样本数目不对称时的SVM模型*)

肖健华　吴今培

(五邑大学智能技术与系统研究所　广东江门529020)

SVM Model with Unequal Sample Number Between Classes

XIAO Jian-Hua　WU Jin-Pei

(Institute of Intelligent Technology & Systems, Wuyi University, Jiangmen 529020)

Abstract The paper briefly introduces the principle of SVM in sorting first. Then using normal distribution as an example, the paper points out that the SVM can not get excellent sorting ability when the sample numbers of different class are different greatly. A method has been designed to compute the penalty parameter to error samples. On the basis of different penalty parameter an advanced SVM algorithm has been developed. Two simulated examples show the new algorithm is feasible.

Keywords Support vector machin, Pattern recognition, Normal distribution, Exponential distribution

1　引言

支持向量机(Support Vector Machine, SVM)是统计学习理论发展的结果，自诞生以来[1,2]，在非线性特征提取、模式识别、函数拟合等方面表现出优良的性能[4~6]，因而受到广泛的重视，目前已成为人工智能领域的研究热点。

然而，由于SVM从诞生到现在不足十年，许多研究工作还没展开[8]，有些已经表现出来的优良特性也没有在数学上得到论证。本文将结合一类样本分布情况，即当两类样本数目相差悬殊时，讨论SVM的分类机制。这种情况在现实中是广泛存在的，如在故障诊断中，正常状态样本数据量总是远比故障状态样本数据量多。

2　基于SVM的模式识别方法

支持向量机是建立在结构风险最小化基础上的，其研究最初是针对模式识别中的线性可分问题，如图1。分割线1(平面)和分割线2(平面)都能正确地将两类样本分开，即都能保证使经验风险最小(为0)，这样的线(平面)有无限多个，但分割线(平面)1离两类样本的间隙最大，称之为最优分类线(平面)。最优分类线(平面)的置信范围最小[3]。

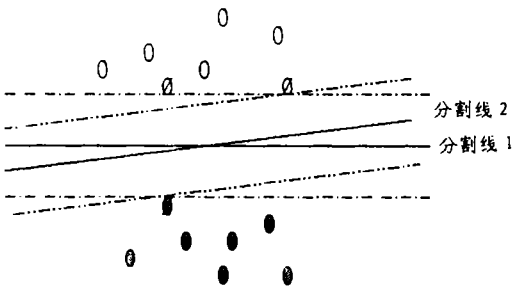


图1　支持向量机原理示意图

设线性可分样本集为(X_i, y_i), $i=1,2,\cdots,n$, $X\in R^d$, $y\in\{-1, 1\}$是类别标号。d维空间中线性判别函数的一般形式为$g(X)=W\cdot X+b$，分类面方程为

$$W\cdot X+b=0 \tag{1}$$

将判别函数归一化，使两类所有样本都满足$|g(X)|\geqslant 1$，这样，分类间隔就等于$2/\|W\|$，因此，求间隔最大变为求$\|W\|$最小。

满足$|g(X)|=1$的样本点，离分类线(平面)距离最小，它们决定了最优分类线(平面)，称之为支持向量(Support Vectors, SV)。图1中带斜线的样本即为SV。

因此，求最优分类面的问题转化为优化问题：

$$\min\phi(W)=\frac{1}{2}\|W\|^2=\frac{1}{2}(W\cdot W) \tag{2}$$

$$s.t\ y_i[(W\cdot X_i)+b]-1\geqslant 0\ i=1,2,\cdots,n \tag{3}$$

本优化问题可以转化为

$$\min Q(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(X_i\cdot X_j)-\sum_{i=1}^{n}\alpha_i \tag{4}$$

$$s.t\ \alpha_i\geqslant 0\ i=1,2,\cdots,n \tag{5}$$

$$\sum_{i=1}^{n}y_i\alpha_i=0$$

并由此可得到最优分类函数为

$$f(X)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i^* y_i(X_i\cdot X)+b^*\} \tag{6}$$

