# 基于距离的样本选择方法及其在实时野点检测中的应用*)

肖健华

(五邑大学智能技术与系统研究所 广东江门529020)

## The Method of Sample Selection Based on Distance and its Application in Outlier Detection

XIAO Jian-Hua

(Institute of Intelligent Technology & Systems, Jiangmen 529020)

**Abstract** The paper introduces the approach of outlier detection based on kernel, and points out the approach is hard to be realized with the increase of sample number. In order to reduce the size of optimization, the sample selection method based on distance is proposed. Through selecting samples, the computation workload and the demand of EMS memory can be decreased greatly. In the end, the real-time demand can be met.

**Keywords** Kernel method, Outlier detection, Sample selection, Pattern recognition

野点(outlier)是显然严重偏离了样本集合中其它观测值的数据点[1]。以往关于野点的研究，主要是将其作为噪声处理，如主成分分析、投影寻踪等。并由此提出各种所谓"稳健"的处理方法，用于克服野点的干扰，甚至将野点数据从数据集中剔除。

然而，正如 Knorr 等人指出[2]："One person's noise is another person's signal"，同一个数据点，对不同的应用对象和应用目的可能产生截然不同的作用。的确，在一些应用领域中，如数据库中的知识发现、机械设备运行状态监测等，野点能提供比正常数据点更多、更重要的信息，是发现新知识、确定新状态的有力手段。

一种常用的野点检测方法是计算包括全部(或几乎全部)正常数据点的最小区域的边界，边界以外的数据点则视为野点。从这一点考虑，野点检测实际上与正常域(正常数据点范围，其中的样本在本文中称为正类样本)的确定属于同一个问题。

在本文中，作者主要从核方法的角度出发，探讨构造正类样本边界的途径。

## 1 核方法下基于边界的野点检测

从模式识别的角度考虑，野点检测实际上可视为一类特殊分类问题[3]。对于一般的分类问题，考虑的是如何将各种类别有效地分开，而在野点检测中，分类的目标是如何准确地描述一类对象(姑且称之为正类)，在此之外大范围的其它对象(称之为负类)则被视为野点，因此，野点检测有时又被称为一类分类问题[4]。

基于边界的野点检测方法本质上是寻找包含全体正类样本的最小球体，球体外即为野点区域。对于样本集

$$\Xi=\{x_1,x_2,\cdots,x_N\} \tag{1}$$

其中本样本都为正类。设将样本集中全体样本完全包围的最小球体的半径为 $R$，球心为 $a$，则满足优化方程

$$\min\ L(R)=R^2 \tag{2}$$

$$s.t.\ R^2-(x_i-a)(x_i-a)^T\geqslant 0 \tag{3}$$

由式(2)和式(3)可定义如下的 Lagrange 函数

$$L(R,a,\Lambda)=R^2-\sum_{i=1}^{N}a_i\{R^2-(x_i-a)(x_i-a)^T\} \tag{4}$$

式中 $a_i\geqslant 0$为 Lagrange 系数。将式(4)对 $R$ 和 $a$ 求偏微分，并令它们等于0，得

$$\sum_{i=1}^{N}a_i=1 \tag{5}$$

$$a=\sum_{i=1}^{N}a_ix_i \tag{6}$$

将式(5)和式(6)代入式(4)，稍作变换，即有优化方程

$$\max\ L=\sum_{i=1}^{N}a_i(x_i\cdot x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j(x_i\cdot x_j) \tag{7}$$

$$s.t.\ \sum_{i=1}^{N}a_i=1\ \ a_i\geqslant 0 \tag{8}$$

实际上，根据 KT 条件，$\Lambda$ 中大部分元素为0，只有一小部分 $a_i>0$，与这些 $a_i$ 对应的样本点决定了边界的构成，这些数据称为支持对象(support objects)。

已知 $\Lambda$，由式(6)即可求出球心 $a$，任选一支持对象由式(9)可求出 $R$

$$R^2-(x_i-a)(x_i-a)^T=0 \tag{9}$$

对于待定状态数据 $z$，令

$$f(z)=(z-a)(z-a)^T=(z\cdot z)-2\sum_{i=1}^{N}a_i(z\cdot x_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j(x_i\cdot x_j) \tag{10}$$

则依据下式可判断 $z$ 是否为野点

$$\begin{cases}f(z)\leqslant R^2 & z\text{ 不为野点}\\ f(z)>R^2 & z\text{ 为野点}\end{cases} \tag{11}$$

设 $x_S$ 为任一支持对象，$d(x,y)$用于描述向量 $x$ 与向量 $y$ 之间的距离。进一步考虑到 $\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j(x_i\cdot x_j)=const.$ 及 $d(x_S,a)=\max d(x_i,a)i=1,2,\cdots,N$，则我们还可以用另外一种方法来判断样本数据是否为野点。令

$$g(z)=(z\cdot z)-2\sum_{i=1}^{N}a_i(z\cdot x_i) \tag{12}$$

则野点的判别方式为

$$\begin{cases}g(z)<g(x_S) & z\text{ 不为野点}\\ g(z)>g(x_S) & z\text{ 为野点}\end{cases} \tag{13}$$

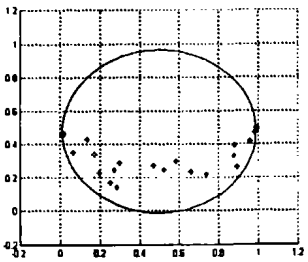
或

*)本文研究得到国家自然科学基金项目的资助(70071023)。肖健华 博士，主要研究人工智能理论与应用。

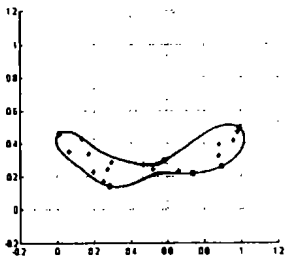
$$\lambda(z)=\begin{cases}\dfrac{g(z)}{g(x_S)}>1 & z\text{ 为野点}\\ <1 & z\text{ 不为野点}\end{cases} \tag{14}$$

显然采用 $g(z)=g(x_S)$ 确定边界，简化了计算，因为它只需要比较 $g(z)$ 与 $g(x_S)$ 的相对大小，而不要求绝对值，省去了计算 $R$ 和 $a$ 的过程。此外表达式(13)和式(14)本身所具有的含义也更明确了。

考虑图1所示正类点分布，所求得的边界列于图1(a)中，图中用小圆圈圈起的点为支持对象，从图中看出，由于采用球形边界，形状单一，且边界区域显得过大，不够紧密，所形成的区域容易进入野点区域。



图(a) 常规方法



图(b) 核方法

图1 正常域边界的形成方法

为此，在式(7)中引入核函数 $(x\cdot y)\to K(x,y)$，则在属性空间中的优化方程为

$$\max L=\sum_{i=1}^{N}a_iK(x_i,x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jK(x_i,x_j) \tag{15}$$

约束条件不变，式(10)和式(12)变为

$$f(z)=K(z,z)-2\sum_{i=1}^{N}a_iK(z,x_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jK(x_i,x_j) \tag{16}$$

$$g(z)=K(z,z)-2\sum_{i=1}^{N}a_iK(z,x_i) \tag{17}$$

特别地，选择高斯径向基核函数

$$K(x,y)=\exp(-\|x-y\|^2/\sigma^2) \tag{18}$$

$\sigma$ 可事先给出[5]。由式(18)有 $K(x,x)=1$，式(15)～式(17)转换为

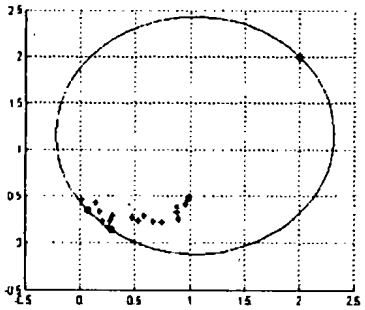
$$\max L=1-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jK(x_i,x_j) \tag{19}$$

$$f(z)=1-2\sum_{i=1}^{N}a_iK(z,x_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jK(x_i,x_j) \tag{20}$$

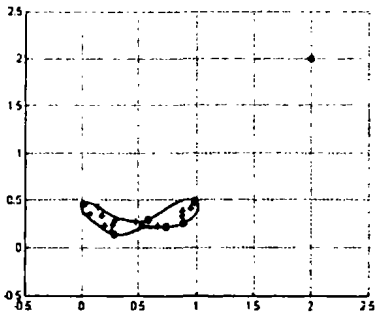
$$g(z)=-\sum_{i=1}^{N}a_iK(z,x_i) \tag{21}$$

