Prediction of Apnea of Prematurity in Neonates using Support Vector Machines and Random Forests

Nikhit Mago*, Shikhar Srivastava, Rudresh D. Shirwaikar, Dinesh Acharya U
Dept. of Computer Science and Engineering
MIT, Manipal University
Manipal, India
*Email: nikhitomatic.mit@gmail.com

Leslie Edward S. Lewis, Shivakumar M
Dept. of Pediatric
KMC, Manipal University
Manipal, India

***Abstract*—Machine Learning has a wide array of applications in the healthcare domain and has been used extensively for analyzing data. Apnea of Prematurity is a breathing disorder commonly observed in preterm infants. This paper compares the usage of Support Vector Machines and Random Forests, which are supervised learning algorithms, to predict Apnea of Prematurity at the end of the first week of the child's birth using data collected during the first three days of neonatal life. This paper also uses an optimization method called Synthesized Minority Oversampling Technique (SMOTE) to resolve the class imbalance problem observed in the data. Principal Component Analysis and one-hot encoding have been implemented for feature extraction and data preprocessing respectively. Among the results obtained, an AUC of 0.72 using the amalgamation of Random Forests and SMOTE is found to be the most accurate model.**

Keywords—Machine Learning; Neonate; SMOTE; Support Vector Machine; Random Forests; AUC; Grid Search

I. INTRODUCTION

A neonate is defined as a new born baby of age less than one month [1]. Babies born before 37 weeks of gestation are called preterm babies [2]. According to a survey conducted by the World Health Organization, around 80% neonates die within the first week of birth [3]. Apnea of Prematurity is defined as the sudden cessation of breathing that lasts at least 20 seconds in infants less than 36 weeks of gestation age [4]. Many machine learning algorithms have been used in the domain of healthcare for hospitalization prediction [5], predicting the survival of breast cancer [6], prognosis of cancer and assessment of risk after surgery [7] and neonatal prediction of diseases [8][9].

The purpose of our work is to use two machine learning algorithms, Support Vector Machines and Random Forests, to predict the Apnea of Prematurity in neonates at the end of Day 7 of neonatal life using physiological and other important predictor variables collected during the first three days of birth. This research work includes data exploration and data visualization techniques to filter out important variables for analysis. Additionally, it also uses Principal Component Analysis for feature extraction and Synthesized Minority Oversampling Technique(SMOTE) to resolve the class imbalance problem. K-fold cross validation, holdout method and ROC curves were used for evaluating model performance. Remainder of this paper is organized as follows. Section II discusses the methodology for the entire research work consisting of data source, data analysis, feature extraction, the algorithms used for prediction and SMOTE. Section III provides the evaluation parameters for the algorithms. Section IV provides the results obtained. Section V presents the discussion and Section VI concludes the paper.

II. METHODOLOGY

A. Data Source

A data set containing 367 observations of neonates was collected from a tertiary hospital. The data set comprised of more than 180 predictor variables and 10 outcome variables, pertaining to number of apnea episodes from the time of admission to discharge. The variable which signifies the number of apnea episodes from Day 4 to Day 7 was chosen to be the final outcome variable. The outcome variable was assigned binary codes, where 0 was assigned to those observations where number of apnea episodes were 0 and 1 was assigned to those observations where number of apnea episodes were greater than 0. This is also called a binary classification problem. Only 20 variables were chosen to be a part of the final data set out of more than 180 variables based on the testimony of the medical experts of the tertiary hospital. These 20 variables were analyzed using scatterplots for numeric variables, and bar charts for categorical variables. The relationship between each of the predictor variables and outcome variable was assessed. This analysis confirmed that the chosen predictors are good predictors for the presence of Apnea of Prematurity. The data set now had 20 predictor variables and 1 outcome variable.

