Online Semi-supervised Learning with Mix-Typed Streaming Features

Di Wu[1*], Shengda Zhuo[2*], Yu Wang[2], Zhong Chen[3], Yi He[4†]

[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China
[2]Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou 510006, China
[3]Department of Computer Science, Xavier University of Louisiana, New Orleans, LA 70125, USA
[4]Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA
wudi.cigit@gmail.com, zhuosd96@gmail.com, yuwang@gzhu.edu.cn, zchen@xula.edu, yihe@cs.odu.edu

Abstract

Online learning with feature spaces that are not fixed but can vary over time renders a seemingly flexible learning paradigm thus has drawn much attention. Unfortunately, two restrictions prohibit a ubiquitous application of this learning paradigm in practice. First, whereas prior studies mainly assume a homogenous feature type, data streams generated from real applications can be heterogeneous in which Boolean, ordinal, and continuous co-exist. Existing methods that prescribe parametric distributions such as Gaussians would not suffice to model the correlation among such *mix-typed* features. Second, while full supervision seems to be a default setup, providing labels to all arriving data instances over a long time span is tangibly onerous, laborious, and economically unsustainable. Alas, a semi-supervised online learner that can deal with mix-typed, varying feature spaces is still missing. To fill the gap, this paper explores a novel problem, named *Online Semi-supervised Learning with Mix-typed streaming Features* (OSLMF), which strives to relax the restrictions on the feature type and supervision information. Our key idea to solve the new problem is to leverage copula model to align the data instances with different feature spaces so as to make their distance measurable. A geometric structure underlying data instances is then established in an online fashion based on their distances, through which the limited labeling information is propagated, from the scarce labeled instances to their close neighbors. Experimental results are documented to evidence the viability and effectiveness of our proposed approach. Code is released in https://github.com/wudi1989/OSLMF.

Introduction

Online learning from doubly-streaming inputs is a new paradigm for data stream analytics that thrives very recently (Zhang et al. 2016; Hou, Zhang, and Zhou 2017; Beyazit, Alagurajah, and Wu 2019; Zhang et al. 2020; He et al. 2021a,b; Wu et al. 2021; Lian et al. 2022; Chen et al. 2022; Wu 2023). Unlike traditional online learning that can deal with data streams residing in a *fixed* feature space only (Aggarwal 2007; Shalev-Shwartz et al. 2011), this new learning paradigm strives to build incremental models with respect to both streaming data and *streaming features*. This allows a more flexible learning environment in which new features can emerge and join the model training process arbitrarily, and pre-existing features may become unobservable or vanish from model during various time spans.

Invited by this flexibility, various domain applications start to model their data in a doubly-streaming format. Consider, for example, a crowd-sensing application, where the mobile users commit their data collectively to train an incremental model that detects air pollution in local areas (Meng et al. 2017; Pan et al. 2017; Schreckenberger et al. 2020). The doubly-streaming property is manifested from the crowd-sensed data streams – the new users joining the sensing effort with upgraded or totally new devices (e.g., cellphones, sensor kits) will lead to new features, while any users who leave (or that some devices fail to commit due to network issues) can incur feature unobservability. To learn from such data streams, a common practice shared by prior studies is to establish the correlation among features, such that the incremental model can: 1) initialize the learning coefficients of any new features using educated guess, expediting convergence with a jump-start when these new features are not described by sufficient data instances, and 2) enjoy a reconstructed information of the unobserved features, leveraging their learned coefficients to improve the prediction performance via online ensembling.

Despite their triumphs, most existing studies are limited by two assumptions. First, the incremental model is trained under *full supervision*, which means that every arriving data instance must be accompanied with a class label. Unfortunately, annotating labels is in general prohibitive, due to the limited manpower and time stretched by the large volume and high velocity of data streams. Second, all features flowing into the model are prescribed to share the same data type, which is often violated in real applications. For example, the features captured by various types of sensor devices are naturally in different data types including Boolean (e.g., rainy or not), ordinal (e.g., $PM_{2.5}$ levels), and continuous (e.g., outdoor temperature). Establishing correlation among such *mix-typed* features is tangibly challenging and cannot be viable by the prior online parametric models that assume, e.g., Gaussian correlation matrices (Agarwal, Chen, and Elango 2010; Balzano, Chi, and Lu 2018; He et al. 2019).

Motivated by this situation, we explore a new learn-

[*]These authors contributed equally.
[†]Corresponding author: Dr. Yi He (yihe@cs.odu.edu)

ing problem, termed *Online Semi-supervised Learning with Mix-typed streaming Features* (OSLMF), striving to make the doubly-streaming data analytics more flexible and applicable. Two challenges are coined in this new problem and shall be solved together: 1) Labels of any arriving data instances are given only *occasionally*, and 2) The feature space describing data is unbounded and varies over time, in which Boolean, ordinal, and continuous features co-exist.

Our main idea to solve the OSLMF problem is to deduce a latent space from the mix-typed streaming features, through respecting the inherent geometric structure among data instances that delivers a reasonably high discriminant power. To realize it, our approach consists of two key build blocks: 1) a copula model that captures data generative marginals from a set of latent and continuous probability densities and forms their correlations, and 2) an online density-peak clustering model that probes the geometric relations underlying data instances so as to propagate the supervision information from the scarcely labeled instances to their neighbors.

Specific contributions of this paper are as follows:

i) This is the first study to explore the online learning problem with mix-typed streaming features and semi-supervision, in which two challenges, namely how the mixed data types and scarce labels can negatively affect the online learning efficacy, are investigated.

ii) A novel algorithm to resolve the new OSLMF problem with copula modeling and density-peak clustering is proposed and elaborated.

iii) Extensive experiments on 14 benchmark datasets are conducted to evidence the viability, effectiveness, and superiority of our proposed algorithm.

Related Work

We relate our OSLMF problem to two research thrusts.

Online Learning from Doubly-Streaming Data is a recent paradigm that generalizes traditional online learning by allowing a non-fixed feature space. Representative studies include (Zhang et al. 2015, 2016; Hou, Zeng, and Hu 2018; Beyazit et al. 2018) which considered a monotonically incremental feature space and (Hou, Zhang, and Zhou 2017; Hou and Zhou 2017; Beyazit, Alagurajah, and Wu 2019; Wu et al. 2019; Hou et al. 2021; Hou, Zhang, and Zhou 2021; He et al. 2021a) that further allows features emerged at previous rounds to become unobservable. These studies create a very flexible thus practical learning environment, as it is often unrealistic to define a set of informative features in advance and hope they can be consistently available over long time spans. Their shared technique is to establish correlation between old and new features such that, when the old features are not observed, their information can be reconstructed to help the learners trained on new features, which are usually weak as they have not seen sufficient data instances, make more accurate predictions.

