Distribution-Based Semi-Supervised Learning for Activity Recognition

Hangwei Qian,[†‡§] Sinno Jialin Pan,[†] Chunyan Miao[†‡£]
[†]School of Computer Science and Engineering
[‡]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
[§]Interdisciplinary Graduate School
[£]Alibaba-NTU Singapore Joint Research Institute
Nanyang Technological University, Singapore
qian0045@e.ntu.edu.sg, {sinnopan, ascymiao}@ntu.edu.sg

Abstract

Supervised learning methods have been widely applied to activity recognition. The prevalent success of existing methods, however, has two crucial prerequisites: proper feature extraction and sufficient labeled training data. The former is important to differentiate activities, while the latter is crucial to build a precise learning model. These two prerequisites have become bottlenecks to make existing methods more practical. Most existing feature extraction methods highly depend on domain knowledge, while labeled data requires intensive human annotation effort. Therefore, in this paper, we propose a novel method, named Distribution-based Semi-Supervised Learning, to tackle the aforementioned limitations. The proposed method is capable of automatically extracting powerful features with no domain knowledge required, meanwhile, alleviating the heavy annotation effort through semi-supervised learning. Specifically, we treat data stream of sensor readings received in a period as a distribution, and map all training distributions, including labeled and unlabeled, into a reproducing kernel Hilbert space (RKHS) using the kernel mean embedding technique. The RKHS is further altered by exploiting the underlying geometry structure of the unlabeled distributions. Finally, in the altered RKHS, a classifier is trained with the labeled distributions. We conduct extensive experiments on three public datasets to verify the effectiveness of our method compared with state-of-the-art baselines.

Introduction

Human activity recognition has spurred a great deal of interest with a wide spectrum of real-world applications, such as security, personalized health monitoring and assisted living (Janidarmian et al. 2017; Bulling, Blanke, and Schiele 2014; Lara and Labrador 2013; Frank, Mannor, and Precup 2010; Avci et al. 2010). Generally, there are two types of scenarios: wireless-sensor-based and video-based. In this work, we focus on wireless-sensor-based activity recognition scenarios. In these scenarios, the data is often in the form of a continuous multivariate time series from multiple sensors. Therefore, the data needs to be divided into segments first, each of which corresponding to a single label. Traditionally, it requires intensive annotation effort with the starting and ending times of each activity. Further, in order to increase the expressiveness of data, feature extraction is commonly applied to each segment. Extracted features are then fed into a classifier to recognize different activities. Note that feature extraction and large amount of labeled training data are crucial in the process, which are discussed in detail hereinafter.

It is well-known that good features can help to discriminate different classes of activities, by increasing the expressiveness of each activity. Generally, feature extraction approaches can be classified into two categories: statistical and structural (Lara and Labrador 2013). Structural features take into account the overall information of the data. For example, SAX method transforms continuous data into discrete symbolic strings (Lin et al. 2007); ECDF method preserves the overall shape and spatial information of time series data (Hammerla et al. 2013; Plötz, Hammerla, and Olivier 2011). Therefore, domain knowledge is highly required for structural features. Statistical features, on the other hand, aim to capture statistical information underlying each time-series segment. There are also around twenty commonly used handcrafted statistical features which are proven to be beneficial practically, including orders of moments (mean, variance, skewness, etc), median, etc (Janidarmian et al. 2017). Major limitations of statistical features are the flexibility of handcrafted features and the involvement of domain knowledge. Recently, Qian, Pan, and Miao (2018) proposed the SMM$_{AR}$ approach, which is capable of automatically extracting all orders of moments as statistical features for activity recognition.

Though SMM$_{AR}$ is able to systematically extract powerful statistical features, as a supervised learning based method, it requires a plethora of labeled data for training. Note that label annotation on a large-scale dataset on sensor readings is a costly process. Therefore, growing research interests have been focused on exploring the trade-off between label ambiguity and human annotation effort. Some researchers focus on efficient annotation strategies to reduce labeling effort, including offline and online strategies (Stikic et al. 2011), such as experience sampling, self-recall and video recording. There also exist several research works applying semi-supervised learning (Zhu 2005) for activity recognition by exploiting unlabeled data, which is supposed to be easy to collect with very low cost, to learn a precise classifier even with a limited number of labeled data (Guan et al. 2007; Stikic, Larlus, and Schiele 2009; Stikic et al. 2011). Most existing semi-supervised learning

methods adopt handcrafted features.

In this paper, we propose a novel semi-supervised learning method, namely Distribution-based Semi-Supervised Learning (DSSL), to free the intensive effort on feature engineering by using the kernel mean embedding technique for distributions (Berlinet and Thomas-Agnan 2011). To elaborate, we treat data stream of sensor readings received in a period as a probability distribution. Modeling input instances as probability distribution is a new and promising machine learning paradigm, and some methods have been successfully developed in the supervised learning manner, e.g., Support Measure Machines (SMMs) (Muandet et al. 2012; 2017). Recently, Qian, Pan, and Miao (2018) proposed a framework based on SMMs for activity recognition, which is known as SMM$_{AR}$. A major advantage of SMM$_{AR}$ over other supervised learning methods for activity recognition is the capability of automatically extracting all the orders of statistical moments as features to represent each input instance. Our proposed method, DSSL, is an extension of SMM$_{AR}$ in the semi-supervised learning manner. Compared with SMM$_{AR}$ and other supervised or semi-supervised learning methods for activity recognition, our contributions are 4-fold:

- Compared with other supervised or semi-supervised learning methods, DSSL is able to represent each instance, i.e., data stream of a period, using all the orders of statistical moments implicitly and automatically, which contains rich information to distinguish activities.
- Compared with SMM$_{AR}$, DSSL relaxes its full supervision assumption, and is able to exploit unlabeled instances to learn an underlying data structure. With the learned structure and a few labeled instances, DSSL is able to learn a precise classifier for activity recognition.
- Most existing works on learning with distributions are supervised. To the best of our knowledge, DSSL is the first attempt on semi-supervised learning with distributions. Moreover, we provide theoretical analysis proving that DSSL is valid for semi-supervised learning in a reproducing kernel Hilbert space (RKHS).
- Extensive experiments are conducted to demonstrate the superior performance of DSSL over a number of state-of-the-art baselines.

