Active Learning Procedure via Sequential Experimental Design and Uncertainty Sampling

Eunsik Park*
Chonnam National University, Gwangju, Republic of Korea - espark02@gmail.com

Jing Wang
Chonnam National University, Gwangju, Republic of Korea

Abstract

Classification is an important task in many fields, including biomedical research and machine learning. Traditionally, a classification rule is constructed based on a bunch of labeled data. Recently, due to technological innovation and automatic data collection schemes, we easily encounter data sets containing large amounts of unlabeled samples. Because labeling each of them is usually costly and inefficient, the way to utilize these unlabeled data in a classifier construction process becomes an important problem. In machine-learning literature, active learning or semi-supervised learning are popular concepts discussed under this situation; classification algorithms recruit new unlabeled subjects sequentially based on the information learned from previous stages of its learning process, and these new subjects are then labeled and included as new training samples. From a statistical aspect, these methods can be recognized as a hybrid of the sequential design and stochastic approximation procedure. In this paper, we study sequential learning procedures for building efficient and effective classifiers, where only the selected subjects are labeled and included in its learning stage. The proposed algorithm combines the ideas of Bayesian sequential optimal design and uncertainty sampling. Computational issues of the algorithm are discussed. Numerical results using both synthesized data and real examples are reported.

Keywords: classification rule; stochastic approximation; D-optimal design; Bayesian estimation.

1. Introduction

Classification is an important task in many fields, including biomedical research, engineering, sociology and many others. The way to construct a classification rule based on a labeled data set is a classical statistical problem. In machine-learning literature, there are several types of learning problems discussed, and depending on how labeled subjects are included into a learning process, they are usually termed as supervised, unsupervised or semi-supervised learning. Recently, due to technical innovation, "big data" has become a buzz phrase in many fields, and we now often encounter data sets that have huge amount of unlabeled data. Hence, the way to utilize these unlabeled data efficiently to construct a classification rule becomes an important problem. Because labeling each unlabeled subject is usually costly and inefficient, a common approach is active learning. This type of learning process will only inquire the label information for the "selected" subjects, which are usually chosen based on the information in the previous learning stages, and then include the newly labeled subjects into its training stage. A learning process will usually continue until a prefixed criterion is reached, such as a prefixed total number of labeled subjects to be used in the training stage.

Moreover, because in an active learning process, subjects are dynamically and sequentially selected, labeled and then added to the training set, this process is naturally related to sequential experimental designs in statistics, where a new observation/experiment is conducted at some particular design points selected, according to the information obtained, using the data gathered up to the current stage. Since data are observed adaptively, these types of methods are also related to the stochastic approximation process. Their original procedure is the Robbins-Monro (RM) procedure and can be viewed as a stochastic version of Newton-Raphson method for nonlinear root-finding problems. Sequential design methods have been intensively studied, and there are even more papers that discussed different modifications of the RM procedure was further modified and their corresponding convergence rates. Recently, the RM procedure to improve its efficiency. This type of procedure is nonparametric in the sense that no parametric model assumption is presumed.

However, the RM procedure can also be derived from a parametric form. For example, using the maximum likelihood estimate (MLE) of a logistic model, a logit-MLE method was proposed for binary data that uses the currently available labeled data to fit a logistic model and then selects the next input with the desired probability based on the fitted logistic model. Because a classification rule construction with an active-learning framework can be formulated as a problem of estimating the threshold boundary between two groups, which can usually be defined by using a probability quantile, it can also be viewed as a stochastic root-finding procedure, as described above. Moreover, logistic models are commonly used models in binary classification problems, and the properties of sequential estimation for generalized linear model (GLM) under general adaptive designs are well studied. Hence, it is natural to construct a binary classification rule, sequentially and adaptively, by putting all of these ingredients together. An active learning algorithm developed in Deng et al.(2009), which combines the logit-MLE of Wu et al.(1985)) and D-optimal design is a successful example. This kind of a method depends on the properties of MLE. Although the existence and uniqueness of MLE can be achieved after a few initial observations, it may still suffer from a severe bias, when the sample size is small, which usually results in an inefficient learning process. In modern literature, Joseph et al. (2007) developed a Bayesian extension of Wu's approach, in which they used the maximum a posterior (MAP) estimates of the parameters of a logistic model, rather than MLEs. Dror and Steinberg(2008) suggested a new sequential experimental design for GLM, in which observations are selected sequentially based on a Bayesian D-optimality criterion and Bayesian estimates of model parameters. These methods motivate us to study a novel modification of Deng et al. (2009).

