基于样本条件价值改进的 Co-training 算法

程圣军 [1] 刘家锋 [1] 黄庆成 [1] 唐降龙 [1]

摘 要 Co-training 是一种主流的半监督学习算法. 该算法中两视图下的分类器通过迭代的方式, 互为对方从无标记样本集中挑选新增样本, 以更新对方训练集. Co-training 以分类器的后验概率输出作为新增样本的挑选策略, 该策略忽略了样本对于当前分类器的价值. 针对该问题, 本文提出一种改进的 Co-training 式算法 — CVCOT (Conditional value-based co-training), 即采用基于样本条件价值的挑选策略来优化 Co-training. 通过定义无标记样本的条件价值, 各视图下的分类器以样本条件价值为依据来挑选新增样本, 以此更新训练集. 该策略既可保证新增样本的标记可靠性, 又能优先将价值较高的富信息样本补充到训练集中, 可以有效地优化分类器. 在 UCI 数据集和网页分类应用上的实验结果表明: CVCOT 具有较好的分类性能和学习效率.

关键词 机器学习, 半监督学习, Co-training, 富信息样本, 条件价值

引用格式 程圣军, 刘家锋, 黄庆成, 唐降龙. 基于样本条件价值改进的 Co-training 算法. 自动化学报, 2013, **39**(10): 1665−1673

DOI 10.3724/SP.J.1004.2013.01665

Conditional Value-based Co-training

CHENG Sheng-Jun[1] LIU Jia-Feng[1] HUANG Qing-Cheng[1] TANG Xiang-Long[1]

Abstract Co-training is one of the major semi-supervised learning methods, which iteratively trains two classifiers under two different views, and uses the predictions of either classifier on the unlabeled examples to augment the training set of the other. In each round of co-training, newly added examples are selected according to the classifier′s posteriori probability output, which neglects examples′ value with respect to the current classifier. This paper proposes an improved co-training style algorithm, termed as CVCOT (conditional value-based co-training), which employs a conditional value-based strategy for selecting candidate training examples. Specifically, the conditional value of unlabeled examples in the co-training process is defined and computed, then it is utilized by either classifier under different views for augmenting the training set of the other. The new strategy can not only guarantee the reliability of the pseudo-labels, but also tends to add more informative examples with higher values to the training sets. Therefore, the classifier under either view will get refined. Experiments on UCI data sets and application to the web page classification task indicate that the CVCOT achieves better classification performance and learning efficiency.

Key words Machine learning, semi-supervised learning, co-training, informative example, conditional value

Citation Cheng Sheng-Jun, Liu Jia-Feng Liu, Huang Qing-Cheng, Tang Xiang-Long. Conditional value-based co-training. *Acta Automatica Sinica*, 2013, **39**(10): 1665−1673

半监督学习是一种研究成熟、应用广泛的机器学习范式, 其关键在于如何利用大量廉价的无标记数据提升分类器性能. 现有的半监督学习方法包括产生式模型 (Generative model)、半监督支撑向量机 (Support vector machines, SVM)、基于图的方法 (Graph-based) 等[1]. Blum 等[2] 提出的协同训练 (Co-training) 发展成为另一种解决半监督学习问题的范式.

Co-training 是一种多视图 (Multi-view) 自助式 (Bootstrap) 算法[3], 即每轮迭代过程中, 算法分别在训练集的两个不同视图下 (如自然分割的属性集) 训练分类器, 各分类器分别从无标记样本池中挑选出一些样本, 加以伪标记 (Pseudo label), 以此作为新增样本来扩充对方训练集. Co-training 的本质在于: 从不同的角度 (互相独立的视图) 看待训练样本集, 可得到不同的特征空间, 在两不同特征空间下训练的分类器可较好地弥补对方的不足. 多视图协同训练的思想已经应用到众多现实领域, 如自然语言处理中句法分析[4] 和名词识别[5]、信息提取[6]、计算机辅助医疗诊断[7] 和垃圾邮件识别[8] 等.

在 Co-training 迭代过程中, 两个分类器互为对

收稿日期 2012-05-08 录用日期 2012-08-02
Manuscript received May 8, 2012; accepted August 2, 2012
国家自然科学基金 (61173087, 61073128), 黑龙江省自然科学基金 (F201021) 资助
Supported by National Natural Science Foundation of China (61173087, 61073128), Natural Science Foundation of Heilongjiang Province (F201021)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

方挑选新增样本, 对于任一视图下的分类器, 其自身新增样本的挑选过程仅由对方根据分类器的后验概率输出来完成, 而本身并不参与其中. 这种被动的挑选策略仅关注新增样本的标记可靠性, 而忽略了新增样本对于当前分类器模型的价值 (默认所有无标记样本的价值相同). 主动学习 (Active learning) 认为, 由于半监督样本存在抽样偏置 (Biased sampling)[9], 无标记样本的价值不同, 无标记样本池中充斥着大量价值很低的冗余和离群样本 (Outlier)[10]. Co-training 所采用的被动挑选策略并不能保证在每次迭代中优先将价值高的样本加入训练集中, 从而导致分类器得不到有效的优化, 影响学习效率.

针对上述问题, 本文提出一种改进的 Co-training 式的算法 — CVCOT (Conditional value-based co-training), 即利用基于样本条件价值的挑选策略来优化 Co-training. 通过比较不同视图下样本的条件价值, 优先将条件价值更高的无标记样本加入训练集中. 该策略不仅保证新增样本的标记可靠性, 而且能够最大化样本的信息度, 从而提高分类器的学习效率, 优化分类表现.

