \end{aligned} \quad (25)$$

According to (10), $F(q(\vec{\alpha}))$ can be further represented by

$$\begin{aligned} F(q(\vec{\alpha})) = \int q(\vec{\alpha}) \\ \times \ln \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} p(\vec{\alpha}_{ik}) \Pi_{t=1}^{T} \mathrm{ID}(\vec{x}_t | \vec{\alpha}_{ik})^{\omega_{ikt}^{C}}}{q(\vec{\alpha})} d\vec{\alpha}. \end{aligned} \quad (26)$$

Then, by maximizing $F(q)$ with respect to $q(\vec{\alpha})$, the optimized variational posterior is derived as

$$q(\vec{\alpha}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \mathrm{Gam}(\alpha_{ikd} | u_{ikd}^{*}, v_{ikd}^{*}), \quad (27)$$

where the hyperparameters u_{ikd}^{*} and v_{ikd}^{*} are given by

$$\begin{aligned} u_{ikd}^{*} &= u_{ikd} + \sum_{t=1}^{T} \omega_{ikt}^{C} \bar{\alpha}_{ikd} \Bigg[\Psi(\sum_{d=1}^{D+1} \bar{\alpha}_{ikd}) - \Psi(\bar{\alpha}_{ikd}) \\ &\quad + \sum_{s \neq d}^{D+1} \bar{\alpha}_{iks} \Psi'(\sum_{d=1}^{D+1} \bar{\alpha}_{ikd})(\langle \ln \alpha_{iks} \rangle - \ln \bar{\alpha}_{iks}) \Bigg] \end{aligned} \quad (28)$$

$$v_{ikd}^{*} = v_{ikd} - \sum_{t=1}^{T} \omega_{ikt}^{C} [\ln x_{td} - \ln(1 + \sum_{d=1}^{D} x_{td})] \quad (29)$$

where the expected values in (28) can be obtained by

$$\bar{\alpha}_{ikd} = \langle \alpha_{ikd} \rangle = \frac{u_{ikd}^{*}}{v_{ikd}^{*}}, \quad (30)$$

$$\langle \ln \alpha_{ikd} \rangle = \Psi(u_{ikd}^{*}) - \ln v_{ikd}^{*}. \quad (31)$$

C. OPTIMIZATION OF Q(S, L)

In this part, we provide the optimization of the joint variational posterior $q(S, L)$ over the state indicator S and the mixture component indicator L. By collecting all the quantities related to $q(S, L)$ from (10), $F(q(S, L))$ can be written as

$$\begin{aligned} F(q(S,L)) = \sum_{S} q(S) \int q(\vec{\pi}) \ln \pi_{s_1} d\pi \\ + \sum_{S} q(S) \int q(A) \sum_{t=1}^{T-1} \ln a_{s_t s_{t+1}} dA \\ + \sum_{S,L} q(S,L) \int q(C) \sum_{t=1}^{T} \ln c_{s_t l_t} dC \\ + \sum_{S,L} q(S,L) \int q(\theta) \sum_{t=1}^{T} \ln \mathrm{ID}(\vec{x}_t | \vec{\alpha}_{s_t l_t}) d\vec{\alpha} \\ - \sum_{S,L} q(S,L) \ln q(S,L) \end{aligned} \quad (32)$$

We can further represent $F(q(S, L))$ by

$$\begin{aligned} F(q(S,L)) = \sum_{S,L} q(S,L) \\ \times \ln \frac{\pi_{s_1}^{*} \prod_{t=1}^{T-1} a_{s_t s_{t+1}}^{*} \prod_{t=1}^{T} c_{s_t l_t}^{*} \mathrm{ID}^{*}(\vec{x}_t | \vec{\alpha}_{s_t l_t})}{q(S,L)}. \end{aligned} \quad (33)$$

where we have

$$\pi_i^{*} = \exp\{\int q(\vec{\pi}) \ln \pi_i d\pi\} = \exp\{\Psi(W_i^{\pi}) - \Psi(\sum_{i=1}^{N} W_i^{\pi})\}, \quad (34)$$

