电子科技大学
University of Electronic Science and Technology of China

Label Distribution Learning

Wei Han

Data Mining Lab,
Big Data Research Center, UESTC
Email： wei.hb.han@gmail.com

Outline

1. Why label distribution learning?

2. What is label distribution learning?
   2.1. Problem Formulation
   2.2. Difference with Related Works

3. How to label distribution learning?
   3.1. Problem Transformation
   3.2. Algorithm Adaptation
   3.3. Specialized Algorithms

4. Evaluation Measurement

5. Application

Section 1

Motivation

Label Ambiguity

"What subjects describe the instance?"

Multi-label Learning

Label Ambiguity

"How to describe the instance?"

some cloud

mostly sky

much water

a bit of building

Label Distribution Learning

More information

数据挖掘实验室
Data Mining Lab

- Multi-label Learning (MLL)

  Thresholding ➡ Positive labels ➡ MLL

  Lose information!

  Trade-off？

- Label Distribution Learning (LDL)
  - Assign a real number to each label
    - Importance
    - Confidence
    - Probability
    - Level
    - ……

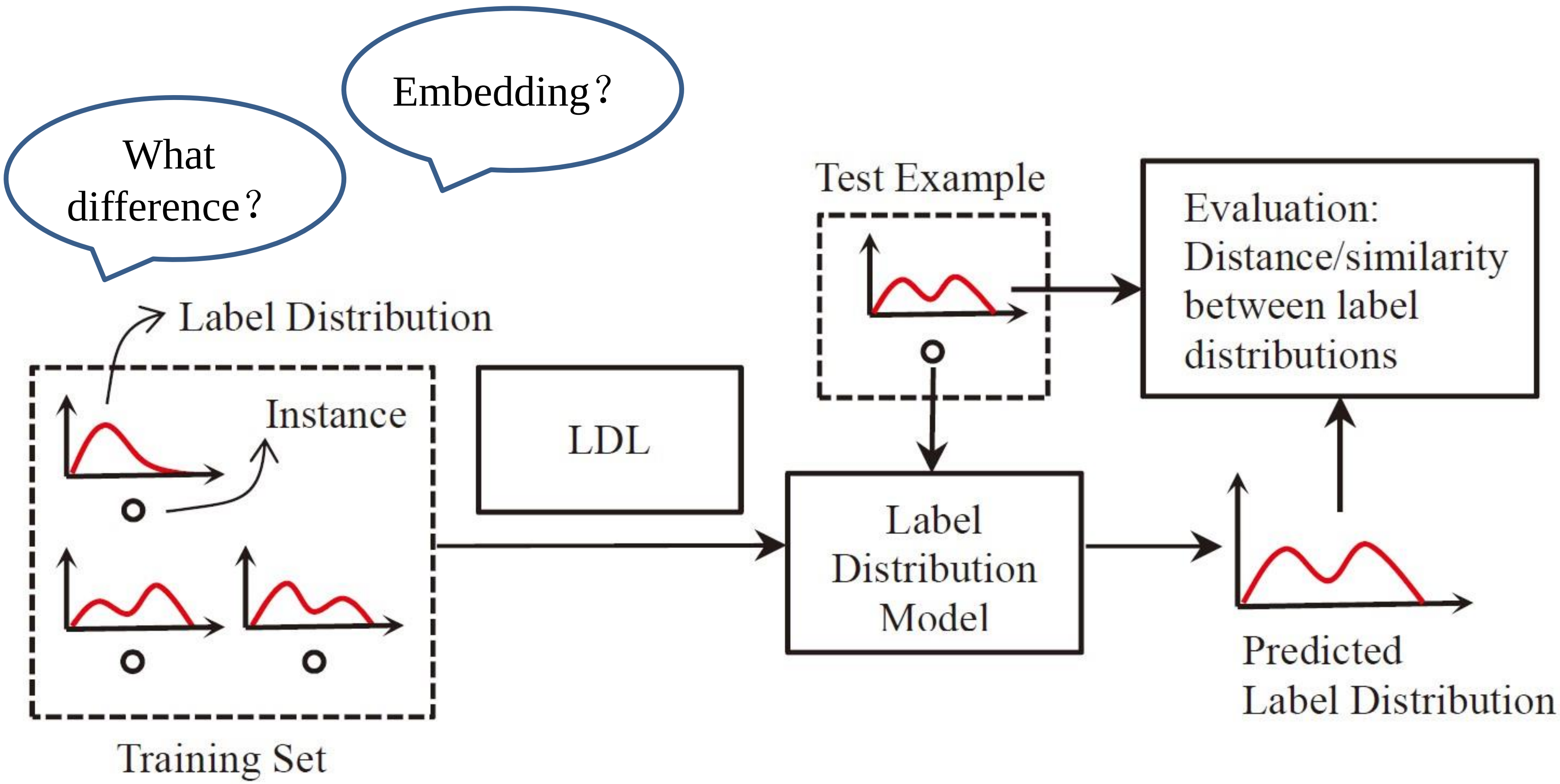
  Keep more, learn more

Section 2
Definition

Problem Formulation

Description degree

A real number d_x^y is assigned to the label y for the instance x

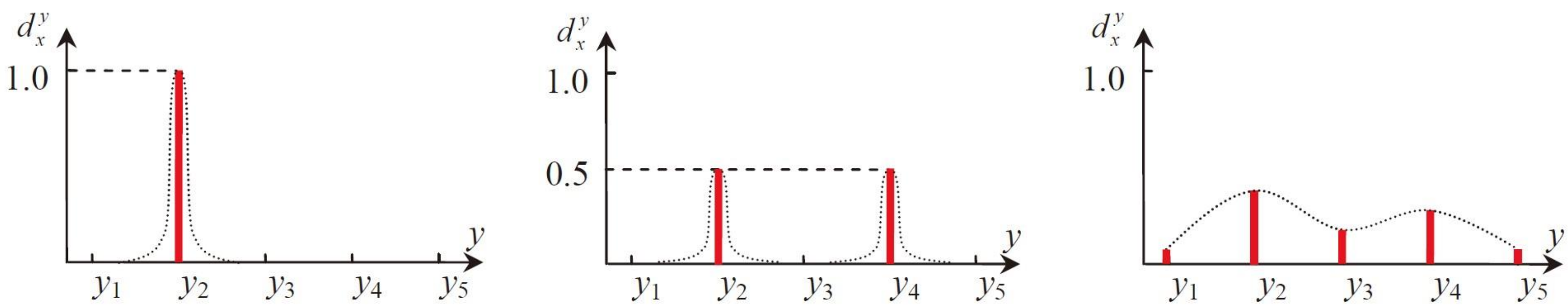


Problem Formulation

