ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 431－440
DOI:10. 16337/j. 1004-9037. 2020. 03. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合数据分布特征的保序学习机

刘忠宝，张志剑，党建飞

(中北大学软件学院,太原,030051)

摘　要：支持向量机(Support vector machine,SVM)作为一种经典的分类方法,已经广泛应用于各种领域中。然而,标准支持向量机在分类决策中面临以下问题:(1)未考虑分类数据的分布特征;(2)忽略了样本类别间的相对关系;(3)无法解决大规模分类问题。鉴于此,提出融合数据分布特征的保序学习机(Rank preservation learning machine based on data distribution fusion, RPLM-DDF)。该方法通过引入类内离散度表征数据的分布特征;通过各类样本数据中心位置相对不变保证全局样本顺序不变;通过建立所提方法和核心向量机对偶形式的等价性解决了大规模分类问题。在人工数据集、中小规模数据集和大规模数据集上的比较实验验证所提方法的有效性。

关键词：类内离散度;支持向量机;大规模数据集;全局保序;核心向量机

中图分类号：TP181　　**文献标志码**:A

Rank Preservation Learning Machine Based on Data Distribution Fusion

LIU Zhongbao, ZHANG Zhijian, DANG Jianfei

(School of Software, North University of China, Taiyuan, 030051, China)

Abstract: As a typical classification method, support vector machine (SVM) has been widely used in various fields. However, the standard SVM faces the following problems in the classification decision: First, it does not consider the distribution characteristics of the classification data; Second, it ignores the relative relationship between sample categories; Third, it can not solve the problem of large-scale classification. In view of this, the rank preservation learning machine based on data distribution fusion (RPLM-DDF) is proposed, in which within-class scatter is introduced to describe the distribution properties, and through the relatively constant position of all kinds of sample data centers, the global sample order remains unchanged. The large-scale classification problem is solved by certifying RPLM-DDF and the duality of the core vector machine. The comparison experiments on the artificial datasets, small-scale datasets and large-scale datasets verity the effectiveness of the RPLM-DDF.

Key words: within-class scatter; support vector machine (SVM); large-scale labeled datasets; global rank preservation; core vector machine(CVM)

引　　言

支持向量机(Support vector machine,SVM)由 Vapnik 和 Corinna 最早提出[1],已经广泛应用于机器

基金项目:国家社会科学基金(19BTQ012)资助项目。
收稿日期:2019-10-10;**修订日期**:2019-11-08

学习、数据挖掘和模式识别等领域，在解决小样本、非线性和高维度的模式识别中表现为速度快、精度高和理论支持清晰等优点[2]。SVM是基于结构风险最小化理论，通过寻找一个最优的超平面得到全局最优解。在SVM提出后，众多学者相继提出了SVM改进算法：范昕炜等提出加权支持向量机（Weighted support vector machine，WSVM）[3]，在实际问题中不同样本在训练时权重不同。针对不同的样本选取不同的惩罚因子，从而提高了小样本分类精度[4-5]。Suykens等[6]提出最小二乘支持向量机（Least squares support vector machine，LSSVM），针对SVM中约束条件所带来的计算复杂、边界定义不清晰等问题，使用等式约束条件代替不等式约束条件[7]。拉格朗日支持向量机（Lagrangian support vector machine，LSVM）[8]提高了在处理大规模线性数据集和中小型规模非线性数据集时的收敛速度。Tsang等提出基于最小包含球（Minimum enclosing ball，MEB）的核心向量机（Core vector machine，CVM）[9]，解决大规模数据集分类时所消耗大量时间和空间成本的问题。Lin等提出模糊支持向量机（Fuzzy support vector machine，FSVM），在实际问题中，数据集通常会存在不同程度的噪声，影响分类效果。为解决噪声问题，引入模糊隶属度来降低噪声和异常值对训练结果的影响[10-11]，Xu等提出一种基于马尔科夫采样的增量支持向量机（Markov resampling incremental support vector machines，MRISVM）[12]来提高运算速度和精度；Liu等提出了一种在线半监督支持向量机（Online semi-supervised support vector machine，OSSVM），提高流数据处理效率[13]；Panja等提出一种最小跨度支持向量机（Minimally Spanned support vector machine，MSSVM）用来减少支持向量的个数，提高训练速度[14]。此外支持向量机的改进算法还有：主从模式下的分布式支持向量机（Master-slave distributed support vector machine，MSDSVM）[15]、Fisher正则化支持向量机（Fisher regularized support vector machine，FRSVM）[16]、变系数支持向量机（Varying coefficient support vector machines，VCSVM）[17]、场域支持向量机（Field support vector machines，FSVM）[18]以及双分布支持向量机（Double distribution support vector machine，DDSVM）[19]。

