

具有N-S磁极效应的最大间隔模糊分类器

刘忠宝, 裴松年, 杨秋翔

(中北大学计算机与控制工程学院 太原 030051)

【摘要】该文提出一种具有N-S磁极效应的最大间隔模糊分类器(MPMMFC)。该方法寻求一个具有N-S磁极效应的最优超平面,使得一类样本受磁极吸引离超平面尽可能近,另一类样本受磁极排斥离超平面尽可能远。针对传统支持向量机面临的对噪声和野点敏感问题,引入模糊技术来降低噪声和野点对分类的影响,从而进一步提高泛化性能和分类效率。通过人工数据集和实际数据集上的实验,证明了MPMMFC的有效性。

关键词 模糊技术; 核方法; 磁极效应; 模式分类

中图分类号 TP181 文献标志码 A doi:10.3969/j.issn.1001-0548.2016.03.012

Maximum Margin Fuzzy Classifier with N-S Magnetic Pole Effect

LIU Zhong-bao, PEI Song-nian, and YANG Qiu-xiang

(School of Computer and Control Engineering, North University of China Taiyuan 030051)

Abstract: Inspired by space geometry and magnetic pole effect theory, a maximum margin fuzzy classifier with N-S magnetic pole (MPMMFC) is proposed in this paper. The main idea is to find an optimal hyperplane based on N-S magnetic pole effect in order to ensure that the distance between one class and the hyperplane is much closer due to pole attractive and the distance between the other class and the hyperplane is much greater due to repulsion. Moreover, due to the traditional support vector machine (SVM) sensitive to noises and outliers, a fuzzy technology is introduced in this paper to reduce the influence of noises and outliers, and the classification efficiencies and generalization performance are improved further. Experimental results on the synthetic datasets and UCI datasets show that the proposed approaches are effective.

Key words fuzzy technology; kernel method; magnetic pole; pattern classification

模式分类是模式识别中的重要内容之一,其通过对有限训练样本的学习得到一个具有较低结构风险和良好泛化能力的分类器^[1]。在当前主流的模式分类方法中,支持向量机(SVM)^[2-4]及其相关变体受到人们的广泛关注,其通过最大化类间间隔实现有效分类。文献[2-4]提出C-SVM方法,该方法基于最大间隔思想在空间中寻找最优超平面将两类分开;文献[5]提出 ν -SVM方法,通过引入参数 ν 来控制支持向量个数的下界和训练误差的上界;文献[6]提出单类支持向量机(one class SVM, OCSVM)试图在高维特征空间中构建最大间隔超平面划分正常数据和噪声数据;文献[7]提出的支持向量数据描述(support vector data description, SVDD)试图在高维特征空间中计算寻求包含所有输入样本的最小包含球划分正常数据和噪声数据;针对SVM以及变体面对噪声

和野点的敏感问题,文献[8]提出模糊支持向量机(fuzzy SVM, FSVM),对不同的样本采用不同的合理的权重系数,在构造目标函数时削弱噪声和野点对分类的影响,从而在一定程度上提高分类性能。

近年来,很多分类新方法在进行分类决策时将类间间隔和类内分布性状考虑在内^[1,9]。文献[10]提出大间隔最小压缩包含球学习机(large margin and minimal reduced enclosing ball, LMMREB),该方法试图寻求两个同心压缩包含球实现类间间隔和类内内聚性的最大化并提高分类性能;文献[11]提出基于熵理论和核密度估计的最大间隔学习机(maximum margin learning machine based on entropy concept and kernel density estimation, MLMEK),MLMEK引入熵和核密度表征分类不确定性和样本分布特征实现分类;文献[12]综合最小包含球和最大间隔思想,提出

收稿日期: 2014-10-08; 修回日期: 2015-03-30

基金项目: 国家社科基金后期资助项目(15FTQ008)

作者简介: 刘忠宝(1981-),男,博士后,副教授,主要从事机器学习方面的研究。

一种用于新奇检测的小球体和大间隔方法(small sphere large margin, SSLM), SSLM在高维特征空间中构建最小包含球包围正常样本实现分类;文献[13]提出一种模糊最大间隔球形结构多类支持向量机(fuzzy maximal-margin spherical-structured multi-class SVM, MSM-SVM), MSM-SVM试图构造正负类间隔最大正类体积最小超球体实现分类。

受以上方法启发,本文在N-S磁极效应理论基础上,结合传统SVM的大间隔思想,提出一种具有N-S磁极效应的最大间隔模糊分类器MPMMFC, MPMMFC在构建最优决策面时,引入模糊性惩罚参数,降低噪声和野点数据对决策面的影响,进一步提高泛化性能。

