计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(11): 2262-2268, 2013

# 一种基于教学模型的协同训练方法

胡菊花 姜 远 周志华
(计算机软件新技术国家重点实验室(南京大学) 南京 210046)
(hujh@lamda.nju.edu.cn)

# A Co-Training Method Based on Teaching-Learning Model

Hu Juhua, Jiang Yuan, and Zhou Zhihua
(*National Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210046)

**Abstract** In many real tasks, there are usually abundant unlabeled data but only a few labeled data, and therefore, semi-supervised learning has attracted significant attention in the past few years. Disagreement-based semi-supervised learning approaches are a kind of state-of-the-art paradigm of semi-supervised learning, where multiple classifiers are generated to label unlabeled instances for each other. Co-training is the first and seminal work in this category. However, during the labeling process, most current co-training style approaches consider only the confidence of the predictor but not any helpfulness for the learner. In this paper, inspired by the real-world teaching-learning system, we propose a teaching-learning model named "TaLe" for co-training, within which the predictor is considered as a teacher who is teaching while the other is the student who is learning. Based on this model, a new variant of co-training algorithm named CoSnT is presented to consider both the confidence of the teacher and the need of the student. Intuitively, the convergence efficiency of co-training can be improved. Experiments on both multi-view and single-view data sets validate the efficiency and even outperformance of CoSnT over both standard co-training algorithm CoTrain that considers only teacher's confidence and CoS algorithm that considers only student's need.

**Key words** semi-supervised learning; disagreement-based; co-training; TaLe model; CoSnT; teaching confidence; learning need

**摘 要** 在很多实际问题中,很容易得到大量未标记数据而较难获取数据的标记;所以半监督学习在过去的10多年中得到了很大的关注.基于不一致性的半监督学习是其中一种十分重要的风范,协同训练是其代表方法.至今为止,大部分协同训练方法在选择未标记示例进行标记时只考虑预测学习器的置信度,而忽视了学习器的需求.受到真实教学系统的启发,提出了一种针对协同训练的教学模型TaLe,其中预测学习器是"教"者,而另一方则为"学"者.进而基于该模型给出了一种新的协同训练方法CoSnT,同时考虑了"教"的置信度和"学"的需求度.实验结果表明CoSnT在收敛效率和泛化性能上都优于标准的协同训练算法.

**关键词** 半监督学习;基于不一致性;协同训练;TaLe模型;CoSnT;"教"置信度;"学"需求度

**中图法分类号** TP181

**收稿日期**:2012-05-09;**修回日期**:2012-08-14

**基金项目**:国家自然科学基金项目(60975043,61021062);江苏省自然科学基金项目(BK2011566);深圳市高性能数据挖掘重点实验室开放课题(CXB201005250021A);百度大规模机器学习与数据挖掘主题研究项目(181215P00524)

**通信作者**:姜 远(jiangyuan@nju.edu.cn)

随着数据收集技术和存储技术的飞速发展,商业、工业、医学等各个领域都产生了海量的数据,但是在许多实际应用问题中,如文本分类、图像检索、医学数据处理,一般很难获取数据的标记,因为标记过程不仅需要耗费大量的人力和时间,而且需要一定的专业知识.如何自动、高效地利用大量廉价的未标记数据来帮助有限的标记数据进行训练,是半监督学习(semi-supervised learning)[1-2]期望解决的问题.

半监督学习主要分为4种方法:基于生成式模型(generative model)的方法[3-4]、基于低密度划分的方法[5-6]、基于图的方法[7-8]和基于不一致性(disagreement-based)的方法[9].特别地,基于不一致性的半监督学习方法得到了很大的关注,因为这类方法不存在其他类别方法的相关问题,如生成式模型方法的模型假设问题、低密度划分方法的直推式(transductive)约束、基于图方法的目标函数非凸问题.基于不一致性的半监督学习方法起源于Blum和Mitchell[10]提出的标准协同训练算法.

协同训练风范[11]已经成为基于不一致的半监督学习方法的一个重要代表.最初的标准协同训练算法假设整个属性集能够被自然地分成两个充分冗余(sufficient and redundant)的视图(view).首先在两个视图上利用标记数据分别训练学习器,然后每个学习器分别对未标记样本进行标记预测,并从中选择置信度较高的示例加入到对方的训练集中来辅助少量标记数据进行学习.

