Activity Recognition via Learning from Distributions

QIAN Hangwei, IGS-LILY NTU

21.Feb.2017

Main Supervisor: Prof. Sinno Jialin Pan
Co-Supervisors: Prof. Lihui Chen, Prof. Chun Yan Miao

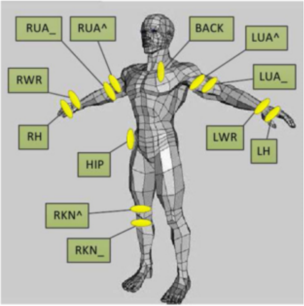
Outline

Outline

Human Activity Recognition

A multi-class classification problem

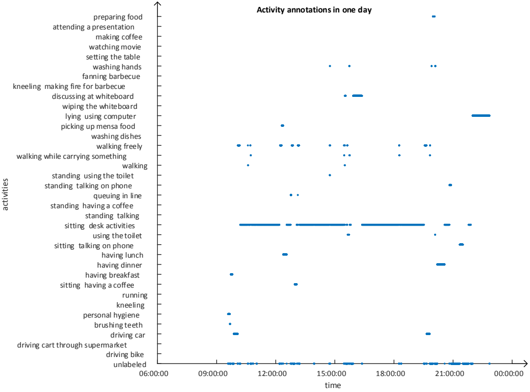
- Input: sensor data (Our focus: on-body sensors)



Human Activity Recognition

A multi-class classification problem

- Input: sensor data (Our focus: on-body sensors)
- Output: activity labels



Human Activity Recognition

A multi-class classification problem

- Input: sensor data (Our focus: on-body sensors)
- Output: activity labels

Tremendous applications:

- eldercare
- healthcare
- smart building
- gaming

Outline

1. Overview
   - Human Activity Recognition
   - Existing Framework
2. Activity Recognition via Kernel Embedding
   - Kernel Mean Embeddings of Distributions
   - Learning from Distributions
3. Experiments Results
4. Conclusion and Future Work

Classification algorithms

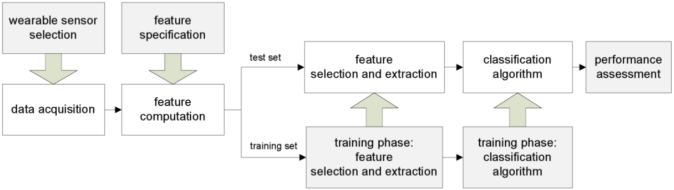


Figure 1: Classical framework of classification

- Supervised: kNN, Naive Bayes, GMM, SVM, etc
- Semi-supervised: co-training, self-training, etc
- Unsupervised: topic modeling, HMM, DP, etc
- Others: fuzzy reasoning, multi-agent-based, etc

Feature selection

Low-level features

- Manual feature engineering, statistical features
- Dimension reduction methods, e.g., PCA, FFT

High-level features

- String matching methods
- Deep learning based methods

Feature selection

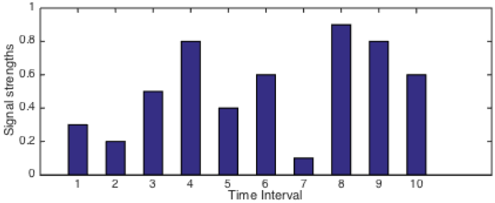
Low-level features

- Manual feature engineering, statistical features
- Dimension reduction methods, e.g., PCA, FFT

High-level features

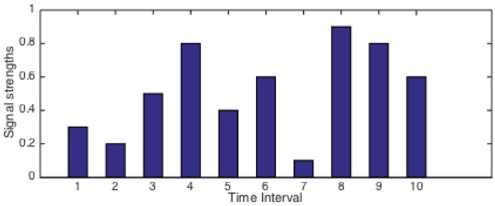
- String matching methods
- Deep learning based methods

Premise: one fixed-length feature vector for each activity



Feature selection

Premise: one fixed-length feature vector for each activity



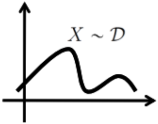
Two simple solutions:

- Frame-level: frame-label pairs
- Segment-level (1entry): mean feature vector to represent one segment

Contributions

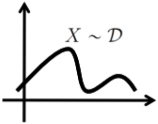
- A new time-series data representation: distribution representations
- A novel feature representation for time-series data based on kernel embedding technique

Intuition



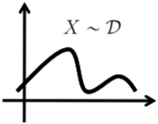
$$\mu_X = \left(\mathbb{E}[X] \right)$$

Problem: Lots of Distributions have the same mean!



$$\mu_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{pmatrix}$$

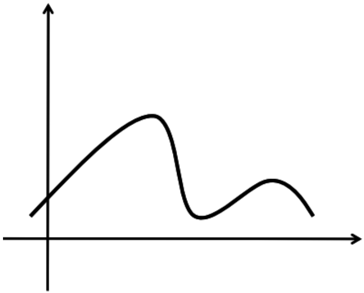
Better, but lots of distributions still have the same mean and variance!



$$\mu_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \\ \mathbb{E}[X^3] \end{pmatrix}$$

Even better, but lots of distributions still have the same first three moments!

