AI in Biomedical Informatics

Project #1: Medical Data Analysis Using Traditional Machine Learning Methods

Szymon Wilk

March 13-20, 2025

Content

1 Introduction 2
2 Tasks 3
3 References 4

1 Introduction

The objectives of this project are as follows:

1. To get familiar with real-life data sets describing selected decision-making problems considered in the Emergency Room (ER) of a children's hospital (triage[1] of selected types of acute pain in children).
2. To prepare the experimental environment (e.g. *Colab*) and analyze provided medical data sets to build decision-making models (classifiers) supporting *triage*.

We provide 4 real-life data sets obtained during prospective and retrospective studies in the ER of a pediatric hospital (Children's Hospital of Eastern Ontario in Ottawa). They describe patients with selected types of acute pain (these are either the most common problems seen in daily practice or the most challenging problems as indicated by cooperating clinical experts) and contain information about the correct initial treatment provided in ER.

A brief description of the data sets is available in Table 1. The class sequence represents the order of the classes in terms of their clinical significance – the most important class is indicated by "!!" (In the case of asthma, two classes are important due to the need for rapid administration of steroids). The original sets contain both qualitative and quantitative (numerical) attributes, with discretization intervals based on expert knowledge. All data sets are imbalanced – the number of objects in the most important class is usually the smallest.

Table 1. Characteristics of the considered datasets

Data set	# objects	# features	# classes	Class sequence	Notes
ap_pro	457	13	3	discharge → observation → consult!!	Abdominal pain Prospective data, the *Observer* column contains physician decisions specified in ER
sp_retro	470	29	3	discharge → clinic → consult!!	Scrotal pain Retrospective data
hp_retro	413	24	3	discharge → xlab → lab_xray_bscan!!	Hip pain Retrospective data
ae_retro	427	48	3	short → long!! →admit!!	Asthma exacerbation Retrospective data, we focus on the decision made after 60 minutes of ER stay

We would like to note that all four sets are characterized by a limited set of features – they correspond to tests performed immediately after the patient is registered in ER (hence, for example, the lack of imaging tests). The models developed from such limited data were to

[1] Triage is a determination of the type of help and its urgency for a given patient. Triage is followed by a diagnosis.

compete with the scoring systems often used in such situations (e.g. MANTRELS for abdominal pain or PRAM for asthma).

2 Tasks

Build *the best possible* decision model (classifier) for each dataset using “traditional” ML techniques [1] (given the small data sizes shallow models are recommended) and evaluate it.

We recommend performing this task in groups of 4 (or smaller), in which one person will analyze the selected data set. For analysis, use available libraries implementing traditional machine learning techniques and data preprocessing techniques (e.g., `scikit-learn`, `imbalanced-learn`). It is also possible to use packages implementing AutoML techniques (e.g., `AutoGluon`).

When performing the analysis, follow these suggestions:

1. Binarize the decision class, where *the positive class* includes the important class or classes (!!), and the negative class includes the other classes.
2. Build a *baseline* classifier (a simple solution using widely adopted techniques, e.g. logistic regression) and a *target* classifier tailored to the decision problem under consideration. A target classifier can include a more complex *pipeline* with steps responsible for preprocessing training data and for fine tuning hyper-parameters.
3. To evaluate the classifiers, use 5-fold stratified *cross validation* repeated 3 times (the suggested sampling seed to 42).
4. During each iteration, do the following:
    a. Evaluate the proposed classifiers (baseline and target) on the test set using the AUPRC (*area under the precision-recall curve*) and AUROC (*area under the ROC curve*) measures. AUPRC is better suited to imbalanced data [2], AUROC complements it.
    b. For each of the classifiers, determine *medium*- and *high-risk* thresholds (for raw numerical outcomes) based on the performance on the training set [3]:
        i. *Medium-risk* threshold = threshold for which *sensitivity* >= 99%,
        ii. *High-risk* threshold = threshold for which *specificity* >= 90%.
    c. Apply thresholds to classify test data using the classifier's continuous response:
        i. If *response* < *medium-risk*, then decision = *negative*
        ii. If *response* >= *high-risk*, then decision = *positive*
        iii. No response by default ("grey area" – moderate risk)
    d. Determine *false-negative rate* (FNR) and *false-positive rate* (FPR) for the test examples classified according to *medium*- and *high-risk* thresholds.
5. After performing all the iterations (3 x 5-fcv), calculate the averaged values:
    a. AUPRC and AUROC
    b. FNR and TNR
    c. Percentage of testing examples classified as *negative* and *positive* using *medium-risk* and *high-risk* thresholds, respectively (it will demonstrate indirectly the size of the gray zone).

After completing the analysis, prepare a short report (it can be an annotated Jupyter notebook) presenting the proposed solution for the problem along with the justification of the solution used and the results obtained. Presentations made by group members should be combined into a single document.

Complete the project by March 27 and send the final report via eKursy.

3 References

1. Brown, K. E., Yan, C., Li, Z., Zhang, X., Collins, B. X., Chen, Y., Clayton, E. W., Kantarcioglu, M., Vorobeychik, Y., & Malin, B. A. (2024). Not the Models You Are Looking For: Traditional ML Outperforms LLMs in Clinical Prediction Tasks. MedRxiv, 2024.12.03.24318400. https://doi.org/10.1101/2024.12.03.24318400
2. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.
3. Than MP, Pickering JW, Sandoval Y, Shah ASV, Tsanas A, Apple FS, Blankenberg S, Cullen L, Mueller C, Neumann JT, Twerenbold R, Westermann D, Beshiri A, Mills NL; MI3 Collaborative. Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. Circulation. 2019 Sep 10;140(11):899-909. doi: 10.1161/CIRCULATIONAHA.119.041980. Epub 2019 Aug 16. PMID: 31416346; PMCID: PMC6749969.