计算机科学
COMPUTER SCIENCE

基于分类不确定性最小化的半监督集成学习算法

何玉林, 朱鹏辉, 黄哲学, Fournier-Viger PHILIPPE

引用本文

何玉林, 朱鹏辉, 黄哲学, Fournier-Viger PHILIPPE. 基于分类不确定性最小化的半监督集成学习算法[J]. 计算机科学, 2023, 50(10): 88-95.

HE Yulin, ZHU Penghui, HUANG Zhexue, Fournier-Viger PHILIPPE. Classification Uncertainty Minimization-based Semi-supervised Ensemble Learning Algorithm [J]. Computer Science, 2023, 50(10): 88-95.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600048

基于分类不确定性最小化的半监督集成学习算法

何玉林[1,2]　朱鹏辉[2]　黄哲学[1,2]　Fournier-Viger PHILIPPE[2]
1 人工智能与数字经济广东省实验室(深圳)　广东 深圳 518107
2 深圳大学计算机与软件学院　广东 深圳 518060

摘　要　半监督集成是将半监督学习与集成学习相结合的一种学习范式,它一方面通过无标记样本来提高集成学习的多样性,同时解决集成学习样本量不足的问题,另一方面集成多个分类器能够进一步提升半监督学习模型的性能。现有的研究从理论和实践两个角度证明了半监督学习与集成学习之间的互益性。针对当前半监督集成学习算法对无标记样本信息利用不完全的缺陷,文中提出了一种新的基于分类不确定性最小化的半监督集成学习(Classification Uncertainty Minimization-Based Semi-Supervised Ensemble Learning,CUM-SSEL)算法,它引入信息熵作为对无标记样本进行打标的置信度评判标准,通过最小化无标记样本打标过程中的不确定性迭代地训练分类器,实现对无标记样本的高效利用,以增强分类器的泛化性能。在标准的实验数据集上对 CUM-SSEL 算法的可行性、合理性和有效性进行了验证,实验表明:随着基分类器的增加,CUM-SSEL 算法的训练呈现收敛的趋势,同时它能够获得优于 Self-Training,Co-Training,Tri-Training,Semi-Boost,Vote-Training,Semi-Bagging 以及 CST-Voting 算法的分类精度。

关键词:半监督集成学习;集成学习;半监督学习;分类不确定性;置信度;信息熵
中图法分类号　TP391

Classification Uncertainty Minimization-based Semi-supervised Ensemble Learning Algorithm

HE Yulin[1,2],ZHU Penghui[2],HUANG Zhexue[1,2] and Fournier-Viger PHILIPPE[2]
1 Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ),Shenzhen,Guangdong 518107,China
2 College of Computer Science & Software Engineering,Shenzhen University,Shenzhen,Guangdong 518060,China

Abstract　Semi-supervised ensemble learning(SSEL) is a combinatorial paradigm by fusing semi-supervised learning and ensemble learning together,which improves the diversity of ensemble learning by introducing unlabeled samples and at the same time solves the problem of insufficient sample size for ensemble learning. In addition,SSEL can improve the generalization capability of classification system by integrating multiple classifiers trained on the highly-credible labeled samples. The existing researches have proved the mutual benefit between semi-supervised learning and integrated learning from both theoretical and practical perspectives. The existing SSEL algorithms are unable to make full use of the unlabeled samples,which limit their prediction capabilities when handling the classification problems with less labeled samples. This paper proposes a novel classification uncertainty minimization-based semi-supervised ensemble learning(CUM-SSEL) algorithm,which introduces the information entropy as the criterion of confidence and uses the characteristics of information entropy to minimize the classification uncertainty in the process of predicting unlabeled samples. The feasibility,rationality and effectiveness of CUM-SSEL algorithm are verified based on a series of persuasive experiments. Experimental results demonstrate that CUM-SSEL is a valid algorithm to deal with the semi-supervised learning problems.

Keywords　Semi-supervised ensemble learning, Ensemble learning,Semi-supervised learning,Classification uncertainty,Confidence,Information entropy

到稿日期:2023-06-05　返修日期:2023-07-28
基金项目:国家自然科学基金面上项目(61972261);广东省自然科学基金面上项目(2023A1515011667);深圳市基础研究重点项目(JCYJ20220818100205012);深圳市基础研究面上项目(JCYJ20210324093609026)
This work was supported by the National Natural Science Foundation of China(61972261), Natural Science Foundation of Guangdong Province (2023A1515011667),Key Basic Research Foundation of Shenzhen(JCYJ20220818100205012) and Basic Research Foundations of Shenzhen (JCYJ20210324093609026).
通信作者:何玉林(yulinhe@gml.ac.cn)

1 引言

半监督学习与集成学习是机器学习的两个重要分支。与传统的有监督学习不同,半监督学习[1-3]不依赖大量有标记数据,它是一种利用大量无标记数据并辅以少量有标记数据的学习方式。而集成学习[4-6]通过一定的规则生成多个学习器,再采用特定的集成策略进行组合,最后综合判断输出最终的预测结果。

