Online Learning in Variable Feature Spaces under Incomplete Supervision

Yi He,[1] Xu Yuan,[1] Sheng Chen,[1] Xindong Wu[2,3]

[1] Center for Advanced Computer Studies (CACS), University of Louisiana at Lafayette, USA
[2] Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, China
[3] Mininglamp Academy of Sciences, Mininglamp Technology, China
{yi.he1, xu.yuan, chen}@louisiana.edu, xwu@hfut.edu.cn

Abstract

This paper explores a new online learning problem where the input sequence lives in an over-time varying feature space and the ground-truth label of any input point is given only occasionally, making online learners less restrictive and more applicable. The crux in this setting lies in how to exploit the very limited labels to efficiently update the online learners. Plausible ideas such as propagating labels from labeled points to their neighbors through uncovering the point-wise geometric relations face two challenges: (1) distance measurement fails to work as different points may be described by disparate sets of features and (2) storing the geometric shape, which is formed by all arrived points, is unrealistic in an online setting. To address these challenges, we first construct a universal feature space that accumulates all observed features, making distance measurement feasible. Then, we use manifolds to represent the geometric shapes and approximate them in a sparse means, making manifolds computational and memory tractable in online learning. We frame these two building blocks into a regularized risk minimization algorithm. Theoretical analysis and empirical evidence substantiate the viability and effectiveness of our proposal.

Introduction

Nowadays, huge volumes of data abound in all domains of human endeavour, calling for learning systems that can process and make decisions in real-time. Online learning is proposed to build such systems: when data streaming in continuously, it initiates the learning process at any time without waiting all data instances to be arrived. Considering that data streams usually span a long time, data instances arriving at different time steps may be described by very disparate features. For this reason, a line of research, aiming to *learn data streams in variable feature spaces*, has been explored to relax the constraint that all data instances must be described by the same set of features (Masud et al. 2010; Gomes et al. 2013; Zhang et al. 2015; Hou, Zhang, and Zhou 2017; Beyazit, Alagurajah, and Wu 2019; He et al. 2019).

Besides, providing full labeling to data instances in an online setting is prohibitive due to the speed, volume, and duration of streaming data. To lift the restriction that all data instances must be labeled, another line of research, which we refer to as *online semi-supervised learning*, has also drawn extensive attention (Goldberg, Li, and Zhu 2008; Goldberg et al. 2011; Chu et al. 2011; Mohamad, Bouchachia, and Sayed-Mouchaweh 2018; Wagner et al. 2018).

Surprisingly, although supporting flexible features and flexible labeling are two important aspects in online learning, no method has been developed to support the both. The goal of this paper is to fill this critical gap. To this end, we explore a new problem, termed *Online learning in Variable feature Spaces under Incomplete Supervision* (OVSIS), where a data stream lives in a feature space being arbitrarily variable with labels rarely given.

A main challenge of solving OVSIS lies to effectively exploit the very limited labeled instances to update the online learner. A plausible idea is to uncover the geometric relations among the input points and propagate the label information from labeled instances to their unlabeled neighbors. However, it suffers from two limitations. First, as different points may live in disparate feature spaces, no metric can be applied directly to measure their distances. Second, the geometric shape is formed by all previously arrived points, so storing them can incur excessive memory overhead.

To overcome the first limitation, we first draw insights from (Hou, Zhang, and Zhou 2017; He et al. 2019) to capture stationarity in a variable feature space. Namely, we construct a universal feature space that embodies all features emerged in arrived instances. We can then project the input points onto this universal feature space and measure distance between them as they now have equal dimensions.

To address the second limitation, we use manifold structures that underlie the universal feature space to embed the point-wise geometric relations. To maintain the "online" property of our system, the manifold is sparsely approximated by a random projection tree (RP-Tree) (Freund et al. 2008). Such an approximation allows to store only a handful of representatives, each of which corresponds to a leaf of the RP-Tree. Each leaf governs a local region on the manifold, representing the points that fall into this region.

Specific contributions of this paper are as follows:

1. This is the first work to explore the OVSIS problem, where the online learner faces a variable feature space and receives incomplete supervision labels.
2. A novel AGDES algorithm is proposed to tackle the OV-

SIS problem, which provably enjoys a tight performance bound than naïve gradient descent methods.

3. We carry out extensive experiments over 10 widely used datasets, and the results demonstrate the viability and effectiveness of our proposal.

This paper proceeds as follows. We in the next section review the related literatures. Then, we formulate the learning problem, spotlight the challenges, and briefly state our thoughts. The objective function, the complete algorithm design, and the theoretical analyses are presented thereafter in sequence. We end by presenting the experimental results and concluding our findings. Proofs, derivation steps, and complexity analysis are deferred to an electronic companion (available as supplementary material).

Related Work

This work brings together two separate research lines in online learning, *i.e.*, learning data streams in variable feature spaces and online learning with incomplete labels. We review the prior works in each line and discuss the relations of our work with them. It is worth pointing out that in concept-drift (Gama et al. 2014; Lu et al. 2018), the statistical properties of features may change as data streaming in, but the number of features carried by each instance is fixed in a priori, which thus differs from our learning problem.

Online Learning in Variable Feature Space. For data streams that are over wide time spans, it is impractical to require all data instances to be described by a fixed set of features. Pioneer efforts have paid attention to adapt online learners to arbitrarily variable feature spaces.

Zhang et al. (2015) allow the arriving instances to carry different sets of features but later instances are assumed to include monotonically more features than the earlier ones. Hou, Zhang, and Zhou (2017); Zhang et al. (2020) place no such assumption but they require to have an overlapping period, during which a batch of consecutive instances must contain all possible features. Such a requirement is too hard to be satisfied in a feature space that varies arbitrarily. As a result, all these methods cannot solve our OVSIS problem.

Recent studies (Beyazit, Alagurajah, and Wu 2019; He et al. 2019) further lift those constraints, allowing the feature space to vary without following any regularity. Despite effective, they stipulate a fully-labeled environment, using labels to capture stationarity in variable feature spaces. These models cannot be updated if an arriving instance is not labeled. Thus, the lower the probability that the labels are given, the slower these models converge. A slow convergence rate then incurs substantial prediction errors in an online setting. Our approach solves this issue by effectively exploiting both labeled and unlabeled data to expedite the learning process, thereby advancing this research line.