Co-training 有着成熟的理论基础[11−13] 和大量实验的验证支持[14−15]. 现有相关改进工作体现在两方面: 1) 将协同训练思想引入到不具备多视图的分类问题中, 由此发展成一类基于集成学习解决半监督问题的方法[16], 如 Goldman 等提出的单视图协同训练[17] 以及周志华等提出的 Tri-training[18] 等; 2) 提高新增样本标记的可靠性. 主要方法为通过数据剪辑技术来剔除训练集中的噪声样本[19−20], 或者利用图的方法来更新样本的标记[21]. 与现有方法改进的角度不同, 本文在多视图框架下采用一种改进的新增样本挑选策略来优化标准协同训练算法, 通过基于样本条件价值的挑选策略来指导各视图下分类器的学习过程. 在 UCI 数据集和基于网页分类的应用上的实验结果表明本文算法能够更有效地利用无标记样本, 可显著地提高算法效率, 较好地优化了学习曲线.

CVCOT 算法的关键在于将主动学习中样本挑选方法应用到 Co-training 中. 主动学习与半监督学习相结合的研究很早就已出现, 通过在半监督学习中嵌入主动学习 (即主动半监督学习), 可以有效地提高分类器的泛化性能, 相关研究有: 1) 单分类器条件下, 主动学习与 EM 或自学习相结合[22−23]; 2) 两分类协同条件下, 主动学习与 Co-training 相结合[24−26]; 3) 多分类器协同条件下, 主动学习与多分类器半监督学习算法相结合[27] 等. 这类算法的主要思想为, 在分类器的每轮迭代中, 挑选部分样本交付人工标记, 利用这些新标记样本来优化半监督学习. 与主动半监督学习算法不同的是, 本文算法并不需要人工查询无标记样本的真实标记, 只是利用主动学习中挑选样本的方法来指导协同训练中新增样本的挑选过程.

1 问题的提出

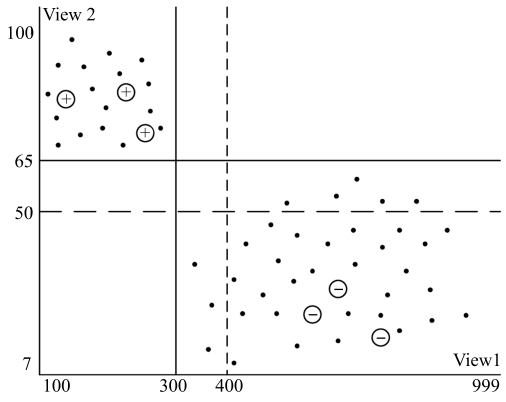
Co-training 解决二元分类问题的一般形式为: 半监督样本集 $L \cup U$ 具备两不同视图 (自然分割的属性集): $X = X^{(1)} \times X^{(2)}$. 分类器模型 f 在 L 的不同视图上分别训练分类器 h_1, h_2, 两分类器互为对方从 U 中挑选样本加以标记, 以此扩充对方的训练集, 即 $L_1 \leftarrow L \cup E_2, L_2 \leftarrow L \cup E_1$, 其中 L_1, L_2 分别对应 h_1, h_2 的训练集, E_1 对应 h_1 为 h_2 挑选的新增样本集, E_2 对应 h_2 为 h_1 挑选的新增样本集. 通过迭代地更新训练集的方式来优化分类器, 直至迭代终止. Co-training 与自学习 (Self-training) 同属于自助式算法, 都是根据分类器的后验概率输出, 选择利用标记置信度最高的样本来更新训练集.

1.1 自学习的困境

自学习[28] 是最早的一种半监督学习算法, 因其简单实用性, 该算法已广泛地应用在自然语言处理等领域. 在自学习的迭代过程中, 分类器对无标记样本进行预测类别, 利用分类结果最为确定的样本更新训练集. 从主动学习的角度来看, 分类不确定性越大, 样本信息度就越高, 其对分类器的价值就越大. 因此自学习的困境在于新增样本的标记可靠性和信息度是不可兼得的, 为保证新增样本的标记可靠性, 自学习只能选取那些分类结果最为确定的无标记样本, 而这类样本对分类器的价值很小. 然而, Co-training 可以很好地避免上述问题. Co-training 中新增样本仍根据分类器的后验概率输出来挑选, 不同的是, Co-training 中各视图下的分类器所对应的新增样本是由对方来挑选的, 这样可避免自学习中所遇到的困境. 文献 [15] 指出在满足视图间条件独立的前提下, Co-training 中新增样本的信息度等同于随机抽样中样本的信息度.

