Statistica Sinica Preprint No: SS-2024-0012	
Title	Optimal Model Averaging for Imbalanced Classification
Manuscript ID	SS-2024-0012
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0012
Complete List of Authors	Ze Chen, Jun Liao, Wangli Xu and Yuhong Yang
Corresponding Authors	Yuhong Yang
E-mails	yangx374@umn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Optimal Model Averaging for Imbalanced Classification

Ze Chen[1], Jun Liao[2], Wangli Xu[2] and Yuhong Yang[3]

[1]*Shandong University,* [2]*Renmin University of China and* [3]*Tsinghua University*

Abstract: Imbalanced data with a high-dimensional input has been widely encountered in many areas of applications. In this situation, it usually becomes essential to reduce redundant variables via model selection to improve the classification performance. However, with a large number of variables, model selection uncertainty is typically very high. To deal with this problem, we present a feasible model averaging procedure based on a cost-sensitive support vector machine (CSSVM) coupled with a cost-sensitive data-driven weight choice criterion for imbalanced classification. Theoretical justifications are provided in two distinct scenarios. When the data exhibits a weak imbalance, we derive a relatively fast uniform convergence rate of the CSSVM solution. In contrast, when the data possesses a strong imbalance, the convergence rate becomes much slower. In both scenarios, an asymptotic optimality of the proposed model averaging approach in the sense of minimizing the out-of-sample hinge loss is established. Moreover, to reduce the computational burden imposed by a large number of candidate models for model averaging, we develop the CSSVM with an L_1-norm penalty to prepare candidate models. Numerical analysis shows the superiority of the proposed model averaging procedure over existing imbalanced classification methods.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Key words and phrases: Asymptotic optimality, Imbalanced data, Model averaging, Uniform convergence rate.

1. Introduction

1.1 Imbalanced Classification Problems

In binary classification, imbalanced data, sometimes called rare event data, occurs when the number of instances of a certain class (the minority class) is significantly smaller than the number of instances of the opposite class (the majority class). The imbalanced classification problems (ICPs) were identified by Yang and Wu (2006) as one of ten challenging problems in data mining research. ICPs have shown up in many real-world applications, such as medical science (Rahman and Davis, 2013), telecommunications (Babu and Ananthanarayanan, 2018), and bioinformatics (Bugnon et al., 2019).

Previous imbalanced learning procedures for solving ICPs can be mainly categorized into data-driven sampling methods and algorithm level methods (Mathew et al., 2017). The data-driven sampling approaches try to balance the class distribution by an under-sampling (of the majority class) technique (see, e.g., Drummond et al. (2003); Liu et al. (2008); Arefeen et al. (2022)) or an over-sampling (of the minority class) technique (see, e.g., Chawla et al. (2002); Douzas and Bacao (2017); Koziarski et al. (2019)) prior to train-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

ing a specific classifier. However, under-sampling based procedures may throw away important information from the data, and over-sampling based methods may produce too many repeated instances that makes the trained classifier behave differently in prediction. Furthermore, theoretical analysis of under- and over-sampling based procedures for parameter estimation are still rare (Wang, 2020).

Cost-sensitive learning is a popular method among algorithm level methods, which assigns different misclassification costs to different classes in the imbalanced data, and it has been applied to different classification modelling systems (see, e.g., Zhou and Liu (2005) for neural networks, Zhang et al. (2018) for deep belief networks, and Zhang (2020) for nearest neighbor). In terms of handling ICPs, the cost-sensitive learning with SVM is popular. However, the existing SVM based cost-sensitive learning approaches (Veropoulos et al., 1999; Yang et al., 2009) may perform poorly in the presence of a large number of redundant variables (Peng et al., 2016).

In cases where the input dimension is high, the classification problem typically presents greater difficulty as a result of the curse of dimensionality (Yang, 2006). In such situations, variable selection is shown to be beneficial for improving the overall classification performance (Liu et al., 2018). Grobelnik (1999) proposed an approach to subset selection based on

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Naive Bayes. Yin et al. (2013) developed two feature selection approaches for high-dimensional imbalanced data based on the class decomposition and Hellinger distance. A backward elimination variable selection approach was investigated by Maldonado et al. (2014). Additionally, Liu et al. (2018) presented an effective feature selection method by optimizing F-measures (Puthiya Parambath et al., 2014).

1.2 Model Averaging

As is now increasingly well-known, one potential drawback of model selection is that it only chooses a single model in the selection process, which ignores possibly high uncertainty (e.g., Draper (1995); Yuan and Yang (2005)). Alternatively, model averaging is being more and more adopted, as it not only significantly reduces the model selection uncertainty but also has the potential to intrinsically improve over the best single model. Here, model selection uncertainty typically involves the instability of the selection outcome in the sense that small changes in the data can lead to significant differences in the chosen models, resulting in unnecessarily high variability in the final estimation or prediction (Yuan and Yang, 2005; Zhang et al., 2013; Nan and Yang, 2014). Model averaging employs continuous weights to combine the estimators or predictions from different models. In contrast,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

model selection can be seen as a special case of model averaging, where the weights are restricted to 0 or 1. As a result, model averaging can reduce the loss of useful information and generally produces more stable estimates or predictions. Intuitively, when two models are very close in terms of selection criteria, using appropriate model weights is often much better than making exaggerated 0-1 decisions (in a winner-takes-all sense) (Yang, 2001).

Many studies have been conducted on model averaging, which can be typically categorized into Bayesian and frequentist approaches. For Bayesian model averaging, Hoeting et al. (1999) provided a detailed review. Regarding frequentist model averaging, there is a lot of work under different model frameworks in the literature, see, e.g., Hansen (2007); Zhang et al. (2020); Chen et al. (2022) for linear regression models, and Yang (2001); Fang et al. (2022); He et al. (2023); Chen et al. (2023) for semiparametric or nonparametric models.

There are also some studies on model averaging that address binary classification problems, such as the optimal model averaging procedure on logistic regression (see, e.g., Zhang et al. (2016c), Ando and Li (2017) and Zhang and Liu (2023)). Recently, Zou et al. (2023) also proposed the support vector classification model averaging method by cross-validation that is asymptotically optimal in the sense of achieving the smallest hinge loss.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

However, these existing methods only focus on balanced data, which are inadequate for handling imbalanced data due to the potential for these methods to exhibit significant bias in predicting minority classes within imbalanced data.

For the high-dimensional imbalanced data, it is not clear how to construct a reliable model averaging method. In this paper, we aim to develop an optimal model averaging method based on CSSVM to handle the ICPs, for which there are two main challenges in contrast to the case in the balanced data setting. First, we need to devise a proper weight choice criterion. In the existing literature on optimal model averaging for balanced data, the misclassification costs of two classes are viewed as the same or similar, which may cause a serious bias in the prediction of the minority class for imbalanced data. Consequently, it becomes crucial to establish a cost-sensitive weight choice criterion and assign misclassification costs of the two classes properly. Second, we hope that the resultant model averaging estimator is asymptotically optimal in the sense of achieving the lowest possible out-of-sample hinge loss. However, the derivation of the asymptotic optimality is much more involved relative to the situation of balanced data since the degree of data imbalance affects the establishment of the asymptotic optimality.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

The main contributions of our paper lie in three aspects. First, we derive a uniform convergence rate of the solution of CSSVM, allowing the model space to diverge. Also, we find that the uniform convergence rate depends on the imbalanced degree of the data. Second, for the ICPs, we propose a CSSVM based model averaging procedure with a cost-sensitive weight choice criterion that applies to imbalanced data. The corresponding asymptotic optimality is established in the sense of minimizing the out-of-sample hinge loss. Note that the closely related optimal model averaging work in Zou et al. (2023) can be regarded as a special case of our work by setting the misclassification costs of the two classes to be equal. We also derive the convergence rate in terms of asymptotic optimality and find that the imbalanced degree of the data make an impact on the convergence rate of optimality. Third, to reduce the computational burden resulting from a large number of candidate models, we present the L_1-norm CSSVM to prepare candidate models.