Label Distribution Constraints:

- $d_x^y \in [0,1]$
- $\sum_y d_x^y = 1$

Loose constraints?



(a) Single-label (b) Multi-label (c) General case

More Details

For the particular instance x_i, the label distribution is denoted by $D_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \cdots, d_{x_i}^{y_c}\}$, where c is the number of possible label values.

Input space : $X \in \mathbb{R}^q$
Complete label set : $Y = \{y_1, y_2, \ldots, y_c\}$:
Training set : $S = \{(x_1; D_1), (x_2; D_2), \cdots, (x_n; D_n)\}$
Output of LDL : $p(y|x; \theta)$, where $x \in X$ and $y \in Y$

Given the training set S, the goal of LDL is to find the θ that can generate a distribution similar to D_i given the instance x_i

More Details

For example, if the Kullback-Leibler divergence is used as the distance measure, then the best parameter vector θ^* is determined by

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i \sum_j \left(d_{x_i}^{y_j} \ln \frac{d_{x_i}^{y_j}}{p(y_j|x_i;\theta)} \right)$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_i \sum_j d_{x_i}^{y_j} \ln p(y_j|x_i;\theta)$$

More Details

Consequently, the simplified equation for SLL is

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \ln p(y(x_i)|x_i;\ \theta)$$

This is actually the maximum likelihood (ML) estimation of θ.

