第9章 离群点分析

目录
CONTENTS

离群点定义与类型

什么是离群点：

- **离群点：是一个数据对象，它显著不同于其他数据对，好像它是被不同的机制产生的一样。**
  - 例如：不同寻常的信用卡交易
- **离群点不同于噪声数据**
  - 噪声数据是被观测变量的随机误差或方差
  - 噪声数据应在离群点检测前被删除
- **离群点产生原因：**
  - 计算的误差或者操作的错误所致
  - 数据本身的可变性或弹性所致

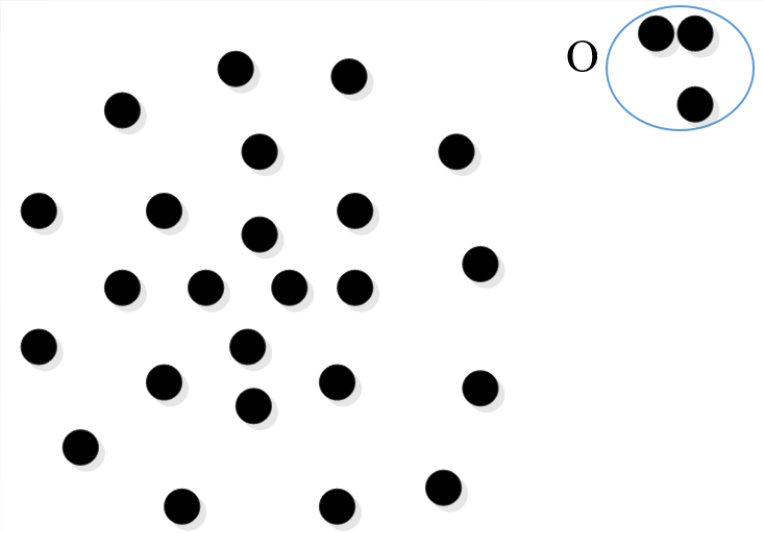


图9-1区域O中的对象为离群点

离群点的类型：

三类: 全局离群点, 条件离群点和集体离群点。

1. 全局离群点 (或点异常)

- 如果一个数据对象显著的偏离数据集的其余部分，则这个数据对象为全局离群点。

2. 条件离群点

- 一个数据对象（样本），如果关于对象的特定情境，它显著偏离其他对象。
- 例如： 多伦多的温度为28° C，这是离群点吗？ (取决于冬天还是夏天)
- 数据对象的属性划分为两组。
  - 情境属性: 定义对象的情境，例如， 时间和地点。
  - 行为属性: 定义对象的特征, 并用来评估对象关于它所处的情境是否为离群点。例如，温度。

离群点的类型：

3. 集体离群点

- 给定一个数据集，数据对象的**一个子集**作为整体显著偏离整个数据集，数据对象的这个子集称为集体离群点。
- 应用：在入侵检测时， 多台计算机不断地相互发送拒绝服务包，则它们可以视为集体离群点，所涉及的计算机可能受到攻击。

数据集可能有多种类型的离群点。

一个对象可能属于多种类型的离群点。

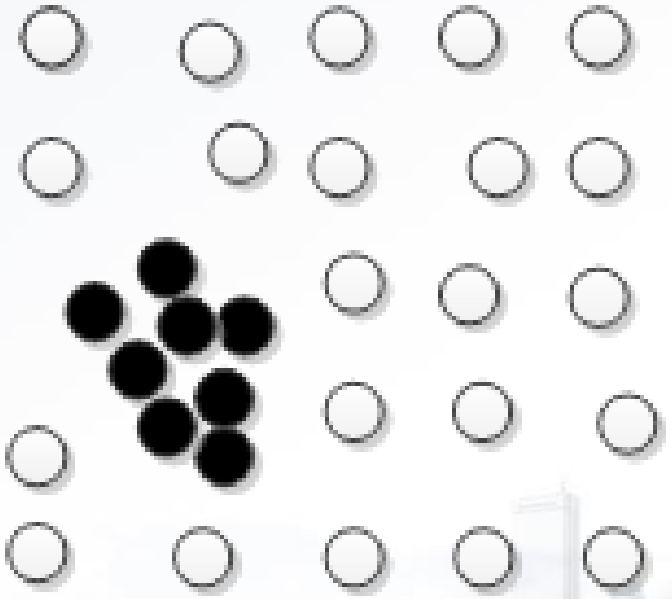


图9-2黑色对象形成集体离群点

Chapter 9.2

离群点检测

– 离群点的检测方法有很多，每种方法在检测时，都会对正常数据对象或离群点做出假设。从这个假设的角度考虑，离群点检测方法可以分为:

1. 基于统计学的离群点检测;
2. 基于近邻的离群点检测;
3. 基于聚类的离群点检测;
4. 基于分类的局部离群点检测。

1. 统计学方法

基于统计分布的检测方法是为数据集构建一个概率统计模型（例如正态、泊松、二项式分布等，其中的参数由数据求得），然后根据模型采用不和谐检验识别离群点。图9.1给出了基于统计分布的检测流程。

