基于图的 Co-Training 网页分类

侯翠琴,焦李成
(西安电子科技大学智能信息处理研究所和智能感知与图像理解教育部重点实验室,陕西西安 710071)

摘 要: 本文充分利用网页数据的超链接关系和文本信息,提出了一种用于网页分类的归纳式半监督学习算法:基于图的 Co-training 网页分类算法(Graph based Co-training algorithm for web page classification),简称 GCo-training,并从理论上证明了算法的有效性. GCo-training 在 Co-training 算法框架下,迭代地学习一个基于由超链接信息构造的图的半监督分类器和一个基于文本特征的 Bayes 分类器. 基于图的半监督分类器只利用少量的标记数据,通过挖掘数据间大量的关系信息就可达到比较高的预测精度,可为 Bayes 分类器提供大量的标记信息;反过来学习大量标记信息后的 Bayes 分类器也可为基于图的分类器提供有效信息. 迭代过程中,二者互相帮助,不断提高各自的性能,而后 Bayes 分类器可以用来预测大量未见数据的类别. 在 Web→KB 数据集上的实验结果表明,与利用文本特征和锚文本特征的 Co-training 算法和基于 EM 的 Bayes 算法相比,GCo-training 算法性能优越.

关键词: 图; 半监督; Co-training; 归纳式; 网页分类
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2009) 10-2173-08

Graph Based Co-Training Algorithm for Web Page Classification

HOU Cui-qin, JIAO Li-cheng
(Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Institute of Intelligent Information Processing, Xidian University, Xi'an, Shaanxi 710071, China)

Abstract: This paper proposes a novel inductive semi-supervised algorithm for web page classification named GCo-training, exploiting texts in web pages and hyperlinks among them. GCo-training iteratively trains two classifiers-a graph-based semi-supervised classifier based on hyperlinks among web pages and a Bayes classifier based on texts in web pages, under the framework of Co-training. On the one hand, the graph-based semi-supervised classifier obtains high accuracy based on a small set of labeled examples through exploiting links among web pages and can augment labeled examples for the Bayes classifier. On the other hand, the Bayes classifier can also provide labeled example for the graph-based classifier after it learning on labeled set augmented by the graph-based classifier. Therefore, the two classifiers help each other and improve their respective performance during the process of training. Finally, the Bayes classifier can classify a large number of unseen examples. We test GCo-training algorithm, Co-training algorithm based on words occurring on web pages and words occurring in hyperlinks and Bayes algorithm based on EM on the Web→KB dataset. Experimental results show GCo-training performs much better than the other algorithms.

Key words: graph; semi-supervised; Co-training; inductive; web page classification

1 引言

互联网上存在海量的网页数据,但标记网页数据需要耗费大量的人力资源,因此利用半监督学习技术,在少量标记数据的基础上,通过挖掘大量未标记数据来提高分类器精度的方法受到学者们和商业界的广泛关注.

学者们提出的大量半监督学习算法都可以用在网页数据上. 如 Transductive SVM[1]和基于 EM 的半监督分类算法[2]. 但它们仅仅利用了网页数据丰富信息中的文本信息,而忽略了网页数据中极具价值的链接信息等. 网页数据中的链接关系可以用图表示,因此基于图的半监督分类算法[3~6]也常用来解决网页分类问题. 基于图的半监督分类算法主要挖掘数据之间的关系,可以只基于少量的标记数据而较精确地预测未标记数据的类别,但由于它是直推式学习算法,只能预测图中未标记数据的类别,对与图不相连接的数据无能为力.

Co-training 族算法[7]是半监督学习领域重要分支之一,要求数据有两组相互独立而又能充分表示数据的特

收稿日期:2008-05-19;修回日期:2009-03-09
基金项目:国家自然科学基金(No. 60602064,No. 60702062);教育部重点项目(No. 108115)国家 973 重点基础研究发展规划(No. 2006CB705707);国家 863 高技术研究发展计划(No. 2007AA12Z223);国家部委科技项目(No. 51307040103);教育部长江学者和创新团队支持计划(No. IRT0645)

征集,并在每个特征集下训练一个分类器.而由于这两组特征是相互独立的,因此一个分类器可以利用另一个分类器对未标记数据的预测来增多训练样本以提高自身的分类精度.Co-training 算法通过迭代使这两个分类器互相帮助,提高各自的分类精度.但它要求每个分类器的初始精度都不能太差,否则迭代很可能会损失分类器的精度.网页数据既可以用网页自身的文本信息表示,又可以用链接网页的锚文本来表示.因此 Co-training 算法可以基于这两组特征集,来对网页分类.这也是最初 Co-training 算法被提出时,验证算法有效性所用的实验.

网页数据独有的超链接关系图为分析数据的类别提供了丰富的信息,而与之相对独立的文本信息是分类算法常用的特征.基于此,本文提出一种 Co-traing 框架下的归纳式半监督分类算法(Graph based Co-training algorithm for web page classification,GCo-training).它在 Co-training 算法框架下迭代地训练两个分类器:基于图的半监督分类器和基于文本特征的 Bayes 分类器.基于图的半监督分类器在少量标记数据情况下对可见数据预测的高精度可以为 Bayes 分类器提供大量的标记信息,而 Bayes 分类器最自信的分类结果也可为基于图的半监督分类器提供帮助.因此通过迭代,二者可互相帮助,提高各自的分类精度.算法学到的 Bayes 分类器能处理大量未见的新网页.本文所提算法充分利用了基于图的半监督学习算法在少量标记数据下预测的高精度特性和 Bayes 分类器对未来样本的分类能力.实验中,用 Web→KB 数据集上对 GCo-training 算法、Co-training 算法和基于 EM 的 Bayes 算法进行大量测试.实验结果表明,本文所提 GCo-training 算法取得了更好的效果.

2 背景知识

由于标记数据类属的耗时性、易错性以及大量未标记类属数据的易得性,半监督学习[8]在近 10 多年来受到广泛关注、得到蓬勃发展,已成为机器学习领域重要分支之一.

