基于邻域原理计算海量数据支持向量的研究*

张文生, 丁 辉, 王 珏

(中国科学院 自动化研究所,北京 100080)

E-mail: zhangws330@sina.com

http://www.ia.ac.cn

摘要: 使用支持向量机理论计算海量数据的支持向量是相当困难的. 为了解决这个问题,提出了基于邻域原理计算支持向量的方法. 在对支持向量机原理与邻域原理比较分析的基础上讨论了以下问题:(1) 构建了从样本空间经过特征空间到扩维空间的复合内积函数,给出计算支持向量的邻域思想;(2) 将支持向量机的理论建立在距离空间上,设计出了计算支持向量的邻域算法,从而把该算法理解为简化计算二次规划的方法;(3) 实验结果说明,邻域原理可以有效地解决对海量数据计算支持向量的问题.

关键词: 支持向量;最优超平面;二次规划;邻域

中图法分类号: TP301 **文献标识码**: A

1995 年,Vapnik 全面阐述了统计学习理论[1,2],其原始动机之一是试图将人工神经网络的研究回归到感知机. 换句话说,如果这个设想成立,基于非线性优化的人工神经网络的研究就可以变换为线性优化的问题. 这对机器学习的研究来说,无疑是十分诱人的. 这就是近几年统计学习理论得到各国研究者与工程师重视的原因[3].

在文献[1,2]中,Vapnik 证明,如果选择满足 Mercer 条件合适的非线性函数(核函数),则一定存在一个非线性映射,将线性不可分的样本集映射到另一个高维线性空间中,使之线性可分. 在理论上,对一个给定的样本集,怎样选择非线性映射,将这个样本集映射到线性可分的高维内积空间,还是一个未解决的问题,本文不准备讨论这个问题. 在计算上,线性可分问题可以转换为借助二次规划求解支持向量(support vector,简称 SV)的问题. 然而,求解二次规划将涉及到对 m 阶矩阵的计算(m 为样本的个数),随着 m 的增大,Burges 等诸多研究者发现,支持向量机(support vector machine,简称 SVM)学习海量样本困难重重. Microsoft 研究院的 Platt 甚至得出以下结论:从计算理论上分析,在个人计算机上,用 SVM 技术处理样本个数的规模界限一般为 4 000 个[4]. 如何在技术上解决这个问题,是本文所讨论的主要内容.

1999 年,我国学者张铃和张钹独立地发表两篇文章[5,6],他们采用球面投影函数作为非线性映射,完成样本点的分类问题,这与 Vapnik 的思想在本质上是相同的[1,2,7]. 但是,Vapnik 将分类问题变换为计算二次规划的最优解问题,而张铃和张钹教授则使用了球邻域的概念,这意味着,将计算分类超平面的问题转换为计算样本点两两之间距离所构成的距离空间上的覆盖问题. 这样,他们所提出的方法完全避开了解二次规划对计算资源要求过高的问题. 本文依据文献[5]的邻域原理,通

* 收稿日期: 2000-06-20; 修改日期: 2000-09-11

基金项目: 国家重点基础研究发展规划 973 资助项目(G1998030508);国家 863 高科技发展计划资助项目(863-306-ZT06-07-1);国家自然科学基金资助项目(79700023);航空基础科学基金资助项目(97J55009)

作者简介: 张文生(1966—),男,河南郑州人,博士生,副教授,主要研究领域为机器学习,神经网络,智能控制;丁辉(1974—),男,山东日照人,硕士生,主要研究领域为人工智能理论及应用;王珏(1948—),男,江苏苏州人,研究员,博士生导师,主要研究领域为人工智能理论和方法,适应性与数据挖掘,人工神经网络.

过构建复合内积函数,提出了计算支持向量的方法,该方法可以有效地解决计算海量数据的支持向量问题.

1 支持向量机原理与邻域原理的比较

一般地说,支持向量机(SVM)理论有以下4个要点:(1) 非线性映射,是这个理论的基础;(2) 对特征空间划分的最优超平面(optimal hyperplane,简称OHP),是SVM的目标;(3) 支持向量(SV),是SVM的结果;(4) 二次规划,是计算SV的手段.为此,我们将对文献[5]所提出的原理与SVM原理进行比较,但是,在以下的比较中,有些涉及到文献[5]的内容,有些将涉及更为广泛的称为邻域原理(见第2节)的内容.

1.1 非线性映射

SVM原理是试图通过选择非线性变换$\phi(x)$,将线性空间D上线性不可分的样本点映射到特征空间H,使得对应的样本点在H中线性可分.根据Mercer定理,选择合适的非线性变换等价于选择满足Mercer定理条件的核函数$K(x,y)$,因此,无须显现地了解非线性映射的具体形式与特征空间的几何性质.

