计算机科学
COMPUTER SCIENCE

基于 DECORATE 集成学习与置信度评估的 Tri-training 算法

王宇飞，陈文

引用本文

王宇飞，陈文. 基于 DECORATE 集成学习与置信度评估的 Tri-training 算法[J]. 计算机科学, 2022, 49(6): 127-133.

WANG Yu-fei, CHEN Wen. Tri-training Algorithm Based on DECORATE Ensemble Learning and Credibility Assessment[J]. Computer Science, 2022, 49(6): 127-133.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100043

基于 DECORATE 集成学习与置信度评估的 Tri-training 算法

王宇飞　陈　文
四川大学网络空间安全学院　成都 610065
(wangyufei079@foxmail.com)

摘　要　Tri-training 是一种基于分歧的半监督学习算法,同时利用了半监督学习和集成学习机制。Tri-training 能有效地利用少量有标记样本和大量无标记样本,通过分类器间的相互协同和迭代来提升模型性能。但是在已标记样本量不足的情况下,Tri-training 生成的初始分类器训练不足,并且在分类器间协同标记的过程中可能产生误标记的噪声数据。针对上述问题,提出了一种结合 DECORATE 集成学习、多样性度量与置信度评估的协同学习算法。该算法基于 DECORATE 集成学习方法,通过添加差异化的人工样本和标记来训练多种偏好的基分类器,以提升分类泛化能力。该算法还基于 JS 散度对分类器进行多样性度量和筛选,以最大化基分类器多样性,同时在迭代过程中基于标签传播算法对伪标记样本进行置信度评估,以减少噪声数据。在 UCI 数据集上进行了分类实验,结果表明,相比 Tri-training 算法及其改进算法,所提算法具有更高的分类准确率和 $F1$ 分数。

关键词:基于分歧的半监督学习;集成学习;置信度评估;多样性度量
中图法分类号　TP181

Tri-training Algorithm Based on DECORATE Ensemble Learning and Credibility Assessment

WANG Yu-fei and CHEN Wen
School of Cyber Science and Engineering,Sichuan University,Chengdu 610065,China

Abstract　Tri-training is a disagreement-based semi-supervised learning algorithm,in which both semi-supervised learning and ensemble learning mechanisms are simultaneously applied. It can improve the model performance by effectively leveraging some labeled samples along with a large amount of unlabeled ones through collaborations and iterations among basic classifiers. However,when the labeled sample size is insufficient,the initial classifiers generated by Tri-training are not sufficiently trained. Furthermore,mislabeled noisy data might be generated during the collaborative labeling process among the classifiers. Aiming at these problems,a collaborative learning algorithm is proposed,which combines DECORATE ensemble learning,diversity measure and credibility assessment. In our method,to improve the generalization performance,multiple preference classifiers are generated based on DECORATE with differentiated artificial data and labels,and the diversities of classifiers are measured and selected by Jensen-Shannon divergence to maxmize the diversity of the classifiers. At the same time,the credibility of the pseudo labeled samples is assessed during the iterations by a label propagation algorithm to reduce the noisy data. The results of classification experiment on UCI data sets demonstrate that the proposed algorithm achieves higher accuracy and $F1$-score than Tri-training algorithm and its improved versions.

Keywords　Disagreement-based semi-supervised learning,Ensemble learning,Credibility assessment,Diversity measure

1　引言

传统的机器学习算法一般可以分为监督学习、无监督学习与半监督学习[1-2]。无监督学习的目的是将观察到的数据点进行无标签的组织与划分,其对数据的标签不作要求;监督学习旨在利用大量的已标记样本进行学习,由学得的模型对新样本进行预测。然而,在众多应用领域,如网络入侵检测、恶意代码检测、医学疾病诊断、行为模式判断等,要获取大量的有标记数据信息,需要专家对样本进行逐一的人工标记,整个过程会耗费大量的人力、物力[3],而大量的无标记样本的

到稿日期:2021-11-03　返修日期:2022-03-02
基金项目:国家重点研发计划(020YFB1805405,2019QY0800);国家自然科学基金(U1736212,61872255,U19A2068);模式识别与智能信息处理四川省高校重点实验室(MSSB-2020-01)
This work was supported by the National Key Research and Development Program of China(020YFB1805405,2019QY0800),National Natural Science Foundation of China(U1736212,61872255,U19A2068) and Key Laboratory of Pattern Recognition and Intelligent Information Processing,Institutions of Higher Education of Sichuan Province(MSSB-2020-01).
通信作者:陈文(wenchen@scu.edu.cn)

获取则相对容易，成本低廉。半监督学习介于监督学习与无监督学习之间，基于数据分布上的模型假设，它能够在已标记样本较少的情况下，利用大量的无标记样本对模型进行迭代更新训练，不断扩充已标记样本集，一定程度上弥补了监督学习因样本量过少造成模型性能不佳的缺陷，提升了模型性能。

