一种提高分类精度的支持向量机 NDR-SVM*)

梁锦锦　刘三阳

(西安电子科技大学数学科学系　西安 710071)

摘　要　提出基于近邻域比率的支持向量机 NDR-SVM。该算法对每个训练样本构造一个近邻域，在此邻域中计算与中心同类的样本占邻域中总样本的比率；根据比率与剔除阈值的大小关系决定邻域中心的取舍，并将所保留的样本带入 SVM 分类. 通过修剪训练集，该算法减弱了噪声对 SVM 泛化能力的影响。实验结果表明，与已有算法相比，NDR-SVM 具有更高的分类精度，大大提高了训练速度。

关键词　支持向量机，近邻域比率，噪声，修剪

New Algorithm NDR-SVM to Improve Accuracy

LIANG Jin-jin　LIU San-yang

(Department of Mathematical Sciences, Xidian University, Xi'an 710071, China)

Abstract　A new algorithm based on neighbor domain rate NDR-SVM is proposed . Construct a neighbor domain for each point in the training set. Figure out the ratio of samples in this sphere, which has the same label with its center, and reserve or delete its center according to the relationship between the ratio and the eliminating range. Then the remained are used for SVM classification. Through pruning the training set, the proposed algorithm decreases the effects of noises on the generalization ability of SVM. Numerical experiments show that, compared with existing algorithm, NDR-SVM has higher accuracy and greatly improves the training speed.

Keywords　SVM, Neighbor domain rate, Noises, Prune

1　引言

支持向量机(Support Vector Machine，简称 SVM)由 Vapnik 等提出，是一种基于统计学习理论的有监督学习方法[1]。它建立在结构风险最小化(Structural Risk Minimization Principle)基础之上[2]，能够较好地解决小样本、高维数、非线形、局部极小等问题，可以有效地进行分类、回归和密度估计等[3]。由于其具有拟合精度高，选择参数少，推广能力强和全局最优等特点，支持向量机成为机器学习领域新的研究热点，并被用于人脸识别、文本分类、手写体识别和蛋白质结构预测等领域。

在对 SVM 的研究中，提高它的分类能力(泛化能力)是所有研究的出发点和归宿。由于目前通用的支持向量机方法存在抗干扰能力差，对噪声敏感等问题[4,5]，适当地处理错分点及噪声成为支持向量机的一个重要论题。文献[6]给出样本的硬剔除方法，根据样本与其最近邻类别指标的同异关系决定其取舍。显然，这样的做法是不精确的。直观的延伸是将其推广到 KNN，但 K 值决定算法的精确程度。

为此，本文构造近邻域比率(NDR)算法修剪训练集：根据邻域中与中心同类的样本比例与剔除阈值的大小关系决定邻域中心的取舍，然后带入支持向量机分类。算法设计简单，容易实现，且邻域由算法自身生成，更具客观性。人造数据及 UCI 标准数据集 German Credit Data 上的仿真实验验证了本文算法的可行性和有效性。

2　支持向量机分类

支持向量机是基于小样本统计学习理论的一种新型机器学习方法。它主要基于以下考虑：(1)基于结构风险最小化原则，最大化分类间隔以得到好的推广能力；(2)算法设计为凸二次规划，避免多解性；(3)采用核化技术实现线形算法的非线形化。根据泛函中的 Mercer 定理，引入核函数将样本映射到高维的特征空间，求解高维空间的线形分类函数可得非线形分类器。由于分类结果只用到数据间内积，整个过程不需知道非线形映射的具体形式，且不增加计算复杂度。

不妨取数据的二分类问题说明支持向量机的分类原理。

记 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 为用于分类的训练集，$x_i\in X=R^n$ 为样本特征，$y_k\in Y=\{1,-1\}$ 表示相应的分类结果属于正类或负类，内积为欧式空间的点积。

