基于特征贡献度加权高斯核函数的粗糙 one-class 支持向量机

田浩兵[1,2] 朱嘉钢[1,2] 陆 晓[2]
(江南大学物联网工程学院 无锡 214122)[1] (江南大学晓山股份联合实验室 无锡 214122)[2]

摘 要 粗糙 one-class 支持向量机(ROCSVM)是一种一类支持向量机,它通过核函数映射,定义上近似超平面和下近似超平面,使得训练样本能根据在粗糙间隔中的位置,自适应地对决策超平面产生影响。由于 ROCSVM 训练集只有正类样本,因此充分挖掘和利用训练样本的分类特征对于提高 ROCSVM 的分类性能有重要意义。为此,提出了一种基于训练样本分类特征贡献度的加权高斯核函数(λ-RBF):先对训练样本做主成分分析(PCA)得到按特征值排序的向量集,以此向量集构造核函数,使得特征值较大的维度在核函数中起较大的作用。在 UCI 标准数据集和仿真数据上的实验结果表明:与一般 RBF 的 ROCSVM 相比,基于 λ-RBF 的 ROCSVM 有着更好的泛化性和更高的识别率。

关键词 粗糙集,一类支持向量机,加权核函数,主成分分析,超平面,过拟合
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.050

WFCD-based Rough Set One-class Support Vector Machine

TIAN Hao-bing[1,2] ZHU Jia-gang[1,2] LU Xiao[2]
(School of IoT Engineering, Jiangnan University, Wuxi 214122, China)[1]
(Co-Laboratory in Hillsun Ltd. of Jiangnan University, Wuxi 214122, China)[2]

Abstract Rough one-class support vector machine(ROCSVM) is a single class SVM. It defines upper approximation and lower approximation hyperplanes by a kernel function mapping, which makes the training samples have an impact on the decision hyperplane adaptively according to the position within the rough margin. Since the ROCSVM only has positive samples, to fully exploit and use the features of the classified training samples have important significance for improving the classification performance of ROCSVM. Thus, we presented a weighted feature-contribution-degree(WFCD) based Gaussian kernel(λ-RBF). First, principal component analysis(PCA) is done to the training set to get vector set sorted by eigenvalues, and then kernel function is constructed based on the vector set, which makes a larger eigenvalue have better effect in the kernel function. Experimental results on UCI standard data sets and simulation data show that compared with the general RBF-based ROCSVM, the λ-RBF based ROCSVM has better generalization and higher recognition rate.

Keywords Rough set, One-class SVM, Kernel function, PCA, Hyperplane, Over-fitting

1 引言

支持向量机(Support Vector Machine, SVM)是最初由 Vapnik-Chervonenkis 等提出的一种用于解决二分类问题的学习算法,它在小样本数据集中显示出特有的优势,并且通过引入核函数,将原始空间中的非线性问题转化为特征空间的线性问题来求解。SVM 是基于结构风险最小化原则的机器学习方法,具有收敛到全局最优、维数不敏感、泛化能力强的优点。

在故障诊断、人脸识别、网络异常检测和文本分类等应用领域,常常遇到如下的 one-class 问题:容易获取大量的目标类模式(或正常数据),但获取少量的非目标模式(或异常数据)非常困难或代价很高。分类的任务是寻找一个最优决策函数以便描述目标类模式所属的区域,并能够将目标类与所有非目标类分割开。Schölkopf 等人将 SVM 推广到 one-class 问题,巧妙地构造最优决策超平面将目标类样本和坐标原点分割开。此外,文献[1,2]提出了类似思想,构造非线性映射空间中的超球体来覆盖目标类样本。

传统的 one-class SVM 对离群点敏感。如文献[3]指出,one-class SVM 采用的 Hinge 损失函数使得它们容易对 outlier产生过拟合现象。针对 one-class 支持向量机过拟合问题,文献[5]将粗糙集理论引入到 one-class 支持向量机,提出了粗糙集 one-class SVM,它能够有效抑制由于 outlier 引起的过拟合问题。

核函数一直是支持向量机重要的研究领域,核函数的选择对支持向量机的分类效果有重要的影响。由于 ROCSVM 训练集只有正类样本,因此核函数的选择对 ROCSVM 的影响尤为重要。基于这个思想,提出了特征贡献度加权高斯核

