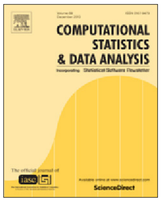
Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Greedy active learning algorithm for logistic regression models

Hsiang-Ling Hsu[a], Yuan-Chin Ivan Chang[b], Ray-Bing Chen[c,*]

[a] National University of Kaohsiung, Taiwan
[b] Academia Sinica, Taiwan
[c] National Cheng Kung University, Taiwan

ARTICLE INFO

Article history:
Received 22 December 2017
Received in revised form 13 August 2018
Accepted 14 August 2018
Available online xxxx

Keywords:
Active learning algorithm
D-efficiency criterion
Forward selection
Graft optimization

ABSTRACT

We study a logistic model-based active learning procedure for binary classification problems, in which we adopt a batch subject selection strategy with a modified sequential experimental design method. Moreover, accompanying the proposed subject selection scheme, we simultaneously conduct a greedy variable selection procedure such that we can update the classification model with all labeled training subjects. The proposed algorithm repeatedly performs both subject and variable selection steps until a prefixed stopping criterion is reached. Our numerical results show that the proposed procedure has competitive performance, with smaller training size and a more compact model compared with that of the classifier trained with all variables and a full data set. We also apply the proposed procedure to a well-known wave data set (Breiman et al., 1984) and a MAGIC gamma telescope data set to confirm the performance of our method.

1. Introduction

To train a classification model, labeled data are essential when a training/testing framework is adopted, and its classification performance relies on both the size and the quality of the training subjects used for learning. In a Big Data scenario, we might easily meet a huge data set; however, the labeled information may be limited, and an abundance of unlabeled subjects are available. To prevent money laundering, Deng et al. (2009) studied the method for building a detection model using bank account data. This is a good example because in this situation, the label of interest (money laundering account) is limited in a regular bank account data set. It would require a huge amount of time and resources to verify whether an account is suspicious or non-suspicious, even though the major parts of the transactions in a bank account should be normal. Efficiently determining the potential risks within a bank account in addition to effectively and efficiently using the unlabeled subjects to improve the classification rule is the key issue, and the concept of active learning can be applied to this situation.

When we train a classifier in an active learning manner, we need to annotate the unlabeled data and recruit them into the training set, which can be done with the information of a model built on the labeled data at the current stage. In the literature, it is observed that people can usually learn a satisfactory model economically with such a procedure (Cohn et al., 1994a; Settles, 2011, 2012). There are many classification performance indexes, and it is clear that this subject selection process may depend on the targeted index (Settles, 2009; Hsu, 2010; Settles, 2011). For example, Culver et al. (2006) studied active learning procedures that maximize the area under the ROC curve (AUC), Long et al. (2010) were interested in the

[*] Corresponding author.
E-mail addresses: hsuhl@nuk.edu.tw (H. Hsu), ycchang@stat.sinica.edu.tw (Y.I. Chang), rbchen@mail.ncku.edu.tw (R. Chen).

https://doi.org/10.1016/j.csda.2018.08.013
0167-9473/© 2018 Elsevier B.V. All rights reserved.

ranking of the data, and Deng et al. (2009) used an active learning study focusing on efficiently selecting the most informative subjects to join the training set to construct an accurate classification model via the experimental design methods. However, when there are many redundant variables (predictors), to have an effective design is difficult and in such a situation, this classification model tends to over-fit the training subjects, which usually leads to high prediction uncertainty. A common approach to improving its prediction stability is to increase the size of the training set at the cost of protracting the training stage. Thus, a procedure that can identify a compact classification model during its training course is preferred.

In this paper, we propose a logistic model-based active learning procedure with a batch sampling to address binary classification problems under big data scenarios. In addition to the subject selection scheme, we also integrate a variable selection step into this procedure for systematically improving the prediction ability and avoiding the over-fitting of our final classification model. Hence, the proposed algorithm is an iterative algorithm, in which we select a batch of new samples at each iteration and then update our binary classification model via a variable selection approach. Both subject selection and variable selection are featured in this novel active learning procedure.