Other Related Work

Limited labeled training data is insufficient to train a good classifier due to the cold start problem of supervised learning. Semi-supervised learning approaches are appealing in practice since they require only a small fraction of labeled training data with a large amount of easily obtained unlabeled data (Chapelle, Schlkopf, and Zien 2010; Zhu 2005). Among existing semi-supervised learning approaches, manifold regularization (Sindhwani, Niyogi, and Belkin 2005) and wrapping kernels using point cloud (Belkin, Niyogi, and Sindhwani 2006) are two classic methods, which incorporates the manifold structure underlying both unlabeled and labeled data into the learning of Support Vector Machines (SVMs).

In the context of activity recognition, Stikic, Larlus, and Schiele (2009) proposed a multi-graph based semi-supervised approach named GLSVM, where each graph propagates different information of activities. Different graphs are then combined to improve label propagation in graphs. After that, an SVM classifier is trained by using both the initially labeled training data and the propagated labels. Matsushige, Kakusho, and Okadome (2015) proposed a semi-supervised kernel logistic regression method for activity recognition, denoted by SSKLR, which extends kernel logistic regression into semi-supervised fashion, and solves the problem by the Expectation-Maximization algorithm. Yao et al. (2016) proposed a robust graph-based semi-supervised method named RSAR to tackle the intra-class variability in activities across different subjects. The RSAR method extracts the intrinsic shared subspace structures from activities with the assumption that intrinsic relationships have invariant properties thus are less sensitive with varying subjects. In (Nazábal et al. 2016), a new Bayesian model is proposed to tackle the scenario with a very low number of sensors. The dynamic nature of human activities are further modeled as a first-order homogeneous Markov chain. Our proposed DSSL is a unified framework that naturally inherits the spirit of learning from distributions and manifold learning.

Preliminaries

Support Measure Machines In supervised learning with distributions, we are given a set of labeled data $\{\mathbf{X}_i, y_i\}_{i=1}^n$, where $\mathbf{X}_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$ and $n_i's$ may vary across different $\boldsymbol{x}_i$. The goal is to learn a classifier f to map $\{\mathbf{X}_i\}$'s to $\{y_i\}$'s. In SMMs (Muandet et al. 2012), each $\mathbf{X}_i$ is mapped to a functional in a RKHS $\mathcal{H}$ via kernel mean embedding (Berlinet and Thomas-Agnan 2011) as $\boldsymbol{\mu}_{\mathbb{P}_i} = \mathbb{E}_{\boldsymbol{x}_{ij}\sim\mathbb{P}_i}[k(\boldsymbol{x}_{ij}, \cdot)]$, where $k(\cdot,\cdot)$ is a characteristic kernel associated with the RKHS $\mathcal{H}$. It has been proven that if the kernel is characteristic, then an arbitrary probability distribution $\mathbb{P}_i$ is uniquely represented by an element $\boldsymbol{\mu}_{\mathbb{P}_i}$ in the RKHS, which implicitly captures all orders of statistical moments of $\mathbf{X}_i$.

The inner product, i.e., a linear kernel, of two distributions, which measures their similarity, can be defined as $\langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle = \frac{1}{n_i n_j} \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} k(\mathbf{x}_{ia}, \mathbf{x}_{jb})$. One can also define a nonlinear kernel of $\boldsymbol{\mu}_{\mathbb{P}_i}$ and $\boldsymbol{\mu}_{\mathbb{P}_j}$ to capture their nonlinear relationships via

$$\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j})_{\tilde{\mathcal{H}}} = \langle \psi(\boldsymbol{\mu}_{\mathbb{P}_i}), \psi(\boldsymbol{\mu}_{\mathbb{P}_j}) \rangle, \tag{1}$$

where $\tilde{k}(\cdot,\cdot)$ is the nonlinear kernel induced by the nonlinear feature map $\psi(\cdot)$, and $\tilde{\mathcal{H}}$ is the corresponding RKHS.

To train a classifier from $\{\mathbf{X}_i\}$'s to $\{y_i\}$'s, SMMs define the optimization problem by learning $f \in \tilde{\mathcal{H}}$ that minimizes the following regularized risk functional

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\mu}_{\mathbb{P}_i}, y_i, f) + \Omega(\|f\|_{\tilde{\mathcal{H}}}), \tag{2}$$

where $\ell(\cdot)$ is the loss function and $\Omega(\cdot)$ is the regularization term. Note that $\tilde{\mathcal{H}} = \mathcal{H}$ if $\tilde{k}$ is linear.

Random Fourier Features Approximation The kernel embedding technique of distributions used in SMMs is computationally expensive as it requires to compute kernel matrices. This makes it impractical in some real-world applications when the size of the dataset is large. To scale up SMMs, Qian, Pan, and Miao (2018) proposed an accelerated version using Random Fourier Features to construct an explicit feature map instead of using the kernel trick. Based on Bochner's Theorem (Rahimi and Recht 2007), a continuous, shift-invariant positive definite kernel $k(\mathbf{x}, \mathbf{x}')$ can be linearized by using the randomized feature map $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^D$ as

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \approx \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}'), \tag{3}$$

where the inner product of explicit feature maps can uniformly approximate the kernel values without the kernel trick, and the random Fourier features are generated by:

$$z_w(\mathbf{x}) = \sqrt{2} cos(w^\top \mathbf{x} + b), \tag{4}$$

where $w \sim p(w)$, which is $k(\cdot, \cdot)$'s Fourier transform distribution on $\mathbb{R}^D$, and b is sampled uniformly from $[0, 2\pi]$. It can be proven that $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}(z_w(\mathbf{x})^\top z_w(\mathbf{x}'))$ for all $\mathbf{x}$ and $\mathbf{x}'$. In practice, D can be small, which enables SMMs to handle large-scale datasets.

