软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(2):279−288 [doi: 10.3724/SP.J.1001.2012.03977]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于最大相关熵准则的鲁棒半监督学习算法*

杨南海[1+], 黄明明[1], 赫 然[1,2], 王秀坤[1]

[1](大连理工大学 计算机科学与技术学院,辽宁 大连 116024)

[2](模式识别国家重点实验室(中国科学院 自动化研究所),北京 100190)

Robust Semi-Supervised Learning Algorithm Based on Maximum Correntropy Criterion

YANG Nan-Hai[1+], HUANG Ming-Ming[1], HE Ran[1,2], WANG Xiu-Kun[1]

[1](School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

[2](National Laboratory of Pattern Recognition (Institute of Automation, The Chinese Academy of Sciences), Beijing 100190, China)

+ Corresponding author: E-mail: yang.nanhai@gmail.com

Yang NH, Huang MM, He R, Wang XK. Robust semi-supervised learning algorithm based on maximum correntropy criterion. *Journal of Software*, 2012,23(2):279−288. http://www.jos.org.cn/1000-9825/3977.htm

Abstract: This paper analyzes the problem of sensitivity to noise in the mean square criterion of Gaussian-Laplacian regularized (GLR) algorithm. A robust semi-supervised learning algorithm based on maximum correntropy criterion (MCC), called GLR-MCC, is proposed to improve the robustness of GLR along with its convergence analysis. The half quadratic optimization technique is used to simplify the correntropy optimization problem to a standard semi-supervised problem in each iteration. Experimental results on typical machine learning data sets show that the proposed GLR-MCC can effectively improve the robustness of mislabeling noise and occlusion as compared with related semi-supervised learning algorithms.

Key words: semi-supervised learning; Gaussian-Laplacian regularized; correntropy; robust; half quadratic optimization

摘 要: 分析了噪声对半监督学习 Gaussian-Laplacian 正则化(Gaussian-Laplacian regularized,简称 GLR)框架的影响,针对最小二乘准则对噪声敏感的特点,结合信息论的最大相关熵准则(maximum correntropy criterion,简称 MCC),提出了一种基于最大相关熵准则的鲁棒半监督学习算法(简称 GLR-MCC),并证明了算法的收敛性.半二次优化技术被用来求解相关熵目标函数.在每次迭代中,复杂的信息论优化问题被简化为标准的半监督学习问题.典型机器学习数据集上的仿真实验结果表明,在标签噪声和遮挡噪声的情况下,该算法能够有效地提高半监督学习算法性能.

关键词: 半监督学习;Gaussian-Laplacian 正则化;相关熵;鲁棒;半二次优化

中图法分类号: TP181 **文献标识码**: A

随着信息时代数据收集和存储能力的极大提高,人们必须面对和处理海量数据.如何对这些数据进行分析和发掘,成为信息领域的共同需求.在实际问题中,如文本分类、生物复杂信息处理、人脸识别等,这些海量数据往往是无标记的,而有标记的数据是十分有限的.一方面,如果只使用少量的标记样本,那么所训练出的学习系

* 基金项目: 国家自然科学基金(60873054, 50909012); 国家教育部高等学校博士学科点专项科研基金(20100041120009)
收稿时间: 2010-05-04; 修改时间: 2010-08-13, 2010-10-11; 定稿时间: 2010-12-09

统很难具有良好的泛化能力[1];另一方面,如果只使用未标记样本,则浪费了标记样本中所提供的有用信息.针对这个问题,1992 年,Merz 等人在利用有标记和无标记数据的分类学习中首次提出了半监督这个概念.它利用大量的无标记数据辅助学习样本,弥补标记数据不足的缺点.目前,在半监督学习领域涌现出了很多算法,如肖宇等人提出的基于近邻传播算法[2]、尹学松等人提出的判别半监督聚类分析[3]、高滢等人提出的 k 均值多关系数据聚类算法[4]等.半监督学习的目标通常分为两类:一类是预测出未标记数据的标签;另一类是归纳出一个决策函数,使得在整个样例空间上错误率最小.显然,归纳决策函数更加复杂和困难.

近年来,半监督领域的一个突出成就是提出了基于图的半监督学习方法.它先将样本点映射成连通带权无向图,然后在构造的图上进行训练.国内外学者针对图的构造等核心问题作了深入的研究并有了较好的应用.陈锦绣等人[5]利用基于图的半监督学习从文本中自动识别出实体之间的关系;Zhao 等人[6]通过构造样本点之间的关系建立图,进而在图上提取出特征向量;Yang 等人[7]结合全局和局部信息构造图,然后进行高效的特征抽取;Zhou 等人[8]利用 k-近邻构造图,提出算法使得图更平滑.但是,实际数据中往往存在噪声.噪声可能来源于人工错误分类,也可能来源于自然环境下的遮挡.数据噪声问题也得到了人们越来越多的关注[9].Pentland 等人提出了模特征空间方法,Naseem 等人在模概念的基础上进一步提出了模线性回归分类方法,Ohba 和 Ikeuchi 提出了特征窗方法来解决部分遮挡问题.但在基于图的半监督学习中,解决方法比较少,其中,文献[10]提出了基于启发式的算法用来解决半监督学习中的数据桥点(bridge point)问题;文献[11]自动计算数据样本之间的相似度,且通过对图的权重矩阵的调整使得算法对桥点不再敏感.虽然文献[10,11]显著提高了半监督学习算法对桥点的鲁棒性,但是它们没有考虑样本的误标记问题[12,13],而且缺少鲁棒理论基础[14].