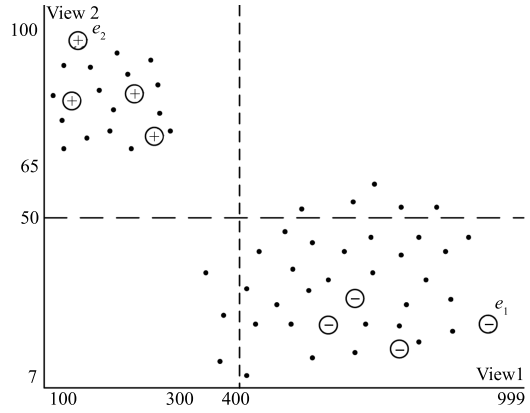
1.2 Co-training 的不足

Co-training 中采用的挑选策略并不能保证新增样本的富信息度. 尤其在实际应用中, 视图间存在一定的相关性, 此时新增样本的信息度会更低. 通过下面一个简单例子可以解释 Co-training 所采用的新增样本挑选策略的局限性. 图 1 表示一个具有二视图的二元分类问题. 样本在每个视图上的特征都为 1 维, 分别属于正反两个类别, 无标记样本由 “•” 表示, 有标记样本由 “◯” 表示 (“+” 和 “−” 分别代表两个不同类别).



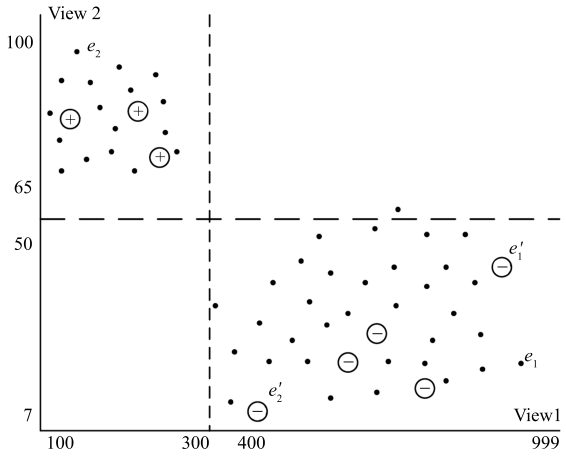
(a) 初始阶段

(a) Initial training stage



(b) Co-training 中一次迭代结果

(b) Retrained after one iteration of co-training



(c) 更优的选择

(c) A better choice

图 1　具有二视图的二元分类问题

Fig. 1　A binary classification problem with 2 views

图 1 (a) 中两条实线分别表示两视图 View 1, View 2 下的目标概念 (Target concept) 的决策界面 (两个特征值最近的不同类别样本间的中心线). 利用初始有标记样本训练出各视图下的分类器 h_1, h_2, 图 1 (a) 中两条虚线分别对应 h_1, h_2 的决策界面. 样本与决策界面之间的距离越大, 分类器对该样本的分类结果越确定, 其标记置信度就越高; 相反, 距离越小, 分类器对该样本的分类结果越不确定, 该样本的信息度就越大, 对决策界面优化的价值越高. Co-training 中两视图下的初始分类器分别对无标记样本进行预测, 各自挑选分类结果最为确定的样本, 加入对方的训练集中. 不失一般性, 假设每轮迭代, 各分类器只挑一个样本. 根据 Co-training 的挑选策略, 标记置信度最高的样本 e_1, e_2 分别被加入到 h_2, h_1 的训练集中. 然而, 分类器的决策界面并没有得到有效的优化, 如图 1 (b) 所示. 这是因为在 View 1 下, e_2 与 h_1 之间的距离也很大, e_2 对于 h_1 的信息度很小, 同理 e_1 对于 h_2 的信息度也很小. 相反, 采用另一种挑选策略, 如图 1 (c) 所示, 挑选 e_1', e_2' 分别来扩充 h_2, h_1 的训练集可以有效地优化它们的决策界面. 尽管 e_1' 的标记置信度略低于 e_1, 但 e_1' 与 h_2 之间的距离要比 e_1 的要小得多, 因此 e_1' 对 h_2 的价值更高; 同理, 相比 e_2, e_2' 对 h_1 的价值更高.

根据上述情形可知, Co-training 采用的新增样本挑选策略只考虑了新增样本的标记可靠性, 而忽视了样本的信息度, 各视图下的分类器被动地接受对方为其挑选的新增样本, 这种策略可能导致分类器的模型得不到有效的优化, 影响学习效率. 采用一种不同的新增样本挑选策略可以有效地避免 Co-training 中存在的问题. 这种策略以无标记样本的条件价值为依据, 不仅考虑新增样本的标记可靠性, 而且最大化样本的价值.

2　CVCOT 算法

CVCOT 算法通过一种基于样本条件价值的挑选新增样本的策略来改进 Co-training. 首先, 给出样本条件价值的定义, 然后给出基于样本条件价值来挑选新增样本的策略, 最后给出算法步骤.

2.1　基于样本条件价值的挑选策略

样本的条件价值 (Conditional value) 在本文中的定义为: 在样本的类别标记已知的前提下, 该样本对于当前分类器模型的价值. 主动学习认为, 样本的价值可通过样本的信息度来衡量, 将富信息样本 (Informative example) 加入训练中可以大大地减小假设空间的大小, 提高学习效率[22]. 对于主动学习, 样本的真实标记 (Ground-truth) 可通过人工查询来获取, 此时无标记样本 x 的条件价值由样本信息度决定: $CV_{al}(x) \propto Inf(x)$, 其中 $Inf(x)$ 表示样本信息度. 而对于半监督学习, 尽管无法获知无标记样本的真实标记, 但可通过分类器输出等途径提供其

"伪标记". 此时, 无标记样本 x 的条件价值可表示为

$$CV_{ssl}(x) \propto Inf(x)P(\hat{y} = groundtruth|x) \quad (1)$$

其中, $P(\hat{y} = groundtruth|x)$ 表示伪标记 $\hat{y}$ 的置信度.

