支撑矢量预选取的双色 Voronoi 图方法 [1]

裴继红 杨 烜*

(深圳大学现代教育技术与信息中心 深圳 518060)
*(深圳大学信息工程学院 深圳 518060)

摘 要 支撑矢量机是在统计学习理论的基础上发展出来的一种新的模式识别方法，在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势．在支撑矢量机中，支撑矢量的选取相当困难，成为其应用的瓶颈问题．该文利用 Voronoi 图在特征空间特有的构造特性，提出了一种预先选取支撑矢量的新方法——双色 Voronoi 图方法．该方法针对数据在空间的分布特性，在训练支撑矢量机以前，利用样本数据的双色 Voronoi 图确定候选的支撑矢量，然后在这些预选的矢量上进行学习．试验证明了该方法的有效性及可行性．

关键词 支撑矢量机， Voronoi 图，双色 Voronoi 图，边界矢量，支撑矢量
中图号 TP391.4

1 引 言

从 20 世纪 60 年代以来， Vapnik 等学者一直致力于有限样本情况下的统计学习理论的研究．到 20 世纪 90 年代中期，针对传统模式识别中存在的问题，提出了一类新的模式识别方法—支撑矢量机 (Support Vector Machine, SVM)[1,2] ． SVM 是一种小样本学习方法，具有很强的推广能力，在解决模式识别的问题中表现出许多特有的优势 [1,3] ，是目前机器学习领域新的研究热点。

但是目前在 SVM 的算法中，基本上都是在优化计算后才能得到支撑矢量 (Support Vector, SV) ．即在优化过程中不仅包含了对 SV 的优化，也包含了对非 SV 的优化，这无疑大大增加了不必要的计算量 (SVM 训练过程中庞大的计算量已成为其应用的瓶颈问题) ．如果能预先选取包含了所有 SV 的侯选矢量集合，且优化计算只针对这些侯选的矢量进行，则将大大减少计算量，提高 SVM 的训练速度．为此，文献 [4] 给出了一种 SV 预选取的中心距离比值法 (CDRM) ，CDRM 方法能较好地选取样本集合的边界矢量作为预选取的 SV ．但是 CDRM 方法在选取边界样本时需要事先判定样本集合是线性可分还是非线性可分，在此基础上利用不同的核函数映射计算内积．另外 CDRM 方法在计算边界样本时还需要人为确定阈值．

数据集合的 Voronoi 图可以很好地表征数据点在空间分布的局部几何特性，它在求解点集或其他几何对象与距离有关的问题时起重要作用，是计算几何中的一个重要概念 [5] ，越来越受到模式识别等研究领域的重视 [5−9] ．从物理意义上来说， SV 就是在样本空间两类的交遇区中，那些靠得最近但又属于不同类的样本 [4] ．

本文提出了一种双色 Voronoi 图方法 (Bi-Color Voronoi diagrams Method, 简称 BCVM) 进行 SV 的预选取．基本思路是：对给定的数据集合构造其 Voronoi 图，在构造 Voronoi 图的过程中对每一个数据点的 Voronoi 多边形的边界进行染色，则在 Voronoi 图构造完成以后，那些边界带有特定颜色的 Voronoi 多边形所在的数据点就是样本集不同类之间的边缘样本， SV 就包含在这些样本矢量中． BCVM 不需要人为干预 (不需要人为输入参数) ，也不需要事先判定样本集合是否线性可分 (事实上判定样本集合是否线性可分的问题往往也比较复杂) ．文中给出的试验证明了这种方法的有效性和可行性．

