On the Generalization of Feature Incremental Learning

Chao Xu[1] , **Xijia Tang**[1] , **Lijun Zhang**[2] and **Chenping Hou**[1*]
[1]College of Science, National University of Defense Technology, Changsha, 410073, China.
[2]Nanjing University, Nanjing, China.
{xcnudt, TXJnudt, hcpnudt}@hotmail.com, {zljzju}@gmail.com

Abstract

In many real applications, the data attributes are incremental and the samples are stored with accumulated feature spaces gradually. Although there are several elegant approaches to tackling this problem, the theoretical analysis is still limited. There exist at least two challenges and fundamental questions. 1) How to derive the generalization bounds of these approaches? 2) Under what conditions do these approaches have a strong generalization guarantee? To solve these crucial but rarely studied problems, we provide a comprehensive theoretical analysis in this paper. We begin by summarizing and refining four strategies for addressing feature incremental data. Subsequently, we derive their generalization bounds, providing rigorous and quantitative insights. The theoretical findings highlight the key factors influencing the generalization abilities of different strategies. In tackling the above two fundamental problems, we also provide valuable guidance for exploring other learning challenges in dynamic environments. Finally, the comprehensive experimental and theoretical results mutually validate each other, underscoring the reliability of our conclusions.

1 Introduction

In recent years, with the vast utilization of Machine Learning (ML) methods in many different applications, Statistical Learning Theory [Weston, 2013; de Mello and Ponti, 2018], which reveals the laws of machine learning, has attracted more and more attention. Among the various research in Statistical Learning Theory, generalization bound [Xu and Zeevi, 2020] plays an important role since it is an index to measure the generalization ability of a model directly. Due to its importance, traditional generalization theories have achieved many solid theoretical results in supervised learning [Lei *et al.*, 2019; Morvant *et al.*, 2012; Antos *et al.*, 2002; Li *et al.*, 2022], semi-supervised learning [El-Yaniv and Pechyony, 2007; Liu and Chen, 2018; Das *et al.*, 2013; He *et al.*, 2021] and unsupervised learning [Li *et al.*, 2019; Li and Liu, 2021; Downey *et al.*, 2010].

Besides the above learning paradigms, we may face more complicated scenarios. In many dynamic environment applications, the data are usually accumulated over time and collected from open and dynamic environments. Thus, the data attributes (features) are incremental and the samples are stored with accumulated feature spaces gradually. For instance, when we deploy sensors in the ecosystem to collect data, in which the signal returned from each sensor corresponds to a feature (old feature). With advancements in observation techniques and sensor technology, new sensors are continuously integrated, generating additional signals (new features) and progressively expanding data feature spaces. This underscores the critical importance of developing learning systems that can effectively adapt to dynamic and evolving environments [Dietterich, 2017]. In this scenario, as shown in Figure 1, the data collection procedure is divided into two stages, i.e., previous and current stages. The corresponding feature spaces include the old feature space $\mathcal{X}_p$, $\mathbf{X}_i^{(1)} \in \mathcal{X}_p, i = 1, 2$ and the new feature space $\mathcal{X}_c$, $[\mathbf{X}_2^{(1)}, \mathbf{X}_2^{(2)}] \in \mathcal{X}_c$.

Under such a data background, feature increment learning [Yang *et al.*, 2022; Gu *et al.*, 2022; Sadreddin and Sadaoui, 2021] has attracted wide attention and inspired a lot of excellent works [Ye *et al.*, 2018; Hou and Zhou, 2018; Xu *et al.*, 2016; Hou *et al.*, 2019; Zhang *et al.*, 2020; Hou *et al.*, 2021]. While existing feature-incremental learning approaches have demonstrated satisfactory performance in various applications, a comprehensive generalization analysis remains absent. This gap arises due to two main challenges: 1) The fundamental i.i.d. assumption of traditional learning is violated in feature-incremental scenarios, rendering conventional generalization theories inapplicable. 2) The diverse strategies employed by different approaches, such as model reuse [Ye *et al.*, 2018] and data tailoring [Hou and Zhou, 2018], complicate unified analysis. Consequently, although these algorithms are intuitively reasonable, rigorous theoretical underpinnings is still limited. To gain a deep understanding of the workings of this complex machine-learning scenario, we focus on two fundamental questions. 1) How to derive the generalization bounds of these feature incremental learning approaches? 2) Under what conditions do these approaches have a strong generalization guarantee?

Aiming at these critical but rarely-studied fundamental problems, We begin by summarizing and refining four strategies, i.e., *feature tailoring, data adaption, model reuse, and data reconstruction*. Subsequently, we derive their generalization

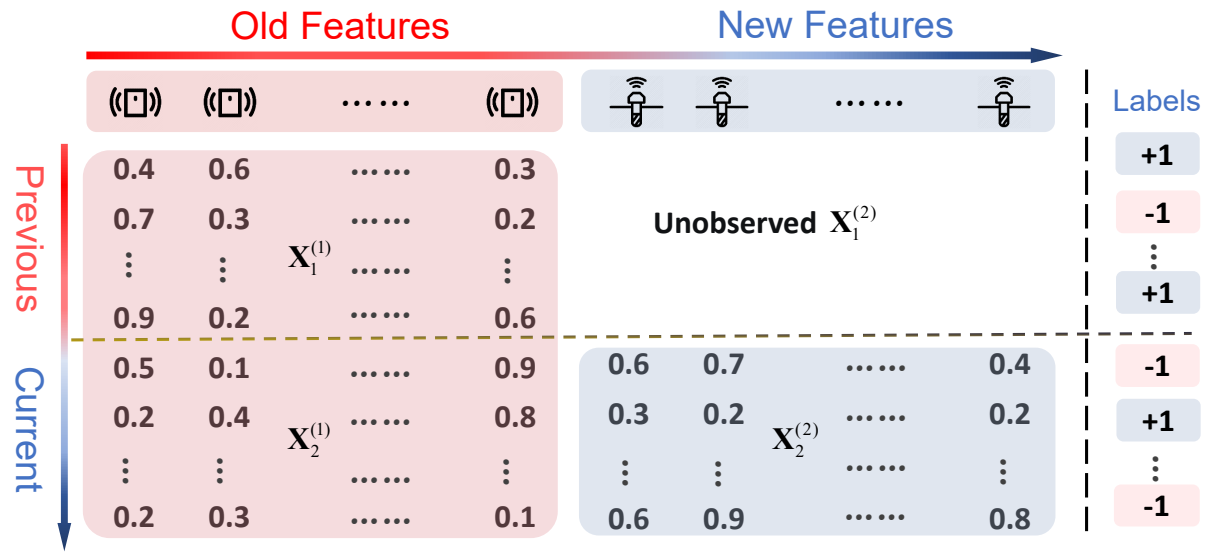


Figure 1: Illustration of incremental features in a dynamic environment. In the ecosystem monitoring task, the data returned by the new sensor (new type features) will accumulate to the previous sensor data (old type features).

bounds, providing rigorous and quantitative insights. The theoretical results reveal that the generalization ability of feature tailoring depends primarily on the predictive power of the old features, while data adaption is influenced by the sample size and the richness of information contained in the current sample. For model reuse, the quality of the pre-trained model plays a critical role, and for data reconstruction, the reconstruction distribution discrepancy is the key determinant. Beyond these theoretical findings, we validate their correctness through numerical comparisons. The main contributions of this paper are summarized as follows.

