基于全局和局部保持的半监督支持向量机

皋 军[1,2,3],王士同[1,3],邓赵红[1,3]
(1.江南大学信息工程学院,江苏无锡 214122; 2.盐城工学院信息工程学院,江苏盐城 224001;
3.浙江大学 CAD&CG 国家重点实验室,浙江杭州 310027)

摘 要: 支持向量机(SVM)作为正则化方法的一个特例在模式识别领域得到了成功地运用,然而传统的SVM方法作为一种有监督的学习方法主要依据最大间隔原则得到决策超平面的法向量,而并没有充分考虑样本内在的几何结构以及所蕴含的判别信息.因此,本文将线性判别分析(LDA)的类内散度和保局投影(LPP)的基本原理引入到SVM中,提出基于全局和局部保持的半监督支持向量机:GLSSVM,该方法在继承传统的SVM方法的特点的基础上,充分考虑样本间具有的全局和局部几何结构,体现样本间所蕴含的局部和全局判别信息,同时满足作为半监督方法的必须依据的一致性假设,从而在一定程度上提高了分类精度.通过在人造数据集和真实数据集上的测试表明该方法具有上述优势.

关键词: 支持向量机; 保局投影; 线性判别分析; 半监督; 一致性假设
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2010) 07-1626-08

Global and Local Preserving Based Semi-supervised Support Vector Machine

GAO Jun[1,2,3], WANG Shi-tong[1,3], DENG Zhao-hong[1,3]
(1. *School of Information Engineering, Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
2. *School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu* 224001, *China*;
3. *State Key Lab. of CAD&CG, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*)

Abstract: The support vector machine (SVM), as one of special regularization methods, has been used successfully in the field of pattern recognition. However, the traditional SVM, a supervised learning method, gets the normal vector of the decision boundary mainly according to the largest interval principle but has not considered the underlying geometric structure and the discriminant information fully. Therefore, a global and local preserving based semi-supervised support vector machine: GLSSVM, is presented in this paper by introducing the basic theories of the locality preserving projections (LPP) and the within-class scatter of linear discriminant analysis (LDA) into the SVM. This method inherits the characteristics of the traditional SVM, fully considers the global and local geometric structure between samples, shows the global and local underlying discriminant information and meets the consistency assumption which the semi-supervised method must coincides with so that the shortcomings of the supervised methods can be overcome and the classification accuracy can be increased. The tests on the artificial and real datasets show the above mentioned advantages of the GLSSVM method.

Key words: support vector machines; locality preserving projection; linear discriminant analysis; semi-supervised; consistency assumption

1 引言

支持向量机(Support Vector Machine, SVM)是数据挖掘中一项新技术[1,2],它借助于统计学习理论和最优化方法解决机器学习问题.该方法已经在众多的模式识别领域得到了成功地运用[3~7].然而传统的SVM作为一种有监督的学习方法,只能在少量的有标号的样本上进行学习,从而在一定程度上导致学习不太充分,同时该方法在学习过程中并没有充分考虑样本之间的几何结构和样本所隐含的判别信息,因此在一定程度上影响了该方法对具体模式进行识别的能力.

为了在一定程度上克服传统SVM方法训练不太充分的问题,文献[8,9]分别提出了直推式支持向量机(Transductive SVM, TSVM)、半监督支持向量机(Semi-Su-

收稿日期:2009-02-14;修回日期:2009-10-28
基金项目:国家863高技术研究发展计划(No.2007AA1Z158, No.2006AA10Z313);国家自然科学基金(No.60903100, No.90820002);国防应用基础研究基金(No.A1420461266);江苏省普通高校研究生科研创新计划(No.CX09B-175Z);浙江大学CAD&CG国家重点实验室开放课题(No.A0802)

per vised SVM, S^3VM).该类支持向量机作为半监督的学习方法,在学习过程中不但要训练有标号的样本,同时还要对大量的无标号的样本进行学习,从而在一定程度上避免了传统 SVM 方法训练不充分的弱点,但和传统 SVM 一样并没有充分考虑训练样本之间的几何结构和所蕴含的判别信息.

近来,为了有效揭示样本内部蕴含的局部几何结构,文献[10~13]分别提出了几种具有一定代表性的流形学习方法:等距映射(Isometric Mapping, Isomap)、局部线性嵌入(Locally Linear Embedding, LLE)和拉普拉斯特征映射(Laplacian Eigenmap, LE)、局部保持投影(Locality Preserving Projections, LPP).特别是 LPP 方法不但可以保持样本间局部几何结构,而且又可以克服其它几种方法难以在新的测试样本上获得低维的投影映射的问题[14],同时容易被非线性嵌入,从而发现高维非线性流形结构.为了充分发挥 LPP 方法长处,Belkin M 等人将 LPP 和传统的正则化方法相结合提出了流形正则化(Manifold Regularization, MR)框架[15],并在此框架下提出半监督支持向量机:拉普拉斯支持向量机(Laplacian Support Vector Machine, Lap-SVM)[15],该方法不但继承了传统 SVM 方法的优点,同时一定程度上克服了训练不充分的缺点,并且在学习过程中充分考虑了样本间的局部几何结构,体现了蕴含在样本中局部的鉴别信息.值得一提的是文献[16,17]通过分析得知线性判别分析(Linear Discriminant Analysis, LDA)[18]中的类内散度矩阵、类间散度和总体散度矩阵确实具有保持了训练样本全局的鉴别信息和全局的几何结构的能力,同时文献[19]指出 LDA 并不能由 LPP 完全替代.从这一层面上讲,Lap-SVM 在一定程度上没有充分考虑保持样本全局的几何结构和判别信息.

