Computational Statistics and Data Analysis 72 (2014) 158–175

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Nonparametric variable selection and classification: The CATCH algorithm

Shijie Tang[a], Lisha Chen[b,*,1], Kam-Wah Tsui[c], Kjell Doksum[c,d]

[a] Millennium Pharmaceuticals, The Takeda Oncology Company, Cambridge, MA, United States
[b] Department of Statistics, Yale University, New Haven, CT, United States
[c] Department of Statistics, University of Wisconsin, Madison, WI, United States
[d] Department of Statistics, Columbia University, New York City, NY, United States

ARTICLE INFO

Article history:
Received 6 April 2012
Received in revised form 23 October 2013
Accepted 23 October 2013
Available online 20 November 2013

Keywords:
Adaptive variable selection
Importance score
Chi-square statistic

ABSTRACT

The problem of classifying a categorical response Y is considered in a nonparametric framework. The distribution of Y depends on a vector of predictors X, where the coordinates X_j of X may be continuous, discrete, or categorical. An algorithm is constructed to select the variables to be used for classification. For each variable X_j, an importance score s_j is computed to measure the strength of association of X_j with Y. The algorithm deletes X_j if s_j falls below a certain threshold. It is shown in Monte Carlo simulations that the algorithm has a high probability of only selecting variables associated with Y. Moreover when this variable selection rule is used for dimension reduction prior to applying classification procedures, it improves the performance of these procedures. The approach for computing importance scores is based on root Chi-square type statistics computed for randomly selected regions (tubes) of the sample space. The size and shape of the regions are adjusted iteratively and adaptively using the data to enhance the ability of the importance score to detect local relationships between the response and the predictors. These local scores are then averaged over the tubes to form a global importance score s_j for variable X_j. When confounding and spurious associations are issues, the nonparametric importance score for variable X_j is computed conditionally by using tubes to restrict the other variables. This variable selection procedure is called CATCH (Categorical Adaptive Tube Covariate Hunting). Asymptotic properties, including consistency, are established.

1. Introduction

We consider classification problems with a large number of predictors that can be numerical or categorical. We are interested in the case in which many of the predictors may be irrelevant for classification, thus the variables useful for classification need to be selected. In genomic research, the classes considered are often cases and controls. Thus variable selection is the same as the important problem of deciding which variables are associated with disease. With modern technology, data of large dimension arise in many scientific disciplines including biology, genomics, astronomy, economics and computer science. In particular, the number of variables can be greater than the sample size, which poses a considerable challenge to the statistical analysis. If only some of the variables are useful for classification, over-fitting is a problem for methods that use all variables, and therefore variable selection becomes critical for statistical analysis.

Methods for variable selection in the classification context include methods that incorporate variable selection as part of the classification procedure. This class includes random forest (Breiman, 2001), CART (Breiman et al., 1984) and GUIDE

[*] Correspondence to: 24 Hillhouse Ave, New Haven, CT 06511, United States. Tel.: +1 203 432 5785; fax: +1 203 432 0633.
E-mail address: lisha.chen@yale.edu (L. Chen).

[1] Corresponding and joint first author.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.024

(Loh, 2009). Random forest assigns an importance score for each of the predictors and one can drop those variables whose importance score fall below a certain threshold. CART, after pruning, will choose a subset of optimal splitting variables to be the most significant variables. GUIDE is a tree-based method particularly powerful in unbiased variable selection and interaction detection. Other research includes methods that incorporate variable selection by applying shrinkage methods with L_1 norm constraints on the parameters (Tibshirani, 1996) that generate sparse vectors of parameter estimates. Wang and Shen (2007) and Zhang et al. (2008) incorporate variable selection with classification based on support vector machine (SVM) methods. Qiao et al. (2008) consider variable selection based on linear discriminant analysis. For these methods, the variable selection is obtained by imposing sparsity on the coefficients in parametric models or in models with prespecified basis functions using L_1 regularization. Hence these methods may not work well when there is a complex relationship between the predictors and the variables to be classified.