我们将半监督学习按照其训练出的学习器能否对未见数据预测分为①:归纳式半监督学习和直推式半监督学习.直推式半监督学习主要包括基于图的半监督学习,而归纳式半监督学习又可根据学习过程中学习器的个数分为:多分类器归纳式半监督学习和单分类器归纳式半监督学习.多分类器归纳式半监督学习的一般假设是各个分类器在未标记数据上的预测应该一致,主要包括利用数据多个特征集的 Co-training 族算法[7,9,10]和在同一特征集上训练多个不同分类器的算法[11].单分类器归纳式半监督学习中学习算法的区别在于它基于的假设不同,如:Transductive SVM[1]是 SVM 在半监督学习模式下的扩展,它约束分类超平面不穿越高密度的数据区域;基于 EM 的半监督分类算法[2]假设数据分布符合特定的概率模型,在此假设基础上,EM 算法同时利用标记数据和未标记数据来估计分布模型的参数,而后用得到的模型预测数据的类别;流形正则框架[12]则假设在流形上相近的两点类别也相近.本文所提算法属于多分类器归纳式半监督学习算法.

2.1 基于图的半监督学习算法

基于图的半监督学习算法的一个基本假设是一致性假设[4],即(1)相邻数据的类别以高概率相同;(2)同一流形上数据的类别以高概率相同;其中(1)是局部假设,也是 k 近邻算法的基本假设,(2)是全局假设.我们用图 1 来说明一致性假设,从局部看图中相邻的点属于同一类;从全局看在同一流形结构上的点属于同一类.虽然节点 a 和节点 b 的距离小于节点 a 和节点 c 之间的距离,但由于节点 a、c 处于同一流形结构上,它们属于同一类.

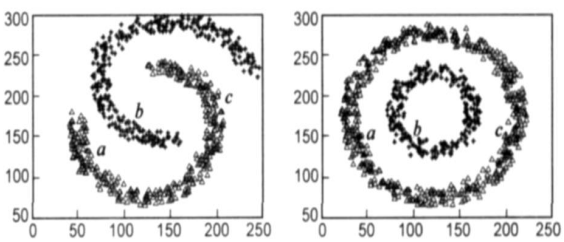


图1 基于图的半监督学习算法基于的一致性假设示意图

给定数据$\{x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_n\}$及前 l 个数据的类别$\{y_1, y_2, \cdots, y_l\}$,其中 $y_i \in \{1, 2, \cdots, C\}$,我们可以用图 $G(V, W)$ 来表示这 n 个数据,其中每个节点 V_i 对应一个数据,$W_{ij} = W_{ji}$表示节点V_i 与节点V_j 之间的相似性.基于图的半监督学习算法的任务就是预测剩下 $n-l$ 个未标记数据的类别.基于一致性假设,它的任务可看作优化一个关于图中节点的函数,使它在相邻节点上光滑变化,而在不同流形结构上突变.但数据是否处在不同的流形结构上无法定量衡量,而前 l 个数据的类别是给定的,因此可以利用给定数据的类别来约束函数在不同流形结构上的变化.基于此原理,Zhou 等人[4]提出如下目标函数:

$$Q(F) = \left(\sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 \right.$$

① 半监督学习是与监督学习相对的一种学习模式,它不仅利用标记数据学习,而且还利用大量未标记数据来进一步提高学习器的性能.而直推式学习是与归纳式学习相对的,只要求预测现有未标记数据,不要求学习器能对够对整个假设空间中所有数据预测.直推式学习与半监督学习的关系一直是机器学习领域争议的热点.本文将能利用未标记数据学习的算法都归为半监督学习,而将直推式学习和归纳式学习看作是半监督学习模式下的两类学习范式.

$$+\mu\sum_{i=1}^{n}\|F_i - Y_i\|^2), \quad (1)$$

其中,C 维行向量 $Y_i=\{Y_{i1}, Y_{i2}, \cdots, Y_{iC}\}$ 表示标记节点 i 的类别,其中 $Y_{ij}=\begin{cases}1, & \text{if } y_i=j\\ 0, & \text{else}\end{cases}$;$C$ 维行向量 F_j 表示算法对节点 j 属于不同类概率的预测,预测节点 j 的类别为:$f_j=\arg\max_k F_{jk}$;$\|a\|$ 是向量的 2-范数;$D_{ii}=\sum_{j=1}^{n}W_{ij}$;$\mu$ 是可调节的正则参数.

公式(1)中,前半部分衡量函数 F 在相邻节点间的变化情况,而后半部分衡量函数对标记数据类别拟合的好坏,参数 μ 控制这两部分的相对权重. 通过推导可得凸函数(1)的最优值为:

$$F^*=\arg\min_{F\in\mathbb{F}} Q(F)=(1-\alpha S)^{-1}Y \quad (2)$$

其中 $S=D^{-1/2}WD^{-1/2}$;D 是只有对角元素为非 0 的矩阵,$D_{ii}=\sum_{j=1}^{n}W_{ij}$;$\alpha=\frac{1}{1+\mu}$

基于图的半监督学习算法首先根据数据之间的相似性关系构造图;而后构造图上的一个函数,如 Blum 和 Chawla 算法[13]中的最小切函数,Zhu 等人算法[3]中的二次能量函数和 Zhou 等人算法[4]中的公式(1)函数;最后优化这个函数,预测未标记数据的类别. 这类算法的核心是构造合理的图,而这类算法的不同之处在于图上构造函数的不同.

3　用于网页分类的基于图的 Co-training 算法

本节首先描述如何基于网页的超链接关系为基于图的半监督学习算法构造合理的相似性图,在此基础上提出基于图的 Co-training 网页分类算法(Graph based Co-training algorithm for web page classification,简称 GCo-training).

3.1　相似性图的构造

基于图的半监督学习算法的核心是图,图中节点之间的链接关系以及图中边的权重是否合理直接影响算法的预测精度. 本节根据 web 数据本身的超链接关系,构造 web 数据的相似性矩阵,即相似性图.

