文章编号:1672 - 058X(2011)05 - 0453 - 05

基于几何算法的支持向量机分类方法

张 瑞 环
(重庆大学 数学与统计学院,重庆 400044)

摘 要:支持向量机是基于统计学习理论的一种学习方法,提出了基于压缩超球体的SVM 分类问题的一种几何方法,具有直观、简单和易于实现的特点;通过实例验证,说明此方法的可行性和有效性,具有一定的推广价值。

关键词:支持向量机;几何算法;压缩凸包;压缩超球体
中图分类号:TP241 **文献标志码**:A

20 世纪 90 年代,Vapnik 等人提出的支持向量机[1-2](support vector machine,SVM)是一种基于统计学习理论的的机器学习方法。它可以克服神经网络[3]分类和传统统计分类方法的许多缺点,适宜于小样本高维度的分类问题,具有较好的泛化能力等特点。因此,支持向量机已经成为解决分类和回归问题的一个行之有效的方法,被成功应用到许多学科领域,如人脸识别[4]、文本分类[5]、故障诊断[6]、基因表示分析[7]等。此处根据直观的几何特征,提出基于几何方法的支持向量机分类算法,它可避免复杂的运算又能够提高计算速度。近年来已有学者研究了支持向量机的几何算法。孔锐提出了基于几何思想的快速支持向量机算法,Mavroforakis M. E. 提出基于简约凸包下的支持向量机分类几何算法,彭新俊等提出基于压缩凸包的 SVM 几何算法及其应用。在此基础上提出压缩超球体的几何方法,该方法具有既保持原几何体的形状,又能克服简约凸包的不足,同时,相对于压缩凸包更加简便易懂的特点,具有较高推广价值。

1 支持向量机

1.1 支持向量机算法

支持向量机较好地实现了结构风险最小化思想,集优化、核函数、推广能力于一身而备受研究者瞩目。支持向量机方法的优点主要体现在以下几个方面:实现了结构风险最小化;算法通过转化为二次规划或线性规划问题求解,可以实现全局最优化;通过非线性变换可将实际问题转换到高维空间,在高维空间中用线性函数实现原空间的非线性问题的解决;通过空间的变换,能有效解决维数问题,并证明转换后的算法其复杂度与维数无关;具有直观的几何意义,可根据几何特征选择较好的学习方法。

1.2 最优分类超平面

给定训练样本集(x_i,y_i),其中:$i=1,2,\cdots,N;x_i\in R^q;y_i\in\{-1,1\}$。正负类分别记为 I^+ 和 I^-。选择映射 Φ 以及惩罚参数 C,构造最优化问题:

$$\begin{aligned}&\min\left[\frac{1}{2}\|w\|^2+c\sum_{i=1}^{N}\xi_i\right]\\&\text{s.t. } y_i[[w,\Phi(x_i)]+b]\geqslant 1-\xi_i\\&(i=1,2,\cdots,N)\xi_i\geqslant 0,c>0\end{aligned}\tag{1}$$

其中,c 为惩罚因子,ξ_i 为松弛变量。

收稿日期:2010 - 10 - 18;**修回日期**:2010 - 11 - 25.
作者简介:张瑞环(1986 -),女,山东菏泽人,硕士研究生,从事支持向量机研究.

一般情况下,引入核函数 $K(x,x')$,并且把该优化问题式(1)写成其对偶形式:

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\alpha_i\alpha_jK(x_i,x_j)\ -\ \sum_{j=1}^{n}\alpha_j\\&\text{s.t.}\ \sum_{i=1}^{n}y_i\alpha_i\ =\ 0\\&0\leqslant\alpha_i\leqslant C,i\ =\ 1,2,\cdots,n\end{aligned}\tag{2}$$

并求解得到最优解 $\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)$ 和最优分类超平面:$f(x)=\sum_{i\in I}\alpha_i^*y_iK(x_i,x)+b$. 其中,$\alpha^*$ 是非负拉格朗日乘子。$K(x,x')=\varphi(x)^{\mathrm{T}}\varphi(x')$ 为核函数,满足 Mercer 条件,常选核函数为高斯核函数。

