数据挖掘
第7章 异常检测

教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn

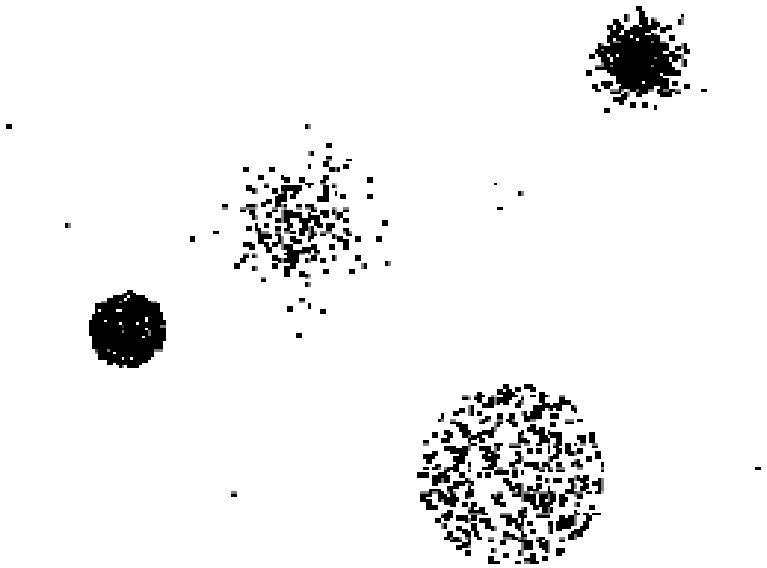
异常/孤立点检测 Anomaly/Outlier Detection

异常检测

- 异常对象：离群点（outlier）
- 偏差检测（deviation detection）、例外挖掘（exception mining）

异常对象往往是相对罕见的

- 总体规模
- 上下文



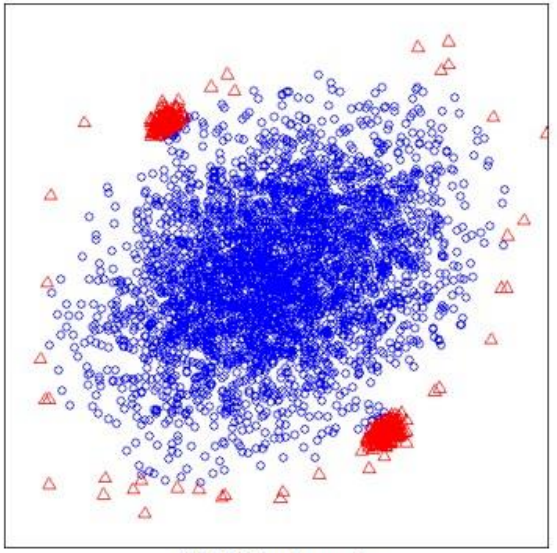
异常/孤立点检测 Anomaly/Outlier Detection

典型应用

- 欺诈检测
- 入侵检测
- 生态系统失调
- 公共卫生
- 医疗

发展

- 改进数据分析
  - 关注正常对象
  - 预处理
- 检测异常
  - 关注异常对象



(a) Original sample
(4096 instances)

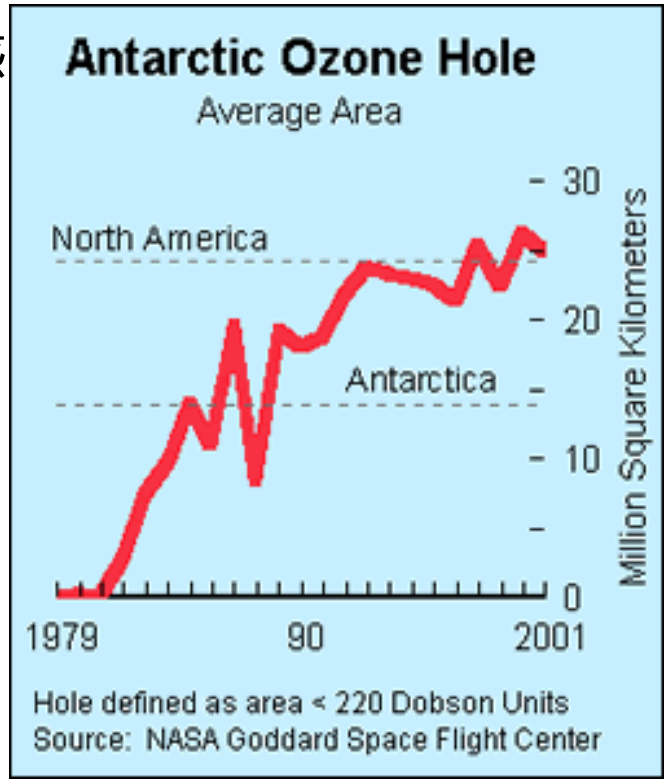
重要案例 Importance of Anomaly Detection

臭氧消耗历史 Ozone Depletion History

1985年，三位研究人员（Farman，Gardinar和 Shanklin）对英国南极调查所收集的数据感到困惑，该数据显示，南极洲的臭氧水平比正常水平下降了10%。

为什么配备有用于记录臭氧水平的仪器的Nimbus 7卫星却没有记录同样低的臭氧浓度?

卫星记录的臭氧浓度非常低，以至于被计算机程序当作异常值（outliers）丢弃了!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

异常的成因 Causes of Anomalies

数据来源于不同的类 Data from different classes

- Measuring the weights of oranges, but a few grapefruit are mixed in

自然变异 Natural variation

- Unusually tall people

数据测量和收集误差 Data errors

- 200 pound 2 year old

噪声与异常点的不同
Distinction Between Noise and Anomalies

噪声是错误的值（erroneous），可能是随机值（random value）或污染对象（contaminating objects）

- 重量记录错误或不同对象（例如者葡萄柚与橙子的混合）

噪声不一定会产生异常值或异常对象

噪音是无趣的

如果不是噪声引起的异常可能会很有趣

噪声（Noise）和异常（anomalies）是相关但截然不同的概念

异常检测技术 Anomaly Detection Techniques

基于模型的技术

- 建立数据模型
- 例如：聚类和回归模型

基于邻近度的技术

- 距离/相似度
- 例如二维或者三维空间

基于密度的技术

- 密度估计：低密度区域
- 数据集具有不同密度区域

基于模型的异常检测
Model-Based Anomaly Detection

建立数据模型并检查

- 无监督（Unsupervised）
  - 异常点（Anomalies）是对模型的拟合能力较差的数据点
  - 异常是扰乱模型（that distort the model）的数据点
  - 例子:
    - 统计分布（Statistical distribution）
    - 聚类（Clusters）
    - 回归（Regression）
    - 几何（Geometric）
    - 图（Graph）
- 监督（Supervised）
  - 异常点被视为稀有类别（rare class）
  - 需要训练数据（带有标签）
- 半监督

问题变体：
Variants of Anomaly Detection Problems

阈值

- Given a data set D, find all data points **x** $\in$ D with anomaly scores greater than some threshold t

Top-n

- Given a data set D, find all data points **x** $\in$ D having the top-n largest anomaly scores