Unfortunately, the prior studies mostly assume a fully supervised learning setting. Without labels, the online learners cannot be updated and the feature correlation is learned slowly, resulting in weakly learned classifiers and erroneously reconstructed features. This can lead to substantial prediction errors. In our OSLMF problem, we strive to build accurate online learners that allow scarce labels, thereby excelling the prior art with a higher level of practicality.

Online Semi-Supervised Learning relieves the label requirement of online learning, with the crux lying in to model and leverage the geometric structure that underlies the data streams. The structure can be either explicit, such as a graph (Zhu, Goldberg, and Khot 2009; Wagner et al. 2018; Huang et al. 2019; Zeisl et al. 2010) defined on a topological space, or implicit, such as a Riemann manifold (Goldberg, Li, and Zhu 2008; Farajtabar et al. 2011; Kumagai and Iwata 2018) or a clustering structure (Dyer, Capo, and Polikar 2013; Yu et al. 2015; Gu et al. 2018) learned from the sequential inputs. The online learners can expedite convergence by leveraging these geometric structures, such as encouraging the nearby instances to share same labels.

However, few semi-supervised online learner thus far has been tailored for doubly-streaming inputs. The main challenge lies in that there is no metric to fairly gauge the distance between pairs of data instances when they are described by different feature spaces, which is the gap that our OSLMF attempt to explore and fulfill. To that end, we propose to use copula model to align the feature spaces by establishing the relationship across various data types including Boolean, ordinal, and continuous, making the distance among data instances arriving along the time horizon measurable, thereby lending our online learner to work well with scarce labels.

The OSLMF Problem

Let $\{(\mathbf{x}_t, y_t) \mid t = 1, 2, \ldots, T\}$ denote an input sequence, of which $\mathbf{x}_t \in \mathbb{R}^{d_t}$ is a d_t-dimensional data vector. In a doubly-streaming setup, we let $d_t \neq d_i$ for any two rounds $t \neq i$ in general. With mix-typed features, we let $\mathbf{x}_t := (\mathbf{x}_C, \mathbf{x}_D)$, where the subscripts C and D denote the continuous and discrete (i.e., Boolean or ordinal) variables, respectively.

At round t, the online learner f_t observes an instance $\mathbf{x}_t$ and returns its prediction $f_t(\mathbf{x}_t)$. Only with a small probability, the ground truth label $y_t \in \{-1, +1\}$ is revealed and the learner suffers a loss $\ell(y_t, f_t(\mathbf{x}_t))$. Based on the loss information, the learner evolves to f_{t+1} and gets ready for the next round. Our goal is to minimize the empirical risk:

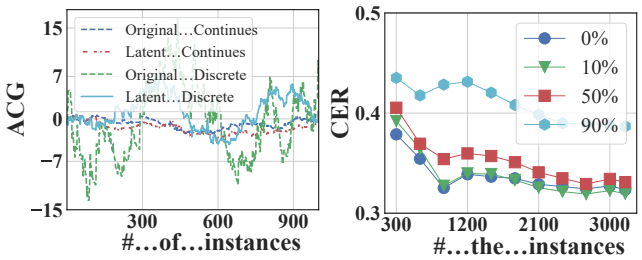
$$R(T) = \frac{1}{l} \sum_{t=1}^{T} \pi(t) \cdot \ell(y_t, f_t(\mathbf{x}_t)), \tag{1}$$

where l denotes the total number of *labeled* instances over T rounds. $\pi(t)$ is an indicator function with $\pi(t) = 1$ for the rounds that reveal label y_t and with $\pi(t) = 0$ otherwise.

Challenges and Our Thoughts

Two challenges (CHs) are manifested from the formulation of the OSLMF problem, described as follows:

CH 1 – Mix-typed streaming features. While the first-order oracle, i.e., the gradient (Cesa-Bianchi and Lugosi 2006) is a common and powerful optimizer of Eq. (1), the features of various data types often span different value



(a) Challenge 1 (b) Challenge 2

Figure 1: Visualization of the two challenges.

scales. Thus, the gradients derived from such different features are *garbled*, where the updating steps suggested by discrete features are in a coarser level of granularity than those by continuous features, leading to more radical updates.

To visualize, a toy example adapted from the "real-stream" dataset in the experiments is illustrated in Figure 1a, which shows the impact of feature type on the derived gradients. In particular, we observe that the the averaged cumulative gradient (ACG) (Schmidt, Le Roux, and Bach 2017) associated with discrete features are oscillating in a *more radical* manner than that with continuous features. The higher the variation of its ACG, the slower the coefficient of that feature is learned. Note, new features are constantly emerging, and initializing their coefficients randomly or as zeros can shift the decision hyperplane in a biased means (He et al. 2021a). Discrete features offering gradients with high variations cannot afford sufficiently updates towards the optimum thus fail to correct the initialization biases. This leads the online learner to make additional erroneous predictions.

CH 2 – Label Scarcity. At the rounds with no label revealed, no risk (loss) is suffered, and hence no gradient is calculated to update the learner as stated in Eq. (1). Intuitively, depending on how scarce the labels are available, the online learner can commit to a low convergence rate, which means that it would take more rounds to converge.

Figure 1b visualizes this intuition, in which Online Convex Programming (Zinkevich 2003) is employed for the learner. The cumulative error rate (CER) that gauges the prediction performance of the learner is illustrated as the curves. We observe that, as the scarcity of labels goes higher, the CER curve tends to stay flat, which indicates a low convergence rate. In online learning, the data instances are presented to the learner in one-pass only. The lower the convergence rate of an online learner, the more the prediction errors that the learner makes compared to a hindsight optimum.

Our Ideas. To overcome the two challenges, we here sketch the two key ideas that motivate our algorithm design. *First*, to tame the radical updates incurred by mix-typed features, we desire a model that can normalize the oscillating gradients over discrete variables into a continuous domain on-the-fly. We advocate the Gaussian copula (GC) (Fan et al. 2017; Hoff et al. 2007; Liu, Lafferty, and Wasserman 2009) that can model complex multivariate distribution of mixed data types in a latent space spanned by continuous normal variables. An online learner trained on this latent space capturing feature correlation enjoys two advantages: 1) any new features can be initialized with educated guess rather than purely random (that incurs bias), and 2) any unobserved feature can be reconstructed such that its learned coefficient can be leveraged to uplift the prediction accuracy.