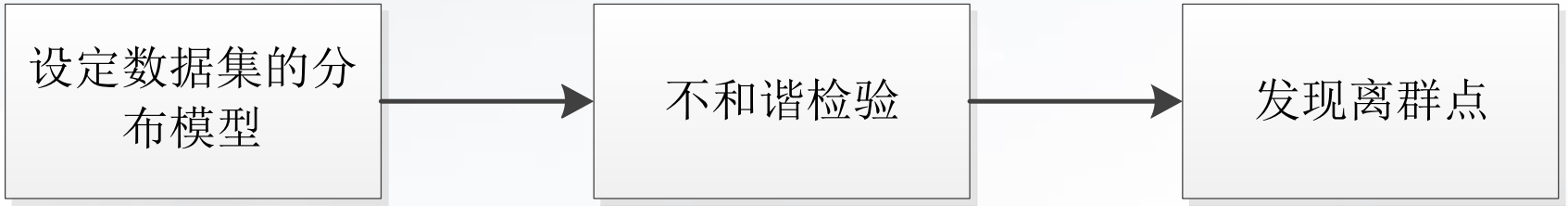


图9.1 基于统计的离群点检测流程

不和谐检验:

- 不和谐检验需要检查两个假设：**工作假设**和**备择假设**。
- **工作假设指的是**如果某样本点的某个统计量相对于数据分布的是**显著性概率充分小**，则认为该样本点是不和谐的，工作假设被拒绝；
- **此时备择假设被采用**，它声明该样本点来自于另一个分布模型。
- 如果某个样本点不符合工作假设，那么认为它是离群点。如果它符合备择假设，认为它是**符合某一备择假设分布的离群点**。

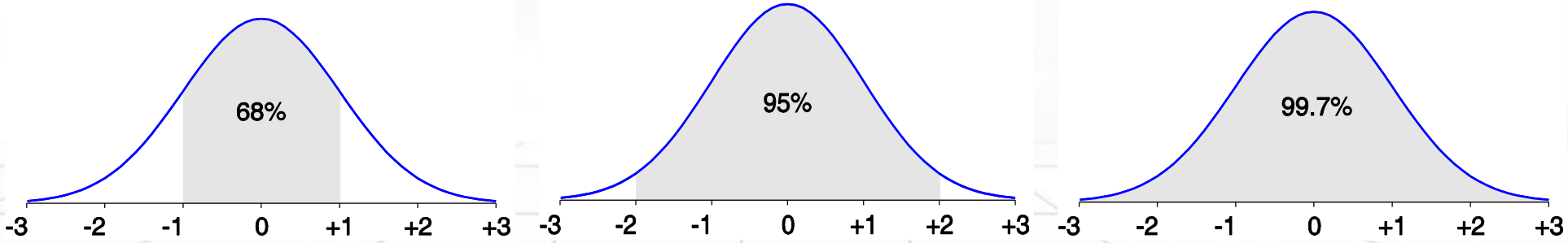
工作假设H为，假设n个对象的整个数据集来自一个初始的分布模型F，即：

$$H: o_i \in F，其中 i=1，2，\ldots，n$$

不和谐检验就是检查对象o_i关于分布F是否显著地大（或小）。

基于正态分布的一元离群点检测

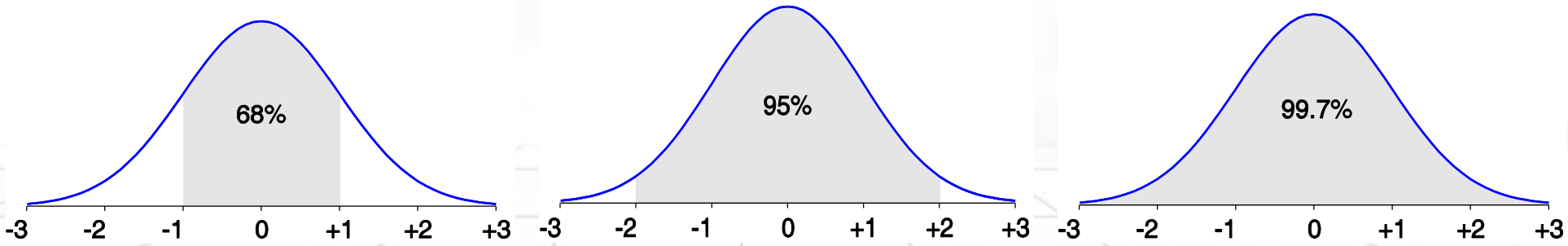
- 正态分布曲线特点：N（μ，σ^2）
- 变量值落在（μ-σ，μ+σ）区间的概率是68.27%
- 变量值落在（μ-2σ，μ+2σ）区间的概率是95.44%
- 变量值落在（μ-3σ，μ+3σ）区间的概率是99.73%



基于正态分布的一元离群点检测

一般的，设属性X取自具有期望值μ，方差σ^2的正态分布N（μ，σ^2），如果属性X满足：$P(|X|\geq C)=\alpha$,

其中C是一个选定的常量，**则X以概率1-α为离群点。**



例9.1　基于统计方法检测年龄离群点

设儿童上学的具体年龄总体服从正态分布，所给的数据集是某地区随机选取的开始上学的20名儿童的年龄。具体的年龄特征如下：年龄={6，7，6，8，9，10，8，11，7，9，12，7，11，8，13，7，8，14，9，12}