Gaussian-Laplacian 正则化(Gaussian-Laplacian regularized,简称 GLR)框架是半监督学习中的经典框架,很多传统的半监督算法都基于该框架[8,15,16].本文分析了噪声对 Gaussian-Laplacian 正则化框架的影响,针对 GLR 框架中的最小二乘准则对噪声敏感的特点[17],结合信息论的相关熵准则,提出了一种基于最大相关熵准则(maximum correntropy criterion,简称 MCC)的鲁棒半监督学习算法(简称 GLR-MCC)**,并证明了算法的收敛性.半二次(half quadratic,简称 HQ)优化算法被用来优化最大相关熵目标函数,在每次迭代中,复杂的信息论目标函数被简化为标准的半监督学习问题,进而提出一种贪婪算法逐步增加目标函数,直至收敛.在典型的机器学习数据集上的仿真实验结果表明,该算法能够有效地处理半监督学习中的误标签噪声和图像噪声.

1 Gaussian-Laplacian 正则化算法

基于图的半监督学习问题的基本设置如下:

给定 n 个数据的集合 $X=\{x_1,\ldots,x_l,x_{l+1},\ldots,x_n\}\in R^d$,其中,$X_l=\{x_i\}_{i=1}^{l}$ 是已经标记的数据,$X_u=\{x_i\}_{i=l+1}^{n}$ 是未标记数据,通常 $l<<n$.标签集合 $L=\{1,\ldots,c\}$,前 l 个点的标签 $\{y_1,\ldots,y_l\}\in L$.对任意 $x_i\in X$,取自固定但通常是未知的分布 $p(x)$.将样本点映射成带权无向图 $G=(V,E)$,其中,V 是顶点集合,对应于整个数据集 X;E 是边集,每条边 $e_{ij}\in E$ 对应于非负权重 w_{ij},它反映 x_i 与 x_j 的相似度.如果 x_i 和 x_j 之间没有边,$w_{ij}=0$.

目前,有两个常用的基本假设来建立预测样本和学习样本之间的关系:聚类假设和流形假设.本质上说,这两类假设是一致的,只是关注的侧重点不同.聚类假设在基于图的半监督学习研究中使用更普遍,聚类假设[18]是指,如果高密度区域的某两个点可以通过区域内某条路径连接,那么这两点拥有相同标签的可能性比较大.Zhou 等人进一步研究了该假设的几何信息,它包含两层含义[8]:(1) 相近点的标签更有可能相同;(2) 相同结构中的点标签相同.前一个假设是局部的,而后者是全局的.根据聚类假设,基于图的半监督算法往往被看作是在图 G 上建立分类函数 f,其中,f 满足:(1) 靠近已经标记的顶点;(2) 整个图平滑.它分别满足聚类假设的两层含义.最小化该目标函数即可得到 f:

$$H(f)=\eta C(x_L)+\beta S(f) \tag{1}$$

** 这里讨论统计学意义下的鲁棒性.算法鲁棒是指算法自动校正由于噪点产生的误差,并利用没有被污染的数据学习.噪点是指那些远离数据中心的点[14].

其中,$C(\cdot)$是损失函数,用于度量预测标签和初始标x_L的差值;$S(f)$是对分类函数f平滑的惩罚,它反映了边缘分布$p(x)$的内在几何信息.我们之所以要惩罚f的几何信息,是因为在半监督学习中仅有很少一部分数据被标记,只通过最小化 f 的损失并不足以训练出一个好的分类器,因此需要先验知识(prior knowledge)帮助训练出一个好的分类器.而 $p(x)$恰恰能够反映出所需要的先验知识.并且,通常假设[19]$p(x)$和 $p(y|x)$之间有直接关系,即 x_1 和 x_2 两点的 $p(x)$相近,则概率分布 $p(y|x_1)$与 $p(y|x_2)$相近.换言之,$p(y|x)$是沿着 $p(x)$几何图形的测地线平滑地变化.

GLR 框架通过最小化以下目标函数寻找最优分类函数 f:

$$f=\arg\min_f \sum_{i=1}^{l}(f_i-y_i)^2+\zeta\sum_{i=1}^{n}\sum_{j=1}^{n}(f_i-f_j)^2w_{ij} \tag{2}$$

其中,前项是损失项,后项是平滑项;$f=[f_1,\ldots,f_l,\ldots,f_n]^T$,$f_i$ 是分类函数,数据集中每个数据都会计算出分类标签;$y=[y_1,\ldots,y_l,\ldots,y_n]^T$,$y_i$ 是数据初始标签;w_{ij} 是图的权重.对平滑项进行处理,很容易得到:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(f_i-f_j)^2w_{ij}=f^TLf \tag{3}$$

$L=D-W$ 是图的 Laplacian 系数,W 是图的权重矩阵,$D=diag\left(\sum_i w_{1i},\ldots,\sum_i w_{ni}\right)$,则公式(2)转化为

$$f=\arg\min_f \sum_{i=1}^{l}(f_i-y_i)^2+\zeta f^TLf \tag{4}$$

通过求导$\partial J/\partial f=0$,得到公式(4)的解:

$$f=(J+\zeta L)^{-1}Jy \tag{5}$$

其中,$J\in R^{n\times n}$ 是对角阵,y 是列向量.

$$J(i,i)=\begin{cases}1, & x_i\text{已标记}\\ 0, & x_i\text{未标记}\end{cases},\ y(i)=\begin{cases}y_i, & x_i\text{已标记}\\ 0, & x_i\text{未标记}\end{cases}.$$

由于公式(2)中的目标函数基于 l^2 范数,因此当噪点出现时,噪点将会显著影响目标函数[17],从而影响未标记样本的标签估计,使算法对未标记样本的标记准确率下降.