Online Learning with Incomplete Labels. Relieving the label requirement in online learning remains an open challenge (Krempl et al. 2014). Existing works fall into two categories. The first category is online active learning (Goldberg et al. 2011; Chu et al. 2011; Lu, Zhao, and Hoi 2016; Hao et al. 2016; Mohamad, Bouchachia, and Sayed-Mouchaweh 2018), where all data instances are unlabeled as they arrived. At any time, the learner decides whether or not to label an appeared instance. The goal here is to train an accurate classifier with a minimized labeling budget.

The second category is online semi-supervised learning, and our approach belongs to this category. The key idea is to exploit the abounded, unlabeled data to improve the classifier trained with scarce labels. The mainstream solutions leverage the topological structure underlying the feature space, such as Riemann manifolds (Goldberg, Li, and Zhu 2008; Farajtabar et al. 2011; Kumagai and Iwata 2018) or similarity graphs (Wagner et al. 2018; Huang et al. 2019), to model the geometric relations among both labeled and unlabeled data points. Through topology edges, the label information can propagate, regularizing the classifier *w.r.t.* the observation that data points scattering in a neighborhood region tend to have same labels.

However, all prior methods in this research line assume that the entire data stream is described by a fixed, known in-advance feature set. Once the feature space varies, the distance metric that these methods rely on for building the point-wise geometric relations fails to work. Our approach does not make this assumption and is thus more general.

The OVSIS problem

This paper uses the following conventions. Bold characters are used for matrices (*e.g.*, $\mathbf{A}$) and vectors (*e.g.*, $\mathbf{a}$). Script typeface is used for sets and spaces (*e.g.*, $\mathcal{A}$). $\|\cdot\|_1$ and $\|\cdot\|_2$ denote ℓ_1- and ℓ_2-norm, respectively and $\langle\cdot,\cdot\rangle$ denotes Euclidean inner product. We define an orthogonal projection as $\Pi_{\mathcal{A}}(\cdot) = \arg\min_{\mathbf{a}\in\mathcal{A}} \|\mathbf{a} - \cdot\|_2$, which takes in a vector and outputs the closest point to it in $\mathcal{A}$.

Problem Statement

Let $\{\mathbf{x}_t \mid t = 1, \ldots, T\}$ denote an input sequence, where $\mathbf{x}_t \in \mathbb{R}^{d_t}$ is a d_t-dimensional vector. At the round t, the learner f_t observes an instance $\mathbf{x}_t$ and gives its prediction. Only with a small probability p_l, the true label $y_t \in \{-1, +1\}$ is revealed. The learner updates f_{t+1} based on the instance $\mathbf{x}_t$ (and y_t if any). Our goal is to find a series of functions $f_1, \ldots, f_T$ that predicts the sequence accurately by minimizing the empirical risk, defined as:

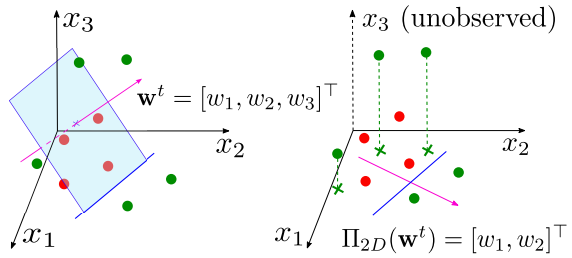
$$R(f) = \frac{1}{l}\sum_{t=1}^{T} \delta(y_t)\ell\big(y_t, f_t(\mathbf{x}_t)\big), \tag{1}$$

where l is the total number of labeled instances over T rounds. $\delta(y_t)$ denotes an indicator function, whose value is 1 if y_t is revealed and 0 otherwise. $\ell(\cdot,\cdot)$ is a convex loss function, such as square loss or logistic loss.

Challenges and Our Thoughts

From Eq. (1), we observe two challenges in solving the OVSIS problem, described as follows.

Challenge I: Feature Space Non-stationarity. The dimension of the input sequence changes over time, so should the learner. However, dynamically adapting the learner to a variable feature space for making accurate predictions is



(a) True Classification (b) Biased Projection

Figure 1: Classification errors caused by a feature space varied from 3D to 2D. (a) Green and red points are correctly separated by a hyperplane; (b) Green points are orthogonally projected onto the 2D space and the hyperplane collapses to a line, where the projected green points are misclassified.

non-trivial. Consider, for example, the feature space changes from 3D to 2D at two consecutive rounds, as shown in Figure 1. Suppose the learner is linear and can correctly classify 3D points at the round t, namely, $f_t(\mathbf{x}_t) = \text{sign}(\mathbf{w}_t^\top \mathbf{x}_t) = \text{sign}(w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3)$. Geometrically, the learner spans a decision (hyper)plane with $\mathbf{w}_t$ being its normal vector, as shown in Figure 1(a).

At the round $t+1$, the same data points arrive, but the feature x_3 does not accompany. To adapt f_t to predict these points, a widely-used method is called Lossless Homogenizing Conversion (LHC) (Masud et al. 2012; Gomes et al. 2013) which collapses the plane to a line, as shown in Figure 1(b). This projection by LHC, however, is biased. Comparing Figures 1(a) and 1(b), we can observe that LHC ignores the information conveyed by the feature x_3, so the points are orthogonally projected to 2D space of x_1, x_2. As a result, many points will cross the decision boundary and will be misclassified. The more intensively the feature space varies, the more likely this kind of misclassification happens. Capturing stationarity in such a variable feature space is a prerequisite for making accurate predictions.

Challenge II: Supervision Label Incompleteness. Existing online learners strictly require to see the true labels at every round, for obtaining informative gradient direction, which allows learners to be updated in an "asymptotically no-regret" fashion. In our OVSIS problem, however, true labels are given only occasionally. If no true label in a round, the learner is not updated according to Eq. (1), since no risk (loss) is suffered. Online learners will therefore take a much larger number of rounds (depending on how scarce the labels are) to converge, which incurs substantial prediction errors.