- We provide a comprehensive generalization analysis of four model design strategies for the feature incremental scenario, deriving their generalization bounds to offer practical guidance for model design.
- We advance rigorous and quantitative comparisons by analyzing the generalization ability of these strategies. Theoretical analysis identifies key factors influencing the tightness of their generalization bounds.
- Comprehensive experimental results corroborate the theoretical findings, enhancing their reliability and demonstrating the feasibility of applying these theoretical insights to model design.

2 Models

2.1 Notations

According to the scenario of this article shown in Figure. 1, the data collection process is divided into two stages, i.e, previous and current stages. The corresponding feature spaces include the previous feature space $\mathcal{X}_p \subseteq \mathbb{R}^{d_1}$ and the current feature space $\mathcal{X}_c \subseteq \mathbb{R}^{d_1+d_2}$. Similarly, we denote the label spaces of the two stages by $\mathcal{Y}_p$ and $\mathcal{Y}_c$, respectively. Due to the invariance of classification task and label space, we consider $\mathcal{Y}_p$ and $\mathcal{Y}_c$ to be identically distributed, denoted as $\mathcal{Y} = \{+1, -1\}$. Specifically, we denote the old features of the previous stage as $\mathbf{X}_1^{(1)}$, the unobserved augmented features as $\mathbf{X}_1^{(2)}$, and the label is $\mathbf{y}_p$. In the current stage, the feature shared by two stages is denoted as $\mathbf{X}_2^{(1)}$, and the augmented feature is denoted as $\mathbf{X}_2^{(2)}$, the label is $\mathbf{y}_c$. Here, $\mathbf{X}_i^{(1)} \in \mathcal{X}_p, i = 1, 2$, $\mathbf{X}_2 = [\mathbf{X}_2^{(1)}, \mathbf{X}_2^{(2)}] \in \mathcal{X}_c$. Without loss of generality, we assume that the data samples in the current stage are global observations and obey distribution $\mathcal{D}_c \triangleq \mathcal{X}_c \times \mathcal{Y}$. The data points in the previous stage are local observations and obey distribution $\mathcal{D}_p \triangleq \mathcal{X}_p \times \mathcal{Y}$. To simplify the presentation, we denote $S_p = (\mathbf{X}_1^{(1)}, \mathbf{y}_p)$ as the samples of previous stage of size n_1, $S_{c_p} = (\mathbf{X}_2^{(1)}, \mathbf{y}_c)$ as the current data of size n_2 that fall into the old feature space, and $S_c = (\mathbf{X}_2, \mathbf{y}_c)$ as the samples in current stage of size n_2.

2.2 Formulations

Firstly, we summarize and refine four strategies for addressing feature incremental problem, i.e., *feature tailoring, data adaption, model reuse, and data reconstruction*, based on the data application modes. The first two serve as baselines, which transform the feature increment learning problem into a traditional learning problem. Model reuse inherit pre-trained models from previous stages by imposing consistency constraints on corresponding local models. Finally, data reconstruction utilizes existing observations to recover unobserved features in the previous stage.

Subsequently, we carry out a model analysis on the four strategies. For illustration, consider the linear classifier. Denoted by $\mathcal{F}$ the hypothesis space, where each linear classifier $f : \mathbf{w}^\top \mathbf{x} \mapsto \mathbb{R}$. Consider a loss function $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ non-negative and Lipschitz continuous. Correspondingly, we denote the hypotheses under the four strategies as $f_i, i = 1, 2, 3, 4$. For the classification task, our goal is to learn a well-generalized classifier for the current data S_c, that is, minimize

the generalization error $R_{\mathcal{D}_c}(f)$. Since there is only empirical data on hand, we optimize the $\hat{R}(f)$ as an approximation.

Lemma 1 (Generalization Error Bound)**.** *Let $\mathcal{L}$ be the family of loss function associated to $\mathcal{F}$, i.e., $\mathcal{L} = \{\mathbf{x} \to \ell(f(\mathbf{x}), \mathbf{y}), f \in \mathcal{F}\}$. Suppose the loss function is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample of size m, the following inequality holds for all $f \in \mathcal{F}$:*

$$R_{\mathcal{D}}(f) \leq \hat{R}_m(f) + 2\mathfrak{R}_m(\mathcal{L}) + \sqrt{\frac{\log(1/\delta)}{2m}}, \tag{1}$$

where $\mathfrak{R}_m(\mathcal{L})$ is Rademacher complexity of loss function class $\mathcal{L}$ associated to $\mathcal{F}$, which can be bounded by using the celebrated Talagrand's lemma [Hahn and BFb, 1976].

Model design and theoretical analysis are carried out around the empirical error and hypothesis space complexity. Inspired by Lemma 1, we will derive the following optimization objective

$$\min_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \alpha \mathcal{H}(f). \tag{2}$$

The first term is the empirical error, $\mathcal{H}(f)$ is a regularization term used to control the complexity of the model, and α is a trade-off parameter.

Strategy 1 (Feature Tailoring)
In the incremental feature scenario, existing classification algorithms cannot be directly applied. One approach is to tailor the features by discarding the incremental features $\mathbf{X}_2^{(2)}$ and training the model using only partial observations. The optimization objective for this strategy is represented as

$$\min_{\mathbf{w},\ (\mathbf{x}_i, y_i) \in S_p \cup S_{c_p}} \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \alpha \left\|\mathbf{w}\right\|_2^2, \tag{3}$$

where $\mathbf{w}$ represents the coefficient vector corresponding to the linear classifier f_1.