在 Co-training 二视图的框架下, 不同视图下样本的分布不同, 因此无标记样本在不同视图下的条件价值不同. 对于二元分类问题, 已知无标记样本 $x = (x^{(1)}, x^{(2)})$, View 1 下分类器 h_1, View 2 下分类器 h_2, 则定义 x 在 View 1 下的条件价值 $CV^{(1)}(x)$ 为: 在 h_2 给出样本 x 伪标记的条件下, 该样本样本对于 h_1 的价值, 如下式所示:

$$CV^{(1)}(x) \propto Inf(x^{(1)})P(\hat{y}^{(2)} = groundtruh|x^{(2)}) \quad (2)$$

其中, $Inf(x^{(1)})$ 为 x 对分类器 h_1 的信息度, $P(\hat{y}^{(2)} = groundtruth|x^{(2)})$ 为分类器 h_2 输出的标记置信度. 同理可得, x 在 View 2 下的条件价值 $CV^{(2)}(x)$:

$$CV^{(2)}(x) \propto Inf(x^{(2)})P(\hat{y}^{(1)} = groundtruth|x^{(1)}) \quad (3)$$

下面通过刻画样本信息度给出 Co-training 中无标记样本条件价值的最终形式.

主动学习中刻画样本信息度的方法有多种[9]. 借鉴主动学习中查询策略理论, 本文通过分类不确定性 (Uncertainty) 和样本代表性 (Representative) 共同来衡量无标记样本的信息度 $Inf(x)$. 该方法认为, 分类不确定性越大、代表性越强, 样本的信息度就越大[29]. 采用 "熵" 来度量分类不确定性. 熵越大, 分类不确定性就越大, 其具体形式如下:

$$H(x) = -\sum_{y} P(y|x) \log P(y|x) \quad (4)$$

其中, y 取值为所有可能的标记. 样本的代表性可用样本区域密度来表示, 区域密度可通过该样本与其他所有样本间的平均距离来度量. 距离越小, 区域密度就越大, 该样本就越具代表性, 因此, 样本区域密度 $Ds(x)$ 可表示为

$$Ds(x) = \mathrm{e}^{-\frac{1}{|D|}\sum\limits_{x_i \in D} distance(x, x_i)} \quad (5)$$

其中, $distance(x, x_i)$ 表示样本 x 与 x_i 之间的距离. 通过考虑样本的代表性可以避免将离群点 (Outlier) 误认为富信息样本. 图 2 中所示 A 点比 B 点的分类不确定性更大, 但是 B 点比 A 点更富信息, 因而 B 点对分类器的价值更高.

结合式 (4) 和式 (5), 样本的信息度可表示为

$$Inf(x) \propto H(x) \times (Ds(x)) \quad (6)$$

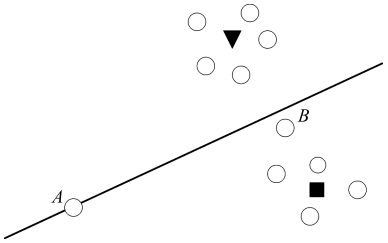


图 2 离群样本示意图

Fig. 2 An illustration of the outlier

另外, 在二元分类问题中, 分类器给出的标记置信度与分类熵成反比, 即分类结果越确定, 标记置信度越高, 分类熵就越低, 因此, 式 (1) 中 $P(\hat{y} = groundtruth|x)$ 可表示为

$$P(\hat{y} = groundtruth|x) \propto (1 - H(x)) \quad (7)$$

结合式 (2), (3), (6) 和 (7) 可得, 在两视图协同训练框架下, 样本 x 在 View 1 下的条件价值为

$$CV^{(1)}(x) \propto H(x^{(1)})(Ds(x^{(1)}))(1 - H(x^{(2)})) \quad (8)$$

同理, 样本 x 在 View 2 下的条件价值为

$$CV^{(2)}(x) \propto H(x^{(2)})(Ds(x^{(2)}))(1 - H(x^{(1)})) \quad (9)$$

因此, 在为不同视图下的分类器挑选新增样本时, 可以分别依据无标记样本在不同视图下的条件价值, 优先将条件价值大的样本补充到训练集中.

2.2 CVCOT 算法步骤

CVCOT 的关键在于如何挑选新增样本, 基于上文给出的方法, CVCOT 算法步骤如下:

输入. 具有二视图的二元分类问题, $D = L \cup U$;

有标记训练集 L, 无标记样本集 U, $X = X^{(1)} \times X^{(2)}$;

分类器模型 f;

n: 算法迭代次数;

k: 每次迭代中挑选的新增样本的数目.

输出. 两视图下分类器 $\hat{h}_1$, $\hat{h}_2$.

初始化. 构造两分类器的初始训练集: $L_1 \leftarrow L$, $L_2 \leftarrow L$;

对 $\forall x_u \in U$, 分别计算其两视图下区域密度: $Ds(x_u^{(1)})$, $Ds(x_u^{(2)})$.

