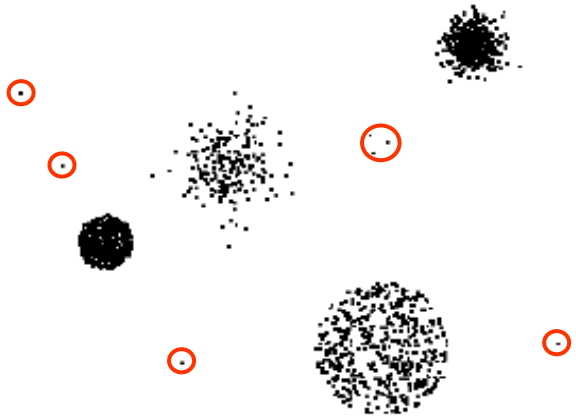
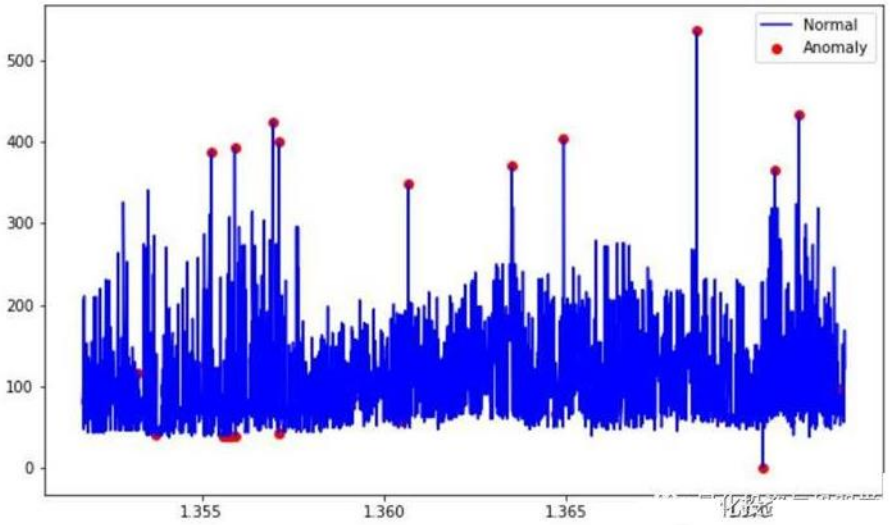
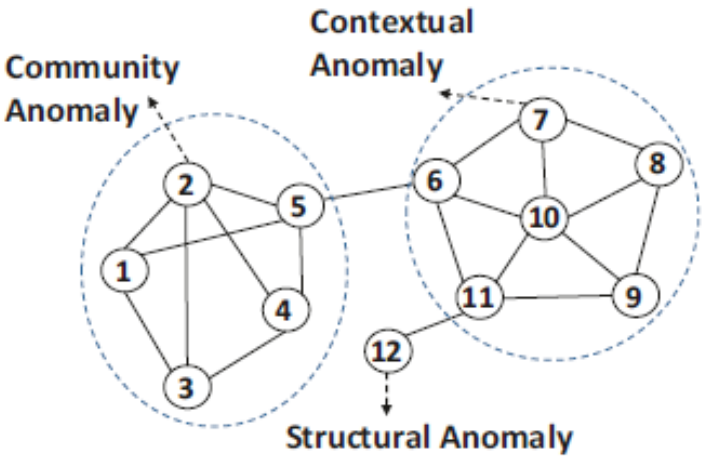
第12章 异常检测

什么是异常(Outlier)？

- Hawkins的定义：异常是在数据集中**偏离大部分数据的数据**，使人怀疑这些数据的偏离并非由随机因素产生，而是产生于完全不同的机制。
- Weisberg的定义：异常是与数据集中其余部分不服从相同统计模型的数据。
- Samuels的定义：异常是足够地不同于数据集中其余部分的数据。
- Porkess的定义：异常是远离数据集中其余部分的数据







ID	f1	f2	f3
1	2	22	48
2	11	23	48
3	3	21	47
4	4	22	46
5	3	24	49
6	12	19	51
7	11	38	55
8	10	19	56
9	13	20	49
10	14	20	49
11	11	21	53
12	12	22	52

异常数据具有特殊的意义和很高的实用价值

- 现有数据挖掘研究大多集中于发现适用于大部分数据的常规模式, 在许多应用领域中，异常数据通常作为噪音而忽略，许多数据挖掘算法试图降低或消除异常数据的影响。而在有些应用领域识别异常数据是许多工作的基础和前提，异常数据会带给我们新的视角。
- 如在欺诈检测中，异常数据可能意味欺诈行为的发生，在入侵检测中异常数据可能意味入侵行为的发生。

异常检测的应用领域

- 电信、保险、银行中的欺诈检测与风险分析
- 发现电子商务中的犯罪行为
- 灾害气象预报
- 税务局分析不同团体交所得税的记录，发现异常模型和趋势
- 海关、民航等安检部门推断哪些人可能有嫌疑
- 海关报关中的价格隐瞒
- 营销定制：分析花费较小和较高顾客的消费行为
- 医学研究中发现医疗方案或药品所产生的异常反应
- 计算机中的入侵检测
- 运动员的成绩分析
- 应用异常检测到文本编辑器，可有效减少文字输入的错误
- ……