The remainder of this article is organized as follows. Section 2 provides a detailed introduction to CSSVM and our model averaging procedure with a cost-sensitive data-driven weight choice criterion, and the corresponding theoretical properties are presented in Section 3. Section 4 discusses some implementation details of the proposed model averaging method. Numerical

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

studies on the proposed method and a real data analysis are presented in Section 5. Concluding remarks are offered in Section 6. Additional numerical results and detailed proofs of the theorems are provided in the Supplementary Materials.

2. Model Setup and Model Averaging

For binary classification, we consider a set of training data $\mathcal{S}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, independently drawn from the distribution of $(\boldsymbol{X}, Y)$, where $X = (1, \tilde{\boldsymbol{X}}^{\mathrm{T}})^{\mathrm{T}} = (1, X_1, \ldots, X_p)^{\mathrm{T}}$, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^{\mathrm{T}} = (1, \tilde{\boldsymbol{x}}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{R}^{p+1}$ and $y_i \in \{1, -1\}$. The minority class instances are denoted $\{(\boldsymbol{x}_i, y_i) : i \in \mathcal{I}^+\}$ with $\mathcal{I}^+ \stackrel{\text{def}}{=} \{i : y_i = 1\}$ and the majority class instances are denoted $\{(\boldsymbol{x}_i, y_i) : i \in \mathcal{I}^-\}$ with $\mathcal{I}^- \stackrel{\text{def}}{=} \{i : y_i = -1\}$. Let $n_1 = |\mathcal{I}^+|$ and $n_2 = |\mathcal{I}^-|$ with $n_1 \leq n_2$, where $|\cdot|$ denotes the number of elements of a set. Here, n_1 and n_2 are random since $n_1 = \sum_{i=1}^n I(y_i = 1)$ and $n_2 = n - n_1$, where $I(\cdot)$ is the indicator function. In addition, to facilitate the study of imbalanced classification data, we consider a triangular array setting by allowing the distribution of Y given X to depend on n. That is,

$$P(Y = 1) = \pi_{1n}, \quad P(Y = -1) = \pi_{2n} = 1 - \pi_{1n},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

and π_{1n} is allowed to tend to zero or remain fixed as $n \to \infty$ in this paper. Denote $\bar{n}_1 \stackrel{\text{def}}{=} E(n_1) = n\pi_{1n}$ and $\bar{n}_2 \stackrel{\text{def}}{=} E(n_2) = n\pi_{2n}$. Our target is to find a classifier Φ that assigns a class y^{new} to be 1 or -1 for a new input $\boldsymbol{x}^{new} = (1, x_1^{new}, \ldots, x_p^{new})^{\mathrm{T}}$.

2.1 Cost-sensitive Support Vector Machine

In this paper, we consider two kinds of imbalanced degrees of the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, namely weak imbalance and strong imbalance. In the following, unless otherwise stated, all limiting processes discussed are as $n \to \infty$.

Definition 1. (Weak Imbalance and Strong Imbalance) The data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is weakly imbalanced with parameter $0 < \zeta < 1/2$ if

$$\frac{\pi_{1n}}{n^{-1/2+\zeta}\log n} \to \infty.$$

The data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is strongly imbalanced with parameter $1 < \tau \leq 2$ if

$$\frac{\pi_{1n}}{(n^{-1/2}\log n)^{\tau}} \to \infty \quad \text{and} \quad \frac{\pi_{1n}}{n^{-1/2}\log n} \leq C,$$

where C is a positive constant.

Clearly, the imbalanced degree of the data with strong imbalance is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

higher than that of the data with weak imbalance. To make this point more clear, we give some illustrative examples. Note that the definitions of weak and strong imbalances are given from an asymptotic perspective. When π_{1n} is fixed (i.e., $\pi_{1n} = c$ for a constant $0 < c < 0.5$), even if π_{1n} is very small (e.g., $\pi_{1n} = 0.00001$), the number of minority class instances remains on the same order as n, asymptotically speaking. Thus, the corresponding data is weakly imbalanced.

Also, if $\zeta = 1/4$ and $\pi_{1n} = n^{-1/4}(\log n)^2$, then the corresponding data is weakly imbalanced with parameter $\zeta = 1/4$. In contrast, if for example $\tau = 5/4$ and $\pi_{1n} = n^{-5/8}(\log n)^2$, then it is strongly imbalanced with parameter $\tau = 5/4$ and any $C > 0$. In practice, given that n is finite, it is challenging to determine whether the data exhibits strong or weak imbalance based on Definition 1. Therefore, we recommend trying both the treatment strategies for strong and weak imbalances on the data (see Section 4 for details).

Denote by $(1 - z)_+ = \max\{1 - z, 0\}$ the hinge loss for $z \in \mathcal{R}$, and $\|\cdot\|$ denotes the Euclidean norm. The standard SVM can be expressed as

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (1 - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \frac{\lambda_n}{2} \|\tilde{\boldsymbol{\beta}}\|^2 \right\}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}} = (\beta_0, \tilde{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$, and $\lambda_n > 0$ is the tuning parame-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

ter. It can be seen from (2.1) that SVM treats the misclassification costs of different classes equally. However, this can result in the SVM being biased towards the majority class when the data is weakly or strongly imbalanced and further lead to poor performance of SVM for the minority class. Thus, to address this issue, for $0 < \bar{C}_n \leq 1$, by assigning a smaller misclassification cost $\bar{C}_n$ to the majority class relative to the minority class, we suggest the CSSVM model that is expressed as the following regularization problem

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \psi_i (1 - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \frac{\lambda_n}{2} \|\tilde{\boldsymbol{\beta}}\|^2 \right\}, \tag{2.2}$$

where $\psi_i = 1$ if $y_i = 1$ and $\psi_i = \bar{C}_n$ if $y_i = -1$. We further estimate β by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \psi_i (1 - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \lambda_n/2 \|\tilde{\boldsymbol{\beta}}\|^2\}$, which can be efficiently solved by quadratic programming algorithms. Thus, for a new input $\boldsymbol{x}^{new}$, the resulting classifier is $\hat{\Phi}(\boldsymbol{x}^{new}) = \mathrm{sgn}((\boldsymbol{x}^{new})^{\mathrm{T}} \hat{\boldsymbol{\beta}})$, where $\mathrm{sgn}(\cdot)$ is the sign function.

Note that CSSVM is similar to the linear weighted SVM studied by Zhang et al. (2016b), but there is one major difference between the two. The misclassification costs specified by weights in the linear weighted SVM are fixed, implying that the probability of $Y = 1$ is fixed. This excludes the scenario of highly imbalanced data. Given that π_{1n} can be allowed to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

go to zero, it is more appropriate that the misclassification cost $\bar{C}_n$ of the majority class should also be permitted to tend to zero.

It is readily seen that the unconstrained regularized empirical loss minimization problem (2.2) is equivalent to the following optimization problem proposed by Veropoulos et al. (1999) (the detailed derivations are presented in Part A of the Supplementary Materials),

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\zeta}} \quad & \left(\frac{1}{2}\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + C^{+}\sum_{i\in\mathcal{I}^{+}}\zeta_i + C^{-}\sum_{i\in\mathcal{I}^{-}}\zeta_i\right) \\ \text{subject to} \quad & y_i\boldsymbol{x}_i^{T}\boldsymbol{\beta} \geq 1-\zeta_i, \\ & \zeta_i \geq 0, i = 1,\ldots,n, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_n)^{\mathrm{T}}$ are slack variables, and $C^{+} > 0$ and $C^{-} > 0$ are regularization parameters for positive and negative classes, respectively. Thus, CSSVM can be regarded as a variant of (2.3). In subsequent theoretical studies, we will mainly focus on the unconstrained problem (2.2).

