DOI:10.7544/issn1000-1239.2021.20190838
58(1):60-69,2021

基于交叉熵的安全 Tri-training 算法

张 永 陈蓉蓉 张 晶
(辽宁师范大学计算机与信息技术学院 辽宁大连 116081)
(zhyong@lnnu.edu.cn)

Safe Tri-training Algorithm Based on Cross Entropy

Zhang Yong, Chen Rongrong, and Zhang Jing
(*School of Computer & Information Technology, Liaoning Normal University, Dalian, Liaoning* 116081)

Abstract Semi-supervised learning methods improve learning performance with a small amount of labeled data and a large amount of unlabeled data. Tri-training algorithm is a classic semi-supervised learning method based on divergence, which does not need redundant views of datasets and has no specific requirements for basic classifiers. Therefore, it has become the most commonly used technology in semi-supervised learning methods based on divergence. However, Tri-training algorithm may produce the problem of label noise in the learning process, which leads to a bad impact on the final model. In order to reduce the prediction bias of the noise in Tri-training algorithm on the unlabeled data and learn a better semi-supervised classification model, cross entropy is used to replace the error rate to better reflect the gap between the predicted results and the real distribution of the model, and the convex optimization method is combined to reduce the label noise and ensure the effect of the model. On this basis, we propose a Tri-training algorithm based on cross entropy, a safe Tri-training algorithm and a safe Tri-training learning algorithm based on cross entropy, respectively. The validity of the proposed method is verified on the benchmark dataset such as UCI (University of California Irvine) machine learning repository and the performance of the method is further verified from a statistical point of view using a significance test. The experimental results show that the proposed semi-supervised learning method is superior to the traditional Tri-training algorithm in classification performance, and the safe Tri-training algorithm based on cross entropy has higher classification performance and generalization ability.

Key words semi-supervised; Tri-training algorithm; cross entropy; convex optimization; sample label

摘 要 半监督学习方法通过少量标记数据和大量未标记数据来提升学习性能.Tri-training 是一种经典的基于分歧的半监督学习方法,但在学习过程中可能产生标记噪声问题.为了减少 Tri-training 中的标记噪声对未标记数据的预测偏差,学习到更好的半监督分类模型,用交叉熵代替错误率以更好地反映模型预估结果和真实分布之间的差距,并结合凸优化方法来达到降低标记噪声的目的,保证模型效果.在此基础上,分别提出了一种基于交叉熵的 Tri-training 算法、一个安全的 Tri-training 算法,以及一种基于交叉熵的安全 Tri-training 算法.在 UCI(University of California Irvine)机器学习库等基准数据集上

收稿日期:2019-12-08;**修回日期**:2020-04-03
基金项目:国家自然科学基金项目(61772252,61902165);辽宁省高等学校创新人才支持计划项目(LR2017044);辽宁省自然科学基金项目(2019-MS-216)
This work was supported by the National Natural Science Foundation of China (61772252, 61902165), the Program for Liaoning Innovative Talents in Universities (LR2017044), and the Natural Science Foundation of Liaoning Province (2019-MS-216).

验证了所提方法的有效性,并利用显著性检验从统计学的角度进一步验证了方法的性能.实验结果表明,提出的半监督学习方法在分类性能方面优于传统的 Tri-training 算法,其中基于交叉熵的安全 Tri-training 算法拥有更高的分类性能和泛化能力.

关键词　半监督学习;Tri-training 算法;交叉熵;凸优化;样本标记

中图法分类号　TP181

传统的分类方法通常使用有标签的数据进行训练.然而,随着人们收集数据能力的不断提升,获得大量的未标记数据样本相对容易,而获取已标记数据样本通常却需要付出昂贵的人力、物力和财力.如何让学习器利用少量的标记数据和大量的未标记数据来提升学习性能,是半监督学习(semi-supervised learning, SSL)[1-2]所要解决的问题.目前常用的半监督学习方法主要包括生成式方法[3]、半监督支持向量机(semi-supervised support vector machine, S3VM)[4-5]、图半监督学习[6]、基于分歧的方法[7]和半监督聚类[8]等.

其中,基于分歧的方法起源于协同训练[9].标准的协同训练算法需要 2 个足够多的冗余视图,即属性可以自然地分为 2 组,且每组在给定类标签的情况下有条件地独立于另一组.然而在实际应用中,并不是所有的数据集都能满足这一特征.为了更加方便地应用到各种常见的数据挖掘场景,Zhou 等人[10]提出了一种协同训练算法 Tri-training,该算法不需要足够的冗余视图,也不需要使用不同的监督学习方法,有效提高了半监督学习的效率.Søgaard[11]提出了一种带分歧的 Tri-training 算法,旨在提升模型的薄弱点,比一般的 Tri-training 算法更具有效率.Saito 等人[12]提出了一种用于无监督域自适应的非对称 Tri-training 算法,通过不对称地使用 3 个神经网络,达到对目标域未标记数据进行标记的目的,从而提高了迁移学习领域的自适应性能.Ruder 等人[13]为了减少 Tri-training 过程中的时间和空间复杂度,把迁移学习思想引入半监督学习,提出了多任务 Tri-training 算法.Ou 等人[14]提出了一种基于正则化局部嵌入的 Tri-training 方法用于高光谱图像分析,解决了奇异值和过拟合的问题,有效提高了高光谱图像的分类精度.Park 等人[15]结合 Tri-training 方法和对抗学习方法构建了一个半监督学习框架,用来检测单个音频剪辑中多个声音事件的设置和偏移.该框架先用 Tri-training 方法标记数据,然后使用对抗学习方法来减少真实数据集和合成数据集之间的域间隙.然而上述 Tri-training 算法依赖于初始的分类器,忽略了标记后产生的噪声标签.研究表明[16],在某些情况下,使用未标记的数据进行半监督学习时会使性能退化,导致一个不安全的学习模型,而在 Tri-training 训练过程中产生的不正确伪标签是性能下降的主要根源.本文提出一种安全的半监督学习方法,即使在使用未标记数据时也不会显著降低学习性能.