对本文后续做以下规定:对于一个包含 N 个模式的二分类问题,给定训练样本集合 $T = \{(x_1, y_1, s_1), (x_2, y_2, s_2), \dots, (x_N, y_N, s_N)\}$ 。其中: $x_i \in R^d (i=1, 2, \dots, N)$ 为输入数据集; $y_i \in \{1, -1\}$ 为类标签,当 $1 \leq i \leq m_1 (m_1 + m_2 = N)$ 时, $y_i = 1$;当 $m_1 + 1 \leq i \leq m_1 + m_2$ 时, $y_i = -1$;且 $s_i (1 \leq i \leq N)$ 为模糊隶属度, $\sigma \leq s_i \leq 1$, σ 为任意小的一个正数。 $y_i = 1$ 含有 m_1 个模式, $y_i = -1$ 含有 m_2 个模式。

1 背景知识

1.1 N-S磁极效应

磁体上磁性最强的部分叫磁极。磁体周围存在磁场,磁体间的相互作用就是以磁场作为媒介的。一个磁体无论多么小都有两个磁极,可以在水平面内自由转动的磁体,静止时指向南方的磁极叫南极(S极),指向北方的磁极叫做北极(N极)。之间呈现同性磁极相互排斥、异性磁极相互吸引的现象。

1.2 支持向量机

支持向量机是一种基于统计学习理论的机器学习方法,核心思想是将结构风险最小化原则引入到分类中,在属性空间中构建最优分类超平面,将两类样本尽可能的分开^[2-3]。

设超平面方程为 $y = \omega^T x + b$,分类间隔为 $2/\|\omega\|$,则最优化问题可描述为如下两种形式。

1) 线性形式

$$\min_{\omega, b, \xi_i} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^N \xi_i \quad (1)$$

$$\text{s.t. } y_i(\omega^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i=1, 2, \dots, N$$

式中, C 为惩罚参数,控制对错分样本的惩罚程度;松弛变量 ξ_i 允许错分样本的存在,一定程度上提高了算法的泛化能力。

2) 非线性形式:

$$\min_{\omega, b, \xi_i} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^N \xi_i \quad (2)$$

$$\text{s.t. } y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i=1, 2, \dots, N$$

式中, $\phi(x_i)$ 为从原始样本空间到高维特征空间的映射。

由Lagrangian定理^[13]将原始问题转化为如下对偶形式:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N \alpha_i \quad (3)$$

$$\text{s.t. } \sum_{i=1}^N \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i=1, 2, \dots, N$$

2 MPMMFC

从物理学角度,MPMMFC可理解为在空间中寻找一个具有磁性的“磁极”分别对两类样本作用,根据样本的磁性不同对两类样本进行分类;从几何角度,MPMMFC可理解为在空间中寻找一个分类超平面,通过计算样本与超平面的关系判断样本类属。

2.1 线性形式

基于上述分析,MPMMFC目标是在样本空间中试图构建一个超平面,使得一类模式离超平面尽可能的近,另一类模式离超平面尽可能的远,该优化问题可描述为如下最优化形式:

$$\min_{\omega, \rho, \xi_i, b} \frac{1}{2} \omega^T \omega - \nu \rho + \frac{1}{\nu_1 m_1} \sum_{i=1}^{m_1} \xi_i s_i + \frac{1}{\nu_2 m_2} \sum_{j=m_1+1}^N \xi_j s_j \quad (4)$$

$$\text{s.t. } \omega^T x_i + b \leq \xi_i, 1 \leq i \leq m_1$$

$$\omega^T x_j + b \geq \rho - \xi_j, m_1 + 1 \leq j \leq N$$

$$\xi \geq 0, \rho \geq 0, \sigma \leq s \leq 1$$

式中, $s_i (1 \leq i \leq N)$ 为模糊隶属度; σ 为任意小的一个正数; ρ 为两类样本间隔;用 $\xi_i s_i$ 代替松弛因子 ξ_i ,使不同样本点在分类时起到不同的作用; ν, ν_1, ν_2 为3个正常数; m_1 和 m_2 分别为两类样本数。

上述最优化问题中 $\frac{1}{2} \omega^T \omega$ 可以确定MPMMFC最优分类面的法向量; $\nu \rho$ 表示两类间隔;

$\frac{1}{\nu_1 m_1} \sum_{i=1}^{m_1} \xi_i s_i$ 和 $\frac{1}{\nu_2 m_2} \sum_{j=m_1+1}^N \xi_j s_j$ 分别表示具有模糊特性的松弛因子,其中模糊特性通过模糊隶属度函数体现,该模糊特性将不同样本区别对待,松弛因子保证算法具有一定的容错性。

