基于流形判别分析的全局保序学习机

张 静[1]，刘忠宝[2]

(1. 中北大学软件学院 太原 030051; 2. 中北大学计算机与控制工程学院 太原 030051)

【摘要】当前主流分类方法在分类决策时无法同时考虑样本的全局特征和局部特征，而且大多算法仅关注各类样本的可分性，往往忽略样本之间的相对关系。为了解决上述问题，提出了基于流形判别分析的全局保序学习机。该方法引入流形判别分析来反映样本的全局特征和局部特征；通过保持各类样本中心的相对关系不变进而实现保持全体样本的先后顺序不变；借鉴核心向量机有关理论和方法，通过建立所提方法与核心向量机对偶形式的等价关系实现大规模分类。人工数据集和标准数据集上的比较实验验证了该方法的有效性。

关 键 词 全局保序； 大规模分类； 流形判别分析； 支持向量机

中图分类号 TP181 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.06.020

Global Rank Preservation Learning Machine Based on Manifold-Based Discriminant Analysis

ZHANG Jing[1] and LIU Zhong-bao[2]

(1. School of Software, North University of China Taiyuan 030051;

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

Abstract In order to solve the problems that many traditional classification methods confronted, a global rank preservation learning machine (GRPLM) based on manifold-based discriminant analysis is proposed in this paper. In GRPLM, the manifold-based discriminant analysis (MDA) is introduced to represent the samples' global and local characteristic; the relative relationship of different class centers is taken into consideration in order to preserve the samples' ranks; the equivalent relation between the QP form of GRPLM and core vector machine (CVM) is analyzed in order to broaden the usage of GRPLM from small- and medium-scale to large-scale. Comparative experiments on several standard datasets verify the effectiveness of the proposed methods.

Key words global rank preservation; large-scale classification; manifold-based discriminant analysis (MDA); support vector machine (SVM)

模式分类是数据挖掘、模式识别、机器学习等领域的研究热点。当前主流的分类方法可归纳为以下几类：1) 决策树。ID3算法通过互信息理论建立树状分类模型，在ID3基础上先后提出C4.5[1]、PUBLIC[2]、SLIQ[3]、RainForest[4]等改进算法；2) 关联规则算法主要包括关联分析算法[5]、多维关联规则算法以及预测性关联规则算法[6]；3) 支持向量机。支持向量机(support vector achine, SVM)[7-10]起初于1995年被提出，随着应用的不断深入，根据不同的应用场合，研究人员先后提出众多改进算法：ν-SVM[11] 、单类支持向量机(one class support vector machine, OCSVM)[12]、支持向量数据描述(support vector data description, SVDD)[13] 、核心向量机(core vector machine, CVM)[14]、Lagrangian支持向量机(Largrangian support vector machine, LSVM)[15]、最小二乘支持向量机(least squares support vector machine, LSSVM)[16]以及光滑支持向量机(smooth support vector machine, SSVM)[17]等；4) 贝叶斯分类器包括半朴素贝叶斯分类器(semi-naive Bayesian classifier)、基于属性删除的选择性贝叶斯分类器(selective Bayesian classifier based on atribute deletion)[18] 、基于懒惰式贝叶斯规则的学习算法(lazy Bayesian rule learning algorithm，LBR)[19] 及树扩张型贝叶斯分类器(tree augmented Bayesian classifier, TAN)[20]等。

上述方法在实际应用中取得良好的分类效果，但它们面临如下挑战：1) 在分类决策时无法同时考

收稿日期：2015－01－25；修回日期：2015－09－25
基金项目：国家自然科学基金(61202311)；山西省高等学校科技创新项目(2014142)
作者简介：张静(1980－)，女，博士生，主要从事智能信息处理方面的研究.