For MLL, the modified equation is

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i \frac{1}{|Y_i|} \sum_{y \in Y_i} \ln p(y|x_i;\ \theta)$$

In fact, this is equivalent to transform the multi-label instances into the weighted single-label instances, and then optimizing the ML criterion based on the weighted single-label instances.

Related Works

数据挖掘实验室
Data Mining Lab

In existing **single-label** or **multi-label** machine learning literatures, an intermediate numerical indicator (e.g., probability, confidence, grade, score, vote, etc.) for each label is not rare.

Differences:

1. Different forms of label

2. Different points of concern

3. Different evaluation measurement

Related Works

Label embedding and **attribute learning** are featured by the intermediate representations for the classes.

Label embedding: Projects the class labels into a subspace

Attribute Learning: Leverage the prior knowledge of attribute-class association to deal with the missing classes (zero-shot learning)

Key point: Each instance is still associated with one class label

Related Works

The basic assumption of **probabilistic label** is that there is only one ‘correct’ label for each instance.

Note also that $\mathrm{d}_{\mathrm{x}}^{\mathrm{y}}$ is not the probability that y correctly labels x, but the proportion that y accounts for in a full description of x.

Fortunately, although not a probability by definition, d_x^y still shares the same constraints with probability

Section 3

Methodology

Category

- Problem Transformation
  Transform an LDL problem into an SLL problem

- Algorithm Adaptation
  Extend existing algorithms to address LDL problem

- Specialized Algorithms
  Directly match the LDL problem

PT-Bayes/PT-SVM

'PT' is the short of 'Problem Transformation'

The core idea is change the training examples into weighted single-label examples

Multi-instance learning?

Resample training set according to the weight of each example. A standard single-label training set including $c \times n$ examples.

The learner must be able to output the confidence/probability for each label y_j. Two representative algorithms, Bayes and SVM, are adopted here for this purpose.

‘AA’ is the short of ‘Algorithm Adaptation’

AA-kNN

Given a new instance x, its k nearest neighbors are first found in the training set. Then, the mean of the label distributions of all the k nearest neighbors is calculated as the label distribution of x.

$$p(y_j|x) = \frac{1}{k} \sum_{i \in N_k(x)} d_{x_i}^{y_j}, (j = 1, 2, \dots, c)$$

AA-BP

The three-layer neural network has q input units, and c output units. The objective function of the BP algorithm is to minimize the sum-squared error. To make sure label distribution constraints, the softmax activation function is used in each output unit.

$$z_j = \frac{\exp(\eta_j)}{\sum_{k=1}^{c} \exp(\eta_k)}, (j = 1, 2, \dots, c)$$

SA-IIS/SA-BFGS

‘SA’ is the short of ‘Algorithm Adaptation’

Assumes the parametric model $p(y|x;\ \theta)$ to be the maximum entropy model:

$$p(y|x;\ \theta) = \frac{1}{Z}\exp\left(\sum_k \theta_{y,k} g_k(x)\right)$$

Where $Z = \sum_y \exp\left(\sum_k \theta_{y,k} g_k(x)\right)$ is the normalization factor, $g_k(x)$ is the k-th feature of x

Then, the objective function of θ is

$$J(\theta) = \sum_{i,j} d_{x_i}^{y_j} \sum_k \theta_{y_i,k} g_k(x_i) - \sum_i \ln \sum_j exp\left(\sum_k \theta_{y_i,k} g_k(x_i)\right)$$

Afterwards, the Improved Iterative Scaling (IIS) and quasi-Newton method BFGS are used to optimize it.

