دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشدآمارزیستی

عنوان :

کاربرد مدل های بوت استرپ تجمعی در ساخت درخت های پیش بینی و نقش آن در افزایش تعمیم پذیری نتایج

توسط :

مرتضی رستمی زاده

استاد راهنما:

دکتر محمدرضا بانشی

استاد مشاور:

دکتر بهشید گروسی

سال تحصیلی : ۱۳۹۴-۱۳۹۳

The application of bootstap aggregation Model in Decision tree models and in providing generalizable results

A Thesis
Presented to
The Graduate Studies

By

Morteza Rostamizadeh

In Partial Fulfillment
Of the requirements for the Degree
Master of Science in:

Biostatistics

Kerman University of Medical Sciences

چکیده پایان نامه

مقدمه و هدف : پیش بینی احتمال وقوع یک نتیجه از اهمیت کلیدی در بسیاری از تحقیقات بالینی برخوردار است. بسیاری از محققان علاقه مند به استفاده از مدل های درختی برای رسیدن به اهداف خود هستند زیرا سهولت نسبی برای تفسیر نتایج طبقه بندی شده در مقایسه با دیگر روش ها از ویژگی های منحصر این مدل ها است. در بسیاری از تحقیقات پزشکی محققان برای پیش بینی احتمال وقوع یک پیامد، مدل درختی خود را بر روی داده مورد نظر برازش می‌دهند و سپس عملکرد مدل را بر روی همان داده ارزیابی می‌کنند. از آنجا که این رویکرد نتایج خوبی به دست می‌دهد، اما نتایج به دست آمده قابلیت تعمیم برای داده مستقل دیگر را ندارند. لذا باید از روش‌های جایگزین استفاده نمود. هدف این مطالعه نشان دادن توانایی روش الگوریتم بوت استرپ تجمعی (بگینگ)، برای بهبود و ثبات مدل درختی کارت است.

مواد و روش‌ها : مدل های درختی کارت و الگوریتم بگینگ روی یک مجموعه داده مربوط به گرایش افراد به جراحی زیبایی، که قبلا بوسیله پرسشنامه درشهرکرمان جمع آوری شده اند، بکاربرده می‌شوند. داده‌های مورد استفاده در این مطالعه شامل۱۲۰۴ نمونه درگروه سنی ۱۴ تا ۶۸ سال می باشند. برای سنجش صحت مدل‌های درختی کارت از درصد های متفاوتی برای جداسازی دو مجموعه داده آموزشی و آزمایشی استفاده می شود که در این مطالعه نیز از درصدهای ۵۰ و ۶۵ و ۷۵ برای مجموعه آموزشی استفاده شده است. سپس مدل بگینگ با ۱۰۰ درخت برازش داده شده روی داده‌ها اعمال شد. به این صورت که نمونه‌های بوت استرپی به عنوان مجموعه آموزشی برای هر درخت به صورت تصادفی و با جایگذاری با حجمی برابر داده‌های اصلی از داده ها استخراج می‌شود. از آنجا که بعضی مشاهدات در نمونه آموزشی قرار ندارند این مشاهدات برای ارزیابی عملکرد هر درخت استفاده می‌شوند. به دلیل وجود مشاهدات تکراری، طبقه هر مشاهده براساس رای اکثریت پیش بینی شده در هر درخت تعیین می‌شود. سپس طبقه نهایی ، طبقه بندی ای خواهد بود که درختان متفاوت آن را پیش بینی کرده اند. برای هر مدل به طور جداگانه میزان حساسیت، ویژگی و دقت پیش بینی کل محاسبه و مقایسه گردید. برازش تمامی مدل ها توسط نرم افزار R انجام شده است.