半监督学习算法大致可以分为基于生成式模型的方法、基于图的方法[4]以及基于分歧的方法等，其中基于分歧的方法的代表为 Co-training 协同学习算法[5]。Co-training 采用两个分类器分别在不同的有标记样本的数据视图上进行训练，然后每个分类器在无标记样本中筛选出置信度高于阈值的部分样本，根据自身分类结果打上标签后，加入另一个分类器的训练集中，用于另一个分类器的迭代更新。Chen 等提出了一种改进的 Co-training 算法[6]，该算法定义了无标记样本的条件价值，在每轮迭代中优先选择条件价值较高的富信息样本加入训练集，然后进行分类器更新。Katz 等提出垂直集成协同学习算法[7]，在算法迭代过程中保存不同迭代轮次的分类器，在不需要额外计算成本的情况下对这些分类器进行集成，提升算法的分类性能。Lu 等提出利用信息熵将数据集划分为信息量相同的两个视图，然后分别采用聚类准则和置信度准则选取未标记样本进行标记[8]。Co-training 系列算法要求数据的两个视图之间满足条件独立性且充分冗余，然而在多数应用场景中，获取的数据很难满足这些条件。因此，Zhou 等在 Co-training 的基础上对其进行扩展，提出了一种新的协同学习算法 Tri-training[9]，该算法结合了 Bagging 集成学习的思想，在已标记样本集上进行 Bootstrap 重采样训练 3 个基分类器。针对无标记样本，基分类器两两组合，为另一个基分类器提供新样本标签，既不要求数据拥有充分和冗余的视图，也不对监督学习算法即基分类器施加约束，适用性更加广泛。Xu 等的研究表明[10]，在缺乏足够标记样本的情况下，Tri-training 与监督学习具有相同的性能。Li 等[11]提出使用 Tri-training 来改进 SVM 算法，3 个分类器分别采用不同的核函数，并将其应用到入侵检测领域。Søgaard[12]提出了带分歧的 Tri-training 算法，即只有两个分类器标签判别相同、第 3 个分类器判别不同的样本，才加入第 3 个分类器的更新训练集，其核心思想是模型只需要在薄弱的地方进行提升，减少了标记开销。Ruder 等[13]结合迁移学习思想提出了多任务 Tri-training 算法，实现了跨模型知识共享和训练加速。Zhang 等提出使用交叉熵代替原算法中的错误率，并结合凸优化方法提出了一种基于交叉熵的安全 Tri-training 学习框架[14]，有效地防止分类性能下降的情况出现。

然而，在初始已标记样本数量较少时，Tri-training 算法采用 Bootstrap 重采样方法生成的 3 个基分类器可能会因训练不足而影响最终的分类性能。这是因为 Bootstrap 重采样的实质是对样本进行多次的重复有放回抽样，进而对样本的总体分布特性进行统计推断。但是 Tri-training 算法仅进行了 3 次重复有放回抽样，在样本量较少时，每次抽样得到的样本子集可能包含多个重复样本，而总体样本集中的大部分样本却没有被抽取到，因此可能与总体样本集差异较大，仅能反映总体样本的少部分数据分布特性，用这样的样本子集训练基分类器会导致泛化性能较差。另外，考虑到基分类器本身的限制以及初始样本集质量等因素，在标记过程中可能会引入较多误标记的噪声数据，特别是当 Tri-training 算法错误率评判指标下降到接近零时，可能会引入大量噪声数据，并随着迭代训练进一步扩大噪声的影响，导致分类器性能下降。

针对上述问题，本文提出了一种结合 DECORATE 集成学习[15]、多样性度量以及置信度评估的 Tri-training 算法 TDDC(Tri-training Algorithm Based on DECORATE Ensemble Learning, Diversity Measure and Credibility Assessment)，本文的主要贡献如下：

(1)提出在生成 Tri-training 的 3 个初始基分类器时，采用 DECORATE 集成学习算法，通过在原始数据集中添加差异化的人工合成样本和标记，来代替传统的 Bootstrap 重采样训练多种偏好的初始基分类器，提高分类性能。

(2)提出基于 JS 散度和无标记样本的基分类器多样性度量方法，最大化基分类器多样性。

(3)将标签传播算法[16]与 Tri-training 算法相结合，对每一轮新标记的样本进行基于标签传播算法的置信度评估，有效地减少噪声数据的影响，同时限制更新训练集样本规模，避免在 Tri-training 算法错误率评判指标很小时导致更新训练集过大，引入大量噪声数据。

本文的结构如下：第 2 节介绍相关工作，包括传统的 Tri-training 算法和 DECORATE 集成学习算法；第 3 节介绍 TDDC 的主要思路；第 4 节通过在 UCI 数据集上的实验验证本文提出的 TDDC 算法的有效性；最后总结全文。

2 相关工作

2.1 Tri-training 算法

Tri-training 算法[9]弥补了 Co-training 算法对数据视图要求苛刻的缺陷，采用 3 个基分类器进行协同学习。该算法同时利用了半监督学习和集成学习机制[17]，具体过程如下：首先，为了避免在同一数据视图下训练分类器造成的趋同现象，在已标记样本集上进行了 3 次 Bootstrap 重采样，生成了 3 个有差异性的初始训练集，用于对 3 个基分类器的初始训练；然后，3 个基分类器之间两两组合，对无标记样本集进行分类并打上伪标签，2 个基分类器标记一致的样本则加入第 3 个基分类器的更新训练集，与已标记样本集一起，用于对基分类器的更新训练，重复迭代这个过程，直到分类器的性能不再提升为止。训练完成后，采用多数投票法输出分类结果。

显然，在算法迭代过程中对无标记样本进行标记时，基分类器两两分类结果相同的样本，并不能确保其伪标签的准确性，很有可能两个分类器同时出现误判，为第 3 个基分类器的更新训练集中添加了标记错误的噪声样本，导致分类性能下降。因此，基于 Angluin 等提出的噪声学习理论[18]，Zhou 等提出[9]当新标记的样本满足式(1)的条件时，可以利用无标记样本来补偿噪声数据带来的影响。

$$|L\cup L^t|\left(1-2\frac{\eta_L|L|+e_i^t|L^t|}{|L\cup L^t|}\right)^2>|L\cup L^{t-1}|\left(1-2\frac{\eta_L|L|+e_i^{t-1}|L^{t-1}|}{|L\cup L^{t-1}|}\right)^2 \tag{1}$$

