基于随机子空间的半监督协同训练算法

王 娇,罗四维,曾宪华
(北京交通大学计算机与信息技术学院,北京 100044)

摘 要: 半监督学习是近年来的一个研究热点. 协同训练(co-training)是利用未标记数据来提高传统监督学习性能的一种半监督学习范式. 本文提出一种基于随机子空间的协同训练算法(RAndom Subspace CO-training,简称为 RASCO). 该算法探讨多视图的协同训练. 用随机判别的理论分析了算法的分类精度和泛化能力. 讨论了随机子空间的维数和个数对分类性能的影响. 在 UCI 数据集上的实验结果表明,与其它同类算法相比,RASCO 算法有较好的性能.

关键词: 半监督学习; 随机子空间; 随机判别; 协同训练; 多视图; RASCO
中图分类号: TP181 **文献标识码**: A **文章编号**: 0372-2112 (2008) 12A-060-06

A Random Subspace Method for Co-Training

WANG Jiao, LUO Si-wei, ZENG Xian-hua
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Semi-supervised learning has received much attention recently. Co-training is a kind of semi-supervised learning method which uses unlabeled data to boost the performance of standard supervised learning algorithms. A novel co-training style algorithm, RASCO (for RAndom Subspace CO-training), is proposed which uses stochastic discrimination theory to extend co-training to multi-view situation. The accuracy and generalizability of RASCO are analyzed. The influences of the parameters of RASCO are discussed. Experiments on UCI data set demonstrate that RASCO is more effective than other co-training style algorithms.

Key words: semi-supervised learning; random subspace; stochastic discrimination; co-training; multi-view; random subspace co-training (RASCO) algorithm

1 引言

传统的机器学习中,为了训练一个分类器,往往需要大量的训练样本. 而在很多实际应用中(如垃圾邮件处理、网页分类等),一方面标记数据(labeled data,有类别信息的数据)数量较少,或获取标记数据的代价昂贵;另一方面有大量的未标记数据(unlabeled data,没有类别信息的数据)可供使用. 半监督学习旨在构建综合利用标记和未标记数据的学习方法,在只有较少标记数据的条件下得到较高的分类精度. 由于半监督学习在数据挖掘、模式识别等众多领域都有着广泛的应用而成为近年来的一个研究热点[1,2].

现有的半监督学习方法包括,基于图的方法(graph-based methods)[3,4],低密度区域分割(low density separation)[5,6],生成式模型(generative model)[7,8]等. Blum 和 Mitchell 提出的协同训练(co-training)[9]算法是另一种半监督学习范式,该算法利用未标记数据来提高传统监督学习的性能. 本文主要在协同训练框架下研究问题.

最初的协同训练算法由 Blum 和 Mitchell[9]提出(为了区别,称为标准协同训练算法),他们假设数据集有两个充分冗余(sufficient and redundant)的视图(view),即,每个属性集都足以训练一个学习器,且给定标记时,每个属性集都条件独立于另一个属性集. 例如,网页分类问题中,每个网页可以由两个视图表示,其中,网页本身包含的信息构成一个视图,指向该网页的超级链接上的信息构成另一个视图. 标准协同训练算法在这两个视图上,利用标记数据分别训练一个分类器,并对未标记数据进行预测. 然后从每个分类器的预测结果中,挑选若干置信度较高的数据,加入另一个分类器的训练集,以便对方用扩大的训练集来进行更新.

Blum 和 Mitchell[9]证明,在充分冗余视图这一条件成立时,标准协同训练算法可以有效地通过利用未标记数据提升学习器的性能. Dasgupta 等人[10]证明,如果标准协同训练中,两个视图间的条件独立性假设成立,那么通过最大化两个分类器对未标记数据预测的一致性,将得到其最小的泛化误差. Balcan 等人[11]进一步研究发

收稿日期:2008-05-08;修回日期:2008-09-19
基金项目:教育部博士点基金(No. 20050004001,No. 200800041049);国家 863 高技术研究发展计划(No. 2007AA01Z168);国家自然科学基金(No. 60773016,No. 60805041,No. 60872082)

现,只要数据分布满足比上述假设弱得多的"扩张性"(expansion)假设,协同训练算法就可以奏效,并分析了协同训练算法的 PAC 可学习性.

Goldman 和 Zhou[12]用两种不同的监督学习方法,从同一个属性集上学习两个不同的分类器,每个分类器都可以把数据空间划分成多个等价类,然后用交叉验证对未标记数据进行标记,并用交叉验证综合两种学习方法形成最终预测. 由于大量使用交叉验证,所以算法具有较高的时间复杂度. Zhou 和 Li[13]提出的 tri-training 算法使用重采样技术(bootstrap sampling),在不同数据集上训练三个分类器,进行协同训练,如果其中任意两个分类器对某个未标记数据的预测相同,则把这个未标记数据、以及对它所属类别的预测,加入到第三个分类器的训练集中. Li 和 Zhou[14]对 tri-training 进行了扩展,提出了可以更换发挥集成学习作用的 Co-Forest 算法. 总的来说,上述算法只使用一个视图,通过不同的监督学习方法或重采样技术,来训练多个分类器,以实现协同训练.

另一些研究者着力于对两视图的协同训练算法进行改进. Nigam 和 Ghani[15]在两个视图上分别用 EM 算法进行类别预测,得到了比标准协同训练算法更好的效果. Brefeld 和 Scheffer[16]在两个视图上分别把 EM 和 SVM 结合起来进行文本分类. Collins 和 Singer[17];以及 Abney[18],通过最小化两个视图对未标记数据预测的不一致,来提升学习算法的性能. Ando and Zhang[19];Johnson and Zhang[20]则用未标记数据学习得到新的特征表示,再用新的特征表示进行两视图的监督学习. 而 Muslea 和 Minton[21]把两个视图学习的思想引入到主动学习之中,取得了很好的效果. 周志华[22]指出,对多视图学习的研究,是今后的重要研究内容.

我们认为,标准协同训练算法中的两个视图,在一些情况下不能满足要求. 比如说,我们知道在机械制图中,完全存在两个视图全都一样的不同的物体,也就是说,对于三维空间,三个特征,任取两个学习认为一致的,可能实际上不一样,这说明了——整体不等于部分之和. 并且,两个视图的协同训练算法通常假设视图类条件独立,这在实际问题中也很难得到保证. 为了解决这两个问题,本文提出一种基于随机子空间的协同训练算法(RAndom Subspace CO-training,简称为 RASCO). 这是第一个把两视图推广到多视图. 其基本思想是,在数据的多个不同视图上分别训练的分类器,能够优势互补,因而可以利用其它分类器的信息,排除自己的不确定性,即通过其它分类器置信的数据,扩大自己的训练集,从而改善各个视图上的分类器性能. 本文用随机判别的理论分析了 RASCO 的分类精度和泛化能力. 讨论了随机子空间的维数和个数对分类性能的影响. 在 UCI 数据集上的实验结果表明,与其它同类算法相比,RASCO 在高维数据上有更好的性能.

2　RASCO 算法

问题的描述:给定标记数据集 $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$,和未标记数据集 $U=\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$,其中 x_i 是 n 维向量,y_i 是它的标记,l 和 u 分别是标记数据集和未标记数据集包含的数据个数. 当 L 较小时,用监督学习算法得到的分类器有时不能令人满意. 如何根据 U 中所蕴涵的信息,来提升学习性能?半监督协同训练的基本思想是:首先用标记数据集 L 和传统监督学习方法,训练多个不同的分类器;然后用每个分类器对未标记数据 U 的预测,扩大其它分类器的训练集 L. 此过程中的关键问题是,如何用 L 构建多个不同的分类器,以及选择哪些未标记数据来扩大训练集.

RASCO 在 L 特征空间的随机子空间中训练分类器,即用数据的不同视图训练得到不同的分类器. 这样构建的不同分类器对不同特征敏感,能够优势互补. 然后用每个分类器对未标记数据进行预测,集成预测结果,选取其中最置信的数据来扩大训练集,如此迭代,从而达到协同训练的目的.

具体地说,为了构建不同的分类器,RASCO 首先把 L 投影到特征空间的随机子空间中. 投影方法如下:假设原数据特征空间为 n 维,随机子空间为 m 维,满足 $m<n$. 设标记数据集有 l 个数据,即 $|L|=l$. 对任意 $p\in L$,可写成 $p=(p_1,p_2,\cdots,p_n)$,将 p 投影到这 m 维张成的空间中,得到的向量可写成 $p_{sub}=(p_{s1},p_{s2},\cdots,p_{sm})$. 由所有 l 个 p_{sub} 组成的向量集合 L_{sub},就是标记数据集 L 在其 m 维随机子空间中的投影. 重复此过程 K 次,得到数据特征空间的 K 个不同视图,$L_{sub_k}(1\leq k\leq K)$.

然后,在每个视图 L_{sub} 上,用传统监督学习方法(如决策树)训练一个分类器 C_{sub}. 这样构建 K 个对不同特征敏感的分类器 $C_{sub_k}(1\leq k\leq K)$.

协同训练是一个迭代的过程,在每次迭代 $j(j=1,2,\cdots)$ 中,为了选择合适的未标记数据来扩大训练集,RASCO 用各个视图上的分类器对未标记数据进行预测,把 K 个预测集成起来,并把集成结果中最置信的数据,作为候选数据.

假设是两类的分类问题(多类的分类问题可做相似的分析),其类别用 P(positive)和 N(negative)表示. 对于任意的 $q\in U$,随机子空间中的分类器对其预测用 $C_{sub_k}(q)$ 表示. $C_{sub_k}(q)=1$ 代表第 k 个分类器对数据点 q 的预测结果是 P 类;$C_{sub_k}(q)=-1$ 代表第 k 个分类器对数据点 q 的预测结果是 N 类. 集成的方法是把 K 个

预测结果取均值, $C=\left(\sum_{k=1}^{K}C_{sub_k}\right)/K$,作为集成分类器的预测. $C(q)\geq 0$,则 q 被判定为 P 类;反之则为 N 类. 把未标记数据中,判定为 P 类的数据置入集合 V_P;判定为 N 类的数据置入集合 V_N. 从集合 V_P 和 V_N 中,分别选取 $|C(q)|$ 值最大的前 θ 个数据,作为最置信的候选数据 U_j.

候选数据集中可能含有被错误标记的数据,直接用其扩大训练集,会因噪声而影响后续分类器的训练. 考虑数据剪辑技术[23],如基于最近邻规则的 Deputation 技术. 对 U_j 中的每个数据,首先按最近邻规则从 $L\cup U_j$ 中选取它的 x 个近邻,然后观察其中是否有 x 个近邻的标记相同,如果有,则将该数据置入集合 L_j. 按文献[23]所述,将 x 和 x 设为 3 和 2. 注意保持类别比例与原数据集一致. 在每次迭代过程 j 中,用扩大的训练集 $(L+L_j)$ 重新训练视图上的分类器. RASCO 算法的主要步骤描述如下:

RASCO 算法步骤

输入:初始标记数据集 L,未标记数据集 U,子空间维数 m,子空间个数 K,监督学习算法 Learn

输出:假设 $C=\left(\sum_{k=1}^{K}C_{sub_k}\right)/K$

(1) 对原数据特征空间,取 K 个 m 维的随机子空间;

(2) 把 L 中的数据分别投影到每个子空间上,得到 $L_{sub_k}(1\leq k\leq K)$;

(3) 在每个子空间上,用 L_{sub_k} 训练分类器 C_{sub_k};

(4) 用每个 C_{sub_k} 对 U 中的数据进行分类,集成分类结果 $C=\left(\sum_{k=1}^{K}C_{sub_k}\right)/K$;

(5) 对任意的 $q\in U$,若 $C(q)\geq 0$ 则把 q 置入集合 V_P,否则置入集合 V_N;

(6) 从 V_P 和 V_N 中,分别选取 $|C(q)|$ 值最大的前 θ 个数据,置入候选集合 U_j;

(7) 对候选集合中的数据进行剪辑,令 $L_j\leftarrow Deputation(U_j)$;

(8) 更新标记数据集,令 $L=L\cup L_j$,$L_j\leftarrow\phi$;

(9) 判断是否满足结束条件,满足则退出;否则转步骤 2.

3 算法分析

随机判别理论[24,25]是 RASCO 算法的理论出发点. 在文献[24]中,用特征空间中点集的决策区域来代表分类器,通过离散随机过程,证明了这些弱分类器可以组合为强分类器. 在文献[25]中,证明了随机判别可以避免过拟合. 文献[26]是随机判别理论的算法实现,文献[27]是其扩展. 本文着重分析 RASCO 算法各随机子空间中的分类器协同训练后的分类精度和泛化能力.

将分类器看成泛函空间中的点,我们先定义一些变量,然后证明满足一定条件的分类器对未标记数据中两个类别的数据有不同的数学期望,这样就可以通过数学期望区别两类数据.

首先设未标记数据集 U 中,P 类数据有 a 个,N 类数据有 b 个,$\forall q\in U$,对分类器 C,定义两个衡量其性能的实值函数:

$$d_P(C)=|\{q:q\text{ 为 }P\text{ 类,且 }c(q)=1\}|/a \quad (1)$$

$$d_N(C)=|\{q:q\text{ 为 }N\text{ 类,且 }c(q)=1\}|/b \quad (2)$$

我们把 $d_P(C)$ 称为正分率,$d_N(C)$ 称为误分率. 最好的情况是

$$d_P(C)=1,\ d_N(C)=0 \quad (3)$$

这时,分类器 C 把未标记数据中所有的 P 类数据判定为 P 类,没有遗漏,把所有的 N 类数据拒识为 P 类,没有错误. $\forall\varepsilon>0$,如果有

$$d_P(C)-d_N(C)>1/\varepsilon \quad (4)$$

我们称分类器 C 为弱分类器,即,其正分率比误分率略高. 定义变量

$$Y(q)=\frac{C(q)}{d_P(C)} \quad (5)$$

定理 1 如果 q 为 P 类,则 $E(Y(q))=1$;如果 q 为 N 类,则 $E(Y(q))<1-(1/\varepsilon)$.

定理 1 的证明参见文献[24]. 命题表明,在两类问题中,对满足式(4)的分类器,P 类数据和 N 类数据有不同的数学期望,这样我们就能通过数学期望对两类数据进行分类.

RASCO 算法在数据特征空间中取随机子空间,在不同的子空间中训练得到不同的分类器. 令 $Z_K(q)=\frac{1}{K}\sum_{i}^{K}Y_i(q)$,其中 $i=1,2,\cdots,K$,K 为子空间个数. 当 K 足够大时,用分类器的均值代替数学期望,得到的分类正确率趋近于 1,即对 $\forall\gamma>0$ 和 $\forall\delta>0$,有:

$$\Pr\left\{|Z_K(q)-E(Y(q))|<\frac{1}{\gamma}\right\}>1-\frac{1}{\delta} \quad (6)$$

注 1:上面证明了满足一定条件的分类器通过协同训练,得到的协同训练分类器的分类精度理论上可以达到任意精度.

注 2:上述结论要满足的一定条件是指各协同训练的分类器要满足式(4). 而决策树、神经网络、或是贝叶斯一般均能保证其正分率比误分率略高,这也使 RASCO 算法在子空间中选择用于训练的分类器时有较多的选择.

注 3:对于分类器个数 K 的取值,如果需要取较大的 K,为了时间上的可容忍性,可以在多台工作站上并行训练分类器,但在下面的实验部分,我们将看到,不需要太大的 K,就能得到比其它同类算法更好的结果.

接下来分析 RASCO 的泛化能力,首先引入文献

[24]中的对偶引理,用 Q 表示所有正例(或反例)的集合,用 S 表示所有满足式(4)的分类器.

定理 2　固定集合 Q 中的数据点 q,用 F_q 表示$\{q\}\times S$;固定集合 S 中的分类器 C,用 G_C 表示 $Q\times\{C\}$. 则 F_q 和 G_C 有相同的概率密度函数.

在 RASCO 算法中,用 $(u_1, u_2, \cdots, u_n)$ 表示函数 $d_P(C)$ 的取值集合;用 $(v_1, v_2, \cdots, v_n)$ 表示函数 $d_N(C)$ 的取值集合. 对于任何正例数据 q(不管它是在标记数据集 L 中,还是在未标记数据集 U 中),F_q 的概率密度函数仅依赖于集合$\{u_i\}$. 因此,通过定理 2 可以得到,定义在 L 中正例数据上的随机变量 G_C,与定义在 U 中正例数据上的随机变量 G_C,有相同的概率密度函数. 所以 $d_P(C)$ 在 L 和 U 上有相同的取值. 对于反例数据和集合$\{v_i\}$,可以用相同的方法分析出,$d_N(C)$ 在 L 和 U 上有相同的取值. 综合上述分析得出,如果用正分率 $d_P(C)$ 和误分率 $d_N(C)$ 来衡量分类器 C,则它在 L 和 U 上有相同的性能.

4　实验与讨论

在实验部分,我们先观察 RASCO 的两个参数(m 和 K)对分类性能的影响,并给出参数的选取策略. 然后把 RASCO 和其它同类算法进行比较.

实验所用数据集来源于 UCI 机器学习公用数据[28]. 数据集的简单描述如表 1. 对于每个数据集,把其 25 %作为测试集 TE,剩余的 75 %作为训练集 TR. 再把 TR 按比例分为标记数据集 L、未标记数据集 U. 注意在此过程中保持各个集合中的数据类别比例不变. 由于数据集的这种分割是随机的,每次产生的 TE、L 和 U 都会不同,所以我们的每个实验在不同的 TE、L 和 U 下运行十次,取十次运行结果的平均值. 分类器使用决策树 J4.8 训练,基于 J4.8 的效率较高.

表 1　实验数据集

data set	no. of samples	no. of features	no. of classed
diabetes	768	8	2
tic-tac-toe	958	9	2
ionosphere	351	34	2
kr-vs-kp	3196	36	2
splice	3190	60	3

4.1　子空间维数 m 的选取

在 diabetes 数据集上的实验结果如图 1,实验所用标记数据的数量在训练集中占 60 %,子空间个数为 30,迭代次数为 10 次. diabetes 数据集的维数为 8,从图 1 中可以看出,错误率在 $m=4$ 时最小,为 20.6 %. 分析其原因是,当 m 取较小的值时,子空间维数太小,每个分类器得到的信息较少,不利用训练分类器;反之,当 m 取较大的值时,子空间的差异度太小,不利于分类器间的协同训练. 在其它数据集上也有相似的现象.

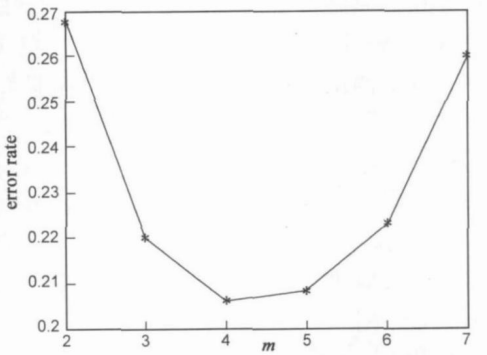

图1　子空间维数m与错误率e的关系曲线

给定数据特征空间维数为 n,我们给出子空间维数 m 的选取策略为:$m=[n/2]$. 这是基于这样选取的组合数最大,即,$C(n, m)$ 当 $n=2m$ 时有最大值. 这意味着,$n=2m$ 时,我们可以构建最多的子空间,训练最多的分类器. 即使我们在实际应用中,不全取这些子空间,但子空间越多,我们有越丰富的选择. 上述实验结果也与这种维数选取策略相符.

4.2　子空间个数 K 对分类精度的影响

在 diabetes 数据集上的实验结果如图 2,实验所用标记数据的数量在训练集中占 60 %,子空间维数 $m=4$,迭代次数为 10 次. 在 8 维的原特征空间中,取 4 维的随机子空间,可能的取法有 $C_8^4=70$ 种,即,可以选取 70 个不同的随机子空间,以训练 70 个不同的分类器. 从图 2 种可以看出,K 从 10 增加到 30 的过程中,错误率下降得最快;而从 30 增加到 70 的过程中,错误率的下降非常缓慢. 这是由于,对 8 维的 diabetes 数据,在其 4 维的随机子空间中训练分类器,30 个分类器的集成,已经能达到较大的预测精度,此时再增加集成分类器的个数,对性能提升的影响不大. 正如在本文第三部分中指出的,不需要取太大的 K,就能达到足够的预测精度.

需要指出的是,虽然 RASCO 集成了 K 个分类器形成最终判别,但与文献[29,30]中的集成半监督学习方

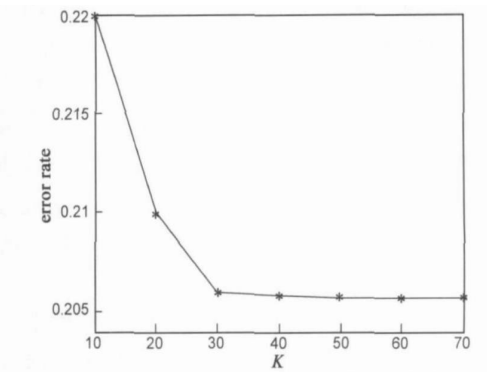

图2　子空间个数K与错误率e的关系曲线

法有本质的不同. 文献[29,30]中的方法先给未标记数据分配一个虚拟的类别标记,然后通过最优化的方法,用标记数据和伪标记数据共同建造集成分类器中的下一个基分类器,强调集成分类器的构造. 与之相比,RASCO 在特征空间的随机子空间中,用传统监督学习方法构建基分类器,算法注重基分类器的协同训练.

4.3 与其它同类算法比较结果

每个实验都用 J4.8 在不同的 TE、L 和 U 的划分下运行十次,取十次结果的平均值. 其中 η 代表标记数据在训练集中所占的比例. 在三个数据集上分别测试 η 为 20 %、40 %、60 % 时算法的错误率 e. 表中,CO 代表 Blum[9] 中提出的 co-training 算法,我们把每个数据集的特征集分成两个视图,以实现 Blum[9] 中的算法. TRI 表示 Zhou[13] 中提出的 tri-training 的算法. Co-Forest 表示 Li[14] 中提出的算法. RASCO 是本文提出的随机子空间上的协同训练算法,其子空间维数 m 的选取按本文实验一中介绍的原则,在 diabetes 上取 4,ionosphere 上取 20,tic-tac-toe 上取 5,集成分类器个数 K 取 30. co-traing 和 RASCO 的迭代次数都为 10 次,得到的结果如表 2 所示.

表 2 基于协同训练的半监督学习算法错误率 e 比较

数据集	η(%)	测试集上的错误率 e(%)			
		CO	TRI	Co-Forest	RASCO
diabetes	20	28.6	28.8	26.1	21.8
	40	27.7	25.2	26.4	21.1
	60	27.2	27.3	25.5	20.6
tic-tac-toe	20	28.5	25.8	27.7	27.4
	40	24.9	16.3	21.7	20.9
	60	17.1	13.3	13.8	12.0
ionosphere	20	15.3	12.1	9.2	11.2
	40	12.8	8.7	7.9	9.5
	60	12.3	8.7	7.5	7.1
kr-vs-kp	20	6.5	2.5	3.5	1.2
	40	5.5	1.8	2.3	0.9
	60	4.7	1.0	1.9	0.7
splice	20	11.3	0.4	9.6	8.5
	40	10.8	8.4	8.5	7.6
	60	9.11	8.2	7.6	6.2

从表 2 中可以看出,在 tic-tac-toe 数据集,20 % 和 40 % 的标记率下,RASCO 的错误率比 tri-training 稍大;在 ionosphere 数据集,20 % 和 40 % 的标记率下,RASCO 的错误率比 Co-Forest 稍大;在其它情况下,RASCO 的错误率均小于 co-training、tri-training 和 Co-Forest 算法. 在总共 5 个数据集的 15 组测试中,RASCO 在其中的 11 组上比其他算法有更小的错误率,说明了在随机子空间中训练的分类器协同训练的有效性.

另外,由于 RASCO 在原特征空间的子空间中训练分类器,当数据集的维数相对于标记数据个数而言比较大时,不充足的训练样本影响学习性能,而取子空间能缓解这个矛盾. 从表 2 也能观察到这一现象,在 kr-vs-kp 和 splice 这两个较高维的数据集上,在 20 %、40 % 和 60 % 这三个标记率下,其错误率 e 都比其它同类算法更低. 这个特点使 RASCO 比其它同类算法更适用于高维数据集的处理.

5 总结与展望

半监督学习旨在利用未标记数据所蕴含的信息,辅助传统的监督学习,以减少对标记数据的需求. 协同训练是半监督学习中的一种重要范式. 本文提出一种基于随机子空间的协同训练算法——RASCO. 据我们所知,该算法是第一个多视图的半监督协同训练算法. RASCO 在数据特征空间的随机子空间中训练分类器,用一个分类器最置信的数据扩大其它分类器的训练集,以此提高分类器性能. 本文用随机判别的理论分析了 RASCO 的分类精度和泛化能力. 讨论了随机子空间的维数和个数对分类性能的影响. 在 UCI 数据集上的实验结果表明,与其它同类算法相比,RASCO 有更好的性能.

下一步的工作中,将把算法应用于数据挖掘和模式识别的一些实际应用中,如网页分类等. 另外,将算法与主动学习结合,用更少的标记数据来训练分类器,也是我们进一步的研究方向.

参考文献:

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

[2] Zhu Xiao-Jin. Semi-supervised Learning with Graphs [D]. Carnegie Mellon University, doctoral thesis, 2005.

[3] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts [A]. Proceedings of the 18th International Conference on Machine Learning [C]. Williamston, MA, 2001. 19 - 26.

[4] Szummer M, Jaakkola T. Partially labeled classification with markov random walks [A]. Advances in Neural Information Processing Systems 14 [C]. Cambridge, MA: MIT Press, 2002. 945 - 952.

[5] Joachims T. Transductive inference for text classification using support vector machines [A]. Proceedings of the 16th International Conference on Machine Learning [C]. New York, USA, 1999. 200 - 209.

[6] Tong S, Koller D. Support vector machine active learning with applications to text classification [A]. Proceedings of the 17th International Conference on Machine Learning [C]. Stanford, US, 2000. 999 - 1006.

[7] Nigam K, McCallum A K, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM [J]. Ma-

chine Learning,2000,39(2 - 3) :103 - 134.

[8] Cozman F G,Cohen I,Cirelo M C. Semi-supervised learning of mixture model[A]. Proceedings of the 20th International Conference on Machine Learning[C]. citeseer,2003. 99 - 106.

[9] Blum A,Mitchell T. Combining labeled and unlabeled data with co-training[A]. Proceedings of the 11th Annual Conference on Computational Learning Theory[C]. Madison,WI,1998. 92 - 100.

[10] Dasgupta S,Littman M,McAllester D. PAC generalization bounds for co-training[A]. Advances in Neural Information Processing Systems 14[C]. Cambridge,MA,MIT Press,2002. 375 - 382.

[11] Balcan M F,Blum A. A PAC-style model for learning from labeled and unlabeled data[A]. Proceedings of the 18th Annual Conference on Computational Learning Theory[C]. citeseer,2005. 111 - 126.

[12] Goldman S,Zhou Y. Enhancing supervised learning with unlabeled data[A]. Proceedings of the 17th International Conference on Machine Learning[C]. San Francisco,CA,2000. 327 - 334.

[13] Zhou Z H,Li M. Tri-training:exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(11) :1529 - 1541.

[14] Li M,Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems,Man and Cybernetics-Part A,2007,37(6) :1088 - 1098.

[15] Nigam K,Ghani R. Analyzing the effectiveness and applicability of co-training[A]. Proceedings of Information and Knowledge management[C]. New York,NY,USA:ACM,2000. 86 - 93.

[16] Breiman U,Scheffer T. Co-EM support vector learning[A]. Proceedings of the 21st International Conference on Machine Learning[C]. citeseer,2004. 121 - 128.

[17] Collins M,Singer Y. Unsupervised models for named entity classifications[A]. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora[C]. College Park,MD,1999. 100 - 110.

[18] Abney S. Bootstrapping[A]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C]. Philadelphia,PA,2002. 360 - 367.

[19] Ando R,Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data[J]. Journal of Machine Learning Research,2005,6,1817 - 1853.

[20] Johnson R,Zhang T. Two-view feature generation model for semi-supervised learning[A]. Proceedings of the 24th International Conference on Machine Learning[C]. New York,NY,USA:ACM,2007. 25 - 32.

[21] Muslea I,Minton S,Knoblock C A. Active learning with multiple views[J]. Journal of Artificial Intelligence Research,2006,(27) :203 - 233.

[22] 周志华. 半监督学习中的协同训练风范[A]. 周志华,王珏 主编. 机器学习及其应用[M]. 2007,北京:清华大学出版社,2007. 259 - 275.

[23] 邓超,郭茅祖. 基于自适应数据剪辑策略的 Tri-training 算法[J]. 计算机学报,2007,30(8) :1213 - 1226.
Deng Chao,Guo Mao-zu. ADE-Tri-training:Tri-training with adaptive data editing[J]. Chinese Journal of Computers,2007,30(8) :1213 - 1226. (in Chinese)

[24] Kleinberg E M. Stochastic discrimination[J]. Annals of Mathematics and Artificial Intelligence,1990,1(1 - 4) :207 - 239.

[25] Kleinberg E M. An overtraining-resistant stochastic modeling method for pattern recognition[J]. Annals of Statistics,1996,4(6) :2319 - 2349.

[26] Kleinberg E M. On the algorithmic implementation of stochastic discrimination[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(5) :473 - 490.

[27] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(8) :832 - 844.

[28] Blake C,Keogh E,Merz C J. UCI repository of machine learning databases[DB/OL]. http://www.ics.uci.edu/-mlearn/MLRepository.html,1998.

[29] D'Alche-Buc F,Grandvalet Y,Ambroise C. Semi-supervised marginboost[A]. Advances in Neural Information Processing Systems 14[C]. MIT Press,2002. 553 - 560.

[30] Bennett K P,Demiriz A,Maclin R. Exploiting unlabeled data in ensemble methods[A]. Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining[C]. Edmonton,Canada,2002. 289 - 296.

作者简介:

王　娇　女,1982 年出生于四川雅安. 2003 年毕业于北京交通大学计算机与信息技术系,现为北京交通大学计算机应用专业硕博连读生. 主要研究方向为机器学习、模式识别.
E-mail:wangjiao0828@163.com

罗四维　男,1943 年出生于北京. 北京交通大学教授、博士生导师. 主要研究方向为神经计算、模式识别.

曾宪华　男,1973 年出生于四川攀枝花. 北京交通大学博士研究生. 主要研究方向为机器学习.