迭代运行 n 次:

步骤 1. 训练分类器: $h_1 \leftarrow f(L_1, X^{(1)})$, $h_2 \leftarrow f(L_2, X^{(2)})$;

步骤 2. 分类器分别对 U 中所有无标记样本进行预测标记;

步骤 3. 对 $\forall x_u \in U$, 通过式 (8) 和式 (9) 分别计算样本在不同视图下的条件价值 $CV^{(1)}(x)$, $CV^{(2)}(x)$;

步骤 4. 从 U 中挑选前 k 个 $CV^{(1)}(x)$ 最大的样本, 结合 h_2 给出的标记, 形成新增样本集 E_{21};

步骤 5. 从 U 中挑选前 k 个 $CV^{(2)}(x)$ 最大的样本, 结合 h_1 给出的标记, 形成新增样本集 E_{12};

步骤 6. 更新样本集: $L_1 \leftarrow L_1 \cup E_{21}$, $L_2 \leftarrow L_2 \cup E_{12}$, $U \leftarrow U - (E_{12} \cup E_{21})$.

CVCOT 与 Co-training 的差异在于: 1) 对于 Co-training, 两视图下的分类器互为对方挑选新增样本, 而本身并不参与其中. 对于 CVCOT, 所有新增样本都是由两分类器来共同挑选, 只不过对于不同视图下的分类器, 挑选结果不同; 2) Co-training 以样本的标记置信度 — 分类器后验概率输出作为挑选样本的依据, 而 CVCOT 以样本的条件价值作为挑选样本的依据.

3 实验结果与分析

本文分别在 UCI 数据集[30] 和基于网页分类的应用进行分类实验. 通过实验验证本文提出的新增样本挑选策略的有效性. 借鉴文献 [18], 本文构造两个多视图自学习衍生算法 Self-training 1、Self-training 2, 与 Co-training 一起作为实验对比算法. Self-training 1 和 Self-training 2 的初始条件与 Co-training 相同, Self-training 1 中两视图下的分类器分别进行自学习, 各自迭代更新自身的训练集, 而 Self-training 2 算法是利用两视图下的分类器判断最一致的无标记样本来进行自学习. 其中, 一致性根据分类器输出的标记置信度相乘来衡量. 通过以上 4 种算法比较来说明样本条件价值在新增样本挑选过程中的重要性. 在测试集上验证时, 所有算法都采用两分类器输出的后验概率相乘的集成方法来进行分类.

3.1 在 UCI 数据集上的实验

表 1 列出 10 个 UCI 数据集的信息. 对每个数据集进行随机划分, 将其 25 % 作为测试集, 剩余 75 % 作为训练集. 有标记样本集的比例为 20 %, 无标记样本集的比例为 80 %. 需要注意的是, UCI 数据集不具备多视图. 本文通过将属性集随机分成两个大小相近的互斥子集的方式来构建两视图. 相关研究表明[15, 17−18]: 由于 UCI 数据集的属性集较为冗余, Co-training 在 UCI 数据集上同样能够取得稳定的分类表现.

3.1.1 实验参数设置

选用朴素贝叶斯 (NB) 作为分类器学习模型. 算法迭代次数 $n = 30$. 为公平对比, 对于所有算法, 每轮迭代中新增样本数量 $k = 20$. 为计算样本区域密度 $Ds(x)$, 该实验采用 Euclidean 尺度来衡量样本间距离. 为得到平均性能, 针对 5 个不同的 L 和 U 的随机划分, 所有算法分别在每个数据集上运行 5 次, 取 5 次独立运行的分类错误率的平均值作为算法在该数据集上的分类错误率. 为说明 CVCOT 中采用熵和区域密度来共同衡量样本信息度的好处, 实验中引入一种本文算法的变体, 即仅以分类熵来衡量样本的信息度, 而不考虑区域密度. 该对比算法记为 Variant.

表 1 UCI 数据集的相关描述

Table 1 Characteristics of the UCI datasets

数据集	属性	样本数目	类别	正例/反例 (%)
Australian	14	690	2	55.5/44.5
Bupa	6	345	2	42.0/58.0
Colic	22	368	2	63.0/37.0
Diabetes	8	768	2	65.1/34.9
German	20	1 000	2	70.0/30.0
Hypothyroid	25	3 163	2	4.8/95.2
Ionosphere	34	351	2	35.9/64.1
Kr-vs-kp	36	3 196	2	52.2/47.8
Sick	29	3 772	2	6.1/93.9
Wdbc	30	569	2	37.3/62.7

3.1.2 实验结果与分析

表 2 列出了 5 种算法在 UCI 数据集上的分类结果. 初始错误率 (Initial) 指迭代之前, 算法在测试集上的分类错误率. 所有算法的初始错误率相同. 最终错误率 (Final) 指迭代终止时, 算法在测试集上的分类错误率. 性能提升比 (Improve) 指最终错误率较初始错误率降低的比例. 表中性能提升比最大值用粗体标记.

根据表 2 可知: 1) 同等迭代次数下, CVCOT 能取得较高的分类性能, 说明 CVCOT 算法的学习效率较高; 2) 协同式算法 (Co-training, CVCOT) 在所有数据集上都有性能提升, 相比之下, 自学习式算法 (Self-training 1, Self-training 2) 在一部分数据集上反而降低分类性能, 这表明协同式算法的分类性能优于自学习式算法; 3) CVCOT 在 7 个数据集上胜出 (获得最大性能提升比), 相比之下, Co-training 胜出 2 次, 并且 CVCOT 在大多数数据集上取得的性能提升比要明显大于 Co-training; 4) 与本文算法的变体 Variant 比较, CVCOT 在大多数数据集上的分类错误率更低, 这说明了在样本的挑选过程中, 通过兼顾样本的区域密度, 即样本代表性, 可以避免将离群样本 (Outlier) 加入训练集中, 从而有效地提高分类性能.

