Feature Selection and Oversampling in Analysis of Clinical Data for Extubation Readiness in Extreme Preterm Infants*

Pascale Gourdeau[1], Lara Kanbar[2], Wissam Shalish[3], Guilherme Sant'Anna[3], Robert Kearney[2], Doina Precup[1]

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 invasive endotracheal 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. To handle this problem, we use SMOTE, an algorithm which creates synthetic examples of the minority class.**

I. INTRODUCTION

The majority of extremely preterm infants (gestational age less than 28 weeks) undergo endotracheal intubation and mechanical ventilation (ETT-MV) after birth in order to survive [1]. ETT-MV is associated with many complications, including Bronchopulmonary Dysplasia (BPD), one of the most serious pulmonary morbidities in preterm infants [2,3]. According to Laughon et al., each week of ETT-MV carries a 2.7-fold increase in the risk of developing BPD [4]. Hence, limiting the duration of ETT-MV is desirable. On the other hand, early extubation has its own hazards, including compromised gas exchange, and ultimately the need for reintubation, which is technically challenging in infants that are very small [5]. Therefore,determining extubation readiness is a major challenge in neonatal intensive care units (NICUs). This decision is often physician-driven and subjective, leading to considerable variations in practice and high rates of extubation failure [6,7,8,9] (10% to 70%, depending on the exact definition of failure).

The goal of our research is to develop a tool to help physicians predict extubation readiness in extremely preterm infants using prospectively collected clinical and physiological data. Commonly, clinical studies work with datasets that do not contain many patients, since data collection is time consuming and expensive. These datasets also contain a very large number of features of interest; for example, in our study, more than a hundred clinical features are recorded per patient. As a consequence, analyzing this data means that we need to work in a feature space whose dimensionality is close to, or even greater than, the number of data points. Automated prediction models, which rely on machine learning approaches, have been investigated with the goal of helping clinicians to take an objective decision about the best time at which to extubate [10,11]. These machine learning techniques are useful for finding patterns in data, but when the data has too many features, fitting the model becomes difficult. On one hand, discarding relevant features hurts the prediction accuracy. On the other hand keeping too many features can lead to overfitting of the model to the datatset on which it is trained, which leads to poor generalization on new, unseen data. Hence, a good mechanism for feature selection is critical in this case. We address this problem by first using clinical knowledge in order to select a fairly large subset of clinical variables, and then using mutual information in order to select a small set of relevant features for further use in classification.

A second major problem which arises in clinical studies is a class imbalance between the number of subjects exhibiting "abnormal" or pathological behaviour (which tends to be small) and the number of subjects in the "healthy" or desirable-outcome class. Usually, the number of pathological examples is much smaller, yet these are cases for which an accurate classification is critical. In the context of our study, the number of babies who fail extubation represents a small proportion of the overall population (25%). Therefore, dealing with this class imbalance is imperative. Since our study population is also small overall, downsampling the majority class is not feasible. Instead, we use an algorithm called SMOTE [12], which creates synthetic examples of the minority class during the training process.

The results obtained on data collected in an ongoing clinical trial show that our feature selection and oversampling approach increase the reliability and accuracy of the classification, but there is room for further improvement.

II. METHODOLOGY

A. Data

All infants admitted to the NICUs at the Royal Victoria Hospital, Jewish General Hospital, Montreal Children's Hospital (Montreal, QC, Canada), Detroit Medical Centre (Detroit, Michigan, USA) and Women and Infants Hospital (Providence, Rhode Island, USA) with a birth weight ≤ 1250 grams and requiring mechanical ventilation are eligible for

*This work was supported by the Natural Science and Engineering Research Council of Canada, the Canadian Institutes of Health Research and McGill University.

[1] School of Computer Science, McGill University, Montreal H3A 0E9, Canada

[2] Department of Biomedical Engineering, McGill University, Montreal H3A 0G4, Canada

[3] Department of Neonatology, McGill University, Montreal, QC H3A 2B4, Canada

this ongoing prospective clinical study (clinicaltrials.gov: NCT01909947). For the patients who are enrolled, heart rate and respiratory measurements are recorded immediately prior to extubation. A clinical database has also been developed for each infant and includes patient demographics (such as gestational age, birth weight, day of life at extubation), Peri-Extubation characteristics (such as ventilator settings and blood gases) and important clinical outcomes (including extubation failure or success). Extubation failure was defined as the need for reintubation within 7 days following extubation. In this paper we analyze a dataset of 120 extremely premature infants whose data has been collected so far. The experimental procedures on the infants described in this paper were approved by the McGill University Health Center Research Ethics Board and by each institution's research ethics committee. Additionally, the written informed consent was obtained from parents.