Second, to aid the label scarcity, we exploit the abundance of unlabelled data instances to deduce a geometric structure underlying the input sequence. On the structure, the instances with similar labels are placed in neighboring regions, while those separated instances are likely to carry disparate labels. To discover such geometric structure, we gauge the distance between paris of instances in the GC-learned latent space and respect their labeling relations as a regularization term. We frame the two ideas into a regularized risk minimization regime with its objective function tailored in the problem statememt.

The Proposed Approach

Overview. Our approach can be conceptually formulated into the following objective functions:

$$\min_{f_1,\ldots,f_T} R(T), \;\; \text{s.t. } \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \text{GC}(\mathbf{z}_t; \text{g}; \Sigma), \tag{2}$$

$$\max_{\text{g};\Sigma} \text{P}[\mathbf{x}_t \mid \mathbf{z}_t; \text{g}^{-1}, \Sigma], \;\; \forall \mathbf{x}_t \in B. \tag{3}$$

In the objectives, Eq. (2) aim at minimizing the semi-supervised learning risk, which posits that the input sequences $\{\mathbf{x}_t\}_{t=1}^T$ are independently drawn from an unknown distribution modeled by a Gaussian copula (GC). Eq. (3) estimates the parameters of GC in a buffer B via an online Expectation-Maximization (EM) process, aiming at discovering a latent representation $\mathbf{z}_t$ (with continuous normals) of each input $\mathbf{x}_t$ (with mix-typed variables). This section delves into model details by scrutinizing the objectives in sequence.

Learning Latent Normals via Gaussian Copula

Gaussian copula (GC) possesses two nice properties as it allows to model the joint distribution underlying the mix-typed features. *First*, by round t-th round ($t > 1$), let $\mathcal{U}_t = \bigcup_1^t \mathbb{R}^t$ represent a universal feature space comprising all features observed so far. Each input $\mathbf{x}_t$ carries a subset of $\mathcal{U}_t$ owing to the feature space variations, where any unobserved features commit to information loss, leading to an inferior learner. GC solves this issue by mapping the observed inputs onto a latent space that contains sufficient statistics for estimating the unobserved features. Such reconstructed information can boost learner to make more accurate online predictions through ensembling (as we will see later in the online ensemble).

Second, GC tames the garbled gradients by its definition:

Definition 1 (GC (Masarotto and Varin 2012))**.** *For* $\forall\, \mathbf{x} \in \mathbb{R}^d$ *that follows the* GC *is a random vector, there is a correlation matrix* Σ *and an element-wise monotone function* $\text{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ *to make that* $\mathbf{x} = \text{g}(\mathbf{z})$ for $\mathbf{z} \sim N_d(\mathbf{0}, \Sigma)$.

As we can see, the latent representation of the input $\mathbf{x}_t := (\mathbf{x}_C, \mathbf{x}_D)$ consists of a set of normal continuous variables $\mathbf{z}_t$

with covariance matrix Σ and zero mean. By training learners on the latent representations, the radical updates are suggested by the continuous rather than discrete features , which is beneficial for a fine-granular search of minimizers.

To enjoy the two aforementioned properties, we explain how to delineate the monotone g and the correlation Σ. Following the prior art (Zhao and Udell 2020; He et al. 2021a), we estimate the discrete variables in $\mathbf{x}_t$ with a monotone cutoff operator taken on probability mass functions, as Σ is invariant to element-wise strictly monotone transformation. Corresponding to a discrete feature $x_i \in \mathbf{x}_D$ with range $|k|$ and mass function $\{p_l\}_{l=1}^{k}$, the mapping is as follows:

$$g_i := \text{cutoff}(z; \mathrm{S}) = 1 + \sum_{s_l \in \mathrm{S}} \mathbb{1}(z > s_l), \tag{4}$$

where $z \in \mathbb{R}$ is continuous normal with cumulative distribution function (CDF) F_z and $\mathrm{S} = \left\{ s_l = F_z^{-1}(\sum_{t=1}^{l} p_t) : l \in |k-1| \right\}$. Then, the latent vector is as $\mathbf{z}_t := \mathrm{g}^{-1}(\mathbf{x}_t) = (\mathrm{g}^{-1}(\mathbf{x}_C), \text{cutoff}^{-1}(\mathbf{x}_D))$, so by the invertibility of monotone mappings. Despite continuous or discrete features, the latent representations have a specified real-value or can yield from Cartesian product of an interval, respectively.

Unobserved feature reconstruction. Note that the dimension of $\mathrm{g}^{-1}(\mathbf{x}_t)$ are equal to that of $\mathbf{x}_t$ but does not match to that of $\mathcal{U}_t$. However, in our OSLMF problem, any feature can change to be unobservable, making the missing entries $\mathbf{x}_M \in \mathcal{U}_t \setminus \mathbb{R}^{d_t}$. For notational succinctness, the observed instance $\mathbf{x}_t$ is denoted as $\mathbf{x}_O$. To obtain a complete latent representation, we reconstruct $\mathbf{z}_t = \phi(\mathbf{x}_t) \in \mathbb{R}^{|\mathcal{U}_t|}$ by establishing relationships between $\mathbf{x}_O$ and $\mathbf{x}_M$. The intuition of our solution is to map the conditional mean vector of the corresponding $\mathbf{z}_M$ based on the marginals of the observed $\mathbf{x}_O$. Such feature reconstruction has two approximation steps: 1) making expectation of the observed $\mathbf{z}_O$ given the observation $\mathbf{x}_O$ and 2) making expectation of the missing $\mathbf{z}_M$ given $\mathbf{z}_O$. The two steps are formulated as follows:

$$\begin{aligned} \tilde{\mathbf{z}}_M &= \mathbb{E}\big[\mathbb{E}[\mathbf{z}_M \mid \mathbf{z}_O, \Sigma] \mid \mathbf{x}_O, \Sigma\big] \\ &= \Sigma_{M,O} \cdot \Sigma_{O,O}^{-1} \cdot \mathbb{E}\left[\mathbf{z}_O \mid \mathbf{x}_O, \Sigma\right], \end{aligned} \tag{5}$$

where $\Sigma_{M,O}$ represents the feature indexes $(\mathbf{x}_M, \mathbf{x}_O)$ corresponding to rows, and $\Sigma_{O,O}$ represents the feature indexes $(\mathbf{x}_O, \mathbf{x}_O)$ corresponding to columns, of sub-matrices of correlation Σ, respectively. Supposing $\tilde{\mathbf{z}}_M$ is an unbiased estimation of $\mathbf{z}_M$, we achieve a complete view of the latent representation $\mathbf{z}_t = (\mathbf{z}_O, \tilde{\mathbf{z}}_M)$. Therefore, we can obtain a reconstructed space of input $\mathbf{x}_t$ by sampling from the copula $\mathrm{GC}(\mathbf{z}_t, \mathrm{g}, \Sigma)$, denoted as $\mathbf{x}_t^{\text{rec}} = (\hat{\mathbf{x}}_O, \hat{\mathbf{x}}_M) \in \mathcal{U}_t$.