到稿日期:2014-07-10 返修日期:2014-10-06 本文受江苏省产学研项目(BY2013015-40)资助。
田浩兵(1988—),男,硕士生,主要研究方向为模式识别、计算机应用,E-mail:howingtian@163.com;朱嘉钢(1957—),男,博士,副教授,主要研究方向为人工智能与模式识别、软件工程;陆 晓(1967—),男,主要研究方向为物联网工程。

函数。对于训练集 X，先对 X 作 PCA 分析，得到训练集的协方差矩阵 $C=PAP^{\mathrm{T}}$，其中 P 是 C 的特征向量矩阵，A 是 C 的特征值矩阵，由此得到训练集 X 的变换矩阵为 $X'=P^{\mathrm{T}}X$。在 λ-RBF 中，核函数由 X' 中各维对应的 RBF 的线性组合构成，且根据 X' 每维对应的特征值的大小分配该维在核函数中的权重，使得特征值较大的维度在核函数中起较大的作用。在 UCI 标准数据集和仿真数据上的实验结果表明：较一般 RBF 的 ROCSVM，基于 λ-RBF 的 ROCSVM 有更好的泛化性和更高的识别率。

2 粗糙集 one-class SVM 和高斯核函数

2.1 粗糙集 one-class SVM

粗糙集 one-class SVM 定义决策超平面的上近似超平面 $H(\omega,\rho_u)$ 和下近似超平面 $H(\omega,\rho_l)$，并将坐标原点与它们之间的距离定义为粗糙间隔。位于上近似超平面之内的样本明确地属于目标类；位于上下近似超平面之间的样本可能是 outlier，也可能属于目标类；而位于下近似超平面之外的样本确定地为 outlier 样本。这样可以更合理地在寻找最优超平面的过程中对不同位置的点给予不同的惩罚。

为了寻找最大的粗糙间隔，求解如下的二次规划：

$$\begin{aligned}&\min_{\omega,\xi,\rho}\frac{1}{2}\|\omega\|^2-(\rho_l+\rho_u)+\frac{1}{vl}\sum_{i=1}^{l}(\xi_i+\delta\xi_i^*)\\&\text{s.t.}\ (\omega\cdot\phi(x_i))\geqslant\rho_u-\xi_i-\xi_i^*,\\&\quad 0\leqslant\xi_i\leqslant\rho_u-\rho_l,\\&\quad \rho_u,\rho_l\geqslant 0,\\&\quad \xi_i^*\geqslant 0,i=1,2,\cdots,l\end{aligned}\tag{1}$$

其中，ξ_i 和 ξ_i^* 是松弛变量，参数 $\delta>1$，表示对于在下边界的点有着更大的惩罚，因为这些点对分割面的影响比其他处的点更大。为了解决这个二次规划问题，首先介绍以下拉格朗日函数：

$$\begin{aligned}L=&\frac{1}{2}\|\omega\|^2-(\rho_u+\rho_l)+\frac{1}{vl}\sum_{i=1}^{l}(\xi_i+\delta\xi_i^*)-\sum_{i=1}^{l}\alpha_i((\omega\cdot\\&\phi(x_i))-\rho_u+\xi_i+\delta\xi_i^*)-\sum_{i=1}^{l}\beta_i(\rho_u-\rho_l-\xi_i)-\sum_{i=1}^{l}\gamma_i\xi_i-\\&\mu_1\rho_l-\mu_2\rho_u-\sum_{i=1}^{l}\eta_i\xi_i^*\end{aligned}\tag{2}$$