2 基于最大熵的鲁棒算法

理论分析和实验结果显示[17],基于相关熵准则的算法能够有效地处理非高斯噪声和大的噪点.为了解决 GLR 框架中最小二乘准则对噪声敏感的问题,结合信息论学习(information theoretic learning,简称 ITL),提出了基于最大相关熵准则的鲁棒半监督学习算法,并在半二次优化技术的基础上提出了贪婪算法,以求解这个非线性优化问题.

2.1 问题描述

近年来,信息论学习在机器学习和计算机视觉领域受到了越来越多的关注,基于信息论的算法已经在鲁棒学习和模式分类中显示了很多的优越性.如在文献[13,20,21]中,Renyi 熵、互信息和相关熵被用作目标函数来解决监督和非监督学习中的问题.相关熵是对两个随机变量 A 和 B 相似度的通用计算方法,定义为[17]

$$V_\sigma(A,B)=E[k_\sigma(A-B)] \tag{6}$$

$k_\sigma(\cdot)$是满足 Mercer 理论[22]的核函数,$E[\cdot]$是期望.相关熵利用核函数,将输入空间非线性地映射到高维空间.与传统的核方法不同,它使每个样本不相关.相关熵准则有坚实的理论基础,且有一些理论性质[23],如对称、非负和有界.与全局度量方式——均方误差(mean square error,简称 MSE)不同,相关熵是局部的,即,相关熵的值主要由核函数沿着直线 $A=B$ 计算,并且属于 Welsch M 估计量[17].它为非高斯信号和噪点的处理提出了新的、鲁棒的解决方案.由于实际问题中数据的联合概率密度往往是未知的,因此对于有限的样本 $\{(a_i,b_i)\}_{i=1}^n$,可得到如下相关关熵的定义[17]:

$$\hat{V}_\sigma(A,B)=\frac{1}{n}\sum_{i=1}^{n}k_\sigma(a_i-b_i) \tag{7}$$

将高斯核函数 $\left(g(a_i,b_i)=\exp\left(-\left(\frac{(a_i-b_i)^2}{\sigma^2}\right)\right)\right)$ 代入到公式(7)中,且分别代入到公式(2)的损失项和平滑项中(损失项中取 $a_i=f_i,b_i=y_i$;平滑项中取 $a_i=f_i,b_i=f_j$),则得到如下的相关熵问题:

$$E(f)=\sum_{i=1}^{l}g(f_i-y_i)+\zeta\sum_{i=1}^{n}\sum_{j=1}^{n}g(f_i-f_j)w_{ij}=\sum_{i=1}^{l}\exp\left(-\frac{(f_i-y_i)^2}{\sigma^2}\right)+\zeta\sum_{i=1}^{n}\sum_{j=1}^{n}\exp\left(-\frac{(f_i-f_j)^2}{\sigma^2}\right)w_{ij} \tag{8}$$

其中,$f=[f_1,\ldots,f_l,\ldots,f_n]^T$,$f_i$ 是求出的标签,最终,每个数据都有求得的标签,且可能与最初标记的标签不同;y_i 是初始标签.根据相关熵[17]属于 Welsch M 估计量,目标函数(8)使用 Welsch M 估计量替换 GLR 中的最小二乘准则.当噪点出现时,相关熵将会更关注局部聚类中心的数据,而忽略噪点.

根据 Karush-Kuhn-Tucker(简称 KKT)条件,对公式(8)中的目标函数求最大值即可得到标签函数 f.为了更好地说明该目标函数能够有效地处理噪声,分别对损失项和平滑项进行讨论.对于损失项,分两种情况:(1) y_i 标记错误,f_i 正确估计,则 $f_i\neq y_i$,把 y_i 看成噪点,期望它对目标函数的贡献比较小;(2) y_i 正确标记,但是由于 y_i 对应的数据具有噪声,以至于 f_i 永远不能被正确评估,因此也期望 f_i 对目标函数的贡献比较小,也就是把 y_i 看成噪点.对于第 2 个平滑项,根据半监督中的聚类假设,期望同类的相接近.若 f_i 和 f_j 属于同一标签,则认为 x_i 和 x_j 潜在地属于同一个聚类,那么 w_{ij} 是合理的,需要在目标函数中给予更多的重视;若 f_i 和 f_j 不属于同一类,那么 x_i 和 x_j 潜在地不属于同一个聚类,因此在目标函数中给予 w_{ij} 一个小的权重.

2.2 优化算法

在信息理论学习中,由于熵往往包含非线性核函数,以至于问题的解往往是非线性的.因此,子空间迭代[21]、半二次优化技术[20,24]、一阶泰勒展开[12]、期望最大化(expectation maximization,简称 EM)方法和共轭梯度法被引入到信息论学习中,用于求解非线性目标函数.本文应用半二次优化技术来求解问题(8).理论分析和实际结果表明,半二次优化技术往往比牛顿梯度法具有更快的收敛速度[24,25].

基于共轭凸函数理论[23],则有:

定理 1. 存在凸函数 $\varphi:R\to R$,满足:

$$g(x,\sigma)=\exp\left(-\frac{x^2}{\sigma^2}\right)=\sup_{p\in R^-}\left(p\frac{x^2}{\sigma^2}-\varphi(p)\right) \tag{9}$$

其中,p 是辅助共轭变量.对于任意一个固定的 x,上式在 $p=-g(x,\sigma)$ 处取得最大值[24].