Section 4
Evaluation Measurement

Distance/Similarity Measurement

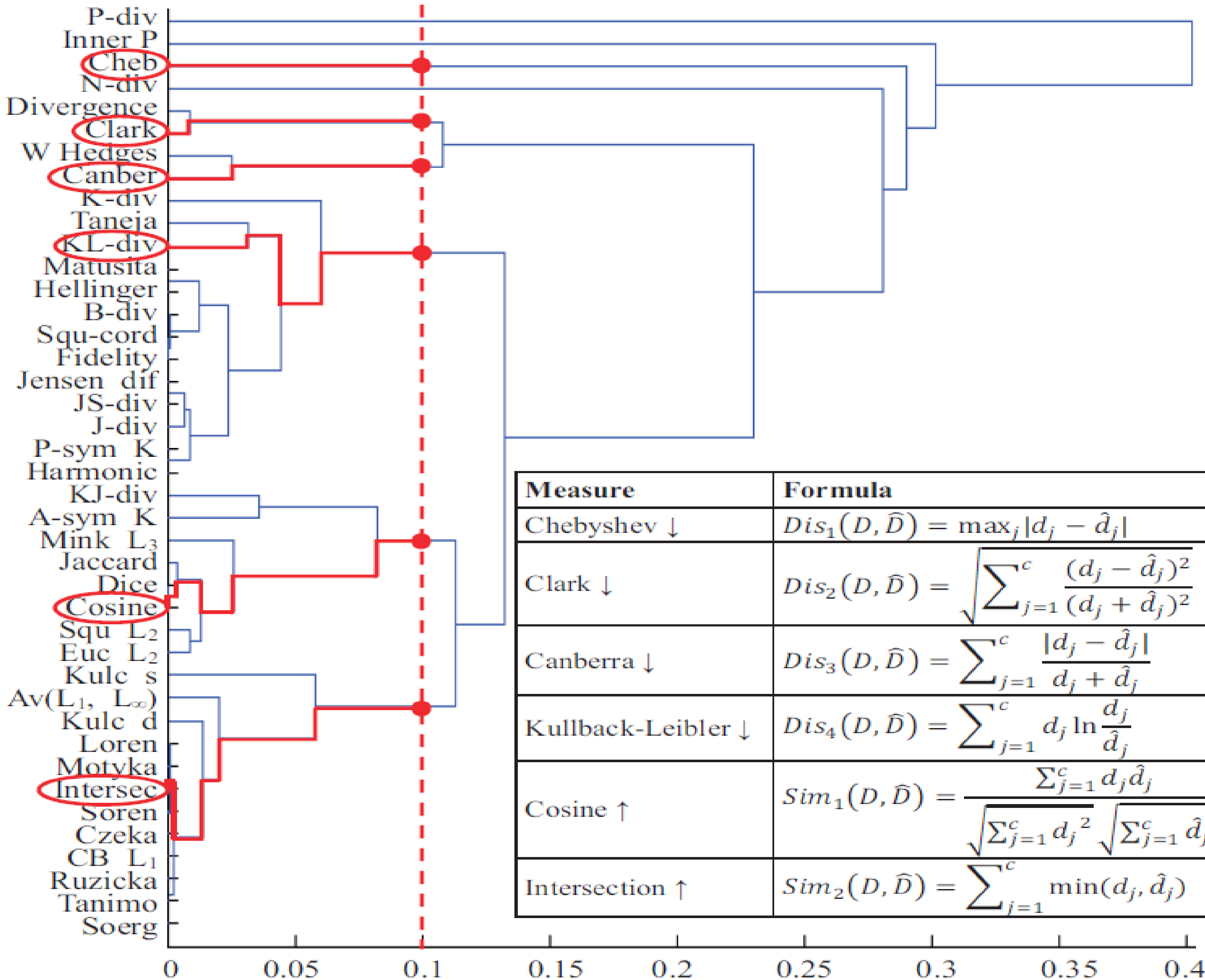
GAN中的度量?

Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions summarized 41 Distance or Similarity Measures from 8 syntactic families.

On a particular dataset, each of the measures may reflect a certain aspect of an algorithm. In order to obtain a set of representative and diverse measures, six measures are finally selected.

