健診データを用いた尿酸値予測モデルの構築

1. Introduction
   1. 健診データが日本には蓄積されていること
   2. 健診データが十分には活用されていないこと
   3. 健診データを用いて他の健診項目を予測することでスクリーニングの手間を省けること
   4. 痛風は心疾患の影響ともなりうることから尿酸値を用いたこと
2. Method
   1. 用いたデータ：健診データ、データセットの分割について説明
   2. データの特徴：他論文の表を参考にする
   3. 用いた手法の説明：GBM、ロジスティック回帰、cuDNN
   4. 用いたソフトウェアの説明：R, Python、Caffe
3. Result
   1. AUCによる比較
4. Discussion
   1. 結果の解釈
   2. 他文献との比較
   3. 健診データを用いた予測モデルの活用の方向性

Objective:

The prediction of uremia has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. For instance, thanks to innovative biomedical technologies, better explanatory prognostic factors are being measured and recorded; thanks to low cost computer hardware and software technologies, high volume better quality data is being collected and stored automatically; and finally thanks to better analytical methods, those voluminous data is being processed effectively and efficiently. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for uremia.

Methods and material:

We used three popular data mining algorithms (cuDNN, RandomForest, Gradient Boosting Machine) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset (more than 400,000 cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes.

Results:

The results indicated that the Gradient Boosting Machine (GBM) is the best predictor with 0.81 AUC on the test dataset, cuDNN came out to be the second with 0.80 AUC and the logistic regression models came out to be the worst of the three with 0.78 AUC.

Conclusion:

Using three novel data-mining algorithm, cuDNN, GBM, randomForest, models to classify the presence of hyperuremia were developed. Using sensitivity analysis on neural network models provided us with the prioritized importance of the prognostic factors used in the study.

Objective: This study presents an effective method of classifying oral malodor from oral microbiota in saliva by using a support vector machine (SVM), an artificial neural network (ANN), and a decision tree. This approach uses concentrations of methyl mercaptan in mouth air as an indicator of oral malodor, and peak areas of terminal restriction fragment (T-RF) length polymorphisms (T-RFLPs) of the 16S rRNA gene as data for supervised machine-learning methods, without identifying specific species producing oral malodorous compounds.

Methods: 16S rRNA genes were amplified from saliva samples from 309 subjects, and T-RFLP analysis was carried out with the DNA fragments. T-RFLP analysis provides information on microbiota consisting of fragment lengths and peak areas corresponding to bacterial strains. The peak area is equivalent to the frequency of a specific fragment when one molecule is selected from terminal fragments. Another frequency is obtained by dividing the number of species-containing samples by the total number of samples. An SVM, an ANN, and a decision tree were trained based on these two frequencies in 308 samples and classified the presence or absence of methyl mercaptan in mouth air from the remaining subject.

Results: The proportion that trained SVM expressed as entropy achieved the highest classification accuracy, with a sensitivity of 51.1% and specificity of 95.0%. The ANN and decision tree provided lower classification accuracies, and only classification by the ANN was improved by weighting with entropy from the frequency of appearance in samples, which increased the accuracy to 81.9% with a sensitivity of 60.2% and a specificity of 90.5%. The decision tree showed low classification accuracy under all conditions.

Conclusions: Using T-RF proportions and frequencies, models to classify the presence of methyl mercaptan, a volatile sulfur-containing compound that causes oral malodor, were developed. SVM classifiers successfully classified the presence of methyl mercaptan with high specificity, and this classification is expected to be useful for screening saliva for oral malodor before visits to specialist clinics. Classification by a SVM and an ANN does not require the identification of the oral microbiota species responsible for the malodor, and the ANN also does not require the proportions of T-RFs.