Classification of Normal and Hypoxic Fetuses using System Identification from Intra-Partum Cardiotocography

Philip A. Warrick PHILIP.WARRICK@MCGILL.CA
Biomedical Engineering Department, McGill University, Montreal, Quebec, Canada

Emily F. Hamilton EMILY.HAMILTON@LMSMEDICAL.COM
LMS Medical Systems, Inc., Montreal, Quebec, Canada.

Robert E. Kearney ROBERT.KEARNEY@MCGILL.CA
Biomedical Engineering Department, McGill University, Montreal, Quebec, Canada

Doina Precup DPRECUP@CS.MCGILL.CA
School of Computer Science, McGill University, Montreal, Quebec, Canada

Keywords: diagnostic decision making, classification, system identification, obstetrics

Abstract

We present a novel approach to classifying normal and hypoxic fetuses during labor, using a mixture of system identification and machine learning methods. We treat uterine pressure from contractions as an input and the fetal heart rate as an output, and fit a non-parametric linear model describing their relationship. We take special steps to deal with noise and guard against overfitting in this step. We use properties of the model as attributes for classification. Our approach shows very promising results on a database of real clinical recordings.

1. Introduction

Oxygen deprivation during labor (hypoxia) is a major problem in obstetrics. It is estimated that between 1 and 7 in 1000 fetuses experience hypoxia that is severe enough to cause fetal death or severe brain injury (ACOG, 2003). Unfortunately, clinicians must rely on indirect measures of oxygen delivery and neurological function in order to assess the fetal state. A standard approach measures maternal uterine pressure (UP) and fetal heart rate (FHR). This pair of signals is called *cardiotocography (CTG)*. An example CTG is presented in Figure 1.

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008.

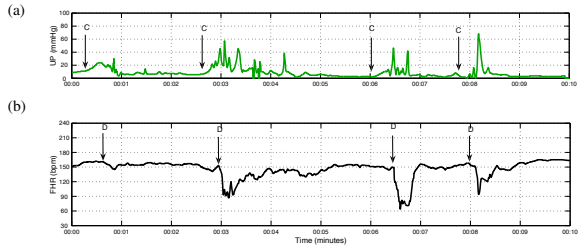

Figure 1. Uterine pressure (a) and fetal heart rate (b) during 10 minutes of labor. Contractions and FHR decelerations are indicated with arrows.

Contractions reduce fetal oxygen supply by compressing the umbilical cord or by diminishing gas exchange in the utero-placental unit, which can have severe consequences if the placenta is already impaired. In response, the fetal heart rate typically decreases, a process known as *FHR deceleration*. There is general consensus among clinicians that deceleration depth, frequency and timing with respect to contractions are indicators of both the strength of the insult and the ability of the fetus to withstand it.

However, physicians are often inconsistent in their interpretation of CTG (Parer et al., 2006). Because significant hypoxia is rare, false alarms are common; moreover, physicians may disregard the more rare, truly abnormal signals. Indeed, approximately 50% of birth-related brain injury cases are deemed preventable, with incorrect CTG interpretation leading the list of causes (Draper et al., 2002;

Ransom et al., 2003). Thus, there is great motivation to find better methods to discriminate between healthy and hypoxic fetuses.

Machine learning techniques can potentially be quite useful in helping physicians perform this difficult task. Previous work in this field has focused on attempting to extract features of the CTG that mimic clinical understanding, and then using classification techniques in combination with these features. Most approaches have addressed contraction and deceleration detection separately (Warrick et al., 2005; Lunghi et al., 2005; Cao et al., 2006; Georgoulas et al., 2006). Once the events are detected, their timing and duration can be determined. Unfortunately, CTG is very noisy. The sensor belt may move on the mother's abdomen, which can cause the signal to become faint, or to get interrupted altogether. The (much lower) heart rate of the mother may interfere with the FHR measurement. Also, CTG is subject to sensor delays (Jezewski et al., 2005). Hence, the feature extraction step is very difficult and its results are imprecise, which makes it very hard to perform successful classification. However, because CTG is universally available in clinical settings, developing automate methods to analyze it is really important.

In this paper, we outline a new approach to this problem, which uses system identification methods, instead of feature extraction, as a pre-processing step. We model the relationship between UP (as an input) and FHR (as an output) (Warrick et al., 2006) using a linear non-parametric model, as explained below. This model inherently contains information about the strength and timing of the FHR response to contractions, in contrast to the feature detection approach where these relations must be explicitly computed as amplitude ratios and time delays, respectively. Based on these models, we create a set of attributes which are then used to classify the fetal state, using decision trees. We obtain very promising results on a database of real cases.

