



Community Similarity Networks

Jindong Wang

2015.09.28

Basics

- **Lane N D, Xu Y, Lu H, et al. Enabling large-scale human activity inference on smartphones using community similarity networks (csn)[C]**//Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011: 355-364.
- **Lane N D, Xu Y, Lu H, et al. Community Similarity Networks[J]**. Personal and ubiquitous computing, 2014, 18(2): 355-368.

Contents

1

Author

2

Motivation

3

Introduction

4

Method

5

Evaluation

6

Conclusion

Author



Nicolas Lane
<http://niclane.org>

- Now: Bell Labs (2015.1--)
- Past: MSRA (Beijing)
- Interests: people--centric sensor data, community, mobile
- 2011 Ph.D. in cs Dartmouth
- Thesis: "*Community--guided Mobile Phone Sensing Systems*"
- Advisors: Andrew T. Campbell and Tanzeem Choudhury
- Best paper (nominee) of UbiComp & MobiSys at 2011, 2012, 2014
- UbiComp 15:3 papers accepted

Motivation(1)

- Sensor-enabled smartphones excellently helps activity recognition
- As user population increase, the difference between people cause the accuracy of classification to degrade quickly
- Population diversity problem
 - Age, sex, behavior patterns...
- Conventional approaches
 - One for all ---- struggle
 - One for one ---- burdensome



Motivation(2)

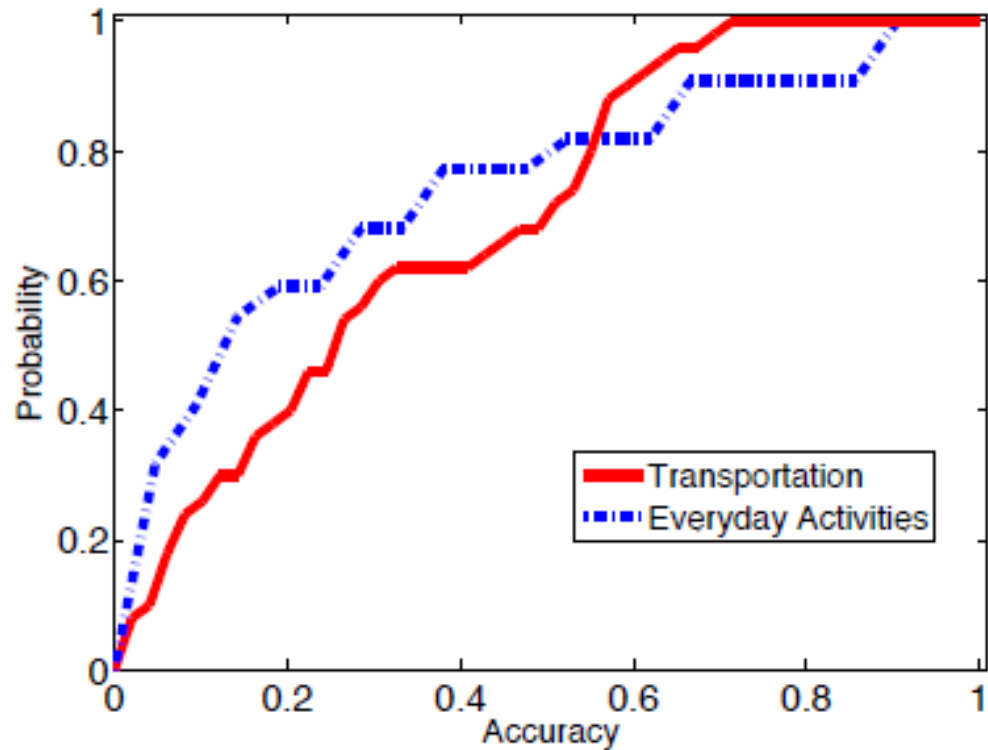
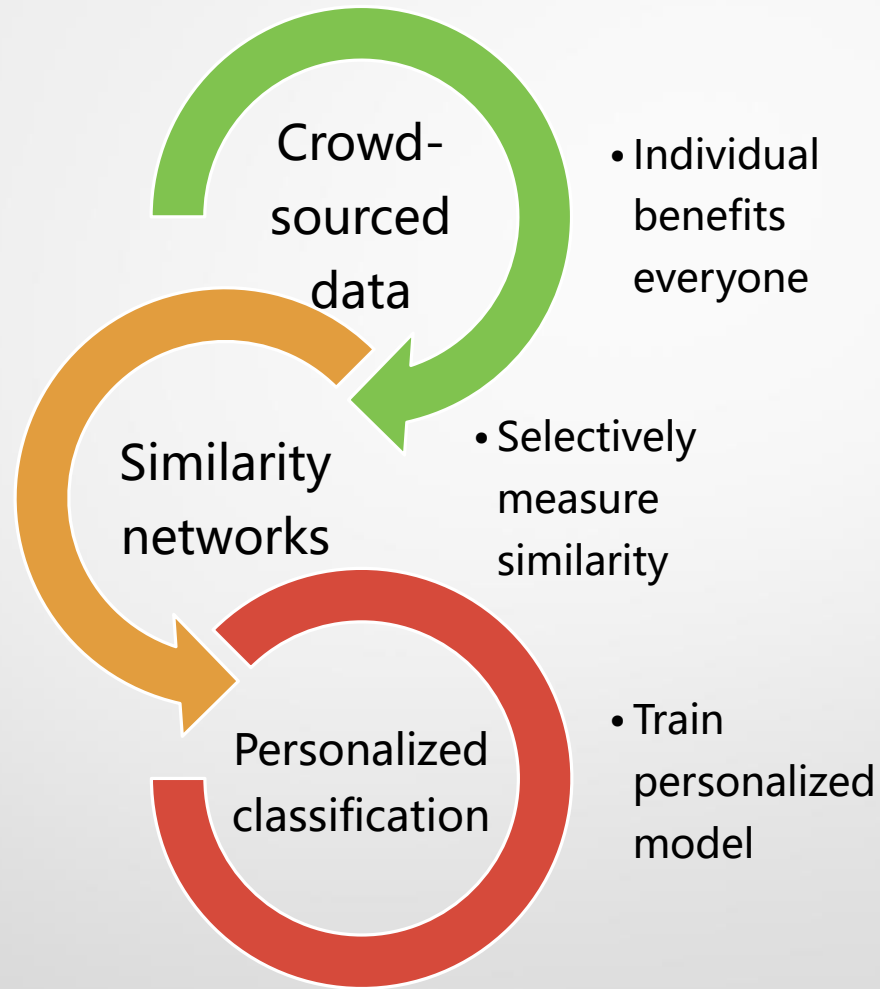


