Distribution-based Semi-Supervised Learning for Activity Recognition (AAAI'19)

Hangwei Qian, Sinno Jialin Pan, Chunyan Miao

Nanyang Technological University, Singapore

November 10, 2018

Outline

Human Activity Recognition

Tremendous applications:

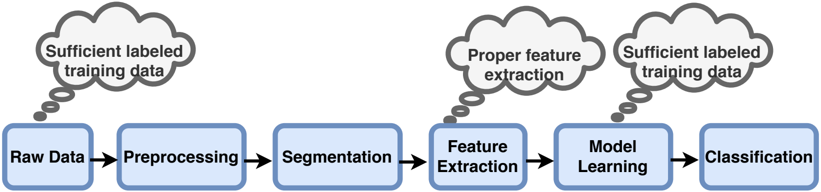
- elderly assistant
- healthcare
- fitness coaching
- smart building
- gaming

Human Activity Recognition

A multi-class classification problem

- Input: wearable onbody sensor data
- Output: activity labels

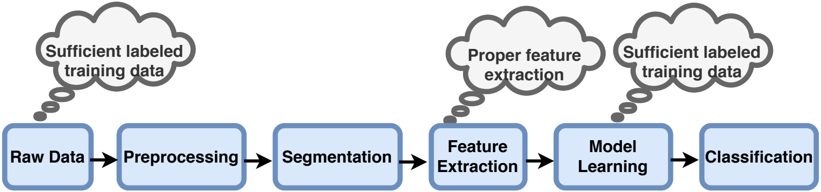
Problem Overview



Two key prerequisites:

1. expressive feature extraction $\rightarrow$ discriminate activities
2. sufficient labeled training data $\rightarrow$ build a precise model

Problem Overview



Two key prerequisites:

1. expressive feature extraction $\rightarrow$ dependent on domain knowledge
2. sufficient labeled training data $\rightarrow$ require a huge amount of human annotation effort

Motivation

1. **Can we extract as many discriminative features as possible, in an automatic fashion?**
   $\rightarrow$ kernel mean embedding of distributions, with NO information loss
   $\rightarrow$ two novel methods **SMM**$_{AR}$ and **R-SMM**$_{AR}$[1]
2. **Can we utilize labeled data as few as possible to alleviate human annotation effort?**
   $\rightarrow$ Distribution-based Semi-Supervised Learning (DSSL)

---

[1] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. **Sensor-based activity recognition via learning from distributions.** In AAAI'18 (oral).

Outline

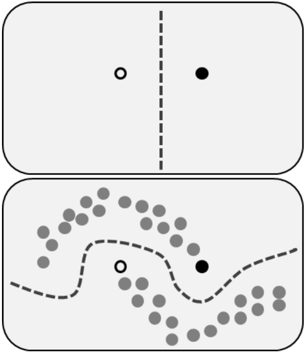
Contribution

DSSL: Distribution-based Semi-Supervised Learning

1. All orders of statistical moments features are extracted implicitly and automatically
2. DSSL relaxes SMM_{AR}'s full supervision assumption, and exploit unlabeled instances to learn an underlying data structure
3. DSSL is the first attempt on semi-supervised learning with distributions, with rigorous theoretical proofs provided.
4. Extensive experiments to show the efficacy of DSSL.

Intuition of DSSL

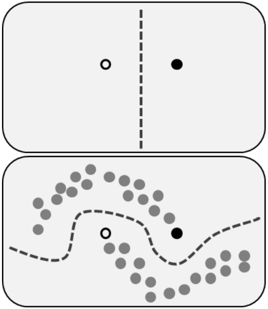
- Label annotation is time-consuming
- Unlabeled data is abundant and informative



Intuition: unlabeled data sheds light on the underlying manifolds of data space

Distribution-based SSL: Main idea

- wrap the data space to reflect the geometry of the data
- modify the similarity measure $\langle f, g\rangle_{\breve{\mathcal{H}}} \triangleq \langle f, g\rangle_{\tilde{\mathcal{H}}} + F(f, g)$
  - data within a manifold (instead of closer Euclidean distance)$\rightarrow$ more similar
  - data with different labels $\rightarrow$ less similar