表 2 5 种算法在 10 个数据集上的分类错误率

Table 2 Classification rate of 5 algorithms on ten different datasets

数据集	Initial	分类错误率									
		CVCOT		Variant		Co-training		Self-training1		Self-training2	
		Final	Improve (%)	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)
Australian	0.238	0.229	**3.9**	0.231	2.9	0.234	1.7	0.234	1.7	0.247	−3.9
Bupa	0.459	0.453	1.3	0.448	2.4	0.448	**2.4**	0.485	−5.7	0.463	−0.9
Colic	0.207	0.176	**17.6**	0.185	10.6	0.203	1.9	0.224	−8.2	0.287	−38.6
Diabetes	0.250	0.245	2.0	0.238	4.8	0.235	**6.0**	0.267	−6.8	0.253	−1.2
German	0.257	0.223	**15.2**	0.236	8.2	0.253	1.6	0.247	3.9	0.264	−2.7
Hypothyroid	0.025	0.021	**19.1**	0.021	16.0	0.024	4.0	0.027	−8	0.022	−12.0
Ionosphere	0.183	0.142	**28.9**	0.148	19.1	0.134	26.8	0.164	10.4	0.155	15.3
Kr-vs-kp	0.144	0.113	**27.4**	0.121	16.0	0.131	9.0	0.125	13.2	0.117	18.7
Sick	0.043	0.034	26.5	0.039	9.3	0.042	2.3	0.041	4.7	0.031	**28.1**
Wdbc	0.071	0.044	**61.4**	0.051	28.2	0.058	18.3	0.049	31.1	0.074	−4.2

总的来说, 采用基于样本条件价值挑选策略的 CVCOT 的性能最优, 这在一定程度上表明了样本信息度对于分类器优化的重要性. 本文采用基于样本条件价值的挑选策略来改进 Co-training, 优先将富信息样本扩充到训练集中, 可以更有效地优化分类器.

3.2 在网页分类上的应用

3.2.1 数据集描述

网页分类数据集源自 Blum 等的 Co-training 论文[2]. 数据集包含 1 051 个来自 4 所大学计算机系主页的网页. 这些网页可分为多个类别, 将 “课程主页” 类别下的网页看作正例, 那些不是 “课程主页” 的网页看作反例, 这样就形成一个二元分类问题, 其中正例占 22 %, 反例占 78 %. 该网页分类问题既可以根据网页本身包含的信息来对网页进行正确分类, 也可以利用链接到该网页的超链接所包含的信息来进行正确分类, 这样的网页数据就有两个充分冗余视图, 刻画网页本身包含的词的属性集构成第一个视图, 而刻画超链接所包含的词的属性集构成第二个视图.

3.2.2 实验参数设置

实验中具体参数设置与文献 [2] 相同. 数据集中 25 % 网页样本作为测试集, 初始有标记样本集 L 仅包含 3 个正例和 9 个反例, 其余全部纳入无标记样本集 U 中. 分类器模型采用朴素贝叶斯, 各分类器每次从 U 中挑选 $k = 12$ 个样本加入训练中, 算法迭代 30 次终止. 基于 Bag of words, 经过 Stemming 和特征选择后, 两视图下属性集的大小分别为 66 和 5. 通过归一化处理, 将基于词频的特征向量转换为各词在词汇表上的分布, 以此将样本间距离转换为计算两分布间的距离. 本文采用一种基于 KL 散度 (KL divergence) 衡量样本间距离:

$$Distance(x_j, x_i) = D(x_j||(\alpha x_i + (1-\alpha)\bar{x})) \quad (10)$$

其中, α 为平滑系数 (本实验中 α 取 0.5). 那么该样本的区域密度可表示为

$$Ds(x) = \mathrm{e}^{-\frac{1}{|D|}\sum\limits_{x_i \in D} D(x_i||(\alpha x+(1-\alpha)\bar{x}))} \quad (11)$$

其中, $D(\cdot||\cdot)$ 是信息论中的一种衡量两个分布之间差异的尺度. 两个分布 $P_1(C)$ 和 $P_2(C)$ 之间的 KL 散度为

$$D(P_1(C)||P_2(C)) = \sum_{j=1}^{|C|} P_1(C)\log\left(\frac{P_1(c_j)}{P_2(c_j)}\right) \quad (12)$$

文献 [23] 指出, KL 散度在描述基于文本特征的样本间距离的效果比 Euclidean 距离等更好.

3.2.3 实验结果与分析

通过 5 次不同的 L 取值, 算法在该数据集上运行 5 次. 表 3 列出 4 种算法的平均分类错误率. 实验结果表明, CVCOT 能有效地利用无标记样本提高分类性能. CVCOT 对分类性能的提升达到 50.7 %, Co-training 为 38.1 %, 两种自学习式算法分别为 18.3 % 和 5.6 %, 这说明了协同式算法在具有二视图的分类问题上的表现要优于自学习式算法. CVCOT 性能优于 Co-training, 表明了改进的新增样本挑选策略能够优化 Co-training 算法.

