Active Testing of Binary Classification Model Using Level Set Estimation

Takuma Ochiai OCHIAI.TAKUMA.N3@S.MAIL.NAGOYA-U.AC.JP
School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Japan

Keiichiro Seno SENO.KEIICHIRO.K0@S.MAIL.NAGOYA-U.AC.JP
Department of Biostatistics, Nagoya University
65 Tsurumai-cho, Showa-ku, Nagoya, Japan

Kota Matsui MATSUI.K@MED.NAGOYA-U.AC.JP
Department of Biostatistics, Nagoya University
65 Tsurumai-cho, Showa-ku, Nagoya, Japan

Satoshi Hara SATOHARA@AR.SANKEN.OSAKA-U.AC.JP
SANKEN, Osaka University
Mihogaoka, Ibaraki, Osaka

Abstract

In this study, we propose a method for estimating the test loss in binary classification model with minimal labeling of the test data. The central idea of the proposed method is to reduce the problem of test loss estimation to the problem of level set estimation for the loss function. This reduction allows us to achieve sequential test loss estimation through iterative labeling using active learning methods for level set estimation. Through multiple dataset experiments, we confirmed that the proposed method is effective for evaluating binary classification models and allows for test loss estimation with fewer labeled samples compared to existing methods.

Keywords: Level Set Estimation, Active Testing

1. Introduction

Labeling cost is an essential problem for effective supervised learning. In supervised learning problems, there are two phases where labeled data is required: the learning phase and the testing phase. For the learning phase, various approaches have been proposed to mitigate the labeling cost, such as active learning (Settles, 2009; Ren et al., 2021) which reduces labeling costs by actively selecting data for labeling, as well as self-supervised learning (Liu et al., 2021), and weakly supervised learning (Sugiyama et al., 2022). On the other hand, the problem of the labeling cost for the testing phase is not fully explored, except for a few seminal studies on active model evaluation (Sawade et al., 2010; Kossen et al., 2021).

The purpose of the testing phase is to estimate the test loss of the model trained in the learning phase. Active model evaluation (Sawade et al., 2010; Kossen et al., 2021) reduces the labeling cost in the testing phase by actively selecting the test data to be labeled. An important question here is how to select the effective test data as the labeling target. Sawade et al. (2010) proposed to sample the test data to be labeled from a probability distribution so that the asymptotic variance of the estimated test

loss is minimized. Kossen et al. (2021) proposed using an unbiased estimator (Farquhar et al., 2021) of the test loss and sampling the test data from a probability distribution so that the variance of the estimated loss is minimized. Although these sampling-based strategies are designed to minimize the variance of the estimate test loss, they sometimes show inefficient behavior in actual problems because they assign non-zero selection probabilities even to ineffective data, i.e., some of the labeling costs can be wasted.

In this study, we consider the problem of estimating the misclassification rate in binary classification problems under the limited labeling budget. To avoid ineffective labeling, we propose to use level set estimation (LSE) (Gotovos, 2013). Level set estimation partitions the set of input points to two subsets, namely the superlevel set and the sublevel set. The superlevel set and the sublevel set contains input points with their corresponding function values above or below of a prescribed threshold, respectively. Level set estimation attains this goal by using an appropriately designed acquisition function to select a fraction of input points and by observing their function values. In this study, we show that the estimation of the binary misclassification rate can be reduced to the problem of level set estimation. By this reduction, we can use the existing acquisition function for level set estimation to the problem of estimating the binary misclassification rate with a small number of labeling, enabling us to avoid ineffective labeling.

2. Level Set Estimation

In this section, we review the level set estimation which constitutes the basis of the proposed method.

