Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2022, 11(4), 387-396
Published Online November 2022 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2022.114040

Hans汉斯

基于协同训练半监督学习的干旱灾害天气预测

王鑫炎，段　勇

沈阳工业大学信息科学与工程学院，辽宁　沈阳

收稿日期：2022年10月10日；录用日期：2022年10月31日；发布日期：2022年11月8日

摘　要

近年来，气象研究领域中存在大量高价值信息的无标签数据，然而对这些无标签数据进行高置信度标记是非常困难的，同时这些数据对于建立准确的气象预测模型又是十分重要的。基于此，本文研究了一种基于协同训练的半监督学习方法并用于干旱灾害天气分析预测，该方法使用重复标记策略，根据训练过程中的数据变化动态的计算置信度阈值作为约束条件，并提出了一种增强协同训练方法来评估气象领域中的无标签样本数据的置信度。为了评估所提出方法的性能进行了实验分析，结果表明，该方法的性能优于原始协同训练方法，有效地提高了分类器的分类精度，并验证了该方法用于干旱灾害天气预测的有效性和显著性。

关键词

灾害天气预测，数据分类，半监督学习，协同训练

Drought Disaster Weather Forecast Based on Co-Training Semi-Supervised Learning

Xinyan Wang, Yong Duan

School of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning

Received: Oct. 10th, 2022; accepted: Oct. 31st, 2022; published: Nov. 8th, 2022

Abstract

In recent years, a large amount of unlabeled data with high-value information exists in the field of meteorological research. However, it is very difficult to mark these unlabeled data with high confidence, and these data are very important for establishing an accurate meteorological prediction model. Based on the situation, this paper studies a semi-supervised learning method based on co-training and used for drought disaster weather analysis and prediction. This method uses a

文章引用: 王鑫炎, 段勇. 基于协同训练半监督学习的干旱灾害天气预测[J]. 人工智能与机器人研究, 2022, 11(4): 387-396. DOI: 10.12677/airr.2022.114040

repeated labeling strategy to calculate the confidence threshold dynamically according to the data changes during the training process as a constraint condition, and proposes an enhanced co-training method to evaluate the confidence of unlabeled sample data in the meteorological field. In order to evaluate the performance of the proposed method, experimental analysis is carried out. The results show that the performance of this method is better than that of the original co-training method, which effectively improves the classification accuracy of the classifier, and verifies the effectiveness and significance of this method for drought disaster weather prediction.

Keywords

Disaster Weather Forecast, Data Classification, Semi-Supervised Learning, Co-Training

1. 引言

现如今世界上气象灾害发生十分频繁，气象干旱灾害造成的损失呈直线上升趋势，直接影响着社会和经济的发展。因此，对气象干旱进行有效地预测，对于流域自然资源条件、地区水资源规划管理、缓解旱情的有害影响等具有重要作用。

然而，海量的气象数据中往往会存在部分无标签样本数据，半监督学习算法能够合理地使用无标签样本数据，并提高模型的学习性能。但是如果大量的无标注样本被贴上错误的标签并用作训练，将导致训练集中存在大量的噪声样本，从而严重影响模型的性能。在相关的研究工作中，Mathuranathan Mayuravaani 等人[1]提出一种基于卷积神经网络的半监督学习方法来识别钢表面缺陷，并利用一种基于边缘的方法来确定预测置信度，根据预测的可信度对样本进行加权，使用加权后的预测标签数据训练，有效地提高模型的精度。Karliane Medeiros Ovidio Vale 等人[2]对自训练算法进行改进，在标记数据过程中应用数据分层的方法，使得数据的代表性和类分布在整个标记过程中保持与最初标记的数据集的比例相同，有效地提高了模型的精度。王宇等人[3]依靠多分类器的互相监督和多分类器标签一致的原理，提出一种多分类器协同训练半监督学习方法，从未标记的样本中选择出置信度高的样本并构建增强样本集，能够最大限度地利用未标记信息，并取得了较好的效果。Zhenlei Li 等人[4]为了更好地利用泛化模型的未标记样本，提出了 Aggressive Growing Mixup 算法，使用标签样本和无标签样本混合训练模型，避免了过拟合，提高了模型训练的性能。Siyan Li 等人[5]为了检测故障事件的类型，提出了一种基于伪标签的时间序列数据半监督模型来检测故障的类型，并在故障分类中表现出优越的性能。Ba Hung Ngo 等人[6]将标记的源样本和少量标记的目标样本分成两个子组进行训练，并提出一种新的半监督自适应框架，框架由未标记的目标数据上训练的视图间模型和几个标记的目标数据上训练的视图内模型组成，使用两种模型协作并充分利用了未标记目标数据的信息，具有最先进的分类性能。Leyu Gao 等人[7]使用半监督学习中的典型自训练算法来评估推特数据的可信度，并利用重复标记策略实现了一种改进的自训练算法，该算法具有更好的分类精度。Karliane Medeiros Ovidio Vale 等人[8]对半监督学习范式进行研究，并对协同训练算法进行改进，提出一种置信度值变化方法，有效地提升了训练的性能。Jia Lu 等人[9]在协同训练算法的基础上提出一种基于熵和多准则的协同训练方法，根据熵将数据集划分为两个信息量相同的视图，并分别采用聚类准则和置信准则选择两个视图中未标记数据，算法有效地解决了高置信度准则并不