仍用图1(a)中数据，现采用核方法求其正常域范围。由图1(b)可看出，通过选择合适的参数，采用核方法后求出的正类区域可以非常紧致，基本上没有剩余空间，因而具有更为优秀的野点识别能力。

此外，在一定范围内，采用核方法确定正常域边界还能排除干扰样本的影响。如图2所示为在图1已有数据点的基础上，增加噪声数据点(2,2)。图(a)为常规方法，由于干扰点的存在，正类区域急剧增大；图(b)所示采用核方法，与图1(b)比较仅在干扰点处增加了一个很小的区域，从而较易通过其它手段将这一孤立区域排除。



(a)常规方法



(b)核方法

图2 干扰信号作用下对边界影响的比较

与支持向量机类似，为了实现错误划分和区域范围之间的折中，可在优化过程中引入松弛变量，此时样本集满足[5]

$$\min L(R)=R^2+C\sum_{i=1}^{N}\xi_i \tag{22}$$

$$s.t.\ (x_i-a)^T(x_i-a)\leqslant R^2+\xi_i$$

$$\xi_i\geqslant 0\quad i=1,2,\cdots,N$$

此时的优化方程为

$$\max L=\sum_{i=1}^{N}a_i(x_i,x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j(x_i,x_j) \tag{23}$$

$$s.t.\ 0\leqslant a_i\leqslant C\quad \sum_{i=1}^{N}a_i=1$$

核方法下的优化方程为

$$\max L=\sum_{i=1}^{N}a_iK(x_i,x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jK(x_i,x_j) \tag{24}$$

约束条件不变。

与式(7)比较，仅是约束条件有了变化。以后的表达式(不论是否引入了核变换)都与前面相同。

## 2 基于距离的样本选择方法及其在实时野点检测中的应用

在一些野点检测过程中，样本点是随着时间逐渐增加的，如机械设备运行状态监测。如果每增加一个正类样本，就采用式(15)或式(24)进行一次优化，进而由式(16)或式(17)确定边界区域，这显然是不现实的，也是不可能的。因为随着机械设备运行时间的推移，正类样本数目不断增加，优化过程所需的时间和存储空间不断增加，最终必然导致优化过程不可能进行。因此有必要对样本进行筛选。

从前面的讨论看出，在由非线性函数 $\Phi$ 形成的核空间 $H$ 下的正常域范围是在特征空间中建立一个以

$$b=\Phi(a)=\Phi(\sum_{i=1}^{N}a_ix_i) \tag{25}$$

为球心，以

$$R=\|\Phi(x_S)-\Phi(a)\| \tag{26}$$

为半径的超球形区域。$x_S$ 为一支持对象。

通常在正常域中心附近的一个区域 $\delta$

$$\delta=\|\Phi(x)-\Phi(a)\|<R \tag{27}$$

内，样本点较为集中，且随着样本数据的增加与边界形成的关系越来越小。这样就可以通过

$$\lambda_{N+1}=\frac{d(x_{N+1},a_N)}{R_N} \tag{28}$$

判断正类样本 $x_{N+1}$ 的加入是否会对正类边界区域产生影响。式(28)中 $d(x_{N+1},a_N)$ 为样本 $x_{N+1}$ 与前 $N$ 个样本形成边界球心 $a_N$ 在属性空间中的距离，$R_N$ 为前 $N$ 个样本形成边界的半径。

设定一阈值 $C_0\in(0,1)$，只有当

$$\lambda_{N+1}>C_0 \tag{29}$$

成立时，我们才认为新增 $x_{N+1}$ 有可能对边界产生影响。基于这种思想，就可以建立图3所示的正类区域，同时也是野点检测方案。

图3中 $D_N$ 为第 $N$ 个样本加入后的样本子集，且该样本子集中的每一个样本 $x$ 都满足条件式(29)，即

$$\frac{d(x,a_N)}{R_N}>C_0 \tag{30}$$

通过后面的算例，我们将看到：随着 $N$ 的增大，对于大部分样本而言，仅需进行距离判断，而不需要作后面的野点判断，从而量级化地减少了计算工作量，也就保证了野点检测的实时性要求。