Strategy 2 (Data Adaption)
Due to the inconsistent feature space dimensions between the two stages of data, existing classification algorithms cannot be directly applied. To address this, a data adaption approach is adopted. Specifically, the data from the previous stage, $\mathbf{X}_1^{(1)}$, is discarded. The optimization objective is formulated as follows

$$\min_{\mathbf{w},\ (\mathbf{x}_i, y_i) \in S_c} \frac{1}{n_2} \sum_{i=n_1}^{n_1+n_2} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \alpha \left\|\mathbf{w}\right\|_2^2. \tag{4}$$

Strategy 1 and Strategy 2 both adapt existing observations to the algorithm, and the difference lies in the way of data tailoring. Intuitively, both strategies are simple and convenient to operate, but part of the valuable observations are wasted.

Strategy 3 (Model Reuse)
The core concept of model reuse is to pre-train a model $\mathbf{w}_0$ during the previous stage and then develop an algorithm to train the classifier for the current stage by leveraging $\mathbf{w}_0$. Specifically, we assume that the model component $\mathbf{w}^1$ shared between the current and previous stages remains consistent with the pre-trained model $\mathbf{w}_0$. Based on this assumption, the optimization objective for strategy 3 is formulated as follows

$$\min_{\substack{\mathbf{w}, b \\ (\mathbf{x}_i, y_i) \in S_c}} \frac{1}{n_2} \sum_{i=1}^{n_2} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \alpha \left\|\mathbf{w}^2\right\|_2^2 + \beta \left\|\mathbf{w}^1 - \mathbf{w}_0\right\|_2^2 \tag{5}$$

Here, $\mathbf{w}^1$ and $\mathbf{w}^2$ represents the vector component of $\mathbf{w}$ corresponding to $\mathbf{X}_2^{(1)}$ and $\mathbf{X}_2^{(2)}$, and α and β are two trade-off parameters.

Strategy 4 (Data Reconstruction)
The main idea of data reconstruction is to use existing observations to reconstruct the unobserved feature $\mathbf{X}_1^{(2)}$ in the previous stage, and then use the reconstructed data as training data to train model. Specifically, the learning task includes data reconstruction and model training. Model training is the same as strategy 1, but the data participated in the training are different, which will not be repeated here. Next, we will mainly discuss data reconstruction.

For demonstration and theoretical analysis, we simply build a reconstruction function $\Phi : \mathbf{X}_\Omega \mapsto \mathbf{X}_1^{(2)}$ that reconstructs $\mathbf{X}_1^{(2)}$, $\mathbf{X}_\Omega$ refers to the existing observations, and $\mathbf{X}_1^{(2)}$ refers to the reconstruction features that are not observed in the previous stage. In this paper, we focus on leveraging the data correlations between the two stages and the feature correlations between the old and new features within the current stage to reconstruct the unobserved features from the previous stage. Based on this, we design the following framework for reconstruction functions Φ

$$\Phi\left(\mathbf{X}_\Omega\right) = \underset{\mathbf{X}_1^{(2)}}{\arg\min}\, \mathcal{L}(\mathbf{D}) + \lambda \mathcal{R}(\mathbf{F}), \tag{6}$$

where λ is a trade-off parameter, $\mathcal{L}(\mathbf{D})$ represents the reconstruction loss of using data correlation to reconstruct the unobserved features, and $\mathcal{R}(\mathbf{F})$ represents the reconstruction loss of using feature correlation.

3 Generalization Ability Analysis

In this part, we will analyze, compare and discuss the generalization ability for the four strategies. First of all, we need to introduce some basic definitions and lemmas.

Definition 1 (Rademacher Complexity [Bartlett and Mendelson, 2001])**.** *Given a function class $\mathcal{F}$. For a function $f \in \mathcal{F}$ and a sample $\mathrm{Z} = (Z_1, Z_2, \cdots, Z_m)$ of size m, $\mathrm{Z} \in \mathcal{Z} = (\mathcal{X}, \mathcal{Y})$. Then, the empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample Z is defined as*

$$\hat{\mathfrak{R}}_{\mathrm{Z}}(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(Z_i) \right], \tag{7}$$

where the random variables σ_i are called Rademacher variables, which obey the uniform distribution on $\{-1, +1\}$.

The Rademacher complexity of $\mathcal{F}$ is the expectation of empirical Rademacher complexity based on the experience of all samples of size m drawn by $\mathcal{D}$

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\mathrm{Z} \sim \mathcal{D}^m} \left[\hat{\mathfrak{R}}_{\mathrm{Z}}(\mathcal{F}) \right]. \tag{8}$$

Definition 2 ($\mathcal{Y}$-Discrepancy [Mohri and Medina, 2012]). *Let $\mathcal{D}_P, \mathcal{D}_Q$ be two distributions over $\mathcal{X}$ and denote by y_P, y_Q the labeling functions over $\mathcal{D}_P$ and $\mathcal{D}_Q$, respectively. Given a hypothesis class $\mathcal{F}$ and the corresponding loss function ℓ, the $\mathcal{Y}$-discrepancy between $(\mathcal{D}_P, y_P)$ and $(\mathcal{D}_Q, y_Q)$ is defined as*

$$disc_Y\left(S_{P_\alpha}, S_Q\right) = \sup_{f \in \mathcal{F}} \left| R_{D_P}\left(f, y_P\right) - R_{D_Q}\left(f, y_Q\right) \right|. \tag{9}$$

As we only have the empirical data on hand, by introducing weights α over the empirical data S_P sampled from $\mathcal{D}_P$, and thus the weighted empirical risk is defined as

$$\hat{R}_{S_{P_\alpha}} = \frac{1}{m} \sum_{i=1}^{m} \alpha_i \ell(f(\mathbf{x}_i), y_{Pi}),$$

the weighted empirical $\mathcal{Y}$-discrepancy is denoted by

$$disc_{\mathcal{Y}}\left(S_{P_\alpha}, S_Q\right) = \sup_{f \in \mathcal{F}} \left| \hat{R}_{S_{P_\alpha}}\left(f, y_P\right) - \hat{R}_{S_Q}\left(f, y_Q\right) \right|. \tag{10}$$

With the definition of weighted empirical $\mathcal{Y}$-discrepancy, the generalization error on $(\mathcal{D}_Q, y_Q)$ can be bounded in terms of the risk over $(\mathcal{D}_P, y_P)$ and their $\mathcal{Y}$-discrepancy.