B. Feature Selection

For the purpose of being inclusive, over 100 clinical features are collected for each patient. In order to focus our attention on the most relevant features, the physicians on our team pre-selected 32 features that they thought could help to predict extubation outcome. These include demographic information (such as birth weight, gestational age, day of life), information about certain medications or pathologies, as well as blood gases. While this step narrowed down the space of features, their number was still too large compared to the number of patients. As a result, a step of automated feature selection was performed, in order to retain only the features that rate most predictive of the extubation outcome. We used mutual information for this step.

Mutual information (MI) quantifies how much we know about a random variable given another random variable [13]. More formally, given random variables X and Y, MI measures the KL-divergence of their joint distribution $p(X,Y)$ with respect to the distribution $p(X)p(Y)$ which would be obtained if the variables were independent:

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

$MI(X,Y) = 0$ means the variables are independent, and the higher MI, the more one variable tells us about the value of the other.

We computed the MI of each of the preselected features with the clinical outcome, and retained all the features whose MI was above a certain threshold.

C. Correcting class imbalance with SMOTE

Synthetic Minority Oversampling Technique (SMOTE) tackles the common problem of imbalanced datasets, in which the class of interest ('abnormal cases') represents a small fraction of the available data [12]. A standard approach to class imbalance is to under-sample the majority class, but this can be undesirable when working with a clinical dataset which is fairly small already. Furthermore, methods based random oversampling, which consist of choosing examples from the minority class at random until a desired class ratio is reached, can bias the classifier, by leaving out important examples or over-emphasizing certain examples just by chance. Methods based on misclassification cost can be used with certain classification methods, but because they require changing the optimization criterion of the learning algorithm, they are not useful for certain very powerful algorithms whose optimization criterion cannot be changed.

For each instance i belonging to the minority class, SMOTE creates an additional synthetic example by taking the difference between its feature vector, x_i, and the feature vector x_k of one of its randomly chosen nearest neighbours k. The number of nearest neighbours considered is a parameter of the algorithm. The result is then multiplied by a random number between 0 and 1 and added to x_i, which results in an artificial example whose features lie on the line segments between i's and k's features. SMOTE thus induces larger regions in which the minority class label is represented. This is advantageous if one assumes a contiguous spatial structure of these examples.

D. Standard Scaling

It is also important to note that the features were passed through a standard scaling algorithm to achieve normalization to 0 mean and unit standard deviation. This is standard practice, as many machine learning algorithms require variables to be in the same range to work reliably.

E. Classification

All of the experiments were carried out using the popular, open-source SciKitLearn library[1], written in Python. We developed a set of scripts in order to carry out parameter optimization for the algorithms in this library, using several machines in parallel.

Crossvalidation is mandatory in order to ensure that a machine learning algorithm generalizes well to unseen data. We performed leave-one-out (LOO) cross-validation, leaving in turn each example out for testing while training on the remaining $n-1$ examples (where $n = 120$). All our statistics are measured by averaging over the test sets.

We used SMOTE with 5 nearest neighbours and doubled the minority class, going from 30 to 60 extubation failures in the dataset. The algorithm was performed at each fold of cross-validation solely on the training set and only original, non-synthetic data was used to test the classifiers.

For classification, we used four algorithms: Logistic regression (LR), Decision trees (DT), and Support Vector Machines with linear and with Gaussian kernels. LR uses the logistic function to compute the relationship between an instance's features and its label, producing a linear decision boundary between the two classes. This linear boundary property makes the classifier easy to understand but often too simple for complicated tasks.

DTs consist of internal nodes where specific features are tested, and leaves, which represent a class label. A new

[1] http://scikit-learn.org/stable/

example will be routed, at each internal node, to one branch, corresponding to the outcome of the test. The example is assigned the label of the leaf it reaches. The learning algorithm determines the tree using criteria for measuring the quality of candidate tests, using measures of the purity of the resulting subsets of training data. DTs are nonlinear classifiers, and are appealing because they allow one to inspect the structure of the tree and decide the importance of each feature.