虑样本的全局特征和局部特征；2) 大多算法仅关注各类样本的可分性，而忽略样本之间的相对关系。GRPLM工作原理为，三类样本在$\boldsymbol{W}_1$方向上的投影顺序为m1 m2 m3，而在$\boldsymbol{W}_2$方向上的投影顺序是m2 m3 m1，假设原空间三类样本的相对关系为m1 m2 m3，则$\boldsymbol{W}_1$方向优于$\boldsymbol{W}_2$方向；3) 无法解决大规模分类问题。鉴于此，提出基于流形判别分析的全局保序学习机(global rank preservation learning machine based on manifold-based discriminant analysis, GRPLM)。该方法通过引入流形判别分析[21]来保持样本的全局和局部特征；在最优化问题的约束条件中增加样本中心相对关系限制保证分类决策时考虑样本的相对关系；通过引入核心向量机(core vector machine, CVM)[14]将所提方法适用范围扩展到大规模数据。

本文后续做如下假设：样本集为$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}\in(\boldsymbol{X}\times\boldsymbol{Y})^N$，其中$x_i\in\boldsymbol{X}=R^N$，$y_i\in\boldsymbol{Y}=\{1,2,\cdots,c\}$，类别数为$c$，各类样本数为$N_i(1,2,\cdots,c)$，$\bar{x}$为所有样本均值，$\bar{x}_i$为第$i$类样本均值。

1 流形判别分析

文献[21]提出流形判别分析MDA，流形判别分析引入基于流形的类内离散度(manifold-based Within-class scatter, MWCS)$\boldsymbol{M}_W$和基于流形的类间离散度(manifold-based between-class scatter, MBCS)$\boldsymbol{M}_B$两个概念，试图利用Fisher准则，通过最大化$\boldsymbol{M}_B$和$\boldsymbol{M}_W$之比获得最佳投影方向。其中，$\boldsymbol{M}_B$和$\boldsymbol{M}_W$的定义如下：

$$\boldsymbol{M}_W=\mu\boldsymbol{S}_W+(1-\mu)\boldsymbol{S}_S \tag{1}$$

$$\boldsymbol{M}_B=\lambda\boldsymbol{S}_B+(1-\lambda)\boldsymbol{S}_D \tag{2}$$

式中，μ和λ为常数并通过网格搜索策略进行选择；$\boldsymbol{S}_w$和$\boldsymbol{S}_B$分别表示类内离散度和类间离散度，其定义同LDA；$\boldsymbol{S}_S$和$\boldsymbol{S}_D$分别表征同类样本和异类样本的局部结构，两者定义如下：

$$\boldsymbol{S}_W=\sum_{i=1}^{c}\sum_{j=1}^{N_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{S}_B=\sum_{i=1}^{c}N_i(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^{\mathrm{T}} \tag{4}$$

$$\boldsymbol{S}_S=\boldsymbol{X}(\boldsymbol{S}'-\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}} \tag{5}$$

式中，$\boldsymbol{S}'$为对角阵且$\boldsymbol{S}'=\sum_j S_{ij}$，$S_{ij}$为同类权重函数：

$$S_{ij}=\begin{cases}\exp(-\left\|x_i-x_j\right\|^2) & y_i=y_j\\ 0 & y_i\neq y_j\end{cases} \tag{6}$$

$$\boldsymbol{S}_D=\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{D})\boldsymbol{X}^{\mathrm{T}} \tag{7}$$

式中，$\boldsymbol{D}'$为对角阵且$\boldsymbol{D}'=\sum_j D_{ij}$，其中$D_{ij}$为异类权重函数：

$$D_{ij}=\begin{cases}\exp(-1/\left\|x_i-x_j\right\|^2) & y_i\neq y_j\\ 0 & y_i=y_j\end{cases} \tag{8}$$

MDA的最优化问题定义如下：

$$J=\max_{\boldsymbol{W}}\frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_B\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}}=\max_{\boldsymbol{W}}\frac{\boldsymbol{W}^{\mathrm{T}}(\lambda\boldsymbol{S}_B+(1-\lambda)\boldsymbol{S}_D)\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}(\mu\boldsymbol{S}_W+(1-\mu)\boldsymbol{S}_S)\boldsymbol{W}} \tag{9}$$

2 GRPLM

2.1 最优化问题

GRPLM利用SVM和MDA分别在智能分类和特征提取方面的优势，在分类过程中将样本的全局特征和局部特征以及样本之间相对关系考虑在内，在一定程度上提高分类效率。GRPLM找到的分类超平面具有以下优势：1) 通过引入流形判别分析来保持样本的全局特征和局部特征；2) 通过最小化基于流形的类内离散度，保证同类样本尽可能紧密；3) 通过保持各类样本中心的相对关系不变进而实现保持全体样本的先后顺序不变。