近年来,半监督学习与集成学习的结合研究受到了广泛的关注。半监督集成学习(Semi-Supervised Ensemble Learning,SSEL)这一概念首次出现在由 Bennett 等[7]提出的 Assemble 方法中,它利用无标记样本和 Boosting 算法来最大化分类边界的间隔。2009 年,为了证明半监督学习和集成学习结合的有效性,Zhou[8]从基于分歧的半监督学习出发,从理论和实践两个不同的角度阐述了半监督学习和集成学习之间的互益性。对于集成学习而言,基分类器应该是准确和多样的,无标记样本可以帮助增加基分类器的多样性,当有很少的标记样本可供训练时,必须利用无标记样本来构建好的集成;对于半监督学习而言,即使无标记样本不能提高单个分类器的性能,也期待分类器集成的性能可以被进一步提高。

对于半监督集成的研究,目前可分为半监督 Boosting 和半监督 Bagging 两类。半监督 Boosting 的核心在于通过多次迭代来增加分类器的性能。在每次迭代过程中,使用当前分类器预测无标记样本的标签,将置信度高的样本加入训练集,当达到指定的学习器个数时,停止迭代,这些分类器最后被集成为一个预测功能强大的分类器。这种方式能够有效提升学习器的性能,但易受到噪声的影响,出现过拟合现象,同时训练时花费的时间开销也较大。半监督 Bagging 的核心是利用重采样的方式训练分类器,构建有差异性的分类器,接着利用迭代机制对模型进行更新。这种方式并行运算效率高,有助于节省时间开销,减少过拟合的风险,但这种方式的精度往往低于半监督 Boosting[9]。

从以上简要的分析中可以发现,无论是半监督 Bagging 还是半监督 Boosting,在对无标记样本的迭代训练完成之后,置信度低的样本并没有被完全利用,而这些置信度低的样本往往包含了额外的信息,能够进一步提升模型的泛化能力。另一方面,在对无标记样本进行预测时,由于模型自身训练不佳,预测时往往带有很大的不确定性,如何最小化这种不确定性也是一个非常关键的问题。

针对上述半监督集成算法存在的问题,本文提出了一种基于分类不确定性最小化的半监督集成学习(Classification Uncertainty Minimization-Based Semi-Supervised Ensemble Learning,CUM-SSEL)算法。首先,为了节省训练时间,CUM-SSEL 算法采用了 Bagging[10]式的集成策略训练多个基分类器;其次,Seedat 等[11]指出,在评估不确定性的方式中,信息熵可以被用于表示预测结果的不确定性。为了有效度量分类过程中的不确定性,本文引入了分类信息熵作为置信度的评判标准,通过不断地迭代训练,可以最小化样本预测过程的不确定性。最后,为了更好地利用无标记样本,预设了一个缓冲池,在迭代过程中,将置信度低的样本加入缓冲池,通过分类器的不断强化来增加缓冲池中样本的置信度,以便更大程度地发挥无标记样本的作用。

本文第 2 章介绍了相关工作;第 3 章对本文用到的预备知识进行了实验验证;第 4 章对本文提出的算法进行了详细展示;第 5 章对算法的可行性、合理性和有效性进行了实验验证;最后总结全文并展望未来。

2 相关工作

2.1 半监督 Boosting

SemiBoost 是一种经典的半监督 Boosting 算法,由 Mallapragada 等[12]提出,它是基于图半监督[13]的一种学习算法。该算法指出在采样时应该将样本间的相似度纳入考量,对样本间相似度较高但目前集成学习器的预测结果不一致性较大的样本设置更大的权重,SemiBoost 在不断迭代的过程中使用集成学习器对无标记样本进行打标,这不仅提高了模型的泛化能力,还提高了模型对相似样本预测的一致性,使模型更加稳定,该过程有效发挥了无标记样本的的作用。

Hou 等[14]提出了一种最大化样本可分性的半监督 Boosting 算法,它从半监督学习的聚类假设与流形假设出发,基于最大化全局散度和最小化局部散度[15]的思想,设计了一种高密度区域局部散度最小、样本空间全局散度最大的准则,在 Gradient Boosting 框架下有效地提高了分类的准确性。

Jafar[16]提出了 MSSBoost 算法,该算法使用了正则化单一顶点作为多类分类问题的新公式,并且使用新的 Boosting 公式将有标记和无标记数据之间的相似性信息与分类器预测相结合,将伪标记分配给无标记数据。研究结果表明,该算法能够有效利用无标记数据中的信息来提高分类器的分类性能。

Chen 等[17]基于协同训练机制提出了一种协同训练目标跟踪算法。该算法利用无标记样本协同训练出的不同视图,在解决目标漂移问题的同时,保持对目标外观的自适应更新,有效解决了目标遮挡和光照变化的问题。

2.2 半监督 Bagging

半监督 Bagging 的典型算法是由 Zhang 等[18]提出的 UDEED 方法,与一般半监督集成算法利用置信度对无标记样本进行预测的方式不同,它一方面定义了函数损失公式保证在已标记样本上经验误差达到最小,另一方面利用无标记样本最大程度地增加学习器间的多样性。

Li 等[19]提出了 Semi-Bagging 方法,该方法采用 Bagging 和 AdaBoost 这两种方式,以扩大训练空间,在训练的过程中,采用最近邻算法和 AdaBoost 作为基学习器对数据进行训练,从而减少扩展训练集中的错误标记,进而提高分类的准确性,并且成功将其应用到了社区问答系统的现实场景中。

Livieris 等[20]提出了 CST-Voting 方法,该方法集成了目前最经典的 3 种半监督算法,即 Self-training[21],Co-training[22]和 Tri-training[23],使用投票法对样本进行预测,并且把该算法应用到了肺部异常检测中。