其中，L^t 表示第 t 轮迭代中基分类器 h_j 和 h_k 为 h_i 标记的更新训练集，e_i^t 代表第 t 轮 h_j 和 h_k 的分类错误率，η_L 代表已标记样本集 L 的噪声上限。一般情况下 η_L 的值非常小，则当 $|L^{t-1}|<|L^t|$ 时，可以得出式(2)，用于判断当前迭代轮次为

h_i 分类器标记的样本能否加入其更新训练集 L' 中。

$$0<\frac{e_i^t}{e_i^{t-1}}<\frac{|L^{t-1}|}{|L^t|}<1 \tag{2}$$

尽管 Tri-training 算法取得了较好的效果，但是在标记过程中仍然可能会引入较多误标记的噪声数据，从而影响分类器性能。为减少噪声数据的影响，并且考虑到已标记样本量较少的情况以及基分类器的多样性等问题，本文在 Tri-training 算法的基础上，结合 DECORATE 集成学习、多样性度量和置信度评估方法，提出了 TDDC 算法。

2.2 DECORATE 集成学习算法

DECORATE 算法[15]是 Melville 等提出的一种集成学习算法，不同于 Bagging 和 Boosting 集成学习算法在已有的训练集上采取一定的策略生成基分类器，DECORATE 算法通过生成人工样本来扩展训练集，从而生成多样化的分类器，具体描述如算法 1 所示。

算法 1 DECORATE 集成学习算法

输入：基分类器 BaseLearner；训练集 L；集成规模 N；最大迭代次数 I；生成的人工样本占训练集的比例 S

输出：集成系统 $C_*=\{C_1,C_2,\cdots,C_n\}$，其中 n 为实际生成的基分类器数目

1. 初始化参数 i←1，trials←1
2. C_1←BaseLearner(L)，C_*←$\{C_1\}$，计算 C_* 对 L 的分类错误率 e
3. while i ＜ N and trials ＜ I do
4. 根据 L 的数据分布生成 S×|L|个人工样本 R
5. 使用 C_* 对 R 进行分类标记，根据分类结果为 R 打上与 C_* 分类结果不同的标签，使新生成的基分类器保持多样性
6. L←L∪R
7. C′←BaseLearner(L)
8. C_*←C_*∪{C′}
9. L←L－R
10. 计算 C_* 对 L 的分类错误率 e′
11. if e′＜=e then
12. i←i+1
13. e←e′
14. else
15. C_*←C_*－{C′}
16. end if
17. trials←trials+1
18. end while

算法 1 中，第 1－2 行在给定的训练集 L 上训练一个基分类器 C_1，得到初始集成系统 $C_*=\{C_1\}$，并计算其在训练集 L 上的分类错误率 e；第 4－5 行根据训练集的数据分布生成 $S\times|L|$ 个近似分布的人工数据 R（对于连续型特征，根据训练集的均值和标准差在对应的正态分布中进行随机采样；对于离散型特征，根据不同数值的频率取值），将 R 交由集成系统 C_* 进行标记。为了使新的基分类器保持多样性，R 的标签要尽量与集成系统 C_* 的分类结果不同，即若某样本类别标签有 1，2，3 这 3 种，集成系统 C_* 分类结果为 1，则将标签 2 或 3 赋予人工数据；第 6－16 行在生成人工数据与标签后，使用 $L\cup R$ 训练新的基分类器 C'，并将 C' 加入集成系统 C_*。为了保证 C' 的加入不会使集成系统性能下降，每轮加入新的基分类器后，集成系统 C_* 都会在 L 上测试分类错误率 e'，若 $e'>e$，则剔除该基分类器。重复迭代此过程，直到集成系统 C_* 达到集成规模或者达到最大迭代次数限制。

DECORATE 算法提供了一种新的生成基分类器的思路，使得集成系统可以在规模更大的数据集上建立。Zhang 等[19]从偏差和方差分解的角度对其有效性进行了解释，并且发现，在训练样本规模较小的情况下，DECORATE 算法性能优于 Bagging 和 Boosting 集成学习算法，并且在训练样本规模较大的情况下，性能也不劣于后两者[15]。

3 TDDC 算法

3.1 基于 DECORATE 集成学习的基分类器生成

Tri-training 在生成初始基分类器时，采用了类似于 Bagging 集成学习的方法，对训练集进行 Bootstrap 重采样得到 3 个有差异性的训练子集，然后在这 3 个子集上训练基分类器，获得 3 个带有分歧的基分类器，避免了 Co-training 对数据视图的严苛要求。

Bootstrap 重采样的实质是对样本进行多次的重复有放回抽样，进而可以对样本的总体分布特性进行统计推断。但是 Tri-training 算法只需要训练 3 个基分类器，因此仅进行了 3 次重复有放回抽样。在已标记样本量较大时，每次抽样得到的样本子集往往能在一定程度上反映总体样本分布。然而，Tri-training 等半监督学习算法的应用场景大多数是在已标记样本较少的情况下，如图 1 所示，此时，每次抽样得到的样本子集可能包含多个重复样本，且重复样本占比较大，而总体样本集中的大部分样本却没有被抽取到，导致样本子集与总体样本集差异较大，仅能反映总体样本的少部分数据分布特性。用这样的样本子集训练基分类器会导致生成的初始基分类器泛化能力较差，从而影响算法的后续迭代，降低最后的协同学习分类精度。