当训练集线形可分时，SVM 的目标就是构造线形最优分类超平面 $w\cdot x+b=0$，要求其将两类样本完全正确地分开，并使分类间隔最大。相应的优化问题为：

$$(P_1)\min \frac{1}{2}\|w\|^2$$

$$s.t.\ y_i(w\cdot x_i+b)\geqslant 1 \quad i=1,\cdots,l$$

当训练集线形不可分时，对样本引入一组松弛变量 $\{\xi_i\}_{i=1}^{l}(\xi_i\geqslant 0)$ 及正比于违反约束的惩罚 C，则有：

$$(P_2)\min \frac{1}{2}\|w\|^2+C(\sum_{i=1}^{l}\xi_i)$$

$$s.t.\ y_i(w\cdot x_i+b)\geqslant 1-\xi_i \quad i=1,\cdots,l$$

*)基金项目：本文受国家自然科学基金(60574075)资助。梁锦锦　博生研究生，主要研究方向为机器学习、数据挖掘和支持向量机；刘三阳教授，博导，主要研究方向为最优化理论方法及其应用、数据挖掘、支持向量机等。

$\xi_i \geqslant 0 \quad i=1,\cdots,l$

当训练集非线形可分时，通过一个非线性映射 $\Phi: x \to \Phi(x)$ 将难于划分的低维样本空间映射到高维特征空间，并引入 Mercer 定理构造核函数（如果函数 $K(x,y)$ 满足 Mercer 条件，则 $K(x,y)=\Phi(x)\cdot\Phi(y)$）。常用的核函数有线形核、多项式核和高斯径向基核等。通过选择适当的核函数可以得到对应高维空间的分类函数。

记 C 为惩罚因子，则支持向量机模型为：

$$(P_3)\min \frac{1}{2}\|w\|^2+C(\sum_{i=1}^{l}\xi_i)$$

$$s.t.\ y_i(w\cdot\varphi\phi(x_i)+b)\geqslant 1-\xi_i \quad i=1,\cdots,l$$

$$\xi_i \geqslant 0 \quad i=1,\cdots,l$$

常通过求解优化问题的对偶规划得到相应的最优解。引入拉格朗日函数，并对权重 w，阈值 b 及松弛 ξ 求微分，则有：

$$L(w,b,\xi,\alpha,\gamma)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i$$

$$-\sum_{i=1}^{l}\alpha_i[y_i(\phi(x_i)\cdot w+b)-1+\xi_i]-\sum_{i=1}^{l}\gamma_i\xi_i \qquad (1)$$

$$\frac{\delta L}{\delta w}=w-\sum_{i=1}^{l}\alpha_i y_i\phi(x_i)=0 \qquad (2)$$

$$\frac{\delta L}{\delta b}=\sum_{i=1}^{l}\alpha_i y_i=0 \qquad (3)$$

$$\frac{\delta L}{\delta \xi}=C-\alpha_i-\gamma_i=0 \qquad (4)$$

将(2)，(3)，(4)式所得结果带入原规划，可得优化问题 (P_3) 的对偶规划 (P_3^*)

$$(P_3^*)\min \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_i y_j\alpha_i\alpha_j K(x_i,x_j)-\sum_{i=1}^{m}\alpha_i$$

$$s.t.\ \sum_{i=1}^{m}y_i\alpha_i=0,$$

$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,m$$

记最优解为 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)'$。称相应于 $\alpha_i>0$ 的样本称为支持向量，它们对构建超平面起着重要的作用，α_i 越大表示样本对超平面的贡献越大；相应于 $\alpha_i=0$ 的点称为非支持向量，它们对构建超平面不起作用。记 SV 为所有支持向量的集合，则最优分类超平面为 $\sum_{i\in SV}\alpha_i^* y_i K(x\cdot x_i)+b^*=0$，其中 $b^*=y_j-\sum_{i\in SV}y_i\alpha_i^* K(x_i,x_j)$ 为分类超平面的域值，从而对任意输入 x，其分类决策为 $f(x)=\mathrm{sign}(\sum_{i\in SV}\alpha_i^* y_i K(x\cdot x_i)+b^*)$。