Figure 2. Classification accuracy varies significantly within a large-scale user population for two datasets, one containing everyday activities and the other transportation modes.

Introduction(1)

- Community similarity networks(CSN)
- A classification system that can be incorporated into mobile sensing system to address the challenge to robust classification caused by PDP
- Make **personalized** model by lowering the user burden with **crowd-sourced data** and **similarity networks**
 - Physical, lifestyle, sensor data

Introduction(2)



Method(1)

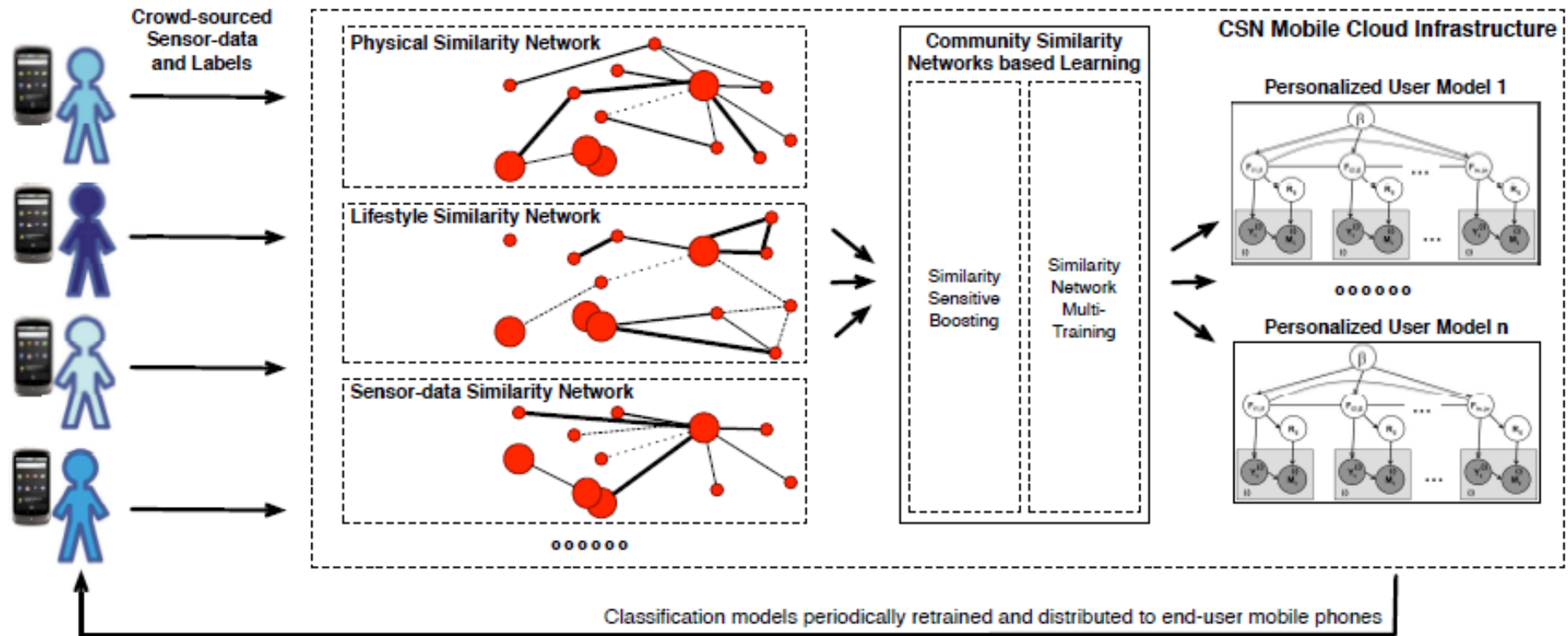
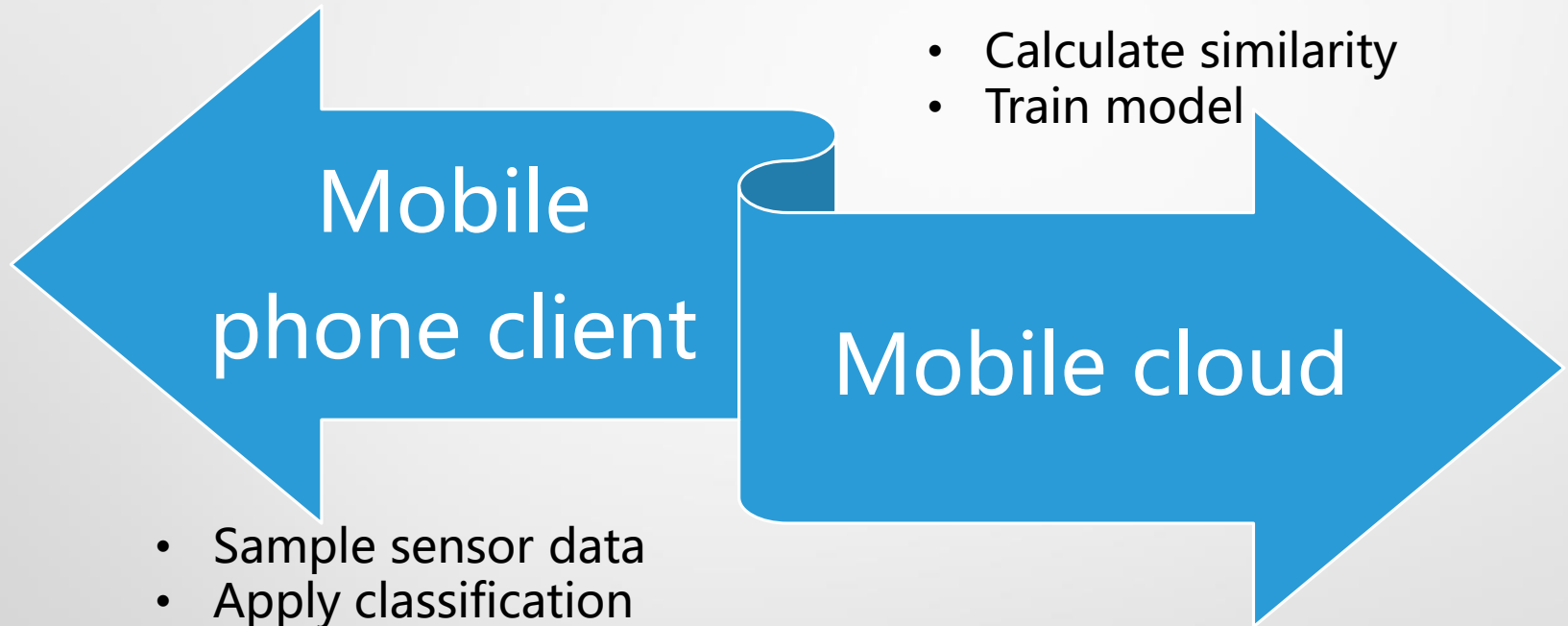



Figure 3. The processing phases within Community Similarity Networks

Method(2)




Method(3)

- Mobile phone client features

A decorative arrow pointing to the right, with a white body and a multi-colored border (green, orange, red, pink, purple) on the top and bottom edges.