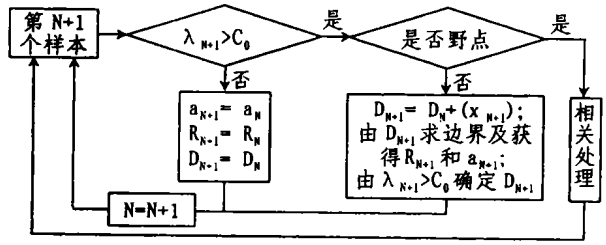



图3 基于距离判别的正常域确定方案

**定理** 已知两个样本数据 $x_1$ 和 $x_2$，经非线性函数 $\Phi$ 作用映射到属性空间 $H$，则这两个向量在特征空间中的 Euclidean 距离为[6]

$$d^H(x_1,x_2)=\sqrt{K(x_1,x_1)-2K(x_1,x_2)+K(x_2,x_2)} \tag{31}$$

证明：$d^H(x_1,x_2)=\|\Phi(x_1)-\Phi(x_2)\|$

$$=\sqrt{[\Phi(x_1)-\Phi(x_2)]^T[\Phi(x_1)-\Phi(x_2)]}$$

$$=\sqrt{\Phi^T(x_1)\Phi(x_1)-\Phi^T(x_1)\Phi(x_2)-\Phi^T(x_2)\Phi(x_1)+\Phi^T(x_2)\Phi(x_2)}$$

$$=\sqrt{\Phi^T(x_1)\Phi(x_1)-2\Phi^T(x_1)\Phi(x_2)+\Phi^T(x_2)\Phi(x_2)}$$

$$=\sqrt{K(x_1,x_1)-2K(x_1,x_2)+K(x_2,x_2)}$$

式(31)给出了属性空间两点之间距离的计算公式，结合式(28)和式(29)就给出了完整的野点检测方案。

在属性空间中，与某点等距离的曲面是一个超球面，然而将之投影回输入空间形成的等距离曲面，则可能为各种形式，如图4。

## 3 算例分析

仿真算例：设在图5所示的椭圆区域(该椭圆区域也是后面各种情况下期望的正常域范围) $x_1^2/100+x_2^2/25=1$ 内，按正态分布 $x_1\sim N(0,10)$ 和 $x_2\sim N(0,5)$ 产生300个正类样本，$x_1$ 和 $x_2$ 相互独立。为方便与上一节中结果对照，本例仍采用高斯径向基核函数

$$K(x,y)=\exp(-\|x-y\|^2/25) \tag{32}$$

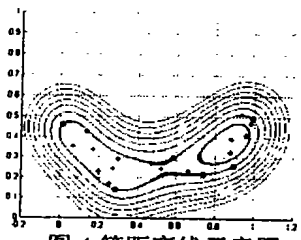


图4 等距离线示意图

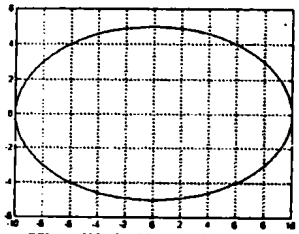
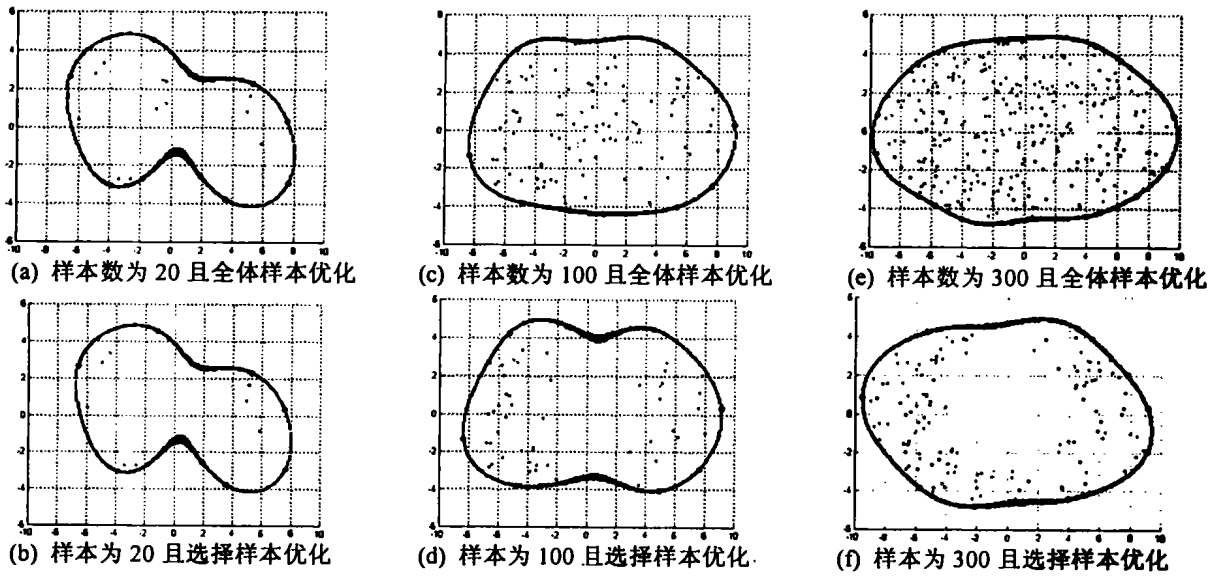


图5 样本数据采样区间

采用两种方法对数据进行处理，结果如图6所示。图(a)、图(c)和图(e)分别为不进行样本选择情况下，样本数目为20，100，300个样本时的正常域区间；图(b)、图(d)和图(f)分别为进行样本选择时的正常域区间。



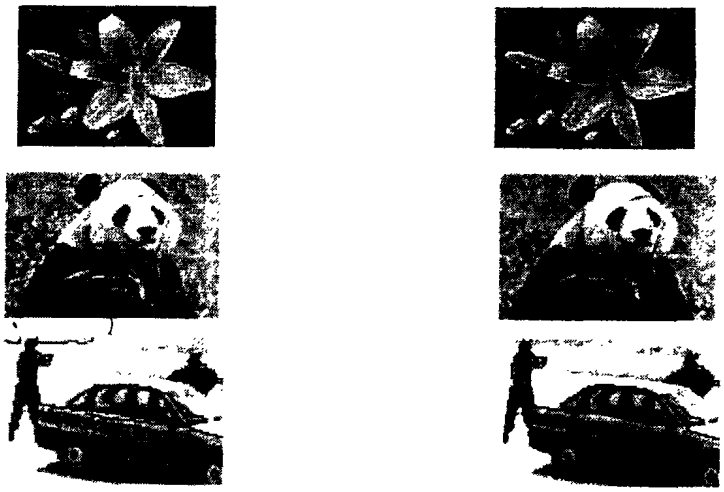
(a) 样本数为 20 且全体样本优化　(c) 样本数为 100 且全体样本优化　(e) 样本数为 300 且全体样本优化

(b) 样本为 20 且选择样本优化　(d) 样本为 100 且选择样本优化　(f) 样本为 300 且选择样本优化

图6 样本选择效果示意图

(下转第173页)

a)原始图像　b)SOM 粗聚类　c)去离散点



d)基于像素数的合并　e)基于颜色的合并

图3　实例的运行过程

(上接第161页)

从对应的图中可以看出，进行样本选择与不进行样本选择所获得的边界区域基本上没有差异，然而进行样本选择却大大减少了计算时间。以300个样本和作者所用的计算机为例，以最后一步作为比较，即从第300个样本输入到该样本判断完毕为止，进行样本选择时所需的时间为10秒，不进行样本选择时所需的时间为88秒，两者之间的比例约为1/9。更为重要的是，随着样本数目的增加，越来越多的样本仅需要判断距离，而不需要进行优化计算，这样所需要的时间就更少了，这一点从图(d)和图(f)可以清晰地看出。

最后说明两点：

(1)为了尽快形成合理的正常域边界，可设定一参数 $N_0$，只有当 $N>N_0$，才对样本进行选择。

(2)式(29)中参数 $C_0$ 的设置，可以动态选定，即 $C_0$ 随着样本数目 $N$ 的增加而增大，但要满足 $C_0\in(0,1)$。

**结论**　本文针对野点检测的实时性要求，提出了基于距离的样本点选择方法，大大地降低了运行时间和内存上的需求。为了方便描述，同时增强数据的可视性，本文的讨论针对二维输入进行，实际上，所得的结论在高维中同样适用。

## 参考文献


1 Anscombe F J. Rejection of Outliers. technometrics, 1960(2): 123～147

2 Knorr E M, Ng R T. Algorithms for Mining Distance-based Outliers in Large Datasets. In: Proc. VLDB, 1998. 392～403

3 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121～167

4 Moya M R, Koch M W. Hostetler L R. One-class Classifier Networks for Target Recognition Applications. In: Proc. World Congress on Neural Networks, Portland, 1993. 797～801

5 Tax D, Duin R. Data Domain Description Using Support Vectors. In: Proc. European Symposium Artificial Neural Networks, 1999. 251～256

6 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