Adaptive Learning for Weakly Labeled Streams

Zhen-Yu Zhang
zhangzy@lamda.nju.edu.cn
National Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing, China

Yu-Yang Qian
qianyy@lamda.nju.edu.cn
National Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing, China

Yu-Jie Zhang
zhangyj@lamda.nju.edu.cn
National Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing, China

Yuan Jiang
jiangy@lamda.nju.edu.cn
National Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing, China

Zhi-Hua Zhou
zhouzh@lamda.nju.edu.cn
National Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing, China

ABSTRACT

In plenty of real-world applications, data are collected in a streaming fashion, and their accurate labels are hard to obtain. For instance, in the environmental monitoring task, sensors are collecting the data all the time. Still, their labels are scarce because the labeling process requires human effort and can conceal annotation errors. This paper investigates the problem of learning with weakly labeled data streams, in which data are continuously collected, and only a limited subset of streaming data is labeled but potentially with noise. This setting is challenging and of great importance but rarely studied in the literature. When the data are constantly gathered with unknown noise on labels, it is quite challenging to design algorithms to obtain a well-generalized classifier. To address this difficulty, we propose a novel noise transition matrix estimation approach for data streams with scarce noisy labels by online anchor points identification. Based on that, we propose an adaptive learning algorithm for weakly labeled data streams via model reuse and effectively alleviate the negative influence of label noise with unlabeled data. Both theoretical analysis and extensive experiments justify and validate the effectiveness of the proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Online learning settings**.

KEYWORDS

weakly supervised learning, data stream, noisy labels

ACM Reference Format:
Zhen-Yu Zhang, Yu-Yang Qian, Yu-Jie Zhang, Yuan Jiang, and Zhi-Hua Zhou. 2022. Adaptive Learning for Weakly Labeled Streams. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539351

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539351

1 INTRODUCTION

In recent years, machine learning algorithms have achieved prolific success in various real-world applications [36]. These approaches, such as deep learning, typically build models on a vast number of data with accurate labels and then deploy them in the testing phase. However, in many real-world applications, data are collected in the stream fashion, and their labels are not always available because the labeling process requires human effort expertise. Meanwhile, the labels might be inaccurate due to manual labeling errors. Therefore, it is desired to facilitate a learning system to online update, respond in real-time, and be robust to weak supervisions [35].

In this paper, we consider the problem of learning with weakly labeled data streams. Specifically, data are collected in the form of a stream, with a small subset of data is observed with labels while the others remain unlabeled. Meanwhile, the observed labels might be inaccurate due to manual labeling errors. This setting is crucial because it occurs in a variety of real-world applications. For instance, the sensors continuously collect the data in environment monitoring tasks or human activity recognition. At the same time, their labels are manually annotated, which is only a small subset of the whole data stream and could be wrongly labeled. A similar situation also occurs in web data classification, where the unlabeled data are accumulating over time. As the limitation on human effort, only a small part of data is labeled and conceals noisy labels.

Most existing approaches consider either offline noisy label learning or data stream learning with accurately labeled data. To obtain a statistically consistent classifier, label noise robust learning approaches typically need the prior knowledge of the noise transition matrix or estimate them with sufficient noisy labeled data in an offline manner [2]. When the noise transition matrix is unknown and the data come in the form of a stream, it is hard to adapt the offline techniques to online estimate the noise transition matrix and recover the true loss of the underlying noise-free distribution. Data stream learning approaches either maintain a single model incrementally [5], or an ensemble of base learners [25]. This line of work requires the ground-truth labels; the rise of label noise and missing will deteriorate the performance if we directly update an incremental model or ensemble the base learners on noisy labeled

data. A few studies consider learning data streams with noisy labels [8], but they typically focus on cleansing outliers and do not explore the unlabeled data, whose number is much larger than that of noisy labeled data in real-world tasks.

Due to the existence of label noise in data streams, we are required to align the learning models trained on weakly labeled data to the noise-free distribution to obtain a statistically consistent classifier rather than a direct ensemble. The central ingredient is estimating the noise transition matrix to recover the underlying true loss from weakly labeled data. A series of offline learning methods are proposed to estimate the noise transition matrix with pre-collected sufficient noisy labeled data. Based on the anchor points assumption [19, 30] or anchor set condition [23], we can calculate the noise transition matrix or distinguish the noise-free distribution from noisy labeled data. However, these approaches typically need to pre-collect a large number of noisy labeled data and then identify the anchor points or estimate the mixture proportion for noise transition matrix approximation in an offline manner.

The challenges arise to estimate the noise transition matrix from data streams with a limited number of noisy labels and a vast number of unlabeled data. This problem turns out quite challenging, and it is non-trivial to take the advance in offline noisy transition matrix estimation approaches to address this problem. On the one hand, when the data arrives in a high-throughput stream, previous methods only explore a part of the data as the stream is generally too large to fully store in the memory. Furthermore, we must update the model online and label the data in real-time. On the other hand, these approaches cannot access label information from unlabeled data, thus cannot leverage the unlabeled data to help to estimate the anchor points and noise transition matrix.

In this paper, we propose an adaptive learning algorithm that leverages several base models and a vast number of unlabeled data to help to estimate the noise transition matrix and derive an online learning model for weakly labeled data streams. We introduce and reuse several base models trained on each weakly labeled data batch to extract the information from high-throughput streams and adaptively update the learning model to make it respond in real-time. The main idea is to leverage the local classifiers to help to identify the anchor points in the data stream and adaptively estimate the noise transition matrix. Then we align the classifiers trained on weakly labeled data to noise-free distribution and reuse them to derive an online model for weakly labeled data streams. We theoretically analyze the proposed algorithm by expected regret. The empirical studies on synthetic data demonstrate the effectiveness of the proposed approach for online noise transition matrix estimation. Experiments on both benchmark datasets and real-world applications validate our approach's superiority.

We summarize our main contributions as follows.

(1) We investigate the problem of learning with noisy labeled and unlabeled data streams, which accommodates many real-world tasks but is rarely studied in the literature.
(2) We propose a stream learning approach, which adaptively constructs and reuses several models to estimate the noise transition matrix and derive an online classifier for weakly labeled streams. We theoretically justify the effectiveness of our proposed algorithm via regret analysis.
(3) We conduct extensive empirical evaluations on synthetic examples, benchmark datasets, and real-world applications to demonstrate the superiority of the proposed algorithm.

2 RELATED WORKS

We first briefly review some related learning scenarios with our task of learning with weakly labeled data streams.

