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

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

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

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

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

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.

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;
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 ≥ 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.

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:

\[
    f_n \approx \sum_{i=0}^{M-1} (h_i \Delta t) u_{n-i-d} = h \ast u_{n-d}
\]

Here, \( \Delta t \) is the sampling period, and \( u_{n-d} \) is the length-\( M \) vector of input signal samples \([u_{n-d-3} \cdots u_{n-d-1} u_{n-d}]\) 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
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.

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 (∼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.

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
Using System Identification from Intra-Partum Cardiotocography for Classification

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 captured 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


Using System Identification from Intra-Partum Cardiotocography for Classification