3 NDR-SVM 分类器

3.1 最近邻分类

最近邻法是一种非常有效的非参数分类算法[7]，或称为消极学习方法。学习过程只简单存贮已知的训练数据；当分类新的测试样本时，一系列相似样本被取出。假定所有样本对应于欧式空间的一个点，通过构造一种距离找出待测样本的一个或 k 个近邻。前者对应于 NN 算法，分类结果是与其最近的样本所属的类别；后者对应于 k-NN 算法，分类结果是这 k 个近邻中多数样本所属的类别。

采用欧式距离度量向量间的距离。给定 $x_i=(x_i^1,x_i^2,\cdots,x_i^n)$，$x_j=(x_j^1,x_j^2,\cdots,x_j^n)$，则点 x_i，x_j 之间的距离为 $\sqrt{\sum_{k=1}^{n}(x_i^k-x_j^k)^2}$。记样本到自身的距离为无穷大，则一个样本的最近邻就是在上述定义下与自身距离最小的样本。

3.2 近邻域比率 NDR 修剪

近邻域比率 NDR 的基本思想是：对训练集中每个样本找出其近邻，如果两者类别不同，则从其近邻出发，找出此样本外的最近邻，并构建某种近邻域。设置一个剔除阈值，判断近邻域中与中心同类的样本比例是否在剔除阈值之内，选择剔除或者不剔除中心样本。具体算法预处理流程如下（其中 C_0 为类别数，$\beta(0<\beta\leqslant 1)$ 为剔除因子，y_i 为第 i 个样本的类别指标）：

步 1：对每个样本 $x_i(i=1,2,\cdots,n)$，计算其最近邻 x_j；

步 2：如果 $y_i=y_j$，转 1 计算下个样本 x_{i+1} 的最近邻；否则 $y_i\neq y_j$，计算 x_j 的最近邻 $x_{i_0}(i_0\neq i)$ 转 3；

步 3：以 x_j 为中心，$\alpha=\max(d_{ij},d_{i_0 j})$ 为半径，构造球形邻域，转 4；

步 4：记球形邻域中与 x_j 同类和异类的样本数目分别为 m,n 转 5；

步 5：如果 $m/m+n<\beta/C_0$，剔除 x_j，否则转 1 考察下个样本 x_{i+1}。

若所有样本都经过考察，停止循环。

3.3 NDR-SVM 分类

经过近邻域比率(NDR)修剪训练集，消除了数据中潜在的异类噪声，保证了数据采集的合理性。用支持向量机对修剪过的训练样本集进行学习，得到分类器 NDR-SVM。

注意到算法的预处理步骤中引入剔除因子 β，其决定了剔除比例。选取较大的 β，较多的样本包括部分有效样本将落在剔除阈值内而被视为噪声剔除，对于较小的 β，较多的样本包括部分噪声将被视为有效样本参予分类，影响分类器的精度。本文通过多次实验确定 β 的最优取值。

4 实验结果

本文实验均在 PC 机(CPU 为 P4，3.06 GHz，内存为 0.99 GB)上，利用 Matlab SVM 软件包及其修改程序进行，具体数据及实验结果如下。

例 1：线形可分无噪数据

随机产生两类完全可分的正态分布点 200 个，每类 100 个。随机抽取输入样本的 1/2 作为训练集，剩余样本作为测试集。分别用 SVM 和 NSVM 分类，并重复随机抽取过程 10 次，取平均结果。实验表明 SVM 和 NSVM 的分类精度均为 100%！这说明对于无噪声数据，该算法并不会剔除任何有效样本，从而保证了剔除的合理性。

例 2：German Credit Data 数据

该数据库有 1000 个样本，+1 类样本 700 个，−1 类样本 300 个，每个样本有 24 个信用信息指标。从中随机抽取 667 个样本作为训练集，其余 333 个样本作为测试集；并重复上述随机抽取过程 10 次，取平均结果。设置相同的参数，本文算法与 NN-SVM 及标准 SVM 的分类结果比较如下。