$$a_{ij}^{*} = \exp\{\int q(A) \ln a_{ij} dA\} = \exp\{\Psi(W_{ij}^{A}) - \Psi(\sum_{j=1}^{N} W_{ij}^{A})\}, \quad (35)$$

$$c_{ik}^{*} = \exp\{\int q(C) \ln c_{ik} dC\} = \exp\{\Psi(W_{ik}^{C}) - \Psi(\sum_{k=1}^{K} W_{ik}^{C})\}. \quad (36)$$

$$\begin{aligned}
&\ln \mathrm{ID}^*(\vec{x}_t|\vec{\alpha}_{s_t l_t}) \\
&\quad = \langle \ln \mathrm{ID}(\vec{x}_t|\vec{\alpha}_{s_t l_t})\rangle_{q(\vec{\alpha})} \\
&\quad = R_{ikd} + \sum_{d=1}^{D}(\bar{\alpha}_{ikd}-1)\ln x_{td} \\
&\qquad - \sum_{d=1}^{D+1}\bar{\alpha}_{ikd}\ln\left(1+\sum_{d=1}^{D}x_{td}\right), \qquad (37)
\end{aligned}$$

$$\begin{aligned}
R_{ikd} &= \ln\frac{\Gamma(\sum_{d=1}^{D+1}\bar{\alpha}_{ikd})}{\Pi_{d=1}^{D+1}\Gamma(\bar{\alpha}_{ikd})} + \sum_{d=1}^{D+1}\bar{\alpha}_{ikd}[\Psi(\sum_{d=1}^{D+1}\bar{\alpha}_{ikd}) \\
&\quad -\Psi(\bar{\alpha}_{ikd})][\langle\ln\alpha_{ikd}\rangle - \ln\bar{\alpha}_{ikd}] \\
&\quad +\frac{1}{2}\sum_{d=1}^{D+1}\bar{\alpha}_{ikd}^2[\Psi'(\sum_{d=1}^{D+1}\bar{\alpha}_{ikd}) - \Psi'(\bar{\alpha}_{ikd})] \\
&\quad \times\langle(\ln\alpha_{ikd} - \ln\bar{\alpha}_{ikd})^2\rangle \\
&\quad +\frac{1}{2}\sum_{h=1}^{D+1}\sum_{g=1,h\neq g}^{D+1}\bar{\alpha}_{ikh}\bar{\alpha}_{ikg}\{\Psi'(\sum_{d=1}^{D+1}\bar{\alpha}_{ikd}) \\
&\quad \times(\langle\ln\alpha_{ikh}\rangle - \ln\bar{\alpha}_{ikh})(\langle\ln\alpha_{ikg}\rangle - \ln\bar{\alpha}_{ikg})\}, \qquad (38)
\end{aligned}$$

$$\begin{aligned}
&\langle(\ln\alpha_{ikd} - \ln\bar{\alpha}_{ikd})^2\rangle \\
&\quad = [\Psi(u_{ikd}^*) - \ln u_{ikd}^*]^2 + \Psi'(u_{ikd}^*). \qquad (39)
\end{aligned}$$

Then, the optimized variational posterior $q(S, L)$ can be obtained by

$$q(S,L) = \frac{1}{Z}\pi_{s_1}^*\prod_{t=1}^{T-1}a_{s_t s_{t+1}}^*\prod_{t=1}^{T}c_{s_t,l_t}^*\mathrm{ID}^*(\vec{x}_t|\vec{\alpha}_{s_t l_t}), \qquad (40)$$

where the normalizing constant Z is given by

$$\begin{aligned}
Z &= \sum_{S,L}\pi_{s_1}^*\prod_{t=1}^{T-1}a_{s_t s_{t+1}}^*\prod_{t=1}^{T}c_{s_t,l_t}^*\mathrm{ID}^*(\vec{x}_t|\vec{\alpha}_{s_t l_t}) \\
&= q(X|\Phi^*). \qquad (41)
\end{aligned}$$

As we compare (41) with (3), we may notice that Z is the approximating likelihood of the optimized model parameters Φ^*, which can be calculated easily according to the forward-backward algorithm [2].