本文从降低误标记样本的角度出发,将交叉熵、凸优化分别与 Tri-training 算法相结合,提出了 3 个不同的算法.实验结果表明,基于交叉熵的安全 Tri-training 算法具有最优的分类性能.针对半监督分类学习任务,本文的主要贡献有 3 个方面:

1) 将交叉熵与 Tri-training 算法相结合,提出一种基于交叉熵的 Tri-training 算法;

2) 使用铰链损失函数,提出了一个安全的 Tri-training 算法;

3) 运用凸优化方法,提出一种基于交叉熵的安全 Tri-training 学习框架.

1　相关工作

本节主要介绍了基于分歧的半监督学习方法 Tri-training 的基本思想、相对熵和交叉熵的相关知识,以及基于凸优化的半监督分类学习方法.

1.1　Tri-training 算法

Tri-training 算法[10]不要求数据集有充足和冗余的视图,通过建立 3 个基分类器,用一种监督学习方法即可实现.基本思想是:首先,通过 Bootstrap 方法重采样原始标记数据集 L,训练得到 3 个基分类器 h_1,h_2,h_3.然后,对未标记数据集 U 中的数据 x 进行标记,只需有 2 个分类器对 x 的标记一致即可,并将该未标记样本及其分类一致的标签加入到另外一个分类器的训练集中,重新训练基分类器,依此类推.重复迭代上述过程,直到 3 个基分类器的性能不再改变为止.最后,采用多数投票法确定最终的分类结果.

在标记过程中,产生的错误标记样本会对结果

造成影响.为了减少标记过程中产生的噪声样本,该算法基于 Angluin 等人[17]的理论结果,根据分类噪声率和每轮训练的分类错误率决定伪标记样本是否用于分类器更新.Zhou 等人[10]证明,如果新标记的训练样本足够多且满足式(1)所设定的约束条件,则基分类器重新训练所得的分类性能会迭代提升.

$$|L\cup L^t|\left(1-2\frac{\eta_L|L|+e_i^t|L^t|}{|L\cup L^t|}\right)^2>|L\cup L^{t-1}|\left(1-2\frac{\eta_L|L|+e_i^{t-1}|L^{t-1}|}{|L\cup L^{t-1}|}\right)^2, \quad (1)$$

其中,L^t 表示第 t 次迭代 h_j 和 h_k 为 h_i 新标记的训练集($j,k\neq i$),e_i^t 表示第 t 次迭代 h_j 和 h_k 的错误率上限,η_L 表示初始训练集 L 的噪声率.通常情况下 η_L 很小,一般假定 $0\leqslant e_i^t\leqslant 0.5$,则当 $|L|>|L^{t-1}|$ 时,由式(1)可推出 $e_i^t|L^t|<e_i^{t-1}|L^{t-1}|$,即有:

$$0<\frac{e_i^t}{e_i^{t-1}}<\frac{|L^{t-1}|}{|L^t|}<1. \quad (2)$$

在 Tri-training 训练过程中,用式(2)来判断由 h_j 和 h_k 给出一致标记的数据所组成的数据集 L^t 是否应该被加入到 h_i 的新训练集中.

Tri-training 不需要数据集有冗余的视图,对基学习器也没有特定的要求,因此成为基于分歧的半监督学习方法中最常用的技术.然而该算法在训练过程中产生的标记噪声,可能会对最后的模型产生不好的影响.为了减少 Tri-training 算法过程中的标记噪声对未标记数据的预测偏差,学习到更好的半监督分类模型,本文结合交叉熵和凸优化方法,提出了一种基于交叉熵的安全 Tri-training 学习框架.

1.2 相对熵与交叉熵

相对熵又称为 KL 散度(Kullback-Leible divergence),是 2 个概率分布间差异的非对称性度量[18].在信息理论中,相对熵可用于表示 2 个概率分布的信息熵的差值.假设在给定样本集 χ 上定义样本真实分布为 P,模型预测分布为 Q,则 P 和 Q 的相对熵或 KL 散度可以定义为[19]

$$D_{\mathrm{KL}}(P\|Q)=\sum_{x\in\chi}P(x)\lg\frac{P(x)}{Q(x)}, \quad (3)$$

D_{KL} 值越小,表示 P 和 Q 的分布越接近.当且仅当 2 个概率分布相同时,相对熵 $D_{\mathrm{KL}}=0$.

交叉主要用于描述两个事件之间的相互关系.Rubinstein[20]提出了交叉熵的概念,用来度量 2 个概率分布的差异性.由式(3)变形可得:

$$\sum_{x\in\chi}P(x)\lg Q(x)=-H(P(x))+\left[-\sum_{x\in\chi}P(x)\lg Q(x)\right]. \quad (4)$$

式(4)等号右端前半部分为分布 P 的熵,后半部分即为交叉熵:

$$H(P,Q)=-\sum_{x\in\chi}P(x)\lg Q(x). \quad (5)$$

交叉熵具有 2 个重要性质:

1) 非对称性.即 $H(P,Q)\neq H(Q,P)$.

2) 非负性.$H(P,Q)\geqslant 0$,且有 $H(P,P)=H(P)$.

在机器学习中,为了评估“学习模型的预测分布 Q”与“训练数据真实分布 P”之间的差距,通常采用 KL 散度来度量,即 $D_{\mathrm{KL}}(P\|Q)$.由式(4)可知,等号右端前半部分 $-H(P)$ 保持不变,因此在优化过程中,只需要考虑交叉熵 $H(P,Q)$.

