IIIT, HYDERABAD

SMAI Project Report

Mortality prediction of ICU patients

Group Members

Abhishek Kumar (201301140)

Saksham Aggarwal (201301111)

Apaar Garg (201301041)

TA Mentor: Aditya Chattopadhyay

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.

Dataset

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

It consists of both static and dynamic patient data:

- Static data includes constant values such as age, height
- Dynamic data includes 25 variables representing lab results of ICU patients and another 6 variables representing their vital statistics including a flag representing whether or not the patient was in the ICU at the time.

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.

Preprocessing

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. The metrics include:

- f(x) = 1 if all derivatives of the feature are non-zero, 0 otherwise
- Difference between first and final value
- First value
- Maximum derivative
- Difference between maximum and minimum derivative
- Maximum value
- Mean derivative
- Mean value
- Absolute difference between median and mean value
- Median of the derivative
- Median value
- K minimum value
- Modal value
- Number values measured
- Lower quartile
- Upper quartile
- Difference between maximum and minimum value
- Signum of the mean derivative
- Standard deviation of the derivative

- Standard deviation
- Sum of values
- Variance
- Variance of derivative

We then removed attributes with more than 400 NAN values. Now, there were 220 features remaining.

Then, we applied PCA to preprocess the data and reduce the dimensions from 220 as far as possible such that more than 99% of the original variance is retained. Top 20 features were obtained after this. After this, the following approaches are followed.

- 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

1.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.

Further preprocessing:

- Fourier descriptors of time dependant variables which vary frequently are calculated.
- Most patients’ records lack physiological variables which is solved by introducing normal values for mean and zero value for standard deviation.

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 have extended 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.

2.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.
- Use Linear Bayes classifier on these after some preprocessing on the data to reduce dimensionality and data repetition.
- Preprocessing:
  - The same PCA approach was followed here too.
- 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 have done 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.

3. A Neural Network Model for Mortality Prediction in ICU

- Twenty-six features were selected after a thorough investigation over the different variables and features.
- 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.
- 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 have implemented the same. We trained multiple Neural nets using fuzzy logic but, use a simpler voting mechanism to assign a class.

4. 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."

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

1. The SVM based approach for this classification problem provides best results when compared with the Linear Bayes and Neural Net based approach.
2. 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.