MPMMFC借鉴N-S磁极效应思想进行分类。从N-S磁极效应角度看,若将分类超平面看作磁极,则

其对第一类吸引, 而对第二类排斥。具体而言, 在上述优化问题中, 约束条件式(4)分别表示两类样本受到磁场作用而产生的不同反应, 即第一类样本距离分类超平面近, 而第二类样本远离分类超平面, ρ 保证两类样本具有良好的可分性。

上述最优化问题的对偶形式为:

$$\begin{aligned} \min_{\alpha \in R^d} & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j^T & (5) \\ \text{s.t.} & 0 \leq \alpha_i \leq \frac{s_i}{v_1 m_1}, 1 \leq i \leq m_1 \\ & 0 \leq \alpha_j \leq \frac{s_j}{v_2 m_2}, m_1 + 1 \leq j \leq N \\ & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 2v \leq \sum_{i=1}^N \alpha_i \end{aligned}$$

证明 根据Lagrangian定理, 上述MPMMFC原始问题的Lagrangian方程为:

$$\begin{aligned} L(\boldsymbol{\omega}, \rho, \xi, b, \alpha, \beta, \lambda) = & \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} - v\rho + \frac{1}{v_1 m_1} \sum_{i=1}^{m_1} \xi_i s_i + \\ & \frac{1}{v_2 m_2} \sum_{j=m_1+1}^N \xi_j s_j + \sum_{i=1}^{m_1} \alpha_i (\boldsymbol{\omega}^T \mathbf{x}_i + b - \xi_i) - \\ & \sum_{j=m_1+1}^N \alpha_j (\boldsymbol{\omega}^T \mathbf{x}_j + b - \rho + \xi_j) - \sum_{k=1}^N \beta_k \xi_k - \lambda \rho \end{aligned} \quad (6)$$

式中, $\alpha_i \geq 0, \beta_k \geq 0, \lambda \geq 0$, 它们分别为Lagrangian乘子, 在 $L(\boldsymbol{\omega}, \rho, \xi, b, \alpha, \beta, \lambda)$ 方程中, 分别对原始变量 $\boldsymbol{\omega}, \rho, \xi, b$ 求偏导并令各偏导方程为0, 可得:

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \Leftrightarrow \boldsymbol{\omega} = - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (7)$$

$$\frac{\partial L}{\partial \rho} = -v + \sum_{j=m_1+1}^N \alpha_j - \lambda = 0 \quad (8)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow 0 \leq \alpha_i \leq \frac{s_i}{v_1 m_1} \quad (9)$$

$$\frac{\partial L}{\partial \xi_j} = 0 \Rightarrow 0 \leq \alpha_j \leq \frac{s_j}{v_2 m_2} \quad (10)$$

$$\frac{\partial L}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^N \alpha_i y_i = 0 \quad (11)$$

将式(7)和式(11)带入到式(6), 可得MPMMFC的对偶形式。

2.2 非线性形式

在非线形情况下, 通过满足Mercer条件的核函数对输入空间进行高维映射, 然后在高维特征空间

中进行模式分类。MPMMFC的核化形式为:

$$\min_{\boldsymbol{\omega}, \rho, \xi, b} \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} - v\rho + \frac{1}{v_1 m_1} \sum_{i=1}^{m_1} \xi_i s_i + \frac{1}{v_2 m_2} \sum_{j=m_1+1}^N \xi_j s_j \quad (12)$$

$$\begin{aligned} \text{s.t.} & \boldsymbol{\omega}^T \phi(\mathbf{x}_i) + b \leq \xi_i, 1 \leq i \leq m_1 \\ & \boldsymbol{\omega}^T \phi(\mathbf{x}_j) + b \geq \rho - \xi_j, m_1 + 1 \leq j \leq N \end{aligned} \quad (13)$$

$$\xi \geq 0, \rho \geq 0, \sigma \leq s \leq 1$$

核化对偶形式为:

$$\min_{\alpha \in R^d} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (14)$$

$$\text{s.t.} 0 \leq \alpha_i \leq \frac{s_i}{v_1 m_1}, 1 \leq i \leq m_1$$

$$0 \leq \alpha_j \leq \frac{s_j}{v_2 m_2}, m_1 + 1 \leq j \leq N \quad (15)$$

$$\sum_{i=1}^N \alpha_i y_i = 0$$

$$2v \leq \sum_{i=1}^N \alpha_i$$

$$\boldsymbol{\omega} = - \sum_{i=1}^N \alpha_i y_i \phi(\mathbf{x}_i)$$

式中, $K(\mu, \nu)$ 为符合Mercer条件的核函数。

2.3 最大间隔的求解方法

在求解完式(12)~式(15)的QP问题后, 考虑两类支持向量集合和 α 集合:

$$SVX_1 = \left\{ \mathbf{x}_i \mid 0 < \alpha_i < \frac{s_i}{v_1 m_1}, 1 \leq i \leq m_1 \right\} \quad (16)$$

$$SVX_2 = \left\{ \mathbf{x}_j \mid 0 < \alpha_j < \frac{s_j}{v_2 m_2}, 1 \leq j \leq N \right\} \quad (17)$$

$$SV\alpha_1 = \left\{ \alpha_i \mid 0 < \alpha_i < \frac{s_i}{v_1 m_1}, 1 \leq i \leq m_1 \right\} \quad (18)$$

$$SV\alpha_2 = \left\{ \alpha_j \mid 0 < \alpha_j < \frac{s_j}{v_2 m_2}, m_1 + 1 \leq j \leq N \right\} \quad (19)$$

根据KKT条件, 对于 SVX_1 , 式(13)变成了一个所有松弛因子为0的等式, 即:

$$\boldsymbol{\omega}^T \phi(\mathbf{x}_i) + b = 0 \quad 1 \leq i \leq m_1 \quad (20)$$

同理对于 SVX_2 , 式(15)也变成了一个等式, 即:

$$\boldsymbol{\omega}^T \phi(\mathbf{x}_j) + b = \rho \quad m_1 + 1 \leq j \leq N \quad (21)$$

因此设 $n_1 = |SVX_1|$, $n_2 = SVX_2$, 由式(20)和式(21)得:

$$\rho^* = \frac{1}{n_2} P_2 - \frac{1}{n_1} P_1 \quad (22)$$

$$b^* = - \frac{1}{n_1} P_1 \quad (23)$$

其中:

$$P_1 = \sum_{\alpha_i \in SV_{\alpha_1}} \sum_{x_i, x_j \in SV_{X_1}} \alpha_i K(x_i, x_j)$$

$$P_2 = \sum_{\alpha_j \in SV_{\alpha_2}} \sum_{x_i, x_j \in SV_{X_2}} \alpha_j K(x_i, x_j)$$

2.4 判别函数

为了判别一个新模式 $\mathbf{x}_i \in R^d$ 的类属,MPMMFC 通过比较该模式与构造出的超平面的距离是否小于 0 来确定该模式的类属。MPMMFC 决策函数为:

$$f = \text{sgn}(\omega^* \phi(\mathbf{x}) + b^*) \quad (24)$$

求得 α 后,可通过式(24)得到 ω^*, b^* 。

3 理论分析

3.1 算法复杂度分析

MPMMFC 解决一个具有线性约束的二次规划问题,其计算对象主要是核函数矩阵,空间复杂度是 $O(N^2)$,其中 N 为训练样本数;其时间复杂度为 $O(N^3)$ 。当面对大规模分类问题时,MPMMFC 的训练时间随着样本数的增加呈指数级增长。因此 MPMMFC 不适用大规模分类问题。目前,一个新的研究成果引起人们的广泛关注:文献[16]提出的核心集向量机(core vector machine, CVM)试图建立最优化问题与最小包含球QP形式的等价性,从而将分类方法的适用范围从中小规模数据推广到大规模数据。限于本文篇幅,下一步的工作是探讨MPMMFC 与最小包含球QP形式的等价性,从而解决MPMMFC 无法进行大规模分类的问题。

3.2 可调参数 ν 性质

本文称对应Lagrangian乘子 $\alpha_i > 0$ 的训练样本 $\mathbf{x}_i (1 \leq i \leq N)$ 为支持向量(SV),对应的松弛变量 $\xi_i > 0 (1 \leq i \leq N)$ 的训练样本 $\mathbf{x}_i (1 \leq i \leq N)$ 为间隔误差(ME)。

定理 1 设:

$$ME_1 = \{s_i | \xi_i > 0, 1 \leq i \leq m_1\} \quad (25)$$

$$ME_2 = \{s_j | \xi_j > 0, m_1 + 1 \leq j \leq N\} \quad (26)$$

$$S_1 = \{s_i | a_i > 0, 1 \leq i \leq m_1\} \quad (27)$$

$$S_2 = \{s_j | a_j > 0, m_1 + 1 \leq j \leq N\} \quad (28)$$

得到如下关系:

$$\frac{1}{m_1} \sum_{s_i \in ME_1} s_i \leq \nu v_1 \leq \frac{1}{m_1} \sum_{s_i \in S_1} s_i \quad (29)$$

$$\frac{1}{m_2} \sum_{s_j \in ME_2} s_j \leq \nu v_2 \leq \frac{1}{m_2} \sum_{s_j \in S_2} s_j \quad (30)$$

