Unsupervised Clustering of Fetal Heart Rate

Humza Ahmed

January 27, 2018

Capstone Project Proposal

Domain Background

Childbirth has been deconstructed into occurring in three general stages [16].The first stage begins with the onset of contractions as the cervix begins to dilate to its maximum size. During this stage uterine contractions are around 30-45 seconds, and irregular in nature occurring around 20 minutes apart. Progressively these contractions become more frequent and get closer to lasting between 60-90 seconds. The second stage begins after the cervix has reached its maximal diameter, and involves regular contractions between 60-90 seconds occurring every 5 or so minutes. During this stage the baby will actively move through the birth canal until it is delivered. The final stage ends with the delivery of the placenta. From the middle of the gestational period through the time of delivery, the early detection of fetal complications can help improve outcomes for both child and parent (e.g. through cesarian section).

Monitoring of fetal state during pregnancy most commonly occurs through the use of Cardiotocography (CTG), or Fetal Electrocardiogram (fECG). Both of these methods have the ability to measure fetal heart rate (FHR) which can indicate forms of fetal distress. However the complexity in making precise judgements based on FHR data has caused division between clinicians about its consistently to improve patient outcomes. In order to standardize clinical practice the American College of Obstetricians and Gynecologists (ACOG) has published a set of guidelines and definitions to facilitate interpretation of FHR data [2]. These definitions in addition to more familiar heart rate characteristics used in adults has allowed clinicians to now describe changes in FHR quantitatively. Although these guidelines have made FHR interpretation more quantitative, they are based upon a review of studies showing mixed results in patient outcomes. These and other studies have even shown an increase in the number of unneeded interventions conducted in healthy participants. [1, 8]. Thus it seems that there is still a need to further solidify our understanding of normal FHR patterns.

The ACOG guidelines in addition to most clinical literature analyzing FHR tends to ignore the temporal differences between Stage I and II of labor. Clinical literature either tends to conduct a global analysis of FHR during labor, or focus on a single stage [4, 12, 18]. It is however known that during uterine contractions FHR shows patterns of deceleration and acceleration [6]. Thus naturally it would be expected that FHR patterns differ between Stage I and II of labor due to the temporal differences in the occurrences of uterine contractions and fetal movement.

Recent machine learning work has found significant differences in model performance and parameter estimates when separating FHR data by labor stage. Spika's work explored the impact of separating data by labor stage on the use of FHR to classify cases of healthy fetuses, and those with acidosis [15]. Not only did creating separate classifiers improve performance, the respective models preferred different input features. In a separate paper, Granero-Belinchon shows that several entropy measure estimates are significantly different when FHR data is separated by labor stage [10]. These observations suggest that the ACOG guidelines may need revision to specify separate recommendations for interpreting FHR data during each stage of labor.

Problem Statement

In summary there is a need to determine if and how normal FHR characteristics differ between Stage I and II of labor. Previous work has been done using machine learning techniques to differentiate heart rate characteristics for both fetuses and adults [3, 5, 7, 10, 11, 13, 14, 15]. Models such as these can be evaluated based on various measures of accuracy/purity and compactness. My objective will be to investigate the link of FHR characteristics to their origin in time within the first two stages of childbirth. This will be done in two main ways after labelling sample FHR time-series as originating from within Stage I or II of labor (i.e. binary labels).

- **Supervised Learning:** I will construct a benchmark decision tree model. This is because information theory estimates of entropy (linked to the concept of information gain) have been shown to be different for FHR in different stages of labor [REF]. I will create an optimized random forest classifier which will also

include unsupervised cluster labels as inputs.. Hopefully this can teach my classifier more about the underlying structure of the data. In either case I will consider classifier performance with an accuracy of > 50% to show a significant link between FHR and labor stage.

- **Unsupervised Learning:** I will evaluate a k-means (benchmark) and a hierarchical clustering model. I will calculate descriptive statistics of each cluster. I will consider the relationship between a cluster and the assigned labels to be significant if the resulting Adjusted Rand Index (ARI) of the clustering is > 0. Other clustering quality metrics such as purity and normalized mutual information will be calculated.