Challenges

$$\langle f, g\rangle_{\breve{\mathcal{H}}} \triangleq \langle f, g\rangle_{\tilde{\mathcal{H}}} + F(f, g) \tag{1}$$

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

1. How to construct the data-dependent kernel by incorporating unlabeled training data?
2. Is the new space valid? Since a RKHS is defined by inner product.
3. How to calculate the loss function given two items are not in the same space?

Challenge 1/3 Construction of kernel

$$\langle f, g \rangle_{\breve{\mathcal{H}}} \triangleq \langle f, g \rangle_{\tilde{\mathcal{H}}} + \langle Sf, Sg \rangle_{\mathcal{V}}, \tag{3}$$

where S is a bounded linear operator.
Denote $\mathbf{f}(\boldsymbol{\mu}) = (f(\boldsymbol{\mu}_{\mathbb{P}_1}), \ldots, f(\boldsymbol{\mu}_{\mathbb{P}_n}))$,

$$\langle Sf, Sf \rangle_{\mathcal{V}} = \mathbf{f}(\boldsymbol{\mu}) M \mathbf{f}(\boldsymbol{\mu})^{\top} \tag{4}$$

Challenge 2/3 Validity of the new space

Theorem 1

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

Proof.

$\forall \boldsymbol{\mu} \in \mathcal{H}, f \in \tilde{\mathcal{H}}, \exists\ C_{\boldsymbol{\mu}} \in \mathbb{R}$, s.t. $|f(\boldsymbol{\mu})| \leq C_{\boldsymbol{\mu}} \|f\|_{\tilde{\mathcal{H}}}$.
$\|S\| = \sup\limits_{f \in \tilde{\mathcal{H}}} \frac{\|Sf\|_{\mathcal{V}}}{\|f\|_{\tilde{\mathcal{H}}}} \leq D. \forall \epsilon > 0, \exists$ an integer $N(\epsilon)$, s.t.
$m > N(\epsilon),\ n > N(\epsilon) \Rightarrow \|f_m - f_n\|_{\tilde{\mathcal{H}}} < \frac{\epsilon}{\sqrt{1+D^2}}$. For any Cauchy sequence in $\breve{\mathcal{H}}$,

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

Challenge 3/3 Loss function calculation

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

Proposition 1

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

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 $\breve{\mathcal{H}}$.

In our case, $M = rL^2$, where L is the commonly-used Laplacian matrix.

Outline

Experimental Setup

labeled training set, unlabeled training set and test set: 0.02:0.1:0.88

Table 1 : 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

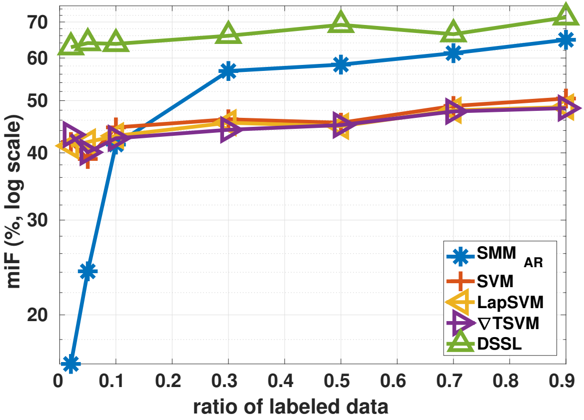
Experimental Results

Table 2 : 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
	▽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**

