Preterm Birth Prediction: Deriving Stable and Interpretable Rules from High Dimensional Data[*]

Truyen Tran   TRUYEN.TRAN@DEAKIN.EDU.AU
Center of Pattern Recognition and Data Analytics
Deakin University, Geelong, VIC 3216, Australia

Wei Luo   WEI.LUO@DEAKIN.EDU.AU
Center of Pattern Recognition and Data Analytics
Deakin University, Geelong, VIC 3216, Australia

Dinh Phung   DINH.PHUNG@DEAKIN.EDU.AU
Center of Pattern Recognition and Data Analytics
Deakin University, Geelong, VIC 3216, Australia

Jonathan Morris   JONATHAN.MORRIS@SYDNEY.EDU.AU
[‡]*Sydney Medical School, The University of Sydney*
St Leonards, NSW 2065, Australia

Kristen Rickard   KRISTEN.RICKARD@SYDNEY.EDU.AU
Clinical and Population Perinatal Health Research
Royal North Shore Hospital, St Leonards, NSW 2065 , Australia

Svetha Venkatesh   SVETHA.VENKATESH@DEAKIN.EDU.AU
Center of Pattern Recognition and Data Analytics
Deakin University, Geelong, VIC 3216, Australia

Abstract

Preterm births occur at an alarming rate of 10-15%. Preemies have a higher risk of infant mortality, developmental retardation and long-term disabilities. Predicting preterm birth is difficult, even for the most experienced clinicians. The most well-designed clinical study thus far reaches a modest sensitivity of 18.2–24.2% at specificity of 28.6–33.3%. We take a different approach by exploiting databases of normal hospital operations. We aims are twofold: (i) to derive an easy-to-use, interpretable prediction rule with quantified uncertainties, and (ii) to construct accurate classifiers for preterm birth prediction. Our approach is to automatically generate and select from hundreds (if not thousands) of possible predictors using stability-aware techniques. Derived from a large database of 15,814 women, our simplified prediction rule with only 10 items has sensitivity of 62.3% at specificity of 81.5%.

---

1. Introduction

Every baby is expected at full term. However, still 10-15% of all infants will be born before 37 weeks as preemies (Barros et al., 2015). Preterm birth is a major cause of infant mortality, developmental retardation and long-term disabilities (Vovsha et al., 2014). The earlier the arrival, the longer the baby stays in intensive care, causing more cost and stress for the mother and the family. Predicting preterm births is highly critical as it would guide care and early interventions.

Most existing research on preterm birth prediction focuses on identifying individual risk factors in the hypothesis-testing paradigm under highly controlled settings (Mercer et al., 1996). The strongest predictor has been prior preterm births. But this does not apply for first-time mothers or those without prior preterm births. There are few predictive systems out there, but the predictive power is very limited. One of the best known studies, for example, achieved only sensitivity of 24.2% at specificity of 28.6% for first-time mothers (Mercer et al., 1996). Machine learning techniques have been used with promising results (Goodwin et al., 2001; Vovsha et al., 2014). For example, an Area Under the ROC curve (AUC) of 0.72 was obtained in (Goodwin et al., 2001) using a large observational dataset.

This paper asks the following questions: can we learn to predict preterm births from a large observational database without going through the hypothesis testing phase? What is the best way to generate and combine hypotheses from data? To this end, this work differs from previous clinical research by first generating hundreds (or even thousands) of potential signals and then developing machine learning methods that can handle many irrelevant features. Our goal is *to develop a method that: (i) derives a compact set of risk factors with quantified uncertainties; (ii) estimates the preterm risks; and (iii) explains the prediction made.*

In other words, we derive a prediction rule to be used in practice. This demands *interpretability* (Freitas, 2014; Rüping, 2006) and *stability* (Yu et al., 2013). While interpretability is self-explanatory (Rüping, 2006), stability refers to the model that is stable under data resampling (Tran et al., 2014; Yu et al., 2013) (i.e., model parameters do not change significantly when re-estimated from a new data sample). Thus stability is necessary for reproducibility and thus must also be enforced.

