Assessing the Performance of Classical and Modern Classification Methods: LR, TR, RF and SVM

Citation:

Assessing the performance of classical and modern classification methods: LR, TR, RF, and SVM., **A. Khademi**, Northeastern Educational Research Association, Trumbull, CT, USA (October 21-23, 2015).

Abstract

A primary purpose in educational testing is achieving optimal decision making results regarding the examinees (admit/reject) or the psychometric aspects of the test (e.g. identifying DIF items or a cut-off score). We assess and compare classification accuracy of the classical logistic regression with tree classification, random forests and support vector machines.

Introduction

Classification requirements in educational settings may occur in two distinct situations: (1) predicting an educational outcome (e.g. admit/reject, pass/fail, average/good/excellent performance) using known examinee attributes (e.g. educational track, GPA, SES), and (2) identifying an instrument or item as featuring or not a certain psychometric characteristic (e.g. identifying a cut-off score on a test or flagging differential functioning items or noninvariant instrument properties). Appropriate statistical methods have been developed in the psychometric literature for both situations. The former problem is usually addressed with logistic regression (or multinomial regression) or discriminant analysis in case of observed scores. The latter problems are addressed traditionally with logistic function families of methods (based on binomial distribution) or latent score methods, such as structural equation modeling and item response theory. Logistic regression is the most widely used classification method in both situations because of its intuitive probability output and the sigmoid curve it produces. Nevertheless, many other classification methods have been developed in recent years that may compete with LR in different regards, including sample size requirements, simultaneous estimation and data completeness requirements. Some of those new classification methods include the lasso, random forests and support vector machines, which have been mainly developed in statistical/machine learning fields but have recently found their way into psychometrics and other applied fields. For

instance, Magis, Tuerlincks, and de Boeck, (2015) have developed a lasso approach for detecting differentially functioning items (DIF), which performs better than logistic regression and Mantel-Haenzsel methods. Gao and Rogers (2011) used classification tree models to verify item difficulty properties in a reading test. Lee and Wollack (2015) used classification tree models and bagging to determine an appropriate cut-off score for placement of students in a mathematics course. Westreich, Lessler, and Funck (2010) used neural networks, support vector machines, decision trees (CART), and meta-classifiers to estimate propensity scores. Researchers in other fields have also extensively used modern classification methods in different data-specific contexts (see for example, Austin, Tu, Ho, Levy, Lee, 2013; Azar, & El-Said, 2014).

Logistic Regression

Logistic regression model uses the logistic function to predict a binary response using maximum likelihood estimation. Logistic regression function with multiple predictors is expressed as:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Classification Tree

Recursive binary splitting is used to classify observations. An observation falls into the most commonly occurring class of training observations in a region.

$$f(x) = \sum_{m=1}^{M} C_m \cdot 1_{(X \in R_m)}$$

Random Forests

Random forests is a tree-based prediction method in which the trees are decorrelated and a random sample of predictors is chosen as split candidates. The random forests predictor is:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B} \sum_{b=1}^{B} T(x; \Theta_b)$$

Support Vector Machines

Support Vector Machines use kernels to produce a nonlinear classification boundary. A kernel is function that quantifies the similarity of two observations. The nonlinear SVM function with kernel is:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

Statement of the Problem

As stated above, alternative classification methods such as classification trees, random forests, and support vector machines have only recently made their way into the psychometric field from statistical learning area. In addition, very few studies have been conducted in the field to assess and compare the performance of these modern classification methods using educational data. Therefore, it is reasonable to assess and compare these new classification methods in the educational research field and compare them with the established methods, such as logistic regression. In this study, we attempt to assess the classification accuracy of the lasso, random forests and support vector machines vis-à-vis logistic regression.

Method

The data for this study includes a subsample of 1,842 (removing 139 rows for missing values) observations on one response variable and six predictor variables from the High School & Beyond dataset for the 10th graders. The original response variable is the math achievement score for 10th grade students from across the US. The outcome variable is then artificially dichotomized to distinguish those who are deemed qualified or ready for a STEM program in college (STEM Ready (categorical).STEM Ready is artificially dichotomized, Yes, if math > 55.01, No if math < 55.00). The predictor variable are science scores, gender (female, male), program type (general, academic, other), school type (public, catholic, private), race (Asian, Hispanic, Black, White) and SES (higher score, higher SES).

Results

Table 1: Variables used in the study.