根据定理 1,公式(8)可以转变为如下增广目标函数:

$$\hat{E}(f,P,Q)=\sum_{i=1}^{l}\left[p_i\frac{(f_i-y_i)^2}{\sigma^2}-\varphi_i(p_i)\right]+\zeta\sum_{i=1}^{n}\sum_{j=1}^{n}\left[q_{ij}\frac{(f_i-f_j)^2}{\sigma^2}-\phi_{ij}(q_{ij})\right]w_{ij} \tag{10}$$

其中,p_i 和 q_{ij} 均为半二次优化中的存储辅助变量.

根据定理 1 可知,对于确定的 f,得到以下等式成立:

$$E(f)=\sup_{P,Q}\hat{E}(f,P,Q) \tag{11}$$

其中,P 是对角阵,满足 $P(i,i)=p_i$;$Q=[q_{ij}]$ 是 $N\times N$ 的矩阵.因此,原目标函数取得最大值处也在该函数中取得最大值:

$$\max_f E(f)=\max_{f,P,Q}\hat{E}(f,P,Q) \tag{12}$$

根据公式(12)以及半二次优化技术,可以将原目标函数分成两个步骤进行求解[24]:

1. 计算辅助变量 P 和 Q:

$$p_i=-\exp\left(-\frac{(f_i^t-y_i)^2}{\sigma^2}\right),q_{ij}=-\exp\left(-\frac{(f_i^t-f_j^t)^2}{\sigma^2}\right).$$

2. 将辅助变量代入公式(12),得到:

$$f^{t+1}=\arg\max_{f}\sum_{i=1}^{l}p_i\frac{(f_i-y_i)^2}{\sigma^2}+\zeta\sum_{i=1}^{n}\sum_{j=1}^{n}q_{ij}\frac{(f_i-f_j)^2}{\sigma^2}w_{ij} \tag{13}$$

公式(13)中的第 2 项可以进一步转换为

$$\sum_{i=1}^{n}\sum_{j=1}^{n}q_{ij}\frac{(f_i-f_j)^2}{\sigma^2}w_{ij}=fL'f \tag{14}$$

其中,$L'=D'-W'$,W'是图的新权重矩阵;$w'_{ij}=p_{ij}w_{ij}$,w_{ij}是 GLR 算法的权重;D'是对角线矩阵,$D'(i,i)=\sum_j W'_{ij}$.

根据 KKT 条件,公式(13)的解为

$$f^{t+1}=(J'+\zeta L')^{-1}J'y \tag{15}$$

其中,$J'=diag([p_1,\ldots,p_n])$;$y$ 是列向量,$y=[y_1,\ldots,y_l,\ldots,y_n]^T$,$y(i)=\begin{cases}y_i, & x_i\text{已标记}\\ 0, & x_i\text{未标记}\end{cases}$.

由公式(15)求出的结果是针对二分类的情况,y 和求出的 f^{t+1} 都是列向量.对于多类,设分为 c 类,则求出的类别函数为

$$F^{t+1}=(J'+\zeta L')^{-1}J'Y \tag{16}$$

其中,F^{t+1} 和 Y 是矩阵,$F^{t+1}=\begin{bmatrix}f'_{11} & f'_{12} & \cdots & f'_{1k}\\ f'_{21} & f'_{22} & \cdots & f'_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ f'_{n1} & f'_{n2} & \cdots & f'_{nk}\end{bmatrix}$,$Y=\begin{bmatrix}y'_{11} & y'_{12} & \cdots & y'_{1k}\\ y'_{21} & y'_{22} & \cdots & y'_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ y'_{n1} & y'_{n2} & \cdots & y'_{nk}\end{bmatrix}$,$k=c$.

由于公式(16)求出的最优解 F^{t+1} 是一个连续变量而不是类别标签,所以提出如下贪婪算法求解一个局部最优解:设 $f'_i=[f'_{i1},\ldots,f'_{ik}]$ 是一个实数行向量,对其离散化得到数据标签,即,行向量中最小值所对应的列数即为该数据类别标签.但由于离散后的解未必是公式(13)的局部最优解,因此采用贪婪策略来计算一个使目标函数上升的解.即,若求得的标签不能使目标函数上升,则保留原来标签,否则替换.将求得的 F^{t+1} 作为下一次迭代的初始标签,迭代至所有的标签都没有变化为止.在每次迭代过程中,GLR-MCC 求解一个标准的 GLR 问题,因此其计算复杂度是原来的 t 倍.GLR-MCC 算法总结如下:

算法 1. 基于最大熵准则的鲁棒半监督(GLR-MCC)学习算法.

输入:$X=\{x_1,\ldots,x_l,x_{l+1},\ldots,x_n\}\in R^d$,其中,$X_l=\{x_i\}_{i=1}^{l}$ 是已标记的数据,$X_u=\{x_i\}_{i=l+1}^{n}$ 是未标记数据.

输出:n 个数据的标签 $f=[f_1,\ldots,f_l,\ldots,f_n]^T$.

步骤 1. 计算样本集 X 的权重矩阵 W.

步骤 2. repeat

(a) 计算 $w'_{ij}=p_{ij}w_{ij}$,并计算 $L'=D'-W'$;

(b) 根据公式(15),计算 J';

(c) 根据公式(16),计算 F^{t+1};

(d) 使用贪婪算法计算新的标签(离散化 F^{t+1});

until 所有点都不需要替换标签.

定理 2. 令 $\hat{E}^t\triangleq\hat{E}(f^t,P^t,Q^t)$ 为算法 1 在第 t 次迭代的损失函数,则算法 1 产生的序列 $\{\hat{E}^t\}_{t=1,2,\ldots}$ 收敛.

