FEATURE SELECTION AND OVERSAMPLING IN ANALYSIS OF CLINICAL DATA FOR EXTUBATION READINESS IN EXTREME PRETERM INFANTS

P. Gourdeau[1], L. Kanbar[2], W. Shalish[3], G. Sant'Anna[3], R. Kearney[2], D. Precup[1]

{[1]School of Computer Science, [2]Department of Biomedical Engineering, [3]Department of Neonatology} McGill University

NSERC CRSNG

McGill

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

ABSTRACT

We present an approach for the analysis of clinical data from extremely preterm infants, in order to determine if they are ready to be removed from endotracheal tube-invasive mechanical ventilation. The data includes over 100 clinical features, and the subject population is naturally quite small. To address this problem, we use feature selection, specifically mutual information, in order to choose a small subset of informative features. The other challenge we address is class imbalance, as there are many more babies that succeed extubation than those who fail. This paper demonstrate how to handle this problem using SMOTE, an algorithm which creates synthetic examples of the minority class.

INTRODUCTION

Motivation

- Most preterm infants (gestational age < 28 weeks) need to undergo endotracheal tube-invasive mechanical ventilation (ETT-IMV) in order to survive
- **Complications associated with prolonged ETT-IMV** include pneumonia, airway trauma, air leaks and bronchopulmonary dysplasia.
- **Early extubation** carries its own **hazards**, including compromised gas exchange and ultimately the need for reintubation (extubation failure).
- **Tradeoff** between limiting duration of ETT-IMV and avoiding reintubation
- Decision to extubate usually physician-driven and **subjective**
- **Goal**: develop tool to assist physicians in **predicting extubation readiness**

Challenges

- **Small dataset** (common for clinical studies): 120 babies
- Many features of interest (>100)
- Good **feature selection** is **critical**
- **Class imbalance**:
  - relatively few pathological examples (the minority class represents ~25% of the data)
  - identifying those examples is crucial

DATA

Patient population

- Infants admitted to neonatal intensive care units (NICUs) in hospitals in the Montreal area, Rhode Island and Detroit.

- The database includes:
  - **patient demographics** (birth weight, gestational age, ...)
  - **peri-extubation characteristics** (blood gases, ventilator settings, ...)
  - **clinical outcomes** (extubation failure)
- **Extubation failure** is defined as the need for reintubation within **7 days** following extubation.

METHODOLOGY

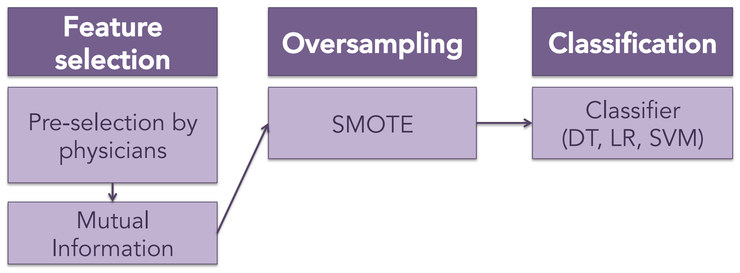


Fig. 1: Pipeline to build a classifier to predict extubation readiness.

Feature Selection (FS)

1) **Physicians' input**:
   - Chose possibly clinically relevant features
2) **Automated FS: mutual information** (MI)
   - MI quantifies how much we know about a random variable given another
   - MI computes the KL-divergence of joint distr. $p(X,Y)$ w.r.t. $p(X)p(Y)$ if X,Y were independent:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

   - We keep features with MI > 0.3 (determined experimentally)

Correcting Class Imbalance – SMOTE

- **SMOTE** addresses the undersized population by creating synthetic examples.
- **Issues with standard approaches:**
  - Undersampling the majority class is not feasible on small datasets
  - Random oversampling can introduce **bias**

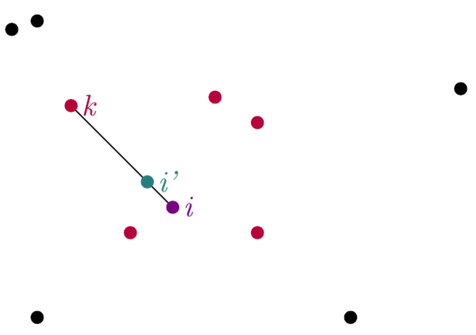


Fig. 2: A visual example of an iteration of SMOTE in 2 dimensions

- **SMOTE:** For each e.g. i in the minority class, we chose k at random from i's m nearest neighbours. To create the new sample i', let x_i and x_k represent i's and k's feature vectors, respectively. Then

$$x_{i'} = x_i + \alpha(x_k - x_i)$$

  where $\alpha \in (0,1)$ is chosen at random.
- We used SMOTE with $m = 5$ to double the minority class, going from 30 to 60 failures

Classification

- **Tool:** open-source *SciKitLearn* library
- **Crossvalidation:** leave-one-out.
  - train on *n-1* examples
  - test on remaining example
- **Algorithms:**
  - **Logistic Regression** (LR): uses logistic function to find the relationship between features and labels.
  - **Decision Trees** (DT): tree with internal nodes as tests on features and leaves as labels.
  - **Support Vector Machines** (SVM): creates a boundary between classes by using kernel functions. We used 2 different kernels: linear (LSVM) and Gaussian (GSVM).