Suppose that a set of d-dimensional points $\mathcal{X}_N = \{\boldsymbol{x}_i\}_{i=1}^N$ is given. For an unknown black-box function f and a prescribed threshold θ, the goal of level set estimation is to partition $\mathcal{X}_N$ to the superlevel set $\mathcal{X}_{\mathrm{up}}$ and the sublevel set $\mathcal{X}_{\mathrm{low}}$ given below.

$$\mathcal{X}_{\mathrm{up}} \equiv \{\boldsymbol{x} \in \mathcal{X}_N \mid f(\boldsymbol{x}) > \theta\}, \quad \mathcal{X}_{\mathrm{low}} \equiv \{\boldsymbol{x} \in \mathcal{X}_N \mid f(\boldsymbol{x}) \leq \theta\}. \tag{1}$$

If we can observe the function values $y_i = f(\boldsymbol{x}_i), i = 1, \ldots, N$ of all the points, this problem can be solved immediately. However, when the evaluation of the function value $y_i = f(\boldsymbol{x}_i)$ is expensive, we are allowed to observe the function values only on a fraction of the points in $\mathcal{X}_N$ (Hozumi et al., 2023). In order to estimate the level set accurately using a limited amount of function evaluations, Gotovos (2013) proposed an active learning algorithm by introducing a Gaussian process prior on the black-box function f. In each step of the algorithm, an input point is selected for function evaluation using an appropriately designed acquisition function, and the Gaussian process prior is updated in a Bayesian manner using the new observation.

2.1 Gaussian Process Models and Level Set Classification Rules

Suppose that a black box function f follows a Gaussian process model (Rasmussen and Williams, 2006) with the mean function $\mu(\boldsymbol{x})$ and kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$, $f \sim \mathcal{GP}(\mu, k)$. Let $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the set of data points whose function value $y_i = f(x_i)$ has been observed. Under these conditions, the predictive mean $\mu_n(\boldsymbol{x}^*)$ and predictive variance $\sigma_n^2(\boldsymbol{x}^*)$ of the unobserved point $\boldsymbol{x}^*$ given $\mathcal{D}_n$ can be expressed as $\mu_n(\boldsymbol{x}^*) = \boldsymbol{k}_*^\top \boldsymbol{K}^{-1} y$ and $\sigma_n^2(\boldsymbol{x}^*) = k_{**} - \boldsymbol{k}_*^\top \boldsymbol{K}^{-1} \boldsymbol{k}_*$, where $\boldsymbol{K} = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$, $\boldsymbol{k}_* = (k(\boldsymbol{x}^*, \boldsymbol{x}_1), k(\boldsymbol{x}^*, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}^*, \boldsymbol{x}_n))^\top$, and $k_{**} = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$.

The most simple rule for classifying the superleve and sublevel set is to apply the threshold θ to the predictive mean $\mu_n(\boldsymbol{x}_i)$ as follows.

$$\hat{\mathcal{X}}_{\text{up}} \equiv \{\boldsymbol{x} \in \mathcal{X}_N \mid \mu_n(\boldsymbol{x}) > \theta\}, \quad \hat{\mathcal{X}}_{\text{low}} \equiv \{\boldsymbol{x} \in \mathcal{X}_N \mid \mu_n(\boldsymbol{x}) \leq \theta\}. \tag{2}$$

2.2 Acquisition Function

In level set estimation, an input point is selected for function evaluation using an acquisition function. The popular acquisition functions are the uncertainty-based functions (Bryan et al., 2005; Gotovos, 2013) and the improvement-based functions (Zanette et al., 2019). In this study, we focus on the uncertainty-based Straddle function (Bryan et al., 2005) defined below,

$$\alpha_{\text{Straddle}}(\boldsymbol{x}) = \beta^{1/2}\sigma_n(\boldsymbol{x}) - |\mu_n(\boldsymbol{x}) - \theta|, \tag{3}$$

where the parameter $\beta > 0$ balances the two terms. In each step of the algorithm, a point $\boldsymbol{x}$ with the largest $\alpha_{\text{Straddle}}(\boldsymbol{x})$ is selected for function evaluation. Here, $\alpha_{\text{Straddle}}(\boldsymbol{x})$ selects a point whose function value $f(\boldsymbol{x})$ is expected to be close to the threshold θ (the second term of (3)) and the uncertainty of the estimated f is large (the first term of (3)).

In this study, we use the following ε-greedy version of the Straddle function as the acquisition function.

$$\alpha = \begin{cases} \alpha_{\text{Straddle}} & \text{w.p. } \varepsilon, \\ \sigma_n^2 & \text{w.p. } 1 - \varepsilon. \end{cases} \tag{4}$$

In this ε-greedy Straddle function, the predictive variance $\sigma_n^2(\boldsymbol{x})$ is selected as the acquisition function with probability $1 - \varepsilon$. In such a situation, a point $\boldsymbol{x}$ that will minimizes the uncertainty of the estimated f_{BB} is selected for function evaluation, which will help improving the estimation accuracy of the level set.

