FETAL HEART RATE CLASSIFICATION BY NON-PARAMETRIC BAYESIAN METHODS

Kezi Yu, J. Gerald Quirk,** and Petar M. Djurić**

*Department of Electrical and Computer Engineering
**Department of Obstetrics/Gynecology, Stony Brook University Hospital
Stony Brook University
Stony Brook, NY 11794, USA
Email: {kezi.yu,petar.djuric}@stonybrook.edu, J.Gerald.Quirk@stonybrookmedicine.edu

ABSTRACT

In this paper, we propose an application of non-parametric Bayesian (NPB) models to classification of fetal heart rate recordings. More specifically, the models are used to discriminate between fetal heart rate recordings that belong to fetuses that may have adverse asphyxia outcomes and those that are considered normal. In our work we rely on models based on hierarchical Dirichlet processes. Two mixture models were inferred from recordings that represent healthy and unhealthy fetuses, respectively. The models were then used to classify new recordings. We compared the classification performance of the NPB models with that of support vector machines on real data and concluded that the NPB models achieved better performance.

Index Terms— Hierarchical Dirichlet process, fetal heart rate, Gaussian mixture models, classification, non-parametric

1. INTRODUCTION

Fetal heart rate (FHR) is routinely monitored during pregnancy to help obstetricians examine fetal health. The main interest is in determining if the heart rate of a fetus points to an inadequate fetal oxygenation during labor. Adverse asphyxia outcomes could be prevented by taking appropriate intervention according to FHR patterns [1]. In clinical practice, FHR signals are evaluated visually by physicians following guidelines published by various institutions such as the National Institute of Child Health and Human Development (NICHD) and the International Federation of Gynecology and Obstetrics (FIGO) [2, 3]. However, a recent study has shown that evaluations of fetal acidosis made by obstetricians have large inter- and intra-variability [4].

With the goal of avoiding subjectivity and maintaining consistency, researchers in signal processing and biomedical engineering have been proposing various automated FHR classification methods. In [5], the authors implemented the naïve Bayes, support vector machines (SVMs) and decision trees as classification algorithms. In [6], artificial neural networks (ANNs) were applied to FHR analysis, with 6 FHR features and 6 clinical parameters as input. A more comprehensive study [7] explored the performance of classical machine learning (ML) classification methods including linear regression and SVMs with different kernels and in combination with feature selection and reduction methods such as random forest (RF) and principle component analysis (PCA).

Despite the vast use and successful application of SVMs and ANNs in many areas, some newly proposed ML techniques have proved to be more flexible and robust. Hierarchical Dirichlet process (HDP) mixture models [8], for instance, free the classical mixture models from a fixed number of mixing components, and they provide a way of processing grouped data jointly. An example of this kind of problem can be described as follows. Consider a collection of newspaper articles. Each article (one group of data) is comprised of certain number of words, which arise from different topics (e.g., economy, politics, sports). Furthermore, let the topics be shared within the whole corpus and let the number of topics in each article vary. The goal is to find all the topics in the corpus and their appearances in the articles, and label each article according to a predefined criterion. In our context, the articles are the FHR recordings, words are segments of the recordings, topics are the components of statistical mixtures of the segments, and labels are defined as *healthy* and *unhealthy*.

In this paper, we extend the application of HDP mixture models to FHR classification. We employ HDP Gaussian mixtures (HDPGMs) to model the collection of FHR tracings of *healthy* and *unhealthy* fetuses, respectively, and infer the structure of the models from real data. Once the models of the two classes are constructed, we use them for classification of new FHR tracings.

The main contribution of this paper lies in the novelty of applying non-parametric Bayesian models to FHR classification. A preliminary analysis of FHR tracings using HDPGMs was presented in [9]. Here, we extend this work to more prac-

This work has been supported by NIH under Award 1R21HD080025-01A1.

tical tasks of classification. In the paper we also show that the proposed methodology has a high potential. Comparison with an SVM method shows better performance of the HDPGMs.

The paper is organized as follows. In the next section, we provide a brief background on the data and then describe in detail the methodology. In Section 3, we present the specifics of our experimental settings. The results of the new method and their comparison with those of an SVM method are provided in Section 4. Finally, we conclude the paper with Section 5.