The Proposed Methodology

Problem Statement

In our project setting of activity recognition, we are given a set of l labeled segments data $\{\mathbf{X}_i, y_i\}_{i=1}^{l}$, and a set of $u = n - l$ unlabeled segments $\{\mathbf{X}_i\}_{i=l+1}^{i=n}$ as training data obtained by applying segmentation methods on the raw data, where $\mathbf{X}_i = [\mathbf{x}_{i1} \ ... \ \mathbf{x}_{in_i}] \in \mathbb{R}^{d \times n_i}$, $y_i \in \{1, ..., L\}$, $l \ll u$, and n_i may vary across different segments. The goal is to make use of both labeled and unlabeled segments to learn a classifier from each segment $\mathbf{X}$ to its corresponding label y.

Following (Qian, Pan, and Miao 2018), each segment $\mathbf{X}_i$, including both labeled and unlabeled, is treated as a *sample* of n_i data points drawn from an unknown distribution $\mathbb{P}_i$. Kernel mean embedding is then applied to map each $\mathbf{X}_i$ to an element $\boldsymbol{\mu}_{\mathbb{P}_i}$ in a RHKS. In practice, to make the learning process more efficient, random Fourier features are used to approximate the nonlinear feature map induced by the kernel of the RKHS via $\boldsymbol{\mu}_{\mathbb{P}_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}(\mathbf{x}_{ij})$. where $\boldsymbol{\mu}_{\mathbb{P}_i} \in \mathbb{R}^D$. Therefore, our goal becomes to learn a classifier $f : \boldsymbol{\mu}_{\mathbb{P}} \to y_i$ from $\{\boldsymbol{\mu}_{\mathbb{P}_i}, y_i\}_{i=1}^{l}$ and $\{\boldsymbol{\mu}_{\mathbb{P}_i}\}_{i=l+1}^{i=n}$.

Distribution-based Semi-Supervised Learning

Borrowing the idea from manifold regularization (Belkin, Niyogi, and Sindhwani 2006) and the technique on warping data-dependent kernels (Sindhwani, Niyogi, and Belkin 2005), we aim to incorporate the underlying manifold structure of both labeled and unlabeled data into the learning of a classifier via warping a RKHS. Specifically, we wrap the RKHS $\tilde{\mathcal{H}}$ defined in (1) to another RKHS $\breve{\mathcal{H}}$ by leveraging unlabeled training segments or distributions to reflect the underlying geometry of $\{\psi(\boldsymbol{\mu}_{\mathbb{P}_i})\}$'s. Notations on different kernels and their corresponding RKHSs used in this paper are summarized in Table 1. The new RKHS $\breve{\mathcal{H}}$ is associated with the new kernel $\breve{k}$, which is data-dependent for semi-supervised learning. We will discuss how to achieve the kernel as well as the resulting new space later. Here, we assume the new kernel $\breve{k}$ is constructed, then the revised optimization problem over $\breve{\mathcal{H}}$ is formulated as

Table 1: Notations of different kernels used in the paper

Kernel	Space	Descriptions
k	$\mathcal{H}$	kernel mean embedding of distributions
$\tilde{k}$	$\tilde{\mathcal{H}}$	kernel on the embedded distributions
$\breve{k}$	$\breve{\mathcal{H}}$	data-dependent kernel constructed based on $\tilde{k}$ for semi-supervised learning

$$f^* = \arg\min_{f \in \breve{\mathcal{H}}} \frac{1}{l} \sum_{i=1}^{l} \ell(\boldsymbol{\mu}_{\mathbb{P}_i}, y_i, f) + \|f\|_{\breve{\mathcal{H}}}^2, \tag{5}$$

where $\ell(\cdot)$ is the loss function. Note the objective function looks similar to that in the supervised learning setting in (2). However, in (5) the RKHS, where the functional to be optimized is $\breve{\mathcal{H}}$, which is influenced by both labeled and unlabeled distributions, while the RKHS in (2) is $\tilde{\mathcal{H}}$, which is defined by labeled distributions only. The new optimization probelm raises a potential problem: f is to be learned in $\breve{\mathcal{H}}$, while the input space of $\boldsymbol{\mu}_{\mathbb{P}_i}$ is $\mathcal{H}$. As these RKHSs are not the same, how to calculate the loss function remains a problem. To sum up, in order to solve the optimization problem (5), three crucial questions need to be answered:

- How to construct the data-dependent kernel $\breve{k}$ by incorporating unlabeled training data?
- Is the new space $\breve{\mathcal{H}}$ valid?
- How to calculate the loss function given $\boldsymbol{\mu}_{\mathbb{P}} \in \mathcal{H}$ and $f \in \breve{\mathcal{H}}$ are not in the same space?

In the following, we investigate the questions one by one.

1) Construction of the Data-dependent Kernel $\breve{k}$ Since unlabeled data may shed light on the underlying structure and manifolds of all data, now the problem becomes how to appropriately construct such a valid RKHS $\breve{\mathcal{H}}$ from $\tilde{\mathcal{H}}$ to achieve so. We first define $\breve{\mathcal{H}}$ to be the space of functionals from $\tilde{\mathcal{H}}$ with the following modified inner product:

$$\langle f, g \rangle_{\breve{\mathcal{H}}} \triangleq \langle f, g \rangle_{\tilde{\mathcal{H}}} + \langle Sf, Sg \rangle_{\mathcal{V}}, \tag{6}$$

where $\mathcal{V}$ is a linear space and $S : \tilde{\mathcal{H}} \to \mathcal{V}$ is a bounded linear operator. The first term in (6) is the common definition of inner product between two functionals, while the second term with the operator S reflects that unlabeled embedded distributions alter our beliefs in the overall structure. Denote by $\mathbf{f}(\boldsymbol{\mu}) = (f(\boldsymbol{\mu}_{\mathbb{P}_1}), ..., f(\boldsymbol{\mu}_{\mathbb{P}_n}))$, we have $\langle Sf, Sf \rangle_{\mathcal{V}} = \mathbf{f}(\boldsymbol{\mu}) M \mathbf{f}(\boldsymbol{\mu})^\top$ with M being a positive semidefinite matrix.

2) Validity of $\breve{\mathcal{H}}$

Theorem 1. *$\breve{\mathcal{H}}$ is a valid RKHS.*

A space is valid if it is bounded and complete.

3) Loss Function Calculation Based on Theorem 1, we have the following propositions.

Proposition 1. $\breve{\mathcal{H}} = \tilde{\mathcal{H}}$.