3 预备知识

3.1 半监督集成学习框架

半监督集成学习 SSEL 经过多年的发展,已经有很多的

改进和优化，但是它的本质并没有变。SSEL 按照半监督学习的思路引入无标记样本，以此来提高集成学习器的多样性。同时又利用大量的无标记样本来解决集成学习样本训练量不足的问题。算法的框架如图 1 所示，执行步骤如下。

步骤 1　利用集成策略从有标记样本 L 中训练 n 个分类器 $C_1, C_2, \cdots, C_n$；

步骤 2　基于分类器 $C_1, C_2, \cdots, C_n$ 对无标记样本集 U 中的每一个样本 $U_i (i=1,2,\cdots,N)$ 进行置信度评判，得到高置信度样本集 $\hat{U}$，N 为数据集中无标记样本的个数；

步骤 3　将高置信度样本集 $\hat{U}$ 加入到有标记样本集 L 中；

步骤 4　将高置信度样本集 $\hat{U}$ 从无标记样本集 U 中删除；

步骤 5　重复步骤 1－步骤 4，直到达到停止准则为止。

不难看出，集成策略、置信度评判标准以及停止准则是 SSEL 算法中非常重要的组成部分：

1）目前主要的集成策略包括 Bagging 和 Boosting 两种方式。

2）置信度评判标准是用于对无标记样本进行打标的一种方式，一个好的置信度评判标准可以有效地挑选出高置信度样本，一个坏的置信度评判标准会导致模型出现过拟合或者欠拟合的现象。目前存在许多不同的置信度评判标准，如利用 KNN 算法的距离、SVM 的边界向量等。

3）为了保证算法能最终停止，需要制定停止准则。最普遍的停止准则是把所有的高置信度样本都筛选出来，直到无标记样本遍历完为止，并加入到训练集中进行模型更新。另一种常见的停止准则是达到迭代次数后停止，无论是否还有剩余无标记样本。

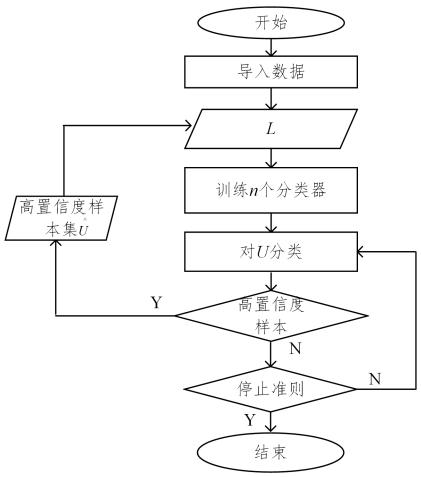


图 1　SSEL 框架示意图

Fig. 1　SSEL framework diagram

3.2　信息熵

信息熵可以度量不确定事物中蕴含的信息量，数据越有序，信息熵越低，反之，数据越混乱，信息熵越高。设 X 是一个离散的随机变量，其定义空间为 R，它的概率分布律函数如下：

$$p_i = p(X=x), x \in R \tag{1}$$

根据信息论可知，离散随机变量 X 的熵为：

$$H(X) = -\sum_{i=1}^{m} p_i \log_2 p_i, i=1,2,\cdots,m \tag{2}$$

其中，m 为 X 的取值数量。信息熵的最大值，可通过求解式(3)所示的优化问题获得。

$$f(p_1, p_2, \cdots, p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i$$
$$\text{s.t.} \sum_{i=1}^{m} p_i = 1 \tag{3}$$

构造拉格朗日函数，具体表达式如下：

$$L(p_i, \lambda) = -\sum_{i=1}^{m} p_i \log_2 p_i + \lambda\left(\sum_{i=1}^{m} p_i - 1\right) \tag{4}$$

计算其关于 p_i 的偏导数：

$$\frac{\partial L(p_i, \lambda)}{\partial p_i} = 0 \Rightarrow p_i = e^{\lambda - 1} \tag{5}$$

将式(5)代入式(3)的约束条件中可得：

$$\sum_{i=1}^{m} e^{\lambda - 1} = 1 \tag{6}$$

通过进一步求解可得熵取最大值时对应的概率值：

$$p_i = \frac{1}{m} \tag{7}$$

将式(7)代入式(3)可得最大熵：

$$H(X) = \log_2 m \tag{8}$$

4　基于分类不确定性最小化的 SSEL 算法

4.1　分类不确定性

在预测过程中，混沌的不确定性估计可能会导致样本的误分类，在迭代训练的过程中，使得错误不断叠加，模型性能不断降低。预测不一定每次都准，但是这种不确定性的最小化能够带来更好的辅助决策。Kendall 等[24]指出，不确定性分为认知不确定性和随机不确定性。

1）认知不确定性是由于模型训练效果欠佳导致的，这种不确定性可以通过增加额外的训练数据来减少，也被称为模型不确定性。而半监督学习正好符合这种特性，因此半监督学习能够很好地解决这种问题。

2）随机不确定性是由于数据标注过程中存在噪声而造成的，因此也被称为数据不确定性，与认知不确定性相比，这种类型的不确定性无法简单地通过收集更多的数据来减少。但是，集成学习却可以通过增加基学习器的个数来很好地处理学习系统的这种随机不确定性。

因此，半监督集成学习通过将半监督学习和集成学习结合来不断地标注无标记数据，以降低学习系统的不确定性。本文将学习系统的不确定性表示为数据分类结果的不确定性，这种分类不确定性既包含了认知的不确定又包含了随机的不确定。下文举一个简单的例子来直观地说明这种不确定性。假设有一个二分类问题，分类器 A 给出概率分布为(0.05,0.95)的标记结果，那么就可以认为它是属于第二类的样本；而如果分类器 B 输出的概率分布为(0.49,0.51)，同样也可以将样本判定为第二类，但是通过上面对信息熵求解的分析可知，分类器 B 对应的输出结果信息熵较大，此时分类器对样本类别的判定比较混沌。在实际应用中，我们更期待分类器能给出像 A 那样较为清晰的预测结果，即分类器对样本的分类不确定性尽可能地小。

4.2　CUM-SSEL 算法

以最大化对无标记样本信息的利用率为切入点，本文

进一步提出了一种基于分类不确定性最小化的半监督集成学习 CUM-SSEL 算法，算法的框架如图 2 所示，执行步骤如下。

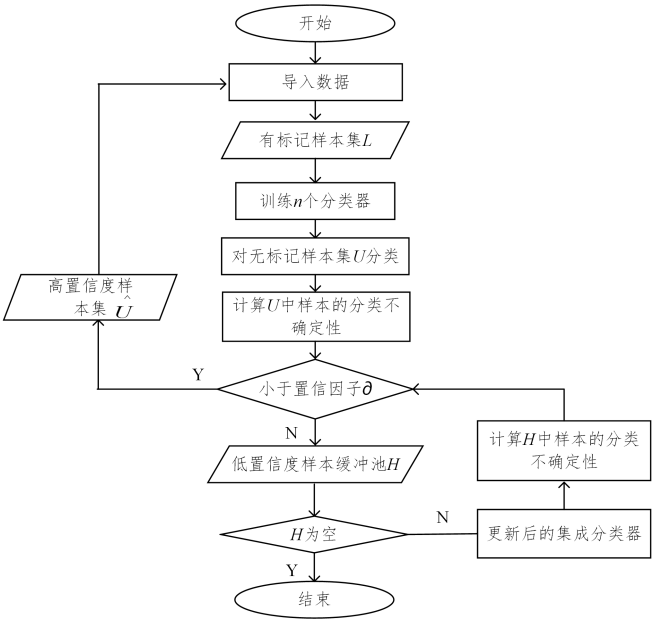


图 2 CUM-SSEL 框架示意图

Fig. 2 CUM-SSEL framework diagram

步骤 1 利用 Bagging 集成策略从有标记样本集 L 中训练 n 个分类器 $C_1, C_2, \cdots, C_n$；

步骤 2 基于分类器 $C_1, C_2, \cdots, C_n$ 对无标记样本集 U 中的每一个样本 $U_i (i=1,2,\cdots,N)$ 进行置信度评判，得到高置信度样本集 $\hat{U}$；

步骤 3 将高置信度样本集 $\hat{U}$ 加入到有标记样本集 L，并对分类器 $C_1, C_2, \cdots, C_n$ 进行更新；

步骤 4 将低置信度样本加入到缓冲池 H；

步骤 5 对缓冲池 H 中的样本进行预测，每一轮更新缓冲池的置信因子 ∂；

步骤 6 重复步骤 3－步骤 5，直到为所有无标记样本打上标签为止。

CUM-SSEL 算法的置信度评判标准是利用信息熵来挑选高置信度和低置信度的样本，样本的概率分布越确定，它的信息熵越低，样本的概率分布越不确定，它的信息熵越高。高置信度的样本用于更新模型，低置信度的样本存于缓冲池中，预设缓冲池的目的就是为了能够在每一轮迭代时更好地利用低置信度的样本。

Breiman[10] 指出，Bagging 策略采用有放回抽样的方式生成数据样本，数据样本与原始数据之间约有 63.2% 的重复，如果基分类器对训练数据分布不敏感，则得到的基分类器之间会很相似，结合之后对泛化能力的提升有限。这类对训练数据不敏感的学习器称为“稳定学习器”。比如，对于 k-近邻分类器这种高稳定分类器，Bagging 并不起作用。因此，Bagging 需要配合不稳定学习器使用，一般来说，学习器越不稳定，效果提升程度越大，而神经网络就是这样一类不稳定学习器。因此，本文使用的基分类器为神经网络，神经网络的输出层通过 Softmax 回归后，分类结果转化为概率分布。样本对应的分类不确定性计算过程如下。利用 n 个基分类器对无标记样本进行预测后，假设有 m 类样本，得到如下分类概率分布：

$$P_1=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{bmatrix} \tag{9}$$

其中：

$$\sum_{i=1}^{m} p_{ki}=1, k=1,2,\cdots,n \tag{10}$$

对每一类样本概率进行求和，得到一个新的概率向量：

$$P_2=[p_1', p_2', \cdots, p_m'] \tag{11}$$

其中：

$$p_i'=\sum_{j=1}^{n} p_{ji}, i=1,2,\cdots,m \tag{12}$$

归一化处理后可得：

$$P_3=[\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m] \tag{13}$$

计算向量 P_3 的信息熵，即：

$$Entropy(P_3)=\sum_{i=1}^{m} \bar{p}_i \log_2 \bar{p}_i \tag{14}$$

将其作为样本的分类不确定性值，其中：

$$\bar{p}_i=p_i' / \sum_{j=1}^{m} p_j', i=1,2,\cdots,m \tag{15}$$

算法 1 给出了 CUM-SSEL 算法的具体流程，其中算法的关键是在对无标记数据集迭代一轮后，对剩余的低置信度样本的利用。在每一轮迭代后，将高置信度的样本加入到训练集对模型进行更新，此时模型的性能得到了进一步的提升。而低置信度的样本加入到缓冲池中，缓冲池中的样本带有很大的不确定性，在模型更新以后，对缓冲池中的低置信度样本进行预测，某些样本在上一轮迭代时是低置信度样本，到下一轮迭代时就可能变为高置信度样本，样本的不确定性会进一步减小，这样就更有把握判断样本的类别。

