دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد رشته آمار زیستی

عنوان:

مقایسه عملکرد مدل‌های درخت بقا و جنگل بقای تصادفی در پیش بینی بقای بیماران سکته حاد قلبی

توسط: ایمان یوسفیان

استاد راهنما: دکتر محمدرضا بانشی

سال تحصیلی: ۱۳۹۴-۱۳۹۳

ت

comparing predictive performance of Survival Tree with Random Survival Forest on the data of patients with Acute Myocardial Infarction

A Thesis
Presented to
The Graduate Studies

By

Iman Yosefian

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

Biostatistics

Kerman University of Medical Sciences
June 2015

چکیده:

مقدمه و اهداف: سادگی تفسیر و عدم محدودیت درساخت مدل‌های درختی منجر به استفاده روز افزون از این مدل‌ها در مطالعات پزشکی می‌شود. اما بیش برازش کردن داده و ناپایداری این مدل‌ها باعث بالا رفتن خطای پیش بینی و عدم تعمیم پذیری نتایج این مدل‌ها به داده‌ای دیگر می‌شود، هرس کردن درخت یکی از روش‌های کاهش خطای پیش بینی و بالا بردن قدرت تعمیم پذیری نتایج درخت است. از طرفی روش‌های گروهی همچون جنگل تصادفی با ساخت چندین مدل درختی در درون خود برروی نمونه‌های بوت استرپ گرفته شده از داده و ترکیب نتایج این مدل‌ها به بالا رفتن تعمیم پذیری نتایج نسبت به یک تک درخت کمک می‌کند. هدف از انجام این مطالعه سنجش میزان توانایی تعمیم پذیری یک درخت و جنگل بقای تصادفی با استفاده از عملکرد پیش بینی این مدل‌ها برروی داده‌ی آموزشی و آزمودنی است.

مواد و روش‌ها: در این مطالعه از اطلاعات مربوط به بقای ۶۰۷ بیمار پذیرش شده با سکته قلبی حاد در بخش CCU بیمارستان امام رضا مشهد استفاده شده است. برای تجزیه و تحلیل داده‌ها ابتدا داده را به دو بخش آموزشی و آزمودنی افراز کرده، سپس سه مدل درخت بقای اشباع شده، درخت بقای هرس شده و جنگل بقای تصادفی را برروی داده‌ی آموزشی برازش داده شده است. برای سنجش میزان درستی پیش بینی این مدل‌ها از آماره‌های C-index و Integraed Brier Score بهره گرفته شده است، با توجه به اختلاف میزان صحت عملکرد پیش بینی در داده‌ی آزمودنی از داده‌ی آموزشی قابلیت تعمیم پذیری دو مدل با یکدیگر مقایسه می‌شود. تمام تحلیل‌های ذکر شده براساس دو نمونه تصادفی ۱۰۰ و ۳۰۰ تایی از داده‌ی اصلی نیز انجام شده است تا اثر تغییر حجم نمونه برروی نتایج سنجیده شود.

یافته‌ها: اختلاف شاخص C-index داده‌ی آزمودنی از داده‌ی آموزشی در مدل درخت بقای اشباع شده (۰/۲۳۷) بسیار بیشتر از این اختلاف در درخت بقای هرس شده (۰/۰۶۶) است، این اختلاف در جنگل بقای تصادفی به مقدار زیادی کاهش می‌یابد (۰/۰۰۶). عملکرد پیش بینی مدل درخت بقای اشباع شده برروی داده‌ی آموزشی براساس شاخص C-index بسیار خوب (۰/۸۷۳) ولی در داده‌ی آزمودنی کاهش محسوسی می‌کند (۰/۶۳۶). در درخت بقای هرس شده شاخص C-index به ترتیب برای داده‌ی آموزشی و آزمودنی برابر ۰/۷۵۷ و ۰/۶۹۱ است همچنین عملکرد مدل جنگی بقای تصادفی در پیش بینی داده‌ی آزمودنی (۰/۷۱۳) تقریبا مشابه داده‌ی آموزشی (۰/۷۰۷) است. همچنین تفاوت آماره IBS داده‌ی آزمودنی از داده‌ی آموزشی براساس مدل درخت بقای اشباع شده، هرس شده و جنگل بقای تصادفی به ترتیب برابر ۰/۱۳۷، ۰/۰۲۵ و ۰/۰۰۱ است. در نمونه‌های ۱۰۰ و ۳۰۰ تایی تفاوت چندانی در نتایج مشاهده نشد.

نتیجه گیری: استفاده از روش‌های گروهی در ساخت مدل‌ها به مراتب عملکرد بهتر یک تک مدل را در برخواهد داشت، خصوصا مواردی که مدل به بی‌ثباتی و عملکرد پیش بینی ضعیف معروف باشد. مدل درخت بقا روش جایگزینی برای مطالعاتی با هدف آنالیز بقا بیماران با سکته قلبی حاد ارائه می‌دهد اما غریزه این مدل برای برازش کامل داده‌ی مورد مطالعه باعث بی‌ثباتی و تعمیم پذیری پایین نتایج این مدل به داده‌ای دیگر است، استفاده از روش جنگل تصادفی این بی‌ثباتی را تا حدود زیادی کاهش می‌دهد و باعث عملکرد بهتر مدل در پیش بینی داده‌ای غیر از داده‌ی مورد مطالعه می‌شود.

واژه‌های کلیدی: درخت بقا، جنگل تصادفی، جنگل بقای تصادفی، درخت اشباع شده، درخت هرس شده و سکته قلبی حاد

Abstaract:

Background & Objective: Easily interpretable and no limitations for constructing the tree models due to use of these models in medical researches. But, instability and low generalization ability of these models lead to define methodes such as pruning and Random Forest. Random Forest that known as ensemble methodes construct several base models and combine these models to achive high generalization ability of a single base model. the goal of this study compare a Saturated Tree, Pruned Tree and Random Forest by using accuracy predictive performance and generalization ability.

Methods: In this study used survival informations of 607 patients admitted to the CCU of imam Reza hospital Mashhad-IRAN with Acute Myocardial Infarction. For analyzing, data set splitted to training set and tet set. Models fitted on training set, then assessed predictive performance this models by using C-index and Integrated Brier Score statistics on training and test sets. Accroding to differece between accuracy prediction on test set and training set of each model compared generalization ability of the two models. For evaluatting impact of data size on results, all the above analyzes done in 100 and 300 sample data.

Results: Difference of C-index statistic on training and test set in Saturated Tree was too much (0.237) in comparison with Pruned Tree (0.066) and RSF (0.006). Performance of Saturated Tree and Pruned tree for prediction on training set was high (0.873 and 0.757) but on test set was low (0.636 and 0.691). While, this difference in RSF reduced, Performance of RSF for prediction on training set (0.7045) and test set (0.7110) are approximately the same. Differences for IBS statistics in Saturated Tree, Pruned Tree and Random Forest were Respectively 0.137, 0.025 and 0.001. By using 100 and 300 random sample of dataset unseen a great difference in results that was mentioned.

conclusion: By using ensemble methods in constructing the base models due to the better performance in comparison to a single base model, specially when the base model is known as instability model that has a low perdictive performance. Random Survival forest as a ensemble method applied in survival tree until decrease instability and increase performance of a single survival tree for prediction future data.

keywords: Survival tree, Random Forest, Random Survival Forest, Saturated Tree, Pruned tree and Acute Myocardial infarction.