Subsequently, we carry out a generalization theoretical analysis of the four strategies. As for the baseline strategies *feature tailoring* and *data adaption*, the generalization bounds are similar to Lemma 1. The difference between the two strategies lies in the form and amount of participating training data. Specifically, comparing the generalization bounds of the two strategies, we take a simple example as illustration. As shown in Figure 2, consider a binary classification task with $\mathcal{Y} = \{+1, -1\}$ and let $d = 2$. We can see that using any local Feature 1 or Feature 2 does not work well for this classification task, since the trained model is underfitting. Correspondingly, the right of Figure 2 shows the generalization error curves of the model trained with local and global features, respectively.

It can be found that the key factors dominating the generalization ability of *feature tailoring* and *data adaption* are the quality of old features and the number of data samples, respectively. This conclusion is straightforward and intuitive, so we will not elaborate further. We next derive generalization error bounds for strategies 2 and 3. Due to space limitations, the details of the proofs are list in the supplementary file.

Theorem 1. *Let $\mathcal{F}_2$ be the family of the hypothesis set, and denote the hypothesis returned by data adaption as f_2. Suppose the loss function is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample of size n_2, the following inequalities hold for all $f_2 \in \mathcal{F}_2$.*

$$R_{\mathcal{D}_c}(f_2) \leq \hat{R}_{n_2}(f_2) + 2L\frac{1}{\sqrt{n_2}}\Lambda M_2 + \sqrt{\frac{\log(1/\delta)}{2n_2}}. \tag{11}$$

Where $\Lambda = \max\{\|\mathbf{x}\| \,|\mathbf{x} \in \mathcal{X}\}$ represents the radius of the feature domain, and $\|\mathbf{w}_2\| \leq M_2$, M_2 represents the radius of the linear hypothesis space. $\mathbf{w}_2$ represents the hypothesis coefficient corresponding to the linear classifier f_2.

Theorem 2. *Let $\mathcal{F}_3$ be the family of the hypothesis set, and denote the hypothesis returned by model reuse as f_3. Suppose the loss function is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample of size n_2, the following inequalities holds for all $f_3 \in \mathcal{F}_3$,*

$$\begin{aligned} R_{\mathcal{D}_c}(f_3) \leq & \hat{R}_{n_2}(f_3) + \sqrt{\frac{\log(1/\delta)}{2n_2}} \\ & + 2L\frac{1}{\sqrt{n_2}}\Lambda\left(\varepsilon + \sqrt{(M_3)^2 - (M_0 - \varepsilon)^2}\right) \end{aligned} \tag{12}$$

Here, $\Lambda = \max\{\|\mathbf{x}\| \,|\mathbf{x} \in \mathcal{X}\}$ represents the radius of the feature domain, and $\|\mathbf{w}_3\| \leq M_3$, M_3 represents the radius of the hypothesis space. $\mathbf{w}_3$ represents the hypothesis coefficient corresponding to the linear classifier f_3. $\|\mathbf{w}_0\| = M_0$, $\mathbf{w}_0$ represents the coefficient vector corresponding to the linear classifier pre-trained by $\left(\mathbf{X}_1^{(1)}, \mathbf{y}_1\right)$. $\left\|\mathbf{w}_3^1 - \mathbf{w}_0\right\| \leq \varepsilon$, $\mathbf{w}_3^1$ represents the vector component of $\mathbf{w}_3$ corresponding to $\mathbf{w}_0$.

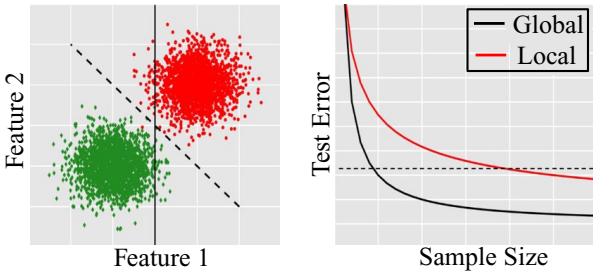


Figure 2: Illustration of the generalization ability of Strategy 1 versus Strategy 2.

As shown in Theorems 1 and 2, the bounds differ from the standard generalization bound as the Rademacher complexity is concretized using the Frobenius norm of matrices. In this way, we can compare the generalization bounds of *data adaption* and *model reuse*. Comparing the generalization bounds of f_2 and f_3, it can be observed that the tightness of the generalization bound is largely influenced by the second term, which corresponds to the Rademacher complexity of the hypothesis space. Therefore, our analysis primarily focuses on comparing this term. In our setting, the current stage feature space retains the old features from the previous stage. Therefore, the classification coefficient $\mathbf{w}_0$ pretrained by the previous stage data should be inherited. It can be obtained that $\mathbf{w}_3^1$ is in the ε-neighborhood of $\mathbf{w}_0$, that is, $\left\|\mathbf{w}_3^1 - \mathbf{w}_0\right\| \leq \varepsilon$, and $\varepsilon \ll M_0$. Besides, compared to $\mathcal{F}_2$, the hypothesis function space $\mathcal{F}_3$ has been constrained due to the adding of $\left\|\mathbf{w}_3^1 - \mathbf{w}_0\right\|_2^2$. Thus, it is natural to assume that $M_3 \leq M_2$. With these mild assumptions, we have the following corollary.

Corollary 1. *Assume that $\left\|\mathbf{w}_3^1 - \mathbf{w}_0\right\| \leq \varepsilon$, $M_3 = \rho M_0$ with $\rho > 1$ and $M_3 \leq M_2$. When $\varepsilon \leq \frac{1}{2}\left(\rho + 1 - \sqrt{(\rho+1)^2 - 2}\right) M_0$, we have*

$$\varepsilon + \sqrt{(M_3)^2 - (M_0 - \varepsilon)^2} \leq M_2. \tag{13}$$

Theorem 1 and Theorem 2 give the generalization error bounds of the hypothesis trained by *data adaption* and *model reuse*. Corollary 1 compared the two generalization upper

Strategies	Main parts	Key factors
Feature Tailoring	$\hat{R}_{n_1+n_2}(f_1)$	Old features' predictive power
data adaption	$2L\frac{1}{\sqrt{n_2}}\Lambda M_2$	Sample size n_2 and the contained information richness
Model Reuse	$\frac{2L\Lambda}{\sqrt{n_2}}\left(\sigma+\sqrt{(M_3)^2-(M_0-\sigma)^2}\right)$	Pre-trained model quality
Data Reconstruction	$disc_{\mathcal{Y}}\left(\hat{S}_{p_\alpha}, S_c\right)$	Reconstruction distribution discrepancy

Table 1: The main parts of generalization bounds of four strategies.

bounds. It can be concluded that the empirical error is a good approximation of the generalization error as the training data tends to infinity. In addition, the generalization error upper bound of f_3 is tighter. In intuitive, *model reuse* reduces the size of the hypothesis space, since $\mathbf{w}_3^1$ is restricted in the ε-neighborhood of $\mathbf{w}_0$. Therefore, we know that the key factor in dominating the generalization ability of the model reuse is the quality of the pre-trained model, that is, the higher the quality of the pre-trained model, the smaller the value of ε and the tighter the generalization bound.