上述思想可表示为如下最优化问题：

$$\min_{\boldsymbol{W}}\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}-\nu\rho \tag{10}$$

$$\text{s.t}\ \ \boldsymbol{W}^{\mathrm{T}}(m_{i+1}-m_i)\geqslant\rho\quad i=1,2,\cdots,c-1 \tag{11}$$

式中，$\boldsymbol{W}$为分类超平面的法向量；ν为常数并通过网格搜索策略选择；ρ为各类样本间隔；$m_i=\frac{1}{N_i}\sum_{k=1}^{N_i}x_k$ $(i=1,2,\cdots,c)$为各类样本均值，c为样本类别数；$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}$表示找到的分类超平面将样本的全局特征和局部特征考虑在内；$\nu\rho$的存在保证各类样本的间隔尽可能大，有利于提高分类精度；式(11)表明GRPLM在分类决策时保持各类样本的相对关系不变。

上述最优化问题的对偶形式如下：

$$\max_{\alpha}\ -\sum_{i=1}^{c-1}\sum_{j=1}^{c-1}\alpha_i\alpha_j(m_{i+1}-m_i)^{\mathrm{T}}\boldsymbol{M}_W^{-1}(m_{j+1}-m_j) \tag{12}$$

$$\text{s.t}\ \ \sum_{i=1}^{c-1}\alpha_i=\nu \tag{13}$$

$$\alpha_i\geqslant 0\quad i=1,2,\cdots,c-1 \tag{14}$$

证明：由Lagrangian定理可得：

$$L(\boldsymbol{W},\rho,\alpha)=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}-\rho-\sum_{i=1}^{c-1}\alpha_i(\boldsymbol{W}^{\mathrm{T}}(m_{i+1}-m_i)-v\rho) \tag{15}$$

$$\frac{\partial L}{\partial \boldsymbol{W}}=0\Leftrightarrow \boldsymbol{W}=\frac{1}{2}\boldsymbol{M}_W^{-1}\sum_{i=1}^{c-1}\alpha_i(m_{i+1}-m_i) \tag{16}$$

$$\frac{\partial L}{\partial \rho}=0\Leftrightarrow \sum_{i=1}^{c-1}\alpha_i=v \tag{17}$$

将式(16)和式(17)带入到式(15)，并去掉与研究无关的常数项可得上述对偶式。

2.2　决策函数

GRPLM的决策函数为：

$$f(x)=\min_{k\in\{1,2,\cdots,c-1\}}\{k:\boldsymbol{W}^{\mathrm{T}}x<b_k\} \tag{18}$$

式中，$b_k=\boldsymbol{W}^{\mathrm{T}}(m_{i+1}+m_i)/2$。

2.3　核化形式

假设映射函数ϕ满足$\phi:x\to\phi(x)$。原最优化问题的核化形式可表示为：

$$\min_{\boldsymbol{W}}\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W^{\phi}\boldsymbol{W}-v\rho \tag{19}$$

$$\text{s.t}\ \ \boldsymbol{W}^{\mathrm{T}}(m_{i+1}^{\phi}-m_i^{\phi})\geqslant\rho\quad i=1,2,\cdots,c-1 \tag{20}$$

式中，

$$m_i^{\phi}=\frac{1}{N_i}\sum_{k=1}^{N_i}\phi(x_k)\ (i=1,2,\cdots,c)$$

$$\boldsymbol{M}_W^{\phi}=\mu\boldsymbol{S}_W^{\phi}+(1-\mu)\boldsymbol{S}_S^{\phi}$$

$$\boldsymbol{S}_W^{\phi}=\sum_{i=1}^{c}\sum_{j=1}^{N_i}N_i(\phi(x_{ij})-m_i^{\phi})(\phi(x_{ij})-m_i^{\phi})^{\mathrm{T}}$$

$$\boldsymbol{S}_S^{\phi}=\sum_{i,j}(\phi(x_i)S_{ii}^{\phi}\phi(x_i^{\mathrm{T}})-\phi(x_i)S_{ij}^{\phi}\phi(x_i^{\mathrm{T}}))=\phi(\boldsymbol{X})(\boldsymbol{S}'^{\phi}-\boldsymbol{S}^{\phi})\phi(\boldsymbol{X}^{\mathrm{T}})$$