The two spaces are the same if each of the space is the subset of the other space. Although the two spaces are the same, the kernels therein are not identical. However, they are connected due to the involvement of unlabeled distributions.

Proposition 2. *$\breve{K} = (I + \tilde{K}M)^{-1}\tilde{K}$, where $\tilde{K}$ with $\tilde{K}_{ij} = \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j})$ is the kernel matrix for $\tilde{\mathcal{H}}$ on $\boldsymbol{\mu}_{\mathbb{P}_i}$'s, and $\breve{K}$ is the kernel matrix in the altered space $\breve{\mathcal{H}}$.*

Note that detailed proofs and derivations of theorems and propositions introduced in this section can be found in the next section. The complexity of the above kernel seems to be a potential problem when the data scales up, since it involves matrix multiplication as well as matrix inversion. However, when conducting experiments on large scale activity recognition datasets, the problem actually is not severe in practice. The reason is that the entries of kernels are dependent on the number of distributions, i.e., number of segments, each containing a repetition of activity, instead of the number of total instances, i.e., one entry for each timestamp equivalent to the product of # sample and # instances per sample. Other feasible solutions to further alleviate this problem include matrix factorization, low-rank approximation (Bach and Jordan 2005), etc. Data selection or feature selection (Nie et al. 2010) can be conducted on training data beforehand to keep a small fraction of key training data. The proposed method can be further developed in an online learning fashion (Hoi, Wang, and Zhao 2014), so that the matrix are maintained in a small scale.

Note that the choice of M is crucial regarding how to properly incorporate unlabeled embedded distributions. In this paper, we set M to be $M = rL^2$, where r is a scalar and $L = D - W$ is the Laplacian matrix, which is widely used in semi-supervised learning (Sindhwani, Niyogi, and Belkin 2005; Belkin, Niyogi, and Sindhwani 2006) to model the geometry structure underlying the data. To be specific, $W_{ij} = \exp\left(-\frac{\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\mathbb{P}_j}\|^2}{2\sigma^2}\right)$ if $\boldsymbol{\mu}_{\mathbb{P}_i}$ and $\boldsymbol{\mu}_{\mathbb{P}_j}$ are connected in the graph, and D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$. Based on the following Theorem 2 (whose derivations are at the end of the paper), the solution for the optimization problem in (5) can be expressed as a linear combination of the functionals $\{\breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}), \cdot\}_{i=1}^{l}$ as

$$f^*(\boldsymbol{\mu}_{\mathbb{P}}) = \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{P}_i}). \tag{7}$$

Theorem 2 (Representer Theorem for the proposed DSSL method)**.** *Given l labeled distributions $\{(\mathbb{P}_1, y_1), ..., (\mathbb{P}_l, y_l)\} \in \mathcal{P} \times \mathbb{R}$, a loss function $\ell : (\mathcal{P} \times \mathbb{R}^2)^l \to \mathbb{R} \cup \{+\infty\}$ and a strictly monotonically increasing real-valued function Ω on $[0, +\infty)$, the minimizer of the regularized risk functional*

$$\ell(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], ..., \mathbb{P}_l, y_l, \mathbb{E}_{\mathbb{P}_l}[f]) + \Omega(\|f\|_{\breve{\mathcal{H}}}), \tag{8}$$

admits an expansion $f = \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)$, where $\alpha_i \in \mathbb{R}$, for $i = 1, ..., l$.

Detailed Proofs

Proof of Theorem 1 Let's start with $\tilde{\mathcal{H}}$ with the kernel $\tilde{k}$. Since $\tilde{\mathcal{H}}$ is a complete Hilbert space, and evaluation functionals therein are bounded, i.e., $\forall \boldsymbol{\mu} \in \mathcal{H}, f \in \tilde{\mathcal{H}}, \exists\, C_{\boldsymbol{\mu}} \in \mathbb{R}$, s.t. $|f(\boldsymbol{\mu})| \leq C_{\boldsymbol{\mu}} \|f\|_{\tilde{\mathcal{H}}}$. Moreover, the bounded operator S is bounded by a constant D, i.e., $\|S\| = \sup_{f \in \tilde{\mathcal{H}}} \frac{\|Sf\|_{\mathcal{V}}}{\|f\|_{\tilde{\mathcal{H}}}} \leq D$. The complete $\tilde{\mathcal{H}}$ means every Cauchy sequence in the space converges to an element in $\tilde{\mathcal{H}}$. Let (f_n) be a Cauchy sequence in $\tilde{\mathcal{H}}$ converging to f, then $\forall \epsilon > 0, \exists$ an integer $N(\epsilon)$, s.t.

$$m > N(\epsilon),\ n > N(\epsilon) \Rightarrow \|f_m - f_n\|_{\tilde{\mathcal{H}}} < \frac{\epsilon}{\sqrt{1 + D^2}}.$$

Now let's turn to $\breve{\mathcal{H}}$. We need to prove the completeness of the space first. According to the definition in Eq. (6), we obtain that for any Cauchy sequence in $\breve{\mathcal{H}}$,

$$\begin{aligned} \|f_m - f_n\|_{\breve{\mathcal{H}}}^2 &= \|f_m - f_n\|_{\tilde{\mathcal{H}}}^2 + \|\boldsymbol{S}(f_m - f_n)\|_{\mathcal{V}}^2 \\ &\leq \|f_m - f_n\|_{\tilde{\mathcal{H}}}^2 + D^2 \|f_m - f_n\|_{\tilde{\mathcal{H}}}^2 \\ \Longrightarrow\ \|f_m - f_n\|_{\breve{\mathcal{H}}} &\leq \sqrt{1 + D^2} \|f_m - f_n\|_{\tilde{\mathcal{H}}} \\ &< \sqrt{1 + D^2} \times \frac{\epsilon}{\sqrt{1 + D^2}} = \epsilon. \end{aligned}$$

Hence $\breve{\mathcal{H}}$ is complete since every Cauchy sequence in $\breve{\mathcal{H}}$ converges to an element in $\breve{\mathcal{H}}$. Moreover, $\breve{\mathcal{H}}$ is bounded based on the property that any Cauchy sequence is bounded (Berlinet and Thomas-Agnan 2011, Lemma 5). This completes the proof.

