DOI:10.11721/cqnuj20210406

半监督自训练方法综述*

吕　佳[1]，李婷婷[2]
(1. 重庆师范大学 计算机与信息科学学院，重庆 401331；2. 武汉大学 信息管理学院，武汉 430072)

摘要:【目的】通过介绍半监督自训练的背景及自训练方法的理论、训练机制、特征属性、比较标准，并梳理自训练方法的应用场景和存在的主要问题，为自训练方法的进一步研究提供参考。【方法】检索半监督自训练方法的相关文献，回顾近些年自训练方法研究领域取得的进展并进行归纳总结。【结果】首先，介绍了半监督自训练方法的背景及自训练方法的理论。然后，从相关研究中梳理了自训练方法训练机制、特征属性、比较标准。在训练机制部分，归类总结了自训练学习过程中训练集的扩充方法；在特征属性部分，分别从自训练方法的置信度度量方式和训练过程中的停止准则进行阐述；在比较标准部分，论述了衡量自训练方法有效性的评价指标。最后，整理归纳了自训练方法的应用场景和主要存在的问题。【结论】半监督自训练方法的理论研究和应用探讨在未来一段时间仍然是机器学习研究的重点和热点之一，本研究对于理解半监督自训练方法的学习机理以及解决实际应用问题等都具有重要的理论价值和现实意义。

关键词:半监督学习；自训练方法；标记样本；无标记样本；噪声
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1672-6693(2021)05-0098-09

近些年来，数据量的爆炸式增长加快了大数据时代的到来，这些数据具有巨大的潜在应用价值，各行各业逐渐意识到数据挖掘的重要性[1]。在现实生活中，收集无标记样本相对容易，而获取大量的标记样本则要困难得多，在数据标注过程中可能会花费大量的人力、物力和财力。由于获取标记样本的成本较高，而获取无标记样本的成本相对较低，如果仅对“昂贵”的标记样本进行学习，而忽略了“廉价”的无标记样本中包含的有效信息，这将造成数据资源的极大浪费[2]。因此，在有标记样本极为缺乏的情况下，利用大量的无标记样本提升分类器泛化性的半监督学习方法成为了机器学习中一个热点研究问题[3-4]。

半监督学习(Semi-supervised learning，SSL)算法可以概括为生成式模型、基于图正则化框架的半监督学习算法、协同训练算法以及自训练方法(Self-Training)。生成式模型将未标记样本的类别概率作为一组缺失参数，然后通过 EM 算法估计标记和模型的参数[5]。基于图正则化框架的半监督学习算法通过训练集中样本结构的相似性度量来构建图进行学习[6-7]。协同训练算法使用两个或者多个分类器在两个或者更多视角上进行学习，这些学习器挑选若干个置信度高的无标记样本辅助学习[8-9]。自训练方法利用少量的标记样本训练初始分类器，然后用该分类器预测无标记样本，进而选择置信度较高的无标记样本扩充标记集。这一过程实际就是分类器使用自身的预测结果来提升自己[10-12]。半监督自训练方法有诸多优点，它不像生成式模型一样需要估计参数，不像基于图的半监督学习方法一样需要构造复杂的图模型，也不像协同训练一样的需要基于特定的假设条件。自训练方法只需要 1 个分类模型、少量的有标记样本和大量的无标记样本，就可以完成复杂的半监督学习任务。

本文主要探讨了半监督自训练方法，该方法是半监督学习中常用的一种著名的分类算法。自训练方法的目的是基于无标记样本的置信度获得 1 个或几个扩充后的标记集，以分类未标记的数据。自训练算法通常包含两个主要步骤：第 1 步，用少量标记数据初始化分类器；第 2 步，用经过训练的分类器预测无标记样本，进而选出置信度较高的无标记样本迭代添加到标记数据中。这两个步骤重复进行，直到达到预设的停止条件，即可训练出一个理想的分类模型。代表性的自训练方法包括自训练与剪辑(Self-Training with editing，SETRED)[13]、自训

* **收稿日期**:2020-11-13　**修回日期**:2021-06-04　**网络出版时间**:2021-06-30 09:38
资助项目:国家自然科学基金(No. 11971084)；重庆市教育委员会“成渝地区双城经济圈建设”科技创新项目(No. KJCX2020024)；重庆市高校创新研究群体资助(No. CXQT20015)；重庆市研究生教育教学改革研究项目(No. YJA203072)；重庆市研究生科研创新项目(No. CYS20241)
第一作者简介:吕佳，女，教授，博士，研究方向为机器学习与数据挖掘，E-mail:jia-lun@163.com
网络出版地址:https://kns.cnki.net/kcms/detail/50.1165.N.20210630.0850.004.html

练最近邻规则使用切边(Self-training nearest neighbor rule using cut edges，SNNRCE)[14]、多标签自训练与剪辑(Multi-label self-training，MLSET)[15]、生成式模型加强自训练策略(Help-training for semi-supervised support vector machines，Help-training)[16]、半监督聚类与自训练集成(Self-training with semi-supervised fuzzy C-mean，STSFCM)[17]、基于数据密度峰的半监督自训练分类(Self-training semi-supervised classification based on density peaks of data，STDP)[18]、基于密度峰值和扩展无参数局部噪声滤波器的自训练方法(Self-training method based on density peak and an extended parameter-free local noise filter，STDPNF)[10]、基于数据密度峰的自训练方法(Self-training method based on an optimum path forest，STOPF)[12]和深度贝叶斯自训练(Deep Bayesian self-training，STDeep Bayesian)[19]等。其中，SETRED、SNNRCE、MLSET 用局部噪声过滤技术去除自训练方法中的噪声样本(被错误标记的样本)。然而，已存在的局部噪声过滤技术依赖参数且仅利用少量的有标记样本。因此，在 STDPNF 中，一个可扩展的无参数的局部噪声过滤技术被提出用于去除自训练方法中的噪声样本。Help-training 利用朴素贝叶斯(Naive Bayes，NB)去发现高置信度的无标记样本，然后辅助训练支持向量机(Support vector machine，SVM)。STSFCM 利用半监督模糊 C 均值聚类去寻找自训练过程中的高置信度样本，然后辅助训练 SVM。STDP 利用著名的密度峰值聚类在训练集上建立一个图结构，然后用该结构去辅助训练给定的分类器。另外，STOPF 是一个无参数的自训练方法，它利用优化路径森林(Optimum path forest，OPF)去辅助自训练方法训练分类器。最近，基于深度朴素贝叶斯的自训练方法 STDeep Bayesian 被提出，通过使用变分推理和现代神经网络体系结构来预测不确定性估计。

近些年来，半监督自训练方法发展非常迅速，但在大多数研究中，并没有对自训练方法进行严格的分析，它们并没有遵循完整的自训练学习框架。目前对自训练的研究中还没有对该方法的训练机制、特征属性、比较标准、实际应用以及主要存在的问题进行清晰的梳理。以上研究的缺失，激励了本文接下来的研究工作。本文首先详细介绍了自训练方法的算法思想、流程，重点梳理了自训练方法的训练机制、特征属性、比较标准等；然后阐述了自训练方法的实际应用场景；最后对自训练方法的主要研究问题进行了总结。

1 半监督自训练方法

SSL 问题定义如下：设有一个标记集 $L=\{(x_i,y_i)\}$，其中：x_i 为标记样本，且 $x_i\in\mathbf{R}^d$，$\mathbf{R}^d$ 表示 d 维实数集，$i=1,2,\cdots,n$，n 为标记样本的个数；y_i 是相对应的标签，$y_i\in\{\omega_1,\omega_2,\cdots,\omega_s\}$，$s$ 为不同标签的数量；未标记的集合 U，$U=\{x_1,x_2,\cdots,x_m\}$，m 为无标记样本的个数，且 $m>n$。$L\cup U$ 组成训练集，SSL 的目的是使用 $L\cup U$，而不是单独使用 L 获得一个稳健的学习模型；待测数据集 T，$T=\{x_1,x_2,\cdots,x_r\}$，r 为待测数据的个数[20-21]。

1.1 自训练方法理论

自训练作为半监督学习方法中的一种，首先用少量的标记样本训练初始分类器，然后用该分类器对无标记样本进行分类，进而选择置信度较高的无标记样本扩充标记样本集，在扩充后的标记样本集中更新分类器，整个训练过程不断迭代直到满足一个预定义的停止标准[22-23]。算法 1 为半监督自训练方法的一般流程，图 1 展示了自训练学习预测过程，表 1 为算法 1 的伪代码。

算法 1　半监督自训练方法

输入：L；U；C；θ；H；I_{Itermax}，其中：C 为迭代过程中使用的分类器，θ 是下一次迭代所选择的未标记样本的数量，H 是选择度量，$S(U_{\text{time}},\theta,C,H)$ 是选择函数，I_{Itermax} 为最大迭代次数。

输出：训练完成的分类器 C。

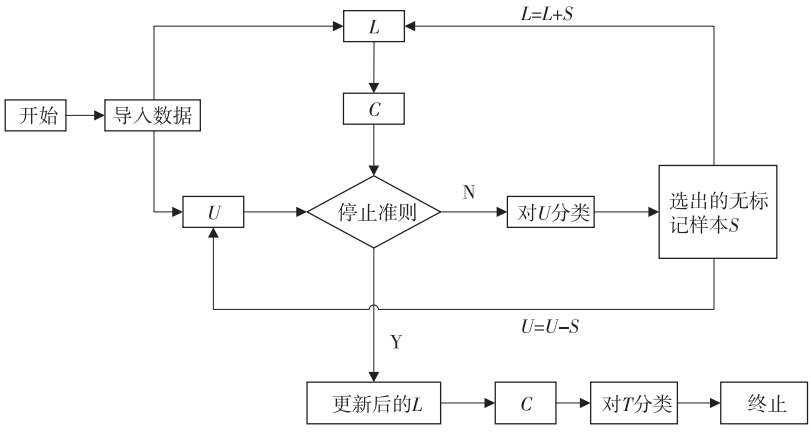

图 1　自训练学习预测框架

Fig. 1　The framework of self-training learning and prediction

表1 算法1伪代码

Tab.1 Pseudocode of algorithm 1

伪代码行号	算法1伪代码
1	$t_{\text{time}}=0$,$L_{\text{time}}=L$,$U_{\text{time}}=U$,其中,t_{time}是迭代次数,L_{time}是在第t_{time}次迭代时的标记样本集,U_{time}是在第t_{time}次迭代时的无标记样本集。
2	While $U!=\varnothing$ 或者 $t_{\text{time}}<I_{\text{Itermax}}$
3	在L_{time}上训练C
4	$S_{\text{time}}=S(U_{\text{time}},\theta,C,H)$, S_{time}为选出的无标记样本集
5	$U_{\text{time}}=U_{\text{time}}-S_{\text{time}}$; $L_{\text{time}}=L_{\text{time}}+S_{\text{time}}$
6	$t_{\text{time}}=t_{\text{time}}+1$
7	end while
8	$L=L_{\text{time}}$,训练分类器C
9	输出训练完成的分类器C

如上所述,可以利用自训练方法对无标记数据进行标记,实现已标记数据的扩充。由于初始标记数据集规模很小,利用该数据集不足以训练出高精度且泛化能力较强的分类器。在自训练过程中,数据被误标记是不可避免的,这些被误标记的样本则会变成训练集中的噪声数据。训练集中的噪声数据会在迭代过程中不断累积,而训练集中样本的有效性却永远不会被质疑。但实际上,训练集中出现噪声样本会直接破坏数据的真实结构信息。因此,噪声样本对分类器的训练是极为不利的,它们会导致分类性能的下降。在随后的迭代过程中不会对新增数据的标记进行判断,噪声样本将一直随附在训练集中,随着迭代过程的继续,分类性能将持续削弱。但是,如果迭代过程中(尤其是迭代的早期)能够找到这些误标记数据并进行修改或去除,则扩充后的训练集的质量会得到提高,分类性能就不会随着迭代的进行而削弱,算法会有更好的分类效果。

1.2 训练机制归类

半监自训练方法迭代地搜索一个或多个扩充的标记集(Extended labeled set, EL)。在训练过程中,可以采用不同机制进行标记集的扩充,本节梳理了自训练过程中EL扩充机制的常见类别(图2)。

1.2.1 添加机制 在自训练过程中可以通过增量、批处理以及修正等机制形成EL。

1) 增量。严格的增量方法从一个扩大的标记集EL=L开始,逐步添加U中置信度最高且满足设定标准的样本。增量机制的关键是如何确定每个无标记样本的置信度预测的方式,即属于每个类的概率。在增量机制中,无标记样本被添加到EL中的顺序决定了学习模型。增量机制的重点是每次迭代中添加样本的数量,文献[24]指出这个参数可以定义为一个常量。而文献[25]则将L中每个类样本数的比例值设置为每一次迭代添加样本的数量。增量机制的优点是在学习阶段比非增量算法更快,缺点是严格的增量算法可能将具有错误预测的样本添加到标记数据集中。因此,在这种情况下,所学习的模型对待测样本的分类性能较低。

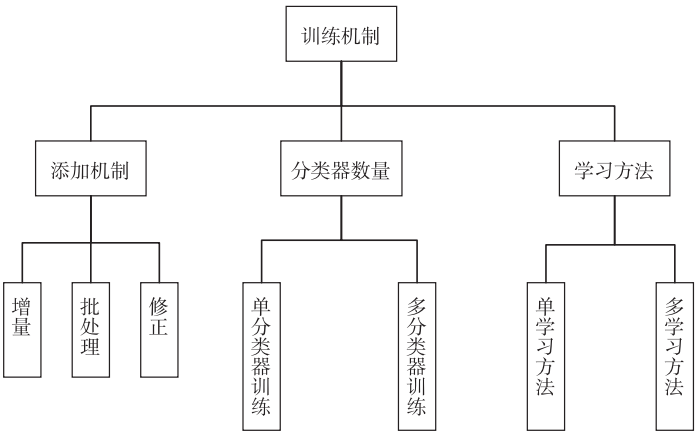

图2 自训练方法训练机制分类

Fig.2 Classification of training mechanism of self-training method

2) 批处理。Li等人[23]采用批处理机制生成EL。批处理机制是指每个无标记样本被添加到EL之前需判断该样本是否满足添加标准,进而将那些满足添加标准的样本立即添加。在批处理机制下,批处理技术不会在学习过程中为每个无标记样本分配一个确定的类。由于批处理技术需要重新排序从标记样本中所获得的模型,因此,该技术比增量算法具有更高的时间复杂度。

3) 修正。修正机制是增量方法主要缺点的解决途径。修正机制从EL=L开始,迭代地添加或删除符合特

定准则的样本,允许对已经执行的操作进行纠正,进而取得较高精度。文献[26-27]结合数据剪辑技术进行噪声过滤,避免了迭代时将噪声样本引入 EL。在噪声样本先验或后清洗阶段,将增量与批处理模式相结合也是一种修正机制。修正机制灵活性较强,但与增量算法和批处理模式相比,该机制需要更高的计算需求。

1.2.2 单(多)分类器训练　自训练方法在标记集扩充阶段可以使用一个或多个分类器。

1) 单分类器训练。在自训练过程中,单分类器训练是指采用单一分类模型迭代搜索 EL。在单分类器训练模型中,每个未标记样本都属于唯一使用的分类器所分配的最可能的类,这些未标记样本的所属概率应该被明确地测量。此外,在单分类器模型中,可以用不同的方法计算置信度,进而选出 EL。例如,在概率模型中,Karlos 等人[28]通过朴素贝叶斯预测中的输出概率来衡量置信度。在最近邻模型中,可以通过近似于距离的置信度来衡量[20]。单分类器方法的主要优点是结构简单,可以更快地计算置信概率。

2) 多分类器训练。在自训练过程中,多分类器训练是指由两个或两个以上的分类模型迭代搜索 EL。在自训练学习中使用多分类器训练的原则是:使用少量样本学习几个弱分类器,进而产生泛化能力更好的强分类器。这些方法在一定程度上由集成方法的成功经验所驱动,如文献[29]。在多分类器训练中,常用两种不同的方法来计算置信度:分类器的一致性和单分类器获得概率的组合。这些技术通常比单分类器训练模型更有效地找出 EL,从而获得更高的精度。多个分类器训练的缺点是该方法的复杂性较高。

1.2.3 单(多)学习方法　除了分类器的数量外,自训练方法中另一个关键问题是学习框架由相同(单)学习方法还是不同(多)学习方法组成,使用的不同学习方法的数量也可以有效地确定 EL。

1) 单学习方法。在自训练过程中,单学习方法是指单个和多个分类器采用同一策略计算样本置信度,进而有效地搜索 EL。采用单学习算法,对分类器的其余属性、视图类型和分类器数量限制条件较低。这种方法的特点是简单,它在很多应用领域中表现良好[24]。

2) 多学习方法。在自训练过程中,多学习方法是指一组不同类型的分类器集成多种策略计算样本置信度,进而有效地搜索 EL。多学习方法在不同学习器呈现的不同假设下工作,利用它们之间的偏差分类,产生局部不同的模型。多学习方法与多分类器训练模型密切相关,该方法本身就是一种多分类器训练方法。因此,将多分类器训练的一般性质外推到多学习中,如文献[20,29-30]从不同角度使用了多学习方法。多学习方法的难点是最合适学习方法的选择问题。

1.3 自训练方法特征

本节梳理了自训练方法训练机制的主要特性,这些特性一定程度上影响了训练过程,从而导致不同的分类效果。

1.3.1 置信度度量　文献[31-32]指出自训练过程中不准确的置信度测量会导致 EL 中添加错误标记的样本,进而造成了自我标记过程的性能退化。置信度度量过程中,可分为简单预测以及协议与组合度量两种方式。

1) 简单预测。在自训练过程中,可通过概率预测得出样本置信度,概率预测可以从训练的学习模型中提取。这一特点在单分类器方法中得到了本质上的体现。例如,概率模型为每个未标记样本返回属于每个类的关联概率。在决策树算法中,置信概率可以估计为作出预测的叶子节点的精度。基于实例的学习方法可以根据样本之间的不同之处来估计概率。

2) 协议与组合。多分类器方法可以根据每个分类器来近似预测样本一致性的概率,以便分类器找到每个无标记样本的最终的置信度,多分类器方法通常使用投票规则确定样本类别。此外,还可以将分类器的一致性与计算概率相结合,建立混合置信度预测模型。

1.3.2 停止准则　停止准则是指用于停止自训练过程的一种设置,它关系着形成的 EL 的规模,这是影响自训练分类性能一个重要的因素[9-10,33]。停止准则对自训练分类性能的影响主要体现在如下 3 个方面:

1) 在自训练方法中,自我标记过程被重复进行,直到所有的无标记样本都被添加到 EL,则停止训练过程。但该方法如果将错误标记的无标记样本添加到 EL 中,则会损害分类性能。

2) 可设置有限的迭代次数,在一个较小的未标记样本池中选择样本,而不是整个未标记集来形成 EL。由于不同数据集的规模不一致,最大迭代次数通常是变动的。该方法已成功地应用于多种自训练方法中,优于以往的停止准则。

3) 当使用的分类器在自我标记过程中不改变所学模型时,可以满足停止准则。这个标准限制了添加到 EL 中的未标记样本的数量。但是,它不能确保错误的未标记样本不会被添加到 EL 中。

1.4 比较标准

在自训练算法评价中,主要有4种指标可以用来比较自训练方法的性能,即分类精度、标记样本数量的影响、噪声容忍度和时间复杂度[34-36]。

1) 分类精度。成功的自训练算法通常能够适当地扩充标记集,进而提高模型的分类精度。

2) 标记样本数量的影响。在自训练方法学习过程中,标记样本的数量决定了学习模型的有效性。一个好的自训练方法应该能够在标记样本数量较少时也能学习到有效的模型。

3) 噪声容忍度。如果将噪声样本添加到标记集,使得未标记数据的分类偏向于不正确的类,进而损害训练效果,造成EL在下一次迭代中含有更大的噪声。这种情况通常发生在训练过程的初始阶段,而且随着无标记样本的减少,损害会更大。在自训练过程中可能会出现两种类型的噪声样本。第一种是由标记样本在类中的分布引起的,这会导致分类器错误地标记一些样本。第二种是原始未标记数据中可能存在异常值。

4) 时间复杂度。目前,已有多种自训练算法,每一种算法都可以达到解决问题的目的,但花费的时间不尽相同。在自训练过程中,算法的时间复杂度主要取决于初始标记样本输入时间,自训练迭代所耗费的时间,算法编译为可执行程序的时间,计算机执行每条指令所需的时间。如果学习阶段需要太长时间,对于生活应用来说可能变得不切实际[25]。因此,算法的时间复杂度也是自训练方法分类性能的重点评价指标之一。

2 自训练方法的应用

自训练方法能够改善和弥补传统分类任务的不足(即有监督分类任务的不足),它利用大量的无标记样本和少量的有标记样本去训练一个分类器,从而节约了标注时间,减少了人工成本。相比半监督学习的其他可适应方法(即让有监督分类方法适应半监督学习环境)[37-39],自训练方法更具有灵活性,因为它能够帮助任何的分类器或者分类任务适应半监督学习环境。

由于自训练方法简单灵活,且不需要构造复杂的数学模型,也不用基于特定的假设条件,在实际生活中应用广泛。目前已被成功地应用在自然语言处理、图像识别、癌症分类等领域。

2.1 自然语言处理

自训练方法已经应用于多项自然语言处理任务。Tran等人[40]提出了一种将人工智能和自训练相结合的方法,通过使用机器标注和人工标注的数据来减少微博流中命名实体识别任务的标注工作量。Hajmohammadi等人[41]提出了一种新的学习模型,该模型将主动学习和半监督自训练方法相结合进行跨语言情感分类。文献[42]使用自训练方法为文本文档中信息提取提供了一个基于子句的框架。

2.2 图像识别

文献[43]提出了一种基于自训练的光谱反射率恢复方法,利用多光谱成像技术准确地重建艺术绘画的光谱图像。Le等人[44]提出了一种鲁棒的、全自动的半自训练系统,可以在具有挑战性的面部图像中同时检测和分割胡须/胡子。文献[45]提出了一种基于自训练方法的系统,该系统适合微调划线去除算法的参数,或者使它们适合于特定的文档图像集合,进而改善学习效果。文献[32]分别采用协同训练和自训练的方法增加标签的数量。协同训练的思想是分别在视觉和振动数据上训练两个分类器,并在彼此的输出上迭代更新分类器。同时,基于视觉分类的自训练方法根据自身的输出重新训练分类器。这两种方法本质上都增加了标签的数量,因此使得地形分类器能够从稀疏的训练数据集中进行操作。

2.3 癌症分类

文献[46]针对基因表达数据进行癌症分类,提出了一种新的低秩表示(LRR)下的自训练子空间聚类算法:SSC-LRR算法。低秩表示首先用于从高维基因表达数据中提取区别特征;然后使用自训练子空间聚类方法生成癌症分类预测。文献[47]将半监督自训练方法应用在乳腺癌智能诊断中,提出了一种结合密度峰值优化模糊聚类的半监督自训练方法。该方法先对无标记样本集进行密度峰值聚类,人工选出聚类中心后,将新的聚类中心作为初始聚类中心进行模糊聚类,从而筛选出有价值的无标记样本。

除了以上重点详述的自训练方法的应用领域,该方法还被有效地的应用在地形识别[32]、高校就业预测[48]、生物基因预测[46]、智能语句识别[49]等实际场景中。

3 自训练方法存在的问题

影响自训练方法性能的关键因素是扩充训练集的机制、基分类器的选择、学习方法的选择、置信度度量以及

停止准则等,这些因素直接或间接影响了训练过程中样本错误标记的概率。在自训练方法中,错误标记是一个非常普遍、棘手甚至不可避免的问题。通常,自训练方法中错误标记的样本被视为噪声样本,这会严重扭曲数据的分布。因此,自训练方法中错误标记的样本会降低学习模型的泛化能力。Adankon 等人[50]为了提升自训练算法的性能,提出了一种新的自训练方法 Help-training,即用 NB[51]辅助训练 SVM[52]。然而,Help-training 的性能很大程度上取决于标记样本的数量和分布。因此,Gan 等人[17]提出了一种改进的算法 STSFCM,其中半监督模糊 C 均值(Semi-supervised fuzzy C-mean, SFCM)[53]被用来辅助训练 SVM。然而,改进后的算法只能在球形分布数据集上得到理想的结果。为了弥补这一缺点,研究者提出了一种基于数据密度峰的自训练方法(Naive Bayes self-training combined density peak clustering and fuzzy C-mean,NBSTDPCFCM)[54],该方法巧妙地结合了基于流形分布数据的密度峰值聚类(Density peak clustering,DPC)[55]算法。因此,NBSTDPCFCM 可以在任意形状的数据集上训练出良好的分类器。虽然 NBSTDPCFCM 表现良好,但仍然存在标签错误的问题[12,54,56]。如果将具有错误预测的样本添加到训练集中,NBSTDPCFCM 的性能将进一步恶化。

许多研究者基于局部噪声滤波器对自训练方法进行研究,如 SETRED[13]、SNNRCE[14]以及 MLSTE[15]。SETRED 和 SNNRCE 使用切边权统计(Cut edge weight statistic, CEWS)[57]来过滤噪声样本,而 MLSTE 使用剪辑最近邻(Edited nearest neighbor,ENN)[58]来删除噪声样本。随后,Triguero 等人[26]发现,当标记数据数量较少时,自训练方法使用现有噪声滤波器的性能较差,即大多数用于自训练方法的局部噪声滤波器需要足够的标记数据才能有效训练。

以上自训练方法研究的关键之处都是如何有效地扩充初始训练集,进而训练泛化性更强的学习模型,只是相互关注的重点不同。此外,在扩充训练集的过程中,高置信度的无标记样本可能会被错误标记,导致自训练方法性能恶化。因此,自训练方法主要待解决的问题如下:

1) 自训练方法受标记样本分布和数量的限制。①当初始标记样本不能代表整个数据分布时,用初始标记样本训练的分类器泛化性较低。这是因为构造的决策边界会偏离真实的决策边界,进而无法有效地归类数据。②当初始有标记样本数量不足时,很难构造有效的分类器。因为构造分类器通常需要足够的标记样本,否则会造成误标记。③当初始有标记样本不足,且不能够代表整个数据集的分布时,就很难有效地发现高置信度无标记样本。如果高置信度无标记样本被错误发现,就会造成误标记。

2) 很难有效地发现高置信度无标记样本。在自训练过程中,高置信度的无标记样本更容易被预测正确。但如果错误地发现高置信度样本则会导致误标记,这些被错误发现的高置信度样本很难被正确预测。因此,在自训练算法中,如何有效地发现高置度无标记样本是一个关键问题。①通常的策略是把基分类器的概率输出作为寻找高置信度样本的方法。然而对于一些经典分类器,比如最近邻、支持向量机等,它们没有概率值输出。②几乎所有的发现高置信度无标记样本的方法都严重依赖于参数,导致算法表现不稳定且应用困难。③几乎所有的发现高置信度无标记样本的方法都需要排序置信度值,排序过程增加了计算时间。④存在的方法难以适应复杂的数据集,比如流形数据。

3) 虽然目前已有较多半监督自训练算法,但这些算法所使用的基分类器都有自身的弱点。现有对半监督自训练算法的理论分析虽然揭示了自训练方法的一些内在机理,但是很多分析都建立在一些较强的假设条件上。因此,在更一般、更接近真实情况条件下如何进行自训练学习,是一个有意义的研究方向。

4 结束语

自训练方法作为半监督分类中最重要的研究领域之一,它利用大量的无标记样本和少量的标记样本训练分类器,节约了标注时间,减少了人工成本,具有重要的意义和实际价值。本文首先介绍了半监督自训练的背景、自训练方法的理论,然后从大量相关研究中梳理了自训练方法训练机制、特征属性、比较标准、应用场景,并整理归纳了自训练方法中主要存在的问题。由此可知,半监督自训练方法的理论研究和应用探讨在未来一段时间仍然是机器学习研究的重点和热点之一。这些研究对于理解半监督自训练方法的学习机理以及解决实际应用问题等都具有重要的理论价值和现实意义。

参考文献:

[1] CHIRANTH H,GRAY K E. Use of machine learning and data analytics to increase drilling efficiency of nearby wells[J].

Journal of Natural Gas Science and Engineering,2017,40:327-335.

[2] 甘海涛.半监督聚类与分类算法研究[D].武汉:华中科技大学,2014.
GAN H T. Research on semi-supervised clustering and classification algorithm[D]. Wuhan: Huazhong University of Science & Technology,2014.

[3] HADY M F A,SCHWENKER F. Semi-supervised learning[J]. Journal of the Royal Statal Society,2006,172(2):530-530.

[4] ZHU X, GOLDBERG A B. Introduction to semi-supervised learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning,2009,3(1):130.

[5] 张宏东.EM算法及其应用[D].济南:山东大学,2014.
ZHANG H D. EM algorithm and applications[D]. Ji'nan: Shandong University,2014.

[6] GAN H,LI Z,WU W,et al. Safety-aware graph-based semi-supervised learning[J]. Expert Systems with Applications,2018,107(1):243-254.

[7] KILINC O, UYSAL I. Gar: an efficient and scalable graph-based activity regularization for semi-supervised learning[J]. Neurocomputing,2018,296(28):46-54.

[8] LEE C H. Multi-label classification of documents using fine-grained weights and modified co-training[J]. Intelligent Data Analysis,2018,22(1):103-115.

[9] 龚彦鹭,吕佳.结合半监督聚类和加权KNN的协同训练方法[J].计算机工程与应用,2019,55(22):114-118.
GONG Y L,LÜ J. Co-training method combined with semi-supervised clustering and weighted K-nearest neighbor[J]. Computer Engineering and Applications,2019,55(22):114-118.

[10] LI J N,ZHU Q S,WU Q W. A self-training method based on density peaks and an extended parameter-free local noise filter for *K* nearest neighbor[J]. Knowledge-Based Systems,2019,184(15):1-12.

[11] LI J N,ZHU Q,WU Q W. An effective framework based on local cores for self-labeled semi-supervised classification[J]. Knowledge-Based Systems,2020,197(12):1-17.

[12] LI J N,ZHU Q S. Semi-supervised self-training method based on an optimum-path forest[J]. IEEE Access,2019,7(1):2169-3536.

[13] LI M,ZHOU Z H. SETRED: self-training with editing[C]//PAKDD'05: Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag,2005:611-621.

[14] WANG Y,XU X,ZHAO H,et al. Semi-supervised learning based on nearest neighbor rule and cut edges[J]. Knowledge Based Systems,2010,23(6):547-554.

[15] WEI Z, WANG H, ZHAO R, et al. Semi-supervised multi-label image classification based on nearest neighbor editing[J]. Neurocomputing,2013,119(7):462-468.

[16] ADANKON M M,CHERIET M. Help-training for semi-supervised support vector machines[J]. Pattern Recognition,2011,44(9):2220-2230.

[17] GAN H,SANG N,HUANG R,et al. Using clustering analysis to improve semi-supervised classification[J]. Neurocomputing,2013,101(4):290-298.

[18] WU D, SHANG M, LUO X, et al. Self-training semi-supervised classification based on density peaks of data[J]. Neurocomputing,2018,275(31):180-191.

[19] RIBEIRO F D S,FRANCESCO C,SWAINSON M,et al. Deep Bayesian self-training[J]. Neural Computing and Applications,2019,7(3):1-18.

[20] 李婷婷,吕佳.基于加权K最近邻改进朴素贝叶斯自训练算法[J].武汉大学学报(理学版),2019,65(5):465-471.
LI T T,LÜ J. Improved naive Bayes self-training algorithm based on weighted K-nearest neighbor[J]. Journal of Wuhan University (Natural Science Edition),2019,65(5):465-471.

[21] 吕佳,黎隽男.结合半监督聚类和数据剪辑的自训练方法[J].计算机应用,2018,38(1):110-115.
LÜ J,LI J N. Self-training method based on semi-supervised clustering and data editing[J]. Journal of Computer Applications,2018,38(1):110-115.

[22] TANHA J,VAN S M,AFSARMANESH H. Semi-supervised self-training for decision tree classifers[J]. International Journal of Machine Learning and Cybernetics,2017,8(1):355-370.

[23] LI T T,LÜ J. Divide-and-conquer ensemble self-training method based on probability difference[J]. Journal of Ambient Intelligence and Humanized Computing,2020(4):1-13.

[24] 董立岩,隋鹏,孙鹏,等.基于半监督学习的朴素贝叶斯分类新算法[J].吉林大学学报(工学版),2016,46(3):884-889.
DONG L Y,SUI P,SUN P,et al. A new naive Bayes classification algorithm based on semi-supervised learning[J]. Journal of

Jilin University (Engineering Edition)，2016，46(3)：884-889.

[25] TRIGUERO I，GARCIA S，HERRERA F，et al. Self-labeled techniques for semi-supervised learning：taxonomy，software and empirical study[J]. Knowledge and Information Systems，2015，42(2)：245-284.

[26] TRIGUERO I，SAEZ J A，LUENGO J，et al. On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification[J]. Neurocomputing，2014，132(132)：30-34.

[27] 黎隽男，吕佳. 结合主动学习与置信度投票的集成自训练方法[J]. 计算机工程与应用，2016，52(20)：167-171.
LI J N，LÜ J. Ensemble self-training method based on active learning and confidence voting[J]. Computer Engineering and Applications，2016，52(20)：167-171.

[28] KARLOS S，FAZAKIS N，PANAGOPOULOU A，et al. Locally application of naive Bayes for self-training[J]. Evolving Systems，2017，8(1)：3-18.

[29] WAN L，HONG Y，HUANG Z，et al. A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks[J]. International Journal of Geographical Information Science，2018，32(11)：2225-2246.

[30] PIROONSUP N，SINTHUPINYO S. Analysis of training data using clustering to improve semi-supervised self-training[J]. Knowledge-Based Systems，2018，143(2)：65-80.

[31] WANG S，WU L，JIAO L，et al. Improve the performance of co-training by committee with refinement of class probability estimations[J]. Neurocomputing，2014，136(8)：30-40.

[32] OTSU K，ONO M，FUCHS T J，et al. Autonomous terrain classification with co- and self-training approach[J]. IEEE Robotics & Automation Letters，2017，1(2)：814-819.

[33] GOSAIN A，SARDANA S. Handling class imbalance problem using oversampling techniques：a review[C]//2017 International Conference on Advances in Computing，Communications and Informatics (ICACCI). Udupi，India：IEEE，2017：79-85.

[34] XIE Q，HOVY E，LUONG M，et al. Self-training with noisy student improves ImageNet classification[EB/OL]. (2019-11-01)[2020-11-13]. https：//arxiv. org/pdf/1911. 0425 2. pdf.

[35] HADIFAR A，STERCKX L，DEMEESTER T，et al. A self-training approach for short text clustering[C]//Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence，Italy：ACL，2019：194-199.

[36] LI J N，ZHU Q S，WU Q W. A parameter-free hybrid instance selection algorithm based on local sets with natural neighbors [J]. Applied Intelligence，2020，50(5)：1527-1541.

[37] ZHAN Y，BAI Y，ZHANG W，et al. A P-ADMM for sparse quadratic kernel-free least squares semi-supervised support vector machine[J]. Neurocomputing，2018，3(306)：37-50.

[38] AMORIM W P，FALCAO A X，PAPA J P，et al. Improving semi-supervised learning through optimum connectivity[J]. Pattern Recognition，2016，4(60)：72-85.

[39] TANG B，HE H. ENN：extended nearest neighbor method for pattern recognition [research frontier][J]. IEEE Computational Intelligence Magazine，2015，10(3)：52-60.

[40] TRAN V C，NGUYEN N T，FUJITA H，et al. A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields[J]. Knowledge Based Systems，2017，12(132)：179-187.

[41] HAJMOHAMMADI M S，IBRAHIM R，SELAMAT A，et al. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples[J]. Information Sciences，2015，317(1)：67-77.

[42] VO D，BAGHERI E. Self-training on refined clause patterns for relation extraction[J]. Information Processing and Management，2017，54(4)：686-706.

[43] XU P，XU H，DIAO C，et al. Self-training-based spectral image reconstruction for art paintings with multispectral imaging[J]. Applied Optics，2017，56(30)：8461-8470.

[44] LE T H N，LUU K，ZHU C，et al. Semi self-training beard/moustache detection and segmentation simultaneously[J]. Image & Vision Computing，2017，58(2)：214-223.

[45] PROKOPIOU K，KAVALLIERATOU E，STAMATATOS E. An image processing self-training system for ruling line removal algorithms[C]//2013 18th International Conference on Digital Signal Processing(DSP). Washington：IEEE，2014：1-6.

[46] XIA C Q，HAN K，QI Y，et al. A self-training subspace clustering algorithm under low-rank representation for cancer classifcation on gene expression data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2017，15(4)：1315-1324.

[47] 罗云松. 半监督自训练算法在乳腺癌分析预测中的应用研究[D]. 重庆：重庆师范大学，2019.
LUO Y S. Application of semi-supervised self-training algorithm in breast cancer analysis and prediction[D]. Chongqing：Chongqing Normal University，2019.

[48] 马茂源.基于改进半监督自训练方法的高校毕业生就业预测应用研究[D].重庆:重庆师范大学,2019.
MA M Y. Research on application of employment guidance for college graduates based on improved semi-supervised self-training method[D]. Chongqing:Chongqing Normal University,2019.
[49] DALVA D,GUZ U,GURKAN H,et al. Effective semi-supervised learning strategies for automatic sentence segmentation[J]. Pattern Recognition Letters,2017,10(105):76-86.
[50] ADANKON M M,CHERIET M. Help-training for semi-supervised support vector machines[J]. Pattern Recognition,2011,44(9):2220-2230.
[51] ZHANG L,JIANG L,LI C,et al. Two feature weighting approaches for naive Bayes text classifiers[J]. Knowledge Based Systems,2016,4(100):137-144.
[52] BOUBOULIS P, THEODORIDIS S, MAVROFORAKIS C, et al. Complex support vector machines for regression and quaternary classification[J]. IEEE Transactions on Neural Networks,2015,26(6):1260-1274.
[53] YIN X,SHU T,HUANG Q. Semi-supervised fuzzy clustering with metric learning and entropy regularization[J]. Knowledge-Based Systems,2012,35(1):304-311.
[54] 罗云松,吕佳.结合密度峰值优化模糊聚类的自训练方法[J].重庆师范大学学报(自然科学版),2019,36(2):96-102.
LUO Y S,LÜ J. Self-training algorithm combined with density peak optimization fuzzy clustering[J]. Journal of Chongqing Normal University (Natural Science),2019,36(2):96-102.
[55] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.
[56] WU D,SHANG M,WANG G,et al. A self-training semi-supervised classification algorithm based on density peaks of data and differential evolution[C]//2018 IEEE 15th International Conference on Networking,Sensing and Control (ICNSC),Zhuhai,China:IEEE,2018:1-6.
[57] MUHLENBACH F,LALLICH S,ZIGHED D A. Identifying and handling mislabelled instances[J]. Journal of Intelligent Information Systems,2004,22(1):89-109.
[58] WILSON,DENNIS L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems Man and Cybernetics,2007,2(3):408-421.

A Summary of Semi-Supervised Self-Training Methods

LÜ Jia[1],LI Tingting[2]
(1. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331;
2. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Purposes**]By introducing the background of semi-supervised self-training, the theory, training mechanism, feature attributes, comparison standards, and sorting out the application scenarios and existing problems of self-training method, it provides reference for further research of self-training method. [**Methods**]The related literatures of semi-supervised self-training method were searched, and the progress in training method research field in recent years was reviewed and summarized. [**Findings**]Firstly, the background and theory of semi-supervised self-training method were introduced. Secondly, the training mechanism, feature attributes and comparison standards of self-training methods were sorted out. In the part of training mechanism, the expansion methods of training set of self-training learning were classified and summarized. In the part of feature attributes, the confidence measurement and stop criteria of self-training method were expounded. In the part of comparison standard, the evaluation index to measure the effectiveness of self-training method was discussed. Finally, the application scenarios and main problems of self-training methods were summarized. [**Conclusions**] The theoretical research and application of semi-supervised self-training method will remain one of the hotspots of machine learning research in the future. This research has important theoretical value and practical significance for understanding the learning mechanism of semi-supervised self-training method and solving practical application problems.

Keywords: semi-supervised learning; self-training method; labeled instances; unlabeled instances; noise instances

(责任编辑 许 甲)