2.2 Model Averaging for CSSVM

In this section, we develop a model averaging method for CSSVM that involves two steps. First, we derive a classifier using CSSVM for each candidate model, where different candidate models include different covariates.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Second, we give a final classifier by a weighted combination of all classifiers produced in the first step, in which a proper choice of weights is crucial.

2.2.1 Candidate Models

Suppose that we have K candidate models, and let $\mathcal{I}_k$ $(k = 1, \ldots, K)$ be the set that contains indices of the covariates under the kth candidate model. Note that $\mathcal{I}_k \subseteq \{1, \ldots, p\}$, and we assume that each candidate model contains the intercept term in this paper. Let $d_k = |\mathcal{I}_k|$ be the number of covariates in the kth candidate model.

Under the kth candidate model, we can obtain a classifier $\hat{\Phi}_k(\boldsymbol{x}_{(k)}^{new}) = \operatorname{sgn}((\boldsymbol{x}_{(k)}^{new})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(k)})$ by $\hat{\boldsymbol{\beta}}_{(k)} = \arg\min_{\boldsymbol{\beta}_{(k)}} \{\sum_{i=1}^{n} \psi_i(1-y_i\boldsymbol{x}_{(k),i}^{\mathrm{T}}\boldsymbol{\beta}_{(k)})_+ + \frac{\lambda_n}{2}\|\tilde{\boldsymbol{\beta}}_{(k)}\|^2\}$, where $\boldsymbol{x}_{(k)}^{new}$ is a (d_k+1)-dimensional vector including the constant one and x_j^{new} $(j \in \mathcal{I}_k)$, $\boldsymbol{x}_{(k),i}$ a (d_k+1)-dimensional vector including the constant one and x_{ij} $(j \in \mathcal{I}_k)$, and $\boldsymbol{\beta}_{(k)} = (\beta_{(k)0}, \beta_{(k)1}, \ldots, \beta_{(k)d_k})^{\mathrm{T}} = (\beta_{(k)0}, \tilde{\boldsymbol{\beta}}_{(k)}^{\mathrm{T}})^{\mathrm{T}}$.

It is natural to consider a set that is the collection of all-subset models as the candidate models, and then the number of potential candidate models is $K = 2^p - 1$. However, this is computationally infeasible when p is large. To handle this problem, one common way is to use nested candidate models (see, e.g., Hansen (2014); Nan and Yang (2014); Zhang et al. (2016c); Zhang et al. (2020); Chen et al. (2023)). Specifically, to order the covariates, we

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

apply the solution path of the following L_1-norm CSSVM,

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \psi_i (1 - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})_+ + \tilde{\lambda}_n \|\tilde{\boldsymbol{\beta}}\|_1 \right\}, \tag{2.4}$$

where $\|\cdot\|_1$ denotes the L_1 norm and $\tilde{\lambda}_n$ is a tuning parameter. The proposal of L_1-norm CSSVM is inspired by L_1-norm SVM (Zhu et al., 2004; Peng et al., 2016) which replaces the L_2-norm penalty with the L_1-norm penalty. The optimization problem of (2.4) can be transformed as a linear programming problem, which will be discussed in Section 4. By decreasing the penalty $\tilde{\lambda}_n$, we can further obtain the solution path of L_1-norm CSSVM, and then sort the covariates based on the order in which they enter the solution path. It is generally considered that the covariates that enter the path earlier are more important than those that enter later. In the end, we construct K nested candidate models, where the kth candidate model contains the first k covariates according to the order in the solution path.

Following a referee's suggestion, to order the covariates, we can also introduce an additional L_1-norm penalty to the objective function without altering its original objective function (2.2). To save space, the details are provided in Part B of the Supplementary Materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

2.2.2 Model Averaging with Cost-sensitive Weight Choice Criterion

Let w_k be the weight of the kth candidate model, and the weight vector $\boldsymbol{w} = (w_1, \ldots, w_K)^{\mathrm{T}}$ belongs to the set $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$. Then, the weighted classifier can be written as $\hat{\Phi}_w(\boldsymbol{x}^{new}) = \mathrm{sgn}(\sum_{k=1}^K w_k (\boldsymbol{x}_{(k)}^{new})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)})$. Let Ψ be a binary variable. For $0 < \bar{C}_n \leq 1$, $\Psi = 1$ if $Y = 1$ and $\Psi = \bar{C}_n$ if $Y = -1$. Ideally, we hope to find an optimal weight vector w that minimizes the out-of-sample hinge loss

$$L_n(\boldsymbol{w}) = E\left\{ \Psi \left(1 - Y \sum_{k=1}^K w_k \boldsymbol{X}_{(k)}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)} \right)_+ \middle| \mathcal{S}_n \right\}, \tag{2.5}$$

where $\boldsymbol{X}_{(k)}$ is a $(d_k + 1)$-dimensional vector including the constant one and X_j $(j \in \mathcal{I}_k)$.

However, we cannot directly minimize (2.5) over $\boldsymbol{w} \in \mathcal{W}$ since the real distribution of $(\boldsymbol{X}, Y)$ is unknown. To choose weights for imbalanced data, we propose to apply the cost-sensitive leave-one-out cross-validation criterion which is defined as

$$SCV_n(\boldsymbol{w}) = \sum_{i=1}^n \psi_i \left(1 - y_i \sum_{k=1}^K w_k \boldsymbol{x}_{(k),i}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)}^{-i} \right)_+, \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

where $\hat{\boldsymbol{\beta}}_{(k)}^{-i}$ is the estimator of $\boldsymbol{\beta}_{(k)}$ with the ith observation deleted under the kth candidate model, that is,

$$\hat{\boldsymbol{\beta}}_{(k)}^{-i} = \arg\min_{\boldsymbol{\beta}_{(k)}} \left\{ \sum_{j \neq i} \psi_j (1 - y_j \boldsymbol{x}_{(k),j}^{\mathrm{T}} \boldsymbol{\beta}_{(k)})_+ + \frac{\lambda_n}{2} \|\tilde{\boldsymbol{\beta}}_{(k)}\|^2 \right\}. \tag{2.7}$$

Thus, the resultant weight estimator is

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} SCV_n(\boldsymbol{w}), \tag{2.8}$$

where $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_K)^{\mathrm{T}}$. Finally, the resultant weighted classifier is given by $\hat{\Phi}_{\hat{\boldsymbol{w}}}(\boldsymbol{x}^{new}) = \mathrm{sgn}(\sum_{k=1}^{K} \hat{w}_k (\boldsymbol{x}_{(k)}^{new})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)})$.

The cost-sensitive weight choice criterion $SCV_n(\boldsymbol{w})$ in (2.6) takes into account the cost of wrong predictions for the minority class. Specifically, by assigning a large weight to the loss function of the minority class, the criterion $SCV_n(\boldsymbol{w})$ is sensitive to the cost of the minority class so as to reduce the case of the minority class being misclassified. However, the cost information associated with two classes is not considered by the following classical leave-one-out cross-validation criterion

$$CV_n(\boldsymbol{w}) = \sum_{i=1}^{n} \left(1 - y_i \sum_{k=1}^{K} w_k \boldsymbol{x}_{(k),i}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)}^{-i} \right)_+, \tag{2.9}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

where $\hat{\boldsymbol{\beta}}_{(k)}^{-i}$ is calculated by (2.7). The numerical results in Section 5 illustrate that our model averaging method with $SCV_n(\boldsymbol{w})$ is indeed superior to that with $CV_n(\boldsymbol{w})$ especially for the data with a high degree of imbalance or a high dimension.

In Part C of the Supplementary Materials, we provide discussions on why the leave-one-out cross-validation criterion is used to select weights instead of K-fold cross-validation in the context of imbalanced data.