In summary, the active learning algorithm for level set estimation in this study can be expressed as the repitition of the following three steps: (i) select the acquisition function α following (4), (ii) select $x = \arg\max_{x \in \mathcal{X}_N} \alpha(x)$, observe its function value $y = f(x)$, and add (x, y) to $\mathcal{D}_n$, (iii) update the Gaussian process prior using $\mathcal{D}_n$.

3. Proposed Method

In this section, we show that the estimation of the binary misclassification rate can be reduced to the problem of level set estimation. By this reduction, we can use the existing acquisition function for level set estimation to the problem of estimating the binary misclassification rate with a small number of labeling, enabling us to avoid ineffective labeling.

3.1 Problem Formulation

We first formalize the estimation problem of the misclassification rate for the binary classification problem $y \in \{0, 1\}$. We consider a binary classification model h that outputs the probability of the class 1 so that $h(\boldsymbol{x}) = \Pr[\boldsymbol{x} \text{ is class } 1] \in [0, 1]$. Suppose that the unlabeled test data $D_{\text{test}} = \{\boldsymbol{x}_i\}_{i=1}^N$ is given. If we know the true label y_i for each $\boldsymbol{x}_i$, we can compute the 0-1 loss as $\mathcal{L}_{01}(h(\boldsymbol{x}), y) = \mathbb{I}[y \neq \mathbb{I}[h(\boldsymbol{x}) > \frac{1}{2}]]$, where $\mathbb{I}$ is the indicator function. The misclassification rate is then given by $\hat{R} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{01}(h(\boldsymbol{x}_i), y_i)$.

Problem 1 (Active Misclassifiation Rate Estimation) *Estimate the binary misclassification rate $\hat{R}$ by labeling only on the subset of D_{test} so that we can reduce the labeling cost.*

3.2 Active Misclassification Rate Estimation Using Level Set Estimation

We now show that the problem of estimating the the binary misclassification rate $\hat{R}$ can be reduced to the problem of level set estimation. For this purpose, we focus on the relationship between the cross-entropy loss and the 0-1 loss. We first recall that the cross-entropy loss is given as $\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y) = -y \log h(\boldsymbol{x}) - (1-y)\log(1-h(\boldsymbol{x}))$. Suppose that when the true class is $y = 0$, we have, $\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), 0) = -\log(1-h(\boldsymbol{x}))$. If $h(x) \leq 0.5$ and the predicted class is 0, $\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), 0) = -\log(1-h(\boldsymbol{x})) \leq \log 2$ holds. If $h(x) > 0.5$ and the predicted class is 1, $\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), 0) = -\log(1-h(\boldsymbol{x})) > \log 2$ holds. By the similar argument for the case when the true class is 1, we can see that the following relationship holds.

$$\begin{cases} \mathcal{L}_{01}(h(\boldsymbol{x}), y) = 0 \Leftrightarrow \mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y) \leq \log 2, \\ \mathcal{L}_{01}(h(\boldsymbol{x}), y) = 1 \Leftrightarrow \mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y) > \log 2. \end{cases}$$

By using the relationship above, we can rewrite the binary misclassification rate as

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}_i), y_i) > \log 2\right]. \tag{5}$$

Here, we define the superlevel set and the sublevel set as follows,

$$\mathcal{X}_{\text{up}} \equiv \{\boldsymbol{x} \in D_{\text{test}} \mid \mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y) > \log 2\}, \quad \mathcal{X}_{\text{low}} \equiv \{\boldsymbol{x} \in D_{\text{test}} \mid \mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y) \leq \log 2\}. \tag{6}$$

The problem of estimating the the binary misclassification rate $\hat{R}$ can then be reduced to the problem of estimating the size of the superlevel set $\mathcal{X}_{\text{up}}$ because $\hat{R} = \frac{1}{N}|\mathcal{X}_{\text{up}}|$. Thus, we can estimate $\hat{R}$ by solving the level set estimation problem with $\mathcal{L}_{\text{CE}}(h(\boldsymbol{x}), y)$ being a black-box function (because y is unknown) and the threshold $\theta = \log 2$.