Our approach is based on ℓ_1-penalized logistic regression (Meier et al., 2008), which is stabilized by a graph of feature correlations (Tran et al., 2014), resulting in a model called Stabilized Sparse Logistic Regression (SSLR). Bootstrap is then utilized to estimate the mode of model posterior as well as to compute feature importance. The prediction rule is then derived by keeping only top k most important features whose weights are scaled and rounded to integers. For estimating the upper-bound of prediction accuracy, we derive a sophisticated ensemble classifier called Randomized Gradient Boosting (RGB) by combining powerful properties of Random Forests (Breiman, 2001) and Stochastic Gradient Boosting (Friedman, 2002).

The models are trained and validated on a large observational database consisting of 15,814 women and 18,836 pregnancy episodes. The SSLR achieves AUCs of 0.85 and 0.79 for 34-week and 37-week preterm birth predictions, respectively, only slightly lower than those by RGB (0.86 and 0.81). The results are better than a previous study with matched size and complexity (Goodwin et al., 2001) (AUC 0.72). A simplified 10-item prediction rule suffers only a small loss in accuracy (AUCs 0.84 and 0.77 for 34 and 37 week prediction, respectively) but has much better transparency and interpretability.

# episodes	18,836	
# mothers	15,814	
# multifetal epis.	500 (2.7%)	
Age	32.1 (STD: 4.9)	
#preterm births:	*<37 weeks*	*<34 weeks*
–total	2,067 (11.0%)	1,177 (6.3%)
–spontaneous	1,283 (62.1%)	742 (63.0%)
–elective	754 (36.5%)	436 (37.0%)

Table 1: Data statistics.

2. Cohort

2.1 Cohort Selection

We acquired a large observational dataset from Royal North Shore (RNS) hospital located in NSW, Australia. The RNS has 18,836 pregnancy episodes collected during the preriod of 2007-2015. The data statistics are summarized in Table 1.

Preterm Birth Definition   A birth is considered preterm if it occurs before 37 full weeks of gestation. There are two types of preterm births – spontaneous and elective (or indicated). Spontaneous births occur naturally without clinical interventions, and this accounts for about 60-70% of all preterm births. In our database, the rules for figuring out spontaneous and elective, as specified by our senior clinician, are:

- Spontaneous: *PPROM* $\vee$ (*Preterm* $\wedge$ *UterineActiv.Adm.* $\wedge \neg PrelaborInterv.$)

- Elective: *Preterm* $\wedge \neg PPROM \wedge \neg UterineActiv.Adm. \wedge$ (*PrelaborInterv.* $\vee$ *CaesareanSect.*)

where PPROM stands for Preterm Premature Rupture of Membranes. This may leave a small portion of births not falling into either category ($\sim 1.5\%$).

Sub-cohort Selection   Our initial analysis revealed a critical, undocumented issue: *covariate shifting* over time, probably due to the evolution of data collection protocol. Using features extracted in the next subsection, we embedded data into 2D using t-SNE (van der Maaten and Hinton, 2008) and visually examined the shifting. Fig. 1 (left) plots data points (episodes) coded in colors which correspond to the year the episodes occurred. The colors change gradually from left (2006, dark blue) to right (2015, bright yellow) with a sudden discontinuity in the middle of 2010. Fig. 1 (middle) amplifies the year 2010. There is also a big shift from 2010 to 2011, as shown in Fig. 1 (right). Years 2011-2015 are more uniformly distributed. For this reason, we will work exclusively on the data collected between 2011-2015 as they are more recent.