相应的统计参数是：均值m=9.1；标准差s=2.3。

如果选择数据分布的阈值q按如下公式计算：$q=m\pm2\times s$，则阈值下限与上限分别为4.5和13.7。

如果将工作假设描述为：儿童上学的年龄分布在阈值设定区间内，则依据不和谐检验，不符合工作假设，即在[4.5，13.7]区间以外的年龄数据都是潜在的离群点，将最大值取整为13，所以年龄为14的孩子可能是个例外。

统计方法的离群点检测的优缺点:

– **优点**

- 建立在非常标准的统计学原理之上，当数据和检验的类型十分充分时，检验十分有效。

统计方法的离群点检测的优缺点:

– **缺点**

①多数情况下，数据的**分布是未知的**或数据**几乎不可能用标准的分布来拟合**，虽然可以使用混合分布对数据建模，基于这种模型开发功能更强的离群挖掘方案，但这种模型更复杂，难以理解和使用。

②当观察到的分布不能恰当地用任何标准的分布建模时，基于统计方法的挖掘不能确保所有的离群点被发现，而且要确定**哪种分布最好**地拟合数据集的代价也非常大。

③即使这类方法在低维（一维或二维）时的数据分布已知，但**在高维情况下**，估计数据对象的分布是极其困难的，对每个点进行分布测试，需要**花费更大的代价**。

2. 基于近邻的离群点检测

- **假定：离群点对象与它最近邻的邻近性显著偏离数据集中其它对象与它们邻近之间的邻近性。**
- **两种方法：**
  1. 基于距离的离群点检测；
  2. 基于密度的离群点检测。

(1)基于距离的离群点检测:

- 如果数据对象集D中大多数对象都远离d，即都不在d的**r-邻域内**，d可视为一个离群点。
- r 是距离阈值， α是分数阈值，如果有 $\frac{\|\{d' \mid dist(d,d') \le r\}\|}{\|D\|} \le \alpha$