Naïve Bayes classification
Markov model smooth

framework

A decorative arrow pointing to the right, with a white body and a multi-colored border (red, pink, purple, green, orange) on the top and bottom edges.

Classification pipeline
Supporting services

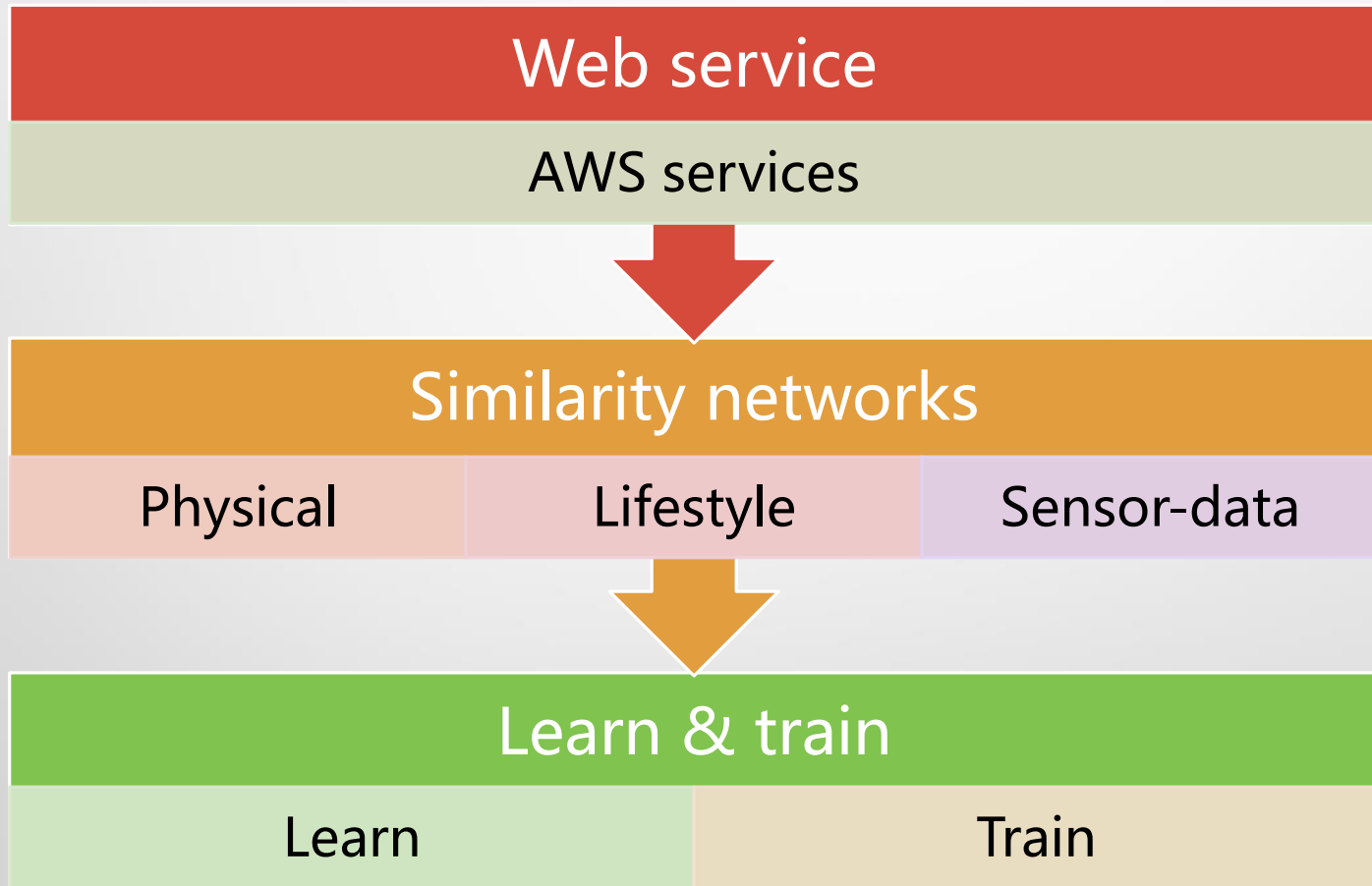
users

A decorative arrow pointing to the right, with a white body and a multi-colored border (purple, green, orange, red, pink, purple) on the top and bottom edges.