The complete algorithm for learning the proposed HMM based on variational Bayes inference is given in Algorithm 1.

D. CONVERGENCE

Since variational Bayes is a generalization of the conventional EM algorithm, the convergence of the proposed variational Bayes learning algorithm can be systematically monitored according to the inspection of the negative free energy $F(q)$ in Eq. (10). Due to the convexity property of the variational posterior distributions, at each iteration of the re-estimating step, the value of the negative free energy is never decreased until it converges to a local maximum.

IV. EXPERIMENTS

In this section, we validate the proposed ID-based HMM, which is referred to as iDHMM, through both synthetic data sets and a real-world application namely network intrusion detection. In both experiments, we set the initial value of the number of mixture components K in the iDHMM to 10. The initial values of the hyperparameters u and v are set to 1 and 0.01, respectively. All the experiments are conducted with 15 repetitions, and the averaged results are reported. Our experiments were performed on MATLAB based on the Windows platform.

Algorithm 1 Variational Bayes Learning of ID-Based HMM

```
## Initialize Hyperparameters ##
$\phi^\pi = [1/N, ..., 1/N]$
$\phi^A = [1/N, ..., 1/N]$
$\phi^C = [1/K, ..., 1/K]$
$v_{ikd} = 0.01, \forall i, k, d$
$u_{ikd} = 1, \forall i, k, d$
Initialize $\alpha$ as: $\alpha_{init} = u/v$
## Initialize parameters of the HMM ##
Initialize responsibilities $\omega^\pi$, $\omega^A$, $\omega^C$ from prior distributions with Eqs. (4), (5) and (6)
Compute $W^\pi$, $W^A$ and $W^C$ with Eqs. (21), (14), (23)
Initialize $\pi$, $A$ and $C$ with Eqs. (34)~(36)
repeat
    ## Variational E-step ##
    Forward: Compute the responsibilities $\omega^\pi$, $\omega^A$ and $\omega^C$ using $\pi$, $A$ and $C$ with Eqs. (15), (22), and (24)
    Update $u$ and $v$ with Eqs. (28) and (29)
    ## Variational M-step ##
    Update $W^\pi$, $W^A$ and $W^C$ using responsibilities $\omega^\pi$, $\omega^A$ and $\omega^C$ with Eqs. (21), (14), (23)
    Update $\pi$, $A$ and $C$ using $W^\pi$, $W^A$ and $W^C$ with Eqs. (34)~(36)
    Backward: Compute the approximating likelihood $Z$ using Eq. (41)
until Convergence is reached
```

A. SYNTHETIC DATA

In the experiment with synthetic data, our goal is to test the effectiveness and accuracy of using variational Bayes for learning the proposed iDHMM. The two simulated data sets are randomly generated as follows. The first synthetic data set (denoted by Data set 1) contains 2 states, where each of the state has 200 data points. A time series of multivariate observations is then formulated based on these data. Specifically, at state 1, the observations at time instances $t = 1 : 200$ are generated according to a mixture of 2 ID distributions $\mathrm{ID}(\vec{x}_t|\vec{\alpha}_{i=1,k=1})$ and $\mathrm{ID}(\vec{x}_t|\vec{\alpha}_{i=1,k=2})$. At state 2, the observations at time instances $t = 200 : 400$ come from another mixture of 2 ID distributions $\mathrm{ID}(\vec{x}_t|\vec{\alpha}_{i=2,k=1})$ and $\mathrm{ID}(\vec{x}_t|\vec{\alpha}_{2=1,k=2})$. The second synthetic data set (denoted by Data set 2) includes 1,600 data points in total with also 2 states, where each state contains 400 instances. The corresponding time series of multivariate observations is formulated as follows. The observations at time instances $t = 1 : 800$ are generated according to a mixture of 2 ID distributions

TABLE 1. The true parameters for generating the synthetic data sets.

		n_k	k	α_1	α_2	α_3	c_k
Data set 1	state 1	100	1	2	10	3	0.5
		100	2	6	15	20	0.5
	state 2	100	1	9	25	12	0.5
		100	2	16	33	11	0.5
Data set 2	state1	400	1	9	15	37	0.5
		400	2	28	2	7	0.5
	state2	400	1	12	30	45	0.5
		400	2	32	50	16	0.5

