CLASSIFICATION OF FETAL HEART RATE SERIES

Shishir Dash[⋆] *Jolene Muscat*[†] *J. Gerald Quirk*[†] *Petar M. Djurić*[⋆]

[⋆] Department of Electrical and Computer Engineering, Stony Brook University, NY
[†]Department of Obstetrics/Gynecology, Stony Brook University Hospital, NY

ABSTRACT

We study the problem of accurate automatic classification of fetal heart rate (FHR) signals using three different classification methods. FHR time series data are segmented into short (15s) spans of data, and features are extracted from them. These features include some established metrics of FHR trends such as acceleration and deceleration durations as well as a new set of features derived from the sequence of beat-to-beat percentage changes of the FHR signals. In total, we use 10 different features and demonstrate the feasibility of using them for classifying short segments into one of two suitably defined classes denoted as normal or abnormal. Classification is achieved using three different methods: support vector machine, a parametric Bayesian method and a non-parametric Bayesian method utilizing a neighbour-counting procedure for class-conditional density estimation. The performances of these methods are demonstrated on a database of physician-annotated recordings from which 580 short epochs of FHR patterns were extracted.

Index Terms— FHR, classification, Bayesian, ROC

1. INTRODUCTION

Pattern classification in the context of biomedical signals has been a problem of interest for many years. With the explosive growth in bioinformatics and biosignal processing research, many novel feature extraction and classification methods have been developed. Some example applications include EEG-based brain-computer interfaces [1] and bioinformatics [2]. A wealth of academic resources may be found for many other related methods and applications.

In this study we consider the problem of automatic or computer-aided diagnosis of fetal heart rate (FHR) signals. Clinically speaking, visual diagnosis of the FHR signals is of critical importance when evaluating the status of pregnancy and delivery. This is because oxygen inadequacy (also termed hypoxia) has a direct effect on the FHR and a timely detection of possible abnormalities in the signal can go a long way to prevention of dangers. However, purely visual assessment of fetal heart rate segments has in the past been proven to have high intra- and inter observer variability [3]. The problem has persisted despite the publication of standardized guidelines, such as those by the National Institute of Child Health and Human Development (NICHD) [4], for the interpretation of FHR patterns. This has led to an alarming increase in the rate of caesarian sections and unnecessary litigation expenses in even simple cases.

Thus, the development of accurate diagnostic algorithms capable of automatic classification of FHR patterns is of paramount importance, and several attempts have been made in this direction. A comprehensive review of such methods has been provided in [5]. The general scheme for classification involves sampling the FHR time series at a suitable rate, dividing it into appropriately sized segments, extracting a feature vector for each segment and using this vector as input to a feature classification system. The extracted features may include power spectral density estimates, (e.g.[6]), morphological features such as number of "accelerations" (increases in FHR), "decelerations" and their corresponding sizes (e.g. [7]), linear features such as mean and variance of FHR over some time period, and nonlinear features including approximate or sample entropy [8]. A recent study has focused on the dynamic relationship between the FHR and the maternal uterine pressure signal, quantifying it as an impulse response function and using the associated gain and phase delay as features capable of discriminating between normal and abnormal cases [9]. Almost all of these algorithms extract the feature vector from fairly long segments of FHR data; typically about 10-20 minutes of data yields one set of features. However, significant nonstationarities in long data sets may decrease the quality of the features. This prompted us to try diagnostic algorithms for short segments of data (here we tried 15s segments) using a new set of features from the beat-to-beat FHR percentage changes. Our idea is to use efficient methods for localized detection of patterns and develop a sequence of classifications to characterize the overall class of long FHR time-series data.

Once feature extraction is implemented, the algorithm must perform the training and classification steps. In the current state-of-the art, support vector machines (SVM's) [10] seem to be the classifiers of choice when it comes to obtaining discriminating functions, mainly because of the relative simplicity of the theory, the convenient extension to nonlinear discriminators through the use of the kernel trick , and high accuracy in several real applications.

Alternatively, one can use the Bayesian approach to classification. In the absence of any training data, one assumes a certain prior probability that a given feature vector belongs to any one of the classes. Then the training set is used to find the class-conditional density of feature-vectors. The final posterior probability is found by taking the product of the class-conditional density and the prior probabilities. The crux of the method is the estimation of the class-conditional densities, for which one may assume parametric (typically Gaussian or mixture Gaussian) or non-parametric forms.