2. Data

We used a database consisting of 264 intrapartum CTG recordings for pregnancies having a birth gestational age greater than 36 weeks and no known genetic malformations. Only records with at least 3 hours of recording were considered. The signals were sampled at discrete time steps, with a frequency of 4Hz.

Each recording was labelled by outcome according to its arterial umbilical-cord base deficit and neonatal indications of neurological impairment. An elevated base deficit measurement is an indicator of metabolic acidosis of sufficient degree to cause neurological injury (Parer et al., 2006). The majority of the recordings were from normal fetuses (221 cases: base deficit $<$ 8 mmol/L); the rest were severely pathological (43 cases: base deficit $\geq$ 12 mmol/L, death or evidence of hypoxic ischemic encephalopathy). The pathological cases are over-represented in the database, compared to their usual incidence in the population.

Because this is real clinical data, it is very noisy. In particular, loss of sensor contact with the abdomen of the mother is a normal occurrence, e.g. if the mother is moving around. To deal with this, we detected when the signal dropped to a very low amplitude. We either merged, bridged or removed the dropouts from consideration, depending on their duration. The details of this process are available in (Warrick et al., 2008). Out of the resulting data, we created 20-minute epochs, with 10 minute overlap between successive epochs. We extracted as many epochs as possible from the beginning of a clean segment. This length of data is a compromise. On one hand, longer epochs typically generate better models. On the other hand, if the epoch is too long, non-stationarity and noise artifacts can negatively affect the results.

3. System identification

The system identification process is summarized by the block diagram in Figure 2.

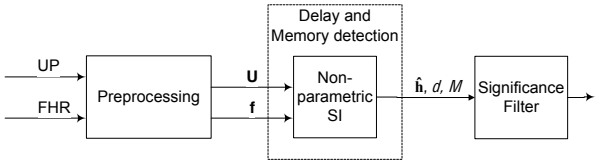

Figure 2. Block diagram of the system identification phase

Let u_n and f_n denote the input (UP) and output (FHR) at time step $n = 1 \ldots N$. In non-parametric linear system identification, the assumption is that f_n can be modeled using a linear combination of the values of the input signal observed over M consecutive time steps. Mathematically, f_n is modeled as the following convolution sum:

$$f_n \approx \sum_{i=0}^{M-1} (h_i \Delta t) u_{n-i-d} = \mathbf{h} * \mathbf{u}_{n-d} \qquad (1)$$

Here, Δt is the sampling period, and $\mathbf{u}_{n-d}$ is the length-M vector of input signal samples $[u_{n-d-M+1} \quad \ldots \quad u_{n-d-1} \quad u_{n-d}]$ used to compute f_n at sample n. The parameter d is called the *delay* and defines the start of the input portion used to compute f_n. For causal (physically realizable) systems, $d \geq 0$. However, in the presence of an input measurement delay, d may be negative (Hunter & Kearney, 1983). In our

case, UP is measured through a pressure transducer sensor, which has a measurement delay estimated between 10 and 80 seconds (Jezewski et al., 2005), so negative as well as positive delays are possible.

The vector of parameters $\mathbf{h}$ (of length M) is called the *impulse response function (IRF)*; together with d, it constitutes the model of the system. The goal of the system identification task is to estimate this model from the CTG data. However, we had to deal with several important challenges: the input is completely uncontrolled and uncalibrated; the input is dominated by the contraction frequency; and the signals are quite noisy. Hence, we had to augment traditional linear system identification methods in a variety of ways.

For each 20-minute epoch, we fitted one model, obtained as a least-squares estimate:

$$\mathbf{h} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{f} \tag{2}$$

where $\mathbf{U}$ is the $N \times M$ matrix formed from $\mathbf{u}_{n-d}$ and $\mathbf{f}$ is the length-N measured output. To suppress noise, we then performed PCA to retain only the most significant IRF principal components (Warrick et al., 2006). The resulting model is denoted by $\hat{\mathbf{h}}$.

PCA works with *fixed* delay and memory parameters, d and M. However, because the input is highly periodic and affected by unknown sensory delays, we needed to determine d and M automatically as well. To do this, we performed a search over values of d and M. The first criterion was that the first and last IRF coefficients have to be close to 0 (hence making sure that the equation (1) is indeed correct). Secondarily, we attempted to ensure that d and M were somewhat consistent between the different epochs. The description of the search procedure is beyond the scope of this paper; we refer the interested reader to (Warrick et al., 2008).