Theorem 3. *Let $\mathcal{F}_4$ be the family of hypothesis set, and denote the hypothesis returned by data reconstruction as f_4. Suppose the loss function is L-Lipschitz, then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample $\hat{S}_p$ of size n_1 and a sample S_c of size n_2, the following inequality holds for all $f_4 \in \mathcal{F}_4$ and any weighted empirical distribution distribution $\mathcal{D}_{p_\alpha}$ over the sample S_p and current distribution $\mathcal{D}_c$ over the sample S_c.*

$$R_{\mathcal{D}_c}(f_4, y_c) \leq \lambda \hat{R}_{\hat{S}_{P_\alpha}}(f_4, y_p) + (1-\lambda)\hat{R}_{S_c}(f_4, y_c) + \lambda disc_{\mathcal{Y}}\left(\hat{S}_{p_\alpha}, S_c\right) + 2L\mathfrak{R}_(n_1+n_2)(\mathcal{F}_4) + \sqrt{\frac{\log(1/\delta)}{2(n_1+n_2)}}. \tag{14}$$

Here, $\hat{\mathcal{D}}_p$ represents the distribution of the reconstructed previous stage data $[\mathbf{X}_1^{(1)}, \hat{\mathbf{X}}_1^{(2)}]$, and $\hat{S}_p$ represents the samples sampled from $\hat{\mathcal{D}}_p$. $\mathcal{D}_c$ represents the distribution of the current stage data, which is the same as the test data distribution, and S_c represents the samples sampled from $\mathcal{D}_c$. λ represents the ratio of samples $\hat{S}_p$ and S_c.

According to Theorem 3, we know that the generalization ability of *data reconstruction* is mainly affected by the distribution discrepancy between the reconstructed data and the current stage data, which is directly related to the performance of the reconstruction function. By observing, a small distribution discrepancy $disc_{\mathcal{Y}}\left(\hat{S}_{p_\alpha}, S_c\right)$ leads to a hypothesis with a tighter generalization error bound and inspires us to consider the distribution discrepancy between the reconstructed data and the observed data when designing the reconstruction function. It should be emphasized that data reconstruction is an independent and arduous task. In this paper, we present a demonstration of designing the reconstruction function based on optimal transport and feature correlation. In summary, we know that the key factor in dominating the generalization ability of the data reconstruction is the reconstruction distribution discrepancy, which plays a key role in the design of reconstruction functions.

Finally, Table 1 summarizes the generalization bounds of the four strategies: *feature tailoring*, *data adaptation*, *model reuse*, and *data reconstruction*. The main parts of the generalization bounds highlight the key factors influencing generalization ability. Illustration 2 compares the generalization performance of *feature tailoring* and *data adaptation*. It reveals that the primary factor for *feature tailoring* is the predictive power of the old features, while for *data adaptation*, it is the number of data samples. Corollary 1 contrasts the generalization bounds of *data adaptation* and *model reuse*, identifying the quality of pre-trained models as the critical determinant for *model reuse*. Theorem 3 examines the generalization ability of *data reconstruction*, concluding that its effectiveness hinges on the reconstruction distribution discrepancy, which is closely tied to the design of the reconstruction function.

4 Mutual Verification Experiments

In this section, soft margin SVM [Cortes and Vapnik, 1995] and logistic regression (LR) [Berger *et al.*, 1996] are applied as demonstrations, aiming to form mutual verification through experiments and theories. Due to space limitation, the optimization objectives and detailed implementation information of the four strategies are provided in the supplementary materials.

4.1 Datasets and Setting

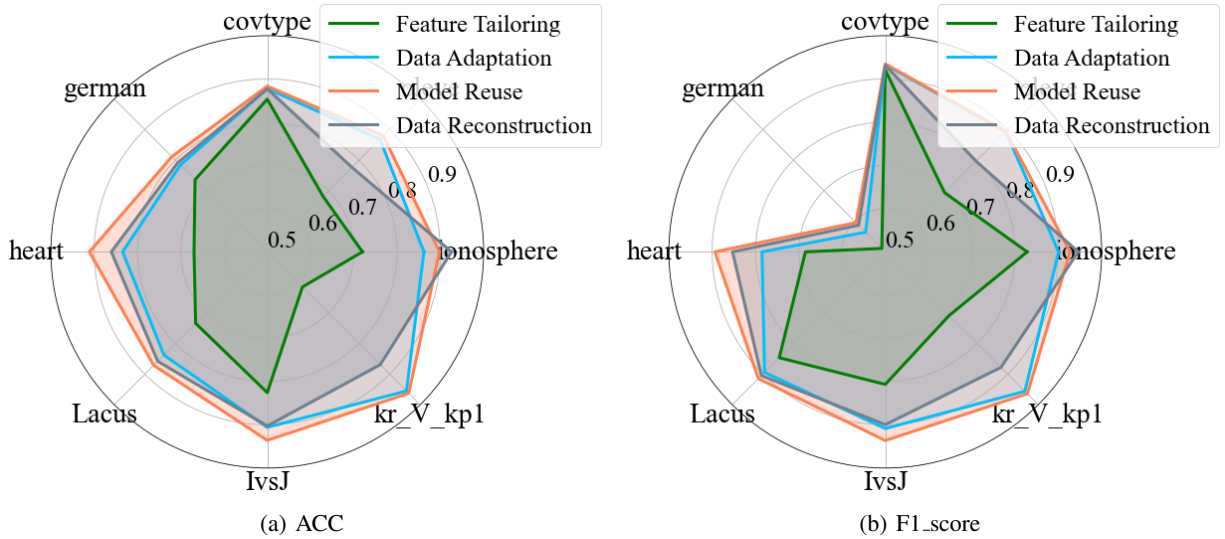
We adopt 8 datasets from UCI Repository [1] and LIBSVM Library [2] to carry out the experiments. As in the feature increment scenario mentioned in this paper, the data collection process is divided into two stages. Specifically, the global feature of the existing data is denoted as $\mathbf{X} = [X^1, \cdots, X^{d_1+d_2}] \subseteq \mathbb{R}^{d_1+d_2}$, where $X^i, i = 1, 2, \cdots, d_1+d_2$ is the i-th feature. Let $\mathbf{X} = [X^1, \cdots, X^{d_1}] \subseteq \mathbb{R}^{d_1}$ be the previous feature and $\mathbf{X} = [X^{d_1}, \cdots, X^{d_1+d_2}] \subseteq \mathbb{R}^{d_1+d_2}$ be the current feature. Furthermore, to obtain the data type that conforms to the scenario in this paper, we tailor the existing binary classification data. Without loss of generality, we let $d_1 = d_2$. Similarly, let $n_2 = n_1/2$ be the amount of data in the previous stage. As for the parameters selection of the algorithms, we conduct K-fold cross-validation on the training set. Specifically, we use the grid search method to obtain the optimal parameter combination, and the search range of each parameter