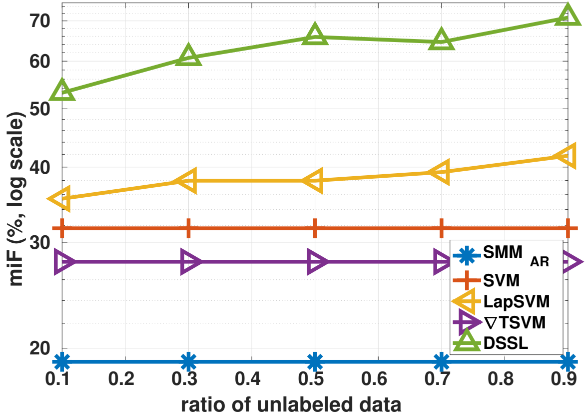
Experiments Analysis (1/3)

Varying ratios of labeled data



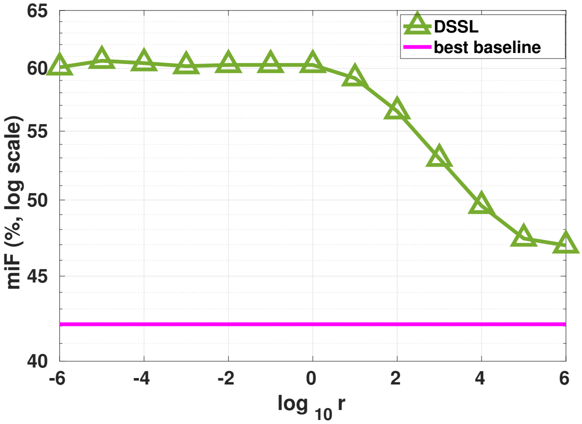
Experiments Analysis (2/3)

Varying ratios of unlabeled data



Experiments Analysis (3/3)

Impact of parameter *r* to the performance



Outline

Conclusion

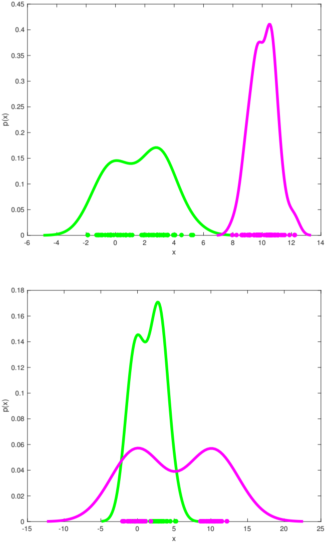
We propose a novel method, i.e., Distribution-based Semi-Supervised Learning (DSSL) for human activity recognition

1. All orders of statistical moments features are extracted implicitly and automatically
2. DSSL relaxes SMM_{AR}'s full supervision assumption, and exploit unlabeled instances to learn an underlying data structure
3. DSSL is the first attempt on semi-supervised learning with distributions, with rigorous theoretical proofs provided.
4. Extensive experiments to show the efficacy of DSSL.

Questions?

Codes will be available in http://hangwei12358.github.io/

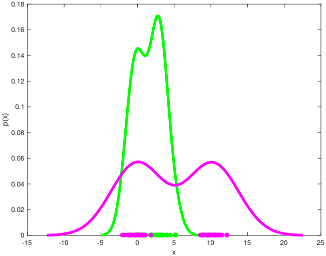
Intuition of Kernel Mean Embedding

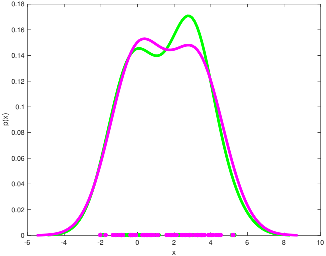


$(\mathbb{E}[x])$ as features

problem: many distributions have the same mean!

Intuition of Kernel Mean Embedding





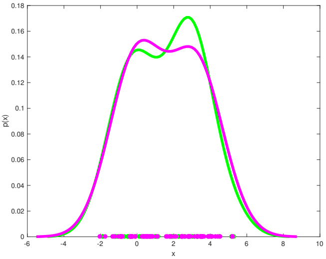
$(\mathbb{E}[x])$ as features

problem: many distributions have the same mean!

$\begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix}$ as features

problem: many distributions have the same mean and variance!

