刘忠宝,张兴芹,王文莉.融合磁极效应和数据分布特征的最大间隔学习机 [J].江西师范大学学报(自然科学版),2023,47(6):645-651.
LIU Zhongbao,ZHANG Xingqin,WANG Wenli.The maximum margin learning machine based on magnetic pole effect and data distribution characteristics [J].Journal of Jiangxi Normal University (Nature Science), 2023,47(6):645-651.

文章编号:1000-5862(2023)-06-0645-07

融合磁极效应和数据分布特征的最大间隔学习机

刘忠宝[1,2,3],张兴芹[1],王文莉[1]
(1.山东外国语职业技术大学信息工程学院,山东 日照 276826;2.北京语言大学语言智能研究院,北京 100083;
3.泉州信息工程学院软件学院,福建 泉州 362000)

摘要:基于几何边界的分类方法是一种典型的智能分类方法,已有的一些方法不仅忽略数据的分布特性,而且没有考虑不同样本对分类结果的影响,因而分类精度有待于进一步提高.鉴于此,受磁极效应启发,该文提出一种新颖的融合磁极效应和数据分布特征的最大间隔学习机.该模型构造的分类超平面距离一类尽可能近,而距离另一类尽可能远,尽量地将这 2 类分开.该模型利用类内离散度和类间离散度来刻画数据分布特征,以期在分类决策时将数据的分布形状考虑在内.此外,模糊隶属度的引入突出了不同样本对分类结果的影响.在 UCI 标准数据上的比较实验表明该方法是有效的.

关键词:分类;磁极效应;数据分布;类内离散度;类间离散度
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2023.06.12

0 引言

智能分类是数据挖掘领域研究的核心问题之一,研究人员提出了一系列智能分类方法并在实践中得到广泛应用.在众多分类方法中,基于几何边界的方法因设计简单、性能优良、适用范围广泛而备受推崇.支持向量机(support vector machine, SVM)[1]是一种典型的基于几何边界的分类方法,该模型通过构建分类超平面将 2 类分开.近年来,研究人员对支持向量机进行了系统研究并提出了若干改进模型.Zhang Wei 等[2]提出一种融合 Relief 特征提取和混合核函数的支持向量机,该模型首先利用 Relief 对训练样本进行特征提取,然后对每个特征加权用于表征特征的重要性,最后引入混合核函数构建支持向量机模型.Ding Hu 等[3]融合随机梯度下降树与支持向量机,用于发现噪声点和离群点.马婷婷等[4]引入非线性超平面思想,提出双参数化间隔支持向量机,该模型较之传统模型具有更优的鲁棒性和泛化能力.李建民等[5]提出融合差分进化和灰狼寻优的支持向量机,该模型利用灰狼寻优算法找到支持向量机的最优参数,差分进化算法保证灰狼寻优高效工作.程凤伟等[6]引入近邻传输思想,在粒计算理论基础上提出一种基于近邻传输的粒度支持向量机,该模型引入竞争机制以提取分类信息,获得了较好的训练效率和泛化性能.支持向量数据描述(support vector data description, SVDD)[7]也是一种典型的基于几何边界的分类方法,该模型通过构造一个球状模型将正常数据与异常数据分开.P. Nguyen 等[8]在 SVDD 基础上通过构造 2 个对称的球状模型进行分类.S. Kim 等[9]将 SVDD 与深度学习模型相融合,用于解决多分类问题,并取得了较好的效果.杨晨等[10]在 SVDD 上引入概率知识,提出基于概率的支持向量数据描述方法;该方法利

收稿日期:2023-03-16
基金项目:福建省社会科学基金(FJ2022A018, FJ2021B126)资助项目.
作者简介:刘忠宝(1981—),男,山西太谷人,教授,博士,博士生导师,主要从事数据分析与处理的研究.E-mail:liuzb@nuc.edu.cn

用 SVDD 训练 2 类样本,引入概率函数计算样本所属类别概率,借助传统 Bagging 集成算法,有效地提高了 SVDD 的分类性能.