[1] http://archive.ics.uci.edu/ml

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Dataset	Strategy	Acc	AUC	F1-Score
ionosphere	Feature Tailoring	0.6706±0.0046	0.6854±0.0047	0.7575±0.0038
	data adaption	0.7013±0.0029	0.7326±0.0043	0.7748±0.0035
	Model Reuse	0.7226±0.0031	0.7431±0.0042	0.8236±0.0025
	Data Reconstruction	**0.7422±0.0032**	**0.7561±0.0022**	**0.8431±0.0041**
cleve	Feature Tailoring	0.6083±0.0034	0.5325±0.0046	0.5657±0.0056
	data adaption	0.6211±0.0056	0.5256±0.0032	0.6012±0.0035
	Model Reuse	**0.6647±0.0034**	**0.5494±0.0043**	**0.6674±0.0046**
	Data Reconstruction	0.5816±0.0036	0.5005±0.0042	0.5544±0.0035
covtype	Feature Tailoring	0.5168±0.0053	0.5971±0.0042	0.6483±0.0035
	data adaption	0.6734±0.0042	0.6544±0.0026	0.7637±0.0035
	Model Reuse	**0.7013±0.0038**	**0.6636±0.0032**	**0.8019±0.0045**
	Data Reconstruction	0.5236±0.0041	0.6037±0.0032	0.6510±0.0065
german	Feature Tailoring	0.7060±0.0036	0.4736±0.0032	0.4968±0.0025
	data adaption	0.7386±0.0041	0.6034±0.0032	0.5477±0.0026
	Model Reuse	**0.7642±0.0051**	**0.6833±0.0062**	**0.5733±0.0045**
	Data Reconstruction	0.6833±0.0034	0.4621±0.0022	0.4708±0.0041
heart	Feature Tailoring	0.7437±0.0026	0.5393±0.0024	0.6976±0.0035
	data adaption	0.8133±0.0043	0.7864±0.0032	0.8196±0.0036
	Model Reuse	**0.8300±0.0069**	**0.8163±0.0032**	**0.8243±0.0046**
	Data Reconstruction	0.7755±0.0040	0.7495±0.0069	0.7685±0.0045
Lacus	Feature Tailoring	0.7060±0.0043	0.6819±0.0043	0.7818±0.0048
	data adaption	0.7471±0.0054	0.6842±0.0054	0.8548±0.0034
	Model Reuse	**0.7880±0.0039**	**0.7134±0.0039**	**0.8609±0.0032**
	Data Reconstruction	0.7123±0.0047	0.6903±0.0047	0.8234±0.0045
IvsJ	Feature Tailoring	0.7627±0.0034	0.7585±0.0042	0.6851±0.0041
	data adaption	0.7827±0.0044	0.7735±0.0032	0.7021±0.0034
	Model Reuse	**0.8027±0.0051**	**0.7923±0.0034**	**0.7321±0.0024**
	Data Reconstruction	0.7717±0.0034	0.7635±0.0046	0.6831±0.0046
kr_V_kp1	Feature Tailoring	0.6954±0.0041	0.5771±0.0032	0.7126±0.0035
	data adaption	0.7064±0.0038	0.6023±0.0031	0.7356±0.0039
	Model Reuse	**0.7664±0.0048**	**0.6634±0.0054**	**0.7826±0.0035**
	Data Reconstruction	0.6864±0.0038	0.5823±0.0032	0.7156±0.0043

Table 2: Comparative experiment results ("↑")(mean±std) of SVM model under the four strategies. The best results on each dataset are bolded.



(a) ACC (b) F1_score

Figure 3: Comparative experiment results ("↑")(mean±std) of LR model under the four strategies.