2. BACKGROUND

2.1. Notation

In the problems of our interest, the observations are organized into groups and assumed exchangeable both within each group and across groups. Specifically, let x_{ji} denote the i-th observation in the j-th group. Then we assume that $x_{j1}, x_{j2}, \dots$ are exchangeable within group j. Furthermore, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ are assumed exchangeable at the group level where $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \dots)$.

We consider that each observation is drawn independently from a mixture model. The symbol θ_{ji} represents the parameters of the mixture component that generated the observation x_{ji}. Finally, $F(\theta_{ji})$ denotes the distribution of x_{ji}.

2.2. Method

The hierarchical Dirichlet process (HDP) mixture model is a non-parametric Bayesian approach to data processing. It was designed to model grouped data where each group is associated with a mixture model and all the mixtures are linked through a hierarchy [8]. Here, we explain the stick-breaking construction of the HDP and its extension to mixture models.

An HDP is a distribution over a set of random probability measures G_j, one for each group. The probability measures G_j have the same base measure, which is denoted by G_0. This measure is the global random probability measure of the hierarchical process. The global measure G_0 is distributed as a Dirichlet process,

$$G_0|\gamma, H \sim \mathrm{DP}(\gamma, H), \tag{1}$$

where γ is the concentration parameter of the process, and H is its base measure. The global measure can be expressed using a stick-breaking representation by

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \tag{2}$$

where δ_{ϕ_k} is a probability measure concentrated at ϕ_k, and $\phi_k \sim H$ independently. The infinite sequence of weights $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ is distributed according to a GEM (Griffiths-Engen-McCloskey) distribution, i.e., $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$. The weights are mutually independent and $\sum_{k=1}^{\infty} \beta_k = 1$. Details on the GEM distribution can be found in [10].

The random measures G_j are conditionally independent given G_0, with distributions

$$G_j|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0), \tag{3}$$

where α is another concentration parameter, and G_0 is the base probability measure. Because G_0 has support given by the points $\boldsymbol{\phi} = (\phi_k)_{k=1}^{\infty}$, each G_j necessarily has the same support as the base measure. Thus, the stick-breaking representation can be written as

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \tag{4}$$

where $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$ and $\sum_{k=1}^{\infty} \pi_{jk} = 1$. Intuitively, one can think of $\boldsymbol{\pi}_j$ being a modified set of weight coefficients of $\boldsymbol{\beta}$ governed by the parameter α.

As for the HDP mixture model, within each group, the mixture components are distributed according to G_j, thus θ_{ji} takes on the value ϕ_k with probability π_{jk}. Let z_{ji} denote the mixture component assignment of x_{ji} such that $\theta_{ji} = \phi_{z_{ji}}$. Given z_{ji}, the distribution of x_{ji} is $F(\phi_{z_{ji}})$. Therefore, the representation of the HDP mixture model can be expressed as

$$\begin{aligned} &\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma), \quad \phi_k \sim H, \\ &\boldsymbol{\pi}_j|\alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \\ &z_{ji}|\boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j, \\ &x_{ji}|z_{ji}, \phi_k \sim F(\phi_{z_{ji}}). \end{aligned} \tag{5}$$

In our experiments, we assume that F is a multi-variate Gaussian distribution. This completes the description of the HDP Gaussian mixture models.

2.3. Database

In our work, we used the open-access CTU-UHB intrapartum cardiotocography (CTG) database [11]. This database contains 552 CTG recordings, each comprising an FHR time tracing and a uterine contraction (UC) signal, both sampled at 4 Hz. Additional information, including maternal data, delivery information, fetal data and fetal outcome data are also available. More details on the data collection can be found in [12].

Real-world FHR tracings inevitably suffer from artifacts and errors. Therefore, preprocessing the data before analysis is necessary. In this paper, we followed the same preprocessing procedures as in our previous work [9]. Samples that were considered to be artifacts were either substituted by linear interpolation or simply discarded. Figure 1 shows an example of an FHR series before and after preprocessing.