分数计算

- Given a data set D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

问题：属性数目
General Issues: Number of Attributes

许多异常（anomalies）是根据单个属性（attribute）定义的

- 高度
- 形状
- 颜色

使用所有属性可能很难找到异常

- 噪声或无关（irrelevant）的属性
- 对象只在某些属性上是异常的

但是，一个异常对象可能在任何单个属性中都不是异常的

问题：异常程度/分数/得分
General Issues: Anomaly Scoring

许多异常检测技术仅提供二分类（binary categorization）

- 对象是异常（anomaly）还是非异常
- 基于分类的异常检测方法（classification-based approaches）尤其如此

其他一些异常检测方法为所有数据点分配分数（score）

- 该分数衡量对象异常的程度或可能性
- 这样可以对对象进行排名

最后，通常需要一个二元决策（binary decision）

- 是否应该标记此信用卡交易?
- 分数仍然有用

有多少异常（anomalies）？

其他问题 Other Issues for Anomaly Detection

一次只识别一个异常或一次识别所有异常（all anomalies）

- 一次一个：屏蔽（Swamping）问题
- 一次多个：泥潭（Masking）问题

评估（Evaluation）

- 如何衡量异常检测效果（measure performance）？
- 有监督与无监督（Supervised vs. unsupervised）的情况

效率（Efficiency）

全局与局部视角：上下文（Context）

- 篮球队

方法介绍

基于统计学的异常检测方法

基于邻近度的异常检测方法

基于密度的异常检测方法

基于簇的异常检测方法

统计方法

基于模型

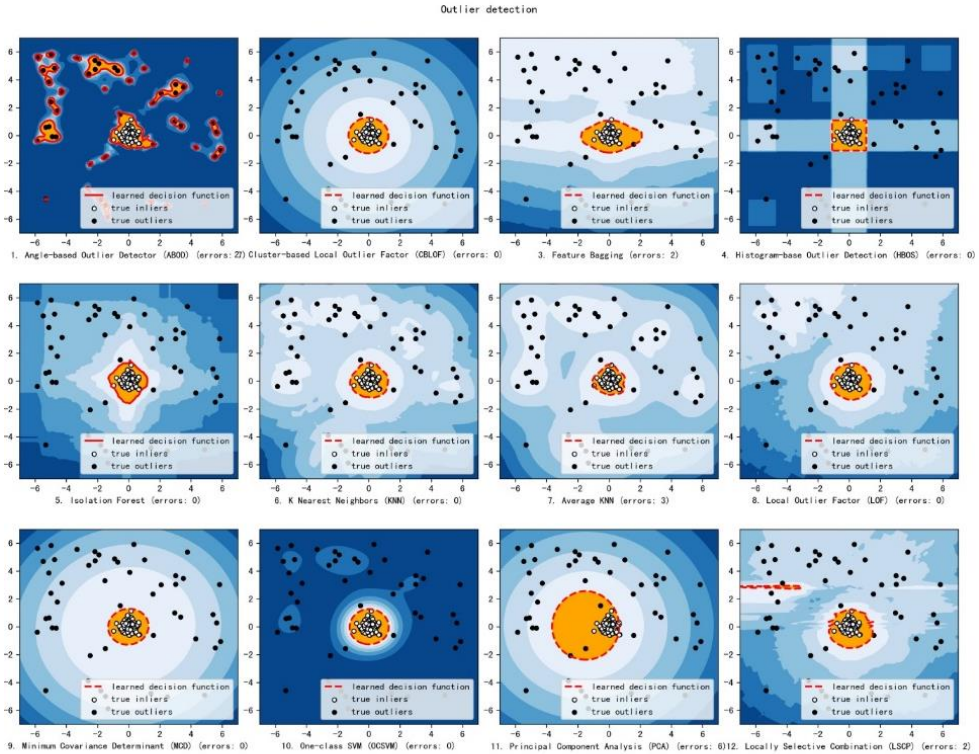
- 概率分布模型

离群点

- 概率低的对象

统计方法的使用依赖于

- 数据分布（Data distribution）
- 模型参数（均值，方差等）
- 期望异常点个数（阈值）



统计方法

基于模型

- 概率分布模型

离群点

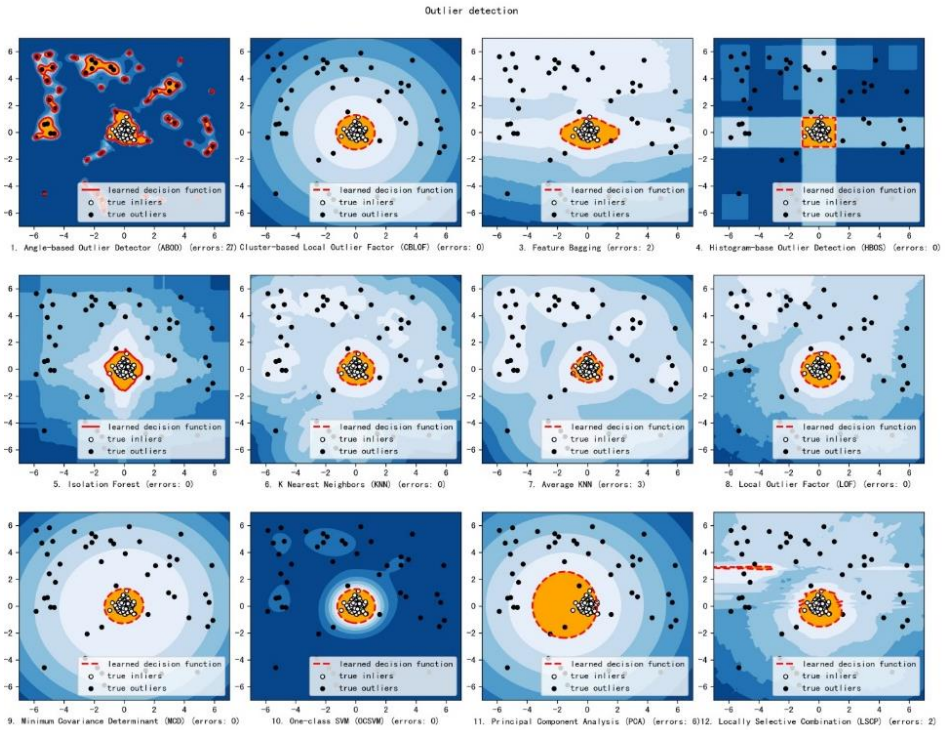
- 概率低的对象

统计方法的使用依赖于

- 数据分布（Data distribution）
- 模型参数（均值，方差等）
- 期望异常点个数（阈值）

问题

- 识别数据集的具体分布
- 使用的属性个数
- 混合分布?