证明:根据半二次优化的性质[24]以及算法 1 在第 2 步选择一个使目标函数增加的方向,有

$$\hat{E}(f^t,P^t,Q^t)\leqslant\hat{E}(f^t,P^{t+1},Q^t)\leqslant\hat{E}(f^t,P^{t+1},Q^{t+1})\leqslant\hat{E}(f^{t+1},P^{t+1},Q^{t+1}).$$

根据相关熵的有界性[17],$\{\hat{E}^t\}_{t=1,2,\ldots}$ 是有界递增的,因此算法 1 收敛. □

3 对比实验和结果分析

3.1 数据集设置

为了验证基于最大相关熵准则的鲁棒算法的有效性,本节通过在不同数据集上的实验展示该算法与 GLR、

线性近邻繁殖算法(linear neighborhood propagation,简称 LNP)[10]算法比较的结果.实验所用数据集是 3 个 UCI (University of California, Irvine)数据集(UCI Machine Learning Repository, School of Information and Computer Science, University of California, Irvine. 2007. http://mlearn.ics.uci.edu/MLRepository.html)、ORL(Olivetti Research Laboratory)人脸数据集(the Oliver Research Laboratory in Cambridge, UK)和 FRGC(face recognition grand challenge)[26]人脸数据集.在 UCI 数据集上,检测两种算法在数据中有误标记情况下的性能;在人脸数据集上,不仅检测数据中存在误标记的情况,而且检测存在噪声图像情况下的性能.同时,测试了不存在噪声情况下的平均准确率.误标记的数据是标记数据中错误分类的数据,造成误标记的原因有多种,如手工误差、不同人对相同数据潜在地划分成不同类别等.图像噪声是由于光照和遮挡等因素产生的.表 1 具体介绍了数据集的特征.

Table 1 Datasets used in experiments

表 1 实验中使用的数据集

数据集	维数	分类数	实例数	数据特点
Balance-Scale	4	3	625	整数
Sonar	60	2	208	实数
Tic-Tac-Toe (TTT)	9	2	958	实数
ORL	1 024	40	400	非负
FRGC	1 024	186	3 720	非负

在下面的实验中,用于训练和测试的样本都是从整个数据集中随机选取的,比较指标是分类准确率;进行 50 次独立随机实验并记录其平均准确率和标准偏差.对于权重矩阵 W 的计算分两步:选择邻居和计算权重.选择邻居存在两种广泛应用的方法:ε-ball 最近邻和 k-最近邻.计算权重的核函数有 3 种:高斯核、逆欧几里德距离和局部线性重构.在 GLR 与 GLR-MCC 进行比较的实验中,我们采用了 k-最近邻(当且仅当 x_i 是 x_j 的 k 个最近的邻居之一时,x_i 是 x_j 的邻居),计算时使用高斯核:

$$W_{ij}=\begin{cases}\exp\left(-\dfrac{\|x_i-x_j\|^2}{\sigma^2}\right), & x_i和x_j是邻居\\ 0, & x_i和x_j非邻居\end{cases}.$$

在下面的实验中,k-近邻计算 W 时取邻居个数为 20,σ取 0.1.在与 LNP 算法比较时,采用了 k-最近邻和局部线性重构法计算权重矩阵.先选取邻居,最近邻参数设置为 40,然后对下式极小化计算矩阵:

$$\xi(W)=\sum_i\left\|x_i-\sum_j W_{ij}x_j\right\|^2 \quad \text{s.t.} \ \sum_j W_{ij}=1.$$

3.2 UCI数据集

UCI 数据集是机器学习中常用的标准测试数据集.本实验中使用了 3 个 UCI 数据集:Balance-scale(http://archive.ics.uci.edu/ml/machine-learning-databases/balance-scale),Sonar(http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar)和 TTT(http://archive.ics.uci.edu/ml/machine-learning-databases/tic-tac-toe).Balance-scale 来自于心理学实验;Sonar 数据集包含来自不同的信号源,包括汽缸的 90 个不同角度和岩石状况的 180 个角度;TTT 编码了一字棋游戏结束时所有可能的配置.从数据集的每类数据中随机地抽取 40%,60%,70%的样本作为训练集,其余的样本作为测试集.为了检测算法修正误标记标签的性能,人工随机设置误标记的数据,修改原标签或者把某些样本归为多类,误标记数据的个数设置为样本的 5%.

表 2 显示了无噪声情况下两种算法的平均准确率,表 3 显示了误标记情况下的平均准确率.由表 1 和表 2 可以看出:无论是否存在噪声,在 3 种 UCI 数据集上,GLR-MCC 算法都比 GLR 算法的学习准确率高;即使在误标记情况下,准确率在 40%的训练集上也达到 75%以上,最好的能达到 90%左右.在无噪声时,由于 GLR-MCC 算法能够自适应地调整权重矩阵,因此 GLR-MCC 算法都能比 GLR 算法能一个更高的准确率;当存在噪声时,两种算法都会降低学习的准确率,但是 GRL 准确率的下降明显大于 GLR-MCC,GLR-MCC 能够更有效地修正误标记数据,减小噪声对学习结果的影响.这是由于当出现误标记时,若标签被正确估计,误标记的样本则会被看成

噪点,对目标函数的贡献较小.该实验结果表明,GLR-MCC 算法具有更好的鲁棒性.

