基于闭凸包收缩的最大边缘线性分类器*

陶 卿[1,2], 孙德敏[3], 范劲松[4], 方廷健[4]

[1](中国科学院 自动化研究所,北京 100080);
[2](中国人民解放军炮兵学院 一系,安徽 合肥 230031);
[3](中国科学技术大学 自动化系,安徽 合肥 230027);
[4](中国科学院 合肥智能机械研究所,安徽 合肥 230031)
E-mail: q_tao@sohu.com
http://www.ia.ac.cn

摘要: SVM(support vector machines)是一种基于结构风险最小化原理的分类技术.给出实现结构风险最小化原理(最大边缘)的另一种方法.对线性可分情形,提出一种精确意义下的最大边缘算法,并通过闭凸包收缩的概念,将线性不可分的情形转化为线性可分情形.该算法与 SVM 算法及其 Cortes 软边缘算法异曲同工,但理论体系简单、严谨,其中的优化问题几何意义清楚、明确.

关 键 词: 闭凸集;收缩;支持向量;最大边缘;分类器
中图法分类号: TP18 文献标识码: A

分类学习算法在神经网络的发展史上起着非常重要的作用.1957 年,美国学者 Rosenblatt 首先提出感知器(perceptron)的概念.它是一个具有单层计算单元的神经网络,并由线性阈值元件组成.感知器在神经网络的研究中有着重要的地位和意义,它提出了自组织、自学习的思想,对能够解决的线性分类问题有一个非常清楚的收敛算法,并从数学角度给出了严格的证明.以后的很多模型(如著名的 BP 模型)都是在这种指导思想下建立的,或者是它的改进和推广[1].由此可见,对分类学习算法的突破可以极大地推进整个神经网络领域以致人工智能的发展.

由于感知器学习算法的初始值可以任意选定,使得由此产生的分离超平面有无穷多种,往往造成了分类超平面严重偏向某一类,即导致了感知器泛化性能不高.另一方面,这种算法未能对在分类中起关键作用的元素进行刻画.当分类结束后添加新的训练样本时,先前已有的运算结果已无作用,网络需重新学习所有样本.可见,这种算法没有真正起到“学习”的作用.

SVM(support vector machines)是由 Vapnik 领导的 AT&T Bell 实验室研究小组提出的一种新的分类方法.1995 年,文献[2,3]的出现是 SVM 诞生的标志.它基于 VC 维(Vapnik-Chervonenkis dimension)理论,是结构风险最小化原理的近似实现,比基于经验风险最小化原理的算法(如 BP 算法、最小二乘法)理论依据更严谨,泛化能力更好.Support Vectors 实际上是训练集的子集,对 Support Vectors 的分类等价于对训练集的分类,这种在分类中起关键作用的元素在信号压缩和特征提取中非常有用.最能体现 SVM 分类性能的是,对线性可分情形,由 SVM 确定的超平面 $wx+b=0$ 正好位于分类集合的“正中间”,此时的结构风险最小化原理表现为最大化 $wx+b=1$ 和 $wx+b=-1$ 的距离,由此使分类超平面到分类集合的距离最大,从而使泛化性能最佳.所以,SVM 算

* 收稿日期: 2000-03-06; 修改日期: 2000-09-20
基金项目: 国家自然科学基金资助项目(60175023);安徽省自然科学基金资助项目(01042304);安徽省优秀青年基金资助项目
作者简介: 陶卿(1965－),男,安徽长丰人,博士,副教授,主要研究领域为神经网络,支持向量机;孙德敏(1939－),男,辽宁新民人,教授,博士生导师,主要研究领域为模式识别与智能系统,控制理论及其应用;范劲松(1967－),男,安徽合肥人,博士,主要研究领域为支持向量机,KDD;方廷健(1939－),男,上海人,教授,博士生导师,主要研究领域为模式识别与智能系统.

法又称为最大边缘算法(maximal margin)[4,5].对线性不可分问题,Cortes 在其博士学位论文[4]中提出一种基于松弛 SVM 中优化问题的软边缘算法.目前,SVM 算法在非线性分类方面也取得了一定的成功和进展[6~8].

我们在文献[9]中对 SVM 算法进行了综述,SVM 算法最后归结为一种二次规划问题,但理论冗长,涉及大量统计学知识,且转化后的规划问题中参数的意义不明显.受到 SVM 算法是最大边缘算法的启发,本文立足于闭凸集间的距离优化,对线性可分情形提出精确意义下的最大边缘算法.而对线性不可分的情形,通过本文提出的闭凸包收缩的概念,将情形归结为线性可分情形.

SVM 的理论体系是以经验风险最小化原理缺乏理论依据为主线,阐述如何建立理论依据较强的和泛化性能好的分类算法,线性可分情形只是说明算法有效的一个实例.本文给出实现结构风险最小化原理(最大边缘)的另一种方法,与前馈神经网络的理论体系相一致,都是首先从线性可分问题着手,逐步向一般情形推广.本文的算法与 SVM 算法异曲同工,但理论体系简单、严谨,几何意义清楚、明确.正是依靠这种几何直观,使得闭凸包适当收缩可以处理线性不可分问题,得到了与 Cortes 的软边缘算法[4]相同的效果.

1 线性可分问题

首先,我们来讨论在线性可分情形下,通过闭凸集间的距离优化,可以得到精确意义下的最大边缘算法.

定义 **1.** 函数 $f(x): R^m \to R$ 称为是线性可分的,若存在 $W \in R^m$ 和 $\theta \in R$,使 $f(x)=\mathrm{sgn}(Wx-\theta)$.

定义 **2.** 集合 Q_1 和 Q_2 称为线性可分的,若存在线性可分函数 $f(x)$,使

$$f(x)=1,\ \forall x \in Q_1;\ f(x)=-1,\ \forall x \in Q_2.$$

定理 **1**[10]**.** 设 $p_1, p_2, ..., p_n$ 是 m 维欧几里德空间的 n 个点,记 Q 为包含 $p_1, p_2, ..., p_n$ 的最小闭凸集,则

$$Q=\{x=\lambda_1 p_1+\lambda_2 p_2+...+\lambda_n p_n : \lambda_1+\lambda_2+...+\lambda_n=1, \lambda_i \ge 0, i=1,2,...,n\}.$$

Q 也称为 $p_1, p_2, ..., p_n$ 的闭凸包或 $p_1, p_2, ..., p_n$ 生成的闭凸集.

根据文献[10]中的 Hahn-Banach 分离定理,容易得到以下定理:

定理 **2.** Q_1 和 Q_2 线性可分的充要条件是它们的闭凸包线性可分.它们的闭凸包的线性可分函数 $f(x)$ 实际上就是 Q_1 和 Q_2 的线性分类器.

从定理 2 可知,对于线性可分情形,考虑集合与考虑它们的闭凸包是等同的.以下设 Q_1, Q_2 分别为 $p_1, p_2, ..., p_n$ 和 $q_1, q_2, ..., q_m$ 的闭凸包,Q_1 和 Q_2 线性可分.

考虑下面的二次优化问题:

$$\begin{cases} \min\left\|\lambda_1 p_1+\lambda_2 p_2+...+\lambda_n p_n-\beta_1 q_1-\beta_2 q_2-...-\beta_m q_m\right\|^2 \\ \lambda_1+\lambda_2+...+\lambda_n=1,\ \beta_1+\beta_2+...+\beta_m=1 \\ \lambda_i \ge 0, \beta_j \ge 0,\ i=1,2,...,n,\ j=1,2,...,m \end{cases}. \tag{1}$$

对于如下的一般的半正定规划二次问题:

$$\begin{cases} \min \dfrac{1}{2} x^T A x + a^T x \\ Dx=b \\ x \in Q \end{cases},$$

文献[11]中提出一种大范围收敛的神经网络模型(见式(2)).由于巧妙地构造了能量函数,它比文献[12]中的模型结构更简单,性能也更优越.我们将利用式(2)来求解式(1)描述的问题.