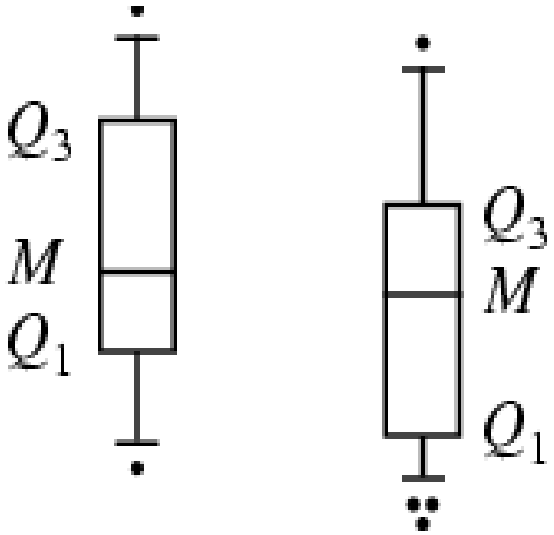


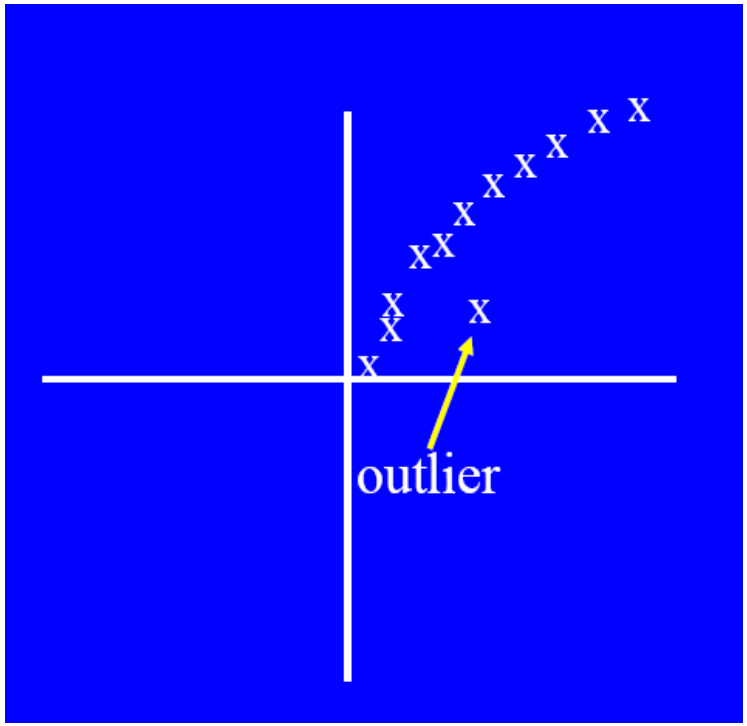
可视化方法Visual Approaches

箱型图或者散点图 Boxplots or scatter plots

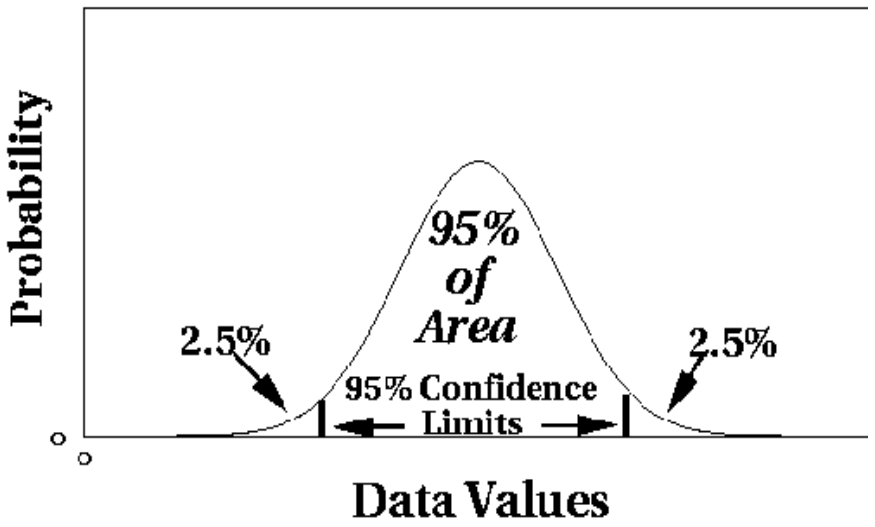
缺点

- 非自动化（not automatic）
- 主观（Subjective）





高斯（正态）分布 Normal Distributions



一维高斯分布

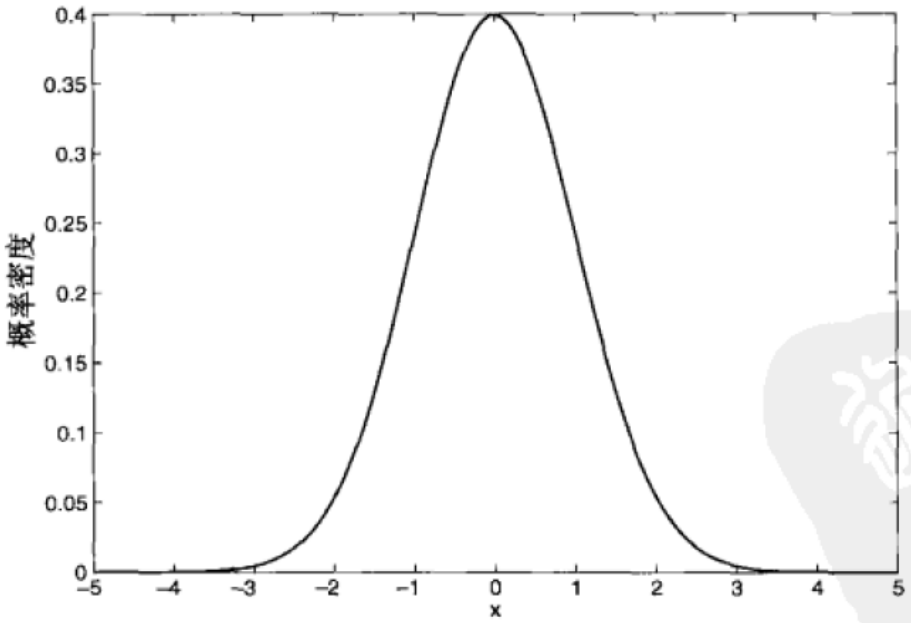
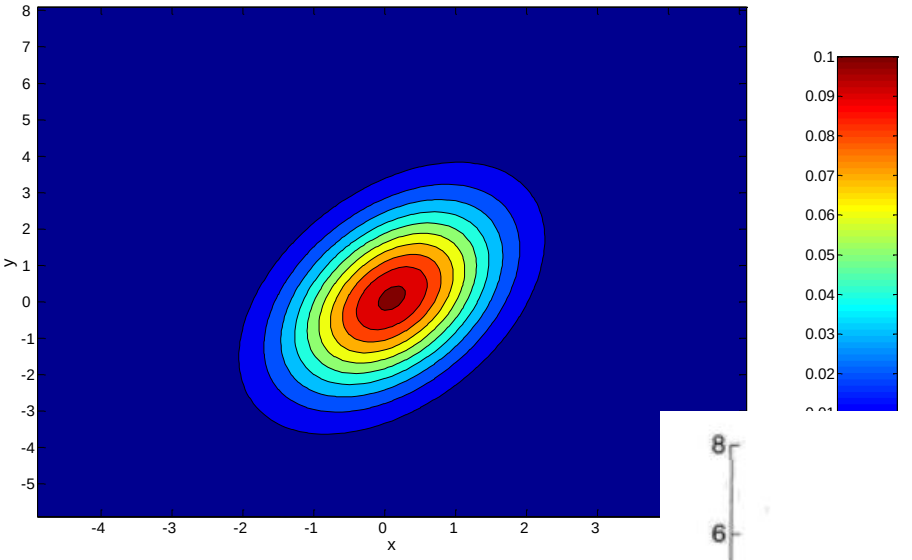


图 10-1　均值为 0，标准差为 1 的高斯分布的概率密度函数

表 10-1　均值为 0，标准差为 1 的高斯分布的样本对(c, α)，$\alpha = prob(|x| \geq c)$

c	$N(0,1)$的α
1.00	0.3173
1.50	0.1336
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

高斯（正太）分布 Normal Distributions



$$mahalanobis(\mathbf{x}, \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})^T$$