B. Data Preprocessing

The data had to be preprocessed before implementing the algorithms. All the categorical variables were converted to binary codes using one-hot encoding where, in a code, a single high is represented as 1 and rest of the bits are represented by 0s. Each code represented a particular category and this method increased the total number of predictor columns to 31. The

978-1-5090-5256-1/16/$31.00 ©2016 IEEE

representation of categorical variables as non-binary integers was not viable as the magnitude of one category cannot be greater or lesser than the magnitude of another category [10]. The missing values of numeric columns were handled by replacing the missing value with the mean of all values corresponding to the particular outcome (0 or 1). Similarly, the missing values of categorical columns were handled by replacing the missing value with the mode of all values corresponding to the particular outcome. For the categorical variables with a lot of missing values, a new category which represented missing values was created. This was done considering that the missing values could have a semantic causality as opposed to bad record keeping, and these missing values could represent information crucial for predictive analysis [11]. Three observations were removed from the data because these observations had exceedingly many missing values with minimal inherent information. The entire data set was normalized using min-max normalization technique to scale the data in 0-1 range for optimization [12]. After applying all the preprocessing techniques, the final data set comprised of 31 predictor columns and 1 outcome column.

C. Feature Extraction

Feature selection deals with the selection of features that prove strong correlation to their respective class labels and is a crucial measure before modelling any machine learning algorithm. Feature selection is useful in better data understanding, visualization and also reduces training time [13]. Feature Extraction is different from Feature selection because a new set of features is created from the original data set which is equal to or less than the size of the original data set, whereas a subset of the original data set is selected in Feature Selection. Feature Extraction leads to reduction in the dimensionality of data where informative and non-redundant features are used for analysis that are also interpretable by humans. Feature Extraction also reduces the over fitting in the data set caused by the high dimensionality of the data just like Feature Selection. The Feature Extraction technique used in this paper is called Principal Component Analysis. Principal Component Analysis is an algorithm that performs dimensionality reduction by reducing to feature components along the axis of maximal variance in the data. It does this by converting a data set with highly correlated features to a set of linearly uncorrelated variables called principal components [14]. This process of forming linearly uncorrelated variables is called orthogonal transformation. Mathematically, PCA finds a surface, which is a vector, to project the data orthogonally. PCA then tends to minimize the sum of squares of distances from the projected surface to the data points. If, a N-Dimensional data set exists which needs to be reduced to a K-Dimensional data set by PCA, then K surfaces or vectors are computed for the data to be projected. PCA will then minimize the sum of squares of distances for each of the K vectors. PCA, therefore allows the retention of principal components that have a high explained variance, with reduction in noise and complexity as compared to the original dataset. In this paper, the number of components in PCA are selected using Minka's Maximum Likelihood Estimation [15] where ratio of explained variance was specified to be >=90% for the number of components taken. This process gave us 12 Principal Components for our analysis.

D. Support Vector Machines

Support Vector Machine is a powerful supervised learning model that facilitates optimal marginal classification in non-linear functional space. It achieves this by maximizing the functional and geometric margins from the selected decision hyperplane. Support Vector Machines can be applied to both regression and classification problems. In this paper, we use SVM as a maximal margin binary classifier on the Gaussian kernel functional space. SVM can be used as both a linear classifier and a non-linear classifier, by applying the kernel trick [16]. For a complex multivariate K-dimensional dataset, it generally isn't possible to linearly separate, and classify the data points in a feature-space of dimensions lower than K [17]. On the other hand, it is almost always possible to separate K-dimensional data on a feature-space of >K dimensions. Further, using special non-linear functions, it is possible to obtain an incrementally better linear separation of classes in a non-linear higher dimensional feature space. SVM uses non-linear functional mappings to efficiently transform the training data to a higher dimensional vector space. In this higher dimensional space, the algorithm searches for an optimal hyperplane to separate the classes in hand by ensuring the observations found close to the margins of the hyperplane, called support vectors, are best separated, thereby maximizing the marginal distance. These support vectors, are tough to classify, but coupled with the outliers provide crucial information about the classification [18].