对于非支持向量满足$\alpha_i=0$，因此，式(6)只需对支持向量进行。b^*可任选一个支持向量，由式(3)的约束条件取等号求出。

对于线性不可分问题，即存在部分样本不满足约束条件式(3)，此时引入非负的松弛变量ξ_i，同时优化问题为：

$$\min\phi(W)=\frac{1}{2}(W\cdot W)+C\sum_{i=1}^{n}\xi_i \tag{7}$$

$$s.t\ y_i[(W\cdot X_i)+b]-1+\xi_i\geqslant 0\ i=1,2,\cdots,n \tag{8}$$

式(7)中常数C起着对错分样本的惩罚作用，实现的是学习机器泛化能力和错分样本数目之间的折中[7]。

此外，在处理非线性问题上，Vapnik还引入了核空间理论：将低维的输入空间数据通过非线性映射函数映射到高维属性空间，将分类问题转化到属性空间进行。可以证明，如果选用适当的映射函数，输入空间线性不可分问题在属性空间将转化为线性可分问题。这种非线性映射函数被称之为核函数(Kernel Function)。实际上，向量的内积$K(X, X_i)=(X\cdot X_i)$本身也可视为一类特殊的核函数。为简化讨论，本文不涉

*)国家自然科学基金资助项目。

及其它核函数，其实，本文获得的结论，可以简单地推广到其它核函数。

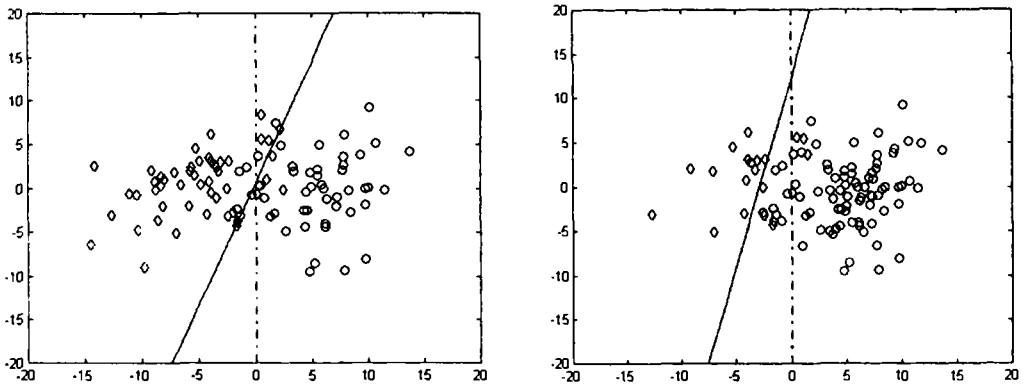
3 SVM 在处理样本数目不对称时所存在的不足

考察图2所示的两维两类样本，它们分别满足正态分布

$$p(x)=\frac{1}{2\pi|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\} \quad (9)$$

其中 $\Sigma=\begin{bmatrix}25 & 0\\ 0 & 25\end{bmatrix}$，对于类别一，$\mu_1=\begin{bmatrix}5\\0\end{bmatrix}$，对于类别二，$\mu_2=\begin{bmatrix}-5\\0\end{bmatrix}$

现在讨论两种情况，一是样本数目相等，各有50个样本，另一种是两类样本数目比例为80:20。图2(a)、(b)中的空心圆圈代表类别一，空心菱形代表类别二。



(a) 样本数目 50:50　　(b) 样本数目 80:20

图2　正态分布下，样本数目比例不同时的 SVM 模型

图2(a)、(b)中的虚线为期望的分类线 $x_1=0$，对于图2(a)，各类样本数目相等，此时基本上能得到合理的分类线，且该分类线随着样本数目的增加，能逐渐逼近期望分类线；对于图2(b)，由于两类样本数目相差悬殊，分类线明显偏向样本少的一方，造成了较大的分类误差。

造成这种现象的根本原因在于式(7)中对不同类别的错误划分采用了相同的惩罚系数 C，使得分类线偏向样本密度小的一方，这样可使错分样本数目减少。

4 新的 SVM 模型

针对样本数目不对称的情况，有三种解决方案可供选择，一是对样本多的一方进行重新采样；二是对样本少的一方增加样本；三是对两类错分样本进行不同的惩罚。显然，前两种方案的目的都是设法使两类样本数目相当，但是，方案一浪费了相当一部分的样本信息，方案二只能增加一些重复的样本，反而增加了存储空间和运算时间的开销。可见，前两个方案都不足取。方案三充分利用了现有的信息，又基本上不增加计算复杂度及在时间、内存上的开销。

采用方案三的关键在于如何确定 $C_i(i=1,2)$ 值。C_i 值太小，表明对样本的错分惩罚太小，可能会导致过多的错分样本，C_i 值太大，表明对样本的错分惩罚大，同样起不到折中的作用。