近年来,一些新方法在分类决策时考虑了数据的分布形状.Hao Peiyi 等[11]引入模糊理论,提出一种模糊球形结构多类支持向量机,该模型构造的一组最小超球模型分别包含 2 类样本,并确保 2 类间隔最大化.陈鹏等[12]在统计数据分布规律的基础上采用 k 均值聚类计算聚类中心,然后采用分段组合方式设计支持向量机分类模型.宋瑞阳等[13]针对孪生支持向量机鲁棒性差、分类性能有限等问题,提出一种基于数据分布特征的加权线性孪生支持向量机,该模型关注数据分布特征对分类超平面位置的影响.R. B. Khanjani 等[14]在已知样本分布的情况下引入机会约束支持向量机(chance-constrained support vector machine, CCSVM),用于判定待测样本的类别.T. Bahraini 等[15]利用与错误分布相关的先验知识,提出一种基于错误分布知识的模糊支持向量机(fuzzy support vector machine based on prior knowledge related to error distribution, PKED-FSVM).

此外,模糊支持向量机相关研究也取得一些进展.顾晓清等[16]为了减少噪声点对分类结果的影响,提出了一种面向大规模噪声数据的软性核凸包支持向量机;该模型一方面保持样本的几何轮廓,另一方面忽略样本几何轮廓附近的噪声点,同时降低了在软性核凸包中噪声的敏感度.周裕群等[17]针对模糊孪生支持向量机噪声仍然敏感、容易过拟合以及不能有效区分支持向量和离群值等问题,提出了一种改进的鲁棒模糊孪生支持向量机;该研究最大贡献在于:将改进的 k 近邻隶属度函数和基于类内超平面的隶属度函数结合,构造了一种新的混合隶属度函数.戴小路等[18]针对基于欧氏距离设计的隶属度函数忽略了样本的总体分布且未考虑样本特征重要性的区分的问题,提出了一种基于加权马氏距离的模糊支持向量机方法.

在上述研究基础上,受磁极效应启发,该文基于文献[19]提出融合磁极效应和数据分布特征的最大间隔学习机(maximum margin learning machine based on magnetic pole effect and data distribution characteristics, MMLM).该模型尝试模拟磁极效应,构造的分类超平面距离一类尽可能近,而距离另一类尽可能远,尽量地将 2 类隔开.与文献[19]相比,该模型的主要优势在于:利用类内离散度和类间离散度来刻画数据分布特征,并在一定程度上提高模型的分类性能.

为了表示方便,本文做以下规定:$\boldsymbol{x}_i(i=1,2,\cdots,N)$ 表示样本,N 表示样本规模;其中 $\boldsymbol{x}_i(i=1,2,\cdots,m_1)$ 表示第 1 类,其类别标签 $y_i=1$;$\boldsymbol{x}_i(i=m_1+1,\cdots,N)$ 表示第 2 类,其类别标签 $y_i=-1$.

1 背景知识

1.1 磁极效应

一个磁体有 2 个磁极,当该磁体在水平面自由转动停止时,指向南方的磁极被称为南极(S 极),指向北方的磁极被称为北极(N 极).磁极效应指的是相同磁性相互排斥、不同磁性互相吸引的现象.若一个待测磁体与该磁体的 S 极相互吸引,则待测磁体为 N 极;若 2 者相互排斥,则待测磁体为 S 极.

1.2 支持向量机

支持向量机的基本思想是基于结构风险最小化原则,在样本空间中找到 1 个分类超平面,确保 2 类之间的间隔最大.令 $y=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+b$ 表示分类超平面,分类间隔表示为 $2/\|\boldsymbol{W}\|$,支持向量机最优化问题表示为

$$\min_{\boldsymbol{W},b,\xi_i}\|\boldsymbol{W}\|^2/2+C\sum_{i=1}^{N}\xi_i,$$

$$\text{s.t. } y_i(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,N,$$

其中 $\boldsymbol{W}$ 表示分类超平面的法向量,C 为惩罚因子,ξ_i 为松弛因子,其保证算法具有一定的容错性.