什么是异常挖掘？

- 异常挖掘可以描述为：给定N个数据对象和所期望的异常数据个数k，发现明显不同的、意外的或与其它数据不一致的前k个对象。
- 异常挖掘问题由两个子问题构成：

  (1) 如何度量异常；

  (2) 如何有效发现异常。

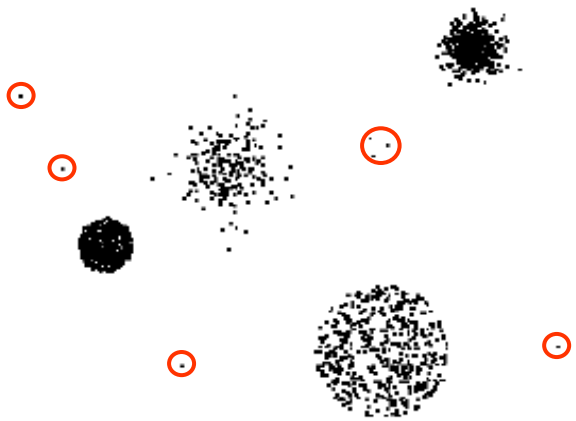
为什么会出现异常数据？

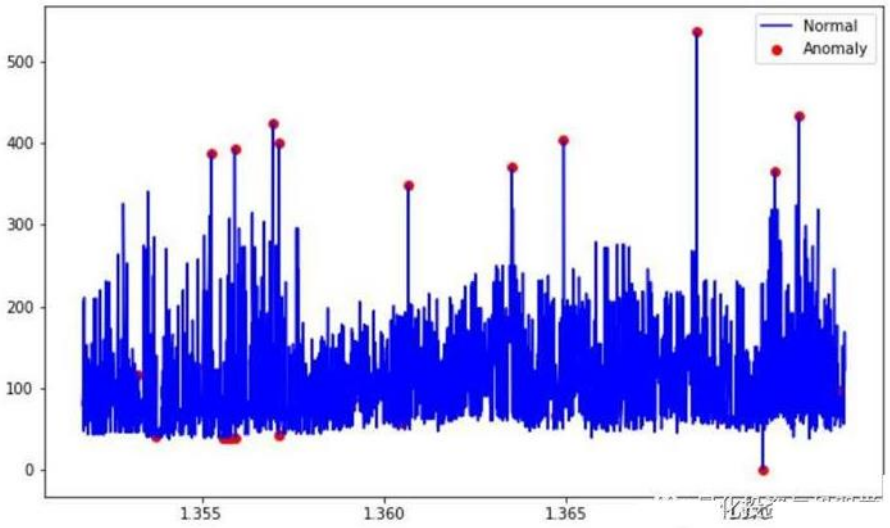
- 测量、输入错误或系统运行错误所致
- 数据内在特性所决定
- 客体的异常行为所致

➢ 由于异常产生的机制是不确定的，异常挖掘算法检测出的“异常数据”是否真正对应实际的异常行为，不是由异常挖掘算法来说明、解释的，只能由领域专家来解释，异常挖掘算法只能为用户提供可疑的数据，以便用户引起特别的注意并最后确定是否真正的异常。对于异常数据的处理方式也取决于应用，并由领域专家决策。

异常数据实例

- 一个人的年龄为-999就可能是由于程序处理缺省数据设置默认值所造成的 ；
- 一个公司的高层管理人员的工资明显高于普通员工的工资可能成为异常数据，但却是合理的数据(如平安保险公司2007年5位高管税后收入超过了1000万元)；
- 一部住宅电话的话费由每月200元以内增加到数千元可能就因为被盗打或其它特殊原因所致；
- 一张信用卡出现明显的高额消费，也许是因为是盗用的卡。

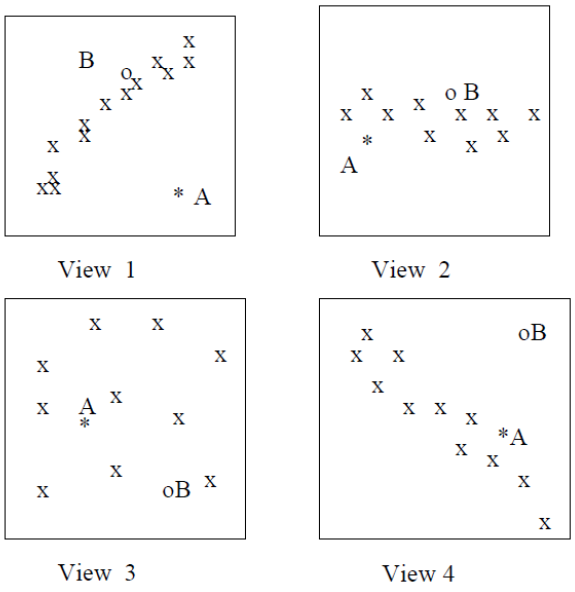




- **异常数据与众不同，但具有相对性**
  - 高与矮，疯子与常人
- **类似术语**
  - Outlier mining，Exception mining：离群点检测、离群点挖掘、例外挖掘和稀有事件挖掘