二维高斯分布

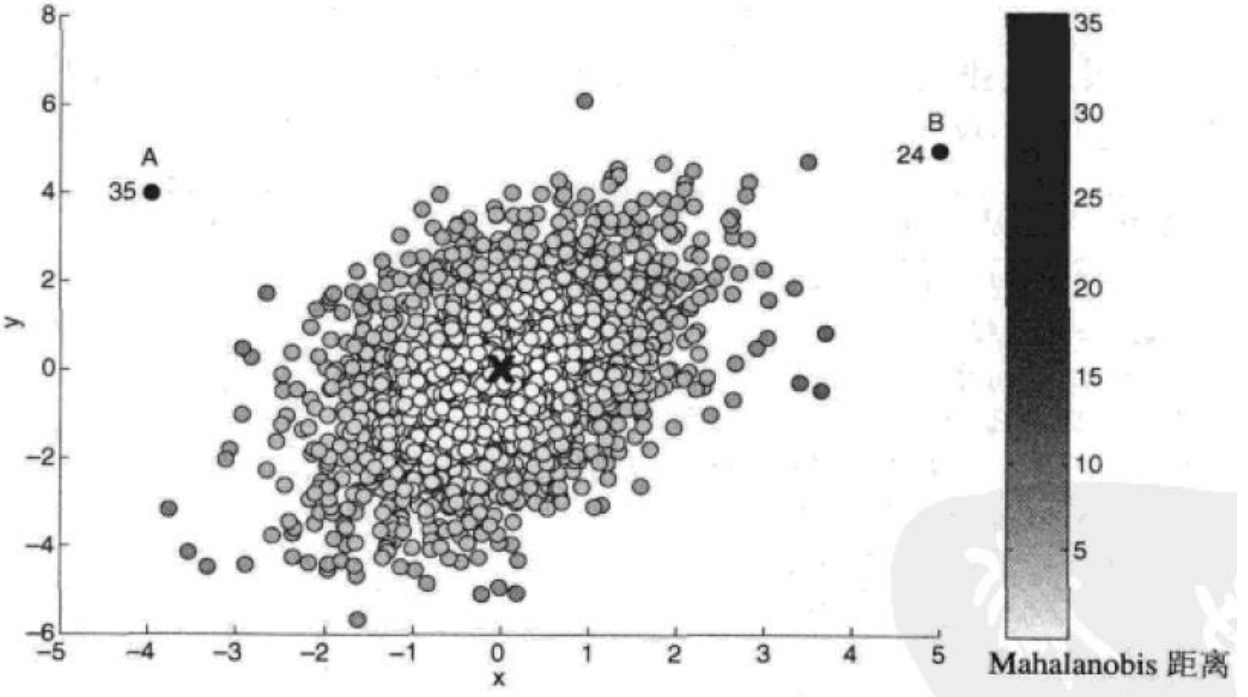


图 10-3　2 002 个点的二维数据集中点到中心的 Mahalanobis 距离

格拉布斯检验法 Grubbs' Test（自学）

Detect outliers in univariate data

Assume data comes from normal distribution

Detects one outlier at a time, remove the outlier, and repeat

- H_0: There is no outlier in data
- H_A: There is at least one outlier

Grubbs' test statistic:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

Statistical-based – Likelihood Approach（自学）

Assume the data set D contains samples from a mixture of two probability distributions:

- M (majority distribution)
- A (anomalous distribution)

General Approach:

- Initially, assume all the data points belong to M
- Let $L_t(D)$ be the log likelihood of D at time t
- For each point x_t that belongs to M, move it to A
  - Let L_{t+1} (D) be the new log likelihood.
  - Compute the difference, $\Delta = L_t(D) - L_{t+1}$ (D)
  - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach（自学）

Data distribution, D = (1 – λ) M + λ A

M is a probability distribution estimated from data

- Can be based on any modeling method (naïve Bayes, maximum entropy, etc)

A is initially assumed to be uniform distribution

Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

优劣势分析
Strengths/Weaknesses of Statistical Approaches

坚实的统计/数学理论基础 Firm mathematical foundation

高效 Can be very efficient

如果分布已知则非常有效 Good results if distribution is known

分布未知则效果差 In many cases, data distribution may not be known

高维数据难以估计分布For high dimensional data, it may be difficult to estimate the true distribution

