嵌入式系统联谊会
www.esbf.org
十周年纪念
2008—2018

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

基于知识共享的机器学习算法研究及其应用

--- 迁移学习

庄福振

中国科学院计算技术研究所

2018 年 11月 24日

大数据以及应用

金融服务业：欺诈检测，用户画像等

能源与公共事业：智能电表分析，资产管理等

运输业，快递：物流优化，派送优化，缓解交通拥堵等

数字媒体：实时广告定位，属性分析等

智慧医疗：病例分析，疾病监测等

零售业：全渠道营销，实时促销等

通讯行业：客户资料货币化，网络流量分析与优化等

司法执法：多点监测，网络安全检测等

举例：大数据分类

Flickr Image

训练/分类

训练

共享知识

图片

人工标记

费时耗力

图片
有标记

ImageNet

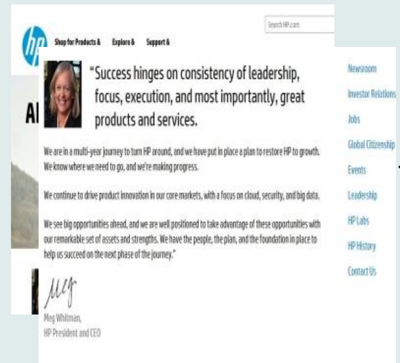
基于知识共享模型

迁移学习

迁移
知识

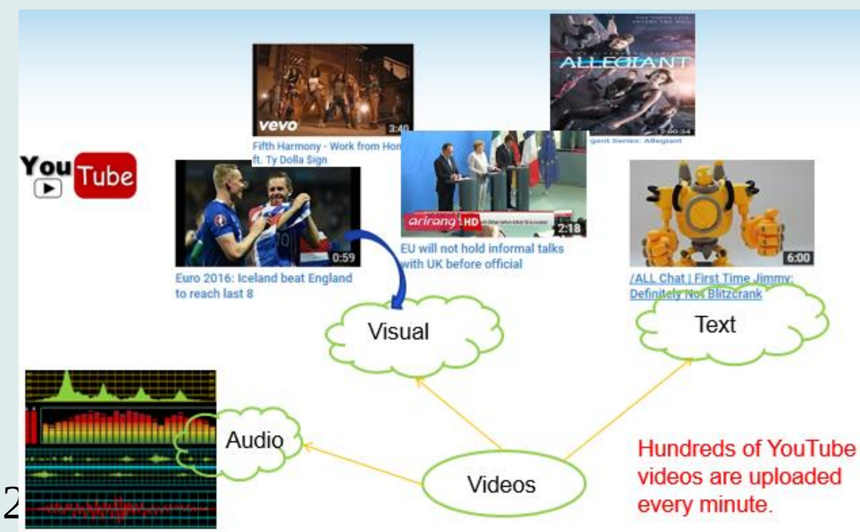
多任务学习

HP 新闻

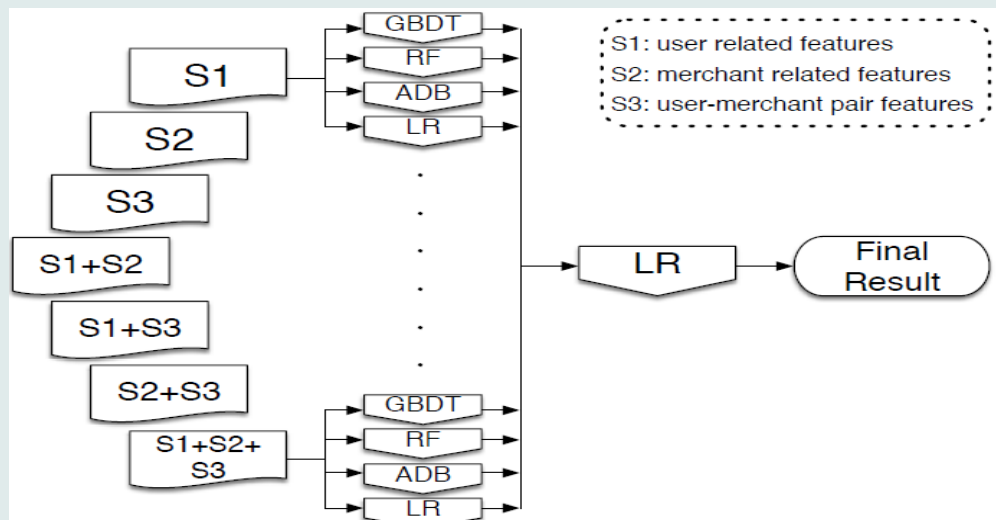


Lenovo 新闻

多视图学习



模型融合



基于知识共享模型区别与联系

INSTITUTE OF COMPUTING TECHNOLOGY

迁移学习 vs. 多任务学习 vs. 多视图学习 vs. 模型融合

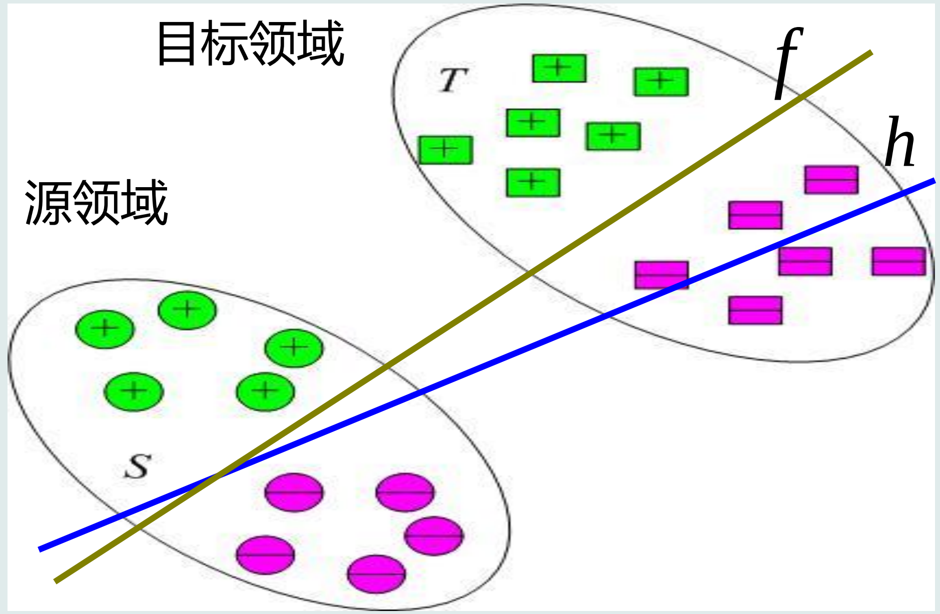
- 迁移学习旨在通过共享知识提升目标领域上的性能，而多任务学习旨在通过共享知识提升所有任务上的总体性能
- 多视图学习旨在充分利用数据多个视图信息，在有限标记数据情况下，提升目标数据上的性能，迁移学习和多任务学习都可以用于多视图学习
- 模型融合是通过共享多个模型的知识，提升目标数据上的性能，这些模型可以来自多个领域，也可以通过采样来自单个领域；模型融合技术可以用于迁移学习、多任务学习以及多视图学习

INSTITUTE OF COMPUTING TECHNOLOGY

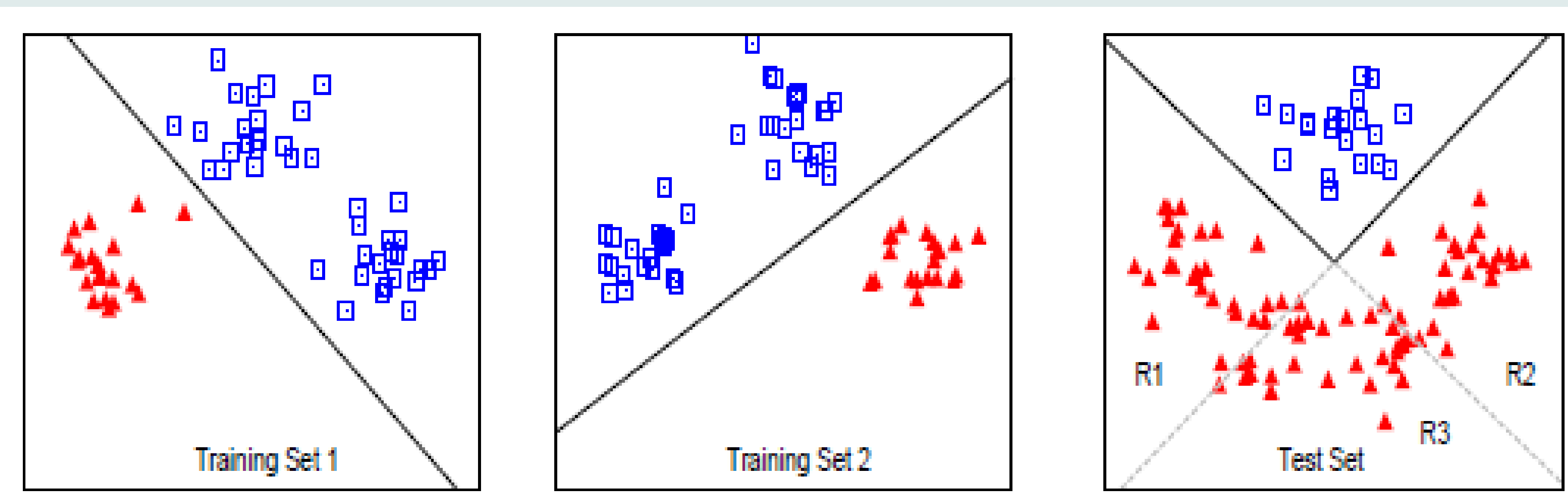
迁移学习

分布不一致性

Example: 1



Example: 2



传统监督机器学习(1/2)

Training Data

Occ	Palm Lines	DragonStar	Fortune?
Prof	long	T	good
Lawyer	short	F	bad
PhD Stu	broken	T	good
Doc	long	F	bad

What if…

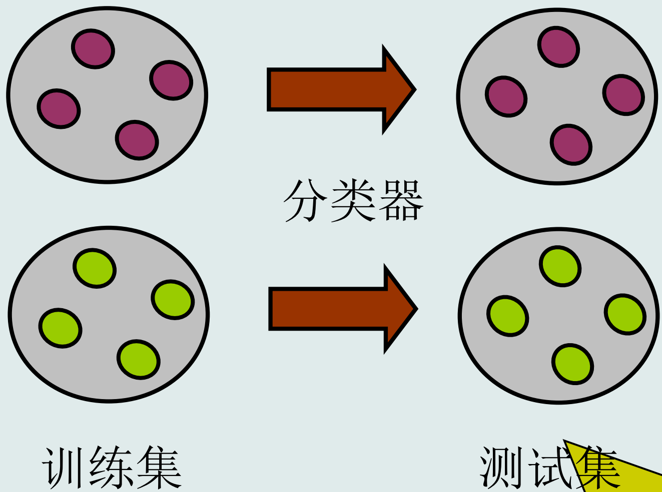
Classifier

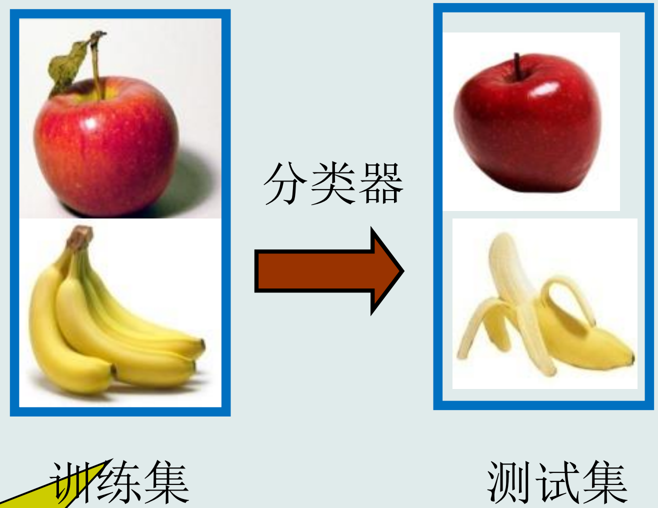
good!