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t}=P\left(x-Ax+D^T y-a\right)-x \\ \dfrac{\mathrm{d}y}{\mathrm{d}t}=-DP\left(x-Ax+D^T y-a\right)+b \end{cases}. \tag{2}$$

设 $\lambda_1^*, \lambda_2^*, ..., \lambda_n^*, \beta_1^*, \beta_2^*, ..., \beta_m^*$ 是问题(1)的一组解,则 Q_1 和 Q_2 的最大边缘线性分类器为过 $\lambda_1^* p_1+\lambda_2^* p_2+...+\lambda_n^* p_n$ 和 $\beta_1^* q_1+\beta_2^* q_2+...+\beta_m^* q_m$ 连线中点且与这条连线垂直的超平面,可用点法式求得其方程.对应于 $\lambda_1^*, \lambda_2^*, ..., \lambda_n^*, \beta_1^*, \beta_2^*, ..., \beta_m^*$ 中非零数的向量称为 $p_1, p_2, ..., p_n$ 和 $q_1, q_2, ..., q_m$ 分类问题中

的 Support Vector.

2 线性不可分情形

定义 **3.** 设 Q 是 $p_1, p_2, ..., p_n$ 的闭凸包,称 $p_0 = \frac{1}{n}p_1 + \frac{1}{n}p_2 + ... + \frac{1}{n}p_n$ 为 Q 的中心;称 $\hat{p}_i = \alpha p_0 + (1-\alpha)p_i$ ($1 \le i \le n$)为 p_i 关于 Q 中心 p_0 收缩率为 $1 > \alpha > 0$ 的收缩;称 $\hat{Q} = \left\{ x = \sum_{i=1}^{n} \lambda_i \hat{p}_i : \lambda_1 + \lambda_2 + ... + \lambda_n = 1, \lambda_i \ge 0, i = 1,2,...,n \right\}$ 为 Q 关于其中心收缩率为 $1 > \alpha > 0$ 的收缩; $\forall y \in R^m$,称 $\hat{Q}_y = \hat{Q} - p_0 + y$ 为 Q 关于 y 收缩率为 $1 > \alpha > 0$ 的收缩.

在上述定义中,如果 $\alpha = 0$,则 $\hat{Q} = Q$;如果 $\alpha = 1$,则 $\hat{Q} = \{p_0\}$.从几何直观上说, p_i 收缩以后距离 Q 的中心应该更近.上面的定义显然满足了这一点(距离为 $(1-\alpha)\|p_i - p_0\|$).当 $\alpha = \frac{1}{2}$ 时, $n = 2$ 时,闭凸集关于其中心收缩的几何意义是非常明显的,此时 Q 为 p_1, p_2 的连线,其中心是 p_1, p_2 连线的中点,而关于 p_0 的收缩 $\hat{Q}$ 为 p_0, p_2 连线中点与 p_0, p_2 连线中点的连线.显然,当 $p_1, p_2, ..., p_n$ 的闭凸包构成单纯复形时,本文所定义的中心与代数拓扑中单纯复形中心的概念是一致的.从定义 3 还可以知道,对一些规则的图形,本文定义的收缩还能保持几何形状不变.

下面将证明,有限生成闭凸集关于其中心的收缩能保持中心的位置不变.为了编程实现和实际上应用的方便,这里给出 $\hat{Q}$ 的另一种描述方式.

定理 **3.** 设 Q 是 $p_1, p_2, ..., p_n$ 的闭凸包, $\hat{Q}$ 为 Q 关于其中心收缩率为 $1 > \alpha > 0$ 的收缩,则 $\hat{Q}$ 的中心与 Q 的中心相同.并且

$$\hat{Q} = \left\{ x = \lambda_1 p_1 + \lambda_2 p_2 + ... + \lambda_n p_n : \lambda_1 + \lambda_2 + ... + \lambda_n = 1, \lambda_i \ge \frac{\alpha}{n}, i = 1,2,...,n \right\}.$$

证明:根据定义 3, $\hat{Q}$ 的中心为 $\sum_{i=1}^{n} \frac{1}{n} \hat{p}_i$.记 Q 的中心为 p_0.

$$\sum_{i=1}^{n} \frac{1}{n} \hat{p}_i = \sum_{i=1}^{n} \frac{1}{n} (\alpha p_0 + (1-\alpha)p_i) = \sum_{i=1}^{n} \frac{\alpha}{n} p_0 + (1-\alpha) \sum_{i=1}^{n} \frac{1}{n} p_i = \alpha p_0 + (1-\alpha)p_0 = p_0.$$

$\forall x \in \hat{Q}$,存在 β_i , $\beta_i \ge 0$, $\sum_{i=1}^{n} \beta_i = 1$,使

$$x = \sum_{i=1}^{n} \beta_i \hat{p}_i = \sum_{i=1}^{n} \beta_i (\alpha p_0 + (1-\alpha)p_i) = \alpha p_0 + \sum_{i=1}^{n} (1-\alpha)\beta_i p_i = \sum_{i=1}^{n} \left(\frac{\alpha}{n} + (1-\alpha)\beta_i \right) p_i .$$

令 $\lambda_i = \frac{\alpha}{n} + (1-\alpha)\beta_i$,显然, $\lambda_i \ge \frac{\alpha}{n}$, $\sum_{i=1}^{n} \lambda_i = 1$.

另一方面, $\forall x \in \left\{ x = \lambda_1 p_1 + \lambda_2 p_2 + ... + \lambda_n p_n : \lambda_1 + \lambda_2 + ... + \lambda_n = 1, \lambda_i \ge \frac{\alpha}{n}, i = 1,2,...,n \right\}$,令 $\beta_i = \frac{\lambda_i - \frac{\alpha}{n}}{1-\alpha}$,由上证明知, $\beta_i \ge 0$, $\sum_{i=1}^{n} \beta_i = 1$,且 $x = \sum_{i=1}^{n} \beta_i \hat{p}_i$. □

设 Q_1 , Q_2 分别为 $p_1, p_2, ..., p_n$ 和 $q_1, q_2, ..., q_m$ 的闭凸包, Q_1 和 Q_2 线性不可分,即 $Q_1 \cap Q_2 \ne \varnothing$.若 Q_1 和 Q_2 的中心相同,可平移 Q_1 或 Q_2 使中心不同,平移点非常容易选取.若 Q_1 和 Q_2 的中心不同,可选取收缩率 α ,满足

$$(1-\alpha)\max\left\{ \max_{1 \le i \le n} \|p_i - p_0\|, \max_{1 \le i \le m} \|q_i - q_0\| \right\} < \frac{1}{2} \|p_0 - q_0\|,$$

同时收缩 Q_1 和 Q_2 ,此时 Q_1 和 Q_2 的收缩必然线性可分.本文与 Cortes 的软边缘算法[4]相对应,重点讨论中心不同的"接近"线性可分情形,下文如果不特别说明,所说的收缩都是关于其中心的.

由于 $\hat{Q}_1$ 和 $\hat{Q}_2$ 是线性可分的,用第 2 节的方法确定的具有最大边缘的超平面即为 Q_1 和 Q_2 的线性分类器.

3 方法应用举例

例 1:考虑下列线性可分问题[13]:

x_1	x_2	Class
3	3	1
1	3	1
2	2.5	1
1	1	–1
3	1	–1
3	2.5	–1
4	3	–1

利用神经网络模型(2)求解规划问题(1)得: $\lambda_1=0.89999$, $\lambda_2=0.0000$, $\lambda_3=0.10001$, $\beta_1=0.0000$, $\beta_2=0.0000$, $\beta_3=0.9004$, $\beta_4=0.09996$. $margin=0.4472$ .Support Vector 和 margin 的结果与文献[13]完全相同,如图 1 所示.

值得一提的是,如果我们采用与文献[13]相同的方法来求解规划问题(1),即用 MATLAB 中的函数 QP, 可得: $\lambda_1=1.0000$, $\lambda_2=0.0000$, $\lambda_3=0.0000$, $\beta_1=0.0000$, $\beta_2=0.0000$, $\beta_3=0.8000$, $\beta_4=0.2000$. $margin=0.4472$.所得到的 Support Vectors 比文献[13]还要少.从图 1 中看,这种结果也是正确的.由方法不同造成不同结果的主要原因是此例中的规划问题(1)有无穷多组解.

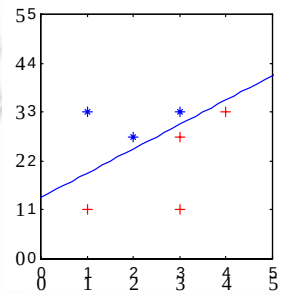

Fig.1 The linear classifier in linearly separable case
图 1 线性可分情形下的最大边缘分类器

例 2:考虑下列线性不可分问题[13]:

x_1	x_2	Class
3	3	1
1	3	1
2	2.5	1
1.5	1.5	1
1	1	–1
3	1	–1
3	2.5	–1
4	3	–1
1	1	–1

取收缩率 $\alpha=\frac{2}{3}$,利用神经网络模型(2)求解规划问题(1)得: $\lambda_1=0.1667$, $\lambda_2=0.1667$, $\lambda_3=0.1667$, $\lambda_4=0.5000$, $\beta_1=0.1333$, $\beta_2=0.1333$, $\beta_3=0.1333$, $\beta_4=0.1333$, $\beta_5=0.4667$.分类结果如图 2 和图 3 所示.

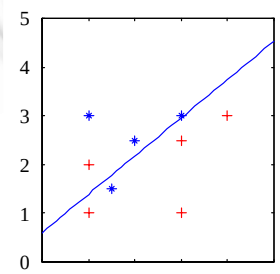

Fig.2 The linear classifier in linearly non-separable case
图 2 线性不可分情形的线性分类器

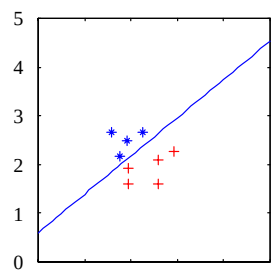

Fig.3 The linear classifier after contraction
图 3 收缩后的线性分类器

4 结 论

SVM是一种基于结构风险最小化原理的非常有发展前景的分类算法.最能说明SVM泛化性能的是线性可分情形下分类超平面的最大边缘性.本文给出实现结构风险最小化原理(最大边缘)的另一种方法.针对线性可分情形,我们提出一种基于闭凸集间距的最大边缘算法,并通过闭凸集的适当收缩来处理线性不可分情形下的线性分类器设计.

与SVM算法的理论体系相比,本文的思路简单、易行,几何意义明确.我们还将结合神经网络和遗传算法[14]考虑本文算法的快速实现,并对非线性分类问题做更深入的研究.

致谢 作者在此感谢审稿人对本文提出了大量有益的建议.

References:

[1] Mehrotha, K., Mohan, C.K., Ranka, S. Elements of Artificial Neural Network. Cambridge, MA: MIT Press, 1997.

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

[3] Cortes, C., Vapnik, V. Support vector networks. Machine Learning, 1995,20:273~297.

[4] Cortes, C. Prediction of generalization ability in learning machines [Ph.D. Thesis]. Department of Computer Science, University of Rochester, 1995.

[5] Osuna, E.E., Freund, R., Girosi, F. Support vector machines: training and applications. Technical Report, AIM1602, Cambridge, MA: MIT Artificial Intelligence Laboratory, 1997.

[6] Cherkassky, V., Mulier, F. Vapnik-Chervonenkis learning theory and its applications. IEEE Transactions on Neural Networks, 1999, 10(5):985~988.

[7] Vapnik, V. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999,10(5):988~999.

[8] Scholkopf, B., *et al*. Imput space versus feature space in kernel-based methods. IEEE Transactions on Neural Networks, 1999, 10(5):1000~1017.

[9] Tao, Qing, Fang, Ting-jian, Fan Jin-song. A kind of new machine learning algorithm: Support Vector Machines. Chinese Pattern Recognition and Artificial Intelligence, 2000,13(3):285~290 (in Chinese).

[10] Yosida, K. Functional Analysis. 5th ed, Berlin: Springer-Verlag, 1978.

[11] Tao, Qing. The neural network based on the constraint domain and its applications in optimization and associative memory [Ph.D. Thesis]. Hefei: University of Science and Technology of China, 1999 (in Chinese).