2.2 Feature Generation

We operate under a hypothesis-free mode – the data is collected in hospital operations. The biggest challenge in generating features from an observational database is to detect and prevent the so-called "leakage" problem (Kaufman et al., 2012). This happens when recorded information implicitly indicates the outcome to be predicted. In pregnancy databases, it could be procedures and tests that are performed late in the course of pregnancy. When it
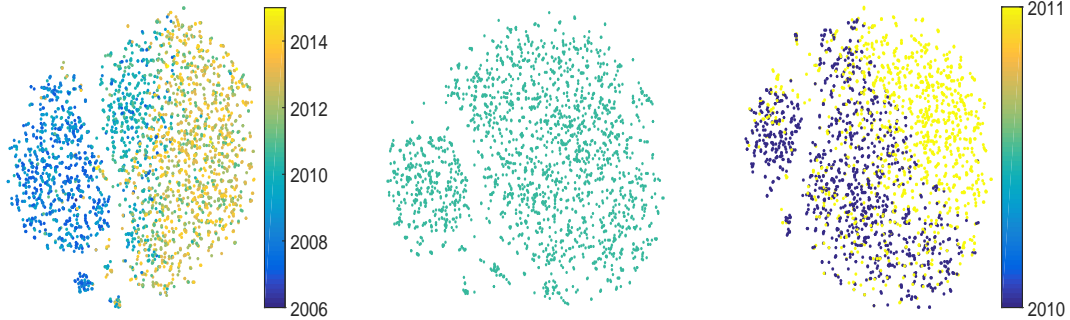
Figure 1: Covariate shifting over time in the RNS dataset (Left: 2006-2015; Middle: 2010, Right: 2010-2011). Each point is a pregnancy episode. Colors represent years of birth. Best viewed in colors.

happens, it may already indicate a full-term birth. With our clinicians and local database experts, we verify the generated feature list to prevent this from happening. We explicitly extracted features that occurred before week 25 of gestation.

During a pregnancy, a woman typically visits the hospital several times before labor. At 20-25 weeks, it is critical to estimate the risks, one of which is the risk of preterm birth. A care model may be decided by clinician depending on the perceived risks. The care allocation contains important information not present in other risk factors, and thus we will treat them separately. A part from real-valued measurements (such as BMI and weights), many features are discrete, for example, a medication name under the data field "medication". Thus discrete features are counts of such discrete symbols. The database also contains a certain amount of free text, documenting the findings by clinicians. For each text field, we extracted unigrams after removing stop words. For robustness, rare features that occurs in less than 1% of data points were removed. Finally, we retained 762 features without textual information, and 2,770 features with textual information.

3. Methods

3.1 Stabilized Sparse Logistic Regression

For a feature vector $\boldsymbol{x} \in \mathbb{R}^p$, we focus on building a linear prediction rule: $f(\boldsymbol{x}) = w_0 + \sum_{i=1}^{p} w_i x_i$, where $w_0, w_1, ..., w_p$ are feature weights. For binary classification, we use logistic regression to define the probability of outcome: $P(y = 1 \mid \boldsymbol{x}) = 1/\left(1 + e^{-f(\boldsymbol{x})}\right)$. To work with a large number of features, we derive a sparse solution by minimizing the following ℓ_1-penalized loss (Meier et al., 2008):

$$L = -\sum_{d=1}^{n} \log P(y^{(d)} \mid \boldsymbol{x}^{(d)}) + \lambda \sum_{i} |w_i| \tag{1}$$

where d denotes data points, and $\lambda > 0$ is the penalty factor responsible for driving weights of weak features toward zeros.

However, sparsity invites instability, that is, the selected features vary greatly if we slightly perturb the dataset (Gopakumar et al., 2014; Tran et al., 2014; Zou and Hastie,

2005). One reason is that when two features are highly correlated, the ℓ_1-penalty will pick one randomly. In healthcare practice, multiple processes and data views are often recorded at the same time, causing a high level of redundancy in observational data. This results in *multiple models (and feature subsets) of equal predictive performance.* This behavior is undesirable because the derived models are unstable to earn trust from clinicians.

An effective way to reduce instability under correlated data is introduced in (Tran et al., 2014). More precisely, we assert that correlated features will have similar weights. This is realized through the following objective function:

$$L = L_0 + \lambda \sum_i \left(\alpha |w_i| + \frac{1-\alpha}{2} \left(w_i - \sum_{j \neq i} S_{ij} w_j \right)^2 \right) \tag{2}$$

where $L_0 = -\sum_{d=1}^n \log P(y^{(d)} \mid \boldsymbol{x}^{(d)})$, $\alpha \in [0,1]$, and $S_{ij} > 0$ is the similarity between features i and j, subject to $\sum_{j \neq i} S_{ij} = 1$. This feature-similarity regularizer is equivalent to a multivariate Gaussian prior of mean $\mathbf{0}$ and precision matrix $(1-\alpha)(\mathbf{S}-\mathbf{I})^\top (\mathbf{S}-\mathbf{I})$ where $\mathbf{I}$ is the identity matrix. Hence minimizing the loss in Eq. (2) is to find the *maximum a posterior* (MAP), where the prior is a product of a Laplace and a Gaussian distributions. In this paper, the similarity matrix $\mathbf{S}$ is computed using the cosine between data columns (each corresponding to a feature). We will refer to this model as Stabilized Sparse Logistic Regression (SSLR).