[1] 2002-06-24 收到， 2002-11-29 改回
国家自然科学基金 (No.60173067) 项目资助

2 支撑矢量机

SVM 的基本思路是在特征空间 (或经过映射的特征空间) 中寻找一个最优的分界面 (超平面)，使推广能力最好 [1,10]。分界面的质量主要取决于它在两类交遇区所处的位置，最优分界面是一个能够使间隔 (margin) 最大的超平面，此处的间隔是指分类超平面到最近的样本的距离。SVM 的分类超平面 (决策函数) 实际上是由其 SV 决定的。SVM 的训练过程是一个二次规划的寻优过程，也就是确定 SV 的过程。已有的 SVM 算法中，这个过程需要在所有的样本数据上进行二次规划后才能得到，这往往需要庞大的计算量为代价。若可以在训练数据集合中找出一个较小的 SV 的候选数据集合，在该集合中做上述优化确定 SV，将可以大大地提高 SVM 的学习速度。

3 支撑矢量预选取的 BCVM 方法

平面上 n 个点的点集 $S=\{p_1,p_2,\cdots,p_n\}$ 的 Voronoi 图是平面域的一个划分，该划分产生的每个子域 $V(p_i)$ 是具有下述性质的点的轨迹：子域内的点 $q(q\neq p_i)$ 与 p_i 的距离小于 S 中其他点与 q 的距离，即 $d(q,p_i)<d(q,p_j)$, $p_j\in S$, $j=1,\cdots,n$, $j\neq i$。一般记 Voronoi 图为 $\mathrm{Vor}(S)$。$\mathrm{Vor}(S)$ 划分平面成 n 个多边形域 $V(p_i)$，每个多边形域 $V(p_i)$ 包含 S 中的一个点，且只包含 S 中的一个点。$\mathrm{Vor}(S)$ 的边是 S 中某点对的垂直平分线上的一条线段或者半直线，并为该点对所在的多边形域所共有。$\mathrm{Vor}(S)$ 域中有的多边形域是无界的。由 $\mathrm{Vor}(S)$ 的定义可以看出，$\mathrm{Vor}(S)$ 给出了特征空间的一种划分，这种划分是基于所给数据集中的每个点在特征空间的分布特性。每个划分子域 $V(p_i)$ 的边界是由与其距离最邻近的一组点决定的。这种划分体现了样本数据在空间的局部邻接分布特性。

3.1 双色 Voronoi 图

对于有 n 个点的数据集 $S=\{p_1,p_2,\cdots,p_n\}$，做如下定义：

定义 1　数据集 S 的 Voronoi 边界色函数是映射 $\phi: S\times S\to B$，且满足以下条件：

(1) $B=\{\mathrm{red},\mathrm{green}\}$ 是数据集 S 的 Voronoi 边颜色指标集合。red 和 green 可以用两个不同的数值表示。例如 red $=1$, green $=-1$。

(2) $\forall p_i\in S$, $p_j\in S$, 则

$$\phi(p_i,p_j)=\begin{cases}\mathrm{red}, & p_i\text{与}p_j\text{属于不同类样本}\\ \mathrm{green}, & p_i\text{与}p_j\text{属于同一类样本}\end{cases}\tag{1}$$

上述定义中，数据集 S 中的点要求，若 $i\neq j$，则 $p_i\neq p_j$。在边界色函数 ϕ 中，一般不考虑 $i=j$，即 $\phi(p_i,p_i)$ 的情况。

定义 2　双色 Voronoi 图 ($\mathrm{BCVor}(S)$) 在定义了边界色函数 ϕ 的数据集 S 上，S 的 $\mathrm{BCVor}(S)$ 是经过边界染色的 $\mathrm{Vor}(S)$。染色过程为：假设数据集 S 的 Voronoi 图为 $\mathrm{Vor}(S)$，在 $\mathrm{Vor}(S)$ 中对应于点 $p_i\in S$ 的 Voronoi 多边形为 $V(p_i)$, $E(p_i,p_j)$ 表示 $V(p_i)$ 的一条边，其中 p_j 是 S 中与 p_i 最近邻近的一个点，则边 $E(p_i,p_j)$ 的颜色为 $\phi(p_i,p_j)$。