Online EM for parameter estimation. The feature reconstruction allows us to optimize the function g and the correlation Σ by gauging the discrepancy between the observed $\mathbf{x}_t := \mathbf{x}_O$ and the reconstructed $\hat{\mathbf{x}}_O$ in a stochastic and online fashion. We first define $g_i^{-1} = \Phi^{-1} \circ F_i$, where Φ is a standard normal CDF and F_i corresponds to the true yet unknown CDF of the i-th feature. A buffer B of arriving instances is employed to empirical estimate F_i as $\hat{F}_i$. The estimator for continuous feature is defined as follows.

$$\hat{g}_i^{-1}(x_i) = \Phi^{-1}\big(H \cdot \hat{F}_i(x_i)\big), \tag{6}$$

where a finite result is ensured by the scale $H = |B|/(|B|+1)$. For discrete features, by swapping out the i-th feature's sample mean for its probability mass p_l^i, we can denote cutoff S^i as a special case of Eq. (6).

$$\mathrm{S}^i = \left\{ \Phi^{-1}\left(\frac{\sum_{t=1}^{|B|} \mathbb{1}\left(\mathbf{x}_t[i] \le l\right)}{|B|+1} \right), l \in |k-1| \right\}, \tag{7}$$

where $\mathbf{x}_t[i]$ indicates the i-th discrete feature of the t-th input. To estimate the correlation Σ, online EM is progressed in the buffer B.

Specifically, by taking the conditional expectation os Σ, we aim to maximize the likelihood that the observed entries (denoted by $\mathbf{X}_O$) of the buffered matrix $\mathbf{X}_B \in \mathbb{R}^{|\mathcal{U}_t| \times |B|}$ can be accurately reconstructed. To disambiguate the notation, the empirical correlation obtained in the precedent round, denoted as $\Sigma^{(t-1)}$, and the objective to be approximated in the current round, denoted as $\hat{\Sigma}$, respectively. The log-likelihood function is the following:

$$\begin{aligned} Q\big(\hat{\Sigma}; \Sigma^{(t-1)}, \mathbf{X}_O\big) &:= \frac{1}{|B|} \sum_{t=1}^{|B|} \mathbb{E}\left[\mathcal{L}(\hat{\Sigma}; \mathbf{x}_t, \mathbf{z}_t) \mid \mathbf{z}_t, \Sigma^{(t-1)} \right] \\ &= \text{const} - \frac{1}{2} \log\det(\hat{\Sigma}) - \frac{1}{2} \operatorname{Tr}\left(\hat{\Sigma}^{-1} G(\Sigma^{(t-1)}, \mathbf{x}_t) \right), \end{aligned} \tag{8}$$

with $\Sigma^{(0)}$ initialized as an identity matrix. To maximize Eq. (8), two steps iterate in a different way as shown below.

E-step. We use Eq. (5) to calculational expectation given $\mathbf{x}_t$ and $\Sigma^{(t-1)}$ in order to express the likelihood $Q(\hat{\Sigma}; \Sigma^{(t-1)}, \mathbf{X}_O)$ in terms of $\hat{\Sigma}$ by substituting $G(\Sigma^{(t-1)}, \mathbf{x}_t) = \mathbb{E}_{t \in B}[\mathbf{z}_t \mathbf{z}_t^\top \mid \mathbf{x}_t, \Sigma^{(t-1)}]$ in Eq. (8).

M-step. Solve $\tilde{\Sigma} = \arg\max_{\Sigma} Q(\Sigma; \Sigma^{(t-1)}, \mathbf{X}_O)$ which, according to the EM theory, is guaranteed to increase the likelihood, *cf.* Chapter 3 in (McLachlan and Krishnan 2007). Then, we use (Cappé and Moulines 2009) to change the correlation in the current round to a harmonic sum of the correlation from the previous round, which is given by $\Sigma^{(t-1)}$ and $\tilde{\Sigma}$. This treatment can generate a $\Sigma^{(1)}, \ldots, \Sigma^{(T)}$ sequence with smooth updates. However, we note that this sequence represents an unconstrained series of monotonically convergent local likelihood maximizers. We use an approximation to fit the empirical maximizer into a normal covariance as follows:

$$\hat{\Sigma} = P_{\mathcal{E}}\big((1-\gamma_t)\, \Sigma^{t-1} + \gamma_t \tilde{\Sigma} \big), \tag{9}$$

with $\gamma_t \in (0, 1]$ being a decaying step size and $P_{\mathcal{E}}(\cdot)$ normalizes the correlation as $\mathbf{D}^{-1/2} \hat{\Sigma} \mathbf{D}^{-1/2}$ with $\mathbf{D} = \text{diag}(\hat{\Sigma})$ (Zhao and Udell 2020; He et al. 2021a).

Learning Data Geometrics via Local Density-Peaks

In the learned latent spaces, any two instances that arrive at different time steps have their feature spaces aligned, hence their distance become measurable. This allows us to uncover the underlying geometric structure of instances, propagating the very limited supervision information from the scarce labelled instances to their neighbors on the structure.

In this work, we harvest the clustering structure to approximate the underlying geometric structure of the data. To respect the online property, a non-iterative, density-peak-based clustering method (Rodriguez and Laio 2014) is employed, with its main process illustrated in Figure 2. Specifically, we characterize each arriving instance $\mathbf{x}_t$ with two

indicators, namely the local density ρ_t and the distance δ_t, defined as:

$$\rho_t = \sum_{\mathbf{x}_i \in B, i \neq t} e^{-\left(d(\mathbf{x}_t, \mathbf{x}_i)/d_{\text{cut}}\right)^2}, \tag{10}$$

$$\delta_t = \begin{cases} \min_{i:\rho_t < \rho_i}(d(\mathbf{x}_t, \mathbf{x}_i)), \text{ others} \\ \max_i(d(\mathbf{x}_t, \mathbf{x}_i)), \forall i, \rho_t \geq \rho_i \end{cases}, \tag{11}$$

where $d(\mathbf{x}_t, \mathbf{x}_i)$ gauges the Euclidean distance between $\mathbf{x}_t$ and $\mathbf{x}_i$ in the reconstructed universal feature space $\mathcal{U}_t$, and d_{cut} is the adaptively adjusted cutoff distance. The distance δ_t is measure between $\mathbf{x}_t$ and any other instance $\mathbf{x}_i$ with a local density higher than ρ_t. Note, we set the cutoff distance d_{cut} as $d_{cut} = \lfloor P_{Arr} \times |B| \times (|B|-1)/2 \rfloor$, where P_{Arr} is empirically set between 1% and 2% (Wu et al. 2018).

