李松,吴润秀,康平,等.基于自适应剪辑与概率参数的 Tri-training 算法[J].江西师范大学学报(自然科学版),2023,47(5):490-496.
LI Song,WU RunXiu,KANG Ping,et al.The ADP-tri-training:tri-training with adaptive editing and probability parameters[J].Journal of Jiangxi Normal University(Nature Science),2023,47(5):490-496.

文章编号:1000-5862(2023)05-0490-07

基于自适应剪辑与概率参数的 Tri-Training 算法

李 松,吴润秀*,康 平,赵 嘉
(南昌工程学院信息工程学院,江西 南昌 330099)

摘要: 半监督学习利用少量标签数据和大量的无标签数据进行学习.Tri-training 是一种基于分歧的半监督分类算法,在进行伪标记时会因误标记而使训练集产生噪声,从而导致算法分类性能下降.为了减少误标记对算法分类性能的影响,该文提出一种基于自适应剪辑与概率参数的 Tri-training 算法(ADPT).新算法利用基于最近邻的 RemoveOnly 数据剪辑技术对触发自适应剪辑策略的标记数据进行噪声识别及剔除,而未触发自适应剪辑策略的标记数据则用概率参数方法对噪声进行识别及剔除.为验证本文算法的分类性能,采用 4 个评价指标,在 9 组 UCI 数据集上进行实验,并与相关算法进行比较.实验结果表明:该算法在准确率、精度、召回率及 $F_{measure}$ 等评价指标上与其他算法相比,具有明显优势.

关键词: 半监督学习;自适应策略;概率参数;三体训练算法
中图分类号: TP 311 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2023.05.08

0 引言

有监督学习需要大量的标签数据进行学习,以保证训练的模型有足够的泛化能力.在生物等领域中获得标签数据是困难且昂贵的,而获得无标签数据却相对简单很多.因此,能够利用少量标签数据和大量无标签数据进行学习的半监督学习(semi-supervised learning,SSL)[1-2]受到了许多科研工作者的关注.目前,半监督学习方法主要包括基于分歧的方法[3]、判别式方法[4-5]、生成式方法[6-7]和基于图的方法[8-9]等.

基于分歧的方法因模型假设、损失函数非凸性以及数据规模影响较小的优点而受到了学者的广泛青睐,该方法起源于 A. Blum 等[10]的协同训练(Co-training).协同训练利用多视图训练 2 个独立的分类器,然后利用训练的分类器互相标记训练集.协同训练数据集需满足 2 个充分冗余且条件独立的视图,然而在实际生活中,这个要求是很难满足的.因此,S. Goldman 等[11]对协同训练进行了改进,改进的协同训练能够在单视图上学习,不需要满足数据集的约束条件.但是改进的协同训练只能用决策树作为基分类器,为了标记置信度必须采用耗时的 10 折交叉验证.为了有效解决此问题,Zhou Zhihua[12]提出了三体训练算法(Tri-training).该算法使用 3 个基分类器,用少数服从多数的思想标记训练集.然而,Tri-training 在进行伪标记时会产生训练集噪声.如何有效减少 Tri-training 伪标记中的训练集噪声是 Tri-training 的研究热点.张永等[13]提出了一种基于交叉熵的安全 Tri-training 算法,该算法首先利用交叉熵替换分类错误率(交叉熵能更好反映模型的预测分布与真实分布的差距),然后利用凸优化的思想减少伪标记的噪声.莫建文等[14]提出了一种基于梯形网络和改进三体训练法的半监督分类,该算法利用熵值法为每一个标记数据分配权重,通过熵和权重的大小判断标记数据是否错误,当标记的数据熵值越小且权重越大时就越安全.邓超等[15]提出了一种基于 Tri-training 和数据剪辑的半监督聚类算法,该算法通过与数据编辑技术相结合对伪标记中的噪音进行修正,降低了伪标记中的噪声.杨艺

收稿日期:2023-05-11
基金项目:国家自然科学基金(52069014)和江西省教育厅科技计划课题(GJJ180940,GJJ201915)资助项目.
通信作者:吴润秀(1971—),女,江西南丰人,教授,主要从事大数据分析与群智能算法研究.E-mail:wrx@nit.edu.cn

等[16]提出了一种基于 Tri-training 的众包标记噪声纠错算法,该算法先通过噪声过滤算法对标记中的噪声进行识别,然后达到将数据集划分为 2 个部分,一部分为干净实例集,另一部分为噪声实例集,最后通过对噪声实例集重新分类,并纠正噪声实例集中的标记.邓超等[17]提出了基于自适应数据剪辑策略的 Tri-training 算法(ADET),该算法先制定自适应剪辑策略,然后通过最近邻数据剪辑技术对噪声进行剔除,从而提升了算法的性能.

上述 Tri-training 改进算法在一定程度上减少了标记噪声,但对伪标记数据集的噪声仍然没有充分剔除,导致算法的分类性能提升效果不明显.针对此问题,本文提出了一种基于自适应剪辑与概率参数的 Tri-training 算法,首先利用自适应剪辑策略对标记的数据集进行噪声剪辑; 而当有未满足自适应剪辑策略的标记数据集时,采用设置概率参数的方法对噪声进行修剪.

1　相关工作

1.1　Tri-training 算法

Tri-training、算法需要将数据集分为 3 部分: 标签数据集 L、无标签数据集 U、测试数据集 T.算法的基本思想如下: 首先通过对 L 使用 Bootstrap 采样方法获取 3 份有差异的训练集; 然后通过差异的训练集训练 3 个基分类器 h_1、h_2 和 h_3,利用基分类器对无标签的数据集 U 中的未标记数据 x 进行标记.用 H_i 表示除 i 以外的分类器集合(如 H_1 表示 h_2 和 h_3),当 H_i 中分类器标记的结果一致时,将数据和标记的结果保存到第 i 个分类器的训练集中,通过新的训练集更新基分类器.重复这一过程,直到分类器性能没有改变为止.最后,通过多数投票的方法对测试数据集进行输出.

由算法的训练过程可知,算法在对无标记数据添加伪标记时容易产生噪声.D. Angluin 等[18]证明了训练集噪声的可学习性,若满足下列条件训练集噪声是可学习的:

$$m \geqslant \frac{2}{\epsilon^2 (1-2\eta)^2} \ln(2N/\delta), \tag{1}$$

其中 m 是训练样例规模,ϵ 是在最坏情况下的分类错误率,η(<0.5)是分类噪声率的上限,N 是假设的数目,δ 是置信度.

令式(1)中 $c=2\mu\ln(2N/\delta)$,式(1)可转换为

$$m=c/(\epsilon^2 (1-2\eta)^2), \tag{2}$$

其中 μ 是使等式成立的因子.

为简化计算,对式(2)变形,可得

$$u=c/\epsilon^2=m(1-2\eta)^2, \tag{3}$$

其中 c 是常数,得到 $u \propto 1/\epsilon^2$.

η_L 表示 L 上的噪声率,则 L 上的误标记数目为 $\eta_L|L|$,e_1^t 表示 H_1 在第 t 轮分类错误率上限.设 H_1 在第 t 轮标记的数目为 z,H_1 标记错误的数目为 z',则第 t 轮分类错误率 $e_1^t=(z-z')/z$,在第 t 轮中 L^t 的误标记数为 $e_1^t|L^t|$,在第 t 轮的分类噪声率为

$$\eta^t=(\eta_L|L|+e_1^t|L^t|)/|L\cup L^t|. \tag{4}$$

将第 t 轮与第 $t-1$ 轮的训练集与噪声率代入式(3),由 $u \propto 1/\epsilon^2$ 知,当代入的结果满足

$$|L\cup L^t|(1-2(\eta_L|L|+e_i^t|L^t|)/|L\cup L^t|)^2 > |L\cup L^{t-1}|(1-2(\eta_L|L|+e_i^{t-1}|L^{t-1}|/|L\cup L^{t-1}|)^2 \tag{5}$$

时,分类器的性能会得到提升,其中 L^t 和 L^{t-1} 表示第 t 轮与 $t-1$ 轮中 H_i 为 h_i 标记的训练集,η_L 表示数据集 L 上的分类噪声率,e_i^t 和 e_i^{t-1} 分别表示在第 t 轮与 $t-1$ 轮中 H_i 的分类错误率.

1.2　自适应剪辑策略

ADET 的剪辑效果通过定义剪辑召回率 r^t 来判断,r^t 通过对 L^t 中数据的剪辑来刻画.设正确移除误标记数为 p,L^t 中实际误标记数为 q,则 $r^t=p/q$,r^t 反映了 L^t 中误标记移除的比例.

在 ADET 中使用自适应剪辑策略对误标记进行噪声剔除时,分类噪声率 η^t 会被改变,用 η_d^t 表示改变后的分类噪声率,其表达式为

$$\eta_d^t=(\eta_L|L|+(1-r^t)e_1^t|L^t|)/\left|L\cup L_d^t\right|, \tag{6}$$

其中 L_d^t 表示 L^t 剪辑后的数据.

在 PAC 可学习理论下,文献[17]证明了如下定理,并通过定理制定了自适应剪辑策略.

定理 1　当 Tri-training 相邻两轮迭代满足 $0<e_1^t/e_1^{t-1}<|L^{t-1}|/|L^t|<1$ 时,则 $\epsilon^t<\epsilon^{t-1}$,即 h_1 分类器的性能会提升.

定理 2　Tri-training 在第 t 轮迭代中,对 L^t 进行剪辑得到 L_d^t,若 $L^t>L_d^t$,且 $(|L^t|-|L_d^t|)/(2e_1^t|L^t|) \leqslant r^t$,则 $\epsilon_d{}^t<\epsilon^t$($\epsilon_d{}^t$ 表示第 t 轮剪辑后的分类错误率),即剪辑后的 h_1 分类器性能比剪辑前的更好.

定理 3　Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若 $t-1$ 轮未进行剪辑,当满足 $0<|L^{t-1}|/|L^t|<1$,$|L^t|>|L_d^t|>|L^{t-1}|$,$0<(1-r^t)e_1^t/e_1^{t-1} \leqslant |L^{t-1}|/|L^t|<1$ 时,用训练集 $L\cup L_d^t$ 和 $L\cup L^{t-1}$ 对 h_1

训练所产生的分类错误率满足 $\epsilon_d{}^t<\epsilon^{t-1}$.

定理 4 Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若 $t-1$ 轮已经进行剪辑,当 $0<|L_d^{t-1}|<|L^t|$,且 $1>|L^{t-1}|/|L^t|\geqslant e_1^t/((1-r^{t-1})e^{t-1})>0$ 成立时,用训练集 $L\cup L^t$ 和 $L\cup L_d^{t-1}$ 对 h_1 训练所产生的分类错误率满足 $\epsilon^t>\epsilon_d^{t-1}$.

定理 5 Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若 $t-1$ 轮已经进行剪辑,当满足 $0<|L_d^{t-1}|/|L^t|<1$,$0<|L_d^{t-1}|<|L_d^t|$,$1>|L^{t-1}|/|L^t|\geqslant(1-r^t)e_1^t/((1-r^{t-1})e_1^{t-1})>0$ 时,用训练集 $L\cup L_d^t$ 与 $L\cup L_d^{t-1}$ 对 h_1 训练所产生的分类错误率满足 $\epsilon_d{}^t<\epsilon_d^{t-1}$.

根据上述定理,制定了如下的自适应剪辑策略:

1) Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若第 $t-1$ 轮剪辑没有触发,且 $|L^t|>|L^{t-1}|$,当定理 1 和定理 2 的充分条件都能成立时,则在第 t 轮时剪辑被触发.

2) Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若第 $t-1$ 轮剪辑没有触发,且 $|L^t|>|L^{t-1}|$,当定理 1 和定理 3 的充分条件都能成立时,则在第 t 轮时剪辑被触发.

3) Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若第 $t-1$ 轮剪辑已被触发,且 $|L^t|>|L_d^{t-1}|$,当定理 2 和定理 4 的充分条件都能成立时,则在第 t 轮时剪辑被触发.

4) Tri-training 在第 $t-1$ 和 t 的相邻两轮迭代中,若第 $t-1$ 轮剪辑已被触发,且 $|L^t|>|L_d^{t-1}|$,当定理 4 的充分条件无法成立且定理 5 的充分条件成立时,则在第 t 轮时剪辑被触发.

5) 除上面 4 种情况外,剪辑都被抑制.

2 基于自适应剪辑与概率参数的 Tri-training算法

2.1 算法原理

ADET 在对第 t 轮的伪标记数据 L^t 进行噪声去除时采用自适应剪辑策略来判断能否对 L^t 中的噪声剔除,当满足自适应剪辑策略定理时,采用基于最近邻的 RemoveOnly 数据剪辑技术对标记数据进行剪辑.然而,在实际运行中发现,ADET 的定理 2~定理 4 的充分条件很难满足,而未满足条件的 L^t 也存在噪声数据,当使用噪声数据对分类器进行训练时会造成算法的性能下降.因此,本文在 ADET 的基础上采用设置概率参数的方法对未进行噪声剔除的 L^t 进行降噪声处理.通过计算未进行噪声剔除的 L^t 预测概率,然后设置概率参数 θ 值对 L^t 进行筛选.当 L^t 的预测概率小于设定的 θ 时被认为是噪声数据,对数据进行删除,反之则认为是安全的数据,对数据进行保留.

2.2 概率参数

基分类器在对无标签的数据集 U 中的未标记数据 x 进行标记时,计算标记数据的预测概率.文献[9]计算标记概率更新标记结果权重,通过权重值降低误标记数据对算法性能的影响.本文将通过改变标记数据的概率参数 θ 值大小来直接筛选误标记数据.当标记结果的概率大于某一概率 θ 值时认为标记结果是正确的,保留标记数据;当小于 θ 值时,则认为标记的数据为误标记数据,将数据进行剔除.

2.3 算法流程

基于自适应剪辑与概率参数 Tri-training 算法的伪代码如图 1 所示.算法的输入:无标记数据集 U,有标记数据集 L,测试集 T.算法的输出:测试集 T 中数据的分类结果 $h(x)$.

在伪代码中,行 1~行 5 首先使用 bootstrap 抽样方法训练出有差异的基分类器,然后初始化分类错误率,剪辑召回率 .行 9~行 13 计算基分类器的分类误差和召回率,以及对无标签数据进行伪标记.行 15~行 17 对满足自适应剪辑策略的伪标记数据进行噪声剔除.行 19~行 22 利用使用自适应剪辑策略去除噪声的数据集训练基分类器.行 24~行 29 先对未满足自适应剪辑策略的数据集采用概率参数的方法去除伪标记中的噪声,然后用去除噪声的训练集训练基分类器.行 32 对测试数据集进行预测输出.

3 实验结果与分析

3.1 实验数据集

为对 ADPT 的有效性进行验证,本文采用 9 个 UCI[20] 机器学习库中的数据集进行测试,数据信息如表 1 所示,每个数据只有正类、负类 2 个类别.为了符合实验需要,将数据分为有标签训练集 L、无标

签训练集 U 和测试数据集 T.其中 L 占数据集总数的 20%,余下数据的 80%为无标签数据集 U 和 20%为测试数据集 T.

1) for $i\in\{1..3\}$ do
2) $S_i\leftarrow$BootstrapSample(L)
3) $h_i\leftarrow$Learn(S_i)
4) $e'_i\leftarrow0.5$; $r_i\leftarrow0$; $de'_i\leftarrow$False
5) end for
6) repeat until none of $h_i(i\in\{1..3\})$ changes for $i\in\{1..3\}$ do
7) $L_i\leftarrow\varnothing$; $L_{di}\leftarrow\varnothing$; $de_i\leftarrow$False; $M_i\leftarrow\varnothing$; update$_i\leftarrow$False
8) $e_i\leftarrow$MeasureErrors(H_i)
9) $r_i\leftarrow$MeasureRecall(H_i)
10) for every $x\in U$ do then
11) if $h_j(x)=h_k(x)$ $(j,k\neq i)$
12) $L_i\leftarrow L_i\cup\{x,h_j(x)\}$
13) end for
14) if L_i meet the adaptive strategy
15) $L_{di}\leftarrow$remove(L_i)
16) $de_i\leftarrow$True
17) end for
18) for $i\in\{1..3\}$ do
19) if $de_i=$True
20) $h_i\leftarrow$Learn($L_i\cup L_{di}$)
21) $e'_i\leftarrow e_i$; $r'_i\leftarrow r_i$;
22) else if update$_i=$True then
23) for every $x\in L_i$ then
24) *if* $\omega_x>\theta$(ω_x 表示 x 的预测概率)
25) $M_i\leftarrow x$
26) end for
27) $h_i\leftarrow$Learn($L\cup M_i$)
28) $e'_i\leftarrow e_i$;
29) end for
30) end repeat
31) Output: $h(x)\leftarrow\arg\max\sum_{h_i(x)=y}1$

图 1　算法的伪代码图

3.2　评价指标

为更好验证 ADPT 的性能,本文采用准确率(A_{ccuracy})、精度(P_{recision})、召回率(R_{ecall})及 F_{measure}值 4 个评价指标对算法进行测试,表 2 是与评价指标相关的混淆矩阵.

表 1　UCI 数据集

数据集	样本数/条	属性数/个	类比例%	
			正类	负类
wdbc	569	30	37.3	62.7
winewhite	4 898	11	33.5	66.5
haberman	306	3	28.4	71.6
spect	267	44	20.6	79.4
electrical	10 000	13	36.2	63.8
Australian	690	14	44.5	55.5
liverdisorder	345	6	42.0	58.0
heart	270	13	44.4	55.6
geraman	1 000	42	30.0	70.0

表 2　混淆矩阵

实际的类	预测的类	
	yes	no
yes	T_P	F_N
no	F_P	T_N

在混淆矩阵中 T_P、T_N 分别表示分类正确的正类与负类,F_P、F_N 表示分类错误的正类和负类.在评价指标中准确率用于统计分类器对元组正确识别的比率;精度计算预测为正类元组中实际为正类元组所占比率;召回率统计正元组中预测为正元组的比率;F_{measure}为精度和召回率的调和均值.这几个指标的数值越接近 1 表明算法性能越好.性能指标的计算公式如下:

$$A_{\text{ccuracy}}=(T_P+T_N)/(P+N),$$

$$P_{\text{recision}}=T_P/(T_P+F_P),$$

$$R_{\text{ecall}}=T_P/(T_P+F_N),$$

$$F_{\text{measure}}=(2P_{\text{recision}}R_{\text{eacall}})/(P_{\text{recision}}+R_{\text{ecall}}).$$

3.3　实验结果分析

为验证 ADPT 的分类性能,将 ADPT 与 Tri-training、文献[13]提出的基于交叉熵的 Tri-training 算法(TCE)、安全的 Tri-training(ST)、基于交叉熵的安全 Tri-training 算法(STCE)和基于自适应数据剪辑的 Tri-training 算法(ADET)[17]进行比较,实验的对比结果如表 3~表 6 所示,表中最后一列为 ADPT 取得最佳结果的 θ 值,每一份数据集在 6 种算法中

的最佳结果用“* ”标出.从表 3~表 6 可以看出:在测试的 9 组数据集中,ADPT 在准确率、精度、召回率、$F_{measure}$等 4 个评价指标占优的数据集分别为 7、7、5、6;在准确率和 $F_{measure}$评价指标上,haberman 数据集的提升效果明显,在精度评价指标上,liverdisorders 和 geraman 数据集的提升效果明显;在召回率评价指标上,spect 和 heart 数据集的提升效果明显.

表 3 不同算法在各数据集上的准确率

数据集	Tri-training	TCE	ST	STCE	ADET	ADPT	θ 值
wdbc	0.937 1	0.951 0	0.944 1	0.958 0	0.957 4	0.971 3*	0.70
winewhite	0.706 9	0.706 1	0.706 9	0.703 7	0.768 6	0.781 2*	0.75
haberman	0.576 9	0.551 3	0.692 3	0.538 5	0.731 1	0.773 2*	0.65
spect	0.746 3*	0.641 8	0.567 2	0.626 9	0.641 5	0.661 5	0.80
electrical	0.995 2	0.994 4	0.994 0	0.996 0	0.999 3	0.999 9*	0.90
Australian	0.803 5	0.820 8	0.826 6	0.849 7	0.861 3	0.867 2*	0.85
liverdisorders	0.540 2	0.563 2	0.551 7	0.597 7	0.601 4	0.627 5*	0.60
heart	0.691 2	0.720 6	0.764 7	0.779 4	0.763 0	0.811 1*	0.90
geraman	0.660 0	0.732 0	0.740 0	0.746 0*	0.729 0	0.734 0	0.85

表 4 不同算法在各数据集上的精度

数据集	Tri-training	TCE	ST	STCE	ADET	ADPT	θ 值
wdbc	0.893 3	0.920 0	0.920 0	0.920 0	0.933 3	0.962 5*	0.65
winewhite	0.506 6	0.554 4	0.615 4	0.565 0	0.700 7	0.728 5*	0.70
haberman	0.250 0	0.312 5	0.312 5	0.375 0	0.538 2	0.556 6*	0.80
spect	0.769 2	0.923 1	0.769 2	1.000 0*	0.611 3	0.661 5	0.70
electrical	0.992 4	0.990 2	0.991 3	0.995 6	0.998 6	1.000 0*	0.80
Australian	0.807 7	0.846 2	0.807 7	0.923 1*	0.854 3	0.878 4	0.80
liverdisorders	0.276 6	0.255 3	0.319 1	0.468 1	0.545 5	0.653 5*	0.60
heart	0.724 1	0.758 6	0.758 6	0.827 6	0.824 8	0.855 0*	0.60
geraman	0.632 4	0.558 8	0.617 6	0.573 5	0.638 1	0.794 8*	0.65

表 5 不同算法在各数据集上的召回率

数据集	Tri-training	TCE	ST	STCE	ADET	ADPT	θ 值
wdbc	0.985 3	0.985 7	0.971 8	1.000 0*	0.949 9	0.959 1	0.75
winewhite	0.524 7	0.521 2	0.520 2	0.517 0	0.532 2	0.571 2*	0.85
haberman	0.160 0	0.172 4	0.277 8*	0.187 5	0.228 2	0.236 7	0.55
spect	0.416 7	0.342 9	0.277 8	0.342 1	0.468 8	0.660 1*	0.90
electrical	0.993 5	0.994 5	0.992 3	0.993 5	0.999 4	0.999 9*	0.85
Australian	0.768 3	0.776 5	0.807 7	0.782 6	0.836 1	0.860 2*	0.70
liverdisorders	0.684 2	0.800 0	0.681 8	0.687 5*	0.362 9	0.512 1	0.70
heart	0.617 6	0.647 1	0.709 7	0.705 9	0.642 0	0.772 6*	0.85
geraman	0.417 5	0.506 7	0.518 5	0.565 2*	0.176 9	0.284 6	0.95

表 6　不同算法在各数据集上的 $F_{measure}$

数据集	Tri-training	TCE	ST	STCE	ADET	ADPT	θ 值
wdbc	0.937 1	0.951 7	0.945 2	0.958 3*	0.941 2	0.956 0	0.70
winewhite	0.515 5	0.537 3	0.563 8	0.539 9	0.603 7	0.630 4*	0.75
haberman	0.195 1	0.222 2	0.294 1	0.250 0	0.307 0	0.498 5*	0.70
spect	0.540 5	0.500 0	0.408 2	0.509 8	0.489 9	0.609 8*	0.90
electrical	0.993 4	0.992 3	0.991 8	0.994 5	0.999 0	0.999 8*	0.90
Australian	0.787 5	0.809 8	0.807 7	0.847 1	0.842 7	0.856 6*	0.70
liverdisorders	0.397 9	0.387 1	0.434 8	0.557 0*	0.396 2	0.491 9	0.80
heart	0.666 7	0.698 4	0.733 3	0.761 9	0.715 5	0.776 2*	0.90
geraman	0.502 9	0.531 5	0.563 8	0.569 3*	0.264 7	0.381 4	0.95

为分析 6 种算法的综合性能,本文引入 Friedman 检验,对准确率、精度、召回率、$F_{measure}$ 等 4 个评价指标进行秩均值检验. Friedman 检验是一种显著性差异检验方法,其秩均值体现了算法的综合性能,秩均值越大算法综合性能越好. Friedman 检验的秩均值如表 7 所示.

表 7　秩均值

评价指标	Tri-training	TCE	ST	STCE	ADET	ADPT
准确率	2.280 0	2.560 0	2.830 0	3.560 0	4.11	5.67*
精度	2.000 0	2.330 0	2.780 0	4.110 0	4.33	5.44*
召回率	2.720 0	3.560 0	3.330 0	3.610 0	3.22	4.56*
$F_{measure}$	2.110 0	2.560 0	3.110 0	4.670 0	3.22	5.33*
均值	2.277 5	2.752 5	3.012 5	3.987 5	3.72	5.25*

从表 7 中可以看出: ADPT 在 4 个评价指标上的秩均值均是最高的,在准确率上获得了最高值 5.67,在召回率上取得了最低值 4.56,但也明显优于其他 5 种算法. 从表 7 的最后一行可以看出: ADPT 的均值 Tri-training 的均值的 2 倍多,比 STCE 提高了 1.262 5.

4　结语

Tri-training 算法虽然对半监督学习性能有一定提高,但是 Tri-training 算法的学习过程会因误标记而产生训练噪声. 为有效降低误标记对 Tri-training 算法的影响,本文提出了一种基于自适应剪辑策略与概率参数的 Tri-training 算法(ADPT) . ADPT 利用自适应剪辑策略先对标记的训练集进行剪辑,当标记的训练集未触发自适应剪辑策略时,利用概率参数对标记的训练集进行噪声识别和剔除. 实验结果表明 ADPT 具有较好的分类性能.

5　参考文献

[1] XU Mengfan, LI Xinghua, LIU Hai, et al. An intrusion detection scheme based on semi-supervised learning and information gain ratio [J]. Journal of computer research and development, 2017, 54(10): 2255-2267.

[2] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning [M]. Cambridge: MIT Press, 2006.

[3] 周志华. 基于分歧的半监督学习 [J]. 自动化学报, 2013, 39(11): 1871-1878.

[4] FISHER R A. The use of multiple measurements in taxonomic problems [J]. Annals of Eugenics, 1936, 7(2): 179-188.

[5] BAUDAT G, ANOUAR F. Generalized discriminant analysis using a kernel approach [J]. Neural Computation, 2000, 12(10): 2385-2404.

[6] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Procee-

dings of the IEEE,1989,77(2):257-286.

[7] SHAHSHAHANI B M,LANDGREBE D A.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.

[8] WANG Fei,ZHANG Changshui.Label propagation through linear neighborhoods [J].IEEE Transactions on Knowledge and Data Engineering,2008,20(1):55-67.

[9] BREVE F,ZHAO Liang,QUILES M,et al.Particle competition and cooperation in networks for semi-supervised learning [J].IEEE Transactions on Knowledge and Data Engineering,2011,24(9):1686-1698.

[10] BLUM A,MITCHELL T.Combining labeled and unlabeled data with co-training [EB/OL].[2022-02-06].https://is.muni.cz/el/1433/jaro2010/PV056/um/12319818/Blum-Mitchell-Cotraining.pdf.

[11] GOLDMAN S,ZHOU Yan.Enhancing supervised learning with unlabeled data [EB/OL].[2022-03-16].http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1B220CE1696AD240EF611FDAF54AC93F?doi=10.1.1.33.2574&rep=rep1&type=pdf.

[12] ZHOU Zhihua,LI Ming.Tri-training:exploiting unlabeled data using three classifiers [J].IEEE Transactions on knowledge and Data Engineering,2005,17(11):1529-1541.

[13] 张永,陈蓉蓉,张晶.基于交叉熵的安全 Tri-training 算法 [J].计算机研究与发展,2021,58(1):60-69

[14] 莫建文,贾鹏.基于梯形网络和改进三训练法的半监督分类 [EB/OL].[2022-03-19].https://doi.org/10.16383/j.aas.c190869.

[15] 邓超,郭茂祖.基于 Tri-Training 和数据剪辑的半监督聚类算法 [J].软件学报,2008,19(3):663-673.

[16] 杨艺,蒋良孝,李超群,等.一种基于 Tri-training 的众包标记噪声纠正算法 [J].电子学报,2021,49(3):424-434.

[17] 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,2007,30(8):1213-1226.

[18] ANGLUIN D,LAIRD P.Learning from noisy examples [J].Machine Learning,1988,2(4):343-370.

[19] 李敦明.基于半监督学习策略的网络异常检测方法研究 [D].上海:华东师范大学,2019.

[20] BACHE K,Lichman M.UCI machine learning repository [EB/OL].[2022-6-30].https://www.researchgate.net/publication/272825857_UCI_Machine_Learning_Repository.

The ADP-Tri-Training: Tri-Training with Adaptive Editing and Probability Parameters

LI Song,WU Runxiu*,KANG Ping,ZHAO Jia

(School of Information Engineering,Nanchang Institute of Technology,Nanchang Jiangxi 330099,China)

Abstract: Semi-supervised learning utilizes a small amount of labeled data and a large amount of unlabeled data for learning.Tri-training is a divergence-based semi-supervised classification algorithm.When pseudo-labeling,Tri-training will cause noise in the training set due to mislabeling,which will reduce the classification performance of the algorithm.In order to reduce the impact of mislabeling on the classification performance of the algorithm,the ADP-Tri-training that is tri-training with adaptive editing and probability parameters (ADPT) is proposed.Firstly the new algorithm uses the nearest neighbor-based RemoveOnly data editing technology to identify and eliminate the noise of the marked data that triggers the adaptive editing strategy,while the marked data that does not trigger the adaptive editing strategy uses the probability parameter method to identify and eliminate noise.In order to verify the classification performance of the algorithm in this paper,four evaluation indicators are used to conduct experiments on 9 groups of UCI datasets,and compare with related algorithms.The experimental results show that the algorithm in this paper has obvious advantages compared with other algorithms in terms of accuracy,precision,recall and $F_{measure}$ indicators.

Key words: semi-supervised learning; adaptive strategy; probability parameter; Tri-training

(责任编辑:冉小晓)