表 4 显示了 LNP 和 GLR-MCC 在误标记噪声下的平均分类准确率.由于 GLR-MCC 可以采用任意形式的权重矩阵,因此,这里我们使用和 LNP 算法相同的权重矩阵计算方式.显然,使用文献[10]中的方法构造权重矩阵比使用最近邻方法构造权重矩阵更有利于体现数据之间的关系,以至于 LNP 和 GLR-MCC 两种方法在大多数情况下的准确率都有提高.但是在错误标记情况下,GLR-MCC 依旧能够比 LNP 取得一个更好的准确率.这是由于 GLR-MCC 以相关熵为目标函数,属于 Welsch 鲁棒估计量;同时,误标记问题并不是桥点问题,因此 LNP 算法并不能很好地处理误标记噪声.

Table 2 Average classification accuracy without noise

表 2 无噪声情况下的平均分类准确率

数据集	GLR			GLR-MCC		
	40%	60%	70%	40%	60%	70%
Balance-Scale	91.10±0.013	91.61±0.014	92.09±0.016	91.77±0.011	92.68±0.015	94.02±0.018
Sonar	79.33±0.038	82.06±0.034	85.24±0.040	80.04±0.039	83.74±0.040	86.02±0.033
TTT	82.35±0.011	83.21±0.012	83.86±0.015	82.55±0.010	83.77±0.016	84.56±0.014

Table 3 Average classification accuracy on mislabeling noise

表 3 误标记噪声下的平均分类准确率

数据集	GLR			GLR-MCC		
	40%	60%	70%	40%	60%	70%
Balance-Scale	89.70±0.011	90.40±0.014	90.38±0.022	91.66±0.010	92.54±0.013	93.93±0.016
Sonar	77.45±0.047	80.32±0.040	83.91±0.038	79.34±0.042	83.29±0.036	85.84±0.037
TTT	81.77±0.013	82.82±0.016	82.87±0.019	82.27±0.012	83.40±0.014	84.37±0.018

Table 4 Average classification accuracy of LNP and GRL-MCC on mislabeling noise

表 4 LNP 和 GLR-MCC 在误标记噪声下的平均分类准确率

数据集	LNP			GLR-MCC		
	40%	60%	70%	40%	60%	70%
Balance-Scale	92.47±0.011	93.32±0.013	95.19±0.008	94.20±0.012	95.40±0.010	96.07±0.013
Sonar	79.33±0.058	80.41±0.031	80.49±0.042	79.44±0.033	80.17±0.045	81.25±0.062
TTT	82.07±0.010	83.01±0.016	83.75±0.020	82.76±0.014	84.32±0.018	84.12±0.022

3.3 ORL和FRGC人脸数据集

为了进一步检验 GLR-MCC 算法的鲁棒性,在 ORL 和 FRGC 两个人脸数据集上进行人脸识别实验,并与 GLR 和 LNP 作对比.ORL 数据集包含 400 张正面人脸图像,共有 40 个人,每人 10 幅灰度图像,是在不同的时间和不同的光照条件下拍摄的,面部表情包括睁眼/闭眼、笑/不笑,角度有左侧、右侧、低头、仰头,人脸细节有戴眼镜/不戴眼镜.在 FRGC 人脸数据库[26]的实验中,一共有 466 个个体,共 8 014 张图像.这些非控制光照的图像包括光照、表情、时间等的变化,因此是较难识别的.本文只选择人脸图像数目大于 20 张的个体[21],这样可以得到 186 个个体,共 3 720 张人脸图像.ORL 和 FRGC 中的每一张图像都被转化成 256 色的灰度图像,并且根据两个眼睛的位置切割成 32×32 个像素.

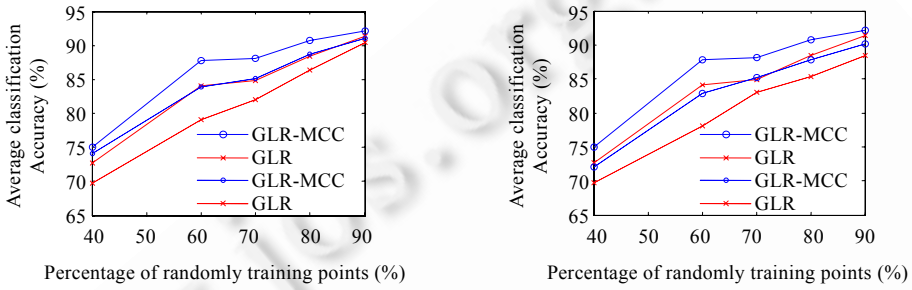
为了验证不同算法的鲁棒性,首先对图像引入人工噪声[12,21].从每个人的图像中随机抽取 10%作为噪声图像,对选中作为噪声的图像在不同位置进行随机矩形遮挡,即重新设置图像中某些部分像素.图 1 显示了 ORL 数据库中剪裁图像和对应的噪声图像.

