第 45 卷 第 4 期 计 算 机 工 程 2019 年 4 月
Vol. 45 No. 4 Computer Engineering April 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)04-0157-06 文献标志码：A 中图分类号：TP181

噪声可容忍的标记组合半监督学习算法

林金钏，艾浩军

（武汉大学 计算机学院，武汉 430072）

摘 要：针对传统机器学习方法在完成分类任务时多数存在人工标记成本较高、泛化能力较弱的问题，提出一种标记组合半监督学习算法。基于集成学习的思想，利用有标记数据训练多个弱模型并进行组合，增强模型的泛化能力。对无标记数据进行预测，生成有噪声的标记并组合建模。在风险最小化的框架下，使模型收敛达到最优。实验结果表明，在 2 种有监督场景下与现有的支持向量机、分类与回归树、神经网络等算法相比，该算法具有较优的泛化能力。

关键词：半监督学习；集成学习；风险最小化；梯度下降；损失函数

中文引用格式：林金钏，艾浩军. 噪声可容忍的标记组合半监督学习算法[J]. 计算机工程，2019，45(4)：157-162，168.

英文引用格式：LIN Jinchuan，AI Haojun. Noise tolerant label combination semi-supervised learning algorithm[J]. Computer Engineering，2019，45(4)：157-162，168.

Noise Tolerant Label Combination Semi-supervised Learning Algorithm

LIN Jinchuan，AI Haojun

（School of Computer Science，Wuhan University，Wuhan 430072，China）

【Abstract】 Traditional machine learning method always needs high cost manual marking process，and exhibits weak ability of generalization in classification task. In order to solve these problems，a label combination semi-supervised learning algorithm is proposed. Taking advantage of the principle of ensemble learning，the algorithm uses the labeled data to train multiple weak learners，and combine them to enhance the generalization ability. Predict the unlabeled data to generate noise labels，and then combine and model these noise labels to make the model more robust. Under the framework of risk minimization，the model converges to the optimal state. Experimental results show that，compared with some existing learning algorithms like Support Vector Machine(SVM)，Classification and Regression Tree(CART)，Neural Network(NN)，the algorithm has relatively good generalization ability.

【Key words】 semi-supervised learning；ensemble learning；risk minimization；gradient descent；loss function

DOI：10.19678/j.issn.1000-3428.0050398

0 概述

随着计算技术、存储技术的快速发展，计算机采集到的数据越来越多，对这些数据的有效分析、挖掘和应用可极大地促进各领域的发展。机器学习是数据分析、挖掘和应用的重要基础。传统的机器学习主要针对监督学习的问题，即对大量有标记的数据建模，用训练好的模型预测未标记数据。在实际任务中，可以很容易地获得未标记数据，但是对这些数据进行标记需要大量的人力和物力。例如，分析医学影像，可以与医院合作获取大量的影像数据，但是对这些影像中的症状进行标记需要专业医生来完成。如果只对少量的标记数据进行监督学习，所得到的模型泛化能力较弱。半监督学习[1]综合使用标记数据和未标记数据，在一定程度上可以增强模型的泛化能力。

目前，半监督学习方法主要包括基于生成式模型的方法[2]、协同训练方法[3]、半监督 SVM(Support Vector Machine)方法[4]、基于图的方法[5]等。基于生成式模型的方法假设所有数据由相同分布产生，将其转化为参数估计的问题，用最大期望(Expectation Maximization，EM)算法进行计算。协同训练方法针对若干个视图进行相互学习，不断将一个视图内最置信的未标记样本加入到另一个视图的标记样集中，从而实现协同训练。半监督 SVM 方法通过调整 SVM 的超平面和未标记数据的标记指派，在所有训练数据(包括有标记和未标记数据)上最大化间隔。基于图的方法用图表示整个数据集，数据的分布信

基金项目：国家重点研发计划(2016YFB0502201)。

作者简介：林金钏(1992—)，男，硕士研究生，主研方向为迁移学习、复杂网络；艾浩军，副教授。

收稿日期：2018-02-02 **修回日期：**2018-03-08 **E-mail：**linjc0418@gmail.com

息和样本点之间的关系都包含在图结构中,在图上进行标记信息的传递。

本文结合集成学习[6]和噪声感知[7]分类方法,提出一种噪声可容忍的标记组合半监督学习(NtLC-SSL)算法。对无标记数据集进行采样,保证采样数据服从整体样本的统计规律,使领域专家进行标记(与主动学习[8]类似)。使用有标记数据训练若干弱模型,对无标记数据进行预测,生成标记数据集并对其噪声进行组合建模。最后将标记噪声模型嵌入传统的分类模型中进行训练。

1 预备知识

1.1 符号说明

令随机变量 $\boldsymbol{x}\in\mathbb{R}^d$,其中,$d$ 为正整数,表示数据的维度。二分类问题的类标签 $y\in\{+1,-1\}$,多分类问题的类标签 $y\in\{1,2,\cdots,k\}$,k 为正整数,表示类别的个数。$\boldsymbol{x}$ 和 y 的联合概率密度函数为 $p(\boldsymbol{x},y)$。在半监督学习中,数据分为有标记数据(S_L)和无标记数据(S_U),定义如下:

$$S_L=\{(\boldsymbol{x}_i^L,y_i^L)\}_{i=1}^{n_L}\overset{i.i.d}{\sim}p_L(\boldsymbol{x},y) \tag{1}$$

$$S_U=\{\boldsymbol{x}_i^U\}_{i=1}^{n_U}\overset{i.i.d}{\sim}p_U(\boldsymbol{x}) \tag{2}$$

其中,$n=n_L+n_U$,表示数据集的数量,$\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{n}$,表示样本点的集合。

令函数 f 表示需要学习的模型,函数 l 表示损失函数,因此,可以定义 l 的风险函数(记为 *l-risk*)为:

$$R_l(f)=E_p[l(f(\boldsymbol{x}),y)] \tag{3}$$

其中,E 表示随机变量的期望,其下标 p 表示随机变量 $\boldsymbol{x}$ 所服从的概率分布。

1.2 集成学习

在机器学习中,监督学习算法在一个给定的假设空间中搜索具有较好泛化能力的模型,但是,需要假设空间包含泛化能力强的模型,这是机器学习中较难把握的一个问题。集成学习方法组合多个假设空间得到一个较优的模型,即组合多个弱模型(Weak Learner,预测效果一般)得到一个强模型(Strong Learner,泛化性能强)。其组合的弱模型可以是同一类的(同质集成学习),也可以是不同类的(异质集成学习)。

典型的集成学习方法,如 bagging[9]、boosting[10]、随机森林,都属于同质集成学习。异质集成学习方法,如 stacking、blending 等,在国际知名的比赛 KDD 和 Kaggle 中取得了较好的预测结果。为了提高集成学习的泛化性能,需要保证对象之间的多样性,例如,bagging 方法通过样本扰动保证,随机森林通过样本扰动和随机特征的选取来保证。

1.3 可容忍噪声的损失函数

当数据集有噪声时,无法保证训练所得模型的可靠性和可信度。因此,在分类问题中需要标记有噪声场景。记有噪声的数据集为:

$$S_\eta=\{(\boldsymbol{x}_i,\hat{y}_i),i=1,2,\cdots,N\} \tag{4}$$

其中,$\hat{y}_i$ 是带噪声的标记,y_i 为实际标记,因此,可以对噪声进行建模:

$$\hat{y}_i=\begin{cases}y_i,\text{概率为}(1-\eta_i)\\ j,j\in\{1,2,\cdots,k\}\text{且}j\neq y_i,\text{概率为}\ \bar{\eta}_{ij}\end{cases} \tag{5}$$

其中,对于任意的样本点 $\boldsymbol{x}_i$,对应的实际标记为 y_i,有 $\sum_{j\neq i}\bar{\eta}_{ij}=\eta_i$。

定义 1(均匀噪声) 对于任意的样本点 $\boldsymbol{x}_i$,$\eta_i=\eta$ 且 $\bar{\eta}_{ij}=\dfrac{\eta}{k-1},\forall j\neq y_i$,其中,$\eta$ 为常数。

定义 2(不均匀噪声) 对于任意的样本点 $\boldsymbol{x}_i$,η_i 和 $\bar{\eta}_{ij}$ 都是 $\boldsymbol{x}_i$ 的函数。简单不均匀噪声是不均匀噪声的一种特殊情况,即 $\bar{\eta}_{ij}=\dfrac{\eta_i}{k-1},\forall j\neq y_i$。

当数据集标记无噪声时,针对特定的损失函数 l,它的风险函数用式(3)表示,记为 $R_l(f)$,令 f^* 为 $R_l(f)$ 的全局最小值。当数据集有噪声时,其服从概率密度分布 p_η,风险函数用式(6)表示,令 f_η^* 为 $R_l^\eta(f)$ 的全局最小值。

$$R_l^\eta(f)=E_{p_\eta}[l(f(\boldsymbol{x}),\hat{y})] \tag{6}$$

在损失函数 l 确定的情况下,对模型进行风险最小化,则模型对噪声可容忍需满足如下条件[11]:

$$P[f^*(\boldsymbol{x})=y]=P[f_\eta^*(\boldsymbol{x})=y] \tag{7}$$

此时,损失函数 l 对噪声是可容忍的,因此,可以得到 $f^*=f_\eta^*$。

2 半监督学习框架

2.1 数据集构建

监督学习算法的有效性取决于是否能够采集到高质量的有标记数据集,但是,对数据集进行标记需耗费大量人力、物力。为了减少成本,只对部分数据进行标记,并结合剩下的无标记数据共同训练模型。因此,需要有效地选择样本进行标记,以增加模型的泛化性能。本文首先使用聚类的方法将数据分成若干个 cluster,然后进行分层抽样,保证数据的统计规律性。

对于一个分类任务,首先需要明确类别的个数 k。根据任务的需求进行数据采集,采集到的样本数据都是无标记的,样本数量为 n。使用 k-means 算法将无标记数据聚合成 k 个 cluster,第 i 个 cluster 样本集合的数量记为 n_i。针对 k 个 cluster 进行分层抽样,总共抽取 $m(m<<n)$ 个样本,每个 cluster 按照比例进行简

单随机抽样,第 i 个 cluster 抽取的样本数量为 mn_i/n 个。将所有随机抽取的样本标记为 S_L,其余样本标记为 S_U。

对于半监督学习的数据集,需要保证标记数据的有效性。标签数据虽然少,但其在统计上必须是无偏的。

2.2　弱模型训练

在国际著名的数据挖掘竞赛 Kaggle 中,集成的方法取得了很好的精度。可见,集成学习方法能够在一定程度上增强模型的泛化能力。本文采用集成学习的思想,在数据集 S_L 上建立多个弱模型,用各个弱模型对数据集 S_U 进行预测。在训练弱模型时,要保证模型的精度,同时避免过拟合。因此,在一般情况下,要选择复杂度较低的模型,或者使用正则化技术降低模型的复杂度。

本文选用的弱模型采用不同的分类算法,如 SVM、随机森林等,来保证各个算法的假设空间存在差异性,从而增强弱模型的多样性[6],提高模型的泛化能力。需要训练的弱模型数量记为 T,需要训练的弱模型集合记为 $\boldsymbol{B}=\{B_i\}_{i=1}^{T}$。将数据集 S_U 作为测试集,并用 $\boldsymbol{B}$ 中所有的弱模型对测试集进行预测。其中,第 i 个弱模型 B_i 对测试集 S_U 的标记结果记为 $Y^{B_i}=\{y_j^{B_i}\}_{j=1}^{n_U}$。所有弱模型对测试集样本的标记结果记为 Y^B。由于弱模型的算法预测过程是独立的,测试集的样本标记之间也是相互独立的。本文规定 $\boldsymbol{B}_i(\boldsymbol{x})$ 表示第 i 个弱模型对样本 $\boldsymbol{x}$ 的预测标记,同理,$\boldsymbol{B}(\boldsymbol{x})$ 表示所有弱模型对样本 $\boldsymbol{x}$ 的预测标记。

2.3　标记噪声建模

在学习弱模型时,使用少部分的有标记数据进行训练,对大部分无标记数据集 S_U 的预测精确度较低。因此,弱模型集合 $\boldsymbol{B}$ 对 S_U 的预测结果集合 Y^B 存在错误标记,本文定义该错误标记为标记噪声。

根据第 1.3 节的描述,噪声可以分为均匀噪声和不均匀噪声。为简化计算,假设标记集合 Y^B 的标记噪声属于均匀噪声,即对于每一个弱模型 B_i,都对应一个标记错误的概率,记为 η^{B_i}。根据第 2.2 节的算法过程,数据集 S_U 的标记集合 Y^B 关于每一个弱模型的标记集合是相互独立的,即 $Y^{B_i}, i=1,2,\cdots,T$ 之间是相互独立的。

对于二分类问题,真实标记 $y\in\{+1,-1\}$,T 个弱模型所预测的标记为 y^B。为了便于计算,假设每个类别的概率是 0.5,则噪声模型为:

$$p_\eta(y^B,y)=\frac{1}{2}\prod_{i=1}^{T}((1-\eta^{B_i})I_{y=y^{B_i}}+\eta^{B_i}I_{y\neq y^{B_i}}) \tag{8}$$

其中,I 为示性函数,如果允许参数 $\eta\in\mathbb{R}^T$ 为随机变量,则可以认为式(8)表示一族标记噪声的生成模型。

以无标记数据集 S_U 为基础,采用最大似然估计的方法对参数 η 进行计算,可以具体化为求解式(9)所示的最优化问题:

$$\hat{\eta}=\underset{\eta}{\operatorname{argmax}}\sum_{\boldsymbol{x}\in S_U}\ln p_{(y^B,y)\sim p_\eta}(y^B=\boldsymbol{B}(\boldsymbol{x}))=\underset{\eta}{\operatorname{argmax}}\sum_{\boldsymbol{x}\in S_U}\ln\left(\sum_{y'\in\{-1,+1\}}p_\eta(\boldsymbol{B}(\boldsymbol{x}),y')\right) \tag{9}$$

其中,$\boldsymbol{B}(\boldsymbol{x})$ 表示所有弱模型对样本点 $\boldsymbol{x}$ 的预测标签,是一个向量。该最优化问题可以使用基于梯度的方法来求解。

对于多分类问题,真实标记 y 使用 one-hot 方法进行编码。每个类别的概率为 $1/k$,因此,噪声模型如式(10)所示,求解参数 η 的方法保持不变。

$$p_\eta(y^B,y)=\frac{1}{k}\prod_{i=1}^{T}((1-\eta^{B_i})I_{y=y^{B_i}}+\eta^{B_i}I_{y\neq y^{B_i}}) \tag{10}$$

2.4　风险最小化

通过基于梯度的方法算出参数 η 的估计值 $\hat{\eta}$ 后,噪声模型能够较精确地描述数据集 S_U 的标签分布情况。本文利用训练好的噪声模型,结合二分类算法 Logistics 回归(多分类问题使用 Softmax 回归)创建分类模型。

引入损失函数分类可校正(classification calibrated)[12]方法,即一个分类器在特定损失函数下的风险值足够小,同时该分类器在 0-1 损失函数下的风险值也足够小。其与第 1.3 节所述的可容忍噪声损失函数相比本质上是一样的。

二分类 Logistics 回归算法所使用的损失函数称为 Logistics 损失:

$$L_{\text{logistic}}(\boldsymbol{x},y)=\ln(1+\mathrm{e}^{-y\cdot f(\boldsymbol{x})}) \tag{11}$$

其中,$(\boldsymbol{x},y)$ 表示一个采样点的值及其对应的标记,$y\in\{+1,-1\}$,$f(\boldsymbol{x})$ 是关于 $\boldsymbol{x}$ 的线性模型,且 $f:\mathbb{R}^d\to\mathbb{R}$,在一般情况下 $f(\boldsymbol{x})=\omega^{\mathrm{T}}\boldsymbol{x}+b$。Logistics 的风险函数可表示为:

$$R_{\text{logistic}}(\boldsymbol{x},y)=E[L_{\text{logistic}}(\boldsymbol{x},y)] \tag{12}$$

其中,R_{logistic} 是期望风险,用来描述整个数据集上的损失。在实际应用中,不可能采集到所有数据,因此,使用经验风险来代替期望风险,根据大数定理,当样本趋近无穷大时,经验风险趋近于期望风险。因此,为了找到期望风险最小的模型,要在样本充分的情况下,最小化经验风险。考虑到样本数量不充分,可能会导致过拟合现象,因此,在经验风险中加入置信风险:

$$R(\omega)=\frac{1}{n_U}\sum_{x\in S_U}E_{(y^B,y)\sim p_\eta}[\ln(1+\mathrm{e}^{-\omega^{\mathrm{T}}xy})\mid y^B=B(\boldsymbol{x})]+\alpha\parallel\omega\parallel^2 \tag{13}$$

其中，$\parallel\omega\parallel^2$ 为置信风险，可减小模型的复杂度，避免过拟合。参数 α 控制置信风险对整体风险的影响程度。

根据风险最小化的原则[13]，求解式(13)的最小值，计算过程如下：

$$\hat{\omega}=\underset{\omega}{\operatorname{argmin}}\,R(\omega) \tag{14}$$

式(14)只是一个基础的 Logistics 回归的变型，因此，可以使用梯度下降的方法求解参数 ω。

对于多分类的 Softmax 回归，文献[12,14-15]证明在多分类的场景下，Logistics 损失是分类可校正的。因此，Softmax 回归的损失函数能够容忍标签的噪声。Softmax 模型如下：

$$\delta_i(\boldsymbol{x})=\frac{\exp(f(\boldsymbol{x})_i)}{\sum_{j=1}^{k}\exp(f(\boldsymbol{x})_j)},i\in\{1,2,\cdots,k\} \tag{15}$$

本文用 one-hot 方法对多分类问题的标签 y 进行编码，所以 $y\in\mathbb{R}^k$，$f:\mathbb{R}^d\to\mathbb{R}^k$，$f(\boldsymbol{x})_i$ 表示 $f(\boldsymbol{x})$ 的第 i 个元素，$\sum_{i=1}^{k}\delta_i=1$。定义 $\boldsymbol{e}_j\in\mathbb{R}^k$，当样本点 $\boldsymbol{x}$ 的标记 y 为 j 时，如果 $\boldsymbol{e}_j$ 的第 i 个元素等于 j，则 $\boldsymbol{e}_{ji}=1$，否则 $\boldsymbol{e}_{ji}=0$(y 和 $\boldsymbol{e}_j$ 是恒等关系)。因此，有 Softmax 的损失函数：

$$L_{\mathrm{Softmax}}(\boldsymbol{x},y)=l(\boldsymbol{x},\boldsymbol{e}_j)=\sum_{i=1}^{k}\boldsymbol{e}_{ji}\ln\frac{1}{\delta_i(\boldsymbol{x})}=\ln\frac{1}{\delta_j(\boldsymbol{x})}=\ln\frac{1}{\delta(\boldsymbol{x})\cdot y} \tag{16}$$

Softmax 的风险函数 $R(\omega)$ 为：

$$R(\omega)=\frac{1}{n_U}\sum_{x\in S_U}E_{(y^B,y)\sim p_\eta}\left[\ln\left(\frac{1}{\delta(\boldsymbol{x})\cdot y}\right)\mid y^B=B(\boldsymbol{x})\right]+\alpha\parallel\omega\parallel^2 \tag{17}$$

同样地，将式(17)带入式(14)，用梯度下降的方法求解参数 ω。对 Logistics(Softmax)模型训练完成就可对新的样本点进行预测。

2.5 算法步骤和模型复杂度分析

本文 NtLC-SSL 算法过程如下：

算法 NtLC-SSL 算法

输入 经过数据清洗的无标记数据集 S

输出 Logistics 模型(Softmax 模型)

1)使用 k-means 算法对数据集 S 进行聚类，质心的个数为 k。针对 k 个 cluster 进行分层抽样(如第 2.1 节所述)。最终生成标记数据集 S_L 和无标记数据集 S_U。

2)取 S_L 为训练集，单独训练模型集合 B 的每一个弱模型，保证其泛化误差。取 S_U 为测试集，输出标记数据集 Y^B。

3)取数据集 Y^B 作为观测值集合，使用基于梯度的方法训练标记噪声模型，如式(9)所示。

4)将标记噪声模型嵌入 Logistics 模型(Softmax 模型)，通过风险最小化框架进行训练。

模型的复杂度决定其是否能够适用于复杂的数据集。简单的模型使用复杂的数据集进行训练，会出现欠拟合的现象。例如，谷歌提出的 GoogLeNet 模型[16]，具有强大的表达能力，在 ILSVRC2012 比赛中，测试误差为 3.08%。因此，需要通过分析 NtLC-SSL 的复杂度，得出该模型的适用场景。

在第 2.4 节中，将训练好的噪声模型嵌入 Logistics 模型(Softmax 模型)中，然后通过梯度下降技术计算模型的参数，最后输出 Logistics 模型(Softmax 模型)。NtLC-SSL 模型与 Logistics 模型(Softmax 模型)的复杂度是类似的，本质上均为广义线性模型。

3 实验结果与分析

3.1 抽样方法的有效性分析

第 2.1 节所阐述的抽样方法，是为了尽量保证弱模型的泛化性能，抽取的样本要尽可能地符合样本总体的统计意义。本文考虑所抽取数据类条件概率的无偏性，采用样本均值作为指标，判断每个类别所对应的样本均值是否与实际总体样本均值近似，即判断：

$$\frac{1}{n_{\mathrm{sample}_i}}\sum_{j=1}^{n_{\mathrm{sample}_i}}\boldsymbol{x}_j^{\mathrm{sample}_i}\approx\frac{1}{n_{C_i}}\sum_{j=1}^{n_{C_i}}\boldsymbol{x}_j^{C_i},i=1,2,\cdots,k \tag{18}$$

其中，$\boldsymbol{x}^{\mathrm{sample}_i}$ 表示所抽取的第 i 个类别的所有样本点，$\boldsymbol{x}^{C_i}$ 表示总体样本第 i 个类别所有的样本点。计算所有类别样本均值之间的误差之和来衡量数据的无偏性：

$$\varepsilon=\frac{1}{k}\sum_{i=1}^{k}(\parallel\bar{X}_{\mathrm{sample}_i}-\bar{X}_{C_i}\parallel)^2 \tag{19}$$

其中，$\bar{X}_{\mathrm{sample}_i}$ 表示所抽取的属于第 i 个类别的样本均值，$\bar{X}_{C_i}$ 表示总体样本中属于第 i 个类别的样本均值。$\parallel X\parallel$ 表示向量的二范数。

本节实验使用 UCI 机器学习数据仓储中的若干个数据集作为基准数据集，包括 Iris、Wine、Breast Cancer、Segment、Handwritten Digits。表 1 为5 个基准数据集的基本参数。

表 1　5 个 UCI 基准数据集的参数

参数	基准数据集				
	Iris	Wine	Breast Cancer	Segment	Handwritten Digits
样本数量	150	178	569	2 310	5 620
特征数量	4	13	32	19	64
类别数量	3	3	2	7	10

将第 2.1 节描述的抽样方法用在 5 个数据集上,由于数据集带有标记,因此省略专家的标记过程。将本文抽样方法与对整体的简单随机抽样方法进行对比,SRS 表示简单随机抽样,NHS 表示本文 NtLC-SSL 分层抽样法。针对每个数据集进行100 次独立实验,对比结果如表 2 所示。

表 2　2 种方法在不同数据集上的抽样结果对比

基准数据集	抽样方法	抽样误差			
		最小值	最大值	均值	方差
Iris	NHS 方法	0.003 88	0.119 04	0.047 84	0.000 76
	SRS 方法	0.009 41	0.176 23	0.055 25	0.001 06
Wine	NHS 方法	0.142 44	0.690 45	0.301 55	0.007 88
	SRS 方法	0.159 94	0.976 80	0.312 79	0.013 79
Breast Cancer	NHS 方法	0.057 07	0.560 95	0.214 26	0.009 58
	SRS 方法	0.046 43	0.584 60	0.230 06	0.013 64
Segment	NHS 方法	0.018 28	0.094 87	0.045 35	0.000 13
	SRS 方法	0.026 00	0.091 45	0.045 91	0.000 12
Handwritten Digits	NHS 方法	0.331 03	0.740 03	0.507 83	0.006 29
	SRS 方法	0.382 84	0.805 41	0.554 23	0.007 01

从表 2 可以看出,与 SRS 方法相比,NHS 方法在平均水平下的抽样误差更小,在数据集 Iris、Wine、Breast Cancer、Segment、Handwritten Digits 上分别减小了 13.41% 、3.6% 、6.87% 、1.22% 、8.37% ,该方法更能抽取符合总体样本统计规律的数据。

3.2　NtLC-SSL 实验分析

为了分析 NtLC-SSL 算法的有效性,本文采用的数据集是有标记数据集,但在运行 NtLC-SSL 算法时忽略标记。引入 SVM、分类与回归树(Classification and Regression Tree,CART)、神经网络(Neural Network,NN)等部分监督学习方法,以及 SemiBoost[17]、Semi-Bagging[18]等部分半监督集成学习方法进行对比实验。分析在监督学习情况下,NtLC-SSL 算法的可行性和优越性。

3.2.1　生成数据和基准数据集的场景分析

为了直观地反映算法的效果,制定一个二分类的任务,采用文献[19]算法生成一个二维特征的数据集,如图 1 所示。NtLC-SSL 算法得到的结果如图 1中的直线所示。将 NtLC-SSL 算法应用在表 1 所述的基准数据集中,并与常见的监督学习方法进行比较,对比结果如表 3 所示,表中加粗数值为最优值。

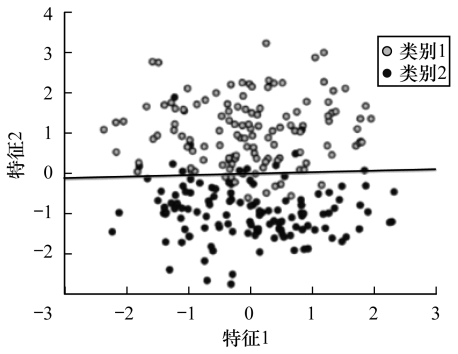

图 1　文献[19]算法生成的二维数据集

表 3　生成数据与基准数据的分类准确率对比

数据集	NtLC-SSL 算法	SVM 算法	CART 算法	NN 算法	Semi Boost 算法	Semi-Bagging 算法
Iris	0.923 3	0.936 7	**0.926 7**	0.916 7	0.908 7	0.857 6
Wine	0.903 0	**0.983 1**	0.949 1	0.981 7	0.893 7	0.812 9
Breast Cancer	0.938 5	0.968 7	0.901 4	**0.982 3**	0.901 2	0.884 1
Segment	0.943 6	0.967 2	0.830 9	**0.973 7**	0.896 1	0.805 4
Handwritten Digits	0.963 0	**0.973 6**	0.688 5	0.947 8	0.913 1	0.867 4
代码生成	0.912 0	0.924 0	**0.968 0**	0.908 0	0.886 3	0.856 5

从表 3 看出,NtLC-SSL 算法的准确率不是最小的,与最优算法相比,差值为 0.01 ~0.09。该算法在某些数据集上的准确率高于 CART 和 NN 算法,差值也只有 0.01 ~0.1。由此,与 SVM、CART 和 NN 这 3 种算法相比,NtLC-SSL 算法的泛化能力在可接受的范围内。与 SemiBoost 和 Semi-Bagging 算法相比,NtLC-SSL 算法的准确率较高,具有更优的泛化能力。

3.2.2　室内定位场景分析

Wi-Fi 指纹室内的定位方法分为 2 个阶段:

1)离线阶段,在预选的参考点(Reference Point,RP)采集 Wi-Fi 接入点(Access Point,AP)的接收信号强度指示(Received Signal Strength Indication,RSSI)值,创建无线图谱(Radio Map,RM)指纹库。

2)在线阶段,用户实时采集 AP 的 RSSI 值,通过模式识别匹配 RM 指纹库,估计用户的位置。每个参考点作为一个类别,在参考点上采集的样本都属于对应参考点的类别。

在离线阶段,需要花费大量的人力和时间来采集原始的 Wi-Fi 信号强度。图 2 所示是某个火车站的一层候车厅,实心点代表参考点,需要志愿者站在该处采集 Wi-Fi 信号并标记样本点,持续时间为 5 min。

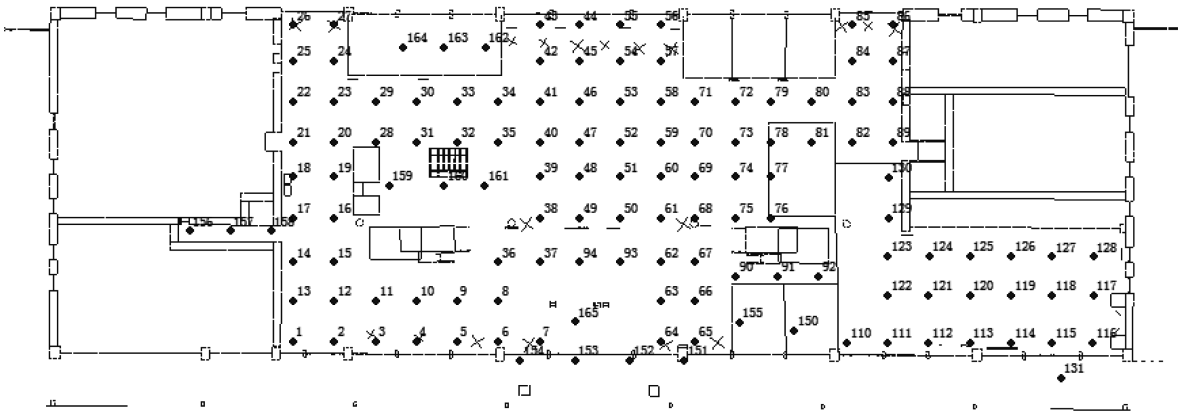

图 2 室内空间参考点示意

指纹库的创建要先规划参考点的分布,才能进行 Wi-Fi 信号采集,即数据集是先有类别,后有样本值。因此,本文 NtLC-SSL 算法创建数据集的方法应用于室内定位时需做适当修改,但其中心思想不变,即保证抽样数据的无偏性。具体地,在带有标记 RP_i ($i=1,2,\cdots,k$)的参考点上进行 Wi-Fi 信号采集,持续时间减小为 10 s,生成有标记数据集 S_L。在所有参考点都采集完成后,设计若干条经过所有参考点的路径,沿着路径匀速行走并采集 Wi-Fi 信号强度,生成无标记数据 S_U。

NtLC-SSL 算法在基于 Wi-Fi 的室内定位数据集上的实验结果如表 4 所示。

表 4 各种算法在室内定位数据集上的分类准确率对比

算法	准确率
NtLC-SSL 算法	0.748 3
SVM 算法	0.807 0
CART 算法	0.529 0
NN 算法	0.907 3
Semi Boost 算法	0.687 5
Semi-Bagging 算法	0.665 3

从表 4 可以看出,NtLC-SSL 算法的准确率分别比 NN 算法和 SVM 算法低了 0.159 0 和 0.058 7,与 CART 算法相比高出 0.219 3,可见,NtLC-SSL 算法的泛化能力相比 CART 算法和 SVM 算法,具有一定的竞争力。与 SemiBoosthe 算法和 Semi-Bagging 算法相比,NtLC-SSL 算法在准确率上有优势。这是由于参考点的数量代表类别的个数(如图 2 所示),因此需要大量的样本和复杂的模型来对其进行建模,使得 NtLC-SSL 算法的复杂度增加。

4 结束语

为完成分类任务,本文提出一种半监督学习的算法 NtLC-SSL。利用集成学习的思想,组合多个弱模型,增强其泛化能力。对弱模型预测的标记噪声进行建模,增强模型的健壮性。实验结果表明,与现有的 SVM、CART、NN 等算法相比,NtLC-SSL 算法具有较好的泛化能力。在某些分类问题中,只需要标记抽取的样本即可取得与其他监督学习算法类似的性能,且成本大幅降低。下一步将改进抽样方法,考虑噪声不均匀的情况,并在风险最小化的框架下改变模型的结构,从而提高算法的精度,拓展其适用范围。

参考文献

[1] 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.

[2] NIGAM K, MCCALLUM A K, THRUN S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 38(2/3): 103-134.

[3] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Conference on Computational Learning Theory. New York, USA: ACM Press, 1998: 92-100.

[4] JOACHIMS T. Transductive inference for text classification using support vector machines[C]//Proceedings of International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1999: 200-209.

[5] ZHU X J, GHAHRAMANI Z, LAFFERTY J D. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 12th International Conference on Machine Learning. Palo Alto, USA: AAAI Press, 2003: 912-919.

[6] ZHOU Z H. Ensemble methods: foundations and algorithms[M]. London, UK: Taylor and Francis Group, 2012.

[7] FRÉNAY B, VERLEYSEN M. Classification in the presence of label noise: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845-869.

[8] 刘康,钱旭,王自强. 主动学习算法综述[J]. 计算机工程与应用,2012,48(34):1-4.

(下转第 168 页)

[11] ZHOU Q Y, PARK J, KOLTUN V. Fast global registration[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 766-782.
[12] 史皓良,吴禄慎,余喆琦,等. 散乱点云数据特征信息提取算法[J]. 计算机工程,2017,43(8):279-283.
[13] 秦绪佳,王建奇,郑红波,等. 三维不变矩特征估计的点云拼接[J]. 机械工程学报,2013,49(1):129-134.
[14] GELFAND N,MITRA N J,GUIBAS L J,et al. Robust global registration[C]//Proceedings of the 3rd Eurographics Symposium on Geometry Processing. Aire-la-Ville, Switzerland:Eurographics Association,2005.
[15] RUSU R B, MARTON Z C, BLODOW N, et al. Persistent point feature histograms for 3D point clouds [EB/OL]. [2017-10-21]. https://ias. in. tum. de/_media/spezial/bib/rusu08ias. pdf.
[16] RUSU R B,BLODOW N,BEETZ M. Fast point feature histograms for 3D registration [C]//Proceedings of IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE Press, 2009: 3212-3217.
[17] 熊风光,蔡晋茹,况立群,等. 三维点云模型中特征点描述子及其匹配算法研究[J]. 小型微型计算机系统,2017,38(3):640-644.
[18] MUJA M,LOWE D. Fast approximate nearest neighbors with automatic algorithm configuration [EB/OL]. [2017-10-21]. http://www. cs. ubc. ca/research/flann/uploads/FLANN/flann_visapp09. pdf.
[19] 王庆璨. 三维点云数据配准算法研究[D]. 西安:西安电子科技大学,2015.
[20] NEWCOMBE R A, IZADI S, HILLIGES O, et al. Kinect fusion: real-time dense surface mapping and tracking [C]//Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality. Washington D. C. ,USA:IEEE Press,2011:127-136.
[21] 刘新. 三维点云数据的配准算法研究[D]. 秦皇岛:燕山大学,2015.
[22] DORAI C,WANG G,JAIN A K,et al. Registration and integration of multiple object views for 3D model construction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(1):83-89.
[23] GAO Y L,FAN J P. Automatic function selection for large scale salient object detection[C]//Proceedings of the 14th ACM International Conference on Multimedia. New York,USA:ACM Press,2006:97-100.
[24] PAULY M, GROSS M, KOBBELT L P. Efficient simplification of point-sampled surfaces [C]//Proceedings of Conference on Visualization. Washington D.C., USA:IEEE Computer Society,2002:163-170.
[25] SERAFIN J, GRISETTI G. NICP: dense normal based point cloud registration[C]//Proceedings of International Conference on Intelligent Robots and Systems. Washington D. C. ,USA:IEEE Press,2015:742-749.
[26] Fast global registration [EB/OL]. [2017-10-21]. https://github. com/IntelVCL/FastGlobalRegistration/tree/master/dataset.

编辑 赵 辉

(上接第 162 页)

[9] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996,24(2):123-140.
[10] SCHAPIRE R E,FREUND Y,BARTLETT P,et al. Boosting the margin:a new explanation for the effectiveness of voting methods [C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers,1997:322-330.
[11] MANWANI N,SASTRY P S. Noise tolerance under risk minimization [J]. IEEE Transactions on Cybernetics, 2013,43(3):1146-1151.
[12] TEWARI A, BARTLETT P L. On the consistency of multiclass classification methods [C]//Proceedings of International Conference on Computational Learning Theory. Berlin,Germany:Springer,2007:143-157.
[13] SHAWE-TAYLOR J,BARTLETT P L,WILLIAMSON R C, et al. Structural risk minimization over data-dependent hierarchies[J]. IEEE Transactions on Information Theory, 1998,44(5):1926-1940.
[14] WESTON J, WATKINS C. Multi-class support vector machines[C]//Proceedings of European Symposia on Artificial Neural Networks. Brussels, Belgium:[s. n.], 1999:83-128.
[15] BARTLETT P L, JORDAN M I, MCAULIFFE J D. Convexity,classification,and risk bounds[J]. Journal of the American Statistical Association, 2006, 101(473): 138-156.
[16] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4,inception-resnet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, USA:AAAI Press,2017:4278-4284.
[17] 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.
[18] LI Y Y,SU L,CHEN J,et al. Semi-supervised question classification based on ensemble learning [C]//Proceedings of International Conference on Swarm Intelligence. Berlin,Germany:Springer,2015:341-348.
[19] GUYON I. Design of experiments for the NIPS 2003 variable selection benchmark[EB/OL]. [2018-01-05]. http://clopinet. com/isabelle/Projects/NIPS2003/Slides/NIPS2003-Datasets. pdf.

编辑 樊丽娜