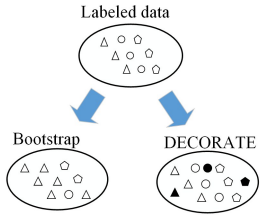

图 1 Bootstrap 和 DECORATE 训练集

Fig. 1 Training sets of Bootstrap and DECORATE

针对上述问题，TDDC 算法将 DECORATE 集成学习算法[15]的基分类器生成方法用于 3 个初始基分类器生成，使用人工数据结合原始数据的方式取代 Bootstrap 重采样来训练基分类器。如图 1 所示，黑色的图形代表生成的人工数据，可以看出，DECORATE 是通过扩展有标签训练集进行基分类器生成，相比 Bagging 采用的 Bootstrap 重采样方法，DECORATE 可以在更大规模的训练集上生成更高精度的基分类器。同时，由于基分类器生成过程中错误率检测机制的存在（见算法 1 第 10－15 行），该方法有效地避免了引入人工数据可能导致的模型性能下降，因此可以在最大程度地利用训练集数据信息的同时获得具有多样性的基分类器，使得基分类器在预测偏好和决策边界上均存在一定的差异性，更有利于 TDDC 的后续迭代训练。

3.2 基于 JS 散度与无标记样本的基分类器多样性度量

集成学习基分类器的多样性是影响集成系统泛化性能的

重要因素之一。若基分类器之间差异较小，则它们的分类结果相似，有可能在同样的样本上分类错误，不能提高集成后的性能；相反，若基分类器之间差异较大，某个基分类器错分的样本就可能被其他基分类器矫正[20]，Melville 等通过实验发现[15]，增加分类器的多样性有助于降低分类错误率，这与 Tri-training 中基分类器间的分歧问题类似。实质上，Tri-training 的 3 个基分类器间相互协同并投票产生最终分类结果就是一种集成学习机制。Wang 等[21]证明，只要协同学习的 PAC 学习器之间存在较大的分歧，就可利用无标记样本通过协同学习算法提升学习器的性能。由此可以得出，使 Tri-training 算法有效的关键之一就是要使 3 个基分类器之间存在一定的分歧。因此我们有必要对基分类器进行多样性度量和筛选，最大化基分类器的多样性。

当使用 DECORATE 算法进行基分类器的生成时，算法本身有一个保持基分类器多样性的机制：人工生成数据 R 的标签要与当前集成系统对 R 的预测结果保持不一致，然后再用于训练基分类器。但是，这个机制只保证了基分类器在人工数据上的分歧。已有的大多数多样性度量方法[22]都是基于已标记样本，由于基分类器在已标记样本上往往已经充分拟合，因此通常它们对其分类的差异不会太大。相较而言，利用大量易于获取的无标记样本进行多样性度量可能更合适[23-24]，因为无标记样本的获取相对容易，并且其从未出现在基分类器的训练集中，所以能更好地体现基分类器的多样性。本文提出结合 JS 散度和无标记样本的多样性度量方法，用于对 DECORATE 生成的基分类器和 TDDC 迭代完成的基分类器的多样性度量。

相对熵又称为 KL 散度(Kullback-Leible Divergence)，常被用于度量两个分布之间的差异性，假设用 X 表示某随机变量，$P(x)$ 和 $Q(x)$ 是 X 上的两个概率分布，则两者之间的 KL 散度定义如式(3)所示：

$$D_{\mathrm{KL}}(P \parallel Q)=\sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)} \tag{3}$$

但是，由于 KL 散度具有非对称性且取值范围较大([0，$+\infty$])，对相似度的判别不够精细，因此 JS 散度被提出。JS 散度是 KL 散度的一种变体，其定义如式(4)所示：

$$D_{JS}(P \parallel Q)=\frac{1}{2}D_{KL}\left(P\middle\|\frac{P+Q}{2}\right)+\frac{1}{2}D_{KL}\left(Q\middle\|\frac{P+Q}{2}\right) \tag{4}$$

可以看出，JS 散度是两个 KL 散度的累加，其很好地解决了 KL 散度非对称的问题，且取值为[0，1]，更加适合用于判别两个分布的相似性。

本文提出将基于 JS 散度与无标记样本的基分类器多样性度量方法与 DECORATE 集成学习算法相结合，用于生成 Tri-training 的基分类器，如算法 2 所示(其他步骤与算法 1 相同，主要修改了 DECORATE 算法的第 7－15 行)。

算法 2 基于基分类器多样性度量的 DECORATE 集成学习算法

输入：基分类器 BaseLearner；训练集 L；集成规模 N；最大迭代次数 I；生成的人工样本占训练集的比例 S；无标记样本集 U

输出：集成系统 $C_*=\{C_1,C_2,\cdots,C_n\}$，其中 n 为实际生成的基分类器数目

1. 初始化参数 i←1，trials←1，U_{js}←0
2. C_1←BaseLearner(L)，C_*←{C_1}，计算 C_* 对 L 的分类错误率 e
3. while i ＜ N and trials ＜ I do
4. 根据 L 的数据分布生成 S×|L|个人工样本 R
5. 使用 C_* 对 R 进行分类标记，根据分类结果为 R 打上与 C_* 分类结果不同的标签，使新生成的基分类器保持多样性
6. L←L∪R
7. C'←BaseLearner(L)
8. C'对无标签数据 U 分类，得到 U_y′
9. C_* 对无标签数据 U 分类，得到 U_y
10. 计算 U_y′和 U_y 的 JS 散度 U'_{js}
11. C_*←C_*∪{C'}
12. L←L－R
13. 计算 C_* 对 L 的分类错误率 e′
14. if e′＜＝e and U'_{js}＞＝U_{js} then
15. i←i＋1
16. e←e′
17. U_{js}←U'_{js}
18. else
19. …

