PROJECT FINAL REPORT: PREDICTING FETAL HEALTH USING CLASSIFICATION WITH CLINICAL DATA

A PREPRINT

Samy Dahman
CS 4774: Machine Learning
University of Virgnia
Charlottesville, VA
sd7fpy@virginia.edu

Drew Parks
CS 4774: Machine Learning
University of Virginia
Charlottesville, VA
dap7tf@virginia.edu

Kaidi Zhang
CS 4774: Machine Learning
University of Virginia
Charlottesville, VA
kz6xc@virginia.edu

May 21, 2021

ABSTRACT

Fetal health has long been an issue the medical world is trying to predict in the early stage of pregnancy as spotting the problem early on can leave time for the health care professionals and parents to prepare for the challenges. Hence, our team decides to use the data set taken from Kaggle, entitled Fetal Health Classification [2], to train and learn the model of classifying fetus' health.

The three categories of the fetus' health are suspect, healthy, and pathological. We tested various methods, including soft voting ensemble learning, random forest, Ada-boosting with decision tree, and artificial neural network. The best result is produced by random forest, which gives an average recall of 0.93 and an average precision of 0.91. Specifically, the recall of pathological class is 0.966, indicating that we can detect over 96% of pathological fetus, which is a very promising result. The result is shown below 1.

Table 1: Metrics of Random Forest Model

Metric/Class	Healthy	Suspect	Pathological
Precision	0.967	0.898	0.903
Recall	0.976	0.828	0.966
f1 score	0.972	0.862	0.933

1 Introduction

The problem we are tackling is fetal-health in the womb. Here, we're trying to determine if a baby is "Normal", "Suspect", or "Pathological". We will do this using a number of different features like number of uterine contractions and fetal movement. This is very much a practical experiment where hospitals in Virginia can determine whether or not a mother needs more sophisticated scans done just based off the basic information of the features. The goal is to create a model that can use the features and correctly classify fetal health and thus allowing hospitals to better understand the fetus' current state in the womb.

The data set contains 2126 entries with over 10+ features on fetal health. The label is stored in the column "fetal health", in which 1, 2, and 3 represent healthy, suspect, and pathological fetus, respectively. The features include general clinical data, such as fetal heart rate, fetal movements, uterine contractions, histograms generated by cardiotocograms, etc. Around 78% of the data is on healthy fetus, and the rest is on suspect and pathological fetus. The incomplete data are removed from the data set before analyzing.

2 Method

In terms of data pre-processing, we did a standard test/train split with 20% test data and 80% training data. We experimented with several different methods that we learnt from class for classifying fetal health. I will outline each in the order that we experimented with them:

1. **Ensemble:** We started with what was presumed to be the best model, ensemble learning. For this, we used a combination of logistic regression, random forest, and support vector machine with a soft voting implementation. We then assessed the performance of each of these by looking at the accuracy of predictions as well as F1 score, precision and recall for each class. From this section, the model that performed the best was actually the Random Forest model, out-competing the ensemble model and the other two that composed the ensemble.
2. **Random Forest:** At this point, we decided to look at the effectiveness of Random Forest on it's own. The hyper-parameter setting we used was n-estimators set to 100 and no max depth. We found the performance of this model to to remain better as seen in the ensemble model. We then used a grid search to find the optimal parameters that is further explained in experiments. Ultimately, the optimal parameters are: max depth of 80, max features 8, max leaf nodes 100, min samples leaf of 2, min samples split of 3, and n estimators of 325.
3. **AdaBoosting:** We then looked at Ada-boosting with a decision tree model of 200 n-estimators and max depth of 2 that performed quite well but not as well as random forest.
4. **Artificial Neural Network:** Lastly, we tried a deep learning model in the hopes that it would surpass the Random Forest model. The ANN had 3 dense layers with ReLu activation in addition to the input layer that was fed the 22 dimensional data as well as the output layer with a softmax activation. We experimented with several values for the number of nodes in the dense layers and found 512 to be optimal.