In this paper, we report on the development of classification methods for short segments of FHR data. This includes development of a new set of features that we expect will be able to accurately capture characteristics of possibly dangerous FHR patterns. For each short segment, this set of features is fed into a classifier such as SVM or Gaussian Bayesian methods as described above. In addition we also use a non-parametric method for class-conditional density estimation (called the "Neighbor Counting" (NC) method) for accurate classification.

2. PROBLEM FORMULATION

We concentrate for now on binary classification of data (e.g., "Normal" vs. "Abnormal") in the supervised case. The FHR signal is

denoted as h_n, where n is the sample number. The features in the database may be viewed as points in m-dimensional feature-space and will be denoted as $\mathbf{x}_i; i \in \{1, 2, \ldots, N\}$, where N is the number of training vectors. For each $\mathbf{x}_i$ we have a corresponding true label y_i which can take values from $\{+1, -1\}$. Training involves finding a function $f(\mathbf{x})$ that is able to separate the features in the two categories with minimum overall cost.

In the following, we describe first the features extracted from the data, including features from the so-called FHR "return" series. Then we outline the proposed non-parametric approach to binary classification via a density estimation method. SVM's and parametric Bayesian methods based on Gaussian assumptions are very well studied and are thus not described here. In the Results section, we demonstrate that these can be used to classify short segments of data in an efficient way.

3. DATA AND FEATURES

Our real database consists of 580 short 15-s epochs of fetal heart rate time series data extracted from clinical recordings of 11 different subjects. The original 20-min Doppler FHR segments, collected in the Department of Obstetric/Gynaecology at Stony Brook University hospital at 4Hz sampling rate, were labeled as normal or abnormal by independent physicians. Out of these 11 recordings, regions of FHR patterns were carefully isolated and labeled as such. Each labeled region was segmented into non-overlapping 15-s epochs. Abnormalities may include decelerating heart rate or low variability, and an epoch was labeled as "abnormal"(the target class, denoted +) if it had either one or both of these attributes. We note that these criteria for abnormality are *not* the only ones used ultimately by physicians, and future work will include the study of more types of abnormality.

Each epoch was denoised using an algorithm similar to [11]. FHR baseline detection was performed using a median-filtering algorithm. We made no differentiation between recordings taken at different gestational ages or stages of delivery.

From each 15-s (60 sample) epoch of data denoted $\mathbf{h} = [h_1, h_2, \ldots, h_{60}]$, several different types of features can be extracted. We focus here on 10 specific features as described in the sequel.

3.1. Features from raw FHR series

From the raw FHR series, we can extract features to quantify the average time the FHR decelerates or accelerates. We consider the following features:

1. Number of FHR samples out of 60, that are above an acceleration threshold ξ_U.
2. Number of FHR samples out of 60 that are below a deceleration threshold ξ_L. Both acceleration and deceleration thresholds were set according to the prescribed guidelines set by NICHD [4].
3. Standard deviation of the FHR series σ_h.

3.2. Features from the FHR return series

The return series r_n is computed as a time series of sample-to-sample percentage changes in the FHR signal h_n. In order to standardize the range of the possible r_n signals, we first center the original signal h_n around a reasonable constant FHR value (here 140 bpm) and then calculate the baseline (b_n) of the centered signal ($\hat{h}_n$). The "unbiased" FHR signal $\tilde{h}_n$ is then obtained by subtracting b_n from h_n. The return series is then obtained for $\tilde{h}_n$ according to:

$$r_n = \frac{\tilde{h}_n - \tilde{h}_{n-1}}{\tilde{h}_n}. \quad (1)$$

From the return series $\mathbf{r} = [r_1, \ldots, r_{60}]$, we can obtain several features. In a separate, as yet unpublished study, we explored the class-separation performance of several different types of features when used individually. These included several statistical moments (since they can quantify the r_n probability distributions) as well as nonlinear features which have been used previously in adult heart rate variability studies. Hypothesis testing was done via the Kolmogorov-Smirnov test. The statistics for the following 6 features were found to be significantly different for the two classes at the $p = 0.05$ level:

1. Total return $S_r = \sum_{n=1}^{60} r_n$.
2. Variance of return data σ_r^2.
3. Skewness of return data $\gamma_r = (E(r - \mu_r)^3)/\sigma_r^3$.
4. Kurtosis of return data $K_r = (E(r - \mu_r)^4)/\sigma_r^4$.
5. Runs ratio of return data ρ_r. This is the number of distinct runs of consecutive increases or decreases of the return series from zero. For instance, the sequence $\{+, +, -, -, +, -, +\}$ has five runs. A higher number of runs indicates higher variability.
6. Shannon Entropy of return data ϵ_r. This feature summarizes the complexity in the return series. We find the histogram of the sequence of r_n's in N_b bins, with frequency in bin k denoted p_k, and then compute ($N_b = 16$ in our implementation):
$$\epsilon_r = -\sum_{k=1}^{N_b} \ln(p_k/60) p_k/60. \quad (2)$$
7. Turning Point Ratio of return data τ_r: A sample z_k from any given sequence $\{z_1, \ldots, z_N\}$ is denoted a turning point if the samples z_{k-1} and z_{k+1} are either both greater than or both smaller than z_k. The turning point ratio τ_z is then defined as the ratio of the number of turning points to the length of the sequence N [12]. To obtain τ_r we first map the r_n data so that for a positive (negative) r_n, $r_n^* = +1(-1)$. Then the number of points in r_n^* preceded and succeeded by r^*'s of opposite signs is computed. This number expressed as a fraction of the total number of samples is τ_r. For example, the sequence $\{+, +, -, -, +, -, ++\}$, has two turning points, and the TPR is 2/8 = 0.25.

Out of these 10 features, we use m at a time for classification. For each value of m there may be many combinations of features, each yielding a different performance. For training and testing, we used the method of 10-fold cross-validation on the full set of 580 feature sets. We utilized the receiver operating characteristic (ROC) method to get classification performance measures for the best combinations of m features. As a performance metric, we used the area under the ROC curve (AUC).

4. PROPOSED CLASSIFICATION APPROACH

We developed a Bayesian formulation for the pattern classification problem, but without assuming any parametric model for the class-conditional likelihoods. After mapping the training feature sets, the

test vector to be classified is mapped into the same feature space. To get an estimate of the target (control) likelihood of the test feature vector, we simply estimate the number of target (control) training samples in the immediate vicinity of the test point and divide this number by the total number of training data from target (control) class $N_+(N_-)$. However, we need to appropriately define the term "immediate vicinity".

The simplest assumption one can make about any feature data is that they arise from a uniform distribution. That is, given any region of hypervolume T in the feature space, the probability that a given feature vector falls in this region is $p = T/V$, where V is the total support hypervolume of the distribution. We define the region as a rectangular cuboid (in m dimensions) whose volume W is directly dependent on the support of each of the feature vectors, as estimated from the training database. If the training feature vector is $\mathbf{x}_i = [x_1^i, x_2^i, \ldots, x_m^i]^T$, for $i \in \{1, \ldots, N\}$, the feature supports and total volume V can be defined as

$$\begin{aligned} S_j &= \max(x_j^1, \ldots, x_j^N) - \\ &\quad \min(x_j^1, \ldots, x_j^N); \quad j = 1, \ldots, m \qquad (3) \\ V &= \prod_{j=1}^{m} S_j. \qquad (4) \end{aligned}$$

Assuming the window of immediate vicinity has width t_j in a direction along the axis of the j-th feature, the corresponding hypervolume is $T = \prod_{j=1}^{m} t_j$. We assume each t_j is directly proportional to the support S_j of the corresponding feature. Given N training vectors, the average number of training vectors mapping inside T is Np. Thus, given some choice of μ, we can calculate the widths as

$$\begin{aligned} t_j &= \gamma S_j \qquad (5) \\ &= S_j \Big(\frac{\mu}{N - k\mu}\Big)^{1/m}, \quad \forall j \in \{1, 2, \ldots, m\}, \qquad (6) \end{aligned}$$

where μ is simply a convenient way to define the window widths for estimating the class-conditional probability of the feature vector, and $0 < k < (1-p)/p$. Given these widths and knowledge of the prior probability $P(y_i)$, we can define the class-conditional and posterior probabilities for some unlabeled test feature vector $\mathbf{x} = [x_1, \ldots, x_m]^T$ as

$$\begin{aligned} P(\mathbf{x}|y_c) &= K_x/K_c, \qquad (7) \\ P(y_c|\mathbf{x}) &\propto P(\mathbf{x}|y_c)P(y_c); \quad c \in \{+1, -1\} \qquad (8) \end{aligned}$$

where K_x denotes the number of training vectors $\mathbf{x}_i$ in class c satisfying $|x_j - x_j^i| \leq t_j, \forall j$, and K_c denotes the total number of training vectors in class c. A threshold dependent decision function can now be defined as follows:

$$f_{NC}(\mathbf{x}; \mu, \rho) = \begin{cases} +1, & \frac{P(y_{+1}|\mathbf{x},\mu)}{P(y_{-1}|\mathbf{x},\mu)} - \rho > 0. \\ -1, & \text{otherwise} \end{cases} \quad . \qquad (9)$$