Further discussion about my approach to this problem will be described in the following sections.

Dataset

I will attempt to cluster FHR for the following time periods during gestation:

- Stage I of Labor
- Stage II of Labor

I will be using the open source Intrapartum CTG dataset available on PhysioNet [9, 17]. This data was all collected at the University Hospital in Brno, Czech Republic. This dataset has continuous FHR time-series readings for at most 90 minutes while also being at most 90 minutes before labor. Recordings also met the following timing criteria for Stage I and II of labor:

- Stage I recordings were at least 30 minutes, and at most 60 minute.
- Stage II recordings to delivery were kept to at most 30 minutes.

There were a total of 552 subject recordings. I will only be including the records of patients having an umbilical pH > 7.15 after birth. This criteria is used to determine if a Fetus has respiratory hypoxia, one of more common disorders that may be detected using FHR monitoring. I will only include subjects that were delivered Vaginally for my main analysis.. Caesarian Sections will not be considered in the main analysis because they may act as confounding variables to normal FHR patterns. I may include them in supplemental exploration.

Furthermore, the organizers of the dataset tried to make it as homogenous as possible by only including full-term infants (>37 weeks of gestation), only including infants free of opiate administration, and only including singleton pregnancies. This will make the included number of participants equal to 417. Furthermore, in order to have samples distinctly from each phase of labor, maintain balanced classes, optimize sample size, and ignore readings that are very close to the beginning or ending of the FHR monitoring I will also do the following.

- Header information details the length of Stage I and II of each patient. This information will be used to exclude time-series data ±5 minutes from the boundary between the two stages.
- Ignore 2.5 minutes from the beginning and end of each time-series.
- Ignore participants with Stage II lasting less than 12.5 minutes.
- Collect an equal number of Stage I and II FHR time-series samples from included participants For participants with lengthy recordings, multiple samples will be acquired. This will occur by randomly selecting a contiguous time-series rounded down to the nearest 5 minutes of Stage II length (e.g. For a participant with a 23 minute long Stage II, I randomly select a 20 minute sample of both Stage I and II to use in my analysis). Ten minute segments will be collected with overlapping windows of 5 minutes.

Final numbers on sample-sizes would be available once I am approved for the project, and look for subjects meeting these requirements. I would expect ~417 (the same number of included participants before looking at time-series requirements) since some subjects would be unusable, but others may allow for the inclusion of multiple samples.

Input Features

I will use a total of eight time and frequency domain features.

Time Domain Features

- **Baseline Heart Rate (BHR)**: Mean heart rate over a particular time period. From initial literature review ~10 minutes.
- **Baseline Heart Rate Variability (HRV)**: Standard deviation from baseline heart rate. From initial literature review ~10 minutes

- **Number of Accelerations (NAC)**: From initial review an acceleration tends to be defined as a rise of 15 beats per minute for at least 15 seconds.
- **Number of Decelerations: (NAD)** From initial review a deceleration tends to be defined as a decrease of 15 beats per minute to a minimum within 30 seconds.

Frequency Domain Features

These will be calculated using the squared absolute value of the Fast Fourier Transform. These are based on the frequencies used in [REF].

- **Power estimated from very low frequency bands (VLFB)** (<0.03Hz)
- **Power estimated from low frequency bands (LFB)** (0.03-0.15Hz)
- **Power estimated from high frequency bands (HFB)** (0.15-Nyquist Frequency)
- **Ratio of LFB/HFB**

Evaluation Metrics

Supervised Learning:

- **Accuracy:** Since my classes are balanced I think accuracy will be a reasonable measure to use. I will consider a model to show a potential link between FHR and Labor Stage if accuracy is > 0.5. This is because at random choosing all samples to be of either class would result in an accuracy of 0.5.
- In order to determine if a particular class is being more misclassified I will report a **confusion matrix** and respective **F1-Score.**

Both of these metrics will be reported as the average result from k-fold Cross validation. I will be using 10 folds.