交叉熵已经广泛应用于组合优化、机器学习等领域.Bosman 等人[21]提出了一种基于交叉熵和平方误差损失函数的梯度随机采样可视化方法,实验结果表明交叉熵损失比二次损失具有更强的梯度和可搜索性.Li 等人[22]在交叉熵损失中加入正则项,提出了一个对偶交叉熵损失函数,用来优化神经网络结构,实验结果表明提出的损失函数提升了分类性能.Lu 等人[23]提出了一种动态加权交叉熵作为语义分割的损失函数,设计了一种加权方法对交叉熵进行加权,并在每一个训练步骤中迭代权重.实验结果表明该方法能有效地提高数据极不平衡情况下的分割精度.Lopez-Garcia 等人[24]将交叉熵和遗传算法相结合,提出了一种模糊规则系统层次结构要素的优化方法,能更好地预测短期交通拥堵状况.交叉熵可直接使用作为损失函数评估模型,当交叉熵最低时,可以认为得到了一个最好的训练模型.

1.3 基于凸优化的半监督学习

在半监督学习中,仅有极少数的标记数据,大部分是未标记数据.因此,人们通常希望利用未标记数据来提高模型的学习性能.但在某些情况下,现有的半监督学习方法比仅使用标记数据的监督学习方法表现更差.为了解决这一问题,Li 等人[25]提出了一种安全的半监督 SVM(safe semi-supervised support vector machine, S4VM)方法,该方法提高了半监督 SVM 的性能.Krijthe 等人[26]提出了一种隐式约束的最小二乘半监督学习方法,该方法隐式地考虑所有可能的未标记数据的标记,同时能最大限度地减少已标记数据的平方损失.Guo 等人[27]提出了一种方案,通过整合几个弱监督的分类器来建立最终的

预测结果,并在噪声学习、域自适应和半监督回归任务上验证了方案的有效性.

2　基于交叉熵的安全 Tri-training 算法

为了提高未标记数据对于半监督分类性能的影响,本文使用凸优化方法,通过半监督分类的凸线性组合给出标记,提高安全半监督分类的准确性.本节首先利用交叉熵在信息差异度量方面的优势,提出了一种基于交叉熵的 Tri-training 算法;其次,提出了一种安全的 Tri-training 学习算法,以提升未标记数据的分类性能;最后,给出了一个基于交叉熵的安全 Tri-training 学习框架.

2.1　基于交叉熵的 Tri-training 算法

机器学习方法期望在训练数据模型上学习到的分布 Q 与真实数据分布 P 越接近越好,然而真实分布是不可知的.假定训练数据是从真实数据中独立同分布采样而得到的,则期望学习到的模型分布 Q 与训练数据分布 P' 近似相同.最小化 Q 和 P' 之间的差异,即最小化 $D_{\mathrm{KL}}(P'\|Q)$.训练数据的分布是不变的,因此,最小化 $D_{\mathrm{KL}}(P'\|Q)$ 等价于求 $H(P',Q)$,也就是 2 个分布之间的交叉熵.交叉熵值越低,表明学习到的模型分布越接近于训练数据分布.

在计算"学习到的模型分布"与"训练数据分布"之间的信息差异量时,交叉熵相比利用误分率来衡量该差异性有更多的优势.本文首先将交叉熵与 Tri-training 算法相结合,提出一种基于交叉熵的 Tri-training 算法,具体描述如算法 1 所示.

算法 1. 基于交叉熵的 Tri-training 算法.

输入:训练集 $D=L\cup U$,其中 L 和 U 分别为已标记数据集和未标记数据集,测试集 T;

输出:测试数据集 T 中数据 x 的标签 $h(x)$.

① for $i=1$ to 3 do
② $S_i\leftarrow Bootstrap(L)$, $h_i\leftarrow Learn(S_i)$;
③ 初始化参数 $e'_i\leftarrow 0.5$ and $|L'_i|\leftarrow 0$;
④ end for
⑤ repeat
⑥ for $i=1$ to 3 do
⑦ $L_i\leftarrow\varnothing$;
⑧ $e_i\leftarrow(H_j+H_k)/2$; /* $H_j,H_k(j,k\neq i)$为基分类器的交叉熵 */
⑨ if $e_i<e'_i$ then
⑩ for U 中每个样本 x do
⑪ if $h_j(x)=h_k(x)$ $(j,k\neq i)$ then
⑫ $L_i\leftarrow L_i\cup\{(x,h_j(x))\}$;
⑬ end if
⑭ end for
⑮ end if
⑯ if $e_i|L_i|>e'_i|L'_i|$ then
⑰ 更新 L_i; /* 根据约束条件式(2),从 L_i 中随机删除若干个数据 */
⑱ end if
⑲ end for
⑳ for $i=1$ to 3 do /* 更新 */
㉑ $S_i\leftarrow L\cup L_i$, $h_i\leftarrow Learn(S_i)$;
㉒ $e'_i\leftarrow e_i$, $|L'_i|\leftarrow|L_i|$;
㉓ end for
㉔ until h_i 的性能没有改变
㉕ $h(x)\leftarrow \mathrm{sgn}\left(\sum_{i=1}^{3}w_ih_i(x)\Big/\sum_{i=1}^{3}w_i\right)$; /* 计算测试集 T 中数据 x 的标签,w_i 为权重 */
㉖ 输出所有的 $h(x)$.

在算法 1 中,行①～④用 Bootstrap 方法从已标记数据集 L 中重采样 3 次,训练得到 3 个基分类器 $h_i(i=1,2,3)$,并初始化其误差参数 e'_i 和无标记学习数据规模 $|L'_i|$.行⑧利用交叉熵估算 h_i 的分类误差 e_i:基分类器 $h_j,h_k(j,k\neq i)$ 分别对标记数据集 L 进行预测,得到预测分布,结合已标记数据集的真实分布,利用式(5)分别计算 h_j 和 h_k 的交叉熵 H_j 和 H_k,得到误差 $e_i=(H_j+H_k)/2$.行⑩～⑭利用任意 2 个基分类器对未标记数据进行分类,如果标记一致,则将该数据及其标签加入到第 3 个基分类器的训练集中.行⑳～㉓利用新生成的训练集对 3 个基分类器重新训练.重复该过程,直到基分类器的性能不再改变为止.最后,根据加权投票规则计算测试数据集的分类标签,其中权重 w_i 表示 3 个基分类器在初始已标记数据集 L 上的分类准确率.