5. RESULTS

Empirical performance analysis was done across all possible feature combinations to find good feature sets. AUC was found for each decision function by varying the corresponding $\boldsymbol{\theta}$ (parameters of the decision function) in a way that the full ranges of sensitivity and specificity are explored. It is denoted $A_{\boldsymbol{\theta}}$. It has been established that the AUC metric is an estimate of the total probability that the value of the decision function of a randomly chosen feature vector from the target class is greater than that for a randomly chosen feature vector from the control class [13]. Thus, the higher the AUC value, the better the method performs (in an average sense).

Table 1: Comparison of classification performance. Higher $A_{\boldsymbol{\theta}}$ values imply better average classifier performance.

Method	m	$A_{\boldsymbol{\theta}}$	Best TPR	Best (1-FPR)
SVM	2	0.59	0.85	0.43
NDG	2	0.60	0.77	0.53
NC	2	0.66	0.69	0.59
SVM	9	0.66	0.78	0.53
NDG	9	0.68	0.64	0.61
NC	9	0.68	0.74	0.60

For the SVM method, we kept the radial basis function scaling factor fixed at 1 and varied the box constraint for soft margins, i.e., $\boldsymbol{\theta} = C$. When using the Bayesian method with Gaussian assumption (n-dimensional Gaussian (NDG) method), we varied the likelihood ratio threshold, i.e., $\boldsymbol{\theta} = \rho$. For the neighbour counting method f_{NC}, we had to vary two parameters, μ and the likelihood ratio threshold, i.e., $\boldsymbol{\theta} = [\mu, \rho]$. However, the analysis was further complicated by the fact that there were a number of possible feature combinations to explore for each value of m. For instance, for $m = 2$, we had a total of $\binom{10}{2} = 45$ feature combinations to sift through in finding the best performance.

Thus, showing the full analysis of the entire feature-combination space for all three classifiers is not possible due to space constraints. Instead we provide ROC curves for two different feature sets of lengths $m = 2$ ($\mathbf{x} = \{\xi_U, \sigma_f\}$) and $m = 9$ ($\mathbf{x} = \{\xi_L, \xi_U, \sigma_f, S_r, \epsilon_r, \rho_r, \sigma_r^2, \tau_r, K_r\}$), respectively. These two feature combinations were found to give good classification performance using all three methods. The corresponding receiver operating characteristic curves are shown in Figs. 1a and 1b respectively. Table 1 summarizes the results. TPR denotes True Positive Rate (also called Sensitivity) while FPR is the False Positive Rate (also called the rate of false alarm). The terms "Best TPR" and "Best $(1 - FPR)$" denote the pair of coordinates $(TPR, 1 - FPR)$ that maximizes the product $TPR(1 - FPR)$.

6. DISCUSSION AND CONCLUSIONS

The above results demonstrate the feasibility of using short epochs for classification of FHR signals. We note that making a final decision on the status of the FHR series incorporates many other factors including the presence of accelerations, variability around the baseline, presence and frequency of contractions (obtained from the uterine contraction signal) along with the use of long segments of cardiotocographic signals. In practice, anywhere between 10-40 minutes of FHR signal may be used by doctors to do a classification. While this has the advantage of utilizing more information, from a signal-processing perspective, it is not very advantageous since the FHR signal, like most biomedical signals, can have significant non-stationarities across long time-scales.

In our proposed approach, we first classify short segments of FHR series, followed by ensemble classification of the sequence of short-classifications. Additionally, we have shown here the possibility of using several features extracted from the FHR return series instead of the raw FHR. Most of the statistical features such as variance, skew and entropy are usually applied on the raw signal. We can see from the results that for a false-alarm rate of 40%, it is possible to achieve 74% sensitivity using 9 out of the 10 considered features.