Intuition of Kernel Mean Embedding



$(\mathbb{E}[x])$ as features

problem: many distributions have the same mean!

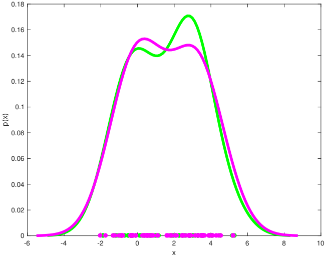
$\begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix}$ as features

problem: many distributions have the same mean and variance!

$\begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \\ \mathbb{E}[x^3] \end{pmatrix}$ as features

problem: many distributions still have the same first 3 moments!

Intuition of Kernel Mean Embedding



$$\mu[P_x] = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \\ \mathbb{E}[x^3] \\ \dots \\ \dots \end{pmatrix}$$

The **infinite dimensional features** should be able to discriminate different distributions!

Kernel Mean Embedding of Distributions

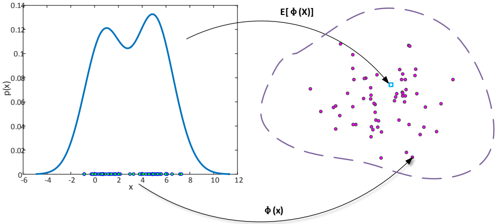


Figure 1 : Illustrations of kernel mean embeddings of a distribution and embeddings of empirical examples

$$\mu[P_x] = E_x[k(\cdot, x)] \tag{6}$$

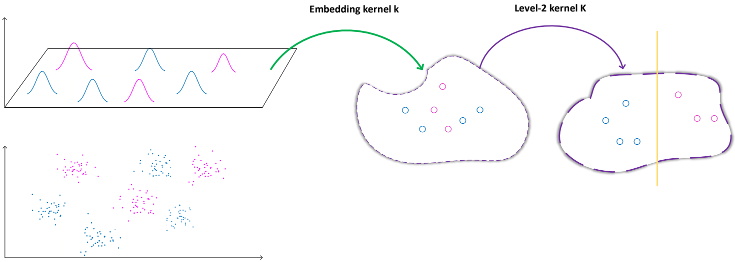
$$\mu[X] = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i) \tag{7}$$

Here $X = \{x_1, ..., x_m\} \overset{i.i.d.}{\sim} P_x$.

SMM$_{AR}$ Framework

$$\langle \hat{\boldsymbol{\mu}}_{\mathbb{P}_x}, \hat{\boldsymbol{\mu}}_{\mathbb{P}_z} \rangle = \tilde{k}(\hat{\boldsymbol{\mu}}_{\mathbb{P}_x}, \hat{\boldsymbol{\mu}}_{\mathbb{P}_z}) = \frac{1}{n_x \times n_z} \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} k(\mathbf{x}_i, \mathbf{z}_j), \tag{8}$$

$$\tilde{k}(\boldsymbol{\mu}_{\mathbb{P}_x}, \boldsymbol{\mu}_{\mathbb{P}_z}) = \langle \psi(\boldsymbol{\mu}_{\mathbb{P}_x}), \psi(\boldsymbol{\mu}_{\mathbb{P}_z}) \rangle \tag{9}$$



Problem Formulation of SMM_{AR}

- Training set: $\{(P_i, y_i)\}, i \in \{1, ..., N\}, x_i \sim P_i, x_i = \{x_{i1}, ..., x_{im_i}\}, y_i \in \{1, ..., L\}$
- Multi-class classifier $\rightarrow$ C_L^2 binary classifiers $f, y = f(\phi(\mu_x)) + b$
- Primal Optimization problem:

$$
\begin{aligned}
&\underset{f,b}{argmin} \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^{N} \xi_i \\
&s.t. y_i = f(\phi(\mu_{x_i})) + b \\
&\quad y_i f(\phi(\mu_i)) \geq 1 - \xi_i, \forall i \\
&\quad \xi_i \geq 0, \forall i
\end{aligned}
\tag{10}
$$