The standard measure used in system identification to evaluate the quality of the models is the *percentage variance accounted for (VAF)*, defined as:

$$\%\text{VAF} = 100 * \left(1 - \frac{\sigma_e^2}{\sigma_f^2}\right), \tag{3}$$

where σ_e^2 denotes the variance of the residual signal, $\mathbf{f} - \hat{\mathbf{f}}$, and σ_f^2 is the variance of the observed output signal. This measures how much of the observed variability in the measured output signal is explained by the output approximation obtained using the model.

Usually, models are further filtered using a threshold on the %VAF measure. However, in our case determining a fixed threshold would be very difficult. Instead, in order to provide more confidence that the model captures true system dynamics, we compared the models computed from the measured FHR to those computed from surrogate FHR signals, generated by the amplitude-adjusted Fourier-transform (AAFT) method (Warrick et al., 2007). AAFT generates signals that look like FHR, but their phase is scrambled, so any timing relationship with the input UP signal should be destroyed. We generated 99 surrogate models for each epoch, and we ranked them and our model in descending order of the %VAF. The model obtained by the system identification procedure was retained only if it was in the top 5 models in this list.

Figure 3 shows as example result from the system identification of a pathological case. As can be seen, the major variability in the signal is captured very well by the model, with only high-frequency components remaining unexplained. The IRF obtained is quite simple.

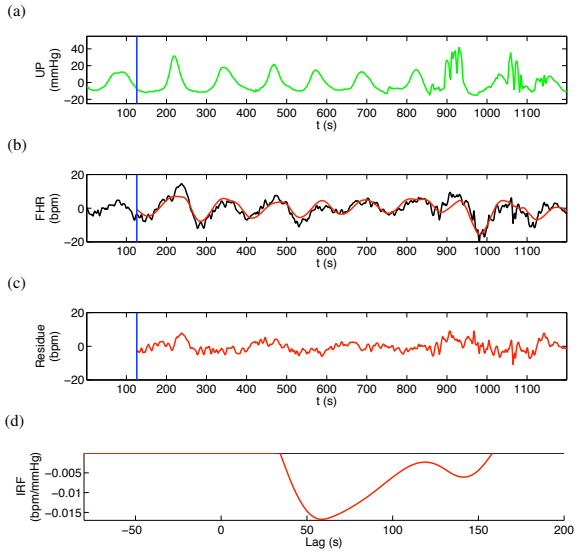

Figure 3. Example epoch of a pathological case. (a) Input UP (b) preprocessed FHR $\mathbf{f}$ (black) and predicted output FHR $\hat{\mathbf{f}}$ (red) (c) residual $\mathbf{f} - \hat{\mathbf{f}}$ (d) IRF model $\hat{\mathbf{h}}$. Vertical blue bars indicate memory length M=126s. The %VAF, delay d and gain G were 69.6%, 34s and 0.50bpm/mmHg, respectively.

4. Classification procedure

We used the models obtained by this system identification step to generate attributes to be used for the classification task. In order to measure the timing of the FHR response to UP, we included as input the delay d. We measured the strength of the response to contractions by the steady-state gain G, defined as the sum of the IRF coefficients: $G = \sum_{i=0}^{M-1} h_i$. Finally, we included as an attribute the %VAF for each model, which measures the success of the system identification step. In our preliminary work (Warrick et al., 2008), all these attributes showed significant

differences between the normal and the pathological cases (using a Kolmogorov-Smirnov distribution test).

We used these attributes and the labels from the database to provide a classification data set. For simplicity and interpretability, we decided to use a decision tree classifier. We used the Weka machine learning library to construct the decision trees. We trained using either data from all three hours, or from the last hour only. The latter is justified by the fact that the fetal state tends to deteriorate with time at unknown rates, so data in the last hour is expected to reflect the fetus in its most distressed state. In both protocols, testing was done on data from all three hours. We used 5 repetitions of 10-fold cross validation. We initially found that training generated majority classifiers due to the class imbalance. We compensated for this by up-sampling the pathological cases, weighting them by the ratio of the class sizes ($\sim 5:1$).