Fig.1 Cropped facial images and their corresponding noisy images on ORL database

图 1 ORL 数据集中的剪裁图像和噪声图像

图 2(a)和图 3(a)分别显示了 GLR,LNP 和 GLR-MCC 在 ORL 和 FRGC 数据集上的实验结果,其中,虚线是两

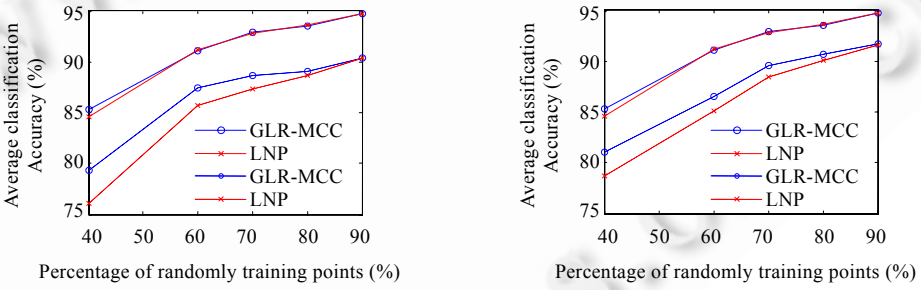
种算法加入噪声后的实验结果.可以看出,在没有噪声的数据集合(如FRGC)上,GLR-MCC相对于LNP在准确率上提高有限.这是由于学习到的权重矩阵已经能够很好地刻画数据之间的结构,而最大相关熵准则(或者鲁棒估计量)是专门针对噪声的方法,因此提高有限.在相关熵中,核宽度(kernel size)参数直接影响算法的鲁棒性.当选择一个较大的核宽度时,MCC趋近于MSE,以至于GLR-MCC等价于GLR.因此,在没有噪声的情况下,两者的结果趋于一致.在图像遮挡噪声下,GLR-MCC和其他算法的准确率都会随着训练样本的增多而提高,且在训练样本所占的百分比不同时,GLR-MCC与其他算法相比,提高的准确率差别明显.在训练集占40%和60%时准确率提高得比较大,60%之后,准确率的提高渐趋平稳.这是由于标记样本较少时,噪声对标签繁殖的影响较大,此时分类准确率就会受到影响.但由于在迭代过程中,预计标签 f_i 和实际标签 y_i 由于噪声产生一个较大的偏差,以至于GLR-MCC算法在每次迭代中给予噪声图像一个较小的权重,噪声图像在训练过程中对目标函数的影响越来越小,从而减弱了噪声图像对标签繁殖结果的影响.训练过程中去除了噪声图像,所以其准确率比其他算法明显提高;而当标记样本增多时,噪声相对标记样本的比例变小,从而减少了对标签繁殖的影响,则GLR-MCC算法和其他算法的分类准确率都较高,GLR-MCC提高得不明显.



(a) 图像噪声 (b) 误标记噪声

Fig.2 Average classification accuracy on ORL database

图2 在ORL数据库上的平均分类准确率



(a) 图像噪声 (b) 误标记噪声

Fig.3 Average classification accuracy on FGRC database

图3 在FRGC数据库上的平均分类准确率

对于误标记实验,在标记的图像中随机抽取10%的图像给予一个随机的错误类别标记[25].图2(b)和图3(b)显示了GLR-MCC和其他算法在ORL和FRGC人脸数据集上进行误标记实验的平均分类准确率.图中横坐标表示训练集所占比例,纵坐标表示平均分类准确率.实验结果表明,GLR-MCC比GLR和LNP取得了更好的分类准确率.虽然数据集不同,但在两个数据集上,GLR-MCC比GLR和LNP的准确率的提高趋于一致.当发生误标记错误时,由于预计标签 f_i 和实际标签 y_i 差别较大,因此在传统GLR算法的目标函数中产生一个较大的损失,从而影响了近邻样本的标记类别;但是对于GLR-MCC算法,由半二次优化求解过程可知,由于相关熵目标函数给予误标记的样本一个较小的权重,以至于误标记样本对目标函数的影响较小,从而在整体上得到一个更好的

学习结果.因此,GLR-MCC 算法能够更有效地处理误标记噪声.以上实验结果表明,GLR-MCC 算法对图像噪声和误标记噪声都具有鲁棒性.

4 结 论

本文提出了一种基于最大相关熵准则的鲁棒半监督学习算法,以提高半监督学习经典框架 GLR 对噪声的鲁棒性.该算法使用最大相关熵准则替换 GLR 框架中的最小二乘准则,在理论上具有更好的鲁棒性[6].为了解决相关熵非线性优化问题,本文在半二次优化技术的基础上提出了一种贪婪迭代算法,并证明了算法的收敛性.在标签噪声和图像噪声上的仿真结果表明,该算法能够有效地提高半监督学习算法的鲁棒性.

References:

[1] Wei J, Peng H. Local and global preserving based semi-supervised dimensionality reduction method. Journal of Software, 2008, 19(11):2833−2842 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2833.htm [doi: 10.3724/SP.J.1001.2008.02833]

[2] Xiao Y, Yu J. Semi-Supervised clustering based on affinity propagation algorithm. Journal of Software, 2008,19(11):2803−2813 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

[3] Yin XS, Hu EL, Chen SC. Discriminative semi-supervised clustering analysis with pairwise constraints. Journal of Software, 2008, 19(11):2791−2802 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2791.htm [doi: 10.3724/SP.J.1001.2008.02791]

[4] Gao Y, Liu DY, Qi H, Liu H. Semi-Supervised *K*-means clustering algorithm for multi-type relational data. Journal of Software, 2008,19(11):2814−2821 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2814.htm [doi: 10.3724/SP.J.1001.2008.02814]

[5] Chen JX, Ji DH. Graph-Based semi-supervised relation extraction. Journal of Software, 2008,19(11):2843−2852 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2843.htm [doi: 10.3724/SP.J.1001.2008.02843]

[6] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning. In: Ghahramani Z, ed. Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). New York: ACM, 2007. 1151−1157. [doi: 10.1145/1273496.1273641]

[7] Yang SH, Zha HY, Zhou SK, Hu BG. Variational graph embedding for globally and locally consistent feature extraction. In: Buntine W, ed. Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases: Part II (ECML PKDD 2009). Berlin: Springer-Verlag, 2009. 538−553. [doi: 10.1007/978-3-642-04174-7_35]

[8] Zhou DY, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. Advances in Neural Information Processing Systems, 2004,16:321−328.