相对于 Tri-training 算法,算法 1 用交叉熵代替了 Tri-training 算法中的分类误差,用加权投票规则确定最终的分类标签,在一定程度上提升了未标记数据的分类性能.

2.2　一个安全的 Tri-training 学习算法

对于半监督分类任务,假设对于未标记的样本得出的 n 个预测结果为 $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$, $\boldsymbol{y}_i\in H^u$, $i=1,2,\cdots,n$,其中 n 为基分类器个数,u 为未标记样本的个数,且在二分类任务中 $H=\{-1,+1\}$.同时,用 $\boldsymbol{y}_0\in H^u$ 表示在使用已标记数据训练的监督模型上所得的分类结果.半监督分类任务的目标是

得到一个安全的预测结果 $\boldsymbol{y}=g(\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\},\boldsymbol{y}_0)$,且结果并不比 $\boldsymbol{y}_0$ 差.用 $\boldsymbol{y}^*$ 表示未标记样本的真实标签,而在实际中 $\boldsymbol{y}^*$ 是未知的,因此将 $\boldsymbol{y}^*$ 看作是基分类器的凸组合 $\boldsymbol{y}^*=\sum_{i=1}^{n}w_i\boldsymbol{y}_i$,并实现以下优化目标[27]:

$$\max_{\boldsymbol{y}\in H^u}\min_{\boldsymbol{w}\in M}\left(\ell\left(\boldsymbol{y}_0,\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)-\ell\left(\boldsymbol{y},\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)\right),\quad(6)$$

其中,$\ell(\cdot,\cdot)$是可用于分类任务的损失函数,$\boldsymbol{w}=(w_1,w_2,\cdots,w_n)$,$w_i$ 为基分类器的非负权重且 $\sum_{i=1}^{n}w_i=1$,M 是反映基分类器重要先验知识的凸组合集.

损失函数是用来估量模型的预测值与真实值的不一致程度.常见的用于分类任务的损失函数包括均方损失、铰链损失、交叉熵损失等.损失函数越小,模型的准确性越好.铰链损失和交叉熵损失都可以转化为凸优化问题,并通过凸优化技巧实现优化,损失函数在分类任务中的优化具有一般普适性[28-29].

铰链损失可以表示为预测结果的线性关系,使得铰链损失作为分类任务中的损失函数具有更好的性能[29].文献[27]证明当 $\ell(\cdot,\cdot)$选用铰链损失时,式(6)在分类任务中可转换为一个凸优化问题.

对于一个分类问题,用 $\boldsymbol{p}=(p_1,p_2,\cdots,p_u)$表示 u 个样本的真实标签,$\boldsymbol{q}=(q_1,q_2,\cdots,q_u)$表示预测标签,铰链损失函数可定义为

$$\ell(\boldsymbol{p},\boldsymbol{q})=\frac{1}{u}\sum_{i=1}^{u}\max\{1-p_iq_i,0\}.\quad(7)$$

当采用铰链损失时,式(6)的最优解 $\hat{\boldsymbol{y}}$ 和 $\hat{\boldsymbol{w}}$ 应该满足:

$$\hat{\boldsymbol{y}}=\operatorname{sgn}\left(\sum_{i=1}^{n}\hat{w}_i\boldsymbol{y}_i\right).\quad(8)$$

根据铰链损失函数的性质,当 $\sum_{i=1}^{n}w_i=1$ 时,有:

$$\ell\left(\operatorname{sgn}\left(\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right),\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)=1-\frac{1}{u}\left\|\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right\|_1.$$

从而式(6)可改写为

$$\min_{\boldsymbol{w}\in M}\ell\left(\boldsymbol{y}_0,\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)-\ell\left(\operatorname{sgn}\left(\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right),\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)=\min_{\boldsymbol{w}\in M}\ell\left(\boldsymbol{y}_0,\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right)+\frac{1}{u}\left\|\sum_{i=1}^{n}w_i\boldsymbol{y}_i\right\|_1-1,\quad(9)$$

其中,M 可表示为$\{\boldsymbol{w}\mid\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_j\boldsymbol{w}\geqslant\boldsymbol{1}\delta,\ \sum_{i=1}^{n}w_i=1,\ w_i\geqslant 0\}$,$\delta$ 为 0~1 之间的一个常量,即为式(9)的约束条件.易知,式(9)是一个凸优化问题,通过对其求解,可安全地利用未标记数据来提高半监督分类算法的鲁棒性.

借助上述思想,我们将安全性融入到 Tri-training 算法中,以此来提高半监督学习算法的准确率和鲁棒性.算法 2 描述了用于半监督分类时安全的 Tri-training 学习算法.

算法 2. 安全的 Tri-training 学习算法.

输入:训练集 $D=L\cup U$,其中 L 和 U 分别为已标记数据集和未标记数据集,测试集 T;

输出:测试数据集 T 中数据 x 的标签 $h(x)$.