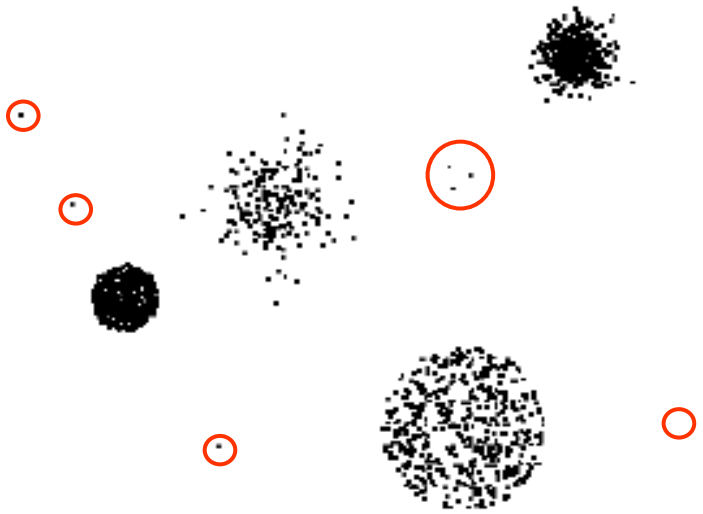
Main Problems主要问题

- 典型正常区域的定义不易
- 正常对象和离群点之间的界线不明确
- 离群点的确切概念随应用领域而异
- 训练/验证已标记数据的可用性：量少且不平衡
- 数据可能包含噪声
- 恶意对手的存在：反检测
- 正常行为不断演变
- 数据维度高

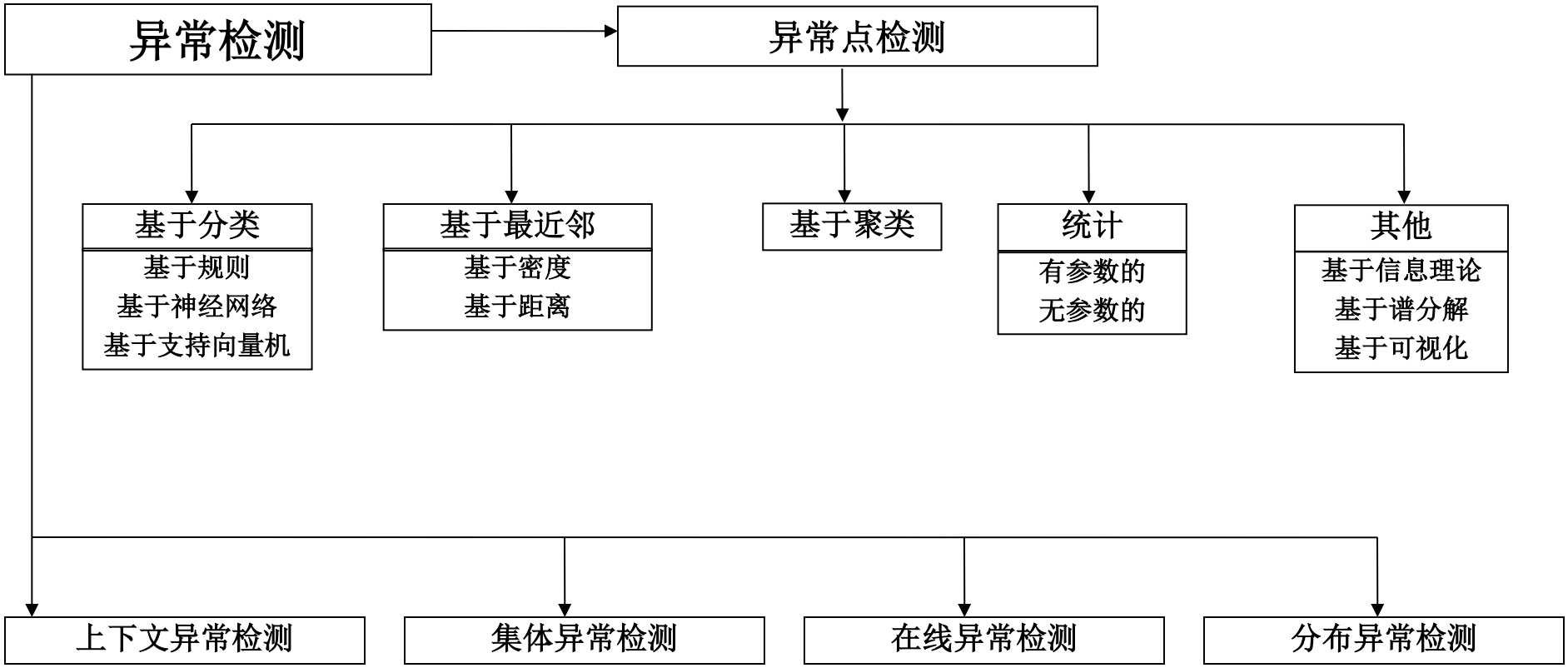


Anomaly Detection Schemes 异常检测方法

- 一般步骤
  - 构建“正常”行为的数据集
    - 数据集可以是针对数据整体的图案或者汇总统计
  - 通过使用“正常”数据集检测异常行为
    - 异常行为是特征与“正常”资料有显著差别的观察对象
- 异常检测方法的类型
  - 分类和聚类
  - 基于统计的方法
  - 基于距离和基于密度的方法
  - 基于图形的方法
  - ……