TABLE 2. The estimated parameters of the synthetic data sets by variational Bayes.

		n_k	k	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{c}_k$
Data set 1	state 1	100	1	2.02	9.80	3.48	0.502
		100	2	6.09	15.79	20.18	0.498
	state 2	100	1	10.22	26.83	13.06	0.508
		100	2	15.80	34.98	10.37	0.492
Data set 2	state1	400	1	8.81	14.58	35.66	0.509
		400	2	29.22	2.17	7.30	0.491
	state2	400	1	12.78	31.40	47.05	0.500
		400	2	32.29	51.01	16.16	0.500

TABLE 3. The estimated initial state probabilities and state transition matrix of each synthetic data set by variational Bayes.

	Initial state probabilities	State transition matrix
Data set 1	[1,0]	$\begin{bmatrix} 0.993 & 0.007 \\ 0.082 & 0.918 \end{bmatrix}$
Data set 2	[0.999,0.001]	$\begin{bmatrix} 0.959 & 0.041 \\ 0.097 & 0.903 \end{bmatrix}$

with different parameters at state 1. At state 2, the observations at time instances $t = 800 : 1600$ are drawn from another mixture of 2 ID distributions. The detailed parameters that were used to generate the two synthetic data sets are given in Table 1.

The parameters of these two synthetic data sets are estimated using variational Bayes as described in Section III. The results of the estimated parameters and the estimated initial state probabilities and state transition matrix of each synthetic data set are demonstrated in Tables 2 and 3. According to these two tables, it is clear that the estimated results obtained by using variational Bayes is very close to the true ones (as shown in Table 1) that were used to generated the data sets, which indicates that the developed variational Bayes approach is able to accurately learn the iDHMM.

It is noteworthy that our variational Bayes learning approach is able to correctly detect the number of mixture models in the synthetic data sets. This is done through the estimation of the mixing coefficients in each state. Although the mixture model in each state was initialized with 10 mixture components, by removing those ones with mixing coefficients that are close to 0 (i.e., than 10^{-5} in our case less) after the convergence, we obtain the optimal number of components for each mixture model with the estimated values of mixing coefficients $\{\hat{c}_k\}$ are shown in Table 2. We can observe that the estimated mixture coefficients $\{\hat{c}_k\}$ in Table 2 are close to those true values $\{c_k\}$ in Table 1. Thus, we can conclude that the mixture coefficients can be accurately estimated through the proposed variational Bayes learning approach.

B. NETWORK INTRUSION DETECTION

During the last decade, a serious risk to the internet and computer networks has been brought by various security threats. An effective approach to detecting different types of attacks is through Intrusion Detection Systems (IDSs). In this experiment, we develop an unsupervised intrusion detection approach to detecting network-based attacks based on the proposed iDHMM. In our approach, the iDHMM with variational Bayes is used to model patterns of normal and intrusive activities.

To validate our intrusion detection approach, we conduct experiments on subsets of the well-known KDD Cup 1999 data set [25]. The original KDD Cup 1999 data set contains about 5 million connection records that was gathered at MIT Lincoln laboratory for the 1998 DARPA intrusion detection evaluation program by simulating attacks on a typical U.S. Air Force LAN. Each data instance of this data set represents an attack or a normal connection that were obtained from the simulated intrusions with 41 features (such as duration, dst bytes, etc.), and each connection denotes a sequence of TCP packets that were acquired from a source IP address to a target IP address under some well-defined protocol. Since the KDD Cup 1999 data set contains connection records that were collected from seven weeks of network traffic, it is considered as time series data in our case, and therefore the HMM is an appropriate choice in this application. In our experiments, we removed the discrete and sparse features (such as protocol_type, service, land, logged_in, etc.) in the data preprocessing stage, and adopted 24 continuous feature attributes.