Offline Learning with Noisy Labels. Learning with noisy labeled data has attracted a lot of attention in recent years. Many statistical consistent algorithms are proposed to address the issue of noisy labels, in which they guarantee the classifier learned from the noisy labeled data to be consistent with the optimal classifier with respect to the data with ground-truth labels [2, 19, 21]. This line of works construct their models based on the prior knowledge of noise transition matrix of or estimating it from sufficient noisy-labeled data first. The estimation of the noise transition matrix is one of the central challenges in the learning tasks with noisy labeled data. A series of assumptions were proposed to estimate the noise transition matrix, e.g., anchor words condition [3], anchor points [19, 30], irreducibility [23]. Specifically, [19] assume there exist some anchor points, i.e., instances belonging to a specific class with probability one. [30] suppose there exist some verified data belonging to the positive class with probability one. [23] make the irreducibility assumption for the noisy labeled data, which says that there exist some anchor points; thus, we can distinguish and recover the noise-free distribution from noisy labeled data. These methods mainly focus on the batch setting, while identifying the anchor points in a data stream and online estimating the noise transition matrix has not yet been well studied.

Besides the statistically consistent approaches, many practical algorithms are proposed to alleviate the negative effect of noisy labels. For example, many approaches are specifically designed to e.g., selecting reliable examples [28], editing labels [18], or adding implicit regularization [12]. Some of them can directly adapt to the learning tasks for data streams. For example, we can filter the data with large loss during the learning procedure [28]. Although these methods obtain satisfactory empirical performance, their general theoretical grantee is unclear.

Label-efficient Online Learning. As the data streams usually contain a vast number of unlabeled data, many studies also exist on training a well-generalized classifier with unlabeled data streams. Some pioneering studies consider online active querying for unlabeled streaming data learning. [5] study label-efficient online learning for prediction with expert advice. They show that actively querying with a constant probability achieves minimax-optimal regret and query complexity for this problem. Subsequent works model the label generation process by a parametric model and obtain regret and query complexity guarantees dependent on the fraction of examples with low margins [1]. Besides the querying mechanism, many semi-supervised learning algorithms are proposed to explore the high-throughput unlabeled data stream with a limited number of labeled data to learn a classifier [10, 27]. However, if the label information is not ground-truth, these algorithms probably converge to an arbitrary result.

Online Learning with Noisy Labels. The conflicts between the fast development of data collecting techniques and the limitation

in human labeling capability pose new challenges to the design of streaming algorithms with noisy labeled data. When the noisy transition matrix is given as prior knowledge, we can rewrite the loss function to recover the risk with respect to the clean labeled data in expectation [21]. With the rewritten loss, we can online update the models and then obtain a statistically consistent classifier for the data streams. However, the noise transition matrix is usually unknown ahead and hardly be estimated accurately before the learning tasks. Some related works make certain assumptions on the noise that the loss is perturbed by a noise distribution with zero-mean and bounded variance, and propose an unbiased estimator for non-linear functions based on oracle querying [6]. Another line of robust boosting is also studied [4, 9]. Based on the assumption that noisy labeled data are with large gradients, a recent work can be adapted to the online learning with noisy labeled data by filtering the instances with large gradient norm [26].

Apart from learning the data stream with weak supervision, researchers also managed to handle even more challenging scenarios where distribution change [31–33], new classes emerge [20, 34] or feature space evolve [14, 15, 29] in the streaming data. To conclude, efficient approaches dealing with weakly supervised streaming data in open environments are very desired in real-world tasks [37].

3 SETTINGS

We first introduce the learning setting of weakly labeled data streams and corresponding notations. In the ordinary supervised multi-class learning, we denote by $\mathcal{D}$ the underlying distribution from which the training data $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ are independently and identically sampled, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, ..., K\}$ is the label space of a multi-class learning task. We assume the streaming data come in the form of mini-batches. At each time $t \in [T]$, we receive a batch of data $\mathbf{B}_t = \{\mathbf{x}_t^1, ..., \mathbf{x}_t^n\}$ of size n. We denote by $\{\mathbf{B}_t\}_{t=1,...,T}$ an stream of size T where each mini-batch comes continuously.

In the setting of noisy labeled and unlabeled data, for each instance $\mathbf{x}_i$, we denote its true label by y_i and the observed noisy label (if available) by $\widetilde{y}_i$. We denote by $\widetilde{\mathbf{B}}_t = \{(\mathbf{x}_t^i, \widetilde{y}_t^i)\}_{i=1}^m \cup \{\mathbf{x}_t^j\}_{j=m+1}^n$ the semi-supervised data batch with m noisy labels. It is the union of noisy labeled data and unlabeled data. The noisy label $\widetilde{y}_t$ is flipped from y_t based on a noise transition matrix M, in which

$$M_{i,j}(\mathbf{x}_t) := \Pr[\widetilde{y}_t = i | y_t = j, \mathbf{x}_t].$$

In this work, we assume the label noise is class-dependent [21], that is, $M(\mathbf{x}_t) = M$ for any $\mathbf{x}_t$ in the feature space. We notice that the noise transition matrix M is *unknown* to the learner.

We consider a family of predictive models $f(\mathbf{w}, \phi(\mathbf{x})) \in \mathbb{R}^K$, where $\mathbf{w} \in \mathcal{W}$ is the parameters, $\phi : \mathcal{X} \mapsto \mathbb{R}^d$ is the transformation function of the original feature $\mathbf{x}$ and $f : \mathcal{W} \times \mathbb{R}^d \mapsto \mathbb{R}^K$ captures the structure of the model. The k-th entry of the model $[f(\mathbf{w}, \phi(\mathbf{x}))]_k$ can been seen as a score for the k-th class and then we determine the final prediction by $\arg\max_{k \in [K]} [f(\mathbf{w}, \phi(\mathbf{x}))]_k$. Such a formulation is general enough to capture various models. For example, we can choose $f(\mathbf{w}, \phi(\mathbf{x})) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle$ to use the linear in parameter model and obtain $\phi(\mathbf{x})$ by the representation learning methods. For notation simplicity, we use $\mathbf{x}$ instead of $\phi(\mathbf{x})$ in the rest paper. The quality of the model is measured by the loss $\ell : \mathbb{R}^K \times \mathcal{Y} \mapsto \mathbb{R}_+$, which can be any convex surrogate loss for classification such as the hingle, squared and logistic loss, etc.

Parameters: Learner's parameter set $\mathcal{W}$, feature space $\mathcal{X}$, loss function ℓ, time horizon T (optional).
At each time instant $t = 1, ..., T$:
1. Observe unlabeled data $\mathbf{B}_t$ and predict by $\mathbf{w}_t \in \mathcal{W}$
2. Receive noisy labels of a small subset in $\mathbf{B}_t$
3. Update the online model from $\mathbf{w}_t$ to $\mathbf{w}_{t+1} \in \mathcal{W}$

Figure 1: Protocol of Learning with Weakly Labeled Streams

Since the data arrives in a high-throughput stream, we require the learning system to run under time and memory constraints because the labels are expected in real-time, and the stream is typically too large to fully store. We formulate our learning protocol of noisy labeled and unlabeled data streams in Figure 1.

4 OUR APPROACH

In this section, we present our adaptive learning approach for weakly labeled data streams. We demonstrate that exploring unlabeled data by several local classifiers plays a significant role in estimating the noise transition matrix, especially when these noisy labeled data in the data stream are scarce.