Anomaly Detection Schemes异常检测方法



3σ原则

- 3σ原则只适用服从正态分布的数据。
- 在3σ原则下，异常值被定义为观察值和平均值的偏差超过3倍标准差的值

$$P(|x-\mu|>3\sigma)\leqslant 0.003$$

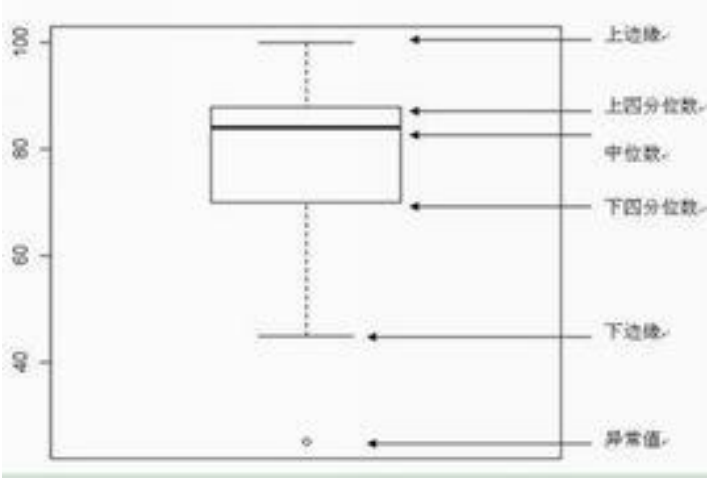
- 在正太分布假设下，大于3σ的值出现的概率小于0.003，属于小概率事件，故可认定其为异常值。

箱型图

- 3σ原则对数据分布有一定限制，而箱型图并不限制数据分布，只是直观表现出数据分布的本来面貌。
- 箱型图识别异常值的结果比较客观，而且判断标准以四分位数和四分位间距为标准，多达25%的数据可以变得任意远而不会扰动这个标准，鲁棒性更强，所以更受大家亲睐。

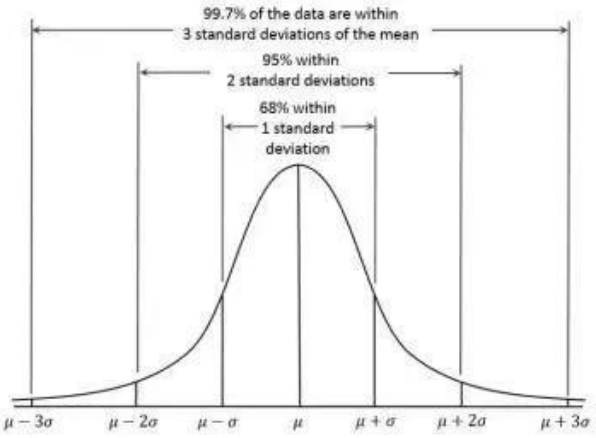
箱型图

- **箱型图识别异常值标准**：异常值被定义为满足下面两个条件之一的值
  - ✓ 大于QU+1.5IQR（或QU+3IQR）
  - ✓ 小于QL-1.5IQR（或QU-3IQR）
- QU是上四分位数，表示全部观察值中有1/4的数据比他大，QL是下四分位数，表示全部数据中有1/4的数据比他小
- IQR是四分位间距，是QU和QL的差，其间包含了观察值的一半。



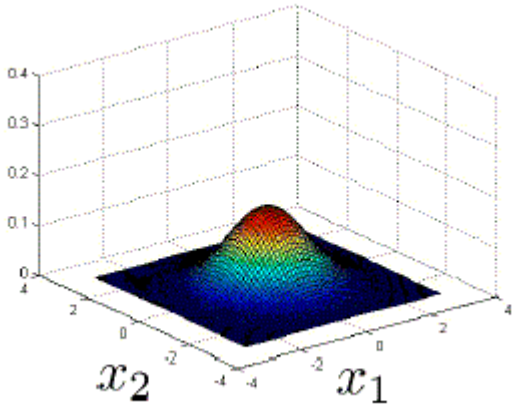
统计方法：单变量/多变量高斯分布

- 总体思想
  - 已知某种统计分布（如高斯分布）
  - 假设所有数据点都由该分布生成（如利用高斯分布的平均值和标准差），进行参数计算
  - 异常值是整体分布产生概率较低的点
- 基本假设
  - 正常数据点遵循（已知的）分布且出现在该模型的高概率区域中
  - 异常值偏离这种分布



统计方法：单变量/多变量高斯分布

- 给定一个由m个n维数据组成的训练集，将训练集转换为n为的高斯分布，通过对m个训练样例的分布分析，得出训练集的概率密度函数，即得出训练集在各个维度上的数学期望 μ 和方差 σ^2。当给定一个新的点，我们根据其在高斯分布上算出的概率p及阈值 ε 进行判断
  - 当p< ε 判定为异常
  - 当p> ε 判定为非异常



$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \qquad \sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

$$p(x) = \prod_{j=1}^{n} p(x_j;\mu_j,\sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$$

存在问题

- 问题：
  - 平均值和标准偏差是根据完整数据集（包括潜在异常值）计算的
  - 平均值和标准偏差的计算对异常值比较敏感。
  - 很难选择恰当的 ε 值。
  - 数据分布未知。

Classification-Based Techniques分类

- 主要思想
  - 基于已标记的训练数据，对正常事件（和(极少)异常事件）构建一个分类模型，以此对每一个新的未知事件进行分类
- 分类模型必须能够处理倾斜(不均衡)的类分布
- 分类
  - 二分类技术
    - 需要了解**正常**类和**异常**类
    - 建立分类，以区分正常事件和已知的异常事件
  - 一分类技术
    - 只需要了解**正常**类
    - 使用改进的分类模型学习正常行为，然后将检测到的偏离正常行为的对象作为异常行为