首先用图 G(V, E)来表示网页数据,其中每个节点 V_i 对应一个网页,E 表示节点间的链接关系,如果节点 V_i 有超链接到节点 V_j,则 $E_{ij}=1$,否则 $E_{ij}=0$. 显然,矩阵 E 是非对称矩阵. 网页数据集中有超链接关系的节点属于同一类的概率比较小. 例如,一个学生主页会链接一些其他同学的主页,也会链接一些著名会议和期刊的主页,甚至还会链接一些有关自己兴趣爱好的网页;很多新闻网页上会有大量的广告链接等. 因此我们需要基于网页数据间的超链接关系来计算它们之间的相似性,使类别相同的个体相似性值大.

受 PageRank 算法[14]的启发,利用每个网页被其它网页链接的情况来计算不同网页之间的相似性,认为被越多相同的网页链接的两个网页,它们越相似. 但由于每个网页向外链接和被链接的网页数都是不同的,如果一个网页向外链接的网页数越多,如门户网站的主页,说明它们之间的关系越弱. 如图 2,节点 f、g 和 h、i 都被两个节点超链接,但不能简单认为节点 f、g 间的相似度与 h、i 间的相似度是一样的. 同样,节点 h、i 和 j 都被节点 c、d 链接,但它们两两之间的相似性也不能简单认为相同. 因此首先规范化链接矩阵 E,使

$$\bar{E}_{ij}=\frac{E_{ij}}{\sqrt{D_i * D_j}} \quad (3)$$

其中 $D_i=\sum_{j,v_i\to v_j}E_{ij}$,$D_j=\sum_{i,v_i\to v_j}E_{ij}$. 而后计算节点间的相似性,

$$W_{ij}=\sum_{k=1}^{n}\bar{E}_{ki} * \bar{E}_{kj} \quad (4)$$

W 矩阵是对称矩阵,因此 Zhou 等人的 LLGC 算法[4]可以在用 W 矩阵构造的图上预测图中未标记节点的类别.

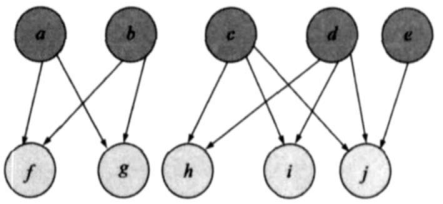


图2　节点f、g和h、i都被两个节点超链接,但不能简单认为节点f、g间的相似度与h、i间的相似度是一样的. 同样,节点h、i和j都被节点c、d链接,但它们两两之间的相似性也不能简单认为相同

3.2　基于图的 Co-training 算法

基于图的半监督算法是利用所有数据间的关系来对数据类别预测的,因此当遇到新的数据时,它就要重新构造图,而当新数据不断涌现时,这种方法是无能为力的. 但在给定数据集下,基于图的半监督算法只需要少量标记数据,通过挖掘数据间的关系,就可达到比较高的预测精度. 受此启发,本节提出用于网页分类的基于图的归纳式半监督学习算法,它充分利用了基于图的半监督学习算法在少量标记数据的基础上就能有高精度预测的优点和 Bayes 分类器处理新样本能力的优点. 所提算法称为基于图的 Co-training 网页分类算法(Graph based Co-training algorithm for web page classification,简称 GCo-training).

基于图的 Co-training 网页分类算法(GCo-training)概述如下:

(1) $t=0$,根据给定的数据构造超链接图 G(V, E),如果节点 V_i 有超链接指向节点 V_j,$E_{ij}=1$;否则 $E_{ij}=0$. 设定算法参数,初始化 Y_i.

(2) 规范化图中边的权重，$\bar{E}_{ij}=\frac{E_{ij}}{\sqrt{D_i * D_j}}$，其中 $D_i=\sum_{j,v_i\neq v_j}E_{ij}$，$D_j=\sum_{i,v_i\neq v_j}E_{ij}$.

(3) 计算相似度矩阵，$W=\bar{E}^{T}\times\bar{E}$.

(4) 用基于 Y_i 的 LLGC 算法预测图中未标记节点的类别，$f_j=\arg\max_k F^{*}_{jk}$.

(5) 用标记数据①x_i 及其类别 y_i 以及未标记数据 x_j 及 f_j，训练基于网页文本特征的 Bayes 分类器.

(6) 用 Bayes 分类器预测图中未标记节点属于各类别的概率，$F_j'=(F_{j1}',F_{j2}',\cdots,F_{jC}')$ 且 $\sum_{k=1}^{C}F_{jk}'=1$.

(7) 找节点 k，$F_{ks}'=\max_{i,j}F_{ij}'$，将节点 k 加入到标记节点集中，令 $Y_{ks}=1$，$Y_{kj}=0$，$j\neq s$.

(8) $t=t+1$，若满足终止条件，输出 Bayes 分类器，结束算法；否则返回步 4.

基于图的 Co-training 算法，首先基于网页数据的链接信息，为 LLGC 算法构造相似度矩阵，而后用给定的标记数据和 LLGC 对未标记数据的预测信息来训练 Bayes 分类器. 由于 LLGC 对未标记数据的预测是有噪音的，在此基础上学习到的 Bayes 分类器精度可能不高. 因此只选 Bayes 分类器预测最自信的一个节点，将其及 Bayes 分类器对该节点的预测结果加入到标记数据集中，为下一次迭代中 LLGC 算法所用. 基于链接信息的 LLGC 算法和基于文本特征的 Bayes 分类器，通过迭代上述过程，互相帮助，不断提高自己的预测精度. 最后，得到的 Bayes 分类器可以处理大量未见样本.