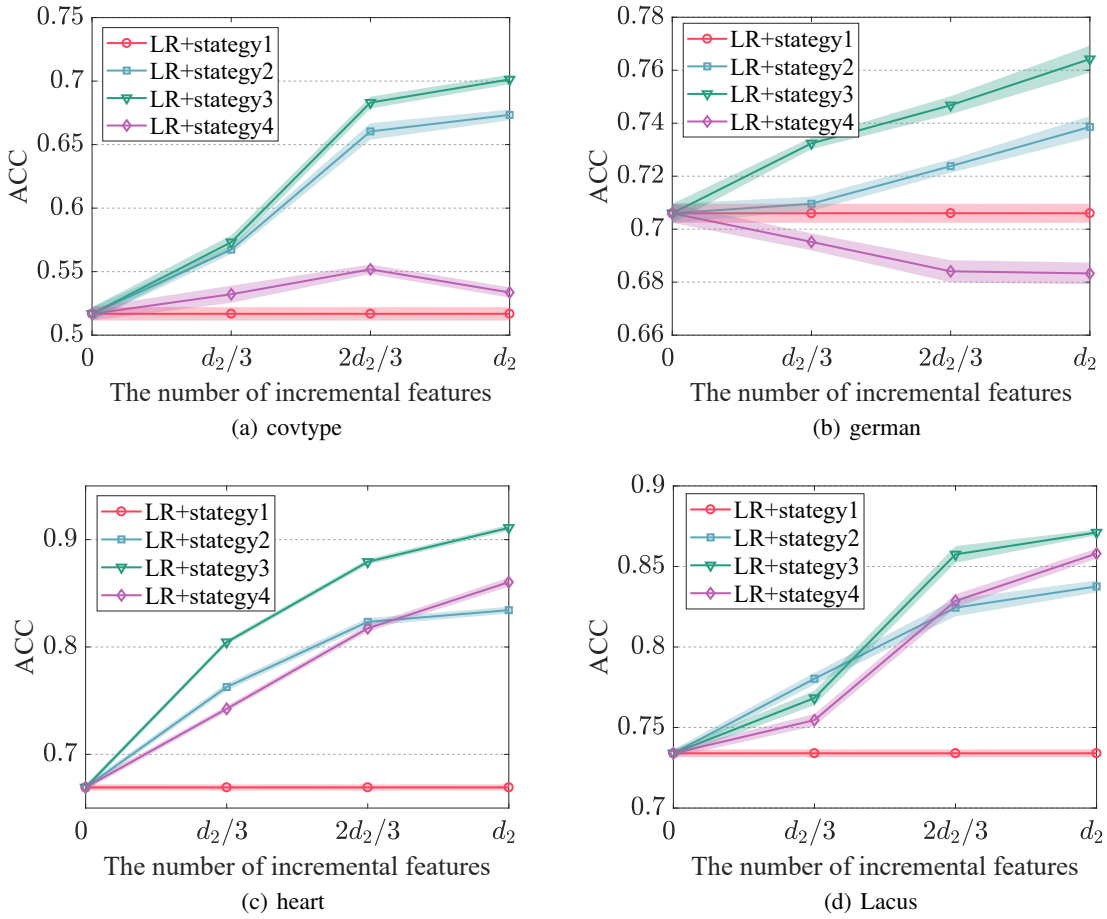


Figure 4: Incremental process experiment results ("↑")(mean±std) of two models under the four strategies.

is $\left[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\right]$. Finally, three common metrics, Accuracy, F1-score, and AUC, were used to evaluate the model performance, and the higher the values of these metrics, the better the performance of the algorithms("↑"). Additionally, we also conducted experiments on several real-world multi-view datasets. Due to space constraints, the experimental results are presented in the supplementary materials.

4.2 Comparative Experiment

In this subsection, we conduct experiments to verify the validity of the above theoretical conclusions. The comparative experiment results are shown in Table 2 and Figure 3. According to the experiment results, it can be concluded that *data adaption* performs better than *feature tailoring* in general, which is consistent with the intuition, that is, the model obtained from *feature tailoring* could be underfitting. In addition, the model performance of *model reuse* has certain advantages over *data adaption*, which is consistent with the conclusion of Corollary 1, that is, the generalization error bound for *model reuse* is tighter than *data adaption*. As for *data reconstruction*, the performance of the model is unstable and generally worse than *model reuse*, since the reconstruction function cannot effectively reduce the distribution discrepancy between the reconstructed data and the observations without prior distribution information about the data. In particular, *data reconstruction* will perform well when the potential data distribution relatively matches the reconstruction function, such as for the dataset 'ionosphere'. That is to say, *data reconstruction* can achieve good generalization ability only when the distribution discrepancy between the reconstructed data and the observations is sufficiently small. This requires incorporating more prior distribution information to construct the reconstruction function.

4.3 Numerical Verification Experiment

In order to verify Corollary 1 numerically, we calculated the numerical value of each variable in Corollary 1 and combined with the condition for comparative verification. Specifically, we further analyze the ε-neighborhood. Denote f_* as the empirical optimal classifier trained by strategy 3, and $\mathbf{w}_*$ as the corresponding coefficient. Let $\hat{\mathbf{w}} = \left[\mathbf{w}_0, \mathbf{w}_*^2\right]$ be the coefficient corresponding to $\hat{f}$. Due to the optimality of the empirical objective at $\mathbf{w}_*$, we have

$$\begin{aligned} &\hat{R}_{n_2}(f_*) + \alpha \left\|\mathbf{w}_*^2\right\|_F^2 + \beta \left\|\mathbf{w}_*^1 - \mathbf{w}_0\right\|_F^2 \\ &\quad \leq \hat{R}_{n_2}(\hat{f}) + \alpha \left\|\mathbf{w}_*^2\right\|_F^2 + \beta \left\|\mathbf{w}_0 - \mathbf{w}_0\right\|_F^2 \end{aligned} \tag{15}$$

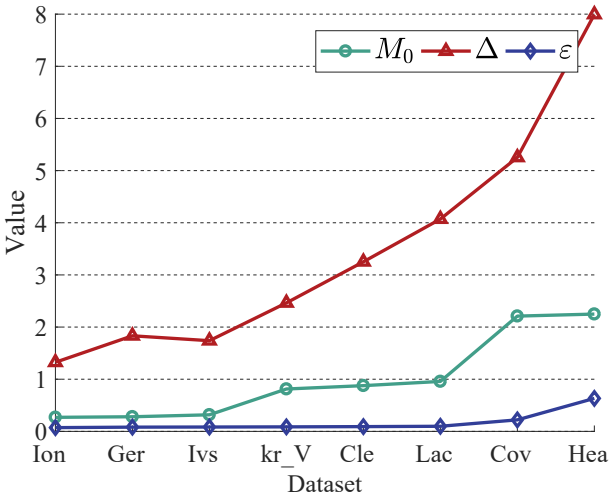


Figure 5: The numerical values of ε, Δ and M_0 produced by the experiments on 8 datasets.

then

$$\left\|\mathbf{w}_*^1 - \mathbf{w}_0\right\|_F^2 \leq \frac{1}{\beta}\left(\hat{R}_{n_2}(\hat{f}) - \hat{R}_{n_2}(f_*)\right)$$

$$\left\|\mathbf{w}_*^1 - \mathbf{w}_0\right\| \leq \sqrt{\frac{1}{\beta}\left(\hat{R}_{n_2}(\hat{f}) - \hat{R}_{n_2}(f_*)\right)}$$

Let

$$\varepsilon = \sqrt{\frac{1}{\beta}\left(\hat{R}_{n_2}(\hat{f}) - \hat{R}_{n_2}(f_*)\right)}$$

$$\Delta = \frac{1}{2}\left(\rho + 1 - \sqrt{(\rho+1)^2 - 2}\right) M_0,$$

we compare the numerical values of ε, Δ and M_0 produced by the experiments. The experimental results are shown in Figure 5. From the experimental results, it can be seen that the condition $\varepsilon \leq \frac{1}{2}\left(\rho + 1 - \sqrt{(\rho+1)^2 - 2}\right) M_0$ is always satisfied.