算法 1 CUM-SSEL 算法

输入：有标记数据集 L；无标记数据集 U；分类器 C_i，i＝1，2，…，n；缓冲池 H

输出：训练后的分类器 C_i，i＝1，2，…，n

1. 基于自助采样法对有标记数据集 L 进行有放回抽样，并对每一个样本子集 L_i 进行训练得到 n 个分类器 C_i。

```
for i=1,2,…,n do
    Li←BootstrapSample(L)
    Ci←learn(Li)
end of for
```

2. 对于每一个无标记样本 $U_i \in U$，对其进行预测后，经过处理得到概率向量 $P_3(1\times m)$，计算其信息熵，若信息熵小于设定的置信因子 ∂，则可以确定样本的类别，打上标签，加入到样本子集 $L_i (i=1,2,\cdots,n)$ 中，更新模型。若信息熵大于置信因子 ∂，则将其放入缓冲池 H 中。

```
for Ui∈U do
  for i=1,2,…,n do
P1←predict_proba(Ci)
  end of for
  for i=1,2,…,m do
    for j=1,2,…,n do
      P2←∑pji,pji∈P1
```

```
    end of for
  end of for
  P3←Normalization(P2)
  if Entropy(P3)<∂
    Li←{Ui, y}
  else
    H←Ui
  end of for
  for i=1,2,…,n do
    Ci←learn(Li)
end of for
```

3. 对于缓冲池 H 中的每个样本，重复步骤 2，每一轮将置信因子∂扩大，当无标记样本全部更新完毕后，算法结束。

CUM-SSEL 属于迭代类算法，若置信因子∂选取过于严格，则可能会导致算法的迭代次数过多，大量的无标记样本被误分类，错误将会不断地堆叠，导致模型的学习效果较差，这种现象在 5.2 节的实验部分也有所体现。因此，为了有效缓解出现这种情况带来的负面影响，在每一轮迭代后，将置信因子∂适当扩大，以减少迭代次数，当置信因子扩大到$\log_2 m$时，所有的无标记样本都会被打上标签，使模型的泛化能力进一步提升，这样算法也最终会停止，其中 m 表示样本的类别。

5 实验验证与分析

5.1 实验数据集和运行环境

本次实验在公开的 UCI 数据集和 KEEL 数据集上开展，数据集信息如表 1 所列。数据集被划分成了 70%的训练集和 30%的测试集。本实验采取准确率(Accuracy)作为模型的评判标准，且所有实验结果均是重复 5 次独立实验的平均结果。

表 1 数据集信息

Table 1 Information of data sets

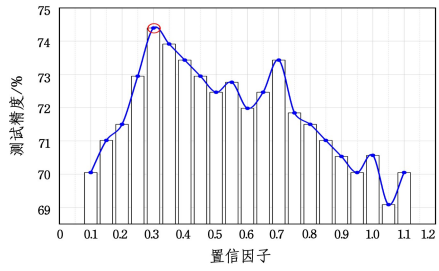
数据集	属性个数	样本个数	类标个数
australian	14	690	2
wdbc	30	569	2
algerian	12	244	2
ionosphere	34	351	2
biodegradation	42	1055	2
messidor_features	20	1151	2
vertebral	6	310	3
forest	27	326	4

实验中的 CUM-SSEL 算法使用 Python 语言实现。所有的实验均在 Intel(R) Core(TM) 3.41 GHz i7-6700 CPU 以及 8GB 内存配置的电脑上进行。

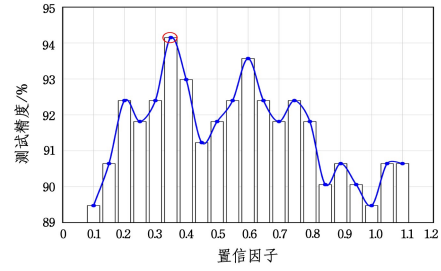
5.2 可行性验证

在 CUM-SSEL 中，如何选取置信因子∂非常重要。本文实验选取了表 1 所列的 australian 和 wdbc 数据集来展示置信因子∂对模型产生的影响，设置已标记样本占总样本的 40%，分类器个数设置为 120。

从图 3 中可以看出，随着置信因子∂持续增加，我们发现在两个数据集中均存在一个置信因子∂，使得模型达到一个较高的测试精度，说明置信因子影响着模型的性能，这表明信息熵作为置信度的评判标准是可行的，验证了 CUM-SSEL 算法的可行性。



(a) australian 数据集



(b) wdbc 数据集

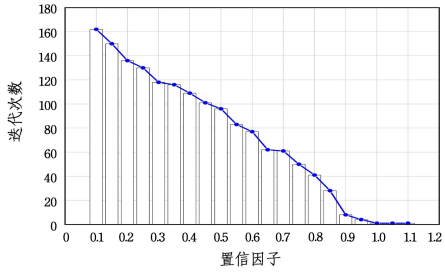
图 3 置信因子对模型精度的影响

Fig. 3 Influence of confidence factor on model accuracy

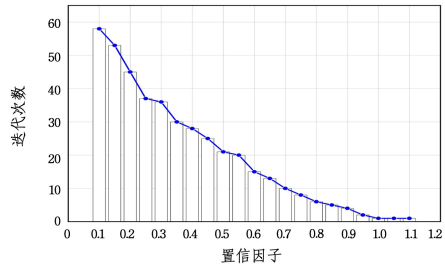
5.3 合理性验证

本实验验证置信因子的设定是否符合理论分析以及集成学习是否能有效地提升模型的性能，进而证实模型的合理性。

设置已标记样本占总样本的 40%，分类器个数为 100。根据理论分析，即由式(8)可知，信息熵的最大值为$\log_2 m$。由图 4 可知，置信因子设置越小，迭代次数就越多，置信因子设置越大，迭代次数就越少，当置信因子达到最大值$\log_2 2=1.0$时，模型只会迭代一轮。这表明，所有的无标记样本在一轮迭代就被全部打上伪标记，并加入到训练集中更新模型，图 3 的结果也反映出，当置信因子到达最大值时，模型精度较低，表明分类过程中出现了较多的错误样本，与理论分析结果相符。



(a) australian 数据集



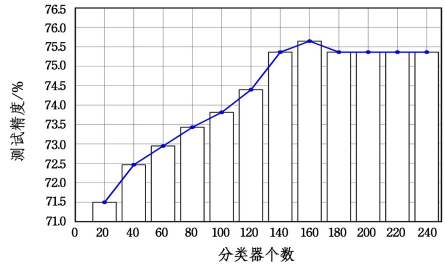
(b) wdbc 数据集

图 4 置信因子对迭代次数的影响

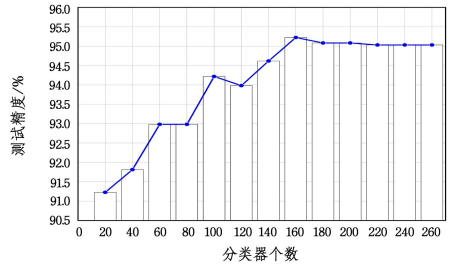
Fig. 4 Influence of confidence factor on iteration number

设置已标记样本占总样本的 40%，置信因子为 0.3。由

5.2 节可知，可以找到一个置信因子使得模型达到一个较高的性能。那么，随着分类器个数的增加，模型的性能能否进一步得到提升，这也是一个非常关键的问题。由图 5 可以直观地看到，随着分类器个数的增加，模型的测试精度基本呈上升趋势，最后趋于收敛状态，验证了 CUM-SSEL 算法的合理性。



(a) australian 数据集



(b) wdbc 数据集

图 5 分类器个数对模型精度的影响

Fig. 5 Influence of the number of classifers on model accuracy

5.4 有效性验证

在本部分实验中，采用测试精度作为评估标准，使用该指标进行消融实验，并且与 3 种半监督学习方法以及 4 种半监督集成算法的性能进行比较，这些算法包括 Self-training[21]，Co-training[22]，Tri-Training[23]，Semi-Boost[12]，Vote-training[25]，Semi-Bagging[19] 以及 CST-Voting[20]。其中，Self-training[21]，Co-training[22]，Tri-Training[23] 属于半监督学习算法；Semi-Boost[12]，Vote-training[25]，Semi-Bagging[19] 以及 CST-Voting[20] 属于半监督集成学习算法。使用的数据集如表 1 所列，每个数据集都划分为有标记训练集、无标记训练集以及测试集。有标记训练数据集占总数据集的 40%，50%，60%，70%。每种算法都是基于有标记和无标记的数据集训练模型，并且在测试集上进行测试，具体的实验结果如表 2—表 5 所列。

表 2 有标记数据占 40%时测试精度的比较

Table 2 Comparison of testing accuracy when the label rate is 40%

数据集	CUM-SSEL	Self-training	Co-training	Tri-Training	Semi-Boost	Vote-training	Semi-Bagging	CST-Voting
australian	**0.7495**	0.6473	0.7149	0.7004	0.7053	0.6859	0.6570	0.7053
wdbc	**0.9467**	0.8889	0.9005	0.9115	0.3684	0.9064	0.8538	0.9183
algerian	**0.7927**	0.6270	0.7704	0.7837	0.7625	0.7837	0.7745	0.7868
ionosphere	**0.8736**	0.8301	0.8641	0.8584	0.6603	0.8301	0.8679	0.8584
biodegradation	**0.8138**	0.7298	0.8075	0.8107	**0.8138**	0.8075	0.8012	0.8075
messidor_features	**0.7167**	0.5346	0.7109	0.7080	0.7080	0.6763	0.6791	0.6994
vertebral	**0.8545**	0.8172	0.8494	0.8408	—	0.8279	0.8301	0.8408
forest	**0.8801**	0.8726	0.8471	0.8484	—	0.8662	0.8675	0.8636

注：—表示算法不能在该数据集上运行，无法得到实验结果。

表 3 有标记数据占 50%时测试精度的比较

Table 3 Comparison of testing accuracy when the label rate is 50%

数据集	CUM-SSEL	Self-training	Co-training	Tri-Training	Semi-Boost	Vote-training	Semi-Bagging	CST-Voting
australian	**0.7369**	0.6667	0.7149	0.7111	0.7149	0.7053	0.6376	0.7101
wdbc	**0.9503**	0.9122	0.9473	0.9421	0.3684	0.9356	0.8947	0.9473
algerian	**0.8196**	0.6284	0.7786	0.8081	0.8108	0.7923	0.7808	0.8060
ionosphere	**0.8831**	0.8396	0.8773	0.8490	0.6742	0.8490	0.7924	0.8679
biodegradation	**0.8328**	0.7456	0.8170	0.802	0.7886	0.8107	0.8138	0.8264
messidor_features	**0.7104**	0.5953	0.7052	**0.7104**	0.7046	0.6763	0.6676	0.6907
vertebral	**0.8439**	0.8279	0.8387	0.8258	—	0.8387	0.8236	0.8215
forest	**0.8738**	0.8407	0.7707	0.8305	—	0.8152	0.8662	0.8356

注：—表示算法不能在该数据集上运行，无法得到实验结果。

表 4 有标记数据占 60%时测试精度的比较

Table 4 Comparison of testing accuracy when the label rate is 60%

数据集	CUM-SSEL	Self-training	Co-training	Tri-Taining	Semi-Boost	Vote-training	Semi-Bagging	CST-Voting
australian	**0.7276**	0.6521	0.6740	0.6724	0.6946	0.7198	0.5990	0.6667
wdbc	**0.9339**	0.9064	0.9239	0.9251	0.3684	0.9298	0.9122	0.9298
algerian	**0.8378**	0.6420	0.7837	0.8243	0.8243	0.7972	0.7103	0.7650
ionosphere	**0.9056**	0.8773	0.8867	0.8962	0.6932	0.8584	0.8018	0.8962
biodegradation	**0.8201**	0.7393	0.8088	0.8120	0.8138	0.8075	0.8107	0.8151
messidor_features	**0.7225**	0.5086	0.6994	0.6892	0.6660	0.6849	0.6965	0.7023
vertebral	**0.8276**	0.7849	0.7741	0.7956	—	0.7849	0.8086	0.7827
forest	**0.8823**	0.8598	0.8662	0.8560	—	0.8280	0.8662	0.8726

注：—表示算法不能在该数据集上运行，无法得到实验结果。

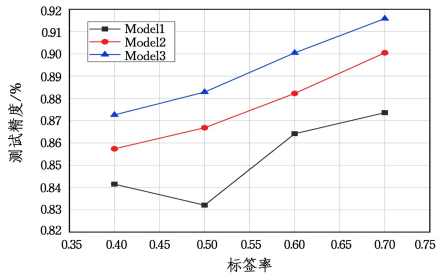
表 5 有标记数据占 70%时测试精度的比较

Table 5 Comparison of testing accuracy when the label rate is 70%

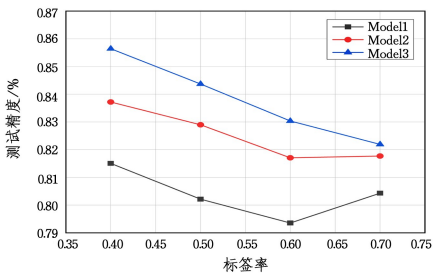
数据集	CUM-SSEL	Self-training	Co-training	Tri-Training	Semi-Boost	Vote-training	Semi-Bagging	CST-Voting
australian	**0.7148**	0.6908	0.6618	0.6676	0.6705	0.6908	0.6473	0.6676
wdbc	**0.9443**	0.9181	0.9176	0.9181	0.3684	0.9203	0.9064	0.9239
algerian	**0.8256**	0.6830	0.7972	0.7824	0.7972	0.7702	0.6986	0.7972
ionosphere	**0.9150**	0.8641	0.9056	0.8962	0.6837	0.8962	0.8584	0.8962
biodegradation	**0.8296**	0.7582	0.8138	0.8044	0.8233	0.8107	0.8264	0.8138
messidor_features	**0.7052**	0.6445	0.6676	0.6861	0.6763	0.6820	0.6763	0.6705
vertebral	**0.8169**	0.7849	0.8064	0.7806	—	0.7956	0.7870	0.7913
forest	**0.8912**	0.8789	0.8853	0.8458	—	0.7898	0.8891	0.8789

注：—表示算法不能在该数据集上运行，无法得到实验结果。

首先，实验选取了 ionosphere 和 vertebral 数据集进行消融实验，从图 6 可以看出，本文的两个改进点对模型性能的提升均发挥了作用。其中 Model1 表示去除缓冲池和信息熵的模型，Model2 表示去除缓冲池的模型，Model3 即为本文模型。在 ionosphere 数据集下，Model2 相对于 Model1 平均约有 2.39%的精度提升，Model3 相对于 Model2 平均约有 1.62%的精度提升。在 vertebal 数据集下，Model2 相对于 Model1 平均约有 2.14%的精度提升，Model3 相对于 Model2 平均约有 1.28%的精度提升。从表 2—表 5 中可以看出，CUM-SSEL 算法在 8 个标准数据集下拥有比其他算法更好的测试精度，从而验证了 CUM-SSEL 算法的有效性。



(a)ionosphere 数据集



(b)vertebal 数据集

图 6 消融实验结果

Fig. 6 Results of ablation experiments

结束语 本文以半监督学习为核心，结合 Bagging 的集成策略，基于信息熵的特性，使样本分类不确定性最小化，提出了一种基于分类不确定性最小化的半监督集成算法。所提算法通过在标准数据集上的实验，验证了它的可行性、合理性和有效性。实验结果表明，信息熵作为置信度的评判标准至关重要，验证了 CUM-SSEL 的可行性；在多个数据集下，随着置信因子的增加，CUM-SSEL 迭代次数逐渐减少；并且随着分类器个数的增加，CUM-SSEL 模型逐渐趋于收敛状态，验证了 CUM-SSEL 的合理性；在与 3 种半监督学习以及 4 种半监督集成学习的性能比较下，CUM-SSEL 具有更高的测试精度，验证了该算法的有效性。本文中还有许多内容需要进一步探讨：

1)分类器的使用。该算法采用的分类器为神经网络，相比其他分类器，神经网络的计算代价更高，花费的时间更长，未来的工作之一就是使用其他的分类器对模型进行优化，进一步减少实验的时间开销，提高模型的运行效率。

2)大数据处理。该算法的运行具有较好的可并行性，有处理大规模数据的潜力，因此未来的工作之一就是将该模型应用在大数据分类问题的处理中。

3)与实际应用结合。半监督学习与集成学习都与现实世界的诸多实际应用有着紧密联系，未来的工作之一会考虑将本文模型用于实际问题的求解中，以检验模型的稳定性和可靠性。

参 考 文 献

[1] MERZ C, CLAIR D S, BOND W. Semi-supervised adaptive resonance theory[C]// Proceedings of IJCNN International Joint Conference on Neural Networks. IEEE, 1992, 3: 851-856.

[2] HADY M, SCHWENKER F. Semi-Supervised Learning [J]. Journal of the Royal Statistical Society, 2006, 172(2): 530.

[3] VAN ENGELEN J, HOOS H H. A survey on semi-supervised learning [J]. MachineLearning, 2020, 109(2): 373-440.

[4] BUHLMANN P, YU B. Analyzing bagging [J]. Annals of Statistics, 2002, 30(4): 927-961.

[5] SCHAPIRE R E. The boosting approach to machine learning: An overview [J]. Lecture Notes in Statistics: Nonlinear Estimation and Classification, 2003, 171: 149-171.

[6] SAGI O, ROKACH L. Ensemble learning: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1249.

[7] BENNETT K P, DEMIRIZ A, MACLIN R. Exploiting unlabeled data in ensemble methods[C]// Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 289-296.

[8] ZHOU Z H. When semi-supervised learning meets ensemble learning[J]. Frontiers of Electrical and Electronic Engineering in China, 2011, 6: 6-16.

[9] DONG X, YU Z, CAO W, et al. A survey on ensemble learning [J]. Frontiers of Computer Science, 2020, 14: 241-258.

[10] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24: 123-140.

[11] SEEDAT N,KANAN C. Towards calibrated and scalable uncertainty representations for neural networks [J]. arXiv: 1911.00104,2019.

[12] 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,2008,31(11): 2000-2014.

[13] LUO Y,ZHU J,LI M,et al. Smooth neighbors on teacher graphs for semi-supervised learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8896-8905.

[14] HOU J,MAO Y,SUN J S. A Semi-supervised Boosting Algorithm for Maximizing Sample Separability [J]. Journal of Nanjing University of Technology,2014,38(5):675-681.

[15] YANG J,ZHANG D,YANG J Y,et al. Globally maximizing,locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(4):650-664.

[16] TANHA J. MSSBoost: A new multiclass boosting to semi-supervised learning [J]. Neurocomputing,2018,314:251-266.

[17] CHEN S,SU S,LI S Z,et al. Cooperative training target tracking algorithm based on online semi-supervised boosting [J]. Journal of Electronics and Information,2014,36(4):888-895.

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

[19] LI Y,SU L,CHEN J,et al. Semi-supervised learning for question classification in CQA [J]. Natural Computing, 2017, 16: 567-577.

[20] LIVIERIS I E,KANAVOS A,TAMPAKAS V,et al. An ensemble SSL algorithm for efficient chest X-ray image classification [J]. Journal of Imaging,2018,4(7):95.

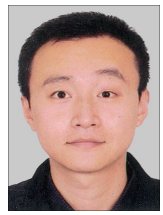
[21] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C]// 33rd Annual Meeting of The Association for Computational Linguistics. 1995:189-196.

[22] BLUM A,MITCHELL T. Combining labeled and unlabeled data with co-training[C]// Proceedings of The Eleventh Annual Conference on Computational Learning Theory. 1998:92-100.

[23] 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.

[24] KENDALL A,GAL Y. What uncertainties do we need in Bayesian deep learning for computer vision? [C]// Proceedings of the 31st Conference on Neural Information Processing Systems. 2017:5580-5590.

[25] GE J,MA T. Semi-supervised learning based on ensemble algorithm[C]// Proceedings of the 29th China Database Academic Conference. 2012:208-213.



HE Yulin, born in 1982, Ph. D, research associate, is a member of China Computer Federation. His main research interests include big data approximate computing technologies, multi-sample statistics theories and methods, and data mining and machine algorithms and their applications.

(责任编辑:喻藜)