To deal with noisy labeled and unlabeled data stream, we first rewrite the loss for noisy labeled data to align the models to noise-free distribution, in which we reweight the noisy labeled instance by a noise transition matrix. Then, we proceed to estimate the noise transition matrix by identifying the anchor points in the data streams with the help of several local classifiers and a vast number of unlabeled data. Finally, we provide our adaptive learning algorithm that reuse the local classifiers and leverage the entire stream history to derive an accurate online model.

4.1 Learning with Noisy Labeled Data

We first consider exploring the noisy labeled data in the stream. In the learning scenario with noisy labeled data, if we simply treat all observed data as accurate, both empirical and theoretical performance will suffer heavily from the label noise. In order to obtain a well-generalized classifier on the test data, we aim to recover the loss with respect to the ground-truth labeled data by instance reweighing for noisy labeled data. Following a similar analysis for the class-dependent label noise in [21], we have

$$\ell(f(\mathbf{w}, \mathbf{x}), y) = \mathbb{E}_{\widetilde{y}} \left[\bar{\ell}(f(\mathbf{w}, \mathbf{x}), \widetilde{y}; M) \right],$$

in which the surrogate loss $\bar{\ell}(\cdot, \cdot; M)$ with noise transition matrix M for noisy labeled data is defined as

$$\bar{\ell}(f(\mathbf{w}, \mathbf{x}), \widetilde{y} = i; M) = \sum_{j \in \mathcal{Y}} \ell(f(\mathbf{w}, \mathbf{x}), y = j) \cdot [(M^\top)^{-1}]_{j,i}, \quad (1)$$

where $[(M^\top)^{-1}]_{j,i}$ is the j-th row and i-th column of matrix $(M^\top)^{-1}$.

The instance reweighing technique demonstrates that we recover the loss of ground-truth labeled data with surrogate loss in Eqn. (1) that reweightes the importance for noisy labeled data. We then turn to the scenario of noisy labeled streams. As the data are collected in a stream, we propose a stream learning algorithm with surrogate gradients for noisy labeled data. Let $\ell(f(\mathbf{w}, \mathbf{x}), y)$ be a

convex function with respect to $\mathbf{w}$. Following a similar analysis for the surrogate loss, the gradients $g(\mathbf{w}, \mathbf{x}, \widetilde{y})$ on noisy labeled data is an unbiased estimator of underlying gradients with respect to ground-truth labeled data, that is,

$$\mathbb{E}_{\widetilde{y}}[g(\mathbf{w}, \mathbf{x}, \widetilde{y})] = \partial_{\mathbf{w}} \ell(f(\mathbf{w}, \mathbf{x}), y),$$

where

$$g(\mathbf{w}, \mathbf{x}, \widetilde{y} = i) = \sum_j \partial_{\mathbf{w}} \ell(f(\mathbf{w}, \mathbf{x}), y = j) \cdot [(M^\top)^{-1}]_{j,i}. \quad (2)$$

Now we are ready to propose our online gradient descent algorithm with unbiased gradients $g(\mathbf{w}, \mathbf{x}, \widetilde{y})$, that is,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot g(\mathbf{w}_t, \mathbf{x}_t, \widetilde{y}_t). \quad (3)$$

where $\eta > 0$ is the learning rate.

In the following, we show that the proposed online gradient descent algorithm for noisy-labeled data streams with the updating procedure in (3) satisfies a low regret in expectation on the ground-truth labeled sequences generate from noise-free distribution.

Theorem 1. *Let $\ell(f(\mathbf{w}, \mathbf{x}), y)$ be convex with respect to $\mathbf{w}$. Denoting by $\mathbb{E}_{1:T}[\cdot]$ the expectation taken over the randomness on the drawn of $\{\widetilde{y}_t\}_{t=1}^T$, the update procedure in* (3) *with learning rate $\eta = D/(L_M G\sqrt{T})$ yields,*

$$\mathbb{E}_{1:T}\left[\sum_{t=1}^T \ell(f(\mathbf{w}_t, \mathbf{x}_t), y_t)\right] - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^T \ell(f(\mathbf{w}, \mathbf{x}_t), y_t) \le L_M D G \sqrt{T}$$

for any underlying clean data sequence $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$, where $D = \max_{\mathbf{w}_1,\mathbf{w}_2\in\mathcal{W}} \|\mathbf{w}_1 - \mathbf{w}_2\|_2$ is the diameter of $\mathcal{W}$ and G is an upper bound on the gradient $G = \max_{\mathbf{w}\in\mathcal{W},\mathbf{x}\in\mathcal{X},y\in\mathcal{Y}} \|\partial_{\mathbf{w}} \ell(f(\mathbf{w}, \mathbf{x}), y)\|_2$. The constant $L_M = \max_{i,j} \left|[(M^\top)^{-1}]_{ij}\right|$ is the upper bound on the every entry of the inverse of the noise transition matrix.

Remark 1. *Theorem 1 demonstrates that through the proposed online gradient descent algorithm* (3) *with unbiased gradients in* (2)*, the difference of the accumulated loss of the learned classifier sequences $\{\mathbf{w}_t\}_{t=1}^T$ and the optimal classifier $\mathbf{w}_*$ on the noise-free distribution decreases in the order of $o(T)$. Thus, we possess a statistically consistent classifier. It is necessary to assume that L_M is bounded, otherwise, there are some classes whose label flipping rate is close to* 1*, which means that we cannot recover the information of those extremely noisy examples. Detailed proofs will be provided in the longer version.*

Remark 2. *The optimal tuning of η requires the knowledge of T. There are approaches in the online learning literature to fixed it [13]. Besides, our experiments show that an empirical value of η is good enough to achieve nice performance for all experimental settings.*

The central ingredient to obtain a well-generalized classifier is the approximation of *noise transition matrix* M. This result motivates us to design an online approach for noise transition matrix estimation from weakly labeled streams.

4.2 Online Estimation for Noise Transition Matrix with Unlabeled Data

In this section, we present an online estimation approach for the noise transition matrix. As the noisy labeled data are scarce, we aim to explore a vast number of unlabeled data in the stream to help to estimate the noise transition matrix, along with noisy labeled data.

We assume that the data streams contain some anchor points for each class, which bridges the gap between the noisy labeled data and the underlying noise-free distribution. Anchor points are instances that belong to a certain class with probability one. For example, some samples must belong to a certain class without any ambiguity. Formally, the anchor points are defined as follows.