式中，$\boldsymbol{S}'^{\phi}$为对角阵且$\boldsymbol{S}'^{\phi}=\sum_j S_{ij}^{\phi}$，其中$S_{ij}^{\phi}$为核同类权重函数：

$$S_{ij}^{\phi}=\begin{cases}\exp(-\left\|\phi(x_i)-\phi(x_j)\right\|^2) & y_i=y_j\\ 0 & y_i\neq y_j\end{cases}$$

原最优化问题的核化对偶形式为：

$$\max_{\alpha}\ -\sum_{i=1}^{c-1}\sum_{j=1}^{c-1}\alpha_i\alpha_j(m_{i+1}^{\phi}-m_i^{\phi})^{\mathrm{T}}(\boldsymbol{M}_W^{\phi})^{-1}(m_{j+1}^{\phi}-m_j^{\phi})$$

$$\text{s.t}\quad\sum_{i=1}^{c-1}\alpha_i=v$$

$$\alpha_i\geqslant 0\quad i=1,2,\cdots,c-1$$

2.4　时间复杂度分析

GRPLM优化问题求解主要包括大小为$N\times N$矩阵的转置运算以及大小为$(c-1)\times(c-1)$Hessian矩阵QP问题求解运算。大小为$N\times N$矩阵的转置运算的时间复杂度为$O(N^2\log(N))$，大小为$(c-1)\times(c-1)$ Hessian矩阵QP问题求解的时间复杂度为$O((c-1)^3)$，GRPLM的时间复杂度为$O(N^2\log(N))+O((c-1)^3)$，由于$c<<N$，则GRPLM的时间复杂度近似表示为$O(c^3)$。

2.5　大规模分类

2.5.1　核心向量机

核心向量机CVM的基本思路是在大规模样本空间中利用一个逼近率为$(1+\varepsilon)$的近似算法找到核心集，该集合较之原始样本具有比更小的规模。更重要的是，文献[14]指出该核心集与样本数无关，只与参数ε有关，该结论有力地支持了CVM在解决大规模分类问题方面的有效性。

2.5.2　最小包含球

最优化问题定义线性形式为：

$$\min\ R^2 \tag{21}$$

$$\text{s.t.}\quad \left\|\boldsymbol{c}-x_i\right\|^2\leqslant R^2\quad i=1,2,\cdots,N \tag{22}$$

式中，$\boldsymbol{c}$为超球体球心；R为超球体半径。非线性形式为：

$$\min\ R^2 \tag{23}$$

$$\text{s.t.}\quad \left\|\boldsymbol{c}-\varphi(x_i)\right\|^2\leqslant R^2\quad i=1,2,\cdots,N \tag{24}$$

式中，$\varphi(x_i)$表示从原始样本空间到高维特征空间的映射。

由Lagrangian定理可得如下对偶形式：

$$\max_{\boldsymbol{\alpha}}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{25}$$

$$\text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\ \boldsymbol{\alpha}\geqslant\boldsymbol{0} \tag{26}$$

式中，$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$，$\boldsymbol{1}=[1,1,\cdots,1]^{\mathrm{T}}$，核函数$\boldsymbol{K}=[k(x_i,x_j)]=[\varphi(x_i)^{\mathrm{T}}\varphi(x_j)]$，$\boldsymbol{0}=[0,0,\cdots,0]^{\mathrm{T}}$。由于$\mathrm{diag}(\boldsymbol{K})$为常数且$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1$，则式(24)可简化为：

$$\max_{\boldsymbol{\alpha}}\ -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{27}$$