3. EXPERIMENTAL SETTINGS

The basic idea of our approach to classification using HDP mixture models is as follows. During a learning stage, two

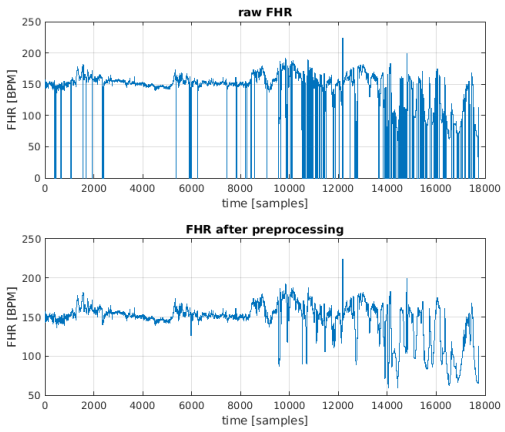

Fig. 1. Comparison between an FHR signal before and after preprocessing.

HDP Gaussian mixture models, $\mathcal{M}_0$ and $\mathcal{M}_1$, were constructed using FHR recordings from *healthy* and *unhealthy* fetuses, respectively (thus, the subscripts '0' and '1' refer to models, distributions, and parameters of *healthy* and *unhealthy* fetuses, respectively). We implemented collapsed Gibbs samplers (proposed in [8]) to infer the parameters of $\mathcal{M}_0$ and $\mathcal{M}_1$. Given a new FHR tracing $\boldsymbol{x}_j$, the decision is made by comparing the likelihoods $p_0 = f(\boldsymbol{x}_j|\mathcal{M}_0)$ and $p_1 = f(\boldsymbol{x}_j|\mathcal{M}_1)$. If $p_0 > p_1$, the FHR series is classified as healthy and vice versa. We note that here we tacitly assume that the priors of the fetuses were equal.

3.1. Dataset

We selected a balanced dataset from the database by the following criteria: an FHR recording was labeled as *unhealthy* if its associated umbilical cord blood pH was less than or equal to a threshold τ_1, and labeled as *healthy* if the pH value was greater than τ_0. There is no consensus of the choice of τ_1, and thus we experimented with both $\tau_1 = 7.05$ as in [5] and $\tau_1 = 7.1$ as in [6, 7]. The number of FHR recordings N in the datasets ended up with 88 and 122 respectively.

In our experiment, only the last 30-minute data of the FHR series were used for analysis. Each recording was divided into non-overlapping segments of l seconds, where l ranged from 10 to 30. Thus, the number of segments in each series was m, where $m = 1800/l$. A feature vector x_{ji} of dimension d was extracted from the ith segment in the jth recording, denoted as $s_{ji}(k), k = 1, \ldots, K$.

3.2. Feature extraction

There is a large body of work on feature extraction from FHR signals. Considering the nature of our model, we selected 14 features in our experiment. The features can be divided into two categories: time domain and frequency domain features.

The time-domain features include the basic statistics of the FHR samples: the mean and standard deviation of $\boldsymbol{s}_{ji} = (s_{ji}(1), \ldots, s_{ji}(K))$. This category also includes standard short-term variability (STV), long-term variability (LTI) defined in [13], short-term irregularity (STI), and long-term irregularity (LTI) in [14]. Finally, on the list of time domain features we also had two standard descriptors of Poincaré plot, SD1 and SD2, as well as complex correlation measure (CCM), proposed in [15]. The frequency-domain features were composed of energies in four frequency bands: very low frequency (VLF: 0-0.06 Hz), low frequency (LF: 0.06-0.3 Hz), medium frequency (MF: 0.3-1 Hz) and high frequency (LF: 1-2 Hz), and the ratio of energies defined by LF/(MF+HF). The frequency-domain features represent the underlying physiological activity of either the mother or the fetus. It is worth noting that there was no consensus on the division of the frequency bands. In our experiments, we used the ranges from [16]. The complete list of features is shown in Table 1.

Table 1. List of all the features.

