Accepted Manuscript

Gradient boosting for high-dimensional prediction of rare events

Rok Blagus, Lara Lusa

PII: S0167-9473(16)30180-3
DOI: http://dx.doi.org/10.1016/j.csda.2016.07.016
Reference: COMSTA 6321

To appear in: *Computational Statistics and Data Analysis*

Received date: 4 March 2016
Revised date: 27 July 2016
Accepted date: 28 July 2016

Please cite this article as: Blagus, R., Lusa, L., Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics and Data Analysis* (2016), http://dx.doi.org/10.1016/j.csda.2016.07.016

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

*Manuscript
Click here to view linked References

Gradient boosting for high-dimensional prediction of rare events

Rok Blagus*, Lara Lusa

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia

Abstract

In clinical research the goal is often to correctly estimate the probability of an event. For this purpose several characteristics of the patients are measured and used to develop a prediction model which can be used to predict the class membership for future patients. Ensemble classifiers are combinations of many different classifiers and they can be useful because combining a set of classifiers can result in more accurate predictions. Gradient boosting is an ensemble classifier which was shown to perform well in the setting where the number of variables exceeds the number of samples (high-dimensional data), however it has not been evaluated for the prediction of rare events. It is demonstrated that Gradient boosting suffers from severe rare events bias, correctly classifying only a small proportion of samples from the rare class. The bias can be removed by using subsampling in combination with appropriate amount of shrinkage but only for a specific number of boosting iterations and for binomial loss function. It is shown that the number of boosting iterations where the rare events bias is removed cannot be estimated efficiently from the training data when the sample size is small. Therefore several corrections for the rare events bias of Gradient boosting are proposed and evaluated by using simulated and real high-dimensional data. It is demonstrated that the proposed corrections successfully remove the rare events bias and outperform the other ensemble classifiers that were considered. Large flexibility and high interpretability of the proposed methods is also illustrated.

*Corresponding author.
Email addresses: rok.blagus@mf.uni-lj.si (Rok Blagus), lara.lusa@mf.uni-lj.si (Lara Lusa)

Download English Version:

https://daneshyari.com/en/article/4949270

Download Persian Version:

https://daneshyari.com/article/4949270

Daneshyari.com