Our Ideas. Our ideas to the two challenges are two-fold. First, to tackle a variable feature space, we construct a *universal feature space* to capture stationary feature information. The universal feature space $\mathcal{U}_t \subseteq \mathbb{R}^{d_1} \cup \mathbb{R}^{d_2} \cup \ldots \cup \mathbb{R}^{d_t}$ at round t is a union of all features carried by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$. We learn a reconstructive mapping $\psi : \mathbb{R}^{d_t} \mapsto \mathcal{U}_t$, such that previously emerged features that are unobserved at the current round are recovered in $\mathcal{U}_t$. Recall the example in Figure 1(b). If x_3 can be recovered, the learner can then exploit the discriminant power of w_3 to make accurate prediction. The biased projection can thus be eliminated.

Second, given scarce labels, to expedite the learning progress, we leverage the unlabeled instances by discovering the geometric relations among all data points. The labeling information then can be propagated through the point-wise geometric relations as a regularization term, which encourages *label smoothness* over the universal feature space – any two data points that have similar predicted labels should be placed in a neighborhood region. Our objective is built upon these two ideas and will be elaborated in the next section.

The Objective Function

Our objective is framed in the scope of regularized risk minimization by taking the following norm:

$$\min_{f,\psi} \frac{1}{l} \sum_{t=1}^{T} \mathcal{L}(y_t, f(\psi(\mathbf{x}_t))) + \lambda_1 \mathcal{H}(\mathbf{x}_t, \psi) + \lambda_2 \Omega(f, \psi), \quad (2)$$

where the learner f and the reconstructive mapping ψ are jointly trained. The first term indicates the supervised loss, suffered by making prediction errors on the universal feature space, and the second term represents the reconstruction error, incurred when the universal feature space is not accurately constructed. The third term represents a manifold regularizer depending on f and ψ. The λ_1 and λ_2 are introduced to absorb the different scales among the three terms. Next, we scrutinize the details of the three terms in Eq. (2).

Construction of Universal Feature Space

The universal feature space $\mathcal{U}_t$ is learned by capturing the relatedness among features. In the OVSIS problem, as the data stream lives in a variable feature space, the features of any two consecutive instances can be different. Capturing a complex feature relatedness based on existing methods (Pan, Yang et al. 2010; Sun 2013) is hence unrealistic. Thus, we follow the spirit of (Hou, Zhang, and Zhou 2017; He et al. 2019) to represent the feature relatedness with linear models.

Let $\mathbf{u}_t := \psi(\mathbf{x}_t) \in \mathbb{R}^{|\mathcal{U}_t|}$ be the reconstruction of $\mathbf{x}_t$ in $\mathcal{U}_t$. We term $\mathbf{u}_t$ as a *universal (feature) vector*. Consider a mean field parameterized by $\mathbf{M} \in \mathbb{R}^{|\mathcal{U}_t| \times |\mathcal{U}_t|}$, over which the likelihood of observing a universal feature u_j in $\mathbf{u}_t$ from an original feature x_i in $\mathbf{x}_t$ is defined as follows:

$$Q(u_j \mid x_i, \mathbf{M}_{i,j}) = \frac{1}{2\sigma} \exp\Big(- \frac{|u_j - \mathbb{E}(u_j)|}{\sigma} \Big), \quad (3)$$

where $\mathbb{E}(u_j) = x_i \cdot \mathbf{M}_{i,j}$, representing a linear mapping relationship between the two features, and σ is a fixed variance of the Laplacian prior (Gerven et al. 2009).

Evidently, if the mapping ψ is correctly learned, the original feature values observed in $\mathbf{x}_t$ should be exactly recovered in $\mathbf{u}_t$. Maximizing the likelihood Q in Eq. (3) is thus equivalent to minimizing the reconstruction error *w.r.t.* $\mathbf{M}$:

$$\begin{aligned} \min_{\psi} \mathcal{H}(\mathbf{x}_t, \psi) &= \min_{\mathbf{M}_t} \|\mathbf{x}_t - \Pi_{\mathbb{R}^{d_t}}(\mathbf{u}_t)\|_2^2 \\ &= \min_{\mathbf{M}_t} \|\mathbf{x}_t - \frac{1}{d_t} \Pi_{\mathbb{R}^{d_t}}(\mathbf{M}_t^\top \mathbf{x}_t)\|_2^2, \end{aligned} \quad (4)$$

where $\mathbf{M}_t = \mathbf{I}_t \mathbf{M}$, and $\mathbf{I}_t \in \{0,1\}^{d_t \times |\mathcal{U}_t|}$ denotes an indicator matrix pinpointing which universal features in $\mathcal{U}_t$ are

also carried by $\mathbf{x}_t$. Represented by $\Pi_{\mathbb{R}^{d_t}}(\cdot)$ an operator that orthogonally projects $\mathbf{u}_t$ onto the $\mathbb{R}^{d_t}$-space, as defined.

With Eq. (4), we observe that our desired mapping $\psi(\cdot)$ is concretely realized by $(1/d_t)(\mathbf{M}_t^\top \cdot)$. Thus, given $\mathbf{x}_t$ at any round, we can project it to $\mathcal{U}_t$ to stabilize its dimension and obtain a universal vector $\mathbf{u}_t = (1/d_t)(\mathbf{M}_t^\top \mathbf{x}_t)$. Intuitively, a universal feature space, if well-constructed, should help a learner f make the least prediction errors. For a round where occasionally the label is given, we leverage this intuition to train the learner f with the mapping ψ jointly. The supervised loss term in Eq. (2) is then concretely defined as:

$$\min_{f,\psi} \mathcal{L}(y_t, f(\psi(\mathbf{x}_t))) = \min_{\mathbf{w}_t, \mathbf{M}_t} \frac{T}{l}\delta(y_t)\ell(y_t, \frac{1}{d_t}\mathbf{w}_t^\top \mathbf{M}_t^\top \mathbf{x}_t), \tag{5}$$

where the ratio T/l is an empirical yet unbiased estimate of the inverse label probability $1/p_l$ (Bubeck and Cesa-Bianchi 2012). We can ad hoc determine this ratio based on the rate at which human can label the data at hand.

Notably, a linear classifier $f(\cdot) = \mathbf{w}_t^\top \cdot$ trained in Eq. (5) is easily compatible with kernels (Kivinen, Smola, and Williamson 2004; Lu et al. 2016). As an example, we could define $f(\cdot) = \sum_{i=1}^{t-1} \alpha_i K(\mathbf{u}_i, \cdot)$, where $K(\cdot,\cdot)$ is a kernel over the universal feature vectors. This extension can empower our system to handle non-linear patterns in data, but the page limitation precludes a detailed discussion. We leave this valuable exploration as a future work.

