Predictive Analytics and Fetal Heart Rate Monitoring: An Analysis

Brandon May

Executive Summary

Whenever a woman goes into labor, the infant's heart rate is usually continuously monitored in the labor and delivery unit. Physicians monitor the heart rate for specific patterns that indicate baby well-being. The question remains, can machine learning be used to analyze these heart rate tracings as good or better than a physician?

This project sought to examine whether this could be true. The data was collected from over 2000 live births and compared to a physician's determination of whether a baby's heart rate tracing pattern was normal, concerning, or pathologic (grossly abnormal requiring intervention). The goal was to have the model predict whether the data indicated a normal, suspect, or pathologic condition.

This data was analyzed, redundant variables were eliminated, and the data was manipulated to use in the machine learning programs. Three models were used: K-Nearest Neighbors, which predicts the target variable by computing how close the data are mathematically to each other. Another method called a Decision Tree was used which breaks down the data into decision points to accurately classify the target. Finally, a Random Forest algorithm was used which is many decision trees factored together.

Both the KNN and Random Forest Methods had high performance indicating that one of these models could be accurate in prediction of the categories of these data. Further analysis

with a larger dataset would need to be performed for validation before initial deployment of the algorithm to be used in conjunction with expert physician assessment.

Abstract:

Neonatal mortality rates have remained steady for the last several years. In the United States, this trend is worrisome. One way of ensuring fetal well-being is via external fetal cardiotocography which measures the heart rate of the infant. Certain findings during labor are reassuring while others are indicative of possible fetal distress. This project sought to determine an accurate and precise machine learning algorithm using data obtained from the UCI Machine Learning database and is composed of various measurements of the heart rate tracings to generate a prediction of normal, suspect, or pathologic findings. Three models were attempted including K Nearest Neighbors, Decision Trees, and Random Forest Models. Both KNN and Random Forest models had the highest F1, Precision, and Recall scores. Given the limited and outdated data points available in this project, further validation of these models would need to be performed with a larger, more diverse dataset before beginning initial deployment to a live birthing unit environment.

Introduction:

Neonatal mortality continues to be a concern in the United States. Despite medical advancements in fetal monitoring, there has not been much progress in affecting neonatal outcomes and between 2014 and 2016, there was no change in perinatal mortality according to the CDC (Gregory *et al.*, 2018). When a woman goes into labor and is admitted to a labor and delivery unit at a hospital, a standard set of protocols are enacted. Typically, the standard of care is continuous fetal heart rate monitoring via external cardiotocography. This fetal heart rate monitor measures the individual heart beats of the infant. A normal heart rate for a fetus is 110-160 beats per minute. At the onset of a contraction, it is considered normal for the heart rate to increase or accelerate. A heart rate decrease, or a deceleration, can be considered pathologic in certain circumstances. It is also normal for there to be a certain amount of beat to beat variability and as long as within a certain range, this is reassuring. Heart rates that are held at a constant rate are more concerning. Some types of heart rate decreases in comparison with a contraction can be concerning and others can be considered normal (Miller, 2020).

With the advent of more sophisticated machine learning techniques, the question remains on if predictive analytics could be used to improve identification of potentially worrisome fetal heart rate tracings. Expert obstetricians are able to interpret these; however, standard tracings can be difficult to read. The heart rate monitor can stop measuring heart beats if there is a shift in infant position and you can even lose the fetal heart rate entirely in which you must resort to a fetal scalp electrode to get an accurate reading. If analytics were able to better identify suspect patterns, perhaps prompt intervention could help improve outcomes.

Some studies have not shown a significant benefit to <u>continuous</u> fetal heart monitoring in reducing infant mortality including large meta analyses in the medical literature. Further, continuous fetal heart rate monitoring has been associated with an increase in instrument-assisted births and Cesarean sections (Alfirevic, 2017). There is some debate in these medical analyses if intervention could have prevented adverse outcomes. Some studies specifically look at the incidence of cerebral palsy, which is typically caused by a lack of oxygen to the fetal brain. One study did show a decrease in perinatal morbidity and mortality, though the evidence overall appears that this may not be beneficial (Blander, 2006).

Therefore, there could be significant benefits including saving lives and preventing morbidity if a successful machine learning model could be used to assist physicians. A successful model would have to have an extremely high degree of accuracy and precision to limit the number of false positives and false negatives. You would not want an algorithm to recommend an intervention when there are no problems in labor; conversely, you do not want a false negative to occur in that fetal well-being was indicated when fetal distress was indeed evident.

Methods:

Data was collected via 2126 fetal cardiotocograms for women undergoing active labor and collected between the years of 1995 and 1998 and interpreted by both an automated system (SIS Porto 2.0) and three expert obstetricians. This data is housed in the University of California-Irvine machine learning database and available for public review (https://archive.ics.uci.edu/ml/datasets/Cardiotocography). The variables included file identifiers, dates, and various measurements including metrics of short-term/long-term variability, histogram characteristics, and heart rates among others. These were used to generate either a 10 point or 3-point grading system. The 3-point grading system was classified as 1 being normal, 2 being suspect, and 3 being pathologic. The 10-point scale included identifiers to types of fetal sleep, acceleration/deceleration patterns, and other pathologic patterns.