Support Vector Machines (SVMs) are a powerful classification algorithm which can provide non-linear decision boundaries. It works by defining a separation or *margin* between the two classes in a high-dimensional feature space, in which instances are likely to become linearly separable. This margin is defined by the instances (vectors) nearest to the decision boundary, i.e., the data points which would modify the boundary, if removed. The goal is to maximize the width of the margin. The trick to producing a non-linear boundary is to use kernel functions, which implicitly compute a dot product of feature vectors:

$$k(x_i, x_j) = \phi(x_i)\phi(x_j)$$

in time that is linear in the size of the inputs x rather than in the size of $\phi(x)$, which is much cheaper computationally. One of the most commonly used is the Gaussian kernel function:

$$k(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}).$$

One can tune the function by changing its width, $\gamma = \sigma^2$[14].

SVMs also have a built-in regularization mechanism, which can be tuned to prevent overfitting. Intuitively, this "knob" allows misclassification in the data, in order to provide a wider margin between the decision boundary and the instances (which leads to more robust classification). This knob is controlled by a parameter, usually called C. Allowing too much misclassification leads to a biased hypothesis, while allowing too little leads to overfitting. It is thus important to find the optimal combination of C and γ. For this purpose, we perform a grid search over these parameters.

III. RESULTS

LR, DTs and linear SVMs had very poor result on our data set, never working better than chance. This is consistent with previous findings in our work [15,16], which showed that linear classifiers are not sufficient in this difficult clinical population. SVM with a Gaussian kernel yields much better results than the other classifiers by allowing a non-linear separation of the data. We therefore only show the results obtained by SVM with feature selection, both with and without oversampling with SMOTE. The Receiver Operative Characteristic (ROC) are plotted on the test data.

A. Feature Selection

Table I shows the features that were obtained as a result of selecting the features with $MI > 0.3$. The selected features relate to baby's maturity (age, weight) and to the blood gasses (HCO_3). Blood gases (pH, HCO_3, PCO_2) are used to diagnose hypoventilation and respiratory acidosis. Hence, these features selected with MI make sense clinically.

TABLE I

MUTUAL INFORMATION SCORES > 0.3

Feature	MI Score
BE[2]	0.409
pCO$_2$[2]	0.385
Birth Weight	0.356
Weight at time of extubation	0.341
HCO$_3$[2]	0.328
Post-conceptual age	0.309

B. Classification without oversampling

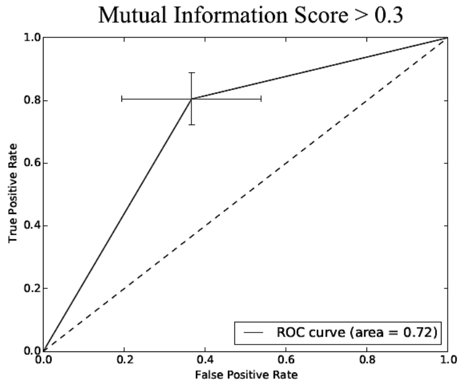

Fig. 1. ROC curve for dataset with features having a MI score > 0.3 without oversampling.

As can be seen in Fig. 1, the classifier obtained is better than chance, but the ROC curve is given by only one point. Hence, although the AUC is reasonable, it is hard to conclude that this classifier is reliable. One would like to obtain several points to have a true curve, and to be able to trade off between false positives and false negatives.

C. Classification with oversampling

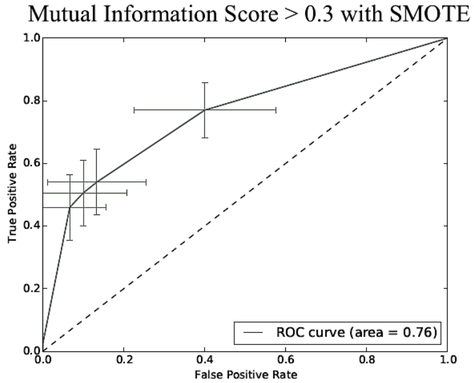

Fig. 2. ROC curve for dataset with features having a MI score > 0.3 with oversampling with SMOTE.

[2] Blood gases measured prior to extubation

Fig. 2 shows thet results of experiments done with oversampled datasets. The AUC is improved, from 0.72 to 0.76. But more importantly, we observe:

- an increase in reliability: there are significantly more points on the ROC curve.
- an decrease of the FPR: several classifiers have lower FPR while often maintaining a high true positive rate.

Both improvements can be attributed to the fact that the addition of synthetic examples makes the decision boundaries more general and increases the coverage of the minority class, thus making it easier to detect.