Classification-Based Techniques分类

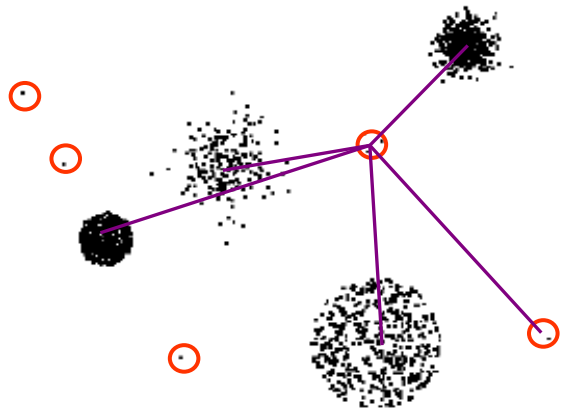
- 优点
  - 二分类技术
    - 模型很容易理解
    - 在多种已知异常对象的检测中具有高精度
  - 一分类技术
    - 模型很容易理解
    - 正常行为可以被准确学习
- 缺点
  - 二分类技术
    - 需要正常类的标记和异常类的标记
    - 不能检测未知的和新兴的异常对象
  - 一分类技术
    - 需要正常类的标记
    - 可能存在高误报率：先前未知(但合法)的数据记录可能被认为是异常的

Clustering-Based Techniques 聚类

- 关键假设
  - 正常数据记录属于大型的、密集的集群，而异常数据记录不属于任何集群或者形成极小的集群
- 按照标签分类
  - 半监督：聚集正常数据，以创建正常行为模式。如果一个新实例不属于或者不靠近任何集群，那么就是异常
  - 无监督：在聚类过程所需步骤之后，需要进行后处理来决定集群的大小，集群间的距离用来判别数据点是否异常
- 应用基于聚类的方法进行异常检测
  - 不适合任何集群的数据记录(集群残差)
  - 小集群
  - 低密度集群或局部异常(远离属于同一聚类的其他点)

Clustering-Based Techniques 聚类

- 基本思想
  - 将数据聚类划分为不同密度的簇
  - 选择小簇中的点作为候选离群点
  - 计算非候选点形成的簇和候选点间的距离
    - 如果候选点距离非候选点形成的簇较远，那么他们是离群点



Clustering-Based Techniques 聚类

- 优点
  - 不需要监督
  - 易适应在线/增量模式，适用于时空数据的异常检测
- 缺点
  - 代价极大：使用索引结构(k-d树，R*树)可能能够减轻该问题
  - 如果正常点不能创建任何簇，那么该方法可能会失败
  - 在高维空间中，数据是稀疏的，任意两个数据记录间的距离可能会非常相似：聚类算法可能不会得到有意义的簇

NN-Based Techniques 最近邻方法

- 关键假设
  - 正常点有近邻，而离群点远离其他节点
- 一般为二步法

  1. 计算每个数据记录和其邻居间的关系
  2. 分析邻居关系，以确定该数据记录异常与否
- 分类
  - 基于距离的方法
    - 离群点是远离其他节点的数据点
  - 基于密度的方法
    - 离群点是低密度区域的数据点

NN-Based Techniques 最近邻方法

- 优点
  - 可以应用于无监督或半监督环境中(对数据分布不作出任何假设)
- 缺点
  - 如果正常点没有足够数量的邻居，该方法可能会失败
  - 计算代价极大
  - 在高维空间中，数据是稀疏的，相似度的概念不能起到很大作用
    - 两个数据记录间的距离会由于稀疏而变得十分相似，以至于每个数据记录都可能被视为潜在的离群点

NN-Based Techniques 最近邻方法

- 基于距离的方法
  - 对于数据集中的点O，如果数据集中至少有p(百分比)的节点到点O的距离超过d，那么就认为O是数据集中的离群点*，记为DB(p, d)
- 基于密度的方法
  - 计算特定区域的局部密度，将低密度区域的实例报为潜在离群点
  - 方法
    - 局部离群因子(Local Outlier Factor, LOF)
    - 连接离群因子(Connectivity Outlier Factor, COF)
    - 多粒度偏差因子(Multi-Granularity Deviation Factor, MDEF)