The primary outcome of this data was the accuracy of the 3-point classifier. The 10-point classifier variables were included in the overall model to see if they impacted the performance of the 3-point classifier predictions. Using R, data was loaded, cleaned, and exploratory analysis was completed analyzing variable distributions, correlations, and relationships. After exploratory data analysis, the data was transitioned over to Jupyter notebooks using Python to run machine learning algorithms. The models were judged based on Precision, Recall, and the F1 score based on imbalances in the target class. K Nearest Neighbors, Decision Tree, and Random Forest algorithms were run with and without the 10-point classification system as predictors.

Results:

Overall, there were 3 variables describing personal case identifiers and dates of measurement, 13 categorical variables, and 24 numerical variables. There was summary information at the end of the raw data that was not explicitly labelled or identified and were removed from analysis. A full listing of the variables and their meanings are attached in the Appendix of this document.

The data was loaded into R Studio and the variables were analyzed by their summary statistics. The 13 categorical variables were recoded as factors for further analysis. Baseline fetal heart rate measurements by both expert and automated determination were both normally distributed. Many of the variables were positively skewed including fetal movements, accelerations, abnormal short and long-term variability, and decelerations. Many of the categorical variables including the classifiers were unbalanced with the majority of the values being within normal limits. The 3-classifier target variable of NSP (Normal, Suspect, or Pathologic) was also unbalanced with the majority of cases being normal and a minority of cases being suspect or pathologic. Variables with skew were deliberately not transformed due to the need for this machine learning algorithm to be predictive with outliers.

Bivariate boxplots were used to compare the target variable (NSP) with the individual variables. In the suspect population, there appeared to be more outliers with higher values in both the beginning and ending measurement variables. Fetal heart rates by both experts and automated analysis appeared to have higher hearts in the pathologic and suspect categories. In the suspect and pathologic categories, the number of accelerations were extremely low with few positive values. The fetal movements/second and the target variable appeared to similar across the classifications. The number of uterine contractions/second appeared to generally be lower in the suspect and pathologic categories. Light decelerations/second appeared to be much higher in the pathologic category. Severe decelerations/second were much higher in the pathologic category. There was a trend towards more percentage of time spent with abnormal short and long-term variability in the suspect and pathologic categories. There was a trend towards much lower histogram minimum values in the pathologic category. Conversely, the histogram minimum value tended to be higher in the suspect category as compared to normal. When comparing the histogram mode, mean, and median values, there was a trend towards more frequent lower values of the fetal heart rate in the suspect category. When looking at the histogram variance, there were higher variances in the pathologic category.

Using a correlation plot, there appeared to be high levels of correlation between the beginning and ending measurements variable in addition to high levels of correlation between expert and automated determination of fetal heart rate in addition to the mode, mean, and median variables. Using a cut-off correlation value of 0.9, the variables LBE, e, and Median were dropped from the dataframe.

Including the 10-point classification system as predictors for the target variable resulted in the highest F1, Precision, and Recall scores for the K Nearest Neighbor, Decision Tree, and Random Forest models.

The best KNN model had best results with number of neighbors 3, uniform weights, used the Minkowski metric, and p of 1, which resulted in an F1 score of 1.00 for Normal, 0.98 for Suspect, and 0.98 for Pathologic.

The Decision Tree model had the best results after hyperparameter tuning using the entropy criterion, minimum splits at 2 and minimum leaf samples of 1 with an F1 score of 0.99 for Normal, 0.94 for Suspect, and 0.98 for Pathologic.

The Random Forest model, being an ensemble of Decision Trees, performed better than the Decision Tree model with the gini criterion, minimum leaf samples of 1, and minimum sample splits of 2. It had an F1 score of 1.00 for normal, 0.98 for suspect, and 0.98 for pathologic.

Interestingly, removal of the 10-point classifier actually caused all three models to have worsening performance.

Conclusions:

The findings of this project indicated that machine learning algorithms employed against fetal heart rate monitoring could potentially be effective in predicting our target variable, whether the fetal heart rate was determined to be normal, suspect, or pathologic.

There are some limitations to this data. It was collected between 1995 and 1998 so it is outdated. In the space of 20 years, there could be a significant difference in maternal obstetric care and demographics. There is also not much information regarding the demographics of the population. A mother with more medical risk factors is expected to have a higher risk pregnancy as opposed to a woman with no medical risk factors. While there can be fetal harm associated with normal pregnancies, a higher risk pregnancy has to be monitored and treated appropriately. Therefore, there needs to be a wide range of data covering a diverse patient population with a variety of medical risk factors to make this algorithm most applicable to real-life labor and delivery units.

There are approximately 2000 records with multiple variables though to deploy a machine learning predictor on a large scale would require a significant more test cases in order to fine tune the machine learning accuracy. As discussed above, to be using an automated machine learning algorithm to categorize fetal heart rate strips would have no room for error. A single misclassification could result in irreversible fetal harm. However, it is promising that the advent of machine learning could help aid expert physicians in risk assessment of questionable fetal heart rate strips.