异常点可能会扰乱分布的参数Anomalies can distort the parameters of the distribution

基于邻近度/距离的方法 Distance-Based Approaches

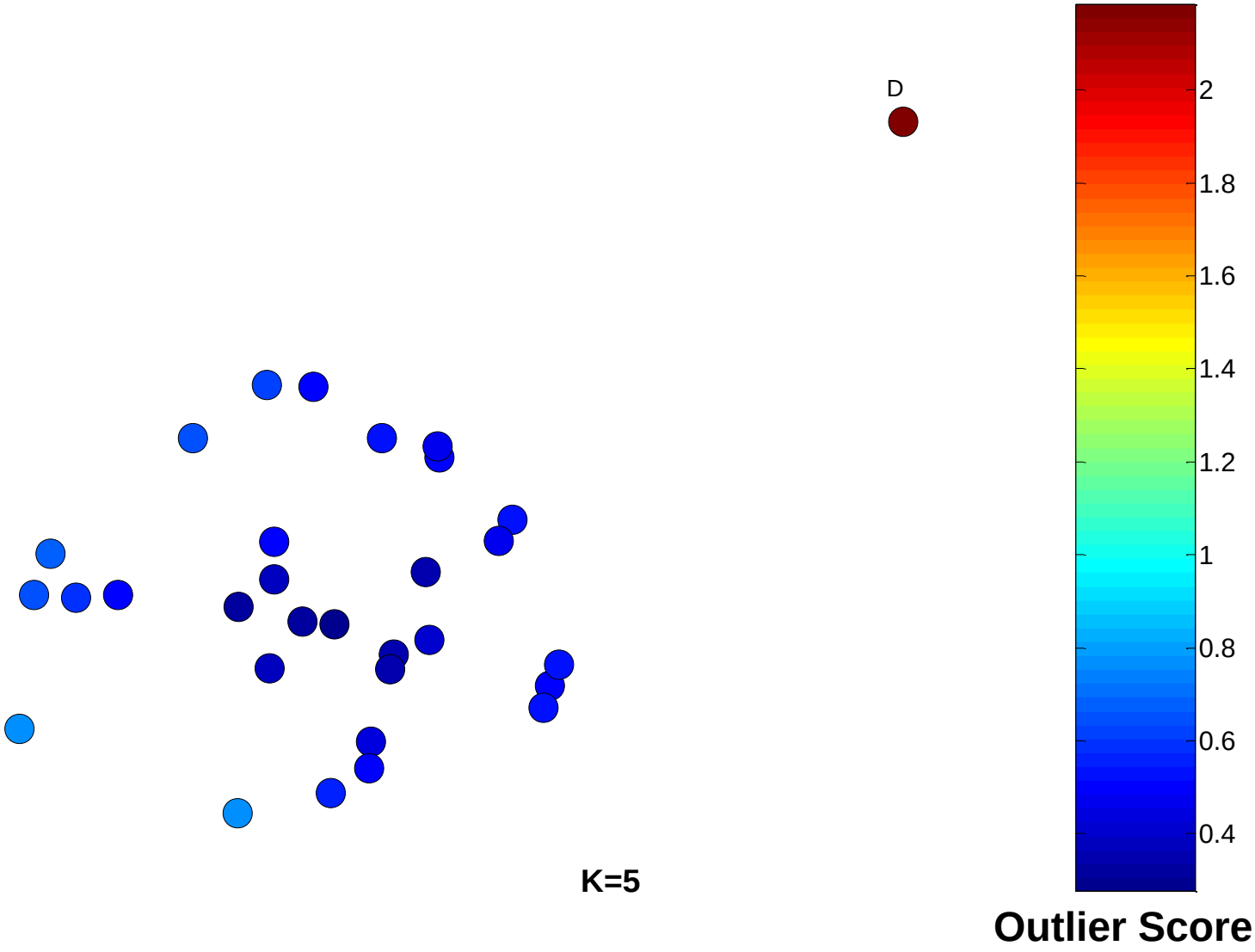
多种方法变体 Several different techniques

思想

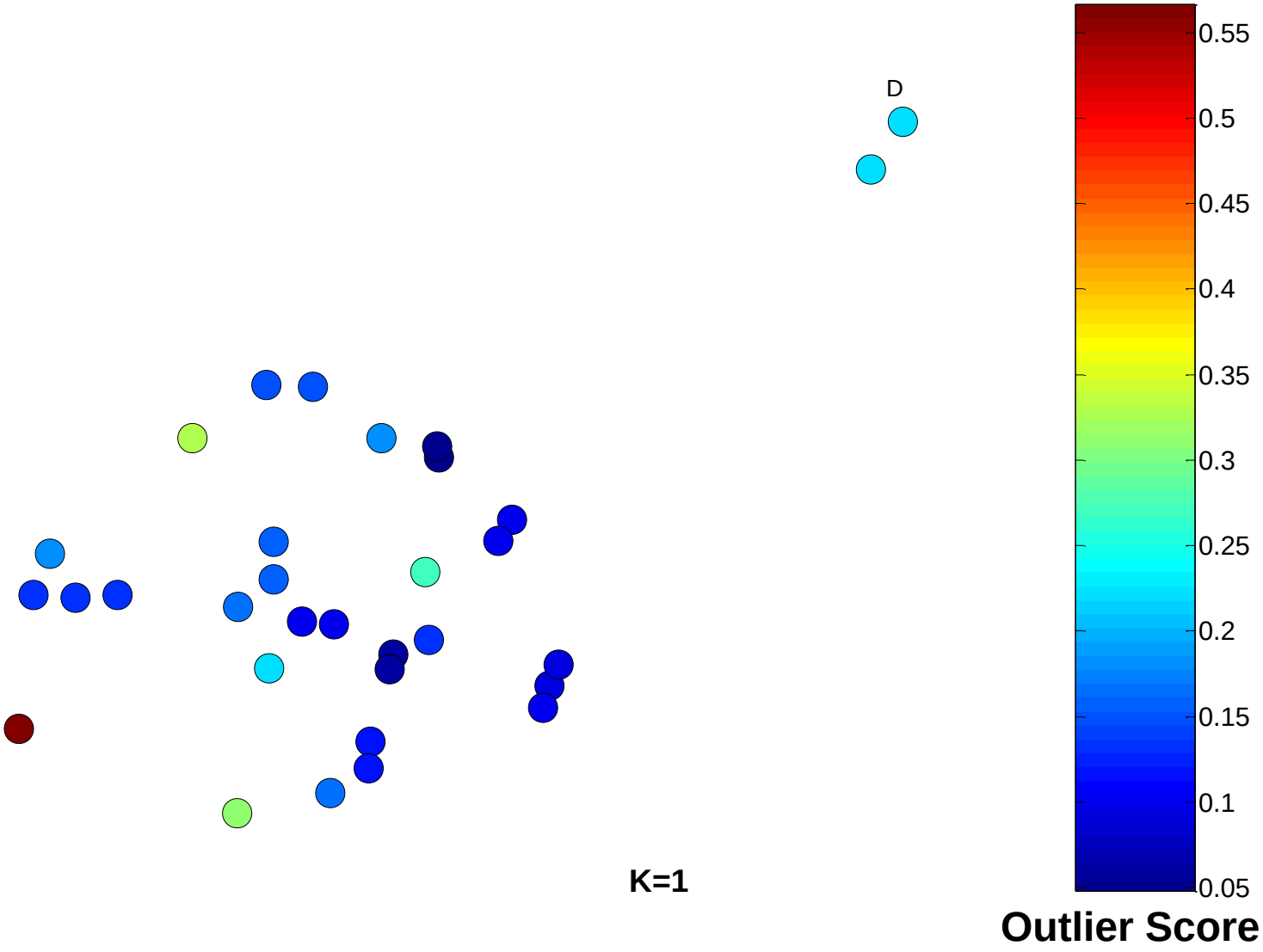
- 一个对象是异常的，如果它远离大部分点。An object is an outlier if a specified fraction of the objects is more than a specified distance away (Knorr, Ng 1998)
  - Some statistical definitions are special cases of this