其中，$\alpha_i\geqslant 0,\beta_i\geqslant 0,\gamma_i\geqslant 0,\eta_i\geqslant 0,\mu_1\geqslant 0,\mu_2\geqslant 0$ 为拉格朗日乘子。根据 KKT 条件，参数满足以下条件：

$$\frac{\partial L}{\partial\omega}=\omega-\sum_{i=1}^{l}\alpha_i\phi(x_i)=0\tag{3}$$

$$\frac{\partial L}{\partial\xi_i}=\frac{1}{vl}-\alpha_i+\beta_i-\lambda_i=0\tag{4}$$

$$\frac{\partial L}{\partial\xi_i^*}=\frac{\delta}{vl}-\alpha_i-\eta_i=0\tag{5}$$

$$\frac{\partial L}{\partial\rho_l}=-1+\sum_{i=1}^{l}\beta_i-\mu_1=0\tag{6}$$

$$\frac{\partial L}{\partial\rho_u}=-1+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}\beta_i-\mu_2\tag{7}$$

$$\alpha_i((\omega\cdot\phi(x_i))-\rho_u+\xi_i+\xi_i^*)=0\tag{8}$$

$$\beta_i(\rho_u-\rho_l-\xi_i)=0\tag{9}$$

$$\gamma_i\xi_i=0,\eta_i\xi_i^*=0,\mu_1\rho_l=0,\mu_2\rho_u=0\tag{10}$$

可以得到这个二次规划的对偶问题如下：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j k(x_i,x_j)\\&\text{s.t.}\ 0\leqslant\alpha_i\leqslant\frac{\delta}{vl},\sum_{i=1}^{l}\alpha_i\geqslant 2\end{aligned}\tag{11}$$

求出对偶规划的最优解 $\alpha=(\alpha_1,\alpha_1,\cdots,\alpha_l)^{\mathrm{T}}$ 后，可以根据 KKT 互补条件式(8)－式(10)对训练样本 x_i 相对于粗糙间隔的位置进行分析。

a. 如果 $\alpha_i=0$，那么 x_i 位于 $H(\omega,\rho_u)$ 之内，满足 $\omega\cdot\phi(x_i)>\rho_u$，对应目标类点。

b. 如果 $0<\alpha_i<\frac{1}{vl}$，那么 x_i 位于 $H(\omega,\rho_u)$ 上，满足 $\omega\cdot\phi(x_i)=\rho_u$，对应上边界上的点，称为上近似界支持向量。

c. 如果 $\alpha_i=\frac{1}{vl}$，那么 x_i 位于 $H(\omega,\rho_u)$ 和 $H(\omega,\rho_l)$ 之间，满足 $\omega\cdot\phi(x_i)=\rho_u-\xi_i$，称为粗糙间隔支持向量。

d. 如果 $\frac{1}{vl}<\alpha_i<\frac{\delta}{vl}$，那么 x_i 位于 $H(\omega,\rho_l)$ 上，满足 $\omega\cdot\phi(x_i)=\rho_l$，对应下边界上的点，称为下近似界支持向量。

e. 如果 $\alpha_i=\frac{\delta}{vl}$，那么 x_i 位于 $H(\omega,\rho_l)$ 之外，满足 $\omega\cdot\phi(x_i)=\rho_l-\xi_i$，其为是非边界支持向量，一般与 outlier 相关。

假设样本 x_j 位于上近似超平面上，即 $0<\alpha_i<\frac{1}{vl}$，则可以由式(8)和式(10)得到：

$$\rho_u=\omega\cdot\phi(x_j)=\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\tag{12}$$

为了增强算法的健壮性，按以下公式计算 ρ_u 的值：

$$\rho_u'=\omega\cdot\phi(x_j)=\frac{1}{l_1}\sum_{j=1}^{l_1}\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\tag{13}$$

其中，l_1 表示落在上支持边界上点的数量，即对应于 $0<\alpha_i<\frac{1}{vl}$ 的点。

同样，可以得到

$$\rho_l'=\omega\cdot\phi(x_j)=\frac{1}{l_2}\sum_{j=1}^{l_2}\sum_{i=1}^{l}\alpha_i k(x_i,x_j)\tag{14}$$

其中，l_2 表示落在下支持边界上点的数量，即对应于 $\frac{1}{vl}<\alpha_i<\frac{\delta}{vl}$ 的点。

基于以上分析，将粗糙 one-class SVM 的决策函数定义为：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i k(x_i,x)-\rho_u')\tag{15}$$

$$f'(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i k(x_i,x)-\rho_l')\tag{16}$$

依据前文的分析，训练样本按照它们在粗糙间隔中的相对位置获得了不同的 α 值(可以分 5 种情况)，从而自适应地对决策超平面产生了影响，并共同决定了它在特征空间中的位置。