Five classes of connections are included in the data sets. Except for the **Normal** connections, there are four different attack types including **DOS**: denial-of-service (e.g. syn flood); **R2L**: unauthorized access from a remote machine (e.g. guessing password); **U2R**: unauthorized access to local superuser (root) privileges (e.g. buffer overflow attack) and **Probing**: surveillance and other probing (e.g. port scanning). Two subsets of the KDD Cup 1999 data set were used to test our intrusion detection approach. Both of the two subsets of the KDD Cup 1999 data set were randomly sampled from the original KDD Cup 1999 data set. In Data set 1, the training set contains 7050 connection records, whereas the test set includes 20000 records. The training set of Data set

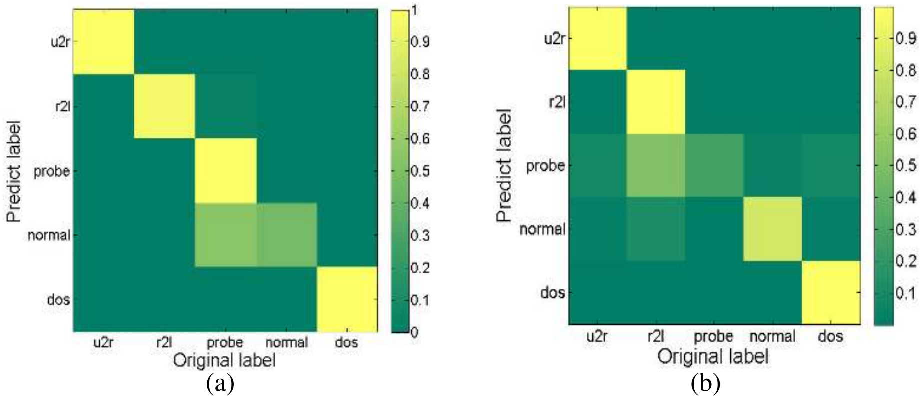

FIGURE 2. The confusion matrices obtained by DHMM on the KDD Cup 1999 data set. (a) KDD Data 1; (b) KDD Data 2.

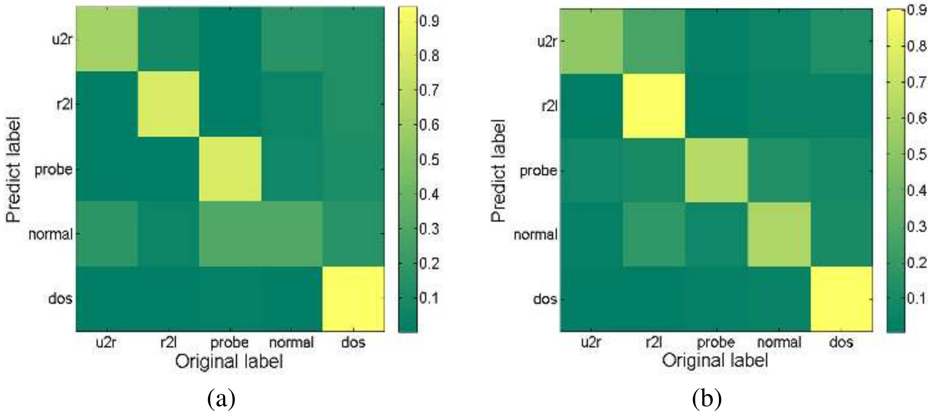

FIGURE 3. The confusion matrices obtained by GHMM on the KDD Cup 1999 data set. (a) KDD Data 1; (b) KDD Data 2.