由上述染色过程可以看出，$\mathrm{BCVor}(S)$ 中的多边形由 red 和 green 两种颜色的边构成。若 $E(p_i,p_j)$ 的颜色为 red，则点对 p_i 与 p_j 属于不同的类，若 $E(p_i,p_j)$ 的颜色为 green，则 p_i 与 p_j 属于相同的类。

由于 Vor(S) 在求解点集或其他几何对象与距离有关的问题时起重要作用，因此许多学者对其性质进行了深入的研究，提出了不少构造 Vor(S) 的理论方法，例如半平面交方法、增量构造方法 [5] 、减量方法 [5] 等等. 对于含有 n 个数据点的集合 S， Vor(S) 构造的时间复杂度为 $O(n\log_2 n)$[5,11] . 而由数据集 S 的 BCVor(S) 构造过程可以看出，它比 Vor(S) 构造的计算步骤仅仅增加了在计算出最近邻近点对的边后，用边界色函数 ϕ 对其染色这一步，这一步时间的增加是可以忽略的. 因此 BCVor(S) 构造的时间复杂度也为 $O(n\log_2 n)$. 在实际 BCVor(S) 的算法中，并不需要先构造出 BCVor(S) ，再对其染色；而可以直接在计算出每一个点的 Voronoi 子域的同时就对其边界进行染色.

3.2 BCVM 方法

由 BCVor 的定义可以看出，那些带有 red 颜色的 Voronoi 多边形所包含的数据点是训练样本集不同类之间的边缘样本，而 SV 是在样本空间两类交遇区中，那些靠得最近但又属于不同类的样本，因此 SV 一定包含在这些样本点中. 本文 SV 预选取的基本思想就是借助所定义的双色 Voronoi 图找出这些数据点. 我们将这种 SV 预选取的方法称为双色 Voronoi 图方法，简称为 BCVM . BCVM 的具体步骤如下：

(1) 对给定 n 个点的训练数据集合 $S=\{x_1,x_2,\cdots,x_n\}$ ，根据集合 S 的类指示函数，定义集合 S 上的 Voronoi 边界色函数 $\phi(x_i,x_j)$.

(2) 利用边界色函数 $\phi(x_i,x_j)$ 在构造数据集合 S 的 Voronoi 图的同时，对每一个 Voronoi 子域的边界进行染色. 得到数据集 S 的双色 Voronoi 图 BCVor(S) .

(3) 在 BCVor(S) 中确定边界带有 red 颜色的 Voronoi 子域所包含的数据点的集合 S_r ， S_r 实际上就是数据集 S 的边缘样本集合，而支撑矢量就包含在边界点集 S_r 中. 因此 S_r 就是 SV 的预选集合.

实际计算中，上述第 (2), 第 (3) 步可以同时一次完成. 即在构造 BCVor(S) 的同时，一旦确定某一个 Voronoi 子域 $V(p_i)$ 的边界为 red 颜色，则将点 p_i 添加到集合 S_r 中，这样在构造完成 BCVor(S) 的同时，也得到了包含 SV 的边缘数据集合 S_r .