  其中||D||是对象的数量个数,分子是满足 $dist(d,d') \le r$ 的点 d' 的个数 。

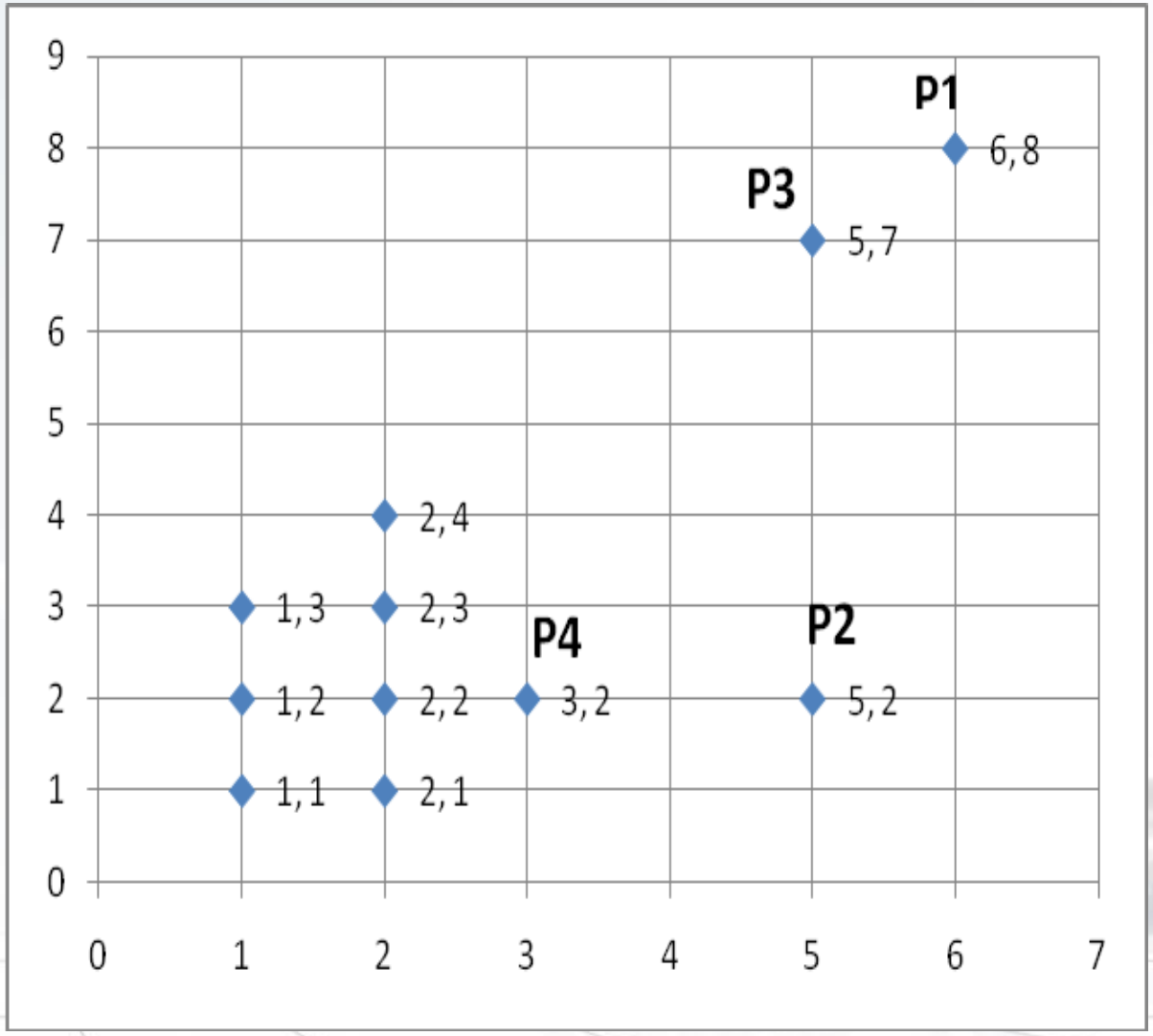
则d是一个D(r, α)离群点。

- 如何计算D(r, α)-离群点：嵌套循环

  对每个对象 d_i （ $1 \le i \le n$ ），计算 d_i 与其它对象之间的距离，统计 r-邻域中其它对象的个数，一旦找到 n×α 个，内循环可以中止，n=||D||。

- 如图所示，$\alpha=2/3$，$r=7$,判断图中d为P1、P2、P3、P4是否是离群点.

$$\frac{\left\|\{d' \mid dist(d,d') \le r\}\right\|}{\|D\|} \le \alpha$$

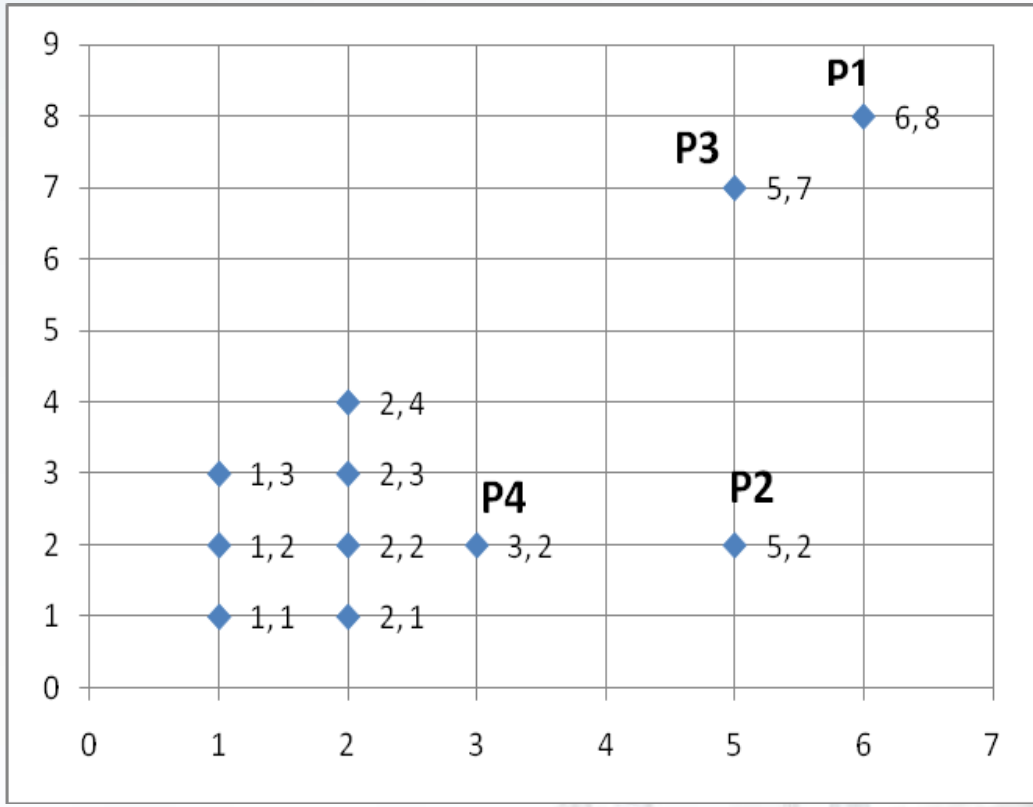


k最近邻算法

假设有一个样本p，用一种距离的度量方法（例如欧氏距离）来计算得到离p最近的k个样本，这k个样本可以组成集合$N(p,k)$。p的离群因子可以定义为：

$$\mathrm{OF1}(p,k)=\frac{\sum_{y\in N(p,k)} distance(p,y)}{|N(p,k)|}$$

其中，$distance(p,y)$表示样本p和y的距离度量，$|N(p,k)|$表示集合$N(p,k)$的大小，即所包含样本的个数。离群因子越大，越有可能是离群点。



当k =2时，使用欧式距离，判断P1、P2哪个点更可能是离群点？

(2)基于密度的离群点检测:

- 基于密度的离群点检测能够检测出基于距离的异常算法所不能识别的一类异常数据——**局部离群点。**
- **局部离群点：**是指一个对象相对于它的局部邻域，特别是关于邻域密度，它是远离的。
- 利用**局部知识**而非全局知识。
- 算法考虑的是**对象与它邻近的密度**，如果一个对象的密度相对于其它的邻近低得多，则视此对象为离群点。