5. Results

Nine of the pathological and eleven of the normal cases contained excessive artifact and had to be completely discarded. Figure 4 shows a sample decision tree. In general, the decision trees constructed were quite consistent among the different folds, and relied heavily on the use of the delay attribute d. This is consistent with clinical understanding, which emphasizes the timing of the FHR deceleration to the contraction. In particular, pathological cases tend to have a delayed response, which is consistently indicated by the decision trees as well.

```
d > 2: P (475.3/175)
d <= 2:
:...VAF > 39.55428: N (246/97.2)
    VAF <= 39.55428:
    :...d <= -22: N (276.8/117.5)
        d > -22:
        :...G <= -0.408663: P (149.5/65.1)
            G > -0.408663: N (280.3/128.1)
```

Figure 4. Sample Decision Tree

Figure 5 shows the average accuracy over each epoch, using the last hour of labor as training data. As shown, the accuracy for the normal class is very high, consistently above 90%. This performance is significantly better than that obtained previously using feature-based methods. The accuracy of the pathological class is much lower, around 50%. However, this result is actually quite relevant from a clinical perspective. In a preliminary study, these cases were presented to clinicians, and their accuracy for the pathological cases was around 30%. We note also that the pathological cases are quite difficult, as they may be caused by different physiological mechanisms. Hence, we do not expect that perfect accuracy can be achieved, even if the system identification can be improved. The accuracy for the pathological class improves close to delivery; this trend also makes sense from a clinical point of view, because the stronger and faster contractions often worsen the condition of the fetus.

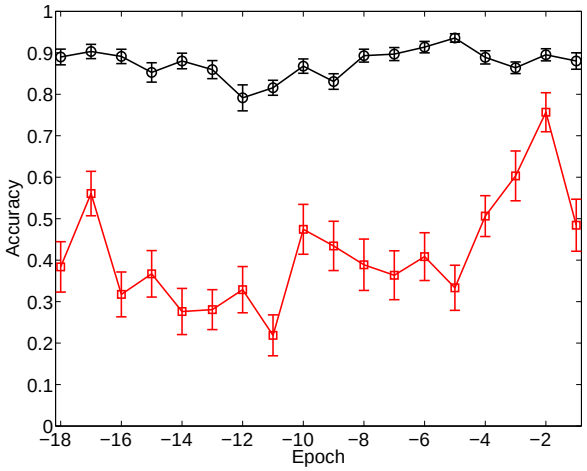

Figure 5. Average classifier accuracy for epochs from normal (top) and pathological (bottom) cases over time (error bars indicate 1 standard error above and below the mean)

For the pathological cases, Figure 6 shows the classifier behaviour by case, rather than by epoch. The cases are ordered by their proportion of correct classifications. We note that due to the fact that the system identification phase may fail, different cases have different numbers of epochs associated with them. For each case, we show the total number of epochs for which a model was generated successfully, the number of epochs correctly classified, how many times the predicted class label changes between the different epochs, as well as the number of times when a change from normal to pathological is predicted. The reason for the latter is that a prediction change from normal to pathological can happen because a problem has occurred with the fetus, and should not necessarily be viewed as a mistake.

6. Discussion

As can be seen, for many cases only a few epochs could be successfully modeled. This is due to the fact that in problematic deliveries, the CTG signal may be interrupted or very noisy due to special procedures taking place. Given the fact that many birth-related brain injury cases could have been prevented by correct CTG interpretation, these results indicate that our system identification and machine