As in a conventional regression analysis, it is well-known that when the degree of dimensionality of the unknown vector of parameters becomes large, its estimated information will be very unstable. Because active learning processes usually rely this type of information, the unstable estimates of parameters will also affect the learning process. In the real example studied in Deng et al. (2009), two variables are selected based on experts' opinions. However, this situation is rare, and there are usually more variables considered for a real example. Thus, the stabilization of a learning process in a high-dimensional case is difficult and important. In this paper, we focus on the higher-dimensional data sets. A Bayesian sequential design is used, and the related computational issues are discussed. In addition, for practical usages, we also study the effects of using different sizes of labeled data sets for an initial training set of an active learning process. In terms of the subject selection during a process, the major difference between a sequential design and an active learning process is that with a sequential design, an experiment will be conducted at the selected points, while in active learning processes with existent unlabeled data, we can only select points near the theoretical ones from an existent data set. Hence, the way of selecting the next point based on the available information plays a key role in an active learning process. Deng et al. (2009) aimed at shortening the distance between the estimated boundary and the true one, such that the subject selection scheme heavily depends on the initial model assumption. In practice, the form of true model is usually unknown. Hence, in order to diminish the effect of model assumptions, we adopt a different design point selection scheme. The advantage of the proposed method will be discussed from both theoretical and practical aspects.

2. Model and Parameter Estimation

Let $\mathbf{x} = (x_1, \ldots, x_p)^T$ be the explanatory vector of subjects and variable $Y = 1$ or $Y = 0$ denote the category that a subject belongs to. Suppose that $P(Y = 1|\mathbf{x}) = F(\mathbf{x})$ is the probability model of $Y = 1$ given $\mathbf{x}$. Also, assume that each variable has a positive relationship with the response; that is, for a larger value of x_j, there is a higher probability of $Y = 1$. Deng et al. (2009) assumed that $F(\mathbf{x})$ had a parametric form

$$F(\mathbf{x}|\boldsymbol{\theta}) = \frac{e^{(z-\mu)/\sigma}}{1 + e^{(z-\mu)/\sigma}}, \tag{1}$$

where $z = \sum_{i=1}^{p} w_i x_i$, $0 < w_i < 1$ for each i, and $\sum_{i=1}^{p} w_i = 1$. Let $\boldsymbol{\theta} = (\mu, \sigma, w_1 \ldots, w_{p-1})^T$ be a vector of $p + 1$ parameters. Also, following model (1), for a given $\mathbf{x}$, Y is a Bernoulli random variable with mean $E(Y|\mathbf{x}) = F(\mathbf{x}|\boldsymbol{\theta})$. Defining $\tilde{\mathbf{x}}^T = (1, \mathbf{x}^T)$ and $\boldsymbol{\beta} = (-\mu/\sigma,\ w_1/\sigma,\ \ldots\ ,\ w_p/\sigma)^T$, we can re-write model (1) as a conventional logistic regression model:

$$F(\mathbf{x}|\boldsymbol{\beta}) = \frac{e^{\tilde{\mathbf{x}}^T\boldsymbol{\beta}}}{1 + e^{\tilde{\mathbf{x}}^T\boldsymbol{\beta}}}. \tag{2}$$

The Fisher information matrix of $\boldsymbol{\beta}$ with a set of design points $d = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$\mathbf{I}(\boldsymbol{\beta}; d) = \mathbf{X}^T \mathbf{W} \mathbf{X}, \tag{3}$$

where $\mathbf{X}$ is the regression matrix with ith row, $i = 1, \ldots, n$ equal to $(1, x_{i1}, \ldots, x_{ip})$, and $\mathbf{W}$ is a diagonal matrix with $w_{ii} = F(\mathbf{x}_i|\boldsymbol{\beta})\,[1 - F(\mathbf{x}_i|\boldsymbol{\beta})]$, $i = 1, ..., n$. It is clear that this information matrix is non-linear in $\boldsymbol{\beta}$ and depends on the unknown $\boldsymbol{\beta}$ only through $\mathbf{W}$.
Suppose that $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are observed labeled data of size n. Using this training set, we obtain MAP estimates of both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_{1,n}, \cdots, \hat{w}_{p-1,n})^T$ and $\hat{\boldsymbol{\beta}}_n = (-\hat{\mu}_n/\hat{\sigma}_n, \hat{w}_{1,n}/\hat{\sigma}_n, \cdots, \hat{w}_{p,n}/\hat{\sigma}_n)^T$ denote these two estimates. Using the current estimates of parameters, the classification rule based on the estimate of F becomes

$$\begin{cases} \hat{F}_n(\mathbf{x}|\hat{\boldsymbol{\beta}}_n) > \gamma, & \text{decide Y=1 ,} \\ \hat{F}_n(\mathbf{x}|\hat{\boldsymbol{\beta}}_n) \leqslant \gamma, & \text{decide Y=0} \end{cases} \tag{4}$$

with an estimated boundary

$$\hat{l}_n(\mathbf{x}) = \{\mathbf{x} = (x_1, \cdots, x_p)^T : \hat{F}_n(\mathbf{x}|\hat{\boldsymbol{\beta}}_n) = \gamma\}, \tag{5}$$

where $\gamma = 0.5$ when there is no extra information, such as $P(Y = 1)$ available. (In general, the cutting point for a logistic classification function is 0.5. However, when there is prior information about the event, such as a prevalence rate in an epidemiology study, the cutting point will usually be adjusted accordingly. This will be discussed later.) Therefore, the active learning problem under this set-up becomes how to recruit a set of training subjects efficiently, such that the final classification function $\hat{F}_n$ will have good prediction power when a learning process is stopped.

3. Subject Selection

Intuitively, in order to have an efficient learning process, we should learn about the most uncertain subjects first. This may improve a classifier in the best way. Thus, when using a probabilistic learning model in an active learning framework, the most commonly used query for getting new data is the uncertainty sampling, where an active learner will query the label information of an instance with a class membership that is least certain. For a binary classification problem, this simply means querying the instance with a membership probability that is closest to 0.5. Thus, in a binary classification case, the uncertainty is usually measured by

$$\mathbf{d}(\mathbf{x}) = \left| \hat{F}_n(\mathbf{x}|\hat{\boldsymbol{\beta}}_n) - \omega \right|, \tag{6}$$

where $\omega = 0.5$.
Let $\mathbf{U}$ be the unlabeled data set. Then, rank points in $\mathbf{U}$ in ascending order based on Equation (6), and an active learning procedure will choose the top ranked point as follows:

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{x} \in \mathbf{U}} \mathbf{d}(\mathbf{x}). \tag{7}$$

Accordingly, choose the one with an estimated probability closest to 0.5 as the next point to be labeled. In high-dimensional cases, there may be a lot of points that have the same or similar $\mathbf{d}(\mathbf{x})$. Therefore, we choose top k_n points as candidates first, where k_n, in our method is decided by a local D-efficiency method by using a locally optimal design. Using Equation (7) as the only criterion cannot provide good estimates of model parameters, and the method of optimal design can be a good supplement to this disadvantage. Thus, let $C = \{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{k_n}\}$ be the set of candidate points that are screened. We then access these candidates further with some optimal experimental design criterion.
The effect of uncertainty sampling becomes obvious when the difference between the sample sizes of two groups is large. This situation happens very often in the problems that aim to detect a set of rare subjects within a large data set or when the population sizes of two groups are uneven. When the true model is exactly linear and the variables for this model are completely known, these two methods are the same. However,

in practice, the form of the true model and the involved variables are usually unknown. For instance, in the example discussed in Deng et al. (2009), the two variables used in their model are selected from a large number of variables by experts, and in fact, the true model may involve other variables. When a model is an approximation with some leftover random errors, the candidate set defined by a Euclidean distance-based method will be very different from the one obtained by using an uncertainty measure. That is, when some perturbation exists, the contour lines can no longer be parallel. Thus, using a perpendicular distance to find a candidate set, as in Deng et al. (2009), cannot be the best choice. That is the reason why we use an uncertainty sampling scheme to define a candidate set first, and then we use a (Bayesian) D-optimal design method to screen out the best subject for parameter estimation. Moreover, when the degree of dimensionality becomes larger, the computation of the determinant of a Fisher information matrix is difficult; when the size of labeled data is small, the information matrix will be either singular or nearly singular, which provides less information for designs. Thus, we adopt a Bayesian D-optimal design, which will stabilize the beginning stages of a learning process.

4. The proposed learning algorithm

Let $n_0 \geq 0$ be labeled data points at the initial stage. Accordingly, the proposed algorithm consists of the following steps:

S1. Compute the MAP estimate $\hat{\boldsymbol{\beta}}_n$ — the posterior estimate of $\boldsymbol{\beta}$ with the currently available labeled data (When $n_0 = 0$, we will use the prior median instead);

S2. Rank the unlabeled data points in $\mathbf{U}$ based on Equation (6). If the estimated posterior probabilities for all points are equal to either 0 or 1, then stop iteration, and use currently estimated $\hat{F}$ as the final classifier; otherwise, go to S3;

S3. Create the candidate set C with the top k_n points based on the ranks in step S2, where k_n is determined based on the local D-efficiency;

S4. Select a new unlabeled point from the set C, according to the following criteria:

(i) If the design points up to the current stage form a nonsingular information matrix of $\boldsymbol{\beta}$, then choose the next point that maximizes ϕ_1 ; that is,

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in C} \phi_1(\mathbf{x}_1, \cdots, \mathbf{x}_n, \mathbf{x}). \tag{8}$$

(ii) If the information matrix is singular, then select the next point from C that maximizes ϕ_1 based on the cumulated n points, k_n-augmentation and the candidate point. That is,

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in C} \phi_1(\mathbf{x}_1, \cdots, \mathbf{x}_n, \tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{k_n}, \mathbf{x}). \tag{9}$$

We consider the case with $\dim(\mathbf{x}) = p \geq 2$, so a Dirichlet distribution is a reasonable prior for $\mathbf{w} = (w_1, \cdots, w_p)^T$. Hence, the following priors are used:

$$\begin{aligned} \mu &\sim \mathrm{N}(\mu_0, \sigma_\mu^2), \quad \sigma \sim \text{Exponential}(\sigma_0), \\ \mathbf{w} &\sim \mathrm{Dir}(\alpha), \text{ where } \alpha = (\alpha_1, \cdots, \alpha_p)^T. \end{aligned} \tag{10}$$

Assume that parameters μ, σ and $\mathbf{w}$ are mutually independent. Then, the posterior distribution of $\boldsymbol{\theta}$, based on the labeled data points $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$, is

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{Y}) \;\propto\; & \prod_{i=1}^{n} \left(\frac{e^{(z_i-\mu)/\sigma}}{1+e^{(z_i-\mu)/\sigma}} \right)^{Y_i} \left(\frac{1}{1+e^{(z_i-\mu)/\sigma}} \right)^{1-Y_i} \\ & \times e^{(\mu-\mu_0)^2/(-2\sigma_\mu^2)} e^{-\sigma/\sigma_0} \left(\prod_{j=1}^{p-1} w_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{p-1} w_j \right)^{\alpha_p - 1} \end{aligned} \tag{11}$$

where $z_i = w_1 x_{i1} + \cdots + w_{p-1} x_{i,p-1} + w_p x_{ip}$, $\sum_{j=1}^{p} w_j = 1$ and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$. Then, the MAP is

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{Y}). \tag{12}$$

5. Conclusions

Active learning selects its own training samples in a sequential manner and requires fewer labeled instances from domain experts, and it still achieves high classification performance. In this paper, we focus on a higher-dimensional case and propose a new subject selection scheme that combines a Bayesian D-optimal design and an uncertainty sampling method. A Bayesian D-optimal design method makes the active learning process more stable in high-dimensional cases, even when the information matrix is nearly singular and therefore, will be more suitable for modern analysis with large data sets. In addition, we also demonstrate that with an initial training set of a small amount of labeled subjects, an active learning process is more stable and efficient in both training time and the size of the labeled data. For uneven group sizes case, we suggest using separate parameters to control uncertainty sampling and adjust the cutting threshold for a better performance. From our numerical studies, we found that the uncertainty measure and the probability of an event might play different roles in an active learning process, especially when the sizes of two groups are uneven. We found that using an uncertainty measure at 0.5 and then adjusting the boundary according to the proportion of group sizes, as that in classical logistic regression models, produces better results in our studies.

These types of methods are suitable for problems with a large amount of unlabeled data available and have great potential for analyzing "big data" problems. From a practical viewpoint, including one new subject at a time is not practical. This is because of not only the computational efficiency but also the operational complexity. This is similar to the situation in clinical trials, in which sampling in a batch, as in a group sequential procedure, is usually preferred. Moreover, labeling an unclassified subject is time-consuming, and there are also some operational costs, such as experts' charge and so on. Hence, the ways of conducting an active learning process with a batch of updated subjects and constructing a classification rule with a satisfactory performance with a given budget constraint are important problems from both practical and theoretical viewpoints.

References

Deng, X.W., Joseph, V.R., Sudjianto, A., Wu, C.F.J., 2009. Active learning through sequential design, with applications to detection of money laundering. Journal of the American Statistical Association 104, 969 - 981.

Dror, H.A., Steinberg, D.M., 2008. Sequential experimental designs for generalized linear models. Journal of the American Statistical Association 103, 288 - 298.

Seettles, B., 2010. Active learning literature survey. Computer Sciendes Technical Report 1648, University of Wisconsin-Madison .