Answer certain questions
Explicitly label data

Method(4)

- Mobile Cloud



Method(5)

- Physical similarity
 - Features:
 - Weight, height, age, Yale Physical Activity Survey & SF-36 physical activity score
 - Metrics:

$$sim(i, j)^{phy} = \exp(-\gamma(\mathbf{x}_i - \mathbf{x}_j)^\top \Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)) \quad (1)$$

where, \mathbf{x}_i and \mathbf{x}_j are the physical vectors for user i and user j , Σ is the covariance matrix and γ is an empirically determined scaling parameter.

Method(6)

- Lifestyle similarity(1)
 - The diversity in how people lives their lives
 - Information:
 - **Mobility**, based on GPS location, tessellated into m distinct square tiles of equal size
 - **diurnal patterns**, a series of timestamps whenever the user is non-stationary, particular hour in a week
 - **distribution of activities**, duration user are supposed to perform certain activities

Method(7)

- Lifestyle similarity(2)
 - Construct 3 histograms for every user & normalization
 - Metrics:

$$sim(i, j)^{life} = \sum_{f \in \mathcal{F}} \mathbf{T}_f(i)^\top \mathbf{T}_f(j) \quad (2)$$

where $\mathbf{T}_f(i)$ is a histogram vector for user i of type f and \mathcal{F} contains each type of lifestyle histogram. Lifestyle similarity between two users is the sum of the inner product of the histograms for each type of lifestyle information used by CSN.

Method(8)

- Sensor-data similarity(1)
 - A pure data-driven approach
 - Features:
 - Accelerometer, microphone and GPS data
 - Metrics:

$$sim(i, j)^{data} = \frac{1}{N_i N_j} \sum_{l=1}^{N_i} \sum_{m=1}^{N_j} sim(x_{il}, x_{jm}) \quad (3)$$

where, $\{x_{il}, l = 1 : N_i\}$ is the data of user i , and $\{x_{jm}, m = 1 : N_j\}$ is the data of user j .

Method(9)

- Sensor-data similarity(2)
 - Problem: **impractical** as the number increases
 - Adopt **Locality Sensitive Hashing**(LSH)
 - A hashing function family can capture the similarity between data
 - Similar data is likely to share the same value after hash mapping:

$$Pr_{h \in \mathcal{H}}[h(x_1) = h(x_2)] = s_{\mathcal{H}}(x_1, x_2) = E_{h \in \mathcal{H}}[s_h(x_1, x_2)] \quad (4)$$

Therein, $x_1, x_2 \in \mathcal{X}$ are two data, \mathcal{H} is a LSH family, h is the hash function sampled from \mathcal{H} , and $s_{\mathcal{H}}$ is a similarity measure of \mathcal{X} , which is induced by the LSH family \mathcal{H} [7].

Method(10)

- Sensor-data similarity(3)
 - Randomly choose B independent 0/1 valued hashing functions $\{h_j\}$ from the random projection for L_2 distance LSH family
 - Form a B -bit hash function $f=(h_1, h_2, \dots, h_B)$
 - The histogram T_f for any user i :

$$\mathbf{T}_f(i) = \sum_{x_{il} \in i} e[f(x_{il})] \quad (5)$$

here, $\{x_{il}, l = 1 : N_i\}$ is data of user i , and $\mathbf{T}_f(i)$ is determined by the hash function f sampled from \mathcal{F} .

- $\{e[l] \mid l \in D\}$ be the standard basis of the $|2B|$ -dimensional vector space

Method(11)

- Sensor-data similarity(4)
 - Thus each element of histogram vector $T_f(i)$ can be regarded as a bin

$$sim(i, j)^{data} = \mathbf{T}_f(i)^\top \mathbf{T}_f(j) \quad (6)$$

To estimate the expectation shown in Eq. 5, we construct several histograms $f \in \mathcal{F}$ for each user and compute an average value using Eq. 6.

Method(12)

- Learning

Similarity-
sensitive
boosting

- Train 3 separate classifiers for each type

Similarity
network
multi-
training

- Semi-supervised approach to recruit additional labels
- Unify the 3 classification models

Method(13)

- Learning: similarity-sensitive boosting
 - Using boosting strategy
 - Impose an additional weight based on **similarity**:

$$weight^{(0)}(x_k) = sim(i, k)$$

where, k indicates the user who produces the data x_k .

