Mortality Prediction of ICU Patients

A Project by
Apaar Garg
Abhishek Kumar
Saksham Aggarwal

Problem Statement

Given time series data of the vital statistics of ICU Patients, we shall predict their likelihood of mortality.

For decades, the number of ICUs has experienced a worldwide increase. The mortality assessment is crucial for making the critical decision of whether to interrupt the life-support treatments when intensive care is considered helpless.

Also, this prediction can help in deciding what treatment process to take.

Workflow

This problem is the subject of a number of machine learning challenges and many solutions which can predict mortality accurately have been identified.

Our project is a comparative study of different machine learning methods and their utility in this task.

- SVM Based - http://www.cinc.org/archives/2012/pdf/0481.pdf
- ANN Based - http://www.cinc.org/archives/2012/pdf/0261.pdf
- Clinical Rules Based - http://www.cinc.org/archives/2012/pdf/0401.pdf
- Linear Bayes Based - http://www.cinc.org/archives/2012/pdf/0473.pdf

Problem Statement

- First, we calculate various metrics reducing the time series data to a singular list of values for each patient. The resulting number of dimensions after this step is 776.
- We then removed attributes with more than 200 NAN values.
- Then, we applied PCA to preprocess the data and reduce the dimensions from 714 as far as possible such that more than 99% of the original variance is retained.
- Before applying PCA, mean centering and scaling was performed to eliminate any possible feature dominating over the others.

Dataset

The dataset is the set used in the XRCU 2015 challenge.

The dataset consists of both static and dynamic patient data.

The static data includes constant values such as age, height

Dynamic data includes 25 variables representing lab results of ICU patients. Another 6 variables represent the vital signs including a flag representing whether or not the patient was in the ICU at the time.

Predicting Mortality of ICU Patients Using Statistics of Physiological Variables and Support Vector Machines

- The paper discusses the use of Support Vector Machines and statistics of physiological variables the to predict the mortality of patients
- First order statistics(Mean and standard deviation) of the variables are provided as input vectors to the SVM.

- Preprocessing:
    - The dataset with PCA applied to it is given as the input to this SVM.

- The SVM by Vapnik minimized the classification error rate while finding the best hyperplane separating the two classes in the feature space.
- The probability risk is calculated using the Gaussian distribution that derives from the distance of classifier to the margin that separates the two classes.

- We plan to extend this idea of using SVM and the first order statistics of the data to the dataset that has been provided and predict the mortality using the features in the dataset.

Linear Bayes Classification for Mortality Prediction

- To predict mortality of ICU patients, a simple linear Bayes classifier is used, for which features were selected using Social Impact Theory based Optimizer.
- Extracts some variables for different time instants along with some common features.
- Use Linear Bayes classifier on these after some preprocessing on the data to reduce dimensionality and data repetition.

- Preprocessing:
  - The original dataset was reduced by performing PCA and getting 99.9% variance.

- Assumes Gaussian class–conditioned distributions with the same covariance matrix for both classes which leads to a linear decision boundary.
- The expectation vector is estimated from training data using the sample mean.
- The covariance is estimated using the sample variables measured during patient's hospitalization at ICU.

- We will be doing somethings similar. We first try to remove features with a lot of missing values. Then we reduce the correlation by removing/merging some features. On this we apply the Linear Bayes Classifier and compare the results with other classifier results.

A Neural Network Model for Mortality Prediction in ICU

- A two-layer neural network with fifteen neurons in the hidden layer was used for classification.
- One hundred voting classifiers were trained and the model's output was the average of the one hundred outputs. A fuzzy threshold was utilized to determine the outcome of each record from the output of the network.

- Preprocessing:
    - Same as done above.

- Sensitivity and precision are used as the most basic tools to evaluate algorithm.
- Since Neural nets can be stuck in local minimas, a “voting” scheme is used. During classification, all of the hundred classifiers predict and intermediate probability for an input and the final output is the average of all these 100 probabilities.

- We will be implementing the same. We will also train multiple Neural nets using fuzzy logic but we any use a simpler voting mechanism to assign a class.

Combining Machine Learning and Clinical Rules to Build an Algorithm for Predicting ICU Mortality Risk

This paper discusses the use of a model specified by a fuzzy inference system based on clinical practice. The model is optimized by using a genetic algorithm.

This lead to firm rules; for example, "If inspired oxygen is stable or increasing and oxygen saturation is decreasing, then the likelihood of mortality is high."

Event 1 Score: 0.40

Event 2 Score: 60

•

Pros:

- Fuzzy rules can be easily understood by clinicians
- Because of this, the rules can be reviewed and feedback can be given.
- This could perhaps help in identifying new trends as well.

Cons:

- Computationally, very heavy.
- Performs noticeably worse than trained Artificial Neural Networks

Results

- Accuracy obtained for Bayes approach: 92.07%
- Accuracy obtained for SVM approach: 93.21%
- Accuracy obtained for Neural Network approach: 90.31%

Conclusion

- The SVM based approach for this classification problem provides best results when compared with the Linear Bayes and Neural Net based approach.

- Applying PCA and preprocessing initially, we reduced the number of dimensions from 714 to 20, thus saving a lot of computation time and reducing redundancy which generates better results. In the real world scenario, this time could prove to be the difference between life and death.

Thank You!