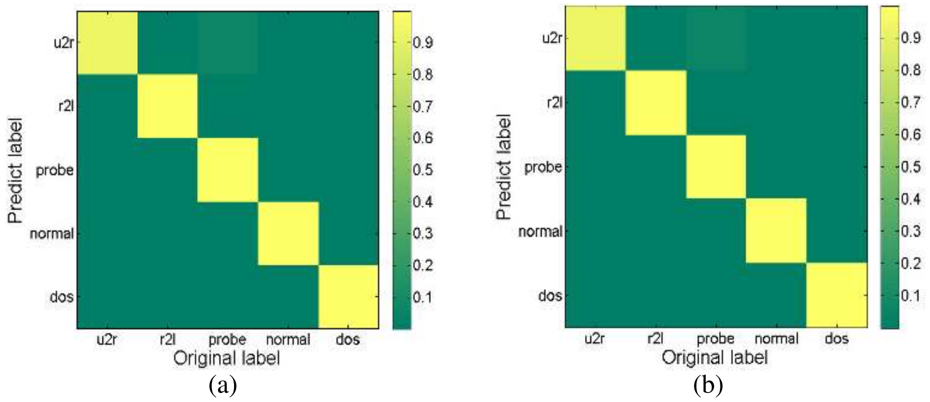

FIGURE 4. The confusion matrices obtained by iDHMM on the KDD Cup 1999 data set. (a) KDD Data 1; (b) KDD Data 2.

TABLE 4. The description of the data sets used in our experiments.

Data set	*Training set*	*Test set*
Data 1	7050	20000
Data 2	19178	100000

2 has 19178 connection records, whereas the test set contains 100000 records. The detailed information regarding our data sets are shown in Table 4.

To demonstrate the advantages of the proposed unsupervised intrusion detection approach, we compare it with two other well-defined HMMs with mixture model that are considered as their emission densities: the HMM with Gaussian mixture model [22] (which is referred to as GHMM) and the HMM with Dirichlet mixture models [21] (which is referred to as DHMM). Both of these two tested HMMs are learnt using variational Bayes. For the GHMM, we use the same settings of the parameters as described in the original paper.

TABLE 5. Classification accuracy of KDD data by different approaches.

Method	*KDD Data 1*	*KDD Data 2*
iDHMM	99.7400%	99.9480%
DHMM	93.2150%	95.8930%
GHMM	89.9255%	95.8051%

For the DHMM, the initial value of the mixing component K is set to 10, the initial values of the hyperparameters u and v are set to 1 and 0.01, respectively. It is also noteworthy that since our approach is unsupervised, class labels of data instances are not included in the training process, but are used for evaluating the accuracy of each approach.

The performance of different intrusion detection approaches on the two tested data set are illustrated in Table 5. Based on the results shown in this table, the proposed iDHMM is able to outperform both DHMM and GHMM for the two data sets, which demonstrated the advantages of using the ID-based HMM for intrusion detection. These results also verify that the ID-based HMM is more suitable than the Dirichlet-based or Gaussian-based HMMs for modeling positive data. Figures 2∼4 demonstrate the confusion matrices obtained by different approaches on the KDD Cup 1999 data set.

V. CONCLUSION

In this work, a variant of the continuous hidden Markov model was proposed for modeling positive sequential data. In contrast with conventional HMMs which often use Gaussian distributions or Gaussian mixture models as the emission probability density, we adopted the ID mixture model as the emission density in our HMM model. This choice was motivated by the better performance of ID mixture model than Gaussian mixture models for modeling positive data. In addition, a convergence-guaranteed approach was developed to learn the proposed ID-based HMM through variational Bayes inference. The effectiveness of the proposed HMM model was verified through both synthetic data sets and a challenging real-world application regarding network intrusion detection. One limitation of the proposed ID-based HMM is that the number of states have to be set manually. In order to tackle this issue, one possible future work is to extend the current HMM with hierarchical Bayesian nonparametric frameworks (such as the hierarchical Dirichlet process [26]), such that the number of states of the HMM will be infinite initially and will be inferred automatically during the learning process.

REFERENCES

[1] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 267–296, Feb. 1989.

[3] P. P. Roy, A. K. Bhunia, A. Das, P. Dey, and U. Pal, "HMM-based indic handwritten word recognition using zone segmentation," *Pattern Recognit.*, vol. 60, pp. 1057–1075, Dec. 2016.

[4] R. Ghosh and P. P. Roy, "Comparison of zone-features for online Bengali and Devanagari word recognition using HMM," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit.*, Oct. 2016, pp. 435–440.

[5] M. Ramalingam and N. A. M. Isa, "Fast retrieval of hidden data using enhanced hidden Markov model in video steganography," *Appl. Soft Comput.*, vol. 34, pp. 744–757, Sep. 2015.