[from Prof. Qiang Yang]

传统监督机器学习(2/2)

- 传统监督学习



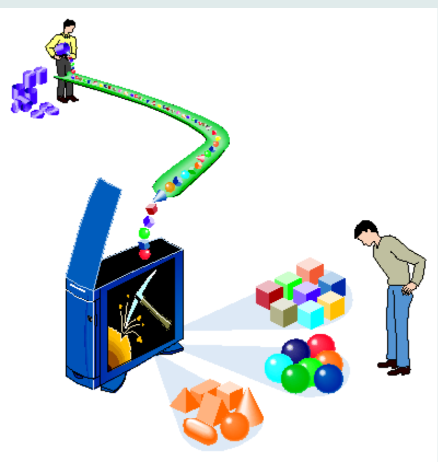


在实际应用中
通常不能满足！

两个基本假设

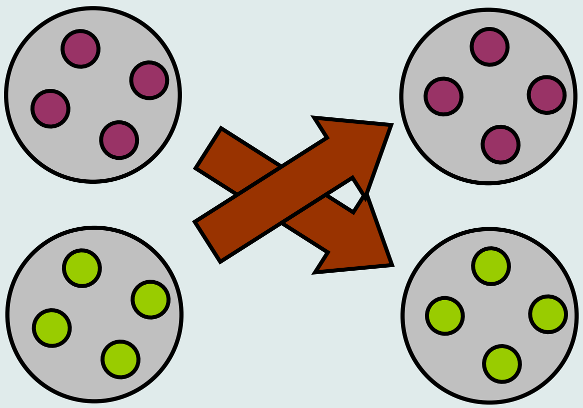
同源、独立同分布

标注足够多的训练样本

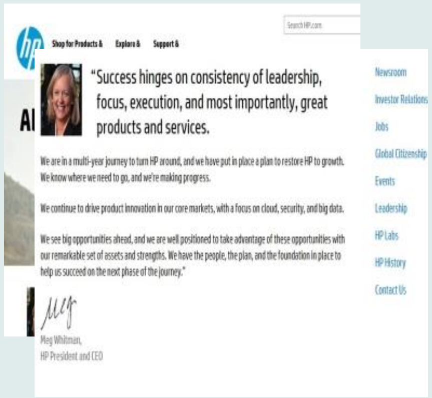


迁移学习

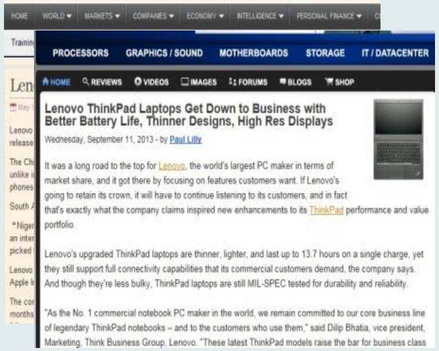
- 实际应用学习场景



HP 新闻

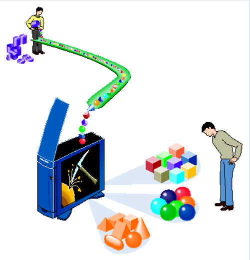


Lenovo 新闻



不同源、分布不一致

人工标记训练样本，费时耗力



迁移学习

– 运用已有的知识对不同但相关领域问题进行求解的一种新的机器学习方法

– 放宽了传统机器学习的两个基本假设

迁移学习场景(1/4)

- 迁移学习场景无处不在

迁移知识

迁移知识

图像分类

- Chess → checkers
- C++ → Java
- Physics → Computer Science

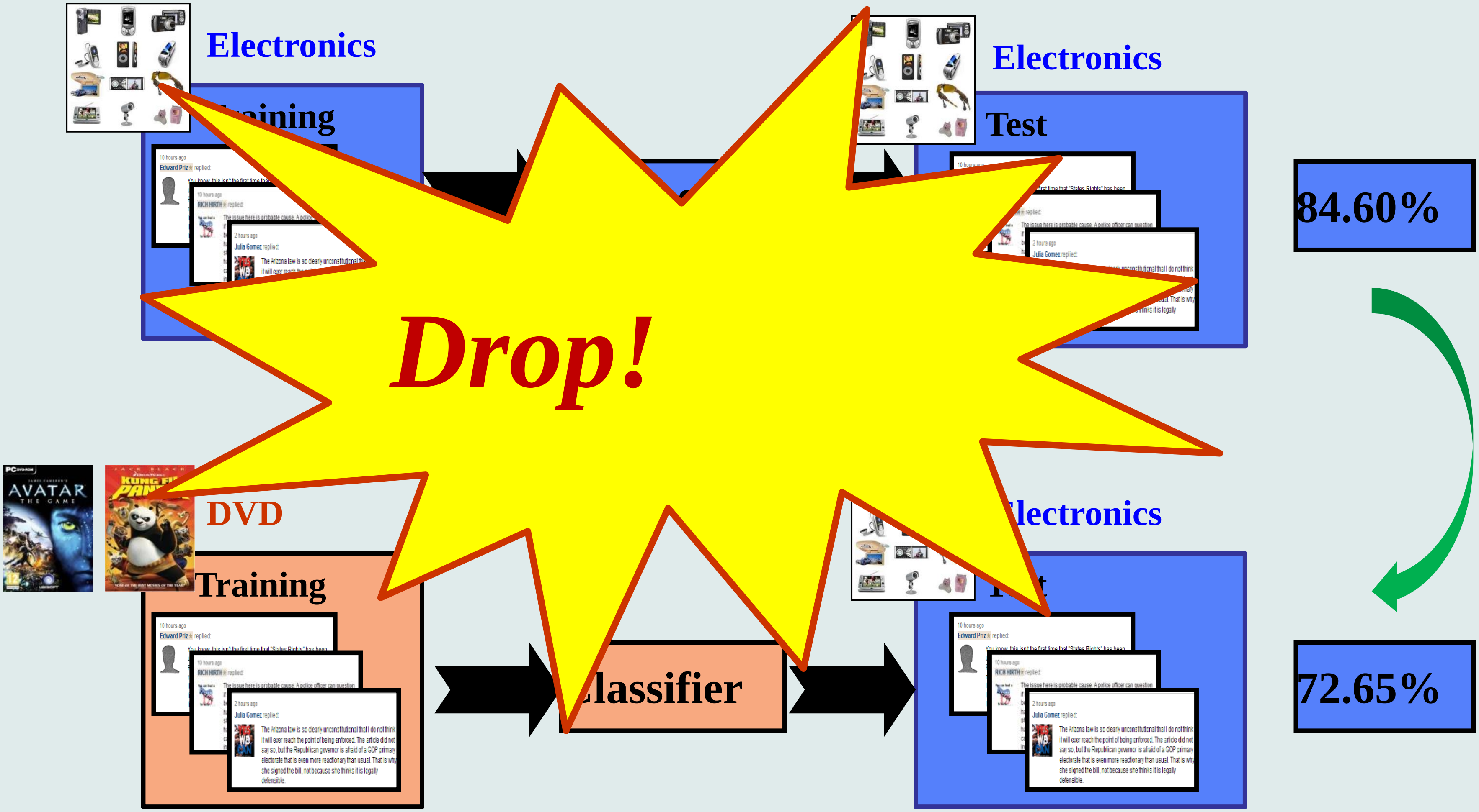
迁移学习场景(2/4)

- 异构特征空间

	Training: Text	Future: Images
Apples	The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family Rosaceae ...	
Bananas	Banana is the common name for a type of fruit and also the herbaceous plants of the genus Musa which produce this commonly eaten fruit ...	

Xin Jin, Fuzhen Zhuang, Sinno Jialin Pan, Changying Du, Ping Luo, Qing He: Heterogeneous Multi-task Semantic Feature Learning for Classification. CIKM 2015 : 1847-1850.

迁移学习场景(3/4)



Electronics
Training
Electronics
Test
84.60%
Drop!
DVD
Training
Electronics
Classifier
72.65%

迁移学习场景(4/4)

Transfer Learning Algorithms

- Concept Learning for Transfer Learning
  - Concept Learning based on Non-negative Matrix Tri-factorization for Transfer Learning
  - Concept Learning based on Probabilistic Latent Semantic Analysis for Transfer Learning
- Transfer Learning using Auto-encoders
  - Transfer Learning from Multiple Sources with Autoencoder Regularization
  - Supervised Representation Learning: Transfer Learning with Deep Auto-encoders
- Robust Transfer Learning
  - Ensemble of Anchor Adapters for Transfer Learning
- Application in Recommender Systems

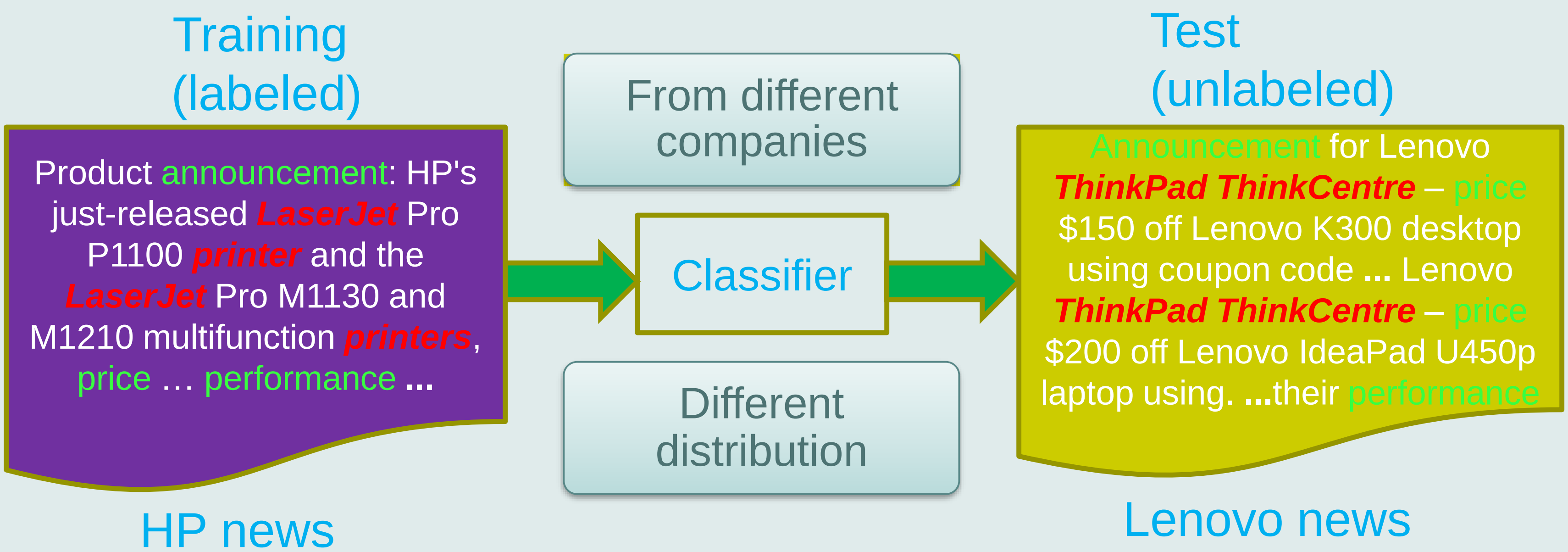
Concept Learning based on Non-negative Matrix Tri-factorization for Transfer Learning

Introduction

- Many traditional learning techniques work well only under the assumption: Training and test data follow the same distribution

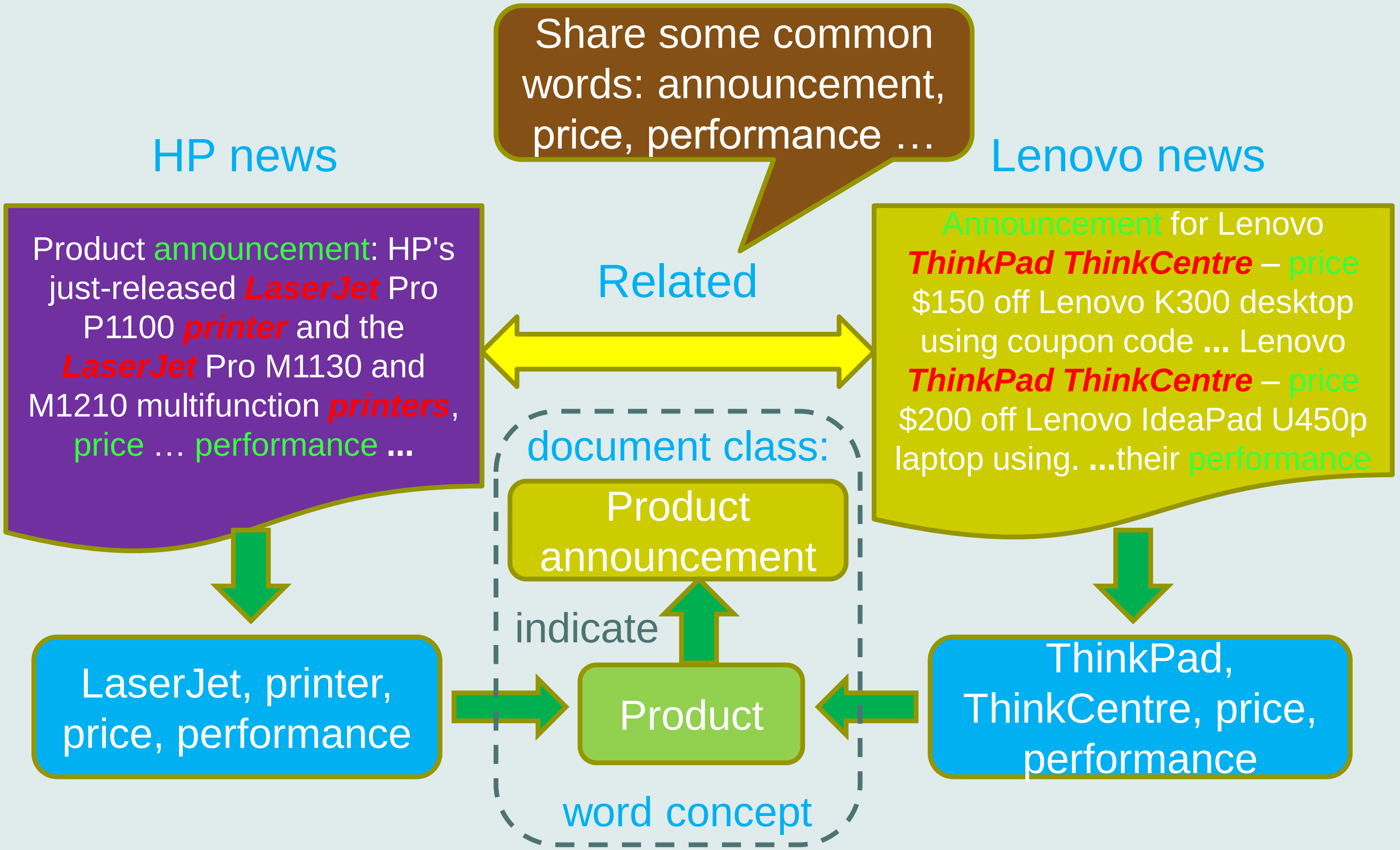
Fail !

Enterprise News Classification: including the classes "Product Announcement", "Business scandal", "Acquisition", … …



Motivation (1/3)

- Example Analysis



Motivation (2/3)

- Example Analysis:

The words expressing the same word concept are domain-dependent

HP	LaserJet, printer, price, performance et al.
Lenovo	Thinkpad, Thinkcentre, price, performance et al.