3. Theoretical Properties

In this section, we show the asymptotic properties of $\hat{\boldsymbol{\beta}}_{(k)}$ and $\hat{\boldsymbol{\beta}}_{(k)}^{-i}$ in both weakly imbalanced and strongly imbalanced cases. Also, the asymptotic optimality of the proposed model averaging method is presented in the sense of minimizing the out-of-sample loss.

We introduce a representation factor $\nu \in [1, 2]$ here. If $\nu = 1$, then the data is weakly imbalanced, and if $1 < \nu \leq 2$, then the data is strongly imbalanced with the parameter $\tau = \nu$. Denote $L_k(\boldsymbol{\beta}_{(k)}) = E\{\Psi(1 - Y\boldsymbol{X}_{(k)}^{\mathrm{T}}\boldsymbol{\beta}_{(k)})_+\}$, where the expectation is calculated with respect to the joint distribution of $(\Psi, Y, \boldsymbol{X}_{(k)}^{\mathrm{T}})^{\mathrm{T}}$. Following Koo et al. (2008) and Zhang et al. (2016b), for the kth candidate model, we define the pseudo-true parameter $\boldsymbol{\beta}_{(k)}^{*} = \arg\min_{\boldsymbol{\beta}_{(k)}} E\{\Psi(1 - Y\boldsymbol{X}_{(k)}^{\mathrm{T}}\boldsymbol{\beta}_{(k)})_+\} = \arg\min_{\boldsymbol{\beta}_{(k)}} L_k(\boldsymbol{\beta}_{(k)})$. Note

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

that if the kth candidate model is identical to the true data generating process, according to Koo et al. (2008) and Zhang et al. (2016b), the pseudo-true parameter is the true parameter. Further, we assume that the pseudo-true parameter $\boldsymbol{\beta}^*_{(k)}$ exists and is unique for each candidate model. Let $S_k(\boldsymbol{\beta}_{(k)}) = -E\{I(1 - Y\boldsymbol{X}^{\mathrm{T}}_{(k)}\boldsymbol{\beta}_{(k)} \geq 0)\Psi Y\boldsymbol{X}_{(k)}\}$ and $H_k(\boldsymbol{\beta}_{(k)}) = E\{\delta(1 - Y\boldsymbol{X}^{\mathrm{T}}_{(k)}\boldsymbol{\beta}_{(k)})\Psi\boldsymbol{X}_{(k)}\boldsymbol{X}^{\mathrm{T}}_{(k)})\}$, where $\delta(\cdot)$ denotes the Dirac delta function and $I(\cdot)$ is the indicator function. Under some appropriate conditions, $S_k(\boldsymbol{\beta}_{(k)})$ and $H_k(\boldsymbol{\beta}_{(k)})$ can be considered as the gradient and Hessian matrix of $L_k(\boldsymbol{\beta}_{(k)})$ (Koo et al., 2008). Let $d_{max} = \max_{1\leq k\leq K} d_k + 1$ where d_k is the number of covariates in the kth candidate model. Denote by f_k^+ and f_k^- the densities of $\tilde{\boldsymbol{X}}_{(k)}$ conditioning on $Y = +1$ and $Y = -1$, respectively, where $\tilde{\boldsymbol{X}}_{(k)}$ is a d_k-dimensional vector including X_j $(j \in \mathcal{I}_k)$.

We write $a_n = \Theta(b_n)$ if there exist $c_1, c_2 > 0$ such that $c_1 b_n \leq a_n \leq c_2 b_n$. Denote by C a generic positive constant. To investigate the asymptotic behavior of the proposed model averaging method, we need the following conditions for our theorems.

Condition 1. f_k^+ and f_k^- are continuous and have common support in $\mathcal{R}^{d_k}$.

Condition 2. $\max_{1\leq j\leq p} |X_j| \leq C < \infty$ almost surely, and $\|\boldsymbol{\beta}^*_{(k)}\| \leq C d_k^{1/2}$ for $k = 1, \ldots, K$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Condition 3. The densities of $\boldsymbol{x}_{(k),i}^{\mathrm{T}}\boldsymbol{\beta}_{(k)}^{*}$ conditioning on $Y=+1$ and $Y=-1$ are uniformly bounded away from zero and have an uniform upper bound C at the neighborhood of $\boldsymbol{x}_{(k),i}^{\mathrm{T}}\boldsymbol{\beta}_{(k)}^{*}=1$ and $\boldsymbol{x}_{(k),i}^{\mathrm{T}}\boldsymbol{\beta}_{(k)}^{*}=-1$, respectively.

Condition 4. Uniformly for $k\in\{1,...,K\}$, $\lambda_{min}(H_k(\boldsymbol{\beta}_{(k)}^{*}))\geq C>0$, where $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of a matrix.

Condition 5. For the tuning parameters λ_n in (2.2), $\lambda_n/(n\pi_{1n}\log p)^{1/2}\leq C<\infty$.

Condition 6. (i) If $\nu=1$, then $\log p=O(n^{\gamma})$, $d_{max}=O\big(n^{(2\zeta-3\gamma)/4}\big)$ and $\log K=O(d_{max}\log n)$ for $0<\gamma<2\zeta/3$, where ζ $(0<\zeta<1/2)$ is defined in Definition 1. (ii) If $\nu>1$, $d_{max}\log p=O((n^{1/2}/\log n)^{\eta})$ and $\log K=O(d_{max}\log n)$ for $0\leq\eta\leq 2/\tau-1$, where τ $(1<\tau\leq 2)$ is defined in Definition 1.

Condition 7. For the misclassification cost $\bar{C}_n$, $\bar{C}_n=\Theta\big((\pi_{1n}/\pi_{2n})^{3/2-1/\nu}\big)$ for $\nu\in[1,2]$.

Condition 8. $\xi_n>0$ and $\xi_n^{-1}n^{-1/2}M_n^{1/2}\xrightarrow{P}0$, where $\xi_n=\inf_{\boldsymbol{w}\in\mathcal{W}}L_n(\boldsymbol{w})$ and $M_n=\pi_{1n}d_{max}^2\log p$.

Most of the above conditions are commonly seen for model selection and model averaging (see, e.g., Zhang et al. (2016a) and Zhang et al. (2016b)). Condition 1 ensures that $S_k(\boldsymbol{\beta}_{(k)})$ and $H_k(\boldsymbol{\beta}_{(k)})$ are well-defined

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

(see Koo et al. (2008) for details). The first part of Condition 2 states that $\max_{1\leq j\leq p}|X_j|$ is bounded almost surely. The second part of Condition 2 requires that the L_2 norm of the pseudo-true parameter $\boldsymbol{\beta}^*_{(k)}$ grows at a rate no faster than d_k. Condition 3 claims that there is enough information around the non-differentiable point of the hinge loss function. Condition 4 requires a lower bound on the smallest eigenvalue of the Hessian matrix at the pseudo-true parameter $\boldsymbol{\beta}^*_{(k)}$. Condition 5 provides the order of the tuning parameters λ_n. Condition 6 puts a restriction on the number of candidate models, the number of covariates and the maximum dimension of candidate models. Condition 7 specifies the misclassification cost in (2.2). It shows that to reduce the case of the minority class being misclassified, we need to assign a larger misclassification cost to the minority class when the data exhibits the higher degree of imbalance. The $\xi_n > 0$ in Condition 8 can be easily satisfied when the data is not completely linearly separable. Condition 8 gives the order of ξ_n that is larger than $n^{-1/2}M_n^{1/2}$ which depends on the degree of imbalance. We give an explanation of rationality about Condition 8 in Part D of the Supplementary Materials.

