An integrated data mining approach to real-time clinical monitoring and deterioration warning

Summary

The authors of the paper present an application of different data mining methods in predicting clinical deterioration in patients. The goal was to utilize different features obtained from detrended fluctuation analysis, spectral analysis, and approximate entropy and cross signal features in data mining methods such as forward feature selection, linear and nonlinear classification algorithms.

How does it build upon, extend, or differ from other work

This work is built upon 2 popular data mining methods (SVM and logistic regression). It only look at the current times series of heart rate and oxygen saturation rate, so other predictors (temperature, blood pressure, and etc.) are ignored. Other works in the field usually aim to predict medical conditions or specific outcome of a disease or injury based on heterogeneous sources of data.

Although the authors mentioned that patient records between 2001 and 2008 were used; no specific number about the amount of data collected was stated; neither was the accuracy or time interval during which patients were observed.

Key ideas in the paper

- Goal: predict mortality/death based from heart-rate and oxygen saturation rate time-series.
- Extract various features from time series and normalize them.
- Use linear correlation and coherence are considered cross-sign features.
- Use SVM and logistic regression for prediction based on features extracted above.
- Avoid over-fitting with forward feature selection.
- Logistic regression with forward feature selection has the best performance.

What is novel or interesting

- The authors aim at hospitals equipped with electronic devices and wireless sensors that have high frequency.
- Kernel SVM outperform linear SVM for clinical data with sparse feature, but it still underperforms to logistic regression.

Positive/Strong Points

- S1: Features collected from DFA, spectral analysis, first order and second order features and cross-signal features are basic but contain virtually almost all information we could get from time-series. The method is therefore very comprehensive.
- S2: So is the evaluation of prediction performance. The authors used AUC, PPV, NPV, sensitivity and specificity.
- S3: Data set imbalanced is overcome by using exploratory undersampling.
- S4: The procedure allows us to interpret the importance of most significant features in predicting mortality.

Negative/Weak Points
- W1: The authors should justify why they preprocess the data set by removing abnormal values. Abnormal values may be an indicator of mortality, so whether they should be excluded is debatable.
- W2: Most features within a single time series are highly correlated.
- W3: How do the authors deal with missing values and/or reading at irregular intervals?
- W4: Health deterioration is actually not recorded in patients’ record. It seems that, only for patients who died, their time series are used to train the health deterioration. Therefore, we might have missed a significant amount of cases where the patients collapsed but recovered.

Research Questions and Points for Discussion
- D1 Should the authors try different boosting techniques to further improve the accuracy?
- D2 How are the features normalized to $[0, 1]$? Since features are extracted from time-series, were local averages used?
- D3 What should be the best prediction performance? AUC, PPV or NPV? Should we aim for having the least number of false negatives?
- D4 What is the computational complexity of the entire procedure? What are some features that do not need to recomputed every time a new data point is read from a sensor?
- D5 How can the procedure alert us in advance whether the patient is going to collapse/mortality occur? What is the time gap between detecting deterioration vs. mortality?
- D6 If other patients’ data such as systolic blood pressure, pulse rate, temperature, age, BMI and etc are added into the model, should we consider other types of correlations between those different time series? Will the feature set grow too large?