在算法中，Bayes 分类器[2]所用特征为网页的文本特征，即词频统计特征. Bayes 分类器对网页 d_i 为 c_j 类预测的概率为：

$$F_{ij}'=P(y_i=c_j\mid d_i;\hat{\theta})=\frac{P(c_j\mid\hat{\theta})*P(d_i\mid c_j;\hat{\theta})}{P(d_i\mid\hat{\theta})}$$

$$=\frac{P(c_j\mid\hat{\theta})\prod_{k=1}^{|d_i|}P(w_{d_{i,k}}\mid c_j;\hat{\theta})}{\sum_{r=1}^{|C|}P(c_r\mid\hat{\theta})\prod_{k=1}^{|d_i|}P(w_{d_{i,k}}\mid c_r;\hat{\theta})}\qquad(5)$$

其中，$|d_i|$ 表示网页 d_i 所含的单词数，$w_{d_{i,k}}$ 表示 d_i 所含的第 k 个单词. 公式(5)基于 bag-of-word 模型，忽略单词的顺序及上下文的信息. 公式(5)中的参数 $\hat{\theta}_{c_j}=P(c_j\mid\hat{\theta})$ 表示 c_j 类文本的概率，而 $\hat{\theta}_{w_t\mid c_j}=P(w_{d_{i,k}}\mid c_j;\hat{\theta})$ 表示 c_j 类文本产生 $w_{d_{i,k}}$ 的概率，它们都是 Bayes 分类器基于标记样本和 LLGC 对未标记样本的预测信息学习到的. 其中，

$$P(c_j\mid\hat{\theta})=\frac{1+\sum_{i=1}^{|n|}P(y_i=c_j\mid d_i)}{|C|+|n|}\qquad(6)$$

$$P(w_t\mid c_j,\hat{\theta})=\frac{1+\sum_{i=1}^{|n|}N(w_t,d_i)\,P(y_i=c_j\mid d_i)}{|M|+\sum_{s=1}^{|m|}\sum_{i=1}^{|n|}N(w_s,d_i)\,P(y_i=c_j\mid d_i)}\qquad(7)$$

与上面定义一致，$|C|$ 表示类别的总数，$|n|$ 表示所有可见数据的总数，即可见网页的个数. 公式(7)中，$|m|$ 表示单词的总数，即特征的个数；$N(w_t,d_i)$ 表示单词 w_t 在网页 d_i 中发生的总次数，而数据 d_i 不属于标记数据集时，

$$P(y_i=c_j\mid d_i)=\begin{cases}1, & \text{if } c_j=f_j\\ 0, & \text{else}\end{cases}\qquad(8)$$

4 GCo-training 有效性分析

Co-training 算法的基本假设是分类器对 (h_1,h_2) 在数据预测上的一致性. 基于此，定义 Co-training 算法下的两个分类器 h_1，h_2 与数据 D 的相容值.

定义 1 分类器对 (h_1,h_2) 与数据 D 的相容值

记 $x=(x_1,x_2)$ 为一满足 D 分布的样本，则分类器对 (h_1,h_2) 与数据 D 的相容值为：$C((h_1,h_2),D)=P_{(x_1,x_2)\sim D}[h_1(x_1)=h_2(x_2)]$.

分类器对与数据的相容值为一个概率值，因此它的取值范围是 $[0,1]$. 由相容值，可得分类器对与数据的不相容值为：

$$\begin{aligned}&iC((h_1,h_2),D)\\&=1-C((h_1,h_2),D)\\&=1-P_{(x_1,x_2)\sim D}[h_1(x_1)=h_2(x_2)]\\&=P_{(x_1,x_2)\sim D}[h_1(x_1)\neq h_2(x_2)].\end{aligned}$$

当数据分布 D 未知而满足 D 分布的数据集 R 已知时，分类器对 (h_1,h_2) 与数据集 R 的相容值为 $\hat{C}((h_1,h_2),R)=P_{(x_1,x_2)\in R}[h_1(x_1)=h_2(x_2)]$，不相容值为 $i\hat{C}((h_1,h_2),R)=P_{(x_1,x_2)\in R}[h_1(x_1)\neq h_2(x_2)]$.

定义 2 分类器对 (h_1,h_2) 的 ε 不相容集 $S_{C,D}(\varepsilon)$

$S_{C,D}(\varepsilon)=\{(h_1,h_2)\in S: iC((h_1,h_2),D)\leq\varepsilon\}$，其中 S 是 (h_1,h_2) 所有可能的集合，$0\leq\varepsilon\leq1$.

基于图的半监督分类器的基本假设是一致性假设[3]. 基于此，定义基于图的半监督分类器 h_1 与数据 $D1$ 的相容值.

定义 3 分类器 h_1 与数据 $D1$ 的相容值

分类器 h_1 与数据 $D1$ 的相容值 $C1(h_1,D1)=P_{x_1,x_2\sim D1}[h_1(x_1)=h_1(x_2)\wedge x_1\in H\wedge x_2\in H]$，其中 H 表示一流形结构.

① 此时的标记数据不仅包括算法执行前给定的标记数据，也包括算法运行过程中，Bayes 算法给定的标记数据.

类似,分类器 h_1 与数据 $D1$ 的不相容值为:

$$\begin{aligned} &iC1(h_1, D_1)\\ &=1-C1(h_1, D_1)\\ &=1-P_{x_1\sim D1, x_2\sim D1}[h_1(x_1)=h_1(x_2)\wedge x_1\in H\wedge x_2\in H]\\ &=P_{x_1\sim D1, x_2\sim D1}[h_1(x_1)\neq h_1(x_2), \wedge x_1\in H\wedge x_2\in H] \end{aligned}$$

定义 4　分类器 h_1 的 ε 不相容集集 $S1_{C1,D}(\varepsilon)$

$S_{C1,D}(\varepsilon)=\{h_1\in S1: iC1(h_1, D_1)\leq\varepsilon\}$,其中 $S1$ 是 h_1 所有可能的集合,$0\leq\varepsilon\leq 1$.

基于定义 2 和定义 4,可得 GCo-training 算法下分类器对 (h_1, h_2) 的 ε 不相容集为:$S_{C,C1,D}(\varepsilon)=\{(h_1, h_2)\in S: iC((h_1, h_2), D)\leq\varepsilon\wedge iC1(h_1, D_1)\leq\varepsilon\}$. 显然,$|S_{C,C1,D}(\varepsilon)|\leq|S_{C,D}(\varepsilon)|$. GCo-training 用分类器 h_2 预测未标记样本的类别,则根据文献[15]中的定理 1 可得如下定理:

定理 1　给定由 u 个未标记样本和 l 个标记样本组成的样本集 R,如果 $u\geq\frac{1}{\varepsilon}[\ln|S|+\ln\frac{2}{\delta}]$, $l\geq\frac{1}{\varepsilon}[\ln|S_{C,C1,D}(\varepsilon)|+\ln\frac{2}{\delta}]$,则在 GCo-training 算法下满足 $\hat{R}(h_2)=\frac{1}{l}\sum_{i=1}^{l}I(h_2(x_i), y_i)=0$ 和 $i\hat{C}((h_1, h_2), R)=0$ 的分类器对 (h_1, h_2) 至少以概率 $1-\delta$ 满足 $R(h_2)=E_{x_i\sim D}(I(h_2(x_i), y_i))\leq\varepsilon$. 其中,

$$I(f(x_i), y_i)=\begin{cases}1, & f(x_i)\neq y_i\\ 0, & f(x_i)=y_i\end{cases}.$$

定理 2　给定 $l=\frac{1}{\varepsilon}[\ln|S_{C,C1,D}(\varepsilon)|+\ln\frac{2}{\delta}]$ 个标记样本和充分多的未标记样本,如果 GCo-training 和 Co-training 下满足 $\hat{R}(h_2)=\frac{1}{l}\sum_{i=1}^{l}I(h_2(x_i), y_i)=0$ 和 $i\hat{C}((h_1, h_2), R)=0$ 的分类器对 (h_1, h_2) 分别以概率 $1-\delta$, $1-\delta_1$ 满足 $R(h_2)=E_{x_i\sim D}(I(h_2(x_i), y_i))\leq\varepsilon$,则 $\delta\leq\delta_1$.

证明　由定理 1 可得:Co-training 中标记样本数 $m_l\geq\frac{1}{\varepsilon}[\ln|S_{C,D}(\varepsilon)|+\ln\frac{2}{\delta_1}]$. 而 $|S_{C,D}(\varepsilon)|\geq|S_{C,C1,D}(\varepsilon)|$,因此当未标记数据充分多,$m_l=\frac{1}{\varepsilon}[\ln|S_{C,C1,D}(\varepsilon)|+\ln\frac{2}{\delta}]$ 时,δ 至少要小于等于 δ_1 才能使 $m_l\geq\frac{1}{\varepsilon}[\ln|S_{C,D}(\varepsilon)|+\ln\frac{2}{\delta_1}]$. 得证.

由定理 2 得 GCo-training 在相同条件下能以更大的概率保证分类器的推广误差上限.

根据 Angluin 和 Laird 的结论[16]:随机抽 m 个样本组成样本集,如果 $m'\geq\frac{2}{\varepsilon^2(1-2\eta)^2}\ln(\frac{2N}{\delta})$,则最小化经验误差的假设 H_i 是 PAC 可学习的,即:$\Pr[d(H_i, H^*)\geq\varepsilon]\leq\delta$. 其中,$\varepsilon$ 是最坏情况下假设的分类误差率,η (<0.5) 是样本噪声的上限,N 是假设的总和,δ 是自信度,$d(H_i, H^*)$ 度量假设 H_i 与真实假设 H^* 之间的差别,$d(H_i, H^*)=\Pr(x\in H_i\cup H^*\wedge x\notin H_i\cap H^*)$.

令常数 $c=2\mu\ln(\frac{2N}{\delta})$,使 $m'=\frac{2c}{\varepsilon^2(1-2\eta)^2}$. 由此可得:训练样本数 m 与样本噪声率成正比例关系,与最大误差率成反比例关系. 因此,当样本的噪声率越大,以相同概率保证同样的分类精度,所需样本越多. 在 GCo-training 算法中,基于图的半监督学习算法预测的高精度保证了其为 Bayes 分类器提供样本的低噪声率,进而保证了 Bayes 分类器的低推广误差.

5　比较仿真实验

为了验证本文所提算法的有效性,我们分别在 Web→KB 数据集[17]上对 GCo-training 算法、基于文本特征和锚文本特征的 Co-training 算法[7]和基于 EM 的 Bayes 算法[2]进行大量仿真实验. 实验结果表明 GCo-training 能取得更好的分类效果.

5.1　实验数据集与评价准则

本文在 Web->KB 数据集上[17]评价算法性能. 分别以 Cornell 大学、Washington 大学、Wisconsin 大学和 Texas 大学的网页集合作为可见数据集①,以 misc 为测试数据集. 每个数据集中包括 student、course、staff、department、faculty、project 和大量的 others 7 类样本,本文只关注 student 类和非 student 类.

首先对网页中的文本信息进行预处理,即去除停用词并对单词进行归根后,去除在整个数据集中发生次数少于或等于 10 次的词根. 同时提取锚文本信息,对锚文本中所有词归根并去除停用词后,去除发生次数少于或等于 3 次的词根. 可见数据集的数据特性如表 1 所示.

表 1　可见数据集的数据特征

	网页总数	Student 类网页数	文本词数	锚文本词数
Cornell	867	128	2825	195
Texas	827	148	2481	160
Wisconsin	1263	156	3518	340
Washington	1205	126	3508	206

测试样本集 misc 中包含 1083 个 student 类样本和 3037 个非 student 类样本. 显然,这是个不平衡分类问题. 本文采用信息检索中一个衡量排序的指标 AUC[18]来评价算法性能. 它的定义如下:

① 可见数据集表示当前算法可看到的样本集,包含有标记的样本没不标记的样本

$$AUC = \frac{S_0 - \frac{n_0(n_0+1)}{2}}{n_0 n_1} \tag{9}$$

其中，n_0、n_1 分别表示正负样本的个数，而 $S_0 = \sum_{i \in positive} idx_i$ 表示将样本按分类器对其预测为正样本的概率从小到大排序，正样本所处名次的叠加. *AUC* 的变化范围是[0,1]，数值越大表明算法的性能越好.

5.2 实验结果

本节比较了 GCo-training 算法、基于文本特征和锚文本特征的 Co-training 算法[7]和基于 EM 的 Bayes 算法[2].