Unsupervised Learning:

Main Metric

- **Adjusted Rand Index (ARI)**: This will be my main measure to determine if clustering is related to my external binary labels of labor stage. ARI will measure if the each model puts pairs of similarly labeled examples within the same clusters. I will consider ARI values > 0 to be significant since this would correspond to better clustering compared to what could be expected by chance.

Supplementary Metrics

These metrics will be used to compare the k-means vs hierarchical models against each other in terms of their ability to cluster. Higher Dunn Index, Purity, and Normalized Mutual Information measures would indicate better clustering characteristics.

- **Dunn Index**: This will be used to determine how compact and well separated the models produce clusters. This can be defined mathematically as the minimum distance between two clusters divided by the maximal diameter of a cluster.
- **Purity:** I will determine the proportion of each cluster coming from time periods within Stage I and II labor respectively.
- **Normalized Mutual Information**: Another metric to see if clustering is really related to the classifying the data by their Stage I or II time origin. I will use the implementation of NMI found in the sklearn.metrics.cluster library. NMI can describe the reduction in entropy of class labels based on if we know cluster labels.

In addition to the above metrics I will provide some descriptive statistics on the input features of the samples within each cluster. These results will be informative even if clusters are found not to be linked to labor stage.

Benchmark Model

Supervised Learning:

As described previously, I will use a decision tree model as my benchmark. This is because measures of entropy have shown differences when looking at FHR in relation to labor stage. I will use a minimum split size of 1/15th the training sample size, and a minimum leaf size of 1/10th the training sample size. I hope that these estimates will limit some potential overfitting by providing enough samples within each leaf/node, and limit the depth of the tree. I will visualize the tree in order to determine what features seemed to be the most important in terms of information gain.

Unsupervised Learning:

A k-means model with two clusters will be used as the benchmark model to which my hierarchical solution will be compared. I have chosen K-means since there seems to be several papers using it to cluster Heart Rate Data [2, 6]. I have chosen K = 2 because I hypotheses that there may be two main clusters each related to examples from Stage I and II of labor respectively.

Solution Statement

Supervised Learning: I will develop a random forest model to optimize my supervised classification of the FHR data. This is because an ensemble of decision trees will still incorporate information gain while helping to limit the occurrence overfitting. I will include unsupervised learning clustering labels as an additional inputs to my optimized model if clustering is significant based on ARI. I will use a model of 50 trees with the same leaf and split requirements as the benchmark. I will include all of my features for each tree. I will examine the feature importance attribute to see if these features differ from the visualization of my benchmark tree.

Unsupervised Learning: I will develop a Hierarchical Clustering model to evaluate against my benchmark k-means model. I believe a hierarchical model should be able to provide clusters that are as homogenous, compact, and separated at least as well as the benchmark model. Furthermore, hierarchical clustering may be able to provide more insight into FHR patterns due to its ability to be easily visualized as a dendogram. This will be especially useful to see what input features differentiate clusters.

Project Design

My project will be consist of 4 major steps. This will include data importing and sorting, computation of input features, development of (un)supervised models,, and lastly solution evaluation.

As mentioned previously, data from the Intrapartum database will be used to build the machine-learning models. I will do this by either using PhysioNet's python library, or find a way to manually import the FHR time-series, and header info into Pandas Dataframes. I will then evaluate the header information to conduct exclusion of participants based on the criteria specified in the Dataset section of this proposal. I will then create a function that has the ability to take in a time-series examine it for my class balance criteria, and extract the appropriate number of time-series samples. I will use this function to loop through all the subjects, and store the extracted samples in a separate DataFrame or Numpy matrix. Separate functions will be created to extract the input features described in the Inputs section of this proposal.

After feature computation, I will build the models and evaluate their performance. The specifics for these models and their evaluation has been described in the Solution Statement and Evaluation Metrics Sections of this proposal. Based on my evaluation metrics I will discuss the success or inability of these models to show a link between FHR and Labor Stage. I will do this by stating if performance met my thresholds of significance. Furthermore, if my supervised models meet my significance thresholds I will discuss how their feature importance compares to observations in clinical literature. If the models do not meet my significance threshold I will at least discuss why this may be occurring due to the

structure of data seen in my unsupervised clustering, and potential improvements that could be made to my supervised classifier.