Although SVM is very time consuming to train, it is a very powerful model due to its ability to form this complex maximum-margin decision hyperplane. SVM also handles over fitting very well as compared to other classifiers [18]. In this paper, we have used the Radial Basis Function or Gaussian kernel with the SVM model, which is given by the formula,

$$K(x, y) = \exp\left(-\gamma \left|\left|x - y\right|\right|^2\right) \quad (1)$$

, where the γ parameter defines how far the influence of a single training example reaches. This parameter is multiplied to the squared Euclidean distance between two feature vectors x and y. An RBF Support Vector Machine has two hyper-parameters that need to optimized, its regularization constant C, and hyper parameter γ. In order to optimize these values, a Grid search was performed on a range of values for the pair (C,γ). Grid Search heuristically finds the optimal value for a pair of hyper parameters from the Cartesian product of their discrete/continuous defined value sets. For every pair of parameters in this grid-space, the model performance was evaluated on the cross-validation set, and the optimal values were selected. A discrete set of values for (C,γ) were assumed, as per [19], for the Grid Search:

$$C \in \{2^{-5} \; to \; 2^{15} \; ; in \; steps \; of \; 2^2\} \quad (2)$$

and,

$$\gamma \in \{2^{-15} \; to \; 2^{3}; in \; steps \; of \; 2^{2}\} \quad (3)$$

E. Random Forests

A way to improve the performance of any machine learning model is to use ensemble methods. Ensemble methods combine many machine learning models to form a robust team for predictions. A combination function is used to reconcile the disagreements among predictions [12]. Ensemble methods reduce the amount of over fitting, as a final prediction is made using the opinions of many models. Due to the nonlinearity and complexity of medical data, ensemble methods are very beneficial for good results as data sets are divided into smaller portions that accurately capture the subtle differences [12]. Random Forest model is a complex ensemble of decision trees generated densely and randomly as a measure against the over-fitting tendency of decision trees. The tree's predictions are reconciled using a voting approach. Random Forests can effectively handle medical data sets with many features because only a small sample of the feature set is used by the ensemble. In training the Random Forest model, to facilitate the hyper-parameter optimization of the number of trees considered, a parameter sweep is performed on the cross-validation set. Discrete parametric values from 1 to 100 were linearly searched, and their performance was evaluated using 10-fold cross validation, to find the optimal number of trees to be selected.

F. Synthesized Minority Over-Sampling Technique(SMOTE)

Synthetic Minority Over-Sampling Technique is used for balancing the outcome classes in a ratio close to 1:1. The outcome classes of the training data set fed to the model for training were imbalanced. The classes were skewed by a 70:30 ratio where 70% of the observations had outcome 0 and 30% of the observations had outcome 1. This could lead to a problem for the classifiers during training. A way to improve model performance is to increase the number of minority class observations. This can be done synthetically by creating new minority class examples. By increasing the number of minority class examples so that they can be in the same ratio as the majority class examples, the model will be able to distinguish the classes better as SMOTE makes the decision boundary more general. NV Chawla et al. [20] proposed a Synthetic Minority Over-Sampling Technique that created synthetic examples rather than simply over-sampling with replacement. Extra training datum was created by performing some operations on the real data. Each of the minority class observations is over-sampled by the introduction of synthesized observations along the line joining any or all of the K nearest neighbors of the nearest minority class. K nearest neighbors less than or equal to 5 of the minority class are randomly chosen depending upon the over-sampling required. If the requirement is to increase the minority class observations K times, then K nearest neighbors will be chosen for each minority class observation and one sample will be created in each direction. Algorithmically, SMOTE creates synthetic observations by subtracting the nearest neighbor from the feature vector that is taken under consideration, multiplying this result by any random value from 0 to 1, and then adding this result to the feature vector under consideration [20]. In the training set, the minority class observations were doubled, i.e. k=2, using SMOTE and the ratio of training data after SMOTE was 55:45, which was fairly balanced.