为分析算法的学习曲线, 图 3 给出在一次运行过程中, 各算法的错误率随着训练过程的迭代变化情况. 图中 “TrainALL” 标签对应在整个训练集上 $(L \cup U)$ 进行监督学习所得的分类错误率.

表 3　4 种算法在网页分类问题上的表现

Table 3　Performance of four algorithms on the web classification problem

分类错误率								
Initial	CVCOT		Co-training		Self-training1		Self-training2	
	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)	Final	Improve (%)
12.6	6.2	50.7	7.8	38.1	10.3	18.3	11.9	5.6 %

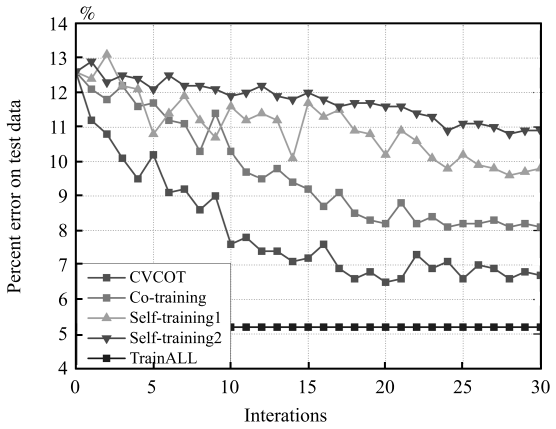


图 3　各对比算法的学习曲线

Fig. 3　The learning curves of comparing algorithms

分析图 3 可知: 1) 协同式算法比自学习式算法更能有效地利用无标记样本; 2) 从首轮迭代开始, CVCOT 的分类错误率一直低于其他 3 种算法, 学习曲线波动相对较少, 可看出 CVCOT 性能较为稳定; 3) CVCOT 算法的学习曲线下降较快, 特别是在前 10 轮迭代中, 这表明 CVCOT 比 Co-training 等算法的学习效率更高, 更能有效地利用无标记样本优化分类器; 4) 相比其他 3 种算法, CVCOT 能较早地收敛, CVCOT 迭代 10 次的分类错误率已经低于 Co-training 迭代 30 次后的水平. CVCOT 的高学习效率归因于其采用的新增样本挑选策略, 该策略优先将那些富信息样本加入训练集中, 能有效地优化分类器.

为分析初始有标记训练集的大小对算法性能的影响, 本文分别在 10 个不同大小的 L ($|L| = \{12, 24, 36, 48, 60, 72, 84, 96, 108, 120\}$) 上进行分类实验, 所有算法各运行 5 次. 图 4 表示在不同大小的初始训练集下, 各算法在测试集上的分类错误率及方差. "Baseline" 标签对应算法的初始分类错误率. 从图中可以看出, 在 $|L|$ 为 12 时, CVCOT 与其他算法在分类错误率上的差距最大. 随着 $|L|$ 的增大, 这种差距逐渐减小, 当 $|L|$ 为 120 时, CVCOT 和 Co-training 的分类错误率基本相近. 这表明 CVCOT 能够在初始训练样本较少的条件下, 取得较为显著的性能优势. 当初始训练样本较多时, 初始分类器的性能较强时, 此时无标记样本的信息度相对较低, 而样本标记置信度为挑选新增样本的主导因素, 此时 CVCOT 近似等同于 Co-training, 这在一定程度上解释了为何随着 $|L|$ 的增大, CVCOT 与 Co-training 性能趋于相近. 另外, 在给定 L 的条件下, CVCOT 和 Self-training 2 的方差较小, Co-training 和 Self-training 1 的方差较大. 这表明本文算法对初始训练样本的选择不敏感, 泛化能力较强.

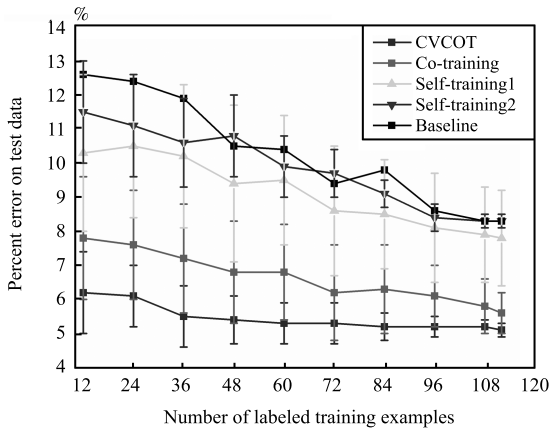


图 4　各算法在不同有标记训练样本大小下的分类结果

Fig. 4　Classification performance of each algorithm under different numbers of labeled training examples

4　结论

本文提出了一种改进的 Co-training 算法. 该算法采用一种基于样本条件价值的挑选策略来指导各视图下分类器的学习过程. 本文从主动学习的角度出发, 给出了无标记样本的条件价值及其定量刻画, 以此形成一种新的样本挑选策略. 该策略不仅保证了新增样本的标记可靠性, 又能优先将价值更大的富信息样本补充到训练集中, 从而有效地提高学习效率. 在 UCI 数据集和网页分类应用上的对比实验表明, 新算法能够显著地提高分类器的学习效率, 有效地利用无标记样本来提升性能. 尤其当初始训练样本较少时, 本文算法的优势更为明显.