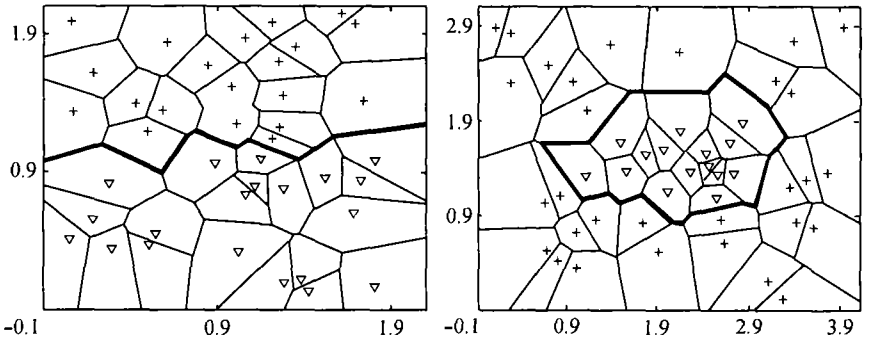
BCVor(S) 的概念给出了一种构造数据集分类边界的方法，这种边界是由不同类的边缘样本点决定的，边界过边缘样本点连线的垂直平分点，距离边缘样本点具有最大的分类间隔 (margin) . 由 BCVor(S) 得到的分类边界是分段线性的，每一段线性分类面在局部来说具有最好的推广性. 因此，SV 包含在这些决定 BCVor(S) 的 red 边界的数据点之中. 另外，由于 BCVor(S) 采用分段线性方法确定了每一个样本数据点周围是否存在与其最近邻近的其他类的样本数据点，即使训练样本集合是非线性可分的， BCVor(S) 也可以很好地确定所有的边缘数据点. 也就是说， BCVM 事先并不要求知道训练样本集合是否线性可分的信息，也就不需要确定用于非线性映射的核函数. 因此， BCVM 在 SV 的预选中，既可用于对样本线性可分的情况，同样也可用于样本非线性可分的情况. 克服了 CDRM[4] 方法需要人为干预 (确定边界阈值) ，以及需要事先判定样本集合是否线性可分，及在非线性可分时构造非线性核函数的问题.

4 仿真试验

为了验证本文算法的可行性和有效性，在文中给出了两个样本集合的仿真实验.

样本集合 Data1 含有 40 个二维数据点，分两类：第一类数据是在顶点分别为 (0,0) (2,0) (0,1) (2,1) 的矩形区域内随机生成的 20 个均匀分布的点；第二类数据是在顶点分别为 (0,1.1) (2,1.1) (0,2.1) (2,2.1) 的矩形区域内随机生成的 20 个均匀分布的点. 样本集合 Data2 有 45 个二维数据点，也分两类：第一类数据是在顶点分别为 (1.1,1.1) (2.9,1.1) (1.1,1.9) (2.9,1.9) 的矩形区域内随机生成的 15 个均匀分布的点；第二类数据是在顶点分别为 (0,0) (4,0) (0,3) (4,3) 的

矩形的内部，以及顶点为 (0.9,0.9) (3.1,0.9) (0.9,2.1) (3.1,2.1) 的矩形的外部共同围成的环形区域内随机生成的 30 个均匀分布的点. Data1 是线性可分的数据集合， Data2 是非线性可分的集合. 图 1 是这两个数据的双色 Voronoi 图，图中 Voronoi 多边形的边界颜色 red 用粗线段表示，边界颜色 green 用细线段表示， BCVor(S) 中带有 red 颜色边的多边形所包含的数据点即为边缘数据点. 经过 BCVM 方法进行 SV 预选取后， Data1 的候选边缘数据集合 SData1 有 12 个点， Data2 的候选边缘数据集合 SData2 有 20 个点.



(a) 线性可分样本集合 Data1　　(b) 非线性可分样本集合 Data2

图 1　两个仿真实验

我们用数据集 Data1 和 SData1 分别对线性可分 SVM 进行单独训练. 对于非线性可分数据集合选用二次多项式 (2) 式作为核函数:

$$K(x, x_i) = (1 + x^T x_i)^2 \tag{2}$$

用 Data2 和 SData2 分别对非线性可分 SVM 进行单独训练.

实验结果: 数据集 Data1 和 SData1 训练后得到了相同的 SV，数据集 Data2 和 SData2 训练后也得到了相同的 SV. 说明了用 BCVM 预选取的边缘数据集合包含了原数据集合的 SV，BCVM 方法是可行的和有效的.