在 TDDC 的 3 个基分类器的生成过程中，每一轮生成基分类器 C' 后，分别用基分类器 C' 和集成系统 C_* 对无标记样本集 U 进行预测，并获得对应的分类结果 U_y 和 U_y'，然后计算两个分类结果分布之间的 JS 散度 U'_{js}，若 U'_{js} 不小于上一轮次的 JS 散度 U_{js}(初始设置为 0)，并且上一轮分类错误率 e 不小于本轮分类错误率 e'，则判定本轮产生的基分类器多样性达标，更新 U_{js} 的值，否则舍弃该分类器，进入下一轮迭代。改进之后，算法能够确保每一轮生成的基分类器与整个集成系统具有一定的差异性，从而有效提高整个集成系统的多样性，有利于降低分类错误率。

由于基分类器之间相互学习，经过迭代之后，基分类器之间会出现趋同现象，导致基分类器间的分歧度减小，从而影响最后的多数投票结果，因此将算法 2 和多样性度量同样用于迭代完成的最终 3 个基分类器上：在迭代结束以后，使用 3 个基分类器对无标记样本集 U 进行分类，得到 3 个分类结果，然后计算两两分类结果之间的 JS 散度，得出平均 JS 散度。若平均 JS 散度小于阈值(取 TDDC 初始阶段生成基分类器过程中的最小 JS 散度)，则对 3 个分类器和各自的最终训练集分别使用算法 2 仅进行一次迭代，生成 3 个具有一定分歧的最终基分类器，用于多数投票。

3.3 基于标签传播算法的置信度评估过程

在传统的 Tri-training 算法中，分类器 h_j 和 h_k 对无标记数据集预测两两一致的样本，若满足噪声学习理论要求，即可加入分类器 h_i 的更新训练集。然而，这并不能保证两个分类器共同标记的伪标签的正确性。若不对伪标记样本进行筛选，采用误标记的噪声样本对分类器进行更新，会使错误积累，最终导致分类器性能提升不明显甚至性能下降。因此，本文提出了一种基于标签传播算法的置信度评估方法，使用更新训练集的伪标记样本对已标记样本进行反向的标签传播，通过比较传播后的标签与真实标签之间的差异度，判定伪标记样本的标记质量，对 TDDC 算法每一轮迭代时的分类器 h_i 的更新训练集进行置信度评估。

标签传播算法不依赖于预定义的优化函数，其核心思想是将已标记样本的标签基于距离相似度迭代传递给邻近样

本，是一种有效的利用无标记样本进行学习的半监督学习算法，适用于本文对更新训练集的置信度评估。对于一个已标记样本 x_i，它与样本 x_j 之间的相似度决定了 x_i 将标签传递给 x_j 的概率。相似度定义如式(5)所示：

$$w_{ij}=\exp\left(-\frac{d_{ij}^2}{\sigma^2}\right)=\exp\left(-\frac{\sum_{d=1}^{D}(x_i^d-x_j^d)^2}{\sigma^2}\right) \tag{5}$$

其中，D 代表样本维数。

在本文使用的标签传播算法中，从已标记样本集中选取部分样本，将其加入分类器 h_i 在迭代过程中的更新训练集，用 $X_L=\{\langle x_1,y_1\rangle,\langle x_2,y_2\rangle,\cdots,\langle x_l,y_l\rangle\}$ 表示，剩余已标记样本集用 $X_U=\{\langle x_1,y_1\rangle,\langle x_2,y_2\rangle,\cdots,\langle x_u,y_u\rangle\}$ 表示，l 和 u 代表两个样本集的规模大小，令 $X=X_L\cup X_U$。式(6)定义了规模为$(l+u)\times(l+u)$的概率传播矩阵 $\boldsymbol{P}$。

$$P_{ij}=\boldsymbol{P}(i\to j)=\frac{w_{ij}}{\sum_{k=1}^{l+u}w_{kj}} \tag{6}$$

其中，P_{ij} 代表将样本 i 的标签传递给样本 j 的概率。

定义$(l+u)\times c$ 的标签矩阵 $\boldsymbol{Y}$，c 表示样本类别数量，Y_{ij} 表示 X 中第 i 个样本 x_i 属于类别 j 的概率，初始时标签矩阵 $\boldsymbol{Y}$ 中前 l 行根据 X_L 中的真实标签设置，其余的值全部设定为 0。

标签传播过程如下：

(1)计算矩阵 $\boldsymbol{P}$ 与矩阵 $\boldsymbol{Y}$ 的乘积 $\boldsymbol{PY}$，令 $\boldsymbol{Y}=\boldsymbol{PY}$；