3 Experiments

We aim to avoid false negatives at all costs so as to avoid diagnosing pathological health conditions as healthy. Such detrimental conditions cannot be allowed to go unnoticed, so our aim is to maximize the recall of the pathological health class. While the easiest way to do this would be to flag every instance as pathological, this of course comes at great cost to the accuracy and precision of the model. This is why the f1 score is a commonly used metric when evaluating a binary classifier. As seen in equation 1, the f1 score is inversely proportional to the sum of the precision (p) and recall (r) inverses, such that f1 approaches one only as both precision and recall approach one.

$$f_1 = \frac{2}{\frac{1}{p} + \frac{1}{r}} \tag{1}$$

While the f1 score is generally a good metric, it gives equal weight to precision and recall, and is above only defined for a binary classifier with a single recall and precision value. We create a new scoring function to fix these issues which we call f2, as seen in equation 2.

$$f_2 = \frac{1+\alpha}{\frac{1}{\langle\vec{p}\rangle} + \alpha\frac{1}{\vec{r}\cdot\vec{w}}} \tag{2}$$

In equation 2, $\langle\vec{p}\rangle$ denotes the average precision, and if the classifier has q different classes to identify the weight vector w is such that $\sum_{i=1}^{q} w_i = 1$. Imposing this normalization condition on $\vec{w}$ ensures that the f2 score remains in the interval [0,1]. In our case we have 3 classes, each with its associated recall to make up the recall three-vector $\vec{r}$. Similarly, the precision values for each class make up the precision vector $\vec{p}$. While the components of $\vec{p}$ are averaged, the dot product of $\vec{w}$ and $\vec{r}$ gives a weighted average of the recall. This weighted average is used to give more weight to the recall of the pathological health class as opposed to the healthy classes. The constant α then acts as a multiplicative factor on the $1/(\vec{r}\cdot\vec{w})$ term such that more weight is given to recall rather than precision when scoring the classifier. Adding this α term to the numerator, this new scoring metric will scale between 0 and 1 for an appropriately normalized weight vector, where a score of 1 indicates perfect recall and precision.

This f2 score was implemented in our fine tuning process by using the *make_scorer* function from *sklearn.metrics*. We conducted a randomized search and then a grid search with cross validation on our model, using *make_scorer* to use equation 2 as our scoring function during the search process. The success of this procedure is shown in table 2, in which the pathological class recall increased by over six percent after searching using $\alpha = 5$ and $\vec{w} = \langle 1, 1, 5\rangle$ (before $\vec{w}$ normalization).

Table 2: Compared Metrics of Initial and Optimized Model

Metric\Class	Healthy	Suspect	Pathological
Recall	-0.006	0.016	0.069
Precision	0.000	0.017	0.007
F1 score	-0.003	0.016	0.037

For each scoring metric in the left column, the values shown in the corresponding row are those metric values of the optimized model minus the metric values of the initial model, for each of the three classes. Note that optimizing using the f2 score has made it so that the most improved metric is the recall of the pathological class, which increased 6.9%! This increased recall was at no cost to the overall accuracy of the model either, as the the overall accuracy increased slightly by 0.2%.

A similar optimization approach was made with an artificial neural net (ANN) developed using the *Keras* open source library. We defined a Python function that could build a sequential model that would make a model of a certain depth, with certain drop out layers and activation functions as well,corresponding to the input passed to this function. A *KerasClassifier* wrapper was then used so that this model could be passed to the *scikit learn GridSearchCV* function. With this implementation, we were able to search over ANNs of different depths, with varying dense layer nodes, drop out ratios, and activation functions. While this practice obtained an overall test accuracy as high as 95% with up to 93% recall on the pathological health class, the random forest classifier performed slightly better on both metrics.

