Diagnosis with Data : Disease Prediction with Machine Learning

Sangyu Shen

UNIVERSITY OF SAN FRANCISCO
Master of Science in Data Science

Abstract

Cerebral necrosis after radiation theraphy for patients with brain metastases is being recognized as a problem more common than previously estimated.

The initial goal of this project is trying to develop a diagnosis model with machine learning techniques to predict the onset of necrosis in order to improve current diagnosis. And the ultimate goal is to better understand the onset of necrosis, reduce its occurrence and generate general guides on radiation therapy dose value for metastases treatment.

Introduction

Brain metastases, or secondary brain tumors are cancer cells that spread from their original site to the brain. They are the most common malignancy occurring within the cranium. In America, about 25% of cancer patients suffered from secondary brain tumors [1].

Treatment of brain metastases is multidisciplinary with radiation forming the cornerstone[2]. However, significant risks are accompanied by radiation therapy.

Cerebral necrosis, which is one of these, is the unnatural death of the cells in an organ or tissue surrounding the tumor. Given that its characteristics on standard imaging are no different that tumor recurrence, traditional diagnosis of necrosis is time-consuming and challenging.

The goal of this project is to leverage the plentiful information of patient data to identify the patients with a higher risk for radiation necrosis, and consequently lead to more effective treatment strategies for brain metastases patients.

Patient Physical Features & Treatment Information
Classification model trained on patient data

⇄

Necrosis Prediction
Based on the risk for necrosis, formulate treatment plan for patients

Data Available

Data that was utilized during this project was contributed by Radiation Oncology Department at UCSF Medical. It consisted of features of over 50,000 metastases as well as over 5,000 patients.

Data preprocessing included feature selection, category encoding and missing value imputation. The date information of patient surgeries was dropped and instead, patient age was calculated; one-hot encoding was applied to represent different tumor primary sites; and since most of the numerical inputs are boolean or discrete, missing values were imputed by mode.

Models Exploration

Since it is a classificaiton problem, three baseline models were explored, including **Logistics Regression**, **Random Forest** and **Gradient Boosting**. After hyperparameter tuning , those three models could archieve negative log-loss of .56, .52 and .43 respectively.

Noticing the fact that the dataset was significantly unbalanced, with over 80% of patients labeling as non-necrosis, we felt the need to adjust the class distribution of our dataset. One technique was under-sampling the majority class, which was the patients without necrosis in this case. Another technique was over-sampling the minority class, which was the patients with necrosis.

Table 1. Model Performances - Log-loss (AUC).

	Logistics Regression	Random Forest	Gradient Boosting
Under-sampling	.90 (.72)	.65 (.81)	.67 (.79)
CC	.63 (.80)	9.83 (.63)	.69 (.72)
SMOTE	.49 (.82)	.33 (.86)	**.20 (.89)**
SME	.71 (.79)	.71 (.80)	.33 (.80)

Re-sampling Techniques

In this project, the explored re-sampling algorithms included Random Under-sampling, **Cluster Centroids Under-sampling (CC)**, Random Over-sampling, **Synthetic Minority Over-sampling (SMOTE)** and **combination sampling (SME)**. The model performances with selected re-sampling methods were shown in Table 1.

Generally speaking, SMOTE outperformed the other re-sampling methods. With data re-sampled by SMOTE, a Gradient Boosting model could achieve a log-loss of .21 and an AUC score of .83. Compared to the baseline models, the model performance was improved significantly by re-sampling.

Cross-validation was then conducted to tun hyper-parametes and to validate the model performance. The mean ROC curve after 5 fold cross-validation was shown in Figure 1.

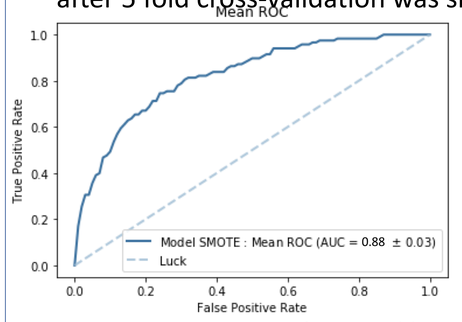

Figure 1. Mean ROC curve of SMOTE Model.

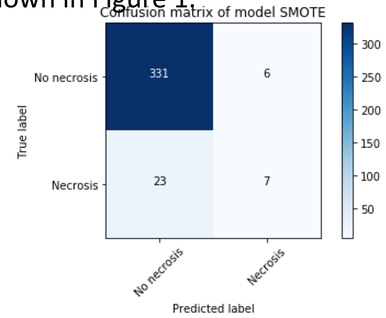

Figure 2. Confusion Matrix of SMOTE Model .

Discussion and Conclusions

As shown in Figure 2., the model had achieve an accuracy of 93%. It performed well in predicting non-necrosis but it still need to be improved to identify necrosis patients.

As suggested by the important features of the model, one possible reason of the vague prediction on necrosis was lacking important information about each metastasis. Current available data was only patient-level. Next step of this project will be collecting metastasis-level data , exploring weighted techniques to handle unbalanced data and then generating suggested radiotherapy dose value for metastases treatment.

Contact

Sangyu Shen
University of San Francisco, MS in Data Science

sangyushen@gmail.com
(415) 307-2257

References

[1] Saha, A., Ghosh, S. K., Roy, C., Choudhury, K. B., Chakrabarty, B., & Sarkar, R. (2013). Demographic and clinical profile of patients with brain metastases: A retrospective study. Asian journal of neurosurgery, 8(3), 157.

[2] Chang, J. E., Robins, H. I., & Mehta, M. P. (2007). Therapeutic advances in the treatment of brain metastases. Clin Adv Hematol Oncol, 5(1), 54-64.

[3] Di Chiro, G., Oldfield, E., Wright, D. C., De Michele, D., Katz, D. A., Patronas, N. J., ... & Kufta, C. V. (1988). Cerebral necrosis after radiotherapy and/or intraarterial chemotherapy for brain tumors: PET and neuropathologic studies. American Journal of Roentgenology, 150(1), 189-197.