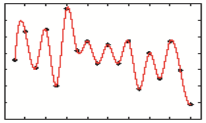
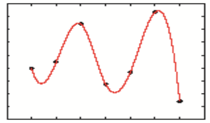
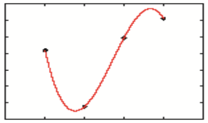
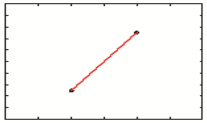
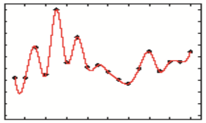
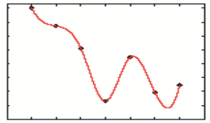
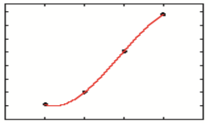
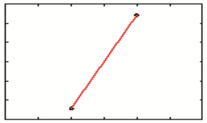
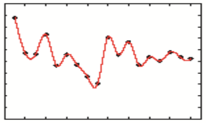
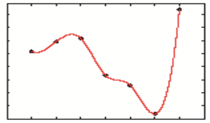
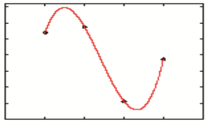
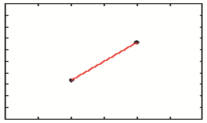
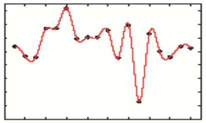
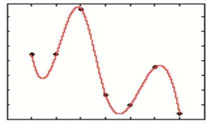
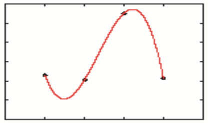
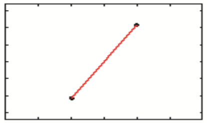
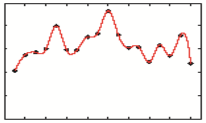
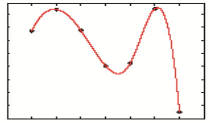
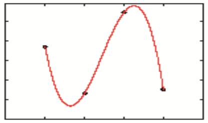
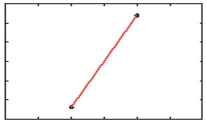
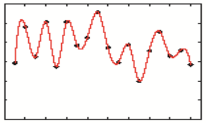
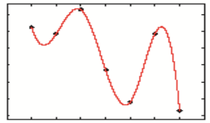
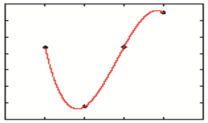
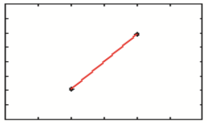
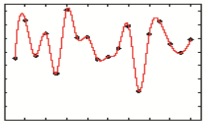
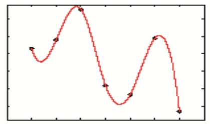
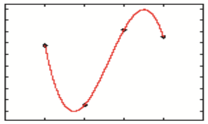
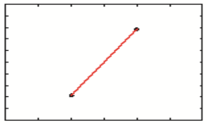
1. The distance between the clusters of any two measures in the set is greater than 0.1 (indicated by the red dash line in the following figure);
2. Each measure in the set comes from a different syntactic family;
3. The selected measures are relatively widely used in the related areas.

Distance/Similarity Measurement



Measure	Formula
Chebyshev ↓	$Dis_1(D,\hat{D}) = \max_j \lvert d_j - \hat{d}_j \rvert$
Clark ↓	$Dis_2(D,\hat{D}) = \sqrt{\sum_{j=1}^{c} \frac{(d_j - \hat{d}_j)^2}{(d_j + \hat{d}_j)^2}}$
Canberra ↓	$Dis_3(D,\hat{D}) = \sum_{j=1}^{c} \frac{\lvert d_j - \hat{d}_j \rvert}{d_j + \hat{d}_j}$
Kullback-Leibler ↓	$Dis_4(D,\hat{D}) = \sum_{j=1}^{c} d_j \ln \frac{d_j}{\hat{d}_j}$
Cosine ↑	$Sim_1(D,\hat{D}) = \frac{\sum_{j=1}^{c} d_j \hat{d}_j}{\sqrt{\sum_{j=1}^{c} {d_j}^2}\sqrt{\sum_{j=1}^{c} {\hat{d}_j}^2}}$
Intersection ↑	$Sim_2(D,\hat{D}) = \sum_{j=1}^{c} \min(d_j, \hat{d}_j)$