***Knorr, Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB-98**

(1)基于距离的NN方法

- 基于距离的方法有两种不同的策略

◆ 第一种策略是采用给定邻域半径，依据点的邻域中包含的对象多少来判定异常；

  ➢ 如果一个点的邻域内包含的对象少于整个数据集的一定比例则标识它为异常，也就是将没有足够邻居的对象看成是基于距离的异常。

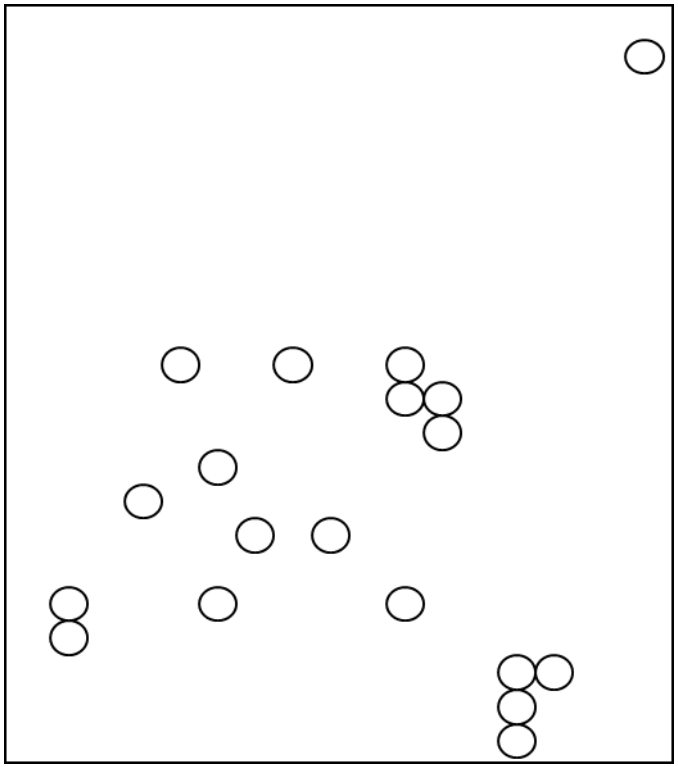
◆ 利用k最近邻距离的大小来判定异常 。

  ➢ 使用k-最近邻的距离度量一个对象是否远离大部分点，一个对象的异常程度由到它的k-最近邻的距离给定 。

  ➢ 这种方法对k的取值比较敏感。如果k太小，则少量的邻近异常点可能导致较低的异常程度。如果k太大，则点数少于k的簇中所有的对象可能都成了异常点。

到k-最近邻的距离的计算

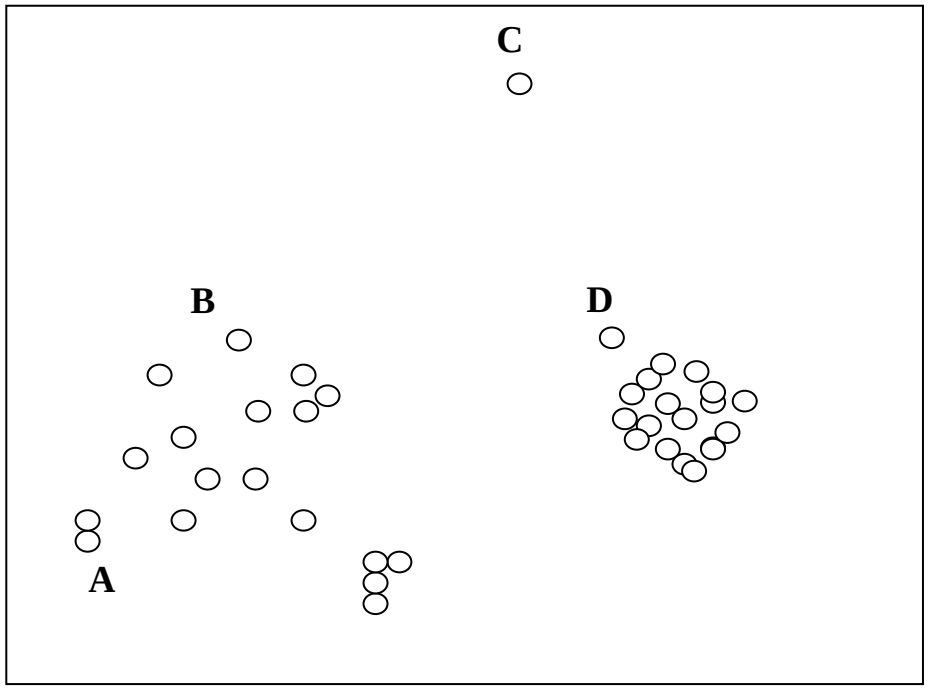
- k-最近邻的距离：
  - 一个对象的异常点得分由到它的k-最近邻的距离给定。
  - 异常点得分的最低值为0，最高值是距离函数的可能最大值----如无穷大



基于距离的异常检测的优缺点

- 优点：
  - 基于距离的异常点检测方案简单
- 缺点：
  - 时间复杂度O(m^2)，不适用于大数据集
  - 不能处理不同密度区域的数据集，因为它使用全局阈值，不能考虑这种密度的变化

不能处理不同密度区域的数据集



当k=5时，哪个点具有最高的异常点得分，B的异常点得分和D的异常点得分哪个低？

基于距离的方法无法检测出类似于 o_2 的异常值

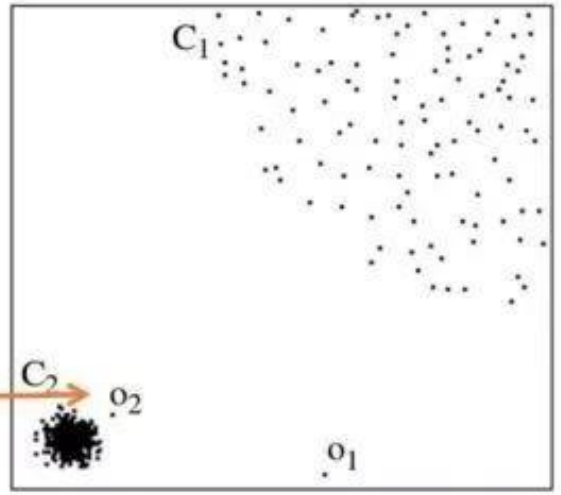
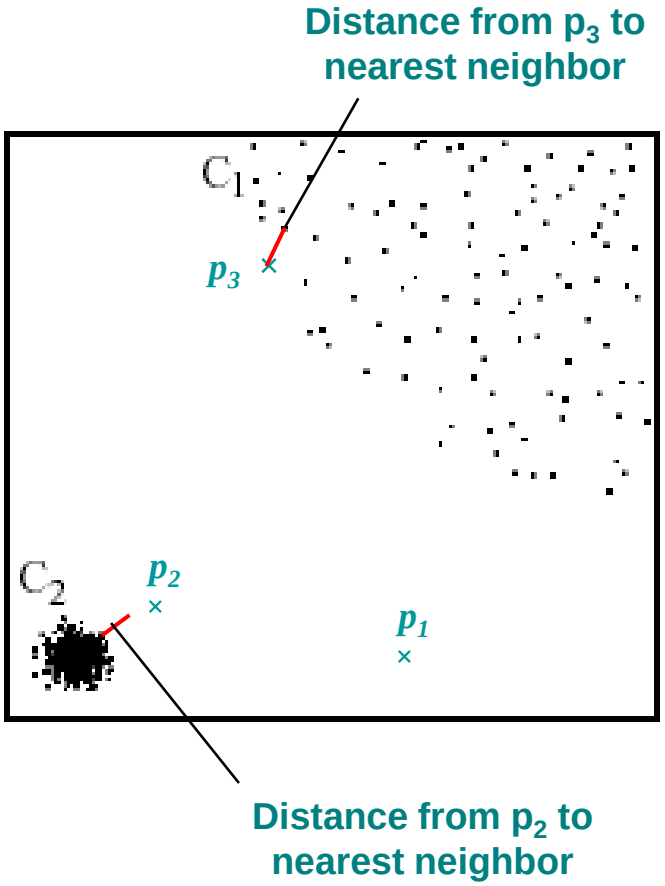


Figure 1: 2-*d* dataset DS1

(2) Local Outlier Factor(LOF)基于密度的NN方法

- 局部离群因子法(Local Outlier Factor, LOF)
  - Example:
    - 在*NN*方法中，p_2 并没有被认为是离群点，而在 *LOF* 方法中发现 p_1 和 p_2 都是离群点
    - NN方法可能认为 p_3 是离群点，但 LOF 方法不会



*** - Breunig, et al, LOF: Identifying Density-Based Local Outliers, KDD 2000.**

基于密度的方法：局部异常因子（LOF）

- **总体思想**
  - 将某一点周围的密度与其局部相邻点周围的密度进行比较
  - 该点与其邻相邻点的相对密度计为异常得分
- **基本假设**
  - 正常数据点的密度与其近邻的密度相近
  - 异常点的密度与其近邻的密度相差较大

基于密度的方法的有关概念

- 对象p的k-距离(k-distance)

  对任意的自然数k，定义p的k-距离(k-distance(p))，为p和某个对象o之间的距离，这里的o满足：

- 1) 至少存在k个对象o’∈D\{p}，使得d(p, o’) ≤ d(p, o)，并且

  2) 至多存在k-1个对象o’ ∈D\{p}，使得d(p, o’) < d(p, o)。

基于密度的方法的有关概念

- 对象p的k-距离邻域($N_{k\text{-distance}}$)

  给定p的k-距离k-distance(p)，p的k-距离邻域包含所有与p的距离不超过k-distance(p)的对象。

$$N_{k-distance}(p) = \{q \mid d(p,q) \le k-distance(p)\}$$

基于密度的方法的有关概念

- 对象p相对于对象o的可达距离

给定自然数k，对象p相对于对象o的可达距离为：

$$reach-dist_k(p,o) = \max\{k-distance(o), d(p,o)\}$$

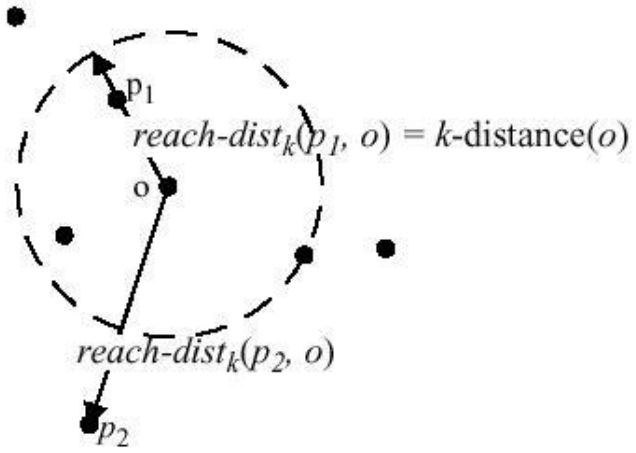


Figure 2: *reach-dist*(p_1,o) and *reach-dist*(p_2,o), for k=4

基于密度的方法的有关概念

- 对象p的局部可达密度(Local Reachable Distance)