Manifold Regularization with Sparsification

Ideally, if all instances are associated with the supervision labels, optimizing Eqs. (4) and (5) may suffice to yield an accurate learner defined on a well-constructed universal feature space. Unfortunately, in our OVSIS problem, labels are incomplete and scarce, resulting in sub-optimal solutions of both the learner f and the mapping ψ. Therefore, we desire an apparatus to exploit the abounded, unlabeled data points, such that the learning process is expedited and the obtained solutions are bettered for all rounds.

A plausible apparatus is to discover the geometric relations underlying the data and use it for regularizing f and ψ. This apparatus fails to work directly on the original data points because their dimensions are different and no distance metric is applicable to measure the geometric relations among those points. Fortunately, by projecting the points onto the learned universal feature space at each round, their dimensions become equal. Discovering the geometric relations among the projected points is hence much easier.

In this work, we trace the manifold structure underlying the universal feature space to embed the point-wise geometric relations. The key idea is straightforward – the data points which are predicted as having similar labels should be placed in a neighborhood region. However, a prominent problem in tracing the manifold is that, for each arriving data point, the distances are measured between this point and all points that have been previously arrived. The distances computation is cumbersome in the first place and, more importantly, storing all arrived data points will soon run out all memory in an online setting. To be practical, it is a must to represent the manifold in a sparse means.

Sparse Approximation via Random Projection. We employ the random projection tree (RP-Tree) (Hegde, Wakin, and Baraniuk 2008; Freund et al. 2008) to sparsify the manifold. Building an RP-Tree can be deemed as an online clustering process, where the size and the number of clusters grow over time to cover the entire manifold. The RP-Tree is updated at each round as follows. As instances arrive, they are sorted into the RP-Tree leaves based on their spatial relations; Once a leaf contains enough instances, it is cut by a hyperplane orienting a random direction, such that the instances in this leaf are split into two subspaces. The RP-Tree then grows by taking these two subspaces as two new leaves.

Without loss of generality, we model the instances fallen in the RP-Tree leaves with Gaussians. At a round t, suppose the RP-Tree has k leaves, where $k \ll t$ and the i^{th} leaf corresponds to a Gaussian centered at $\boldsymbol{\mu}_i \in \mathbb{R}^{|\mathcal{U}_t|}$. Here, we estimate $\{\boldsymbol{\mu}_i\}_{i=1}^k$ incrementally over time and ignore the covariance structures to take the computational advantage. Packing all the ideas above, the manifold regularizer with sparse approximation is analytically defined as follows.

$$\min_{f,\psi} \Omega(f,\psi) = \min_{\mathbf{w}_t, \mathbf{M}_t} \sum_{i=1}^{k} \Big(\frac{1+\hat{y}_{\boldsymbol{\mu}_i}\hat{y}_t}{4} D_{\mathcal{U}_t}(\boldsymbol{\mu}_i, \mathbf{x}_t)^2 + \frac{1-\hat{y}_{\boldsymbol{\mu}_i}\hat{y}_t}{4}\big(\max\{0, m - D_{\mathcal{U}_t}(\boldsymbol{\mu}_i, \mathbf{x}_t)\}\big)^2 \Big), \tag{6}$$

where $\hat{y}_t$ and $\hat{y}_{\boldsymbol{\mu}_i}$ denote the predicted labels of the input $\mathbf{x}_t$ and the i^{th} leaf representative $\boldsymbol{\mu}_i$, respectively, whose value assignments are decided by f and ψ together and shall be discussed in detail later on. Being $m > 0$ a margin, it defines a radius around any leaf representative. The distance from an arriving data point to any leaf representative in $\mathcal{U}_t$ is measured by $D_{\mathcal{U}_t}(\boldsymbol{\mu}_i, \cdot) = \|\boldsymbol{\mu}_i - \psi(\cdot)\|_2$.

Algorithm Design and Analysis

By plugging Eqs. (4), (5), and (6) back into Eq. (2), our solution to the OVSIS problem is unified into one objective of minimizing the regularized risk formulated as below.

$$R_t(f,\psi) := \frac{T}{l}\delta(y_t)\ell\big(y_t, f(\psi(\mathbf{x}_t))\big) + \lambda_1 \|\mathbf{x}_t - \Pi_{\mathbb{R}^{d_t}}(\psi(\mathbf{x}_t))\|_2^2 + \lambda_2 \sum_{i=1}^{k} \Big(\frac{1+\hat{y}_{\boldsymbol{\mu}_i}\hat{y}_t}{4} D_{\mathcal{U}_t}(\boldsymbol{\mu}_i, \mathbf{x}_t)^2 + \frac{1-\hat{y}_{\boldsymbol{\mu}_i}\hat{y}_t}{4}\big(\max\{0, m - D_{\mathcal{U}_t}(\boldsymbol{\mu}_i, \mathbf{x}_t)\}\big)^2 \Big), \tag{7}$$

where $R_t(f,\psi)$ denotes the *instantaneous regularized risk* at the t^{th} round and, to shorten notation, is written as R_t. Denoted by $f(\cdot) = \mathbf{w}_t^\top \cdot$ and $\psi(\cdot) = (1/d_t)((\mathbf{I}_t\mathbf{M})^\top \cdot)$ the classifier and the reconstructive mapping at the t^{th} round, respectively, with the subscript t omitted when the context is clear. We note a desirable property of R_t as follows.

Proposition 1. *Let $\mathcal{F} \subseteq \mathbb{R}^{|\mathcal{U}_t|}$ and $\Psi \subseteq \mathbb{R}^{|\mathcal{U}_t| \times |\mathcal{U}_t|}$ be two convex sets. R_t is bi-quasi-convex w.r.t. $f \in \mathcal{F}$ and $\psi \in \Psi$.*

A straightforward algorithm can be deduced from Proposition 1 by following the common steps of solving a bi-convex program (Gorski, Pfeuffer, and Klamroth 2007): (i)

Algorithm 1: The AGDES Algorithm

Initialize: $f \in \mathcal{F}$, $\psi \in \Psi$, $p = 0.5$, and $R_T^{\text{ori}} = R_T^{\text{rec}} = 0$.