III. EVALUATING MODEL PERFORMANCE

A. Holdout Method

The holdout method is a technique for dividing the entire data set into training and testing set. The training set is used to generate or train a model which is then evaluated by the test set. The training and test sets should be mutually exclusive [21] and should be randomly sampled. The test set in no way should be allowed to influence the model generated by the training set [12]. Model performance can be evaluated using the predictions of the test set and the labels of the test set for classification problems. In this paper, 80% of the data set was used for training the classifier and 20% of the data set was used for testing purposes. Out of the 364 observations in hand, 290 belonged to the training set and 64 were in the test set. The 80% of the training data also went through K-Fold Cross Validation as explained next.

B. K-fold Cross Validation

K-Fold Cross Validation is a technique that improves the holdout method. It reduces the variance of the resulting estimate. The training data set is divided into K random subsets that hold equal amount of data [21]. The holdout method is repeated K times for each of these K subsets. Every time, one of the K subsets is used as a test set and the remaining K-1 subsets are used to train the model. Measures of accuracy like, ROC AUC, Accuracy and Kappa are used to report the validity of Cross Validation. After the process of training and validating the data is done K times, the average performance of all k folds is reported. K-Fold Cross Validation approach is a little computationally expensive if the value of K is large. In this paper, 10- Fold Cross Validation was applied on the training data for the models. 10 different folds were created at random by the process of sampling from the training data set that consisted of 10% of the training data each. ROC curves explained later were used to assess performance on the cross validation set.

C. Confusion Matrix

It is a matrix of the outcomes of actual class and predicted class for a classification problem shown in Fig. 1.

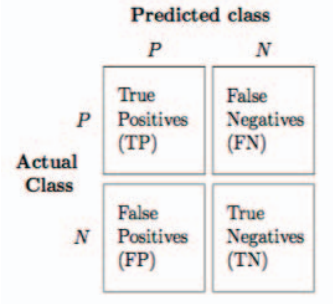

Fig. 1. Confusion Matrix

D. Sensitivity and Specificty

Sensitivity is defined as the proportion of the positive class observations that were correctly classified by a model. It is formulated as the number of True Positives divided by the sum of True Positives and False Negatives. It is also called the True Positive Rate. Specificity is defined as the proportion of negative class observations that were correctly classified by a model. It is formulated as the number of True Negatives divided by the sum of True Negatives and False positives.

E. ROC Curves

Also known as Receiver Operating Characteristic curves, these curves are plots between Sensitivity on the Y-axis and (1-Specificity) on the X axis. The sensitivity is also called the True Positive Rate and 1- specificity is also called the False Positive rate. The goal of a ROC curve is to assess performance of a model using Area Under the Curve or AUC. If the AUC = 0.5, the line will be represented by the x=y line and the model will have no predictability. The AUC will be greater than 0.5 for the lines above this line and the AUC will be lesser than 0.5 for the lines below the x=y line. The ideal AUC of any model should be 1.0, represented by the Y-axis.

IV. RESULTS

Area under the curve from the ROC curves was reported to assess the performance of the Support Vector Machine and Random Forests models on the test set. Each of the two training sets underwent 10-fold cross validation. Principal Component Analysis was performed on the data set for feature extraction where 12 Principal Components were used for analysis. Fig. 2. shows the results of this implementation using ROC curves and specifies the AUC for each of the two algorithms.

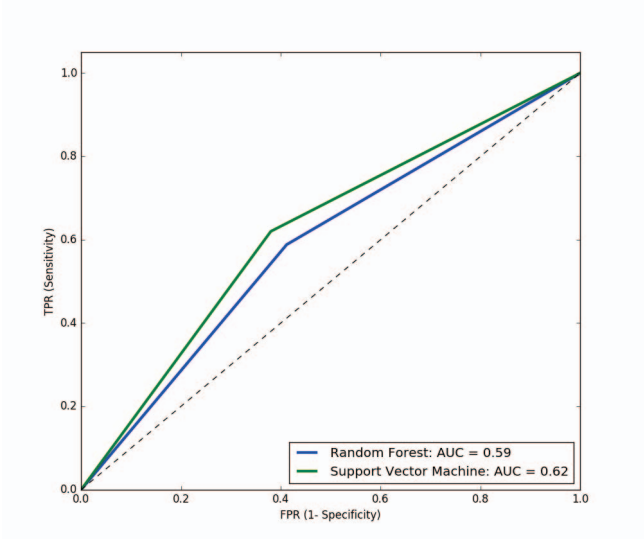

Fig. 2. ROC curve for SVM and Random Forests

Table I. shows the optimization parameters for both the algorithms.

