دانشگاه علوم پزشکی کرمان
مرکزی

دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد رشته آمار زیستی

عنوان:

مقایسه ویژگی های مدل شبکه عصبی مصنوعی و مدل درختی در تشخیص عوامل موثر بر تزریق مواد در زندانیان

توسط: اعظم رستگاری

استاد راهنما: دکتر محمدرضا بانشی

استاد مشاور: دکتر علی اکبر حق دوست

سال ۱۳۹۱

Comparison of characteristics of Neural Network Analysis and Decision Tree Modelon Prediction of factors influence drug injection in prison

A Thesis

Presented to

The Graduate Studies

By

Azam Rastegari

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in:

Biostatistics

Kerman University of Medical Sciences

December 2012

چکیده پایان نامه:

مقدمه: امروزه از مدل های آماری در حیطه پزشکی برای اهداف گوناگونی از قبیل پیش بینی، تشخیص بیماری، تعیین تأثیر یک مداخله ، بررسی اثربخشی هزینه ها، بهبود کیفیت مراقبت های بهداشتی و درمانی، تخصیص بیماران به گروه های مختلف از نظر خطر بروز و تخصیص یک روش یا مداخله درمانی به گروه خاصی از بیماران و غیره استفاده می شود.

در حال حاضر برای انواع مطالعات با هر نوع پیش فرض و تحت هر شرایطی، روش های آماری مناسبی وجود دارد که محققین می توانند با توجه به شرایط مطالعه، مناسب ترین روش را برای تحلیل انتخاب کنند. یکی از مسائل اصلی در مطالعات پزشکی تعیین متغیرهای مستقلی هستند که بیشترین تأثیر را بر روی پیامد دارند. روش های کلاسیک برای چنین مطالعاتی استفاده از مدل های رگرسیونی می باشد.

روش های رگرسیونی یکی از روش های کلاسیک (استاندارد) برای مدلسازی رابطه بین متغیر وابسته و متغیر مستقل می باشد. در این روش معمولاً از P-value برای تعیین معنی داری متغیرها استفاده می شود. مدل های رگرسیونی معمولاً برای حجم زیاد داده محدودیت هایی دارند. به طوری که معنی داری آماری در این گونه مطالعات بسیار تحت تأثیر حجم نمونه و توان آزمون قرار می گیرد. در مدل های رگرسیونی خطی باید فرض نرمال بودن متغیر وابسته برقرار باشد. همچنین برای تفسیر فاصله اطمینان ها که یکی از مزایای مدل های رگرسیونی هستند نیز به این فرض نیاز است.

یکی دیگر از فرض هایی که باید قبل از استفاده از مدل های رگرسیونی بررسی شود تساوی واریانس توزیع متغیر وابسته در گروه های متغیر مستقل می باشد. علاوه بر این برای استفاده از مدل های رگرسیونی فرض بر این است که متغیر های مستقل نسبت به یکدیگر مستقل اند. که همیشه این گونه نیست مخصوصاً در مورد داده های پزشکی که اغلب متغیرها به هم وابسته هستند.

با توجه به موارد مذکور محققین باید با روش های جایگزین این گونه مدل ها آشنا شوند تا در هنگام مواجهه با چنین شرایطی بتوانند مناسب ترین مدل را انتخاب کنند. از جمله روش های جایگزین، می توان به مدل مدل های درختی و شبکه عصبی مصنوعی نام برد.

بنابراین در این مطالعه خواستار مقایسه ویژگی های این دو مدل و بررسی عوامل موثر بر آن ها هستیم.

روش کار: برای برازش مدل ها در مطالعات معتبر علمی از درصد های متفاوتی برای ایجاد دو مجموعه داده آموزشی و آزمایشی استفاده شده است. و در این مطالعه نیز این تنوع در نظر گرفته شده و برای تولید این دو مجموعه داده از چهار درصد متفاوت به صورت زیر استفاده شده است (۵۵٪ ، ۶۵٪ ، ۷۵٪ ، ۸۵٪).

برای برازش مدل درختی از روش های کارت و چید استفاده شده است. برای انتخاب بهترین مدل شبکه عصبی مصنوعی برای هر چهار مجموعه داده ساخته شده از داده اصلی مدل های متفاوتی با تعداد نرون های ۵، ۱۰، ۱۵ و ۲۰ در لایه پنهان برازش داده شده است. و همچنین برای بررسی تأثیر میزان نرخ یادگیری در مراحل آموزش و نقش آن در میزان خطای مدل از مقدارهای متفاوت ۰/۱، ۰/۳، ۰/۵، ۰/۷ و ۰/۹ برای نرخ یادگیری استفاده شده است. برای هر مدل به طور جداگانه میزان حساسیت، ویژگی، سطح زیر منحنی ROC و میانگین مجموع مربعات خطا (MSE) محاسبه گردید. برازش تمامی مدل ها توسط نرم افزار SPSS17 انجام شده است.