因此,本文将 LDA 中的类内散度和 LPP 的基本原理引入到 SVM 中提出基于全局和局部保持的半监督支持向量机(GLSSVM).该方法有如下优势:继承了传统 SVM 方法特色的同时还在一定程度上避免了学习不充分的缺陷,并且符合半监督学习方法必须依据的一致性假设[20];该方法是首次同时将 LDA 中的类内散度、LPP 的基本原理引入到支持向量机中,从而在一定程度上不但可以保持样本内在局部几何结构,同时还可以在一定程度上保持样本的全局几何结构,体现蕴含于样本的全局判别信息;该方法还可以很容易地进行非线性嵌入,得到非线性方法:Ker-GLSSVM,以发现具有高维非线性特征的流形结构.

2　流形正则化(MR)框架

传统的正则化方法[21](比如 SVM 等方法)已经得到了广泛的运用,然而通过研究发现该类方法更关注于分类函数的光滑性,而不太关注隐含于样本中的分类信息,从而在一定程度上没有充分利用样本内在的几何结构.因此,文献[15]将流形学习方法的观点引入到传统的正则化方法并结合再生核 Hilbert 空间(Reproducing Kernel Hilbert Space, RKHS)[22]相关性质构造了非线性的流形正则化(MR)框架:

$$\min_{f\in H_K}\frac{1}{l}\sum_{i=1}^{l}V(x_i,y_i,f(x_i))+\gamma_A\|f\|_K^2+\gamma_I\|f\|_I^2 \quad (1)$$

其中

损失函数 $V(x_i,y_i,f(x_i))=\begin{cases}0 & \text{if } y_if(x_i)\geqslant 1\\ 1-y_if(x_i) & \text{otherwise}\end{cases}$,

正则单元 $\|f\|_K^2$ 用于控制分类器的复杂性,而单元 $\|f\|_I^2$ 则反映的是样本分布的内在流形结构[15].

根据式(1)对应的流形正则化框架并结合 LPP 的基本原理,就可以得到非线性的半监督的拉普拉斯支持向量机(Lap-SVM)[15]所对应的原始优化问题:

$$\min_{\alpha\in R^{l+u},\xi\in R^l}\frac{1}{l}\sum_{i=1}^{l}\xi_i+\gamma_A\alpha^TK\alpha+\gamma_I\alpha^TKLK\alpha \quad (2)$$

$$\text{s.t. } y_i\Big(\sum_{j=1}^{l+u}\alpha_jK(x_i,x_j)+b\Big)\geqslant 1-\xi_i,\ i=1,\cdots,l$$

其中 l,u 分别代表训练样本中有标号样本和无标号样本数,$K(,)$表示 Mercer 核函数,L 是拉普拉斯矩阵[13],$K=(k_{ij})_{l+u,l+u},\forall\ k_{ij}=K(x_i,x_j)$.

然而式(2)在一定程度上并不能直接反映 Lap-SVM 方法具有的特色和不足,因此我们可以根据 Representer Theorems[22]容易得到 Lap-SVM 相对应的线性的形式:

$$\min_{\omega\in R^n,\xi\in R^l}\frac{1}{l}\sum_{i=1}^{l}\xi_i+\gamma_A\|\omega\|^2+\gamma_I\omega^TXLX^T\omega \quad (3)$$

$$\text{s.t. } y_i((\omega,x_i)+b)\geqslant 1-\xi_i,\ i=1,\cdots,l$$

其中 X 包括 l 个有标号样本和 u 个无标号样本,ω 是超平面的法向量.

从式(3)可以明显看出 Lap-SVM 方法只是在 C-SVM 方法的基础上增加了正则化单元 $\omega^TXLX^T\omega$,而该单元不但保证 Lap-SVM 方法在选择决策超平面时满足最大间隔的原则,同时保持了样本内在的局部流形结构.但式(3)也在一定程度上说明了 Lap-SVM 方法存在的缺限,即在一定程度上没有充分考虑保持样本内在的全局结构和全局信息.因此,本文将 LDA 中的类内散度矩阵和局部流形单元一并引入传统的 SVM,提出基于全局和局部保持的半监督支持向量机:GLSSVM,该方法在确定决策超平面时不但会考虑数据分布的局部结构,同时还会考虑保持各类类内最紧.

3　基于全局和局部保持的半监督支持向量机:GLSSVM

根据以上分析得知,因为 LDA 方法中的各类内散

度单元在一定程度上可以保持样本内在的全局几何结构和全局的判别信息[16,17],因此本文的 GLSSVM 方法通过引入 LDA 中的类内散度矩阵以达到在一定程度上保持样本内在的全局几何结构是合理的.

3.1 线性 GLSSVM 方法

定义 1[18] 假设有 l 个样本组成的样本集 $D=\{x_1,\cdots,x_l\}$, $\forall x_i\in R^n$,它们分别属于 2 个不同的类 (C_+,C_-),其中大小为 l_k 样本子集 D_k 属于第 k 类,给定分类决策平面的法向量 ω,则类内散度为:$\omega^T S_w\omega$.其中:类内散度矩阵 $S_w=\sum_{k=1}^{2}\sum_{x\in D_k}(x-u_k)(x-u_k)^{\mathrm{T}}$,均值 $u_k=\frac{1}{l_k}\sum_{x\in D_k}x,(k=1,2)$.