3.2 Deriving Prediction Rule and Risk Curve

The prediction rule and risk curve are generated using the following algorithm. Given number of retained features k and number of bootstraps B, the steps are as follows:

1. **Bootstrap model averaging**: B SSLR models (Sec. 3.1) are estimated on B data bootstraps. The feature weights are then averaged. Together with the MAP estimator in Eq. (2), this model averaging is closely related to finding a mode of the parameter posterior in an approximate Bayesian setting. This procedure is expected to further improve model stability by simulating data variations (Wang et al., 2011).

2. **Feature selection**: Features are ranked based on importance, which is averaged weight × feature standard-deviation (Friedman and Popescu, 2008). This measure of feature importance is insensitive to feature scale, and encodes the feature strength, stability and entropy. Top k features are kept.

3. **Prediction rule construction**: Weights of selected features are linearly transformed and rounded to sensible integers. For example, the weights range from 1 to 10 for positive weights, and from -10 to -1 for negative weights. The prediction rule has the following form: $f^*(\boldsymbol{x}) = \sum_{j=1}^k \eta_j x_j$ where η_j are non-zero integers. We shall refer to features with positive weights as *risk factors*, and those with negative weigths as *protective factors*.

4. **Risk curve**: The prediction rule is then used to score all patients. The scores are converted into risk probability using univariate logistic regression. This produces a risk curve.

3.3 Randomized Gradient Boosting

State-of-the-art classifiers are often ensembles such as Random Forests (RF) (Breiman, 2001) and Stochastic Gradient Boosting (SGB) (Friedman, 2002). To test how simplified sparse linear methods may fare against complex ensembles, we develop a hybrid RF/SGB called Randomized Gradient Boosting (RGB) that estimates the outcome probability as follows:

$$P(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp\left(-\sum_t \beta_t h_t(\boldsymbol{x}_t)\right)}$$

where $\beta_t \in (0, 1)$ is a small learning rate, $\boldsymbol{x}_t \in \boldsymbol{x}$ is a feature subset, and each $h_t(\boldsymbol{x}_t)$ is a regression tree, which is added in a sequential manner as in (Friedman, 2002). Following (Breiman, 2001), and each non-terminal node is split based on a small random subset of features.

4. Results

4.1 Evaluation Approach/Study Design

The data is randomly spitted into two parts: 2/3 for training and 1/3 for testing. To be consistent with the practice of clinical research, for the test data *we maintain balanced classes* through under-sampling of the majority class. The parameters for Stabilized Sparse Logistic Regression (SSLR) in Eq. (2) are set as $\alpha = 0.5$, and $\lambda = 5$. The Randomized Gradient Boosting (RGB, Sec. 3.3) had 500 decision trees learnt from a learning rate of 0.03, and each tree had 256 leaves at most. Each tree is trained on a random subset of $m = \lfloor p/3 \rfloor$ features, and each node split is based on a random sub-subset of $\lceil m/3 \rceil$ features.

For performance measures, we report sensitivity (recall), specificity, NPV, PPV (precision), F-measure ($2\times$recall$\times$precision/(recall + precision)) and AUC. Except for AUC, the other measures depend on the decision threshold at which the prediction is made, that is we predict $\hat{y} = 1$ if $P(y = 1 \mid \boldsymbol{x}) \geq \tau$ for threshold $\tau \in (0, 1)$. We chose the threshold so that sensitivity matches specificity in the training data.