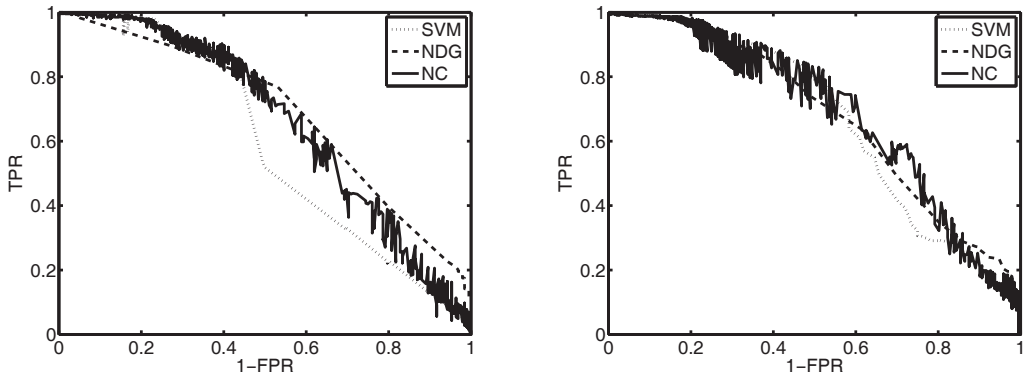

Fig. 1: Receiver operation characteristic curve for the m-feature combinations from analysis of FHR data for all 3 methods, with (a) $m = 2$ and (b) $m = 9$. TPR = True positive rate (sensitivity); FPR = False Positive Rate.

When we studied class-separation performance using hypothesis testing methods on individual features, we also analyzed the effect of using segment lengths varying from 10s to 1 minute. It was observed that the features from the raw FHR signal (i.e., ξ_U, ξ_L, σ_h) were significantly separated for all segment lengths. For return features, different results were obtained for different segment lengths. The two nonlinear features were significant separators for all segment lengths, while the statistical moments like total return and variance were significant separators of shorter segment lengths (up to 30s). However, since we were using non-overlapping epochs and a fixed amount of real data, the total number of feature sets was different for different segment lengths (fewer training sets of 1 minute length). In our judgement, it makes sense to use results from bigger training sets as a basis for choosing features, which is why we used 15s epoch lengths.

To further improve the results, we need to study (a) feature extraction from different durations of FHR and (b) a bigger database of supervised training data. However, we note that complete visual annotation of large durations of FHR data segmented into many shorter segments may not be feasible, and thus we also need to develop algorithms for unsupervised or semi-supervised training. In addition, we plan to include features extracted from the uterine pressure, as input to the classifier.

7. REFERENCES

[1] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, pp. R1–, 2007.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, p. 2507, 2007.

[3] E. Blix, O. Sviggum, K. Koss, and P. Øian, "Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 110, no. 1, pp. 1–5, 2003.

[4] G. Macones, G. Hankins, C. 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.

[5] V. Chudáček, "Automatic analysis of intrapartum fetal heart rate," Ph.D. dissertation, Czech Technical University, Prague, 2011.

[6] A. Reddy, M. Moulden, and C. W. Redman, "Antepartum high-frequency fetal heart rate sinusoidal rhythm: computerized detection and fetal anemia," *American Journal of Obstetrics and Gynecology*, vol. 200, no. 4, pp. 407.e1–407.e6, 2009.

[7] J. Pardey, M. Moulden, and C. W. Redman, "A computer system for the numerical analysis of nonstress tests," *American Journal of Obstetrics and Gynecology*, vol. 186, no. 5, pp. 1095–1103, 2002.

[8] J. Spilka, V. Chudáček, M. Koucky, L. Lhotska, M. Huptych, P. Janku, G. Georgoulas, and C. Stylios, "Using nonlinear features for fetal heart rate classification," *Biomedical Signal Processing and Control*, vol. In Press, Corrected Proof, 2011.

[9] P. Warrick, E. Hamilton, D. Precup, and R. Kearney, "Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 4, pp. 771 –779, 2010.

[10] C. Campbell and Y. Ying, *Learning with Support Vector Machines*, R. J. Brachman and T. Dietterich, Eds. Morgan & Claypool Publishers, 2011.

[11] D. Ayres-de Campos, C. Costa-Santos, J. Bernardes *et al.*, "Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the SisPorto® multicentre validation study," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 118, no. 1, pp. 52–60, 2005.

[12] S. Dash, K. Chon, S. Lu, and E. Raeder, "Automatic real time detection of atrial fibrillation," *Annals of biomedical engineering*, vol. 37, no. 9, pp. 1701–1709, 2009.

[13] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.