- Sim(i,k):edge weight between two individuals
- Those similar to user I will be weighted highly
- Weak classifier be replaced with feasible ones

Method(14)

- Learning: similarity network multi-training
 - Boosting ensures good different types of models
 - Exploit the strengths of each network by multi-training
 - Multi-training: semi-supervised training algorithm to utilize multiple complementary views of the same labeled training data to generate additional labels

Method(15)

- Learning: similarity network multi-training
 - Start with 3 classifiers
 - Iterative process when classifiers are used to label data
 - Then each classifier is retrained
 - Result labels are accepted only agreement made
 - Stops when conditions are made(number of recruited labels)

Evaluation(1)

Two large real-world datasets

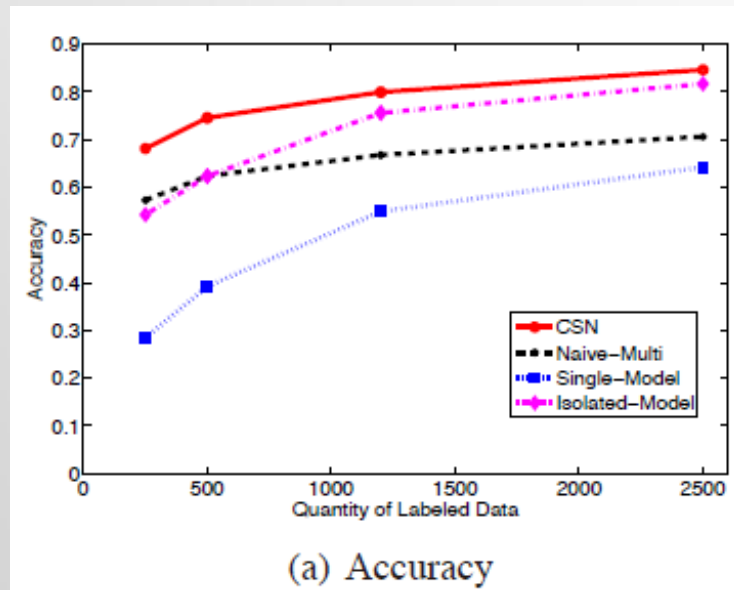
- Everyday activities (41; Nexus One; walk; meeting)
- Transportation (51;bike, bus, car, walk)

Three benchmarks

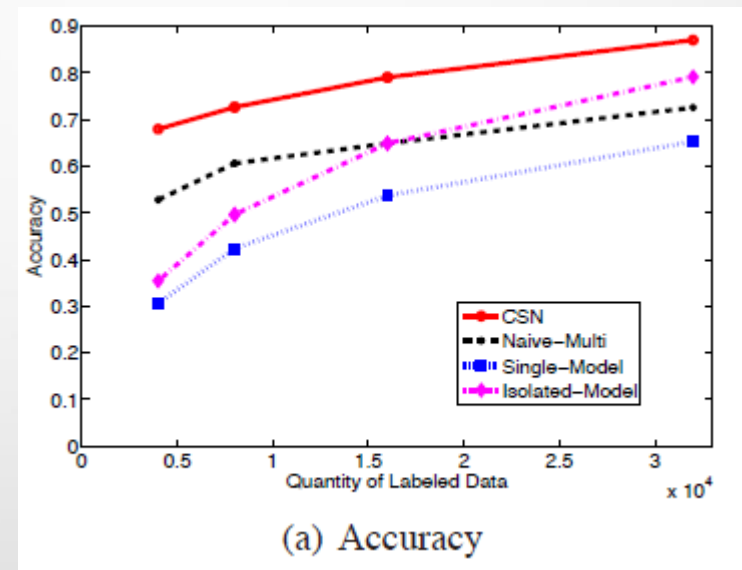
- **Single** (Same model for all users)
- **Isolated** (every user has own isolated model)
- **Naïve-multi** (exactly use full CSN)

Evaluation(2)

- Robust classification with low user burden
 - CSN is more robust with lower user burden
 - Achieve higher accuracy and evenly distributed