Category	Feature
Time	Mean, standard deviation STV, STI, LTV, LTI, SD1, SD2, CCM
Frequency	VLF, LF, MF, HF, ratio

3.3. Dimensionality reduction

Before training the model, we reduced the dimension of the feature space from 14 to q by way of principle component analysis (PCA). Since the ranges of values in each dimension varied largely, we first scaled these values to lie in the interval $(-1, 1)$. An example of a PCA result of all the data when the number of recordings is $N = 88$ and the number of segments per recording is $m = 180$ is shown in Fig. 2. The gray bars are the explained variance ratios of each principle component, and the blue line corresponds to the cumulative variance ratio. According to the preliminary analysis of all the data, we observed that in most cases, the first four principle components would retain 95% variance. Thus we chose $q = 2, 3$, and 4.

3.4. Model priors

An HDP mixture model has two concentration parameters, γ and α, as described in Section 2.2. In [8], the authors provided an auxiliary posterior sampling scheme for γ and α. In our experiments, the concentration parameters were given gamma priors, $\gamma \sim \text{gamma}(1, 1)$ and $\alpha \sim \text{gamma}(10, 1)$, and they were sampled during inference. Thus, we did not need to specify any parameters before learning the model. The only variables that we had to choose were the segment length l and the reduced dimension of the feature space q.

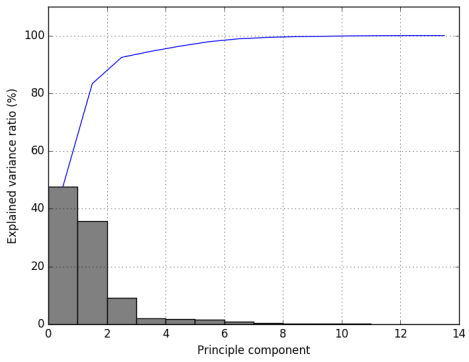

Fig. 2. An example of the PCA results.

3.5. Performance assessment

The performance of the classifiers was quantified by the weighted relative accuracy (WRA) [17]. WRA is an unbiased accuracy measure defined as WRA = $4 \times \text{cost} \times (\text{TPR} - \text{FPR})/(1 + \text{cost})^2$, where TPR and FPR denote the true and false positive rates. In this study, we assigned the cost to be equal to 1.

We used the 5-fold stratified cross-validation (CV) method for performance assessment, and the results were averaged across each iteration.

4. RESULTS

In this section, we provide the performance of HDPGMs and compare it to that of an SVM-based method, which achieved the best performance both in [5] and [7].

4.1. The HDPGM-based method

As described in Section 3.1, we experimented with two different thresholds τ_1 that delineate the unhealthy and the healthy groups of fetuses. By setting $\tau_1 = 7.05$, the number of FHR recordings used for analysis $N = 88$, and for $\tau_1 = 7.1$, $N = 122$. After segmentation, feature extraction and PCA, the dataset was transformed to N groups of data, each group having m observations of dimension q. In each iteration of the CV, 80% of the recordings were used as training data, and the rest were for validation.

The true negative rate (TNR) and true positive rate (TPR) were also calculated from the testing dataset. The classification performance of HDPGMs are shown in Table 2, with the best performance highlighted in bold font.

4.2. The SVM-based method

In testing the SVM-based method, we used the same two datasets. We first extracted the same 14 features described in Section 3.2 from the last 30-minute data of each FHR series, and then we scaled them to the range $(-1, 1)$. The scaled feature vectors were fed to the SVMs. The performance metrics were averaged across the 5-fold cross-validation. The SVM-based method had two free parameters: the cost C and γ. We searched for the optimal combination of these parameters and the results obtained with them are shown in Table 3.

When we compare these results with those in Table 2, we note that in both cases, for $\tau_1 = 7.05$ and $\tau_1 = 7.1$, the proposed method outperformed the SVM-based method.

Table 2. Performance of HDPGMs

N	q	l	TPR	TNR	WRA
		10 sec	**0.753**	**0.844**	**0.597**
	2	20 sec	0.708	0.822	0.531
		30 sec	0.681	0.844	0.525
		10 sec	0.706	0.800	0.506
88	3	20 sec	0.636	0.844	0.480
		30 sec	0.655	0.867	0.522
		10 sec	0.700	0.733	0.433
	4	20 sec	0.656	0.800	0.456
		30 sec	0.642	0.703	0.344
		10 sec	0.637	0.753	0.390
	2	20 sec	0.606	0.769	0.376
		30 sec	**0.654**	**0.754**	**0.408**
		10 sec	0.654	0.704	0.358
122	3	20 sec	0.655	0.721	0.376
		30 sec	0.554	0.803	0.356
		10 sec	0.622	0.720	0.342
	4	20 sec	0.604	0.738	0.342
		30 sec	0.587	0.719	0.306

Table 3. Peformance of SVMs.