[12] Xia, Y. A new neural network for solving linear and quadratic programming problems. IEEE Transactions on Neural Networks, 1996,7(6):1544~1547.

[13] Gunn, S. Support vector machine for classification and regression. Technical Report, Image Speech and Intelligent Systems Group Report 9805, University of Southampton, 1998.

[14] Tao, Qing, Cao, Jin-de, Sun, De-min, *et al*. The dynamic genetic algorithm based on the neural network with constraints. Journal of Software, 2001,12(3):462~467 (in Chinese).

附中文参考文献:

[9] 陶卿,方廷健,范劲松.一种新的机器学习算法:Support Vector Machines.模式识别与人工智能,2000,13(3):285~290.

[11] 陶卿.基于约束区域的神经网络模型及其在优化和联想记忆中的应用[博士学位论文].合肥:中国科学技术大学,1999.

[14] 陶卿,曹进德,孙德敏,等.基于约束区域神经网络的动态遗传算法.软件学报,2001,12(3):462~467.

附 录

设 Q_1, Q_2 分别为 $p_1, p_2, ..., p_n$ 和 $q_1, q_2, ..., q_m$ 的闭凸包, Q_1 和 Q_2 线性可分. Q_1 和 Q_2 线性分类器的设计可归结为下列二次规划问题[2~5]:

$$\left.\begin{aligned}&\min_{w,k}\frac{1}{2}\|w\|^2,\\&s.t.y_i(w\cdot x_i+b)\ge 1,i=1,2,...,m+n.\end{aligned}\right\} \quad (a)$$

这里 $x_i\in\{p_1,...,p_n,q_1,...,q_m\}$.当 $x_i\in\{p_1,...,p_n\}$时, $y_i=1$;当 $x_i\in\{q_1,...,q_m\}$时, $y_i=-1$.

当 Q_1和 Q_2 线性不可分时,规划问题(a)是无解的.对这种情形下的线性分类器设计问题,一般都是将误分模式的个数作为目标函数进行优化而得到权值 w 和阈值 b,但这些算法的计算复杂性以 l^d 形式增长[4],其中 l 是模式的个数, d 是输入空间的维数.Cortes 在文献[4]中提出一种将规划问题(a)进行松弛的软边缘算法:

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m+n}\zeta_i$$

$$s.t.\ y_i(w\cdot x_i+b)\ge 1-\zeta_i,\zeta_i\ge 0,i=1,2,...,m+n.$$

其中 C 是预先给定的常数.这种算法以 $\sum_{i=1}^{m+n}\zeta_i$ 表示误分的测度,与输入空间的维数无关,以全局观念考虑了分类问题.以这种方法确定的超平面也称为广义最优超平面[13].

Maximal Margin Linear Classifier Based on the Contraction of the Closed Convex Hull*

TAO Qing[1,2], SUN De-min[3], FAN Jin-song[4], FANG Ting-jian[4]

[1](*Institute of Automation, The Chinese Academy of Sciences, Beijing* 100080, *China*);

[2](1*st Department, Artillery Academy of PLA of China, Hefei* 230031, *China*);

[3](*Department of Automation, University of Science and Technology of China, Hefei* 230027, *China*);

[4](*Hefei Institute of Intelligent Machines, The Chinese Academy of Sciences, Hefei* 230031, *China*)

E-mail: q_tao@sohu.com

http://www.ia.ac.cn

Abstract: The SVM (support vector machines) is a classification technique based on the structural risk minimization principle. In this paper, another method is given to implement the structural risk minimization principle. And an exact maximal margin algorithm is proposed when classification problem is linearly separable. The linearly non-separable problem can be changed to separable linearly by using the proposed concept of the contraction of a closed convex set. The method in this paper has the same function and quality as SVM and Cortes' soft margin algorithm, but its theoretical system is simple and strict, and the geometric meaning of its optimization problem is very clear and obvious.

Key words: closed convex set; contraction; support vectors; maximal margin; classifier

* Received March 6, 2000; accepted September 20, 2000

Supported by the National Natural Science Foundation of China under Grant No.60175023; the Natural Science Foundation of Anhui Province of China under Grant No.01042304; the Excellent Youth Foundation of Anhui Province of China