RESULTS

Feature Selection

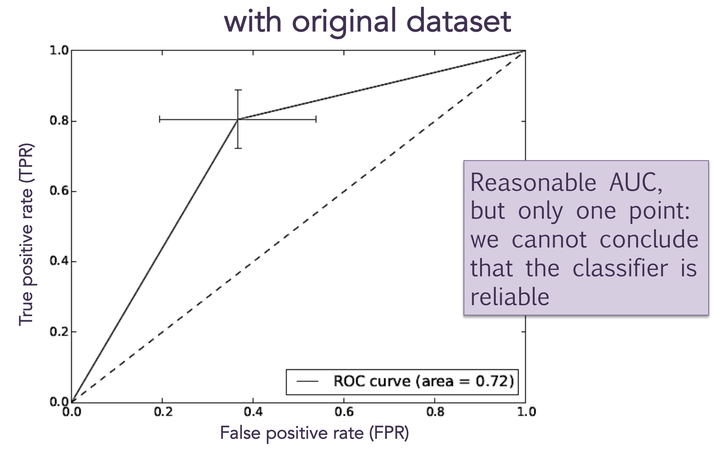
FEATURE	MI SCORE
Base excess*	0.409
Partial pressure of carbon dioxide*	0.385
Birth Weight	0.356
Weight at time of extubation	0.341
Bicarbonate*	0.328
Post-conceptual age	0.309

Table 1: features with MI score > 0.3 *Blood gases measured prior to extubation

Oversampling and Classification

Due to their simplicity as classifiers, LR, DT and LSVM had poor results.

ROC curve for Gaussian SVM classification with original dataset



ROC curve for Gaussian SVM classification with dataset oversampled using SMOTE

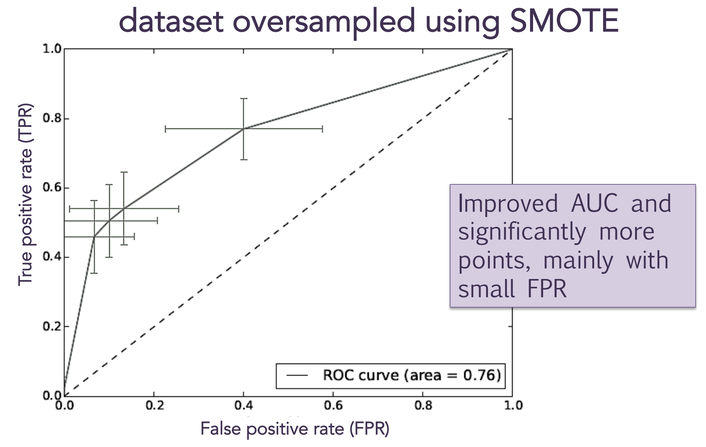


Fig. 3, 4: ROC curves for GSVM, without and with oversampling with SMOTE

DISCUSSION & CONCLUSION

- **SMOTE** is responsible for:
  - → **Increase in reliability**
  - → **Decrease of FPR**
- It makes the decision boundary **more general** and **increases coverage** of the minority class
- SMOTE should be considered by practitioners using **very imbalanced datasets**
- Still room for **improvement**:
  - → build a second classifer for physiological data
  - → use the mixture of experts model to take advantage of both classifiers
  - → probability instead of binary output
  - → increase the size of the dataset

MAIN REFERENCES

[1] Walsh M., et al. (2007) *A clusterrandomized trial of benchmarking and multimodal quality improvement to improve rates of survival free of bronchopulmonary dysplasia for infants with birth weights of less than 1250 grams.* Pediatrics 119, pp. 876-890.
[2] Snijders C., et al. (2011) *Incidents associated with mechanical ventilation and intravascular catheters in neonatal intensive care: exploration of the causes, severity and methods for prevention.* Archives of Disease in Childhood - Fetal and Neonatal Edition 96, pp. F121-F126.
[3] N.V. Chawla, et al. 2002. *SMOTE: synthetic minority over-sampling technique.* J. Artif. Int. Res. 16, 1 (June 2002), 321-357.
[4] Cover, Thomas M., and Joy A. Thomas. *Entropy, Relative Entropy and Mutual Information.* Elements of Information Theory. New York: J. Wiley, 1991. 12-23.

37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25-29 August 2015.