4.2 Visual Examination

To visually examine the difficulty of the prediction problem, we embed data points (episodes) into 2D using t-SNE (van der Maaten and Hinton, 2008). Fig. 2a plots data points coded in colors corresponding to preterm or full-term. There is a small cluster mostly consisting of preterm births, and a big cluster in which preterm births are randomly mixed with term births. This suggests that there are no simple linear hyper-planes that can separate the preterm births from the rest.

4.3 Prediction Results

We investigate multiple settings: observed features only (Obs), observation with booking & care allocation (Obs+care), and observation with textual information (Obs+care+text). Table. 2 reports sensitivity, specificity, NPV, PPV, and F-measure by SSLR and RGB. The sensitivity for SSLR ranges from 0.698–0.734 at specificity of 0.643–0.732. The sensitivity for RGB is between 0.621–0.720 at specificity of 0.740–0.841. The F-measures for both classifiers are comparable in the range of 0.693–0.732.

Table 3 reports the AUC for different settings (for spontaneous births only and all cases). For spontaneous births, the highest AUC of 0.773 is achieved by RGB using all available
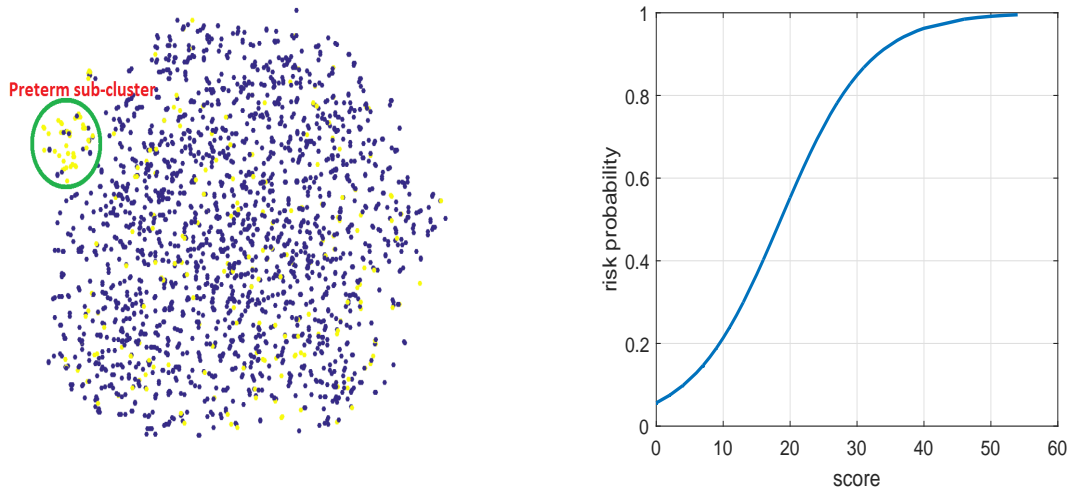
(a) Distribution of term/preterm births. Each point is an episode. Bright color represent preterm births (less than 37 full weeks of gestation). Best viewed in colors.

(b) Risk curve for prediction rule in Table 7.

Figure 2: Preterm birth distribution and estimated risk curve.

	Obs		**Obs+care**		**Obs+care+text**	
	SSLR	RGB	SSLR	RGB	SSLR	RGB
Sensitivity	0.723	0.621	0.734	0.644	0.698	0.720
Specificity	0.643	0.820	0.711	0.841	0.732	0.740
NPV	0.690	0.675	0.719	0.693	0.699	0.717
PPV	0.679	0.783	0.726	0.809	0.731	0.743
F-measure	0.700	0.693	0.730	0.717	0.714	0.732

Table 2: Classifier performance for 37-week preterm births. Obs = observed features without care allocation. SSLR = Stabilized Sparse Logistic Regression, RGB = Randomized Gradient Boosting.

Outcome	**Algo.**	**Obs**	**Obs+care**	**Obs+care+text**
Spontaneous	SSLR	0.717	0.744	0.754
	RGB	0.750	0.761	0.773
All	SSLR	0.764	0.791	0.790
	RGB	0.782	0.804	0.807

Table 3: AUC for 37-week preterm. See Tab. 2 for legend explanation.