We organize the rest of this paper as follows. Section 2 presents the details of the subject selection and variable selection steps, and then we describe our active learning algorithm with an integrated subject and variable selection steps. Section 3 presents numerical results, where in addition to the simulation studies, we apply our algorithm to a well-known wave data set used in Breiman et al. (1984) and a MAGIC gamma telescope data set. We present a brief discussion and conclusion in Section 4.

2. Methodology

We consider a pool-based active learning procedure as studied in Lewis and Gale (1994) and assume that to obtain those unlabeled data is cheap and to query their label information is expensive. Hence, we should rationally select the unlabeled subjects from this large pool for being labeled to reduce the overall cost of model learning. We state the general framework of the pool-based active learning methods as follows.

```
Initialization: An initial labeled training set and a pool of unlabeled data
repeat
    Learn the current model based on the current labeled training data.
    Select points from unlabeled set via a query strategy framework based on the current learned model.
    Query labels for these selected points and update the training set.
until The stopping criterion is satisfied;
```

It is known that when there is a lengthy vector of variables, to train a classifier with an entire set of variables may diminish its prediction power and protract its training time; a compact classification model is usually preferred when the model is sparse. Hence, in our algorithm, we also emphasize the variable selection strategy in its learning process in addition to the common subject selection as in active learning procedures reported in the literature.

The variable selection frame has two common approaches: forward selection and backward elimination (Whitney, 1971). Because active learning procedures will usually start from a small size of training samples and keep accumulating according to a pre-specified selection rule, we can only obtain satisfactory estimation results for a small number of parameters. Hence, in our study, we use the forward selection scheme that increases the size of the variable set by adding a new variable to the current model at a time and adopt a greedy selection approach. Based on the characters discussed above, we propose a modified active learning algorithm, which integrates both batch-subject selection and greedy variable selection features together, and we refer to this Greedy AcTivE learning algorithm as GATE learning (or GATE) algorithm throughout this paper. Basically, we implement a variable selection step once we have an updated training set, and in each iteration of GATE, we add more labeled samples, and re-justify our classification model.

2.1. Logistic model for binary classification

Let Ξ_S denote the index set of the whole sample points and Ξ_s be the current training index set with labeled data. Thus, $\Xi_s^c = \Xi_S \setminus \Xi_s$ is the pool of the unlabeled data. In this section, we focus on how to identify the batch unlabeled subjects from Ξ_s^c for binary classification based on a logistic model and, meanwhile, propose a two-stage query procedure by putting the uncertainty sampling with the optimal design criterion together. Afterward, we introduce a greedy forward selection to update the current model by selecting a candidate variable from $\Xi_v^c = \Xi_V \setminus \Xi_v$, where Ξ_V is the index set of the whole variables and Ξ_v denotes the index set of the current active variables in the logistic model.

Assume that the ith individual variate $Y_i \in \{0, 1\}$ is a binary variable with

$$P(Y_i = 1) = p_i = E(Y_i) \text{ and } P(Y_i = 0) = 1 - p_i,$$

and let the feature values of the ith subject be $\mathbf{x}_{i,p} = (x_{i,j})^\top,\ i \in \Xi_s,\ j \in \Xi_v$ whose dimension is equal to p. Then, we can fit this data set with a logistic regression model below :

$$p_i = F(\mathbf{x}_{i,p}|\boldsymbol{\beta}_p) = \frac{\exp\{\boldsymbol{\beta}_p^\top \mathbf{x}_{i,p}\}}{1 + \exp\{\boldsymbol{\beta}_p^\top \mathbf{x}_{i,p}\}},$$

Please cite this article in press as: Hsu H., et al., Greedy active learning algorithm for logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.013.