ν, ν_1, ν_2 同时控制着支持向量的下界和间隔误差的上界。

证明 由于 $\sigma \leq s_i \leq 1$, 则式(14)约束条件变为:

$$0 \leq \alpha_i \leq \frac{s_i}{\nu_1 m_1} \leq \frac{1}{\nu_1 m_1}, 1 \leq i \leq m_1 \quad (31)$$

根据Kuhn-Tucher定理,对偶变量与约束的乘积在鞍点处为0,即 $\lambda \rho = 0$, 当 $\rho > 0$, 由式(8)得:

$$-v + \sum_{j=m_1+1}^N \alpha_j = 0 \Rightarrow \sum_{j=m_1+1}^N \alpha_j = v \quad (32)$$

由式(13)和式(26)得 $\sum_{i=1}^{m_1} \alpha_i = v$, 当 $\xi_i > 0$ 时,

$a_i = \frac{s_i}{\nu_1 m_1}$ 对于所有正类间隔误差成立,则有:

$$\sum_{i=1}^{m_1} \alpha_i = v \geq \frac{1}{\nu_1 m_1} \sum_{s_i \in ME_1} s_i \quad (33)$$

由式(26)知,每个正支持向量至多贡献 $\frac{s_i}{\nu_1 m_1}$, 则得:

$$\sum_{i=1}^{m_1} \alpha_i = v \leq \frac{1}{\nu_1 m_1} \sum_{s_i \in S_1} s_i \Rightarrow v \leq \frac{1}{\nu_1 m_1} \sum_{s_i \in S_1} s_i \quad (34)$$

由式(28)和式(29)可得 $\frac{1}{m_1} \sum_{s_i \in ME_1} s_i \leq \nu v_1 \leq \frac{1}{m_1} \sum_{s_i \in S_1} s_i$,

对于负类,同理可得式(25)。

4 实验结果与分析

实验的目的是验证MPMMFC和C-SVM、 ν -SVM、OCSVM在UCI数据集上的有效性,实验环境为2.90 GHz Pentium CPU, 2 G RAM, Redhat Enterprise Linux Server 6.0 及matlab2013a。实验选取的核函数为高斯核函数(RBF):

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right) \quad (35)$$

式中, σ 为训练样本平均范数的平方根。

目前,隶属度函数的构造方法有很多,一般模糊隶属度函数主要有两种:一种是基于样本到类中心的距离来度量模糊隶属度的大小;另一种是通过密度来度量模糊隶属度的大小。本文采用文献[14]提出的基于K近邻法思想的模糊隶属度函数,通过紧密度来度量模糊隶属度大小的方法。

定义数据点与点之间的距离为:

$$d_{ij} = |\mathbf{x}_i - \mathbf{x}_j| \quad i, j \in l, i \neq j$$

$$d_{i1} \leq d_{i2} \leq \dots \leq d_{i(l-1)} \quad (36)$$

紧密度的隶属度定义为:

$$\begin{cases} b_i = 1 / \sum_{j=1}^k d_{ij} \\ B = \max(b_1, b_2, \dots, b_l) \\ s_i = b_i / B \end{cases} \quad (37)$$

实验数据集包括一个人工数据集以及11个UCI标准数据集, 其中 m_1 代表第一类样本数, m_2 代表第二类样本数。数据集详细信息如表1所示。

表1 实验中所采用的UCI数据集

数据集	样本总数	m_1	m_2	维数
banana	200	100	100	2
blood	748	178	570	4
breast	569	212	357	30
iris	100	50	50	4
liver	345	145	200	6
seeds	210	70	140	7
glass	146	70	76	9
balance-scale	576	288	288	4
monks	432	216	216	6
spectf	269	214	55	44
heart	267	212	55	44
pima	768	500	268	8

4.1 实验参数设置

本文所提算法MPMMFC、C-SVM、 ν -SVM、OCSVM和SSLM的分类精度和参数选择密切相关, 目前参数选择的方法主要有: 单一验证估计、留一法和K倍交叉验证法等, 本文采取5倍交叉验证法。