Intuition



$$\boldsymbol{\mu}_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \\ \mathbb{E}[X^3] \\ \cdots \\ \cdots \end{pmatrix}$$

(not exactly, but right idea...)

- But the vector is infinite……..how do we compute things with it?????

Outline

Kernel Embeddings

Kernel[1]

The $\mathcal{X}$ is a valid set. A kernel **k**: $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ promises the existence of a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \times \mathcal{H}, s.t.$

$$\forall x, x\prime \in \mathcal{X}, \mathbf{k}(x, x\prime) = \langle \phi(x), \phi(x\prime) \rangle_{\mathcal{H}}.$$

Reproducing kernel Hilbert space (RKHS)[2]

Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a valid set $\mathcal{X}$. A function $\mathbf{k} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if **k** satisfies:

•$\forall x \in \mathcal{X}, \mathbf{k}(\cdot, x) \in \mathcal{H}$

•$\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, \mathbf{k}(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

Kernel Mean Embeddings of a Distribution

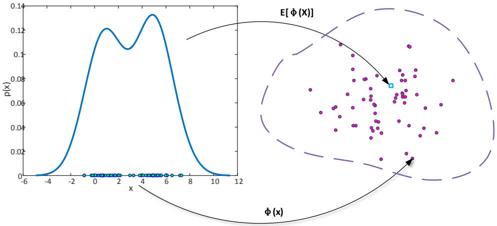


Figure 2: Illustrations of kernel mean embeddings of a distribution and embeddings of empirical examples

$$\mu[P_x] = E_x[k(\cdot, x)] \tag{1}$$

$$\mu[X] = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i) \tag{2}$$

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

Kernel Mean Embeddings of a Distribution

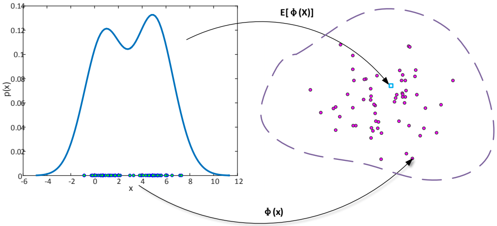


Figure 2: Illustrations of kernel mean embeddings of a distribution and embeddings of empirical examples

Theorem [2]

Assume that $\|f\|_\infty \leq R$ for all $f \in \mathcal{H}, \|f\|_\mathcal{H} \leq 1$. Then with probability at least $1 - \delta$,

$\|\mu[P_x] - \mu[X]\| \leq 2R_m(\mathcal{H}, P_x) + R\sqrt{-m^{-1}log(\delta)}$

Kernel Mean Embeddings of Distributions

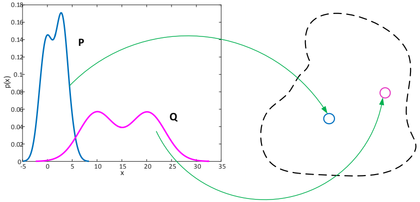


Figure 3: Illustration of the kernel mean embedding of two different distributions

Injectivity[2]

A universal kernel k can promise an injective mean map $\mu : P_x \rightarrow \mu[P_x]$.

Outline

Learning from Distributions

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

$$\tilde{k}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \langle \psi(\boldsymbol{\mu}_i), \psi(\boldsymbol{\mu}_j) \rangle \tag{2}$$

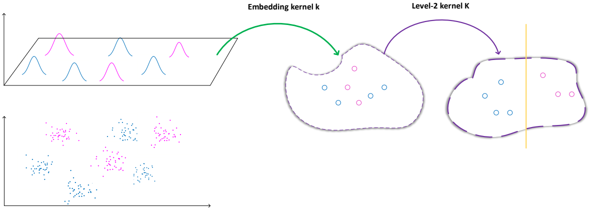


Figure 4: Illustration of the two level kernel learning framework based on the distributional data.