Input : Parameters λ_1, λ_2, c, and γ.

for $t = 1, \ldots, T$ **do**

- Receive instance $\mathbf{x}_t$ and update RP-Tree ;
- Split $\bar{f}$ and $\tilde{f}$ from f based on $\mathbf{x}_t$;
- Predict the label as $\text{sign}(\hat{y}_t)$ using Eq. (8) ;
- Reveal the true label y_t with small probability p_l;
- $R_T^{\text{ori}} \mathrel{+}= \ell(y_t, \langle \bar{f}, \mathbf{x}_t \rangle)$, $R_T^{\text{rec}} \mathrel{+}= \ell(y_t, \langle \tilde{f}, \tilde{\mathbf{x}}_t \rangle)$;
- Suffer the instantaneous risk R_t using Eq. (7);
- Reweight coefficient p using Eq. (9) with $\tau = 2\sqrt{2 \ln 2/T}$;
- Update $f \leftarrow f - \eta_t \nabla_f R_t$ and $\psi \leftarrow \psi - \eta_t \nabla_\psi R_t$;
- Truncate $\mathbf{w}_t$ using Eq. (10) based on c and γ;

Decomposing Eq. (7) into two quasi-convex subproblems *w.r.t.* f and ψ, respectively; (ii) Alternating between the two subproblems, minimizing over one while keeping the other one fixed. We term this straightforward algorithm as *Naïve Alternating Gradient Descent* (NAGD).

Unfortunately, this NAGD algorithm fails to work perfectly in our context due to two issues. First, the new features would augment the hypothesis space of f and ψ ceaselessly, leading to inferior performance of naïve gradient descent optimizers. Second, as the universal feature space includes all emerged features, it can soon grow to an unmanageably large dimension. Storing all features in $\mathcal{U}_t$ and computing the reconstructive mapping ψ at every round lead to considerable memory and computational overheads. To overcome these two issues, we devise a novel algorithm and have its performance bound analyzed in sequence as follows.

Our Algorithm

We consider to 1) improve the learning performance via *ensemble prediction*; and 2) bound the maximal dimension of $\mathcal{U}_t$ via the *sparse truncation*, with the details as follow.

Ensemble Prediction. At the initial rounds, since few instances has been seen, the mapping ψ may not be sufficiently learned. Given an input $\mathbf{x}_t$, its reconstructed universal vector $\psi(\mathbf{x}_t)$ may be inaccurate, which can in turn negatively affect the classifier f being jointly trained on it. To aid this process, instead of simply defining the prediction of $\mathbf{x}_t$ in an inner product form, namely, $\langle f, \psi(\mathbf{x}_t) \rangle$, we separate the features being observed in the input $\mathbf{x}_t$ from those features being unobserved yet reconstructed in $\mathcal{U}_t$, ensembling two base predictions:

$$\hat{y}_t = p\langle \bar{f}, \mathbf{x}_t \rangle + (1-p)\langle \tilde{f}, \tilde{\mathbf{x}}_t \rangle, \tag{8}$$

where $\tilde{\mathbf{x}}_t = \psi(\mathbf{x}_t) \setminus \mathbf{x}_t \in \mathbb{R}^{|\mathcal{U}_t| - d_t}$ carries the reconstructed universal features that are not observed in the input $\mathbf{x}_t$. Define $\bar{f} \in \mathbb{R}^{d_t}$ and $\tilde{f} \in \mathbb{R}^{|\mathcal{U}_t| - d_t}$ as the classifiers corresponding to $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$, respectively, and $[\bar{f}, \tilde{f}] \equiv f$.

The intuition is to let the ensemble coefficient p decide the impact of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$ in making predictions, eliminating the prediction errors caused by noises in the reconstructed universal features. Denoted by $R_T^{\text{ori}} = \sum_{t=1}^T \delta(y_t) \ell(y_t, \langle \bar{f}, \mathbf{x}_t \rangle)$ and $R_T^{\text{rec}} = \sum_{t=1}^T \delta(y_t) \ell(y_t, \langle \tilde{f}, \tilde{\mathbf{x}}_t \rangle)$ the cumulative risks suffered by making predictions on $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$ over T rounds, respectively. At the round $T+1$, the coefficient p in Eq. (8) is updated based on the risk exponentials (Cesa-Bianchi and Lugosi 2006; Hou, Zhang, and Zhou 2017), *i.e.*,

$$p = e^{-\tau R_T^{\text{ori}}} / (e^{-\tau R_T^{\text{ori}}} + e^{-\tau R_T^{\text{rec}}}), \tag{9}$$

where $\tau = 2\sqrt{2 \ln 2/T}$ is a turned parameter.

Sparse Truncation. To restrict the dimension of $\mathcal{U}_t$ in a manageable size, we prune less informative features from $\mathcal{U}_t$. In the context of linear classifiers, the feature informativeness is associated with the weight vector $\mathbf{w}_t$ – the larger the value of w_i, the more informative the i^{th} feature (Li et al. 2017). To better distinguish the numerical values of the feature weights, we project the classifier $\mathbf{w}_t$ onto an ℓ_1-ball

$$\mathbf{w}_t \leftarrow \min\{1, c/\|\mathbf{w}_t\|_1\} \mathbf{w}_t, \tag{10}$$

such that most values of its elements are concentrated to its several largest elements. The positive parameter c controls the sparseness of the projected classifier during optimization. We could then enforce the learner to retain at most γ features by simply dropping the small weighted features (Wang et al. 2013; Zhang et al. 2015).

The ensemble prediction and sparse projection together deliver a novel algorithm, named *Alternating Gradient Descent with Ensemble prediction and Sparse truncation* (AGDES), with main steps summarized in Algorithm 1.

Performance Bound

We borrow the *regret* from online convex programming (Zinkevich 2003) to analyze the performance of our AGDES algorithm. Let (f^*, ψ^*) be the static optimum over T rounds, namely, $f^*, \psi^* = \arg\min_{f,\psi} \sum_{t=1}^T R_t$, which can be obtained in a hindsight only. The regret is bounded by the difference between the empirical risk suffered by our AGDES algorithm and that by this offline optimal solution.