References

1. Alfirevic, Z., Devane, D., & Gyte, G. M. (2006). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev*, *3*(3).
2. American College of Obstetricians and Gynecologists. (2009). ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. *Obstetrics and gynecology*, *114*(1), 192.
3. Baldzer, K., Dykes, F. D., Jones, S. A., Brogan, M., Carrigan, T. A., & Giddens, D. P. (1989). Heart rate variability analysis in full-term infants: spectral indices for study of neonatal cardiorespiratory control. *Pediatric research*, *26*(3), 188.
4. Beard, R. W., Filshie, G. M., Knight, C. A., & Roberts, G. M. (1971). The significance of the changes in the continuous fetal heart rate in the first stage of labour. *BJOG: An International Journal of Obstetrics & Gynaecology*, *78*(10), 865-881.
5. Chan, H. L., Fang, S. C., Ko, Y. L., Lin, M. A., Huang, H. H., & Lin, C. H. (2006). Heart rate variability characterization in daily physical activities using wavelet analysis and multilayer fuzzy activity clustering. *IEEE Transactions on Biomedical Engineering*, *53*(1), 133-139.
6. Electronic Fetal Heart Monitoring: Healthwise Medical Information on eMedicineHealth. (n.d.). Retrieved January 27, 2018, from https://www.emedicinehealth.com/script/main/art.asp?articlekey=129181
7. Ferrario, M., Signorini, M. G., & Magenes, G. (2009). Complexity analysis of the fetal heart rate variability: early identification of severe intrauterine growth-restricted fetuses. *Medical & biological engineering & computing*, *47*(9), 911-919.
8. Freeman, R. K. (2002). Problems with intrapartum fetal heart rate monitoring interpretation and patient management. *Obstetrics & Gynecology*, *100*(4), 813-826.
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).
10. Granero-Belinchon, C., Roux, S. G., Abry, P., Doret, M., & Garnier, N. B. (2017). Information Theory to Probe Intrapartum Fetal Heart Rate Dynamics. *Entropy*, *19*(12), 640.
11. Inbarani, H. H., Banu, P. N., & Azar, A. T. (2014). Feature selection using swarm-based relative reduct technique for fetal heart rate. *Neural Computing and Applications*, *25*(3-4), 793-806.
12. Langer, B., Carbonne, B., Goffinet, F., Le Gouëff, F., Berkane, N., & Laville, M. (1997). Fetal pulse oximetry and fetal heart rate monitoring during stage II of labour. 1. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, *72*(1), S57-S61.
13. Phongsuphap, S., Pongsupap, Y., Chandanamattha, P., & Lursinsap, C. (2008). Changes in heart rate variability during concentration meditation. *International journal of cardiology*, *130*(3), 481-484.
14. Spilka, J. (2013). Complex approach to fetal heart rate analysis: A hierarchical classification model. *Czech Technical University, Faculty of Electrical Engineering, Prague*, 35-47.
15. Spilka, J., Leonarduzzi, R., Chudácek, V., Abry, P., & Doret, M. (2016, November). Fetal Heart Rate Classification: First vs. Second Stage of Labor. In *Proceedings of the 8th International Workshop on Biosignal Interpretation, Osaka, Japan* (pp. 1-3).
16. Stages of labor and birth: Baby, it's time! (2016, June 22). Retrieved January 26, 2018, from https://www.mayoclinic.org/healthy-lifestyle/labor-and-delivery/in-depth/stages-of-labor/art-20046545
17. Václav Chudáček, Jiří Spilka, Miroslav Burša, Petr Janků, Lukáš Hruban, Michal Huptych, Lenka Lhotská. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth* 2014 **14**:16.
18. Williams, K. P., & Galerneau, F. (2003). Intrapartum fetal heart rate patterns in the prediction of neonatal acidemia. *American Journal of Obstetrics & Gynecology*, *188*(3), 820-823.