实验中所有参数的选择通过网格搜索策略来选取。对于核函数(高斯径向基), σ^2 在网格 $\{\sigma^2/8, \sigma^2/4, \sigma^2/2, \sigma^2, 2\sigma^2, 4\sigma^2, 8\sigma^2\}$ 中搜索选取。对于C-SVM, 惩罚参数C在 $\{0.01, 0.03, 0.05, 0.08, 0.1, 0.1, 0.5, 1, 5, 10\}$ 中搜索选取, 对于 ν -SVM, 参数 ν 在 $\{0.01k, 0.1k\}$ 中搜索选取, k 为1~9之间的整数; 对于OCSVM, 参数 ν 在网格 $\{0.001, 0.002, 0.004, 0.008, 0.1, 0.2, 0.4, 0.8, 0.9, 1\}$ 中搜索选取的; 对于SSLM, 参数 ν 在网格 $\{5, 10, 20, 30, 40, 50, 60, 70, 80\}$ 中选取, ν_1 和 ν_2 在 $\{0.001, 0.01\}$ 中搜索; 对于MPMMFC, 根据参数定理, 参数 ν 在网格 $\{1, 3, 5, 8, 10, 13, 15, 20, 25, 30, 35, 40, 45, 50, 60, 80\}$ 中搜索, ν_1 和 ν_2 在 $\{0.001, 0.01, 0.1\}$ 中搜索, k 在网格 $\{0, 1, 2, 3, 4, 5\}$ 中选择。

4.2 实验结论

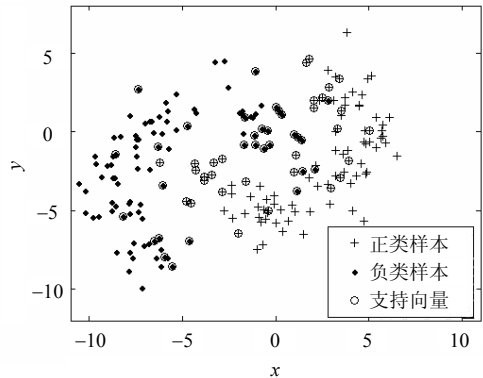
通过执行5倍交叉验证来搜索优化参数值, 并采用g-means度量来评价性能, 所有实验独立执行10次, 实验结论取平均值, 采用几何度量方法 $g = \sqrt{a^+ a^-}$ 评价算法性能, 其中 a^+ 和 a^- 分别为正类和负类的分类精度。该方法同时考虑了正类和负类的分类效果而

被广泛用于处理不平衡数据集问题。

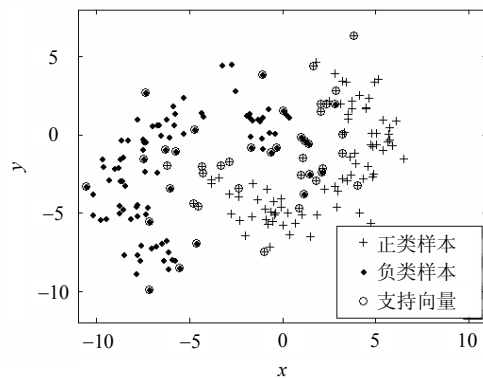
4.2.1 人工数据集

首先采用人工数据集——banana数据集来比较MPMMFC算法和C-SVM、 ν -SVM的性能优劣。实验参数及实验结果如图1所示, 图中横纵坐标为二维空间的xy轴坐标。

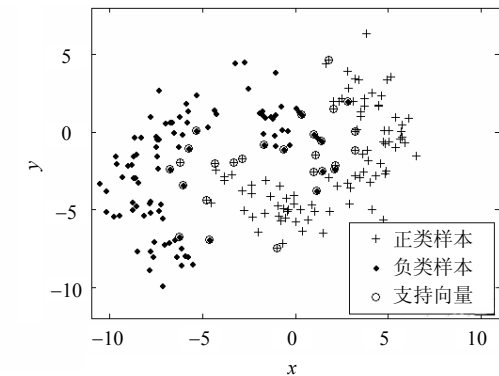
从图1a~1c可以看出, MPMMFC在人工香蕉型数据集上的支持向量数量相对于C-SVM、 ν -SVM要少, 而且在分类性能上也有较高的准确率。