总是有效的问题，更好地发挥协同训练的互补作用。Karliane Medeiros Ovidio Vale 等人[10]对自训练和协同训练方法进行改进，通过不同的计算置信度方式和每次迭代中选择标签的策略，提出了 FlexCon-G、FlexCon 和 FlexCon-C 三种方法，并且三种方法都优于原始的自训练和协同训练方法。

本文针对气象领域广泛存在的无标签样本数据，利用基于协同训练的半监督学习方法来合理使用无标签样本数据，在原始协同训练方法(OCT, Original Co-Training) [11]的基础上，使用重复标记策略，根据训练过程中的数据变化动态的计算置信度阈值作为约束条件，并使用数据样本增强方法，提出了一种增强协同训练方法(ECT, Enhance Co-Training)来评估气象领域中的无标签样本数据的置信度，不仅可以解决标签样本数据不足的问题，而且可以有效地提高模型的学习性能，分类的精度和效率。

2. 基于特征工程的干旱灾害数据处理

本文的干旱灾害天气原始数据集来自于开发的公共数据平台，其中一共包含了 24 万条数据。原始数据集给出了每条数据的类别，即干旱级别，由于在实际自然现象中干旱等级最高的两类情况出现较少，在数据集中对应的样本也远少于其他等级，因此将原始数据中的这两类样本合并为 1 类。处理后的数据样本为 5 个类，代表的干旱情况级别分别为：None (无干旱)、D0 (异常干燥)、D1 (中度干旱)、D2 (严重干旱)、D3 (极度干旱)，其中每条数据样本都是特定时间点的干旱级别。同时给出了 WS10M 风速 10 米(m/s)、T2M 2 米的温度(℃)、TS 地球皮肤温度(℃)等 18 种能够反映气象要素的属性特征，其中每个属性特征都包含了 90 天的数据。接下来根据本文所研究的问题对干旱灾害数据集如下操作。

2.1. 数据处理

由于原始干旱灾害数据中包含空值数据，因此需要查看数据集数据，删除包含空值的记录，主要是针对包含空值的数据占总体比例较低，删除这些数据对于数据整体影响不大。

原始干旱灾害的数据集中干旱类别特征为非数字类型，而实际上机器学习模型需要的数据是数字型的，只有数字类型才能进行计算。因此，对干旱类别特征使用 Label-encoding 进行编码，Label-encoding 将原始特征值编码为自定义的数字标签，就是用标签进行编码，给特征变量自定义数字标签，进而量化特征。

2.2. 数据特征构造与特征选择

对于原始干旱灾害数据集的每一个属性特征，都由 90 天的气象数据构成，而这些数据无法作为机器学习的训练集直接使用，因此本文根据干旱灾害数据集每个属性特征 90 天的气象数据，分别计算 30 天、60 天、90 天气象数据的均值，并统计其最大值、最小值和中位数，进而得到新的属性特征集合。之后使用 XGBoost 权重和 Spearman 相关系数对新的特征集合进行特征选择，根据 XGBoost 权重选择出大于平均权重的特征，对 Spearman 相关系数绝对值排序并选取前 50%的属性特征，最后将依据 XGBoost 权重和 Spearman 相关系数选择的特征取并集作为最终模型训练所使用的属性特征。

2.3. 基于 SMOTE 与 Tomek Link 相结合算法平衡数据样本

通过对干旱灾害数据分析发现不同的干旱类别数据数量存在严重的类别不均衡。训练数据集样本类别严重不平衡会严重影响模型的预测效果，在训练的过程中预测模型会倾向于识别多数类样本。因此，需要对数据样本类别不平衡的训练数据集进行数据平衡处理。传统的随机欠采样方法删除样本的随机性可能会丢失含有重要信息的样本。随机过采样方法简单地复制少数类的样本可能会导致严重的过拟合，粗暴地合成少数类样本可能还会引入无意义的甚至有害的新样本。因此，本文采用 SMOTE 与 Tomek Link

相结合算法，可将过采样和欠采样技术相结合进行混合重采样，其基本思想是增加样本集中少数类样本的个数，减少样本集中多数类样本的个数，以此来降低不平衡度。

SMOTE 过采样算法[12]首先计算少数类样本 x 到少数类样本集中所有样本的欧氏距离，得到 K 个近邻样本，然后对每一个少数类样本 x，从其 K 个近邻样本中选取一部分样本 x_n，之后根据下式(1)构建新的样本。

$$x_{new} = x + rand(0,1) * |x - x_n| \tag{1}$$

Tomek Link 欠采样算法[13]首先找出一对属于不同类别且距离最近的样本，这对样本互为最近邻，然后将数据集中每对这样的样本进行删除，使得数据集中互为最近邻的样本对都属于同一类别。

使用 SMOTE 过采样算法会产生许多新的样本，在这些样本中会产生重叠的样本对，这时使用 Tomek Link 欠采样算法将这些难以区分的样本对进行清洗，减少数据集中的噪声数据。SMOTE 与 Tomek Link 相结合的算法使得数据集各类别的分布更加均衡。

图 1 显示了采用 SMOTE 与 Tomek Link 相结合的算法进行各类别数据平衡前后的数据分布。对干旱灾害数据集使用 SMOTE 与 Tomek Link 相结合的算法处理之后，数据集样本更加均衡，更加有利于分类器的分类。

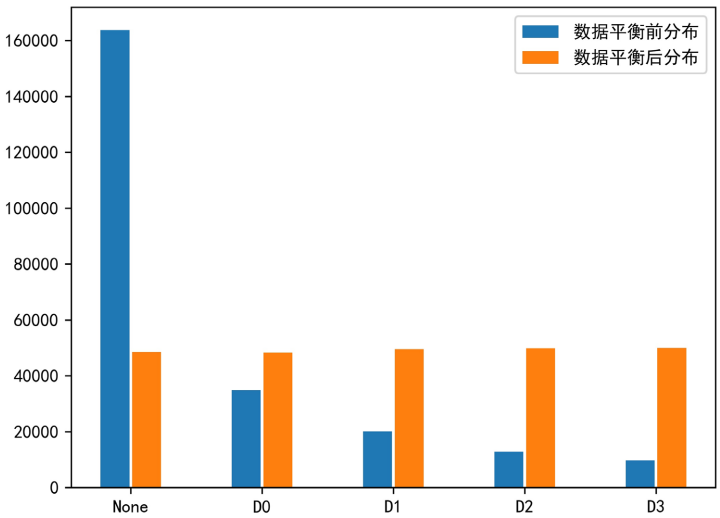

Figure 1. Quantitative distribution of drought disaster data before and after balance
图 1. 干旱灾害各类别数据平衡前后数量分布

3. 增强协同训练方法

在实际应用中，由于干旱等灾害天气领域的数据包含大量高信息价值的无标签数据样本，为了合理地使用这些无标签样本数据，本文对半监督学习中的原始协同训练方法(OCT)进行改进，提出了一种增强协同训练方法(ECT)对无标签数据进行分类标记，并将选择的置信度值高的无标签样本数据及其标签预测值加入到训练集中，以期提高模型训练的性能和精度。

3.1. 原始协同训练方法

协同训练方法是机器学习中半监督学习的主要方法之一，它通过多学习者协作来探索无标记数据中的有效信息。协同训练方法由 Blum 和 Mitchell [11]提出，它通过迭代分类无标签数据集并将高置信度值的预测实例添加到初始标签数据集来扩充标签数据集，协同训练方法构建两个互斥的视图，生成