1.3 线性判别分析

线性判别分析(linear discriminant analysis, LDA)引入类内离散度和类间离散度的概念,在 Fisher 准则基础上建立最优化问题.对于 2 类问题,2 类离散度定义如下.

类内离散度为

$$\boldsymbol{S}_W=\sum_{i=1}^{c}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)^{\mathrm{T}}/N,$$

其中 c 表示样本类别数,令 $c=2$,x_{ij} 表示第 i 类的第 j 个样本,N 表示样本总数,N_i 表示第 i 类样本数,$\overline{\boldsymbol{x}}_i$ 表示第 i 类样本均值.

类间离散度为

$$\boldsymbol{S}_B=\sum_{i=1}^{c}N_i(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}/N,$$

其中 $\bar{x}$ 表示所有样本均值.

LDA 找到的投影方向 $\boldsymbol{W}$ 满足同类样本距离尽可能近,而异类样本距离尽可能远.在 Fisher 准则基础上建立如下优化问题:

$$J(\boldsymbol{W}_{\mathrm{opt}})=\max_{\boldsymbol{W}}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{W})/(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{W}).$$

上述优化问题的求解可转化为求解 $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{W}=\lambda\boldsymbol{W}$ 特征向量组成的矩阵 $\boldsymbol{W}$.

2　最大间隔学习机(MMLM)

2.1　优化问题

由磁极效应可知,一个磁体受到 N 极吸引,则必然受到 S 极排斥.受该理论启发,MMLM 试图在样本空间中找到一个“磁体”将 2 类样本分开,若该“磁体”对一类样本吸引,则 2 者距离尽可能地近;若该“磁体”对另一类样本排斥,则 2 者距离尽可能地远.该“磁体”即为 MMLM 构造的分类超平面.基于上述思想,提出融合磁极效应和数据分布特征的最大间隔学习机(MMLM).与支持向量机类似,该模型试图在样本空间中找到 1 个分类超平面将 2 类分开,但不同之处在于,该模型找到的分类超平面与一类样本尽可能近且与另一类尽可能远;此外,该模型引入类内离散度和类间离散度以表征样本的分布特征,引入模糊隶属度以突出不同样本对分类结果的影响.上述思想可描述为如下优化问题:

$$\min_{\boldsymbol{W},\rho,\xi,b}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{S}_W-\boldsymbol{S}_B)\boldsymbol{W}/2-\nu\rho+\sum_{i=1}^{m_1}\xi_i s_i/(\nu_1 m_1)+\sum_{j=m_1+1}^{N}\xi_j s_j/(\nu_2 m_2),\tag{1}$$

$$\text{s.t.}\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+b\leqslant\xi_i,1\leqslant i\leqslant m_1,\tag{2}$$

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j+b\geqslant\rho-\xi_j,m_1+1\leqslant j\leqslant N,\tag{3}$$

$$\xi\geqslant 0,\rho\geqslant 0,\sigma\leqslant s\leqslant 1.$$

其中 $\boldsymbol{W}$ 表示分类超平面的法向量,ρ 表示 2 类间隔,ξ_i 表示松弛因子,s_i 表示模糊隶属度函数,m_1 和 m_2 分别表示 2 类样本规模,σ 为大于 0 且小于 1 的任意数,ν、ν_1、ν_2 为模型参数,3 者均大于 0.

在目标函数(1)中,$\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{S}_W-\boldsymbol{S}_B)\boldsymbol{W}/2$ 用于确定 MMLM 分类超平面的法向量,其中 $\boldsymbol{S}_W$ 和 $\boldsymbol{S}_B$ 用来描述样本的分布性状;$\nu\rho$ 表示类间间隔;$\sum_{i=1}^{m_1}\xi_i s_i$ 和 $\sum_{j=m_1+1}^{N}\xi_j s_j$ 表示松弛因子,与 SVM 不同的是,该松弛因子引入模糊隶属度函数,该函数给样本赋予不同权重,以期松弛因子有一定的容错性.约束条件(2)和(3)表明各类样本在“磁体”(即分类超平面)作用下的不同反应,式(2)表示与“磁体”极性相异的样本受到“磁体”吸引,故 2 者之间距离较近;式(3)表示与“磁体”极性相同的样本受到“磁体”排斥,故 2 者之间距离较远.