Results

Real	 18 Labels	 7 Labels	 4 Labels	 2 Labels
PT-Bayes	 {0.048, 0.552, 1.885, 0.036} {0.960 0.891}	 {0.042, 0.230, 0.492, 0.016} {0.984, 0.929}	 {0.170, 0.516, 0.911, 0.128} {0.907, 0.786}	 {0.192, 0.310, 0.421, 0.086} {0.939, 0.808}
PT-SVM	 {0.006, 0.105, 0.368, 0.0012} {0.9988, 0.980}	 {0.040, 0.190, 0.359, 0.011} {0.990, 0.949}	 {0.041, 0.108, 0.188, 0.006} {0.995 0.954}	 {0.018, 0.025, 0.036, 0.001} {0.999, 0.982}
AA-*k*NN	 {0.005, 0.091, 0.282, 0.0009} {0.9991, 0.9844}	 {0.028, 0.151, 0.381, 0.007} {0.993, 0.945}	 {0.030, 0.080, 0.124, 0.003} {0.997, 0.970}	 {0.058, 0.085, 0.118, 0.007} {0.994, 0.942}
AA-BP	 {0.008, 0.156, 0.583, 0.003} {0.997, 0.968}	 {0.014, 0.076, 0.172, 0.002} {0.998, 0.975}	 {0.036, 0.085, 0.145 0.004} {0.996, 0.964}	 {0.069, 0.102, 0.143, 0.010} {0.991, 0.931}
SA-IIS	 {**0.004**, **0.083**, 0.283, **0.00077**} {**0.99923**, 0.9843}	 {0.013, 0.072, 0.160, 0.002} {0.998, 0.977}	 {0.016, 0.042 0.067 0.0008} {0.9992, 0.984}	 {0.012, 0.018, 0.025, 0.0003} {0.9997, 0.988}
SA-BFGS	 {0.006, 0.086, **0.260**, 0.00081} {0.99919, **0.986**}	 {**0.012, 0.056, 0.120, 0.001**} {**0.999, 0.983**}	 {**0.014, 0.034, 0.055, 0.0006**} {**0.9994, 0.987**}	 {**0.007, 0.010, 0.014, 0.0001**} {**0.9999, 0.993**}

Experimental Results on the Artificial Dataset

Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS
Chebyshev ↓	0.080(3)	0.653(6)	0.086(4)	0.101(5)	0.0767(2)	**0.0766**(1)
Clark ↓	**0.341**(1)	1.135(6)	0.382(4)	0.520(5)	0.349(2)	0.352(3)
Canberra ↓	**0.488**(1)	1.823(6)	0.564(4)	0.699(5)	0.489(2)	0.495(3)
Kullback-Leibler ↓	0.030(3)	1.482(6)	0.035(4)	0.066(5)	**0.029**(1)	0.030(2)
Cosine ↑	0.990(3)	0.377(6)	0.989(4)	0.983(5)	0.99116(2)	**0.99120**(1)
Intersection ↑	0.920(3)	0.347(6)	0.914(4)	0.899(5)	0.9233(2)	**0.9234**(1)
Avg. Rank	2.33	6.00	4.00	5.00	1.83	1.83
Running Time (ms)	22 / 45	391 / 1,153	0 / 79,961	101,568 / 149	1,168 / 33	187 / 33

Meanwhile, in 15 public real-world datasets[1], SA-BFGS obtains all the first place under six measures.

1. The datasets and the Matlab code of the LDL algorithms are available at web site: http://cse.seu.edu.cn/PersonalPage/xgeng/LDL/index.htm

Section 5

Application

Where are the label distribution?