Remark 1. From Condition 5, it is seen that λ_n can be taken to be zero or very close to zero. According to Condition 6, it is deduced that $d_{max} < n$. Therefore, the parameters are estimable for the candidate models without

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

the penalty term by the following optimization problem

$$\hat{\boldsymbol{\beta}}_{(k)} = \arg\min_{\boldsymbol{\beta}_{(k)}} \frac{1}{n}\sum_{i=1}^{n} \psi_i (1 - y_i \boldsymbol{x}_{(k),i}^{\mathrm{T}} \boldsymbol{\beta}_{(k)})_+. \tag{3.1}$$

Note that, with a finite sample, the optimization problem (3.1) may have multiple minimizers since the objective function is piecewise linear. In this case, $\hat{\boldsymbol{\beta}}_{(k)}$ can be chosen to be any minimizer. Our theoretical results still hold, that is, $\hat{\boldsymbol{\beta}}_{(k)}$ converges to the pseudo-true parameter $\boldsymbol{\beta}^*_{(k)}$ as $n \to \infty$. Under the assumption of the uniqueness of $\boldsymbol{\beta}^*_{(k)}$, the uniqueness of the minimizer $\hat{\boldsymbol{\beta}}_{(k)}$ of the optimization problem (3.1) is not essential on the basis of our theoretical techniques (see Zhang et al. (2016b) for more discussions). Thus, our approach allows λ_n to be zero or very close to zero.

Theorem 1. *Under Conditions 1–7, we have*

$$\max_{1\le k\le K} \|\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}^*_{(k)}\| = O_p\left(\left(\frac{d_{max}\log p}{n\pi_{1n}^{2-2/\nu}}\right)^{\frac{1}{2}}\right). \tag{3.2}$$

Further, we also have

$$\max_{1\le i\le n}\max_{1\le k\le K} \|\hat{\boldsymbol{\beta}}^{-i}_{(k)} - \boldsymbol{\beta}^*_{(k)}\| = O_p\left(\left(\frac{d_{max}\log p}{n\pi_{1n}^{2-2/\nu}}\right)^{\frac{1}{2}}\right). \tag{3.3}$$

Theorem 1 shows the uniform convergence rates of $\hat{\boldsymbol{\beta}}_{(k)}$ and $\hat{\boldsymbol{\beta}}^{-i}_{(k)}$ in the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

presence of weakly and strongly imbalanced data.

Remark 2. In the case where the data is balanced, which is a special case of weak imbalance in our work, Zou et al. (2023) proved the uniform convergence rate of the estimated parameters. In their proof, the misclassification costs for both classes are set to be equal. The misclassification cost $\bar{C}_n$, which depends on the degree of data imbalance π_{1n}, is key for deriving the uniform convergence rate. The first challenge in our proof is to determine the order of $\bar{C}_n$ such that the uniform convergence rate obtained under weak imbalance is consistent with that derived for balanced data. Moreover, when the data is highly imbalanced (i.e., strong imbalance), a natural question arises: whether the estimated parameters still converge and at what rate? The second challenge is to derive the convergence of the estimated parameters and determine the convergence rate under strong imbalance, which also involves determining the appropriate order of $\bar{C}_n$.

Based on the uniform convergence rates, we can further prove the asymptotic optimality of the proposed model averaging method. To achieve this, we first present the following lemma.

Lemma 1. *Under Conditions 1–7, we have*

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|L_n(\boldsymbol{w})-E\left\{\Psi\left(1-Y\sum_{k=1}^{K}w_k\boldsymbol{X}_{(k)}^{\mathrm{T}}\boldsymbol{\beta}_{(k)}^{*}\right)_{+}\right\}\right|=O_p\left(\left(\frac{\pi_{1n}d_{max}^2\log p}{n}\right)^{\frac{1}{2}}\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} |SCV_n(\boldsymbol{w}) - nL_n(\boldsymbol{w})| = O_p\big(\big(n\pi_{1n}d_{max}^2 \log p\big)^{1/2}\big).$$

Lemma 1 reflects the degree of approximation between $L_n(\boldsymbol{w})$ and $E\{\Psi(1-Y\sum_{k=1}^{K} w_k \boldsymbol{X}_{(k)}^{\mathrm{T}}\boldsymbol{\beta}_{(k)}^{*})_+\}$, as well as between $SCV_n(\boldsymbol{w})$ and $nL_n(\boldsymbol{w})$, which will be applied in the following asymptotic optimality theorem.

Theorem 2. *Under Conditions 1–8, we have*

$$\frac{L_n(\hat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})} = 1 + O_p\left(\frac{n^{-1/2}M_n^{1/2}}{\xi_n}\right) \xrightarrow{P} 1,$$

where $\xi_n = \inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})$ *and* $M_n = \pi_{1n}d_{max}^2 \log p$.

Theorem 2 shows that the proposed model averaging method is asymptotically optimal in the sense of achieving the lowest possible out-of-sample hinge loss. Moreover, we attain the convergence rate of the asymptotic optimality, which involves π_{1n} that varies depending on the degree of data imbalance. From the convergence rate, it becomes apparent that the asymptotic optimality is easier to achieve if ξ_n is larger.

Note that there is a screening step in the proposed model averaging procedure based on L_1 norm CSSVM, but Theorem 2 gives the asymptotic optimality without the screening step. We have also established the asymptotic optimality property after the screening step. To save space, the details

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

are included in Part E of the Supplementary Materials.

4. Implementation

In this section, we discuss some implementation details of the proposed model averaging method, such as the choices of $\bar{C}_n$ and λ_n, the solution of L_1-norm CSSVM, as well as the calculation of the optimal weights $\hat{w}$ in (2.8).

First, for the choice of $\bar{C}_n$, based on Condition 7 required for our optimality theorem, we recommend taking $\bar{C}_n = (n_1/n_2)^{1/2}$ for the weak imbalance. In the context of strong imbalance, considering the case of the highest degree of data imbalance (i.e., $\tau = 2$), we suggest employing $\bar{C}_n = n_1/n_2$. Whenever there is difficulty deciding between weak and strong imbalance, it is advisable to consider both $\bar{C}_n = n_1/n_2$ and $\bar{C}_n = \sqrt{n_1/n_2}$ as experimental options in practice. As for the choice of λ_n, we propose a data-driven strategy. Specifically, for a given value of λ_n, let $\hat{\boldsymbol{w}}(\lambda_n) = (\hat{w}_1(\lambda_n), \ldots, \hat{w}_K(\lambda_n))^{\mathrm{T}}$ be the estimated weights by (2.8), and

$$SCV_n(\hat{\boldsymbol{w}}(\lambda_n)) = \sum_{i=1}^{n} \psi_i \left(1 - y_i \sum_{k=1}^{K} \hat{w}_k(\lambda_n) \boldsymbol{x}_{(k),i}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)}^{-i}\right)_+ .$$

Then, we can conduct a gird of values for λ_n: $\Lambda = \{\lambda_{n,1}, \ldots, \lambda_{n,m}\}$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

select λ_n by $\hat{\lambda}_n = \arg\min_{\lambda_n \in \Lambda} SCV_n(\hat{\boldsymbol{w}}(\lambda_n))$.

Second, to derive the linear programming formulation of the L_1-norm CSSVM, we first introduce the slack variables $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^{\mathrm{T}}$ and $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_p)^{\mathrm{T}}$. With the slack variables, the optimization problem of (2.4) can be transformed as the following linear programming problem (Zhu et al., 2004):

$$\min_{\boldsymbol{\zeta}, \boldsymbol{\vartheta}, \boldsymbol{\beta}} \left(\sum_{i \in \mathcal{I}^+} \zeta_i + \bar{C}_n \sum_{i \in \mathcal{I}^-} \zeta_i + \frac{\tilde{\lambda}_n}{2} \sum_{j=1}^{p} \vartheta_j \right)$$

$$\text{subject to } \ \zeta_i \geq 0, \ \zeta_i \geq 1 - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \quad i = 1, \ldots, n,$$

$$\vartheta_j \geq \beta_j, \ \vartheta_j \geq -\beta_j, \quad j = 1, \ldots, p.$$

The above standard linear programming problem can be efficiently solved by several R packages, such as `lpSolve` and `linprog`.