尽管上述几种方法在特定领域中均有良好的分类效果，但仍然面临一些挑战：(1)分类过程未考虑数据样本内部的分布特征，造成了数据资源的浪费，无法进一步提升分类性能。(2)分类结果忽视了各类样本的相对关系。假设特征空间中有3类样本，样本的先后顺序为m_1，m_2，m_3，分类结果应尽量保证3类样本的相对顺序不变。如图1所示，3类样本投影在W_1方向上顺序为m_1，m_2，m_3，投影在W_2方向上的顺序为m_1，m_3，m_2，从样本保序性角度看，投影方向W_1优于W_2。(3)无法解决大规模分类问题。因此，本文提出融合数据分布特征的保序学习机（Rank preservation learning machine based on data distribution fusion，RPLM-DDF），该方法引入线性判别分析（Linear discriminant analysis，LDA）中的类内离散度S_W用以表征数据的分布特征；将各类样本中心相对关系考虑到最优化问题中，来确保分类结果依然保持相对顺序不变；引入核心向量机来保证RPLM-DDF对大型数据集依然可用。

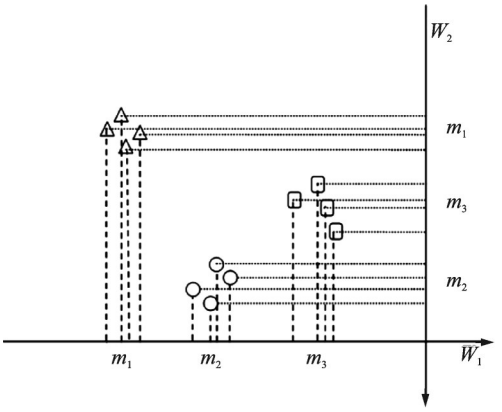


图1 RPLM-DDF工作示意图

Fig.1 RPLM-DDF working diagram

1 相关理论

1.1 最优化问题

假设样本集为$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}\in(X,Y)^N$，其中：$x_i\in X$，$y_i\in Y$；$\bar{x}_i$代表第$i$类样本

均值,X表示所有样本特征的集合;y_i代表第i类,Y表示所有类别;$N_i(i=1,2,\cdots,c)$为各类样本规模,c为类别数,N为样本总规模。

RPLM-DDF使数据样本的类内离散度S_W尽可能小,并通过各类样本的中心位置顺序不变来保持各类样本的顺序不变。RPLM-DDF的最优化问题可描述为

$$\min_{\boldsymbol{W}} \frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\frac{1}{2}\beta\boldsymbol{W}^{\mathrm{T}}S_W\boldsymbol{W}-\nu\rho \tag{1}$$

$$\text{s.t. } \boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i)\geqslant\rho \quad i=1,2,3,\cdots,c-1 \tag{2}$$

式中:$\boldsymbol{W}$为分类超平面的法向量,β为平衡因子,ρ为各类样本间距,ν为常数用来制约ρ,使得$\nu\rho$达到最好的约束效果。$\boldsymbol{m}_i=\frac{1}{N}\sum_{k=1}^{N_i}\boldsymbol{x}_k$ $(i=1,2,3,\cdots,c)$为各类样本的均值。S_W定义为:$S_W=\sum_{i=1}^{c}P(C_i)\sum_{\boldsymbol{x}\in C_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}}$,其中$C_i(i=1,2,3,\cdots,c)$表示第$i$类样本集合,$P(C_i)=N_i/N$ $(i=1,2,3,\cdots,c)$。

由Lagrangian定理得

$$L(\boldsymbol{W},\rho,\alpha)=\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\frac{1}{2}\beta\boldsymbol{W}^{\mathrm{T}}S_W\boldsymbol{W}-\nu\rho-\sum_{i=1}^{c-1}\alpha_i\left(\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i)-\rho\right) \tag{3}$$

式中Lagrangian乘子$\alpha_i\geqslant 0$。

L分别对$\boldsymbol{W}$,ρ求偏导并令偏导数等于0,可得

$$\frac{\partial L}{\partial \boldsymbol{W}}=\boldsymbol{W}+\beta S_W\boldsymbol{W}-\sum_{i=1}^{c-1}\alpha_i(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i)=0\Rightarrow \boldsymbol{W}=(1+\beta S_W)^{-1}\sum_{i=1}^{c-1}\alpha_i(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i) \tag{4}$$

$$\frac{\partial L}{\partial \rho}=-\nu+\sum_{i=1}^{c-1}\alpha_i=0\Rightarrow \sum_{i=1}^{c-1}\alpha_i=\nu \tag{5}$$

将式(4)—(5)代入式(1),可得对偶形式

$$\max_{\alpha}\ \sum_{i=1}^{c-1}\sum_{j=1}^{c-1}\alpha_i\alpha_j(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i)^{\mathrm{T}}(1+\beta S_W)^{-1}(\boldsymbol{m}_{j+1}-\boldsymbol{m}_j) \tag{6}$$

$$\text{s.t. } \alpha_i=v\ \alpha_i\geqslant 0 \quad i=1,2,3,\cdots,c-1 \tag{7}$$

1.2 判别函数和时间复杂度

RPLM-DDF的判别函数为

$$f(\boldsymbol{x})=\min_{k\in\{1,2,\cdots,c-1\}}\{k:\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}<b_k\} \tag{8}$$

式中$b_k=\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{m}_{i+1}-\boldsymbol{m}_i)/2$。

RPLM-DDF的求解主要包含大小为$N\times N$阵的转置运算,其时间复杂度为$O(N^2\log(N))$;大小为$(c-1)\times(c-1)$ Hessian矩阵二次规划(Quadratic programming, QP)问题的求解运算,时间复杂度为$O((c-1)^3)$。所以RPLM-DDF的时间复杂度为$O(N^2\log(N)+(c-1)^3)$,但是$c\ll N$,则RPLM-DDF的时间复杂度可用$O(c^3)$近似表示。

1.3 非线性形式

1.3.1 核化形式

诸多实际问题中的数据在原始空间往往线性不可分,使用非线性映射函数将原始空间样本数据映射到高维空间中,原本在低维空间线性不可分的数据,转化为高维空间线性可分的数据。

假设映射函数φ满足$\varphi:x\rightarrow\varphi(x)$条件,RPLM-DDF最优化问题的非线性形式可表示为

$$\min_{W} \ \frac{1}{2}W^{\mathrm{T}}W+\frac{1}{2}\beta W^{\mathrm{T}}S_W^{\varphi}-\nu\rho \tag{9}$$

$$\text{s.t.} \ \ W^{\mathrm{T}}(\boldsymbol{m}_{i+1}^{\varphi}-\boldsymbol{m}_i^{\varphi})\geqslant\rho \quad i=1,2,3,\cdots,c-1 \tag{10}$$

式中：$\boldsymbol{m}_i^{\varphi}=\frac{1}{N_i}\sum_{k=1}^{N_i}\varphi(\boldsymbol{x}_k)\ (i=1,2,3,\cdots,c-1)$，$S_W^{\varphi}=\sum_{i=1}^{c}P(C_i)\sum_{x=C_i}(\boldsymbol{x}-\boldsymbol{m}_i^{\varphi})(\boldsymbol{x}-\boldsymbol{m}_i^{\varphi})^{\mathrm{T}}$。

同理可得上述优化问题的核化对偶形式为

$$\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{c-1}\sum_{j=1}^{c-1}\alpha_i\alpha_j(\boldsymbol{m}_{i+1}^{\varphi}-\boldsymbol{m}_i^{\varphi})^{\mathrm{T}}(1+\beta S_W^{\varphi})^{-1}(\boldsymbol{m}_{j+1}^{\varphi}-\boldsymbol{m}_j^{\varphi}) \tag{11}$$

$$\text{s.t.} \ \ \sum_{i=1}^{c-1}\alpha_i=\nu,\ \alpha_i\geqslant 0 \quad i=1,2,3,\cdots,c-1 \tag{12}$$

1.3.2 核函数形式

通过引入核函数，无需知道非线性变换函数$\varphi(\boldsymbol{x})$的具体形式及参数，高维空间中的内积可以通过核函数直接运算，升维后算法复杂度没有随着维度增加而增加。但是使用核函数时由于$\varphi(\boldsymbol{x})$是未知的，所以无法直接求解S_W^{φ}和$\boldsymbol{m}_i^{\varphi}$，因此不能直接求解式(11)和式(12)的对偶问题，故提出一种方案解决上述问题。以下推论均假设只有两类数据。

原始最优化问题转化为

$$\min_{W} \ \frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\frac{1}{2}\beta\boldsymbol{W}^{\mathrm{T}}S_W\boldsymbol{W}-\nu\rho \tag{13}$$

$$\text{s.t.} \ \ y_i(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i)\geqslant\rho \tag{14}$$

根据再生核希尔伯特空间(Reproducing kernel Hilbert space，RKHS)的性质，$\boldsymbol{W}$可以写成$\boldsymbol{W}=\sum_{i=1}^{N}\alpha_i\varphi(\boldsymbol{x}_i)$，$K(\boldsymbol{x},\boldsymbol{x}')=<\varphi(\boldsymbol{x})\varphi(\boldsymbol{x}')>$。

式(13)中

$$\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{1}{2}\alpha^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}^{\mathrm{T}}\alpha \tag{15}$$

式中：$\boldsymbol{Y}$为对角矩阵，$\boldsymbol{Y}=\mathrm{diag}(y_1,y_2,\cdots,y_N)$；$\boldsymbol{G}$为一个由核函数内积组成的$N\times N$矩阵，即

$$\boldsymbol{G}=\begin{bmatrix}K(\boldsymbol{x}_1,\boldsymbol{x}_1) & K(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_N)\\ K(\boldsymbol{x}_2,\boldsymbol{x}_1) & K(\boldsymbol{x}_2,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_2,\boldsymbol{x}_N)\\ \cdots & & & \\ K(\boldsymbol{x}_N,\boldsymbol{x}_1) & K(\boldsymbol{x}_N,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_N,\boldsymbol{x}_N)\end{bmatrix}_{N\times N}$$

则$\boldsymbol{W}^{\mathrm{T}}S_W\boldsymbol{W}$可以表示为

$$\boldsymbol{W}^{\mathrm{T}}S_W^{\varphi}\boldsymbol{W}=\boldsymbol{W}^{\mathrm{T}}\left(\sum_{i=1}^{c}P(C_i)\sum_{x\in C_i}(\varphi(\boldsymbol{x})-\boldsymbol{m}_i^{\varphi})(\varphi(\boldsymbol{x})-\boldsymbol{m}_i^{\varphi})^{\mathrm{T}}\right)\boldsymbol{W} \tag{16}$$

式中$\boldsymbol{m}_i^{\phi}$为高维度特征空间中第i类的样本均值，表示为$\boldsymbol{m}_i^{\varphi}=\frac{1}{N_i}\sum_{x\in C_i}\varphi(\boldsymbol{x})$。

取第一类为例

$$\boldsymbol{W}^{\mathrm{T}}\sum_{x\in C_i}(\varphi(\boldsymbol{x})-\boldsymbol{m}_1^{\varphi})(\varphi(\boldsymbol{x})-\boldsymbol{m}_1^{\varphi})^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{K}_1(I_1-L_1)\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\alpha} \tag{17}$$

式中：$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\alpha_3,\cdots,\alpha_N)$；$\boldsymbol{I}_1$为$N_1$阶单位矩阵；$\boldsymbol{L}_1$为$N_1\times N_1$阶$\frac{1}{N_1}$填充的矩阵；$\boldsymbol{Y}$为对角矩阵，定义为$\boldsymbol{Y}=\mathrm{diag}(y_1,y_2,y_3,\cdots,y_{N1})$。

$$\boldsymbol{K}_1=\begin{bmatrix} K(\boldsymbol{x}_1,\boldsymbol{x}_1^{(C_1)}) & K(\boldsymbol{x}_1,\boldsymbol{x}_2^{(C_1)}) & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_{N_1}^{(C_1)}) \\ K(\boldsymbol{x}_2,\boldsymbol{x}_1^{(C_1)}) & K(\boldsymbol{x}_2,\boldsymbol{x}_2^{(C_1)}) & \cdots & K(\boldsymbol{x}_2,\boldsymbol{x}_{N_1}^{(C_1)}) \\ \vdots & \vdots & & \vdots \\ K(\boldsymbol{x}_N,\boldsymbol{x}_1^{(C_1)}) & K(\boldsymbol{x}_N,\boldsymbol{x}_2^{(C_1)}) & \cdots & K(\boldsymbol{x}_N,\boldsymbol{x}_{N_1}^{(C_1)}) \end{bmatrix}_{N\times N_1}$$

同理,第二类可得

$$\boldsymbol{W}^{\mathrm{T}}S_W^{\varphi}\boldsymbol{W}=P(C_i)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{K}_i(\boldsymbol{I}_i-\boldsymbol{L}_i)\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\alpha} \tag{18}$$

将式(15),(18)代入式(13)可得

$$\min_{\alpha}\ \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}[\boldsymbol{G}+\beta P(C_1)\boldsymbol{K}_1(\boldsymbol{I}_1-\boldsymbol{L}_1)\boldsymbol{K}_1^{\mathrm{T}}+\beta P(C_2)\boldsymbol{K}_2(\boldsymbol{I}_2-\boldsymbol{L}_2)\boldsymbol{K}_2^{\mathrm{T}}]\boldsymbol{Y}\boldsymbol{\alpha}-\nu\rho \tag{19}$$

$$\text{s.t.}\ \ y_i(\sum_{j=1}^{N}\alpha_j y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)+b)\geqslant\rho\qquad\rho\geqslant 0 \tag{20}$$

令$\boldsymbol{Q}=\boldsymbol{Y}[\boldsymbol{G}+\beta P(C_1)\boldsymbol{K}_1(\boldsymbol{I}_1-\boldsymbol{L}_1)\boldsymbol{K}_1^{\mathrm{T}}+\beta P(C_2)\boldsymbol{K}_2(\boldsymbol{I}_2-\boldsymbol{L}_2)\boldsymbol{K}_2^{\mathrm{T}}]\boldsymbol{Y}$,代入式(19)可得

$$\min\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}-\nu\rho \tag{21}$$

$$\text{s.t.}\ \ y_i\left(\sum_{j=1}^{N}\alpha_j y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)+b\right)\geqslant\rho\quad\rho\geqslant 0 \tag{22}$$

令非线性RPLM-DDF的Lagrangian函数为

$$L(\boldsymbol{\alpha},b,\rho,h,g)=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}-\nu\rho-\sum_{i=1}^{N}h_i\left\{y_i\left[\sum_{j=1}^{N}\alpha_j y_j K(x_j,x_i)+b\right]-\rho\right\}-\sum_{i=1}^{N}g_i\rho \tag{23}$$

式中Lagrangian乘子$h_i\geqslant 0$,$g_i\geqslant 0$。

L分别对b和ρ求偏导并令偏导数等于0,可得

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{N}h_i y_i=0 \tag{24}$$

$$\frac{\partial L}{\partial\rho}=-\nu+h_i-g_i\Rightarrow\begin{cases}g_i=h_i-\nu\\0\leqslant h_i\leqslant\nu\end{cases} \tag{25}$$

将式(24)—(25)代入式(23)中,可得

$$L(\boldsymbol{\alpha},\boldsymbol{h})=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h} \tag{26}$$

L对$\boldsymbol{\alpha}$求偏导并令偏导数等于0,有

$$\frac{\partial L}{\partial\alpha}=\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h}=0\Rightarrow\boldsymbol{\alpha}=\boldsymbol{Q}^{-1}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h} \tag{27}$$

将式(27)代入式(26),可得

$$\min\frac{1}{2}(\boldsymbol{Q}^{-1}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h})^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{Q}^{-1}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h})-(\boldsymbol{Q}^{-1}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h})^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}\boldsymbol{h}=\min\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}[\boldsymbol{Y}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y}(\boldsymbol{Q}^{-1})\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}]\boldsymbol{h} \tag{28}$$

$$\text{s.t.}\ \ \sum_{i=1}^{N}h_i y_i=0\quad 0\leqslant h_i\leqslant\nu \tag{29}$$

由KKT条件,可得

$$h_i\left\{y_i\left[\sum_{j=1}^{N}\alpha_j y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)+b\right]-\rho\right\}=0 \tag{30}$$

最终决策函数为

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{31}$$

式中 $b=\dfrac{1}{|SV_s|}\left[y_i-\sum_{j=1}^{N}\alpha_j y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)\right]$, $SV_s=\{(\boldsymbol{x}_i,y_i)|0<h_i\leqslant\rho, i=1,2,3,\cdots,N\}$。

1.4 大规模分类问题

为了将所提方法推广到大规模数据集,引入核心向量CVM,通过建立RPLM-DDF与最小包含球对偶形式的等价关系,从而保证RPLM-DDF能够解决大规模分类问题。核心向量机把QP问题的求解转化为计算最小包含球(Minimum enclosing ball,MEB)问题。利用逼近率为$(1+\varepsilon)$的近似算法得到核心集。核心集的规模远远小于原始数据规模,从而降低了算法时间、空间复杂度。大规模数据集的实验表明,核心向量机与标准的SVM拥有相似的精度,但速度更快,可处理更大规模的数据集。

最小包含球最优线性形式为

$$\min\ R^2 \tag{32}$$

$$\text{s.t.}\quad \left\|\boldsymbol{c}-\boldsymbol{x}_i\right\|^2\leqslant R^2\quad i=1,2,3,\cdots,N \tag{33}$$

式中c为超球体的球心,R为超球体的半径。

核化形式为

$$\min\ R^2 \tag{34}$$

$$\text{s.t.}\quad \left\|\boldsymbol{c}-\varphi(\boldsymbol{x}_i)\right\|^2\leqslant R^2\quad i=1,2,3,\cdots,N \tag{35}$$

式中$\varphi(x_i)$表示从低维空间到高维空间的映射。

由Lagrangian定理可得

$$\max_{\boldsymbol{\alpha}}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{36}$$

$$\text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1\qquad \boldsymbol{\alpha}\geqslant\boldsymbol{0} \tag{37}$$

式中 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]^{\mathrm{T}}$, 核函数 $\boldsymbol{K}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]=[\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)]$, $\boldsymbol{1}=[1,\cdots,1]^{\mathrm{T}}$, $\boldsymbol{0}=[0,\cdots,0]^{\mathrm{T}}$。因为$\mathrm{diag}(\boldsymbol{K})$等于常数并且$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1$,所以式(36)可简化为

$$\max_{\boldsymbol{\alpha}}\ -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{38}$$

令$\theta=\boldsymbol{Y}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y}(\boldsymbol{Q}^{-1})^{-1}\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}$,RPLM-DDF的QP问题可转换为

$$\min\boldsymbol{h}^{\mathrm{T}}\boldsymbol{\theta}\boldsymbol{h} \tag{39}$$

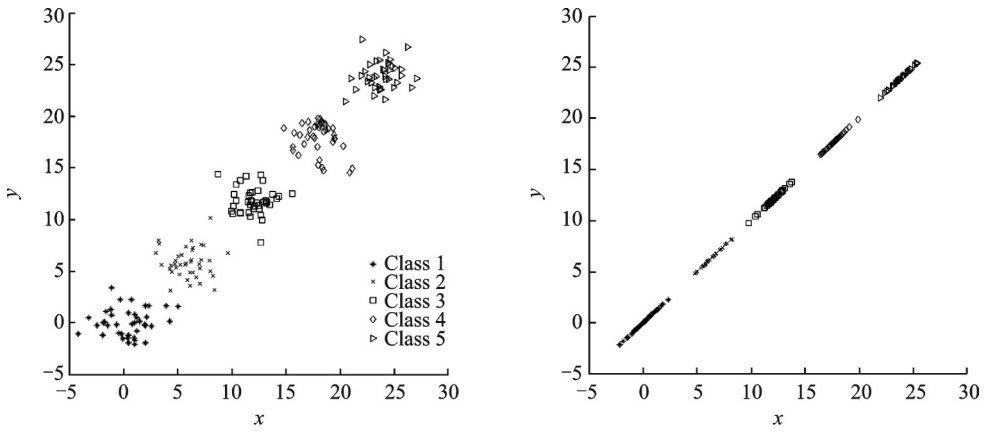
$$\text{s.t.}\quad \boldsymbol{h}^{\mathrm{T}}\boldsymbol{1}=1\quad \boldsymbol{h}\geqslant\boldsymbol{0} \tag{40}$$

式中$\boldsymbol{1}=[1,\cdots,1]^{\mathrm{T}}$, $\boldsymbol{0}=[0,\cdots,0]^{\mathrm{T}}$。RPLM-DDF与最小包含球形式等价,故RPLM-DDF可以使用MEB来解决大规模分类问题。

2 实验分析

2.1 人工数据集

人工生成5类数据,各类样本40个,各类中心点分别是(0,0),(6,6),(12,12),(18,18)和(24,24),标准差为2,并服从Gaussian分布。生成数据集如图2(a)所示,通过RPLM-DDF求得方向向量为$\boldsymbol{W}$,将生成数据投影到$\boldsymbol{W}$后如图2(b)所示。由图2可知,RPLM-DDF具有良好可分性,并且保持原始数据位置相对顺序不变。



(a) 人工实验数据集
(a) Artificial data set

(b) 实验结果
(b) Experimental results

图2 人工数据集及实验结果

Fig.2 Artificial data set and experimental results

2.2 中小型数据集

实验所需数据集如表1所示。选取60%数据作为训练集,剩余40%数据作为测试集。实际问题中经常使用的核函数有:线性核函数,多项式核函数,高斯核函数和Sigmoid核函数。不同的核函数在不同应用环境下表现各异,实验结果如图3所示。

表1 实验数据集

Table 1 Experimental data set

Dataset	Instances number	Dimension	Class number
Iris	150	4	3
Liver	345	7	2
Glass	214	10	5
Wine	178	13	3
German	1000	20	2

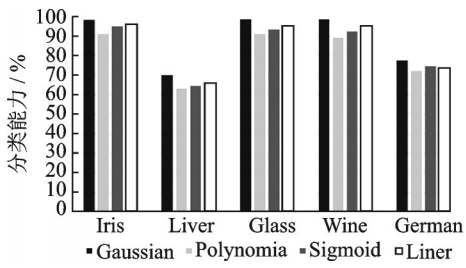


图3 核函数与实验结果

Fig.3 Kernel function and experimental results

由图3可以看出,与线性核函数、多项式核函数和Sigmoid核函数相比,基于高斯核函数的RPLM-DDF在实验数据集上具有更优的分类能力,因此实验选取高斯核函数。

实验采用交叉验证的方法。将RPLM-DDF与SVC(Support vectors classification),KNN(K-nearest neighbor)、朴素贝叶斯(Naive Bayes, NB)、决策树(Decision tree, DT)和多层感知器(Multi-layer perceptron, MLP)进行比较实验。使用网格搜索方法,在恰当的范围划分网格并遍历网格内所有点进行取值得到参数。γ在{0.001, 0.01, 0.1, 1 ,5, 10}中选择;惩罚参数C在{0.01, 0.05, 0.1, 0.5, 1, 5, 10}中选择;ν在{0.01, 0.1, 0.5, 1, 3, 5, 10}中选择;δ在$\{\bar{x}/2\sqrt{2}, \bar{x}/2, \bar{x}/\sqrt{2}, \bar{x}\sqrt{2}, 2\bar{x}, 2\sqrt{2}\bar{x}\}$中选择;最近邻数$K$在{1, 2, 3, 5, 10, 15, 20}中选择,$\bar{x}$是训练样本数据的平均范数平方根。MLP共4层全连接层,每层神经元数量为[128, 64, 64, Class Number],使用SGD优化器,学习率为0.001。实验参数如表2所示,实验结果如表3所示。

表2 实验参数

Table 2 Experimental parameters

Dataset	SVC	KNN	RPLM-DDF
Iris	$\gamma=1, C=1$	K=2	$\delta=\sqrt{2}\,\bar{x}, \nu$=0.1
Liver	$\gamma=0.01, C=0.5$	K=20	$\delta=\bar{x}/2\sqrt{2}, \nu$=0.1
Glass	$\gamma=0.1, C=1$	K=5	$\delta=\bar{x}/2\sqrt{2}, \nu$=0.5
Wine	$\gamma=0.001, C=5$	K=1	$\delta=\bar{x}/\sqrt{2}, \nu$=0.1
German	$\gamma=0.1, C=0.5$	K=20	$\delta=\bar{x}/\sqrt{2}, \nu$=0.1

表3 中小规模数据集对比实验结果

Table 3 Experimental results of small and medium-sized datasets %

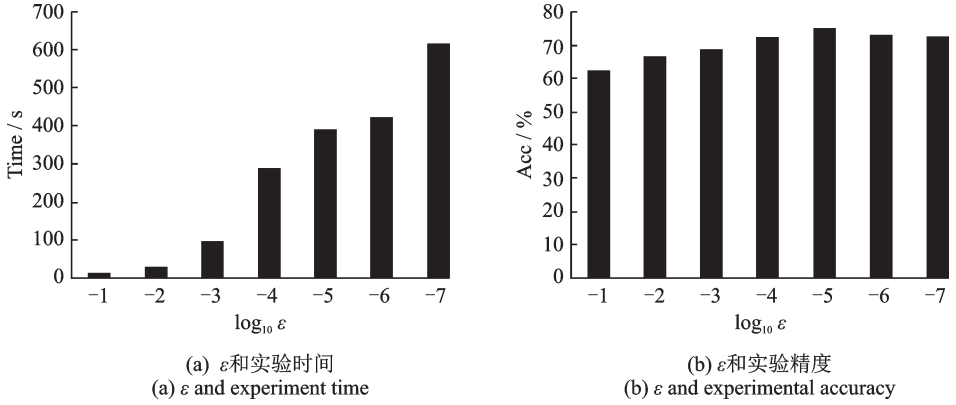
Dataset	SVC	KNN	NB	DT	MLP	RPLM-DDF
Iris	96.67	**98.33**	95.00	93.33	96.67	**98.33**
Liver	64.49	67.39	60.86	63.48	64.35	**70.28**
Glass	97.67	97.67	90.69	87.38	93.46	**98.83**
Wine	69.44	68.05	93.26	95.56	86.11	**98.61**
German	71.25	71.25	72.50	69.00	75.17	**77.50**
Average	79.90	80.54	82.46	81.75	83.15	**88.71**

由表3可以看出,RPLM-DDF较之SVC,KNN,NB,DT和MLP,在平均分类性能上精度更高。在Iris数据集中KNN和RPLM-DDF表现相当,在Liver,Glass,Wine和German数据集中,与SVC,KNN,NB,DT和MLP传统分类方法相比,RPLM-DDF的分类效果更优。

2.3 大型数据集

2.3.1 ε参数对实验的影响

实验采用Bank Marketing DataSet数据集,共有45 211个样本,17维描述信息,共分为两类。60%的数据集作为训练样本,剩余数据集作为测试样本。ε将在$\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$中选取。$\varepsilon$对实验时间影响如图4(a)所示,$\varepsilon$对实验精度Acc的影响如图4(b)所示。



(a) ε和实验时间 (a) ε and experiment time

(b) ε和实验精度 (b) ε and experimental accuracy

图4 ε对实验RPLM-DDF的影响

Fig.4 Effect of ε on experiment RPLM-DDF

由图4可知,ε越小所需训练时间越长,但不是ε越小精度越高。选取合适的ε值可以减少训练时间,并达到最高精度。

2.3.2 性能分析

将数据集的20%,40%,60%和80%作为训练集,并从剩余数据任取500个作为测试集。实验结果如表4所示。由表4可以看出,随着训练样本的增加,RPLM-DDF分类精度呈上升趋势。训练时间随着训练样本的增加而增加,但是RPLM-DDF能在有限的时间内高精度地完成分类任务。

表4 RPLM-DDF对大规模数据分类结果

Table 4 RPLM-DDF classification results of large-scale data

Datasets size	Abalone		Bank		California	
	Acc/%	Time/s	Acc/%	Time/s	Acc/%	Time/s
20	61.46	80.12	63.68	156.32	46.03	243.84
40	70.21	130.45	67.32	278.53	54.58	403.47
60	75.36	173.26	71.58	295.72	60.26	672.94
80	76.14	197.63	77.04	331.18	64.57	734.28

3 结束语

针对SVM的不足,本文提出RPLM-DDF方法。RPLM-DDF主要优势在于:(1)在考虑最优化问题时将类内结构融合起来,合理有效地利用这种信息,提高了算法分类精度;(2)较好地保持了数据的相对关系不变;(3)基于核心向量机使RPLM-DDF支持大规模分类问题。在人工数据集、中小规模数据集和大规模数据集上实验表明与传统分类方法相比,所提方法具有更优的分类能力。然而,RPLM-DDF的分类结果依赖于实验参数的选取,如何更加高效地选择最优参数是下一步研究的重点。

参考文献:

[1] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
ZHANG Xuegong. Introduction to statistical learning theory and support vector machines[J]. ACTA Automatica Sinica, 2000, 26(1): 32-42.

[2] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.
WANG Haiyan, LI Jianhui, YANG Fenglei. Overview of support vector machine analysis and algorithm[J]. Application Research of Computers, 2014, 31(5): 1281-1286.

[3] 范昕炜, 杜树新, 吴铁军. 可补偿类别差异的加权支持向量机算法[J]. 中国图象图形学报, 2003, 8(9): 1037-1042.
FAN Xinwei, DU Shuxin, WU Tiejun. Weighted support vector machine based classification algorithm for uneven class size problems[J]. Journal of Image and Graphics, 2003, 8(9): 1037-1042.

[4] 汪廷华, 田盛丰, 黄厚宽. 特征加权支持向量机[J]. 电子与信息学报, 2009, 31(3): 514-518.
WANG Tinghua, TIAN Shengfeng, HUANG Houkuan. Feature weighted support vector machine[J]. Journal of Electronics and Information Technology, 2009, 31(3): 514-518.

[5] ABIDINE M B, FERGANI B. News schemes for activity recognition systems using PCA-WSVM, ICA-WSVM, and LDA-WSVM[J]. Information, 2015, 6(3): 505-521.

[6] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[7] CHEN L, ZHOU S. Sparse algorithm for robust lssvm in primal space[J]. Neurocomputing, 2017, 275: 2880-2891.

[8] MANGASARIAN O L, MUSICANT D R. Lagrangian support vector machines[J]. The Journal of Machine Learning Research, 2001, 1(3): 161-177.

[9] 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(2): 363-392.

[10] LIN C F, WANG S D. Fuzzy support vector machines[J]. IEEE Trans Neural Netw, 2002, 13(2): 464-471.

[11] 李娜, 孙乐, 胡一楠, 等. 模糊型支持向量机及其在入侵检测中的应用[J]. 科技创新与应用, 2018,11: 154-158.
LI Na, SUN Le, HU Yinan, et al. Fuzzy support vector machine and its application in intrusion detection[J]. Technology Innovation and Application, 2018, 11: 154-158.

[12] XU J, XU C, ZOU B, et al. New incremental learning algorithm with support vector machines[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2018, 49(11): 2230-2241.

[13] LIU Y, XU Z, LI C. Online semi-supervised support vector machine[J]. Information Sciences, 2018, 439/440: 125-141.

[14] PANJA R, PAL N R. MS-SVM: Minimally spanned support vector machine[J]. Applied Soft Computing, 2018, 64: 356-365.

[15] CHEN Q, CAO F. Distributed support vector machine in master-slave mode[J]. Neural Networks, 2018, 101: 94.

[16] ZHANG L, ZHOU W D. Fisher-regularized support vector machine[J]. Information Sciences, 2016, 343: 79-93.

[17] LU X, DONG F, LIU X, et al. Varying coefficient support vector machines[J]. Statistics & Probability Letters, 2018, 132: 107-115.

[18] HUANG K, JIANG H, ZHANG X Y. Field support vector machines[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(6): 454-463.

[19] CHENG F, ZHANG J, LI Z, et al. Double distribution support vector machine[J]. Pattern Recognition Letters, 2017, 88: 20-25.

作者简介:

刘忠宝(1981-),男,教授,研究方向:数据挖掘、信息资源管理, E-mail: liuzb@nuc.edu.cn。

张志剑(1994-),男,硕士研究生,研究方向:数据挖掘、人工智能。

党建飞(1993-),男,硕士研究生,研究方向:数据挖掘、知识发现。

(编辑:刘彦东)