The centroids of clusters can thus be determined intuitively: an instance that has a high ρ_t (hence surrounded by a large number of neighbors) and a high δ_t (hence placed far away from any other likely centroids) is deemed as a centroid. Figure 2a illustrates the decision graph that selects centroids by their harmonic mean of ρ_t and δ_t. Empirical and theoretical evidences were documented in (Rodriguez and Laio 2014) to substantiate that the centroids and their corresponding clusters selected through this density-peak criterion is on a par with those iterative clustering methods such as online K-means (Hosseini, Gholipour, and Beigy 2016; Din et al. 2020), yet renders a much higher computational efficiency and better fit to an online learning setting.

Label propagation via geometric structure. After determining the cluster centroids, we can construct the geometric structure by letting each $\mathbf{x}_t$ point to its nearest instance $\mathbf{x}_i$ that has a higher ρ_t, thereby forming a directed graph as shown in Figure 2b. To propagate the very limited supervision information, we conduct self-training on the learned structure. The crux lies in an iterative selection of unlabelled instances that can be predicted correctly with high confidence. Such instance can be naturally deemed as that pointed by the labelled instances. If no labeled instance is pointing to it, then the propagating path further traces back with a depth-first search, until one label is found. To scale up for streaming inputs, our online learner proceeds as follows. As $\mathbf{x}_t$ arrives, we merge it into B and pop up a previously buffered instance $\mathbf{x}_i$ that is not labeled. We select $\mathbf{x}_i$ as the most confidently predicted instance, so by the constructed geometric shape. Compared to the prior semi-supervised online learners (He et al. 2021a; Din et al. 2020) that pop up and predict the oldest instance in B (where the prediction on $\mathbf{x}_t$ is actually made at the $t + B$-th time step, with a B prediction delay), our approach enjoys a higher online efficiency as the prediction delay of any arriving instance $\mathbf{x}_t$ is *at most* B. Specifically, in the worst case, $\mathbf{x}_t$ cannot be confidently predicted and is not popped up until it becomes the oldest instance in B, our approach shares a B prediction delay as previous studies. In more ideal cases, $\mathbf{x}_t$ is pointed by a labeled instance and is hence predicted at the same or nearby round as it appears.

Online Ensembling for Expedited Convergence

Thus far, we have described how to build classifier to make on-the-fly predictions with a buffer B on the feature space

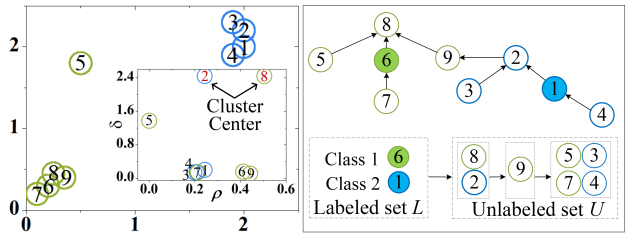


Figure 2: An example of learning data geometrics with 9 instances in a 2D feature space. (a) The orginal data distribution, and the inserted figure is the decision graph to yield instances with high ρ and δ. (b) The learned geometric structure and its self-training processes where colored solid circles ('6' and '1') are initial labelled instances. The labeling information is propagated through the learned topological structure, from the labeled instances to their immediate neighbors ('8' and '2') at first, and gradually to more far-away instances.

reconstructed with the learned GC model. We denote this classifier as f_O, with 'O' indicating that it is trained on the original, mix-typed features. Let $y_O = \langle f_O, \mathbf{x}_t^{\text{rec}} \rangle$ conceptually denote the prediction. Despite the latent space learned by GC is intermediately leveraged as $\mathbf{x}_t^{\text{rec}} = \text{g}(\mathbf{z}_t)$, its power is not fully compiled into the model as the latent space is with continuous (normal) variables only thus may lead to faster convergence. However, this gives rise to a trade-off, as any newly emerge feature being not described by a sufficient number of instances can create inaccurate latent representations. It is hence desireable to train the online learners that can enjoy the continuous merit of the learned latent space on the one hand, but will not suffer from its inaccuracy on the other. This motivates us to leverage online ensemble learning that two base classifiers trained on both original and latent feature spaces jointly suggest more accurate predictions.

Let $y_Z = \langle f_Z, \mathbf{z}_t \rangle$ denote the prediction made on the latent representation of $\mathbf{x}_t$. The ensemble prediction is $\hat{y}_t = \alpha_1 \cdot y_O + \alpha_2 \cdot y_Z$, with $\alpha_1 + \alpha_2 = 1$. The values of α_1 and α_2 determine the importance of the two base classifiers f_O and f_Z, respectively. Let $R_O(T) = \sum_{t=1}^{T} \pi(t) \cdot \ell(y_t, y_O)$ and $R_Z(T) = \sum_{t=1}^{T} \pi(t) \cdot \ell(y_t, y_Z)$ be the cumulative risks suffered by f_O and f_Z over T rounds, respectively. Then, at the round T +1, α_1 is updated based on the risk exponentials as follows(He et al. 2019; Cesa-Bianchi and Lugosi 2006):

$$\alpha_1 = e^{-\mu R_O(T)} / (e^{-\mu R_O(T)} + e^{-\mu R_Z(T)}) \tag{12}$$

where $\mu = 2\sqrt{2 \ln 2/T}$ is a tuned parameter.

Experiments

This section documents empirical evidence to substantiate the effectiveness of our proposed OSLMF algorithm.

Datasets. Our evaluation are conducted on 14 datasets, including 13 from the UCI repository (Asuncion 2007) and one from the Massive Online Analysis (MOA) (Bifet et al. 2018) that simulates a real streaming setup. The evaluated

Dataset	#Inst.	#Feat.	Dataset	#Inst.	#Feat.
wpbc	198	33	dna	949	180
ionosphere	351	34	german	1000	24
wdbc	569	30	splice	3190	60
australian	690	14	kr-vs-kp	3196	36
credit-a	690	15	magic04	19,020	10
wbc	699	9	a8a	22,696	123
diabetes	768	8	stream	10,000	1000

Table 1: Statistics of the studied datasets.

datasets span diverse application domains, such as education, finance, etc. Table 1 summarizes their statistics.