2 几何算法

Bennett 等人通过对偶问题的形式讨论了支持向量机的几何解释,结果表明支持向量机等价于求解最大间隔的分类决策超平面。由此可知,关于分类问题,其目标是最优超平面的求解问题,而支持向量机只是一种求解途径。对于两类可分情况,可直接通过几何原理简易地求出其最大间隔分类超平面。Keerthi 等提出了求解可分情形下的支持向量机的几何算法,同时对不可分情形,详细讨论了如何引入变换方法将问题转变成可分情形。下面引入简约凸包、压缩凸包和压缩超球体的概念。

2.1 简约凸包[8]

设集合 I 由 k 个不同点组成,对于 $0<\mu<1$,I 的简约凸包 $\mathrm{KCH}(I,\mu)$ 定义为:$\mathrm{KCH}(I,\mu)=\{x:x=\sum_{i=1}^{k}a_ix_i,x_i\in I,\sum_{i=1}^{k}a_i=1,0\leqslant a_i\leqslant\mu\}$。图 1 给出了不同的 μ 值对应于不同的简约凸包的形态。当 $\mu=1$ 时,就是常规的凸包。当 $\mu<1$ 时,凸包随着 μ 的大小变化有着不同的压缩程度。针对不同的分类问题,可通过适当调整 μ 的大小,将两个部分重叠的凸包转化成可分情形。SVM 的直观解释:可分情形 SVM 等价于求两类训练样本构成的凸包间的最近点对,最大间隔超平面为这两个点之间的中垂平面。而不可分情形的 SVM 等价于求两个简约凸包之间的最近点对,并求两点之间的中垂平面。但简约凸包存在着改变了原有空间的几何形状的缺陷,为此文献[9]提出压缩凸包概念。

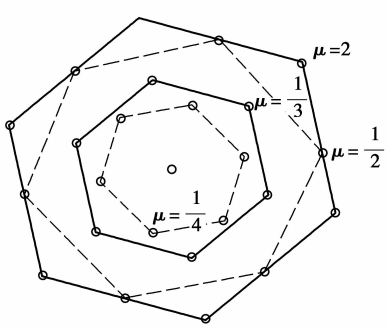

图 1 简约凸包

2.2 压缩凸包[9]

设集合 I 由 k 个不同点组成,对实数 $0<\lambda<1$,I 的压缩凸包 $\mathrm{CCH}(I,\lambda)$ 定义为:$\mathrm{CCH}(I,\lambda)=\{x:x=\sum_{i=1}^{k}a_i\hat{x}_i,\hat{x}_i=(1-\lambda)x^c+\lambda x_i,x_i\in R^q,0\leqslant a_i\leqslant 1,\sum_{i=1}^{k}a_i=1,x^c=\frac{1}{k}\sum_{i=1}^{k}x_i\}$。由此可知压缩凸包中的每一点都是通过原始数据集合 X 中的相应点朝着其重心的方向压缩得到,没有改变原始特征集合的几何形状,从而能保持训练数据的许多特征。并且 λ 的值越小,其压缩程度就越大(图 2)。

文献[9]还讨论了 CCH 的许多优良性质,如压缩后的重心位置、极点数保持不变。这为研究 SVM 的几何算法提供的理论支撑。

2.3 压缩超球体

设集合 I 由 k 个不同点组成,对实数 $0<\alpha<1$,I 的压缩超球体 $\mathrm{CCB}(I,\alpha)$ 定义为:

$$\mathrm{CCB}(I,\alpha)\ =\ \{x:\|x-x^c\|\ \leqslant\alpha R,x,x_i\in R^q,R\ =\ \max_{1\leqslant i\leqslant k}\|x_i-x^c\|\ ,x^c\ =\ \frac{1}{k}\sum_{i=1}^{k}x_i\}$$

其中 $\|x-x'\|$ 表示两空间点的欧氏距离,x^c 表示超球体的重心,R 表示球半径。

图 3 显示压缩超球体仍然没有改变原始数据的几何形状,与压缩凸包所不同的是,经过超球体的压缩变换后的任意点是由原始数据的二次函数形式表达,而不是线性的形式,并且重心也不改变。该几何算法的