2.5.3　GRPLM与MEB关系

令$\beta=\alpha/v$，GRPLM的QP形式可转化为：

$$\max_{\boldsymbol{\beta}}\ -\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta} \tag{28}$$

$$\text{s.t}\quad \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{1}=1$$

$$\boldsymbol{\beta}\geqslant 0$$

式中，$\boldsymbol{K}=[(m_{i+1}-m_i)^{\mathrm{T}}\boldsymbol{M}_W^{-1}(m_{i+1}-m_i)]$，$\boldsymbol{1}=[1,1,\cdots,1]^{\mathrm{T}}$，$\boldsymbol{0}=[0,0,\cdots,0]^{\mathrm{T}}$。GRPLM与MEB的对偶形式等价，则可利用CVM解决大规模分类问题。

GRPLM-CVM算法描述如下：

1) 初始化参数；$B(\boldsymbol{c},R)$：球心为$\boldsymbol{c}$，半径为R的最小包含球；S_t：核心集；t：迭代次数；ε：终止参数。

2) 对于$\forall z$有$\varphi(z)\in B(\boldsymbol{c}_t,(1+\varepsilon)R)$，则转到步骤6)，否则转到步骤3)；

3) 如果$\varphi(z)$距离球心c_t最远，则$S_{t+1}=S_t\cup\{\varphi(z)\}$；

4) 寻找最新最小包含球$B(S_{t+1})$，并设置：$\boldsymbol{c}_t=\boldsymbol{c}_{B(S_{t+1})}$，$R_t=R_{B(S_{t+1})}$；

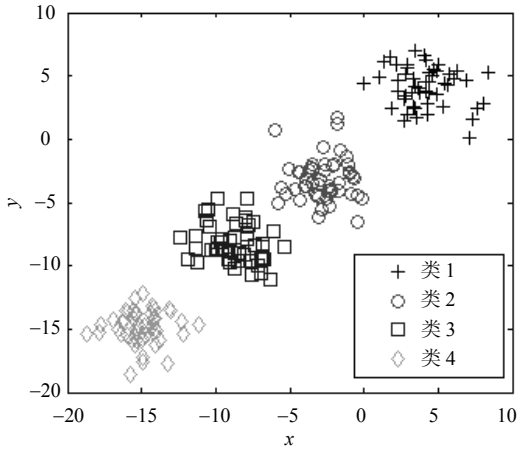
5) t=t+1并转到步骤2)；

6) GRPLM对核心集进行训练并得到决策函数。

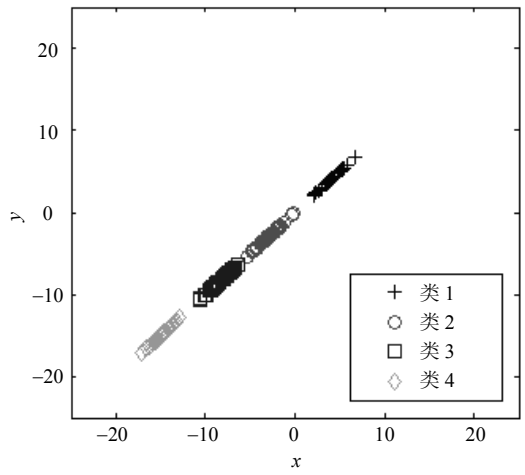
3 实验分析

3.1 人工数据集

人工生成4类服从Gaussian分布的数据集，各类样本50个，其中心点分别为(4,4)、(–3,–3)、(–9,–9)、(–15,–15)，均方差均为2.5。人工数据集如图1a所示。将上述数据集投影到GRPLM找到的方向向量可得图1b，GRPLM中参数v选取1。



a. 人工数据集



b. 实验结果

图1 人工数据集及实验结果

由图1b可以看出，GRPLM找到的方向向量能较好地保持原始数据的相对关系不变，且具有良好的可分性。

3.2 中小规模数据集

实验数据集见表1所示，分别选取实验数据集中各类样本的60%作为训练样本，剩余样本用作测试。

表1 中小规模数据集

数据集	样本数	类别数	样本维数
Wine	178	3	13
Iris	150	3	4
Liver	345	2	7
Glass	270	7	9
Pima	768	2	8