Proof of Proposition 1 Firstly, we decompose $\breve{\mathcal{H}}$ to two orthogonal parts as

$$\breve{\mathcal{H}} = span\{\breve{k}(\boldsymbol{\mu}_{\mathbb{P}_1}, \cdot), ..., \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_l}, \cdot)\} \oplus \breve{\mathcal{H}}^{\perp},$$

where $\breve{\mathcal{H}}^{\perp}$ vanishes at all labeled embedded distributions, i.e.,

$$\forall f \in \breve{\mathcal{H}}^{\perp}, i \in \{1, ..., l\}, f(\boldsymbol{\mu}_{\mathbb{P}_i}) = 0. \tag{9}$$

Accordingly $Sf = 0$, which means $\langle f, g \rangle_{\breve{\mathcal{H}}} = \langle f, g \rangle_{\tilde{\mathcal{H}}}, \forall f \in \breve{\mathcal{H}}^{\perp}, g \in \breve{\mathcal{H}}$. Moreover,

$$\begin{aligned} f(\boldsymbol{\mu}_{\mathbb{P}}) = \langle f, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\breve{\mathcal{H}}} &= \langle f, \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\tilde{\mathcal{H}}} \\ &= \langle f, \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\tilde{\mathcal{H}}} + \langle Sf, S\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\mathcal{V}} \\ &= \langle f, \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\breve{\mathcal{H}}}. \end{aligned}$$

Thus, we have

$$\forall f \in \breve{\mathcal{H}}^{\perp},\ \langle f,\ \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) - \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \rangle_{\breve{\mathcal{H}}} = 0. \tag{10}$$

That is $\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) - \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \in (\breve{\mathcal{H}}^{\perp})^{\perp}$. By substituting (9) into (10), we obtain $\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot) \in (\breve{\mathcal{H}}^{\perp})^{\perp},\ \forall i$, which means

$$span\{\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\}_{i=1}^{l} \subseteq span\{\breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\}_{i=1}^{l}. \tag{11}$$

Secondly, we decompose $\tilde{\mathcal{H}}$ as $\tilde{\mathcal{H}} = span\{\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\}_{i=1}^{l} \oplus \tilde{\mathcal{H}}^{\perp}$. Similarly, we have

$$\langle f,\ \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot) \rangle_{\tilde{\mathcal{H}}} = 0,\ \forall f \in \tilde{\mathcal{H}}^{\perp},\ \forall i \in \{1, ..., l\}.$$

As $Sf = 0$, we have $\langle f, g\rangle_{\tilde{\mathcal{H}}} = \langle f, g\rangle_{\breve{\mathcal{H}}}$, and

$$\begin{aligned} f(\boldsymbol{\mu}_{\mathbb{P}}) = \langle f, \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\tilde{\mathcal{H}}} &= \langle f, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\breve{\mathcal{H}}} \\ &= \langle f, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\tilde{\mathcal{H}}} + \langle Sf, S\breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\mathcal{V}} \\ &= \langle f, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\tilde{\mathcal{H}}}. \end{aligned}$$

Therefore, we have $\langle f,\ \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) - \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\tilde{\mathcal{H}}} = 0$. Since $f \in \tilde{\mathcal{H}}^{\perp}$, it becomes $\langle f, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\tilde{\mathcal{H}}} = 0$, i.e., $\breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) \in (\tilde{\mathcal{H}}^{\perp})^{\perp}$. Therefore, we have

$$span\{\breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\}_{i=1}^{l} \subseteq span\{\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\}_{i=1}^{l}. \tag{12}$$

Finally, by considering both (11) and (12), we conclude that the two spans are the same. This completes the proof.

Proof of Proposition 2 Based on Proposition 1, we have

$$\breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) = \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) + \sum_{j=1}^{n} \beta_j(\boldsymbol{\mu}_{\mathbb{P}})\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot), \tag{13}$$

where the coefficients β_j depend on $\boldsymbol{\mu}_{\mathbb{P}}$. If we can obtain the exact formulation for β_j, then we can derive relations between two spaces by explicit forms. To find β_j, we use a system of linear equations generated by evaluating $\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)$ at $\boldsymbol{\mu}_{\mathbb{P}}$:

$$\begin{aligned} &\langle \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot), \breve{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot)\rangle_{\breve{\mathcal{H}}} \\ &= \langle \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot),\ \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) + \sum_{j=1}^{n} \beta_j(\boldsymbol{\mu}_{\mathbb{P}})\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot)\rangle_{\breve{\mathcal{H}}} \\ &= \langle \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot),\ \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \cdot) + \sum_{j=1}^{n} \beta_j(\boldsymbol{\mu}_{\mathbb{P}})\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot)\rangle_{\tilde{\mathcal{H}}} + \tilde{\boldsymbol{k}}_{\boldsymbol{\mu}_{\mathbb{P}_i}}^{\top} M\boldsymbol{g}, \end{aligned}$$

where $\tilde{\boldsymbol{k}}_{\boldsymbol{\mu}_{\mathbb{P}_i}}^{\top} = \left(\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_1}), ..., \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_n})\right)$ and $\boldsymbol{g}$ consists of components $\boldsymbol{g}_i = \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{P}_i}) + \sum_{j=1}^{n} \beta_j(\boldsymbol{\mu}_{\mathbb{P}})\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \boldsymbol{\mu}_{\mathbb{P}_i})$. Then we have the following linear equation for the coefficients $\beta(\boldsymbol{\mu}_{\mathbb{P}}) = (\beta_1(\boldsymbol{\mu}_{\mathbb{P}}), ..., \beta_n(\boldsymbol{\mu}_{\mathbb{P}}))^{\top}$:

$$-M\tilde{\boldsymbol{k}}_{\boldsymbol{\mu}_{\mathbb{P}}} = (I + M\tilde{K})\beta(\boldsymbol{\mu}_{\mathbb{P}}). \tag{14}$$

Based on (13) and (14), we obtain the following explicit form for $\breve{k}(\cdot, \cdot)$:

$$\breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j}) = \tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j}) - \tilde{k}_{\boldsymbol{\mu}_{\mathbb{P}_i}}^{\top}(I + M\tilde{K})^{-1}M\tilde{k}_{\boldsymbol{\mu}_{\mathbb{P}_j}}.$$

The above equation can be written in the following concise matrix form:

$$\breve{K} = \tilde{K} - \tilde{K}(I + M\tilde{K})^{-1}M\tilde{K}. \tag{15}$$

It can be shown that by applying the Sherman-Morrison-Woodbury (SMW) identity, (15) can be further rewritten as

$$\breve{K} = (I - \tilde{K}(I + M\tilde{K})^{-1}M)\tilde{K} = (I + \tilde{K}M)^{-1}\tilde{K}. \tag{16}$$

This completes the proof.

