CS 760: Machine Learning Spring 2022

Predict Adverse Clinical Outcomes and Biological Age Using Medical Data

AUTHORS: Lingyun Xiao, Deming Lu

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Machine learning techniques have become increasingly popular in healthcare industry as powerful tools that complement traditional methods for clinical decision-making and risk stratification. In this project, we explore a real dataset on *Opportunistic Cardiometabolic Screening* and predict death as well as biological age of each patient with CT data. In particular, we construct two ML models using Logistic Regression and Neural Network, and treat highly imbalanced dataset with strategies of oversampling, undersampling and focal loss. We find that the balanced accuracy increases significantly by treating the imbalanced data set properly. Also, our result suggests that by introducing additional clinical data, the accuracy is further boosted. Finally, we predict each patient's biological age with linear regression and neural network, and find that both approaches tend to compress the actual chronological ages into the range of middle age.

1 Introduction

Cutting-edge techniques in the area of artificial intelligence and machine learning have gained tremendous popularity in the healthcare industry in the last decade. On one hand, conventional clinical research often suffers from intensive labor work, expensive conduction of experiments and unforeseen biases that could potentially compromise its successful implementation. On the other hand, The rapid advancements in the ML discipline has made profound changes and improvements in healthcare by training large-scale clinical data and using statistical tools to extract insights from these data.

After the performances of CT scans and clinical practices, incidental data is collected for each patient, thereby constituting a rich, comprehensive data set. Nevertheless, these CT and clinical data often receives less attention and goes to waste after these practices. For example, in the study of metabolic syndromes by [Pickhardt et al., 2020], the authors collected new CT and clinical data from *Opportunistic Cardiometabolic Screening* as well as negative clinical outcomes associated with each patient. Therefore, by employing ML methods, we *reuse* these extensive and valuable data and perform various informative trials, such as *prediction* of adverse clinical outcomes. In particular, we make use of **logistic regression** and **neural network** as two baseline ML models in supervised learning for the prediction task. We first use CT data only perform the classification task, and then include features from the clinical data. As a consequence, we noticed that the accuracy unanimously improves after incorporating clinical data, which suggests that clinical data contains essential information for predicting clinical outcomes.

Also, we have noticed the fact that the group of patients with adverse outcomes (minority class) is significantly smaller than the group without such outcomes (majority class). As a consequence, the raw data suffers from high *imbalance*, which could falsely indicates an abnormally high accuracy rate. Thus, we introduce three methods to alleviate the bias driven by the imbalance: **undersampling**, **oversampling** and **focal loss**.

The former two methods balance the data *size* of the majority and minority groups in our classification task, while the latter approach reduces the *cross-entropy loss* for well-classified samples and further drive the main focus toward the minority class. Furthermore, the conventional accuracy is replaced by a novel construction of **balanced score** in order to further reduce the impact from unbalanced data. As a result, we have found that all three methods lead to better prediction accuracy.

Finally, we use the CT data again to perform a regression task to perform a prediction of a patient's biological age. In particular, we assume that patients with similar CT data are more likely to have similar clinical syndromes as well as biological age. Therefore, we use **linear regression** and **neural network** to use features from CT data and infer the biological age to capture both linear and nonlinear relationships between features and the outcome.

2 Related Work

We list a few related works on conducting prediction tasks using CT and clinical data.

1. [Yao et al., 2021] use logistic regression for clinical data features selection, while a convolutional neural network was used to extract CT images features. The prediction model was established by integrating the above two kinds of features for PVO prediction, and the proposed methods were evaluated using fourfold cross-validation. The result suggests that by using both types of data, the accuracy score is significantly higher than the case where only a single type of data is considered. We also obtain the similar result, but we further incorporate the data *imbalance* into the analysis in order to make the prediction more robust.
2. [Raihan et al., 2021] uses four different machine learning methods XGBoost, Adaboost, Logistic Regression as well as Random Forest to diagnose chronic renal disease. Similar to the previous paper, the authors have not considered strategies to mitigate data imbalance.
3. [Shiri et al., 2021] performed univariate analysis to determine the most predictive features among all imaging and clinical data, and the authors use various accuracy scores such as ROC, AUC and accuracy. On the contrary, we suspect that data imbalance could invalidate the predictability of these conventional scores, and therefore define the balanced accuracy score to capture such drawback.