上述优化问题的对偶形式为

$$\min_{\alpha\in\mathbf{R}^d}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{S}_W-\boldsymbol{S}_B)^{-1}\boldsymbol{x}_j/2,$$

$$\text{s.t.}\ 0\leqslant\alpha_i\leqslant s_i/(\nu_1 m_1),1\leqslant i\leqslant m_1,$$

$$0\leqslant\alpha_j\leqslant s_j/(\nu_2 m_2),m_1+1\leqslant j\leqslant N,$$

$$\sum_{i=1}^{N}\alpha_i y_i=0,\ \sum_{i=1}^{N}\alpha_i\geqslant 2\nu,$$

其中 α_i 为拉格朗日乘子且 $\alpha_i\geqslant 0$.

证　根据拉格朗日定理,定义原优化问题的拉格朗日方程如下:

$$L(\boldsymbol{W},\rho,\xi,b,\alpha,\beta,\lambda)=\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{S}_W-\boldsymbol{S}_B)\boldsymbol{W}/2-\nu\rho+\sum_{i=1}^{m_1}\xi_i s_i/(v_1 m_1)+\sum_{j=m_1+1}^{N}\xi_j s_j/(v_2 m_2)+\sum_{i=1}^{m_1}\alpha_i(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+b-\xi_i)-\sum_{j=m_1+1}^{N}\alpha_j(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j+b-\rho+\xi_j)-\sum_{k=1}^{N}\beta_k\xi_k-\lambda\rho,\tag{4}$$

其中 $\alpha_i\geqslant 0,\beta_k\geqslant 0,\lambda\geqslant 0$ 为拉格朗日乘子.

分别对 $\boldsymbol{W},\rho,\xi,b$ 求偏导,并令偏导等于 0,有

$$\partial L/\partial\boldsymbol{W}=0\Leftrightarrow\boldsymbol{W}=-(\boldsymbol{S}_W-\boldsymbol{S}_B)^{-1}\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i,\tag{5}$$

$$\partial L/\partial\rho=-\nu+\sum_{j=m_1+1}^{N}\alpha_j-\lambda=0,\tag{6}$$

$$\partial L/\partial\xi_i=0\Rightarrow 0\leqslant\alpha_i\leqslant s_i/(\nu_1 m_1),\tag{7}$$

$$\partial L/\partial\xi_j=0\Rightarrow 0\leqslant\alpha_j\leqslant s_j/(\nu_2 m_2),\tag{8}$$

$$\partial L/\partial b=0\Leftrightarrow\sum_{i=1}^{N}\alpha_i y_i=0.\tag{9}$$

将式(5)~(9)代入到式(4)可得原优化问题的对偶形式.

2.2　重要参数求解

在优化问题中的参数 ρ 和 b 对于分类模型至关重要,其求解过程如下.

根据 KKT 条件,对于支持向量,式(2)变为一组

松弛因子为 0 的等式,有

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+b=0,1\leqslant i\leqslant m_1.$$

同理式(3)变为

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_j+b=\rho,m_1+1\leqslant j\leqslant N.$$

根据简单的数学变换,可得参数 ρ 和 b 的值:

$$\rho=\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{x}_j-\boldsymbol{x}_i), \tag{10}$$

$$b=-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i. \tag{11}$$

将 $\boldsymbol{W}=-(\boldsymbol{S}_W-\boldsymbol{S}_B)^{-1}\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i$ 代入式(10)和式(11),可得

$$\rho=-(\boldsymbol{S}_W-\boldsymbol{S}_B)^{-1}\left(\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_j-\sum_{i=1}^{m_1}\alpha_i y_i\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_i\right),$$

$$b=-(\boldsymbol{S}_W-\boldsymbol{S}_B)^{-1}\sum_{i=1}^{m_1}\alpha_i y_i\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_i.$$

2.3 决策函数

判断一个待测样本 $\boldsymbol{x}$ 的类属,需要通过比较其与分类超平面之间的距离.决策函数定义如下:

$$f(\boldsymbol{x})=\mathrm{sgn}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+b), \tag{12}$$

其中 sgn(·)为符号函数.