基于文本特征和锚文本特征的 Co-training 算法按[7]设置参数. 基于 EM 的 Bayes 算法按公式(6)和(7)估计 Bayes 分类器的参数，但 $P(y_i = c_j | d_i)$ 是 Bayes 分类器在上一次迭代中，用公式(5)预测可见样本集中网页 d_i 属于 c_j 类的概率. 在 GCo-training 中，LLGC 中 α 的值设为 0.8，而所有算法总的迭代次数都设为 10. 学习 Bayes 分类器中参数的时间复杂度为 $O(m * n)$，Bayes 分类器预测未标记数据类别概率的复杂度为 $O((n-l) * C)$，m 为特征数，l 为标记样本数，n 为可见样本数. 而 LLGC 算法迭代一次的时间复杂度为 $O(n * n * C)$，C 为类别数. 则 GCo-training 算法总的时间复杂度为 $O(K * (K' * n * n * C + m * n + (n-l) * C))$，$K$ 为算法总的迭代次数，K 为 GCo-training 算法每代 LLGC 的迭代次数. 而基于文本特征和锚文本特征的 Co-training 算法的时间复杂度为：$O(K * (m * n + m' * n + (n-l) * C))$，$m$ 为锚文本特征数. 基于 EM 的 Bayes 算法时间复杂度为：$O(K * (m * n + (n-l) * C))$.

在各个可见数据集中，分别以 1%、2%、3%、4%、5%、6%、7%、8%、9%和 10%的概率随机选定标记样本，并保证 student 类和非 student 类中都至少有一个标记样本. 在每种概率情况下，每种算法在各个可见数据集上都独立运行 30 次，统计 AUC 的平均值，结果见表 2～5.

表 2 以 Texas 大学数据集为可见数据集，各算法在不同标记概率下对 misc 中数据预测结果的比较

标记概率	EM Bayes	Co-training	GCo-training
1%	0.5302	0.5737	0.7962
2%	0.4932	0.5833	0.8075
3%	0.5760	0.6471	0.8177
4%	0.5938	0.6682	0.8162
5%	0.5810	0.6281	0.8141
6%	0.6491	0.6815	0.8248
7%	0.6285	0.6416	0.8127
8%	0.6208	0.7036	0.8231
9%	0.6381	0.7136	0.8231
10%	0.6072	0.7128	0.8224

表 3 以 Cornell 大学数据集为可见数据集，各算法在不同标记概率下对 misc 中数据预测结果的比较

标记概率	EM Bayes	Co-training	GCo-training
1%	0.5912	0.5694	0.8070
2%	0.5695	0.6026	0.8070
3%	0.6197	0.6077	0.8172
4%	0.5598	0.6477	0.8178
5%	0.6061	0.6717	0.8233
6%	0.6134	0.6845	0.8229
7%	0.6418	0.6892	0.8200
8%	0.6277	0.7036	0.8213
9%	0.6285	0.7157	0.8211
10%	0.6750	0.6976	0.8223

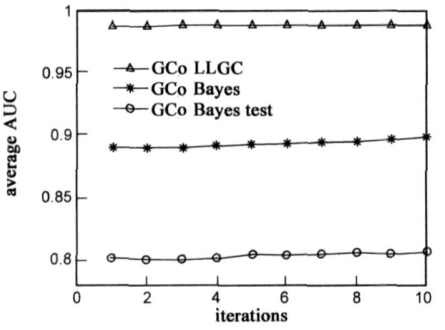
表 4 以 Wisconsin 大学数据集为可见数据集，各算法在不同标记概率下对 misc 中数据预测结果比较

标记概率	EM Bayes	Co-training	GCo-training
1%	0.6028	0.5437	0.7565
2%	0.6088	0.5609	0.7586
3%	0.5997	0.5418	0.7603
4%	0.6274	0.5871	0.7644
5%	0.6262	0.6122	0.7604
6%	0.6257	0.5938	0.7694
7%	0.6073	0.6474	0.7733
8%	0.6248	0.6520	0.7622
9%	0.6475	0.6561	0.7731
10%	0.6232	0.6724	0.7698

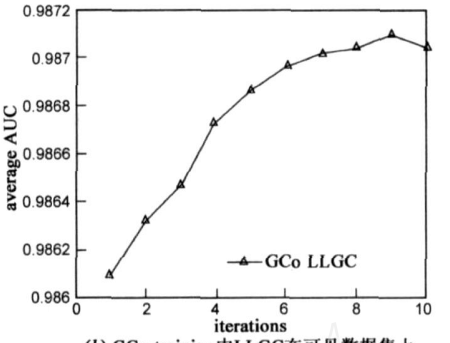
表 5 以 Washington 大学数据集为可见数据集，各算法在不同标记概率下对 misc 中数据预测结果比较

标记概率	EM Bayes	Co-training	GCo-training
1%	0.5830	0.5502	0.5534
2%	0.5735	0.5759	0.5663
3%	0.5998	0.5923	0.5968
4%	0.5854	0.5921	0.6168
5%	0.6102	0.6084	0.6167
6%	0.5371	0.6274	0.6212
7%	0.5863	0.6112	0.6291
8%	0.5874	0.6312	0.6317
9%	0.5908	0.6190	0.6454
10%	0.6258	0.6391	0.6532

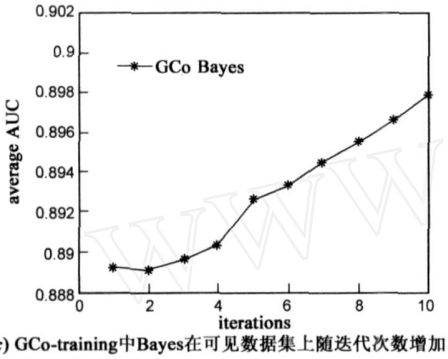
从统计结果(表 2～5)我们可以发现，以 Texas、Cornell 及 Wisconsin 大学数据集为可见数据集时，GCo-training 算法有明显的优势. 以 Texas 大学为可见数据集时 GCo-training 平均优于 Co-training 25.07%，优于基于 EM 的 Bayes 38.68%；以 Cornell 大学为可见数据集时 GCo-training 平均优于 Co-training 24.75%，优于基于 EM 的 Bayes 33.72%；而以 Wisconsin 大学为可见数据集时 GCo-training 以 26.74%平均优于 Co-training，以 30.28%平均优于基于 EM 的 Bayes. 从统计结果也可发现，Co-training 的性能优于基于 EM 的 Bayes 算法性能. 而以 Washington 大学数据集为可见数据集时，在多数情况下 GCo-training 略优于 Co-training 和基于 EM 的 Bayes 算法. 比较在不同标记概率下的结果，可以进一步发现随着标记数据的增多，分类器的预测结果呈稳步提高趋势. 这与我们的直觉相一致，也进一步说明了标记数据类别的作用.