We propose a nonparametric method for variable selection called Categorical Adaptive Tube Covariate Hunting (CATCH) that performs well as a variable selection algorithm and can be used to improve the performance of available classification procedures. The idea is to construct a nonparametric measure of the relational strength between each predictor and the categorical response, and to retain those predictors whose relationship to the response is above a certain threshold. The nonparametric measure of importance for each predictor is obtained by first measuring the importance of the predictor using local information, and then combining such local importance scores to obtain an overall importance score. The local importance scores are based on root chi-square type statistics for local contingency tables.

In addition to the aforementioned nonparametric feature, the CATCH procedure has another property: it measures the importance of each variable conditioning on all other variables thereby reducing the confounding that may lead to selection of variables spuriously related to the categorical variable Y. This is accomplished by constraining all predictors but the one we are focusing on. For the case where the number of predictors is huge ($d \gg n$), this can be done by restricting principal components for some types of studies. See Remark 3.5.

Our approach to nonparametric variable selection is related to the EARTH algorithm (Doksum et al., 2008) which applies to nonparametric regression problems with a continuous response variable and continuous predictors. It measures the conditional association between a predictor i and the response variable conditional on all the other predictors $\{j\}_{j\neq i}$, by constraining the $\{j\}_{j\neq i}$ variables to regions called tubes. The local importance score is based on a local linear or a local polynomial regression. See also Miller and Hall (2010) who later considered variable selection procedure similar to EARTH. The contribution of the current paper is to develop variable selection methods for the classification problem with a categorical response variable and predictors that can be continuous, discrete or categorical.

Our CATCH algorithm can be used as a variable selection step before classification. Any classification method, preferably nonparametric, can be used after we single out the important variables. In particular, SVM and random forest are statistical classification methods that can be used with CATCH. We show in simulation studies that when the true model is highly nonlinear and there are many irrelevant predictors, using CATCH to screen out irrelevant or weak predictors greatly improves the performance of SVM and random forest.

The CATCH algorithm works with general classification data, with both numerical and categorical predictors. Moreover, the CATCH algorithm is robust when the predictors interact with each other, especially when numerical and categorical predictors interact, e.g., for hierarchical interactive association between the numerical and categorical predictors. We present a model with a reasonable sample size and predictor dimension for which random forest has a relatively low chance of detecting the significance of the categorical predictor. The CART and GUIDE algorithms, with pruning, can find the correct splitting variables including the categorical one, but yield relatively high classification errors. Moreover CART and GUIDE have trouble choosing the splitting predictors in the correct order. The CATCH algorithm achieves higher accuracy in the task of variable selection. This is due to the importance scores being conditional as illustrated in the simulation example in Section 4.1.

The rest of the paper will proceed as follows. In Section 2 we introduce importance scores for univariate predictors. In Sections 3.1–3.4 we extend these scores to multivariate predictors, and in Section 3.5, we introduce the CATCH algorithm. In Section 4, we use simulation studies to show the effectiveness of the CATCH algorithm. A real example is provided in Section 5. In Section 6, we provide some theoretical properties to justify the definition of local contingency efficacy, and in Section 7, we show asymptotic consistency of the algorithm.

2. Importance scores for univariate classification

Let $(X^{(i)}, Y^{(i)}),\ i = 1, \dots, n$, be independent and identically distributed (i.i.d.) as $(X, Y) \sim P$. We first consider the case of a univariate X, then use the methods constructed for the univariate case to construct the multivariate versions.

2.1. Numerical predictor: local contingency efficacy

Consider the classification problem with a numerical predictor. Let

$$\Pr(Y = c|x) = p_c(x), \quad c = 1, \dots, C, \qquad \sum_{c=1}^{C} p_c(x) \equiv 1. \tag{2.1}$$