对于一个给定的测试点 x_t，从式(15)和式(16)获得值 $f(x_t)$ 和 $f'(x_t)$ 后，可以按以下定义得出样本点的类标签：

(1)如果 $f(x_t)>0$，则样本点 x_t 明确归入正类，即目标类。

(2)如果 $f'(x_t)<0$，则样本点 x_t 明确归入负类。

(3)如果 $f(x_t)<0$ 且 $f'(x_t)>0$，则样本点 x_t 可能属于负类，现有给出的信息不足以判定它的类别。

2.2 高斯核函数(RBF)

支持向量机主要考虑的问题是如何用 Margin 来训练线性机，但它依赖于数据的预处理，即在更高维数的空间中来表达模式，并且通常比原来的特征空间的维数高很多。这也是

现实模式分类中解决原始空间中样本线性不可分的问题而提出的，在支持向量机中，相似性和相似程度是用内积进行估价的，这些内积强烈依赖于映射的选择，选择不同的映射就意味着对相似性和相似程度的不同估价标准。所以在解决实际问题中，映射的选择是很重要的，支持向量机的核函数就反映了这种映射关系，核函数的选择直接影响到 SVM 分类器性能的优劣，常需要根据具体的问题构造相应的核。

在模式分类问题中，无需知道核函数的具体形式，只需知道样本在高维空间中存在的相似关系，即只知道样本之间的内积关系(核矩阵)。因此把核函数问题转化为核矩阵的问题来处理，这样，模式分类问题就变得比较直观且易于处理。

传统的 RBF 核函数的表达式如下：

$$k(x_i,x_j)=\mathrm{e}^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}} \tag{17}$$

对于给定训练样本 $X=\{x_i\mid x_i\in R^d,i=1,2,\cdots,l\}$，SVM 中对应的高斯核矩阵为：

$$K_{matrix}=\begin{bmatrix}k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_l)\\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_l)\\ \vdots & \vdots & \vdots & \vdots\\ k(x_l,x_1) & k(x_l,x_2) & \cdots & k(x_l,x_l)\end{bmatrix} \tag{18}$$

3 特征贡献度加权 RBF

3.1 主成分分析

PCA 是统计分析方法中的一种重要方法，它利用一小部分主成分来解释数据协方差结构，而这些主成分是原始变量的线性组合。

已知 $x=\{x_1,x_2,\cdots,x_n\}$ 是 d 维随机向量，期望 $p=E(x)$，它由样本的平均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}(x_i)$ 估计；协方差矩阵 $C=Cov(x)$，它由样本的协方差阵 $S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}$ 估计。协方差矩阵可被分解成 $C=PAP^{\mathrm{T}}$，其中 $P=(p_1,p_2,\cdots,p_d)$ 是 C 的特征向量矩阵(也称因子载荷矩阵)，$A=diag(\lambda_1,\lambda_2,\cdots,\lambda_d)$ 是 C 的特征值矩阵。则 x 的主成分转换矩阵定义为 $x'=P^{\mathrm{T}}(x-p)$，主成分 x_i' 的贡献率为 $\lambda_i/\sum_{k=1}^{p}\lambda_k$，$i=1,2,\cdots,d$。

3.2 特征贡献度加权 RBF

支持向量机超平面的定义依赖核函数的映射，核函数的选择和变化对支持向量机分类准确率的影响很大。鉴于此，在传统 RBF 的基础上对核函数加以改进，定义适合粗糙集 one-class SVM 的核函数。

随机生成两类 15 维符合高斯分布的数据点集 $X_1\sim N(3.5,1)$ 和 $X_2\sim N(1,1)$，对其作 PCA 分析，分别绘出映射后的各样本点在特征贡献度最高的两个维度和特征贡献度最低的两个维度的分布情况，如图 1、图 2 所示。