设两类样本集 Ξ_1,Ξ_2，样本数目分别是 N_1,N_2，$N_1>N_2$。合理的 C_i 值可通过对其不断逼近获得，在逼近过程中，应使下式近似成立，

$$\frac{1}{2}(W\cdot W)=\sum_{i=1}^{n}C_i\xi_i \quad (10)$$

其中 $C_i=\begin{cases}C_1 & X_i\in\Xi_1\\ C_2 & X_i\in\Xi_2\end{cases}$

同时 C_1,C_2 满足

$$\frac{C_1}{C_2}=\frac{N_2}{N_1} \quad (11)$$

式(11)表明，对不同类别错误划分的惩罚与样本数目成反比。$C_i(i=1,2)$ 的初值可通过方案一获得。从上面的讨论可归纳出 SVM 新算法：

1)对样本多的一方，按等概率重新采样，设数目为 $N'_1=N_2$；

2)选择较大的系数 C，对样本集 $\{N'_1=N_2\}$ 进行 SVM 设计，获取 W,ξ_i 值；

3)由式(10)和式(11)计算此时的 C_1,C_2；

4)在新的 C_1,C_2 值下，对样本集 $\{N_1,N_2\}$ 进行 SVM 设计，获取 W,ξ_i 值；

5)由式(10)和式(11)计算此时的 C_1,C_2；

6)判断 C_1,C_2 的变化情况，若变化小于所设定的阈值，则7)，否则返回4)；

7)获得最终的 SVM 模型，结束。

一般经过3个循环，即可获得最终的 SVM 模型。第6)步中的判断条件可采用判断式：

$$\frac{(C_{11}-C_{10})^2+(C_{21}-C_{20})^2}{C_{11}^2+C_{21}^2}<\delta \quad (12)$$

δ 为一百分比常数，如5%。$C_{i1},C_{i0}(i=1,2)$ 分别表示当前时刻与前一时刻的 C_i 值。

5 算例分析

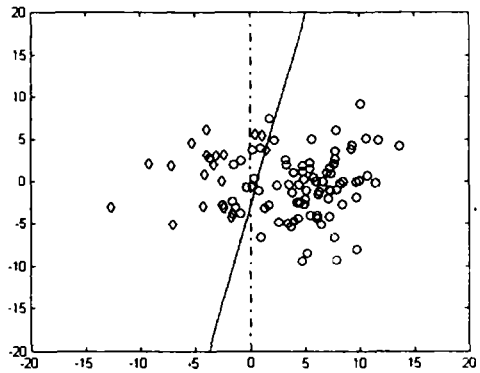


图3　正态分布下的 SVM 新模型

为说明新算法的有效性，现举两个算例。

算例1　两维两类样本，分别满足式(9)所示的正态分布，

所有参数与前面相同，样本数目比例依然为80:20。

采用上面介绍的新算法，并在计算过程中使

$$\frac{C_1}{C_2}=\frac{1}{4} \tag{13}$$

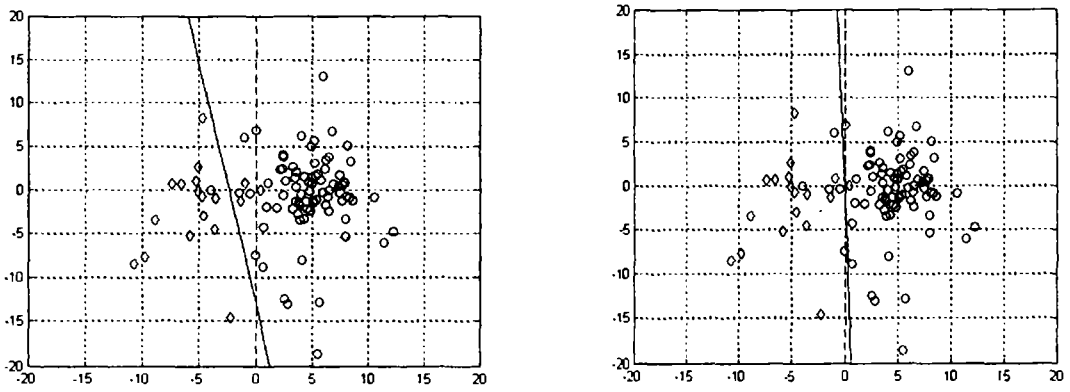
图3为运算结果，实线为实际分类线，点划线为期望的分类线。将图3与图2(b)比较，不难发现，新分类线更为接近于期望的分类线。