度量一个对象是否远离大部分点的一种最简单的方法是使用到k最近邻的距离

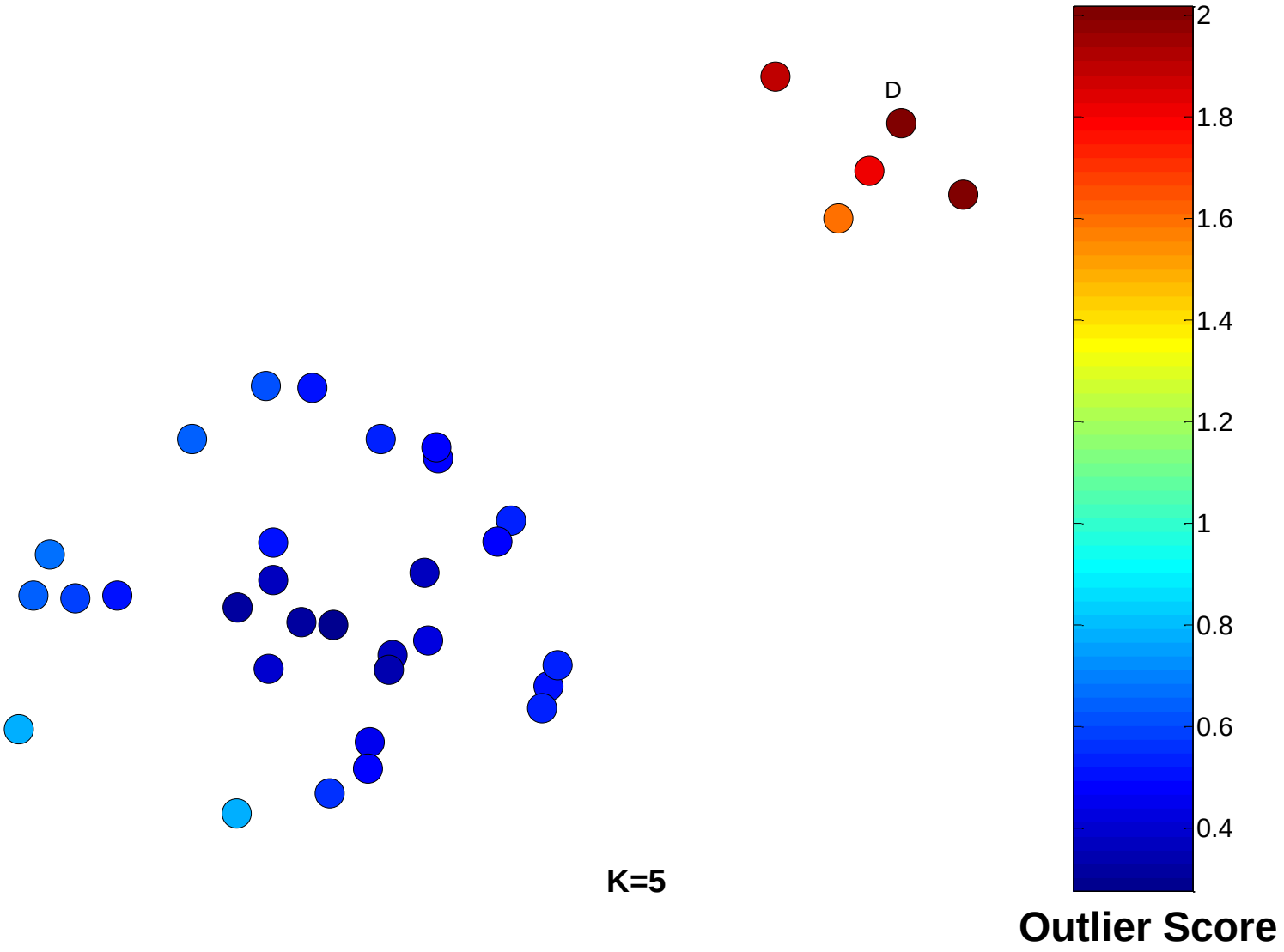
Five Nearest Neighbor - One Outliers



One Nearest Neighbor - Two Outliers



Five Nearest Neighbors - Small Cluster



优劣势分析
Strengths/Weaknesses of Distance-Based Approaches

简洁 Simple

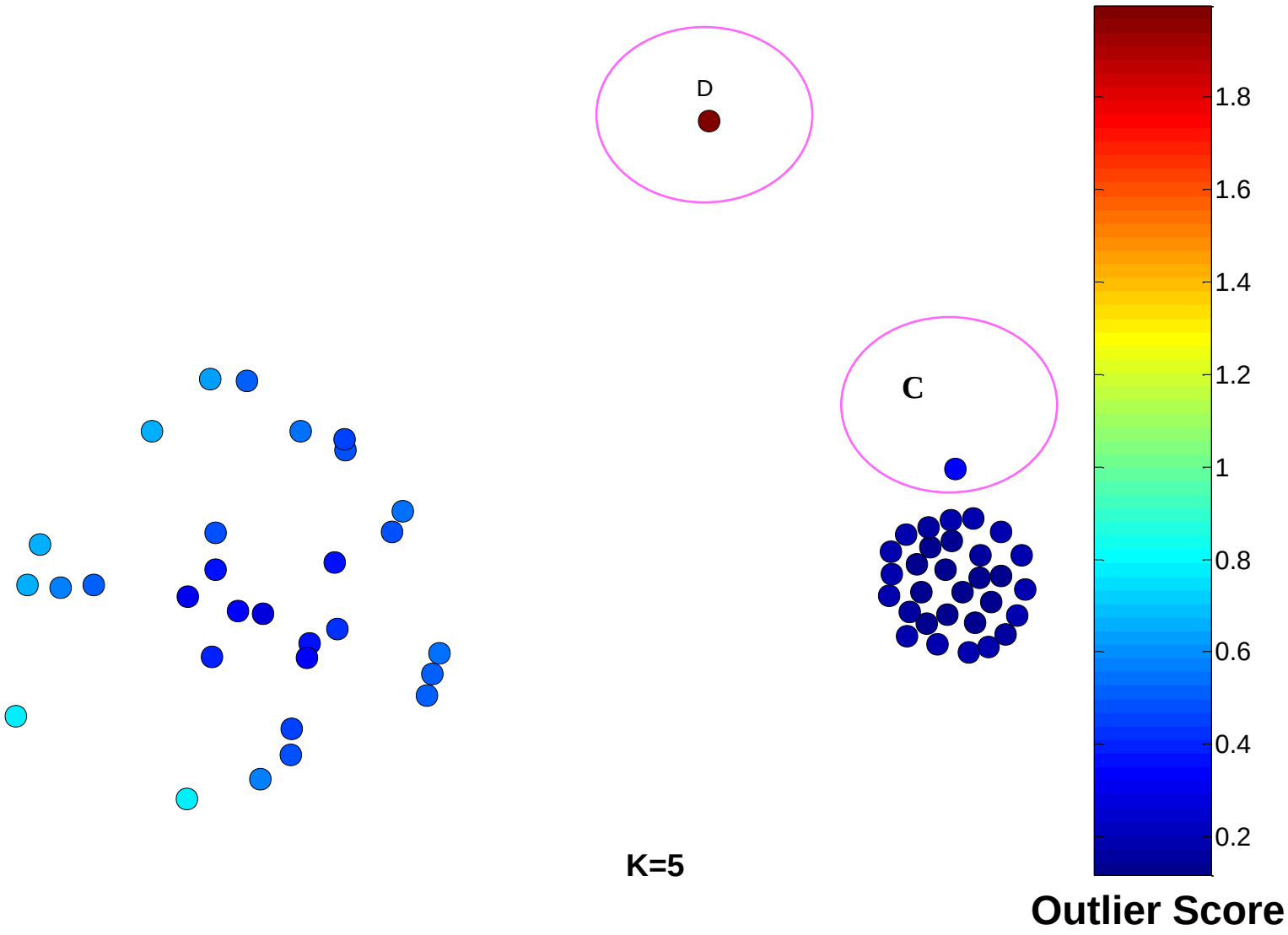
代价/复杂度高 Expensive – $O(n^2)$

参数敏感 Sensitive to parameters

密度敏感 Sensitive to variations in density

高维数据中距离意义不大 Distance becomes less meaningful in high-dimensional space

Five Nearest Neighbors - Differing Density



基于密度的方法 Density-Based Approaches

基于密度的离群点（Density-based Outlier）:一个对象的离群点得分是该对象周围密度的逆。

- K近邻
  - 到第k个邻居的距离的逆：Inverse of distance to kth neighbor
  - 到k个邻居的平均距离的逆：Inverse of the average distance to k neighbors
- DBSCAN definition

上述定义无法同时处理具有不同密度的区域

- If there are regions of different density, this approach can have problems

Relative Density

考虑一个点的密度相对于其k近邻的密度

- Consider the density of a point relative to that of its k nearest neighbors

$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

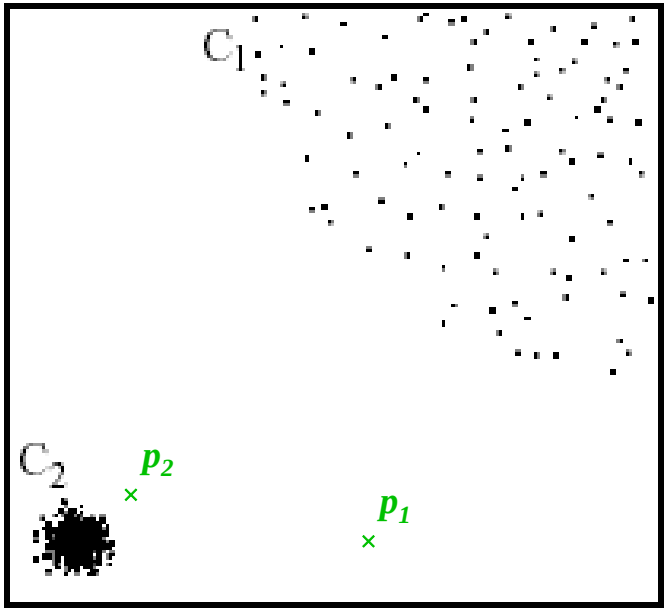
1: {k is the number of nearest neighbors}
2: **for all** objects $\mathbf{x}$ **do**
3: Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
4: Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
5: **end for**
6: **for all** objects $\mathbf{x}$ **do**
7: Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation 10.7.
8: **end for**