3 Dataset

We obtain the raw dataset from the work on *Opportunistic Cardiometabolic Screening* by Perry Pickhardt (Department of Radiology, UW-Madison) and others. The main dataset consists of the record of 9223 patients and could be parsed into the following three major categories:

3.1 CT Data

The CT data subset is a collection of **numerical** data on each patient's Bone measure, Fat measures, Muscle measures, Aortic Calcification and Liver fat. The total number of features used in subsequent analysis is 11. This is the primary dataset our analysis uses to conduct prediction and inference tasks.

3.2 Clinical Data

Clinical data is a mixed collection of **numerical** and **categorical** data. The subset contains anonymized Case ID info and Clinical F/U interval [days from CT]; individual traits such as BMI, sex, age, smoking/drinking habits; professional diagnostic scores such as FRS Score and Fracture risk assessment score, and finally a binary variable of the occurrence of Metabolic Syndrome. We exclude the case ID info and F/U intervals, and use the additional features to check if the accuracy scores have improved.

3.3 Clinical Outcomes

The data on clinical outcomes contains both **numerical** and **categorical** data. All the clinical outcomes are adverse, including Death, Cardiovascular events, Pathologic/Osteoporotic fracture, Alzheimer's and Cancer. Also, the dates after CT indicate the time window of the occurrences of such adverse outcomes. In our subsequent analysis, we focus on *death* as the only outcome of interest.

4 Approach

4.1 Data cleaning and Preprocessing

Given the highly heterogeneous features with missing values as well as notable data imbalance, data cleaning and preprocessing is a critical procedure that must be done appropriately before we train our models. The subsections below summarize all the strategies we have used to alleviate bias from the raw dataset on our predictions.

Removing Invalid and Missing Values We observe that there exist some missing entries and invalid input such typo string or NaN values. Given that we have a large-scale dataset, it turns out that the percentage of missing value is relatively low. Also, there are often multiple missing values within one sample, so we suspect that these samples could be meaningless. Given the imbalance of the dataset, we suspect that replacing the missing numeric entries by the MEAN of that column would not result in a significant improvement on the predictability of the data. Moreover, a popular choice of substituting the categorical entries by the MODE of each category would result in *higher* imbalance. Therefore, we have dropped all the observations with at least one missing entry, and the final dataset consists of medical records of 7051 patients.

Numerical features For numerical features, we use **min-max normalization** to avoid drastically different scales of features. The specific normalization is defined below. The discrepancy in the absolute magnitudes of data should not have any impact on our analysis, and the *relative* magnitude within each individual feature should be weighted equally.

$$X_{new} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Categorical features As for categorical features, most of them (e.g. sex, tobacco) are features with only two categories, so we simply switch them into a binary variable of 0 and 1. For the adverse outcome of death, we identify the dead patients and assign a positive value to them, while treating the rest of the patients as

alive. Provided that what we have is a medical record, interpolating the adverse outcomes, especially *death* is inappropriate for empty entries. For all the other categorical features with more than two classes, we have found that the distribution of each class is highly skewed, and many classes have negligible observations available. Therefore, we have transformed them into simple binary variables in order to eliminate the bias from such skewness.

Data Split After data cleaning and preprocessing, the final dataset is divided into two subsets: the training set and the testing set. The training set is primarily used to train our models and not used in subsequent analysis, while the testing set, strictly separated from the training set, is used to evaluate the performance of our analysis. Ideally, we could have some *new* data as a testing set and make our current testing set as the *validation* set. However, since we are not comparing our results to work by others, we simply follow the 80-20 ratio to split training and testing datasets.