[6] S. Kayte, M. Mundada, and J. Gujrathi, "Hidden Markov model based speech synthesis: A review," *Int. J. Comput. Appl.*, vol. 130, no. 3, pp. 35–39, 2015.

[7] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[8] J. Novoa, J. Fredes, V. Poblete, and N. B. Yoma, "Uncertainty weighting and propagation in DNN–HMM-based speech recognition," *Comput. Speech Lang.*, vol. 47, pp. 30–46, Jan. 2018.

[9] F. Forbes and N. Peyrard, "Hidden Markov random field model selection criteria based on mean field-like approximations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1089–1101, Sep. 2003.

[10] S. P. Chatzis and D. I. Kosmopoulos, "A variational Bayesian methodology for hidden Markov models utilizing student's-t mixtures," *Pattern Recognit.*, vol. 44, no. 2, pp. 295–306, 2011.

[11] W. Fan, N. Bouguila, J.-X. Du, and X. Liu, "Axially symmetric data clustering through Dirichlet process mixture models of Watson distributions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1683–1694, Jun. 2019.

[12] W. Fan, H. Sallay, and N. Bouguila, "Online learning of hierarchical Pitman–Yor process mixture of generalized Dirichlet distributions with feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 9, pp. 2048–2061, Sep. 2017.

[13] T. Bdiri and N. Bouguila, "Positive vectors clustering using inverted Dirichlet finite mixture models," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1869–1882, 2012.

[14] T. Bdiri and N. Bouguila, "Bayesian learning of inverted Dirichlet mixtures for SVM kernels generation," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1443–1458, 2013.

[15] W. Fan and N. Bouguila, "Nonparametric hierarchical Bayesian models for positive data clustering based on inverted Dirichlet-based distributions," *IEEE Access*, vol. 7, pp. 83600–83614, 2019.

[16] W. Fan, C. Hu, J. Du, and N. Bouguila, "A novel model-based approach for medical image segmentation using spatially constrained inverted Dirichlet mixture models," *Neural Process. Lett.*, vol. 47, no. 2, pp. 619–639, 2018.

[17] G. G. Tiao and I. Cuttman, "The inverted Dirichlet distribution with applications," *J. Amer. Stat. Assoc.*, vol. 60, no. 311, pp. 793–805, 1965.

[18] K. N. Shu, T. Krishnan, and G. J. Mclachlan, "The EM algorithm," in *Handbook of Computational Statistics: Concepts and Methods*. Berlin, Germany: Springer, 2012, pp. 139–172.

[19] M. Oudelha and R. N. Ainon, "HMM parameters estimation using hybrid Baum–Welch genetic algorithm," in *Proc. Int. Symp. Inf. Technol.*, vol. 2, Jun. 2010, pp. 542–545.

[20] S. Cheshomi, S. Rahati-Q, and M.-R. Akbarzadeht-T, "Hybrid of chaos optimization and Baum–Welch algorithms for HMM training in continuous speech recognition," in *Proc. Int. Conf. Intell. Control Inf. Process.*, Aug. 2010, pp. 83–87.

[21] E. Epaillard and N. Bouguila, "Variational Bayesian learning of generalized Dirichlet-based hidden Markov models applied to unusual events detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1034–1047, Apr. 2019.

[22] S. Ji, B. Krishnapuram, and L. Carin, "Variational Bayes for continuous hidden Markov models and its application to active learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 522–532, Apr. 2006.

[23] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[24] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 324–335, 2010.

[25] P. Aggarwal and S. K. Sharma, "Analysis of KDD dataset attributes—Class wise for intrusion detection," *Procedia Comput. Sci.*, vol. 57, pp. 842–851, Jan. 2015.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

RU WANG received the B.E. degree from the Department of Computer Science and Technology, Huainan Normal University, China, in 2018. She is currently pursuing the M.Sc. degree with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. Her research areas include machine learning and mixture models.

WENTAO FAN received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Concordia University, Montreal, QC, Canada, in 2009 and 2014, respectively. He is currently an Associate Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. His research interests include machine learning, computer vision, and pattern recognition.