文献[5]是显现地给出了特征空间的形式——球面空间,也即使用球函数将样本空间中的样本点映射到一个半球面上.显然,写出这个球函数对应的核函数是不困难的,因此,可以将这个方法理解为SVM的一个特例.

1.2 最优超平面

按照SVM原理,在特征空间上,使用最优超平面(OHP)对样本集进行线性划分.如果在特征空间上样本集不能被线性划分,该原理允许存在分类误差,但是应保证用一个超平面来划分样本集.文献[5]则要求不能有分类误差,并且如果使用一个超平面无法完全划分,则可以允许使用多个超平面来划分,这与Widrow的Madaline考虑是一致的[8].

1.3 支持向量(SV)

SVM理论追求学习具有泛化能力最强的目标,实现的过程是首先求出SV,然后求出OHP.对SVM理论,SV是OHP:$(w\cdot x)-b=0$距离最近的样本点,并且同一类的SV与OHP距离完全相等,不同类的SV与OHP距离不一定相等.

对文献[5],由于采用特殊映射,使得超平面$(w\cdot x)-b=0$中的项$w\cdot x$可以用来度量两点之间的距离,因而超平面方程可以理解为以w为球心,以b为半径球邻域边界的方程.使得SV可以理解为距离最优球邻域边界最近的那些样本点,这与SVM对SV的理解是完全一致的.另外,由于文献[5]允许通过多个超平面(球邻域)来划分样本点,所以,对应于每一个超平面,有一组SV,全部SV集为每一组SV的并集.值得注意的是,SV保持距离各自最优球邻域的边界最近,但是,同一类样本点相应的最近距离值不再保持完全相等.

1.4 二次规划计算与邻域计算

SVM与文献[5]所使用的方法都要计算一个$m\times m$内积构成的矩阵M(SVM理论中目标函数的Hessian矩阵,简称Hessian矩阵).这个矩阵的每一个元素,可以理解为样本集合中两个样本之间距离的测量.

SVM求解SV和OHP是通过求解二次规划来实现的,求解二次规划的核心是对矩阵M进行化简,这涉及到解一个线性方程组的问题.尽管样本点线性可分,如果选择的核函数不合适,则这个

方程组有无穷多解或无解,需要使用优化方法求出二次规划的近似最优解,从而获得 SV 的近似最优解.

而文献[5]则采用了一种与 SVM 理论完全不同的方法,由于该方法把 Hessian 矩阵的元素理解为两两样本之间距离的测量,从而可以引入邻域来简化 Hessian 矩阵.这意味着,用文献[5]的方法无须考虑 Hessian 矩阵的复杂计算,这恰恰是这个方法大大节省计算资源的原因.

在一定条件下,本文将文献[5]中特定的非线性映射扩展到任意非线性映射,由此,才能把文献[5]推广以用于求解 SV.

2 计算 SV 的邻域原理

在本节中,我们试图建立求 SV 的邻域原理,其目的是用邻域的方法求 SV.由于邻域的概念涉及到距离,而在 SVM 理论中,讨论 OHP 问题仅用到特征空间中向量的内积.在特征空间中,两个向量的内积与其距离没有一定的序关系(见第 2.2 节),为了建立内积与距离之间的序关系,必须对样本点作适当的限制.为此,我们将特征空间扩维,在扩维后的空间中定义内积和距离,使得样本点之间的内积关系能够真正表征它们之间的距离关系.

2.1 基本假设条件

假定 m 个训练样本:$K=\{x^t(i)\mid t=1,2,i=1,\ldots,m\}$,其中样本点的坐标 $x^t(i)\in D$,不失一般性,假设样本点的类别 $t\subset\{1,2\}$.选定合适的核函数 $K(x,y)$,线性空间 D 中的样本点经过非线性变换 $\phi(x)$ 映射到特征空间 H 上线性可分,但是,非线性变换 $\phi(x)$ 未知.$x^t(i)$ 经过 $\phi(x)$ 映射到 H 空间上的 $h^t(i)$,简记为 $h^t=\phi(x^t(i))$.不失一般性,我们假设 $D=R^{n_0}$,$H=R^{n_1}$,$n_1>n_0$,n_0 和 n_1 均为自然数.

由于样本集合 $L=\{h^t(i)\mid t=1,2,i=1,\ldots,m\}$ 在 H 空间上线性可分,那么,一定存在一个 H 上的 OHP:$(w\cdot x)-b=0$ 将样本点分成两部分.由 Mercer 定理可知,$K(x,y)$ 是 H 上的内积,H 作为 Hilbert 空间可以定义距离,在此距离意义下,满足无误差地划分两类样本点,并且最近样本点与 OHP 的距离最大.