4. Experiment

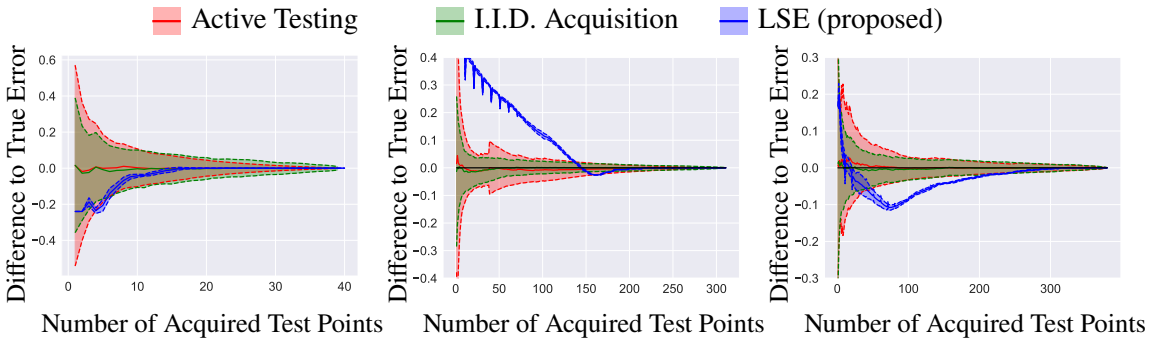
In this section, we examine the efficacy of the proposed method for active misclassification rate estimation.

4.1 Experimental Setup

Datasets We used three binary classification datasets, which are artificial data, Breast Cancer Wisconsin (Original) (Wolberg, 1992), and Diabetes Schulz et al. (2006). The latter two datasets are obtained from the UCI Machine Learning Repository and the LIBSVM data sets.

We created the artificial dataset as follows. The input x is one-dimensional and it distributes uniformly in the interval $[0, \pi/2]$. If $\sin x \geq 0.5$, its corresponding outcome $y = 1$, and $y = 0$ otherwise. We used 10 pairs of (x, y) for training a classifier h, and another 40 inputs as the unlabeled test data D_{test}.

Breast Cancer and Diabetes are real-world datasets with nine-dimensional input and 683 data points, and eight-dimensional inputs and 768 data points, respectively. In real-world problems,



(a) Artificial Data (b) Breast Cancer (c) Diabetes

Figure 1: Result: The solid line represents the average result of each method, and the colored shadow represents the standard deviation of the results.

it is typically the case that the training and testing data follow different distributions, i.e., there is a distribution shift. To simulate such a realistic scenario, we intentionally added distribution shifts for these two datasets. For Breast Cancer, we selected the data points with the features `Clump_Thickness` $\leq$ median for the training data, and the remaining data as the testing data. As the result, 372 and 311 data points are selected for the training and testing data, respectively. For Diabetes, we selected the data points with the features `Glucose` $\geq$ median for the training data, and the remaining data as the testing data. As the result, 388 and 380 data points are selected for the training and testing data, respectively.

Binary Classification Model We adopted linear logistic regression as the binary classification model h to be evaluated. The misclassification rate in the test data were 0.24, 0.09, and 0.16 for artificial data, Breast Cancer, and Diabetes, respectively.

Level Set Estimation Setup For level set estimation, we used the logarithm of the cross-entropy loss as the black-box function $f = \log \mathcal{L}_{\mathrm{CE}}$. This is because, while the cross-entropy loss takes a value close to zero for most of correctly classified data, it takes a large value for incorrect data. If we use the cross-entropy loss as it is as the black-box function, it tends to be a less smooth function with large bumps near the classification boundary. Such a less smooth functions are difficult to be fit by the Gaussian process. The use of logarithm makes the black-box function more smoother, enabling effective fitting by the Gaussian process.

Baseline Methods We compared the proposed method with the following two methods, *I.I.D. Acquisition* and *Active Testing* (Kossen et al., 2021). I.I.D. Acquisition selects the next point to be labeled uniformly at random from the unlabeled test points. The misclassification rate is estimated as the simple average of the 0-1 loss computed on labeled test points. Active Testing selects the next point to be labeled with probability proportional to its estimated 0-1 loss. The misclassification rate is estimated using the unbiased estimator of Farquhar et al. (2021).

4.2 Result

For each dataset, we ran each method 100 times with different random seeds. Figure 1 shows the difference of the true missclassification rate and the estimated one over different numbers of labeled

test points. We can find three important observations from the figure. First, I.I.D Acquisition and Active Testing provides unbiased estimate of the true misclassification rate, while the proposed method provide biased estimate particularly when the number of acquired points is small. Second, although the proposed method is biased, it almost converges to the true misclassification rate after a certain number of labels. Third, while I.I.D Acquisition and Active Testing tends to incur large variance, the variance of the proposed method is considerably small. From these observations, we can conclude that the proposed method can provide a better estimate of the misclassification with small variance a certain number of labels, compared to the baseline methods.

Table 1 shows the average number of labels (± std.) required for estimating the misclassification rate up to 0.1% tolerance. It is evident from the table that the proposed method required the smallest number of labels on average for accurate estimation. In particular, on artificial data and Breast Cancer, the proposed method required around three times less number of labels, showing the significant improvement over the baseline methods. On Diabetes, although the improvement is marginal, the proposed method attained the smallest number on average.

Table 1: The average number of labels (± std.) required for estimating the misclassification rate up to 0.1% tolerance.

	Active Testing	I.I.D. Acquisition	LSE (proposed)
Artificial	37.4 ± 2.8	39.0 ± 0.0	14.3 ± 1.3
Breast Cancer	280.8 ± 30.0	307.4 ± 0.9	214.5 ± 0.5
Diabetes	375.4 ± 6.1	377.3 ± 0.5	307.5 ± 5.6

5. Conclution

In this study, we proposed a method for estimating the binary misclassification rate with a small amount of labeling. In the proposed method, we reduced the problem to the level set estimation problem by utilizing the relationship between the cross entropy loss and 0-1 loss. By this reduction, we can apply the existing acquisition function for level set estimation to the problem of estimating the binary misclassification rate with a small number of labeling, enabling us to avoid ineffective labeling.

The experiment results confirm the effectiveness of the proposed method. The results show that, although the proposed method provides a biased estimate of the misclassification rate, the estimate almost converges to the true misclassification after a certain number of labels, which is typically far smaller than the existing methods.

Acknowledgments

KM was supported by JSPS KAKENHI Grant Number 23H03456, 20K19871, and 19H04071. SH was supported by JSPS KAKENHI Grant Number 20K19860, 23H03456, and JST, PRESTO Grant Number JPMJPR20C8.

References

Brent Bryan, Robert C Nichol, Christopher R Genovese, Jeff Schneider, Christopher J Miller, and Larry Wasserman. Active learning for identifying function threshold boundaries. *Advances in neural information processing systems*, 18, 2005.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.

Alkis Gotovos. Active learning for level set estimation. Master's thesis, Eidgenössische Technische Hochschule Zürich, Department of Computer Science, 2013.

Shota Hozumi, Kentaro Kutsukake, Kota Matsui, Syunya Kusakawa, Toru Ujihara, and Ichiro Takeuchi. Adaptive defective area identification in material surface using active transfer learning-based level set estimation. *arXiv preprint arXiv:2304.01404*, 2023.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. Springer, 2006.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 951–958, 2010.

Leslie O Schulz, Peter H Bennett, Eric Ravussin, Judith R Kidd, Kenneth K Kidd, Julian Esparza, and Mauro E Valencia. Effects of traditional and western environments on prevalence of type 2 diabetes in pima indians in mexico and the us. *Diabetes care*, 29(8):1866–1871, 2006.

Burr Settles. Active learning literature survey. *Technical Report.*, 2009.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine learning from weak supervision: An empirical risk minimization approach*. MIT Press, 2022.

WIlliam Wolberg. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5HP4Z.

Andrea Zanette, Junzi Zhang, and Mykel J Kochenderfer. Robust super-level set estimation using gaussian processes. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18*, pages 276–291. Springer, 2019.