Third, similar to the optimization problem of (2.4), the optimization problem of (2.8) for weights can also be solved by the constrained linear program as follows.

$$\min_{\boldsymbol{\zeta}, \boldsymbol{w}} \left(\sum_{i \in \mathcal{I}^+} \zeta_i + \bar{C}_n \sum_{i \in \mathcal{I}^-} \zeta_i \right)$$

$$\text{subject to } \ \zeta_i \geq 0, \ i = 1, \ldots, n, \quad w_k \geq 0, \ k = 1, \ldots, K,$$

$$\sum_{k=1}^{K} w_k = 1, \quad \zeta_i \geq 1 - y_i \sum_{k=1}^{K} w_k \boldsymbol{x}_{(k),i}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(k)}^{-i}, \ i = 1, \ldots, n.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

By solving the optimization problem, we derive the optimal weights $\hat{w}$ in (2.8).

5. Numerical Analysis

In this section, we take into account different simulation settings to evaluate the performance of the proposed model averaging method. Additionally, we compare it with several existing methods which are listed below:

- MASCV$_1$ and MASCV$_2$: the proposed model averaging method with the cost-sensitive weight choice criterion $SCV_n(\boldsymbol{w})$ in (2.6), and the values of $\bar{C}_n$ are taken as n_1/n_2 and $(n_1/n_2)^{1/2}$, respectively.

- MACV$_1$ and MACV$_2$: the proposed model averaging method with the weight choice criterion $CV_n(\boldsymbol{w})$ in (2.9) and different values of $\bar{C}_n$.

- CSSVM$_1$ and CSSVM$_2$: the standard linear CSSVM in (2.2), and different values of $\bar{C}_n$.

- SMOTE: apply the synthetic minority over-sampling technique in Chawla et al. (2002) to balance the data, and then use standard linear SVM.

- OVER: create possibly balanced samples by random over-sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

minority examples (Lunardon et al., 2014), and then use standard linear SVM.

- UNDER: create possibly balanced samples by random under-sampling majority examples (Lunardon et al., 2014), and then use standard linear SVM.

- OVUN: create possibly balanced samples by combination of over-sampling and under-sampling (Lunardon et al., 2014), and then use standard linear SVM.

SMOTE can be implemented by the R package `smotefamily`, and OVER, UNDER and OVUN can be obtained from the R package `ROSE`. For the imbalanced data, accuracy is not a good measure. For instance, when a dataset has 10 positive examples and 190 negative examples, the accuracy of a method that identifies all instances as negative classes will be 95%, but it will be completely ineffective as a classifier for discovering the positive cases. In order to efficiently evaluate the performance of the different methods on the imbalanced data, as suggested by Kubat et al. (1997), we consider the G-measure that is the geometric mean of sensitivity and specificity. Another important evaluation criterion is the area under the ROC curve (AUC) (Bradley, 1997). A higher G-measure (or AUC) value

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

indicates better classification performance in an overall sense.

5.1 Simulation Study

In this subsection, we consider the following data generation process: $\tilde{\boldsymbol{x}}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i \in \mathcal{I}^+$ and $\tilde{\boldsymbol{x}}_i \sim N(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i \in \mathcal{I}^-$, where $\boldsymbol{\mu} = (0.6, \ldots, 0.6, 0, \ldots, 0)^{\mathrm{T}}$ with first q elements being 0.6, and $\boldsymbol{\Sigma} = (\sigma^{|i-j|})_{p\times p}$ with $\sigma = 0.3$. Further, we take $n_1 = n\pi_{1n}$, where $0 < \pi_{1n} < 1/2$, and then $n_2 = n(1-\pi_{1n})$. Clearly, the π_{1n} value reflects the imbalanced degree of the generated data. According to the regularization problem (2.2) and its equivalent optimization problem (2.3), it can be inferred that $\bar{C}_n = C^-/C^+$ and $\lambda_n = \bar{C}_n/C^-$. Hence, for the choice of the tuning parameter λ_n, we perform a gird of values: $\Lambda = C_{tun} \times \bar{C}_n$, where $C_{tun} = \{0.1, 0.5, 1, 1.5, 2\}$ in this subsection.

In Examples 1–5, we consider a sample size of $n = 5000$ for the testing data. All results of Example 1–5 are based on 100 replications. In each example, we generate K candidate models according to the L_1-norm CSSVM, where the tuning parameters $\tilde{\lambda}_n$ are evenly spaced within the range $[0.001, 10]$, with $\max\{50, p/5\}$ values in total. Furthermore, we take $K = p$ if $p \leq [n^{1/2}]$ and $K = [n^{1/2}]$ otherwise, where $[n^{1/2}]$ denotes the smallest integer not less than $n^{1/2}$.

Example 1. We evaluate the performance of all methods for ICPs with a

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

small number of covariates. We set $p = 10$ and $q = 5$. Four different imbalanced degrees for the values of π_{1n} are considered: $\pi_{1n} = \{0.05, 0.1, 0.2, 0.4\}$.

The results of Example 1 are given in Table 1, and the maximal G-measure and AUC values are highlighted in bold for each scenario. From Table 1, we can see that MASCV_1 always yields the largest G-measure and AUC values in nearly all cases. An exception is the scenario $\pi_{1n} = 0.1$, in which MASCV_1 performs slightly worse than MACV_1 by an almost negligible margin. MACV_1 has the same performance as MASCV_1 in the majority of cases concerning G-measure and AUC, but for the case with the high degree of imbalance and the small sample size, i.e., $n = 200, \pi_{1n} = 0.05$, MACV_1 exhibits inferior performance relative to MASCV_1. We can also find in Table 1 that MASCV_2 performs almost as well as MASCV_1 and is the second best method when the data has a low degree of imbalance ($\pi_{1n} = 0.4$). Note that MASCV_1 has larger AUC values than OVER, UNDER and OVUN by a small margin in all cases. However, in terms of G-measure values, MASCV_1 usually does so by a large margin.

To save space, the performance of various methods on AUC and G-measure in Examples 2–5 are included in Part F of the Supplementary Materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Table 1: The averaged G-measure and AUC values of various methods for Example 1.