(2)标准化 $\boldsymbol{Y}$ 的每一行，替换 $\boldsymbol{Y}$ 的前 l 行为 X_L 中的真实标签值；

(3)重复步骤 1 和 2，直到矩阵收敛或者达到迭代次数。

当标签传播迭代停止时，统计 $X_U'=\{\langle x_1,y_1'\rangle,\langle x_2,y_2'\rangle,\cdots,\langle x_u,y_u'\rangle\}$，$y_i'$ 为迭代结束时 X_U 中第 i 个样本 x_i 隶属概率最大的类别标签。计算已标记样本集 X_U 的真实标签与 X_U' 传播结果标签 y_i' 的累积差异即准确率。若准确率低于阈值(取迭代过程中标签传播结果准确率最大值)，则说明本次分类器 h_i 的更新训练集中存在较多误标记的噪声样本，应该舍弃该更新训练集，重新进行筛选；若准确率高于阈值，则结合原始已标记训练集对分类器 h_i 进行更新训练。

传统 Tri-training 算法中，错误率定义如式(7)所示：

$$e=\frac{s}{t} \tag{7}$$

其中，s 表示分类器 h_j 和 h_k 同时预测错误的样本数，t 表示 h_j 和 h_k 预测结果相同的样本数。用于错误率计算的数据集为已标记样本集。当迭代进行到后期时，3 个基分类器对已标记样本集拟合良好，因此可能会出现错误率 e 非常小的情况，根据式(2)和式(7)可知，此时更新训练集的样本量可能会非常大甚至等于整个无标记样本集的数量，这会引入大量的噪声，使基分类器性能下降，因此本文算法中限制更新训练集的最大规模等于已标记样本集的规模，避免上述情况的发生。

3.4 TDDC 算法流程

本文所提的 TDDC 算法如算法 3 所示。

算法 3 TDDC 算法

输入：训练集 L；无标签样本集 U；测试集 T；基分类器 Base-Learner；DECORATE 最大迭代次数 I；生成的人工样本占训练集的比例 S

输出：测试集分类结果 Res

```
1. 初始化 e_i'←0.5，|L_i'|←0，使用算法 2 生成 3 个基分类器
2. repeat
3.     for i←1 to 3 do
4.         L_i←∅，update_i←False
5.         根据式(7)计算错误率 e_i
6.         if e_i < e_i' then
7.             for U 中每个样本 x do
8.                 if h_j(x)=h_k(x)(j,k≠i) then
9.                     L_i←L_i∪{x,h_j(x)}
10.                end if
11.            end for
12.            if |L_i| > |L_i'| then
13.                update_i←True，根据噪声学习理论从 L_i 中删除部分样本，同时控制 L_i 样本规模
14.                对 L_i 进行基于标签传播算法的置信度评估，若评估未通过，则跳转到上一步重新获取 L_i
15.            end if
16.        end if
17.    end for
18.    for i←1 to 3 do
19.        if update_i=True then
20.            h_i←BaseLearner(L∪L_i)，e_i'←e_i；|L_i'|←|L_i|
21.    end for
22. until h_i 性能无改进
23. 基于 JS 散度与无标记样本，对 3 个基分类器进行多样性度量，若多样性不满足要求，对 3 个分类器和 L∪L_i 分别使用算法 2 仅进行一次迭代，生成 3 个新的基分类器
24. 使用 3 个基分类器对 T 进行分类，并采用多数投票法输出最终分类结果 Res
```

TDDC 算法中，第 1 行采用算法 2 生成了 3 个具有分歧的基分类器，并进行了参数的初始化；第 5 行根据式(7)计算了分类器在训练集 L 上的错误率，第 7－15 行中，分类器 h_i 和 h_k 对无标记样本集 U 进行了预测，并为两者预测结果相同的样本打上伪标签，结合噪声学习理论以及 3.3 节提出的规模控制方法，获取更新训练集 L_i，并对 L_i 进行基于标签传播算法的置信度评估，若准确率大于阈值，则更新基分类器，否则重新筛选 L_i；第 18－21 行对每轮迭代的分类器进行更新；第 23 行对迭代完成的基分类器进行多样性度量，若多样性不满足要求，则使用算法 2 生成新的基分类器；第 24 行执行多数投票，输出最终分类结果。

4 实验结果与分析

4.1 数据集与实验设置

为了验证 TDDC 算法的有效性，实验选取了表 1 所列的 UCI 机器学习数据集[25]进行实验。

表 1 本文实验使用的数据集信息

Table 1 Data sets used in experiment

data set	size	attribute	Class
wine	178	13	3
australian	690	14	2
ionosphere	351	34	2
dermatology	366	34	6
heart	303	13	2

实验在 Windows10，Python3.7.3 环境下进行，基分类器

采用 scikit-learn 机器学习库决策树分类器，分类器参数采用默认参数，DECORATE 算法的人工数据比例参数 S 设置为 0.2，最大迭代次数 I 设置为 50。实验中将数据集划分为 20%的测试集和 80%的训练集，从训练集中分别选取 5%，10%，15%，20%的样本作为已标记训练集，以测试在不同已标记训练集比例下算法的表现，将剩余的样本作为算法迭代过程中的无标签样本集。实验对比算法为 Tri-training[9] 和文献[12]中的算法 Tri-D，以及文献[14]中基于交叉熵的 Tri-training 算法 TCE。

4.2 实验结果分析

每个数据集的实验结果均取 20 次实验的平均结果，以保证结果的准确性。实验评价指标采用准确率 accuracy 和 $F1$ 值 $F1$-score。实验结果如表 2－表 5 所列。

表 2 准确率(已标记样本比例：5%)

Table 2 Accuracy(with 5% labeled samples)