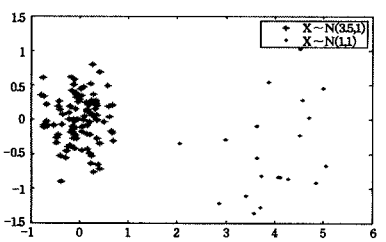

图 1 样本点在对应特征值最高的两个维度中的分布

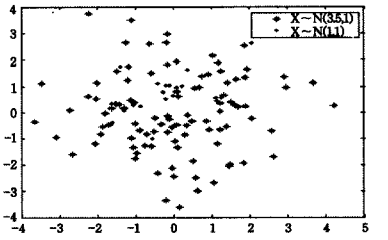

图 2 样本点在对应特征值最低的两个维度中的分布

从图 1、图 2 中可以看出，对于贡献度较高的维度，两类样本的区分度非常高，当贡献度较低时，样本就较难区分了。基于这个特性，提出了空间特征加权 RBF。对于 PCA 变换后的数据集 $X=\{x_i\mid x_i\in R^d,i=1,2,\cdots,l\}$，$\lambda$-RBF 的定义为：

$$k(x_i,x_j)=C_1\mathrm{e}^{(x_i^1-x_j^1)^2}+C_2\mathrm{e}^{(x_i^2-x_j^2)^2}+\cdots+C_d\mathrm{e}^{(x_i^d-x_j^d)^2} \tag{19}$$

其中，x_i^d、x_j^d 表示 x_i、x_j 的第 d 维，定义向量 λ 为数据集 X 经过主成分分析(PCA)后得到的特征值向量，特征值从大到小排列，$C_i(i=1,2,\cdots,d)$ 的定义如下：

$$C_i=\frac{d\lambda_i}{\sum_{j=1}^{d}\lambda_j} \tag{20}$$

根据文献[6]，核函数有如下性质：

若 $k_1,k_2,\cdots,k_n$ 是核函数，则

(1)$\alpha k_i,\alpha\geqslant 0,i=1,2,\cdots,n$ 是核函数；

(2)$k_1+k_2+\cdots+k_i(i=1,2,\cdots,n)$ 是核函数。

由以上性质可知，λ-RBF 为有效的核函数，说明如下。显然

$$\begin{aligned}&k_1(x_i^1,x_j^1)=\mathrm{e}^{-\frac{1}{2\sigma^2}(x_i^1-x_j^1)^2}\\ &k_2(x_i^2,x_j^2)=\mathrm{e}^{-\frac{1}{2\sigma^2}(x_i^2-x_j^2)^2}\\ &\quad\vdots\\ &k_d(x_i^d,x_j^d)=\mathrm{e}^{-\frac{1}{2\sigma^2}(x_i^d-x_j^d)^2}\end{aligned} \tag{21}$$

均为输入向量为一维的高斯核函数。由于 $-\frac{1}{2\sigma^2}$ 为常数，式(21)可表示为：

$$\begin{aligned}&k_1(x_i^1,x_j^1)=C_1\mathrm{e}^{(x_i^1-x_j^1)^2}\\ &k_2(x_i^2,x_j^2)=C_2\mathrm{e}^{(x_i^2-x_j^2)^2}\\ &\quad\vdots\\ &k_d(x_i^d,x_j^d)=C_d\mathrm{e}^{(x_i^d-x_j^d)^2}\end{aligned} \tag{22}$$

根据上述核函数性质式(1)，以上各式均为核函数，其中 C_i 为可调节的参数常量；再根据上述性质式(2)，将逐式相加所得到的式(23)也是核函数。

$$k(x_i,x_j)=C_1\mathrm{e}^{(x_i^1,-x_j^1)^2}+C_2\mathrm{e}^{(x_i^2,-x_j^2)^2}+\cdots+C_d\mathrm{e}^{(x_i^d,-x_j^d)^2} \tag{23}$$

核函数反映了输入数据之间的相似性，在 λ-RBF 的每一项中，对 x_i,x_j 的每一维单独计算相似性，并且为贡献率较大的维度分配较大的权重。

4 实验与分析

为了检验本文提出的 λ-RBF 粗糙 one-class 支持向量机在处理分类问题时的性能，在仿真数据和 UCI 数据库的 Pen-based Handwritten Digits(PHD)数据集上进行了数值实验。