定义 2 假设有 l 个分属于 2 个不同的类 (C_+,C_-)有标号样本和 u 个无标号样本组成的样本集 $X_{l+u}=\{x_1,\cdots,x_l,x_{l+1},\cdots,x_{l+u}\}$,其中 l 个有标号样本对应的类内散度为:$\omega^T S_w\omega$,数据集 X_{l+u} 对应邻接图的拉普拉斯矩阵为 L[13],则线性 GLSSVM 方法对应的优化问题为:

$$\min_{\omega,b,\xi}\frac{1}{l}\sum_{i=0}^{l}\xi_i+\frac{\gamma_A}{2}\omega^T S_W\omega+\frac{\gamma_I}{2}\omega^T X_{l+u}LX_{l+u}^T\omega \tag{4}$$

$$\text{s.t. } y_i((\omega,x_i)+b)\geqslant 1-\xi_i,\ i=1,\cdots,l \tag{5}$$

其中,$\gamma_A\geqslant 0,\gamma_I\geqslant 0$.

定理 1 线性 GLSSVM 方法原始优化问题(4)、(5)对偶问题为:

$$\min_{\beta}\frac{1}{2}\beta^T H\beta-1^T\beta \tag{6}$$

$$\text{s.t. }\sum_{i=1}^{l}y_i\beta_i=0,0\leqslant\beta_i\leqslant\frac{1}{l} \tag{7}$$

其中 $H=(h_{ij})_{l\times l},h_{ij}=\gamma_i\gamma_j x_i^T(\gamma_A S_W+\gamma_I X_{l+u}LX_{l+u}^T)^+x_j,1=(1_1,\cdots,1_l)^{\mathrm{T}}$.

证明 式(4)、(5)对应的拉格朗日函数为:

$$\begin{aligned}L(\omega,b,\beta,\gamma)=&\frac{1}{2}\omega^T(\gamma_A S_W+\gamma_I X_{l+u}LX_{l+u}{}^T)\omega\\&+\frac{1}{l}\sum_{i=1}^{l}\xi_i\\&-\sum_{i=1}^{l}\beta_i(y_i(\omega^T x_i+b)-1+\xi_i)\\&-\sum_{i=1}^{l}\gamma_i\xi_i\end{aligned} \tag{8}$$

其中,$\beta=(\beta_1,\cdots,\beta_l),\gamma=(\gamma_1,\cdots,\gamma_l)$是拉格朗日系数.

根据 Karush-Kuhn-Tucker(KKT)[22]条件:

$$\frac{\partial L}{\partial\omega}=0$$

$$\Rightarrow\omega=\sum_{i=1}^{l}\beta_i\gamma_i(\gamma_A S_W+\gamma_I X_{l+u}LX_{l+u}^T)^+x_i \tag{9}$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{l}\beta_i\gamma_i=0 \tag{10}$$

$$\frac{\partial L}{\partial\xi_i}=0\Rightarrow\gamma_i=\frac{1}{l}-\beta_i \tag{11}$$

将式(9)、(10)、(11)回代到式(8)定理成立.

按照传统 SVM 方法计算偏移量的原理可以直接得到 GLSSVM 方法的偏移量 b:

$$b=\frac{1}{l}\sum_{i}^{l}(\gamma_i-\sum_{j=1}^{l}\gamma_j\beta_j x_j^T(\gamma_A S_W+\gamma_I X_{l+u}LX_{l+u}^T)^+x_i) \tag{12}$$

其中,$(\ )^+$ 表示广义逆矩阵.

由此,可以得到线性 GLSSVM 算法:

Input: l 个有标号样本 $\{(x_i,y_i)\}_{i=1}^{l}$, u 个无标号样本 $\{x_j\}_{j=l+1}^{l+u}$

Output: 分类决策函数 $g(x)$

Step1: 根据 $\{(x_i,y_i)\}_{i=1}^{l}$,通过定义 1 构造类内散度矩阵 S_W;

Step2: 根据 K-近邻方法构造 $l+u$ 个样本节点的邻接图 G,并计算得到权矩阵 W[13];

Step3: 定义拉普拉斯矩阵 $L=D-W$,其中 D 是对角矩阵,

$D_{ii}=\sum_{j=1}^{l+u}W_{ij}$;

Step4: 选择参数 γ_A,γ_I,根据定理 1 求解拉格朗日系数 β;

Step5: 根据式(9)计算决策超平面法向量 ω;

Step6: 根据式(12)计算偏移量 b;

Step7: 输出分类函数 $g(x)=\omega^T x+b$.

以上算法表明由于需要计算类内散度矩阵 S_W,GLSSVM 方法与 Lap-SVM 方法相比具有较高的空间复杂度($O(n^2)$)和时间复杂度($O(n^3)$),特别是在处理高维小样本数据时尤为明显.为了在一定程度上提高本文方法执行效率,在测试高维数据时首先使用 PCA 方法对数据进行相应的预处理.需要强调的是从式(9)可以看出决策超平面法向量 ω 不仅仅和有标号的样本相关,还决定于无标号样本,这一点在构造非线性 GLSSVM 方法时具有非常重要的意义.

3.2 非线性 GLSSVM: Ker-GISSVM

当样本内在几何结构呈现出高维非线性流形时,线性 GLSSVM 方法是没有办法得到非线性流形结构的,因此本文提出非线性 Ker-GLSSVM 方法.然而通过分析得知在非线性化过程中,必须要将类内散度矩阵 S_W 转化成图拉普拉斯矩阵的形式[16,17].

根据定义 1 假设 l 个有标号样本 $X_l=\{x_i\}_{i=1}^{l}$分属于 2 类,其中每类样本数为 l_k,则 l 个样本对应的 K-近邻邻接图 G 的权矩阵 W 按如下方式定义[16,17]:

$$W_{ij}=\begin{cases}1/l_k\text{ if }x_i\text{ and }x_j\text{ both belong to }k\text{ class}\\0\qquad\qquad\text{otherwise}\end{cases} \tag{13}$$

则 S_W 对应的图拉普拉斯矩阵形式为[16,17]:

$$S_W=X_l L_W X_l^T \tag{14}$$

其中,$L_W = I - W$,I为单位矩阵.