2.4 算法描述

MMLM 的算法流程如下.

输入:训练样本集 X_{Train}.

输出:在测试样本集 X_{Test} 中样本所属类别.

步骤 1 将实验数据集分为训练样本集 X_{Train} 和测试样本集 X_{Test};

步骤 2 利用拉格朗日乘子法将原优化问题转化为对偶形式;

步骤 3 在训练样本集 X_{Train} 上训练得到分类超平面的法向量 $\boldsymbol{W}$;

步骤 4 计算得到如式(12)所示的决策函数;

步骤 5 判断在测试样本集 X_{Test} 中样本的类属,并计算得到分类精度.

3 实验设计与分析

3.1 实验设计

实验的目的是通过与传统支持向量机 SVM 以及当前主流方法比较,如 APG_SVM[6]、SCH-SVM[7]、多层感知机(Multilayer Perceptron, MLP)、文献[19]所提模型,以验证所提方法 MMLM 的有效性.APG_SVM 融合邻传输思想和粒度计算理论,通过竞争机制更有效地提取重要的分类信息进行 SVM 训练.SCH-SVM 忽略样本在核空间几何轮廓内部上的噪声,在有效约减训练样本的同时能够保持较高的分类精度.MLP 由输入层、隐藏层、输出层组成,实验用到的隐藏层为单层结构,激活函数采用 ReLU(·).实验环境为 2.90 GHz Pentium CPU,2 G RAM,Redhat Enterprise Linux Server 6.0 及 matlab 2013a.选取在 UCI 开放数据集中的 liver、breast、glass、balance-scale、monks、heart 数据集作为实验数据.上述实验数据如表 1 所示.分别将 40%、60%、80%的实验数据集作为训练样本,其余部分用作测试.

表 1 实验数据集

实验数据集	样本总数	第 1 类样本数	第 2 类样本数	维度
liver	345	145	200	6
breast	569	212	357	30
glass	146	70	76	9
balance-scale	576	288	288	4
monks	432	216	216	6
heart	267	212	55	44

在实验中选用距离函数作为模糊隶属度函数,其定义如下:

$$s(\boldsymbol{x}_i)=1-\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|/R,$$

其中 $\bar{\boldsymbol{x}}$ 表示类中心,R 表示类半径且 $R=\max_i\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|$.

利用 10 折交叉验证和网格搜索法来确定实验中的参数.其中 MMLM 和文献[19]所提模型的参数 ν 在网格{1,5,10,15,20,25,30}中搜索,ν_1、ν_2 分别在网格{0.001,0.010,0,050,0.100,0.500}中搜索;在 SVM 中的参数 C 在网格{0.01,0.05,0.10,0.50,1.00,5.00,10.00}中搜索;在 APG_SVM 中的参数 C 在网格{1,5,10,50,100,500,1 000,5 000}中搜索,调和参数 p_a 在网格{0.1,0.2,0.3,0.4,0.5}中搜索;在 SCH-SVM 中的分组参数 P 和 V 分别在网格{10^4,5×10^4,10^5}和{10^3,2×10^3,3×10^3,4×10^3,5×10^5}中搜索.实验参数列入表 2,实验结果记录于表 3~表 5,其中表内数值为分类精度.