Evaluation Protocol. We follow the same protocol of prior studies (Zhang et al. 2016; He et al. 2019) to simulate two types of streaming feature dynamics. 1) Trapezoidal Data Streams, in which later inputs tend to carry incrementally more features. we split the original datasets into ten chunks, where in the i-th chunk only the first $i*10\%$ features would be retained, i.e., the first data batch will retain the first 10% features and so forth. 2) Capricious Data Streams, in which new features appear and old features fadeaway over time arbitrarily, we randomly remove 50% features in each arriving instance randomly. To simulate a semi-supervised learning environment, 50% labels are randomly removed from the datasets. Cumulative error rate (CER) is employed to measure the algorithm performance, which counts the ratio of error predictions over all instances seen so far.

Results. Table 2 and Figures 3-4 documents the experimental results. We compare OSLMF with four rival models, FOBOS (Singer and Duchi 2009), OMR (Goldberg, Li, and Zhu 2008), OLSF (Zhang et al. 2016), and OVFM (He et al. 2021a), aiming to answer questions (**Q1** – **Q4**) as follows.

Q1. *Does our OSLMF outperform the state-of-the-arts?*

To better analyze the results in Table 2, we make statistical analyses of the loss/win, the Wilcoxon signed-ranks test (p-value) (Demšar 2006), and the Friedman test (F-rank) (Demšar 2006). The statistical results are attached at the bottom of the table. We make three observations from the results. *First*, OSLMF achieves the best accuracy performance (the lowest CER) on most cases and loses to its competitors in five out of 84 settings only. *Second*, all the p-value are smaller than 0.05, which verifies that OSLMF has significantly better prediction accuracy than all the comparison models at a 95% confidence level. *Third*, according to F-rank, the performance of OVFM and OLSF tie and are both inferior to our OSLMF, followed by OMR and lastly FOBOS. Such result coincides with the design of these competitors, where OVFM and OLSF have tailored learning machinism to deal with a varying feature space while OMR posits fixed features. FOBOS reprents a baseline without any speical design for feature space dynamics nor label scarcity, thus ends up with the worst accuracy performance.

Q2. *How actually can GC tame mix-typed features online?*

The comparison between our OSLMF and OMR amounts to the answer. Whereas both of them leveraged the geometric structure of data to realize online semi-supervised learning, OMR does not allow inputs with mix-typed streaming features. We observe that OSLMF has a lower CER than OMR by a ratio of 23.1%. Also, from Figures 3, OSLMF arrives at 60.56%, 38.97%, 0.41%, and 76.63% lower CER than those of OMR on the datasets of australian, german, kr-vs-kp, and a8a, respectively – the datasets naturally with mix-typed features. These findings substantiate that the effectiveness of GC in dealing with mix-typed streaming features, which helped our OSLMF to attain superior prediction performance.

Q3. *Can density-peaks profile geometric-structure of data?*

The comparison between our OSLMF and OVFM amounts to the answer, as OVFM assumes fully labeled data streams. Also, we compare their ultimate CERs in Table 2 and trends of CERs over time in the setting of capricious data streams as shown in Figure 3. We observe that OSLMF has a lower CER than OVFM throughout the online learning process. Its ultimate CERs are 23.92%, 29.28%, 13.93%, and 54.97% lower than that of OVFM on the datasets of australian, german, kr-vs-kp, and a8a, respectively. These observations verify that by propagating the limited labeling information on the learned data geometrics via local density-peaks, our OSLMF enjoys the abundance of unlabelled data, which are leveraged to uplift its prediction accuracy.

Q4. *Does online ensembling yield better accuracy?*

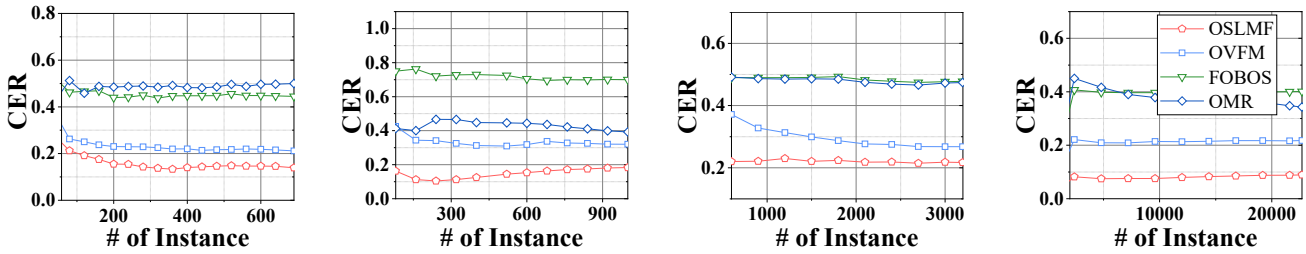
To empirically investigate how OSLMF adaptively controls the combination of the two base classifiers, we monitor the changes of the ensembling coefficient α_1, as plotted in Figure 4. The pattern of α_2 is symmetric to α_1 as they add up to 1 thus is omitted to keep succinct. We observe that the variation patterns of α_1 differs across datasets, while the accuracy performance of OSLMF remains its increasing trend (with a decreasing CER). The inconsistent patterns of α_1 necessitates the ensembling, as the higher/lower its value, the more/less important the classifier trained in the observed feature space, while it is next to foresee when this classifier prevail the other classifier trained on the latent normal space. Our ensemble strategy allows the coefficients α_1 and α_2 learned from the streaming inputs, thus lifts the overhead of choosing the better classifier in a prior. To further investigate, we ablate the proposed algorithm by plotting the CER of using observed features to make predictions only, yielding an simplified algorithm named OSLMF-F. We observe that OSLMF-F is inferior to the ensemble OSLMF with a consistently higher CER (thus lower accuracy).

Conclusion

In this paper, we strive to push the boundary of online learning from doubly-streaming data. A new learning problem named OSLMF is explored, which imposes no assumption on the feature types nor the learning labels, thereby excelling the prior studies that make assumptions on either or both in terms of flexibility and applicability. To tame the mix-typed streaming features, we leverage the Gaussian copula