Density-based: LOF approach（自学）

For each point, compute the density of its local neighborhood

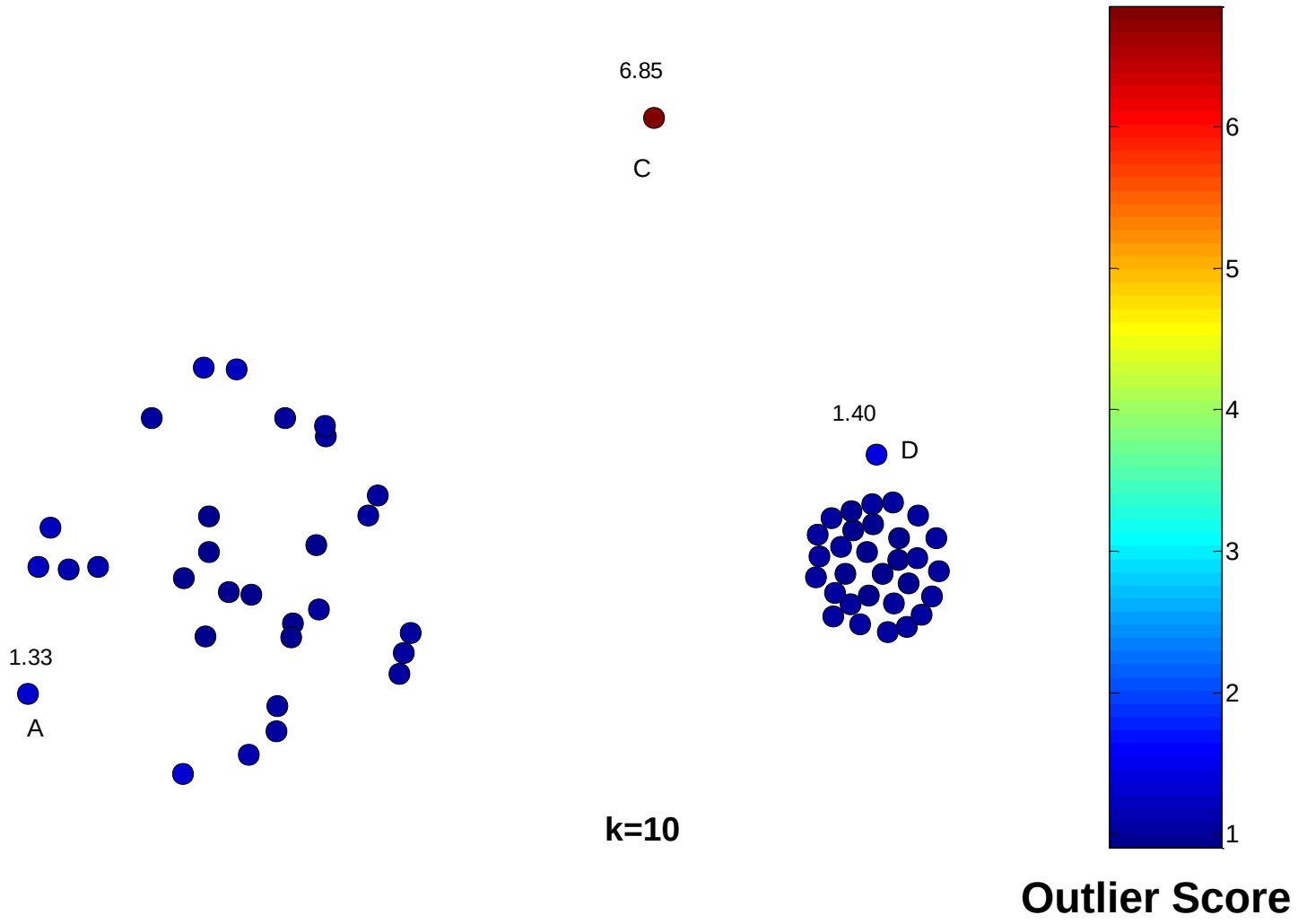
Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors

Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Relative Density Outlier Scores



Strengths/Weaknesses of Density-Based Approaches

简洁 Simple

复杂度高 Expensive – $O(n^2)$

参数敏感 Sensitive to parameters

高维空间问题

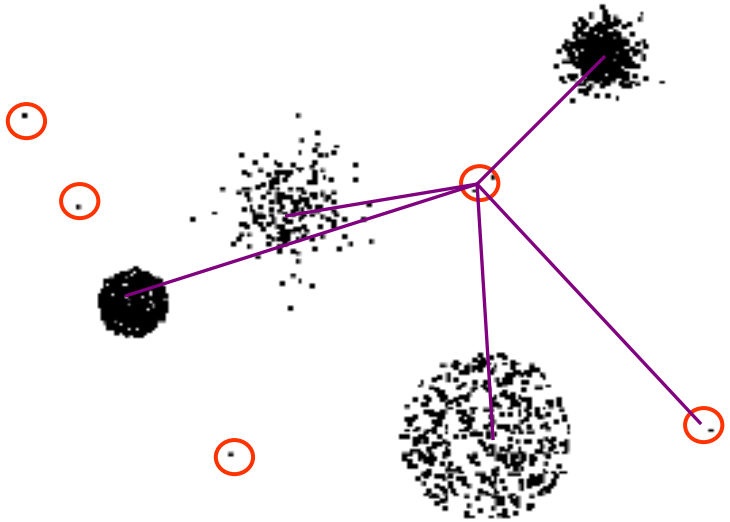
- Density becomes less meaningful in high-dimensional space

基于聚类的技术 Clustering-Based Approaches

聚类分析发现强相关的对象组，而异常检测发现不与其他对象强相关的对象。

方法1：丢弃远离其他簇的小簇

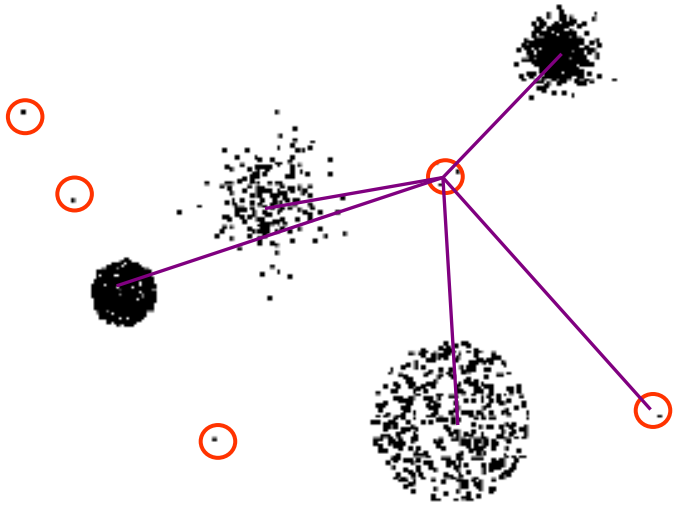
方法2：聚类所有对象，然后评估对象属于簇的程度



基于聚类的技术 Clustering-Based Approaches

基于聚类的离群点（Clustering-based Outlier）：An object is a cluster-based outlier if it does not strongly belong to any cluster