TABLE I. OPTIMIZATION PARAMETERS FOR THE CROSS-VALIDATION SET

Algorithm	Value
Support Vector Machine- value of (C,Y)	(0.03125,0.125)
Random Forests- Grid Searched value of No of trees	69

The models reported poor accuracies after the implementation of the algorithms. Synthesized minority oversampling was performed on the training data as explained in the previous sections and the models were again cross validated using 10-fold cross validation, and finally evaluated on the test set. Fig. 3. shows the results after SMOTE was run on the training data using ROC curves.

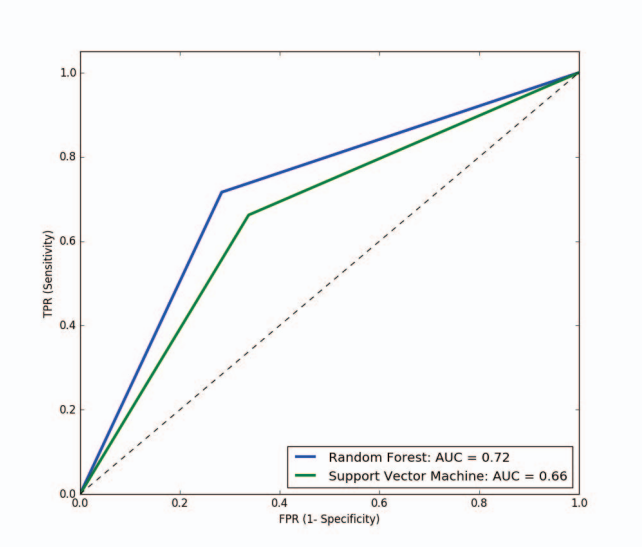

Fig. 3. ROC curve for SVM and Random Forests after SMOTE

Table II. shows the optimization parameters for both the algorithms after SMOTE.

TABLE II. OPTIMIZATION PARAMETERS FOR THE CROSS-VALIDATION SET AFTER SMOTE

Algorithm	Value
Support Vector Machine- value of (C,Y)	(0.03125,0.03125)
Random Forests- Grid Searched value of No of trees	39

V. DISCUSSION

The data set collected was preprocessed using one-hot encoding and min-max normalization. Missing values for both the numeric and categorical variables were filled using average and mode respectively. All of the predictor variables were analyzed using scatterplots and bar charts to validate the importance of these variables. The final clean data set was reviewed by the medical community and granted access for further predictive analysis of Apnea of Prematurity in neonates. The important features were extracted using Principal Component Analysis and 12 principal components were used

for the machine learning algorithms. After dividing the data set into training and testing sets using the hold out method, k-fold cross validation was applied on the training sets to reduce over fitting of data and the two machine learning algorithms applied were evaluated by the test sets using ROC curves. In Fig. 1, SVM and Random Forests reported low AUCs on their test sets, which were 0.62 and 0.59 respectively. The low efficiencies of the two models have two causalities: Firstly, the medical data set was observed to be complex, noisy and non-linear, after the initial data analysis. The inherent assumption of PCA, to decompose features along the axis of maximal variance, may be detrimental to such intricate noisy medical data with less linearity and minimal patterns. Secondly, the training set was imbalanced by a ratio of 70:30 majority class to minority class. This consequently means that the presence of few minority class observations made it even more difficult for the machine learning algorithms to classify accurately. This problem was resolved using the SMOTE technique explained above, where the minority class observations were doubled and as a result, the two classes became fairly balanced. After the application of this technique, the AUC of Random Forests increased significantly from 0.59 to 0.72, whereas the AUC of Support Vector Machines increased slightly from 0.62 to 0.66 shown in Fig. 2. Random Forests served well on this complex and non-linear data set after applying SMOTE on the training set.