表 2　实验参数

实验数据集	SVM	APG_SVM	SCH-SVM	文献[19]模型	MMLM
liver	$C=0.01$	$C=1\ 000, p_a=0.2$	$P=1\times10^4, V=2\times10^3$	$\nu=5, \nu_1=0.10, \nu_2=0.001$	$\nu=5, \nu_1=0.10, \nu_2=0.01$
breast	$C=0.10$	$C=1\ 000, p_a=0.3$	$P=1\times10^4, V=3\times10^3$	$\nu=5, \nu_1=0.01, \nu_2=0.010$	$\nu=5, \nu_1=0.01, \nu_2=0.01$
glass	$C=0.05$	$C=500, p_a=0.1$	$P=1\times10^5, V=1\times10^3$	$\nu=15, \nu_1=0.05, \nu_2=0.010$	$\nu=20, \nu_1=0.10, \nu_2=0.01$
balance-scale	$C=0.50$	$C=500, p_a=0.2$	$P=5\times10^4, V=2\times10^3$	$\nu=10, \nu_1=0.10, \nu_2=0.500$	$\nu=5, \nu_1=0.01, \nu_2=0.10$
monks	$C=5.00$	$C=1\ 000, p_a=0.2$	$P=5\times10^4, V=4\times10^3$	$\nu=5, \nu_1=0.05, \nu_2=0.010$	$\nu=1, \nu_1=0.01, \nu_2=0.01$
heart	$C=1.00$	$C=1\ 000, p_a=0.1$	$P=1\times10^4, V=1\times10^3$	$\nu=1, \nu_1=0.50, \nu_2=0.500$	$\nu=1, \nu_1=0.10, \nu_2=0.10$

表 3　40%训练样本的实验结果

实验数据集	SVM	SCH-SVM	APG_SVM	文献[19]模型	MLP	MMLM
liver	0.608 7	0.715 0	0.768 1	0.758 5	0.584 5	0.801 9
breast	0.756 6	0.806 5	0.852 9	0.862 2	0.700 9	0.879 8
glass	0.715 9	0.784 1	0.806 8	0.818 2	0.724 1	0.852 3
balance-scale	0.780 3	0.861 3	0.878 6	0.878 6	0.780 3	0.890 2
monks	0.602 3	0.664 1	0.664 1	0.683 4	0.598 5	0.702 7
heart	0.687 5	0.768 8	0.787 5	0.818 8	0.700 0	0.837 5
平均精度	0.691 9	0.766 6	0.793 0	0.803 3	0.681 4	0.827 4

表 4　60%训练样本的实验结果

实验数据集	SVM	SCH-SVM	APG_SVM	文献[19]模型	MLP	MMLM
liver	0.681 6	0.768 1	0.811 6	0.787 8	0.644 9	0.862 3
breast	0.828 9	0.894 7	0.934 2	0.933 9	0.801 8	0.916 7
glass	0.775 9	0.844 8	0.879 3	0.879 3	0.793 1	0.896 6
balance-scale	0.839 1	0.917 4	0.947 8	0.947 8	0.813 0	0.947 8
monks	0.659 0	0.739 9	0.728 3	0.739 9	0.659 0	0.774 6
heart	0.738 3	0.803 7	0.831 8	0.850 5	0.717 0	0.878 7
平均精度	0.753 8	0.828 1	0.855 5	0.856 5	0.737 8	0.879 5

表 5　80%训练样本的实验结果

实验数据集	SVM	SCH-SVM	APG_SVM	文献[19]模型	MLP	MMLM
liver	0.710 1	0.782 6	0.833 3	0.822 5	0.710 1	0.840 6
breast	0.841 8	0.909 9	0.947 3	0.951 6	0.805 3	0.960 4
glass	0.794 8	0.871 8	0.888 9	0.906 0	0.827 6	0.923 1
balance-scale	0.846 0	0.930 6	0.958 8	0.961 0	0.834 8	0.974 0
monks	0.699 4	0.760 1	0.760 1	0.783 2	0.709 3	0.823 7
heart	0.752 3	0.836 4	0.859 8	0.873 8	0.735 8	0.897 2
平均精度	0.774 1	0.848 6	0.874 7	0.883 0	0.770 5	0.903 2