$p = 10$	$n = 200$, $\pi_{1n} = 0.05$		$n = 200$, $\pi_{1n} = 0.1$		$n = 200$, $\pi_{1n} = 0.2$		$n = 200$, $\pi_{1n} = 0.4$	
	G-measure	AUC	G-measure	AUC	G-measure	AUC	G-measure	AUC
MASCV_1	**0.785**	**0.791**	0.826	0.827	**0.835**	**0.836**	**0.846**	**0.846**
	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
MACV_1	0.775	0.785	**0.828**	**0.829**	**0.835**	**0.836**	**0.846**	**0.846**
	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
CSSVM_1	0.759	0.771	0.819	0.823	0.831	0.832	0.842	0.842
	(0.005)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
MASCV_2	0.707	0.743	0.785	0.799	0.820	0.825	0.844	0.845
	(0.005)	(0.003)	(0.004)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
MACV_2	0.643	0.704	0.728	0.760	0.816	0.822	0.844	0.844
	(0.010)	(0.006)	(0.008)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)
CSSVM_2	0.727	0.752	0.787	0.799	0.819	0.823	0.840	0.842
	(0.005)	(0.003)	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
SMOTE	0.760	0.773	0.818	0.821	0.829	0.830	0.837	0.838
	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
OVER	0.184	0.744	0.167	0.819	0.167	0.827	0.160	0.837
	(0.002)	(0.012)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
UNDER	0.227	0.751	0.180	0.811	0.173	0.822	0.159	0.839
	(0.005)	(0.008)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
OVUN	0.194	0.740	0.172	0.817	0.173	0.823	0.166	0.830
	(0.003)	(0.012)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
$p = 10$	$n = 300$, $\pi_{1n} = 0.05$		$n = 300$, $\pi_{1n} = 0.1$		$n = 300$, $\pi_{1n} = 0.2$		$n = 300$, $\pi_{1n} = 0.4$	
	G-measure	AUC	G-measure	AUC	G-measure	AUC	G-measure	AUC
MASCV_1	**0.809**	**0.812**	0.842	**0.844**	**0.840**	**0.840**	**0.849**	**0.849**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
MACV_1	0.808	0.811	**0.843**	**0.844**	**0.840**	**0.840**	**0.849**	**0.849**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
CSSVM_1	0.793	0.798	0.836	0.837	0.839	0.839	0.846	0.846
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
MASCV_2	0.731	0.761	0.803	0.813	0.836	0.830	0.847	0.847
	(0.004)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
MACV_2	0.673	0.722	0.745	0.772	0.821	0.826	0.847	0.847
	(0.008)	(0.005)	(0.005)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
CSSVM_2	0.747	0.770	0.803	0.814	0.824	0.828	0.845	0.845
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
SMOTE	0.794	0.800	0.837	0.838	0.837	0.837	0.840	0.841
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
OVER	0.180	0.793	0.160	0.835	0.160	0.837	0.156	0.842
	(0.002)	(0.005)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
UNDER	0.210	0.782	0.168	0.826	0.166	0.832	0.154	0.844
	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
OVUN	0.189	0.787	0.163	0.832	0.166	0.830	0.160	0.838
	(0.003)	(0.005)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)

Note: The standard errors are given in parentheses, and the maximal G-measure and AUC values are highlighted in bold for each scenario.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

5.2 Real Data Analysis

The proposed procedure is applied to a diabetes health indicators dataset and a human activity dataset that are both available from the UCI Irvine Machine Learning Repository (`https: //archive.ics.uci.edu`). For the sake of space,the detailed introduction and analysis about the human activity dataset are presented in Part G of the Supplementary Materials. Here, we focus on the diabetes health indicators dataset. This dataset contains 253680 observations with 35346 (13.9%) diabetes instances and 218334 (86.1%) no diabetes instances. There are 21 predictors including dummy variables blood pressure, high cholesterol, cholesterol check, smoker, stroke, heart disease or attack, physical activity, fruits, vegetables, heavy drinkers, health care coverage, doctor cost, sex, difficulty walking, and non-dummy variables BMI, health level, mental health, physical health, age, education and income. Here, in consideration of the computational cost, we use a randomly chosen subset of 20000 observations to evaluate the competing approaches, while preserving the same imbalance ratio (13.9%) as the original dataset.

To assess the performance of each method, we still use G-measure and AUC, and the parameters are set as in Section 5.1, such as $\bar{C}_n$ and the tuning parameters λ_n, and the number of candidate models K. The preparation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

of candidate models is based on the L_1-norm CSSVM. From the remaining 233680 observations, we randomly select $n_{\text{train}} = 200,300$ instances as the training data to fit a classifier based on each approach and then use the chosen subset of 20000 observations as the testing data to evaluate the performance of each method. Note that the imbalanced degree of the training data is also maintained as 13.9%. Then, by repeating the above steps 100 times, the averaged G-measure and AUC values can be obtained. The detailed results for different approaches are provided in Table 2.

It is observed from Table 2 that MASCV_1 always results in the best performance in terms of both G-measure and AUC with the MACV_1 coming in a close second in all cases. With the increase of n_{train}, the AUC and G-measure values of MASCV_1 not only gradually increase but also continue to be the largest compared to its competitors. Moreover, OVER, UNDER and OVUN have a quite poor performance in terms of the G-measure. To sum up, the proposed CSSVM based model averaging approach emerges as more preferred.

6. Conclusion

Imbalanced classification problems present a challenge in both theory and methodology. In this work, in the CSSVM framework, a model averag-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Table 2: The averaged G-measure and AUC values of various methods for the real data.

$n_{\text{train}} = 200$	MASCV_1	MACV_1	CSSVM_1	MASCV_2	MACV_2
AUC	**0.700**	0.691	0.671	0.588	0.510
	(0.002)	(0.002)	(0.006)	(0.005)	(0.003)
G-measure	**0.698**	0.684	0.667	0.427	0.058
	(0.002)	(0.003)	(0.002)	(0.017)	(0.018)
	CSSVM_2	SMOTE	OVER	UNDER	OVUN
AUC	0.635	0.670	0.659	0.671	0.650
	(0.002)	(0.003)	(0.007)	(0.002)	(0.002)
G-measure	0.589	0.663	0.317	0.322	0.325
	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
$n_{\text{train}} = 300$	MASCV_1	MACV_1	CSSVM_1	MASCV_2	MACV_2
AUC	**0.710**	0.704	0.695	0.607	0.509
	(0.002)	(0.002)	(0.001)	(0.005)	(0.003)
G-measure	**0.707**	0.701	0.693	0.487	0.049
	(0.002)	(0.002)	(0.001)	(0.018)	(0.014)
	CSSVM_2	SMOTE	OVER	UNDER	OVUN
AUC	0.648	0.691	0.693	0.689	0.678
	(0.002)	(0.001)	(0.001)	(0.001)	(0.006)
G-measure	0.604	0.688	0.301	0.306	0.307
	(0.004)	(0.002)	(0.001)	(0.002)	(0.001)

Note: The standard errors are given in parentheses, and the maximal G-measure and AUC values are highlighted in bold for each scenario.

ing technique with a cost-sensitive weight selection criterion is proposed. For a theoretical understanding, we introduce the notations of weak and strong imbalances, which play a key role in deriving asymptotic results. Theoretical results include the uniform convergence rates of the CSSVM solutions in diverging model spaces. In particular, for the data with weak (or strong) imbalance, a faster (or slower) uniform convergence rate is re-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

vealed. In addition, the proposed model averaging procedure is proved to asymptotically achieves the smallest possible out-of-sample hinge loss, with the corresponding convergence rate varying based on the degree of data imbalance. In the situation where the number of candidate models is large, a model screening strategy reliant on the L_1-norm CSSVM is introduced. The simulation results strongly favor the proposed model averaging procedure in comparison with existing methods. Extending our model averaging method to kernel-based SVMs is a promising direction for future research

Supplementary Material

The proofs of all theoretical results, the justifications of conditions, and additional numerical results are provided in the Supplementary Material document.

Acknowledgments

We truely appreciate the constructive suggestions made by three reviewers. We also thank Co-Editor Yi-Hau Chen and the AE for advices on revising our work. The work of Ze Chen is supported by the Postdoctoral Fellowship Program of CPSF (No. GZC20231478) and the China Postdoctoral Science Foundation (No. 2024M761782). The work of Jun

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Liao is partially supported by the National Natural Science Foundation of China (No. 12001534 and 11971323). The work of Wangli Xu is supported by Beijing Natural Science Foundation (No. Z200001), Public Health & Disease Control and Prevention, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China (No. 2023PDPC), and the MOE Project of Key Research Institute of Humanities and Social Sciences (No. 22JJD910001).

REFERENCE

Ando, T. and Li, K.C. (2017). "A weight-relaxed model averaging approach for high-dimensional generalized linear models." *The Annals of Statistics*, **45**, 2654–2679.

Arefeen, M.A., Nimi, S.T., and Rahman, M.S. (2022). "Neural network-based undersampling techniques." *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **52**, 1111–1120.

Babu, S. and Ananthanarayanan, N. (2018). "Enhanced prediction model for customer churn in telecommunication using EMOTE." In "International Conference on Intelligent Computing and Applications," pages 465–475.

Bradley, A.P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition*, **30**, 1145–1159.