Proof of Theorem 2 Any functional $f \in \breve{\mathcal{H}}$ can be uniquely decomposed into a component $f_{\boldsymbol{\mu}}$ in the space spanned by the kernel mean embedding $f_{\boldsymbol{\mu}} = \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)$, and a component $f_{\perp}$ orthogonal to it, i.e., $\langle f_{\perp}, \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot)\rangle = 0, \forall j \in \{1, ..., l\}$. Therefore, we have

$$f = f_{\boldsymbol{\mu}} + f_{\perp} = \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot) + f_{\perp}.$$

Thus, for all j, we can further induce that

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_j}[f] &= \left\langle \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot) + f_{\perp},\ \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot) \right\rangle \\ &= \left\langle \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot), \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_j}, \cdot) \right\rangle. \end{aligned}$$

This indicates the loss function term in (8) does not depend on $f_{\perp}$. Besides, the second term $\Omega(\cdot)$ in (8) is strictly monotonically increasing, so we have

$$\begin{aligned} \Omega(\|f\|_{\breve{\mathcal{H}}}) &= \Omega\left(\left\|\sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot) + f_{\perp}\right\|_{\breve{\mathcal{H}}}\right) \\ &= \Omega\left(\sqrt{\left\|\sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\right\|_{\breve{\mathcal{H}}}^2 + \|f_{\perp}\|_{\breve{\mathcal{H}}}^2}\right) \\ &\geq \Omega\left(\left\|\sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)\right\|_{\breve{\mathcal{H}}}\right), \end{aligned}$$

where the equality holds if and only if $f_{\perp} = 0$. Therefore, the first term in (8) is independent of $f_{\perp}$ and the second term reaches its minimum when $f_{\perp} = 0$. Consequently, any minimizer must take the form $f = f_{\boldsymbol{\mu}} = \sum_{i=1}^{l} \alpha_i \breve{k}(\boldsymbol{\mu}_{\mathbb{P}_i}, \cdot)$. This completes the proof.

Experiments

We conduct experiments on 3 sensor-based activity datasets. The statistics are listed in Table 2. Skoda records 10 gestures in car maintenance scenarios with 20 acceleration sensors being put on the arms of the subject (Stiefmeier, Roggen, and Tröster 2007). Each gesture is repeated around 70 times. The transitions between two gestures are labeled as Null class, which are also considered as activities. WISDM uses accelerometer sensors embedded in the phones to collect six regular activities: jogging, walking, ascending stairs, descending stairs, sitting and standing (Kwapisz, Weiss, and Moore 2010). HCI composes of gestures with the hand describing different shapes: a circle, a square, a pointing-up triangle, an upside-down triangle, and an infinity symbol (Förster, Roggen, and Tröster 2009). Each gesture is recorded over 50 repetitions, and about 5 to 8 seconds per repetition. Null class exists as well in HCI dataset.

Experimental Setup

Following the criteria in (Qian, Pan, and Miao 2018), we adopt both micro-F_1 score (miF) and weighted macro-F_1

Table 2: Statistics of datasets used in experiments.

Datasets	# Sample	# Instances per sample	# Feature	# Class
Skoda	1,447	68	60	10
HCI	264	602	48	5
WISDM	389	705	6	6

score (maF) to evaluate the performance of different methods. All the reported results are the average values together with the standard deviation over 6 random splits for training and testing. Each dataset is randomly split into 3 subsets: labeled training set, unlabeled training set and test set. Each subset is set to contain activities of all classes. We set the ratio to be 0.02:0.1:0.88 and fix $r = 100$. The impact of differentiating r will be discussed later. Different from experimental setups in existing papers that set labeled data's ratio to be quite large (Matsushige, Kakusho, and Okadome 2015; Stikic, Larlus, and Schiele 2009), we deliberately set the labeled data's ratio to be extremely small. Hence, our method requires fewer labels and thus more practical with regards to applicability in reality. Evaluations are conducted on the test set. We adopt RBF kernels for all the kernels used in the experiments.

Baselines We compare the proposed DSSL method with the following state-of-the-art methods.

- State-of-the-art supervised methods with various features:
  - SVMs (Chang and Lin 2011): as SVM is a vectorial-based classifier, we use mean, variance, etc to generate a feature vector for each segment.
  - SAX-a (Lin et al. 2007) treats data as strings, and structural features are extracted. We follow the settings in (Lin et al. 2007) with no dimension reduction. The parameter alphabet_size range is $a \in \{3, 6, 9\}$.
  - ECDF-d (Hammerla et al. 2013; Plötz, Hammerla, and Olivier 2011) extracts d descriptors from each sensor's each dimension. $d \in \{5, 15, 30, 45\}$.

  Note that the overall shape and spatial features besides the mean and variance features are concatenated before applying the SVM classifier.
- State-of-the-art supervised method based on distributions, SMM$_{AR}$ (Qian, Pan, and Miao 2018).
- Classic vectorial-based semi-supervised methods:
  - LapSVM (Belkin, Niyogi, and Sindhwani 2006) is an extension of SVM with manifold regularization.
  - $\bigtriangledown$TSVM (Chapelle and Zien 2005) is a Transductive SVM by using gradient descent for training. As this is a transductive approach rather than a truly semi-supervised learning approach, we make the test data available in the training phase of this method.
- State-of-the-art semi-supervised methods specifically designed for activity recognition:
  - SSKLR (Matsushige, Kakusho, and Okadome 2015) is a semi-supervised kernel logistic regression method with Expectation-Maximization algorithm.
  - GLSVM (Stikic, Larlus, and Schiele 2009) is a multi-graph method where each graph captures different aspects of the activities.