Definition 1 (Anchor Points). *Given an instance $(\mathbf{x}_i, y_i)$, if $\Pr[Y = y_i|X = \mathbf{x}_i] = 1$, then we call $(\mathbf{x}_i, y_i)$ an anchor point.*

Based on the definition of anchor points, we have the following equation for each anchor point $(\mathbf{x}_i, y_i = c)$ of class c,

$$\Pr[Y = \widetilde{y}_i = j|X = \mathbf{x}_i] = \sum_{c=1}^C M_{c,j} \Pr[Y = y_i = c|X = \mathbf{x}_i] = M_{i,j}.$$

The left hand of the above equation is the class-posterior probabilities of the noisy labels on the anchor points, which can be approximated by the classifier learned on noisy labeled stream with soft-max. Therefore, to estimate the noise transition matrix, we require an online model that directly learns the noisy labeled data and simultaneously identify the anchor points in the data stream.

For the estimation of class-posterior probabilities $\Pr[\widetilde{y}_i = j|X = \mathbf{x}_i]$ on the anchor points, we employ an online model $\widetilde{\mathbf{w}}_t \in \mathcal{W}$ with a semi-supervised learning algorithm that is directly trained from the noisy labeled data and unlabeled data such as the Pseudo-labeling algorithm [16]. The challenge arises in the online identification of anchor points. Previous approaches typically train a classifier on sufficient noisy labeled data and set the most high-confidence instances as anchor points [7, 19]. However, when most data are unlabeled and come as a stream, adapting the offline mechanism to the data stream can be inaccurate for anchor points identification. It is also not feasible to first collect sufficient noisy labeled data to train a model and then identify the anchor points because we require the stream learning system to online update and respond in real-time. In the following, we propose a ensemble-based online approach of anchor points identification with data streams.

We would like to introduce the following guiding principle for online anchor points identification, that is,

Assumption 1 (pseudo-consistent). *For an instance $\mathbf{x}_t \in \mathcal{X}$, if several unbiased models that trained on their weakly supervised data make the same prediction (pseudo label) for this instance with high confidence, then we consider it as an anchor point.*

Assumption 1 is specially designed for identifying anchor points in the weakly labeled data stream. We maintain several classifiers trained with surrogate loss in (1) and a handful of candidate anchor points for selection. We consider that the anchor points should obtain the same prediction with high confidence from diverse weak classifiers as they are non-ambiguous.

Based on assumption 1, we propose the Anchor Points Identification (API) algorithm for the streaming data and *adaptively* estimate the noise transition matrix with the help of unlabeled data. Suppose the transition matrix in last round is available, we then train a new classifier for the next batch with the surrogate loss in (1) and update the transition matrix. For each data batch $\widetilde{\mathbf{B}}_t$, we train a local classifier $\mathbf{w}_t^l \in \mathcal{W}$ (the superscript l indicates that the model is trained by this local batch) based on previous estimated noise

Algorithm 1 Anchor Points Identification (API)

1: Maintain a model pool of size K and a data pool of candidate anchor points $\mathbb{A}$, initial M^0 and confidence threshold σ
2: **for** $t = 1, ..., T$ **do**
3: Receive $\widetilde{\mathbf{B}}_t$ and $\widetilde{\mathbf{w}}_t$
4: Train a local classifier $\mathbf{w}_t^l$ from $\widetilde{\mathbf{B}}_t$ by (4)
5: Add $\mathbf{w}_t^l$ to model pool and data with high confidence larger than σ in $\widetilde{\mathbf{B}}_t$ considered by $\mathbf{w}_t^l$ to the anchor points set $\mathbb{A}$
6: Find anchor points in the candidate set by pseudo-consistent check and output the noise transition matrix M^t by (5)
7: Return noise transition matrix M^t and local classifiers
8: **end for**

Algorithm 2 Adaptive Learning for Weakly Labeled Streams

1: Choose learning rate $\eta \geq 0$, representation function $\phi(\cdot)$, let $\mathbf{w}_1$ and $\widetilde{\mathbf{w}}_1$ be any point in $\mathcal{W}$
2: **for** $t = 1, ..., T$ **do**
3: Receive $\mathbf{B}_t$ and output predictions
4: Receive noisy labels and formulate $\widetilde{\mathbf{B}}_t$
5: Update $\widetilde{\mathbf{w}}_{t-1}$ to $\widetilde{\mathbf{w}}_t$ by a semi-supervised learning algorithm directly trained on noisy labeled data and unlabeled data
6: Sent $\widetilde{\mathbf{B}}_t$ and $\widetilde{\mathbf{w}}_t$ to the API Algorithm 1
7: Query API Algorithm 1 for M^t and local classifiers
8: Obtain $\mathbf{w}_{t+1}$ by (6) and (7)
9: **end for**

transition matrix M^{t-1} and current batch of data, that is,

$$\mathbf{w}_t^l = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{m} \bar{\ell}(f(\mathbf{w}, \mathbf{x}_t^i), \widetilde{y}_t^i; M^{t-1}) + \Omega(\mathbf{w}, \widetilde{\mathbf{B}}_t). \tag{4}$$

where $\Omega(\mathbf{w}, \widetilde{\mathbf{B}}_t)$ is the regularizer of model on the unlabeled data. Any semi-supervised learning algorithm, e.g., pseudo-labeling, can train the local classifier with surrogate loss in (1).

We maintain a model pool that contains the most recent local classifiers trained on each data batch, and a candidate anchor points set $\mathbb{A}$. We initial the model and data pool as an empty set and initial the noise transition matrix M^0 as an identity matrix. At each time t, we denote by the model pool as $\{\mathbf{w}_{t-K+1}^l, ..., \mathbf{w}_t^l\}$ of size K. After obtaining a new local classifier $\mathbf{w}_t^l$, we insert it into the model pool and drop the most previous one.

Based on $\{\mathbf{w}_\tau^l\}_{\tau=t-K}^t$, we then update the anchor points set. We first find the data in $\widetilde{\mathbf{B}}_t$ with high confidence (larger than threshold σ) considered by $\mathbf{w}_\tau^l$, and insert them to the anchor points set $\mathbb{A}$. Then we perform the pseudo-consistent check to update the anchor points set. Specifically, we reserve the data points with the sample pseudo label in the anchor points set, and drop those data in the candidate set that do not meet this assumption.

After updating the anchor points set $\mathbb{A}$, we then calculate the noise transition matrix M_t. For each class with anchor points $\{(\mathbf{x}_i, y_i = c)\}_{i=1,...,n_c}$, we estimate the noise transition matrix by

$$M_{c,j}^t = \frac{1}{n_c} \sum_{i=1}^{n_c} \Pr[\widetilde{y} = j | X = \mathbf{x}_i], \tag{5}$$

in which the noisy class-posterior probability is approximated by $\widetilde{\mathbf{w}}_t$ with soft-max, which is directly trained on the noisy labeled and unlabeled data. We summarize the API algorithm in Algorithm 1.

4.3 Adaptive Learning via Model Reuse

In this part, we introduce the proposed Adaptive learning algorithm for weakly labeled Streams (AdaStreams). We aim to leverage the entire stream history and reuse local models to derive an online model that labels a new data in real-time, only storing a tiny fraction of data as anchor points and several recent local models.