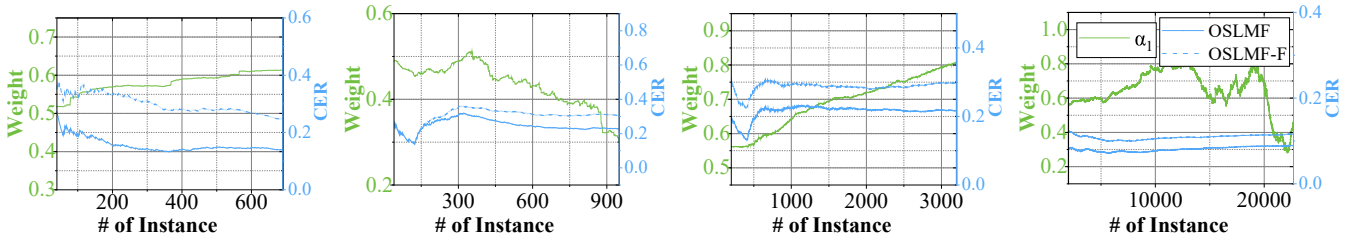
	Trapezoidal Data Streams				Capricious Data Streams			
Dataset	FOBOS	OMR	OLSF	OSLMF	FOBOS	OMR	OVFM	OSLMF
wpbc	.237 ± .000	.345 ± .000	.366 ± .001	**.235 ± .003**	.248 ± .000•	.320 ± .000•	.309 ± .000•	.567 ± .001
ionosphere	.342 ± .000	.443 ± .000	.230 ± .000	**.225 ± .000**	.479 ± .000	.418 ± .000•	.269 ± .000•	.466 ± .000
wdbc	.577 ± .000	.460 ± .000	.347 ± .000	**.187 ± .000**	.628 ± .000	.399 ± .000	.113 ± .000	**.110 ± .000**
australian	.497 ± .000	.491 ± .000	.486 ± .000	**.356 ± .000**	.455 ± .000	.492 ± .001	.255 ± .000	**.194 ± .000**
credit-a	.445 ± .000	.415 ± .000	.312 ± .000	**.186 ± .000**	.445 ± .000	.484 ± .000	.484 ± .000	**.416 ± .000**
wbc	.345 ± .000	.394 ± .000	.455 ± .000	**.219 ± .000**	.162 ± .000	.461 ± .000	.072 ± .000	**.059 ± .000**
diabetes	.349 ± .000	.376 ± .000	.331 ± .000	**.170 ± .000**	.349 ± .000	.426 ± .000	.399 ± .000	**.331 ± .004**
dna	.518 ± .000	.496 ± .000	.499 ± .000	**.462 ± .000**	.511 ± .000	.496 ± .000	.282 ± .000	**.229 ± .000**
german	.300 ± .000	.381 ± .000	.407 ± .000	**.227 ± .000**	.700 ± .000	.372 ± .000	.321 ± .000	**.227 ± .000**
splice	.500 ± .000	.493 ± .000	.375 ± .000	**.311 ± .000**	.519 ± .000	.400 ± .001	.498 ± .000	**.424 ± .000**
kr-vs-kp	.482 ± .000	.523 ± .000	.239 ± .000	**.221 ± .000**	.478 ± .000	.242 ± .000	.280 ± .000	**.241 ± .000**
magic04	.665 ± .000	.529 ± .000	.374 ± .000	**.348 ± .000**	.689 ± .000	.438 ± .000	.317 ± .000	**.091 ± .000**
a8a	.375 ± .000	.482 ± .003	.273 ± .004	**.179 ± .001**	.401 ± .003	.368 ± .001	.191 ± .001	**.086 ± .001**
stream	.615 ± .000	.472 ± .000	.233 ± .000	**.230 ± .000**	.621 ± .000	.471 ± .000	.231 ± .000	**.224 ± .000**
loss/win	0/14	0/14	0/14	0/42	1/13	2/12	2/12	5/37
p-value	.0005	.0005	.0005	−−	.0008	.0015	.0071	−−
F-rank	3.286	3.124	2.500	1.000	3.357	3.071	2.286	1.285

Table 2: The comparison results on cumulative error rates. We repeated the experiment 10 times for each dataset, averaged the cumulative error rate (CER), and calculated the variance of the 10 times values. Experimental results (CER ± Variance) for 14 data sets in the case of trapezoidal and capricious data streams. • indicates the cases that OSLMF loses the comparison.



(a) australian (b) german (c) kr-vs-kp (d) a8a

Figure 3: CER trends of four methods in capricious data streams.



(a) australian (b) dna (c) kr-vs-kp (d) a8a

Figure 4: (Capricious) Temporal variation of ensemble weight α_1 and CERs of OSLMF and its ablation variant OSLMF-F .

to model the correlation between discrete and continuous variables in a latent normal space, so as to mitigate radical updates for fast convergence. To exploit the scarce labeled instances, we uncover the geometric structure underlying the arriving instances via density-peak clustering, so as to propagate the labeling information to their unlabeled neighbors. Extensive experiments on 14 benchmark datasets are conducted to evidence the viability, effectiveness, and superiority of our proposed algorithm. The results substantiated that our proposed algorithm significantly outperforms the state-of-the-art competitors. In the future, we plan to make the hyper-parameters of our proposed algorithm self-adaptive through evolutionary computation algorithms.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 62176070, U21A20463, and 62106024, in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber research & development, innovation, and workforce development (visit cyberinitiative.org, for more information about CCI), and in part by the Guangzhou Basic and Applied Basic Research Project (202201020221).

References

Agarwal, D.; Chen, B.-C.; and Elango, P. 2010. Fast online learning through offline initialization for time-sensitive recommendation. In *KDD*, 703–712.

Aggarwal, C. C. 2007. *Data streams: models and algorithms*, volume 31. Springer.

Asuncion, A. 2007. Uci machine learning repository, university of california, irvine, school of information and computer sciences. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*.

Balzano, L.; Chi, Y.; and Lu, Y. M. 2018. Streaming pca and subspace tracking: The missing data case. *Proceedings of the IEEE*, 106(8): 1293–1310.

Beyazit, E.; Alagurajah, J.; and Wu, X. 2019. Online learning from data streams with varying feature spaces. In *AAAI*, volume 33, 3232–3239.

Beyazit, E.; Hosseini, M.; Maida, A.; and Wu, X. 2018. Learning simplified decision boundaries from trapezoidal data streams. In *International Conference on Artificial Neural Networks*, 508–517. Springer.

Bifet, A.; Gavaldà, R.; Holmes, G.; and Pfahringer, B. 2018. *Machine Learning for Data Streams with Practical Examples in MOA*. MIT Press. https://moa.cms.waikato.ac.nz/book/.

Cappé, O.; and Moulines, E. 2009. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 593–613.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Chen, F.; Wu, D.; Yang, J.; and He, Y. 2022. An Online Sparse Streaming Feature Selection Algorithm. *arXiv preprint arXiv:2208.01562*.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7: 1–30.

Din, S. U.; Shao, J.; Kumar, J.; Ali, W.; Liu, J.; and Ye, Y. 2020. Online reliable semi-supervised learning on evolving data streams. *Information Sciences*, 525: 153–171.

Dyer, K. B.; Capo, R.; and Polikar, R. 2013. Compose: A semisupervised learning framework for initially labeled nonstationary streaming data. *IEEE transactions on neural networks and learning systems*, 25(1): 12–26.

Fan, J.; Liu, H.; Ning, Y.; and Zou, H. 2017. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2): 405–421.