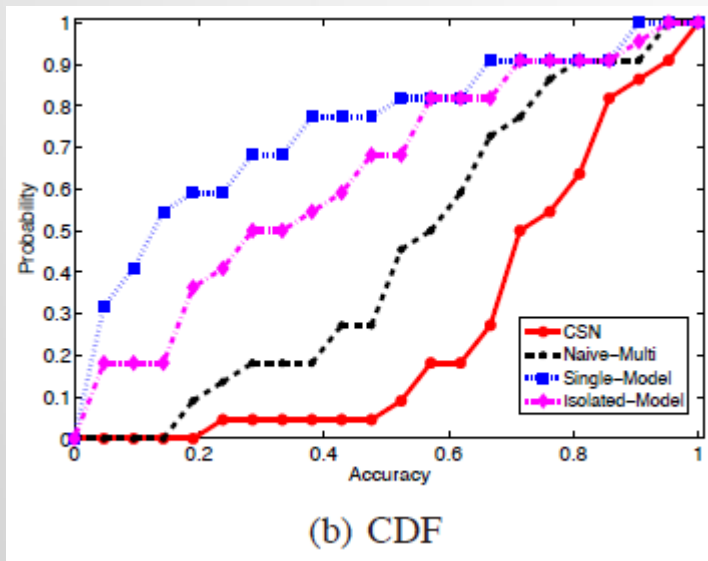
Everyday activities



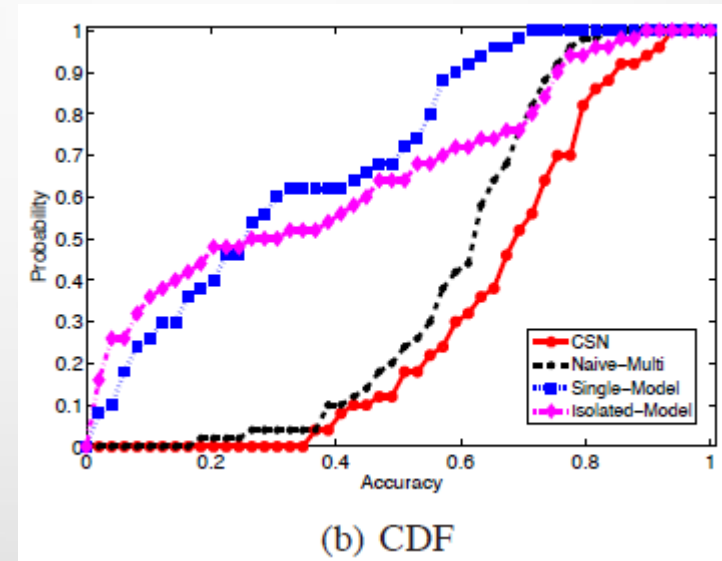
Transportation

Evaluation(3)

- Robust classification with low user burden



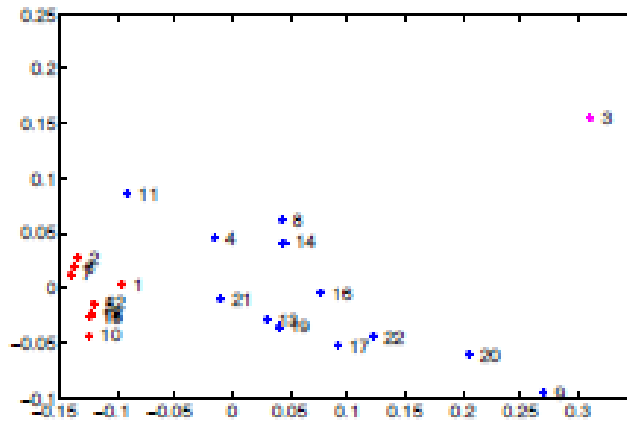
Everyday activities



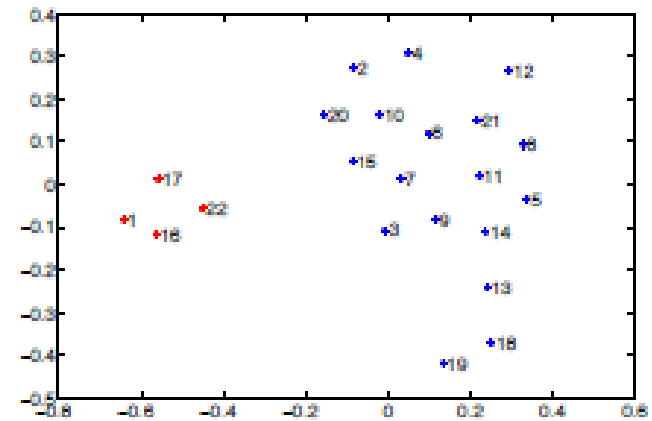
Transportation

Evaluation(4)

- Benefits of leveraging similarity networks
 - Collected additional demographic information from 22 from everyday activities to test if CSN captures meaningful differences between people