4.2 Strategies to Treat Imbalanced Data [He and Garcia, 2009]

Data Imbalance Typically, it is natural that there exists some degree of data imbalance in a dataset, and such imbalance may not affect the performance of the classification if the degree has not gone to the extreme. However, in the particular scenario of death prediction, we have seen that only around 4% patients are dead some time after the CT scans. Therefore, if we simply inform the machine to predict that a random pool of 96% of all the patients will survive, such naive prediction would actually result in a high accuracy score. Furthermore, even for the samples with positive outcome (i.e. death), placing too much *weight* on negative samples will also lead to inaccurate prediction.

Sampling In order to treat data imbalance, several techniques have been developed for the specific purpose to balance majority and minority classes in a dataset. One of the most popular techniques is sampling, which is summarized below:

1. Oversampling: Random Oversampling involves supplementing the training data with multiple copies of the minority classes. Therefore, we view oversampling as a technique that balances out the weights placed on majority and minority classes. In our implementation, we randomly choose patients with the label *death* and add them to the training set as new samples.
2. Undersampling : On the one hand, oversampling augments the original dataset by introducing additional samples of minority classes. On the other hand, random undersampling removes data from the majority class of the original data set, and also balances out the weights. Similar as above, we also randomly select a set of majority class examples and eliminate them from our dataset.

One key difference between oversampling and undersampling is obvious: Undersampling involves *deletion* of samples from the majority class. Therefore, it is possible that our trained classifier would ignore some essential information pertained to the majority class. In our implementation, we we random delete a certain number of samples in the majority class (survived) in order to reduce the chance of systematically dispose important insights from the available data.

Focal Loss Finally, the third strategy is to use a focal loss function, which is an improvement from the conventional binary cross entropy loss function. Focal loss includes a modulating term to the cross entropy loss in order to shift the focus to difficult, misclassified examples. In fact, it is a dynamically scaled cross entropy loss, where the scaling factor gradually decays to zero as confidence in the correct class increases.

Intuitively, this scaling factor can automatically down-weight the contribution of "well-classified" examples during training and instead places more weights on hard samples.

The equation of focal loss is as follows:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

which includes an additional term to the standard cross entropy criterion:

$$CE(p_t) = -\log(p_t)$$

In general, in the healthcare discipline, patients suffer from adverse outcomes such as death tend to be more informative than healthy patients with no distinct syndromes. Thus, by transforming the traditional cross entropy loss to the focal loss, we are able to place more weights on positive samples as the minority class, and such method also aligns with intuition in clinical practices.

4.3 Logistic Regression and Neural Network

In this project, we use two different machine learning techniques to predict death of a patient: logistic regression and neural network.

The primary reason for using logistic regression is due to its simplicity and fast training speed in a binary classification task, as seen in our analysis. Also, logistic regression does not make any assumption for the distribution of each feature, which is ideal in our setting because the distribution of adversely affected patients in the healthcare industry is difficult to obtain. Furthermore, our the scale of our dataset far exceeds the number of features used in the prediction task, which is less inclined to overfitting the data. In particular, we implement Logistic Regression using Scikit-Learn package.

However, the major limitation of the logistic regression approach is its assumption of linearity between the outcome and features, which might not necessarily be true. In fact, it is difficult to imagine that the adverse clinical outcome (i.e. death) is affected by some linear combination of features in CT scans and clinical trials, since clinical practices fundamentally too challenging to extract such simple relationship. Therefore, we also choose to use Neural Network in order to capture the non-linearities that could exist in the relationship between features and outcomes. In order to implement Neural Network in practice, we use the pytorch package, which is widely used in the practical implementations of NN models in the industry. The initialization of our network is summarized in Table1:

Table 1: Hyperparameters of neural network in prediction and regression

Hyperparameters	**Values**
No of layers	3
Neurons in each layer	64
Activation function	Relu, Sigmoid
Optimizer	Adam Optimizer
Loss function	Cross Entropy, Focal Loss

5 Results

5.1 Death Prediction

Balanced Score As emphasized in previous sections, data imbalance puts our analysis under severe risk of overconfidence. Therefore, instead of using the conventional accuracy score to test our prediction results, we define the following **Balanced Score** to evaluate the trained models:

$$S = \frac{1}{2}(\frac{TP}{TP+FP} + \frac{TN}{TN+FN}) \times 100$$

As shown from the definition, the updated Balanced Score is close to the conventional Accuracy score when the data is balanced, but place more weights on the minority class when the dataset suffers from severe imbalance.