We adaptively estimate the noise transition matrix and update the online model with weakly labeled data simultaneously. At each time, after we make the prediction and receive the noisy labeled data $\widetilde{\mathbf{B}}_t$, we update $\widetilde{\mathbf{w}}_t$ by a semi-supervised learning algorithm with noisy labeled data and unlabeled data to approximate the noisy class-posterior probabilities. Then we send $\widetilde{\mathbf{w}}_t$ and $\widetilde{\mathbf{B}}_t$ to Algorithm 1 to estimate the noise transition matrix M^t and update the recent local classifiers $\{\mathbf{w}_\tau^l\}_{\tau=t-K,...,t}$. Finally, enlightened by the idea of online learning with predictable sequence [22], we propose the following updating procedure, to reuse the recent local classifiers learned from unlabeled data with weighted decay and the online model that learned on the noisy labeled data to derive an accurate one,

$$\widehat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \eta\langle \mathbf{w}, \frac{1}{m}\sum_{i=1}^{m} g(\mathbf{w}_t, \mathbf{x}_t^i, \widetilde{y}_t^i)\rangle + \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}_t\|_2^2 \tag{6}$$

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \eta\langle \mathbf{w}, \frac{1}{m}\sum_{i=1}^{m}\sum_{\tau=t-K}^{t} \alpha_\tau \cdot g(\mathbf{w}_\tau^l, \mathbf{x}_t^i, \widetilde{y}_t^i)\rangle + \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}_{t+1}\|_2^2 \tag{7}$$

where the unbiased gradients $g(\cdot,\cdot,\cdot)$ is parametrized by M^t and $\{\alpha_\tau\}_{\tau=t-K,...,t}$ with $\alpha_\tau \geq 0$ is the weight sequence decays in exponential manner. The gradients is estimated based on all labeled data each round. We summarize the algorithm in Algorithm 2.

We propose the following theorem to show that, without storing the entire stream, the proposed algorithm simultaneously exploits the noisy labeled and unlabeled data, and derive a statistically consistent classifier with respect to the noise-free distribution.

Theorem 2. *Under the same assumptions as Theorem 1 and setting the learning rate* $\eta = D/\sqrt{1 + \sum_{t=1}^{T} \|g_t - g'_{t-1}\|_2^2}$, *the update procedure in Eqn.* (6) *and* (7) *yields,*

$$\mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \ell(f(\mathbf{w}_t, \mathbf{x}_t), y_t)\right] - \min_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \ell(f(\mathbf{w}, \mathbf{x}_t), y_t)$$
$$\leq D\sqrt{1 + \mathbb{E}_{1:T}[\sum_{t=1}^{T} \|g_t - g'_{t-1}\|_2^2]}$$

for any underlying clean data sequence $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$. *The notations* $g_t = \frac{1}{m}\sum_{i=1}^m g(\mathbf{w}_t, \mathbf{x}_t^i, \widetilde{y}_t^i)$ *and* $g'_t = \frac{1}{m}\sum_{i=1}^m \sum_\tau \alpha_\tau \cdot g(\mathbf{w}_\tau^l, \mathbf{x}_t^i, \widetilde{y}_t^i)$.

Remark 3. *Theorem 2 shows that the expected regret can be bounded by the sum of gradients gap* $\|g_t - g'_{t-1}\|_2^2$ *between the online model and the weighted reuse of local unbiased classifiers. The local classifier is aligned and reused with unbiased loss and unlabeled data, which serves as regularization during the learning process. When the unlabeled data are sufficient, and the data stream is rather stable,*

recent local classifiers trained on unlabeled data can provide a good approximation for the next gradient, minimizing the regret's upper bound. We notice that this exploration of unlabeled data is safe, in that for the worst-case scenario, we recover back to the $O(\sqrt{T})$ regret in Theorem 1. Although our theoretical results require an accurate estimation of the noise transition matrix, we demonstrate in the experiments that our proposed adaptive estimation is empirically successful. Detailed proofs will be provided in the longer version.

The proposed update mechanism makes the whole learning system be online update, labeling new data in real-time, and robust to weakly-supervised data with only a small storage cost.

5 EXPERIMENTS

In this section, we examine the performance of the proposed algorithm on synthetic, benchmark datasets and real-world applications. Specifically, we aim to answer the following questions:

- **Q1:** Does the proposed approach approximate the optimal classifier on the noise-free distribution by learning with weakly labeled data streams? Does it find the anchor points in the data stream and correctly estimate the noise transition matrix under different types of label noise?
- **Q2:** Does the proposed algorithm outperform other contender approaches in various benchmark applications?
- **Q3:** Does the proposed algorithm show effectiveness on a real-world task with complex and unknown label noise?

5.1 Experimental Settings

In this part, we describe the datasets, the label noise simulation process, and the data stream construction details.

Benchmark datasets. We conduct experiments on 8 benchmark datasets from various real-world applications, including flight delay prediction[1] (Weather), electricity price change prediction (Electricity), spam email recognition (Spam), RFID location detection[2] (RFID), human activity recognition (HAR, HHAR, WISDM-AR) and online gender recognition[3] (Portraits). Unless otherwise noted, other datasets can be found in the UCI repository[4]. The data stream length varies from 940 to 54,872, and the class number varies from 2 to 6. Some of the datasets have a slight distribution change issue such as the Weather and Portraits datasets. We summarize the brief statistics of benchmark datasets in Table 1.

Label noise types. Since the labels in these datasets are ground-truth, we simulate various types of label noise by corrupting the labels of training and validation set with two types of label noises, which accommodates various real-world label noise:

- **Uniform noise:** The labels have a noise rate of p to be uniformly flipped to other classes. This type of label noise simulates the random annotation noise.
- **Pair noise:** Labelers are assumed to make mistakes only within the most similar pair classes. More specifically, labels have a noise rate of p to flip to a random pair class. This type of label noise simulates two confusing classes.

[1]http://users.rowan.edu/~polikar/nse.html
[2]http://www.lamda.nju.edu.cn/data_RFID.ashx
[3]http://people.eecs.berkeley.edu/~shiry/projects/yearbooks/yearbooks.html
[4]https://archive.ics.uci.edu/ml/datasets.php

Table 1: Brief statistics of benchmark datasets

Dataset	# Length	# Dim	Dataset	# Length	# Dim
Weather	18,159	8	RFID	940	150
Electricity	45,312	8	WISDM-AR	1,207	315
HAR	1,607	128	Spam	9,324	500
HHAR	54,872	128	Portraits	37,921	3,072

Real-world dataset with unknown label noise. We also conduct the experiments on a real-world animal recognition dataset ANIMAL-10N with human annotation noise[5] [24]. This dataset contains confusing animals with a total of 55,000 images. The images are crawled from several online search engines, including Bing and Google with the noisy labels. The label noise arise in annotation error with an unknown noise rate. 5,000 images are checked with ground-truth labels to test the algorithm's performance.