(单位：%)

data set	Single	Tri	Tri-D	TCE	TDDC
wine	70.42	70.83	70.69	74.44	78.61
australian	76.16	75.14	75.72	74.93	79.38
ionosphere	70.85	74.51	75.92	73.66	76.20
dermatology	70.58	71.22	71.08	72.70	79.19
heart	67.95	68.36	69.51	68.03	70.74

表 3 准确率(已标记样本比例：10%)

Table 3 Accuracy(with 10% labeled samples)

(单位：%)

data set	Single	Tri	Tri-D	TCE	TDDC
wine	77.50	79.31	80.56	80.41	83.19
australian	77.93	79.64	77.39	80.00	80.62
ionosphere	78.10	79.23	79.44	77.25	82.46
dermatology	83.18	81.55	82.23	83.38	86.69
heart	68.61	69.34	68.93	72.05	71.39

表 4 准确率(已标记样本比例：15%)

Table 4 Accuracy(with15% labeled samples)

(单位：%)

data set	Single	Tri	Tri-D	TCE	TDDC
wine	82.50	84.72	83.19	85.28	86.53
australian	80.43	81.05	81.34	81.67	83.15
ionosphere	83.73	84.51	85.00	82.75	86.76
dermatology	87.30	87.57	89.05	87.36	89.66
heart	69.59	70.74	69.67	70.41	73.44

表 5 准确率(已标记样本比例：20%)

Table 5 Accuracy(with 20% labeled samples)

(单位：%)

data set	Single	Tri	Tri-D	TCE	TDDC
wine	84.58	85.83	87.64	85.69	89.72
australian	81.12	82.03	82.32	82.03	83.30
ionosphere	85.14	86.76	85.77	87.25	88.38
dermatology	90.54	91.01	91.35	89.12	92.57
heart	71.80	71.97	72.96	72.13	74.26

图 2 给出了在各个数据集上算法的 $F1$ 值，其中 Single 指没有利用无标记样本集、只在已标记样本集上训练的单分类器。

由于 TDDC 在生成基分类器时采用了 DECORATE 集成学习方法，结合了多样性度量手段，在人工样本生成和分类器筛选阶段增加了时间开销，并且迭代过程中对伪标记样本进行了置信度评估，因此不可避免地增加了算法的计算复杂度，导致 TDDC 的训练时间比 Tri-training 长。在 australian 数据集上进行了 20%已标记样本情况下的 1 000 轮训练，结果显示，Tri-training 总耗时为 17.8 s，TDDC 总耗时为 54.4 s。

从表 2－表 5 可以看出，TDDC 算法在各个数据集上都取得了较高的分类准确率，除了在 heart 数据集 10%已标记样本比例的情况下略低于 TCE，其他实验指标均高于对比算法。特别是当已标记训练样本较少时，准确率提升最明显。具体地，在 5%已标记样本比例下，5 个数据集的实验结果显示，相比 Tri-training 及其改进算法 Tri-D 和 TCE，TDDC 的准确率提高了 1.23%～8.11%，在 10%已标记样本比例下的准确率也有明显的提升。随着已标记样本比例的增加，算法准确率呈现上升趋势，在 15%和 20%已标记样本比例下，相比对比算法，TDDC 在各个数据集上的准确率仍然最高。从图 2 的曲线图可以更加直观地得出，本文提出的 TDDC 算法在各个数据集上的 $F1$ 值与准确率相似。在已标记样本比例较少时，相比单分类器学习，对比算法有时会出现性能下降的情况，而此时 TDDC 仍然表现良好，表明了将 DECORATE 算法和基于 JS 散度与无标记样本的多样性度量方法用于基分类器生成的有效性，较好地解决了半监督学习场景下已标记样本不足时性能提升困难的问题。

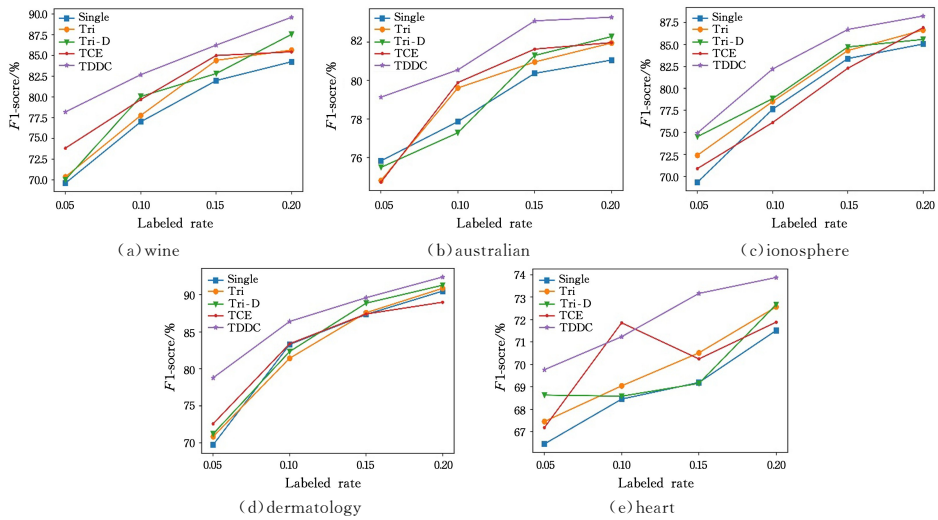

图 2 不同数据集 $F1$ 值

Fig. 2 $F1$-score of different data sets

在已标记样本比例为 5%时,TDDC 性能提升最多,相比 Tri-training 及其改进算法,其在 wine 和 dermatology 数据集上 $F1$ 值最高提高了 8.19%和 7.94%,而在另外 3 个数据集上最高提升了2.58%~4.42%;在 10%已标记样本比例下,除了 heart 数据集,在其他各数据集上最高也提升了 3.27%~6.09%;在 15%和 20%已标记样本比例下,TDDC 的 $F1$ 值也高于其他算法,且有不同程度的提升。可以看出,在已标记样本比例逐渐增加时,TDDC 仍然有较明显的性能提升,说明了基于标签传播算法的置信度评估和规模控制方法的有效性。综上所述,TDDC 算法是一种有效的半监督学习算法。