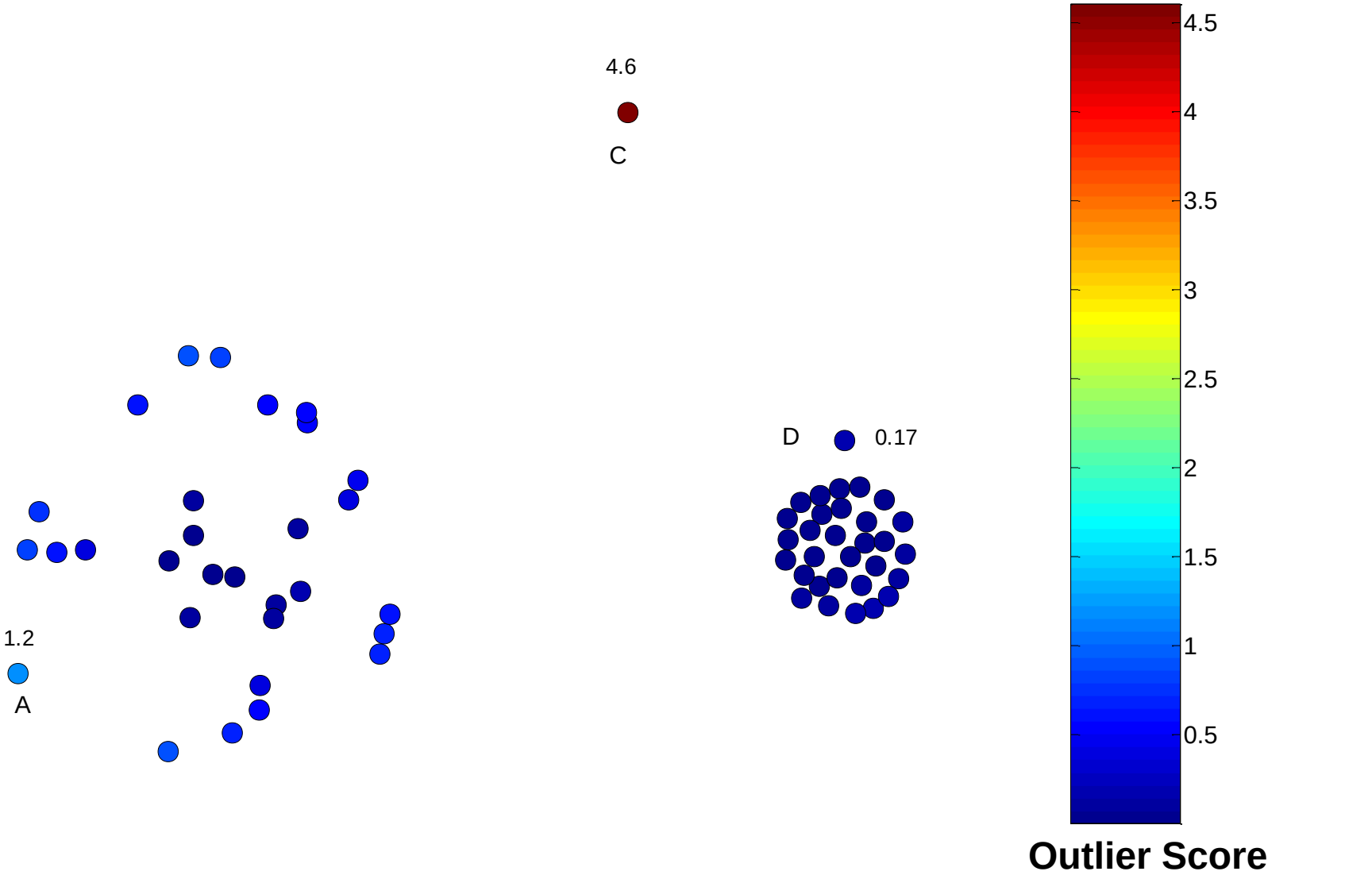
- 基于原型：For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
- 基于密度For density-based clusters, an object is an outlier if its density is too low
- 基于图For graph-based clusters, an object is an outlier if it is not well connected



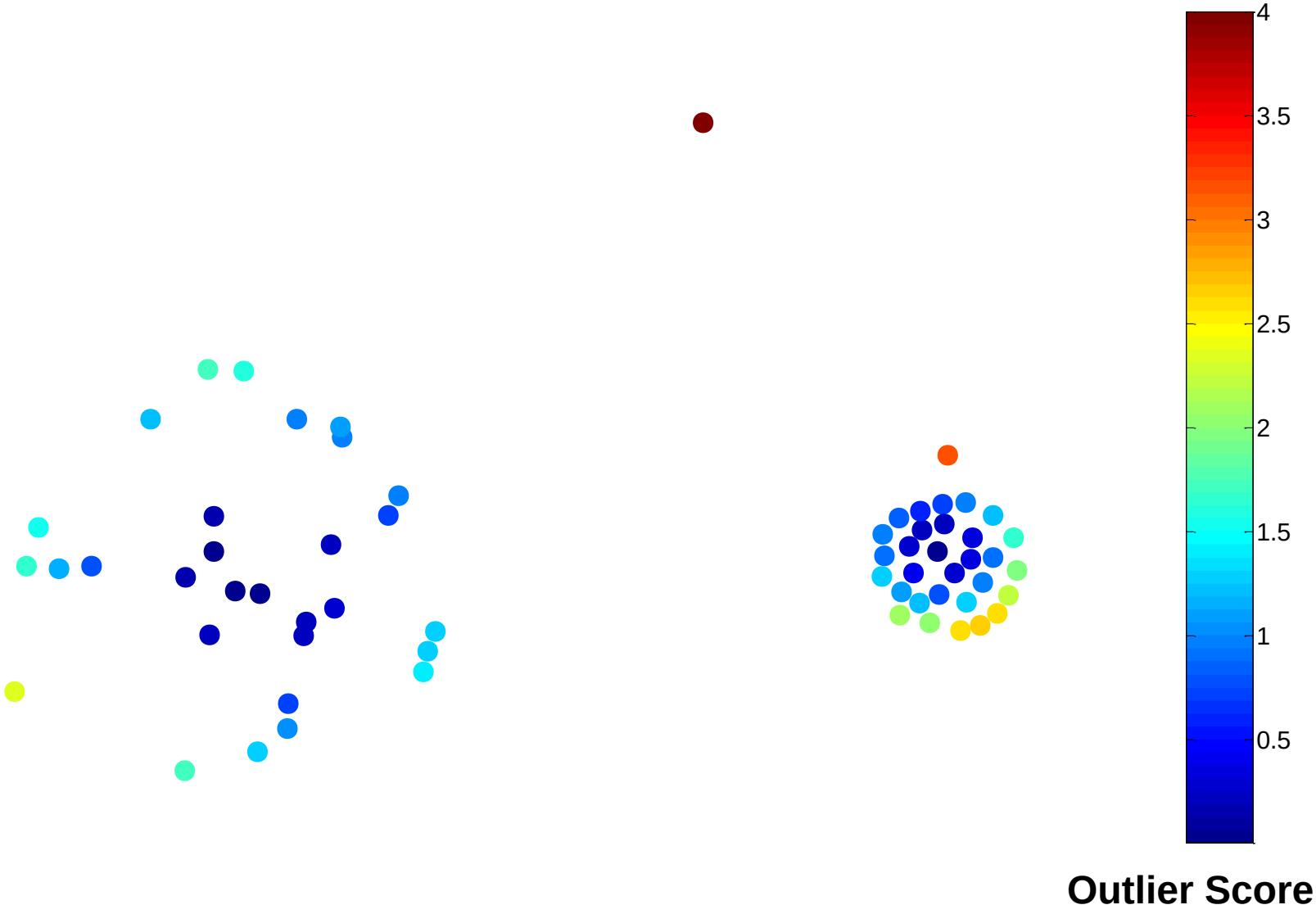
其它问题：

- 异常值的影响
- 簇的个数的影响

Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



一些问题

离群点对初始聚类的影响

- 结果是否有效
- 对象聚类，删除离群点，对象再次聚类
- 没有最优解的保证

一些问题

使用簇的个数

- 对象是否被认为是离群点可能依赖于簇的个数
- 对不同的簇个数重复该分析或者找出大量小簇

Strengths/Weaknesses of Distance-Based Approaches

简洁 Simple

多种可用的聚类技术 Many clustering techniques can be used

使用哪种聚类技术 Can be difficult to decide on a clustering technique

簇的个数的确定Can be difficult to decide on number of clusters

异常点会影响簇 Outliers can distort the clusters

谢谢！

数据挖掘
教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn