具有渐进局部学习特性的多色 Voronoi 分类器设计 [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: 本文提出了一种多色 Voronoi 分类器 MCVC. MCVC 在学习样本上有好的边界推广性，随着样本数量的增加 MCVC 的分类面可以逼近任意的分类函数. MCVC 具有好的局部特性，对新加样本的训练只影响其周围的局部性态，不会对全局产生大的影响，可以克服神经网络方法对样本的过学习问题. 实验表明 MCVC 对于线性和非线性分类问题都具有最优分类面.

关键词: Voronoi 图，多色，分类器设计
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 1009-5896(2004)10-1613-07

Design of Multicolor Voronoi Classifier with Gradually Local Learning Ablility

Pei Ji-hong Yang Xuan*

(*Modern Education Tech. and Info. Center, Shenzhen Univ., Shenzhen* 518060, *China*)
*(*School of Info. and Eng., Shenzhen Univ., Shenzhen* 518060, *China*)

Abstract A novel MultiColor Voronoi Classifier (MCVC) is proposed, which can be applied to linear and nonlinear classification problems. MCVC has sound ability to expend classification plane between samples. With increment of samples, it can be shown that the classification plane of MCVC can close to any classification function. MCVC has very good local ability too. When new learning sample is added, only local classification planed is modified and the whole classification characteristics are not modified greatly. So MCVC can solve the overfitting problem of neural network. Experiments show that MCVC is feasible to linear classification and nonlinear classification problems.

Key words Voronoi diagrams, Multi-color, Classifier design

1 引言

在模式识别中，模式分类器的设计是需要解决的基本问题之一. 在分类器的设计中，基于贝叶斯决策理论的分类器是一种经验误差最小化原则下的分类器 [1]，在样本数量趋于无穷大时，达到期望最优，但在小样本情况下无法保证好的推广性. 而基于统计学习理论的支撑矢量机 (SVM) 分类器在理论上可以解决小样本的学习问题 [2−4]，SVM 的分类面距不同类的支撑矢量有最大的分类间隔 (margin). 但 SVM 方法并没有给出如何设计最优分类器的具体方法，另外支撑矢量计算问题往往是一个耗时的过程.

本文提出了一种多色 Voronoi 分类器 (MCVC) 的设计方法. MCVC 是特征空间的一种划分，这种划分是基于学习样本的空间分布得到的. MCVC 的边界是由不同类的边缘样本点决定的，分界面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. MCVC 的分类面随着数据点 (样本点) 的逐渐增加，逐渐变得越来越精细化，从而使分类结果也逐渐精

[1] 2003-06-10 收到，2003-11-20 改回
国家自然科学基金 (No.60173067) 资助课题

细化. MCVC 的渐进构造特性和局部构造特性正好模拟了人的学习过程，特别是对特例的学习过程. MCVC 可以解决小样本的学习问题，其学习速度和分类速度都很快，可以解决支撑矢量计算耗时的问题. 另外，MCVC 的构造过程与学习样本的分布特点无关，可以适用于具有不同分布特点的样本学习问题. 由于 Voronoi 图在计算几何中有很好的研究基础可以借鉴 [5,6]，所以这种分类器具有很好的理论与实际应用价值.

2 多色 Voronoi 图的构造及性质

Voronoi 图是一种重要的几何结构，在求解点集或其他几何对象与距离有关的问题时具有重要作用 [5]. 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 Vor($S$), Vor($S$) 划分平面成 n 个多边形域 $V(p_i)$，每个多边形域 $V(p_i)$ 都表示距离 p_i 比距离其他点更近的点的分布范围. 在 Voronoi 图的基础上，我们可以构造多色 Voronoi 图如下：

对于 S 中的每个学习样本，若两个学习样本属于同一类，则其多边形域的颜色相同；属于不同类的学习样本的多边形域颜色各不相同. 假设两个学习样本的多边形域共用某个边界，若这两个样本属于同一类，则该边界为绿色 (green)；否则为红色 (red). 经过染色的 Voronoi 图称为多色 Voronoi 图，记为 CVor(S).

由上述构造过程可以看出，如果学习样本共有 m 类，则多色 Voronoi 图就有 m 种颜色，而 CVor(S) 中的多边形由红和绿两种颜色的边构成. CVor(S) 具备 Voronoi 图的所有几何性质. 下面四个性质对于分类问题比较重要.

性质 1 在 CVor(S) 中，边界颜色都是 green 的 Voronoi 多边形所在的点的集合为 S_g，则集合 $S_r=S-S_g$ 的多色 Voronoi 图 CVor(S_r) 与原 CVor(S) 具有相同的 red 边界，同时具有形状相同的色块区域.

性质 2 给定新的点 $p_{\text{new}}\notin S$，假设 p_{new} 落在 CVor(S) 的 Voronoi 多边形 $V(p_i)$ 中. 若 $V(p_i)$ 的边界颜色都是 green，且 p_{new} 与 p_i 属于同一类，则集合 $S_{e1}=S+\{p_{\text{new}}\}$ 的多色 Voronoi 图 CVor(S_{e1}) 与 CVor(S) 具有相同的 red 边界，同时具有形状相同的色块区域.

性质 3 给定含有 n 个数据点集合 S 的 CVor(S)，给定查询点 $q(q\notin S)$，确定 q 落入哪个 Voronoi 多边形域中的时间复杂度为 $O(\log_2 n)$. 构造 CVor(S) 的时间复杂度为 $O(n\log_2 n)$.

性质 1 和性质 2 对于我们构造分类器具有重要的意义. 因为只需要保留具有红色边界的学习样本点，则重新构造 Voronoi 图后得到的色块区域具有相同的形状.

3 多色 Voronoi 分类器 (MCVC) 的设计步骤

给定数据集合 $S=\{x_1,x_2,\cdots,x_n\}$，以及集合上的类标记集 $C=\{c_1,c_2,\cdots,c_m\}$. 数据集 S 是划分为 m 类的有监督学习样本，要求设计一个有 m 类分类能力的分类器. 设计步骤如下：

(1) 构造数据集合 S 的 Voronoi 图.

(2) 根据多色 Voronoi 图的定义生成数据集合 S 的多色 Voronoi 图 CVor(S). 保存多色 Voronoi 图 CVor(S) 的备份.

(3) 在 CVor(S) 中合并同色区域，得到数据集 S 的边界点集 S_r，构造由边界点集 S_r 决定的多色 Voronoi 图 CVor(S_r).

(4) CVor(S_r) 即为从数据样本集合 S 学习得到的模式分类器—— MCVC.

$\mathrm{CVor}(S_r)$ 的 red 边界构成了 MCVC 的分类曲面．由 MCVC 的构造过程看出，其算法效率主要取决于 $\mathrm{CVor}(S)$ 的构造过程，而 $\mathrm{CVor}(S)$ 的构造复杂度为 $O(n\log_2 n)$，这表明 MCVC 的学习效率为 $O(n\log_2 n)$．一般来说在样本数据集合 S 较大的情况下，集合 S_r 中数据点的数量远远少于数据集合 S 中数据点的数量，因此，得到的分类器一般具有较简单的空间结构．

MCVC 对未知样本数据点 p 的分类过程就是 p 点在 $\mathrm{CVor}(S_r)$ 中的最近邻近查询过程，即查询给定的样本点落在那一个色区的问题．在得到 MCVC 后实际上已获得点集 S_r 对样本空间进行划分的 Voronoi 图，在特征空间为二维和三维的情况下，已经证明可以在 $O(\log_2 n_r)$ 时间内完成查询 [7,8]，也就是说 MCVC 对未知模式的识别时间为 $O(\log_2 n_r)$，即 MCVC 的分类算法效率可以达到 $O(\log_2 n_r)$．由于 MCVC 中使用的是经过合并简化的 Voronoi 图，在数据集合较大时，MCVC 中边界点集 S_r 中的数据点数 n_r 一般来说远远少于用于学习的样本数据集合 S 的点的数量 n，因此分类速度是可以满足要求的．

4 MCVC 对新样本渐进局部学习

在实际模式分类问题中经常会遇到渐进学习的问题，即在给定一组样本完成分类器设计后，如果对分类器的分类效果不满意，则需要加入新的样本进行训练．MCVC 是一种具有很好的渐进学习特性的分类器．下面说明 MCVC 对新样本的学习过程．

假定设计原始分类器的样本集合为 S，得到的分类器为 $\mathrm{MCVC_{old}}$，新增加的样本的集合为 S_{new}．则 MCVC 对新样本的学习实际上是构造数据点集为 $S+S_{\mathrm{new}}$ 的多色 Voronoi 图 $\mathrm{CVor}(S+S_{\mathrm{new}})$，进而其边缘数据集 $(S+S_{\mathrm{new}})_r$ 的多色 Voronoi 图 $\mathrm{CVor}\ (S+S_{\mathrm{new}})_r$．而 $\mathrm{CVor}(S+S_{\mathrm{new}})_r$ 就是添加新的样本集后，经过增强训练得到的新的多色 Voronoi 分类器．可以采用 Voronoi 图构造的联机增量算法来完成上述 MCVC 的增强训练．下面以增加一个新的训练样本点 p_{new} 为例说明．具体过程为

(1) 调出备份的数据集合 S 的多色 Voronoi 图 $\mathrm{CVor}(S)$．

(2) 在 $\mathrm{CVor}(S)$ 中对 p_{new} 进行最近邻近查询，确定 p_{new} 落在 $\mathrm{CVor}(S)$ 的 Voronoi 多边形的位置．这一步的时间复杂度为 $O(\log_2 n)$．假设 p_{new} 落在 $\mathrm{Vor}(p_i)$ 内．$\mathrm{Vor}(p_i)$ 多边形是由 n_i 个超平面构成的凸超多面体，即点 p_i 周围有 n_i 个最近邻近点 $p_{n_1},p_{n_2},\cdots,p_{n_i}$．

(3) 在 $\mathrm{CVor}(S)$ 中修改 Voronoi 多边形 $\mathrm{Vor}(p_i)$．这一步的时间复杂度与 p_i 周围最近邻近点的数目 n_i 有关，为 $O(n_i)$．修改后的图即为数据集合 $S+\{p_{\mathrm{new}}\}$ 的多色 Voronoi 图 $\mathrm{CVor}(S+\{p_{\mathrm{new}}\})$．

(4) 保留 $\mathrm{CVor}(S+\{p_{\mathrm{new}}\})$ 的备份．

(5) 合并 $\mathrm{CVor}(S+\{p_{\mathrm{new}}\})$ 的同色区域，得到边界点的 $\mathrm{CVor}(S+\{p_{\mathrm{new}}\})_r$，即为经过新样本增强训练后的新的模式分类器—— $\mathrm{MCVC_{new}}$．

在上述增强学习算法中，如果由第 2 步得到的 $\mathrm{Vor}(p_i)$ 的边界均为 green，则增强学习算法只需要做完第 4 步．实际上此时 $\mathrm{MCVC_{new}}=\mathrm{CVor}(S+\{p_{\mathrm{new}}\})_r=\mathrm{CVor}(S_r)$．这一点是由性质 2 保证的．而完成第 4 步得到 $\mathrm{CVor}(S+\{p_{\mathrm{new}}\})$ 的备份则保证了在以后的增强学习中可以得到更加合理的分类面．在新增训练集合 S_{new} 的样本点数 $n_z>1$ 时，可以进行 n_z 次单点增强学习，最终得到 $\mathrm{CVor}(S+S_{\mathrm{new}})$，再由 $\mathrm{CVor}(S+S_{\mathrm{new}})$ 得到 $\mathrm{MCVC_{new}}=\mathrm{CVor}(S+S_{\mathrm{new}})_r$．

5 MCVC 的特性

由分类器 MCVC 的设计过程可以看出，MCVC 的分类面是由分段线性函数构成的超平面组成的．这些超平面是与训练集 S 的边缘样本点具有最远分类间隔的垂直平分面组成的．因此在现有样本上有好的边界推广性，特别是对小样本具有较好的学习能力．在小样本的情况下，MCVC 的性能与 SVM 的性能基本一致．随着样本数量的增加可以证明 MCVC 的分类面可以

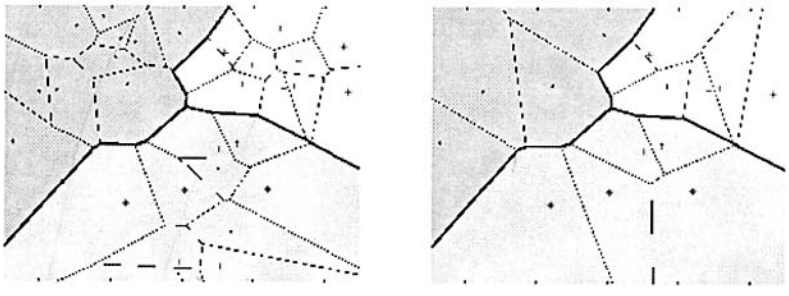
逼近任意的分类函数. MCVC 可以看成是一种与模型无关的分类器，既可以用于线性可分的情况，也适用于线性不可分样本的分类. MCVC 的学习过程与样本的分布特点无关，可以适用于具有不同分布特点的样本学习.

由 MCVC 对新样本的增强学习算法可以看出， MCVC 具有好的局部特性，对新加样本的训练只影响其周围的局部性态，不会对全局产生大的影响. 因此 MCVC 很自然地克服了神经网络方法对样本的过学习问题. 这种学习特性与人脑对新增信息的学习特性很类似.

MCVC 具有清晰的几何结构，便于对其进行分析和理解. MCVC 的这种结构对识别一些异常特定目标或特例来说具有好的分类识别特性. 例如对于网络的入侵检测问题，数量很少的可疑的入侵目标可能占有特征空间很小的区域，这时用常规的分类器设计方法不好处理. 而对于 MCVC 则原则上可以很好地完成这一任务.

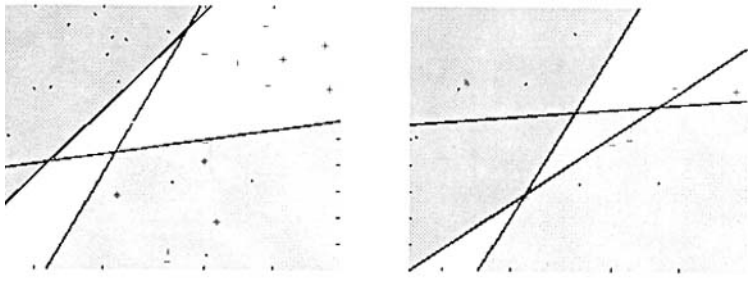
6 实验与结论

图 1(a) 是在二维特征空间随机生成的具有三类的样本集合 data3(有 30 个样本点) 的多色 Voronoi 图，其中，同类样本点的多边形域用相同的颜色标注，同类样本之间的边界用 green(细实线) 标注，不同类样本之间的边界用 red(粗实线) 标注. 图 1(b) 是构造的 MCVC 分类器. 可以看出，去掉那些边界都是 green 的样本点之后，利用剩下的 15 个边界样本点得到的多色 Voronoi 图的色块形状与原多色 Voronoi 图的色块形状是相同的，即分类面由边界控制点决定，而与类内的样本点无关. 图 2(a) 是利用感知器算法学习的得到 data3 的线性分类面，图 2(b) 是利用感知器算法学习的得到 15 个边界样本的线性分类面，可以看出，感知器的分类面与样本有密切关系， 30 个学习样本和 15 个学习样本的分类面有较大差异，而 MCVC 的分类面在保证 15 个样本点是边界点的前提下可以得到相同的分类面.



(a) data3 的多色 Voronoi 图　　(b) 训练后的 MCVC 分类器

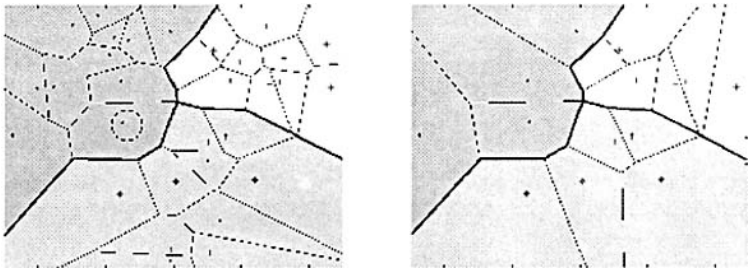
图 1



(a) 利用感知器算法得到的 data3 的线性分类面　　(b) 利用感知器算法确定的 15 个边界样点的分类面

图 2

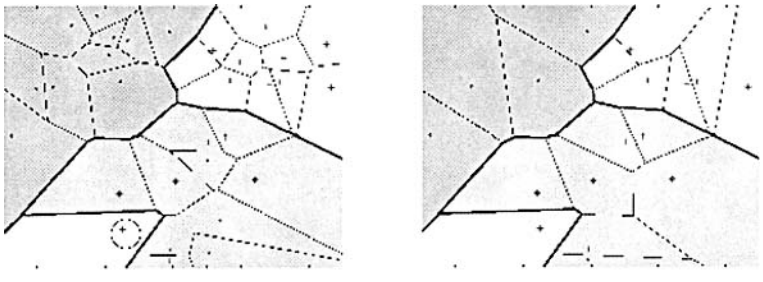
图 3(a) 是增加了一个学习样本点 (属于左上角一类，画圆圈者) 后得到的多色 Voronoi 图. 由于该样本点是一个边界控制点，因此影响分类面的形状，而对于类内的局部结构，该样本点的影响并不大. 图 3(b) 是对应的增强学习后的 MCVC 分类器，其分类面由于新的样本点的影响而发生了局部变化. MCVC 的分类面只取决于少数的控制点，只要控制点确定下来，分类面就可以确定下来；而利用感知器等线性判决函数得到的分类面依赖于多数的样本点.



(a) 增加一个学习样本之后构造的多色 Voronoi 图　　(b) 增强训练后的 MCVC 分类器

图 3

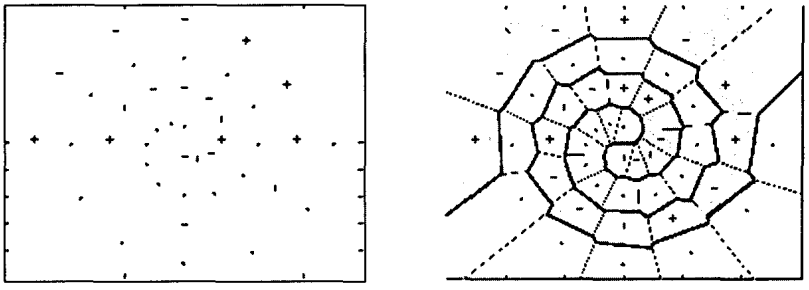
图 4(a) 是增加了一个学习样本点 (属于右上角一类，画圆圈者) 后得到的多色 Voronoi 图. 可以看到该样本点落入了另一类样本点中，其分类面形成一个“岛”的情况. 图 4(b) 是对应的增强学习后的 MCVC 分类器，其分类面实际上描述了一种对特例学习后的分类情况. 对于这种样本点的分布特点，线性分类函数无法确定分类面，只能利用非线性判决函数来确定，而这一过程往往计算非常复杂. 而 MCVC 可以很好地确定分类面，其算法的复杂度并没有因为线性可分或是线性不可分而有所区别.



(a) 增加一个学习样本之后构造的多色 Voronoi 图　　(b) 增强训练后的 MCVC 分类器

图 4

为了对 MCVC 分类器的分类能力进行检验，我们利用双螺旋分布的数据样本进行分类. 图 5 (a) 是 44 个双螺旋分布的标准学习样本，我们利用这些标准样本进行学习，构造出 MCVC 分类器，如图 5(b) . 从图 5(b) 可以看出 MCVC 的分类面可以很好地表现出双螺旋的特点，并且双螺旋的学习样本采样越密集，MCVC 的分类面越逼近双螺旋的形状. 这进一步表明，只要学习样本可以反映样本点的分布特点，并且学习样本足够多，MCVC 可以逼近任意形状的分类面.



(a) 双螺旋学习样本 (b) 双螺旋 MCVC 分类器

图 5

表 1 双螺旋分类样本正确分类率

试验次数	1	2	3	4	5	6	7	8	9	10
螺旋 1	0.97	1.0	1.0	0.91	1.0	0.81	1.0	0.88	1.0	0.97
螺旋 2	1.0	0.97	1.0	0.97	1.0	0.97	1.0	0.94	0.97	1.0

我们在标准双螺旋样本上加高斯噪声，构造双螺旋的分类样本，利用这些分类样本检验 MCVC 的分类性能. 我们每次随机产生 62 个分类样本，其中 31 个属于其中一个螺旋，另外 31 个属于另一个螺旋. 这里需要说明的是，我们实验用的两个螺旋中心的距离 $d = 5$ ，而高斯噪声方差 $\sigma = 1.4$. 表 1 是 10 次实验的分类结果，从表中可以看出在分类样本的噪声较强的情况下，其正确分类率是比较理想的.

在分类器的设计中，如何对小样本进行学习、如何对特例进行学习，以及分类器的渐进局部学习等等问题是分类器设计中的一个研究热点. 本文提出的分类器 MCVC 及其设计方法，可以较好地解决上述问题. MCVC 分类器不需要定义特定的目标函数，通过学习样本的多色 Voronoi 图来确定类间的分类面. 其分类速度只与分类边界的控制点个数有关，其值小于样本点的个数. 而对于大量学习样本的情况，分类界面的控制点个数一般远远小于学习样本的个数，因此 MCVC 训练后对未知样本的分类效率是很高的. 同时 MCVC 对新增加的训练样本具有快速局部学习能力，这种局部学习能力克服了神经网络方法常出现的过学习问题，很类似于人对概念的学习. MCVC 既适用于线性可分情况，也适用于线性不可分的情况，是一种与模式无关的最优分类器，可广泛用于各种数据类型的分类问题.

需要说明的是 Voronoi 图的理论是在二维空间的计算几何的基础上发展起来的，当数据空间扩展到高维时，每个二维的 Voronoi 多边形就扩展为高维空间中的 Voronoi 凸超多面体. 由于 MCVC 的构造与 Voronoi 图的构造方法密切相关，因此如果在高维空间中存在有效的 Voronoi 图构造方法，那么相应地就有高维 MCVC 的构造方法. 但是目前对于高维 Voronoi 图构造方法的研究还处于初步阶段 [9−11] ，在三维空间中已经找到了 Voronoi 图构造方法，并且可以证明其算法效率是 $O(n\log_2 n)$ ，因此 MCVC 可以有效地扩展到三维空间中使用. 但在四维以上的高维空间中还没有一个成熟的构造方法，其构造算法复杂度也没有明确的结论. 相应地，我们提出的 MCVC 一般不直接在高维空间中解决问题，而是将高维数据投影到三维或二维空间，然后再利用 MCVC 来解决问题.

参 考 文 献

[1] 蔡元龙. 模式识别. 西安：西安电子科技大学出版社， 1986 年，第 3 章： 67–101.

[2] 边肇祺，张学工. 模式识别. 北京：清华大学出版社， 2000 年，第 7 章： 161–173; 第 13 章： 284–303.

[3] Vapnik V N, 张学工译. 统计学习理论的本质. 北京：清华大学出版社， 2000 年，第 0 章： 1–10.

[4] Vapnik V N. An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 1999, 10(5): 988–999.

[5] 周培德. 计算几何—算法分析与设计. 北京：清华大学出版社，2000 年，第 4 章，88–132；第 10 章：236–271.

[6] Fortune S. A sweepline algorithm for Voronoi diagrams. *Algorithmica*, 1987, 2(2): 153–174.

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

[8] Shamos M I, Hoey D. Closest-point problems. Proc. 16th IEEE Ann. Symp. on the Foundations of Computer Science, CA, USA, 1975: 151–162.

[9] Amato N M, Preparata F P. An NC parallel 3D Convex hull algorithm. In: Proc. 19th Annual ACM Symp. Comput. Geom., San Diego, CA, USA, 1993: 289–297.

[10] Amato N M, Goodrich M T, Ramos E A. Parallel algorithms for higher-dimensional convex hulls. In: Proc. 35th Annual IEEE Symp. Found. Comput. Sci., Santa Fe, NM, USA, 1994: 386–694.

[11] Amato N M, Goodrich M T, Ramos E A. Computing faces in segment and simplex arrangements. In: Proc. 27th Annual ACM Symp. Theory Comput., Las Vegas, Nevada, USA, 1995: 672–682.

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

杨　烜： 女，1969 年生，博士后，副教授，研究方向为智能信息处理、模式识别与人工智能、图像融合、数据融合.