Finally, it is important to note that all babies who underwent extubation were thought to be ready to be extubated by the physicians, so the fact that some of the failure cases are correctly caught by the automated approach is quite significant.

IV. DISCUSSION

We succeeded to build reliable and relatively accurate classifiers using clinical data by performing state-of-the-art feature selection and oversampling techniques. Classification has a lot to gain from using mutual information to choose features that influence class labels and unbalanced datasets can greatly benefit from synthetic oversampling techniques such as SMOTE. It is key to note that linear classifiers are not sufficient for this task. We anticipate that SMOTE would be a very useful tool for others who work with highly unbalanced dataset.

While the results are positive, there is a lot of room for improvement. We believe that physiological data recorded prior to extubation contains crucial information to predict extubation readiness in neonates. Using this data un the classification is the subject of ongoing work in our research group.

As our data collection is ongoing, it is imperative to repeat the experiments once we have a larger dataset to allow us to tackle overfitting problems more efficiently and have smaller standard error bars. A larger dataset would also allow us to consider more features while maintaining the reliability of our classifier, which could ultimately improve its performance.

We are also planning on implementing a mixture of experts model (MEM) to benefit from both the clinical and physiological (time series) data. The MEM allows us to build a complex classfier from simple learners by choosing the weight for each expert given an input instance. This is done by what we call a gating function, which is learned in the model. In our case, the experts will be the classifiers for the time series and the clinical data.

Finally, our current model outputs a binary prediction - success or failure. We ultimately want to build a predictive model where the output is the confidence or probability that the extubation will be successful.

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. PM:17473087

[2] Miller J.D., Carlo W.A. (2008) Pulmonary Complications of Mechanical Ventilation in Neonates Clinics in Perinatology, Volume 35, Issue 1, Pages 273-281.

[3] 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.

[4] Laughon MM, Langer JC, Bose CL, et al. Prediction of Bronchopulmonary Dysplasia by Postnatal Age in Extremely Premature Infants. American Journal of Respiratory and Critical Care Medicine.

[5] Sant'Anna GM, Keszler M. Weaning infants from mechanical ventilation. Clinics in perinatology.

[6] P .G. Davis, D.J. Henderson-Smart ”Nasal continuous positive airways pressure immediately after extubation for preventing morbidity in preterm infants.” Cochrane Database Syst Rev pp. CD000143, 2003.

[7] F. Hermeto, M. N. Bottino, K. Vaillancourt, and G. M. SantAnna, ”Implementation of a respiratory therapist-driven protocol for neonatal ventilation: impact on the premature population”, Pediatrics, vol.123, pp. e907- 16, May 2009.

[8] Giaccone A, Jensen E, Davis P, Schmidt B. Definitions of extubation success in very premature infants: a systematic review. Archives of disease in childhood Fetal and neonatal edition. 2014;99(2):F124-127

[9] Berger J, Mehta P, Bucholz E, Dziura J, Bhandari V. Impact of early extubation and reintubation on the incidence of bronchopulmonary dysplasia in neonates. American journal of perinatology.

[10] Mueller M, et al. Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modeling. Pediatr Res. 2004;56:118.

[11] Mikhno A, Ennett CM. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. In: Ann. Int. Conf. of the IEEE Eng. Med. Biol. Soc. IEEE Engineering in Medicine and Biology Society; 2012. p. 509497.

[12] N.V. Chawla, et al. 2002. ”SMOTE: synthetic minority over-sampling technique”. J. Artif. Int. Res. 16, 1 (June 2002), 321-357.

[13] Cover, Thomas M., and Joy A. Thomas. ”Entropy, Relative Entropy and Mutual Information.” Elements of Information Theory. New York: J. Wiley, 1991. 12-23.

[14] Berwick, R. ”An Idiot's Guide to Support Vector Machines (SVMs).” Massachusetts Institute of Technology. N.p., 2008. Web. 11 Dec. 2014. http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf.

[15] Precup, D., et al. ”Prediction of Extubation Readiness in Extreme Preterm Infants Based on Measures of Cardiorespiratory Variability.” in Eng. in Med. and Biol. Soc., 2012 Ann. Int. Conf. of the IEEE, San Diego, USA, 2012, pp.5630-633.

[16] C. A. Robles-Rubio, K. A. Brown and R. E. Kearney, ”Automated Unsupervised Respiratory Event Analysis”, in Proc. 33rd IEEE Ann. Int. Conf. Eng. Med. Biol. Soc., Boston, USA, 2011, pp.3201-3204.