The association between word concepts and document classes is domain-independent

Motivation (3/3)

INSTITUTE OF COMPUTING TECHNOLOGY

- Further observations:
  - Different domains may use same key words to express the same concept (denoted as *identical concept*)
  - Different domains may also use different key words to express the same concept (denoted as *alike concept*)
  - Different domains may also have their own distinct concepts (denoted as *distinct concept*)
- The identical and alike concepts are used as the shared concepts for knowledge transfer
- We try to model these three kinds of concepts simultaneously for transfer learning text classification

Preliminary Knowledge

- Basic formula of matrix tri-factorization:

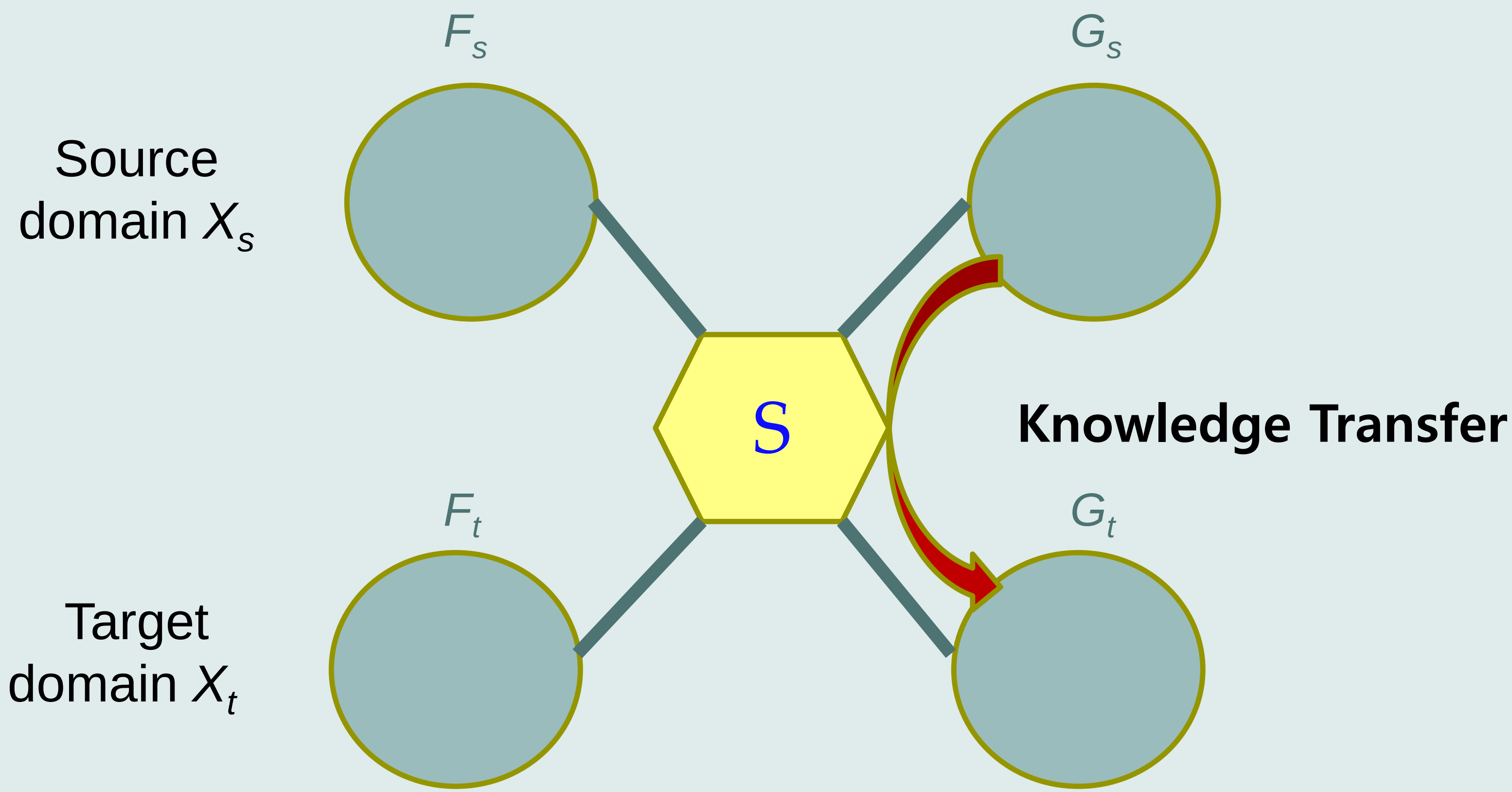
$$X_{m\times n} = F_{m\times k_1} S_{k_1\times k_2} G^T_{n\times k_2}$$

where the input X is the word-document co-occurrence matrix

F	denotes concept information, may vary in different domains
S	indeed is the association between word concepts and document classes, may retain stable cross domains
G	denotes the document classification information

Previous method - MTrick in SDM 2010 (1/2)

- Sketch map of MTrick



- Considering the alike concepts

MTrick (2/2)

- Optimization problem for MTrick

$$
\begin{aligned}
&\min_{F_s,G_s,S,F_t,G_t} ||X_s - F_s S G_s^T||^2 + \frac{\alpha}{n_s} \cdot ||G_s - G_0||^2 \\
&\quad + \beta \cdot ||X_t - F_t S G_t^T||^2, \\
&\text{s.t.} \quad \sum_{j=1}^{k_1} F_{s(ij)} = 1, \sum_{j=1}^{k_2} G_{s(ij)} = 1, \\
&\qquad \sum_{j=1}^{k_1} F_{t(ij)} = 1, \sum_{j=1}^{k_2} G_{t(ij)} = 1,
\end{aligned}
\tag{2.10}
$$

the association *S* is shared as bridge to transfer knowledge

G_0 is the supervision information

- Dual Transfer Learning (Long et al., SDM 2012), considering identical and alike concepts

Triplex Transfer Learning (TriTL) (1/5)

INSTITUTE OF COMPUTING TECHNOLOGY

- Further divide the word concepts into three kinds:

$$
\begin{aligned}
X_{m\times n} &= F_{m\times k} S_{k\times c} G_{n\times c}^{T} \\
&= [F_{m\times k_1}^{1}, F_{m\times k_2}^{2}, F_{m\times k_3}^{3}] \begin{bmatrix} S_{k_1\times c}^{1} \\ S_{k_2\times c}^{2} \\ S_{k_3\times c}^{3} \end{bmatrix} G_{n\times c}^{\top}
\end{aligned}
$$

F^1, identical concepts; F^2, alike concepts; F^3, distinct concepts

- Input: s source domain $X_r (1 \le r \le s)$ with label information, t target domain X_r $(s+1 \le r \le s+t)$
- We propose Triplex Transfer Learning framework based on matrix tri-factorization (TriTL for short)

TriTL (2/5)

- Optimization Problem

$$\min_{F_r,S_r,G_r} \mathcal{L} = \sum_{r=1}^{s+t} ||X_r - F_r S_r G_r^\top||^2$$

$$= \sum_{r=1}^{s+t} ||X_r - [F^1, {F^2}_r, {F^3}_r] \begin{bmatrix} S^1 \\ S^2 \\ {S^3}_r \end{bmatrix} G_r^\top||^2$$

$$s.t. \sum_{i=1}^{m} {F^1}_{[i,j]} = 1, \sum_{i=1}^{m} {F^2}_{r\,[i,j]} = 1,$$

$$\sum_{i=1}^{m} {F^3}_{r\,[i,j]} = 1, \sum_{j=1}^{c} G_{r\,[i,j]} = 1.$$

F^1, S^1 and S^2 are shared as the bridge for knowledge transfer across domains

The supervision information is integrated by G_r ($1 \le r \le s$) in source domains

TriTL (3/5)

INSTITUTE OF COMPUTING TECHNOLOGY

- We develop an alternatively iterative algorithm to derive the solution and theoretically analyze its convergence

$$S^1{}_{[i,j]} \leftarrow S^1{}_{[i,j]} \cdot \sqrt{\frac{[\sum_{r=1}^{s+t} F^{1\top} X_r G_r]_{[i,j]}}{[\sum_{r=1}^{s+t} (F^{1\top} A_r G_r + F^{1\top} B_r G_r + F^{1\top} C_r G_r)]_{[i,j]}}}, \tag{22}$$

$$S^2{}_{[i,j]} \leftarrow S^2{}_{[i,j]} \cdot \sqrt{\frac{[\sum_{r=1}^{s+t} F^{2\top}_r X_r G_r]_{[i,j]}}{[\sum_{r=1}^{s+t} (F^{2\top}_r B_r G_r + F^{2\top}_r A_r G_r + F^{2\top}_r C_r G_r))]_{[i,j]}}}, \tag{23}$$

$$S^3{}_{r[i,j]} = S^3{}_{r[i,j]} \cdot \sqrt{\frac{[F^{3\top}_r X_r G_r]_{[i,j]}}{[F^{3\top}_r C_r G_r + F^{3\top}_r A_r G_r + F^{3\top}_r B_r G_r]_{[i,j]}}}, \tag{24}$$

$$F^1{}_{[i,j]} \leftarrow F^1{}_{[i,j]} \cdot \sqrt{\frac{[\sum_{r=1}^{s+t} X_r G_r S^{1\top}]_{[i,j]}}{[\sum_{r=1}^{s+t} (A_r G_r S^{1\top} + B_r G_r S^{1\top} + C_r G_r S^{1\top})]_{[i,j]}}}, \tag{19}$$

$$F^2{}_{r[i,j]} \leftarrow F^2{}_{r[i,j]} \cdot \sqrt{\frac{[X_r G_r S^{2\top}]_{[i,j]}}{[B_r G_r S^{2\top} + A_r G_r S^{2\top} + C_r G_r S^{2\top}]_{[i,j]}}}, \tag{20}$$

$$F^3{}_{r[i,j]} \leftarrow F^3{}_{r[i,j]} \cdot \sqrt{\frac{[X_r G_r S^{3\top}_r]_{[i,j]}}{[C_r G_r S^{3\top}_r + A_r G_r S^{3\top}_r + B_r G_r S^{3\top}_r]_{[i,j]}}}, \tag{21}$$

$$G_{r[i,j]} \leftarrow G_{r[i,j]} \cdot \sqrt{\frac{[X_r^\top F_r S_r]_{[i,j]}}{[G_r S_r^\top F_r^\top F_r S_r]_{[i,j]}}}. \tag{25}$$

TriTL (4/5)

- Analysis of Algorithm Convergence

  - According to the methodology of convergence analysis in the two works [Lee et al., NIPS'01] and [Ding et al., KDD'06], the following theorem holds.

Theorem (Convergence): *After each round of calculating the iterative formulas, the objective function in the optimization problem will converge monotonically*.

TriTL (5/5)

- Classification on target domains
  - When $1 \leq r \leq s$, G_r contains the label information, so we remain it unchanged during the iterations

    when x_i belongs to class j, then $G_{r(i,j)}=1$, else $G_{r(i,j)}=0$
  - After the iteration, we obtain the output G_r ($s+1 \leq r \leq s+t$), then we can perform classification according to G_r

$$\arg\max_j G_{r_{(i,j)}}.$$

Data Preparation (1/3)

- Sentiment Classification, four domains: books, dvd, electronics, kitchen
  - Randomly select two domains as sources, and the rest as targets, then 6 problems can be constructed
- 20Newsgroups

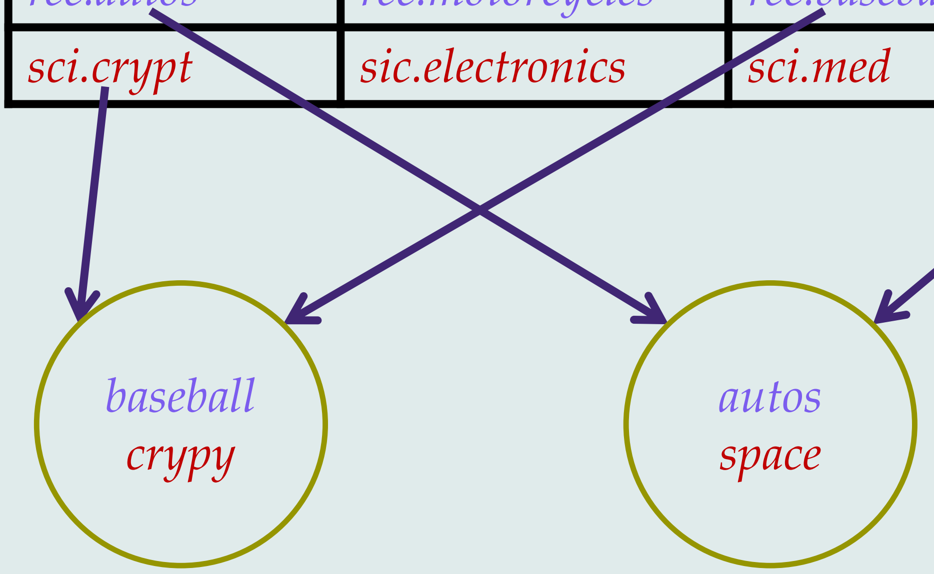
rec	*rec.autos*	*rec.motorcycles*	*rec.baseball*	*rec.hockey*
sci	*sci.crypt*	*sic.electronics*	*sci.med*	*sci.space*
comp	*comp.graphics*	*comp.sys.ibm.pc.hardware*	*comp.sys.mac.hardware*	*comp.windows.x*
talk	*talk.politics.misc*	*talk.politics.guns*	*talk.politics.mideast*	*talk.religion.misc*

  - Four top categories, each top category contains four sub-categories

Data Preparation (2/3)

- Construct classification tasks (Traditional TL)

rec **+**	*rec.autos*	*rec.motorcycles*	*rec.baseball*	*rec.hockey*
sci **-**	*sci.crypt*	*sic.electronics*	*sci.med*	*sci.space*