[9] Li M, Zhang ZX, Huang KQ, Tan TN. Robust visual tracking based on simplified biologically inspired features. In: Bayoumi M, ed. Proc. of the 16th IEEE Int'l Conf. on Image Processing (ICIP 2009). Piscataway: IEEE Press, 2009. 4061−4064. [doi: 10.1109/ICIP.2009.5413456]

[10] Wang F, Zhang CS. Label propagation through linear neighborhoods. IEEE Trans. on Knowledge and Data Engineering, 2008, 20(1):55−67. [doi: 10.1109/TKDE.2007.190672]

[11] Wang F, Zhang CS. Robust self-tuning semi-supervised learning. Neurocomputing, 2007,70(16-18):2931−2939. [doi: 10.1016/j.neucom.2006.11.004]

[12] He R, Hu BG, Yuan XT. Robust discriminant analysis based on nonparametric maximum entropy. In: Zhou ZH, ed. Proc. of the 1st Asian Conf. on Machine Learning: Advances in Machine Learning (ACML 2009). Berlin: Springer-Verlag, 2009. 120−134. [doi: 10.1007/978-3-642-05224-8_11]

[13] Yuan XT, Hu BG. Robust feature extraction via information theoretic learning. In: Danyluk A, ed. Proc. of the 26th Annual Int'l Conf. on Machine Learning (ICML 2009). New York: ACM, 2009. 1193−1200. [doi: 10.1145/1553374.1553526]

[14] Huber P. Robust Statistics. New Jersey: Wiley, 1981.

[15] Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs. Lecture Notes in Computer Science, 2004,3120:624−638. [doi: 10.1007/978-3-540-27819-1_43]

[16] Zhu XJ, Ghahramani Z, Lafferty Z. Semi-Supervised learning using gaussian fields and harmonic functions. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning (ICML 2003). AAAI Press, 2003. 912−919.

[17] Liu WF, Pokharel PP, Principe JC. Correntropy: Properties and applications in non-gaussian signal processing. IEEE Trans. on Signal Processing, 2007,55(11):5286−5298. [doi: 10.1109/TSP.2007.896065]

[18] Chapelle O, Weston J, Schölkopf B. Cluster kernels for semi-supervised learning. Advances in Neural Information Processing System, 2003,15:585−592.

[19] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006,7(11):2399−2434.

[20] He R, Hu BG, Zheng WS, Guo YQ. Two-Stage sparse representation for robust recognition on large-scale database. In: Fox M, Poole D, eds. Proc. of the 24th AAAI Conf. on Artificial Intelligence (AAAI 2010). AAAI Press, 2010. 475−480.

[21] He R, Hu BG, Yuan XT, Zheng WS. Principal component analysis based on nonparametric maximum entropy. Neurocomputing, 2010,73(10-12):1840−1852. [doi: 10.1016/j.neucom.2009.12.032]

[22] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[23] Tyrrell RR. Convex Analysis. New Jersey: Princeton: Princeton Press, 1970.

[24] Nikolova M, Ng MK. Analysis of half-quadratic minimization methods for signal and image recovery. Society for Industrial and Applied Mathematics, 2005,27(3):937−966. [doi: 10.1137/030600862]

[25] Yuan XT. Algorithm study on non-parametric kernel density clustering and feature extraction [Ph.D. Thesis]. Beijing: the Chinese Academic of Sciences, 2009 (in Chinese with English abstract).

[26] Philips PJ, Flynn PJ, Sruggs T, Bowyer KW, Bowyer KW, Jin Chang, Hoffman K, Marques J, Jaesik Min, Worek W. Overview of the face recognition grand challenge. In: Hebert M, Kriegman D, eds. Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005). Washington: IEEE Computer Society, 2005. 947−954. [doi: 10.1109/CVPR.2005.268]

附中文参考文献:

[1] 韦佳,彭宏.基于局部与全局保持的半监督维数约减方法.软件学报,2008,19(11):2833−2842. http://www.jos.org.cn/1000-9825/19/2833.htm [doi: 10.3724/SP.J.1001.2008.02833]

[2] 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报,2008,19(11):2803−2813. http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

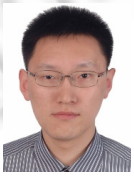
[3] 尹学松,胡恩良,陈松灿.基于成对约束的判别型半监督聚类分析.软件学报,2008,19(11):2791−2802. http://www.jos.org.cn/1000-9825/19/2791.htm [doi: 10.3724/SP.J.1001.2008.02791]

[4] 高滢,刘大有,齐红,刘赫.一种半监督 K 均值多关系数据聚类算法.软件学报,2008,19(11):2814−2821. http://www.jos.org.cn/1000-9825/19/2814.htm [doi: 10.3724/SP.J.1001.2008.02814]

[5] 陈锦绣,姬东鸿.基于图的半监督关系抽取.软件学报,2008,19(11):2843−2852. http://www.jos.org.cn/1000-9825/19/2843.htm [doi: 10.3724/SP.J.1001.2008.02843]

[25] 袁晓彤.非参数核密度聚类与特征提取算法研究[博士学位论文].北京:中国科学院研究生院,2009.

杨南海(1970－),男,四川江津人,博士生,讲师,CCF 会员,主要研究领域为机器学习,数据挖掘.



赫然(1979－),男,博士,助理研究员,CCF 会员,主要研究领域为信息论学习,鲁棒学习,计算机视觉.

黄明明(1985－),女,硕士,主要研究领域为模式识别,数据库系统.

王秀坤(1945－),女,教授,博士生导师,主要研究领域为数据库,决策支持系统.