4.1 仿真数据实验

随机生成两类符合高斯分布的数据点集：$X_1\sim N(3.5,1)$

和 $X_2 \sim N(1,1)$，X_1 中有 100 个数据样本，X_2 中有 20 个数据样本。在训练模型的过程中，将 X_1 作为目标类即正类，将 X_2 作为负类。如图 3 所示，x 轴表示生成样本的维度，y 轴表示分类的平均识别率，有 * 的线表示使用 λ-RBF 后的 Rough one-class SVM 的识别率，有。的线表示使用传统 RBF 后的 one-class SVM 的识别率，两者均是在数据经过 PCA 后取贡献率大于 90%的维度分类，有 · 的线表示没有经过 PCA 的数据进行 Rough one-class SVM 分类。

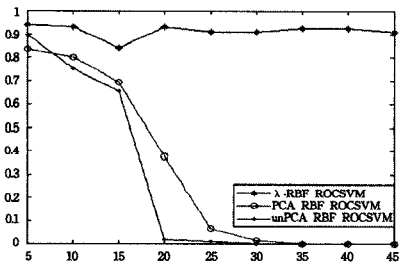

图 3 λ-RBF ROCSVM、传统 RBF ROCSVM 及无 PCA RBF ROCSVM 识别性能的对比

从图 3 可以看出，在数据维度小于 15 维时，λ-RBF ROCSVM 相对于其他两类 SVM 的优势并不明显，而 PCA RBF ROCSVM 和 unPCA RBF ROCSVM 的分类性能相似，并且分类性能都随着维度的增加逐渐下降；当维度超过 15 维时，后两种分类算法迅速下降，而 λ-RBF ROCSVM 的分类性能没有太大浮动，特别是当维度超过 20 维后，从图中可以看到后两种的分类性能已经很低了。因此可以得知，当分类数据为高维度样本时，λ-RBF ROCSVM 对样本数据的维度有着很强的泛化性，λ-RBF one-class SVM 分类性能不随样本数据维度的变化而产生太大的波动，而基于传统 RBF 函数的 Rough one-class SVM 和不经过 PCA 的 Rough one-class SVM 识别性能在处理高维数据时分类性能较差。

4.2 UCI 标准数据实验

为了进一步验证 λ-RBF 在 Rough one-class SVM 的分类问题中有着良好的性能，在 UCI 数据库的 PHD 数据集上进行了数值实验。数据集的基本情况如表 1 所列，实验共包含 10 类数据样本，每个样本由 16 个特征描述。

表 1 PHD 数据集的基本信息

	类别									
	0	1	2	3	4	5	6	7	8	9
训练集	780	779	780	719	780	720	720	778	719	719
测试集	363	364	364	336	364	335	336	364	336	336

实验过程中对每一类样本分别用 RBF 核函数和 λ-RBF 核函数 Rough one-class SVM 训练分类器，两者均是在数据经过 PCA 后取贡献率大于 90%的维度分类，同样还用未经过 PCA 的原始数据作分类比较；此外分类器的参数 υ 和 δ 以及核参数 σ 均采用 5-fold 交叉验证的方法选取。表 2 给出了基于 RBF 和 λ-RBF 以及没有经过 PCA 的 Rough one-class SVM 在 PHD 数据集上的实验结果，分别统计每类识别过程中的目标样本识别率(PR)、异常样本识别率(NR)以及总的样本的识别率(AR)。实验中每种模型使用 5-fold 交叉验证选取的参数如表 3 所列。

表 2 RBF 和加权 RBF 型 one-class SVM 分类识别率对比(%)

类别	λ-RBF ROCSVM			PCA RBF ROCSVM			unPCA RBF ROCSVM		
	AR	PR	NR	AR	PR	NR	AR	PR	NR
0 类	96.68	84.30	98.12	95.74	84.02	97.10	94.57	81.82	96.04
1 类	90.57	91.76	90.43	71.07	79.40	70.10	72.76	83.57	71.51
2 类	92.97	91.21	93.17	90.85	89.84	90.97	89.77	91.76	89.53
3 类	96.08	94.35	96.27	92.05	95.83	91.65	93.48	87.09	91.07
4 类	95.77	72.25	98.50	90.99	79.95	92.28	94.94	84.89	96.11
5 类	93.42	83.58	94.47	88.19	85.97	88.43	85.16	91.64	84.48
6 类	96.68	82.44	98.20	92.82	81.85	93.99	92.22	83.93	93.11
7 类	92.65	72.80	94.96	83.25	85.71	82.96	91.42	84.34	92.25
8 类	93.17	78.57	94.72	83.25	78.27	83.78	86.36	90.48	85.93
9 类	94.03	77.68	95.76	85.16	82.14	85.48	88.71	83.33	89.28