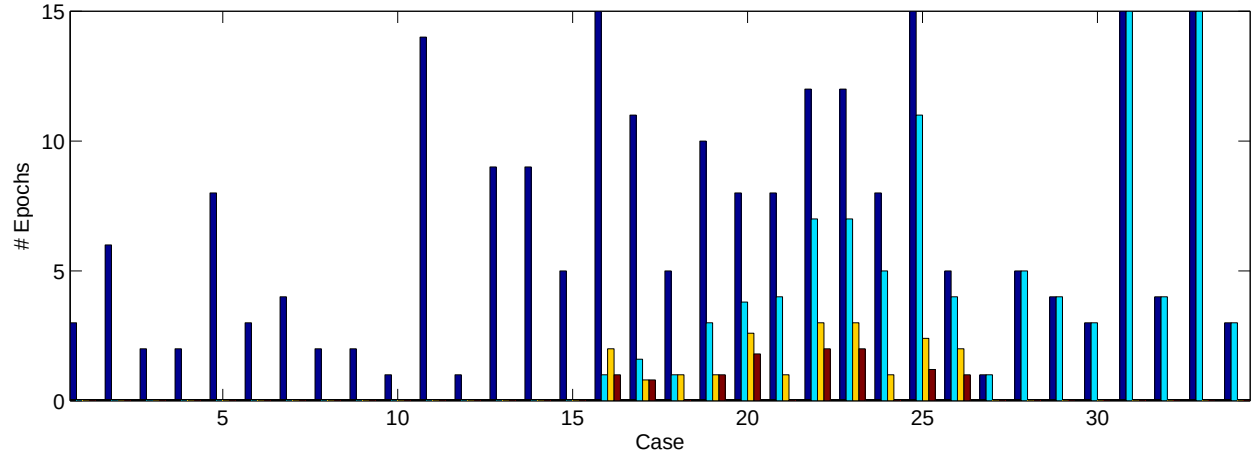

Figure 6. Classification of pathological cases (ordered by proportion correct). For each case, the vertical bars show the number of epochs tested (blue), the number correct (cyan), the number of prediction transitions (yellow) and the number of normal-to-pathological prediction transitions (red).

learning approach could be very helpful for this challenging diagnostic problem, and in particular, that it could help clinicians diagnose problematic cases earlier. We emphasize that this is a *fully automated* procedure, which can easily be deployed with the software associated with current CTG monitors which are in clinical use. This makes our approach unique among the state-of-art methods in this subfield.

The classification data set we use is still quite limited in the number of features included. We are currently working on including the baseline value of the heart rate (a very important clinical indicator which is not captures in the current data set), as well as features based on the *change* in the models over time. We also plan to try a broader slate of classifiers and compare their results.

In order to further assess the utility of this approach, we are working on evaluating the obtained classifiers on a database of "intermediate" cases. These are fetuses showing signs of hypoxia at birth, but which are not severely pathological and eventually recover. We also point out that the data we currently use has been collected from different hospitals, and not through a systematic clinical study. This contributes to a wide variation in the quality of the recordings. A clinical study could significantly improve the current performance.

Acknowledgments

The authors acknowledge the financial support of this work by LMS Medical Systems, Inc. and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

ACOG (2003). *Neonatal encephalopathy and cerebral palsy: Defining the pathogenesis and pathophysiology*. ACOG Task force on Neonatal Encephalopathy and Cerebral Palsy.

Cao, H., Lake, D., Ferguson, J.E., I., Chisholm, C., Griffin, M., & Moorman, J. (2006). Toward quantitative fetal heart rate monitoring. *Biomedical Engineering, IEEE Transactions on*, *53*, 111–118.

Draper, E., Kurinczuk, J., Lamming, C., Clarke, M., James, D., & Field, D. (2002). A confidential enquiry into cases of neonatal encephalopathy. *Arch Dis Child Fetal Neonatal Ed*, *87*, F176–F180.

Georgoulas, G., Stylios, D., & Groumpos, P. (2006). Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *Biomedical Engineering, IEEE Transactions on*, *53*, 875–884.

Hunter, I. W., & Kearney, R. E. (1983). Two-sided linear filter identification. *Medical & Biological Engineering & Computing*, *21*, 203–209.

Jezewski, J., Horoba, K., Matonia, A., & Wrobel, J. (2005). Quantitative analysis of contraction patterns in electrical activity signal of pregnant uterus as an alternative to mechanical approach. *Physiological Measurement*, 753.

Lunghi, F., Magenes, G., Pedrinazzi, L., & Signorini, M. (2005). Detection of fetal distress though a support vector machine based on fetal heart rate parameters. *Computers in Cardiology, 2005* (pp. 247–250).

Parer, J. T., King, T., Flanders, S., Fox, M., & Kilpatrick, S. J. (2006). Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association? *Journal of Maternal-Fetal & Neonatal Medicine*, *19*, 289–294.

Ransom, S., Studdert, D., Dombrowski, M., Mello, J., & Brennan, T. (2003). Reduced medicolegal risk by compliance with obstetric clinical pathways: A case-control study. *Obstet Gynecol*, *101*.

Warrick, P., Hamilton, E., & Macieszczak, M. (2005). Neural network based detection of fetal heart rate patterns. *Neural Networks, 2005. Proceedings. 2005 IEEE International Joint Conference on* (pp. 2400–2405).

Warrick, P. A., Hamilton, E. F., Precup, D., & Kearney, R. E. (2008). Identification of the dynamic relationship between intra-partum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *Under review*.

Warrick, P. A., Kearney, R. E., Precup, D., & Hamilton, E. F. (2006). System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses. *Computers in Cardiology 2006. Proceedings.* (pp. 937–940).

Warrick, P. A., Kearney, R. E., Precup, D., & Hamilton, E. F. (2007). Time progression of a parametric impulse response function estimate from intra-partum cardiotocography for normal and hypoxic fetuses. *Computers in Cardiology 2007* (pp. 693–696).