由此可以构造非线性的Ker-GLSSVM方法原始优化问题.

定理2　非线性Ker-GLSSVM方法的原始优化问题为:

$$\min_{\alpha,b,\xi} \frac{1}{l}\sum_{i=0}^{l}\xi_i + \frac{\gamma_A}{2}\alpha^T K_l L_W K_l^T \alpha + \frac{\gamma_I}{2}\alpha^T K_{l+u} L K_{l+u}\alpha \quad (15)$$

$$\text{s.t. } y_i\left(\sum_{j=1}^{l+u}\alpha_j K(x_i,x_j)+b\right)\geqslant 1-\xi_i,\ i=1,\cdots,l \quad (16)$$

其中 $K(,)$表示 Mercer 核函数

证明　假设 l 个有标号样本 $X_l=\{x_i\}_{i=1}^{l}$,u 个无标号样本 $X_u=\{x_j\}_{j=l+1}^{l+u}$,则 $X_{l+u}=(X_l,X_u)$.考虑一非线性函数 ϕ 将样本 x 投影到特征空间 $F(\phi: x\to\phi(x))$,而在解决实际问题时并不需要求解非线性函数 ϕ,则原始样本空间中的数据矩阵在特征空间 F 中可以重新表示为:

$$X_l^\phi=\{\phi(x_i)\}_{i=1}^{l},\ X_u^\phi=\{\phi(x_j)\}_{j=l+1}^{l+u},\ X_{l+u}^\phi=(X_l^\phi,X_u^\phi).$$

通过对线性 GLSSVM 方法的决策超平面法向量 ω 分析得知,法向量 ω 不但与有标号样本有关,同时也联系着无标号样本,并结合 Representer Theorems[22] 可以将特征空间中的非线性决策超平面法向量表示为:$\omega^\phi=\sum_{i=1}^{l+u}\alpha_i\phi(x_i)$,其中 $\alpha=(\alpha_1,\cdots,\alpha_{l+u})^{\mathrm{T}}$ 表示权值矢量.则特征空间中各正则单元可以表示为:

$$\begin{aligned}&\omega^{\phi T}S_W^\phi\omega^\phi=\alpha^T X_{l+u}^{\phi T}X_l^\phi L_W X_l^{\phi T}X_{l+u}^\phi\alpha=\alpha^T K_l L_W K_l^T\alpha\\&\omega^{\phi T}X_{l+u}^\phi L X_{l+u}^{\phi T}\omega^\phi=\alpha^T X_{l+u}^{\phi T}X_{l+u}^\phi L X_{l+u}^{\phi T}X_{l+u}^\phi\alpha=\alpha^T K_{l+u}LK_{l+u}\alpha\end{aligned} \quad (17)$$

其中 K_l 是$(l+u)\times l$矩阵,K_{l+u}是$(l+u)\times(l+u)$矩阵.

结合式(17)和式(4)、(5)定理成立.

定理 3　非线性 Ker-GLSSVM 方法原始优化问题(15)、(16)的对偶问题为:

$$\min_{\beta}\frac{1}{2}\beta^T H_{ker}\beta-1^T\beta \quad (18)$$

$$\text{s.t. }\sum_{i=1}^{l}y_i\beta_i=0,0\leqslant\beta_i\leqslant\frac{1}{l} \quad (19)$$

其中 $H_{ker}=YK_l^T(\gamma_A K_l L_W K_l^T+\gamma_I K_{l+u}LK_{l+u})^+K_lY$,$Y=diag(y_1,\cdots,y_l)$,$1=(1_1,\cdots,1_l)^{\mathrm{T}}$.

证明　(15)(16)对应的拉格朗日函数为:

$$\begin{aligned}L(\alpha,\beta,\gamma,\xi)=&\frac{1}{2}\alpha^T(\gamma_A K_l L_W K_l^T+\gamma_I K_{l+u}LK_{l+u})\alpha\\&+\frac{1}{l}\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\beta_i\left(y_i\left(\sum_{j=1}^{l+u}\alpha_j K(x_j,x_i)+b\right)-1\right.\\&\left.+\xi_i\right)-\sum_{i=1}^{l}\gamma_i\xi_i\end{aligned} \quad (20)$$

其中,$\beta=(\beta_1,\cdots,\beta_l)$,$\gamma=(\gamma_1,\cdots,\gamma_l)$是拉格朗日系数.

根据 KKT条件:

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{l}\alpha_i y_i=0 \quad (21)$$

$$\frac{\partial L}{\partial\xi_i}=0\Rightarrow\gamma_i=\frac{1}{l}-\beta_i \quad (22)$$

将式(21)、(22)代入式(20)得到新的拉格朗日函数:

$$\begin{aligned}L(\alpha,\beta,\gamma,\xi)=&\frac{1}{2}\alpha^T(\gamma_A K_l L_W K_l^T+\gamma_I K_{l+u}LK_{l+u})\alpha\\&-\sum_{i=1}^{l}\beta_i\left(y_i\left(\sum_{j=1}^{l+u}\alpha_j K(x_j,x_i)+b\right)-1\right)\\=&\frac{1}{2}\alpha^T(\gamma_A K_l L_W K_l^T+\gamma_I K_{l+u}LK_{l+u})\alpha\\&-\alpha^T K_l Y\beta+1^T\beta\end{aligned} \quad (23)$$

其中,$Y=diag(y_1,\cdots,y_l)$,$1=(1_1,\cdots,1_l)^{\mathrm{T}}$.