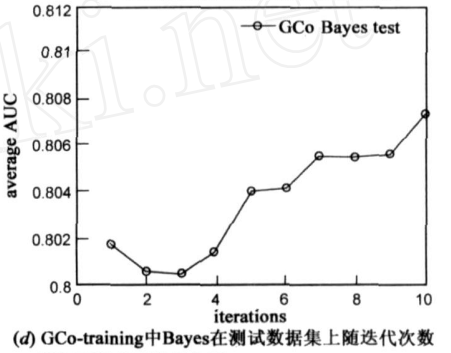
(*a*) 整体示意图



(*b*) GCo-training中LLGC在可见数据集上随迭代次数增加预测结果的变化图

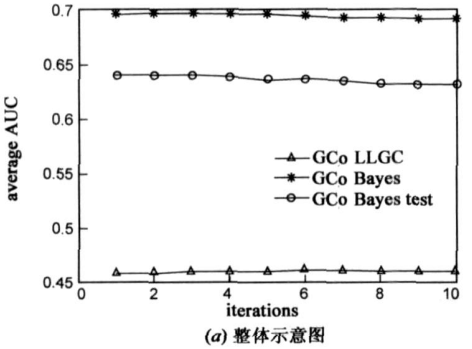


(*c*) GCo-training中Bayes在可见数据集上随迭代次数增加预测结果的变化图

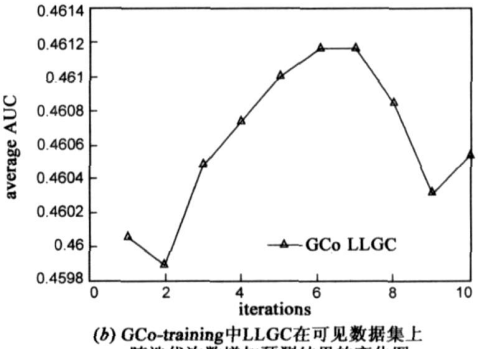


(*d*) GCo-training中Bayes在测试数据集上随迭代次数增加预测结果的变化图

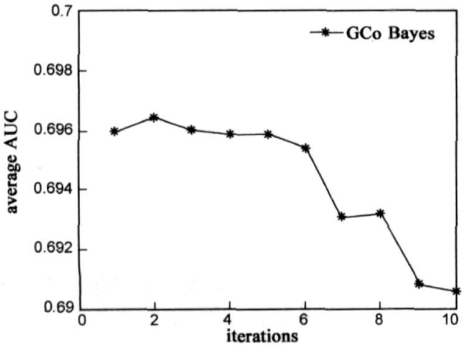
图3　以Texas大学为可见数据集, 在标记概率为2%情况下, GCo-training中LLGC在可见数据集上和Bayes在可见数据集及测试数据集上预测结果随迭代次数增加的变化图



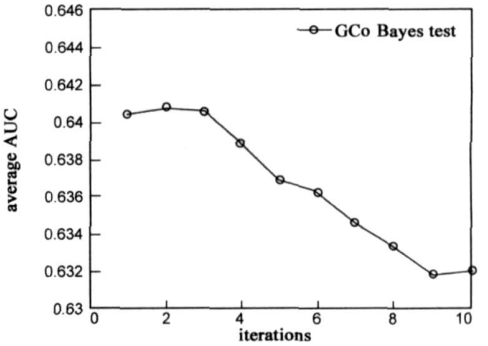
(*a*) 整体示意图



(*b*) GCo-training中LLGC在可见数据集上随迭代次数增加预测结果的变化图



(*c*) GCo-training中Bayes在可见数据集上随迭代次数增加预测结果的变化图



(*d*) GCo-training中Bayes在测试数据集上随迭代次数增加预测结果的变化图

图4　以Washington大学为可见数据集, 在标记概率为8%情况下, GCo-training中LLGC在可见数据集上和Bayes在可见数据集及测试数据集上预测结果随迭代次数增加的变化图

为了更清晰地了解本文算法在迭代过程中,对可见数据及未见数据预测结果的变化,我们统计以 Texas 大学为可见数据集,标记概率为 2 %,以 Washington 大学为可见数据集,标记概率为 8 %情况下,算法对可见数据和测试数据平均预测结果随着迭代次数变化的情况.从图 3 中,我们可以发现随着迭代次数的增加,基于图的半监督学习算法 LLGC 在可见数据集上的预测结果逐步变好,同时 Bayes 算法在可见数据集及测试数据集上的预测结果也稳步提高.这表明 LLGC 和 Bayes 在迭代过程中,互相帮助逐步提高自己的预测精度.从图 4 中可发现随着迭代次数的增加,GCo-training 中 LLGC 的预测值呈先增后减趋势,而 Bayes 分类器在可见数据和预测数据上的预测结果都呈下降趋势.这与 LLGC 在 Washington 大学可见数据集上的预测 AUC 值仅为 0.46 有关.LLGC 的错误预测值在迭代过程中逐步放大,导致基于 LLGC 在可见数据集上预测结果学习的 Bayes 分类器随着算法的迭代预测精度呈下降趋势.

6 结论

本文提出了一种用于网页分类的归纳式半监督学习算法 GCo-training,并从理论上证明了算法的有效性.从数据角度看,GCo-training 充分利用了网页间的链接信息和网页本身所含的文本信息;从学习算法方面看,GCo-training 充分利用了基于图的半监督学习算法对图中数据预测的高精度性和 Bayes 分类器归纳式处理样本的能力.仿真实验表明这种新方法性能优越,在网页分类问题上有一定的应用潜力.当 LLGC 预测精度较高时,GCo-training 算法性能在迭代过程中稳步提高,但当 LLGC 性能很差时,GCo-training 可能会退化.因此防止 GCo-training 在迭代过程中性能退化的理论分析部分还待进一步研究和挖掘,这也是我们下一步工作的目标.另外,研究在 Co-training 框架下结合基于图的半监督学习算法与其他的归纳式分类器,如 SVM 等,也是我们下一步的研究工作.

参考文献:

[1] T Joachims. Transductive inference for text classification using support vector machines[A]. Proceedings of the 16th International Conference on Machine Learning[C]. San Fransisco: Norgan Kaufmann,1999. 200 - 209.

[2] K Nigam,A McCallum,S Thrun,T Mitchell. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning,2000,39:103 - 134.

[3] X Zhu,Z Ghahramani,J Lafferty. Semi-supervised learning using gaussian fields and harmonic functions[A]. Proceedings of the 20th International Conference on Machine Learning[C]. New York:AAAI Press,2003. 912 - 919.

[4] D Zhou,O Bousquet,T Lal,J Weston,B Scholkopf. Learning with local and global consistency[A]. Advances in Neural Information Processing System 16[C]. Cambridge: MIT Press, 2004. 321 - 328.

[5] D Zhou,B Scholkopf,T Hofmann. Semi-supervised learning on directed graphs[A]. Advances in Neural Information Processing System 17[C]. Cambridge:MIT Press 2005. 1633 - 1640.

[6] D Zhou,J Huang,B Scholkopf. Learning from labeled and unlabeled data on directed graph[A]. Proceedings of the 22nd International Conference on Machine Learning[C]. New York: ACM Press,2005. 1041 - 1048.

[7] A Blum,T Mitchell. Combining labeled and unlabeled data with Co-training[A]. Proceedings of the 11th Annual Conference on Computational Learning Theory[C] New York: ACM Press, 1998. 92 - 100.

[8] X Zhu. Semi-Supervised Learning Literature Survey[R]. Technical Report 1530,Department of Computer Sciences,University of Wisconsin,Madison. 2005.

[9] K Niga,R Ghani. Analyzing the effectiveness and applicability of Co-training[A]. Proceedings of the 17th International Conference on Machine Learning[C]. San Fransisco:Norgan Kaufmann,2000. 86 - 93.

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

[11] M Culp,G Michailidis. An iterative algorithm for extending learners to a semi-supervised setting[J]. Journal of Computational Graphics and Statistics,2008,17(3):1-27.

(下转第 2219 页)

作者简介:

侯翠琴 女,1983 年生于山西左权,西安电子科技大学博士研究生,主要研究方向为结构数据学习、半监督学习、特征选择、稀疏学习及数据挖掘等.

E-mail:houcuiqin@hotmail.com

焦李成 男,1959 年生于陕西白水,西安电子科技大学教授,博士生导师,主要研究方向为智能计算、机器学习和智能信息处理等.发表多部专著,在国内外刊物上发表论文 100 余篇.

E-mail:lchjiao@mail.xidian.edu.cn

representation of multiphase flow and area preserving curvature flow[J]. Commun. Math. Sci,2008,6(1):125 - 148.

[3] T Chan,L Vese. Active contours without edges[J]. IEEE Image Proc,2001,10(2):266 - 277.

[4] Li C,Xu C,Gui C,M D Fox. Level set evolution without re-initialization:A new variational formulation[A]. IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1[C]. San Diego:IEEE Computer Society Press,2005. 430 - 436.

[5] V Caselles,R Kimmel,G Sapiro. Geodesic active contours[J]. International Journal of Computer Vision,1997,22(1):61 - 79.

[6] Zhao HK,T Chan,B Merriman,S Osher. A variational level-set approach to multiphase motion [J]. Journal of Computational Physics,1996,127:179 - 195.

[7] D Mumford,J Shah. Optimal approximation by piecewise smooth functions and associated variational problems [J]. Communications on Pure and Applied Mathematics,1989,42(5):577 - 685.

[8] R Malladi,J A Sethian,B C Vemuri. Shape modeling with front propagation:A level set approach[J]. IEEE Trans on PMAI,1995,17(2):158 - 175.

[9] Han X,Xu CY,J L Prince. A topology preserving level set method for geometric deformable models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2003,25(6):755 - 768.

[10] J Gomes,O Fangeras. Reconciling distance functions and levelsets [J]. Visiual Communic and Image Representation,2000,1(11):209 - 223.

[11] 王大凯,侯榆青,彭进业. 图像处理的偏微分方程方法[M]. 北京:科学出版社,2008.

作者简介:

何　宁　女,1970 年 7 月生于辽宁盘锦,首都师范大学数学科学学院博士研究生,主要研究方向为 CT 理论与应用及数字图像处理.
E-mail:hening_cnu@sina.com

张　朋　男,1957 年 3 月生于陕西榆林,首都师范大学数学科学学院教授、博士生导师,主要从事 CT 理论与应用及应用数学方面的研究与教学工作. 先后主持国家和省部级科研项目 10 余项,曾获亚洲 CT 科技进展荣誉奖杯、教育部优秀骨干教师奖、北京市科学技术一等奖等多项奖励.

(上接第 2180 页)

[12] M Belkin,P Niyogi,V Sindhwani. Manifold regularization:a geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research,2006,7:2399 - 2434.

[13] A Blum,S Chawla. Learning from labeled and unlabeled data using graph mincuts[A]. Proceedings of the 18th International Conference on Machine Learning[C]. San Fransisco:Norgan Kaufmann,2001. 19 - 26

[14] L Page,S Brin,R Motwani,T Winograd. The Pagerank Citation Ranking:Bring Order to the Web[R]. Technical Report, Stanford University,1998.

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

[16] S Angluin,P Laird. Learning from noisy examples[J]. Machine Learning,1988,2(4):343 - 370.

[17] M Craven,D DiPasquo,D Freitag,A McCallum,T Mitchell,K Nigam,S Slattery. Learning to construct knowledge bases from the world wide web[J]. Artificial Intelligence,2000,118(1/2):69 - 113.

[18] C Ling,J Huang,H Zhang. AUC:a Statistically consistent and more discriminating measure than accuracy[A]. Proceedings of the 18th International Joint Conference on Artificial Intelligence[C]. San Fransisco:Morgan Kaufmann,2003. 329 - 341.