Problem Formulation

- 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:

$$\begin{aligned} Objective : \underset{f,b}{argmin} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. y_i = f(\phi(\mu_{x_i})) + b \\ y_i f(\phi(\mu_i)) \geq 1 - \xi_i, \forall i \\ \xi_i \geq 0, \forall i \end{aligned} \tag{3}$$

Problem Formulation

- Dual Optimization Problem:

$$L(f, b, \alpha, \beta) = \frac{1}{2}\|f\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{y_i(f(\phi(\mu(x_i))) + b) - 1\} - \sum_i \beta_i\xi_i \tag{4}$$

$$\frac{\partial L}{\partial f} = 0 \Rightarrow f = \sum_{i=1}^{N}\alpha_i y_i \phi(\mu(x_i))$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{5}$$

$$\frac{\partial L}{\partial \xi} = 0 \Rightarrow \alpha_i = C - \beta_i$$

Problem Formulation

- Substitute Eq.(5) into Eq.(4):

$$
\begin{aligned}
L(\alpha, \beta) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \tilde{k}(\mu_n, \mu_m) \\
s.t. 0 \leq \alpha_i \leq C \\
\beta_i \geq 0 \\
\sum_{i=1}^{N} \alpha_i y_i = 0
\end{aligned} \tag{6}
$$

- $y_{new} = f(\phi(\mu_{x_{new}})) + b = \sum_{i=1}^{N} \alpha_i y_i \phi(\mu_i)^T \phi(\mu_{x_{new}}) + b$

Summary of Learning from Distributions Methods

Methods	Main Advantages	Main Disadvantages
Define new kernels	Easy to implement	Designed specifically for certain distributions and application domains
Divergence estimators	can be calculated in closed form	lack the consistency analysis; assume the training distributions are Gaussians
Kernel-kernel estimator	take the whole distributions as input, scalar values as output, and address the consistency	Strong assumptions on input space and kernels; need density estimation; the input distributions can affect the convergence rate.
Double-basis estimator	easy to scale up	need density estimation;strong assumptions on input space
Distribution to distribution regression	Inputs and outputs are both distributions	need density estimation
MERR method	Prove the consistency under mild conditions, more general; no need of kernel density estimation as an intermediate step	Data are assumed to be i.i.d.
SMM	First to propose a learning algorithm on probability distributions	lack the analysis of consistency and computational cost
OCSMM	Group anomaly detection	computational expensive
Fastfood + MERR	interdisciplinary applications; fast	Approximations to obtain the embeddings
Latent SMM	solve the bag-of-words data classification problem	Some parameters need to be fixed beforehand
Learning with additional distributions	incorporate the unlabeled data	need kernel density estimation
Distributional data learning	encode distributional data into tuples of attribute values by aggregation/generative methods/discriminative methods	cannot capture all the information in the distributions

Experiments Setup

Data Sets

Datasets	Features	Instances	Classes	Subjects
Opportunity	113	674926	4	4
Skoda	60	696975	10	1

Baseline Methods

Methods	Algorithms
Frame-based	SVM kNN
Segment-based	1entry+SVM 1entry+kNN miFV

Performance Metric

Evaluation in 2 aspects: frame, event F1 score:

- microaverage F1 score (miF)

$$\begin{aligned} miF &= 2 \times \frac{precision_{all} \times recall_{all}}{precision_{all} + recall_{all}} \\ precision_{all} &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \\ recall_{all} &= \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \end{aligned} \tag{7}$$

- weighted macroaverage F1 score (maF)

$$maF = 2 \times \sum_i w_i \frac{precision_i \times recall_i}{precision_i + recall_i} \tag{8}$$

Experiments Results

Table 1: Experiments results on two data sets (unit: %)

Methods	Algorithms	Opportunity miF_segment	Opportunity maF_segment	Skoda miF_segment	Skoda maF_segment
	Proposed	**77.286**	**75.673**	**98.895**	**98.887**
Segment-based	1entry+SVM	74.036	72.997	92.279	92.213
	1entry+kNN	75.381	75.186	91.588	91.518
	miFV	25.577	24.009	61.877	55.086
	Proposed	**77.286**	**75.673**	**98.895**	**98.887**
Frame-based	SVM	75.914	75.423	81.819	77.062
	kNN	74.333	73.952	92.618	92.044

Experiments Results

Table 1: Experiments results on two data sets (unit: %)

Methods	Algorithms	Opportunity miF_segment	Opportunity maF_segment	Skoda miF_segment	Skoda maF_segment
	Proposed	**77.286**	**75.673**	**98.895**	**98.887**
Segment-based	1entry+SVM	74.036	72.997	92.279	92.213
	1entry+kNN	75.381	75.186	91.588	91.518
	miFV	25.577	24.009	61.877	55.086
	Proposed	**77.286**	**75.673**	**98.895**	**98.887**
Frame-based	SVM	75.914	75.423	81.819	77.062
	kNN	74.333	73.952	92.618	92.044