Experimental Results

Overall Experimental Results The experimental results are presented in Table 3. The proposed DSSL consistently performs the best on all datasets. DSSL outperforms all the other methods by 5.6%, 17.7%, and 14.4% respectively on three datasets in terms of miF. This favorably indicates the effectiveness of the proposed DSSL. Note that in Table 3, the performances of the comparison methods on WISDM are much worse than those on the other two datasets. This may be due to the data complexity caused by the large number of subjects in WISDM. On datasets Skoda and HCI, the performance ranking is DSSL $>$ SMM$_{AR}$ $>$ SVMs $\approx$ ECDF $>$ SAX, which reveals that 1) distribution-based methods are more capable of distinguishing different activities; 2) feature extraction plays an important role and string-based data representation in SAX is not that proper for activity data compared to ECDF; 3) with the increase of descriptor d, the performance of ECDF is increasing in HCI dataset while decreasing in Skoda and WISDM, meaning ECDF may be task-dependent. However, note that SMM$_{AR}$ performs the worst on WISDM dataset, which illustrates that distribution-based methods are more dependent on the number of labeled data than vectorial-based methods. This indeed reflects the motivation of our proposed method. Nevertheless, DSSL does not suffer from this limitation ascribed to its semi-supervised fashion. For semi-supervised methods, the ranking is DSSL $>$ LapSVM $\approx$ GLSVM $\approx$ $\bigtriangledown$TSVM $>$ SSKLR, which demonstrates the prevalence of graph-based methods over logistic regression method for activity data.

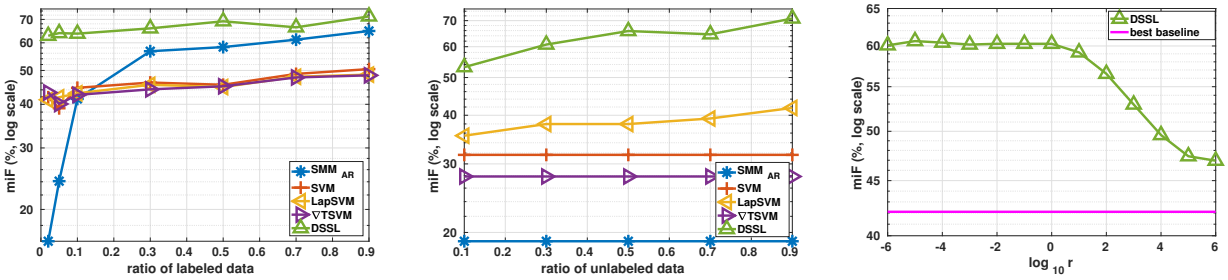
Impact of Ratio of Labeled Data To analyze the impact on the proportion of labeled training data, we conduct experiments on WISDM dataset. We fix the ratio of test data and unlabeled training data to be 20% and 20% respectively, and alter the ratio of labeled training data to be {0.02, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9} of the rest 60% data. The results are depicted in Figure 1(a). DSSL performs the best under all the ratios. When more labeled training data becomes available, all methods perform better. Moreover, distributional-based method (SMM$_{AR}$) has larger performance enhancement than vectorial-based methods, which further verifies the superiority of learning from distributions.

Impact of Ratio of Unlabeled data We investigate the influence of unlabeled data by fixing the ratio of labeled training data and test data to be 1% and 20%, respectively, and modifying unlabeled training data to be {0.1, 0.3, 0.5, 0.7, 0.9} of the remaining 79% data. Note that supervised methods (SMM$_{AR}$, SVMs) and transductive methods ($\bigtriangledown$TSVM) perform the same under this setting, while the performances of semi-supervised methods keep increasing with more unlabeled training data as shown in Figure 1(b).

Impact of parameter r In previous experiments, we fix $r = 100$. Here we conduct sensitivity test on r. As indicated

Table 3: Experimental results on 3 activity datasets (unit: %).

Methods		Skoda		HCI		WISDM	
		miF	maF	miF	maF	miF	maF
Vectorial-based supervised	SVMs	85.7±1.8	42.5±0.9	69.7±9.6	69.6±9.4	41.5±5.2	39.6±6.8
	SAX_3	39.6±6.3	18.7±2.9	36.0±3.0	34.7±2.5	34.6±1.4	30.6±1.2
	SAX_6	37.2±6.1	18.6±2.8	39.7±7.3	38.4±7.9	34.9±3.0	30.5±5.0
	SAX_9	40.3±6.5	19.9±3.2	39.8±8.7	37.0±9.2	33.6±2.9	28.8±5.8
	ECDF_5	84.2±2.1	41.6±1.0	67.7±10.1	67.6±9.1	42.1±6.3	40.5±7.7
	ECDF_15	79.8±1.5	39.2±0.7	68.4±10.4	68.5±9.6	39.4±3.3	36.2±5.7
	ECDF_30	72.6±1.2	35.4±0.3	68.6±11.1	68.7±10.5	37.7±2.5	32.6±4.9
	ECDF_45	65.7±2.5	31.5±1.3	68.6±11.4	68.6±10.8	36.4±1.4	31.3±3.6
Vectorial-based semi-supervised	LapSVM	89.7±2.1	44.6±1.2	76.1±4.8	76.3±4.7	40.1±3.8	34.5±3.5
	$\bigtriangledown$TSVM	85.9±2.7	84.8±2.8	75.4±11.5	75.5±11.2	41.3±5.6	39.4±6.9
	SSKLR	25.4±19.3	12.1±2.5	24.2±17.2	18.1±10.1	24.6±17.0	17.3±9.9
	GLSVM	89.7±2.1	44.5±1.2	75.7±5.8	75.7±5.7	40.4±3.8	33.9±4.0
Distribution-based supervised	SMM_{AR}	93.2±0.9	93.1±1.0	82.2±13.4	78.9±18.4	20.5±3.3	11.7±3.9
Distribution-based semi-supervised	DSSL	**98.8±0.5**	**98.8±0.5**	**99.9±0.2**	**99.9±0.2**	**56.5±5.1**	**55.6±5.0**



(a) Varying ratios of labeled data. (b) Varying ratios of unlabeled data. (c) Impact of r to the performance.