算例2 两维两类样本，分别服从指数分布

$$p(X)=\frac{1}{2}\lambda\exp[-\lambda\|X-\mu\|] \tag{13}$$

式中 $\lambda=4$，对于类别一，$\mu_1=\begin{bmatrix}5\\0\end{bmatrix}$，样本数目为80；对于类别二，$\mu_1=\begin{bmatrix}-5\\0\end{bmatrix}$，样本数目为20。

图4(a)，(b)分别画出了在一般 SVM 模型和 SVM 新模型下的分类线，从图中可以看出，新模型克服了样本数目差异所造成的影响。



(a) 一般 SVM 模型　　(b) SVM 新模型

图4　指数分布下，样本数目比例不同时的 SVM 模型比较

结论　本文简要地分析了 SVM 的分类原理，指出在样本数目相差较悬殊时，SVM 不能获得良好的分类能力。为此，本文提出了对不同类的错分样本进行不同惩罚的方法，两个算例的运算结果也充分说明，这种修正是有效的。

为简化描述，本文采用最为简单的向量内积作为核函数，设计上，整个算法对所有的核函数都是适用的。

参 考 文 献

1 Vapnik V. The Nature of Statistical Learning Theory. New York:Springer-Verlag,1995

2 Osuna E,Freund R,Girosi F. Support Vector Machine:Training and Applications. AI Memo 1602. MIT Artificial Intelligence Laborabory,1997

3 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery,1998,2(2):121～167

4 Campbell C. Kernel Methods:A survey of Current Techniques. http://citeseer.nj.nec.com/compbell00kernel.html

5 Schölkopf B,Smola A,Müller K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation,1998,10:1299～1319

6 Ruiz A,López-de-Teruel P E. Nonlinear Kernel-Based Statistical Pattern Analysis. IEEE Trans. on Neural Networks,2001,12(1):16～32

7 张学工. 关于统计学习理论与支持向量机. 自动化学报,2000,26(1):32～42

8 肖健华,樊可清,吴今培,等. 应用于故障诊断的 SVM 理论研究. 振动、测试与诊断,2001,21(4):258～262

(上接第139页)

10 Medvidovic N, Taylor R N. A Classification and Comparison Framework for Software Architecture Description Languages. IEEE Transaction on Software Engineering,2000,26(1):70～93

11 Tracz W. LILEANNA:A Parameterized Programming Language. In:Proc. Second Int'l Workshop Software Reuse,Mar. 1993. 66～78

12 Garlan D,Monroe R,Wile D. ACME:An Architecture Description Interchange Language. In:Proc. CASCon'97,Nov. 1997

13 Medvidovic N,Rosenblum D S,Taylor R N. A Language and Environment for Architecture-Based Software Development and Evolution. In:Proc. 21st Int'l Conf. Software Eng. (ICSE'99),May,1999. 44～53

14 OMG AD/97. 08. 05,1997,United Modeling Language Specification,version 1.1,Object Management Group

15 Metayer D L. Describing software architecture styles using graph grammars. IEEE Transactions on Software Engineering,1998,24(7):521～533

16 Inveradi P,Wolf A L. Daniel Yankelevich,Behavioral Type Checking of architectural Component Based on Assumptions. From:http://www.sei.cmu.edu/publications/cu-cs-861-98.ps

17 Yuan Chun,Chen Yiyun. Software Architecture Evolution by Multiset Transformation. In:Proc. of Intl. conf. on Software:Theory and Practice,Aug. 2000. 236～243

18 云晓春,方滨兴. 基于构件设计的正确性验证. 小型微型计算机系统,1999,20(5):330～334

19 Kruchten P B. the 4+1 View Model of Architecture. IEEE Software,1994. 42～50

20 饶若楠,尤晋元. 基于统一建模语言的分布式企业信息系统软件体系结构模型. 上海交通大学学报,2001,34(7):979～982

21 邓勇,丁峰,沈均毅. 基于 UML 的软件体系结构建模方法研究. 小型微型计算机系统,2001,22(10):1206～1209

22 Rational,Reusable Asset Specification—— Architectural Description Standard:[Technique Report]. from *http://www.Rational.com/publications*