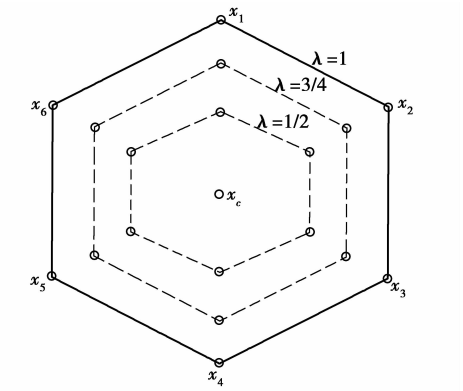

图 2　压缩凸包

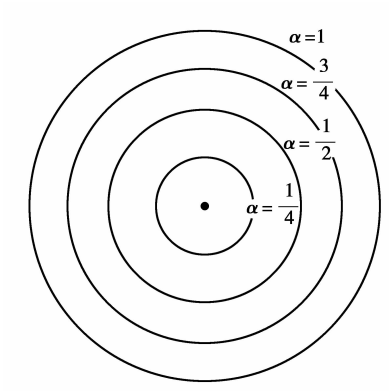

图 3　压缩超球体

关键是适当地选取压缩系数 α,使得两个超球体相切或相离,从而将不可分的分类问题转化成为可分情形。该算法克服了不易确定特征空间中样本集凸包的极点问题。假设两类别的数据集合分别记为 I^+ 和 I^-,$I^+ = \{x_i \mid y_i = 1, i = 1,2,\cdots,n_+\}$,$I^- = \{x_j \mid y_j = -1, j = 1,2,\cdots,n_-\}$,确定两类别的重心为:

$$x^{c+} = \frac{1}{n^+}\sum_{i=1}^{n^+} x_i, x^{c-} = \frac{1}{n^-}\sum_{j=1}^{n^-} x_j \tag{3}$$

计算 I^+ 和 I^- 中的任意点到各自重心之间的最远距离:

$$R^- = \max_{1\leqslant j\leqslant n^-} \| x_j - x^{c-} \|, R^+ = \max_{1\leqslant i\leqslant n^+} \| x_i - x^{c+} \| \tag{4}$$

记两集合的重心 x^{c+} 与 x^{c-} 之间的距离为 d,即:

$$d = \| x^{c+} - x^{c-} \| \tag{5}$$

选取压缩系数 $\alpha = \dfrac{d}{R^+ + R^-}$ 和压缩后的两个超球体的半径分别为 αR^+ 和 αR^-。显而易见可证 $\alpha R^+ + \alpha R^- = \alpha(R^+ + R^-) = \dfrac{d}{R^+ + R^-}.(R^+ + R^-) = d$ 该表达式为两球体相切的必要条件。

如果 $\alpha = \dfrac{d}{R^+ + R^-} > 1$,两超球体一定可分;反之,$\alpha < 1$ 时,两超球体一定不可分,需要进行超球体压缩。确定两超球体的切点 x^0,构造超平面:$\omega^* x + b^* = 0$,其中 $\omega^* = (x^{c+} - x^{c-})'$,$b^* = -(x^{c+} - x^{c-})' x^0$。

即该超平面与两球心的连线垂直,则该超平面即为所求的最优分类超平面。

3　算法设计及步骤

算法步骤:

1) 由式(3)和式(4)分别计算正负两类 I^+ 和 I^- 的类中心 x^{c+} 和 x^{c-} 以及类中心点 x^{c+},x^{c-} 之间的距离 d。