Outcome	Algo.	Obs	Obs+care
Spontaneous	SSLR	0.806	0.828
	RGB	0.841	0.849
All	SSLR	0.834	0.850
	RGB	0.857	0.862

Table 4: AUC for 34-week preterm. Tab. 2 for legend explanation.

Outcome	weeks	Risk factors only	W/Protect. factors	SSLR
Spontaneous	34	0.804	0.823	*0.828*
	37	0.725	0.728	*0.743*
Elective/spontaneous	34	0.816	0.837	*0.850*
	37	0.757	0.767	*0.784*

Table 5: AUC of prediction rules with care information. The left column is the SSLR with all factors for reference. Risk factors are those with positive weights, whereas protective factors have negative weights.

information. For all births, the highest AUC is 0.807, also by RGB with all information. Overall, RGB fares slightly better than SSLR in AUC measure. Care information, such as booking and allocation of care, has a good predictive power.

Table 4 reports the AUC for 34-week prediction. For spontaneous births prediction, the largest AUC of 0.849 is achieved by RGB using care information. For both elective and spontaneous births, the largest AUC is 0.862.

4.4 Prediction Rules

Prediction rules are generated using the procedure described in Sec. 3.2. Table 5 reports the predictive performance of generated rules with 10 items. Generally the performance drops by several percent points. Using protective factors (those with negative weights) is better, suggesting that they should be used rather than discarded. Table 6 lists the items and their associated weights (with standard deviations) for the case *without* care information. The top three risk factors are *multiple fetuses*, *cervix incompetence* and *prior preterm births*. Other risk factors include domestic violence, history of hypertension, illegal use of marijuana, diabetes history and smoking. Likewise, Table 7 reports for the case *with* care information, which plays an important roles as risk factors.

Fig. 2b shows the risk curve estimated from the first prediction rule in Table 7 (without protective factors). When the score is 0, there is still a 5.3% chance of preterm. That says that the risk factors here only account for 50% of preterm births. When the score is 10 (e.g., with twins), the risk doubles.

5. Discussion and Related Work

We have presented methods for predicting preterm births from high-dimensional observational databases. The methods include: (i) discovering and quantifying risk factors, and (ii) deriving simple, interpretable prediction rules. The main methodological novelties are (a) the use of stabilized sparse logistic regressions (SSLR) for deriving stable linear prediction models, and

Risk factor	Score (±Std)
1. Number of fetuses at 20 weeks >= 2	10 (±0.7)
2. Cervix shortens/dilates before 25wks	8 (±1.3)
3. Preterm pregnancy	3 (±0.7)
4. Domestic violence response: deferred	2 (±0.8)
5. Hist. Hypertension: essential	2 (±1.2)
6. Illegal drug use: Marijuana	2 (±1.0)
7a. Hist. of Diabetes Type 1	2 (±1.0)
8a. Daily Cigarette: one or more	2 (±0.9)
9a. Prescription 1st Trimester: insulin	1 (±0.8)
10a. Baby Aboriginal Or Tsi: yes	1 (±0.7)
7b. Ipc Gen. Confident: sometimes	-2 (±0.7)
8b. Ultrasound Indication :other	-2 (±0.8)
9b. Ipc Emotional Support: yes	-3 (±0.5)
10b. Ipc Generally Confident: yes	-3 (±0.5)

Table 6: 10-item prediction rule, *without* care information. (The first rule with items [1-6,7a-10a] (risk factors only) achieves AUC 0.702; the second rule with items [1-6; 7b-10b] (risk+protective factors) achieves AUC 0.743).

Risk factor	Score (±Std)
1. Number of fetuses at 20 weeks >= 2	10 (±0.9)
2. Cervix shortens/dilates before 25wks	9 (±1.7)
3. Allocated Care: private obstetrician	6 (±0.8)
4. Booking Midwife: completed birth	5 (±1.2)
5. Allocated Care: hospital based	4 (±0.6)
6. Preterm pregnancy	3 (±0.8)
7. Illegal drug use: Marijuana	2 (±0.9)
8. Hist. Hypertension: essential	2 (±1.4)
9a. Dv Response: deferred	2 (±1.2)
10a. Daily Cigarette: one or more	1 (±1.1)
9b. Ipc Emotional Support: yes	-3 (±0.6)
10b. Ipc Generally Confident: yes	-3 (±0.5)

Table 7: 10-item prediction rule, *with* care information. Dv: domestic violence. The first rule with items [1-8,9a-10a] (risk factors only) achieves AUC 0.757; the second rule with items [1-6; 7b-10b] (risk+protective factors) achieves AUC 0.767.