Data stream construction. Different from the offline learning scenario, we perform the algorithm on data streams. For each benchmark data stream, we consider they are coming as mini-batches of size 100, with 10% of them being noisy labeled with uniform or pair label noise, 80% of them being unlabeled data and 10% of them being test data. For the synthetic example and ANIMAL-10N dataset, these data are non-temporal, we simulate a stream by generating random permutations of the data points or the images in mini-batches and *one-passly* perform the algorithm. We conduct all the experiments for 5 trials and use the overall mean and standard deviation of predictive accuracy as measurement, which is the ratio between the number of correct predictions and the stream length.

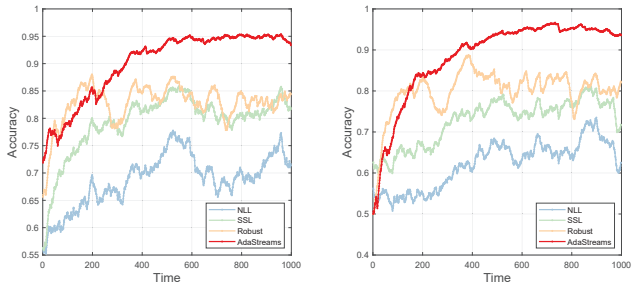
5.2 Synthetic Example

In this part, we aim to answer **Q1**. We use a synthetic example as an intuitive illustration of the advantage of our proposed algorithm. We generate a synthetic classification dataset of size 1,000 with four classes. We generate 250 data from a Gaussian distribution with different mean vectors for each class. We set 10% data as labeled data, 10% data as test data, while the remaining data as unlabeled. To simulate the noisy labels in the data stream, we flip their ground-truth labels by both uniform label noise and pair label noise with a noise rate of $p = 0.4$. We use a two-layer neural network as the classifier to perform the stream learning task.

Approximate the optimal classifier. We first compare the proposed AdaStreams algorithm with three comparators. Specifically, the first baseline method is the noisy label learning (NLL) method that considers the noisy labeled data as accurate ones and directly performs an online gradient descent algorithm. The second baseline method is a semi-supervised learning method (SSL) [16] that ignores the issue of label noise but it explores the unlabeled data. We further compare the proposed approach with a robust noisy label learning approach Co-teaching+ [28] (Robust) that filters the large loss data to obtain a robust classifier.

We report the accuracy curves of our proposed algorithm and other approaches under symmetric and pair label noise in Figure 2. The results show that the proposed AdaStreams algorithm converges to the optimal classifier on the noise-free distribution. The NLL and SSL methods that consider the noisy labeled data as correct

[5]https://dm.kaist.ac.kr/datasets/animal-10n/



(a) Accuracy curve with symmetric noise (b) Accuracy curve with pair noise

Figure 2: Accuracy curve under different types of label noise.

ones fail the learning task. The label noise robust algorithm Robust shows a relatively good performance; even initially, its performance is comparable with our proposed AdaStreams algorithm. However, the performance of the Robust algorithm declines later and is not stable, while our proposed algorithm maintains a good performance over time. This phenomenon indicates that instances with large loss are not always noisy labeled data. The filter mechanism based on loss value can lose important information during stream learning.

Find the anchor points. We then compare the proposed online anchor points identification method with the offline method that considers the data with the highest confidence as the anchor points [7, 19]. We denote this approach as Highest-Confidence (HC). The only difference between the HC approach and the AdaStreams algorithm is their identified anchor points set. The HC approach considers the instances assigned with the highest confidence by $\widetilde{\mathbf{w}}_t$ trained on the noisy labeled data as anchor points. Both of the HC and AdaStreams approaches estimate the noise transition matrix based on the class-posterior probability approximated by $\widetilde{\mathbf{w}}_t$.

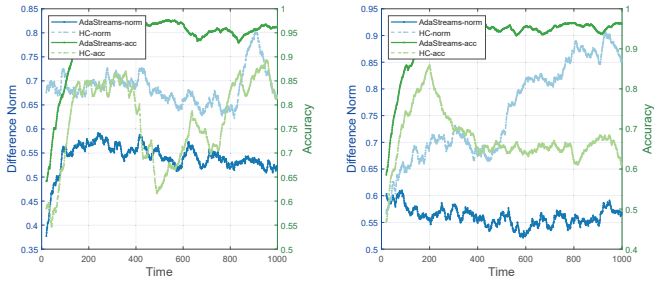
We introduce the L_2 norm of the difference between the estimated noise transition matrix and the true one as a measure. At each time, we calculate $\|M - M^t\|_2$ to measure the correctness of noise transition matrix estimation. The lower $\|M - M^t\|_2$ is, the better we recover the true noise transition matrix based on the proposed online anchor points identification method.

We plot the accuracy curves and the norm of the noise transition matrix difference in Figure 3. We can see that our proposed algorithm achieves a lower difference norm $\|M - M^t\|_2$ over time and obtain better and more stable performance. This result is because we better identify the anchor points than simply adapt offline methods to the stream case. We can further observe that the difference norm $\|M - M^t\|_2$ of the HC approach increases with time under both symmetric and pair label noise. This phenomenon indicates that we cannot directly adapt approaches for anchor points identification in the offline scenario to the streaming setting. The performance will decrease over time with a poor estimation of anchor points.

To conclude, the empirical studies on synthetic data show that the proposed AdaStreams algorithm approximates the optimal classifier in the noise-free distribution. Moreover, the proposed API mechanism successfully estimates the noise transition matrix.

5.3 Comparisons on Benchmark Datasets

In this part, we aim to answer **Q2**. We compare the proposed approach with other state-of-the-art methods on 8 benchmark datasets



(a) $\|M - M^t\|_2$ at each time stamp and accuracy curve under symmetric noise (b) $\|M - M^t\|_2$ at each time stamp and accuracy curve under pair noise

Figure 3: Difference norm $\|M - M^t\|_2$ and accuracy curve under different types of label noise. The lower $\|M - M^t\|_2$ is, the better we approximate the true noise transition matrix.

under uniform noise and pair noise, respectively. For each dataset, we simulate the uniform label noise $p = 0.4$ and pair label noise $p = 0.4$. We randomly take 10% data as noisy labeled ones in the data stream while considering the remaining data as unlabeled.

We compare the proposed algorithm with six contenders, including a noisy label learning method, a semi-supervised learning method, and four label noise robust semi-supervised stream learning algorithms. The two baseline methods are:

- **S-CoT+** [28]: Co-teaching+ is a SGD-based algorithm for learning with noisy labels. We adapt it to the streaming setting, named S-CoT+, in which we perform co-teaching+ at each data batch with the last classifier as initialization.
- **TLP** [27]: it is a graph-based semi-supervised learning method designed for data streams. TLP maintains a small synopsis of stream that can be quickly updated with new examples.

