Electronic Health Records (EHR) of patients to foretell the presence of a cardiovascular event within six months

Bardh Prenkaj[a], Paola Velardi[b]
[a]prenkaj@di.uniroma1.it
[b]velardi@di.uniroma1.it

February 2, 2023

Abstract

In recent years the use of electronic health records (EHR), which are digital, real-time, patient-centred databases that make health information available instantly and securely to authorised users, has led to the development of various artificial intelligence-based systems capable of extracting relevant information and performing different tasks. To mention a few popular applications, EHR data have been exploited by Machine Learning algorithms to support caregivers in prognostics and diagnostics, to allow patients' stratification in precision cohorts, to suggest personalised therapies, to predict the risk of complications, and more. Despite the large amount of published research on the use of deep learning (DL) algorithms for risk prediction from EHR, state-of-the-art systems are unable to reliably cope with many sources of complexity, such as data sparsity, lack of standards, latent temporal dependencies, irregular time intervals, errors and inconsistencies in the data entry process, and absence of precise interpretability mechanisms, to name a few.

1 Dataset description

The dataset contains information about patient health records distributed as described in Table 1

Table 1: Description of the tables used in this project. All tables are in CSV files.

Table name	Table description
anagraficapazientiattivi	Contains patient-related biography and personal info
prescrizionidiabetenonfarmaci	Represents diets assigned to patients and blood glucose controls
prescrizionidiabetefarmaci	prescriptions of diabetes drugs
prescrizioninondiabete	Prescriptions of NON-diabetes drugs
esamistrumentali	Medical tests
esamilaboratorioparametri	Laboratory tests
diagnosi	Diagnosis tests

All these tables have an ID representing a patient, a pair **(idana, idcentro)**, and every event will correspond to a single date on which that event happened.

2 Task 1

Since the dataset contains many heterogeneous data, this task is about preparing (pre-processing) all the above tables containing **ONLY** active patients. For each step report the number of patients remaining and the class distribution. This task requires you to complete the following action items:

1. *Select events of interest* – we want only patients with at least one cardiovascular event in their trajectories.
2. *Invalid feature cleaning* - check for dates and time intervals $[x, y]$ such that $y < x$ and not $x \leq y$. Check for years that do not make sense (e.g., events before the birth of a particular patient).

3. *Remove patients with all dates in the same month* - we only want patients that have a long trajectory of examinations and diagnoses.

4. *Modify the actual ranges of* ***esamilaboratorioparameteri*** - see Table 2.

5. *Cohort selection and label definition* - use only those patients that after all the previous steps contain at least two events before calculating the label. Let $\mathcal{P} = \{p_1, \ldots, p_n\}$ be the set of all patients in the dataset. Let $d(e_k^i)$ be the date of the last event e_k for patient $p_i \in \mathcal{P}$. The label of the patient p_i is calculated as follows:

$$y(p_i) = \begin{cases} 1 & \text{if, within } d(e_k^i) - 6 \text{ months, } p_i \text{ has a cardiovascular event} \\ 0 & \text{otherwise} \end{cases}$$

   Eliminate the patients that have a trajectory shorter than or equal to 6 months.

6. **Concentration** - consider other cleaning strategies that improve the dataset's quality. How do you measure the quality before and after performing your cleaning strategy?

Table 2: The true ranges of the AMD/STITCH codes.

Code	Descriptive name	True range
AMD004	Systolic blood pressure	$40 \leq x \leq 200$
AMD005	Diastolic blood pressure	$40 \leq x \leq 130$
AMD007	Fasting blood glucose	$50 \leq x \leq 500$
AMD008	HbA1c	$5 \leq x \leq 15$
AMD009	Creatininemia	Not available
AMD111	Microalbuminuria	Not available
STITCH001	BMI	Not available
STITCH002	LDL Cholesterol	$30 \leq x \leq 300$
STITCH003	Non-HDL Cholesterlo	$60 \leq x \leq 330$
STITCH004	eGFR MDRD	Not available
STITCH005	eGFR CKD-EPI	Not available

3 Task 2

Since the dataset is imbalanced, this task is about balancing the class distribution. This task requires you to complete the following action items:

1. *Class imbalance #1* - not all patients will have a cardiovascular event within the stabilised six-month period. Thus, we would expect that the class distribution is highly imbalanced. For each patient p_i such that $y(p_i) = 0$, eliminate the last six months of history to avoid giving the model prediction hints into the future. For each patient p_i such that $y(p_i) = 1$, create m copies $\{p_i^1, \ldots, p_i^m\}$ such that all the cardiovascular events in the last six months of p_i^j $\forall i \in [1, |\mathcal{P}|]$ $\forall j \in [1, m]$ are eliminated, and the other events are shuffled and cancelled at random. In this way, you have a sort of balancing criterion (i.e., up-sampling the minority class).

2. *Class imbalance #2* - Action item #1 isn't going to be sufficient for balancing purposes. Propose your balancing strategy - possibly an advanced approach - and evaluate vanilla-LSTM, T-LSTM[1] Baytas et al. (2017), and PubMedBERT[2] Gu et al. (2020) on the balanced version of the dataset.

3. **Concentration** - given a particular patient series $p_i \in \mathcal{P}$, how do you model the heterogeneous intervals $\Delta(e_j, e_{h+1})$ between an event e_h and e_{h+1}? Notice that these inter-event gaps are different for each patient, but, more importantly, they are not the same within the same patient. Evaluate your interval modelling on the balanced dataset obtained from step 2 on vanilla-LSTM, T-LSTM, and PubMedBERT.

[1] https://github.com/illidanlab/T-LSTM

[2] Search for PubMedBert at huggingface and you can choose which model suits your needs. E.g., https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext.

4 Task 3 (up to +5 bonus points for the concentration item)

This task is about devising your prediction strategy taking into consideration the choices you made in the previous two tasks. This tasks requires you to complete the following action items:

1. Some macro-events e_h^i are represented as visits within a patient trajectory $p_i \in \mathcal{P}$. These visits might contain other micro-events $e_h^i = \{e_{h,1}^i, \ldots, e_{h,l}^i\}$ including lab examinations, autonomous glucose tests, and so on. While the order of the macro-events $e_h^i \in p_i$ is important - i.e., the timestamp h indexes the events in time - the order of the micro-events $e^i h, j \in e_h^i \ \forall j \in [1, l]$ is not essential in being represented in the learnt latent space of your prediction model. Devise a prediction strategy that does not consider the order of the micro-events to perform predictions. *Remember that you can implement a method available in the literature. You don't need to propose something novel!*
2. **Concentration** - perform a Bayesian optimisation on your proposed model and test it against vanilla-LSTM and T-LSTM with default parameters.

5 How do I submit the project?

Each deliverable must contain the following directory structure:

- (student1_id)_(student2_id)_(student3_id)_EHR
  - src
    - task1
      - main.py (or main.ipnyb)
    - task2
      - main.py (or main.ipnyb)
    - task3
      - main.py (or main.ipnyb)
  - res
    - (student1_id)_(student2_id)_(student3_id)_report.pdf
  - data (directory containing the dataset tables in CSV)

Please DO NOT deliver the folder data.

In each task folder, you'll have your files and the main python file, which executes and reproduces the same results as you have reported. The res subfolder must contain your written report. Only PDFs generated from LaTeX source codes will be accepted. To this end, we'll attach the template in a zip file that you can import on Overleaf and proceed with the writing. The above directory structure contains only the project's skeleton; feel free to add more files and directories as you see fit. You should attach a zip file with the name **(student_ids)_ML22_23.zip** in an email to **prenkaj[AT]di[DOT]uniroma1[DOT]it** with the subject line **[ML Project] Student (student_ids) delivery: Focus on Task X**. The deadline for delivery is three days before the date on which you'll take the written exam. In other words, the possible dates for delivery are the **8th of January for those that on the 12th of January will take the test; and on the 11th of February for those that the 15th of February will take the test**.

Notice that you can deliver your project also in February and take the written exam in January. Obviously, your final results will be registered in the February exam intake.

References

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 65–74).

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... Poon, H. (2020). *Domain-specific language model pretraining for biomedical natural language processing.*