表 1 不同算法的分类结果比较

算法	精度	时间(秒)
SVM(C=1,linear)	76.53%	109.9
SVM(C=1,poly,d=1)	75.02%	107.3
SVM(C=1,rbf,σ=1)	70.67%	96.5
NN−SVM((C=1,linear))	75.07%	50.6
NN−SVM(C=1,poly,d=1)	73.46%	45.1
NN−SVM(C=1,rbf,σ=1)	72%	36.5
NDR−SVM(C=1,linear)	75.68%	58.4
NDR−SVM(C=1,poly,d=1)	76.88%	46.8
NDR−SVM(C=1,rbf,σ=1)	72.67%	58.3

(下转第 183 页)

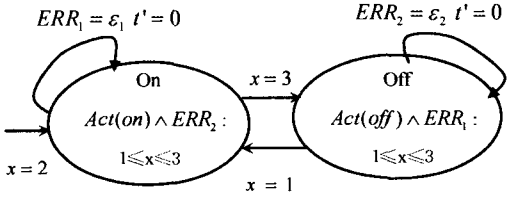
函数，迁移后的 Act 关系保持不变，因此迁移目标位置仍是原位置 v。依这种方式修改后的自动机模型记为 $Approx_\varepsilon(H,n)$。显然在每个位置上有 $|error|\leqslant\varepsilon$，但在离散迁移过程中引入了累加误差。$Approx_\varepsilon(H,n)$的构造过程可形式化定义为：

定义 8($Approx_\varepsilon(H,n)$的构造) 给定自动机 H 的 $Approx(H,n)=(L,X,Lab,E,Init,Inv,Act,J,U)$，构造其关于给定值 ε 的精化模型 $Approx_\varepsilon(H,n)$为

$Approx_\varepsilon(H,n)=(L',X',Lab',E',Init',Inv',Act',J',U')$使其满足：

(a) $L'=L, X'=X, Lab'=Lab$。(d) $\forall l'\in L', Inv'(l)=Inv(l)_\wedge\wedge Err_i(l)\leqslant\varepsilon$。(e) $Init'(l)=Init(l)$。(f) $E'=E\cup\{(l,\tau,l)\mid l\in L\}$。(g) $Act'(l)=Act(l)$。(h)任意 $e\in E'$，如果 $e\in E$，则 $j'(e')=j(e')$，如果 $e'=(l,\tau,l), j'(e)\equiv j(e)_\wedge\wedge Err_i(l)=\varepsilon\wedge t'=0\wedge stable(X)$。(i) $U'=U$。

定理 4 由 $Approx_\varepsilon(H,n)$的构造过程有 $H\preccurlyeq Approxe(H,n)\preccurlyeq Approx(H,n)$



$Axt(on)\equiv x'\leqslant x+(5-x)t-1/2(5-x)t^2+1/6(5-x)t^3\wedge x'\geqslant x+(5-x)t-1/2(5-x)t^2+1/6(5-x')t^3\ ERR_1=1/6|x'-x|t^3$

$Act(off)\equiv x'\leqslant x-xt+1/2xt^2-1/6x't^3\wedge x'\geqslant x-xt+1/2xt^2-1/6xt^3$

$ERR_2=|1/6x't^3-1/6xt^3|$

图 2 H 的 $Approx_\varepsilon(H,n)$自动机，其中 $\varepsilon_1=\varepsilon_2=0.5$

例 3：在例 1 中所描述的混合自动机按照上述方法构造的三阶近似多项式自动机如图 2 所示。假定 $\varepsilon_1=\varepsilon_2=0.5$，其中模态 off 的多项式近似轨迹如图 3 所示。由图 3 可看出两条多项式曲线间囊括了原指数曲线，近似可通过增大多项式阶数或减少 ε 值来逼近原曲线。