یافته ها: در بین تمام مدل های شبکه عصبی مصنوعی برازش داده شده با تعداد متفاوت نرون در لایه پنهان و میزان نرخ های یادگیری متفاوت و روی چهار مجموعه داده، کمترین میانگین مجموع مربعات خطا مربوط به مدل است که روی مجموعه داده آموزشی ۶۵ درصد برازش داده شده است(۰/۳۱۱). که این مدل دارای ۵ نرون در لایه پنهان می باشد. همچنین کمترین مجموع مربعات خطا در مدل های درختی مربوط به مدل چید می باشد(۰/۳۱۸) که روی مجموعه داده ۶۵٪ برازش داده شده است.

نتیجه گیری: باید در نظر داشت که امروزه روش های آماری بسیار متنوعی برای تجزیه و تحلیل داده ها و برای انواع هدف ها موجود است. اما برای استفاده از تمامی این مدل ها باید دقیقاً به بررسی شرایط حاکم بر داده ها و پیش فرض های مورد نیاز مدل ها پرداخته شود. و سپس با توجه به نوع هدف و کاربرد آن به انتخاب روش آماری مناسب اقدام شود. بدون شک هر مدل آماری نقاط قوت و ضعفی دارد که باید به آن ها توجه شود. در این مطالعه مشاهده شد که مدل شبکه عصبی مصنوعی بهتر از مدل های درختی عمل کرد اما باید به این نکته بسیار مهم توجه داشت که مدل شبکه عصبی مصنوعی از عوامل بسیاری که کاربر می تواند در ساختار شبکه به اختیار خود انتخاب کند، تأثیر می پذیرد. از جمله این عوامل عبارت اند از تعداد لایه پنهان و تعداد نرون های این لایه، نوع تابع محرک در لایه های پنهان و خروجی، میزان نرخ یادگیری و اندازه مومنتم و غیره. در صورتی که مدل درختی گزینه های زیادی برای تغییر ندارد.

Abstract

Introduction:

Nowadays the statistical models are used in medicine for different goals such as prediction, classification, assessment of intervention effect, Cost Effectiveness analysis, assessment of improvment in the the quality of health care, and allocation of patients to different groups according to risk.

For different study designs and under different circumstances, there is an appropriate statistical method. So researches can select the best model for analysis. One of the main issues in medical studies is to identify the variables that have the greatest impact on dependent variable. Traditional statistical methods for these studies are regression models.

One of the traditional methods to model the relationship between the dependent variable and the independent variable are regression models. In these methods P-value is used for specification of statistical significant. However, regression models are useful when eniguh data is available. When number of events, relative to number of variables, is low, regression models is not a powerful tool. Another assumption that we must check before use the regression model is equality of variances in different group of dependent variable. In addition the independent variables should be independent from each other. However, the data might not always follow these assumptions. According to these issues researchers should be familiar with the alternative models such as tree-based and neural network models. The aim of this study is to compare the performance of the tree-based and neural network models.

Method:

To fit the models in we used four different percentages of data as training set (55%, 65%, 75%, and 85%).

We constructed two different trees: CART and CHAID model. To select the best model of neural network we used different number of neuron in the hidden layer (5, 10, 15, and 20). Also To evaluate the effect of learning rate on the errors of the models we use different value for learning rate (0.1, 0.3, 0.5, 0.7, and 0.9). For each model we calculate sensitivity, specificity, area under the ROC curve and mean square error (MSE). All analyses are done using SPSS version 17.

Result:

Among all neural network models (with different neuron in hidden layer and different learning rate), the best t model was thoseused 65% of data as training data set. We have seen that the lowest mean square error (0.311) when 5 neurons were used in the hidden layer.. Finally for tree based model, CHAID model was fitted on 75% traming data set With the lowest mean square error (0.318)

Conclusion: Nowadays there are very different statistical methods to analyze data with different goals but at first we must explore the characteristics of data and model assumptions. Depending on the purpose of the study, we must use appropriate statistical methods, and consider the strengths and weaknesses of the models. In this study the neural network model works better than tree models. We should emphasize that many factors have influence on the performance of neural network model, for example: number of neurons in hidden layer or the value of learning rate, activation function, and momentum factor.