使用式(23)对权矢量求偏导数为:

$$\frac{\partial L}{\partial\alpha}=0\Rightarrow\alpha=(\gamma_A K_l L_W K_l^T+\gamma_I K_{l+u}LK_{l+u})^+K_lY\beta \quad (24)$$

将式(24)代入到式(23)定理成立.

依据同样原理可以得到 Ker-GLSSVM 方法的偏移量 b_{ker}为:

$$b_{ker}=\frac{1}{l}\sum_{i=1}^{l}\left(y_i-\sum_{j=1}^{l+u}\alpha_j K(x_j,x_i)\right) \quad (25)$$

因此,可以得到非线性 Ker-GLSSVM 方法:

Input:l 个有标号样本$\{(x_i,y_i)\}_{i=1}^{l}$,u 个无标号样本$\{x_j\}_{j=l+1}^{l+u}$

Output:分类决策函数 $f(x)$

Step1:根据$\{(x_i,y_i)\}_{i=1}^{l}$并使用式(13)、(14)构造类内散度矩阵 S_W 对应的图拉普拉斯矩阵 L_W;

Step2:根据 K-近邻方法构造 $l+u$ 个样本节点的邻接图 G,并计算得到权矩阵 W[13];

Step3:定义拉普拉斯矩阵 $L=D-W$,其中 D 是对角矩阵,$D_{ii}=\sum_{j=1}^{l+u}W_{ij}$;

Step4:选择适当的核函数 $K(,)$,构造矩阵 K_l、K_{l+u};

Step5:选择参数 γ_A,γ_I,根据定理3求解拉格朗日系数 β;

Step6:根据式(24)计算权值矢量 α;

Step7:根据式(25)计算偏移量 b_{ker}

Step8:输出分类函数 $f(x)=\sum_{j=1}^{l+u}\alpha_j K(x_j,x)+b_{ker}$.

需要说明的是半监督方法一般需要满足所谓的一致性(Consistency)假设[20],即:(a)相近的点可能具有相同的标号;(b)点与点具有相似的结构.由于本文的GLSSVM方法中引入了类内散度 $\omega^T S_w\omega$,从而使得类内样本在决策超平面上的投影更加紧密,也就是说当不同样本在决策超平面上的投影如果尽可能相近时,则可以认为这些样本属于同类,这一点显然满足假设(a);同时在本文方法中保留了保持流形结构单元 $\omega^T XLX^T\omega$,从而可以保证假设(b),这一点与 Lap-SVM 方法相同.因此,从这一层面上讲本文的方法是一种比较合理的半监督学习方法.

4 实验

为了说明本文方法在保持样本内在的全局和局部结构特点，将 GLSSVM 方法分别在人造数据集(团状数据集、流形结构数据集 two-moons)、真实数据集：UCI 数据集(http://www.ics.uci.edu/~mlearn/MLRepository.html)、人脸识别数据集(http://www.cs.uiuc.edu/homes/dengcai2/)进行测试，并与相应的方法进行比较.通过测试人造数据集来说明本文方法在寻找决策超平面过程中所依据的基本原理和基本方法；测试真实数据集来表明本文方法作为一种半监督分类方法的有效性，反映方法保持样本内在结构的效果.在测试过程中，为了更好地确定本文 GLSSVM、Ker-GLSSVM 的参数，本文使用交叉验证的方法.

4.1 测试人造数据集

人造数据集经常被用来测试算法效果[15,23]，本文使用两种不同风格(团状、流形)的人造数据集来说明本文方法构造决策超平面基本原理.

4.1.1 测试团状数据集

为了说明本文 GLSSVM 方法在寻找分类决策超平面时充分考虑了类内散度 $\omega^T S_w \omega$ 的作用，即在一定程度上可以保持样本内在的全局结构和全局鉴别信息，同时说明本文方法在分类决策超平面时同传统 SVM、Lap-SVM 方法的差异性.我们通过构造如图 1(a)的由 100 个样本组成的数据集，其中 40 个正类样本、60 个负类样本，同时将所有样本都作为有标号样本，即在有监督的情形下进行测试.为了更好地选取 K-近邻参数 k、核宽度参数 σ、参数 γ_A 和参数 γ_I，我们采用 10-折交叉验证.该测试过程分别设定上述参数的选取范围：$\sigma = \gamma_A = \gamma_I = [2^{\wedge}-7, 2^{\wedge}-4, 2^{\wedge}-1, 1, 2^{\wedge}3, 2^{\wedge}7, 2^{\wedge}10]$、$k = 1$.三种方法所对应的决策超平面如图 1($b$)、($c$)、($d$).

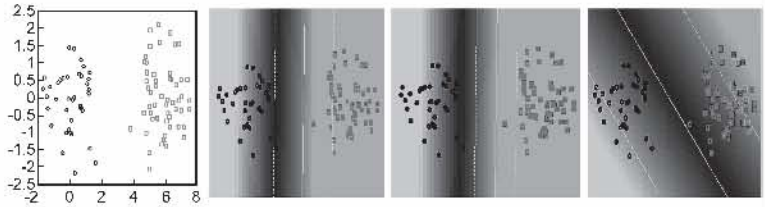


图1 三种方法对应的决策超平面

从图 1 可以看出：

(1)图 1(b)、(c)说明当测试数据集在一定程度上不具备较为明显的流形结构时，Lap-SVM 方法和传统 SVM 方法在选择决策超平面时依据的原则是相似的；

(2)图 1(c)和图 1(d)可以看出 GLSSVM 方法与 Lap-SVM 方法所得的决策超平面具有较为明显的不相似性.这说明在处理不具备较为明显的流形结构数据时，Lap-SVM 方法只能依据最大间隔的原则去构造相应的决策超平面法，而由于本文的 GLSSVM 方法中引入了类内散度矩阵，从而在构造决策超平面时不但会考虑最大间隔原则，同时充分考虑类内散度最小，即各类在决策超平面所对应的法向量上投影要尽量的紧密，以达到在某种程度上保持样本内在的局部几何结构的同时，还有效地保持了样本内在的全局结构和反映全局鉴别信息.

