数 据 挖 掘

Data Mining

主讲: 张仲楠 教授

厦門大學
XIAMEN UNIVERSITY

异常检测

目 录

1. 概述

1. 基本概念

- 异常检测的目标是发现一些不符合正常规律或行为的对象。
- 异常对象被称作离群点(outlier)，因为在数据的散点图中，它们的分布远离其他的数据点。
- 异常检测也称为偏差检测(deviation detection)，因为异常对象的某些属性值明显偏离该属性的期望值或典型的属性值。
- 异常检测也称为例外挖掘(exception mining)，因为异常在某种意义上是例外的。

1. 概述

2. 应用场景

- 欺诈检测
  - 通过寻找窃贼的购买模式或注意一些不同于常见行为的变化来检测窃贼。
- 入侵检测
  - 通过监视系统和网络的异常行为来检测入侵。
- 医疗和公共卫生
  - 不寻常的症状或检测结果可能预示着潜在的健康问题。
- 航空安全
  - 从飞机传感器记录中鉴定异常事件。

目 录

2.异常检测问题的特性

1. 异常的定义

- **定义9.1**：异常是指不符合正常实例分布的观测，比如与某种分布下大多数实例不相似。
- 这个定义没有假设这种分布是可以用已知统计分布术语轻易表达的。事实上，这种困难正是许多异常检测方法使用非统计学方法的原因。
- 可以按照看到一个对象的概率或某种更极端的方法来对数据对象排序。出现的概率越低，这个对象越可能是一个异常数据。
- 造成异常现象的原因有很多：噪声，对象来自不同的分布，对象仅仅是这个分布中极少出现的。

2.异常检测问题的特性

2. 数据的性质

- 输入数据的性质在决定选择合适的异常检测技术中起着重要的作用。
- 输入数据的一些共有特性包括属性的数量和类型，以及描述每个数据实例的表示方式。
- **单变量或多变量**
  - 如果数据包含单一属性，一个对象是否异常的问题只依赖于这个对象的属性值是否是异常的。
  - 如果一个数据对象使用多个属性表示，可能在某些属性上有异常的数值，但在其他的属性上表现正常。
  - 一个对象即使在每个单独属性上的数值都不是异常的，整体而言它也可能是异常对象。

2.异常检测问题的特性

2. 数据的性质

- **记录数据或邻近度矩阵**
  - 表示一个数据集合最常见的方式是使用记录数据或它的变体，比如数据矩阵，在矩阵中每个数据实例都使用相同的属性进行描述。
  - 一些异常检测方法使用邻近度矩阵(proximity matrix)，矩阵中的每一个数值表示两个实例间的接近程度(相似或不相似)。

2.异常检测问题的特性

2. 数据的性质

■标签的可用性

- ■ 一个数据实例的标签表明这个实例是正常的还是异常的。
- ■ 如果有一个每个数据实例都带有标签的训练数据集，那么异常检测的问题就可以转化为监督学习(分类)问题。
- ■ 因为异常数据非常稀少，获得异常类的类标是非常具有挑战性的。
- ■ 本质上大多数异常检测问题是无监督的。
- ■ 给定一个输入数据集，从正常实例中区分出异常实例是非常具有挑战性的。

2.异常检测问题的特性

2. 数据的性质 --- 异常的两个主要性质

- 数量相对较少
  - 因为异常数据是不常见的，所以大部分输入数据中正常的实例占了绝大多数。
  - 在大多数异常检测技术中，常将输入数据集用作正常类的不完美表示。
  - 一些异常检测方法提供了指定输入数据中异常值期望数量的机制，这类方法可以处理具有大量异常的数据。
- 稀疏分布
  - 与正常的对象不同，异常数据通常和其他实例不相关，因此在属性空间中分布稀疏。
  - 有一些异常检测方法专门用来寻找聚类异常，通常假设这些异常数据的数量很小或者离其他实例的距离很远。

2.异常检测问题的特性

3. 如何使用异常检测

- 任何通用的异常检测方法都有两种不同的方法可以使用。
- 在第一种方法中，给定的输入数据中既包含正常实例又包含异常实例，我们需要在这些数据中找到异常实例。本章中所有的异常检测方法都可以在这种设定中使用。
- 在第二种方法中，我们还获得了需要识别为异常的测试实例（每次出现一个）。大多数异常检测方法（除了少数例外）都能够使用输入数据集来提供新测试实例的输出。

目 录

3.统计方法

1. 基本概念

- 统计方法使用概率分布(比如，高斯分布)对正常类进行建模。
- 这些分布的一个重要特征是它们把每一个数据实例和一个概率值进行关联，表示这个实例从分布中生成的可能性有多大。
- 异常数据被认为是那些不太可能从正常类的概率分布中生成的实例。
- 有两种类型的模型可以用来表示正常类的概率分布:
  - 参数模型使用那些熟知的统计分布族，这些分布需要从数据中进行参数估计。
  - 非参数模型则非常灵活，并且直接从得到的数据中学习正常类的分布。