Performances We evaluate our trained models by comparing the balanced scores on the same testing set. In the following table, LG denotes Logistic Regression, NN means Neural Network. CT and CD are the CT data and Clinical data. The first two columns summarize the balanced scores for predictions using CT data *only*, and the last two columns combine both CT and clinical data to make predictions. We have multiplied the balanced scores by 100.

Balanced Score				
-	LG+CT	NN+CT	LG+CT+CD	NN+CT+CD
No other operation	51	50	54	53
Oversampling	60	63	73	75
Undersampling	55	62	72	76
Focal Loss	-	63	-	76

Table 2: Performance on CT/CT+CD datasets using classification strategies of LR and NN

The histogram below (Figure 1) visualizes the results and makes them more intuitive to look at:

By comparing the score in the above table, we make the following observations:

1. All three strategies (Oversampling, Undersampling and Focal Loss) have significantly alleviated the negative impact of data imbalance.
2. The inclusion of clinical data that complements CT data improves the prediction of death outcome significantly across all methods.
3. Neural Network shows a slightly better performance than Logistic Regression.

5.2 Biological Age Prediction

The next *regression* task on biological age prediction is another type of prediction task that deviates from the classification in previous sections. As for the estimation of biological age, we make the assumption that patients with similar CT data should also have same implied biological ages, as they are more likely to share similar body characteristics. In particular, we choose to use **Linear Regression** and **Neural Network** to combine various features in CT data, and infer the biological ages from them. Also, we evaluate the performances by computing the residual loss for both linear regression and neural network.