2) 分别计算集合 I^+ 和 I^- 中每个点到其中心点 x^{c+},x^{c-} 的距离,并取最大距离为 R^+ 和 R^-。

3) 如果 $R^- + R^+ \leqslant d$,则可判断两类别 I^+ 和 I^- 一定线性可分,即取 $\alpha = 1$,不使用压缩。否则,令 $\alpha = \dfrac{d}{R^- + R^+}$,分别求得集合 I^+ 和 I^- 的如下形式的压缩超球体:

$$\mathrm{CCB}(I^+,\alpha) = \{X: \| X - X^{c+} \| \leqslant \alpha R^+, R^+ = \max_{1\leqslant i\leqslant n^+} \| X_i - X^{c+} \|, X^{c+} = \frac{1}{n^+}\sum_{i=1}^{n^+} X_i\}$$

$$\mathrm{CCB}(I^-,\alpha) = \{X: \| X - X^{c-} \| \leqslant \alpha R^-, R^- = \max_{1\leqslant i\leqslant n^-} \| X_i - X^{c-} \|, X^{c-} = \frac{1}{n^-}\sum_{i=1}^{n^-} X_i\}$$

显然,经过压缩超球体变换后的数据是线性可分的。

4)构造最优分类超平面。首先,确定两超球体的相切点 $x^0 = \frac{R^-}{R^- + R^+}x^{c+} + \left(1 - \frac{R^-}{R^- + R^+}\right)x^{c-}$,其次,根据所求得的定点 x^0 构造最优分类超平面:$\omega^* x + b^* = 0$,其中 $\omega^* = (x^{c+} - x^{c-})'$,$b^* = -(x^{c+} - x^{c-})'x^0$。由最优分类超平面得到分类的决策函数 $f(x) = \mathrm{sgn}(\omega^* x + b^*)$。

4 实例验证

为了检验提出的几何算法的实际效果,从 UCI 机器学习数据库中下载关于鸢尾花的数据和伽马望远镜数据集进行分析检验。其中鸢尾花的数据特征集合由萼片长度、萼片宽度、花瓣长度和花瓣宽度 4 个特征指标组成,伽马望远镜数据集由长度、宽度等 10 个特征指标组成,样本容量分别为 200 和 19 020,训练样本数分别为 80 和 12 000,测试样本数分别 80 和 7 020。首先对其数据进行初步分析。以其中两个特征指标和类别属性 3 个因素分别构造三维散点图,如图 4,5 所示。

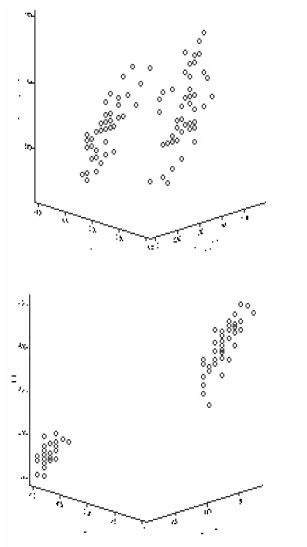

图 4 鸢尾花的散点图

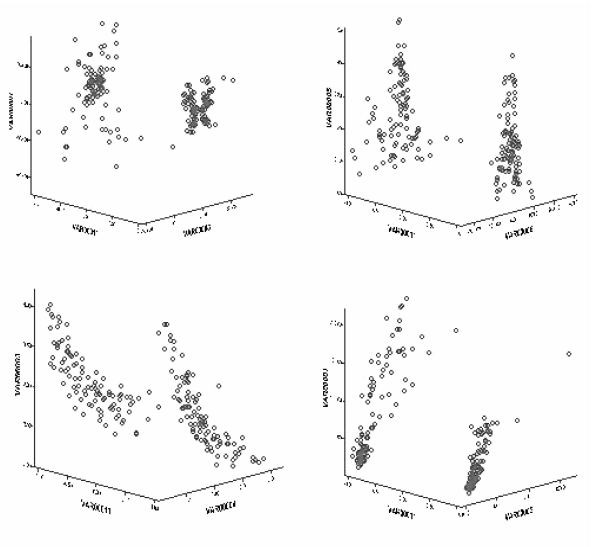

图 5 伽马望远镜的散点图

由散点图直观判断数据集的分类特征。图 4,5 显示这两个数据集属于基本可分情况。只需对原始数据集进行适当压缩,可得到完全可分的状况。具体计算结果如表 1,2 所示。由此计算得决策函数为:$f(x) = \mathrm{sgn}((1.043\,333, -0.66, 2.86, 1.106\,667)'x - 12.436\,02)$,由此计算得决策函数为:$f(x) = \mathrm{sgn}((-23.323\,7, -10.081, -0.119\,84, 0.009\,788, 0.000\,917, 21.069\,7, 20.362\,94, 0.075\,332, -24.631\,2 - 10.071\,2)'x + 4\,395.34)$对测试数据的验证结果如表 3:

表 1 鸢尾花的基本数据结果

鸢尾花	相关数据
正类集的重心	(6.07, 2.79, 4.333 333, 1.353 333)
负类集的重心	(0.266 7, 3.45, 1.473 33, 0.246 67)

表 2 伽马望远镜的基本数据结果

伽马望远镜	相关数据
正类集的重心	(43.48, 18.55, 2.78, 0.38, 0.22, 2.62, 17.29, 0.14, 18.77, 190.46)
负类集的重心	(70.80, 28.64, 2.90, 0.38, 0. 0.21, −18.45, −3.07, 0.06, 43.40, 200.53)

表 3 测试数据的结果验证

	鸢尾花	伽马望远镜
测试数据量	80	7 020
分类正确率	99.8%	68.42%

通过两个实例的验证,可以看出,不管是小样本还是大样本,提出的压缩超球体的方法都能达到相对理想的效果。但是每种算法都会有其不足之处,本文的算法虽然行之有效,但是分类精度稍有不足。对于精度要求较高的问题,此方法的研究还有待改进。

参考文献:

[1] VAPNIK V. The Natural of Statistical Learning Theory [M]. New York:Springer, 1995

[2] VAPNIK V. Statistical Learning Theory [M]. New York:Wiley, 1998

[3] RIPLEY B. Pattern Recognitionand Neural Networks [M]. Cambridge:Cambridge University Press, 1996

[4] 周志明, 陈敏. 支持向量机的人脸识别方法[J]. 咸宁学院学报, 2003(3):35-38

[5] 刘祥楼, 张淼, 刘得军,等. 基于支持向量机的文本分类方法[J]. 大庆石油学院学报, 2008(2):48-50

[6] 王自强, 段爱玲, 张德贤. 基于支持向量数据描述的高效异常数据检测算法[J]. 吉林大学学报:工学版 , 2009(2):67-68

[7] 刘青, 杨小涛. 基于支持向量机的微阵列基因表达数据分析方法[J]. 小型微型计算机系统 , 2005(3):25-30

[8] MAVROFORAKIS M E, THEODORIDIS S. Ageometric approach to support vector machine(SVM) classification[J]. IEEE Trans Neural Netw,2007,17(3) :671-682

[9] 彭新俊,王翼飞. 基于 CCH 的 SVM 几何算法及其应用[J]. 应用数学和力学, 2009(1):78-80

[10] KEERTHI S, SHEVADE S, BHATTACHARYYA C,etal. A fast iterative nearest point algorithm for support vector machine classifier design[J]. IEEETransNeuralNetw,2000,11(1) :124-136

[11] BENNETT K, BREDENSTEINER E. Geometry in learning[A]. In:Geometry at Work[C]. Washington, DC:Mathematical Association of America,1998,132-145

[12] 张兢,侯旭东,吕和胜. 基于朴素贝叶斯和支持向量机的短信智能分析系统设计[J]. 重庆理工大学学报:自然科学版, 2010,24(1):77-80

Support Vector Machine Classification Method Based on Geometry Algorithm

ZHANG Rui-huan

(School of Mathematics and Statistics, Chongqing University, Chongqing 400044, China)

Abstract: Support vector machine is a kind of learning method based on statistical learning. A kind of geometry algorithm based on SVM classification method of compressed hyper sphere is pointed out and this algorithm has the characteristics of directviewing, simplicity and easy implementation. Tested by examples, this algorithm is feasible and effective and has certain popularization ability.

Key words: support vector machine; geometry algorithm; reduced convex hull; compressed hyper sphere

责任编辑:李翠薇