We also compare the proposed AdaStreams with four label noise robust semi-supervised stream learning algorithms, that is,

- **SIIS in Last Round** [11]: SIIS is a graph-based SSL algorithm. It emphasizes the leading eigenvectors of the Laplacian matrix associated with small eigenvalues, such that this method constructs a label noise robust graph and propagates labels on this graph. We perform the SIIS algorithm in the last data batch and use it to predict the next batch.
- **SIIS-Ensemble** [11]: This is an ensemble version of SIIS approaches for data streams. For a fair comparison, we maintain K SIIS models in the most recent batches and predict the next data batch by majority voting.
- **S-PL-CoT+** [16, 28]: To exploit the unlabeled data in the data stream, we first perform the Pseudo-labeling algorithm to assign each unlabeled data a pseudo label, then consider them as noisy labels and then perform S-CoT+ algorithm.
- **DIVIDEMIX** [17]: This is a noisy labeled learning algorithm that considers the high confidence data as clean ones and takes the remaining data as unlabeled. We add the unlabeled data in their proposed framework.

We also provide a comparison algorithm that estimates the noise transition matrix by the classifier trained on noisy labeled data, to test the superiority of our ensemble-based mechanism of alleviating the label noise in the data stream.

Table 2: Performance comparisons on benchmark datasets. On each dataset, 5 test runs were conducted and the average accuracy as well as standard deviation are presented, with the best one emphasized in bold.

(a) Symmetric label noise with $p = 0.4$

Dataset	S-CoT+	TLP	SIIS-L	SIIS-E	S-PL-CoT+	DIVIDEMIX	Ada-HC	AdaStreams
Weather	58.04 ± 2.14	59.24 ± 1.42	61.65 ± 2.18	63.51 ± 0.73	58.87 ± 2.24	**65.56 ± 1.75**	50.50 ± 5.85	65.18 ± 1.75
Electricity	54.80 ± 5.43	56.21 ± 1.52	54.54 ± 2.17	52.16 ± 3.13	55.05 ± 5.59	57.25 ± 0.74	52.41 ± 6.23	**57.30 ± 0.29**
HAR	61.60 ± 2.83	74.55 ± 4.01	76.03 ± 1.01	75.51 ± 1.32	60.33 ± 3.47	74.39 ± 0.88	79.21 ± 0.96	**79.43 ± 1.25**
HHAR	52.11 ± 1.19	54.33 ± 3.23	54.32 ± 2.25	**56.67 ± 0.64**	50.93 ± 1.35	55.31 ± 1.06	55.52 ± 2.03	56.18 ± 1.81
RFID	54.70 ± 1.48	73.03 ± 2.33	66.06 ± 0.73	69.05 ± 0.72	54.68 ± 1.42	64.57 ± 0.76	74.25 ± 2.44	**75.41 ± 1.99**
WISDM-AR	57.74 ± 4.47	67.37 ± 4.82	73.09 ± 2.15	74.41 ± 1.38	56.54 ± 6.06	73.63 ± 1.19	74.21 ± 1.79	**74.71 ± 0.78**
Spam	83.54 ± 1.79	87.27 ± 3.32	81.72 ± 1.42	82.37 ± 1.21	84.22 ± 1.87	76.90 ± 6.53	90.09 ± 0.64	**90.13 ± 0.44**
Portraits	67.32 ± 0.14	66.61 ± 1.23	–	–	66.77 ± 0.64	67.24 ± 0.57	72.36 ± 0.75	**75.01 ± 0.42**

(b) Pair label noise with $p = 0.4$

Dataset	S-CoT+	TLP	SIIS-L	SIIS-E	S-PL-CoT+	DIVIDEMIX	Ada-HC	AdaStreams
Weather	55.36 ± 6.96	58.42 ± 2.57	66.14 ± 2.40	66.08 ± 2.51	57.70 ± 5.53	66.19 ± 2.00	66.96 ± 3.57	**67.77 ± 1.45**
Electricity	52.64 ± 6.52	55.81 ± 4.09	54.64 ± 0.74	55.85 ± 1.82	54.72 ± 5.21	57.08 ± 1.49	53.69 ± 4.98	**57.87 ± 0.51**
HAR	59.34 ± 5.41	58.61 ± 1.78	81.96 ± 1.24	82.04 ± 2.65	60.57 ± 7.78	**82.68 ± 2.40**	80.73 ± 1.25	78.48 ± 1.54
HHAR	52.80 ± 1.62	54.89 ± 2.79	55.03 ± 1.02	55.12 ± 1.32	51.34 ± 1.79	55.16 ± 1.75	55.46 ± 1.40	**55.75 ± 2.75**
RFID	55.13 ± 2.27	66.80 ± 5.36	69.45 ± 1.83	70.95 ± 1.66	54.80 ± 1.77	65.17 ± 1.03	71.89 ± 3.14	**72.41 ± 1.31**
WISDM-AR	56.84 ± 4.12	67.12 ± 4.51	71.48 ± 3.27	72.68 ± 2.65	53.60 ± 4.67	72.25 ± 2.25	72.91 ± 1.34	**73.73 ± 2.23**
Spam	83.50 ± 1.24	88.27 ± 2.12	86.04 ± 2.41	87.14 ± 2.40	84.50 ± 1.17	78.95 ± 5.40	90.02 ± 0.52	**90.36 ± 0.56**
Portraits	65.61 ± 0.36	66.50 ± 1.12	–	–	66.91 ± 0.45	67.91 ± 0.28	71.85 ± 0.46	**74.47 ± 0.39**

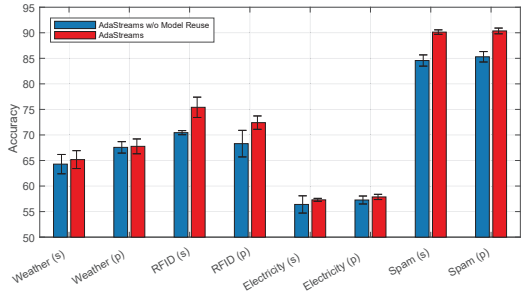
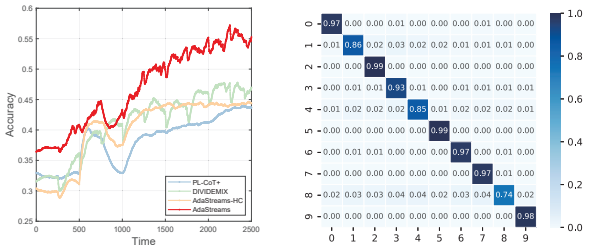


Figure 4: Performance comparisons on benchmark datasets. '(s)' denotes symmetric noise and '(p)' denotes pair noise.



(a) Accuracy curve with contenders. (b) Estimated noise transition matrix.

Figure 5: 5(a) is the accuracy comparison on the ANIMAL-10N dataset. 5(b) is the estimated noise transition matrix by the AdaStreams algorithm. The estimated average noise rate is 7.7% and the real noise rate is about 8% [24].