结束语 本文对基于分歧的半监督学习算法进行了总结,分析了其优缺点,并提出了一种结合 DECORATE 集成学习、多样性度量和置信度评估的 Tri-training 算法 TDDC。实验结果表明,TDDC 算法在分类性能上优于传统的 Tri-training 算法,较好地应对了半监督学习场景下已标记样本不足时,算法性能提升困难的问题。

在未来的研究中,应注重如何高效地挑选对模型提升最有价值的样本,以及如何进一步提升模型的抗噪能力。

参 考 文 献

[1] GONG S,ZHAO C. Intrusion detection system based on classification[C]// IEEE International Conference on Intelligent Control. IEEE,2012:78-83.

[2] MAZEL J,CASAS P,LABIT Y,et al. Sub-Space clustering,Inter-Clustering Results Association & anomaly correlation for unsupervised network anomaly detection[C]// 7th International Conference on Network and Service Management(CNSM 2011). IEEE,Paris,France,2011:1-8.

[3] ZHOU Z H,LI M. Semi-supervised learning by disagreement [J]. Knowledge & Information Systems,2010,24(3):415-439.

[4] ZHU X J,GHAHRAMANI Z,LAFFERTY J D. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions [C]// Machine Learning,Proceedings of the Twentieth International Conference(ICML 2003). Washington,DC,USA. 2003:912-919.

[5] BLUM A,MITCHELL T. Combining Labeled and Unlabeled Data with Co-Training[C]// Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison:ACM,1998:92-100.

[6] CHEN S J,LIU J F,HUANG Q C,et al. Conditional Value-based Co-training[J]. Acta Automatica Sinica,2013,39(10):1665-1673.

[7] KATZ G,CARAGEA C,SHABTAI A. Vertical Ensemble Co-Training for Text Classification[J]. ACM Transactions on Intelligent Systems and Technology,2017,9(2):1-23.

[8] LU J,GONG Y. A co-training method based on entropy and multi-criteria[J]. Applied Intelligence,2021,51(6):1-14.

[9] 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.

[10] XU G,ZHAO J,HUANG D. An improved social spammer detection based on tri-training[C]// 2016 IEEE International Conference on Big Data(Big Data). IEEE,2016:4040-4042.

[11] LI J,WEI Z,LI K. A Novel Semi-supervised SVM Based on Tri-training for Intrusition Detection[J]. Journal of Computers,2010,5(4):638-645.

[12] SØGAARD A. Simple semi-supervised training of part-of-speech taggers[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL,2010:205-208.

[13] RUDER S,PLANK B. Strong Baselines for Neural Semi-supervised Learning under Domain Shift[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL,2018:1044-1054.

[14] ZHANG Y,CHEN R R,ZHANG J. Safe Tri-training Algorithm Based on Cross Entropy[J]. Journal of Computer Research and Development,2021,58(1):60-69.

[15] MELVILLE P,MOONEY R J. Creating diversity in ensembles using artificial data[J]. Information Fusion,2005,6(1):99-111.

[16] ZHU X J,GHAHRAMANI Z. Learning from labels and unlabeled data with label propagation[J]. Tech Report,2002,3175(2004):237-244.

[17] ZHOU Z H. Disagreement-based Semi-supervised Learning[J]. Acta Automatica Sinica,2013,39(11):1871-1878.

[18] ANGLUIN D,LAIRD P. Learning From Noisy Examples[J]. Machine Learning,1988,2(4):343-370.

[19] ZHANG C X,WANG G W,ZHANG J S. An empirical bias-variance analysis of DECORATE ensemble method at different training sample sizes[J]. Journal of Applied Statistics,2012,39(3/4):829-850.

[20] SUN B,WANG J D,CHEN H Y,et al. Diversity measures in ensemble learning[J]. Control and Decision,2014(3):385-395.

[21] WANG W,ZHOU Z H. Analyzing Co-training Style Algorithms [C]// European Conference on Machine Learning. Springer-Verlag,2007:454-465.

[22] KUNCHEVA L I,WHITAKER C J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy[J]. Machine Learning,2003,51(2):181-207.

[23] CHU R,WANG M,ZENG X Q,et al. A New Diverse Measure in Ensemble Learning Using Unlabeled Data[C]// 2012 Fourth International Conference on Computational Intelligence,Communication Systems and Networks(CICSyN). IEEE,2012:18-21.

[24] ZHANG M L,ZHOU Z H. Exploiting unlabeled data to enhance ensemble diversity[J]. Data Mining and Knowledge Discovery,2013,26(1):98-129.

[25] DUA D,GRAFF C. UCI Machine Learning Repository [DB/OL]. [2019-12-10]. https://archive.ics.uci.edu/ml/.

WANG Yu-fei, born in 1996, postgraduate. His main research interests include semi-supervised learning, cyber security and data mining.

CHEN Wen, born in 1983, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include network security, information hiding and data mining.

(责任编辑:李亚辉)