(b) the use of bootstrap model averaging for distill simple prediction rules in an approximate Bayesian fashion. To estimated the upper-bound of model accuracy for given data, we also introduced Randomized Gradient Boosting, which is a hybrid of Random Forests (Breiman, 2001) and Stochastic Gradient Boosting (Friedman, 2002).

FINDINGS

For 37-week preterm births, the highest AUC using RGB is in the range 0.78 using only observational information, and in the range 0.80-0.81 with care information (booking + allocation decision). The SSLR is slightly worse with the AUC in the range of 0.76 with only observational information, and AUC of 0.79 with care decision.Thus, care information has a good predictive power. This is expected since it encodes doctor's knowledge in risk assessment. It is also likely to be available later in the course of pregnancy. The results are better than a previous study with matched size and complexity (Goodwin et al., 2001) (AUC = 0.72). Simplified prediction rules with only 10 items suffer some small loss in accuracy. The AUCs are 0.74 and 0.77 with and without care information, respectively. The payback is much better in transparency and interpretability.

RELATED WORK

Preterm birth prediction has been studied for several decades (de Carvalho et al., 2005; Goldenberg et al., 1998; Iams et al., 2001; Macones et al., 1999; Vovsha et al., 2014). Most existing research either focuses on deriving individual predictive factors, or builds prediction model under highly controlled data collection. Three most common known risk factors are: prior preterm births, cervical incompetence and multiple fetuses. These agree with our findings (e.g., Table 6). Data mining approaches that leverage observational databases have been attempted in (Goodwin et al., 2001) and (Vovsha et al., 2014) showing a great promise.

Clinical prediction rules have been widely used in practice (Gage et al., 2001). A popular approach is logistic regression with scaled and rounded coefficients. *Model stability* has been studied in biomedical prediction (Austin and Tu, 2004; He and Yu, 2010; Gopakumar et al., 2014; Tran et al., 2014). The machine learning community has worked and commented on *interpretable prediction rules* in multiple places (Bien et al., 2011; Carrizosa et al., 2016; Emad et al., 2015; Freitas, 2014; Huysmans et al., 2011; Rüping, 2006; Vellido et al., 2012; Wang et al., 2015). There are been applications to biomedical domains (Haury et al., 2011; Song et al., 2013; Ustun and Rudin, 2015). In (Ustun and Rudin, 2015), the authors seek to derive a sparse linear integer model (SLIM) where the coefficients are linear. The simplification of complex models is also known as *model distillation* (Hinton et al., 2015) or model compression (Bucilua et al., 2006). Most current work in model distillation focuses on deep neural networks, which are hard to interpret.

LIMITATIONS

The models derived in this paper are subject to the quality of data collected. For example, the covariate shift problem can occur within a hospital over time, as pointed out in Sec. 2.1. This study is also limited to data collected just for the pregnancy visits. There may be more predictive information in the electronic medical records. However, initial inquiry revealed that since pregnant women are relatively young, the medical records are rather sparse.

Conclusion

The methods presented in this paper to derive stable and interpretable prediction rules have shown promises in predicting preterm births. The accuracy achieved is better than those reported in the literature. As the classifiers are derived directly from the hospital database, they can be implemented to augment the operational workflow. The prediction rules can be used in paper-form as a check-list and a fast look-up risk table.

References

Peter C Austin and Jack V Tu. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of clinical epidemiology*, 57(11):1138–1146, 2004.

Fernando C Barros, Aris T Papageorghiou, Cesar G Victora, Julia A Noble, Ruyan Pang, Jay Iams, Leila Cheikh Ismail, Robert L Goldenberg, Ann Lambert, Michael S Kramer, et al. The distribution of clinical phenotypes of preterm birth syndrome: implications for prevention. *JAMA pediatrics*, 169(3):220–229, 2015.