N	C	γ	TPR	TNR	WRA
88	3	0.1	0.650	0.867	0.517
122	1	0.1	0.556	0.836	0.392

5. CONCLUSION

In this paper, we proposed hierarchical Dirichlet process Gaussian mixture models for classification of FHR tracings. These models belong to the family of non-parametric models, and they have the nice feature of being almost free from selecting model parameters. The proposed method was implemented on a reduced feature space obtained after PCA analysis of 14 features. We compared the performance of the proposed method to that of an SVM-based method on real tracings. The results show that the proposed method outperformed the SVM-based method.

6. REFERENCES

[1] E. Chandraharan and S. Arulkumaran, "Prevention of birth asphyxia: Responding appropriately to cardiotocograph (ctg) traces," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 21, no. 4, pp. 609–624, 2007.

[2] G. A. Macones, G. D. Hankins, C. Y. Spong, J. Hauth, and T. Moore, "The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: Update on definitions, interpretation, and research guidelines," *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, vol. 37, no. 5, pp. 510–515, 2008.

[3] G. Rooth, A. Huch, and R. Huch, "FIGO news: Guidelines for the use of fetal monitoring," *International Journal of Gynecology and Obstetrics*, vol. 25, pp. 159–67, 1987.

[4] L. Hruban, J. Spilka, V. Chudáček, P. Janků, M. Huptych, M. Burša, A. Hudec, M. Kacerovskỳ, M. Kouckỳ, M. Procházka, et al., "Agreement on intrapartum cardiotocogram recordings between expert obstetricians," *Journal of Evaluation in Clinical Practice*, vol. 21, no. 4, pp. 694–702, 2015.

[5] J. Spilka, V. Chudáček, M. Kouckỳ, L. Lhotská, M. Huptych, P. Janků, G. Georgoulas, and C. Stylios, "Using nonlinear features for fetal heart rate classification," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 350–357, 2012.

[6] A. Georgieva, S. J. Payne, M. Moulden, and C. W. Redman, "Artificial neural networks applied to fetal monitoring in labour," *Neural Computing and Applications*, vol. 22, no. 1, pp. 85–93, 2013.

[7] L. Xu, C. W. Redman, S. J. Payne, and A. Georgieva, "Feature selection using genetic algorithms for fetal heart rate analysis," *Physiological Measurement*, vol. 35, no. 7, pp. 1357, 2014.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 2012.

[9] K. Yu, J. G. Quirk, and P. M. Djurić, "Fetal heart rate analysis by hierarchical Dirichlet process mixture models," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 709–713.

[10] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, pp. 639–650, 1994.

[11] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, and L. Lhotská, "Open access intrapartum ctg database," *BMC Pregnancy and Childbirth*, vol. 14, no. 1, pp. 1, 2014.

[12] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[13] H. Gonçalves, A. P. Rocha, D. Ayres-de Campos, and J. Bernardes, "Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery," *Medical and Biological Engineering and Computing*, vol. 44, no. 10, pp. 847–855, 2006.

[14] J. De Haan, J. Van Bemmel, B. Versteeg, A. Veth, L. Stolte, J. Janssens, and T. Eskes, "Quantitative evaluation of fetal heart rate patterns: I. Processing methods," *European Journal of Obstetrics & Gynecology*, vol. 1, no. 3, pp. 95–102, 1971.

[15] C. K. Karmakar, A. H. Khandoker, J. Gubbi, and M. Palaniswami, "Complex correlation measure: a novel descriptor for Poincaré plot," *Biomedical Engineering Online*, vol. 8, no. 1, pp. 1, 2009.

[16] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 365–374, 2003.

[17] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, pp. 1137–1145.