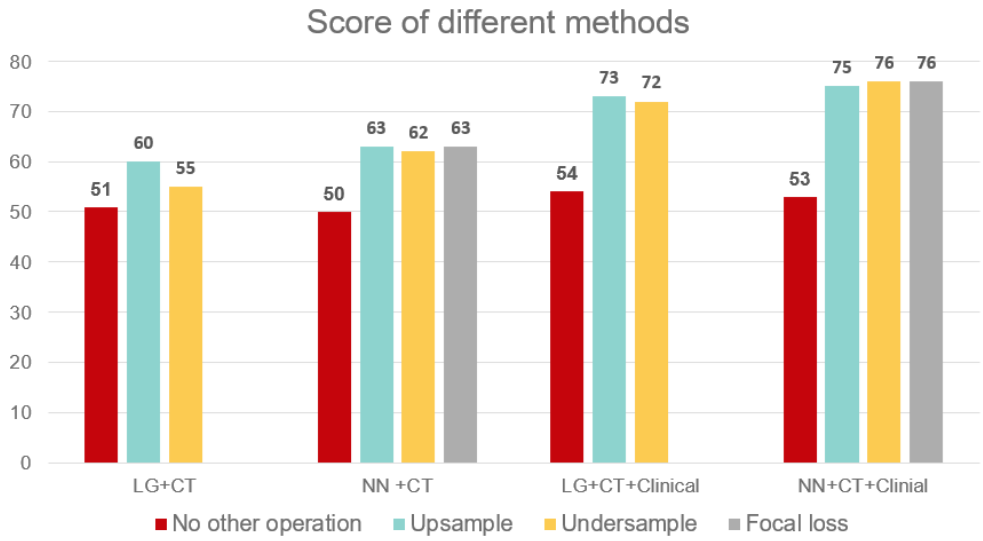


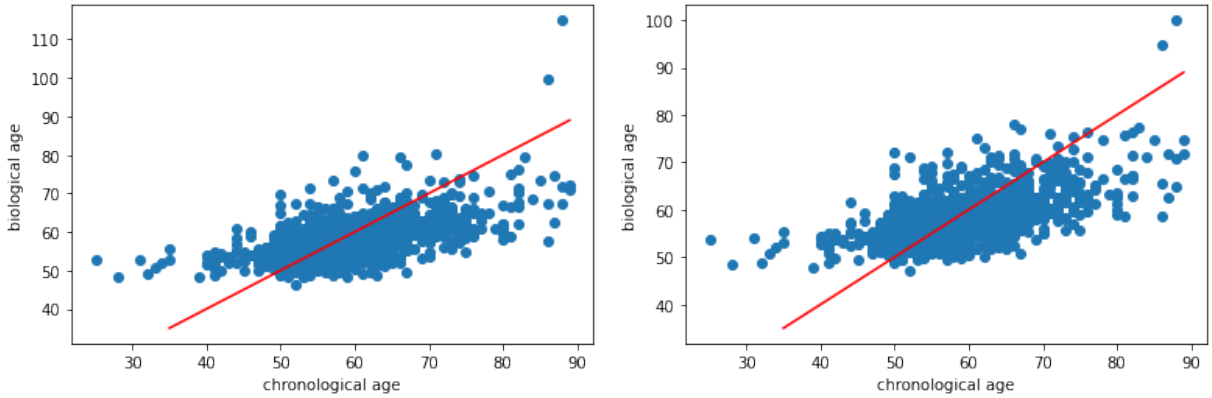
Figure 1: Histogram on performances of various classification methods

Figure 2 shows the results we obtain from the regression task. The red lines indicate the cases where chronological age coincides with the biological age. As a result, the performance of Linear Regression and Neural Network are similar. Both approaches "shrink" the chronological ages to fit into middle ages, which means younger population tends to have higher biological age, and older population is more likely to have lower biological age.

6 Conclusion and Future Work

In this project, we demonstrate that unused CT and clinical data contain insightful information in the prediction of adverse clinical outcomes as well as the inference of biological age of a patient. By introducing strategies to treat the highly imbalanced data set, utilizing logistic regression and neural network in supervised learning to perform the classification task, and using linear regression to conduct statistical inference, we have demonstrated that a number of features in the CT and clinical data have considerable potential to be integrated into routine clinical practices.

In the future, one promising direction is to obtain substantial data on various types of clinical outcomes and conduct an improved *multinomial* logistic regression. In this case, ML techniques could further support physicians with a more specified and accurate diagnosis. Also, as more clinical data becomes available, we would be able to test our algorithms on these new data and compare with other ML methods in a more systematic way. Finally, as for the regression task on biological age prediction, *kNN regression* could provide alternative insights into the composition of similar groups of patients, which could also lead to more accurate and efficient clinical practices. We look forward to reviewing all possibilities to extend our current work.



(a) Linear Regression, loss= 255

(b) Neural Network, loss= 250

Figure 2: Biological Age Regression

References

[He and Garcia, 2009] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

[Pickhardt et al., 2020] Pickhardt, P., Graffy, P., Zea, R., Lee, S., Liu, J., Sandfort, V., and Summers, R. (2020). Opportunistic screening for metabolic syndrome in asymptomatic adults utilizing fully automated abdominal ct-based biomarkers. *American Journal of Roentgenology*, 216.

[Raihan et al., 2021] Raihan, M. M. S., Ahmed, E., Karim, A., Azam, S., Raihan, M., Akter, L., and Hassan, M. M. (2021). Chronic renal disease prediction using clinical data and different machine learning techniques. In *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, pages 1–5. IEEE.

[Shiri et al., 2021] Shiri, I., Sorouri, M., Geramifar, P., Nazari, M., Abdollahi, M., Salimi, Y., Khosravi, B., Askari, D., Aghaghazvini, L., Hajianfar, G., et al. (2021). Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest ct images in covid-19 patients. *Computers in biology and medicine*, 132:104304.

[Yao et al., 2021] Yao, Z., Hu, X., Liu, X., Xie, W., Dong, Y., Qiu, H., Chen, Z., Shi, Y., Xu, X., Huang, M., et al. (2021). A machine learning-based pulmonary venous obstruction prediction model using clinical data and ct image. *International Journal of Computer Assisted Radiology and Surgery*, 16(4):609–617.