4 Results

The superior model with both the highest overall accuracy and the highest recall on the third class is the optimized random forest classifier. As multi-classification results may be depicted in a variety of ways, for the clarity of our results the confusion matrix may be seen in equation 3, with the classes ordered healthy, suspect, then pathological. Common scoring metrics extracted from this confusion matrix are displayed in table 3.

$$confusion = \begin{pmatrix} 325 & 6 & 2 \\ 10 & 53 & 1 \\ 1 & 0 & 28 \end{pmatrix} \tag{3}$$

Table 3: Optimized Random Forest Model Test Results

Metric\Class	Healthy	Suspect	Pathological
Recall	0.976	0.828	0.966
Precision	0.967	0.898	0.903
F1 score	0.972	0.862	0.933
Accuracy	0.976	0.828	0.966

For each scoring metric in the left column, the values shown in the corresponding row are those metric values of the optimized model for each of the three classes. The overall test accuracy of this model is recorded at 0.953%.

The overall accuracy is high enough to be useful for it's intended purpose, at over 95%. It should be noted however, that as there is not an even number of classes, this accuracy should be compared to the 78% accuracy which would be obtained from always assuming the fetal health to be healthy. The recall for the pathological class on the test set is also remarkably high, at just under 97%, meaning exceptionally few pathological health fetuses will be mis-classified. Code for this project may be found here.

5 Conclusion

Overall, the results of our machine learning project exemplify the idea that the capability of data is much higher than we believe. In the sense that the data collected from this relatively inexpensive scan called the cardiotocogram leads to far better modelling capability than we had believed. That's all due to the power of data analysis methods like machine learning. We set out doing this with the hypothesis that we would be able to predict fetal health to large extent, values over 80% recall and 85% accuracy. However, we exceeded this for all our classes as outlined in the results section above. Thus, we have accepted our hypothesis and now begin to examine the implications for the state of Virginia. Health care costs in Virginia and the United States as a whole is much higher than most of the developed world. A report done in 2020 says that "In 2018, the U.S. spent nearly twice as much on health per person as comparable countries ($10,637 compared to $5,527 per person, on average)."[4]. The cardiotocogram is a relatively inexpensive scan that obtains all the features we used. This scan can be done in Virginia hospitals to help mothers gain insight on the health of their babies in a costly manner that is more inclusive of lower income people.

However, there were limitations and shortcomings to our experiments. One possible shortcoming may have been that the current data contains 10 features about the histogram generated by cardiotocograms out of total 21 features, implying that the model may put too much emphasis on the histogram and overfitting the training set. Removing redundant features may help train a more robust model in the real world application. For example, the data contains both histogram mean and median, we can remove one of them and see how the model performs. Meanwhile, although the random forest model generates great metrics with the current data, it might be due to overfitting as the data set is small. Also, as data set gets larger, the amount of space taken by the forest will grow dramatically. Therefore, testing more models with different selection metrics, such as alternative architectures for the artificial neural networks, soft voting classifiers, and naive Bayes classifiers, might result in more robust and applicable models. There also are more sophisticated deep learing techniques that may be above the scope of our class that could have been used on the data-set as the performance from the artificial neural net we used was surprisingly low compared to the Random Forest model. Another thing we could've done was stratified our sampling better as the distribution was around 77% healthy babies.

6 Contributions

This section outlines what each team member did. Firstly, Kaidi did all the data-cleaning and setting up for the machine learning methods that we were going to use. She also wrote both the "Abstract",and "Introduction" sections. Samy implemented the machine learning models, deciding to test a variety as described in the "Methods" section which he authored, then recommending one to optimize. He also authored the "Conclusion" section along with Kaidi. Drew handled the optimization and scoring of the best chosen model from Samy. He also describes this in the "Preliminary Experiments" section as well as the "Results" section.

References

[1] Akhan Akbulut, Egemen Ertugrul, and Varol Topcu. Fetal health status prediction based on maternal clinical history using machine learning techniques In *Computer Methods Programs Biomed*, 2018 Jun 14.

[2] Fetal Health Classification In *Kaggle*, 2021 Nov.

[3] Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318

[4] Cox, Nisha Kurani and Cynthia. “What Drives Health Spending in the U.S. Compared to Other Countries.” *Peterson-KFF Health System Tracker*, 28 Sept. 2020, www.healthsystemtracker.org/brief/what-drives-health-spending-in-the-u-s-compared-to-other-countries/: :text=In