两个互补分类器，并利用不同的选择策略来选择两个视图上的无标签样本数据。协同训练方法流程如图 2 所示。

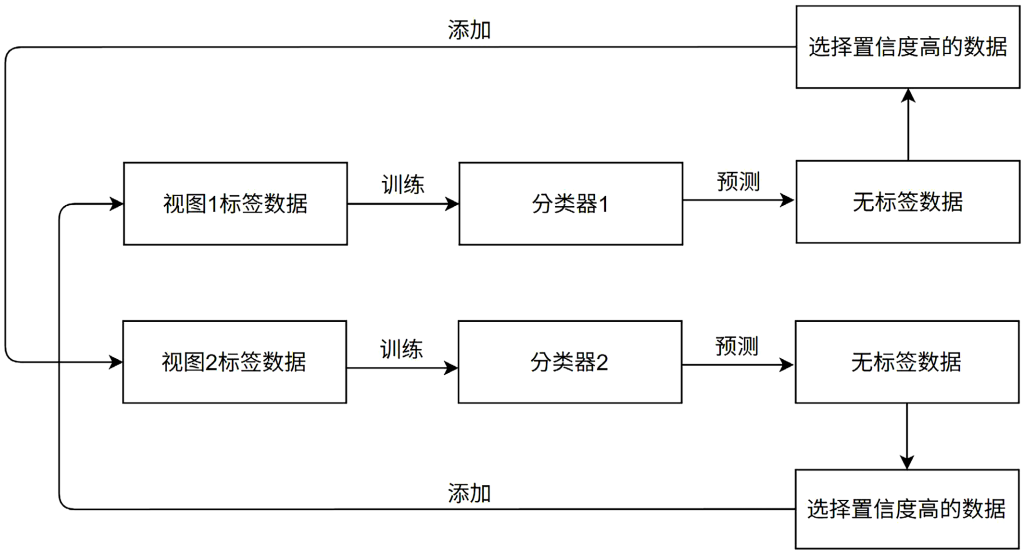

Figure 2. Co-training method process
图 2. 协同训练方法流程

在协同训练方法中，两个分类器起互补作用，第一个分类器的预测用于扩充第二个分类器的标签数据集，第二个分类器的预测用于扩充第一个分类器的标签数据集。该方法首先创建两个标签数据集 L_1 和 L_2，然后基于两个标签数据集训练两个分类器 h_1 和 h_2，使用分类器 h_1 和 h_2 对无标签数据集预测，然后将分类器 h_1 的预测中高置信度的样本实例(P 个正标记和 N 个负标记)添加到标签数据集 L_2 中，同样，将分类器 h_2 的预测中高置信度的样本实例(P 个正标记和 N 个负标记)添加到标签数据集 L_1 中，重复上述过程，直到无标签数据集为空。

算法 1 协同训练半监督学习算法

输入：标签数据集 L，无标签数据集 U
输出：分类器 h_1，分类器 h_2
1. 将标签数据集 L 划分为训练子集 L_1、L_2
2. *repeat*:
3. 分别基于训练子集 L_1、L_2 训练分类器 h_1 和 h_2
4. 分别使用分类器 h_1 和 h_2 对无标签数据集 U 进行预测分类。
5. 将 h_1 分类选择的高置信实例(P 个正标记和 N 个负标记)添加到训练子集 L_2
6. 将 h_2 分类选择的高置信实例(P 个正标记和 N 个负标记)添加到训练子集 L_1，
7. 从无标签数据集 U 中删除被分类器 h_1 和 h_2 选择的实例。
8. *until* $\{U\} = \varnothing$

3.2. 增强协同训练方法

本文对协同训练方法进行改进，使用随机森林(Random Forest)和 K 近邻(K-Nearest Neighbor)算法作为分类器，通过在训练过程中使用重复标记策略，根据训练过程中的数据变化动态的计算置信度阈值，并使用数据样本增强方法，能够有效地限制噪声数据的选择，提高分类器的泛化性能。

3.2.1. 重复标记策略

在常规的协同训练方法中，将选择出具有高置信度的无标签数据及其预测标签值扩充到训练数据集中，然后这些选定的无标签数据将从原始无标签数据集中删除。然而由于初始训练集较小，获得的分类器的分类精度不高，在训练过程中可能存在错误标记。因此，本文使用了重复标记策略，那些被选定的无标签数据不会从原始无标签数据集中删除，而是在每次迭代中重复标记，以确保在后续的迭代中纠正标记错误的样本。

3.2.2. 动态计算置信度阈值方法

协同训练过程中选择高置信度无标签样本时，将预测置信度大于置信度阈值的无标签样本及其对应的预测标签值扩充到训练集中。然而，如果设置的置信度阈值过小，则可能会选择过多错误标记的样本实例，从而为模型引入大量的噪声数据。如果设置的置信度阈值过大，则可能丢失过多正确标记的样本实例，从而降低模型的泛化能力。因此，本文使用了动态计算置信度阈值的方法，根据训练过程中的数据变化，动态的计算置信度阈值，并根据实际情况选择置信度高的无标签样本数据。动态计算置信度阈值主要基于四个方面：上一次迭代的置信度阈值、分类器的训练集精度(分类器使用先前迭代中标记的实例作为训练集，并将最初标记的数据集作为测试集，计算分类器的训练集精度值)、分类器的测试集精度(分类器使用先前迭代中标记的实例作为训练集，并根据测试集计算分类器的测试集精度值)、上一次迭代中标记的实例的百分比，基于以上三个方面，计算其平均值作为当前的置信度阈值。

$$P\left(t_{i+1}\right)=\frac{1}{4}\left(P\left(t_i\right)+\frac{1}{\left|L_t\right|}\sum_{j=1}^{\left|L_t\right|}prec\left(s_j\right)+\frac{1}{\left|L_t\right|}\sum_{k=1}^{\left|L_t\right|}prec\left(s_k\right)+\frac{\left|L_t\right|}{\left|D_u\right|}\right) \tag{2}$$

上式(2)为置信度阈值计算公式，其中 $P\left(t_{i+1}\right)$ 是当前迭代中的置信度值，$P\left(t_i\right)$ 是上一次迭代中的置信度阈值，$\left|L_t\right|$ 是上一次迭代选择的无标签数据样本数，$prec\left(s_j\right)$ 是训练集 s_j 在上一次迭代中的预测值，$prec\left(s_k\right)$ 是测试集 s_k 在上一次迭代中的预测值，$\left|D_u\right|$ 是无标签数据集的实例总数。

3.2.3. 数据样本增强方法

在选择无标签数据样本的过程中本文使用了数据样本增强的方法，在协同训练过程中，分类器 h_1、h_2 分别对无标签数据集进行预测，并根据计算得到的置信度阈值分别选择出高置信度的无标签数据样本，然后选择出分类器 h_1、h_2 同时选中的无标签数据样本，根据这些无标签数据样本的预测标签值进行扩充训练集。如果分类器 h_1、h_2 对当前无标签数据样本预测的预测值相同，则将当前无标签数据样本及其标签预测值加入到训练集 L_1、L_2 中。如果分类器 h_1、h_2 对当前无标签数据样本预测的预测值不同，则根据当前无标签数据样本的预测值计算确定度，确定度值越大，则说明分类器对当前无标签数据样本预测正确的可靠性越高，反之，确定度值越小，则说明分类器对当前无标签数据样本预测正确的可靠性越低。因此，如果根据分类器 h_1 预测值计算的确定度值大于根据分类器 h_2 预测值计算的确定度值，则将当前无标签数据样本及其标签预测值加入到训练集 L_2 中，如果根据分类器 h_2 预测值计算的确定度值大于根据分类器 h_1 预测值计算的确定度值，则将当前无标签数据样本及其标签预测值加入到训练集 L_1 中。使用数据样本增强的方法可以有效提高分类器的泛化能力，提高分类的精度。

$$Cer\left(h_k\right)=P\left(c_i\left|x\right.\right)+\frac{1}{N-1}\sum_{i\neq j}P\left(c_i\mid x\right) \tag{3}$$

上式(3)是分类器对当前无标签数据样本预测值的确定度计算公式，根据软标签预测值进行计算。其中 $Cer\left(h_k\right)$ 是分类器 h_k 的确定度，$P\left(c_i\left|x\right.\right)$ 是当前无标签数据样本 x 被分在 c_i 类的概率，N 表示共有 N 个类别，$\sum_{i\neq j}P\left(c_i\mid x\right)$ 表示当前无标签数据样本 x 被分在其余类别的概率总和。

算法 2 增强协同训练半监督学习算法

输入：标签数据集 L，训练子集 L_1、L_2，无标签数据集 U，置信度值 P_1、P_2
输出：分类器 h_1，分类器 h_2
1. 将标签数据集 L 划分为训练子集 L_1、L_2
2. *repeat*:
3. 分别基于训练子集 L_1、L_2 训练分类器 h_1、h_2
4. 使用分类器 h_1、h_2 对无标签数据集 U 进行预测分类得到数据集 U_1、U_2
5. 根据公式(2)对分类器 h_1、h_2 分别计算置信度阈值 P_1、P_2
6. 对分类器 h_1 预测得到的 U_1 选择出大于置信度阈值 P_1 的数据集 U_1'
7. 对分类器 h_2 预测得到的 U_2 选择出大于置信度阈值 P_2 的数据集 U_2'
8. 选择出 U_1' 和 U_2' 中相同的样本，得到样本集 U'
9. 将分类器 h_1、h_2 对样本集 U' 预测结果相同的数据添加到训练子集 L_1、L_2
10. 将分类器 h_1、h_2 对样本集 U' 预测结果不同的数据根据公式(3)计算确定度 Cer_1、Cer_1
11. 如果 $Cer_1 > Cer_1$，将当前预测的样本实例添加到训练子集 L_2。
12. 如果 $Cer_1 < Cer_1$，将当前预测的样本实例添加到训练子集 L_1。
13. *until iteration* = *MaxIterationor* $\{U\} = \{U'\}$

4. 实验结果与分析

本文对干旱灾害天气数据集进行实验，使用随机森林(Random Forest)和 K 近邻(K-Nearest Neighbor)算法作为分类器，为验证本文提出的增强协同训练方法(ECT)的有效性和可行性，同时使用 UCI 公开的 Nursery 数据集来进行补充实验，并将使用 ECT 获得的结果与原始协同训练方法(OCT)获得的结果进行比较，接下来，本文从性能和统计分析的角度介绍获得的结果。

实验 1 对比了在干旱灾害天气数据集上，使用本文提出的 ECT 方法和传统的 OCT 方法的预测效果。

表 1 展示了在干旱灾害天气数据集上，根据初始标签训练集的百分比：5%、10%、15%、20%、25%和 30%，分别使用 ECT 方法与 OCT 方法进行训练并获得的准确率。

Table 1. Accuracy evaluation using OCT and ECT on training datasets of different scales
表 1. 使用 OCT 和 ECT 方法在不同比例训练数据集上的准确率评估效果

算法	5%	10%	15%	20%	25%	30%
OCT	56.90	67.18	73.24	76.95	79.69	81.88
ECT	60.53	69.06	74.17	77.51	79.98	81.98

通过分析表 1，可以得到，在按照不同比例划分的初始标签训练集的情况下，本文提出的 ECT 方法比 OCT 方法获得了更高的精度。此外，这两种方法的准确率随着初始标签训练集的百分比增加而增加，在标签训练集中使用 5%的实例时，初始标签训练集合通常非常小，OCT 方法无法对无标签数据集进行有效的预测分类，而 ECT 方法相比 OCT 方法有较高的性能提升。当初始标签训练集的百分比从 5%增加到 30%时，由于可以学习的标签数据的信息增多，两种方法在性能上的差异并不大。

本文提出的 ECT 方法中使用了动态变化的置信度阈值，在对无标签数据预测并根据置信度阈值选择时，不会选择置信度低于置信度阈值的无标签数据样本，只会选择可靠性和置信度高的无标签数据样本及其标签预测值，因此，这些被选择的数据样本对分类器的预测有着提升的作用，提高了协同训练预测过程中的分类能力。

图 3 显示了在标记过程中使用 ECT 和 OCT 方法的无标签数据集的百分比。从图中可以看到，ECT 方法选择了初始无标签数据集的 70%到 85%的数据样本，而 OCT 方法选择了整个初始无标签数据集。实验结果表明，选择的无标签数据样本中如果存在低置信度的数据样本可能会降低协同训练方法的性能。因此，ECT 方法能够限制部分无标签数据样本，并选择能够对模型产生积极影响的无标签数据样本，起到提高模型分类精度的作用。

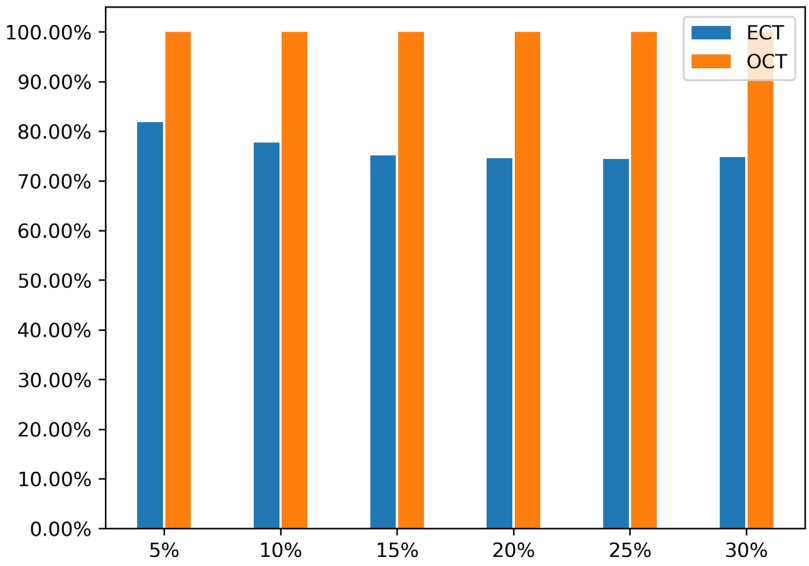

Figure 3. Percentage of instances marked by different methods
图 3. 不同方法标记实例所占百分比

为验证本文提出的 ECT 方法有效性和通用性，实验 2 作为补充实验。实验 2 对比了在 Nursery 数据集上，使用本文提出的 ECT 方法和传统的 OCT 方法的预测效果。Nursery 数据集是托儿所对招生申请进行分类的数据集，包含了 12,960 条样本数据，每条样本数据含有 8 种特征属性，样本数据类别分为 5 类。

表 2 展示了在 Nursery 数据集上，根据初始标签训练集的百分比：5%、10%、15%、20%、25%和 30%，分别使用 ECT 方法与 OCT 方法进行训练并获得的准确率。通过分析表 2，可以得到，在按照不同比例划分的初始标签训练集的情况下，使用本文提出的 ECT 方法比 OCT 方法获得了更高的精度。

Table 2. Accuracy evaluation using OCT and ECT on training datasets of different scales
表 2. 使用 OCT 和 ECT 方法在不同比例训练数据集上的准确率评估效果

算法	5%	10%	15%	20%	25%	30%
OCT	82.90	86.93	90.72	93.26	94.83	95.94
ECT	87.29	88.43	93.75	94.78	96.78	96.94

图 4 显示了 Nursery 数据集在标记过程中使用 ECT 和 OCT 方法选择的无标签数据集的百分比。从图中可以看到，ECT 方法选择了初始无标签数据集的 60%到 90%的数据样本，而 OCT 方法选择了整个初始无标签数据集。本文提出的 ECT 方法相较于 OCT 方法在选择样本时，可以有效地避开噪声数据，选择置信度值更高的数据样本，有效地提高了分类器的分类能力。

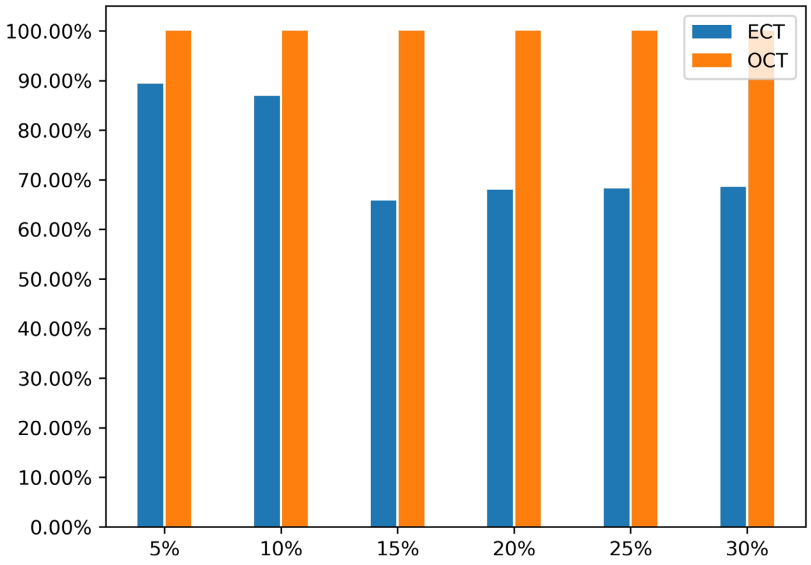

Figure 4. Percentage of instances marked by different methods
图 4. 不同方法标记实例所占百分比

5. 结论

本文提出了一种增强协同训练方法(ECT)。如前所述，原始协同训练方法(OCT)在迭代运行过程中，有一个固定选择样本数的限制，用来定义每次迭代需要选取的无标签样本数据及其预测标签值，但是它在迭代过程中可能选择低置信度的无标签样本数据，这些数据会降低训练方法的性能。本文提出的 ECT 方法对比 OCT 方法进行了改进，在训练过程中使用重复标记策略，根据训练过程中的数据变化动态的计算置信度阈值作为约束条件，并使用数据样本增强方法来选择无标签样本数据及其标签预测值，在训练过程中 ECT 方法可能不会使用无标签数据集的全部数据，这意味着 OCT 方法选择的部分置信度值低的样本数据会对分类器的预测产生消极的影响，而 ECT 只选择置信度值高的样本数据，这对分类器的预测产生积极的作用。因此，增强协同训练方法(ECT)可以有效地提升分类器的分类能力，提高分类器的分类精度。

参考文献

[1] Mayuravaani, M. and Manivannan, S. (2021) A Semi-Supervised Deep Learning Approach for the Classification of Steel Surface Defects. 2021 10*th International Conference on Information and Automation for Sustainability* (*ICIAfS*), Negambo, 11-13 August 2021, 179-184. https://doi.org/10.1109/ICIAfS52090.2021.9606143

[2] Vale, K.M.O., Canuto, A.M.P., Gorgônio, F.L., *et al.* (2019) A Data Stratification Process for Instances Selection in Semi-Supervised Learning. 2019 *International Joint Conference on Neural Networks* (*IJCNN*). Budapest, 14-19 July 2019, 1-8. https://doi.org/10.1109/IJCNN.2019.8851946

[3] 王宇, 李延晖. 一种基于协同训练半监督的分类算法[J]. 华中师范大学学报(自然科学版), 2021, 55(6): 1020-1029.

[4] Li, Z., Hong, Z. and Zheng, H. (2021) Aggressive Growing Mixup: A Faster and Better Semi-Supervised Learning Approach. 2021 *IEEE* 3*rd International Conference on Frontiers Technology of Information and Computer* (*ICFTIC*), Greenville, 12-14 November 2021, 278-284. https://doi.org/10.1109/ICFTIC54370.2021.9647353

[5] Li, S., Zhang, Y., Yang, F., *et al.* (2021) Fault Classification in Transmission Network with Semi-supervised Learning Method. 2021 *IEEE PES Innovative Smart Grid Technologies Europe* (*ISGT Europe*), Espoo, 18-21 October 2021, 1-6. https://doi.org/10.1109/ISGTEurope52324.2021.9640006

[6] Ngo, B.H., Kim, J.H., Chae, Y.J., *et al.* (2021) Multi-View Collaborative Learning for Semi-Supervised Domain Adaptation. *IEEE Access*, **9**, 166488-166501. https://doi.org/10.1109/ACCESS.2021.3136567

[7] Gao, L., Shah, S., Assery, N., *et al*. (2021) Semi-Supervised Self Training to Assess the Credibility of Tweets. 2021 *IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking* (*ISPA/BDCloud/SocialCom/SustainCom*), New York City, 30 September 2021-3 October 2021, 1532-1537. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00206

[8] Vale, K.M.O., Gorgônio, F.L., Araújo, Y.N., *et al*. (2020) A Co-Training-Based Algorithm Using Confidence Values to Select Instances. 2020 *International Joint Conference on Neural Networks* (*IJCNN*), Glasgow, 19-24 July 2020, 1-7. https://doi.org/10.1109/IJCNN48605.2020.9206621

[9] Lu, J. and Gong, Y. (2021) A Co-Training Method Based on Entropy and Multi-Criteria. *Applied Intelligence*, **51**, 3212-3225. https://doi.org/10.1007/s10489-020-02014-6

[10] Vale, K.M.O., Gorgônio, A.C., Flavius Da Luz, E.G., *et al*. (2021) An Efficient Approach to Select Instances in Self-Training and Co-Training Semi-Supervised Methods. *IEEE Access*, **10**, 7254-7276. https://doi.org/10.1109/ACCESS.2021.3138682

[11] Blum, A. and Mitchell, T. (1998) Combining Labeled and Unlabeled Data with Co-Training. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 92-100. https://doi.org/10.1145/279943.279962

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., *et al*. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[13] Tomek, I. (1976) Two Modifications of CNN. *IEEE Transactions on Systems Man & Cybernetics*, **SMC-6**, 769-772. https://doi.org/10.1109/TSMC.1976.4309452