(a) Physical

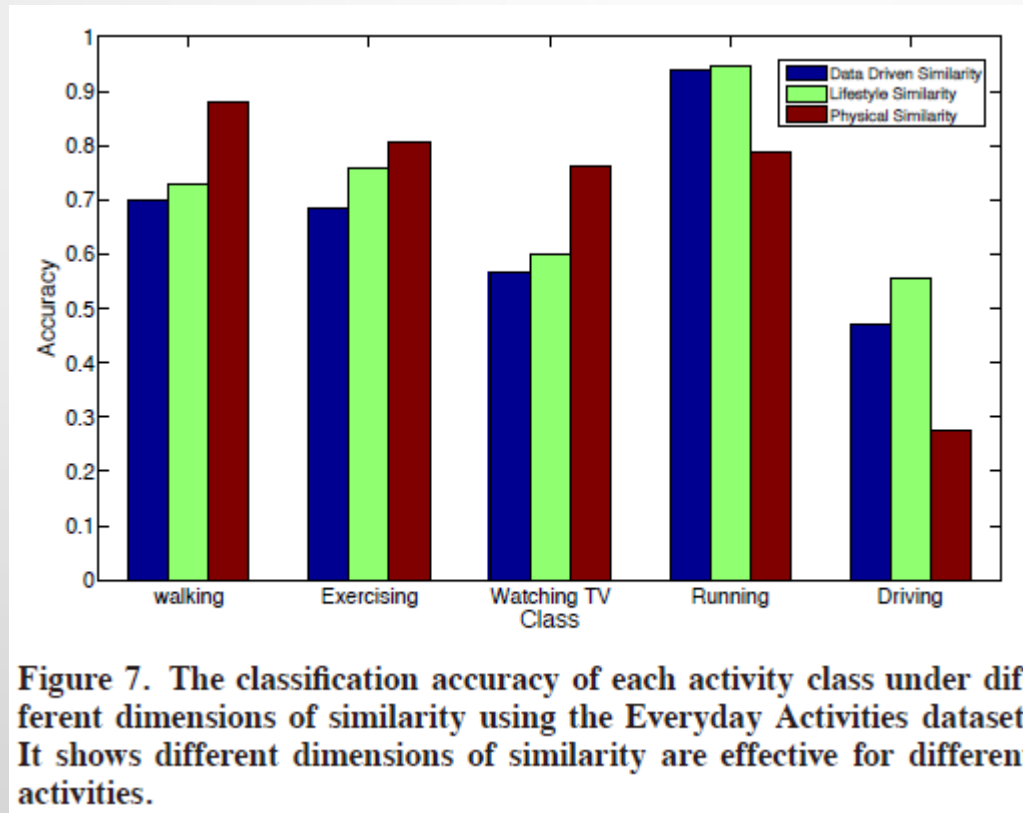


(b) Lifestyle

Figure 6. MDS projection of physical and lifestyle similarity networks used by CSN.

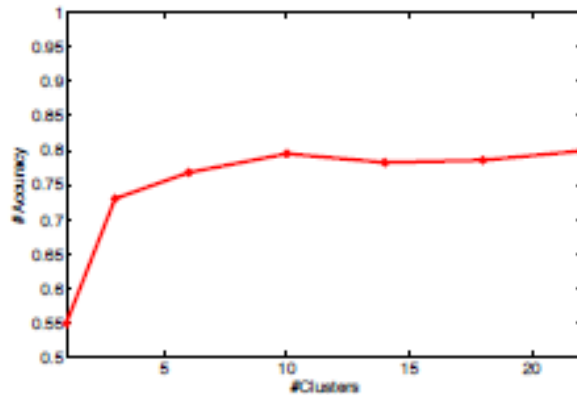
Evaluation(5)

- Benefits of leveraging similarity networks
 - Distances between points in these figures are proportional to differences in similarity.

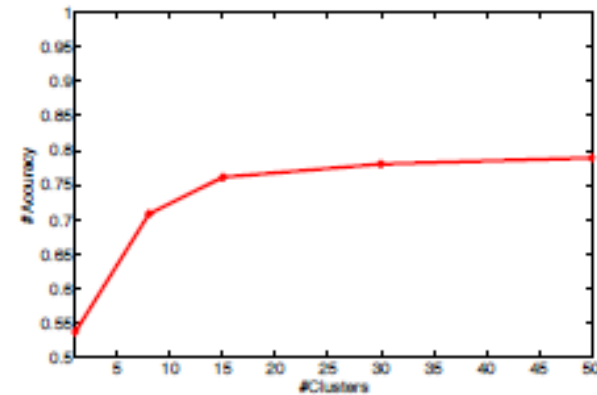


Evaluation(6)

- Cloud scalability with low phone overhead
 - Upload strategy: not until recharging
 - Using clustering to reduce training time
 - 400 GB of data, 200/9/3 minutes for 3 similarity types



(a) Everyday Activities



(b) Transportation

Figure 8. The accuracy of CSN when we group the users into different number of clusters under both datasets.

Conclusion

- Contributions
 - First system to propose embedding inter-person similarity within the training process
 - Support the extraction of similarity networks from raw data and end-user input
 - Evaluate with two large real-world datasets
- Drawbacks
 - Clustering technique not helpful when #clusters grows
 - Only on mobile phone



Q & A

Thank you



Community Similarity Networks

Jindong Wang

2015.09.28