Figure 1: Performance of DSSL on WISDM under different settings (in miF).

in Fig. 1(c), the performance of DSSL on test data keeps stable when $r \in [10^{-6}, 1]$. When r becomes larger, the performance of DSSL begins to decrease. This observation indicates that r balances the tradeoff between labeled and unlabeled data. Larger r implies stronger emphasis on unlabeled data. More importantly, under all different r values, DSSL consistently outperforms all other methods. Fig. 1(c) shows the best baseline, i.e., ECDF_5 in WISDM's case.

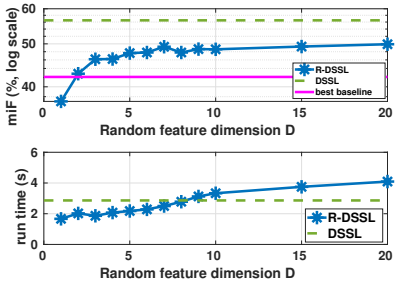


Figure 2: Impact of D to the performance on WISDM.

Impact on random Fourier feature (RFF) dimension D We analyze how R-DSSL accelerates DSSL with D-dimensional explicit statistical features. The experiments are conducted on a Linux server with Intel(R) Xeon(R) E5-2695 2.40GHz CPU. As shown in Fig. 2, R-DSSL steadily outperforms the best baseline when $D \geq 2$. Note that R-DSSL performs slightly worse than DSSL due to its approximation nature, however it requires less computational run time when $D < 8$ compared to DSSL.

Conclusion

In this paper, we propose a semi-supervised learning framework, DSSL, for sensor-based activity recognition problems. The proposed DSSL naturally embeds automatic feature extraction and classification in a semi-supervised learning manner. Extensive experiments are conducted on three activity datasets to demonstrate the superiority of DSSL compared with a number of state-of-the-art methods.

Acknowledgments

This research is supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017),

and the Interdisciplinary Graduate School, Nanyang Technological University under its Graduate Research Scholarship. Sinno J. Pan thanks the support from the NTU Singapore Nanyang Assistant Professorship (NAP) grant M4081532.020.

References

Avci, A.; Bosch, S.; Marin-Perianu, M.; Marin-Perianu, R.; and Havinga, P. J. M. 2010. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *ARCS Workshops*, 167–176.

Bach, F. R., and Jordan, M. I. 2005. Predictive low-rank decomposition for kernel methods. In *ICML*, 33–40.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Berlinet, A., and Thomas-Agnan, C. 2011. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* 46(3):33:1–33:33.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol* 2(3):27:1–27:27.

Chapelle, O., and Zien, A. 2005. Semi-supervised classification by low density separation. In *AISTATS*.

Chapelle, O.; Schlkopf, B.; and Zien, A. 2010. *Semi-Supervised Learning*. The MIT Press.

Förster, K.; Roggen, D.; and Tröster, G. 2009. Unsupervised classifier self-calibration through repeated context occurences: Is there robustness against sensor displacement to gain? In *ISWC*, 77–84.

Frank, J.; Mannor, S.; and Precup, D. 2010. Activity and gait recognition with time-delay embeddings. In *AAAI*.

Guan, D.; Yuan, W.; Lee, Y.; Gavrilov, A.; and Lee, S. 2007. Activity recognition based on semi-supervised learning. In *RTCSA*, 469–475.

Hammerla, N. Y.; Kirkham, R.; Andras, P.; and Ploetz, T. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *ISWC*, 65–68.

Hoi, S. C. H.; Wang, J.; and Zhao, P. 2014. LIBOL: a library for online learning algorithms. *J. Mach. Learn. Res.* 15(1):495–499.

Janidarmian, M.; Fekr, A. R.; Radecka, K.; and Zilic, Z. 2017. A comprehensive analysis on wearable acceleration sensors in human activity recognition. *Sensors* 17(3):529.

Kwapisz, J. R.; Weiss, G. M.; and Moore, S. 2010. Activity recognition using cell phone accelerometers. *SIGKDD Explorations* 12(2):74–82.

Lara, O. D., and Labrador, M. A. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials* 15(3):1192–1209.

Lin, J.; Keogh, E. J.; Wei, L.; and Lonardi, S. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Min. Knowl. Discov.* 15(2):107–144.

Matsushige, R.; Kakusho, K.; and Okadome, T. 2015. Semi-supervised learning based activity recognition from sensor data. In *GCCE*, 106–107.

Muandet, K.; Fukumizu, K.; Dinuzzo, F.; and Schölkopf, B. 2012. Learning from distributions via support measure machines. In *NIPS*, 10–18.

Muandet, K.; Fukumizu, K.; Sriperumbudur, B. K.; and Schölkopf, B. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* 10(1-2):1–141.

Nazábal, A.; Garcia-Moreno, P.; Artés-Rodríguez, A.; and Ghahramani, Z. 2016. Human activity recognition by combining a small number of classifiers. *IEEE J. Biomed. Health Inform.* 20(5):1342–1351.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. Q. 2010. Efficient and robust feature selection via joint l2,1-norms minimization. In *NIPS*, 1813–1821.

Plötz, T.; Hammerla, N. Y.; and Olivier, P. 2011. Feature learning for activity recognition in ubiquitous computing. In *IJCAI*, 1729–1734.

Qian, H.; Pan, S. J.; and Miao, C. 2018. Sensor-based activity recognition via learning from distributions. In *AAAI*.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. In *NIPS*, 1177–1184.

Sindhwani, V.; Niyogi, P.; and Belkin, M. 2005. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, 824–831.

Stiefmeier, T.; Roggen, D.; and Tröster, G. 2007. Fusion of string-matched templates for continuous activity recognition. In *ISWC*, 41–44.

Stikic, M.; Larlus, D.; Ebert, S.; and Schiele, B. 2011. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(12):2521–2537.

Stikic, M.; Larlus, D.; and Schiele, B. 2009. Multi-graph based semi-supervised learning for activity recognition. In *ISWC*, 85–92.

Yao, L.; Nie, F.; Sheng, Q. Z.; Gu, T.; Li, X.; and Wang, S. 2016. Learning from less for better: semi-supervised activity recognition via shared structure discovery. In *UbiComp*, 13–24.

Zhu, X. 2005. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison.