半监督集成学习综述

蔡 毅[1,2] 朱秀芳[1,2] 孙章丽[1,2] 陈阿娇[3]

(北京师范大学地表过程与资源生态国家重点实验室 北京 100875)[1]

(北京师范大学资源学院 北京 100875)[2] (湖南师范大学资源与环境科学学院 长沙 410081)[3]

摘 要 半监督学习和集成学习是目前机器学习领域中两个非常重要的研究方向,半监督学习注重利用有标记样本与无标记样本来获得高性能分类器,而集成学习旨在利用多个学习器进行集成以提升弱学习器的精度。半监督集成学习是将半监督学习和集成学习进行组合来提升分类器泛化性能的机器学习新方法。首先,在分析半监督集成学习发展过程的基础上,发现半监督集成学习起源于基于分歧的半监督学习方法;然后,综合分析现有半监督集成学习方法,将其分为基于半监督的集成学习与基于集成的半监督学习两大类,并对主要的半监督集成方法进行了介绍;最后,对现有研究进了总结,并讨论了未来值得研究的问题。

关键词 半监督学习,集成学习,半监督集成学习,boosting,Bagging,泛化性能

中图法分类号 TP181 **文献标识码** A

Semi-supervised and Ensemble Learning: A Review

CAI Yi[1,2] ZHU Xiu-fang[1,2] SUN Zhang-li[1,2] CHEN A-jiao[3]

(State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China)[1]

(College of Resources Science and Technology, Beijing Normal University, Beijing 100875, China)[2]

(College of Resources and Environment Science, Hunan Normal University, Changsha 410081, China)[3]

Abstract Semi-supervised learning (SSL) and ensemble learning are two important paradigms in the field of machine learning research. SSL attempts to achieve strong generalization by exploiting both labeled and unlabeled instances, while ensemble learning aims to improve the performance of weak learner by making use of multiple classifiers. SSL ensemble learning is a novel paradigm which can improve the generalization performance of classifier by combining SSL and ensemble learning. Firstly the development process of SSL ensemble learning was analyzed and it was found that SSL ensemble learning is derived from disagreement-based SSL. Then, classify SSL Ensemble learning methods were classified into two categories: SSL-based ensemble learning and ensemble-based SSL. A detailed description for the main methods of SSL Ensemble learning was given. Finally, the current research status of SSL ensemble learning was summarized and some issues which are worth of further study were given.

Keywords Semi-supervised learning, Ensemble learning, Semi-supervised ensemble learning, Boosting, Bagging, Generalization performance

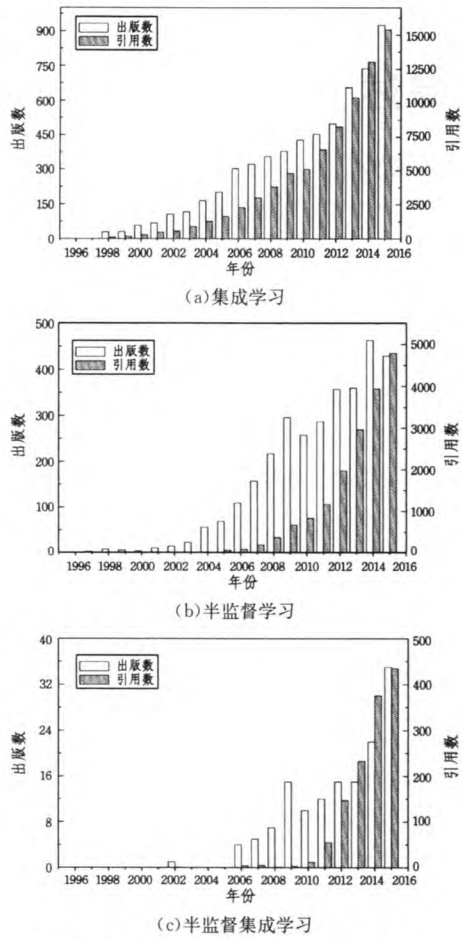
1 引言

机器学习作为人工智能的重要研究内容,经过近半个世纪的发展,现今已和模式识别、数据挖掘、统计学习、计算机视觉、自然语言处理等多个领域相互影响、交织发展,无论是理论上还是实践上都取得了巨大的进展,并广泛地应用于文本分类、语音识别、图像解译、医学诊断等多个领域。

半监督学习和集成学习目前是机器学习领域中两个较为重要的研究方向。通过 Web Of Science 平台(www.webofknowledge.com),分别以关键词"Ensemble Learning","Semi-supervised Learning"和"'Ensemble' and 'Semi-supervised'"为主题进行文献检索,检索时间为 2016 年 08 月 17 日。统计结果发现,无论是集成学习(见图 1(a))还是半监督学习(见图 1(b))的文献发表数量和引文数量,在近 20 年(1996 年—2015 年)都呈"指数型"趋势增长。相对于半监督学习和集成学习而言,目前半监督集成学习(见图 1(c))的研究体量较少,但是半监督集成研究作为一个交叉研究分支所表现出的优势正日益引起学者的关注,相关的文献出版量与引文数量也在逐年稳步提升。半监督学习与集成学习的思想几乎同时期产生[1-3],但是在很长一段时间内二者平行发展,并未有太多交集。Zhou[4]认为导致该现象的原因可能是:1)半监督学习流派认为只要利用好足够多的无标记样本,就能将学习器做得足够好,因此不需要多个学习器;2)集成学习流派认为只要有多个学习器就能通过集成学习将弱学习器转为强学习器,不需要额外的无标记样本。而实际上,半监督和集成学习可以相互帮助,进一步改善学习器的性能。

本文受国家自然科学青年基金项目(41401479),高分辨率对地观测重大专项(民用部分)(02-Y30B06-9001-13115)资助。

蔡 毅(1992—),男,硕士生,主要研究方向为遥感图像模式识别,E-mail:charyee@126.com;朱秀芳(1982—),女,副教授,硕士生导师,主要研究方向为遥感应用,E-mail:zhuxiufang@bnu.edu.cn;孙章丽(1986—),女,博士生,主要研究方向为水文遥感应用,E-mail:sunzhangli@gmail.com;陈阿娇(1991—),女,硕士生,主要研究方向为气象水文数据挖掘,E-mail:ajchen0807@163.com。



(a)集成学习

(b)半监督学习

(c)半监督集成学习

图 1 1996 年－2015 年文献出版与引文数

半监督集成学习是一种同时利用半监督学习与集成学习的新学习方式，主要研究如何将半监督学习与集成学习有机结合，从而进一步提高整体分类器的性能。通过分析国内外已发表的文献，发现目前鲜有研究对半监督集成这一学习方式进行完整系统地综述，本文将对半监督学习与集成学习组合学习方式的产生及发展进行阐述，对现有半监督集成学习的组合方式进行分析，对目前主流的半监督集成方法进行介绍，最后对半监督集成的问题及发展方向进行展望。

2 半监督集成的起源

通过分析半监督集成的发展过程，在直观上可以发现二者的结合并非偶然。图 2 示出了半监督学习、集成学习以及半监督集成的发展过程。图 2(a)－图 2(c) 3 部分分别对应表示半监督学习、半监督集成学习以及集成学习的相关内容。

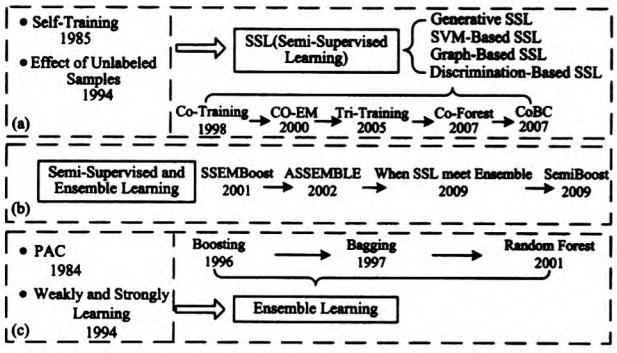


图 2 半监督集成学习发展过程

2.1 半监督学习

半监督学习是指学习器不依赖外界交互、自动地利用未标记样本和已标记样本来提升学习性能的学习方式[5]。半监督学习是在已标记样本稀少、获取困难，人工标注费时费力，但未标记样本却充足易得的客观现实条件下产生的[6]。无标记样本在学习器建模中能发挥作用的根本原因在于他们和已标记样本都是独立同分布地采样于相同的数据源。在半监督学习中，利用无标记样本的增益信息时主要基于“聚类假设”和“流行假设”，而这两种假设的本质都是指“相似的样本拥有相似的输出”[5]。

半监督学习的发展过程见图 2(a)。半监督学习最早可追溯到 1985 年的自训练学习[6-7]。1992 年，Merz 等[8-9]第一次使用“Semi-Supervised”一词。1994 年，Shahshahani 等[2]指出了使用未标记样本有助于减缓小样本下的“Hughes”现象，确立了无标记样本或半监督学习的价值和地位。根据目前主流分类方式，半监督学习主要分为基于生成式的半监督学习、基于 SVM 的半监督学习、基于图的半监督学习和基于分歧的半监督学习四大类[5-6]。也有人把自训练半监督方法单独列为一类[10-11]，自训练半监督可看作是早期利用无标记样本的一种原始框架，当学习器使用经验风险最小原则和 0-1 损失函数时，理论上无标记样本对于学习器没有帮助[6,8]，噪声和可分性较差的类别容易在算法迭代过程中导致误差累积，恶化学习器性能[5,12]。

2.2 集成学习

集成学习的主要思路是先通过一定的规则生成多个学习器，再采用某种集成策略进行组合，最后综合判断输出最终结果。一般而言，通常所说的集成学习中的多个学习器都是同质的“弱学习器”，基于该弱学习器，通过样本集扰动、输入特征扰动、输出表示扰动、算法参数扰动等方式生成多个学习器，进行集成后获得一个精度较好的“强学习器”[5]。随着集成学习研究的深入，其广义的定义[13]逐渐被学者们所接受，它是指对多个学习器集合采用学习的方式，而不对学习器性质加以区分。根据这一定义，多学习器系统(multi-classifier system)、多专家混合(mixture of experts)以及基于委员会的学习(committee-based learning)等多个领域都可以纳入到集成学习中[14]。但目前仍然以同质分类器的集成学习研究居多。

集成学习的主要发展过程见图 2(c)。集成学习的理论基础是 PAC 理论[15]和强可学习与弱可学习理论[3]。集成学习的理论基础表明强可学习器与弱可学习器是等价的[14]，因此可以寻找方法将弱可学习器转换为强可学习器，而不必去直接寻找较难发现的强可学习器。目前具有代表性的集成学习方法有 Boosting，Bagging，随机森林。以二分类问题为例，假如有 N 个分类器相互独立，错误率都为 P，使用简单的投票法组合分类器，其分类器的错误率为[5,16]：

$$P_{error}=\sum_{k=0}^{N/2}\binom{N}{k}(1-p)^k p^{N-k} \tag{1}$$

从式(1)可看出，$P<0.5$ 时，错误率 P_{error} 随 N 增大而减少。如果每个分类器的错误率都小于 0.5，且相互独立，那么集成学习器个数越多，错误率越小，N 无穷大时，错误率为 0。另外以表 1 为例[5]，√代表分类正确，×代表分类错误，集成方法为简单投票。从表 1 中可看出，只有当个体分类器“好而不同”时，集成才能发挥作用。因此，集成学习的关键在于多个学习器的两个指标，即“精确性”和“多样性”。

表 1

	测试例 1	测试例 2	测试例 3		测试例 1	测试例 2	测试例 3		测试例 1	测试例 2	测试例 3
h1	√	√	×	h1	√	√	×	h1	√	×	×
h2	×	√	√	h2	√	√	×	h2	×	√	×
h3	√	×	√	h3	√	√	×	h3	×	×	√
集成	√	√	√	集成	√	√	×	集成	×	×	×
	集成提升性能				集成不起作用				集成起负作用		

2.3 半监督集成学习

2.3.1 基于分歧的半监督学习

半监督集成学习最早可追溯到 Co-training 算法[17]，该方法又称为协同训练方法，如图 2(a)所示。Co-training 算法基于这样的前提假设：对于学习目标而言，存在两个条件独立且充分(Sufficient)冗余(Redundant)的视图，在每个视图上利用已有的已标记样本训练学习器，利用训练好的学习器对部分置信度较高的未标记样本进行标记，然后将这些伪标记样本提供给对方视图的学习器作为已标记样本使用，反复迭代至两个学习器都不再更新为止。该方法为半监督学习开辟了新的分支，即基于分歧的半监督学习，并成为了这一方向的代表性方法。在该思路的启发下，Nigam 提出 CO-EM 方法[18]，该方法将 EM 生成式模型扩展成两视图来联合进行参数估计的半监督学习。随后 Brefeld[19] 又提出了将 CO-EM 扩展至使用 SVM 作为基础学习器的变种方法。

对于 Co-training 方法而言，若能满足其算法的前提假设，利用未标记样本可将学习器性能提升到任意高[20]，但是现实中却难以找到满足前提假设的条件(存在充分冗余的多视图)。为解决这一问题，周志华等[20]发展了基于单视图的 Tri-training 方法，又称为三体训练法[21]。该方法在单个视图下训练 3 个学习器，通过 3 个学习器对未标记样本进行标记，以“少数服从多数”的原则决定未标记样本的最终伪标记，并将伪标记样本加入已标记样本集中重新训练“少数”学习器。LI 等[22]又将 Tri-training 理论推广至决策树，提出了 Co-Forest协同森林法。Hady 等[23]利用多个存在分歧的学习器集成来进行半监督学习，提出了 CoBC 方法框架。

Co-training 与 Tri-training 作为两种典型代表方法，在基于分歧的半监督学习范式的发展中起到重要作用。国内外学者在此基础上又提出了多种其他分歧半监督方法，并将其广泛应用到多个领域[17,24-25]。更重要的是，人们逐渐认识到使用多个学习器能够为半监督学习带来一定的好处。基于分歧的方法作为半监督学习与集成学习研究之间的纽带，使集成学习与半监督学习的交叉研究越来越多[21,26]。

2.3.2 半监督集成学习的理论

半监督集成学习理论的研究晚于其方法本身的研究。在 2002 年，Bennett 和 Demiriz 提出了 ASSEMBLE 方法[27](见图 2(b))，第一次使用了“Semi-supervised Ensemble”即半监督集成这一概念。实际上，ASSEMBLE 是在 SSMBoost 方法[28]的基础上改进而来，这两种方法在本质上接近，都利用无标记样本和集成学习技术来最大化分类决策边界的间隔(Margin)[29]。但由于当时研究者对半监督学习与集成学习之间的互益性的理解有限，因此直到近些年其理论与应用研究得到进一步进展后才引起业界重视。2009 年，SemiBoost 方法[29]的提出进一步夯实了半监督集成这种新的学习方式的基础。

2009 年，Zhou[4] 发表了《When Semi-supervised Learning Meets Ensemble Learning》，以基于分歧的半监督方法为例，从理论和实践角度论证了集成学习和半监督学习之间的互益性。Zhou 认为[4]：对集成学习而言，可通过利用半监督学习引入无标记样本来解决集成学习有标记样本量不足的问题，且半监督学习还可以增加集成学习中学习器的多样性；对于半监督学习而言，集成学习能进一步降低半监督学习的泛化误差，且使得半监督学习的收敛速度更快。

从模型的输入、构建、输出三位一体的整体角度来看，可以将集成学习和半监督学习看作是现今成熟的监督学习方法向不同方向进行扩展的两种范式。半监督是从模型的输入角度进行扩展，将有标记的和大量未标记样本的组合作为训练样本；而集成学习则是在输出结果上进行扩展，综合考虑不同学习器的输出结果。因此从直观上看，将半监督学习和集成学习进行结合的动机是非常合理的。

3 半监督集成学习的分类

本文认为半监督与集成学习的组合学习方式源于基于分歧的半监督学习思路。半监督集成的组合学习新方式主要是探究如何将半监督学习和集成学习二者有机结合，取长补短，充分利用二者优势，最终达到提高整体学习器的性能的目的。根据组合方式的不同将其分为基于半监督的集成学习与基于集成学习的半监督学习两大类。半监督集成与集成半监督的判断标准是训练完成后用于分类的结果学习器是否为单一学习器，若最终用来预测的结果学习器是单一学习器，则为集成半监督学习，否则为半监督集成学习。

3.1 基于半监督的集成学习

基于半监督的集成学习方法如表 2 所列。

表 2　基于半监督的集成学习方法

类别	代表方法	方法描述	串行/并行
半监督 Boosting	SSMBoost[28]	利用无标记样本最大化 Boosting 分类间隔	串行
	ASSEMBLE[27]	利用无标记样本最大化 Boosting 分类间隔	串行
	SemiBoost[29]	通过 Boosting 方式划分并集成多个半监督分类器	串行
	SemiBoost-CR[10]	在 SemiBoost 基础上，同时选取置信度较高和置信度较低的未标注样本，以对训练集进行扰动，提高集成学习器多样性	串行
半监督 Bagging	Semi-Bagging[30]	通过 Bagging 方式划分并集成多个半监督分类器	并行
	Extends RF to SSL[31]	将随机森林中引入无标记样本计算	并行
半监督随机子空间	Rel-RASCO[32]	将相关随机子空间方法和 Co-training 相结合	并行
	SSC-RSDR[33]	利用随机子空间方法对基于图的半监督方法进行降维，最终进行集成输出	并行

基于半监督的集成学习在本质上是一种集成方法，主要利用半监督学习的思路引入无标记样本来增加集成学习器的多样性，同时又可以通过无标记样本来解决集成学习训练样本不足的问题。此外，还可以将无标记样本用于集成学习器的多样性判断。根据集成学习的分类可以将基于半监督的集成学习分为半监督 Boosting、半监督 Bagging、半监督随机子空间三大类。

3.1.1 半监督 Boosting

Boosting 方法是一类算法思想的统称。它的主要思想是先利用初始训练集训练一个基学习器，再根据基学习器的精度调整对应的样本权重，使得分错的样本得到更多的关注。然后再利用调整后的样本分别训练下一轮学习器，重复迭代至满足指定次数 T，并将 T 组基学习器加权组合。常见的 Boosting 类方法有 AdaBoost 算法[34]、Gentleboost 算法[35]、Logitboost 算法[36]等。而半监督 Boosting 方法就是为了将半监督技术引入至 Boosting 中，使其能够利用未标记样本增强集成学习性能。图 3 是半监督 Boosting 的一个示意框图，A_i 表示半监督学习器，α_i 表示第 i 个学习器的权重。该过程可看成是 Boosting 与半监督学习在横、纵方向上的整合。横向上，首先采用 Boosting 的思想按先后次序生成多个分类器，而这若干个分类器都是在纵向上通过利用半监督学习方式训练得到的；然后通过线性加权模型将多个分类器组合为集成学习器。

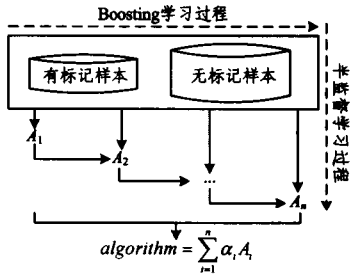


图 3 半监督 Boosting 过程示意图

SemiBoost[29]是一种典型的半监督 Boosting 方法，该方法同时基于半监督流行假设与聚类假设，同时利用有标记样本、无标记样本和相似矩阵作为输入。该方法利用相似矩阵挑选无标记样本，并使用当前的集成学习器对无标记样本进行标记作为伪标记样本，最后将伪标记样本加入训练集以参与下一次迭代，直到集成分类器满足输出条件才退出迭代过程。该算法的关键在于：1)在每次迭代时，如何选择最佳的无标记样本参与下一次模型训练；2)采取何种策略对无标记样本进行标记。SemiBoost 框架[29]并不意味着给出一个确定的高精度学习器，而是为提升学习器的性能提供一个良好框架。SemiBoost 框架能显著地提升基学习器的精度，在使用少量标记样本的前提下，借助大量无标记样本，使用该框架能将基学习器提升到与现有监督分类器相当的水平，而且其算法稳定性在实际分类中非常有价值。

唐焕玲等[10]在 SemiBoost 的基础上进行改进，提出了 SemiBoost-CR 框架，该分类模型利用最大差距和 K 近邻两种方法进行未标记样本置信度度量，在利用置信度筛选样本时另辟蹊径，同时选取置信度较高和置信度较低的未标注样本，分别以不同的策略将其加入到训练样本中，以对训练集进行扰动。其实证实验显示，SemiBoost-CR 方法能够有效提升 Naïve Bayesian 分类器性能，并成功应用到文本分类领域[10]。

此外，对于在线学习(Online Learning)而言，在每次更新学习器时容易引入误差，并最终可能导致学习器性能的恶化。半监督 Boosting 是一种非常适合于在线学习的学习方式，因为 Boosting 是一种顺序学习的方式，而半监督又可以不断地利用新加入的无标记样本对学习模型进行修正，所以半监督 Boosting 可以同时利用二者的优点帮助解决在线学习的问题。目前已有一些研究将半监督 Boosting 应用于在线学习中[37-40]。

3.1.2 半监督 Bagging

半监督 Bagging 旨在将 Bagging 方法扩展至半监督范畴。Bagging 方法是一种并行的集成学习方法，其核心在于利用放回抽样从初始训练集中随机抽取样本组成训练样本，初始训练集中样本在整个迭代过程中可能出现多次也可能不出现，每轮迭代用于训练的样本之间相互独立。与半监督 Boosting 方法不同，半监督 Bagging 的迭代过程基于 Bagging 的基本思想，如图 4 所示。其中，A_i 表示半监督学习器，y 表示对应样本的真实标签，在横向上通过 Bagging 抽样方法[41]抽样选出 n 套有标记样本与未标记样本集，然后在纵向上分别利用 n 组标记样本与无标记样本集独立进行半监督训练，挑选精度最高的组合进行集成作为结果分类器。

Li 等[30]提出 semi-Bagging 方法来解决社区问答系统(Community Question Answering，CQA)的文本分类问题。semi-Bagging 属于半监督 Bagging 范畴。该算法的基本步骤如下：同时使用有标记样本和无标记样本作为输入，通过 Bagging 抽样方法从有标记样本中抽取多组训练集，以 KNN 分类器作为基学习器组成初始集成学习器；再通过初始集成学习器确定无标记样本的伪标记，并利用伪标记样本更新原标记样本集；最后利用新的标记样本集进行 Bagging 集成学习器的训练。

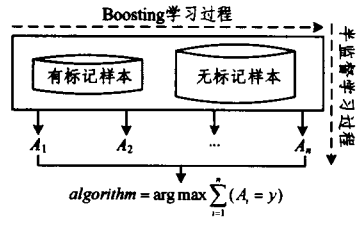


图 4 半监督 Bagging 过程示意图

Semi-Bagging 伪代码[30]如下：

输入：标记样本集 L，未标记样本集 U

过程：1. Bagging=Bagging. buildclassfier(L)

2. KNN=KNN. Buildclassifier(L)

3. for each x∈U do

if $h_1(x)=h_2(x)$

then $L=L\cup\{x,h_1(x)\}$

end for

Semi-Bagging=Bagging. buildclassifier(L)

输出：$h(x)=\arg\max\sum h_i(x)=y$

半监督 Bagging 方法继承了 Bagging 方法的优点，集成学习器的多个子学习器之间的计算相互独立，因此可以通过并行运算加快数据处理速度，非常有助于节省时间开销。半监督 Bagging 方法易于改造，使得计算机能够进行并行计算，将其应用到大规模处理任务中。这种并行方法效率较高，但

精度低于半监督 Boost 方法[16]。

3.1.3 半监督随机子空间

半监督随机子空间是一种将随机子空间与半监督学习结合的方法。随机子空间通过随机抽样抽取样本的不同特征，达到生成多样化的学习器的目的。无论是在数学领域还是机器学习领域，随机子空间都得到了广泛的研究[42-44]，著名的随机森林模型[45]就是基于随机子空间和 Bagging 抽样方式，通过对训练样本的特征集在行和列上进行采样来增加学习器多样性的决策树集成分类器。在半监督集成领域，许多学者对基于半监督的随机森林方法进行了研究，这方面的研究极少有固定的研究框架，一般是利用半监督技术带来的无标记样本信息，采取不同的方式对随机森林的各方面问题进行完善与优化。

Leistner 等[31]在分析随机森林与半监督学习各自优缺点的基础上，使用确定性退火算法优化样本（包含标记样本与未标记样本）的类间边界，将随机森林方法扩展到半监督领域。该方法不需要使用传统的“one-verse-all”或者“one-verse-one”的方法便可将二类分类半监督算法直接应用到多类分类问题上。实证实验表明，单纯的半监督算法在迭代一定次数后误差不降反升，而半监督随机森林算法的误差可随着迭代次数的增加而下降[31]。

Liu 等[32-33]分析出小样本下的随机森林的瓶颈在于树的节点划分操作，于是引入半监督学习并加入无标记样本来辅助随机森林的节点划分。该方法能避免因较差的初始值和局部最优导致的过度拟合问题。半监督随机子空间方法的计算复杂度较低，且受初始值影响较小，算法鲁棒性较高[32-33]。此外，Xia 等[46]利用半监督特征提取技术（SemiSupervised Feature Extraction，SSFE）构建了多个随机森林分类器，并在此基础上对分类结果通过多数投票的方式进行集成，其集成结果精度要高于只使用 SSFE 技术或者随机森林方法的结果精度，这也从侧面说明了将未标记样本引入集成学习带来的好处。

3.2 基于集成的半监督学习

基于集成学习的半监督方法本质上可以理解为半监督方法。将集成学习应用于半监督主要有两种做法：1）通过集成学习控制无标记样本的标注过程来减少未标记的不确定性；2）融入集成学习来优化半监督学习中的分类决策边界间隔。基于集成的半监督方法如表 3 所列。

表 3　基于集成的半监督方法

类别	代表方法	方法描述	串行/并行
使用集成学习减少无标记样本风险	Tri-training[20]	使用 3 个分类器投票决定无标记样本的标签	并行
使用集成学习优化半监督学习分类间隔	Rough set based on semi-ensemble[47]	粗糙集，集成学习优化负标签的边界	串行
	Information-theoretic Regularized Semi-Boost[39]	将信息论指标（信息熵和互信息）作为正则化项引入半监督 Boosting 中	串行

3.2.1 使用集成学习减少无标记样本风险

很多研究发现，引入无标记样本后可能导致原有学习器性能恶化[48]。在半监督学习器的训练过程中，为了控制预测无标记样本时所存在的风险，学者们常常会对标注过程进行风险控制，例如对无标记样本与有标记样本之间的相似程度进行度量，选择置信度高的无标记样本作为伪标记样本参与学习器的迭代更新；而另一种更为常用的方法则是利用集成学习的思想对标注过程进行风险控制，其过程如图 5 所示。该过程的关键在于对无标记样本进行预测时选择多个学习器集成判断无标记样本的最终标记。总体而言，在利用集成学习方法辅助半监督学习方面的方法研究较少，主要是存在如下矛盾：半监督学习的应用场景是有标记样本不足的情况，然而传统的集成学习本身就需要大量的标记样本进行训练，所以在利用集成学习减少未标记样本在标记注过程中所引入的不确定信息时存在着先后矛盾问题。

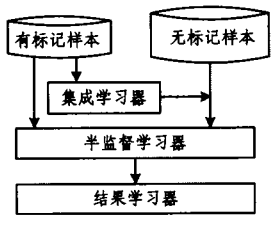


图 5　使用集成学习减少无标记样本的风险

Zhou 等[26]为了解决 Co-training 在现实中难以找到充分冗余的视图而提出了一种改进的协同训练方法 Tri-training。Tri-training 方法的一个突出特点是在半监督学习过程中采用了 3 个学习器，利用 Bootstrap Sampling 放回重复采样法将原始训练集分为 3 个部分，在每部分训练样本集上分别训练学习器。每个未标记样本都由其他两个学习器进行预测，若其他两个学习器预测结果一致，则认为该结果具有较好的置信度，将其加入第 3 个学习器的训练样本集中，并更新第 3 个学习器。对待分类样本进行预测时，同时使用 3 个学习器进行投票集成判断输出结果。虽然从最后的结果分类器看 Tri-training 似乎是一种集成学习器，但实际上在选择最后用来预测的结果分类器时，既可以采用 3 个学习器集成预测，也可以任选三者之一用来预测。所以从该方法的初衷与训练过程来看，一般认为 Tri-training 是一种利用集成学习思路来辅助无标记样本的标注过程的半监督学习方法。此外，Vote-training[49]，ENSSL[50]等也可认为是通过集成学习来降低无标记样本的标注过程的风险的方法。

这类方法之所以能取得较好的效果，本文认为主要得益于：1）沿用了集成学习中使用的放回重复采样方法（例如 bootstrap sampling 自助采样法[51]）和结果集成方法（例如投票、加权平均、Stacking）；2）在学习器的数目上，该类方法相对于普通集成学习器而言只利用少量的异质性学习器进行集成，判断无标记样本的伪标记，再将伪标记样本添加到半监督学习的过程中。这样减少了集成学习器的数目，降低了对有标记样本数目的要求。通过样本采样、属性特征扰动等生成的集成方法能够增强无标记样本标记过程的安全性，控制半监督学习过程中的风险，进一步提升半监督集成学习的精度。

3.2.2 使用集成学习优化半监督学习分类间隔

对于一般的分类问题而言，另一种基于集成学习的半监督方法是在半监督学习原有的基础上，通过引入集成学习进一步优化其分类边界间隔。分类边界间隔问题一直以来都是统计学习理论领域的重要研究问题。采取类似于半监督学习

中利用规则化项对学习器进行约束的方式，在原有集成半监督的目标函数上添加了规则化项，以缓解半监督过度拟合问题。该方法对于纯半监督学习而言，本质上是增加了集成学习对半监督学习器的分类边界间隔进行优化的过程。

这类研究的过程过于灵活，与其他类别的半监督集成方法的界限常常非常模糊，这类方法通常需要引入其他方法或者技术来解决特定的问题。例如，Shi 等[47]将粗糙集引入解决只有正类标记样本与无标记样本(不存在负类标记样本)的二类分类问题中，使用 3 个学习器进行集成，利用分类边界间隔改善负类标记的边界，即通过粗糙集与集成学习使正类标记的边界在特征空间中尽可能收敛于其真实边界。Zheng 等[39]将基于有标记样本的对数损失函数项和利用无标记样本计算的信息论指标作为正则化项，并将其引入半监督 Boost 中，分别选择信息熵和互信息两种信息论指标作为正则化项。其研究结果表明，引入两种信息论指标的方法的目标函数的收敛速度均快于常用的 AdaBoost 方法，而且减小了误差。Chen 等[52-53]提出规则化半监督 Boosting 方法，利用常用的损失函数优化框架，将局部光滑规则化项引入半监督的 Boost 算法中，这一做法能够提升学习器的整体泛化性能，加快训练速度。

将附加的测度项(例如粗糙集、信息熵等)加入目标函数中，本质上是利用更多的外界测量指标通过集成学习进一步优化半集成半监督学习器模型，使其进一步满足于真实分布，提高了分类器的精度，但是另一方面也增加了目标函数的求解复杂度，使得算法泛化性能下降。

结束语 总体而言，半监督集成学习属于机器学习领域中较新的研究方向，半监督集成学习作为半监督学习和集成学习的交叉学习方式，充分利用了其二者的优势，在多个应用领域取得了较好的效果。在整体分析半监督集成学习的现状的基础上，本文对未来的研究有如下思考：

(1)根据实际应用特点，因地制宜选择半监督集成方法。

各类方法存在较大的差异，难以找到统一的半监督集成研究框架，所以应根据应用场景的特点选择合适的半监督集成方法才能发挥其优势与作用。例如，在线学习与离线学习在训练样本的选择上是截然不同的。对于在线学习(Online Learning)，学习器的更新容易引入误差，并最终可能导致学习器性能的恶化，使用半监督 Boosting 的同时利用无标记样本和 Boosting 算法的优势可以让在线学习有效避免这一问题。除了在线数据外，现实生活中更多的是离线数据处理，其特点是数据处理量和训练所需的样本量非常大，训练样本的标记过程费时费力。半监督随机子空间和 Semi-Supervised Random Bagging 都是并行处理的算法，因此可以通过此类算法加快数据处理速度。

(2)进一步深入基于集成学习的半监督研究，特别是同质性集成向异质性集成转变。

深入分析基于半监督的集成学习与基于集成的半监督学习两大类方法，不难发现目前基于半监督的集成方法居多，其中又以半监督 Boosting 方法研究最广，然而基于集成学习的半监督学习方法较为稀少。以后的研究中，有必要对其他的半监督集成方法做进一步的研究。特别是探讨如何利用集成学习的机制进行控制或提高半监督学习的安全性。例如，利用同质性集成学习中的各类扰动方法结合异质性学习器的集成方法来辅助半监督学习是有益的方向。

(3)利用无标记样本同时辅助集成学习的训练过程和多样性度量。

学习器多样性是集成学习难以回避的问题，尽管目前还难以从理论上明确论证学习器多样性的作用，但是多样性度量俨然已经成为评价集成学习器的重要指标[54-56]。目前绝大多数的集成方法的训练过程和多样性度量都依赖于大量的有标记样本，而现实中有标记样本不足往往成为集成学习发挥优势的瓶颈。基于半监督的集成学习旨在解决集成学习在训练过程中有标记样本不足的问题，如果在将来的研究中不仅仅考虑如何通过半监督方式减少集成学习对有标记样本的依赖程度，同时又能利用无标记样本对集成学习器进行显式或隐式的多样性度量，则能更大程度地挖掘无标记样本作用。

参 考 文 献

[1] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[2] SHAHSHAHANI B M, LANDGREBE D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[3] KEARNS M, VALIANT L. Cryptographic limitations on learning Boolean formulae and finite automata[J]. Journal of the ACM (JACM), 1994, 41(1): 67-95.

[4] ZHOU Z. When Semi-supervised Learning Meets Ensemble Learning[M]// Benediktsson J A, Kittler J, Roli F. Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 529-538.

[5] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[6] 张晨光，张燕. 半监督学习[M]. 北京：中国农业科学技术出版社，2013.

[7] ERGER J O. Statistical decision theory and Bayesian analysis [M]. Springer Science & Business Media, 2013.

[8] CHAPELLE O, SCHÖLKOPF B, ZIEN A. Semi-Supervised Learning [M]. Cambridge, Massachusetts, USA: The MIT Press, 2006.

[9] MERZ C J, CLAIR D S, BOND W E. Semi-supervised adaptive resonance theory (smart2)[C]// IEEE. 1992: 851-856.

[10] 唐焕玲，鲁明羽. 利用置信度重取样的 SemiBoost-CR 分类模型[J]. 计算机科学与探索，2011(11): 1048-1056.

[11] 李亚楠. 基于 Self-training 的步态识别研究[D]. 济南：山东大学，2013.

[12] 谭琨. 高光谱遥感影像半监督分类研究[M]. 徐州：中国矿业大学出版社，2014.

[13] OPITZ D, MACLIN R. Popular ensemble methods: An empirical study[J]. Journal of Artificial Intelligence Research, 1999, 11: 169-198.

[14] 张燕平，张玲. 机器学习理论与算法[M]. 北京：科学出版社，2012.

[15] VALIANT L G. A theory of the learnable[J]. Communications of the ACM, 1984, 27(11): 1134-1142.

[16] 夏俊士. 基于集成学习的高光谱遥感影像分类[D]. 徐州：中国矿业大学，2013: 138.

[17] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]// 11th Conference on Computational Learning Theory. ACM, 1998: 92-100.

[18] NIGAM K, GHANI R. Analyzing the effectiveness and applicability of co-training[C]// ACM. 2000: 86-93.

[19] BREFELD U, SCHEFFER T. Co-EM support vector learning [C]// International Conference on DBLP. 2004: 16.

[20] ZHOU Z, LI M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[21] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013(11): 1871-1878.

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

[23] HADY M F A, SCHWENKER F. Co-Training by Committee: A Generalized Framework for Semi-Supervised Learning with Committees[J]. Int. J. Software and Informatics, 2008, 2(2): 95-124.

[24] HADY M F A, SCHWENKER F, PALM G. Semi-supervised learning for tree-structured ensembles of RBF networks with co-training[J]. Neural Networks, 2010, 23(4): 497-509.

[25] 邓超, 郭茂祖. 基于自适应数据剪辑策略的 Tri-training 算法[J]. 计算机学报, 2007(8): 1213-1226.

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

[27] BENNETT K P, DEMIRIZ A, MACLIN R. Exploiting unlabeled data in ensemble methods[C]// Acm Int. Conf. Knowledge Discovery & Data Mining. 2002: 289-296.

[28] GRANDVALET Y, AMBROISE C. Semi-supervised margin-boost[M]// Advances in Neural Information Processing Systems. 2001: 553-560.

[29] MALLAPRAGADA P K, JIN R, JAIN A K, et al. Semiboost: Boosting for semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2000-2014.

[30] LI Y, SU L, CHEN J, et al. Semi-supervised Question Classification Based on Ensemble Learning[M]// Advances in Swarm and Computational Intelligence. 2015: 341-348.

[31] LEISTNER C, SAFFARI A, SANTNER J, et al. Semi-supervised random forests[C]// IEEE Conf. on Computer Vision. IEEE. 2009: 506-513.

[32] LIU X, SONG M, TAO D, et al. Semi-supervised node splitting for random forest construction[C]// CVPR 2013. 2013: 492-499.

[33] LIU X, SONG M, TAO D, et al. Random Forest Construction with Robust Semisupervised Node Splitting[J]. IEEE Transactions on Image Processing, 2015, 24(1): 471-483.

[34] FREUND Y S R E. Experiments with a new boosting algorithm [M]. San Francisco, California: Morgan Kaufmann, 1996.

[35] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Additive logistic regression: a statistical view of boosting[J]. The Annals of Statistics, 2000, 28(2): 337-407.

[36] CAI Y, FENG K, LU W, et al. Using LogitBoost classifier to predict protein structural classes[J]. Journal of Theoretical Biology, 2006, 238(1): 172-176.

[37] CHEN S, ZHU S, YAN Y. Robust visual tracking via online semi-supervised co-boosting[J]. Multimedia Systems, 2015: 1-17.

[38] ZEISL B, LEISTNER C, SAFFARI A, et al. On-line semi-supervised multiple-instance boosting[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE. 2010: 1879.

[39] ZHENG L, WANG S, LIU Y, et al. Information theoretic regularization for semi-supervised boosting[C]// Acm Sigkdd Internation Conference on Knowledge Discovery and Data Mining. 2009: 1017-1026.

[40] GRABNER H, LEISTNER C, BISCHOF H. Semi-supervised on-line boosting for robust tracking[M]// Computer Vision-ECCV 2008. Springer, 2008: 234-247.

[41] 杜培军, 夏俊士, 薛朝辉, 等. 高光谱遥感影像分类研究进展[J]. 遥感学报, 2016, 20(2): 236-256.

[42] TAO D, TANG X, LI X, et al. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1088-1099.

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

[44] SKURICHIAN M, DUIN R P. Bagging, boosting and the random subspace method for linear classifiers[J]. Pattern Analysis & Applications, 2002, 5(2): 121-135.

[45] LIAW A, WIENER M. Classification and regression by randomForest[J]. R News, 2002, 2(3): 18-22.

[46] XIA J, LIAO W, CHANUSSOT J, et al. Improving random forest with ensemble of features and semisupervised feature extraction[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(7): 1471-1475.

[47] SHI L, MA X, XI L, et al. Rough set and ensemble learning based semi-supervised algorithm for text classification[J]. Expert Systems with Applications, 2011, 38(5): 6300-6306.

[48] BELLAL F, ELGHAZEL H, AUSSEM A. A semi-supervised feature ranking method with ensemble learning[J]. Pattern Recognition Letters, 2012, 33(10): 1426-1433.

[49] 王轶初. 基于集成学习的半监督学习算法研究[D]. 西安: 西安电子科技大学, 2011.

[50] 葛荐. 基于集成算法的半监督学习研究[D]. 南京: 南京信息工程大学, 2012.

[51] EFRON B, TIBSHIRANI R J. An introduction to the bootstrap [M]. CRC Press, 1994.

[52] CHEN K, WANG S. Regularized boost for semi-supervised learning[C]// Conference on Neural Information Processing Systems. 2008: 281-288.

[53] CHEN K, WANG S. Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(1): 129-143.

[54] CUNNINGHAM P, CARNEY J. Diversity versus quality in classification ensembles based on feature selection[M]. Springer, 2000: 109-116.

[55] COZMAN F G, COHEN I, CIRELO M. Unlabeled Data Can Degrade Classification Performance of Generative Classifiers[C]// IEEE Asia-pacific Service Computing Conference. 2002: 327-331.

[56] 孙博, 王建东, 陈海燕, 等. 集成学习中的多样性度量[J]. 控制与决策, 2014(3): 385-395.