VI. CONCLUSION

This research work uses two supervised learning algorithms to predict Apnea of Prematurity in neonates during the first 7 days of their life, and thereby attempts to automate the disease diagnosis process. In preprocessing, the conversion of categorical variables to binary coded variables using one-hot encoding is vital to data and feature consistency. It is found that the inherent bias of PCA to reduce along maximal variance, causes it to perform poorly on the noisy, non-linear medical data, and is detrimental to predictive performance on both models. Furthermore, the class-imbalance problem inherent in most medical datasets, is resolved effectively by applying Synthetic Minority Oversampling Technique, to the training data. Consequently, both models, trained with SMOTE, significantly outperform, the respective models trained with PCA. Random Forest model, trained on the SMOTE dataset, most effectively models the data, whilst preventing overfitting, and gives the highest predictive performance.

ACKNOWLEDGMENT

Authors are deeply indebted to Manipal Institute of Technology and NICU, Kasturba Hospital for providing an opportunity to develop and demonstrate this research work.

REFERENCES

[1] King, J.R., Kimberlin, D.W., Aldrovandi, G.M. and Acosta, E.P., 2002. Antiretroviral pharmacokinetics in the paediatric population. *Clinical pharmacokinetics*, *41*(14), pp.1115-1133.

[2] Goodwin, L. and Maher, S., 2000, March. Data mining for preterm birth prediction. In *Proceedings of the 2000 ACM symposium on Applied computing-Volume 1* (pp. 46-51). ACM.

[3] Aggarwal, R., Singhal, A., Deorari, A.K. and Paul, V.K., 2001. Apnea in the newborn. *The Indian Journal of Pediatrics*, *68*(10), pp.959-962.

[4] Zhao, J., Gonzalez, F. and Mu, D., 2011. Apnea of prematurity: from cause to treatment. *European journal of pediatrics*, *170*(9), pp.1097-1105.

[5] Dai, Wuyang, et al. "Prediction of hospitalization due to heart diseases by supervised learning methods." *International journal of medical informatics* 84.3 (2015): 189-197.

[6] Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), pp.113-127.

[7] Oliveira, T., Barbosa, E., Martins, S., Goulart, A., Neves, J. and Novais, P., 2013, June. A prognosis system for colorectal cancer. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (pp. 481-484). IEEE.

[8] Mikhno, A. and Ennett, C.M., 2012, August. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5094-5097). IEEE.

[9] Precup, D., Robles-Rubio, C.A., Brown, K.A., Kanbar, L., Kaczmarek, J., Chawla, S., Sant'Anna, G.M. and Kearney, R.E., 2012, August. Prediction of extubation readiness in extreme preterm infants based on measures of cardiorespiratory variability. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5630-5633). IEEE.

[10] Jaeger, T. Florian. "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models." *Journal of memory and language* 59.4 (2008): 434-446.

[11] Zumel, Nina, John Mount, and Jim Porzak. *Practical data science with R*. Manning, 2014.

[12] Lantz, Brett. *Machine learning with R*. Packt Publishing Ltd, 2013.

[13] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

[14] Richardson, M., 2009. Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*.

[15] Thomas P. Minka: Automatic Choice of Dimensionality for PCA. NIPS 2000: 598-604

[16] Aizerman, A., Braverman, E.M. and Rozoner, L.I., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, *25*, pp.821-837.

[17] Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller. "Kernel principal component analysis." *International Conference on Artificial Neural Networks*. Springer Berlin Heidelberg, 1997.

[18] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[19] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

[20] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.

[21] Kohavi, R., 1995, August. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).