Personalized colorectal cancer survivability prediction with machine learning methods

Samuel Li
Department of Computer Science, Princeton University
seli@princeton.edu

Abstract

Keywords: Cancer survivability prediction, SEER, machine learning, personalized medicine, imbalanced classification

1. Introduction

2. Background and Methodology

Machine Learning Pipeline

Data Source

Ethnicity	**Not Survived**	**Survived**
White	53343	222204
Hispanic	6762	30813

Data Preprocessing

Models

We selected the following classifiers to predict survivability. Models were selected based on preliminary test performance as well as models previously reported in literature.

- *Logistic Regression*

3. Results

Model	Hispanic	White	Mixed
Logistic Regression	.859	.872	.87
Random Forest	.855	.865	.849
AdaBoost	.859	.871	.859
Neural Network	.873	.875	.856

Table 1:

Logistic Regression		Random Forest		AdaBoost	
Hispanic	**White**	**Hispanic**	**White**	**Hispanic**	**White**
Histology	Histology	Metastasis	Metastasis	Extension	Extension
Extension	Lymph node inv.	Stage	Stage	Histology	Age
Metastasis	Extension	Age	Age	Age	Histology
Surgery site	Surgery site	No surg. reason	No surg. reason	Tumor size	Positive nodes
Diagnostic conf.	Metastasis	Positive nodes	Positive nodes	Positive nodes	Metastasis
Lymph node inv.	Diagnostic conf.	Surgery site	Surgery site	Metastasis	Tumor size
No surg. reason	Primary site	Tumor size	Tumor size	Surgery site	Surgery site

Table 2:

Model	Hispanic	White
Logistic Regression	.628	.683
Weighted Logistic Regression	.783	.8
Undersampled Logistic Regression	.79	.8
Random Forest	.623	.631
Weighted Random Forest	.782	.775
Undersampled Random Forest	.787	.796

Table 3: EMPHASIZE this is G-mean

4. Discussion

5. Conclusion

Here are two sample references: [1, 2].

References

[1] R. Al-Bahrani, A. Agrawal, A. Choudhary, Survivability prediction of colon cancer patients using neural networks, Health informatics journal (2017) 1460458217720395.

[2] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, International journal of methods in psychiatric research 20 (1) (2011) 40–49.