5 结 论

本文利用数据集的 Voronoi 图在样本空间的划分特性，提出了一种支撑矢量 (SV) 预选取的方法——双色 Voronoi 图方法 (BCVM). BCVM 根据样本数据点的局部邻域特性，在构造用于训练 SVM 的样本数据集的 Voronoi 图时，对每一个数据点的 Voronoi 多边形的边界进行染色，得到数据集的双色 Voronoi 图 (BCVor(S)). 在 BCVor(S) 中具有 red 边的 Voronoi 多边形所在的数据点是样本集合的边缘数据点，边缘数据集合中完全包含了原数据集合的 SV. 边缘数据点的数量一般来说远远少于整个数据集合中样本点的数量，用这些边缘数据点对 SVM 进行训练可以大大加快 SVM 的学习速度. BCVM 方法具有很好的几何直观性，不需要事先判定所给的数据集合是线性可分还是非线性可分，因此 BCVM 的适应性比较广泛.

参 考 文 献

[1] V. N. Vapnik 著，张学工译，统计学习理论的本质 [M]. 北京，清华大学出版社， 2000 年， 11–12.

[2] V. N. Vapnik, An overview of statistical learning theory, IEEE Trans. on Neural networks, 1999, 10(5), 988–999.

[3] 边肇祺，张学工，模式识别 [M]，北京，清华大学出版社，2000 年，161–176, 284–305 .

[4] 焦李成，张 莉，周伟达，支撑矢量预选取的中心距离比值法，电子学报，2001, 29(3), 383–386.

[5] 周培德，计算几何—算法分析与设计 [M]，北京，清华大学出版社，2000 年，88–130, 236–271.

[6] F. Aurenhammer, Voronoi diagrams—a survey of a fundamental geometric data structure, ACM Comput. Survey, 1991, 23(3), 345–405.

[7] G. W. Rogers, J. Solka, D. S. Malyevac, C. E. Priebe, A self-organizing network for computing *a posteriori* conditional class probability, IEEE Trans. on Systems, Man and Cybernetics, 1993, 23(6), 1672–1682.

[8] N. K. Bose, A. K. Garga, Neural network design using Voronoi diagrams, IEEE Trans. on Neural Networks, 1993, 4(5), 778–787.

[9] C. Gentile, M. Sznaier, An improved Voronoi-diagram-based neural net for pattern classification, IEEE Trans. on Neural Networks, 2001, 12(5), 1227–1234.

[10] 阎平凡，张长水，人工神经网络与模拟进化计算 [M]，北京，清华大学出版社，2000 年，60–95.

[11] D. Chen, Efficient geometric algorithm on the EREW PRAM, IEEE Trans. on Parallel Distrib. Syst., 1995, 6(1), 41–47.

PRE-EXTRACTING SUPPORT VECTOR FOR SUPPORT VECTOR MACHINE USING BI-COLOR VORONOI DIAGRAMS

Pei Jihong Yang Xuan*

(*Modern Educational Technology & Info. Center, Shenzhen Univ., Shenzhen* 518060, *China*)

*(*School of Information and Eng., Shenzhen University, Shenzhen* 518060, *China*)

Abstract Support Vector Machines (SVMs) are a new generation learning system based on recent advances in statistical learning theory. SVMs have many well features that make them attractive for small samples, nonlinear and high dimensional pattern recognition. However, choice of Support Vectors(SVs) is difficult in SVMs, which is a bottleneck problem. In this paper, a novel method using bi-color Voronoi diagram is proposed to pre-extract SVs based on Voronoi diagram. Considering the distribution feature of samples space, this method determinates SVs based on the bi-color Voronoi diagram before training SVMs. Learning is based on these pre-extracted vectors. Experiments show that this method is feasible and effective.

Key words Support vector machine, Voronoi diagrams, Bi-color Voronoi diagrams, Margin vector, Support vector

裴继红： 男，1966 年生，副教授，博士，主要研究兴趣：模式识别、图像分析与理解、智能信息处理、模糊集理论、智能人机交互.

杨 烜： 女，1969 年生，副教授，博士后，长期从事图像处理、计算机视觉、图像融合等方向的研究.