迄今为止,协同训练在理论分析、算法改进和实际应用上都得到了很大的发展.理论分析工作有文献[10,12-15],其中Wang和Zhou[14]证明了只要两个最初的学习器有足够的差异,那么通过协同训练利用未标记数据提高学习性能是能够得到保证的;之后,他们进一步给出了协同训练的充分必要性定理[15],解决了12年以来协同训练的充要条件一直不清楚的问题.同时为了适应不同实际问题的需求,协同训练算法得到了很多改进[16-20],其中Zhou和Li[19-20]首次将协同训练改进运用到半监督回归问题中.协同训练及其变体已经被广泛应用于各类实际问题中[18,21-24],其中Zhou等人[23]将协同训练和主动学习[25]进行了结合.

然而,大部分现有的协同训练方法在选择未标记样本进行标记的过程中,一般只考虑预测学习器的置信度而忽视了学习器的需求.直观上来讲,选择置信度高的样本进行标记能够避免太多的错误标记,而同时选择对学习器有帮助的样本则可以提高学习效率.Zhou等人[23]通过主动学习来寻找最具信息的未标记样本,但需要人工干预.

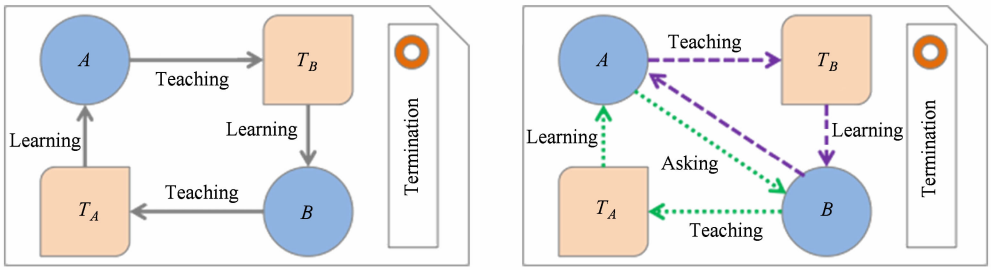
本文受到真实教学场景的启发,提出了一种针对协同训练的教学模型(teaching-learning model,TaLe).在该模型中预测学习器的标记过程被看作"教(teaching)"的过程,而在增大的训练集上进行重新训练则被看作"学(learning)"的过程.其实在协同训练过程中,协同训练的双方可能存在这样的差异性:学习器很难确信判断类别的点,学习器能够以较高的置信度加以标记,直观上来说对这样的点加以标记加入到训练集中更有助于学习器的学习,从而提高协同训练的收敛效率.

基于TaLe模型,本文提出了一种既考虑"教"置信度又考虑"学"需求度的协同训练方法CoSnT.在CoSnT的每一轮中,"学"者会向"教"者询问自己最有需求的样本标记,"教"者则会根据这一请求选择不仅自己置信度较高而且对"学"者有帮助的样本来标记.实验结果表明,CoSnT算法对比于标准协同训练算法(只考虑"教"置信度)和CoS算法(只考虑"学"需求度),不仅提高了收敛效率,泛化性能也较优.

## 1 TaLe模型

设想协同训练在真实的教学场景下,其实不难发现,每一轮双方轮流进行着"教"与"学"的过程.其中对未标记样本进行选择标记来增加对方训练集就是"教"的过程,而学习器在训练集上重新训练就是"学"的过程,所以协同训练过程基本都可以表示为如图1(a)所示的教学模型TaLe.两个学习器和通过各自的训练集进行不断的"教"与"学",直到达到模型预定的停止条件"Termination".

在真实的教学场景中不难发现,学生由于自身知识水平的限制可能会遇到自己当前难以解决或者不确定的问题,于是他们会向老师提出疑问.通过老师针对性的解答学生不仅得到了问题的解决,也提高了自身的能力;更重要的是这些积极主动的学生一般进步更快.受这一现实场景启发,很容易得到图1(b)所示的教学模型,其中协同训练每一轮包括两个教学环(不同线型),每个环中包含询问(asking)、教(teaching)、学(learning)3个步骤.基于此教学模型,在协同训练的过程中增加学习器根据自身需求询问的步骤而不是盲目接受标记信息,直观上来说能够提高协同训练的效率.




(a) Standard co-training (b) Add "asking" step

Fig. 1 TaLe model (Through and, learner, do co-training until "Termination").

图 1 教学模型

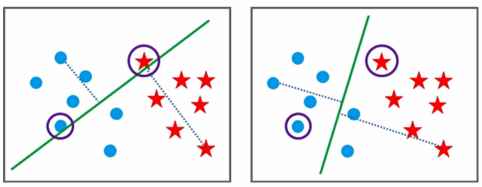
## 2 CoSnT 方法

### 2.1 相关概念

本文用 $\mathcal{X}_1$ 和 $\mathcal{X}_2$ 分别表示协同训练的两个视图,从而构成属性空间 $\mathcal{X}=\langle\mathcal{X}_1,\mathcal{X}_2\rangle$,目标概念为 $\mathcal{Y}\in\{0,1\}$,0 和 1 表示负类和正类. 标记样本 $(\boldsymbol{x},y)$ 则可表示为 $(\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle,y)$,其中 $\boldsymbol{x}_1\in\mathcal{X}_1$, $\boldsymbol{x}_2\in\mathcal{X}_2$, $y\in\mathcal{Y}$. 给定少量标记样本 $\mathcal{L}=\{(\langle\boldsymbol{x}_{1i},\boldsymbol{x}_{2i}\rangle,y_i)\}(i=1,\cdots,\ell)$ 和大量未标记样本 $\mathcal{U}=\{(\langle\boldsymbol{x}_{1j},\boldsymbol{x}_{2j}\rangle)\}(j=1,\cdots,u)$,协同训练期望通过利用大量未标记样本来帮助少量标记样本进行训练得到一个足够好的学习器,从而对未见示例进行分类.

TaLe 模型中"教"的置信度在文献[10]中明确定义为示例归属于特定类别的后验概率,即假设"教"者对于示例 $\boldsymbol{x}$ 的分类为 $c$,分类置信度 $Conf(\boldsymbol{x})=P(c|\boldsymbol{x})$. 那么应该如何定义"学"的需求度呢?

如图 2 所示的一些未标记样本分别用圆和星代表它们的真实标记,协同训练的中间分类器可能给出图 2(a)中较弱的一个分类界面. 显然地,这个分类器对于离分类界面较远的点具有较高的置信度,而对于离分类界面最近的两个圈出的点几乎是无法判别类别的. 如果"教"者此时能够对这两个点进行正确分类来增加训练集,那么这个学习器通过重新训练,可能会给出图 2(b)所示的较强分类界面;也就是说,对于图 2(a)的分类器来说,图中圈出的两个点是它当前最难区分的,也是对其学习能力的提高有关键帮助的.



(a) Before (b) Later

Fig. 2 Learner's need analysis with the circle and star as the ground-truth label.

图 2 学习器需求分析

由此可以定义当前学习器对于示例 $\boldsymbol{x}_i$ 的需求度为:

**定义 1.** 需求度. 学习器 $f=\boldsymbol{w}\times\boldsymbol{x}+b$ 对于样本 $\boldsymbol{x}_i$ 的需求度为 $Need_f(\boldsymbol{x}_i)=1-|\boldsymbol{w}\times\boldsymbol{x}_i|$;特别地,如果学习器 $f$ 是贝叶斯类分类器,则为 $Need_f(\boldsymbol{x}_i)=1-|0.5-Conf_f(\boldsymbol{x}_i)|$.

不难发现该想法借鉴于基于 margin 的主动学习[25]和结合主动学习的协同训练[23]. 值得指出的是,文献[23]研究的是将半监督学习和主动学习进行结合,在学习过程中需要用户额外提供真实的类别标记;而本文则只是考虑半监督学习,并由协同训练的学习器本身提供标记信息. 显然地,在获取同样数量标记样本的情况下,文献[23]的方法性能会更好,但是它需要额外消耗用户的时间、专业知识等资源.

### 2.2 CoSnT

考虑到协同训练的双方可能存在这样的差异性:学习器 $A$ 很难确信判断类别的点,学习器 $B$ 能够以较高的置信度加以标记;直观上来说这样的点更有助于学习器 $A$ 的学习,从而提高协同训练的收敛效率. 本文基于 TaLe 模型提出了一种新的协同训练算法(co-training with smart student and teacher, CoSnT). 在 CoSnT 中,智能"学"者知道自己当前对于数据池中每一个样本的需求程度,从而向"教"者提出标记请求,而智能"教"者能够正确判断对哪些点进行标记对"学"者学习性能的提高最有帮助.

由此 CoSnT 算法如图 3 所示,在协同训练的每一轮中,首先"学"者 $S$ 对于数据池 $\mathcal{U}'$ 中的所有样本

进行需求度计算,得到$\{Need_s(\boldsymbol{x}_j(j=1,\cdots,\mu)\}$来向"教"者 $T$ 提出标记请求;然后"教"者给出 $\mathcal{U}'$ 中所有样本分类的置信度$\{Conf_T(\boldsymbol{x})=P(0|\boldsymbol{x})|T(\boldsymbol{x})=0\}$或$\{Conf_T(\boldsymbol{x})=P(1|\boldsymbol{x})|T(\boldsymbol{x})=1\}$;进而"教"者对正点和负点分别进行"教"置信度排序和"学" 需求度排序,最后根据平均排序最优来选择未标记样本进行标记.值得指出的是该算法最后根据文献[10]的做法对两个视图上的分类器进行了简单的概率组合.

CoSnT algorithm.

1) Input:

$\mathcal{L}$——labeled training examples;

$\mathcal{U}$——unlabeled training examples;

$\mu,k$——pool size, number of co-training iterations;

$p,n$——# positive and # negative points chosen in each iteration.

2) Process:

Create a pool $\mathcal{U}'$ by choosing $\mu$ examples at random from $\mathcal{U}$

Loop for $k$ iterations:

① Use $\mathcal{L}_1=\{(\boldsymbol{x}_{1i},y_i)\}(i=1,\cdots,\ell)$ to train a classifier $h_1$;

② Use $\mathcal{L}_2=\{(\boldsymbol{x}_{2i},y_i)\}(i=1,\cdots,\ell)$ to train a classifier $h_2$;

③ Allow $h_1$ to rank examples from $\mathcal{U}'$ according to $Need_{h_1}$ and $Conf_{h_1}$;

④ $h_1$ chooses the first average ranked $p$ positive and $n$ negative to label, respectively;

⑤ Allow $h_2$ to do the similar teaching step as $h_1$;

⑥ Add these self-labeled examples to $\mathcal{L}$;

⑦ Randomly choose $2p+2n$ examples from $\mathcal{U}$ to replenish $\mathcal{U}'$.

3) Output: $h_1,h_2$ and combine $(h_1,h_2)$.

Fig. 3 Pseudo-code of CoSnT algorithm.

图 3 CoSnT 算法

类似地,CoSnT 算法在标记的过程中,"教"者会简单地对"学"者提供的需求度最高的那些点进行标记,不管它本身对应标记的置信度是否高.

## 3 实验结果

为了验证 CoSnT 算法的收敛效率,本文在多视图和单视图数据上分别进行了实验测试.

### 3.1 实验配置

本文主要将 CoSnT 应用于 1 个典型的双视图数据"*course*"[10],以及 3 个随机选出的 UCI 数据集上[26]:"*mushroom*","*credit-a*","*cylinder-bands*".其中"*course*"是 1 个网页数据集,每个网页由 2 个视图表示:"*pages*"和"*links*"分别是网页本身的内容和网页对应超链接的内容.这个数据集一共包含 1 051 个网页,其中 230 个是正样本;根据正负样本的比例,实验设定初始标记样本为 3 个正样本和 9 个负样本,并设定图 3 算法中 $p=1,n=3$. UCI 数据集主要用于单视图数据的实验测试,3 个数据集包含的样本总数和正样本数目分别为 8124/3916,690/241,540/312.根据正负样本的比例,设定初始标记样本均为 3 个正样本和 3 个负样本,并设定 $p=1,n=1$.对每个数据集,随机选择 25% 的数据作为测试数据,然后从剩下的 75%中随机选择预设数目的正点和负点作为初始标记数据加入到 $\mathcal{L}$ 中,剩下的点全部加入到未标记数据集 $\mathcal{U}$ 中,并以 30 次重复实验在测试数据上的平均错误率作为评价标准.同时实验设定数据池的大小 $\mu=75$,协同训练的轮数 $k=50$;协同训练基分类器主要运用了 WEKA[27] 中的 NaiveBayes 和 RBFNetwork.所有实验配置主要依据标准协同训练算法[4].

### 3.2 多视图数据

在多视图数据集"*course*"上以基分类器 NaiveBayes 进行了实验测试,比较了标准协同训练算法 CoTrain,CoS 和 CoSnT 的性能.

如图 4(a)所示,通过观察每个视图上的分类器和它们的组合分类器在测试数据集上平均错误率,随着协同训练轮数增加的变化情况,不难发现 CoS 算法由于最初学习器较弱,加入的错误标记过多导致了性能的恶化;但是 CoSnT 因为同时考虑了协同训练双方的需求度和置信度,对比于标准协同训练算法在收敛效率和泛化性能上都有所提高.

### 3.3 单视图数据

在单视图数据上,以 WEKA[27] 中属于不同系列的 NaiveBayes 和 RBFNetwork 作为基分类器进行协同训练.实验结果如图 4(b)~(d)所示,不难发现 CoSnT 在每个数据集上收敛效率和泛化性能的优越性.

值得指出的是,在"*cylinder-bands*"数据集上,当标准协同训练方法出现性能持平或下降时,CoSnT 能够缓解这种性能下降的趋势甚至使得性能增强.同时发现 CoS 在其余两个数据集上的性能并没有强烈的恶化趋势,在"*cylinder-bands*"上的性能甚至要优于标准协同训练算法.出现这一现象的可能原因是最初两个弱分类器已经满足了这样的差异性:一个分类器无法判断类别的点,另一个分类器有着较高的分类置信度.但是没有置信度的约束,同样会导致加入更多的错误标记使得学习性能弱于 CoSnT 算法.

另一方面需要指出的是,基于分类器本身的学习性能会影响协同训练在不同数据集上的学习性能;同时这也是协同训练的优势所在,可以根据应用的不同需求来选择合适的基分类器.本文 CoSnT 算法的优势在于无论选择何种基分类器均能通过利用未标记样本显著提升学习性能.

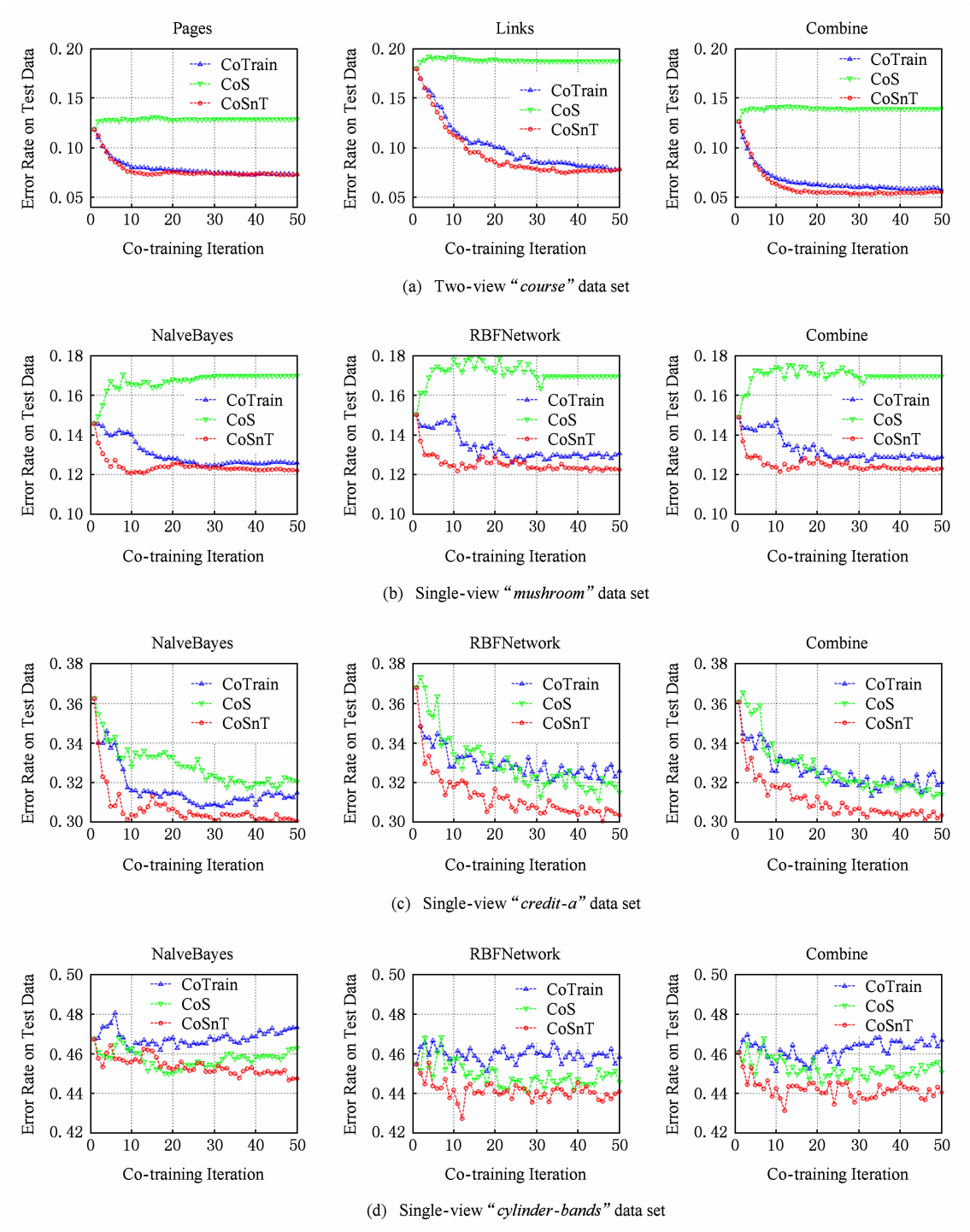



Fig. 4 Prediction error rate comparison.

图 4 协同训练双方和它们组合分类器的预测平均错误率比较

## 4 总 结

本文受到真实教学场景的启发,针对协同训练提出了一种新的教学模型 TaLe,并基于此模型利用协同训练双方的特定差异:"学"者无法确信判断类别的点,"教"者有着较高的分类置信度,提出了一种新的协同训练算法 CoSnT. CoSnT 不同于以往的协

同训练算法,在协同训练的过程中不仅考虑了“教”者的置信度,还考虑了“学”者的需求度,从而使得学习更加高效.在双视图和单视图数据集上的实验结果表明 CoSnT 算法在收敛效率和泛化性能上都要优于标准协同训练算法.

值得指出的是,由学习器本身提供的标记信息未必准确,因此需要考虑这些样本的可靠性,这是未来需要进一步研究的工作.

## 参 考 文 献


[1] Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning [M]. Cambridge, MA: MIT Press, 2006

[2] Li Yufeng, Huang Shengjun, Zhou Zhihua. Regularized semi-supervised multi-label learning [J]. Journal of Computer Research and Development, 2012, 49(6): 1272-1278 (in Chinese)
(李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习算法[J]. 计算机研究与发展, 2012, 49(6): 1272-1278)

[3] Shahshahani B, Landgrebe D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon [J]. IEEE Trans on Geo-Science and Remote Sensing, 1994, 32(5): 1087-1095

[4] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labelled and unlabelled data [C] // Advances in Neural Information Processing System 9. Cambridge, MA: MIT Press, 1997: 571-577

[5] Joachims T. Transductive inference for text classification using support vector machines [G] //Bratko I, Dzerski S, eds. Proc of ICML'99. San Francisco: Morgan Kaufmann, 1999: 200-209

[6] Chapelle O, Zien A. Semi-supervised learning by low density separation [C/OL] //Cowell R G, Ghahramani Z, eds. Proc of AISTATS'05. 2005 [2012-02-01]. http://www.gatsby.ucl.ac.uk/aistats/aistats2005_eproc.pdf

[7] Zhu J, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions [G] //Fawcett T, Mishra N, eds. Proc of ICML'03. Menlo Park: AAAI, 2003: 912-919

[8] Zhou Dengyong, Bousquet O, Lal T N, et al. Learning with local and global consistency [G] //Thrun S, Saul L, Scholkopf B, eds. Advances in Neural Information Processing System 16. Cambridge, MA: MIT Press, 2004: 321-328

[9] Zhou Zhihua, Li Ming. Semi-supervised learning by disagreement [J]. Knowledge and Information Systems, 2010, 24(3): 232-257

[10] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training [C] //Bartlett P L, Mansour Y, eds. Proc of COLT'98. New York: ACM, 1998: 92-100

[11] Zhou Zhihua. Co-training paradigm of semi-supervised learning [G] //Zhou Zhihua, Wang Jue, eds. Machine Learning and Its Application. Beijing: Tsinghua University Press, 2007: 259-275 (in Chinese)
(周志华. 半监督学习中的协同训练风范[G] //周志华, 王珏. 机器学习及其应用. 北京: 清华大学出版社, 2007: 259-275)

[12] Dasgupta S, Littman M, McAllester D. PAC generalization bounds for co-training [G] //Dietterich T G, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing System 14. Cambridge, MA: MIT Press, 2002: 375-382

[13] Balcan M F, Blum A, Yang Ke. Co-training and expansion: Towards bridging theory and practice [G] //Saul L K, Weiss Y, eds. Advances in Neural Information Processing System 17. Cambridge, MA: MIT Press, 2005: 89-96

[14] Wang Wei, Zhou Zhihua. Analyzing co-training style algorithms [G] //LNCS 4701: Proc of ECML'07. Berlin: Springer, 2007: 454-465

[15] Wang Wei, Zhou Zhihua. A new analysis of co-training [G] //Getoor L, Scheffer T, eds. Proc of ICML'11. New York: ACM, 2010: 1135-1142

[16] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training [C] //Proc of ACM CIKM'00. New York: ACM, 2000: 86-93

[17] Goldman S, Zhou Yan. Enhancing supervised learning with unlabeled data [G] //Langley P, eds. Proc of ICML'00. San Francisco: Morgan Kaufmann, 2000: 327-334

[18] Hwa R, Osborne M, Sarkar A, et al. Corrected co-training for statistical parsers [C/OL] //Proc of ICML'03 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. 2003 [2012-02-01]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.3124

[19] Zhou Zhihua, Li Ming. Semi-supervised regression with co-training [G] //Kaelbling L P, Saffiotti A, eds. Proc of IJCAI'05. Denver, CO: Professional Book Center, 2005: 908-913

[20] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[21] Pierce D, Cardie C. Limitations of co-training for natural language learning from large data sets [G/OL] //Lee L, Harman D, eds. Proc of EMNLP'01. 2001 [2012-02-01]. http://www.cs.cornell.edu/home/llee/emnlp/proceedings.html

[22] Mavroeidis D, Chaidos K, Pirillos S, et al. Using tritraining and support vector machines for addressing the ECML-PKDD 2006 Discovery Challenge [G/OL] //Bickel S, eds. Proc of ECML-PKDD'06 Discovery Challenge Workshop. 2006: 39-47. [2012-02-01]. http://www.ecmlpkdd2006.org/discovery_challenge_proceedings.pdf

[23] Zhou Zhihua, Chen Kejia, Dai Hongbin. Enhancing relevance feedback in image retrieval using unlabeled data [J]. ACM Trans on Information Systems, 2006, 24(2): 219-244

[24] Guo Qi, Chen Tainshi, Chen Yunji, et al. Effective and efficient microprocessor design space exploration using unlabeled design configurations [G] //Walsh T, eds. Proc of IJCAI'11. Menlo Park: AAAI, 2011: 1671-1677

[25] Balcan M F, Broder A, Zhang T. Margin based active learning [G] //LNCS 4539: Proc of COLT'07. Berlin: Springer, 2007: 35-50

[26] Asuncion A, Newman D J. UCI machine learning repository [OL]. [2012-02-01]. http://archive.ics.uci.edu/ml/datasets.html

[27] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques [M]. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005

**Hu Juhua**, born in 1986. Master. Student member of China Computer Federation. Her main research interests include machine learning and data mining, especially learning with labeled and unlabeled examples.

**Jiang Yuan**, born in 1976. Professor. Member of China Computer Federation. Her main research interests include artificial intelligence, machine learning, data mining, information retrieval and so on.

**Zhou Zhihua**, born in 1973. Professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include artificial intelligence, machine learning, data mining, pattern recognition and so on(zhouzh@nju.edu.cn).