- <u>They come with the original data</u>
  - Emotion Distribution [Zhou, Xue and Geng, ACMMM'15]
- <u>They come from the prior knowledge</u>
  - Facial Age Estimation [Geng, Yin and Zhou, TPAMI'13]
- <u>They are learnt from the data</u>
  - Relative Labeling-Importance Aware Multi-label Learning
    [Li, Zhang and Geng, ICDM'15]

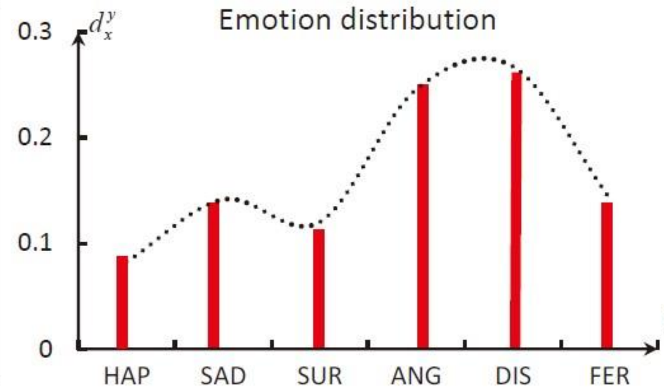
Come with Original Data

- Emotion Distribution via Facial Expressions

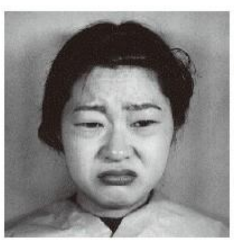
[Zhou, Xue and Geng, ACMMM’15]

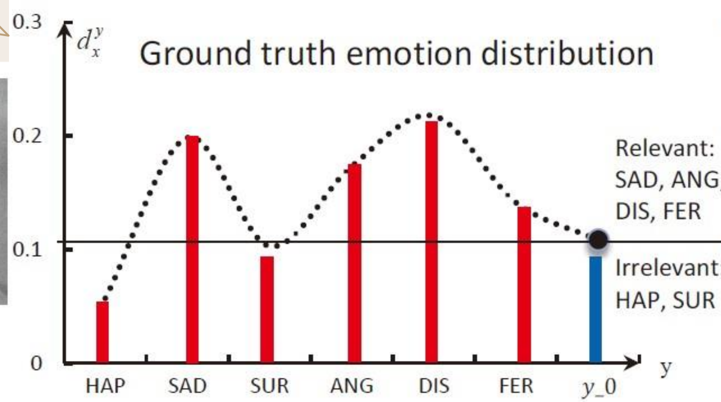
Emotion Distribution Recognition

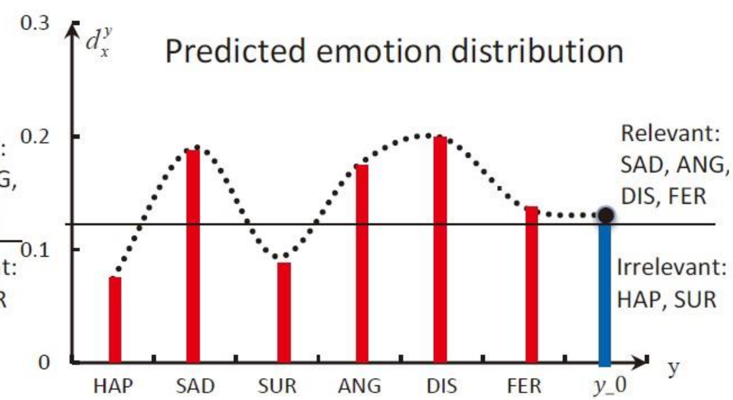
emotion	score	Multi-label
HAP	1.35	-1
SAD	2.32	-1
SUR	1.97	-1
ANG	4.03	1
DIS	4.39	1
FER	2.35	-1