References

1 Chapelle O, Schölkopf B, Zien A. *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006

2 Blum A, Mitchell T. Combining labeled and unlabeled data

with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. Wisconsin, MI: ACM, 1998. 92−100

3 Zhu X J. Semi-supervised Learning Literature Survey, Computer Science Technical Report 1530. University of Wisconsin Madison, USA, 2008

4 Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. Pittsburgh, PA, 2001. 1−9

5 Steedman M, Osborne M, Sarkar A, Clark S, Hwa R, Hockenmaier J, Ruhlen P, Baker S, Crim J. Bootstrapping statistical parsers from small datasets. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Budapest, Hungary: Association for Computational Linguistics Stroudsburg, 2003. 331−338

6 Li M, Li H, Zhou Z H. Semi-supervised document retrieval. *Information Processing and Management*, 2009, **45**(3): 341−355

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

8 Mavroeidis D, Chaidos K, Pirillos S, Vazirgiannis M. Using tri-training and support vector machines for addressing the ECML-PKDD 2006 discovery challenge. In: Proceedings of the 2006 ECML-PKDD Discovery Challenge Workshop. Berlin, Germany, 2006. 39−47

9 Settles B. Active Learning Literature Survey, Computer Science Technical Report 1648, University of Wisconsin-Madison, USA, 2009

10 Singh A, Nowak R D, Zhu X J. Unlabeled data: now it helps, now it doesn′t. *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2008. 1513−1520

11 Dasgupta S, Littman M L, McAllester D. PAC generalization bounds for co-training. *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2001. 375−382

12 Balcan M, Blum A, Yang K. Co-training and expansion: towards bridging theory and practice. *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2005. 89−96

13 Wang W, Zhou Z H. A new analysis of co-training. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010. 1135−1142

14 Du J, Ling C X, Zhou Z H. When does cotraining work in real data? *IEEE Transactions on Knowledge and Data Engineering*, 2011, **23**(5): 788−799

15 Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: Proceedings of the 9th ACM International Conference on Information and Knowledge Management. McLean, VA: ACM, 2000. 86−93

16 Zhou Z H, Li M. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 2010, **24**(3): 415−439

17 Goldman S A, Zhou Y. Enhancing supervised learning with unlabeled data. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2000. 327−334

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

19 Li M, Zhou Z H. SETRED: self-training with editing. In: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hanoi, Vietnam: Springer-Verlag, 2005. 611−621

20 Deng Cao, Guo Mao-Zu. ADE-Tri-training: tri-training with adaptive data editing. *Chinese Journal of Computers*, 2007, **30**(8): 1213−1226
(邓超, 郭茂祖. 基于自适应数据剪辑策略的 Tri-training 算法. 计算机学报, 2007, **30**(8): 1213−1226)

21 Zhang M L, Zhou Z H. CoTrade: confident co-training with data editing. *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics*, 2011, **41**(6): 1612−1626

22 Chen Rong, Cao Yong-Feng, Sun Hong. Multi-class image classification with active learning and semi-supervised learning. *Acta Automatica Sinica*, 2011, **37**(8): 954−962
(陈荣, 曹永锋, 孙洪. 基于主动学习和半监督学习的多类图像分类. 自动化学报, 2011, **37**(8): 954−962)

23 MaCallum A, Nigam K. Employing EM in pool-based active learning for text classification. In: Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1998. 350−358

24 Muslea I, Minton S, Knoblock C A. Active + Semi-supervised learning = Robust multi-view learning. In: Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia: Morgan Kaufmann Publishers Inc, 2002. 435−442

25 Muslea I, Minton S, Knoblock C A. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 2006, **27**(1): 203−233

26 Zhou Z H, Chen K J, Dai H B. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 2006, **24**(2): 219−244

27 Li M, Zhang H Y, Wu R X, Zhou Z H. Sample-based software defect prediction with active and semi-supervised learning. *Automated Software Engineering*, 2012, **19**(2): 201−230

28 Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1995. 189−196

29 Lewis D D, Gale W A. A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: Springer-Verlag, 1994. 3−12

30 Asuncion A, Newman D J. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml/datasets.html, January 10, 2010

程圣军 哈尔滨工业大学计算机科学与技术学院博士研究生. 主要研究方向为数据挖掘, 机器学习, 模式识别. 本文通信作者. E-mail: hitwer@gmail.com
(**CHENG Sheng-Jun** Ph. D. candidate at the School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers data mining, machine learning, and pattern recognition. Corresponding author of this paper.)

刘家锋 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为模式识别, 机器学习, 图像处理, 图像理解与机器视觉.
E-mail: jefferyliu@hit.edu.cn
(**LIU Jia-Feng** Associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers pattern recognition, machine learning, image processing, image understanding, and computer vision.)

黄庆成 哈尔滨工业大学计算机科学与技术学院副教授. 主要研究方向为模式识别, 智能机器人, 多 agent 系统.
E-mail: huangqc@hit.edu.cn
(**HUANG Qing-Cheng** Associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers pattern recognition, intelligent robot, and multi-agent system.)

唐降龙 哈尔滨工业大学计算机科学与技术学院教授. 中国计算机学会高级会员. 主要研究方向为光学字符识别, 生物特征识别, 图像处理及模式识别.
E-mail: tangxl@hit.edu.cn
(**TANG Xiang-Long** Professor the at School of Computer Science and Technology, Harbin Institute of Technology, and senior member of China Computer Federation. His research interest covers OCR, biometrics, image processing, and pattern recognition.)