یافته‌ها : در این مطالعه، نتایج حاصل نشان داد که وقتی عملکرد مدل های درختی کارت روی مجموعه آموزشی ارزیابی می‌شود. شاخص‌های اندازه گیری شده حساسیت، ویژگی و دقت کل نسبت به حالتی که عملکرد هر مدل روی داده آزمایشی متناظر با آن مدل بررسی می‌شود، بیشتر است . وقتی کل داده به عنوان مجموعه آموزشی انتخاب شد مقدار شاخص های حساسیت و دقت مدل به ترتیب برابر ۰۶ و ۰.۵۹ به دست امد. برای مجموعه آموزشی۵۰ درصد میزان حساسیت و دقت کلی به ترتیب برابر ۰۶و۰.۵۹ به دست آمد. در حالی که این مقادیر برای مجموعه آزمایشی متناظر با آن مدل به ترتیب برابر۰.۵۲و ۰.۵۳ به دست آمد. همچنین برای مجموعه آموزشی ۶۵ درصد میزان حساسیت و دقت مدل به ترتیب برابر ۰.۵۱ و ۰.۵۹ بود که برای مجموعه آزمایشی متناظر این مقادیر به صورت ۰.۴۵ و ۰.۵۴ به دست آمد. برای الگوریتم بگینگ میزان دقت کلی برابر ۰.۵۶ و حساسیت مدل برابر ۰.۵۵ به دست آمده است. پایداری نتایج به دست آمده از الگوریتم بگینگ، قابل استفاده برای داده های مستقل است.

نتیجه گیری : در این بررسی مشخص گردید که در پیش بینی گرایش افراد به جراحی زیبایی، از نظر پارامترهای حساسیت، ویژگی و دقت کل مدل بگینگ نسبت به مدل درختی کارت وضعیت بهتری دارد. زیرا وقتی از مدل کارت استفاده می‌شود شاخص‌های عملکرد در مجموعه داده جدید کاهش می یابند. در حالی که وقتی از مدل، بگینگ استفاده می شود جواب های پایدارتری به دست می آید که برای داده‌های مستقل نیز این نتایج به دست می‌آید. که در کل استفاده از نتایج مدل بگینگ برای داده های مستقل آینده منطقی تر هستند.

واژه های کلیدی : مدل درختی کارت، بوت استرپ تجمعی، بگینگ

Abstract

Background & Objective : Prediction of the possibility of a outcome is of key importance in many medical studies. Many researchers are willing to use tree models to achieve their goals, because the relative ease of interpreting classified results compared with other methods is an exclusive feature of this model. In many medical studies, to predict the probability of an outcome, researchers fit their models to the intended data and then evaluate the performance of the model on the same data. This approach produces good results, but the obtained results are not extensible to other independent data. The aim of this study was to demonstrate the ability of Bootstrap aggregating (bagging) in improving and stabilizing the CART model.

Methods: CART tree model and bagging algorithm were applied on a dataset of ٤٠٤ items related to people's attitude toward cosmetic surgery, which had been previously gathered by means of questionnaire in Kerman, Iran. To evaluate the accuracy of CART tree models, different percentages of training set and test set are used. The percentages of training sets used in this study were ٥٠, ٦٥ and ٧٥. The bagging model was fitted to ١٠٠ trees and then applied on the data as bootstrap samples were randomly extracted from the data as training datasets for each tree, while replacing with amounts equal to the original data. Since some observations are not present in the training dataset, they are used in the evaluation of each tree's performance. Final group prediction for each subject was determined following majority voting.For each model, sensitivity, specificity and the total prediction accuracy were separately calculated and compared. The fitting of all models was carried out by R software.

Result: When the whole data was used the overall accuracy was ٥٩٪. In the test data set and Bagging, accuracy reduced to ٥٣٪ and ٥٦٪. Corresponding figures in terms of sensitivity were ٦٠٪, ٥٢٪, and ٥٥٪.

Conclusion: We have seen that Bagging corrects the performance estimates for over optimization. Bagging method produces statistics which has higher chance for external validity.

Keywords: CART, external validity, bootstrap aggregating, data splitting, Bagging