4.1.2 测试流形数据集(two-moons)

数据集 two-moons(见图 2、图 3)经常被用来测试一些流形学习方法[16,23]，图 2(a)、图 3(a)分别表示训练样本和预测样本.通过使用 Lap-SVM 和本文 Ker-GLSSVM 方法分别测试该数据集来说明 Ker-GLSSVM 方法在保持非线性流形结构和全局结构的性能.

实验设计：该训练样本集有 200 个 2 维数据组成(正类有 107 个样本，负类有 93 个样本)，在测试前随机各选取 40 个样本作为有标号样本集(即 $l = 80$)，剩余的 120 个样本作为无标号样本(即 $u = 120$).在测试上述两种方法的过程中，为了更好地选取 K-近邻参数 k、核宽度参数 σ、参数 γ_A 和参数 γ_I，同样使用 10-折交叉验证法，其中上述参数选取的范围为：σ、γ_A、γ_I 同实验 4.1.1、$k = [1,5,15,40,60]$.测试结果见图 2、图 3.

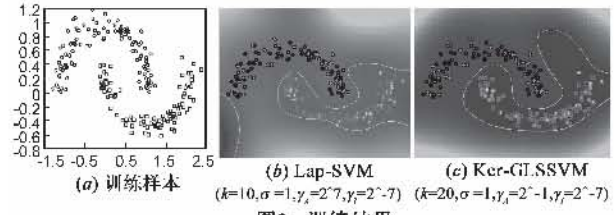


图2 训练结果

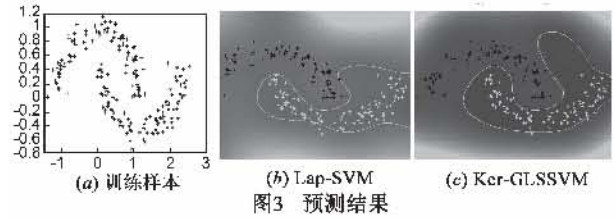


图3 预测结果

从图 2、图 3 可以看出：

(1)图 2 表明的是两种方法在做半监督学习所得到

的训练结果,其中图中重色点为有标号样本,白色的线表示各自算法对应的决策超平面.图 2 可以充分说明,两种方法在保持该数据集内在的局部非线性流形结构具有显著的效果.

(2)从图 2(b)、(c)并结合图 3(b)、(c)所表示的预测功能,表明 Ker-GLSSVM 方法在保持局部流形性能要略好于 Lap-SVM 方法.这是因为当将原始输入空间通过非线性映射到某一个特征空间(一般为 RKHS)时,一个根本目的就是把原线性空间中不可分问题转换成在特征空间中的线性可分问题,也就是说在特征空间中进行的操作还是线性的,因此,当把 two-moons 这种非线性可分的数据映射到某一个特征空间后,Ker-GLSSVM 方法、Lap-SVM 方法在该特征空间进行的还是线性变换.从这一层面上讲,由于本文的 Ker-GLSSVM 方法引进了类内散度,这就导致该方法在寻找决策超平面时不但要考虑数据内在的局部流形结构,同时还要保证类内散度最小,即一定程度上保持样本内在的全局几何结构和全局判别信息,而这一点 Lap-SVM 方法一定程度上是不具备的.

4.2　测试真实数据集

为了更全面地说明本文 GLSSVM、Ker-GLSSVM 作为一种半监督的分类方法具有的分类性能,在本阶段分别来测试 UCI 数据集、人脸识别数据集:Yale 数据集、ORL 数据集,同时与两种基于流形学习的半监督方法:SDA(Semi-Supervised Discriminant Analysis)[14]、Lap-SVM 进行对比,以增加本文算法性能的说服力.

4.2.1　测试 UCI 数据集

UCI 数据集经常被用来测试算法的分类精度[23],在该测试阶段我们抽取该数据集的 8 个 2 分类的数据子集:Ionoshere、Heart、Glass2、Sonar、Monk2、Liver_disorder、Auatralian、German_org 来分别测试线性、非线性 SDA,线性、非线性 Lap-SVM 和 GLSSVM、Ker-GLSSVM.

实验设计:为了测试三种半监督学习方法,本文将被测试的数据集中的训练样本集分成有标号和无标号两种形式(见表 1),同时为了更好的选取算法的参数,我们选用 5-折交叉验证方法,其中 SDA、Lap-SVM 和本文的 GLSSVM、Ker-GLSSVM 方法对应的参数选取为:k、σ、γ_A、γ_I 同实验 4.1.2,而 SDA 方法参数 α 取值范围同参数 σ,同时 SDA 方法最后采用最近邻分类器.实验结果见表 2.