2.2 内积与距离之间的关系

在线性空间 H 上,定义一种内积 (x,y),$x,y\in H$.在内积的基础上,我们引入范数 $\|x\|=\sqrt{(x,x)}$,$x\in H$.进而可以在 H 上定义距离 $\rho(x,y)=\|x-y\|=\sqrt{(x,x)+(y,y)-2(x,y)}$,这样,线性空间 H 就成为了距离空间.

易见,$\forall\ x^1,y^1,x^2,y^2\in H$,当 $(x^1,y^1)>(x^2,y^2)$ 时,不一定有 $\rho(x^1,y^1)\geqslant\rho(x^2,y^2)$,或者 $\rho(x^1,y^1)\leqslant\rho(x^2,y^2)$ 恒成立.即使是 $x^1=x^2$,上述结论也未必成立,因此,在 SVM 理论中,由二次规划目标函数的 Hessian 矩阵中一个元素 $K(x^i,x^j)$ 的大小关系不能推出 $\rho(x^i,x^j)$ 的大小关系.

为了使两个向量内积的大小与距离的大小建立保序关系,必须对向量作必要的限制,最简单、最直观的方法就是限制向量的范数相等.

2.3 球面变换

在内积空间 H 上,根据 D 中向量 x,y 与 H 中向量的对应关系,可以定义 H 上的内积 $K(x,y)$,进而可以定义 H 上的范数.我们将 H 空间上的样本集 $L=\{h^t(i)\mid t=1,2,i=1,\ldots,m\}$ 映射到扩维空间 F 中正半球面 S^{n_1} 上,此时,扩维空间 $F=R^{n_1+1}$.

定义 1. 设 H_1 是空间 H 的有界子集,定义一个一一变换 $T:H\to S^{n_1}$,对 $\forall\ h\in H_1$,有 $T(h)=$

$(h,\sqrt{d^2-\|h\|^2})$,其中$\|h\|$为h在空间H中的范数,$d=\max\{\|h\|\ |h\in H_1\}$,称变换$T$为球面变换.

特别地,空间H中样本集L的最大半径为$R=\max\{\|h^i(i)\|\ |h^i(i)\in L\}$,$h^i(i)$经过球面变换$T$成为$s^i(i)$,其中$s^i(i)=(\phi(x^i(i)),\sqrt{R^2-\|\phi(x^i(i))\|^2})$.

2.4 R^{n_1+1}上向量的内积

由于S^{n_1}在线性空间R^{n_1+1}中,为了衡量S^{n_1}两点的内积,必须定义R^{n_1+1}中的内积.

定理1. 设$h^i=\phi(x)$和$h^j=\phi(y)\in H$,$x,y\in D$,内积空间H中向量h^i与h^j的内积为$K(x,y)$,任取$s^i,s^j\in F$,其中$s^i=(h^i,x_{n+1})$,$s^j=(h^j,y_{n+1})$,$x_{n+1},y_{n+1}\in R^1$.令$(s^i,s^j)=K(x,y)+x_{n+1}y_{n+1}$,则$(s^i,s^j)$是$R^{n_1+1}$上的内积,简记为$S(s^i,s^j)=(s^i,s^j)$.

由内积定义的3个条件容易验证定理1成立.

定义2. 我们将R^{n_1+1}上的内积$S(s^i,s^j)$限制在半球面S^{n_1}上,使得S^{n_1}上的任意两点有内积.我们称这种受限制的内积为球面内积函数.

按照$\rho(x,y)$的定义(见第2.2节),S^{n_1}上的任意两点可以定义距离$\rho(s^i,s^j)$,其中$s^i,s^j\in S^{n_1}$.

2.5 求SV的邻域原理

首先,定义扩维空间F上的复合内积$S\circ K(x,y)$.根据上述结论和符号规定,特征空间H借助原空间D可以定义内积$K(x,y)$,扩维空间F借助特征空间H可以定义内积$S(s^i,s^j)$,由此,扩维空间F有一个复合内积$S\circ K(x,y)$.而该复合内积$S\circ K(x,y)$对应于D到F的一个核函数,从而扩维空间F的内积满足Vapnik关于SVM核函数的一切性质.但是,这个核函数同时又满足距离与内积之间的保序关系.

其次,定义半球面S^{n_1}上的球邻域.扩维空间F有了内积,按照内积可以定义距离,使得扩维空间F成为距离空间,在半球面S^{n_1}上定义球邻域为$w^1\cdot s-b_1>0(w^1,s\in S^{n_1},b_1\in R^1)$,其中$w^1$为球心,$s$为半球面$S^{n_1}$上的点,$b_1$为半径.