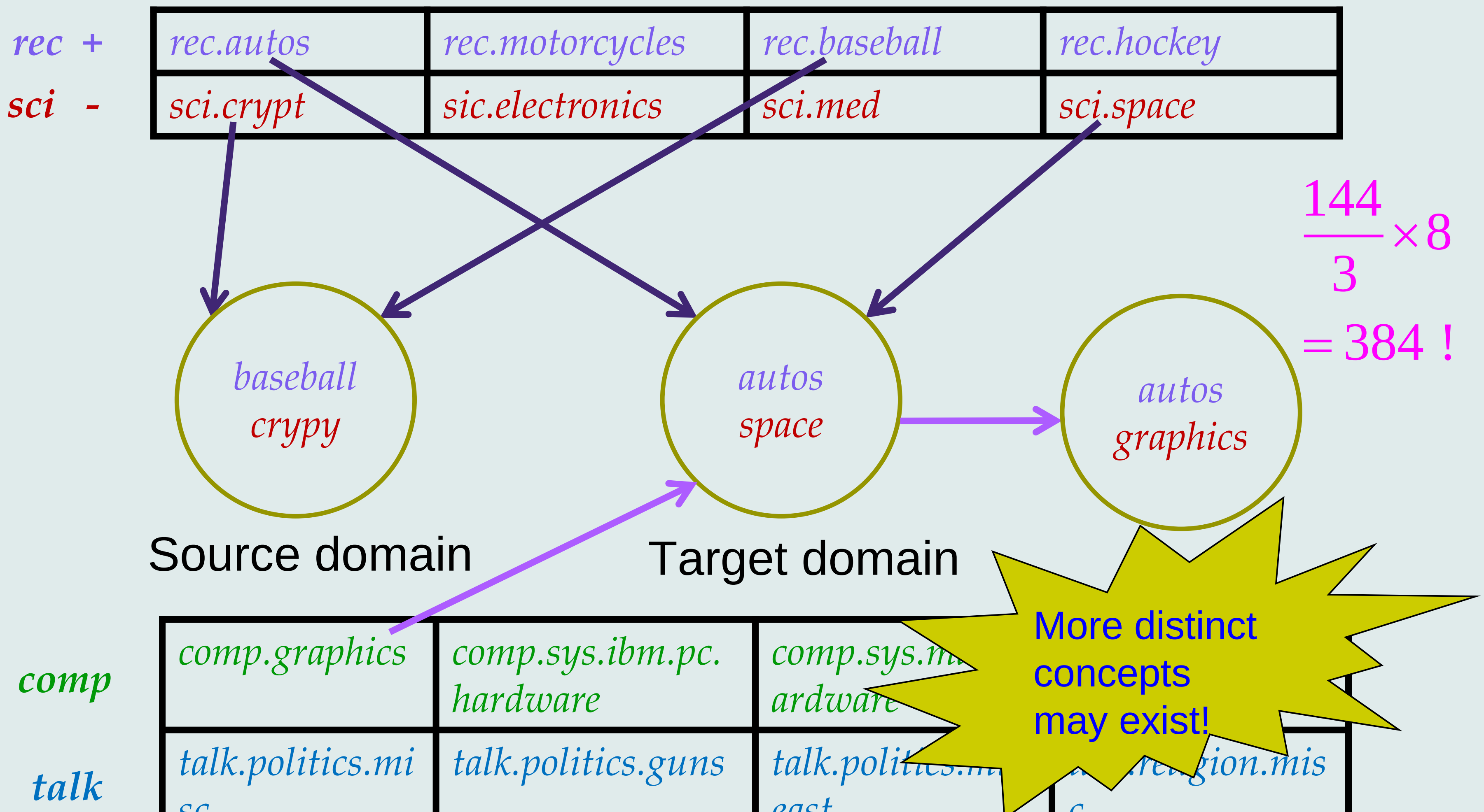
Source domain

Target domain

- For the classification problem with one source domain and one target domain, we can construct 144 ($P_4^2 \cdot P_4^2$) problems

Data Preparation (3/3)

- Construct new transfer learning problems



Compared Algorithms

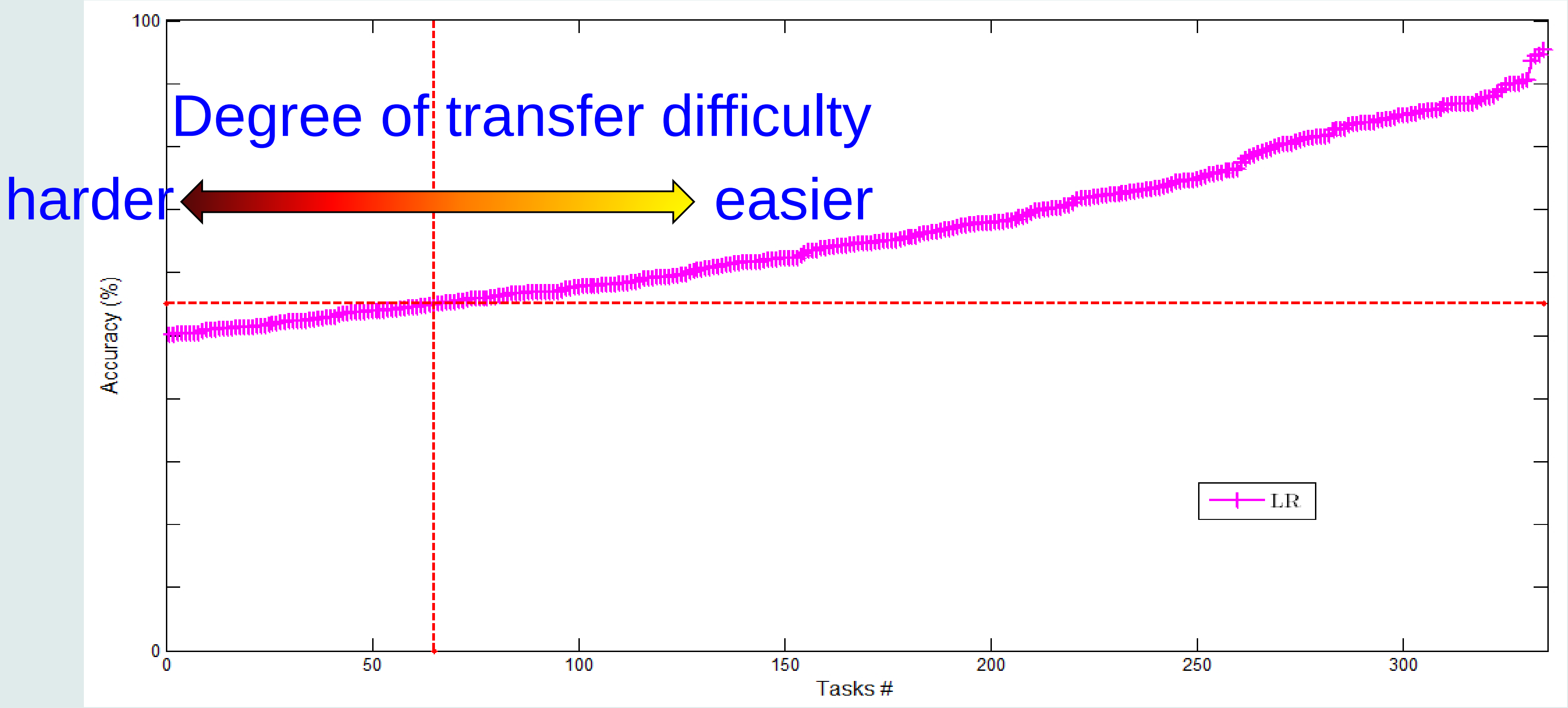
- Traditional learning Algorithms
  - Supervised Learning:
    - Logistic Regression (LR) [David et al., 00]
    - Support Vector Machine (SVM) [Joachims, ICML'99]
  - Semi-supervised Learning:
    - TSVM [Joachims, ICML'99]
- Transfer learning Methods:
  CoCC [Dai et al., KDD'07], DTL [Long et al., SDM'12]

	Alike	Identical	Distinct
CoCC [5]		√	
MTrick [9]	√		
DKT [11]	√		
DTL [12]	√	√	
TriTL	√	√	√

- Classification accuracy is used as the evaluation measure

Experimental Results (1/3)

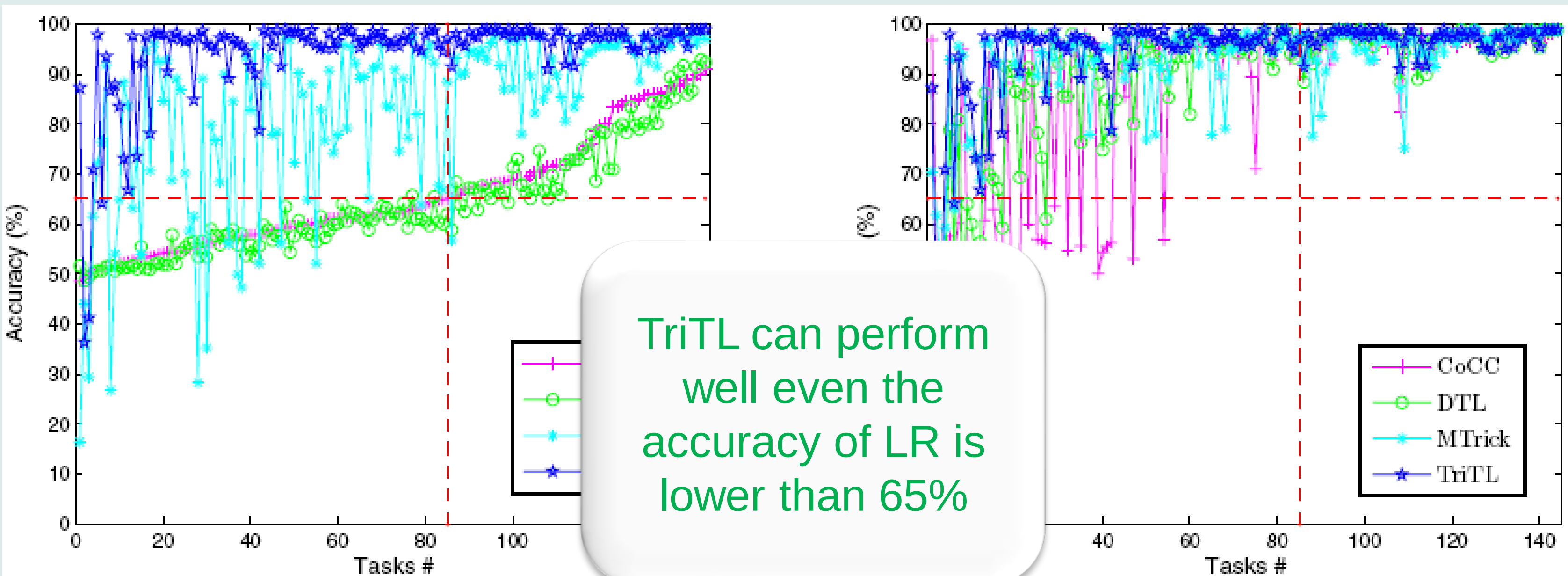
- Sort the problems with the accuracy of LR



- Generally, the lower of accuracy of LR can indicate the harder to transfer, while the higher ones indicate the easier to transfer

Experimental Results (2/3)

- Comparisons among TriTL, DTL, MTrick, CoCC, TSVM, SVM and LR on data set *rec vs. sci* (144 problems)



TriTL can perform well even the accuracy of LR is lower than 65%

Data Set		LR	SVM	TSVM	CoCC	DTL	MTrick	TriTL
rec vs. sci	*Lower*	57.41	56.78	75.73	79.69	84.29	90.44	**92.23**
	Higher	75.77	73.48	91.66	96.18	96.56	95.53	**97.19**
	Total	65.57	64.20	82.81	87.02	89.75	92.70	**94.43**

Experimental Results (3/3)

- Results on new transfer learning problems, we only select the problems, whose accuracies of LR are between (50%, 55%] (Only slightly better than random classification, thus they might be much more difficult).
- We obtain 65 problems

Table 6: Average Performances (%) on 65 Much Harder Transfer Learning Tasks

LR	SVM	TSVM	CoCC	DTL	MTrick	TriTL
52.45	51.81	74.32	69.66	75.34	78.45	**80.93**

- TriTL also outperforms all the baselines

Experimental Results (4/5)

Source domain: S (*rec.autos*, *sci.space*)，

Target domain: T(*rec.sport.hockey*, *talk.politics.mideast*)

Alike concept	Topic 1	S	**cars**,drew,brakes,centerline,tek,brake,**car**,**speed**,uokmax,com,bird,ford,**clutch**,virginia,convertible,wv,sho,uoknor,taurus,callison
		T	show,coverage,andrew,msu,eos,baltimore,**play**,ca,tom,pat,**ice**,**game**,caps,francis,**baseball**,overtime,night,stats,jagr,**espn**
	Topic 2	S	police,rocks,chintan,amin,**road**,**vw**,**gas**,purdue,**gt**,cactus,lehigh,**driving**,**accident**,**mph**,wagon,**auto**,uiuc,insurance,**car**,**cars**
		T	sweden,**sport**,emotional,ca,blues,friedman,skins,next,prism,kevin,jersey,mask,gatech,gtd,**goalie**,hrivnak,capitals,**fan**,mike,go

Distinct concept	S	Topic 1	**planet**,observations,teflon,tommy,cacs,srl,baalke,**mars**,gov,higgins,**jpl**,**nasa**,**temperature**,**planets**,kelvin,dseg,ti,smiley,**jupiter**,**hst**
		Topic 2	glen,oz,kelvin,**planetary**,mercury,**saturn**,**nasa**,radiation,ti,phil,mccall,gov,fraering,sun,**jpl**,**mars**,ron,**jupiter**,fnal,baalke
	T	Topic 1	**israelis**,ncsu,mcrcim,igc,sexual,shostack,brad,marc,quote,davidsson,**istanbul**,dog,cute,idf,favors,das,bu,**gaza**,pro,cpr
		Topic 2	hernlem,hasan,**isreal**,**civilians**,**istanbul**,**hamas**,mcgill,**lebanese**,elias,diesel, wagon, nissan, mileage, byte, saturn, toyota, si, cars, car, db

Conclusions

- Explicitly define three kinds of word concepts, i.e., identical concept, alike concept and distinct concept
- Propose a general transfer learning framework based on nonnegative matrix tri-factorization, which simultaneously model the three kinds of concepts (TriTL)
- Extensive experiments show the effectiveness of the proposed approach, especially when the distinct concepts may exist

Transfer Learning from Multiple Sources with Autoencoder Regularization

Motivation(1/2)

Electronics	Video Games
Compact; easy to operate; very **good** picture, **excited** about the quality; looks **sharp**!	A very **good** game! It is **action packed** and full of **excitement**. I am very much **hooked** on this game.

- **Transfer learning based on original feature space may fail to achieve high performance on Target domain data**
- **Due to the success of representation learning by deep learning. We consider the autoencoder technique to collaboratively find a new representation of both source and target domain data**

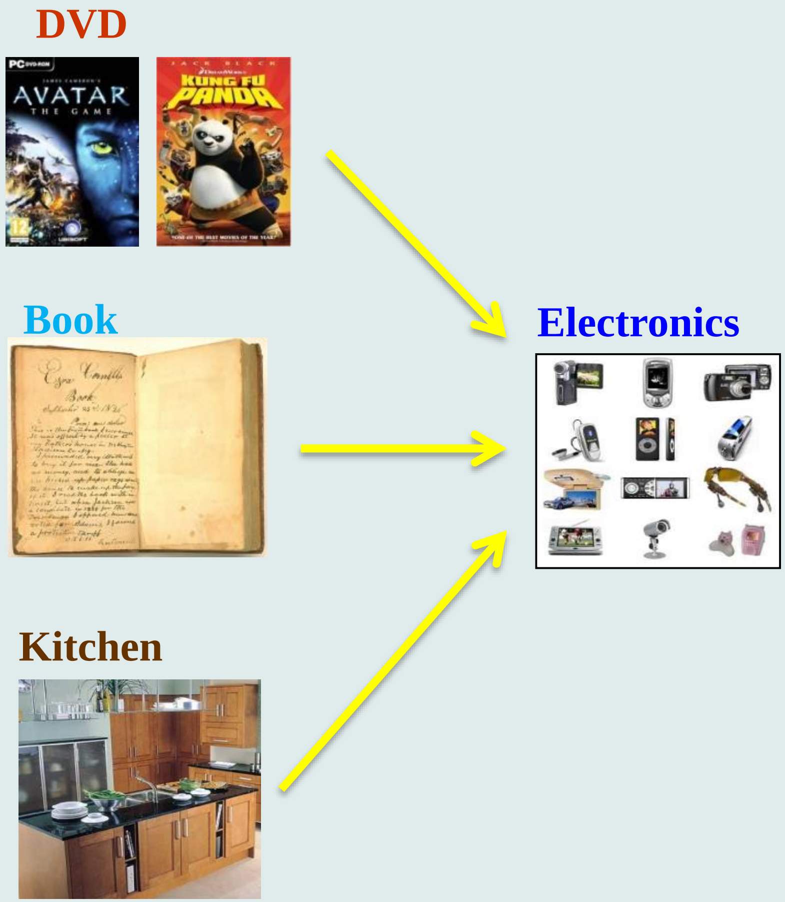
Motivation(2/2)

DVD

Electronics

- **Previous methods often transfer from one source domain to one target domain**

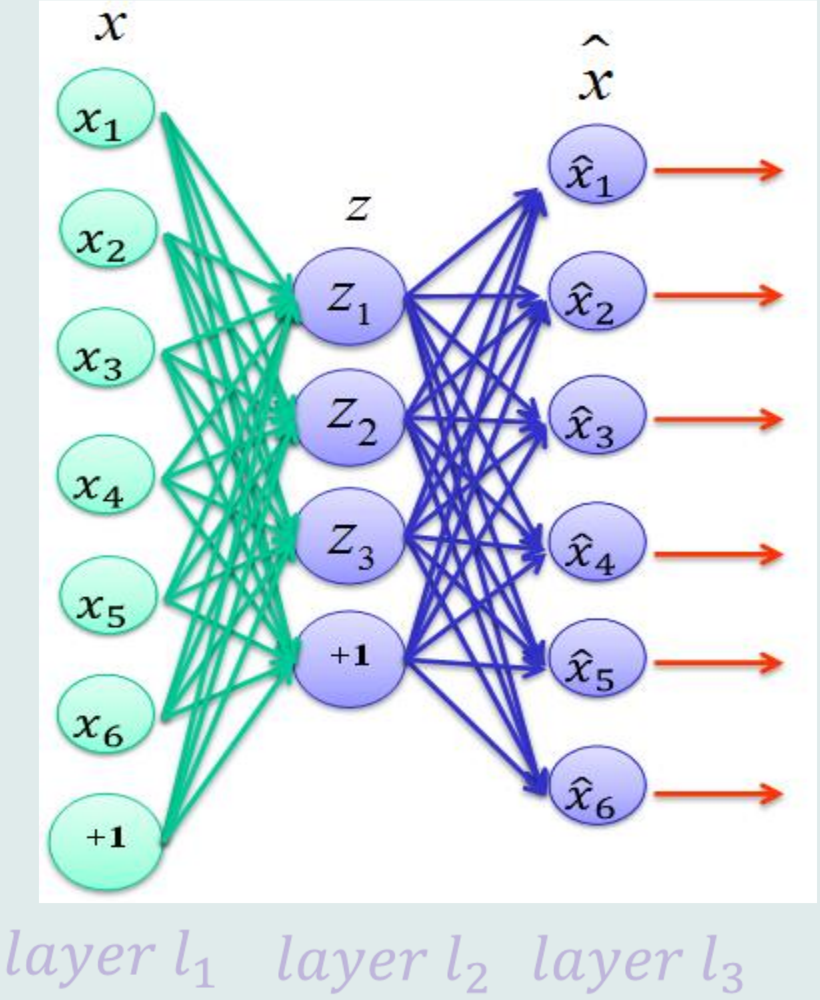
Motivation(2/2)



- Previous methods often transfer from one source domain to one target domain
- We consider the consensus regularized framework for learning from multiple source domains

We propose a transfer learning framework of consensus regularization autoencoders to learn from multiple sources

Autoencoder Neural Network



$$z = h(\mathbf{W}x + b),$$

$$\hat{x} = g(\mathbf{W}'z + b'),$$

- **Minimizing the reconstruction error to derive the solution:**

$$\min_{\mathbf{W}, b, \mathbf{W}', b'} = \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2.$$

where *h, g* are nonlinear activation function, e.g., Sigmoid function, for encoding and decoding

Consensus Measure-(1/3)

- **Example:** three-class classification problem, three classifiers predict instances

Source 1: $D_1 \longrightarrow f_1$

Source 2: $D_2 \longrightarrow f_2$

Source 3: $D_3 \longrightarrow f_3$

Constraint

	f_1	f_2	f_3
x_1	1	1	1
x_2	3	3	3
x_3	2	2	2
x_4	2	3	1
x_5	3	1	3
x_6	1	2	3

Consensus Measure-(2/3)

INSTITUTE OF COMPUTING TECHNOLOGY

- **Example:** three-class classification problem, prediction on instance x

$$\left.\begin{aligned} p_1 &= (1, 0, 0) \\ p_2 &= (1, 0, 0) \\ p_3 &= (1, 0, 0) \end{aligned}\right\} \xrightarrow{Average} (1,0,0) \xrightarrow{-Entropy} C_e = -E(1,0,0) = 0$$

Minimal entropy, Maximal Consensus

$$\left.\begin{aligned} p_1 &= (1, 0, 0) \\ p_2 &= (0, 1, 0) \\ p_3 &= (0, 0, 1) \end{aligned}\right\} \xrightarrow{Average} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \xrightarrow{-Entropy} C_e = -E\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

Maximal entropy, Minimal Consensus

- **Entropy based Consensus Measure (Luo et al., CIKM'08)**

$$\psi(x; \{\theta_i\}_{i=1}^{r}) = -\sum_{c \in C} \bar{p}(c) \log \frac{1}{\bar{p}(c)}, \qquad \bar{p}(c) = \frac{1}{r} \sum_{i=1}^{r} p_i(c),$$

θ_i is the parameter vector of classifier i, C is the class label set

Consensus Measure-(3/3)

INSTITUTE OF COMPUTING TECHNOLOGY

- For simplicity, the consensus measure for binary classification can be rewritten as

$$\psi(x; \{\theta_i\}_{i=1}^{r}) = (\bar{p} - (1 - \bar{p}))^2 = (2\bar{p} - 1)^2.$$

since their effects on making the prediction consensus are similar.

- In this work, we impose the consensus regularization to autoencoders, and try to improve the learning performance from multiple source domains

Some Notations

INSTITUTE OF COMPUTING TECHNOLOGY

- **Source domains**

  Given r source domains: $\mathcal{D}_S^{(1)}, \cdots, \mathcal{D}_S^{(r)}$, i.e.,

  $\mathcal{D}_S^{(j)} = \left\{x_{S_i}^{(j)}, y_{S_i}^{(j)}\right\}_{i=1}^{n_j}$, $y_{S_i}^{(j)} \in \{-1, 1\}$.

  The first corresponding data matrix is $X_S^{(1)}$

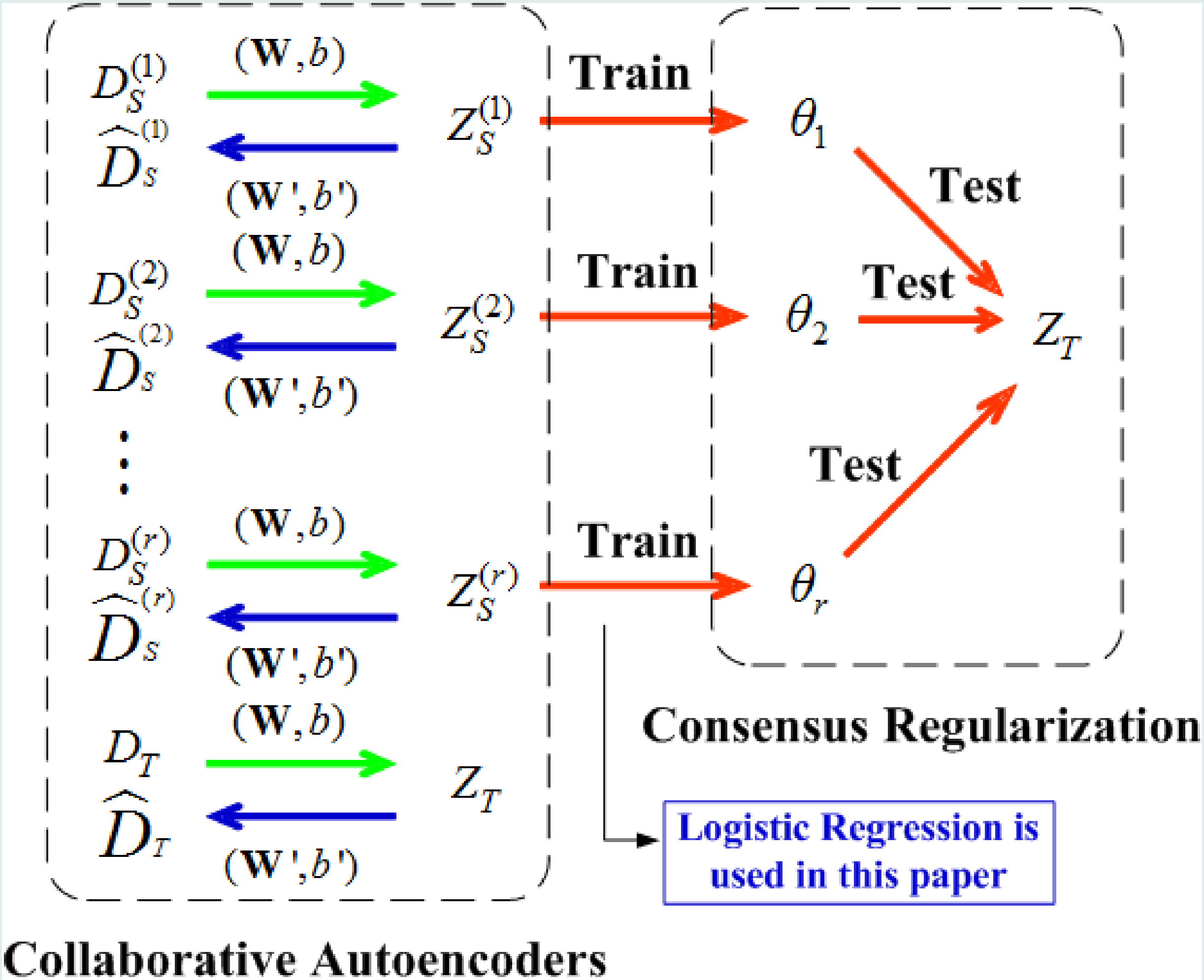
- **Target domain**

  $\mathcal{D}_T = \{x_{T_i}, y_{T_i}\}_{i=1}^{n}$

  The corresponding data matrix is X_T

- **The goal is to train a classifier f to make precise predictions on $\mathcal{D}_T$.**

Framework of CRA



- The data from all source and target domains share the same encoding and decoding weights
- The classifiers trained from the new representation are regularized to predict the same results on target domain data

Optimization Problem of CRA

INSTITUTE OF COMPUTING TECHNOLOGY

➢ The optimization problem:

$$\min_{\Theta,\Theta',\{\theta_j\}} \mathcal{J} = \epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) + \gamma\Omega(\Theta, \Theta')$$
$$+\alpha\ell(z_S, y_S; \{\theta_j\}) - \beta\psi(z_T; \{\theta_j\}),$$

Reconstruction Error

$$\epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) = \sum_{j=1}^{r}\sum_{i=1}^{n_j} \|x_{S_i} - \hat{x}_{S_i}\|^2 + \sum_{i=1}^{n} \|x_{T_i} - \hat{x}_{T_i}\|^2$$

Optimization Problem of CRA

INSTITUTE OF COMPUTING TECHNOLOGY

- **The optimization problem:**

$$\min_{\Theta,\Theta',\{\theta_j\}} \mathcal{J} = \epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) + \gamma\Omega(\Theta, \Theta') + \alpha\ell(z_S, y_S; \{\theta_j\}) - \beta\psi(z_T; \{\theta_j\}),$$

Consensus Regularization

$$\psi(z_T; \{\theta_j\}) = \sum_{i=1}^{n} \left\| 2\frac{\sum_{j=1}^{r} \sigma(\theta_j^\top z_{T_i})}{r} - 1 \right\|^2$$

Optimization Problem of CRA

➢ The optimization problem:

$$\min_{\Theta,\Theta',\{\theta_j\}} \mathcal{J} = \epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) + \gamma\Omega(\Theta, \Theta') + \alpha\ell(z_S, y_S; \{\theta_j\}) - \beta\psi(z_T; \{\theta_j\}),$$

The total loss of source classifiers over the corresponding source domain data with the hidden representation

$$\ell(z_S, y_S; \{\theta_j\}) = \sum_{j=1}^{r}\left(-\sum_{i=1}^{n_j} \log\sigma(y_{S_i}^{(j)}\theta_j^\top z_{S_i}^{(j)}) + \lambda\theta_j^\top\theta_j\right)$$

Weigh decay term

$$\Omega(\Theta, \Theta') = (\|\mathbf{W}\|^2 + \|b\|^2 + \|\mathbf{W}'\|^2 + \|b'\|^2)$$

The Solution of CRA

- We use the gradient descent method to derive the solution of all parameters

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{\partial \mathcal{J}}{\partial \mathbf{W}}, \qquad b \leftarrow b - \eta \frac{\partial \mathcal{J}}{\partial b},$$

$$\mathbf{W}' \leftarrow \mathbf{W}' - \eta \frac{\partial \mathcal{J}}{\partial \mathbf{W}'}, \quad b' \leftarrow b' - \eta \frac{\partial \mathcal{J}}{\partial b'},$$

$$\boldsymbol{\theta}_j \leftarrow \boldsymbol{\theta}_j - \eta \frac{\partial \mathcal{J}}{\partial \boldsymbol{\theta}_j},$$

η is the learning rate. The time complexity is $O(rnmk)$

The output: the encoding and decoding parameters, and source classifiers with latent representation.

Target Classifier Construction

INSTITUTE OF COMPUTING TECHNOLOGY

Two Scheme:

- Train the source classifiers based on $Z_S^{(1)}, Z_S^{(2)}, \cdots, Z_S^{(r)}$ and combine them as

$$f_T(x_T) = \frac{1}{r}\sum_{j=1}^{r} \sigma\left(\theta_j^\top (\sigma(\mathbf{W}x_T + b))\right)$$, where $$\sigma(u) = \frac{1}{1+e^{-u}}.$$

- Combine all the source domain data as Z_S and train a unified classifier using any supervised learning algorithms, e.g., SVM, Logistic Regression (LR).
- The two accuracies are denoted as CRA_v and CRA_u , respectively

Data Sets-(1/2)

- **Image Data (from Luo et al., CIKM08) (Some examples)**

A_1 A_2 A_3 A_4

A

B

B_1 B_2 B_3 B_4

Three sources:

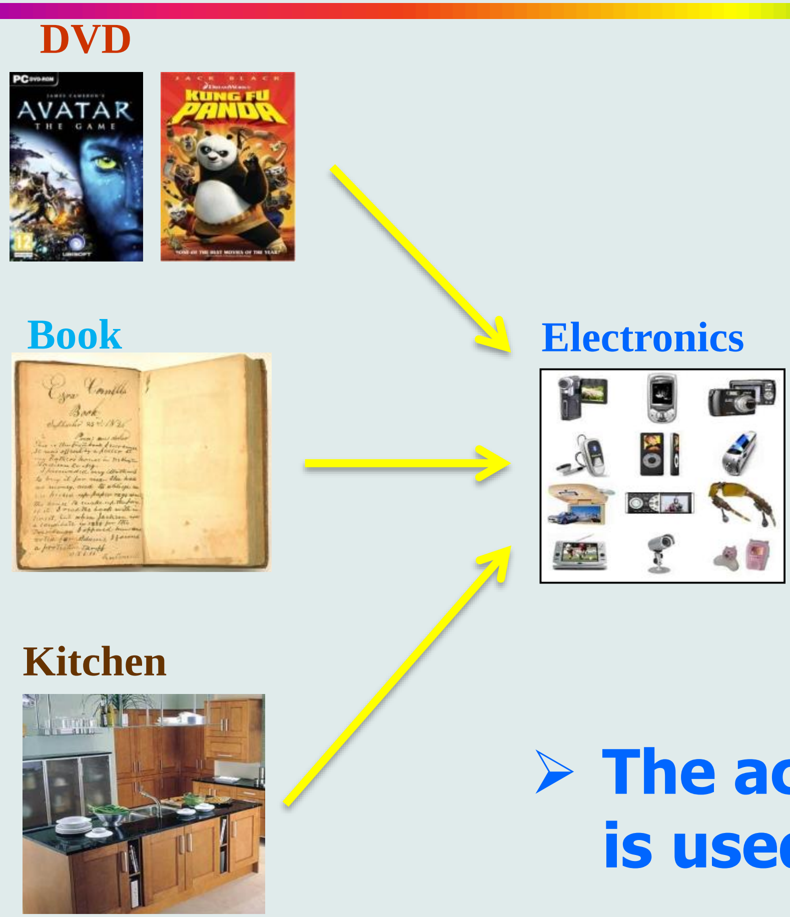
A_1B_1 A_2B_2 A_3B_3

Target domain:

A_4B_4

- **Totally, 96 ($4 \cdot P_4^4$) 3-source vs 1-target domain (3 vs 1) problem instances can be constructed for the experimental evaluation**

Data Sets-(2/2)



- Sentiment Classification (from Blitzer et al., ACL07)
- Four 3-source vs 1-target domain classification problems are constructed

- The accuracy on target domain data is used as the evaluation measure
- Both SVM and LR are used to train classifiers on the new representation

All Compared Algorithms

INSTITUTE OF COMPUTING TECHNOLOGY

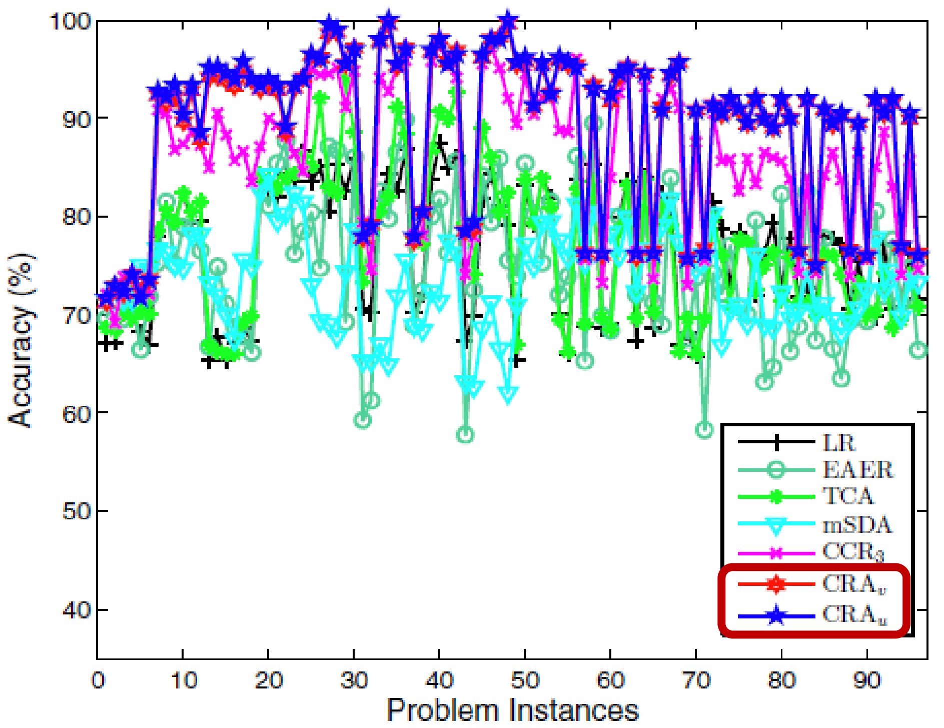
- **Baselines**
  - **Supervised learning on original features: SVM [Joachims, ICML'99], Logistic Regression (LR) [David et al., 00]**
  - **Embedding method based on autoencoders (EAER) [Yu et al., ECML'13]**
  - **Marginalized Stacked Denoising Autoencoders (mSDA) [Chen et al., ICML'12]**
  - **Transfer Component Analysis (TCA) [Pan et al., TNN'11]**
  - **Transfer learning from multiple sources (CCR_3) (Luo et al., CIKM'08)**
- **Our method: CRA_v and CRA_u**

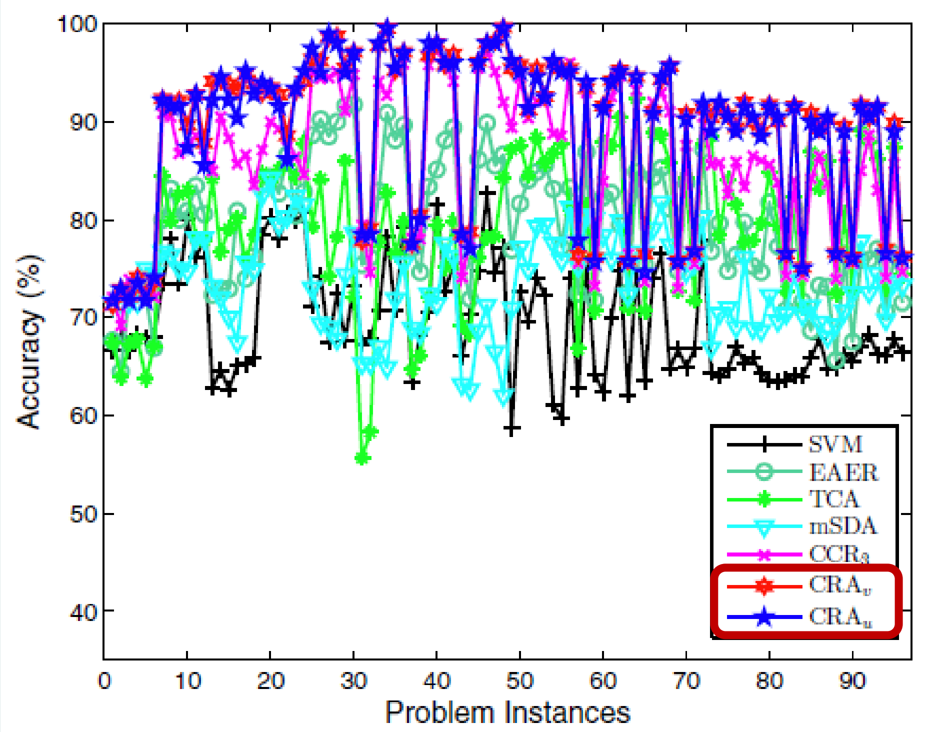
For the methods which can not handle multiple sources, we train the classifiers from each source domain and merged data of all sources (r+1 accuracies). Finally, maximal, mean and minimal values are reported.