3.2.1. 核函数对分类结果的影响

GRPLM的分类性能受核函数选择的影响。本实验将Gaussian核函数、Polynomial核函数、Sigmoid核函数、Epanechnikov核函数，分别带入GRPLM核化形式中来考察GRPLM的分类性能。实验结果如图2所示。

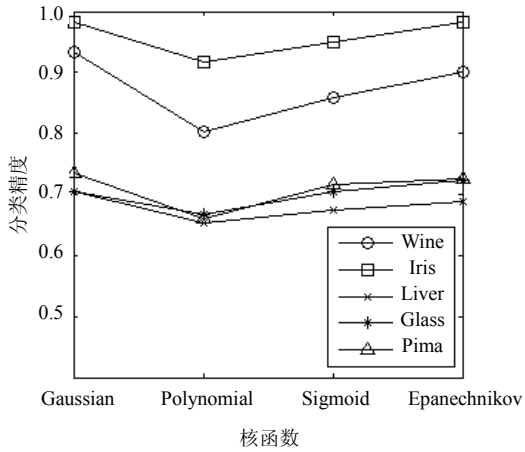


图2 核函数对分类结果的影响

由图2可以看出：在Wine、Liver、Glass、Pima数据集上，选取Gaussian核函数的GRPLM分类精度最优，选取Epanechnikov核函数的GRPLM分类精度次之，选取Sigmoid核函数和Polynomial核函数的GRPLM分类精度分别排在后两位；选取Gaussian核函数和Epanechnikov核函数的GRPLM在Iris数据集上分类精度相同且均具有最优的分类能力。

后续实验选取核函数的依据是：Sigmoid核函数在特定参数下与Gaussian核函数具有近似性能；Polynomial核函数参数较多Polynomial核函数参数较多且较难确定；Gaussian核函数和Epanechnikov核函数在实际中均广泛被使用。但从方便计算角度考虑，后续实验选用Gaussian核函数。

3.2.2　比较实验

本节通过与多类支持向量机[22]、K近邻算法比较实验验证GRPLM的有效性。本文实验K取5。Multi-class SVM的最优化表达式如下：

$$\min \ \frac{1}{2}\sum_{m=1}^{c}\|\boldsymbol{W}_{\boldsymbol{m}}\|^2+C\sum_{i=1}^{N}\sum_{m\neq y_i}\xi_i^m \tag{29}$$

$$\text{s.t.} \quad \boldsymbol{W}_{y_i}{}^{\mathrm{T}}x_i+b_{y_i}\geqslant \boldsymbol{W}_{\boldsymbol{m}}{}^{\mathrm{T}}x_i+b_m+2-\xi_i^m \tag{30}$$

$$\begin{aligned}&\xi_i^m\geqslant 0 \quad i=1,2,\cdots,c\\ &m,y_i\in\{1,2,\cdots,c\},m\neq y_i\end{aligned} \tag{31}$$

GRPLM分类精度与参数选取有关。参数通过网格搜索策略选择。Gaussian核函数的方差δ在网格$\{\bar{x}/2\sqrt{2},\bar{x}/2,\bar{x}/\sqrt{2},\bar{x},\sqrt{2}\bar{x},2\bar{x},2\sqrt{2}\bar{x}\}$中搜索选取，其中$\bar{x}$为训练样本平均范数的平方根；多类支持向量机中，惩罚因子C在网格{0.01, 0.05, 0.1, 0.5, 1, 5, 10}中搜索选取；GRPLM中，参数v在网格{0.1, 0.5, 1, 5, 10}中搜索选取。Gaussian核函数方差δ、Multi-class SVM的惩罚因子C、GRPLM的参数v均通过5倍交叉验证法获得。实验参数选取如表2所示，实验结果如表3所示。

表2　实验参数表

数据集	参数	
	Multi-class SVM	GRPLM
Wine	C=0.01, $\delta=\bar{x}/\sqrt{2}$	v=0.1, $\delta=\sqrt{2}\bar{x}$
Iris	C=0.01, $\delta=\bar{x}/\sqrt{2}$	v=0.5, $\delta=\bar{x}/2\sqrt{2}$
Liver	C=0.1, $\delta=\sqrt{2}\bar{x}$	v=0.1, $\delta=\bar{x}/2\sqrt{2}$
Glass	C=0.5, $\delta=\bar{x}/2$	v=1, $\delta=\bar{x}/2$
Pima	C=0.01, $\delta=\bar{x}/2\sqrt{2}$	v=0.1, $\delta=\bar{x}/2$

表3　中小规模数据集实验结果

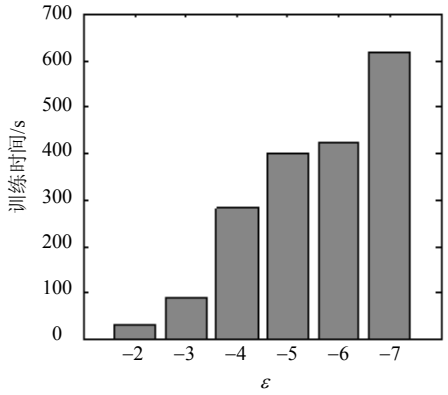
数据集	分类精度/%		
	Multi-class SVM	KNN	GRPLM
Wine	88.76	83.10	93.26
Iris	98.31	95.00	98.31
Liver	63.77	65.21	70.43
Glass	62.96	61.11	70.37
Pima	66.40	66.40	73.57
平均	76.04	74.16	81.19

由表3可以看出：从平均分类性能看，与Multi-class SVM和KNN相比，GRPLM在UCI数据集上具有更优的分类精度。具体而言，在Wine、Liver、Glass、Pima数据集上GRPLM的分类精度好于Multi-class SVM和KNN；在Iris数据集上GRPLM和Multi-class SVM分类精度相当且略高于KNN。综上所述，GRPLM在中小规模数据集上能较好地完成分类任务。

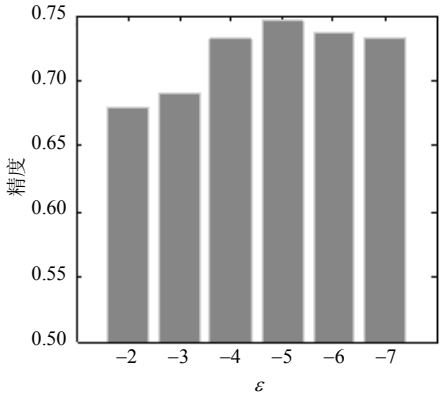
3.3　大规模数据集

3.3.1　终止参数ε对分类结果的影响

实验采用Bank数据集，实验选取60%的数据集用作训练，剩余的用于测试。CVM中的参数ε在网格$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$中选取。GRPLM-CVM的效率受到参数$\varepsilon$取值的影响。具体情况参如图3所示。



a. ε与样本训练时间的关系



b. ε与分类精度的关系

图3　终止参数ε对GRPLM-CVM的影响

图3反映的结论是参数ε影响到算法的训练时间以及分类精度。不失一般性，实验选取ε=10^{-6}。

3.3.2　GRPLM-CVM分类性能分析

实验采用文献[23]提供的数据集。实验训练样本分别取数据集的20%、40%、60%、80%，从剩下样本中任取500个作为测试样本。实验结果如表4所示。

表4　GRPLM-CVM分类结果

训练样本量/%	Abalone		Bank		California		Census	
	分类精度/%	时间/s	分类精度/%	时间/s	分类精度/%	时间/s	分类精度/%	时间/s
20	63.46	78.93	67.54	145.53	48.51	235.30	60.03	248.84
40	72.11	127.65	70.08	255.12	57.32	394.08	62.58	311.47
60	76.14	160.03	73.66	258.78	62.78	651.90	70.20	506.94
80	77.87	186.96	78.82	305.13	65.57	705.23	75.44	840.28

由表4可以看出：随着训练样本规模的增大，GRPLM-CVM分类精度和训练时间呈上升趋势，即训练样本规模影响GRPLM-CVM的分类性能。从分类效果看，GRPLM-CVM基本上能在有限时间内较好地完成分类任务。

4 结 论

针对当前分类方法面临的不足，提出基于流形判别分析的全局保序学习机GRPLM。该方法的主要优势在于：1) 进行分类决策时同时考虑样本的全局特征和局部特征；2) 具有保持样本相对关系不变的特性；3) 能在一定程度上解决传统分类器面临的大规模分类问题。与传统分类器的比较实验表明本文所提方法在分类性能方面具有一定优势。但GRPLM仍存在一定问题，如其分类能力对参数的选取较为依赖，如何提高参数选取效率对GRPLM分类性能的提升至关重要，这也是下一步的工作。

参 考 文 献

[1] QUINLAN J R. C4.5: Programs for Machine Learning[M]. San Francisco: Morgan Kaufmann Publishers, 1993.

[2] RASTOGI R, SHIM K. Public: a decision tree classifier that integrates building and pruning[C]//Proc of the Very Large Database Conference (VLDB). New York: [s.n.], 1998: 404-415.

[3] MEHTA M, AGRAWAL R, RISSANEN J. SLIQ: a fast scalable classifier for data mining[C]//Proc of International Conf Extending Database Technology(EDBT'96). France: [s.n.], 1996: 18-32.

[4] GEHRKE J, RAMAKRISHNAN R, GANTI V. Rainforest: a framework for fast decision tree construction of large datasets[J]. Data Mining and Knowledge Discovery, 2000, 4(2-3): 127-162.

[5] LIU B, HSU W, MA Y. Integrating classification and association rule[C]//Proc of the 4th International Conf on Knowledge Discovery and Data Mining. New York, USA: AAAI Press, 1998: 80-86.

[6] LI W M, HAN J, JIAN P. CMAR: Accurate and efficient classification based on multiple class association rules[C]//Proc of IEEE International Conf on Data Mining. Washington D C: IEEE Computer Society, 2001: 369-376.

[7] YIN X, HAN J. Classification based on predictive association rules[C]//SIAM International Conf on Data Mining. San Francisco: [s.n.], 2003: 331-335.

[8] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[9] 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.
DENG Nai-yang, TIAN Ying-jie. Support vector machine: Theory, algorithm and development[M]. Beijing: Science Press, 2009.

[10] PAL M, FOODY G M. Feature selection for classification of hyper spectral data by SVM[J]. IEEE Trans on Geoscience and Remote Sensing, 2010, 48(5): 2297-2307.

[11] SCHOLKOPF B, SMOLA A, BARTLET P. New support vector algorithms[J]. Neural Computation, 2000, 12: 1207-1245.

[12] SCHOLKOPF B, PLATT J, SHAWE-TAYLOR J, et a1. Estimating the support of high-dimensional distribution[J]. Neural Computation, 2001, 13: 1443-1471.

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

[14] 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.

[15] MANGASARIAN O, MUSICANT D. Lagrange support vector machines[J]. Journal of Machine Learning Research, 200l(1): 161-177.

[16] SUYKENS J A, VANDEWALLE J. Least squares support vector machines classifiers[J]. Neural Processing Letters, 1999, 19(3): 293-300.

[17] LEE Y J, MANGASARIAN O. SSVM: a smooth support vector machines[J]. Computational Optimization and Applications, 2001, 20(1): 5-22.

[18] LANGLEY P, SAGE S. Introduction of selective Bayesian classifier[C]//Proc of the 10th Conf on Uncertainty in Artificial Intelligence. Seattle: Morgan Kaufmann Publishers, 1994: 339-406.

[19] ZHENG Z H, WEBB G I. Lazy Bayesian rules[J]. Machine Learning, 2000, 32(1): 53-84.

[20] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learing, 1997, 29(2): 131-163.

[21] 刘忠宝, 潘广贞, 赵文娟. 流形判别分析[J]. 电子与信息学报, 2013, 35(9): 2047-2053.
LIU Zhong-bao, PAN Guang-zhen, ZHAO Wen-juan. Manifold-based discriminant analysis[J]. Journal of Electroincs & Information Technology, 2013, 35(9): 2047-2053.

[22] WESTON J, WATKINS C. Multi-class support vector machines[R]. London: Department of Computer Science, Royal Holloway University of London Technology, 1998.

[23] LIN L, LIN H T. Ordinal regression by extended binary classification[J]. Advanced in Neural Information Processing Systems, 2007, 19: 865-872.

编 辑 蒋 晓