最后,在扩维空间F上求SV和OHP.按照Vapnik对SV和OHP的定义,我们不在特征空间H上求SV和OHP,而是在扩维空间F中求SV和OHP.把样本点限制在球面S^{n_1}上,用求SV的邻域算法(见第3节)求SV和最优球邻域:$w\cdot s-\varphi>0$,进而在扩维空间F上求出OHP:$w\cdot s-\varphi=0$.

我们把上述思想称为SV的邻域原理.

2.6 S^{n_1}上样本的线性可分性

由于原样本点在H上线性可分,我们来考察经过球面变换T后,对应的样本点在S^{n_1}上的可分性.

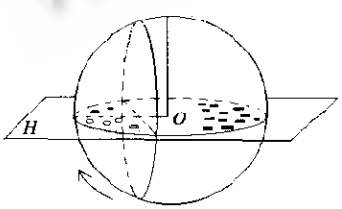

Fig.1 Covering by sphere neighborhood

图1 用球邻域覆盖

定理2. 两类样本集在H空间上线性可分,那么,经过变换T,在正半球面S^{n_1}上一定存在一个球邻域将两类样本分开.

证明:因为两类样本集在H空间上线性可分,所以在特征空间H中存在分离超平面$P:(w\cdot x)-b=0$.由于定义1中的变换T为一一变换,那么,我们可以通过不断转动P得到球面S^{n_1}上的一个球邻域,并且该球邻

域覆盖某一类所有的样本点(如图1所示),记该球邻域为 $w^1 \cdot s-b_1>0(w^1,s\in S^{n_1},b_1\in R^1)$. □

2.7 球邻域的性质

由基本假设,我们选择适当的核函数,使两类样本集在 H 上线性可分.为了用邻域原理求SV,我们必须研究球邻域的性质,为此,首先给出一个定理.

定理3. 若样本集在 H 空间上线性可分,那么,覆盖某一类全部样本点体积最小的超球一定存在且惟一.

证明:我们指出:要求覆盖某一类全部样本点体积最小的超球 B,等价于求覆盖某一类全部样本点半径 R 最小的超球 B.因此,要证明此定理,只要证明满足如下非线性规划的解存在,并且惟一即可.

$$(\mathrm{NP})\qquad \min F(R,a)=R^2,$$
$$\text{s.t.}\quad (x^i-a)^T\cdot(x^i-a)\leqslant R^2,\quad x^i\in K_1,$$

其中 $\overline{K_1}$ 为第1类样本投影到球面 S^{n_1} 上所构成的点集.我们引入Lagrange乘子 $\alpha_i\geqslant 0$,可以得到

$$L(R,a,\alpha)=R^2-\sum_i \alpha_i\{R^2-(x^i-a)^T\cdot(x^i-a)\}.$$

对 R,a 求偏导数可得 $\sum_i \alpha_i=1$,$a=\sum_i \alpha_i x^i$.代入上式,这样,式(NP)变为二次规划问题:

$$(\mathrm{QP})\qquad \max L(\alpha)=\sum_i \alpha^i(x^i,x^i)-\sum_{i,j}\alpha^i\alpha^j(x^i,x^j),$$
$$\text{s.t.}\quad \sum_i \alpha_i=1,\quad \alpha_i\geqslant 0.$$

因为式(QP)的Hessian矩阵正定(Mercer定理),所以 α_i 存在且惟一.我们取球心为 $a=\sum_i \alpha_i x^i$,半径 $R=\sqrt{\max\{(x^i-a)^T\cdot(x^i-a)\mid x^i\in K_1\}}$,则本定理成立. □

事实上,定理3在几何意义上揭示了最小超球 B 与球面 S^{n_1} 相交圆成一个圆周,该圆是超球 B 的大圆,圆心为超球的球心 M,并且圆周上至少有一个某类样本点,将 OM 延长到球面 S^{n_1} 上,得到 S^{n_1} 惟一一点,记为 ω^1(如图2所示).

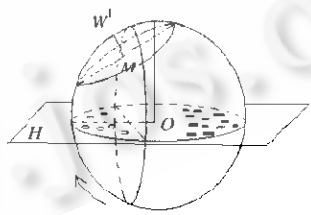

Fig. 2 The minimal hypersphere and the maximal circle

图2 最小超球与最大圆

有了定理3,我们就很容易证明下面的定理.

定理4. 样本点在特征空间 H 上线性可分,经过球面投影 T 将样本点映射到 S^{n_1},那么,球面上覆盖某一类所有样本,并且面积最小的球邻域存在且惟一.