Experimental Results-(1/2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

➢ Results on 96 image classification problems





	LR	SVM	LR		SVM		mSDA	CCR$_3$	CRA$_v$	LR	SVM
			EAER	TCA	EAER	TCA				CRA$_u$	CRA$_u$
Max	83.9	81.7	83.2	84.2	85.6	85.2	83.1	87.5			
Min	65.0	56.0	62.3	66.8	71.3	69.8	64.6	83.5	**89.2**	**89.4**	88.9
Mean	76.1	69.6	74.9	77.0	79.4	79.1	73.5	85.9			

Experimental Results-(2/2)

➢ Results on 4 sentiment classification problems

Tasks		LR	SVM	LR EAER	LR TCA	SVM EAER	SVM TCA	mSDA	CCR_3	CRA_v	LR CRA_u	SVM CRA_u
tar.book	Max	79.3	78.4	67.8	68.5	73.0	66.2	**82.3**	78.6			
	Min	71.0	71.5	57.0	58.9	69.3	59.3	77.6	78.2	79.2	79.2	79.1
	Mean	75.7	74.9	63.0	64.2	70.9	62.8	**79.9**	78.4			
tar.kitchen	Max	85.6	85.4	78.9	75.2	77.5	73.1	84.7	**86.1**			
	Min	76.4	74.9	71.0	64.2	75.9	64.7	81.4	85.6	85.9	**86.3**	85.8
	Mean	81.0	80.5	76.6	69.4	76.7	68.7	83.5	85.9			
tar.elec.	Max	83.9	83.1	74.2	72.9	72.8	70.5	**85.2**	79.3			
	Min	73.5	73.0	68.5	60.7	69.4	59.4	74.4	75.4	84.1	**84.7**	82.4
	Mean	78.7	78.9	70.8	67.1	71.2	65.2	81.0	75.6			
tar.dvd	Max	79.7	79.5	69.5	68.5	70.8	67.4	**82.3**	80.2			
	Min	73.6	72.2	56.5	61.4	67.7	61.3	78.2	79.7	80.6	**81.1**	80.8
	Mean	77.0	75.9	65.1	65.2	69.0	64.3	80.3	80.1			
Average	Max	82.1	81.6	72.6	71.3	73.5	69.3	**83.7**	81.1			
	Min	73.6	72.9	63.2	61.3	70.6	61.2	77.9	79.7	82.5	**82.8**	82.0
	Mean	78.1	77.5	68.9	66.5	72.0	65.3	81.2	80.5			

Conclusions

INSTITUTE OF COMPUTING TECHNOLOGY

- The well known representation learning technique autoencoder is considered, and we formalize the autoencoders and consensus regularization into a unified optimization framework

- Extensive comparison experiments on image and sentiment data are conducted to show the effectiveness of the propose algorithm

Cross-domain Novelty Seeking Trait Mining for Recommendation

Motivation (1/2)

用户序列行为建模：

用户在不同领域具有各种行为序列

$$O_1^s \rightarrow O_2^s \rightarrow O_3^s \rightarrow O_4^s \rightarrow ?$$

$$O_1^t \rightarrow O_2^t \rightarrow O_3^t \rightarrow O_4^t \rightarrow ?$$

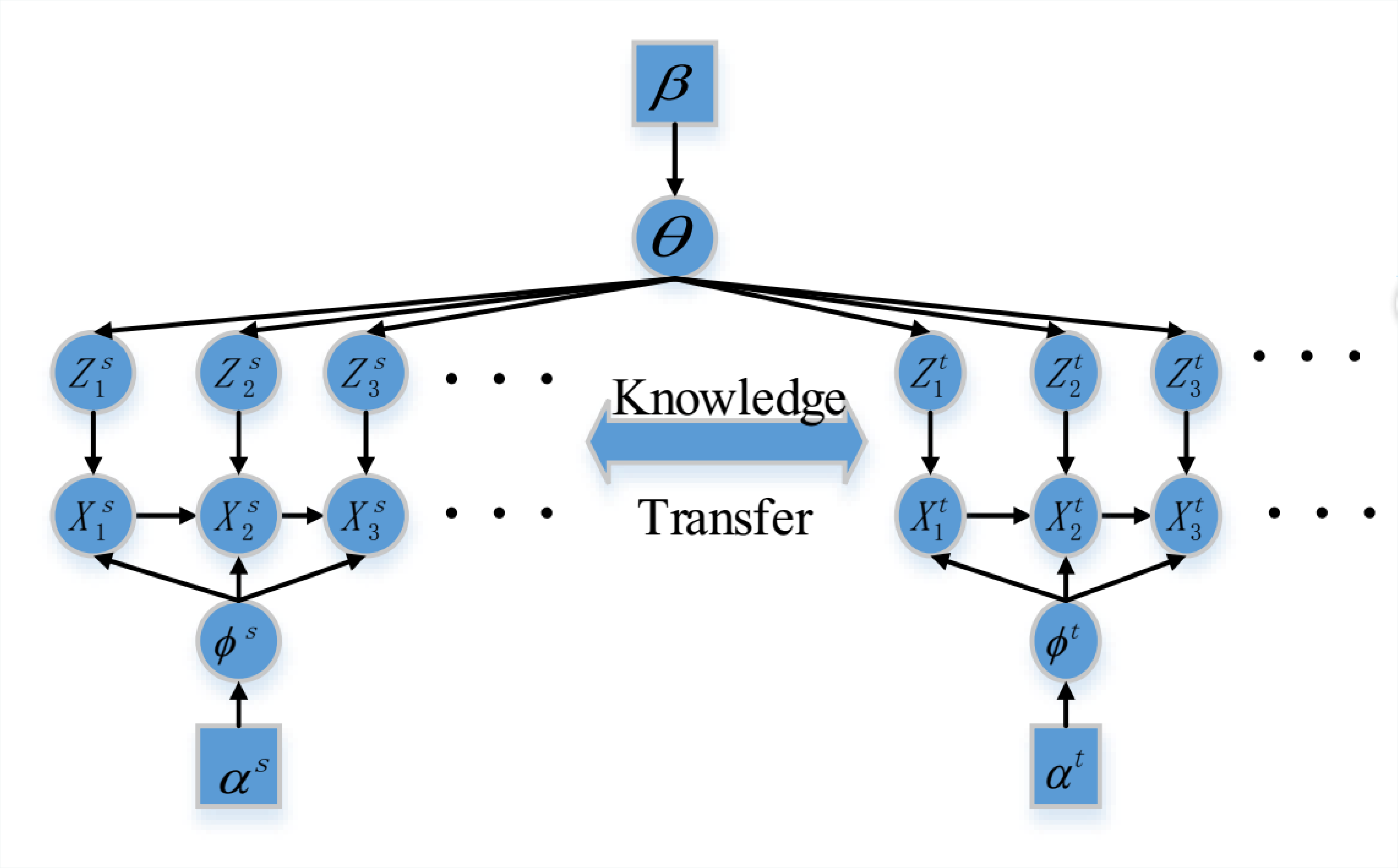
用户在各种领域有自己的行为序列数据，如何对下一个行为进行预测？

当目标领域数据较少的情况下，如何借助不同领域的数据迁移知识，辅助目标领域的行为预测？

Motivation (2/2)

新颖性建
模和例子

Cross-domain Model

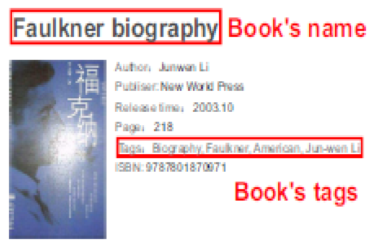


Data Set & Experimental Results

(a) An user's watching list of movies.

(b) An example of movie's information.

(c) An example of music's information.



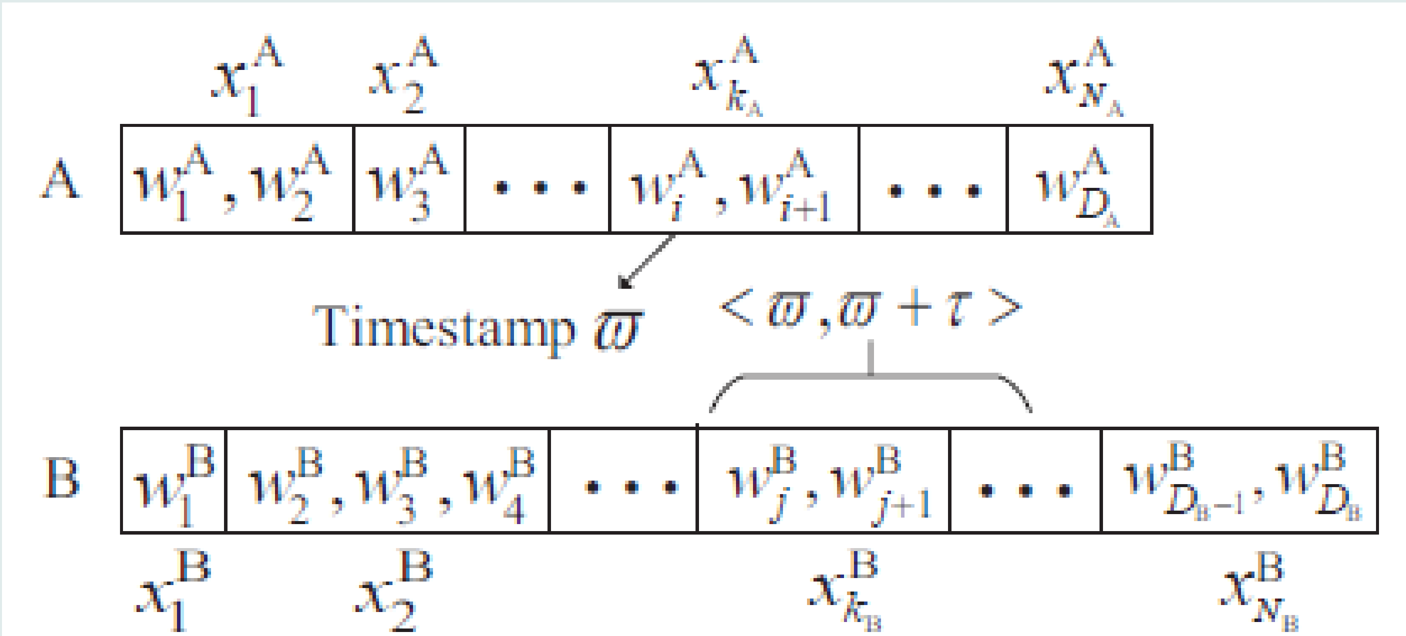
(d) An example of book's information.

Table 2: The statistics of seven pairs of data sets

Source → Target	Statistics	
Movie_category → Music_tags Music_tags → Movie_category	#user	1,653
	#Movie_category	368,446
	#Ave_Movie_category	222.90
	#Music_tags	9,229
	#Ave_Music_tags	5.58
Movie_tags → Music_tags Music_tags → Movie_tags	#user	1,653
	#Movie_tags	317,742
	#Ave_Movie_tags	192.22
	#Music	9,229
	#Ave_Music_tags	5.58
Movie_dir → Music_tags Music_tags → Movie_dir	#user	1,653
	#Movie_tags	373,164
	#Ave_Movie_tags	225.75
	#Music_tags	9,229
	#Ave_Music_tags	5.58
Music_tags → Book_tags Book_tags → Music_tags	#user	423
	#music_tags	2,474
	#Ave_music_tags	5.8
	#Book_tags	25,342
	#Ave_Book_tags	59.91

		OF(OF_U)	MC(MC_U)	NSM(NSM_U)	CDNST
$A \rightarrow B$	MRR	0.1601(0.1522)	0.2015(0.1779)	0.3128(0.3017)	**0.3623**
	nDCG@15	0.2153(0.2047)	0.2677(0.2299)	0.3821(0.3673)	**0.4363**
	p@3	0.1044(0.0937)	0.1409(0.1203)	0.2822(0.2736)	**0.3325**
$B \rightarrow A$	MRR	0.3982(0.2413)	0.4135(0.2575)	**0.5644**(0.3180)	0.5014
	nDCG@15	0.4998(0.3279)	0.5125(0.3715)	**0.6489**(0.3945)	0.5687
	p@3	0.3373(0.2100)	0.3649(0.2241)	**0.5488**(0.2992)	0.4797

Result Analysis (1/2)



$$Sim(A_u \rightarrow B_u) = \frac{1}{N_p} \sum_{i=1}^{D_A} \sum_{w_j^B \in W_i^B} Sim(v_i^A, v_j^B),$$

$$Sim(A \rightarrow B) = \frac{1}{N} \sum_{u} \frac{1}{S} im(A_u \rightarrow B_u).$$

Result Analysis (2/2)

Table 6: The Relatedness on 7 Pairs of Data Sets

Music_tags →Movie_category	Music_tags → Movie_tags	Music_tags → Movie_dir	Music_tags → Book_tags	Book_tags → Movie_category	Book_tags → Movie_tags	Book_tags → Movie_dir
0.3125	0.4891	0.3559	0.3559	0.3217	0.1650	0.3329
Movie_category →Music_tags	Movie_tags → Music_tags	Movie_dir → Music_tags	Book_tags → Music_tags	Movie_category → Book_tags	Movie_tags → Book_tags	Movie_dir → Book_tags
0.2559	0.3290	0.2704	0.1794	0.2008	0.05228	0.1835
0.3901 ↑	0.4956 ↑	0.4490 ↑	0.3979 ↑	0.3139 ↑	0.1633 ↑	0.3314 ↑

		Movie_category → Music_tags	Movie_tags → Music_tags	Movie_dir → Music_tags	Book_tags → Music_tags	Movie_category → Book_tags	Movie_tags → Book_tags	Movie_dir → Book_tags
MRR	OF	0.5093	0.5093	0.5093	0.5149	0.2483	0.2483	0.2483
	OF_U	0.3033	0.3004	0.3990	0.1736	0.1895	0.1581	0.1655
	MC	0.5303	0.5303	0.5303	0.5275	0.2588	0.2588	0.2588
	MC_U	0.3921	0.3174	0.2969	0.1869	0.2414	0.1853	0.1822
	NSM	**0.6842**	**0.6842**	**0.6842**	**0.6891**	**0.4031**	**0.4031**	**0.4031**
	NSM_U	0.1808	0.3994	0.3935	0.3522	0.2129	0.3198	0.3676
	CDNST	0.6745	0.5755	0.5659	0.6628	0.3616	0.3347	0.3350
nDCG@15	OF	0.6145	0.6145	0.6145	0.6584	0.3323	0.3323	0.3323
	OF_U	0.4192	0.4068	0.3410	0.4398	0.2539	0.2134	0.2210
	MC	0.6291	0.6291	0.6291	0.6670	0.3443	0.3443	0.3443
	MC_U	0.5019	0.4234	0.4055	0.4504	0.3173	0.2531	0.2491
	NSM	**0.7599**	**0.7599**	**0.7599**	**0.7657**	**0.4989**	**0.4989**	**0.4989**
	NSM_U	0.2266	0.4706	0.4882	0.4480	0.2681	0.4005	0.4595
	CDNST	0.7442	0.6125	0.5990	0.7353	0.4549	0.4326	0.4025

Result Analysis (2/2)

Table 6: The Relatedness on 7 Pairs of Data Sets

Music_tags →Movie_category	Music_tags → Movie_tags	Music_tags → Movie_dir	Music_tags → Book_tags	Book_tags → Movie_category	Book_tags → Movie_tags	Book_tags → Movie_dir
0.3125	0.4891	0.3559	0.3559	0.3217	0.1650	0.3329
Movie_category →Music_tags	**Movie_tags → Music_tags**	**Movie_dir → Music_tags**	**Book_tags → Music_tags**	**Movie_category → Book_tags**	**Movie_tags → Book_tags**	**Movie_dir → Book_tags**
0.2559	0.3290	0.2704	0.1794	0.2008	0.05228	0.1835
0.3901 ↑	0.4956 ↑	0.4490 ↑	0.3979 ↑	0.3139 ↑	0.1633 ↑	0.3314 ↑

		Movie_category → Music_tags	Movie_tags → Music_tags	Movie_dir → Music_tags	Book_tags → Music_tags	Movie_category → Book_tags	Movie_tags → Book_tags	Movie_dir → Book_tags
MRR	OF	0.5093	0.5093	0.5093	0.5149	0.2483	0.2483	0.2483
	OF_U	0.4056	0.4273	0.3926	0.4672	0.1519	0.1721	0.1367
	MC	0.5303	0.5303	0.5303	0.5275	0.2588	0.2588	0.2588
	MC_U	0.4295	0.4412	0.4093	0.4620	0.1982	0.2073	0.1842
	NSM	0.6842	0.6842	0.6842	0.6891	0.4031	0.4031	0.4031
	NSM_U	0.6018	0.6294	0.5987	0.6341	0.3492	0.3263	0.3392
	CDNST	**0.7054**	**0.7122**	**0.6946**	**0.7067**	**0.4183**	**0.4942**	**0.4128**
nDCG@15	OF	0.6145	0.6145	0.6145	0.6584	0.3323	0.3323	0.3323
	OF_U	0.5561	0.5726	0.5437	0.5983	0.2572	0.2834	0.2163
	MC	0.6291	0.6291	0.6291	0.6670	0.3443	0.3443	0.3443
	MC_U	0.5836	0.5982	0.5727	0.6068	0.3064	0.2985	0.3183
	NSM	0.7599	0.7599	0.7599	0.7657	0.4989	0.4989	0.4989
	NSM_U	0.6283	0.6429	0.6157	0.6548	0.4851	0.4746	0.4954
	CDNST	**0.7749**	**0.7826**	**0.7652**	**0.7857**	**0.5163**	**0.5089**	**0.5281**

Conclusions

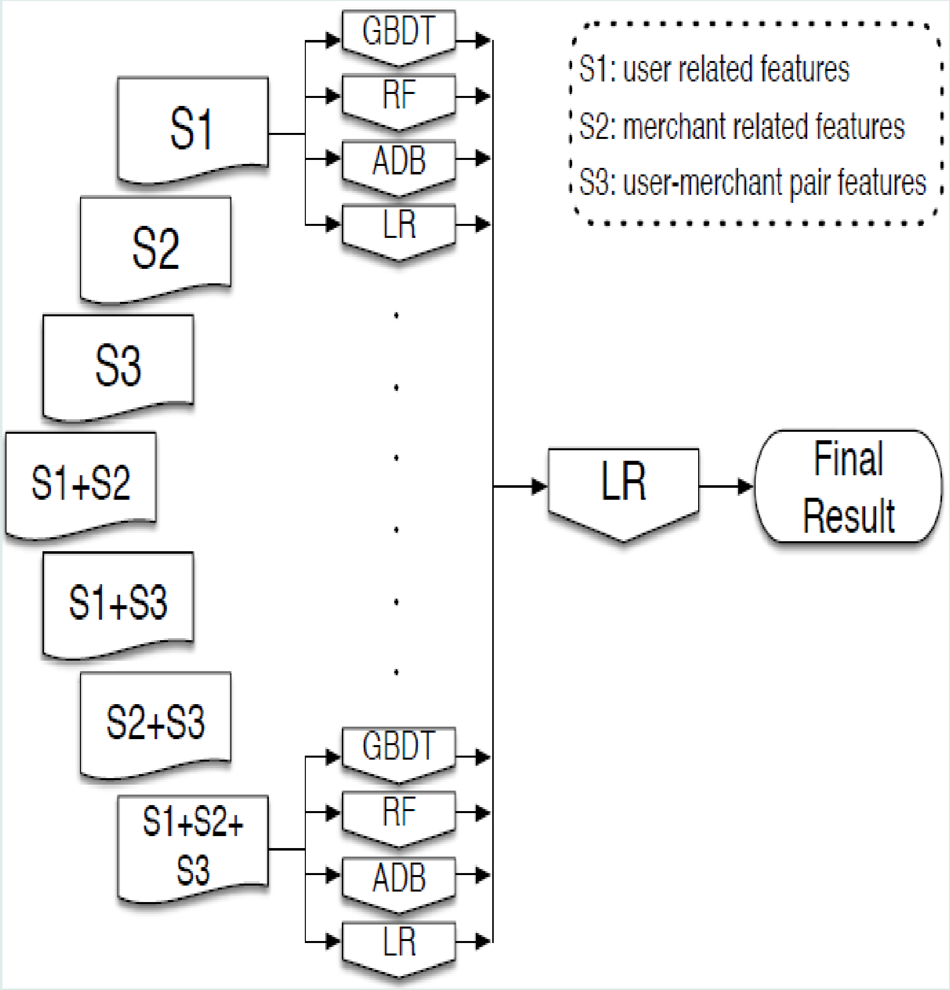
- We propose a new sequential transfere learning for recommendation
- We proposed a new cross-domain recommendation algorithm, in which the novelty-seeking trait of users are shared across source and target domains for effective knowledge
- We define an effective relatedness measure to judge when CDNST can work
- Extensive experiments conducted on real-world data sets demonstrate the effectiveness of CDNST

基于模型融合的电商平台潜在重复购买用户预测

IJCAI 2015数据挖掘竞赛任务

- 阿里巴巴提供数据：
  - 商家有时候在特殊的日子做大的促销活动，吸引更多的买家
  - 很多被吸引的买家只在该商家购买一次，这些促销活动带来的效应非常短
  - 提供天猫上商家、买家丰富的日志信息
- 任务
  - 识别、预测哪些买家可以转化为重复购买的买家非常重要
  - 提供6个月历史日志数据，预测哪些买家可以在同一个商家重复购买

- 设计方案：



IJCAI 2015数据挖掘竞赛第一名

Methods	AUC in local evaluation
GBDT	0.688379
Random Forest	0.688377
AdaBoost	0.683360
Logistic Regression	0.681488
Model Ensemble	0.691793
Model+Feature Ensemble	**0.692564**

Leaderboard

Rank	Score	Nickname	Best Submission (GMT +8)
1	0.711373	hrem	2015-06-21 02:17:39
2	0.711287	LeavingSeason	2015-06-21 08:22:31
3	0.710163	FAndy&kimiyoung&Neo	2015-06-21 08:58:24
4	0.709877	偏执狂小江	2015-06-21 08:32:49
5	0.709464	OneP	2015-06-21 02:03:28
6	0.709070	parameicnm	2015-06-21 07:49:53
7	0.708976	9*STAR	2015-06-18 23:00:39
8	0.706037	senochow	2015-06-20 22:50:41

References

[1] **Fuzhen Zhuang**, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, Zhongzhi Shi: Exploiting Associations between Word Clusters and Document Classes for Cross-Domain Text Categorization. ***SDM 2010***, pp. 13-24.

[2] **Fuzhen Zhuang**, Ping Luo, Zhiyong Shen, Qing He, Yuhong Xiong, Zhongzhi Shi, Hui Xiong: Collaborative Dual-PLSA: mining distinction and commonality across multiple domains for text classification. ***CIKM 2010***, pp. 359-368.

[3] **Fuzhen Zhuang**, Ping Luo, Hui Xiong, Yuhong Xiong, Qing He, Zhongzhi Shi: Cross-Domain Learning from Multiple Sources: A Consensus Regularization Perspective. ***IEEE Trans. Knowl. Data Eng***. 22 (12): 1664-1678 (2010).

[4] **Fuzhen Zhuang**, Ping Luo, Zhiyong Shen, Qing He, Yuhong Xiong, Zhongzhi Shi, Hui Xiong: Mining Distinction and Commonality across Multiple Domains Using Generative Model for Text Classification. ***IEEE Trans. Knowl. Data Eng***. 24 (11): 2025-2039 (2012).

[5] **Fuzhen Zhuang**, Ping Luo, Changying Du, Qing He, Zhongzhi Shi: Triplex transfer learning: exploiting both shared and distinct concepts for text classification. ***WSDM 2013***, pp. 425-434.

References

[6] **Fuzhen Zhuang**, Ping Luo, Peifeng Yin, Qing He, Zhongzhi Shi.: Concept Learning for Cross-domain Text Classification: a General Probabilistic Framework. ***IJCAI 2013***, pp.1960-1966.

[7] **Fuzhen Zhuang**, Xiaohu Cheng, Sinno Jialin Pan, Wenchao Yu, Qing He, Zhongzhi Shi: Transfer Learning with Multiple Sources via Consensus Regularized Autoencoders. **ECML/PKDD** (3) 2014 : 417-431.

[8] **Fuzhen Zhuang**, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, Qing He: Supervised Representation Learning: Transfer Learning with Deep Autoencoders. **IJCAI 2015**.

[9] **Fuzhen Zhuang**, Dan Luo, Xin Jin, Hui Xiong, Ping Luo, Qing He: Representation Learning via Semi-Supervised Autoencoder for Multi-task Learning. **ICDM 2015**: 1141-1146.

[10] **Fuzhen Zhuang**, Ping Luo, Sinno Jialin Pan, Hui Xiong, Qing He: Ensemble of Anchor Adapters for Transfer Learning. **CIKM 2016**: 2335-2340.

[11] **Fuzhen Zhuang**, Zhou Yingmin, Fuzheng Zhang, Xiang Ao, Xing Xie and Qing He: Sequential Transfer Learning: Cross-domain Novelty Seeking Trait Mining for Recommendation. **WWW 2017**.

References

[12] Baoyu Jing, Chenwei Lu, Deqing Wang, **Fuzhen Zhuang**, Cheng Niu: Cross-Domain Labeled LDA forText Classification. ICDM 2018.

[13] Yongchun Zhu, **Fuzhen Zhuang***, Deqing Wang: Aligning Domain-specific Distribution and Classifier for Cross-domain Classification from Multiple Sources. **AAAI 2019**.

References

[1] Zhao Zhang , **Fuzhen Zhuang*** , Meng Qu , Fen Lin , Qing He : Knowledge Graph Embedding with Hierarchical Relation Structure. EMNLP 2018 : 3198-3207.

[2] Fuzhen Zhuang , Dan Luo , Nicholas Jing Yuan , Xing Xie , Qing He : Representation Learning with Pair-wise Constraints for Collaborative Ranking. WSDM 2017 : 567-575.

[3] **Fuzhen Zhuang**, Dan Luo, Xin Jin, Hui Xiong, Ping Luo, Qing He: Representation Learning via Semi-Supervised Autoencoder for Multi-task Learning. **ICDM 2015** : 1141-1146.

[4] Xin Jin, **Fuzhen Zhuang**, Sinno Jialin Pan, Changying Du, Ping Luo, Qing He : Heterogeneous Multi-task Semantic Feature Learning for Classification. **CIKM 2015** : 1847-1850.

[5] Xin Jin, Ping Luo, **Fuzhen Zhuang**, Jia He, Qing He: Collaborating between Local and Global Learning for Distributed Online Multiple Tasks. **CIKM 2015** : 113-122.

[6] Xin Jin, **Fuzhen Zhuang**, Hui Xiong, Changying Du, Ping Luo, Qing He: Multi-task Multi-view Learning for Heterogeneous Tasks. **CIKM 2014** : 441-450.

References

INSTITUTE OF COMPUTING TECHNOLOGY

[7] Xin Jin, **Fuzhen Zhuang**, Shuhui Wang, Qing He, Zhongzhi Shi: Shared Structure Learning for Multiple Tasks with Multiple Views. **ECML/PKDD (2) 2013** : 353-368.

[8] Changying Du , **Fuzhen Zhuang** , Qing He , Zhongzhi Shi : Multi-task Semi-supervised Semantic Feature Learning for Classification. **ICDM 2012** : 191-200.

[9] Changying Du , Jia He , **Fuzhen Zhuang** , Yuan Qi , Qing He : Nonparametric Bayesian Multi-Task Large-margin Classification. ECAI 2014 : 255-260.

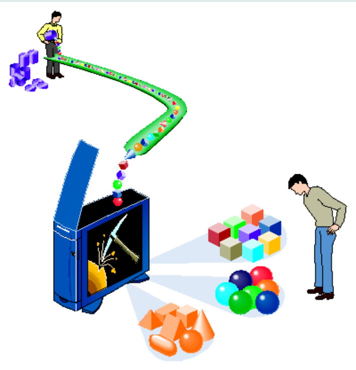
[10] Jia He , Changying Du , **Fuzhen Zhuang** , Xin Yin , Qing He , Guoping Long : Online Bayesian Max-Margin Subspace Multi-View Learning. IJCAI 2016 : 1555-1561.

[11] Zhao Zhang , **Fuzhen Zhuang** , Zheng-Yu Niu , Deqing Wang , Qing He : MultiE: Multi-Task Embedding for Knowledge Base Completion. CIKM 2018 : 1715-1718.

[12] Xiao Zhang , Wenzhong Li , Vu Nguyen , **Fuzhen Zhuang*** , Hui Xiong , Sanglu Lu : Label-Sensitive Task Grouping by Bayesian Nonparametric Approach for Multi-Task Multi-Label Learning. IJCAI 2018 : 3125-3131.

谢谢！

Q&A



Source codes:

http://mldm.ict.ac.cn/platform/pweb/download.htm

http://www.intsci.ac.cn/users/zhuangfuzhen/