- miFV: Null class interruption and imbalanced multi-class

Choices of kernels

Table 2: miF_event of proposed method on Skoda data set with different kernels

		$\tilde{k}(\cdot,\cdot)$ LIN	POLY3	RBF	SIG
$k(\cdot,\cdot)$	LIN	91.4300	91.3852	91.3632	28.6446
	POLY3	98.1202	98.0728	98.1556	92.0938
	RBF	98.1422	90.8818	**98.8950**	98.3728
	SIG	87.7026	87.0830	90.4140	90.4176

Choices of kernels

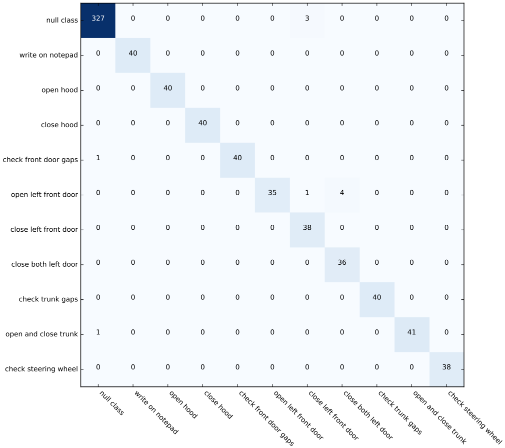
Table 2: miF_event of proposed method on Skoda data set with different kernels

		$\tilde{k}(\cdot,\cdot)$			
		LIN	POLY3	RBF	SIG
$k(\cdot,\cdot)$	LIN	91.4300	91.3852	91.3632	28.6446
	POLY3	98.1202	98.0728	98.1556	92.0938
	RBF	98.1422	90.8818	**98.8950**	98.3728
	SIG	87.7026	87.0830	90.4140	90.4176

PD kernels better than non-PD kernels

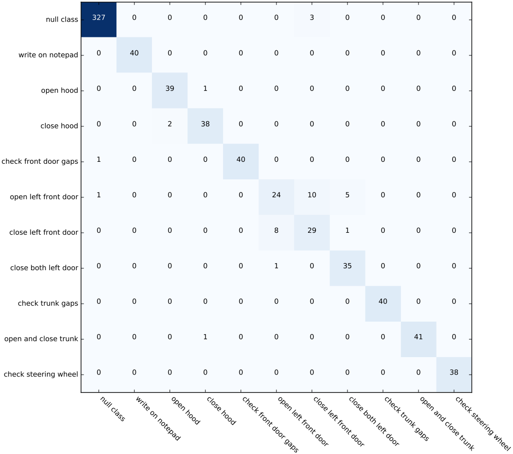
Interclass similarity

Confusion matrix of proposed method



Interclass similarity

Confusion matrix of 1entry+SVM method



Interclass similarity

Confusion matrix of SVM method

	null class	write on notepad	open hood	close hood	check front door gaps	open left front door	close left front door	close both left door	check trunk gaps	open and close trunk	check steering wheel
null class	329	1	0	0	0	0	0	0	0	0	0
write on notepad	0	40	0	0	0	0	0	0	0	0	0
open hood	0	0	40	0	0	0	0	0	0	0	0
close hood	0	0	0	40	0	0	0	0	0	0	0
check front door gaps	1	0	0	0	40	0	0	0	0	0	0
open left front door	2	0	0	0	0	0	0	38	0	0	0
close left front door	2	0	0	0	0	0	2	34	0	0	0
close both left door	0	0	0	0	0	0	0	36	0	0	0
check trunk gaps	0	0	0	0	0	0	0	0	40	0	0
open and close trunk	1	0	0	0	0	0	0	0	0	41	0
check steering wheel	0	0	0	0	0	0	0	0	0	0	38

Conclusion

- A new way to represent activity data: probability distributions (compared to commonly fixed-length vectors)
- A novel feature representation for time-series activity data via kernel embedding
  - Extract more discriminative features from sensor data
  - Robustness to Null class and imbalanced class cases
  - Handle interclass similarity problems
- Experiments results illustrate the efficacy of the proposed method

Future Work

- Extend the framework to more applications
- Extend the framework for large-scale data
- Develop the semi-supervised and unsupervised learning framework for the activity recognition problems

References I

- Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.
- Alex Smola et al. "A Hilbert space embedding for distributions". In: *International Conference on Algorithmic Learning Theory*. Springer. 2007, pp. 13–31.