证明:由定理2和定理3容易得出定理4成立. □

3 计算SV的邻域算法

为了表达简洁,我们引入记号:$I(t)=\{i\mid y^i=y^t\}$,$K_t=\{x^i\mid i\in I(t)\}$.$d^1(i)=\max_{i\in I(t)}\{s^i,s^t\}$,$d^2(i)=$

$\min_{j\in I(0)}[\langle s^j,s^j\rangle>\max_{j\in I(0)}\langle s^j,s^j\rangle]$, $\varphi_i=\frac{1}{2}\{d^1(i)+d^2(i)\}$, $c_i(t)=\{s\mid\langle s,s^j\rangle>\varphi_i\}$, $w^j=s^j$, 其中 $t=1,2$. 由于样本点经过 T 映射到球面 S^{n_1} 上, 所以本节在 R^{n_1+1} 中讨论问题.

3.1 邻域算法的几何解释

由于样本集在半球面 S^{n_1} 上线性可分, 因此一定存在以 S^{n_1} 上某点为中心的球邻域, 使得该球邻域覆盖某一类所有样本点, 同时不覆盖异类样本点. 在 R^{n_1+1} 中, SV 只能从两类样本集中选取, 如果真正把 OHP 求出后再求 SV, 那么, 求 SV 的过程相当复杂. 在样本集规模很大的情况下, 我们换一个角度考虑问题, 在第 1 类样本集 K_1 中选定一个样本点 x^j, 分别求 x^j 与 K_1 和 K_2 中所有样本点的内积, 然后, 以 x^j 为中心作球邻域 $c_j(1)$. 如果 $c_j(1)$ 覆盖 K_1 中所有样本点, 那么, $c_j(1)$ 的边界 $\langle s,s^j\rangle-\varphi_j=0$ 可以看作一个分类超平面. 现在要问 $c_j(1)$ 的边界是否为 OHP? 如果 $c_j(1)$ 的边界是 OHP, 根据 $d^1(i)$, $d^2(i)$ 分别表示异类样本点中与 x^j 最近的距离、同类样本点中与 x^j 最远的距离, 按下面算法求出的 x_{1sv} 和 x_{2sv} 就是 SV. 如果 $c_j(1)$ 的边界不是 OHP, 则 x_{1sv} 和 x_{2sv} 作为 SV 的近似, $c_j(1)$ 的边界作为近似 OHP.

图 3(a)为给定两类样本的分类问题, 圆圈表示第 1 类, 黑点表示第 2 类. 用下面介绍的 SV 邻域算法, 经过图 3(b)~(d)到图 3(e), 图 3(e)表示求得 SV 集合(在原来的点上画叉表示). 图 3(f)表示用 SVM 原理求得 SV 集合(在原来的点上画叉表示).

Fig. 3 The process of classification using sphere neighborhood

图 3 用球邻域分类的过程

根据 SV 邻域原理的思想, 我们给出求 SV 的 3 种邻域算法, 这 3 种算法的复杂度都是多项式的. 局部 SV 邻域算法与样本点的次序有关, 但是算法速度相当快; 近似最优 SV 邻域算法与样本点的次序无关, 求出的 SV 的个数较少, 计算速度也比较快; SV 邻域算法不但具有前两个算法的优点, 而且还考虑到样本集的泛化能力. 当然, 邻域原理还有很多变形, 本节我们只列举其中 3 种典型算法.

不妨记全体样本点的 SV 集合为 V, 并且将其初始化为 $V=\varnothing$. $Card(x)$ 表示求集合 x 元素个数的函数.

3.2 算法 1(局部 SV 邻域算法)

(1) 任选 $x^j\in K_1$, 计算 $d^1(i)$, $d^2(i)$.

(2) 分别计算 $x_{1sv}\in K_1$ 和 $x_{2sv}\in K_2$, 使 $d^2(i)$, $d^1(i)$ 成立, $V=V\cup\{x_{1sv}\cup\{x_{2sv}\}$, 令 $K_1=K_1\backslash c_j(1)$.

(3) 如果 $K_1\neq\varnothing$, 则 K_1 和 K_2 互换, 返回(1); 如果 $K_1=\varnothing$, 则停止.

3.3 算法 2(近似最优 SV 邻域算法)

(1) 任选 $x^j\in K_1$, 计算 $d^1(i)$, $d^2(i)$. 令 $n_j(1)=Card(c_j(1)\cap K_1)$, $n_{j_1}=\max_{x^j\in K_1}n_j(1)$.

(2) 计算 $x_{1sv}\in K_1$, $x_{2sv}\in K_2$, 分别使 $d^2(j)$, $d^1(j)$ 成立, $V=V\cup\{x_{1sv}\cup\{x_{2sv}\}$, 令 $K_1=K_1\backslash c_j(1)$.

(3) 如果 $K_1\neq\varnothing$,则 K_1 和 K_2 互换,返回(1);如果 $K_1=\varnothing$,则停止.

3.4 算法 3(SV 邻域算法)

(1) 任选 $x^i\in K_1$,计算 $d^1(i)$,$d^2(i)$. 令 $n_i(1)=Card(c_i(1)\cap K_1)$,$n_{i_1}=\max\limits_{x^i\in K_1} n_i(1)$.

(2) 计算 $x_{1sv}\in K_1$,$x_{2sv}\in K_2$,分别使 $d^2(i_1)$和 $d^2(i_1)$成立,则 K_1 的预选 SV 集 $V_1=\{x_{1sv}\}\cup\{x_{2sv}\}$.

(3) 对 K_2 重复(1)和(2)的操作,求出 K_2 的预选 SV 集 V_2.

(4) 如果 $\max\limits_{x^i\in K_1} n_i(1)=\max\limits_{x^j\in K_2} n_j(2)$,则计算 $d(t)=\max\limits_{K_1,K_2}\{d^1(i)+d^2(i)\}$,选择 K_t 类,$V=V\cup V_t$;否则,计算 $n_{i_t}(t)=\max\{\max\limits_{x^i\in K_1} n_i(1),\max\limits_{x^j\in K_2} n_j(2)\}$,选择 K_t 类,$V=V\cup V_t$.

(5) 不妨设(4)中选中 K_t,令 $K_t=K_t\backslash c_{i_t}(t)$.

(6) 如果 $K_t\neq\varnothing$,则返回(1);如果 $K_t=\varnothing$,则停止.

3.5 小 结

对上面 3 种算法进一步分析,我们可以看到下面几个问题:

其一,对于用邻域算法求 SV 而言,由于球邻域的中心 w^i 不是在全部球面 S^{n-1} 上选取,而是在全部样本点中选取,因此,尽管样本点线性可分,并且每一类有一个球邻域覆盖该类所有样本点,然而邻域算法求出的 SV 和分类超平面也有可能是次优解.

其二,在特征空间 H 上,如果用一个超平面不能线性划分两类样本点,即一个球邻域不能完全覆盖某一类样本点,上述邻域算法可以实现多个超平面划分样本点的思想,求出多个分离超平面和相应的多组 SV,这个思想与 50 年代 Widrow 提出的 Madline 是一致的[8]. 对上述邻域算法,如果有多个分离超平面($k>1$),我们可以按文献[6]中的算法的第 2 步,解系数为 0 或 1 的简单线性不等式组,容易求出一组权值和阈值,得到一个近似 OHP.

其三,SV 的邻域原理是一种思想,在使用中有很多技巧和变形,该思想把寻找覆盖问题最优解这一 NP 完全问题转化为求可行解这一 P 问题. 本文的目的在于用这种思想,把 Vapnik 二次规划求解问题转化为邻域算法求解 SV 问题.

其四,我们使用邻域原理的主要原因在于,该原理有全局寻优和局部寻优等多种变形算法,对于大规模数据,该方法计算速度极快,同时对计算机资源要求很低. 通过大量实验我们发现,邻域原理求出的 SV 普遍满足问题的要求.

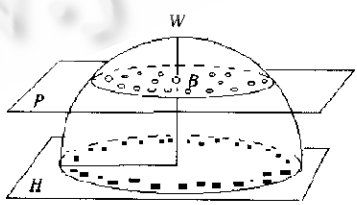

Fig. 4 Interpretation of SV and OHP in sphere neighborhood method

图 4 用球邻域法解释支持向量和最优超平面

其五,有时对给定的核函数,样本点在特征空间中线性不可分,经过特征空间扩维和样本点进行球面变换,能够使得样本点线性可分,进而,按照邻域算法能求出分类超平面和 SV. 此时,我们很容易用邻域原理求出 SV 和 OHP(如图 4 所示). 当然,如果对核函数选择不合适,或者 SVM 方

法与球邻域法所选择的核函数不一致,那么球邻域法求出的SV不一定与SVM理论求出的SV一致(见第4.2节).

4 实验结果分析

我们已经提出了基于邻域原理的简化求分离超平面的方法,本节根据核函数的异同,对小规模、中型规模和大规模样本数据应用两种方法求SV.在核函数相同的情况下,通过计算可以发现:在小样本情况下,邻域原理与Vapnik提到的SVM方法求出的SV基本相同;在中型样本情况下,邻域原理与Vapnik提到的SVM方法求出的SV有微小差别;在大样本情况下,对非线性变换映射后不可分的样本集,Vapnik提出的SVM方法难以精确求出SV,而邻域原理方法则很容易求出近似最优的SV.我们看到,对于海量数据(2×10^7),邻域原理方法求解SV计算速度极快,同时对计算机资源要求很低,而SVM原理不具有这种优点,由此可见,对大规模样本点进行分类,求SV的邻域算法得到的结果令人满意.在核函数不同的情况下,用两种方法求解各有优缺点.