Bugnon, L.A., Yones, C., Milone, D.H., and Stegmayer, G. (2019). "Deep neural architectures

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

for highly imbalanced data in bioinformatics." *IEEE Transactions on Neural Networks and Learning Systems*, **31**, 2857–2867.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, **16**, 321–357.

Chen, Z., Liao, J., Xu, W., and Yang, Y. (2023). "Multifold cross-validation model averaging for generalized additive partial linear models." *Journal of Computational and Graphical Statistics*, **32**, 1649–1659.

Chen, Z., Zhang, J., Xu, W., and Yang, Y. (2022). "Consistency of BIC model averaging." *Statistica Sinica*, **32**, 1–6.

Douzas, G. and Bacao, F. (2017). "Self-organizing map oversampling (SOMO) for imbalanced data set learning." *Expert Systems with Applications*, **82**, 40–52.

Draper, D. (1995). "Assessment and propagation of model uncertainty." *Journal of the Royal Statistical Society, Series B*, **57**, 45–97.

Drummond, C., Holte, R.C., et al. (2003). "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling." In "Workshop on Learning from Imbalanced Datasets II," volume 11, pages 1–8.

Fang, F., Li, J., and Xia, X. (2022). "Semiparametric model averaging prediction for dichotomous response." *Journal of Econometrics*, **229**, 219–245.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Grobelnik, M. (1999). "Feature selection for unbalanced class distribution and Naive Bayes." In "International Conference on Machine Learning," pages 258–267.

Hansen, B.E. (2007). "Least squares model averaging." *Econometrica*, **75**, 1175–1189.

Hansen, B.E. (2014). "Model averaging, asymptotic risk, and regressor groups." *Quantitative Economics*, **5**, 495–530.

He, B., Ma, S., Zhang, X., and Zhu, L. (2023). "Rank-based greedy model averaging for high-dimensional survival data." *Journal of the American Statistical Association*, **118**, 2658–2670.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). "Bayesian model averaging: A tutorial." *Statistical Science*, **14**, 382–401.

Koo, J.Y., Lee, Y., Kim, Y., and Park, C. (2008). "A bahadur representation of the linear support vector machine." *Journal of Machine Learning Research*, **9**, 1343–1368.

Koziarski, M., Krawczyk, B., and Woźniak, M. (2019). "Radial-based oversampling for noisy imbalanced data classification." *Neurocomputing*, **343**, 19–33.

Kubat, M., Matwin, S., et al. (1997). "Addressing the curse of imbalanced training sets: one-sided selection." In "International Conference on Machine Learning," pages 179–186.

Liu, M., Xu, C., Luo, Y., Xu, C., Wen, Y., and Tao, D. (2018). "Cost-sensitive feature selection by optimizing F-measures." *IEEE Transactions on Image Processing*, **27**, 1323–1335.

Liu, X.Y., Wu, J., and Zhou, Z.H. (2008). "Exploratory undersampling for class-imbalance

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

learning." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**, 539–550.

Lunardon, N., Menardi, G., and Torelli, N. (2014). "ROSE: a package for binary imbalanced learning." *R Journal*, **6**, 79–89.

Maldonado, S., Weber, R., and Famili, F. (2014). "Feature selection for high-dimensional class-imbalanced data sets using support vector machines." *Information Sciences*, **286**, 228–246.

Mathew, J., Pang, C.K., Luo, M., and Leong, W.H. (2017). "Classification of imbalanced data by oversampling in kernel space of support vector machines." *IEEE Transactions on Neural Networks and Learning Systems*, **29**, 4065–4076.

Nan, Y. and Yang, Y. (2014). "Variable selection diagnostics measures for high-dimensional regression." *Journal of Computational and Graphical Statistics*, **23**, 636–656.

Peng, B., Wang, L., and Wu, Y. (2016). "An error bound for L_1-norm support vector machine coefficients in ultra-high dimension." *Journal of Machine Learning Research*, **17**, 8279–8304.

Puthiya Parambath, S., Usunier, N., and Grandvalet, Y. (2014). "Optimizing F-measures by cost-sensitive classification." In "Advances in Neural Information Processing Systems," pages 2123–2131.

Rahman, M.M. and Davis, D.N. (2013). "Addressing the class imbalance problem in medical datasets." *International Journal of Machine Learning and Computing*, **3**, 224–228.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

Veropoulos, K., Campbell, C., Cristianini, N., et al. (1999). "Controlling the sensitivity of support vector machines." In "International Joint Conference on Artificial Intelligence," volume 55, pages 55–60.

Wang, H. (2020). "Logistic regression for massive data with rare events." In "International Conference on Machine Learning," volume 119, pages 9829–9836.

Yang, C.Y., Yang, J.S., and Wang, J.J. (2009). "Margin calibration in SVM class-imbalanced learning." *Neurocomputing*, **73**, 397–411.

Yang, Q. and Wu, X. (2006). "10 challenging problems in data mining research." *International Journal of Information Technology & Decision Making*, **5**, 597–604.

Yang, Y. (2001). "Adaptive regression by mixing." *Journal of the American Statistical Association*, **96**, 574–588.

Yang, Y. (2006). "Comparing learning methods for classification." *Statistica Sinica*, **16**, 635–657.

Yin, L., Ge, Y., Xiao, K., Wang, X., and Quan, X. (2013). "Feature selection for high-dimensional imbalanced data." *Neurocomputing*, **105**, 3–11.

Yuan, Z. and Yang, Y. (2005). "Combining linear regression models: When and how?" *Journal of the American Statistical Association*, **100**, 1202–1214.

Zhang, C., Tan, K.C., Li, H., and Hong, G.S. (2018). "A cost-sensitive deep belief network for imbalanced classification." *IEEE Transactions on Neural Networks and Learning Systems*,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

30, 109–122.

Zhang, S. (2020). "Cost-sensitive KNN classification." *Neurocomputing*, **391**, 234–242.

Zhang, X. and Liu, C.A. (2023). "Model averaging prediction by K-fold cross-validation." *Journal of Econometrics*, **235**, 280–301.

Zhang, X., Lu, Z., and Zou, G. (2013). "Adaptively combined forecasting for discrete response time series." *Journal of Econometrics*, **176**, 80–91.

Zhang, X., Wu, Y., Wang, L., and Li, R. (2016a). "A consistent information criterion for support vector machines in diverging model spaces." *Journal of Machine Learning Research*, **17**, 466–491.

Zhang, X., Wu, Y., Wang, L., and Li, R. (2016b). "Variable selection for support vector machines in moderately high dimensions." *Journal of the Royal Statistical Society, Series B*, **78**, 53–76.

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016c). "Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models." *Journal of the American Statistical Association*, **111**, 1175–1790.

Zhang, X., Zou, G., Liang, H., and Carroll, R.J. (2020). "Parsimonious model averaging with a diverging number of parameters." *Journal of the American Statistical Association*, **115**, 972–984.

Zhou, Z.H. and Liu, X.Y. (2005). "Training cost-sensitive neural networks with methods address-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0012

ing the class imbalance problem." *IEEE Transactions on Knowledge and Data Engineering*, **18**, 63–77.

Zhu, J., Rosset, S., Tibshirani, R., and Hastie, T. (2004). "1-norm support vector machines." In "Advances in Neural Information Processing Systems," volume 16, pages 49–56.

Zou, J., Yuan, C., Zhang, X., Zou, G., and Wan, A.T. (2023). "Model averaging for support vector classifier by cross-validation." *Statistics and Computing*, **33**, 117–139.

Ze Chen

Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, China.

E-mail: chze96@sdu.edu.cn

Jun Liao

Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China.

E-mail: junliao@ruc.edu.cn

Wangli Xu

Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China.

E-mail: wlxu@ruc.edu.cn

Yuhong Yang (corresponding author)

Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China.

E-mail: yyangsc@tsinghua.edu.cn