表 1　Ionoshere、Heart、Glass2、Sonar、Monk2、Liver_disorder、Auatralian、German_org 数据集

Datasets	Training Samples Sets		Test Samples Sets	The Number of Features
	The Number of Labeled Samples	The Number of Unlabeled Samples		
Ionoshere	60	170	117	33
Heart	60	120	90	13
Glass2	50	55	55	9
Sonar	20	115	70	60
Monk2	50	115	432	6
Liver_disorder	80	150	115	6
Auatralian	150	310	460	14
German_org	50	230	666	24

表 2　SDA、Lap-SVM 和本文的 GLSSVM、Ker-GLSSVM 测试精度的比较

Datasets	SDA		Lap-SVM		GLSSVM	Ker-GLSSVM
	Linear	Non-linear	Linear	Non-linear		
ionoshere	0.76078	0.86325	0.75213	0.91453	0.78632	0.94017
heart	0.6667	0.7667	0.78889	0.8778	0.81111	0.8889
glass2	0.70909	0.74545	0.67272	0.76346	0.6909	0.76346
sonar	0.68571	0.72857	0.7	0.82857	0.7	0.77143
monk2	0.65972	0.6898	0.66435	0.73379	0.66667	0.74669
liver_disorder	0.54782	0.64348	0.67826	0.72174	0.66957	0.73913
Auatralian	0.65652	0.67173	0.64783	0.66739	0.66949	0.69111
German_org	0.5270	0.57832	0.55843	0.57057	0.56879	0.6021

根据表 1、表 2 可以得到如下结论:

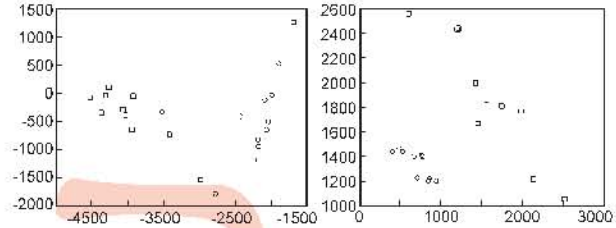
(1)从表 2 可以看出三种半监督学习方法在处理半监督分类时在一定程度上都有较强的学习能力,这是因为根据表 1 可以看出每一个测试数据集中有标号训练样本占训练样本的比例相对都比较低,而在表 2 中反映的测试精度相对来说还是比较理想的,因此,从这一层面上讲,上述三种方法还是有效的.同时表 2 还反映出当每一种算法引入核技巧(本文使用的是 Gauss RBF 核函数)后,非线性方法同线性方法相比在精度上都有较为明显的提高,这说明在具体使用这些半监督方法

时引入核技巧是合适的.

(2)表 2 还表明本文的 GLSSVM、Ker-GLSSVM 方法同 SDA、Lap-SVM 方法相比在绝大多数测试集上具有较好的分类效果,这可以在一定程度上说明本文方法在引入类内散度时并没有影响分类的效果,反而在一定程度上提高了分类精度,从而可以说明本文方法在保持样本内在的局部鉴别信息的同时,还充分考虑了样本内在的全局结构,反映了样本间的全局鉴别信息.

4.2.2 测试人脸图像数据集

人脸图像数据集呈现出明显的非线性流形结构(见图 4),因此被众多的流形学习方法用来作为测试数据集[10~13,15,16,23,24],以反映流形学习方法的有效性.为此,该测试阶段只使用 SDA、Lap-SVM 的非线性形式同本文的 Ker-GLSSVM 方法作比较,以说明本文方法在处理具有流形结构的人脸图像数据方面所具有的有效性,我们选择 Yale、ORL 人脸图像数据集作为测试集,其中 Yale 数据集(32×32)是包含 15 个类别的人脸数据,同一类中有 11 种不同表情的人脸数据(见图 5);ORL 数据集(32×32)包含 40 个类别的人脸数据,同一类中有 10 种不同表情的人脸数据(见图 6).



(a) Yale 1vs2 (b) ORL 1vs2

图4 两数据子集的二维数据表示

图5 Yale数据集中某一类中所有图像

图6 ORL数据集中某一类所有图像

实验设计:因为人脸图像数据集是一种典型的小样本高维数据集,为了降低算法运行过程中由于矩阵奇异性引起的计算误差,该测试过程首先使用 PCA 方法对上述两个人脸图像数据集进行特征降维,即将原图像的 1024 维降为 9 维(因为根据测试分析降为 9 维后所得的新数据集可以保持原数据集 90% 以上的信息量),同时在测试时一般选取的训练样本略多于维数 9,这样在一定程度上会降低 SDA、Lap-SVM、Ker-GLSSVM 方法求解相应矩阵逆矩阵过程中所产生的误差.为了测试上述三种非线性方法,该测试过程随机在 Yale、ORL 数据集中各抽取三个 2 分类的数据子集,即 Yale_1vs2(第 1 类与第 2 类,下同)、Yale_6vs10、Yale_4vs15、ORL_1vs2、ORL_7vs32、ORL_23vs35,其中每个 Yale 数据子集包含 10 个训练样本(4 个有标号样本,6 个无标号样本)和 12 个预测样本,而 ORL 数据子集包含 10 个训练样本(4 个有标号样本,6 个无标号样本)和 10 个预测样本.由于各数据子集的训练样本偏少,该测试使用留一法交叉验证,其中 SDA、Lap-SVM、Ker-GLSSVM 方法中的参数设定为:σ、γ_A、γ_I、α 同实验 4.2.1,$k=[1,2,3]$,(实验结果见表 3).

根据表 3 反映了本文的 Ker-GLSSVM 方法在处理人脸图像数据集时具有较好的分类效果,这充分说明了当处理具有明显流形结构的人脸图像数据时,本文方法不但可以在一定程度上保持局部的流形结构,同时还可以保证类内散度最小,从而可以保持人脸数据内在的全局结构和全局的鉴别信息.