a. C-SVM实验: $\sigma^2 = \sigma^2/4, C=0.5$, 准确率: 96.49%



b. ν -SVM实验: $\sigma^2 = \sigma^2/4, \nu=0.01$, 准确率98.03%



c. MPMMFC实验: $\sigma^2 = \sigma^2/6, \nu=1, \nu_1=0.006, \nu_2=0.1$, 准确率: 100%

图1 香蕉型数据集模式分类结果

4.2.2 UCI数据集

通过UCI数据集来评价MPMMFC和C-SVM、 ν -SVM、OCSVM及SSLM的性能。一类模式和二类

模式最优实验参数、实验结果分别如表2和表3所示。

表2 二类模式的分类结果

数据集	算法		
	C-SVM	ν -SVM	MPMMFC
blood	$C=0.5,$ $\sigma^2=\sigma^2*8$	$\nu=0.01,$ $\sigma^2=\sigma^2*8$	$\nu=3, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2*2$
	98.13±1.47%	99.22±0.21%	98.79±0.28%
breast	$C=0.1, \sigma^2=\sigma^2/8$	$\nu=0.03,$ $\sigma^2=\sigma^2/4$	$\nu=5, \nu_1=0.01,$ $\nu_2=0.01, \sigma^2=\sigma^2/8$
	89.44±2.24%	90.03±2.17%	92.33±2.72%
iris	$C=0.01,$ $\sigma^2=\sigma^2*8$	$\nu=0.04,$ $\sigma^2=\sigma^2*8$	$\nu=15, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2*4$
	97.24±1.30%	98.21±0.52%	98.56±1.45%
liver	$C=0.01,$ $\sigma^2=\sigma^2/8$	$\nu=0.5,$ $\sigma^2=\sigma^2/8$	$\nu=5, \nu_1=0.1,$ $\nu_2=0.001, \sigma^2=\sigma^2/2$
	65.94±2.96%	72.04±2.71%	74.90±3.88%
seeds	$C=0.5,$ $\sigma^2=\sigma^2*4$	$\nu=0.01,$ $\sigma^2=\sigma^2*8$	$\nu=1, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2*4$
	92.67±1.93%	96.39±1.47%	96.24±1.31%
glass	$C=0.05,$ $\sigma^2=\sigma^2*8$	$\nu=0.9,$ $\sigma^2=\sigma^2*8$	$\nu=20, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2*4$
	84.45±1.07%	85.15±0.90%	87.37±1.21%
balance-scale	$C=0.5, \sigma^2=\sigma^2/4$	$\nu=0.1,$ $\sigma^2=\sigma^2/8$	$\nu=5, \nu_1=0.01,$ $\nu_2=0.1, \sigma^2=\sigma^2/8$
	92.97±0.40%	94.67±0.55%	96.75±0.51%
monks	$C=5, \sigma^2=\sigma^2/2$	$\nu=0.1,$ $\sigma^2=\sigma^2/2$	$\nu=1, \nu_1=0.001,$ $\nu_2=0.001, \sigma^2=\sigma^2/2$
	66.81±0.42%	65.50±0.14%	73.80±0.35%

表3 一类模式的分类结果

数据集	算法		
	OCSVM	SSLM	MPMMFC
spectf	$\nu=0.8,$ $\sigma^2=\sigma^2*2$	$\nu=10, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2*4$	$\nu=1, \nu_1=0.1,$ $\nu_2=0.001, \sigma^2=\sigma^2*4$
	72.97±2.81%	78.12±1.96%	78.66±2.19%
heart	$\nu=0.01,$ $\sigma^2=\sigma^2*8$	$\nu=10, \nu_1=0.1,$ $\nu_2=0.001, \sigma^2=\sigma^2*4$	$\nu=3, \nu_1=0.01,$ $\nu_2=0.01, \sigma^2=\sigma^2*4$
	78.99±4.03%	80.34±4.16%	82.24±3.92%
pima	$\nu=1,$ $\sigma^2=\sigma^2/8$	$\nu=30, \nu_1=0.1,$ $\nu_2=0.01, \sigma^2=\sigma^2/4$	$\nu=25, \nu_1=0.01,$ $\nu_2=0.1, \sigma^2=\sigma^2/4$
	67.15±0.97%	67.27±1.25%	67.68±1.34%

从表2和表3可以看出,和其他几种方法(C-SVM, ν -SVM, OCSVM, SSLM)相比,MPMMFC在二类分类和一类分类上都取得了较好或相近的性能,在breast、liver、glass、balance-scale、monks、spectf、heart等数据集上,MPMMFC相对于传统的算法具有明显的性能优势;而在iris、pima等数据集上,MPMMFC和传统分类方法分类精度基本相当;在blood、seeds数据集上,MPMMFC的分类性能略逊于传统算法 ν -SVM,但分类精度基本可以接受。综上所述,MPMMFC在核函数映射后的高维空间,通过模糊隶属度给不同的样本增加不同的权重系

数,使得构造的超平面不仅能实现二类模式分类,而且还能解决单类模式分类问题。

5 结束语

受大间隔思想和N-S磁极效应理论,本文提出具有N-S磁极效应的最大间隔模糊分类器。该方法借鉴N-S磁极效应理论,在样本空间中寻找一个最优超平面,使得一类样本受磁极吸引离超平面尽可能近,另一类样本受磁极排斥离样本尽可能远,而且通过引入模糊技术,根据不同样本的贡献赋予不同的权重进一步提高算法的分类性能。人工数据集和UCI数据集实验结果表明,与传统分类方法C-SVM、 ν -SVM、SSLM、OCSVM相比,MPMMFC在解决二分类以及单类问题上具有一定的优势。

参考文献

- [1] KOBAYASHI C, MEHRYAR M, FEMANDO P. Gaussian margin machines[C]//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater Beach Florida: Journal of Machine Learning Research, 2009, 5: 105-112.
- [2] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.
- [3] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 95-135.
- [4] LI Hang. Statistical learning methods[M]. Beijing: Tsinghua University Press, 2012: 95-135.
- [5] 邓乃扬, 田英杰. 支持向量机——理论、算法与拓展[M]. 北京: 科学出版社, 2009.
- [6] DENG Nai-yang, TIAN Ying-jie. Support vector machine-theory, algorithms and extension[M]. Beijing: Science Press, 2009.
- [7] SCHOLKOPF B, SMOLA A, BARTLET P. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207-1245.
- [8] SCHOLKOPF B, SMOLA A. Learning with kernels[M]. Cambridge: MIT, 2002.
- [9] TAX D M J, DUIN R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.
- [10] LIN C F, WAN S D. Fuzzy support vector machine[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.
- [11] SHIVASWAMY P K, JEBARA T. Maximum relative margin and data-dependent regularization[J]. Journal of Machine Learning Research, 2010, 11(2): 747-788.
- [12] 陶剑文, 王士同. 大间隔最小压缩包含球学习机[J]. 软件学报, 2012, 23(6): 1458-1471.
- [13] TAO Jian-wen, WANG Shi-tong. Large margin and minimal reduced enclosing ball learning machine[J]. Journal of Software, 2012, 23(3): 1458-1471.
- [14] 刘忠宝, 王士同. 基于熵理论和核密度估计的最大间隔学习机[J]. 电子与信息学报, 2011, 33(9): 2187-2194.

(下转第239页)

- Proceedings of the 4th International Symposium on Information Processing in Sensor Networks. [S.l.]: IEEE, 2005.
- [10] 王瑞锦. 复杂环境下的无线传感器网络定位关键技术研究[D]. 成都: 电子科技大学, 2013.
WANG Rui-jin. Key techniques of localization in wireless sensor networks under complex environment[D]. Chengdu: University of Electronic Science and Technology of China, 2013.
- [11] ZHANG W, CAO G. DCTC: Dynamic convoy tree-based collaboration for target tracking in sensor networks[J]. IEEE Transactions on Wireless Communications, 2004, 3(5): 1689-1701.
- [12] JI X, ZHA H, METZNER J J, et al. Dynamic cluster structure for object detection and tracking in wireless ad-hoc sensor networks[C]//IEEE International Conference on Communications. [S.l.]: IEEE, 2004, 7: 3807-3811.
- [13] 鲍微. 无线传感器网络中能量高效的目标跟踪问题研究[D]. 合肥: 中国科学技术大学, 2009.
BAO Wei. Research on energy-efficient target tracking in WSNs[D]. Hefei: University of Science and Technology of China, 2009.
- [14] ABRAHAM I, DOLEV D, MALKHI D. LLS: a locality aware location service for mobile Ad hoc networks[C]// Proceedings of the 2004 Joint Workshop on Foundations of Mobile Computing. [S.l.]: ACM, 2004.
- [15] SARKAR R, GAO J. Differential forms for target tracking and aggregate queries in distributed networks[J]. IEEE/ACM Transactions on Networking, 2013, 21(4): 1159-1172.
- [16] SARKAR R, YIN X, GAO J, et al. Greedy routing with guaranteed delivery using ricci flows[C]//International Conference on Information Processing in Sensor Networks, IPSN 2009. [S.l.]: IEEE, 2009: 121-132.

编辑 叶芳

(上接第232页)

- LIU Zhong-bao, WANG Shi-tong. A maximum margin learning machine based on entropy concept and kernel density estimation[J]. Journal of Electronics & Information Technology, 2011, 33(9): 2187-2194.
- [12] WU Ming-rui, YE Jie-ping. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.
- [13] HAO P Y. A new fuzzy maximal-margin spherical-structured multi-class support vector machine[C]// Proceedings of the 2013 International Conference on Machine Learning and Cybernetics. Tianjin: Machine Learning and Cybernetics (ICMLC), 2013, 1: 241-246.
- [14] 张祥, 徐光佑, 肖小玲. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报, 2006, 17(5): 951-958.
ZHANG Xiang, XU Guang-you, XIAO Xiao-ling. Fuzzy support vector machine based on affinity among samples[J]. Journal of Software, 2006, 17(5): 951-958.
- [15] QIN Chuan-dong, LIU San-yang. Fuzzy smooth support vector machine with different smooth functions[J]. Journal of Systems Engineering and Electronics, 2012, 23(3): 460-466.
- [16] TSANG I W, KWOK J T, CHEUNG P M. Core vector machines: Fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005(6): 363-392.

编辑 叶芳