Farajtabar, M.; Shaban, A.; Rabiee, H. R.; and Rohban, M. H. 2011. Manifold coarse graining for online semi-supervised learning. In *ECML-PKDD*, 391–406. Springer.

Goldberg, A. B.; Li, M.; and Zhu, X. 2008. Online manifold regularization: A new learning setting and empirical study. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 393–407. Springer.

Gu, B.; Yuan, X.-T.; Chen, S.; and Huang, H. 2018. New incremental learning algorithm for semi-supervised support vector machine. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1475–1484.

He, Y.; Dong, J.; Hou, B.-J.; Wang, Y.; and Wang, F. 2021a. Online Learning in Variable Feature Spaces with Mixed Data. In *ICDM*, 181–190. IEEE.

He, Y.; Wu, B.; Wu, D.; Beyazit, E.; Chen, S.; and Wu, X. 2019. Online learning from capricious data streams: a generative approach. In *IJCAI*.

He, Y.; Yuan, X.; Chen, S.; and Wu, X. 2021b. Online Learning in Variable Feature Spaces under Incomplete Supervision. In *AAAI*, volume 35, 4106–4114.

Hoff, P. D.; et al. 2007. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1): 265–283.

Hosseini, M. J.; Gholipour, A.; and Beigy, H. 2016. An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams. *Knowledge and information systems*, 46(3): 567–597.

Hou, B.-J.; Yan, Y.-H.; Zhao, P.; and Zhou, Z.-H. 2021. Storage Fit Learning with Feature Evolvable Streams. In *AAAI*.

Hou, B.-J.; Zhang, L.; and Zhou, Z.-H. 2017. Learning with feature evolvable streams. *NeurIPS*, 30.

Hou, B.-J.; Zhang, L.; and Zhou, Z.-H. 2021. Prediction With Unpredictable Feature Evolution. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10.

Hou, C.; Zeng, L.-L.; and Hu, D. 2018. Safe classification with augmented features. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2176–2192.

Hou, C.; and Zhou, Z.-H. 2017. One-pass learning with incremental and decremental features. *IEEE transactions on pattern analysis and machine intelligence*, 40(11): 2776–2792.

Huang, C.; Li, P.; Gao, C.; Yang, Q.; and Shao, J. 2019. Online Budgeted Least Squares with Unlabeled Data. In *ICDM*, 309–318. IEEE.

Kumagai, A.; and Iwata, T. 2018. Learning dynamics of decision boundaries without additional labeled data. In *KDD*, 1627–1636.

Lian, H.; Atwood, J. S.; Hou, B.; Wu, J.; and He, Y. 2022. Online Deep Learning from Doubly-Streaming Data. In *ACM Multimedia*.

Liu, H.; Lafferty, J.; and Wasserman, L. 2009. The nonparanormal: Semiparametric estimation of high dimensional

undirected graphs. *Journal of Machine Learning Research*, 10(10).

Masarotto, G.; and Varin, C. 2012. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6: 1517–1549.

McLachlan, G. J.; and Krishnan, T. 2007. *The EM algorithm and extensions*. John Wiley & Sons.

Meng, Y.; Jiang, C.; Quek, T. Q.; Han, Z.; and Ren, Y. 2017. Social learning based inference for crowdsensing in mobile social networks. *IEEE Transactions on Mobile Computing*, 17(8): 1966–1979.

Pan, Z.; Yu, H.; Miao, C.; and Leung, C. 2017. Crowdsensing air quality with camera-enabled mobile devices. In *AAAI*, volume 31, 4728–4733.

Rodriguez, A.; and Laio, A. 2014. Clustering by fast search and find of density peaks. *science*, 344(6191): 1492–1496.

Schmidt, M.; Le Roux, N.; and Bach, F. 2017. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2): 83–112.

Schreckenberger, C.; Glockner, T.; Stuckenschmidt, H.; and Bartelt, C. 2020. Restructuring of Hoeffding Trees for Trapezoidal Data Streams. In *ICDM*, 416–423. IEEE.

Shalev-Shwartz, S.; et al. 2011. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2): 107–194.

Singer, Y.; and Duchi, J. C. 2009. Efficient learning using forward-backward splitting. *Advances in Neural Information Processing Systems*, 22.

Wagner, T.; Guha, S.; Kasiviswanathan, S.; and Mishra, N. 2018. Semi-supervised learning on data streams via temporal label propagation. In *International Conference on Machine Learning*, 5095–5104. PMLR.

Wu, D. 2023. *Robust Latent Feature Learning for Incomplete Big Data*. Springer Nature press.

Wu, D.; He, Y.; Luo, X.; Shang, M.; and Wu, X. 2019. Online feature selection with capricious streaming features: A general framework. In *2019 IEEE International Conference on Big Data (Big Data)*, 683–688. IEEE.

Wu, D.; He, Y.; Luo, X.; and Zhou, M. 2021. A latent factor analysis-based approach to online sparse streaming feature selection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(11): 6744–6758.

Wu, D.; Shang, M.; Luo, X.; Xu, J.; Yan, H.; Deng, W.; and Wang, G. 2018. Self-training semi-supervised classification based on density peaks of data. *Neurocomputing*, 275: 180–191.

Yu, Z.; Luo, P.; You, J.; Wong, H.-S.; Leung, H.; Wu, S.; Zhang, J.; and Han, G. 2015. Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28(3): 701–714.

Zeisl, B.; Leistner, C.; Saffari, A.; and Bischof, H. 2010. Online semi-supervised multiple-instance boosting. In *CVPR*, 1879–1879. IEEE.

Zhang, Q.; Zhang, P.; Long, G.; Ding, W.; Zhang, C.; and Wu, X. 2015. Towards mining trapezoidal data streams. In *2015 IEEE International Conference on Data Mining*, 1111–1116. IEEE.

Zhang, Q.; Zhang, P.; Long, G.; Ding, W.; Zhang, C.; and Wu, X. 2016. Online learning from trapezoidal data streams. *IEEE Transactions on Knowledge and Data Engineering*, 28(10): 2709–2723.

Zhang, Z.-Y.; Zhao, P.; Jiang, Y.; and Zhou, Z.-H. 2020. Learning with feature and distribution evolvable streams. In *ICML*, 11317–11327. PMLR.

Zhao, Y.; and Udell, M. 2020. Missing value imputation for mixed data via gaussian copula. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 636–646.

Zhu, X.; Goldberg, A. B.; and Khot, T. 2009. Some new directions in graph-based semi-supervised learning. In *ICME*, 1504–1507. IEEE.

Zinkevich, M. 2003. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *ICML*, 928–936.