(2)基于密度的离群点检测:

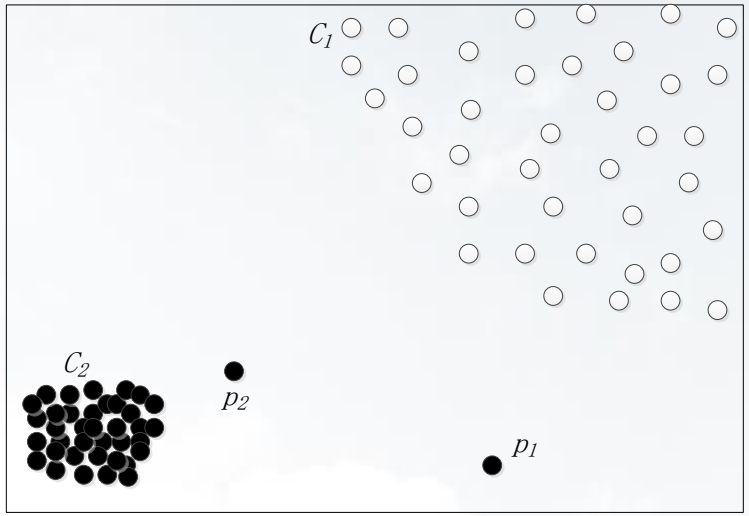


图9.2 基于密度的局部离群点检测的必要性

聚类簇C_1属于低密度区域，聚类簇C_2属于高密度区域。依据传统的基于密度的离群点检测算法，C_1中任何一个数据点q与其近邻的距离大于数据点p_2与其在C_2中的近邻的距离，数据点p_2会被看作是正常点，当然能检测出数据点p_1是离群点。

(2)基于密度的离群点检测:

图9.2中，**p2相当于C2的密度来说是一个局部离群点**，这就形成了基于密度的局部离群点检测的基础。此时，评估的是一个对象是离群点的程度，这种“离群”程度就是作为对象的**局部离群因子**（LOF），然后计算 。

数据集中的数据点x和x_i，x到x_i的**可达距离**reach_dist$_k(x, x_i)$定义为

$$reach\_dist_k(x, x_i) = \max\{dist_k(x_i), dist(x, x_i)\}$$

其中，$dist_k(x_i)$指数据点x_i到其第k个近邻的距离，$dist(x, x_i)$指数据点x和x_i的距离。通常，距离度量**选用欧式距离**，而且x到x_i的可达距离$reach_dist_k(x, x_i)$与x_i到x的可达距离$reach_dist_k(x_i, x)$**一般并不相同**。

(2)基于密度的离群点检测:

- **局部可达密度**

  对象x的局部可达密度定义为x的k最近邻点的**平均可达距离**的**倒数，反映出距离越大，密度越小**:

$$lrd_k(x) = \frac{k}{\sum_{x_i \in KNN(x)} reach\_dist_k(x, x_i)}$$

- **局部离群因子（LOF）**表征了x是离群点的程度，定义如下:

$$LOF_k(x) = \frac{\sum_{x_i \in KNN(x)} \frac{lrd_k(x_i)}{lrd_k(x)}}{k}$$

(2)基于密度的离群点检测:

– **结论**

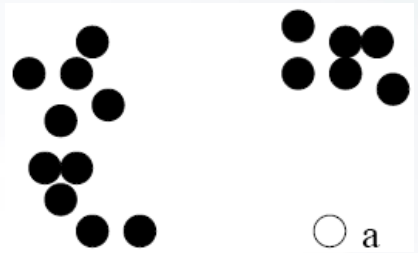
- LOF算法计算的离群度不在一个通常便于理解的范围[0,1]，而是**一个大于1的数，并且没有固定的范围**。而且数据集通常数量比较大，内部结构复杂，LOF极有可能因为取到的近邻点属于不同数据密度的聚类簇，使得计算数据点的近邻平均数据密度产生偏差，而得出与实际差别较大甚至相反的结果。

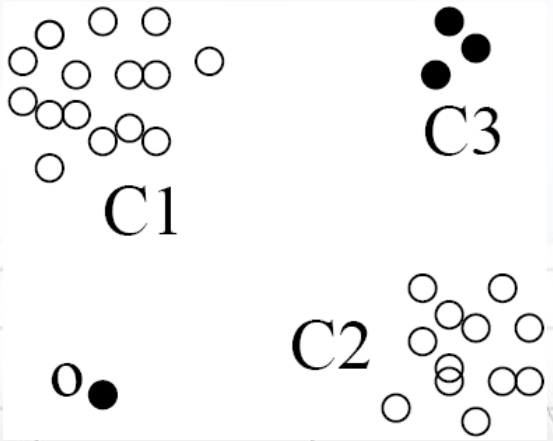
– **优点**

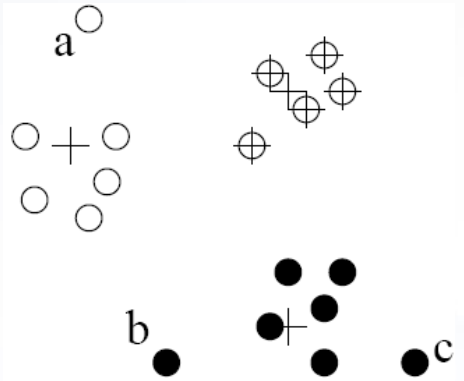
- 通过基于密度的局部离群点检测就能**在样本空间数据分布不均匀的情况下**也可以准确发现离群点。

3. 基于聚类的方法

- 该对象属于某个簇吗？如果不，则它被识别为离群点。
- 该对象与最近的簇之间的距离很远吗？如果是，则它是离群点。
- 该对象是小簇或稀疏簇的一部分吗？如果是，则该簇中的所有对象都是离群点。







基于聚类的离群点检测挖掘方法有两种：

(1)基于对象离群因子法

- 假设数据集D被聚类算法划分为k个簇$C=\{C_1, C_2, \ldots, C_k\}$，对象$p$的**离群因子（Outlier Factor）** $OF1(p)$定义为**p与所有簇间距离的加权平均值**：

- $$OF1(p) = \sum_{j=1}^{k} \frac{|C_j|}{|D|} \cdot d(p, C_j) \quad (9\text{-}6)$$

- 其中，$d(p, C_j)$表示对象p与第j个簇C_j之间的距离。$|C_j|$是簇C_j的样本数，$|D|$是数据集D的样本数。

(1)基于对象离群因子法：

- 两阶段离群点挖掘方法如下：

① 对数据集D采用一趟聚类算法进行聚类，**得到聚类结果**$C=\{C_1, C_2, \ldots, C_k\}$

② 计算数据集D中所有**对象p的离群因子**OF1(p)，及其平均值Ave_OF和标准差Dev_OF，满足条件：$OF1(p) \geq Ave\_OF+\beta \times Dev\_OF(1 \leq \beta \leq 2)$ 的对象判定为离群点。通常取β=1或1.285。

(1)基于对象离群因子法：

例9.2基于对象的离群因子法

对于图9-5所示的二维数据集，比较点p_1(6，8)，p_2(5，2)，哪个更有可能成为离群点。假设数据集经过聚类后得到聚类结果为C={C_1，C_2，C_3}，图中红色圆圈标注，三个簇的质心分别为：C_1(5.5，7.5)、C_2(5，2)、C_3(1.75，2.25)，试计算所有对象的离群因子。

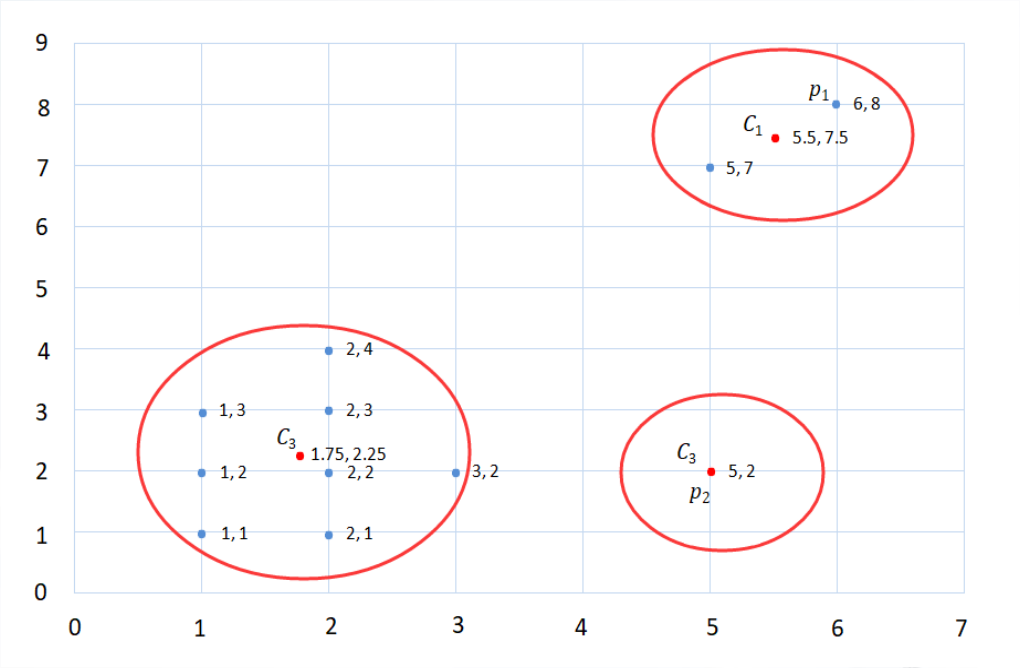


图9-5 基于聚类的离群点检测二维数据集

(1)基于对象离群因子法：

解：根据对象p的离群因子（Outlier Factor）$OF1(p)$的定义，对于p_1点有：

$$OF1(p_1)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(p_1,\ C_j)=\frac{8}{11}\sqrt{(6-1.75)^2+(8-2.25)^2}+$$

$$\frac{1}{11}\sqrt{(6-5)^2+(8-2)^2}+\frac{2}{11}\sqrt{(6-5.5)^2+(8-7.5)^2}=5.9$$

对于p_2有：

$$OF1(p_2)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(p_2,\ C_j)=\frac{8}{11}\sqrt{(5-1.75)^2+(2-2.25)^2}+$$

$$\frac{1}{11}\sqrt{(5-5)^2+(2-2)^2}+\frac{2}{11}\sqrt{(5-5.5)^2+(2-7.5)^2}=3.4$$

可见，点p_1较p_2更可能成为离群点。

(1)基于对象离群因子法：

同理可求得所有对象的离群因子，结果如表9-1所示。

进一步求得所有点的离群因子平均值：

$Ave\_OF=2.95$，

标准差：　$Dev\_OF=1.3$，

假设$\beta = 1$，则阈值：

$$E = Ave\_OF + \beta \times Dev\_OF = 2.95 + 1.3 = 4.25,$$

离群因子大于4.25的对象可视为离群点，p_1离可视为离群点。

表9-1 离群因子表

X	Y	OF1
1	2	2.2
1	3	2.3
1	1	2.9
2	1	2.6
2	2	1.7
2	3	1.9
6	8	5.9
2	4	2.5
3	2	2.2
5	7	4.8
5	2	3.4

(2)基于簇的离群因子法：

假设数据集D被聚类算法划分为k个簇$C=\{C_1, C_2, \cdots, C_k\}$，簇$C_i$离群因子（Outlier Factor）$OF2(C_i)$定义为**簇$C_i$与其他所有簇间距离的加权平均值**：

$$OF2(C_i)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(C_i, C_j) \quad (9\text{-}7)$$

$d(C_i, C_j)$为C_i簇与簇C_j的质心（中心）的距离。如果一个簇离几个大簇的距离都比较远，则表明该簇偏离整体比较远，其离群因子也较大。$OF2(C_i)$度量了C_i偏离整个数据集的程度，其值越大，说明C_i偏离整体越远。

(2)基于簇的离群因子法：

基于簇的离群因子离群点检测算法描述如下：

①聚类：对数据集D进行聚类，得到聚类结果$C=\{C_1, C_2, \cdots, C_k\}$；

②确定离群簇：计算**每个簇**$C_i(1\leq i\leq k)$**的离群因子**$OF2(C_i)$，按$OF2(C_i)$递减的顺序重新排列$C_i(1\leq i\leq k)$，求满足

$$\sum_{j=1}^{k}\frac{|C_j|}{|D|}\geq\varepsilon\ (0<\varepsilon<1)\quad(9\text{-}8)$$

的最小下标b，将簇C_1，C_2，…，C_b标识为“**outlier**”类（**即每个对象均看成离群**），而将C_{b+1}，C_{b+2}，…，C_k标识为“**normal**”类（**即其中每个对象均看成正常**）。

(2)基于簇的离群因子法：

例9. 3基于簇的离群因子法

对于图9-5所示的二维数据集，聚类后得到三个簇$C=\{C_1, C_2, C_3\}$，簇心分别为：$C_1(5.5, 7.5)$、$C_2(5, 2)$、$C_3(1.75, 2.25)$。