4.1 SV的邻域算法与SVM算法选择相同的核函数

我们选择Vapnik介绍的常用核函数(多项式函数$d=2$,径向基函数$\sigma=1$)进行实验,结果基本相同,下面仅给出多项式函数($d=2$)的结果.

4.1.1 对小规模样本实验的结果(见表1)

Table 1 The experimental results of small scale samples

表1 小规模样本集的实验结果

Number of samples①	Linear separability②	Number of SV (neighborhood method)③	Number of SV (SVMs method)④	Number of different SVs⑤
29	Linear separable (Fig. 5(a))⑥	3	3	0
30	Linear separable (Fig. 5(b))⑦	4	4	0
25	Linear nonseparable (Fig. 5(c))⑧	3	3	0
23	Linear nonseparable (Fig. 5(d))⑨	8	8	0

①样本点个数,②线性可分性,③SV个数(邻域法),④SV个数(SVM法),⑤不同的SV个数,⑥线性可分(图5(a)),⑦线性可分(图5(b)),⑧线性不可分(图5(c)),⑨线性不可分(图5(d)).

4.1.2 对中型规模复杂样本实验的结果

对于印度的S. S. Keerthi等人所提供的194个样本点的Spiral螺线(线速度改变量$=-0.063651$,角速度改变量$=0.195476876$,圈数$=3$)和其他一些数据进行实验,结果见表2.

Table 2 The experimental results of middle scale samples

表2 中规模样本的实验结果

Number of samples①	Linear separability②	Number of SV (neighborhood method)③	Number of SV (SVMs method)④	Number of different SVs⑤
194	Linear nonseparable (Fig. 5(e))⑥	82	82	0
1 500	Linear separable (Fig. 3)⑦	9	8*	2

* represents the approximate optimal solution to solve the quadratic programming by using Wolfe method and Platt's SMO method⑧.

①样本点个数,②线性可分性,③SV个数(邻域法),④SV个数(SVM法),⑤不同的SV个数,⑥线性不可分(图5(e)),⑦线性可分(图3),⑧*表示用Wolfe方法和Platt的SMO方法求解二次规划得到的近似最优解.

4.1.3 对大规模样本实验的结果

用UCI标准数据库构造10^4个Spiral螺线生成的样本点,圈数$=9$,实验结果见表3.

Table 3 The experimental results of large scale samples

表 3 大规模样本的实验结果

Number of samples①	Linear separability②	Number of SV (neighborhood method)③
1×10^4	Linear separable (Fig. 5(e))④	2 116
2×10^7	Linear separable (Fig. 5(f))⑤	5

①样本点个数,②线性可分性,③SV 个数(邻域法),④线性不可分(图 5(e)),⑤线性可分(图 5(f)).

上述结果可以直观地用图 5 来表示.

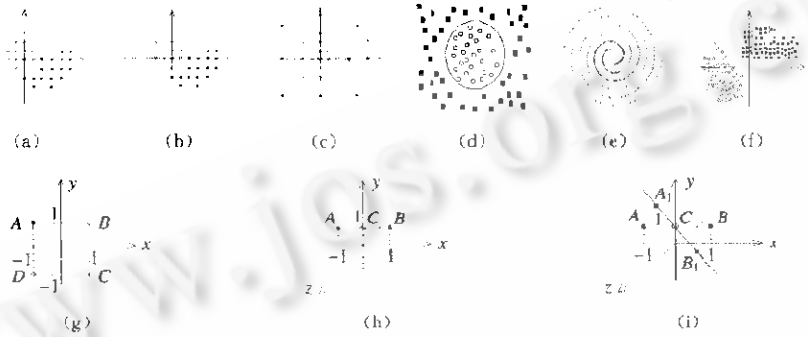

Fig. 5 Illustration of the experimental results

图 5 实验结果的图示

4.2 SV 的邻域算法与 SVM 算法选择不同的核函数

我们选择 XOR 问题作为原空间上的样本点,如图 5(g)所示,用球邻域法选择球面内积函数 $K_{sph}(x,y)=x_1x_2+y_1y_2+\sqrt{(R^2-x_1^2-x_2^2)(R^2-y_1^2-y_2^2)}$,非线性变换 $T((x_1,x_2))=(x_1,x_2,\sqrt{R^2-x_1^2-x_2^2})$,SVM 理论选择核函数 $K_{svm}(x,y)=x_1x_2+y_1y_2+x_1x_2y_1y_2$,非线性变换为 $f((x_1,x_2))=(x_1,x_2,x_1x_2)$. 前者在高维特征空间中线性不可分,后者在高维特征空间中线性可分.我们选择另外一组样本点,直线上的 3 点,$K_1=\{(1,1),(-1,1)\}$,$K_2=\{(0,1)\}$,如图 5(h)和(i)所示,核函数、非线性变换选择同上.实验结果见表 4.

Table 4 The experimental results of different kernel functions by using two methods

表 4 不同核函数用两种方法的实验结果

Number of samples⑥	Property of the problems⑦	Number of SV (neighborhood method)⑧	Number of SV (SVMs method)⑨
4	XOR (Fig. 5(g))⑩	(Linear separable⑫) 4	(Linear nonseparable⑬) 4
3	Three points on one line (Fig. 5(h), Fig. 5(i))⑪	(Linear separable) 3	(Linear nonseparable) 3

①样本点个数,②问题的性质,③SV 个数(邻域法),④SV 个数(SVM 法),⑤XOR(图 5(g)),⑥线性可分,⑦线性不可分,⑧直线上的 3 点(图 5(h)和图 5(i)).

5 结 论

本文首先比较分析了 SVM 原理与邻域原理,然后,把 SVM 原理建立在距离空间上,设计出基于邻域原理的计算 SVM 的邻域算法,并进行了实验分析.通过上述讨论,我们发现,在线性可分的情况下,邻域原理和 SVM 理论都可以解决分类问题,求出 SV 和最优超平面,但是,在大规模样本的情况下,用邻域原理近似计算 SV 问题显得方便、有效.

事实上,邻域原理求 SV 的过程是求一组近似 SV 的过程,该过程本质上是简化 SVM 理论中

二次规划目标函数的 Hessian 矩阵.该方法不但几何意义明确,而且计算速度快(该算法复杂度是多项式的),每次可以消掉内积矩阵的多行多列,使得计算机内存开销小.

值得注意的是,本文所讨论的问题的前提是存在一个合适的核函数,样本点在特征空间中线性可分,而选取合适的核函数是实现 SVM 的核心问题,一般认为,求核函数需要知道样本点适当的先验知识.应当指出,样本点在特征空间中线性不可分的情况下,邻域原理方法也能很好地解决分类问题,求出近似 OHP.

References:

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

[2] Vapnik, V. N. Statistical Learning Theory. New York: John Wiley & Sons, Inc., 1998.

[3] Cherkassky, V., Mulier, F., Vapnik, V. N., *et al*. Special issue on VC learning theory and its applications. IEEE Transactions on Neural Networks, 1999,10(5):985~1098.

[4] Scholkopf, B., Burges, J. C., Smola, A. J. Advances in Kernel Methods Support Vector Learning. Cambridge, MA: MIT Press, 1999.

[5] Zhang, Ling, Zhang, Bo. A geometrical representation of McCulloch-Pitts neural model and its applications. IEEE Transactions on Neural Networks, 1999,10(4):925~929.

[6] Zhang, Ling, Zhang, Bo. Neural networks based on classifiers for a vast amount of data. In: Ning, Zhong, Zhou, Li-zhu, eds. Proceedings of the 3rd Pacific-Asia Conference PAKDD-99, Methodologies for Knowledge Discovery and Data Mining. Beirlin: Springer-Verlag, 1999. 238~246.

[7] Zhang, Bo, Zhang, Ling, Wu, Fu-chao. Programming based learning algorithms of neural networks with self-feedback connections. IEEE Transactions on Neural Networks, 1995,6(3):771~775.

[8] Widrow, B., Winter, R. G. Layered neural nets for pattern recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1988,36(3):1109~1118.

Study on Computing the Support Vectors of Massive Data Based on Neighborhood Principle*

ZHANG Wen-sheng, DING Hui, WANG Jue

(Institute of Automation, The Chinese Academy of Sciences, Beijing 100080, China)

E-mail: zhangws330@sina.com

http://www.ia.ac.cn

Abstract: It is quite difficult to compute the support vectors of massive data using the theory of support vector machine. To solve this problem, a method is brought forward to compute support vectors based on the neighborhood principle in this paper. Several questions are discussed based upon comparison and analysis of the support vector machine theory and the neighborhood principle as below: (1) The inner product function from the sample space to the dimension-expand space via the feature space is constructed, and the neighborhood principle of computing the support vectors is presented; (2) Vapnik's support vector machine theory is constructed on the distance space, the algorithm is designed to compute support vectors, and the algorithm is regarded as a method to reduce the computation of quadratic programming; (3) The experimental results show that the neighborhood principle can solve the problem of support vector computation of massive data effectively.

Key words: support vector; optimal hyperplane; quadratic programming; neighborhood

* Received June 20, 2000; accepted September 11, 2000

Supported by the National Grand Fundamental Research 973 Program of China under Grant No. G1998030508; the National High Technology Development Program of China under Grant No. 863 306 ZT06 07 1; the National Natural Science Foundation of China under Grant No. 79700023; the Aeronautic Basic Science Foundation of China under Grant No. 97J55009