One finding that bears mentioning is that the variable for repeated decelerations were all 0's meaning there were no cases with repetitive decelerations, which could be concerning.

The vast majority of the records were read as normal. A smaller minority were rated as suspect and the smallest proportion was pathological. For practitioners, this is a good thing because this means that less pathologic conditions were identified which is beneficial for an infant's health. However, for prediction, the algorithm will have to be altered to adequately assess suspect and pathologic conditions since assumption of normality would be expected by chance to the distribution.

Further, on exploratory analysis of the input variables to the target variable of NSP, there were several trends that are consistent with known associations of pathologic categories. There was higher variance of values in the histograms of the fetal heart rates in the pathologic category. There also tended to be higher amounts of time with percentage of time spent with abnormal short and long-term variability. The number of uterine contractions/second were lower in the pathologic category. The number of decelerations were higher in the pathologic category and FHR tended to be lower in the pathologic categories as well. This validates the common findings that are associated with pathologic fetal heart rates.

Both the KNN and Random Forest Models performed almost equally well. Using the 10-point classifier, which could be used as a target itself, helped predict our 3-class target variable better than without it. One possible explanation is that some of the information in the 10-point classifier could help the algorithm point to a suspect or pathologic target class and provides more information for the algorithm to make a judgment.

Further model deployment would be further testing with the categories used in this analysis on a larger subset of patients using the SIS Porto 2.0 system (or a newer version if available). There would need to be thousands more datapoints and patients analyzed using these criteria to further validate the model with expert Obstetrician oversight to interpret the results in accordance with current clinical guidelines and standard of care. This is required to determine the target variable on whether a fetal heart tracing is normal, suspect, or pathologic which can only be known after successful delivery of an infant. Another future direction would be to accumulate more data using the same measurements in this project but directly compare Random Forest vs. KNN models directly against one another. Given that Random Forest models are an agglomeration of decision trees, mirroring the decision points that clinicians make when interpreting a fetal heart tracing, this may be a more appropriate algorithm to use in the future.

Acknowledgements:

Special thanks to Bellevue University and Amirfarrokh Iranitalab, Ph.D. for support for this project. I would also like to thank the University of California – Irvine for hosting potential datasets in their UCI Machine Learning Database. The data was available and cleaned and made easy for use to attempt fitting machine learning algorithms and validate predictive accuracy.

Appendix A - Variable Descriptions:

FileName = Case Identifier

SegFile = Case Identifier

Date = Date of Measurement

B = Starting Point of Measurement

E = Ending Point of Measurement

LBE = Baseline Fetal Heart Rate (Determined by Expert)

LB = Baseline Fetal Heart Rate (Automated)

AC = Number of Accelerations/Second

FM = Number of Fetal Movements/Second

UC = Number of Uterine Contractions/Second

DL = Number of Light Decelerations/Second

DS = Number of Severe Decelerations/Second

DP = Number of Prolonged Decelerations/Second

ASTV = Percent of Time with Abnormal Short-Term Variability

MSTV = Mean Value of Short-Term Variability

ALTV = Percent of Time with Abnormal Long-Term Variability

MLTV = Mean Value of Long-Term Variability

Width = Width of Fetal Heart Rate Histogram

Min = Minimum of Fetal Heart Rate Histogram

Max = Maximum of Fetal Heart Rate Histogram

Nmax = Number of Histogram Peaks

Nzeros = Number of Histogram Zeros

Mode = Histogram Mode

Mean = Histogram Mean

Median = Histogram Median

Variance = Histogram Variance

Tendency = Histogram Tendency (-1 = Left Asymmetric, 0 = Symmetric, 1 = Right Asymmetric)

CLASS – FHR Pattern Class Code

A = Calm Sleep

B = REM Sleep

C = Calm Vigilance

D = Active Vigilance

AD = Accelerative/Decelerative Pattern (Stress Simulation)

DE = Decelerative Pattern (Vagal Stimulation)

LD = Largely Decelerative Pattern

FS = Flat Sinusoidal Pattern (Pathologic State)

SUSP = Suspect Pattern

NSP = Fetal State Class Code (1 = Normal, 2 = Suspect, 3 = Pathologic)

References Cited:

Alfirevic, Z., Gyte, G. M., Cuthbert, A., & Devane, D. (2017). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database of Systematic Reviews*. doi: 10.1002/14651858.cd006066.pub3

Blander, C. L. (2006). Intrapartum Electronic Fetal Heart Rate Monitoring and the Prevention of Perinatal Brain Injury. *Obstetrics & Gynecology, 108*(6), 1554.doi: 10.1097/01.aog.0000250069.31615.15

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gregory, E. C., M.P.H., Drake, P., M.S., & Martin, J. A., M.P.H. (2018). Lack of Change in Perinatal Mortality in the United States, 2014-2016. *NCHS Data Brief No. 316*.

Miller, D. A., MD. (2020, May 01). Intrapartum fetal heart rate monitoring: Overview (V. Berghella MD & V. A. Barss MD, FACOG, Eds.). Retrieved June 11, 2020, from https://www.uptodate.com/contents/intrapartum-fetal-heart-rate-monitoring-overview.