- **Ada-HC**: This method adapts previous noise transition matrix estimation methods [7, 19] to stream setting, this comparator estimates the anchor points by the most confidence data considered by the classier on noisy labels.

Overall results. We report the comparison results with state-of-the-art contenders in Table 2. The results show that the proposed algorithm successfully addresses the stream learning task and outperforms other approaches. Overall, the AdaStreams outperforms both robust noisy label learning baselines and robust semi-supervised learning methods. Compared with two baseline algorithms (S-CoT+ and TLP), AdaStreams achieves higher accuracy and better stability, indicating the need of aligning the learning model from weakly labeled data to noise-free distribution. Compared with four robust semi-supervised stream learning approaches, the AdaStreams algorithm achieves a very promising performance, because it estimates the noise transition matrix and obtains a statistically consistent classifier. Compared with the offline mechanism that considers the high confidence data as anchor points (Ada-HC), the AdaStreams attains higher accuracy on almost all datasets, which shows the superiority of the ensemble-based anchor points identification mechanism.

Effectiveness of model reuse. In the proposed AdaStreams algorithm, we explore the unlabeled data in the high-throughput stream by reusing the previous local classifiers as regularization. We report the average accuracy comparison on four benchmark datasets with two types of label noise of the AdaStreams with and without model reuse in Figure 4. We observe that reusing the models that explore the unlabeled data improves the overall performance under both symmetric and pair label noise scenarios. This results show the effectiveness of model reuse to explore the unlabeled data in a high-throughput stream through local classifiers on batches.

5.4 Real-world Application

We answer **Q3** in this part. We conduct the experiments on a real-world animal recognition dataset ANIMAL-10N with unknown annotation noise. We report the accuracy comparison with other state-of-the-art contenders in Figure 5(a). We observe that the proposed AdaStreams has a very promising performance compared with the PL-CoT+ and DIVIDEMIX methods. AdaStreams-HC adapt

offline noise transition estimation method by considering the high confidence data as anchor points, which achieves similar performance with two robust online semi-supervised learning methods. The AdaStreams algorithm shows a better performance, implying the superiority of our online anchor points identification method.

We also report the estimated noise transition matrix in Figure 5(b). Previous empirical studies show that the label noise rate in the ANIMAL-10N dataset is about 8% [24]. As shown in Figure 5(b), the estimated average noise rate is about 7.7%, which is very similar to the underlying noise rate in this real-world task.

6 CONCLUSION

In this paper, we study the problem of learning from weakly labeled data streams. We design a novel ensemble-based approach to explore the unlabeled data to identify the anchor points in the data stream and adaptively estimate the noise transition matrix. Based on that, we propose a stream learning algorithm that simultaneously exploits the noisy labeled data and unlabeled data via model reuse, and obtain a statistically consistent classifier on noise-free distribution. Our proposed AdaStreams algorithm is equipped with nice guarantees: by expected regret analysis, we theoretically justify the usefulness of unlabeled data and demonstrate that the proposed algorithm satisfies a low regret in expectation. We conduct extensive experiments on synthetic, benchmark datasets as well as a real-world application, demonstrating the superiority and robustness of the proposed algorithm.

7 ACKNOWLEDGMENTS

This research was supported by NSFC (62176117, 61921006). Yu-Jie Zhang is now at the University of Tokyo. The authors thank Peng Zhao and Long-Fei Li for insightful discussions. We are also grateful to the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1220–1228, 2013.
[2] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2012.
[4] Joseph K Bradley and Robert E Schapire. Filterboost: Regression and classification on large datasets. *Advances in Neural Information Processing Systems*, 20, 2007.
[5] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.
[6] Nicolo Cesa-Bianchi, Shai Shalev Shwartz, and Ohad Shamir. Online learning of noisy data with kernels. In *Proceedings of the 23rd Annual Conference Computational of Learning Theory (COLT)*, pages 218–230, 2010.
[7] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1789–1799, 2020.
[8] Fang Chu, Yizhou Wang, and Carlo Zaniolo. An adaptive learning approach for noisy data streams. In *Proceedings of the 4th International Conference on Data Mining (ICDM)*, pages 351–354, 2004.
[9] Yoav Freund. A more robust boosting algorithm. *arXiv preprint arXiv:0905.2138*, 2009.
[10] Andrew B Goldberg, Ming Li, and Xiaojin Zhu. Online manifold regularization: A new learning setting and empirical study. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 393–407, 2008.
[11] Chen Gong, Hengmin Zhang, Jian Yang, and Dacheng Tao. Learning with inadequate and incorrect supervision. In *Proceedings of the 17th International Conference on Data Mining (ICDM)*, pages 889–894, 2017.
[12] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor W Tsang, Ya Zhang, and Masashi Sugiyama. Masking: a new perspective of noisy supervision. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 5841–5851, 2018.
[13] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.
[14] Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1416–1426, 2017.
[15] Chenping Hou and Zhi-Hua Zhou. One-pass learning with incremental and decremental features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2776–2792, 2018.
[16] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning*, volume 3, page 896, 2013.
[17] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.
[18] Ming Li and Zhi-Hua Zhou. Setred: Self-training with editing. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 611–621, 2005.
[19] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.
[20] Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618, 2017.
[21] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pages 1196–1204, 2013.
[22] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Conference On Learning Theory (COLT)*, volume 30, pages 993–1019, 2013.
[23] Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2052–2060, 2016.
[24] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5907–5915, 2019.
[25] W Nick Street and YongSeog Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 377–382, 2001.
[26] Tim van Erven, Sarah Sachs, Wouter M Koolen, and Wojciech Kotlowski. Robust online convex optimization in the presence of outliers. In *roceedings of the 34th Annual Conference Computational of Learning Theory (COLT)*, pages 4174–4194, 2021.
[27] Tal Wagner, Sudipto Guha, Shiva Kasiviswanathan, and Nina Mishra. Semi-supervised learning on data streams via temporal label propagation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5095–5104, 2018.
[28] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7164–7173, 2019.
[29] Zhen-Yu 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)*, pages 11317–11327, 2020.
[30] Zhen-Yu Zhang, Peng Zhao, Yuan Jiang, and Zhi-Hua Zhou. Learning from incomplete and inaccurate supervision. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[31] Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1 – 45, 2021.
[32] Peng Zhao, Xinqiang Wang, Siyu Xie, Lei Guo, and Zhi-Hua Zhou. Distribution-free one-pass learning. *IEEE Transaction on Knowledge and Data Engineering*, 33:951 – 963, 2021.
[33] Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 12510–12520, 2020.
[34] Peng Zhao, Yu-Jie Zhang, and Zhi-Hua Zhou. Exploratory machine learning with unknown unknowns. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 10999–11006, 2021.
[35] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.
[36] Zhi-Hua Zhou. *Machine Learning*. Springer Nature Singapore, 2021.
[37] Zhi-Hua Zhou. Open-environment machine learning. *arXiv preprint arXiv:2206.00423v2*, 2022.