Multi-Emotion Recognition



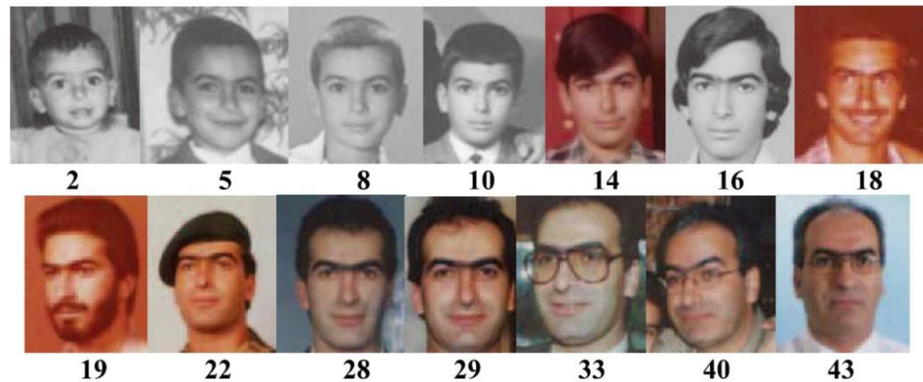




Come from Prior Knowledge

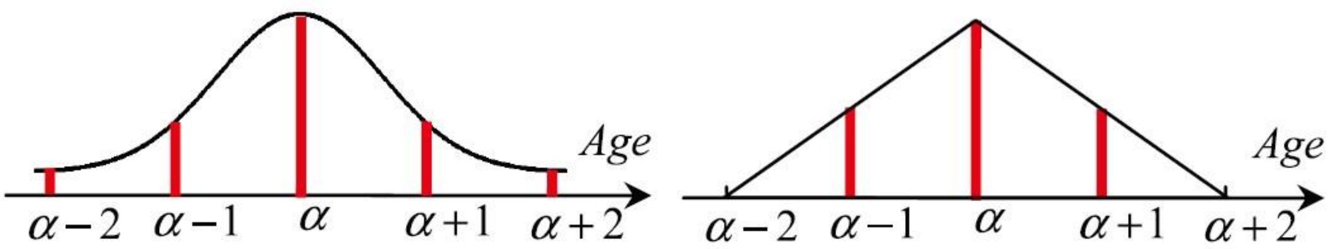
- Facial Age Estimation

[Geng, Yin, and Zhou, TPAMI'13]
[Geng, Smith-Miles, and Zhou, AAAI'10]



Prior Knowledge

- Aging is a slow and gradual progress
- The faces at close ages look quite similar



(a) (b)

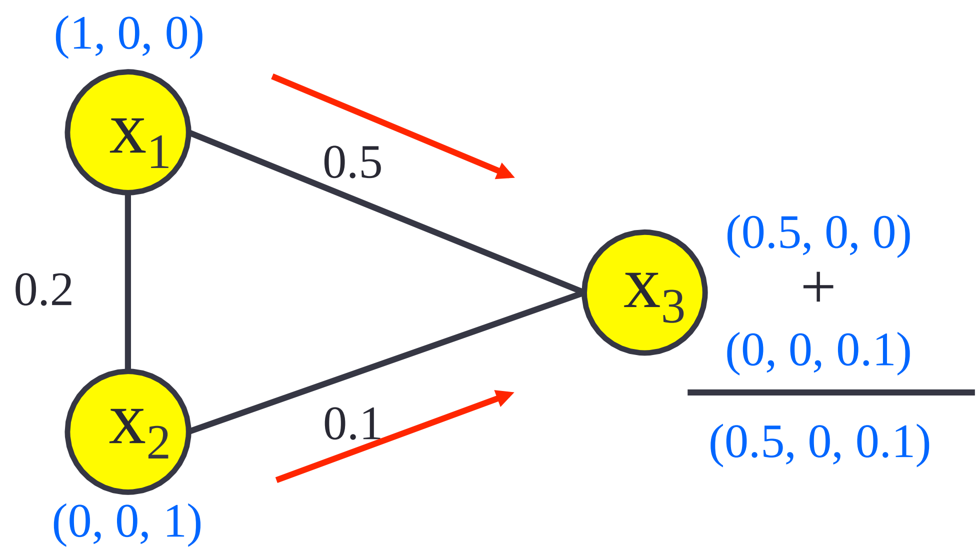
- Centered at the chronological age
- Highest at the chronological age and gradually decrease on both sides

Relative Labeling-Importance Aware Multi-label Learning

[Li, Zhang and Geng, ICDM’15]

- Implicit Relative Labeling-Importance

  Label Propagation on the Training Set



Thank you