3.2　实验分析

由表 3~表 5 可以看出:从各模型的平均精度看,随着训练样本规模的增大,平均精度呈上升趋势.其主要原因是:经标注的训练样本蕴含了重要的

特征信息,通过学习更大规模的训练样本,能够更好地发现样本的潜在特征,有助于提高模型的分类精度.此外,各模型在不同数据集上的实验结果呈现一定的规律性,以如表 4 所示的实验结果为例,具体分析如下:除 breast 数据集外,MMLM 在 liver、glass、balance-scale、monks、heart 等数据集上均具有最优的分类精度.在 liver 数据集上,MMLM 分类性能最优,APG_SVM 次之,MLP 最差;MMLM 分类精度分别比 SVM、SCH-SVM、APG_SVM、MLP、文献[19]模型的分类精度高 0.180 7、0.094 2、0.050 7、0.217 4、0.074 5.在 glass 数据集上,MMLM 分类性能最优,APG_SVM 和文献[19]模型次之,MLP 最差;MMLM 分类精度分别比 SVM、SCH-SVM、APG_SVM、MLP、文献[19]模型的分类精度高 0.120 7、0.051 8、0.017 3、0.103 5、0.017 3.在 heart 数据集上,MMLM 分类性能最优,文献[19]模型次之,MLP 最差;MMLM 分类精度分别比 SVM、SCH-SVM、APG_SVM、MLP、文献[19]模型的分类精度高 0.140 4、0.075 0、0.046 9、0.161 7、0.028 2.在 balance-scale 数据集上,MMLM、APG_SVM、文献[19]模型分类性能均最优,达到 0.947 8,MLP 最差;它们的分类精度分别比 SVM、SCH-SVM、MLP 的分类精度高 0.108 7、0.030 4、0.134 8.在 monks 数据集上,MMLM 分类性能最优,文献[19]模型次之,SVM 和 MLP 最差;MMLM 分类精度分别比 SVM、SCH-SVM、APG_SVM、MLP、文献[19]模型的分类精度高 0.115 6、0.034 7、0.046 3、0.115 6、0.034 7.在 breast 数据集上,APG_SVM 分类性能最优,文献[19]模型次之,MLP 最差;MMLM 分类精度比 APG_SVM 和文献[19]模型的分类精度分别低 0.017 5、0.017 2.从平均精度看,MMLM 的平均分类性能最优,达到 0.879 5,分别比 SVM、SCH-SVM、APG_SVM、文献[19]模型、MLP 的分类精度高 0.125 7、0.051 4、0.024 0、0.023 0、0.141 7.MMLM 在大部分数据集上取得了最优的分类性能,其主要原因是:与 SVM 和 APG_SVM 相比,MMLM 引入了模糊隶属函数,通过为每个样本赋予不同权重来表征不同样本对分类结果的影响,从而具有更优的分类性能.MMLM、文献[19]模型、SCH-SVM、MLP 在分类决策时均考虑不同样本对分类结果的影响,但与文献[19]模型、SCH-SVM、MLP 相比,MMLM 引入类间离散度和类内离散度来表征训练样本的分布性状,这在一定程度上提高了分类精度.

4 总结

智能分类方法在实践中得到广泛应用,但已有方法往往忽略数据的分布性状,且没有考虑不同样本对分类结果的影响.鉴于此,受磁极效应启发,提出一种新颖的融合磁极效应和数据分布特征的最大间隔学习机.在该模型中,利用在线性判别分析中的类内离散度和类间离散度来表征数据的分布性状,利用模糊隶属度来表征不同样本对分类结果的影响.在 UCI 标准数据集上的比较实验表明,所提模型在分类精度方面具有一定优势.然而,该模型对参数较为依赖,一个更优的参数选择方法值得进一步研究,这将是笔者下一步的工作.

5 参考文献

[1] CORTES C,VAPNIK V.Support-vector networks [J].Machine Learning,1995,20(3):273-297.

[2] ZHANG Wei CHEN Junjie.Relief selection and parameter optimization for support vector machine based on mixed kernel function [J].International Journal of Performability Engineering,2018,14(2):280-289.

[3] DING Hu, XU Jinhui. Random gradient descent tree: a combinatorial approach for SVM with outliers [EB/OL]. [2022-06-17]. https://www. xueshufan. com/publication/2593094871.