3.统计方法

2. 使用参数模型

- 一些常用的参数模型被广泛用于描述许多类型的数据集，这些模型包括高斯分布、泊松分布和二项分布。其中涉及的参数需要从数据中学习，例如，高斯模型需要从数据中确定均值和方差这两个参数。
- 参数模型能够非常有效地表示正常类的行为，尤其当知道正常类服从某个特定分布时。通过参数模型计算的异常分数具有较强的理论性质，可用于分析异常分数并评估其统计显著性。
- 将在一元和多元环境下讨论用高斯分布对正常类进行建模。

3.统计方法

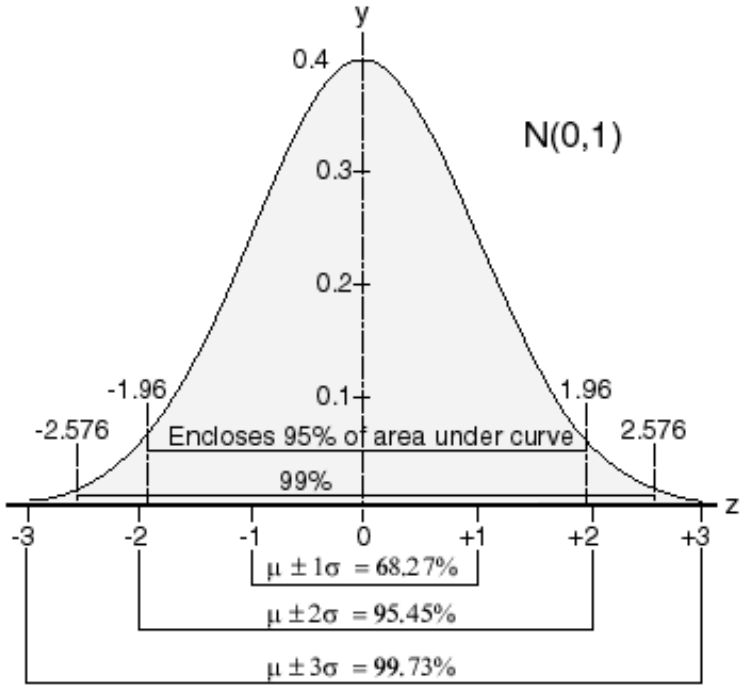
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu)^2$$

2. 使用参数模型 --- 一元高斯分布

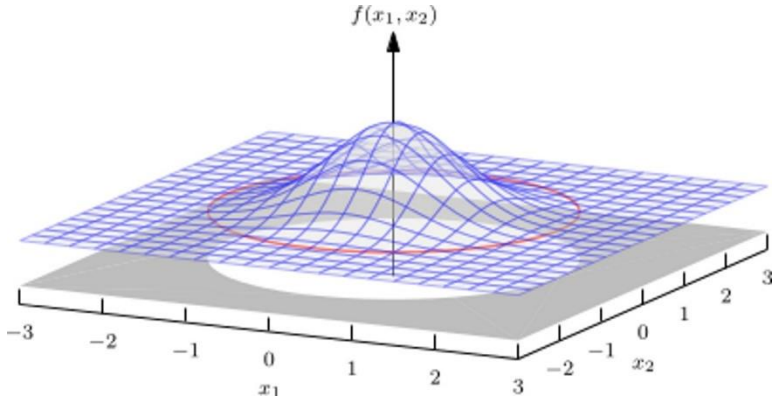
- 高斯分布有两个参数μ和σ，它们分别表示均值和标准差，并且使用符号$N(\mu, \sigma)$表示。
- 高斯分布中一个点x的概率密度函数$f(x)$表示为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- x离分布中心越远，$f(x)$的值越小，因此可以使用点x到原点的距离作为异常分数。
- 这个距离值有一个概率解释，可以用来评定x是一个异常点的置信度。