Theorem 1 (Asymptoticity of AGDES)**.**

$$\sum_{t=1}^T R_t - \sum_{t=1}^T R^* \leq \sqrt{2T \ln 2} \tag{11}$$

with $R^* = R_t(f^*, \psi^*)$ being the minimal risk which applies the static optimum to the input $\mathbf{x}_t$ at every round.

It follows that $\lim_{T \to \infty} (\sqrt{2T \ln 2}/T) = 0$. Therefore, compared to a hindsight optimum, AGDES is *asymptotically no-regret*. To see the superiority of the AGDES algorithm explicitly, we also analyze the performance of NAGD, which can be deemed as a reduced version of AGDES by taking off the ensemble prediction and sparse truncation.

Theorem 2 (Regret Bound of NAGD)**.**

$$\sum_{t=1}^T R_t - \sum_{t=1}^T R^* \leq (D_{\mathcal{F}} + D_\Psi)\sqrt{T} + (\sqrt{T} - \frac{1}{2})G, \tag{12}$$

where $D_{\mathcal{F}}$ and D_Ψ are the diameters of $\mathcal{F}$ and Ψ, respectively. $G := \max_{f_t \in \mathcal{F}} \|\nabla_{f_t} R_t\|_2^2 + \max_{\psi_t \in \Psi} \|\nabla_{\psi_t} R_t\|_2^2$.

Dataset	#Inst.	#Feat.	Dataset	#Inst.	#Feat.
ionosphere	351	34	kr-vs-kp	3196	36
wdbc	569	30	HAPT	10,929	561
australian	690	14	magic04	19,020	10
diabetes	768	8	IMDB	25,000	7500
dna	949	180	CCYS	33,000	355

Table 1: Statistic information of the 10 datasets

So by Theorem 1 and Theorem 2, we remark:

Remark 1. *AGDES can provably perform better and more robust than NAGD because it enjoys a tighter regret bound, which is independent to the scalar terms (whose magnitudes grow over time), being invariant to a variable feature space.*

Experiments

This section aims to experimentally validate whether our solution is viable and effective to the OVSIS problem. Our evaluations are conducted on 10 datasets, including 8 from the UCI repository (Dua and Karra Taniskidou 2017) and 2 from the datasets of IMDB (Maas et al. 2011) and CCYS (He et al. 2021). The evaluated datasets span diverse application domains, such as economy, education, bioinformatics, *etc*. Table 1 summarizes their statistics.

For the UCI datasets, we simulate a variable feature space by following the setting as in (He et al. 2019). That is, we randomly remove at most 50% of features from each instance, while leaving the remaining as the observed features. For IMDB and CCYS, the instances are already described by various sets of features. Details of the two datasets are referred to their respective literatures.

To simulate label scarceness, we randomly mask 70% labels (*i.e.*, $p_l = 0.3$). The predicted labels are saved for all rounds and, once the input sequence ends, all true labels are revealed and the accuracy is calculated by comparing how many the predicted labels are the same as the true labels.

We take three online learning competitors, OCO (Zinkevich 2003), OMR (Goldberg, Li, and Zhu 2008) and OCDS (He et al. 2019), along with our proposed AGDES for experiments. To validate the tightness of our bound (Theorem 1), NAGD is also evaluated. Our evaluation aims to answer the following four research questions.

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

Table 2 presents the results of performance comparison in terms of classification accuracy. From the table, We have three observations. *First*, our AGDES achieves the best performance with an averaged accuracy of 74.3%. The statistical evidence proves that it outperforms all the compared methods across 10 datasets, with a 14% performance improvement in average. *Second*, compared to OMR and OCDS, which respectively address label scarcity and feature space dynamics, our AGDES and NAGD outperform OMR by ratios of 12.1% and of 6.6%, respectively, and excel OCDS by ratios of 12.3% and of 6.9%, respectively. This result substantiates that the two building blocks in our approach, *i.e.*, the universal feature space construction and the manifold regularization, are indispensable and can work together to make online learners more flexible and applicable. *Third*, OCO performs the worst across datasets. The reason is that it makes prediction based on the observed features only while requiring labels at all rounds for updating its learner. Our approach remarkably exceeds its performance by addressing both the feature space dynamics and label incompleteness. In particular, our AGDES and NAGD algorithms win OCO in 19 out of 20 settings.

Q2. *Can manifold regularizer improve the learning accuracy when the supervision labels are scarce?*

The answer is revealed in Figure 2, which presents the trending of *average cumulative risk* (ACR), defined as $ACR = (1/T)\sum_{i=1}^{T} \|y_t - \hat{y}_t\|_2^2$ at the round T. Essentially, the steeper the ACR trend decreases, the faster a corresponding method converges. From the results, we observe that OCDS, which does not tackle label scarcity, suffers from flat convergence rate and ends up with sub-optimal solutions.

The performance of OMR, which also applies the manifold regularizer, is mixed. On one hand, it enjoys a steeper learning curve than OCDS in several datasets. On the other hand, OMR's results are inferior to OCDS' in two datasets: dna (Figure 2b) and IMDB (Figure 2c). The reason is that these datasets are high-dimensional, with each having a much larger number of observed features in total. The input sequences in these datasets are thus described by a more intensively variable feature space, which OMR cannot handle well as it prescribes a fixed set of features.

We observe that for all datasets, both our AGDES and NAGD outperform OCDS in terms of both convergence rate and the final accuracy. This result witnesses the usefulness of manifold regularizer in improving the prediction performance when the labeling information is incomplete.

Q3. *Does AGDES outperform NAGD? This directly tests the tightness of our bound in Theorem 1.*

From Table 2 and Figure 2, we observe that 1) AGDES enjoys a higher accuracy in almost all datasets than NAGD; 2) AGDES converges to significantly lower ACR values; and 3) in most learning rounds, AGDES adheres to the ACR values that are smaller than NAGD. These observations indicate that AGDES can always trace the better feature subset and assign larger weights to the subset comprising either the original or the recovered features in accordance with the predictions they made via the ensemble prediction strategy. To conclude, AGDES indeed holds a tighter performance bound, thereby empirically performing better.

Q4. *How robust is our algorithm to the parameters?*

The objective Eq. (7) is governed by two parameters λ_1 and λ_2 that need to be determined in an ad hoc way. We investigate the impact of parameter values on our approach.

The parameter λ_1 decides how slack the reconstruction errors are tolerated. The parameter λ_2 controls how strongly the learner preserves the manifold structure. We grid search λ_1 in $\{1\text{e}{-3}, 5\text{e}{-3}, \ldots, .075\}$ and λ_2 in

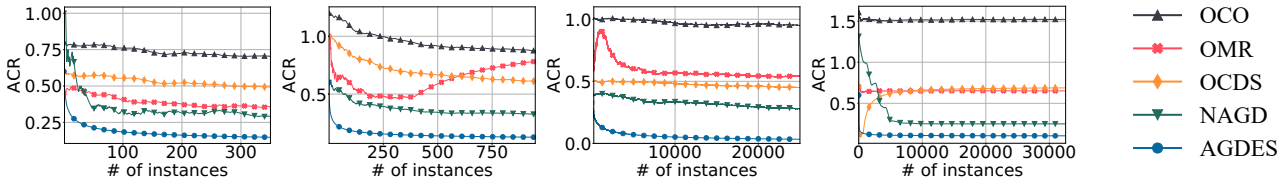
Dataset	OCO	OMR	OCDS	NAGD	**AGDES**
ionosphere	$.616 \pm .012$ •	$.742 \pm .017$ •	$.703 \pm .012$ •	$.772 \pm .010$	$\mathbf{.795 \pm .015}$
wdbc	$.700 \pm .018$ •	$.722 \pm .007$ •	$.751 \pm .011$	$.765 \pm .008$	$\mathbf{.783 \pm .007}$
australian	$.623 \pm .013$ •	$.729 \pm .006$	$.641 \pm .008$ •	$.732 \pm .008$	$\mathbf{.742 \pm .010}$
diabetes	$.612 \pm .009$ •	$.705 \pm .004$	$.654 \pm .006$ •	$.707 \pm .009$	$\mathbf{.725 \pm .008}$
dna	$.510 \pm .006$ •	$.528 \pm .002$ •	$.647 \pm .005$ •	$.695 \pm .005$ •	$\mathbf{.741 \pm .004}$
kr-vs-kp	$.608 \pm .006$ •	$.695 \pm .005$ •	$.667 \pm .008$ •	$.701 \pm .009$ •	$\mathbf{.763 \pm .007}$
HAPT	$.657 \pm .003$ •	$.683 \pm .001$ •	$.703 \pm .003$ •	$.714 \pm .003$ •	$\mathbf{.766 \pm .002}$
magic04	$.611 \pm .005$ •	$.725 \pm .003$	$.634 \pm .003$ •	$\mathbf{.744 \pm .004}$	$.732 \pm .003$
IMDB	$.555 \pm .017$ •	$.575 \pm .015$ •	$.617 \pm .006$ •	$.633 \pm .009$ •	$\mathbf{.706 \pm .009}$
CCYS	$.510 \pm .021$ •	$.521 \pm .010$ •	$.592 \pm .006$ •	$.601 \pm .007$ •	$\mathbf{.672 \pm .006}$
NAGD: w/t/l	9 / 1 / 0	4 / 6 / 0	5 / 5 / 0	—	1 / 4 / 5
AGDES: w/t/l	10 / 0 / 0	7 / 3 / 0	9 / 1 / 0	5 / 4 / 1	—

Table 2: Experimental results (Mean Accuracy $\pm$ Standard Deviation) on 10 datasets, where a random permutation is applied on the input sequence to repeat each experiment 5 times. The best results are bold. • indicates our approach has a statistically significant better performance than the compared methods (hypothesis supported by *paired t-tests* at 95% significance level). The win/tie/loss counts for our NAGD and AGDES algorithms are summarized in the last two rows.

$\{1.5\text{e}{-5}, 5\text{e}{-5}, \ldots, .085\}$, and present corresponding accuracy of AGDES in Figure 3. We observe that the optimal value of λ_2 varies across different datasets. Fortunately, as applied in many other manifold learning methods, we can tune this parameter by buffering a small validation set from the data stream (Zhu et al. 2018; Ma et al. 2018; Guo, Mao, and Zhang 2019). Indeed, making manifold regularization parameters self-adaptive is under active research but remains an open problem (Kang, Peng, and Cheng 2017; Zhao et al. 2017). We leave the exploration of this meaningful problem in the future work. Encouragingly, the optimal value of λ_1 on different datasets stably scatters around 0.01 (the tick around the middle of λ_1-axis). This robustness gives us the confidence on adopting such an empirical value in practice.

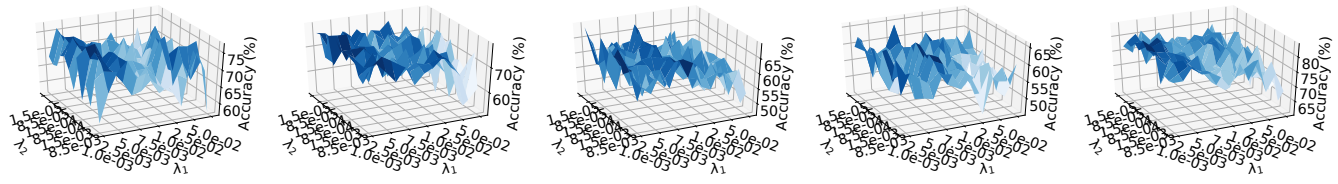
Conclusion

This paper aims to push the boundaries of online learning as the investigated problem assumes no restrictions on feature space dynamics or training instance labeling, while previous efforts always place restrictions on either or both. We thus make online learning techniques more applicable to real-world applications. Our solution constructs a universal feature space for capturing stationarity feature information and exploits unlabeled instances to expedite the learning process by uncovering the manifold structure underlying data. A theoretical analysis reveals performance advantages of our solution, and extensive empirical evaluation further substantiates the viability and effectiveness of our proposal.



(a) ionosphere (b) dna (c) IMDB (d) CCYS Legend

Figure 2: The trends of average cumulative risk (ACR) of OCO, OMR, OCDS, NAGD, and AGDES.



(a) ionosphere (b) dna (c) IMDB (d) CCYS (e) HAPT

Figure 3: Surface of classification accuracy *w.r.t.* λ_1 and λ_2. The darker the color, the higher the corresponding accuracy.

Acknowledgments

We thank AAAI 2021 reviewers for their constructive feedback. We thank WWW 2021 reviewers for their constructive feedback. This work was supported in part by the US National Science Foundation (NSF) under grants 1652107, 1763620, 1948374, and 1750886. The work of Dr. Xindong Wu was supported by the National Natural Science Foundation of China (NSFC) under grant 91746209 and the National Key Research and Development Program of China under grant 2016YFB1000901. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agencies.

References

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

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning* 5(1): 1–122. ISSN 1935-8237. doi:10.1561/2200000024. URL http://dx.doi.org/10.1561/2200000024.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press. doi:10.1017/CBO9780511546921.

Chu, W.; Zinkevich, M.; Li, L.; Thomas, A.; and Tseng, B. 2011. Unbiased online active learning in data streams. In *KDD*, 195–203.

Dua, D.; and Karra Taniskidou, E. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

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.

Freund, Y.; Dasgupta, S.; Kabra, M.; and Verma, N. 2008. Learning the structure of manifolds using random projections. In *NeurIPS*, 473–480.

Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; and Bouchachia, A. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46(4): 1–37.

Gerven, M. V.; Cseke, B.; Oostenveld, R.; and Heskes, T. 2009. Bayesian source localization with the multivariate Laplace prior. In *NeurIPS*, 1901–1909.

Goldberg, A. B.; Li, M.; and Zhu, X. 2008. Online manifold regularization: A new learning setting and empirical study. In *ECML-PKDD*, 393–407. Springer.

Goldberg, A. B.; Zhu, X.; Furger, A.; and Xu, J.-M. 2011. Oasis: Online active semi-supervised learning. In *AAAI*, 362–367.

Gomes, J. B.; Gaber, M. M.; Sousa, P. A.; and Menasalvas, E. 2013. Mining recurring concepts in a dynamic feature space. *IEEE Transactions on Neural Networks and Learning Systems* 25(1): 95–110.

Gorski, J.; Pfeuffer, F.; and Klamroth, K. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical methods of operations research* 66(3): 373–407.

Guo, H.; Mao, Y.; and Zhang, R. 2019. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3714–3722.

Hao, S.; Zhao, P.; Lu, J.; Hoi, S. C.; Miao, C.; and Zhang, C. 2016. Soal: Second-order online active learning. In *ICDM*, 931–936. 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*, 2491–2497.

He, Y.; Wu, B.; Wu, D.; Beyazit, E.; Chen, S.; and Wu, X. 2021. Toward Mining Capricious Data Streams: A Generative Approach. *IEEE Transactions on Neural Networks and Learning Systems* 32(3): 1228–1240.

Hegde, C.; Wakin, M.; and Baraniuk, R. 2008. Random projections for manifold learning. In *NeurIPS*, 641–648.

Hou, B.-J.; Zhang, L.; and Zhou, Z.-H. 2017. Learning with feature evolvable streams. In *NeurIPS*, 1417–1427.

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

Kang, Z.; Peng, C.; and Cheng, Q. 2017. Clustering with adaptive manifold structure learning. In *ICDE*, 79–82. IEEE.

Kivinen, J.; Smola, A. J.; and Williamson, R. C. 2004. Online learning with kernels. *IEEE transactions on signal processing* 52(8): 2165–2176.

Krempl, G.; Žliobaite, I.; Brzeziński, D.; Hüllermeier, E.; Last, M.; Lemaire, V.; Noack, T.; Shaker, A.; Sievi, S.; Spiliopoulou, M.; et al. 2014. Open challenges for data stream mining research. *ACM SIGKDD explorations newsletter* 16(1): 1–10.

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

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2017. Feature Selection: A Data Perspective. *ACM Computing Surveys (CSUR)* 50(6): 1–45.

Lu, J.; Hoi, S. C.; Wang, J.; Zhao, P.; and Liu, Z.-Y. 2016. Large scale online kernel learning. *The Journal of Machine Learning Research* 17(1): 1613–1655.

Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; and Zhang, G. 2018. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* 31(12): 2346–2363.

Lu, J.; Zhao, P.; and Hoi, S. C. 2016. Online passive-aggressive active learning. *Machine Learning* 103(2): 141–183.

Ma, J.; Wu, J.; Zhao, J.; Jiang, J.; Zhou, H.; and Sheng, Q. Z. 2018. Nonrigid point set registration with robust transformation learning under manifold regularization. *IEEE transactions on neural networks and learning systems* 30(12): 3584–3597.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *ACL*, 142–150. Portland, Oregon, USA. URL http://www.aclweb.org/anthology/P11-1015.

Masud, M. M.; Chen, Q.; Gao, J.; Khan, L.; Han, J.; and Thuraisingham, B. 2010. Classification and novel class detection of data streams in a dynamic feature space. In *ECML-PKDD*, 337–352. Springer.

Masud, M. M.; Chen, Q.; Khan, L.; Aggarwal, C. C.; Gao, J.; Han, J.; Srivastava, A.; and Oza, N. C. 2012. Classification and adaptive novel class detection of feature-evolving data streams. *IEEE Transactions on Knowledge and Data Engineering* 25(7): 1484–1497.

Mohamad, S.; Bouchachia, A.; and Sayed-Mouchaweh, M. 2018. A bi-criteria active learning algorithm for dynamic data streams. *IEEE transactions on neural networks and learning systems* 29(1): 74–86.

Pan, S. J.; Yang, Q.; et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10): 1345–1359.

Sun, S. 2013. A survey of multi-view machine learning. *Neural Computing and Applications* 23(7-8): 2031–2038.

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

Wang, J.; Zhao, P.; Hoi, S. C.; and Jin, R. 2013. Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering* 26(3): 698–710.

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

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, N.; Zhang, L.; Du, B.; Zhang, Q.; You, J.; and Tao, D. 2017. Robust dual clustering with adaptive manifold regularization. *IEEE Transactions on Knowledge and Data Engineering* 29(11): 2498–2509.

Zhu, B.; Liu, J. Z.; Cauley, S. F.; Rosen, B. R.; and Rosen, M. S. 2018. Image reconstruction by domain-transform manifold learning. *Nature* 555(7697): 487–492.

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