Jacob Bien, Robert Tibshirani, et al. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

C. Bucilua, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006.

Emilio Carrizosa, Amaya Nogales-Gómez, and Dolores Romero Morales. Strongly agree or strongly disagree?: Rating features in Support Vector Machines. *Information Sciences*, 329:256–273, 2016.

Mario Henrique Burlacchini de Carvalho, Roberto Eduardo Bittar, Maria de Lourdes Brizot, Carla Bicudo, and Marcelo Zugaib. Prediction of preterm delivery in the second trimester. *Obstetrics & Gynecology*, 105(3):532–536, 2005.

Amin Emad, Kush R Varshney, and Dmitry M Malioutov. A semiquantitative group testing approach for learning interpretable clinical prediction rules. In *Proc. Signal Process. Adapt. Sparse Struct. Repr. Workshop, Cambridge, UK*, 2015.

Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

Brian F Gage, Amy D Waterman, William Shannon, Michael Boechler, Michael W Rich, and Martha J Radford. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *Jama*, 285(22):2864–2870, 2001.

Robert L Goldenberg, Jay D Iams, Brian M Mercer, Paul J Meis, Atef H Moawad, RL Copper, Anita Das, Elizabeth Thom, Francee Johnson, Donald McNellis, et al. The preterm prediction study: the value of new vs standard risk factors in predicting early and all spontaneous preterm births. nichd mfmu network. *American Journal of Public Health*, 88 (2):233–238, 1998.

Linda K Goodwin, Mary Ann Iannacchione, W Ed Hammond, Patrick Crockett, Sean Maher, and Kaye Schlitz. Data mining methods find demographic predictors of preterm birth. *Nursing research*, 50(6):340–345, 2001.

Shivapratap Gopakumar, Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Stabilizing high-dimensional prediction models using feature graphs. *IEEE Journal of Biomedical and Health Informatics*, 2014.

Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210, 2011.

Zengyou He and Weichuan Yu. Stable feature selection for biomarker discovery. *Computational biology and chemistry*, 34(4):215–225, 2010.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

JD Iams, RL Goldenberg, BM Mercer, AH Moawad, PJ Meis, AF Das, SN Caritis, M Miodovnik, MK Menard, GR Thurnau, et al. The preterm prediction study: can low-risk women destined for spontaneous preterm birth be identified? *American journal of obstetrics and gynecology*, 184(4):652–655, 2001.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

George A Macones, Sally Y Segel, David M Stamilio, and Mark A Morgan. Prediction of delivery among women with early preterm labor by means of clinical characteristics alone. *American journal of obstetrics and gynecology*, 181(6):1414–1418, 1999.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

BM Mercer, RL Goldenberg, A Das, AH Moawad, JD Iams, PJ Meis, RL Copper, F Johnson, E Thom, D McNellis, et al. The preterm prediction study: a clinical risk assessment system. *American journal of obstetrics and gynecology*, 174(6):1885–1895, 1996.

Stefan Rüping. *Learning interpretable models.* PhD thesis, Dortmund, Techn. Univ., Diss., 2006, 2006.

Lin Song, Peter Langfelder, and Steve Horvath. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5, 2013.

Truyen Tran, Dinh Phung, Wei Luo, and Svetha Venkatesh. Stabilized sparse ordinal regression for medical risk stratification. *Knowledge and Information Systems*, 2014. DOI: 10.1007/s10115-014-0740-4.

Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, pages 1–43, 2015.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172. Citeseer, 2012.

Ilia Vovsha, Ashwath Rajan, Ansaf Salleb-Aouissi, Anita Raja, Axinia Radeva, Hatim Diab, Ashish Tomar, and Ronald Wapner. Predicting preterm birth is not elusive: Machine learning paves the way to individual wellness. In *2014 AAAI Spring Symposium Series*, 2014.

Jialei Wang, Ryohei Fujimaki, and Yosuke Motohashi. Trading interpretability for accuracy: Oblique treed sparse additive models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1245–1254. ACM, 2015.

Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The annals of applied statistics*, 5(1):468, 2011.

Bin Yu et al. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.