表 3 非线性 SDA、非线性 Lap-SVM 和本文的 Ker-GLSSVM 方法测试结果

Datasets		SDA (non-linear)	Lap-SVM (non-linear)	Ker-GLSSVM
Yale	Yale_1vs2	0.83333	0.58333	0.91667
	Yale_6vs10	1	0.91667	0.6667
	Yale_4vs15	0.75	0.6667	0.83333
ORL	ORL_1vs2	0.9	0.8	1
	ORL_7vs32	0.7	0.8	0.8
	ORL_23vs35	0.8	0.8	0.9

5 总结

本文根据传统的 SVM 方法存在的不足,将 LDA、LPP 的基本原理引入到传统的 SVM 方法中,提出基于全局和局部保持的半监督支持向量机(GLSSVM),该方法不但会在一定程度上克服传统 SVM 方法训练不充分的缺陷,同时还会充分考虑样本内在的全局和局部几何结构,反映样本间的全局和局部的判别信息,而且从理论上分析了该方法满足半监督方法必须依据的一致性原则.诚然,该方法由于引入了类内散度矩阵,从而使得该方法在处理实际问题时要保证一定数量的有标号样本,同时会在一定程度上提高算法的时间、空间复杂度,因此,如何克服算法上述不足将是我们以后研究的方向.

参考文献:

[1] Vanpanik V. Statistical Learning theory[M]. NewYork: Wiley

Press,1998.
[2] Scholkopf B, Smola A. Learning with Kernels-Support Vector Machine, Regularization, Optimization, and Beyond[M]. Cambridge, MA: MIT Press, 2002.
[3] Pontil M, Verri A. Support vector machine for 3D object recognition[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1998, 20(6): 673 – 646.
[4] 李建中,杨昆,高宏.考虑样本不平衡模型无关的基因选择方法[J].软件学报,2006,17(7):1485 – 1493.
Li Jianzhong, Yang Kun, Gao Hong. Model-free gene selection method by considering unbalanced samples[J]. Journal of Software, 2006, 17 (7): 1485 – 1493. (in Chinese)
[5] 邓林,马尽文,裴健.秩和基因选取方法及其肿瘤诊断中的应用[J].科学通报,2004,49(15):1652 – 1657.
Deng Lin, Ma Jinwen, Pei Jian. Rank sum method for related gene selection and its application to tumor diagnosis[J]. Chinese Science Bulletin, 2004, 49(15): 1652 – 1657. (in Chinese)
[6] Tefas A, Kotropoulos C, Pitas I. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(7): 735 – 746.
[7] Zafeiriou S F, Tefas A, Pitas I. Minimum class variance support vector machines[J]. IEEE Trans on Image Processing, 2007, 16 (10): 2551 – 2564.
[8] Joachims T. Transductive Inference for Text Classification using Support Vector Machines[A]. Proc. ICML-99[C]. San Fransisco: Morgan Kaufmann, 1999.
[9] Fung G, Mangasarian O L. Semi-supervised support vector machines for unlabeled data classification[J]. Optimization Methods and Software , 2001, (15): 29 – 44
[10] Tenenbaum J B, Silva V D, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290: 2319 – 2323.
[11] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290: 2323 – 2326.
[12] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373 – 1369.
[13] He X F, Niyogi P. Locality Preserving Projections[OL]. http://peples.cs.uchicago.edu/xiaofei/LPP NIPS03.pdf.
[14] Cai D, He X F, Han J W. Semi-Supervised Discriminant Analysis[OL]. http://www.cs.uiuc.edu/homes/dengcai2/ publication/ conference/iccv07 dengcai SDA.pdf.
[15] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from examples[J]. Journal of Machine Learning Research, 2006, (7): 2399 – 2434.
[16] He X F, Yan S C, Hu Y X, et. al. Face recognition using Laplacian faces[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328 – 340.
[17] Wang H X, Chen S B, Hu Z L, Zeng W M. Locality-preserved maximum information projection[J]. IEEE Trans on Neural Networks, 2008, 19(4): 571 – 585.
[18] 边肇祺,张学工.模式识别[M].第二版,北京,清华大学出版社,2001.
Bian Z Q, Zhang X G. Pattern Recognition[M]. Beijing: TsingHua University Press, 2001. (in Chinese)
[19] Yan S C, Xu D, Zhang B Y, et al. Graph embedding and extensions : a general framework for dimensionality reduction [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40 – 51.
[20] Zhou D Y, Bousquet O, Lal T, et. al. Learning with local and global consistency[A]. Proc. NIPS' 2003[C]. Vancouver: MIT Press, 2003.
[21] Chen Z, Haykin S. On different facets of regularization theory [J]. Neural Comput, 2002, 14(12): 2791 – 2846.
[22] Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem[A]. Proc. COLT' 2001[C]. Amsterdam: Springer Press, 2001. 416 – 426.
[23] Xue H, Chen S C, Yang Q. Discriminatively regularized least-squares classification[J]. Pattern Recognition, 2009, 42(1): 93 – 104.

作者简介:

鼐　军　男,副教授,中国计算机学会会员.1971 年 10 月出生于江苏阜宁.1996 年、2004 年分别在四川大学、南京航空航天大学获得理学学士、工学硕士学位,2007 年进入江南大学信息工程学院攻读博士学位,从事数据挖掘、人工智能、模式识别有关研究工作.
E-mail: gjdlin@yahoo.cn

王士同　男,教授、博士生导师、中国计算机学会高级会员.1964 年 4 月出生于江苏邗江.1984 年、1987 年在南京航空航天大学获得工学学士、硕士学位.现为江南大学信息工程学院院长,从事人工智能、模式识别、模糊系统、医学图像处理和生物信息学方面的研究工作.

邓赵红　男,1981 年生,讲师,博士,主要研究方向为人工智能、机器学习和生物信息学.