① $h_0\leftarrow Learn(L)$;

/* 用 L 训练基准分类器 h_0 */

② 调用 $Tri\text{-}training(L,U)$训练基分类器 h_i $(i=1,2,3)$;

③ $\boldsymbol{y}_0\leftarrow h_0(T)$,$\boldsymbol{y}_i\leftarrow h_i(T)$ $(i=1,2,3)$;

/* 利用 h_0,h_i 对 T 进行初步预测 */

④ 构造如(9)所示的目标函数:

$$\min_{\boldsymbol{w}\in M}\ell\left(\boldsymbol{y}_0,\sum_{i=1}^{3}w_i\boldsymbol{y}_i\right)+\frac{1}{u}\left\|\sum_{i=1}^{3}w_i\boldsymbol{y}_i\right\|_1;$$

/* u 为测试集 T 中的数据个数 */

⑤ 求解该优化问题并得到最优解 w_i^*;

⑥ $h(x)\leftarrow\operatorname{sgn}\left(\sum_{i=1}^{3}w_i^*h_i(x)\right)$;

/* 计算 T 中数据 x 的标签 */

⑦ 输出所有的 $h(x)$.

首先,算法 2 的行①在标记数据集 L 上训练了一个基准分类器 h_0,行②调用了传统的 Tri-training 算法,得到 3 个训练好的基分类器 h_i $(i=1,2,3)$;其次,行③利用基准分类器 h_0 和 3 个基分类器 h_i 对测试数据集 T 进行初步预测,得到基准预测结果 $\boldsymbol{y}_0$ 和候选预测结果 $\boldsymbol{y}_i$;然后,行④⑤按照文献[27]的思想构建一个凸优化问题,并求解该优化问题,得到最优权重 w_i^*;最后,输出未标记数据的标签.

Tri-training 算法在基分类器预测分类结果后,采用多数投票方法来确定未标记数据的标签,未充分考虑每个分类器自身的强弱,可能降低未标记数据的分类性能,甚至低于直接用基准分类器 h_0 对未标记数据分类得到的分类性能.与 Tri-training 算法相比,算法 2 分别用基准分类器和 Tri-training 对未标记数据进行学习,利用其初始预测结果构建一个优化问题,从而提升未标记数据的分类性能.

2.3 基于交叉熵的安全 Tri-training 学习框架

为了减少算法 2 过程中产生的噪声标记,进一步降低 Tri-training 的误分率,本文将交叉熵用于

安全 Tri-training 算法中,提出基于交叉熵的安全 Tri-training 算法,如算法 3 所示.算法 3 不仅利用交叉熵来估算每个基分类器的分类误差,而且还基于初始预测结果构建了一个优化问题来求解最优解.

算法 3. 基于交叉熵的安全 Tri-training 算法.

输入:训练集 $D=L\cup U$,其中 L 和 U 分别为已标记数据集和未标记数据集,测试集 T;

输出:测试数据集 T 中数据 x 的标签 $h(x)$.

① $h_0\leftarrow Learn(L)$;

/* 用 L 训练基准分类器 h_0 */

② $h_i\leftarrow Learn(Bootstrap(L))$, 初始化参数 $e_i'\leftarrow 0.5$ 和 $|L_i'|\leftarrow 0$ $(i=1,2,3)$;

/* 重采样 L 训练 3 个基分类器 h_i */

③ repeat

④　for $i=1$ to 3 do

⑤　　$L_i\leftarrow\varnothing$;

⑥　　$e_i\leftarrow(H_j+H_k)/2$; /* H_j, $H_k(j,k\neq i)$为基分类器的交叉熵 */

⑦　　if $e_i<e_i'$ then

⑧　　for U 中每个样本 x do

⑨　　if $h_j(x)=h_k(x)$ $(j,k\neq i)$ then

⑩　　$L_i\leftarrow L_i\cup\{(x,h_j(x))\}$;

⑪　　end if

⑫　　end for

⑬　　end if

⑭　　if $e_i|L_i|>e_i'|L_i'|$ then

⑮　　更新 L_i; /* 根据约束条件式(2),从 L_i 中随机删除若干个数据 */

⑯　　end if

⑰　end for

⑱　for $i=1$ to 3 do /* 更新 */

⑲　　$S_i\leftarrow L\cup L_i$, $h_i\leftarrow Learn(S_i)$;

⑳　　$e_i'\leftarrow e_i$, $|L_i'|\leftarrow|L_i|$;

㉑　end for

㉒ until h_i 的性能没有改变

㉓ $\boldsymbol{y}_0\leftarrow h_0(T)$, $\boldsymbol{y}_i\leftarrow h_i(T)$ $(i=1,2,3)$;

㉔ 构造目标函数:

$$\min_{w\in M}\left(\ell\left(\boldsymbol{y}_0,\sum_{i=1}^{3}w_i\boldsymbol{y}_i\right)+\frac{1}{u}\left\|\sum_{i=1}^{3}w_i\boldsymbol{y}_i\right\|_1\right);$$

㉕ 求解该优化问题并得到最优解 w_i^*;

㉖ $h(x)\leftarrow \operatorname{sgn}\left(\sum_{i=1}^{3}w_i^*h_i(x)\right)$;

/* 计算 T 中数据 x 的标签 */

㉗ 输出所有的 $h(x)$.

3　实验结果与分析

3.1　实验数据集

为了验证本文方法的有效性,实验选取 12 个 UCI(University of California Irvine)机器学习库[30]中的数据集和 1 个入侵检测数据集 UNSW-NB15[31],共 13 个数据集,这些数据集仅包含 2 类,数据集基本信息如表 1 所示.其中,入侵检测数据集 UNSW-NB15 仅分为正常类和异常类两大类,且选取该数据集的 10%进行实验.

Table 1　Experimental Datasets

表 1　实验数据集

数据集	样本数	属性数	类比例/%	
			正类	负类
wdbc	569	30	37.3	62.7
winewhite	4898	11	33.5	66.5
abalone	4177	8	32.1	67.9
haberman	306	3	28.4	71.6
spect	267	44	20.6	79.4
chronic	400	14	37.5	62.5
adult	32561	14	24.1	75.9
electrical	10000	13	36.2	63.8
Australian	690	14	44.5	55.5
liverdisorders	345	6	42.0	58.0
heart	270	13	44.4	55.6
german	1000	24	30.0	70.0
UNSW-NB15	175341	42	31.9	68.1

本文针对半监督学习,所需要的数据集仅含有少量的标记数据,其余大部分为未标记数据,而本文的 13 个数据集都带有标记.为了构造半监督学习任务,模拟现实中仅含有少量标记样本的情况,实验中选取 20%的样本作为测试集 T,余下的全部为训练集 D,从 D 中选取 20%作为已标记样本集 L,80%作为未标记样本集 U.

3.2　实验结果分析

本文提出了 3 个算法,包括基于交叉熵的 Tri-training 算法(Tri-training algorithm based on cross entropy, TCE)、安全的 Tri-training 算法(safe Tri-training algorithm, ST)和基于交叉熵的安全 Tri-training 算法(safe Tri-training algorithm based on cross entropy, STCE),与 Tri-training 算法进行实验对比.实验中采用反向传播神经网络(back propagation, BP)作为基分类器.

本文在混淆矩阵的基础上对算法进行性能评价.对于二分类问题,混淆矩阵如表 2 所示:

Table 2 Confusion Matrix

表 2 混淆矩阵

真实类别	预测类别	
	正类	负类
正类	*TP*(true positives)	*FN*(false negatives)
负类	*FP*(false positives)	*TN*(true negatives)

表 2 中 TP,FP,TN,FN 分别表示真正类、假正类、真负类、假负类的数量.召回率 *Recall*、精度 *Precision*、特效性 *Specificity*、*F* 值 *F-measure*、*G* 均值 *G-means*、准确率 *Accuracy* 等性能指标的计算公式分别为

$$Recall=TP/(TP+FN), \tag{10}$$

$$Precision=TP/(TP+FP), \tag{11}$$

$$Specificity=TN/(TN+FP), \tag{12}$$

$$F\text{-}measure=\frac{2\times Precision\times Recall}{Precision+Recall}, \tag{13}$$

$$G\text{-}means=\sqrt{Recall\times Specificity}, \tag{14}$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}. \tag{15}$$

本文选取了 *F-measure*,*G-means*,*Accuracy*,*Precision*,*Recall* 这 5 个指标来评价算法性能,实验结果如表 3~7 所示:

Table 3 *F-measure*

表 3 *F-measure* 值

数据集	Tri-training	TCE	ST	STCE
wdbc	0.937 1	0.951 7	0.945 2	**0.958 3**
winewhite	0.515 5	0.537 3	**0.563 8**	0.539 9
abalone	0.570 3	**0.575 6**	0.572 0	0.572 0
haberman	0.195 1	0.222 2	**0.294 1**	0.250 0
spect	**0.540 5**	0.500 0	0.408 2	0.509 8
chronic	0.962 0	0.926 8	**1.000 0**	0.938 3
adult	**0.619 0**	0.610 0	0.610 9	0.614 2
electrical	0.993 4	0.992 3	0.991 8	**0.994 5**
Australian	0.787 5	0.809 8	0.807 7	**0.847 1**
liverdisorders	0.393 9	0.387 1	0.434 8	**0.557 0**
heart	0.6667	0.698 4	0.733 3	**0.761 9**
german	0.5029	0.531 5	0.563 8	**0.569 3**
UNSW-NB15	0.943 2	0.941 9	0.940 8	**0.943 6**

注:黑体值为最佳性能指标值.

Table 4 *G-means*

表 4 *G-means* 值

数据集	Tri-training	TCE	ST	STCE
wdbc	0.938 2	0.951 2	0.943 8	**0.958 6**
winewhite	0.641 4	0.644 2	**0.650 7**	0.642 4
abalone	0.700 6	0.688 2	0.692 6	**0.724 9**
haberman	0.351 8	0.365 7	**0.476 3**	0.383 1
spect	**0.622 6**	0.576 3	0.500 9	0.584 9
chronic	0.962 7	0.929 3	**1.000 0**	0.940 1
adult	**0.790 2**	0.783 7	0.783 6	0.786 0
electrical	0.995 1	0.994 4	0.99 37	**0.995 5**
Australian	0.801 0	0.818 9	0.824 7	**0.851 3**
liverdisorders	0.584 9	**0.641 2**	0.588 3	0.612 4
heart	0.687 3	0.716 8	0.758 6	**0.775 9**
german	0.588 6	0.647 8	0.662 4	**0.689 0**
UNSW-NB15	0.940 6	0.935 8	0.936 6	**0.943 0**

注:黑体值为最佳性能指标值.

Table 5 *Accuracy*

表 5 准确率

数据集	Tri-training	TCE	ST	STCE
wdbc	0.937 1	0.951 0	0.944 1	**0.958 0**
winewhite	**0.706 9**	0.706 1	**0.706 9**	0.703 7
abalone	0.789 5	0.779 9	0.783 7	**0.803 8**
haberman	0.576 9	0.551 3	**0.692 3**	0.538 5
spect	**0.746 3**	0.641 8	0.567 2	0.626 9
chronic	0.970 3	0.940 6	**1.000 0**	0.950 5
adult	**0.839 6**	0.835 9	0.836 0	0.837 4
electrical	0.995 2	0.994 4	0.994 0	**0.996 0**
Australian	0.803 5	0.820 8	0.826 6	**0.849 7**
liverdisorders	0.540 2	0.563 2	0.551 7	**0.597 7**
heart	0.691 2	0.720 6	0.7647	**0.779 4**
german	0.660 0	0.732 0	0.740 0	**0.746 0**
UNSW-NB15	0.918 4	0.916 8	0.915 0	**0.918 9**

注:黑体值为最佳性能指标值.

Table 6 *Precision*

表 6 精度

数据集	Tri-training	TCE	ST	STCE
wdbc	0.893 3	**0.920 0**	**0.920 0**	**0.920 0**
winewhite	0.506 6	0.554 4	**0.615 4**	0.565 0
abalone	0.570 3	**0.609 4**	0.589 8	0.535 2
haberman	0.25 00	0.312 5	0.312 5	**0.375 0**
spect	0.769 2	0.923 1	0.769 2	**1.000 0**
chronic	**1.000 0**	**1.000 0**	**1.000 0**	**1.000 0**

Continued (Table 6)

数据集	Tri-training	TCE	ST	STCE
adult	**0.542 2**	0.534 0	0.535 5	0.538 6
electrical	0.992 4	0.990 2	0.991 3	**0.995 6**
Australian	0.807 7	0.846 2	0.807 7	**0.923 1**
liverdisorders	0.276 6	0.255 3	0.319 1	**0.468 1**
heart	0.724 1	0.758 6	0.758 6	**0.827 6**
german	**0.632 4**	0.558 8	0.617 6	0.573 5
UNSW-NB15	0.995 4	0.991 3	0.993 4	**0.997 6**

注:黑体值为最佳性能指标值.

Table 7　*Recall*

表 7　召回率

数据集	Tri-training	TCE	ST	STCE
wdbc	0.985 3	0.985 7	0.971 8	**1.000 0**
winewhite	**0.524 7**	0.521 2	0.520 2	0.517 0
abalone	0.570 3	0.545 5	0.555 1	**0.614 3**
haberman	0.160 0	0.172 4	**0.277 8**	0.187 5
spect	**0.416 7**	0.342 9	0.277 8	0.342 1
chronic	0.926 8	0.863 6	**1.000 0**	0.883 7
adult	**0.721 3**	0.711 4	0.711 0	0.714 6
electrical	0.993 5	**0.994 5**	0.992 3	0.993 5
Australian	0.768 3	0.776 5	**0.807 7**	0.782 6
liverdisorders	0.684 2	**0.800 0**	0.681 8	0.687 5
heart	0.617 6	0.647 1	**0.709 7**	0.705 9
german	0.417 5	0.506 7	0.518 5	**0.565 2**
UNSW-NB15	0.896 2	**0.897 2**	0.893 6	0.895 2

注:黑体值为最佳性能指标值.

从表 3～7 可以看出,大部分数据集在 STCE 算法上表现较好,其中在 *F-measure*, *G-means*, *Accuracy*, *Precision* 这 4 个指标上,分别在 7,7,8,9 个数据集上取得了最优的性能.ST 算法在 *F-measure*, *G-means*, *Accuracy* 这 3 个指标上表现仅次于 STCE 算法.直观上看,在 *Recall* 指标上,ST 算法在 4 个数据集上表现最好,而 STCE 和 Tri-training、TCE 算法相似,都在 3 个数据集上取得了较好性能.相对而言,仅有极个别数据集在 TCE 和 Tri-training 算法上表现最优.

为了进一步分析 4 个算法的性能,我们引入统计显著性检验[32],从统计学的角度来比较分析算法性能.本文采用 Friedman 检验进行计算,通过重复测量方差分析(analysis of variance, ANOVA)来比较各个算法的平均排名.其中,重复测量 ANOVA 是一种测试 2 个以上相关样本均值之间差异的统计方法.表 8 显示了本文提出的 3 种算法与 Tri-training 算法经过 Friedman 检验后的等级排名,取值越低等级越高.

Table 8　Average Rank After Friedman Test

表 8　Friedman 检验后的平均等级

评价指标	Tri-training	TCE	ST	STCE
F-measure	2.999 99	2.923 07	2.500 00	1.576 92
G-means	2.923 08	2.923 08	2.538 46	1.615 38
Accuracy	2.576 92	3.000 00	2.576 91	1.846 15
Precision	2.884 62	2.884 62	2.500 04	1.730 77
Recall	2.499 96	2.500 04	2.769 23	2.230 77
均值	2.776 91	2.846 16	2.576 93	1.799 99

由表 8 可以看出,基于交叉熵的安全 Tri-training 算法 STCE 在 5 项评价指标上均取得了最高的等级,最低取值为 1.576 92,最高取值也只有 2.230 77,比传统的 Tri-training 算法的最低取值 2.499 96 还要低.根据平均等级可以得到 4 个算法的分类性能,从高到低依次为:STCE, ST, Tri-training, TCE,即 STCE 算法的分类性能最好,ST 算法次之,但也明显好于 Tri-training 和 TCE 算法.Tri-training 算法和 TCE 算法性能相当,表明仅用交叉熵替代误分率作为 Tri-training 更新的条件,并不能显著改善半监督学习的分类性能,而利用凸优化可以在一定程度上提高半监督学习的分类性能,将交叉熵和凸优化方法相结合,可得到更好的分类性能.

我们进一步使用 Holm 检验[32]来对最佳排名方法 STCE 与其他方法进行比较,并选择 STCE 作为控制方法,取置信水平 $\alpha=0.05$ 进行测试,结果如表 9 所示:

Table 9　Holm Test Results

表 9　Holm 检验结果

算法	i	概率值 p_i	Holm 测试值 α/i	假设 $\alpha=0.05$
TCE	3	0.009 80	0.016 66	拒绝 STCE
Tri-training	2	0.009 81	0.025	拒绝 STCE
ST	1	0.068 31	0.05	接受 STCE

表 9 中的 $p_i(i=1,2,3)$ 是由 Holm 检验计算出来的概率值,用来度量否定原假设的证据.概率值越低,否定原假设的证据越充分.检验从概率值 p_3 开始,由于 $p_3=0.009\,80$ 小于 Holm 测试值 $\alpha/3=0.016\,66$,因此拒绝假设.同样,第 2 个假设也被拒绝.对于最后一个假设,由于 $p_1>0.05$,故接受该假设.

结果表明,STCE 算法优于 TCE 算法和传统的 Tri-training 算法,然而 STCE 算法和 ST 算法之间没有统计学上的差异.这个结论和表 8 得出的结论相一致.

4 总结与展望

传统的 Tri-training 算法能有效提升半监督学习的分类性能.为了进一步降低 Tri-training 算法过程中产生的误标率,提高算法性能,获得良好的半监督分类模型,本文分别提出了基于交叉熵的 Tri-training 算法(TCE)、安全的 Tri-training 算法(ST)和基于交叉熵的安全 Tri-training 算法(STCE).实验结果表明,提出的 STCE 方法具有最好的分类性能.然而,本文并没有考虑到半监督学习过程中数据集不平衡的问题,这将是我们下一步的研究内容和方向.

参 考 文 献

[1] Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning [M]. Cambridge, MA: MIT Press, 2006

[2] 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 (in Chinese)
(许勐璠, 李兴华, 刘海, 等. 基于半监督学习和信息增益率的入侵检测方案[J]. 计算机研究与发展, 2017, 54(10): 2255-2267)

[3] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labeled and unlabeled data [C] //Proc of the 9th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 1996: 571-577

[4] Bennett K P, Demiriz A. Semi-supervised support vector machines [C] //Proc of the 11th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 1998: 368-374

[5] Joachims T. Transductive inference for text classification using support vector machines [C] //Proc of the 16th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1999: 200-209

[6] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts [C] //Proc of the 8th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 19-26

[7] Zhou Zhihua. Disagreement-based semi-supervised learning [J]. ACTA Automatica Sinica, 2013, 39(11): 1871-1878 (in Chinese)
(周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11): 1871-1878)

[8] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding [C] //Proc of the 19th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2002: 19-26

[9] Blum A. Combining labeled and unlabeled data with co-training [C] //Proc of the 11th Annual Conf on Computational Learning Theory. New York: ACM, 1998: 92-100

[10] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifiers [J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(11): 1529-1541

[11] Søgaard A. Simple semi-supervised training of part-of-speech taggers [C] //Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 205-208

[12] Saito K, Ushiku Y, Harada T. Asymmetric tri-training for unsupervised domain adaptation [C] //Proc of the 34th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2017: 2988-2997

[13] Ruder S, Plank B. Strong baselines for neural semi-supervised learning under domain shift [C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1044-1054

[14] Ou Depin, Tan Kun, Du Qian, et al. A novel tri-training technique for the semi-supervised classification of hyperspectral images based on regularized local discriminant embedding feature extraction [J]. Remote Sensing, 2019, 11(6): 654

[15] Park H, Yun S, Eum J, et al. Weakly labeled sound event detection using tri-training and adversarial learning [J]. arXiv preprint, arXiv:1910.06790, 2019

[16] Li Yufeng, Liang Deming. Safe semi-supervised learning: A brief introduction [J]. Frontiers of Computer Science, 2019, 13(4): 669-676

[17] Angluin D, Laird P. Learning from noisy examples [J]. Machine Learning, 1988, 2(4): 343-370

[18] Kullback S, Leibler R A. On information and sufficiency [J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86

[19] MacKay D J C. Information Theory, Inference, and Learning Algorithm [M]. Cambridge, UK: Cambridge University Press, 2003

[20] Rubinstein R Y. Optimization of computer simulation models with rare events [J]. European Journal of Operational Research, 1997, 99(1): 89-112

[21] Bosman A S, Engelbrecht A, Helbig M. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions [J]. arXiv preprint, arXiv: 1901.02302, 2019

[22] Li Xiaoxu, Yu Liyun, Chang Dongliang, et al. Dual cross-entropy loss for small-sample fine-grained vehicle classification [J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4204-4212

[23] Lu Sheng, Gao Feng, Piao Changhao, et al. Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data [C] //Proc of Int Conf on Artificial Intelligence and Advanced Manufacturing. Piscataway, NJ: IEEE, 2019: 230-233

[24] Lopez-Garcia P, Onieva E, Osaba E, et al. A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy [J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(2): 557-569

[25] Li Yufeng, Zhou Zhihua. Towards making unlabeled data never hurt [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 175-188

[26] Krijthe J H, Loog M. Implicitly constrained semi-supervised least squares classification [C] //Proc of the 14th Int Symp on Intelligent Data Analysis. Berlin: Springer, 2015: 158-169

[27] Guo Lanzhe, Li Yufeng. A general formulation for safely exploiting weakly supervised data [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 3126-3133

[28] Li Yufeng, Guo Lanzhe, Zhou Zhihua. Towards safe weakly supervised learning [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019 [2019-06-12]. https://ieeexplore.ieee.org/document/8735810

[29] Rosasco L, Vito E D, Caponnetto A, et al. Are loss functions all the same [J]. Neural Computation, 2004, 16(5): 1063-1076

[30] Dua D, Graff C. UCI Machine Learning Repository[DB/OL]. [2019-12-10]. http://archive.ics.uci.edu/ml

[31] Tavallaee M, Bagheri E, Lu Wei, et al. A detailed analysis of the KDD cup 99 data set [C] //Proc of the IEEE Symp on Computational Intelligence for Security and Defense Applications. Piscataway, NJ: IEEE, 2009: 53-58

[32] Demšar J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research, 2006, 7(1): 1-30

Zhang Yong, born in 1975. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data mining, machine learning, and intelligent computing.

张　永, 1975 年生.博士,教授,博士生导师, CCF 高级会员.主要研究方向为数据挖掘、机器学习和智能计算等.

Chen Rongrong, born in 1994. Master candidate. Her main research interests include machine learning and data mining.

陈蓉蓉,1994 年生.硕士研究生.主要研究方向为机器学习和数据挖掘.

Zhang Jing, born in 1984. PhD. Member of CCF. Her main research interests include pattern recognition and machine learning.

张　晶,1984 年生.博士,CCF 会员.主要研究方向为模式识别和机器学习.