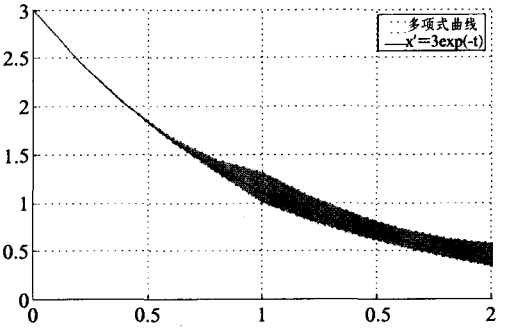


图 3 模态 off 的近似动态轨迹

结束语 本文将线性 phase-portrait 近似推广到多项式 phase-portrait 近似，叙述了如何自动构造混合自动机的多项式 phase-portrait 近似，及如何精化近似模型。本文所基于的模型其向量场形式限于 $\dot{x}_i=\varphi(X)$，$\varphi(X)\in poly(X)$。对于更一般的形式，在以后的文章中进行进一步的研究。另外，在做抽象精化时，如何消除累积误差也是今后的研究方向。

参 考 文 献

[1] Alur R, Courcoubetis C, Halbwachs N, et al. The algorithmic analysis of hybrid systems. Theoretical Computer Science, 1995, 13: 3-34

[2] Henzinger P, Ho H, Wong-Toi H. Algorithmic analysis of nonlinear hybrid systems. IEEE Transactions on Automatic Control, 1998, 43: 540-554

[3] Franzle M. Analysis of hybrid systems: An ounce of realism can save infinity of states // Proceedings of Computer Science Logic. LNCS, 1999, 1683: 126-140

[4] Tiwari A, Khanna G. Series of abstractions for hybrid automata//Tomlin C, Greenstreet M R, eds. HSCC, LNCS, 2002, 2289: 465-478

[5] Henzinger T A, Kopke P W, Puri A, et al. What's decidable about hybrid automata? Journal of Computer and System Sciences, 1998, 57: 94-124

[6] Lanotte R, Tini S. Taylor Approximation for Hybrid Systems // Tomlin C, Greenstreet M R, eds. HSCC, LNCS, 2005, 3414: 402-416

[7] Chutinan A, Krogh B H. Verification of polyhedral-invariant hybrid automata using polygonal flow pipe approximation// Morari M, Thiele L, eds. HSCC, LNCS, 1999, 1569: 76-90

(上接第 168 页)

从表 1 可以看出，对于 German Credit Data 取相同的参数，三种核函数下该算法的分类精度均有不同程度的提高，且训练时间明显降低！实验中还发现，取相同的核参数，三种核函数下的分类精度均对惩罚参数 C 的变化不敏感。这说明算法有效剔除并降低了噪声对分类的影响！

结束语 针对传统支持向量机分类对噪声与野值敏感的问题，本文提出基于近邻域比率的支持向量机 NDR-SVM。通过近邻域中与中心同类的样本比率与剔除阈值的大小关系决定邻域中心的取舍，并用支持向量机对修剪后的训练集进行学习。由于球心与半径均由算法自身确定，近邻域客观反映了数据分布。通过设定合适的剔除因子 $\beta(0<\beta\leqslant 1)$，该方法能够有效剔除影响分类性能的噪声和野值。仿真实验验证了该方法的可行性和有效性，但剔除因子 β 的设定是关键。下一步的工作是，如何根据具体问题给出剔除因子的一个合适取值。

参 考 文 献

[1] Vapnik V. The nature of statistical learning theory [M]. Springer-Verlag, 1995

[2] Amaris S, Wu S. Improving support machine classifier by modifying kernel function [J]. Neural Networks, 1999, 12: 783-789

[3] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167

[4] Guyon I, Maticn, Vladmrnvapnik. Discovering Information Patt-erns and Data Cleaning [M]. Cambridge: MA MIT Press, 1996: 181-203

[5] Zhang X G. Using Class-center Vectors to Build Support Vector Machines [Z]// Proc. IEEE NNSP'99. Wisconsin, USA, 1999: 3-11

[6] 李红莲，王春花，袁保宗. 一种改进的支持向量机 NN-SVM. 计算机学报，2003，26(8)：1015-1020

[7] Cover T M, Hart P E. Nearest neighbor pattern classification [J]. In Trans IEEE Inform Theory, 1967, TT-13: 21-27