表 3 5-fold 交叉验证参数选取

类别	λ-RBF ROCSVM			PCA RBF ROCSVM			unPCA RBF ROCSVM		
	δ	υ	σ	δ	υ	σ	δ	υ	σ
0 类	12	0.25	−3	4	0.05	0	9	0.05	2
1 类	4	0.1	−4	8	0.45	0	16	0.45	0
2 类	64	0.2	−3	4	0.05	0	64	0.05	1
3 类	32	0.05	−1	32	0.05	8	4	0.05	10
4 类	4	0.05	−5	8	0.35	0	8	0.05	2
5 类	8	0.1	−3	32	0.1	0	8	0.3	2
6 类	32	0.1	−4	2	0.05	1	16	0.05	1
7 类	2	0.05	−4	16	0.05	1	8	0.15	1
8 类	64	0.3	−5	64	0.65	0	32	0.1	2
9 类	2	0.1	−4	4	0.1	1	8	0.2	1

观察实验结果可以发现，基于 λ-RBF 的 Rough one-class SVM 总的测试样本平均识别率为 94.2%，而 PCA RBF ROCSVM 和 unPCA RBF ROCSVM 总的平均识别率分别为 87.34%和 88.94%。在正类样本识别方面，λ-RBF Rough one-class SVM 对于不同的测试类样本分类性能有一定的波动，但在整体上优于其他两类算法。在对于异常类的检测方面，从实验数据可以看出，λ-RBF Rough one-class SVM 有着较强的稳定性，平均识别率为 95.46%，而后两者的平均识别率分别为 87.67%和 88.93%，从 UCI 数据集实验可以看出，λ-RBF Rough one-class SVM 的分类性能优于其他两种算法。同时，在网格搜索法参数寻优时也发现，对于每一组参数，基于 λ-RBF 的 Rough one-class SVM 有着相对稳定的识别率。

结束语 本文根据核函数特点，在粗糙集 one-class SVM 和 RBF 的基础上构造了一种特征贡献度加权高斯核函数，这种核函数能充分挖掘和利用训练样本的分类特征，增强分类性能。针对一分类问题，λ-RBF 除在 ROCSVM 上有性能的提高外，还可以推广到传统的 one-class SVM。如何将其推广到其他的核函数，将是下一步的重点研究内容。

参 考 文 献

[1] Tax D M J, Duin R P W. Support Vector Data Description [J]. Machine Learning, 2004, 54(1): 45-66

[2] Campbell C, Bennett P. A Linear Programming Approach to Novelty Detection [M]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001

(下转第 246 页)

BLOG上的加速比优化较大，在处理器数目为16时，有36%的优化，在ENRON上有19%的优化。

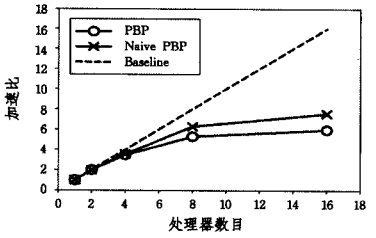

图6 KOS数据集上的加速比

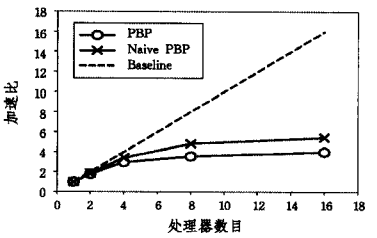

图7 BLOG数据集上的加速比

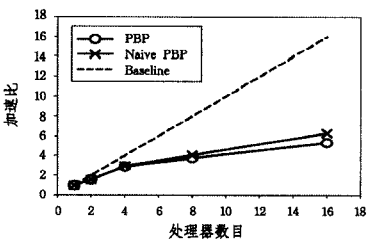

图8 ENRON数据集上的加速比

结束语 本文介绍了并行LDA模型的一般架构，提出朴素并行LDA算法，从计算时间和通信时间两个角度对并行LDA模型进行分析，给出了基于单词影响因子、阈值，以及增大通信间隔的方法来改善并行LDA在时间上的消耗。通过对比实验，选择最优的阈值参数以及通信间隔参数，使得并行LDA的加速比显著提高，并在精度上得到一定保证，精度损失在1%以下。