4.4 Impact of Incremental Features

To investigate the effect of incremental features on model performance, we conducted experiments by progressively increasing the number of incremental features from 0 to d_2. Specifically, we selected $d_2/3$, $2d_2/3$, and d_2 as representative points to illustrate performance trends. Results in Figure 4 reveal the following: (1) *Feature tailoring* performance depends only on the observed features from the previous stage. (2) For *data adaptation* and *model reuse*, performance generally improves with more incremental features, as additional features provide more information, alleviating underfitting. (3) In contrast, *data reconstruction* shows inconsistent performance, with degradation on some datasets due to increased reconstruction complexity as the feature count grows.

5 Conclusion

In this paper, we focus on the feature increment learning problem, for which theoretical analysis is still limited. To gain a deep understanding of the workings of this complex machine learning problem, we first summarize and refine four strategies, i.e., *feature tailoring, data adaption, model reuse, and data reconstruction*, based on the data application modes. Furthermore, we carry out research on the generalization theory of these four typical strategies in feature incremental scenarios. Specifically, we propose a common procedure and analyze the generalization ability of these four common data application strategies, and make a horizontal comparison of them to derive rigorous and quantitative conclusions. In addition, a series of experiments prove that the theory in this paper is effective, which is helpful to guide the model design through the theory.

Acknowledgments

This work was partially supported by the National Key Research and Development Program (No. 2022ZD0114803), the NSF for Distinguished Young Scholars under Grant No. 62425607, the Key NSF of China under Grant No. 62136005. Chenping Hou is the corresponding author.

References

[Antos *et al.*, 2002] András Antos, Balázs Kégl, Tamás Linder, and Gábor Lugosi. Data-dependent margin-based generalization bounds for classification. *J. Mach. Learn. Res.*, 3:73–98, 2002.

[Bartlett and Mendelson, 2001] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT*, volume 2111 of *Lecture Notes in Computer Science*, pages 224–240. Springer, 2001.

[Berger *et al.*, 1996] Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguistics*, 22(1):39–71, 1996.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[Das *et al.*, 2013] Shubhomoy Das, Travis Moore, Weng-Keen Wong, Simone Stumpf, Ian Oberst, Kevin McIntosh, and Margaret M. Burnett. End-user feature labeling: Supervised and semi-supervised approaches based on locally-weighted logistic regression. *Artif. Intell.*, 204:56–74, 2013.

[de Mello and Ponti, 2018] Rodrigo Fernandes de Mello and Moacir Antonelli Ponti. *Machine Learning - A Practical Approach on the Statistical Learning Theory*. Springer, 2018.

[Dietterich, 2017] Thomas G. Dietterich. Steps toward robust artificial intelligence. *AI Mag.*, 38(3):3–24, 2017.

[Downey *et al.*, 2010] Doug Downey, Oren Etzioni, and Stephen Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction. *Artif. Intell.*, 174(11):726–748, 2010.

[El-Yaniv and Pechyony, 2007] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its

applications. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007, Proceedings*, volume 4539 of *Lecture Notes in Computer Science*, pages 157–171. Springer, 2007.

[Gu *et al.*, 2022] Shilin Gu, Yuhua Qian, and Chenping Hou. Incremental feature spaces learning with label scarcity. *ACM Trans. Knowl. Discov. Data*, 16(6):106:1–106:26, 2022.

[Hahn and BFb, 1976] M. G. Hahn and BFb. *Probability in Banach Spaces*. Probability in Banach Spaces, 1976.

[He *et al.*, 2021] Haiyun He, Hanshu Yan, and Vincent Y. F. Tan. Information-theoretic generalization bounds for iterative semi-supervised learning. *CoRR*, abs/2110.00926, 2021.

[Hou and Zhou, 2018] Chenping Hou and Zhi-Hua Zhou. One-pass learning with incremental and decremental features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2776–2792, 2018.

[Hou *et al.*, 2019] Chenping Hou, Ling-Li Zeng, and Dewen Hu. Safe classification with augmented features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2176–2192, 2019.

[Hou *et al.*, 2021] Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Learning with feature evolvable streams. *IEEE Trans. Knowl. Data Eng.*, 33(6):2602–2615, 2021.

[Lei *et al.*, 2019] Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Trans. Inf. Theory*, 65(5):2995–3021, 2019.

[Li and Liu, 2021] Shaojie Li and Yong Liu. Sharper generalization bounds for clustering. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6392–6402. PMLR, 2021.

[Li *et al.*, 2019] Feijiang Li, Yuhua Qian, Jieting Wang, Chuangyin Dang, and Liping Jing. Clustering ensemble based on sample's stability. *Artif. Intell.*, 273:37–55, 2019.

[Li *et al.*, 2022] Jian Li, Yong Liu, and Weiping Wang. Convolutional spectral kernel learning with generalization guarantees. *Artif. Intell.*, 313:103803, 2022.

[Liu and Chen, 2018] Chao Liu and Di-Rong Chen. Generalization error bound of semi-supervised learning with ^{1} regularization in sum space. *Neurocomputing*, 275:1793–1800, 2018.

[Mohri and Medina, 2012] Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. *CoRR*, abs/1205.4343, 2012.

[Morvant *et al.*, 2012] Emilie Morvant, Sokol Koço, and Liva Ralaivola. Pac-bayesian generalization bound on confusion matrix for multi-class classification. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[Sadreddin and Sadaoui, 2021] Armin Sadreddin and Samira Sadaoui. Incremental feature learning for infinite data. *CoRR*, abs/2108.02932, 2021.

[Weston, 2013] Jason Weston. Statistical learning theory in practice. In Bernhard Schölkopf, Zhiyuan Luo, and Vladimir Vovk, editors, *Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik*, pages 81–93. Springer, 2013.

[Xu and Zeevi, 2020] Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *CoRR*, abs/2011.06186, 2020.

[Xu *et al.*, 2016] Chang Xu, Dacheng Tao, and Chao Xu. Streaming view learning. *CoRR*, abs/1604.08291, 2016.

[Yang *et al.*, 2022] Yanyan Yang, Degang Chen, Xiao Zhang, Zhenyan Ji, and Yingjun Zhang. Incremental feature selection by sample selection and feature-based accelerator. *Appl. Soft Comput.*, 121:108800, 2022.

[Ye *et al.*, 2018] Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, and Zhi-Hua Zhou. Rectify heterogeneous models with semantic mapping. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1904–1913. PMLR, 2018.

[Zhang *et al.*, 2020] Zhenyu Zhang, Peng Zhao, Yuan Jiang, and Zhi-Hua Zhou. Learning with feature and distribution evolvable streams. In *Proceedings of the 37th International Conference on Machine Learning,ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 11317–11327. PMLR, 2020.