[4] 马婷婷,杨志霞,叶俊佑.鲁棒双参数化间隔支持向量机 [J].计算机工程与应用,2022,58(9):74-82.

[5] 李建民,陈慧,杨冬芹,等.改进 GWO 优化 SVM 的服务器性能预测 [J].计算机工程与设计,2019,40(11):3099-3105,3163.

[6] 程凤伟,王文剑.基于近邻传输的粒度 SVM 算法 [J].计算机科学与探索,2020,14(7):1194-1199.

[7] TAX D M J,DUIN R P W.Support vector data description [J].Machine Learning,2004,54(1):45-66.

[8] NGUYEN P, TRAN D. Repulsive-SVDD classification [C]// Proceedings of the 19th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, May 19-22,2015,Ho Chi Minh City, Vietnam, Switzer-

land: Springer Cham,2015:277-288.
[9] KIM S,CHOI Y,LEE M.Deep learning with support vector data description [J].Neurocomputing,2015,165:111-117.
[10] 杨晨,王婕婷,李飞江,等.基于概率的支持向量数据描述方法 [J].计算机应用,2019,39(11):3134-3139.
[11] HAO Peiyi.A new fuzzy maximal-margin spherical-structured multi-class support vector machine [EB/OL]. [2022-06-17]. https://ieeexplore. ieee. org/document/6890475? denied=.
[12] 陈鹏,刘爽,左莉,等.基于数据分布规律的分段组合支持向量机研究 [J].微电子学与计算机,2015,32(3):94-99.
[13] 宋瑞阳,孟华,龙治国.基于数据分布特征的线性孪生支持向量机 [J].计算机科学,2019,46(S1):407-411.
[14] KHANJANI-SHIRAZ R, BABAPOUR-AZAR A, HOSSEINI-NODEH Z, et al. Distributionally robust joint chance-constrained support vector machines [J]. Optimization Letters,2023,17(2):299-332.
[15] BAHRAINI T, GHAZI S, YAZDI H S. Toward optimum fuzzy support vector machines using error distribution [J]. Engineering Applications of Artificial Intelligence, 2020, 90:103545.
[16] 顾晓清,倪彤光,姜志彬,等.面向大规模噪声数据的软性核凸包支持向量机 [J].电子学报,2018,46(2):347-357.
[17] 周裕群,张德生,张晓.一种改进的鲁棒模糊孪生支持向量机算法 [J].计算机工程与应用,2023,59(1):140-148.
[18] 戴小路,汪廷华,周慧颖.基于加权马氏距离的模糊多核支持向量机 [J].计算机科学,2022,49(S2):302-306.
[19] 刘忠宝,裴松年,杨秋翔.具有 N-S 磁极效应的最大间隔模糊分类器 [J].电子科技大学学报,2016,45(2):227-232,239.

The Maximum Margin Learning Machine Based on Magnetic Pole Effect and Data Distribution Characteristics

LIU Zhongbao[1,2,3], ZHANG Xingqin[1], WANG Wenli[1]

(1.School of Information Engineering,Shandong Vocational and Technical University of International Studies,Rizhao Shandong 276826,China;2.Institute of Language Intelligence,Beijing Language and Culture University,Beijing 100083,China;3.School of software,Quanzhou University of Information Engineering,Quanzhou Fujian 362000,China)

Abstract:The geometric boundary based classification method is one of typical classification methods.The existed methods often neglect the data distribution and influence of different samples to the classification result,therefore,their classification accuracies can't be greatly improved.In view of this,inspired by magnetic pole effect theory,a novel classification method named maximum margin learning machine based on magnetic pole effect and data distribution characteristics (MMLM) is proposed in this paper.In this model,the hyperplane is close to one class and far away from another.The within-class scatter and between-class scatter is introduced to describe the data distribution characteristics.Meanwhile,the fuzzy membership function is used to reflect the importance of different samples.The comparative experiments on the UCI standard datasets verify the effectiveness of the proposed method.

Key words:classification;magnetic pole effect;data distribution;within-class scatter;between-class scatter

(责任编辑:冉小晓)