EXTENDED ABSTRACT

Journal of Labor and Childbirth

Predicting Preterm Birth from Electrical Impedance Spectroscopy: A Comparative Study Using Different Machine Learning Algorithms

David Tian

Department of Automatic Control and Systems Engineering, The University of Sheffield, United Kingdom

Background

Preterm birth (PTB) refers to labour before 37 weeks of pregnancy. PTB can cause new-born deaths and long-term diseases for the survived babies. The cost of care for surviving premature babies during their first year of life could be much higher than for term babies. Therefore, predicting and preventing PTB accurately early during pregnancy can effectively improve birth outcomes. However, the current approaches: assessment of cervical length and determination of cervicovaginal fetal fibronectin, have low predictive accuracies for women with low PTB risk and nulliparous women. Electrical impedance spectroscopy (EIS) is a technology concerned with applying a sinusoidal current to a sample under test to measure its impedance over a specified frequency range. EIS has been a powerful technique for measuring the electrical properties of materials. In machine learning, the class distribution of a training set critically affects the performance of the model trained in that a model trained on an imbalanced training set tends to bias towards predicting the majority class for unseen patterns. An EIS dataset collected for PTB prediction is often imbalanced in that the majority of the patients in the dataset are on-term and only contains a small number of preterm patients.

Methods

This work proposes a machine learning methodology for training PTB predictors on an imbalanced cervical EIS training set collected during 20 to 22 weeks of pregnancy. The methodology integrates 3 pre-processing algorithms (random sampling with replacement, polynomial feature transformation and information gain feature selection) with 4 classification algorithms: logistic regression, single-layer neural networks, Naïve Bayes and random forest, respectively. The EIS dataset was randomly split into training (66%) and test (34%) sets using stratified sampling. The methodology was run 100 times on different training and test sets. Each time the 4 classifiers were trained on the training set and evaluated on the corresponding test set, respectively. After 100 iterations of the methodology, the best training set of the 100 training sets was selected according to the highest performance of the test set. Finally, the neural network which had been trained using the best training set was re-trained iteratively on the best training set to further improve the best performance. The electrical impedance Z is defined as a complex number:

$$\begin{aligned} Z &= |Z| \cdot e^{j\cdot(Arg(Z))} \\ &= |Z|\cos(\omega t + \Phi) + \boldsymbol{j} \cdot |Z|\sin(\omega t + \Phi) \\ &= X + \boldsymbol{j} \cdot Y \end{aligned}$$

where |Z| is the amplitude of Z; Arg(Z) is the anticlockwise angle +Θ from the positive real axis to Z;

$-2\prod \leq \Theta \leq 2\prod$ in radians; $Arg(Z)=\omega t + \Phi$; ω is the frequency of Z in radians/second; t is the time in seconds (s); Φ is the phase of Z ($-\prod \leq \Phi \leq \prod$ in radians and $\prod$=180 degrees). An EIS consists of 14 electrical impedances (EIs) measured at 14 frequencies. To predict PTB of woman, EIS is measured at the cervix a number of times. The average EIS of all the EIS measurements is used as a training instance. The amplitudes (14 features) and the phases (14 features) of the 14 EIs are the 28 features for machine learning.

The *random sampling with replacement algorithm* is applied to each class of an imbalanced EIS training set to create a *balanced training set*. Given an imbalanced EIS training set and the size k of the balanced training set, the algorithm randomly draws $\frac{k}{2}$ instances from each class with replacement; then, it merges the selected instances of both classes into a balanced EIS training set of size k. In this study, after iterative experimentation, k was set to 1984.

Polynomial features transformation constructs polynomial features of a given degree d from the 28 original features. The polynomial features have higher importance in classification of data than the original features. The polynomial features of a given degree d consists of all the products of the original features of the dataset up to d. For a dataset of n features, there are $\frac{(n+d)!}{n!d!}$ polynomial features of d. In this study, polynomial features of degree 4 were constructed. For a training set of 28 features, there are 35960 polynomial features of degree 4. The high dimensionality of the polynomial features causes serious overfitting of classifiers to the training set and training time would be too extremely long.

In order to overcome the overfitting and speed up training, *information gain feature selection* is performed on the polynomial features to select the most important features only. Information gain measures the non-linear dependency between a feature and the class variable of a dataset. In this study, the top-30 ranked polynomial features by information gain are selected to reduce a training set.

Logistic regression, single-layer neural networks, naive Bayes and random forest were trained on the reduced training set and

evaluated on the reduced test set respectively. After 100 iterations of the methodology, the best training set of the 100 training sets was selected according to the highest performance of the test set. Finally, the neural network which had been trained using the best training set was re-trained iteratively on the best training set to further improve the best performance.

Results

The best training set of the 100 training sets corresponds to the best testing performance of area under Receiver Operating Characteristic Curve (AUC) 0.75 of logistic regression (Table 1). A single-layer neural network of 5 neurons improved the best testing performance to AUC 0.78 after iterative re-training of the network on the best training set.

Table 1: The performance of the classifiers over 100 iterations of the methodology

Algorithm	**Training AUC (mean, min, max)**	**Testing AUC Mean (min, max)**
Logistic regression	0.76 (0.71, 0.83)	0.62 (0.53, **0.75**)
Random forest	1 (1, 1)	0.50 (0.35, 0.66)
Naïve Bayes	0.53 (0.48, 0.62)	0.44 (0.34, 0.52)
Neural Networks	0.52 (0.47, 0.65)	0.55 (0.52, 0.64)

Conclusion

The proposed methodology is an effective approach for predicting the PTB of women from their cervical EIS dataset collected during 20 to 22 weeks of pregnancy. A single-layer neural network re-trained on the best training set of the methodology has achieved the highest predictive performance compared with logistic regression, random forest and Naïve Bayes.