Variable	**Attributes**
Science	**Min: 10.28, Max: 34.68, Mean: 2292, SD:6.29**
SES	**Min: -2.23, Max: 2.30, Mean: 0,256, SD: 0.817**
Gender	**Female: n= 917, Male: n= 925**
Race	**Asian (n=210), Black (n=181), Hispanic (n = 255), White (n=1196)**
Program	**Academic (n=1000), General (n=566), Other (n=276)**
School	**Catholic (n=151), Private (n=432), Public (n=1259)**
STEM Ready (response)	**Yes = 678, No = 1164**

Table 2: Logistic regression confusion table on test data (mean accuracy= .8469).

	No	Yes	Total
No	511 (True negative)	77 (False positive)	588
Yes	64 (False negative)	269 (True positive)	333
Total	575	346	921

Table 3: Tree classification confusion table on test data (mean accuracy= .800).

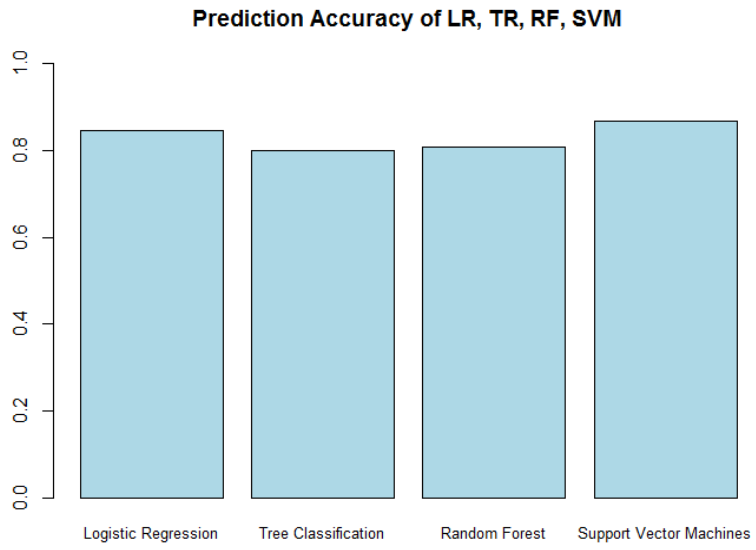
	No	Yes	Total
No	522 (True negative)	90 (False positive)	612
Yes	65 (False negative)	244 (True positive)	309
Total	587	334	921

Table 4: Random forest confusion table on test data (mean accuracy= .8067; trees grown=500; OOB error rate= 19.33; package: randomForest_4.6-12).

	No	Yes	Total
No	502 (True negative)	83 (False positive)	585
Yes	95 (False negative)	241 (True positive)	336
Total	597	324	921

Table 5: Support Vector Machines confusion table on test data (mean accuracy= 0.8677; gamma=0.001, cost=100, CV= 10-fold; kernel=radial, # of support vectors= 317)

	No	Yes	Total
No	384 (True negative)	211(False positive)	595
Yes	178 (False negative)	148 (True positive)	326
Total	562	359	921



Educational Importance of the Study

Researchers in the educational setting need to arrive at sensitive decisions for high-stakes situations, such as placement, determining a cut-off score and identifying DIF items. Different statistical methods have been developed to aid in the decision making process. However, not all of these methods may prove accurate and optimal in educational settings. The results of the present study helps researchers, policy makers and practitioners in the educational testing field to choose a classification method with the highest accuracy rate for educational purposes. As the results in this study show, logistic regression holds strongly compared to other classification methods. However, we can see that the Support Vector Machines outperforms LR. Therefore, SVM can be a promising algorithm to be deployed in the psychometrics field.

References

Austin, P. C., Tu J. V., Ho, J., E., Levy, D. & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology, 66, pp. 398-407.

Azar, A. T. & El-Said, S. A., (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Computational and Applications, 24, pp. 1163-1177.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(1), 77–104. doi: 10.1177/0265532210364380

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. New York: Springer.

Lee, C. & Wollack, J. (2015). *Using Classification Tree Models and Bagging to Determine Course Placement. Paper presented at Graduate Student Research Session, the 77th Annual Meeting of the National Council on Measurement in Education, April 15-19, 2015, Chicago, Illinois, USA.*

Magis, D., Tuerlincks, F. & de Boeck, P. (2015). Detection of differential item functioning using the lasso approach. Journal of Educational and Behavioral Statistics, 40(2).

United States Department of Education. National Center for Education Statistics. High School and Beyond, 1980: Sophomore and Senior Cohort Third Follow-up (1986). ICPSR08896-v3. Ann Arbor, MI: Inter-university Consortium for Political and Social Research[distributor], 2014-01-21. http://doi.org/10.3886/ICPSR08896.v3

Westreich, D., Lessler, J., & Funck, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology, 63(8). pp. 826-833.