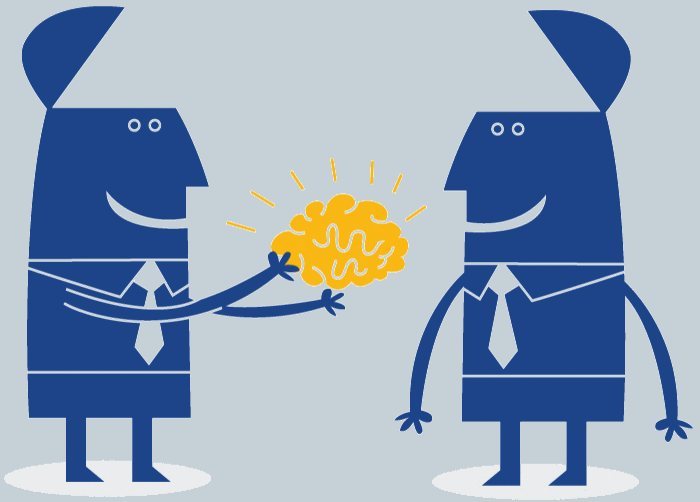
迁移学习

INTEGRATION OF GLOBAL AND LOCAL METRICS FOR DOMAIN ADAPTATION LEARNING

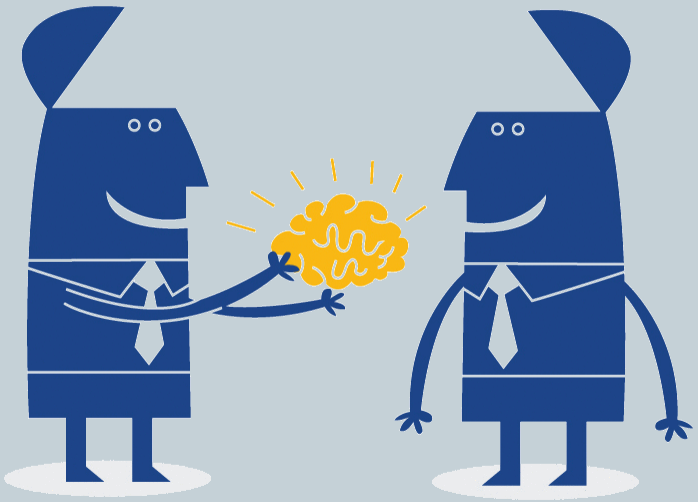


报告人：徐骏捷
Gods_Dusk@miriding.com

什么是迁移学习

- 迁移学习是运用已存有的知识对不同但相关领域问题进行求解的新的一种机器学习方法.

- 迁移学习的目标是将从一个应用场景中学到的知识，用来帮助新的应用场景中的学习任务



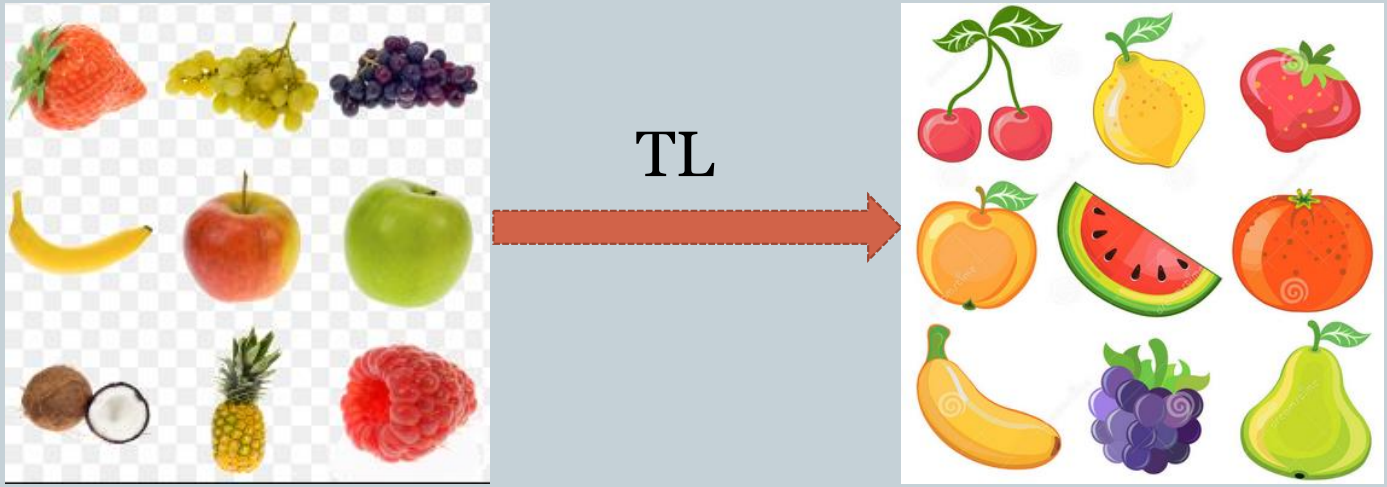
什么是迁移学习

- 人类具有迁移学习的能力

什么是迁移学习

- 机器的迁移学习任务

利用源域当中学习到的知识来提升目标域学习的速度或效果。



什么是迁移学习

- 逐渐减少训练精力的过程

Supervised Classification

什么是迁移学习

- 逐渐减少训练精力的过程

Supervised Classification

Semi-supervised Learning

什么是迁移学习

- 逐渐减少训练精力的过程

Supervised Classification

Semi-supervised Learning

Transfer Learning

为什么要迁移学习

- 现实中：训练数据和测试数据通常不服从同一种分布
- 训练消耗：已拥有大量带标签的数据或者已训练的分类器
- 效率：训练更快

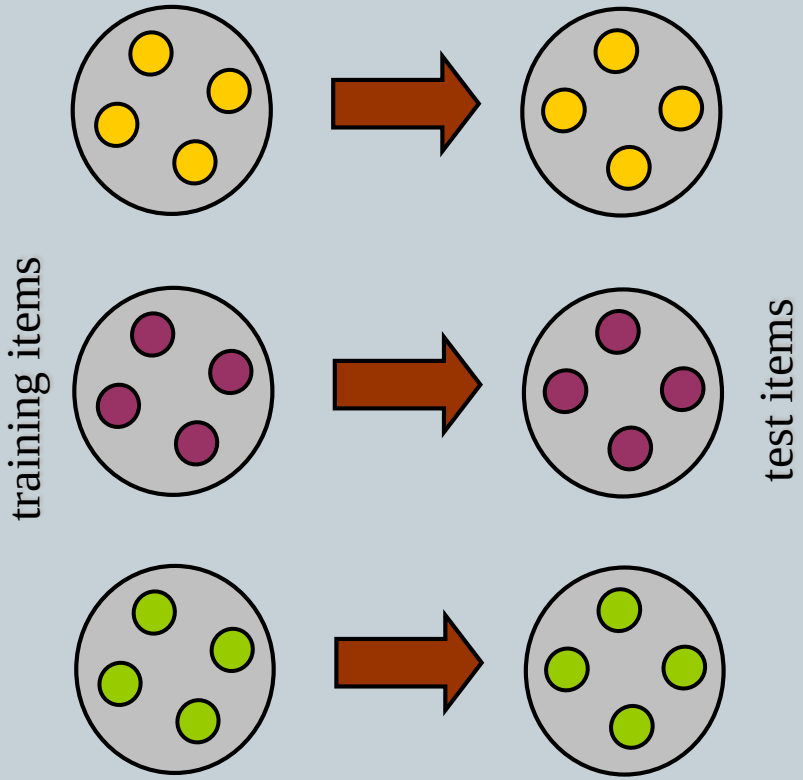
传统机器学习VS迁移学习

- 传统分类学习：两个基本的假设：
  - (1) 用于学习的训练样本与新的测试样本满足独立同分布的条件
  - (2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型

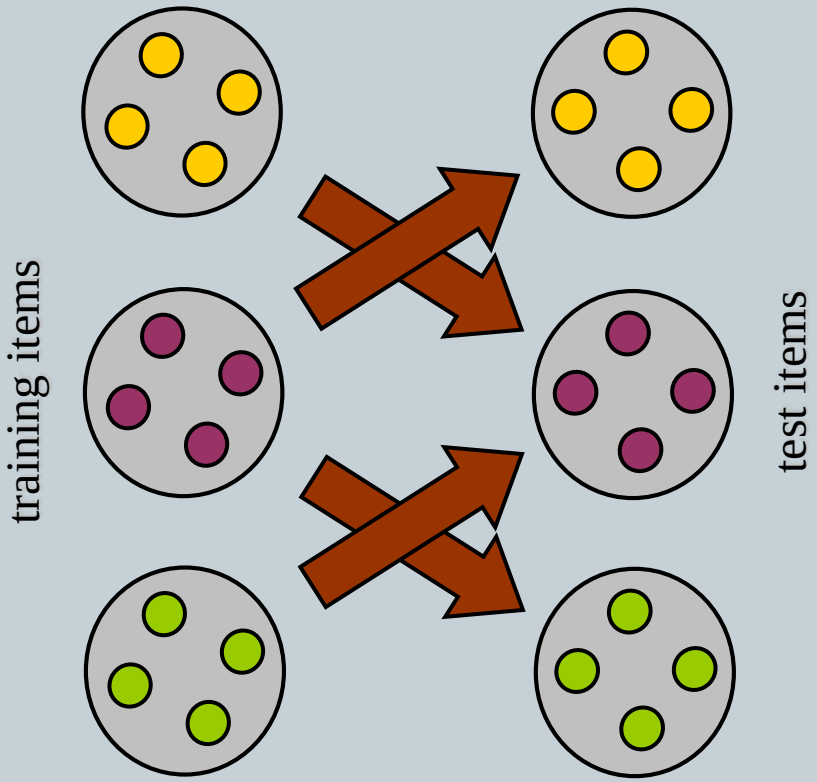
- 迁移学习：放宽假设，迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题

传统机器学习VS迁移学习

- 传统机器学习只适用于自己的领域

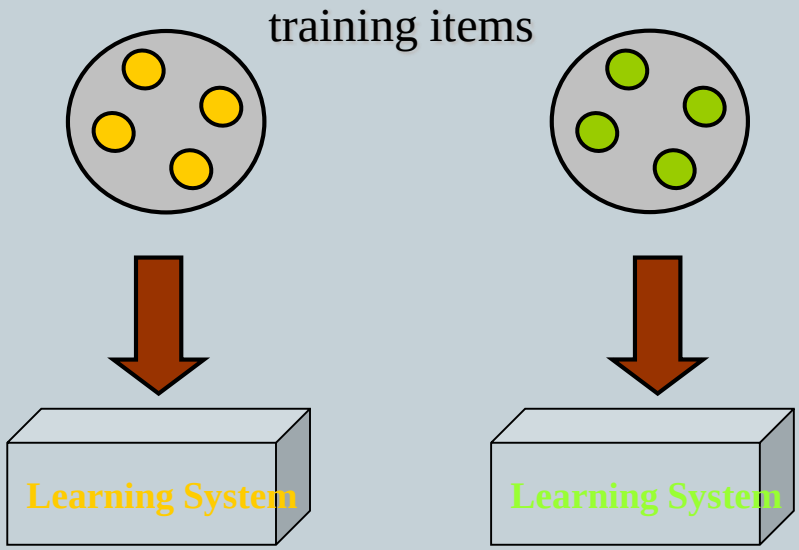


迁移学习可以交叉适用

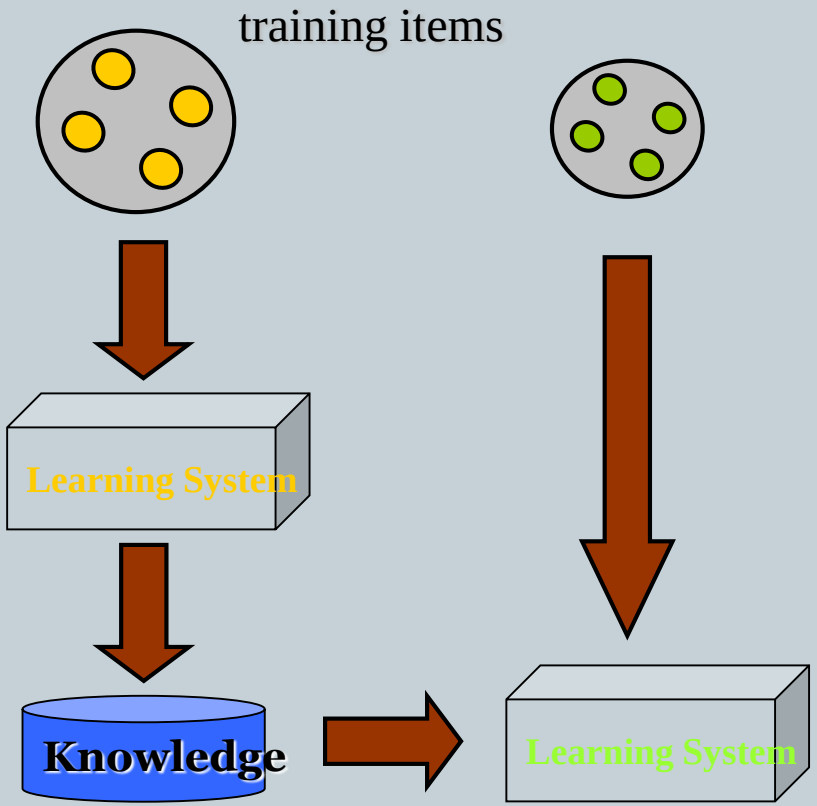


传统机器学习VS迁移学习

- 传统机器学习过程



迁移学习过程



迁移学习类别

- 源领域和目标领域样本是否标注以及任务是否相同：
  - 归纳迁移学习
    - 目标领域中有少量标注样本
  - 直推式迁移学习
    - 只有源领域中有标签样本
  - 无监督迁移学习
    - 源领域和目标领域都没有标签样本
- 迁移学习方法采用的技术划分：
  - 基于特征选择的迁移学习算法研究
    - 识别出源领域与目标领域中共有的特征表示
  - 基于特征映射的迁移学习算法研究
    - 共同映射到低维特征空间
  - 基于权重的迁移学习算法研究
    - 选择那些对目标领域分类有利的训练样本

迁移学习应用研究

- 目前,迁移学习典型的应用方面的研究主要包含有:
- 文本处理
  - 联合聚类方法，迁移贝叶斯分类器，双重迁移模型
- 图像分类
  - 翻译迁移学习方法，异构迁移学习方法
- 协同过滤
  - 特征子空间的迁移学习方法
- 基于传感器的定位估计

- Reuters-21578
- *Maximum Mean Discrepancy Embedding*

Data Set	SVM		TSVM	
	original	MMDE	original	MMDE
people vs place	0.519(0.039)	0.654(0.021)	0.553(0.025)	0.666(0.036)
orgs vs people	0.661(0.021)	0.722(0.034)	0.694(0.026)	0.726(0.033)
orgs vs places	0.670(0.025)	0.709(0.021)	0.704(0.035)	0.743(0.036)

迁移学习常用资源

- 文本挖掘数据集：20Newsgroups, SRAA, Reuters-21578
- 垃圾邮件过滤数据集：
  www.ecmlpkdd2006.org/challenge.html
- **WiFi**定位数据集：
  www.cse.ust.hk/~qyang/ICDMDMC2007
- 情感分类数据集：
  www.cis.upenn.edu/~mdredze/datasets/sentiment
- 加州大学伯克利分校的一些学者提供了一个关于迁移学习的MATLAB工具包：
  http://multitask.cs.berkeley.edu/

DA via Transfer Component Analysis

- 源域数据：$D_S = \{X_S, Y_S\}$；目标域数据：$D_T = \{X_T, Y_T\}$，所有数据构成$X = X_S \cup X_T$。设n = $|X|$。
- 所做的假设：

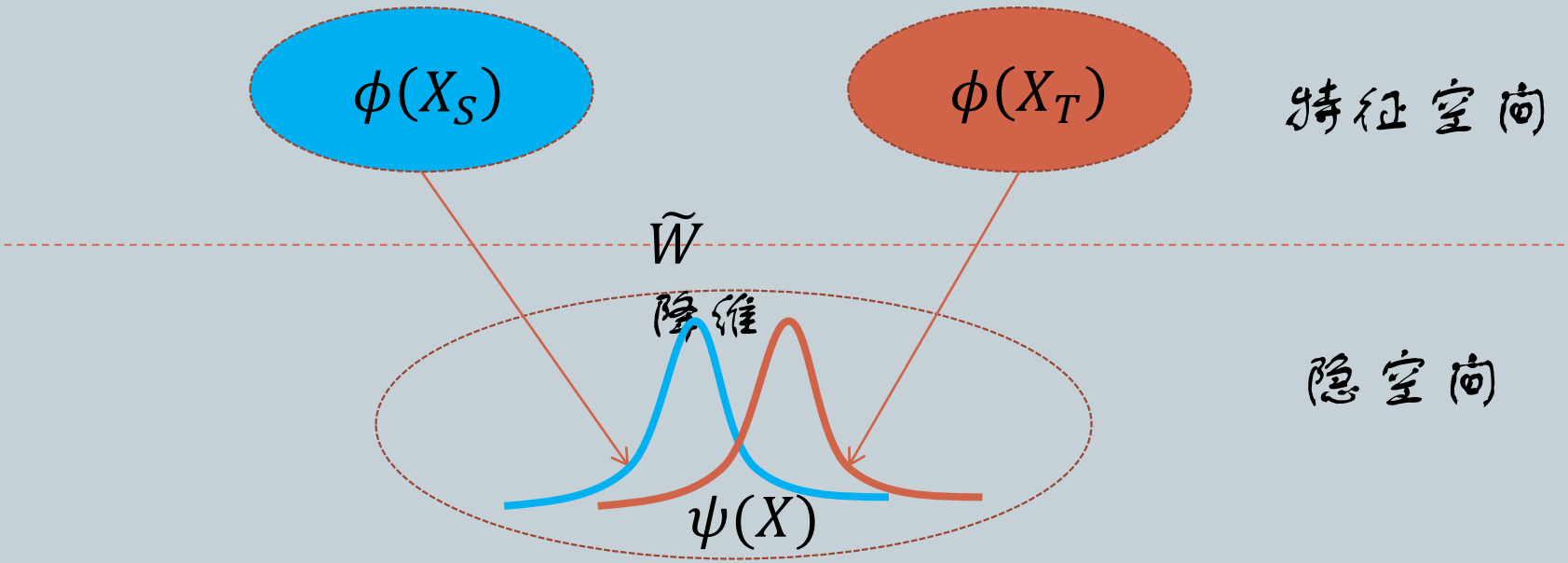
$$P(X_S) \neq P(X_T)$$
$$P(Y_S|\psi(X_S)) = P(Y_T|\psi(X_T))$$

其中，ψ将样本映射到隐空间中。

DA via Transfer Component Analysis

- 特征映射：ϕ，对应核函数κ。
- 降维映射：$\tilde{W}$。

$$\psi = \phi \circ \tilde{W}$$



TCA

- 经验核映射

设核函数κ在数据集X上构成的核矩阵为K，即$\mathrm{K_{ij}} = \kappa(X_i, X_j)$，则其经验核映射为：

$$\phi_n = K^{-\frac{1}{2}} \begin{pmatrix} \kappa(x_1, x) \\ \vdots \\ \kappa(x_n, x) \end{pmatrix}$$

性质：ϕ_n在X上的核矩阵同样是K。（对比ϕ_n与ϕ）

- ψ的核矩阵

考虑$n \times m$的降维矩阵$\tilde{W}$，可以推出ψ在X上的核矩阵为：

$$\tilde{K} = \left(KK^{-1/2}\tilde{W}\right)\left(\tilde{W}^T K^{-1/2} K\right) = KWW^T K$$

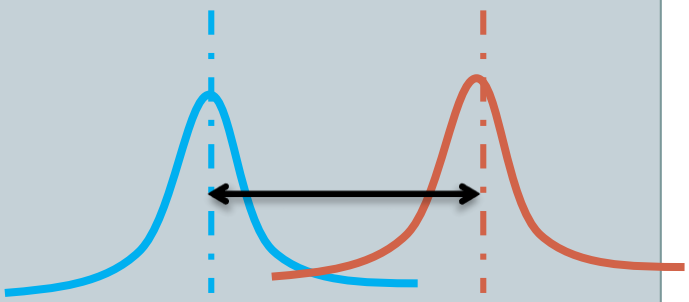
其中，$W = K^{-1/2}\tilde{W}$。

TCA

- 目标：最优化MMD距离。

$$\mathrm{MMD} = \left\| \overline{\psi(X_S)} - \overline{\psi(X_T)} \right\|^2$$

将此目标展开变成一系列隐空间样本的内积的和：$\psi(x_i) \cdot \psi(x_j)$。



设$L_{ij} \in \mathbb{R}^{n \times n}$只有$(i,j)$位置有值1，则

$$\mathrm{tr}\left(\mathrm{AL_{ij}}\right) = \mathrm{A_{ji}}$$

因此，可以写出L使得：

$$MMD = \mathrm{tr}\left(\tilde{K}L\right)$$
$$= \mathrm{tr}(W^T KLKW)$$

TCA

- 加入正则化项和约束条件后得到目标函数：

$$\min \operatorname{tr}\left(\mathrm{W}^{\mathrm{T}}\mathrm{W}\right)+\mu \operatorname{tr}(W^{T}KLKW)$$
$$s.t.W^{T}KHKW=I$$

其中，$H=I-\frac{1}{n}\mathbf{1}$，$W^{T}KHKW$定义了中心化矩阵，该约束用来避免$W=0$，防止把数据都映射到同一点上。

- 解析解

通过拉格朗日法可以把上述优化问题转换为KFD形式，从而得到其解析解：

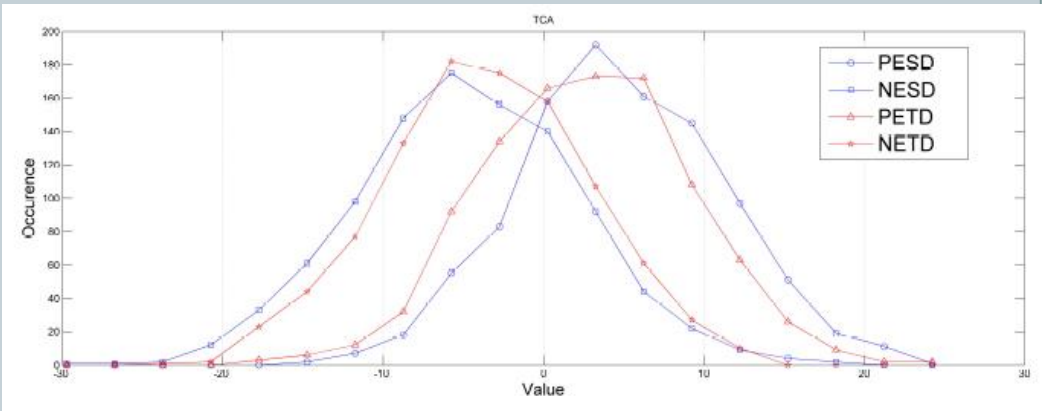
$$W=(I+\mu KLK)^{-1}KHK$$

TCA

- 在求出最优W之后，就可以计算每个样本映射到隐空间后的向量。通过对隐空间中的带标记数据构建分类器。

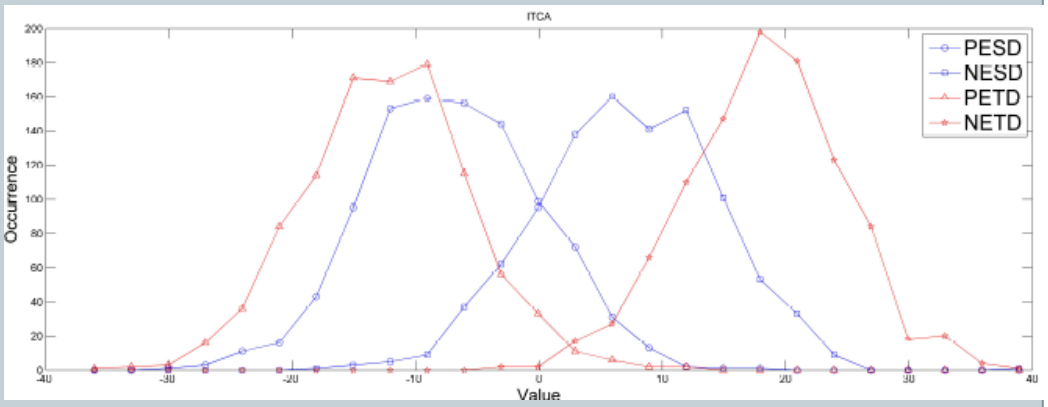
IGLDA

- 考虑源域中数据含有标签，那么我们不仅可以最小化MMD距离，还可以最小化类内距离。



- 优点：

1. 不仅使得全局上两个分布尽量相似，
2. 还使得类内距离尽量小，增加了样本的可分性。



IGLDA

- 设源域数据的类别构成集合：C。则类内距离Intra-Class Distance定义为：

$$\mathrm{ICD} = \sum_{C_k} \sum_{x_i, x_j \in C_k} \overline{\left\| \psi(x_i) - \psi(x_j) \right\|^2}$$

展开后同样是内积形式，所以得到类似的目标函数：

$$\min \mathrm{tr}\left(\mathrm{W}^{\mathrm{T}}\mathrm{W}\right) + \mu\, \mathrm{tr}(W^T KLKW) + \lambda \mathrm{tr}(W^T KL_{ICD}KW)$$
$$s.t. W^T KHKW = I$$

- 解析解

$$W = (I + \mu KLK + \lambda KL_{ICD}K)^{-1} KHK$$

实验结果

- Cross-Domain Text Classification

Methods	Exp.Dim.	people vs. places		orgs vs. people		orgs vs. places	
		ACC	SD	ACC	SD	ACC	SD
SVM		0.5096	0.0179	0.5010	0.0152	0.5286	0.0122
TCA+SVM	d=5	0.5790	0.0223	0.7522	0.0122	0.7168	0.0170
	d=10	0.6271	0.0194	0.7625	0.0127	0.6833	0.0279
	d=20	0.6003	0.0310	0.7735	0.0231	0.6873	0.0196
	d=30	0.5890	0.0188	0.7684	0.0230	0.6605	0.0213
SSTCA+SVM	d=5	0.6585	0.0261	0.6563	0.0428	0.6721	0.0414
	d=10	**0.6976**	0.0837	0.7156	0.0302	0.6755	0.0280
	d=20	0.6538	0.0626	0.6915	0.0243	0.6543	0.0401
	d=30	0.6151	0.0496	0.6954	0.0146	0.6610	0.0374
IGLDA+SVM	d=5	0.5991	0.0224	0.7701	0.0199	**0.7279**	0.0217
	d=10	0.6065	0.0379	0.7746	0.0207	0.6795	0.0225
	d=20	0.6122	0.0377	0.7726	0.0143	0.6767	0.0286
	d=30	0.6020	0.0269	**0.7803**	0.0149	0.6882	0.0258

参考文献

- Sinno Jialin Pan, Qiang Yang, A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering* (**IEEE TKDE**)
- http://apex.sjtu.edu.cn/apex_wiki/Transfer%20Learning
- 庄福振,罗平,何清,史忠植.迁移学习研究进展.软件学报,2015,26(1):26-39. http://www.jos.org.cn/1000-9825/