参考文献

[1] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[2] Hofmann T. Probabilistic latent semantic indexing[C]// Special Inspector General for Iraq Reconstruction. 1999: 50-57

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[C]// Neural Information Processing Systems. 2001: 601-608

[4] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004, 101(1): 5228-5235

[5] Zeng J, Cheung W K, Liu J. Learning topic models by belief propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5): 1121-1134

[6] 都志辉，等. 高性能计算并行编程技术——MPI并行程序设计[M]. 北京：清华大学出版社，2001
Du Zhi-hui, et al. High performance computing parallel programming technology——MPI parallel program design[M]. Peking: Tsinghua University Press, 2001

[7] Newman D, Asuncion A U, Smyth P, et al. Distributed inference for latent dirichlet allocation[C]// Neural Information Processing Systems. 2007

[8] Asuncion A U, Smyth P, Welling M. Asynchronous distributed learning of topic models[C]// Neural Information Processing Systems. 2008: 81-88

[9] Wang Y, Bai H, Stanton M, et al. Plda: Parallel latent dirichlet allocation for large-scale applications[C]// AAIM. 2009: 301-314

[10] Liu Z, Zhang Y, Chang E Y, et al. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing [J]. ACM TIST, 2011, 2(3): 1-18

[11] Zhai K, Boyd-Graber J L, Asadi N, et al. lda: a flexible large scale topic modeling package using variational inference in mapreduce[C]// WWW. 2012: 879-888

[12] Yan F, Xu N, Qi Y. Parallel inference for latent dirichlet allocation on graphics processing units[C]// Neural Information Processing Systems. 2009: 2134-2142

(上接第242页)

[3] Wang L, Jia H D, Li J. Training Robust Support Vector Machine with Smooth Ramp Loss in the Primal [J]. Neurocomputing, 2008, 71: 3020-3025

[4] Lin C F, Wan g S D. Fuzzy Support Vector Machine [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471

[5] Xu Yi-tian, Liu Chun-mei. A rough margin-based one class support vector machine[J]. Neural Comput & Applic, 2013, 22: 1077-1084

[6] Bishop C M. Pattern Recognition and Machine Learning [M]. Cambridge: Springer, 2007: 291-320

[7] 王磊，杨一帆，周启海. 粗糙one-class支持向量机[J]. 计算机科学，2009，36(9)：242-245
Wang Lei, Yang Yi-fan, Zhou Qi-hai. Rough Set based One-class Support Vector Machine[J]. Computer Science, 2009, 36(9): 242-245

[8] 秦玉平，王祎，伦淑娴，等. 基于超椭球支持向量机的兼类文本分类算法[J]. 计算机科学，2013，40(11A)：98-100
Qin Yu-ping, Wang Yi, Lun Shu-xian, et al. Multi-label Text Classification Algorithm Based on Hyper Ellipsoidal SVM[J]. Computer Science, 2013, 40(11A): 98-100

[9] Jeong Y-S, Kang I-H, Jeong M-K, et al. A New Feature Selection Method for One-Class Classification Problems [J]. IEEE Transactions on Systems, Man, and Cybernetics—part c: Applications and Reviews, 2012, 42(6): 1500-1509

[11] Asharaf S, Shevade S K, Murty M N. Rough support vector clustering[J]. Pattern Recogn, 2005, 38(10): 1779-1783

[12] Xu Y. Classification algorithm based on feature selection and samples selection [J]. 6th International Symposium on Neural Networks(ISNN 2009). 2009

[13] Bicego M, Figueiredo M A T. Soft clustering using weighted one-class support vector machines [J]. Pattern Recognition, 2009, 42(1): 27-32

[14] Schölkopf B, Platt J, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution [J]. Neural Comput, 2001, 13(7): 1443-1471

[15] Junejo I N, Bhutta A A, Foroosh H. Single-class SVM for dynamic scene modeling [J]. Signal, Image and Video Processing, 2013, 7(1): 45-52

[16] Shawe-Taylor J, Cfistianini N. Kernel methods for pattern analysis [M]. Cambridge: Cambridge University Press, 2004