按照欧氏距离计算簇之间的距离，分别为：

$$d(C_1, C_2)=\sqrt{(5.5-5)^2+(7.5-2)^2}=5.52$$

$$d(C_1, C_3)=\sqrt{(5.5-1.75)^2+(7.5-2.25)^2}=6.45$$

$$d(C_2, C_3)=\sqrt{(5-1.75)^2+(2-2.25)^2}=3.26$$

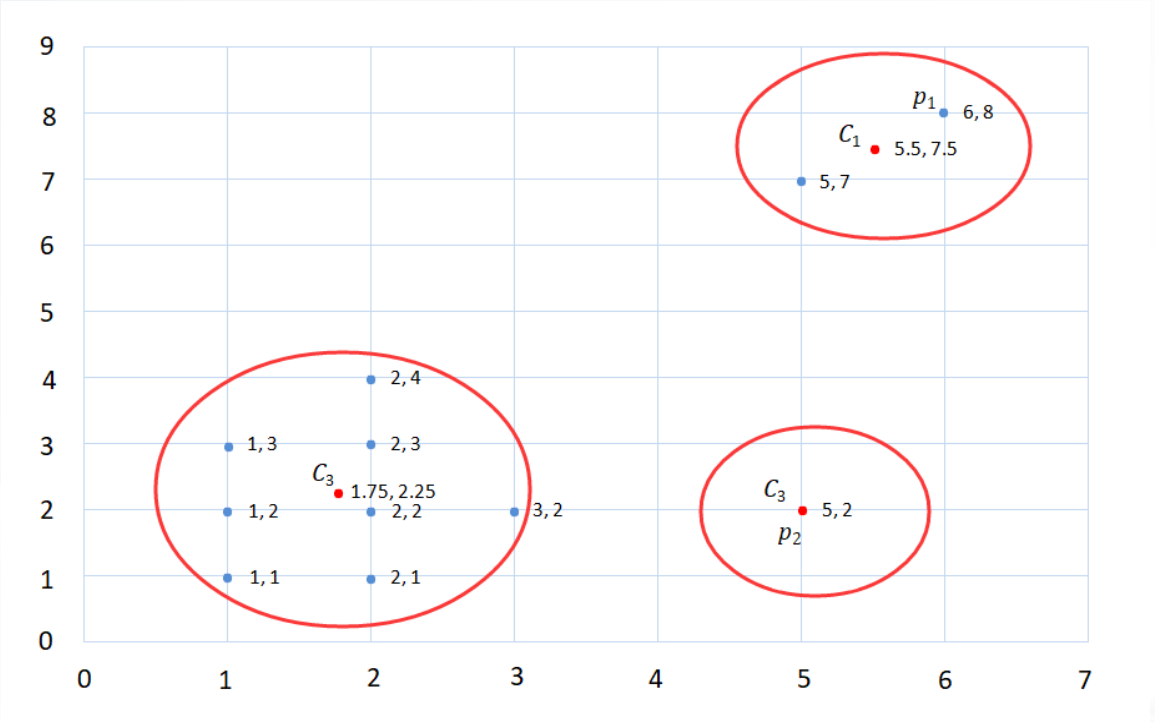


图9-5 基于聚类的离群点检测二维数据集

(2)基于簇的离群因子法：

例9. 3基于簇的离群因子法

进一步计算三个簇的离群因子，具体如下：

$$OF2(C_1)=\frac{1}{11}d(C_1，C_2)+\frac{8}{11}d(C_1，C_3)=\frac{1}{11}\times 5.52+\frac{8}{11}\times 6.45=5.19$$

$$OF2(C_2)=\frac{2}{11}d(C_2，C_1)+\frac{8}{11}d(C_2，C_3)=\frac{2}{11}\times 5.52+\frac{8}{11}\times 3.26=3.37$$

$$OF2(C_3)=\frac{2}{11}d(C_3，C_1)+\frac{1}{11}d(C_3，C_2)=\frac{2}{11}\times 6.45+\frac{1}{11}\times 3.26=1.47$$

可见簇C_1的离群因子最大，其中包含的对象判定为离群点，与例9. 2得到的结论相同。

3. 基于聚类的方法

- 基于聚类的离群点检测方法具如下优点：
  ① 它们可以检测离群点，而不要求数据是有标号的，即它们以**无监督方式检测**。它们对许多类型的数据都有效..
  ② 簇可以看做数据的概括，一旦得到簇，基于聚类的方法只需要把对象与簇进行比较，以确定该对象是否是离群点。这一**过程通常很快**，因为与对象总数相比，簇的个数通常很小。
- 基于聚类的方法的缺点是，它的有效性**高度依赖于所使用的聚类方法**。这些方法对于离群点检测而言可能不是最优的。对于大型数据集，聚类方法通常开销很大，这可能成为一个瓶颈。

4. 基于分类的方法

- 使用基于分类检测离群点的时候，分类器可以使用前面介绍的常用的分类器，如SVM、KNN、决策树等。
- 为解决正常数据和离群点数据分布的不均衡，可以使用一类模型进行分类。简单来说就是**构建一个描述正常数据的分离器，不属于正常的数据就是离群点**。

例9.2 使用SVM检测离群点。

在图9.3中，三个圆圈内的样本是正常数据，圆圈外的数据是离群点。可以使用圆圈内的正常数据训练一个决策边界，通过这个边界就可以区分数据是正常数据还是非正常数据—离群点。即，如果给定的新对象在正常类的**决策边界内**，则**被视为正常的**；如果新对象在**边界外**，则**被视为离群点**。这样就不需要训练离群点数据模型，避免了由于数据分布不均衡造成的分类器准确率低的现象。

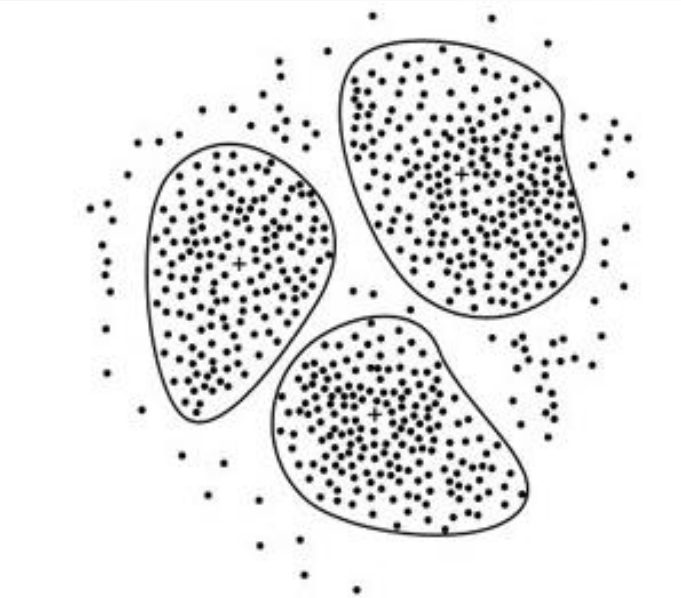


图9.3 使用SVM检测离群点数据样本