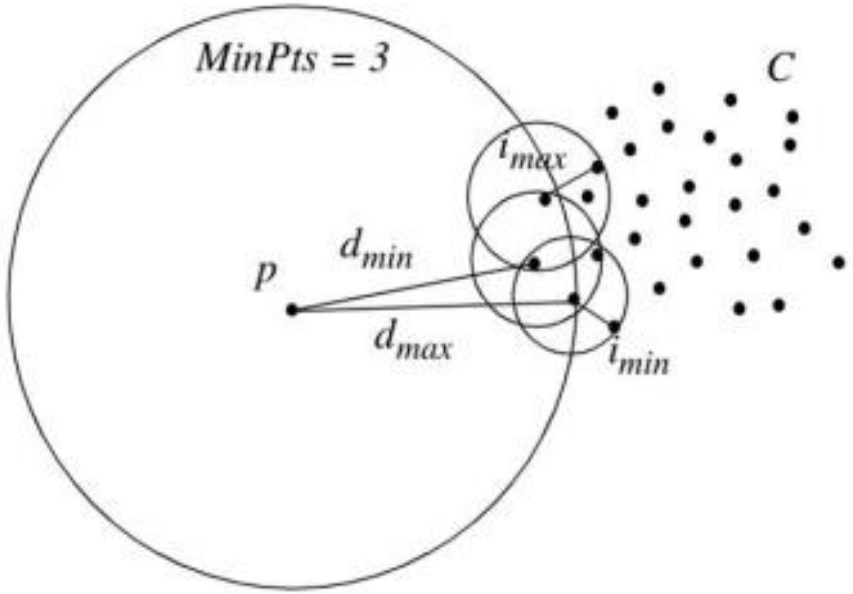
对象p的局部可达密度为对象p与它的MinPts-邻域的平均可达距离的倒数

$$lrd_{MinPts}(p) = 1 \Big/ \left(\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p,o)}{|N_{MinPts}(p)|} \right)$$

基于密度的方法的有关概念

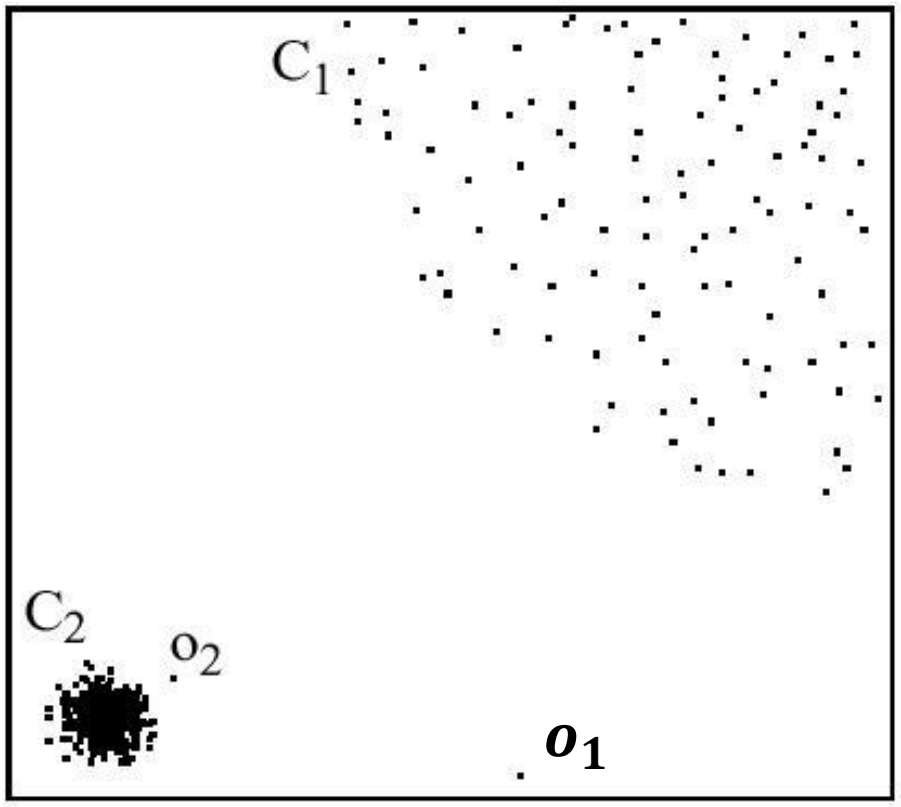
- 对象p的局部异常因子(Local Outlier Factor)

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$



局部异常的性质

- 对象p的局部异常因子表示p的异常程度，局部异常因子愈大，就认为它更可能异常；反之则可能性小。
- 簇内靠近核心点的对象的*LOF*接近于1，那么不应该被认为是局部异常。而处于簇的边缘或是簇的外面的对象的LOF相对较大，如右图中对象o_1，o_2。



局部异常因子计算

- 第一步先产生所有点的MinPts-邻域（同时得到MinPts-距离），并计算到其中每个点的距离；
  - 对低维数据，可以利用网格（Grid）来作k-NN查询，整个计算时间为 O(n)；
  - 对中维或中高维数据，必须采用索引结构如X-树等，使得作k-NN查询的时间为O(logn) ，整个计算时间为 O(n logn)；
  - 对特高维数据，索引结构不再有效，时间复杂度提高到$O(n^2)$。
- 第二步计算每个点的局部异常因子。

算法小结

- 基于统计的异常检测应用主要局限于科研计算，这主要是因为必须事先知道数据的分布特征这就限制了它的应用范围。
- 基于距离的算法跟基于统计的算法相比，不需要用户拥有任何领域知识，在概念上更加直观。更重要的是，距离异常更接近Hawkins的异常本质定义。
- 基于密度的异常观点比基于距离的异常观点更贴近Hawkins的异常定义，因此能够检测出基于距离异常算法所不能识别的一类异常数据——局部异常。局部异常观点摈弃了以前所有的异常定义中非此即彼的绝对异常观念，更加符合现实生活中的应用。