3.统计方法

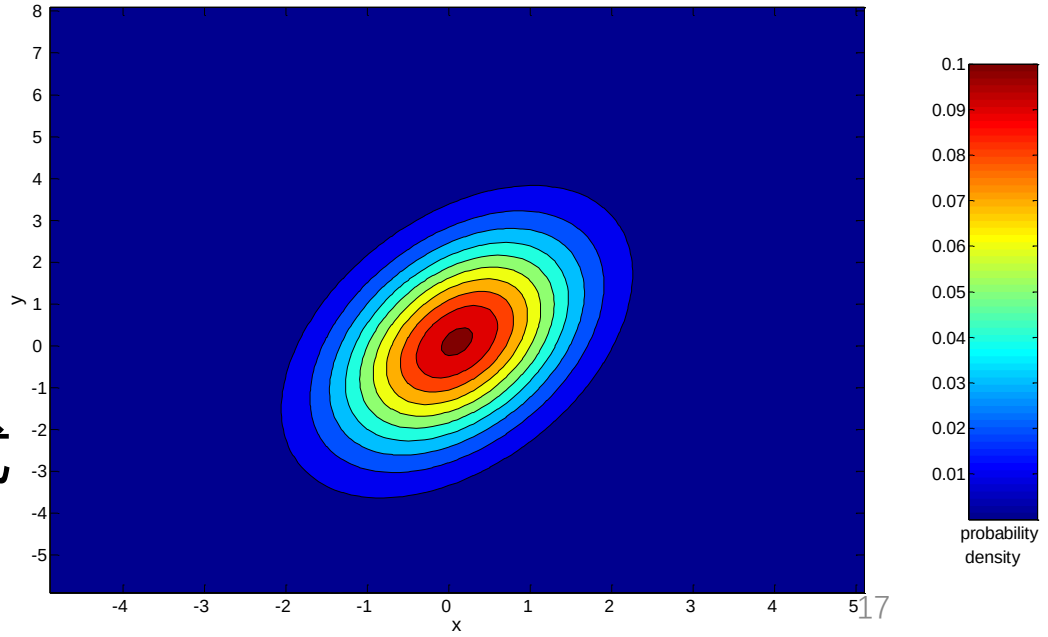


2. 使用参数模型 --- 多元高斯分布

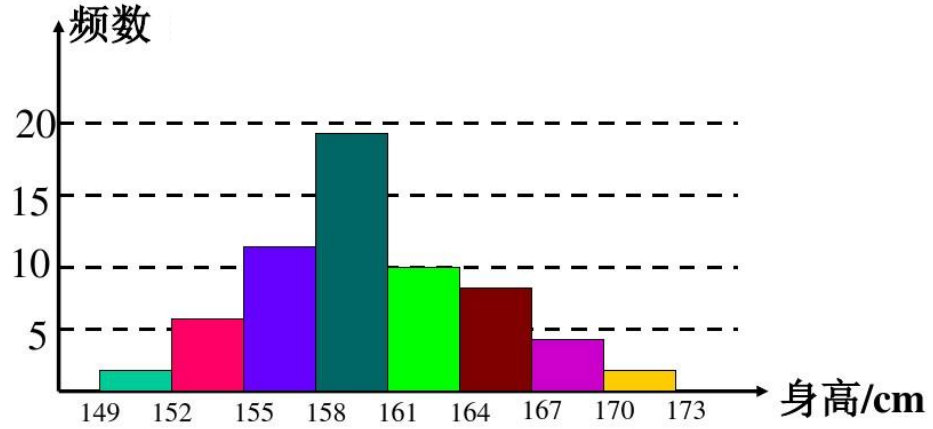
- 对于由两个或多个连续属性组成的数据集，可以使用多元高斯分布对正常类建模。
- 多元高斯分布 $N(\mu, \Sigma)$包括两个参数：均值向量μ和协方差矩阵Σ，它们需要从数据中估计得到。
- 点$\boldsymbol{x}$的概率密度函数$N(\mu, \Sigma)$表示为：

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}}$$

- n是$\boldsymbol{x}$的维数，$|\Sigma|$表示协方差矩阵的行列式



3.统计方法



3. 使用非参数模型

- 简单的对正常类建模的非参数方法是建立正常数据的直方图。
- 如果数据包含单一连续的属性，那么可以使用等宽离散技术来构造属性的不同范围的容器。
- 之后，可以检查一个新的实例是否落在了直方图的容器中。如果它没有落在任何一个容器中，则可以认为它是一个异常实例。
- 否则，可以使用实例所在容器频率的倒数作为它的异常分数。这个方法称为基于频率的或基于计数的异常检测方法。

目 录

4.基于邻近度的方法

1. 基本概念

- 基于邻近度的方法把那些远离其他对象的实例认定为离群点。
- 该方法依赖的假设是，正常实例是相关的并且彼此接近，而异常的实例与其他实例不同，因此与其他实例的距离相对较远。
- 因为许多基于邻近度的技术都基于距离度量，因此也称为基于距离的异常检测技术。
- 基于邻近度的方法都是无模型异常检测技术，因为它们没有构建一个明确的正常类模型用于计算异常分数。
- 它们利用每个数据的局部视角计算其异常分数。
- 它们比统计方法更加通用，因为确定一个对数据集有意义的邻近度度量通常比确定其统计分布更容易。

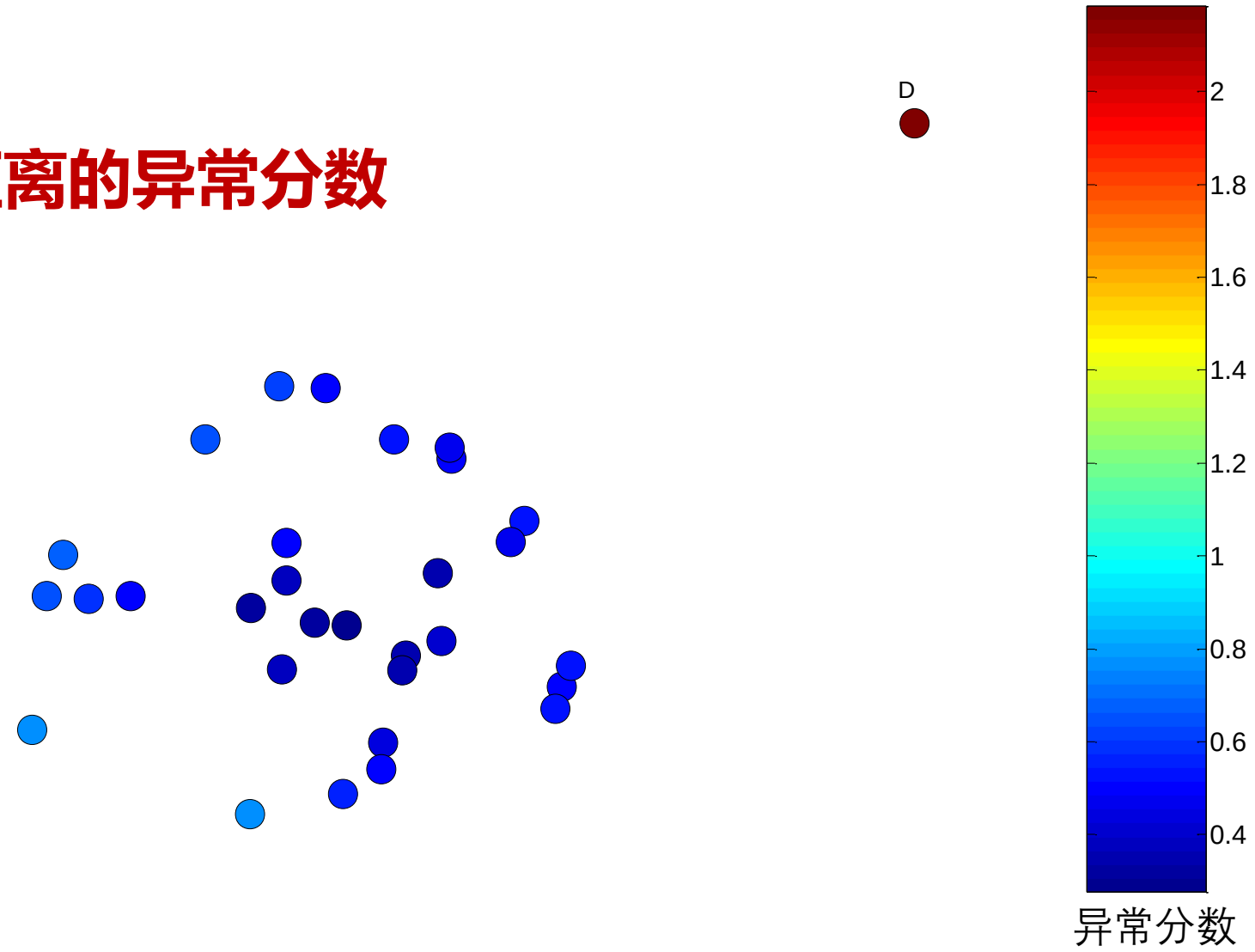
4.基于邻近度的方法

2. 基于距离的异常分数

- 定义数据实例x的基于邻近度的异常分数的最简单方法之一是使用它到第k个最近邻的距离 $dist(x, k)$。
- 如果一个实例 x 有许多其他实例位于它的附近(正常类的特性)，$dist(x, k)$的值将会很小。
- 一个异常实例x将会和它的k-近邻实例有非常远的距离，因此$dist(x, k)$的值也会比较大。

4.基于邻近度的方法

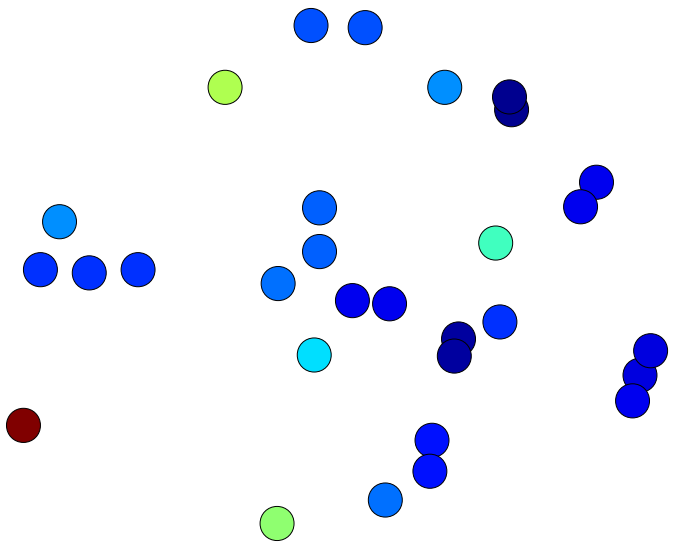
2. 基于距离的异常分数

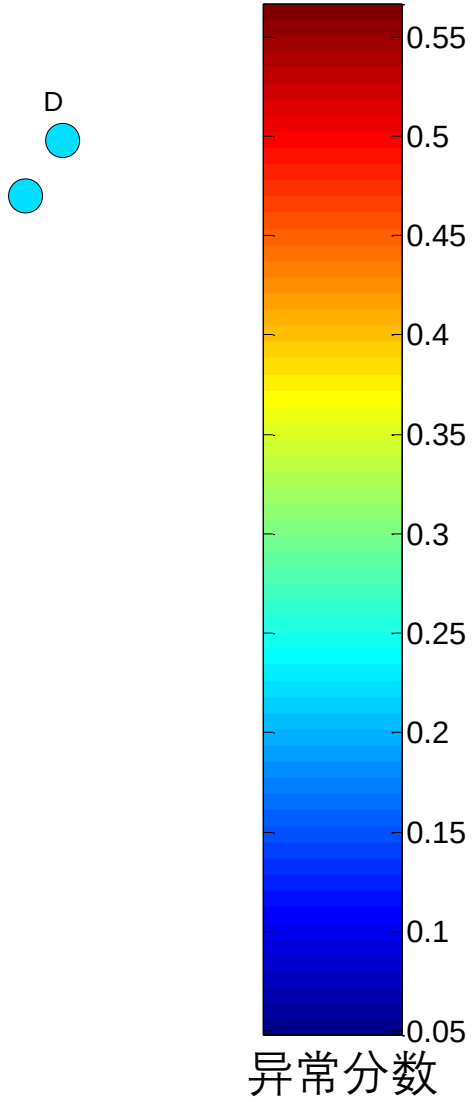


基于距离的5-近邻异常分数($dist(\boldsymbol{x},\ 5)$)

4.基于邻近度的方法

2. 基于距离的异常分数



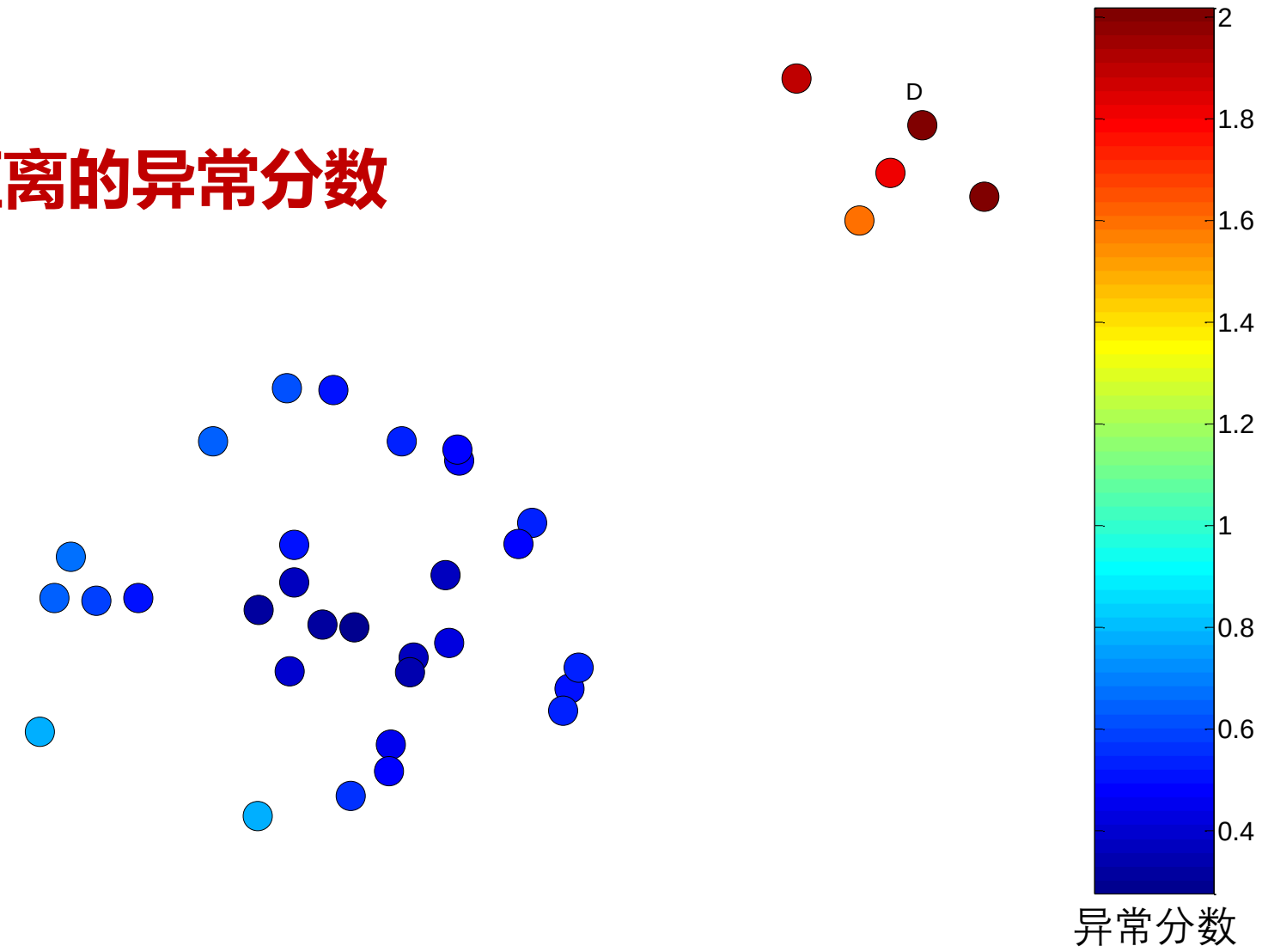


$dist(x,\ k)$会对k的取值非常敏感。如果k值太小，例如1，那么少量的离群点彼此靠近可以显示出低的异常分数。

基于距离的1-近邻异常分数($dist(\boldsymbol{x},\ 1)$)

4.基于邻近度的方法

2. 基于距离的异常分数



如果 k太大，那么在小于k个对象的簇中的所有对象都可能变成异常。

基于距离的5-近邻异常分数($dist(\boldsymbol{x},\ 5)$)

4.基于邻近度的方法

2. 基于距离的异常分数

- 另一种基于距离的异常分数是取前k个最近邻距离的平均值$avg.dist(\boldsymbol{x}, k)$，其对$k$的选择更具有鲁棒性。
- 事实上，$avg.dist(\boldsymbol{x}, k)$作为一种可靠的基于邻近度的异常分数被广泛应用在多个应用中。

4.基于邻近度的方法

3. 基于密度的异常分数

- 一个实例周围的密度可以定义为$\frac{n}{V(d)}$，其中n 是指距离它在一个特定距离d 内的实例数，$V(d)$是邻域的体积。
- 由于$V(d)$对于给定的d 是恒定的，所以实例周围的密度通常用固定距离d内的实例数量n 来表示。此定义类似于 DBSCAN 聚类算法所使用的定义。
- 从基于密度的观点来看，异常是在低密度区域中的实例。因此，一个异常在距离d内具有的实例个数比正常的实例更少 。

4.基于邻近度的方法

3. 基于密度的异常分数

- 在基于密度的度量中选择参数 d是具有挑战性的。
  - 如果d太小，那么许多正常实例会错误地显示低密度值。
  - 如果d太大，那么许多异常可能具有类似于正常实例的密度。

密度定义为$\frac{n}{V(d)}$

- 基于距离和基于密度的异常分数是相反的关系。这可以用来定义如下的密度度量，其基于两个距离度量 $dist(\boldsymbol{x},\ k)$和$avg.dist(\boldsymbol{x},\ k)$：

$$density(\boldsymbol{x},k) = \frac{1}{dist(\boldsymbol{x},k)}$$

$$avg.density(\boldsymbol{x},k) = \frac{1}{avg.dist(\boldsymbol{x},k)}$$

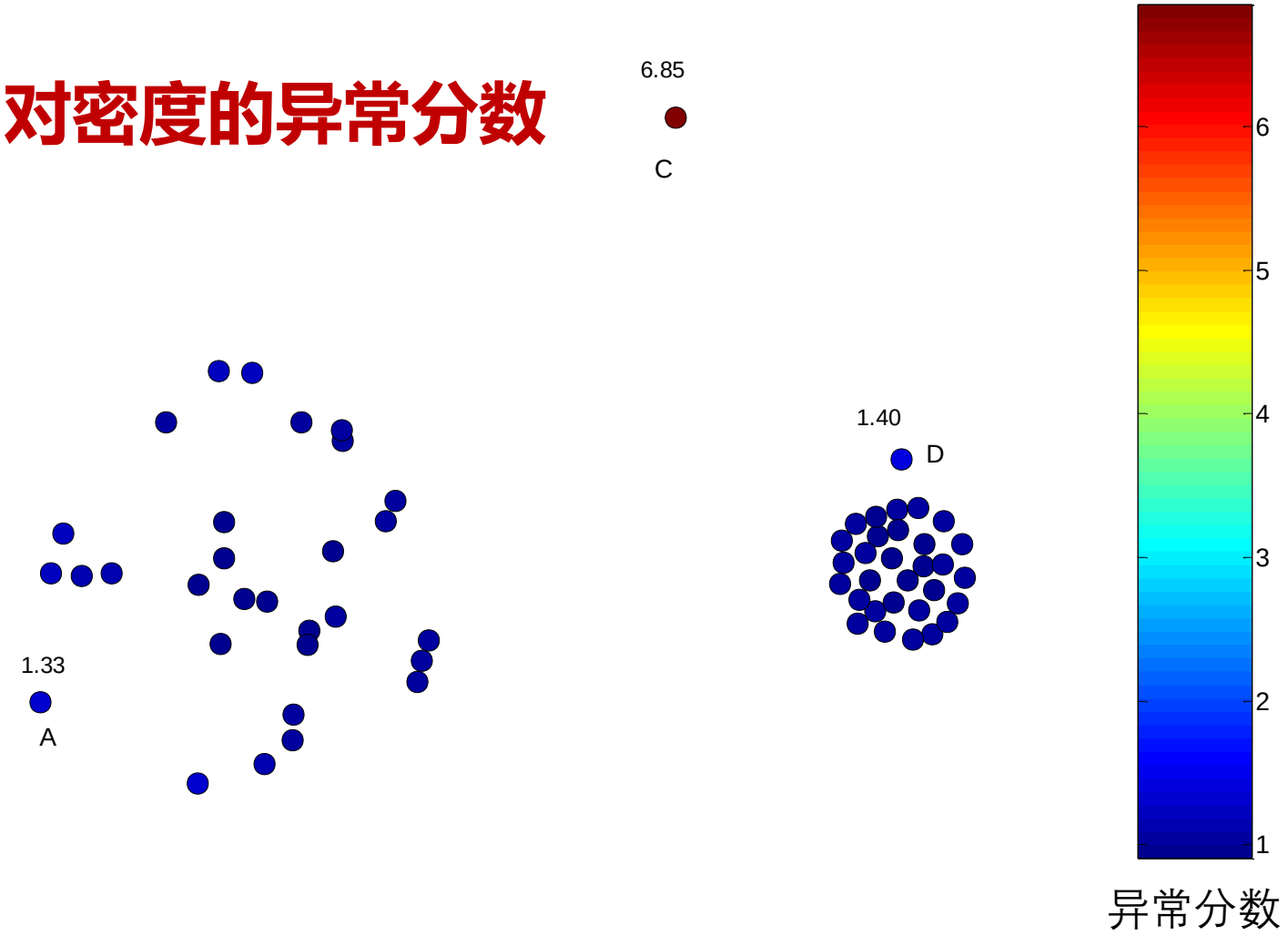
4.基于邻近度的方法

4. 基于相对密度的异常分数

- 上述基于邻近度的方法只考虑单个实例的局部性来计算其异常分数。
- 在数据包含不同密度的区域的情况下，这样的方法无法正确地识别异常，因为正常位置的概念会随着区域的变化而变化。

4.基于邻近度的方法

4. 基于相对密度的异常分数



D的分数远低于松散点簇中的许多点

基于距离的5-近邻异常分数($dist(\boldsymbol{x},\ 5)$)

4.基于邻近度的方法

4. 基于相对密度的异常分数

- 为了正确地识别上述数据集中的异常点，需要一个与相邻实例的密度相关的密度概念。
  - 例如，上图中的点D比点A有一个更高的绝对密度，但是它的密度却比它最邻近的点低
- 对于一个点x，有一种方法就是计算它的 k-近邻（y_1到y_k）平均密度和x的密度的比率，如下：

$$\text{相对密度}(\boldsymbol{x}, k) = \frac{\frac{1}{k}\sum_{i=1}^{k} density(\boldsymbol{y_i}, k)}{density(\boldsymbol{x}, k)}$$

- 当点附近的平均密度明显高于点的密度时，它的相对密度就会很高。

目 录

5.基于聚类的方法

1. 概述

- 基于聚类的异常检测方法使用簇来表示正常类。
- 这依赖于这样的假设：正常实例彼此接近，因此可以被分组成簇。
- 异常点则为不符合正常类的簇的实例，或者出现在与正常类的簇相距很远的小簇中的实例。
- 基于聚类的方法可以分为两种类型：
  - 将小簇视为异常。
  - 一个点如果没有很好地符合聚类将被定义为异常，通常由该点到簇心的距离来度量。

5.基于聚类的方法

2. 发现异常簇

- 这种方法假定在数据中存在簇异常，其中异常以小规模的紧密组出现。
- 当异常从同一异常类中产生时，就会出现聚类异常。
- 异常簇通常为小簇，因为异常在自然界中是罕见的。
- 由于异常不符合正常的模式或行为，所以异常也被期望会远离正常类的簇。
- 检测异常簇的一种基本方法是对总体数据进行聚类，并标记大小太小或与其他簇相距太远的簇 。

5.基于聚类的方法

3. 发现异常实例

- 从聚类的角度来看，另一种描述一个异常的方式是该实例不能被任何正常簇解释。
- 一个异常检测的基本方法是，首先聚类所有数据(主要包括正常实例)，然后评估每个实例属于其各自簇的程度。
  - 例如，如果使用K均值聚类，一个实例到它所属簇的质心的距离表示它属于该簇的程度。
- 因此，远离它们各自簇的质心的实例可以被识别为异常。

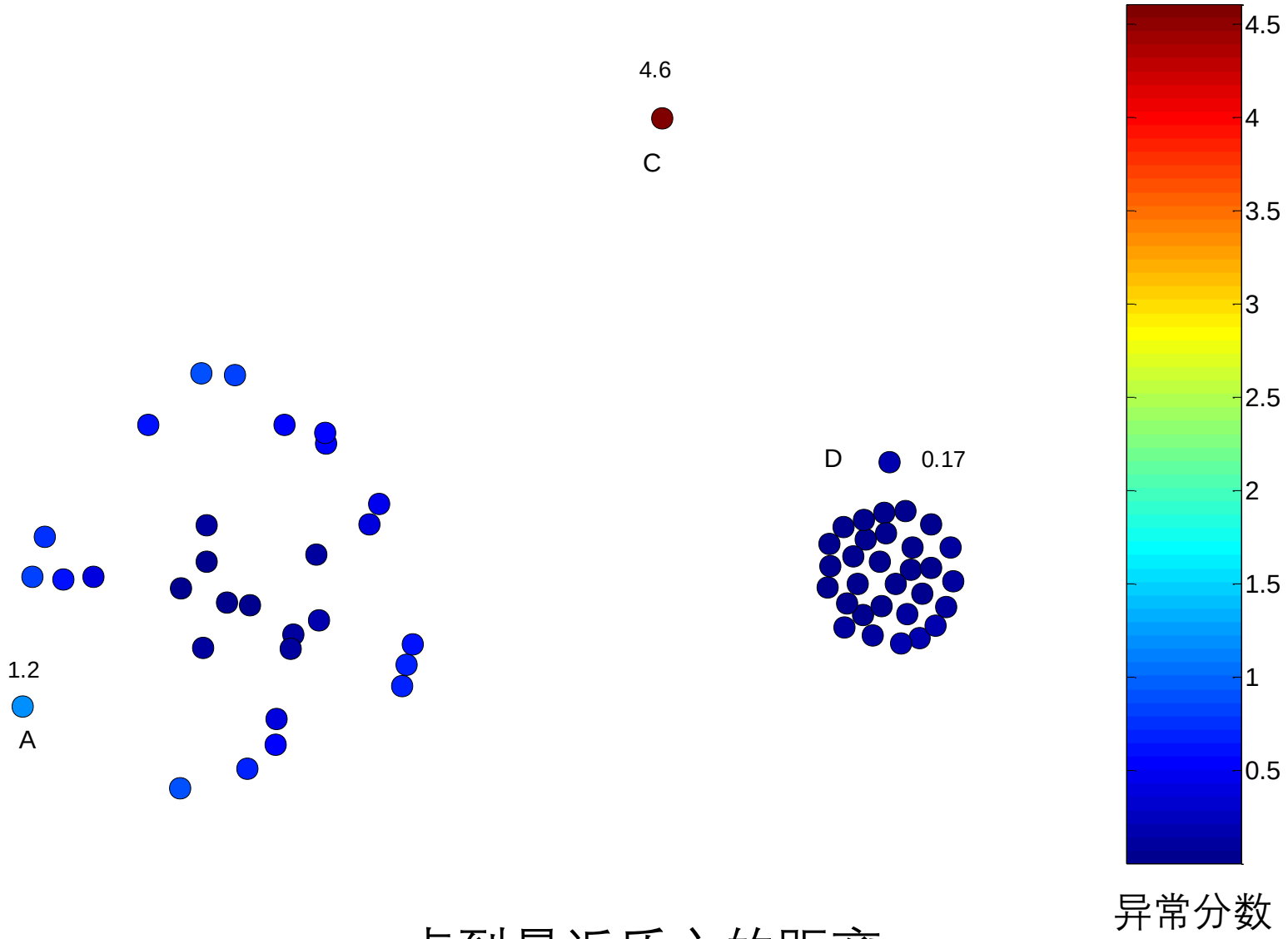
5.基于聚类的方法

3. 发现异常实例 --- 评估一个对象属于一个簇的程度

- 对于基于原型的簇，有几种方法来评估一个实例属于一个簇的程度。
- 一种方法是度量实例与其簇原型之间的距离，并将其视为该实例的异常分数。
- 如果这些簇的密度不同，那么我们可以构造一个异常分数，该分数可以就实例到簇中其余实例的距离而言，度量实例到簇原型的相对距离。
- 另一种可能是假设簇可以用高斯分布精确地建模，那就使用马氏距离作为度量方式。

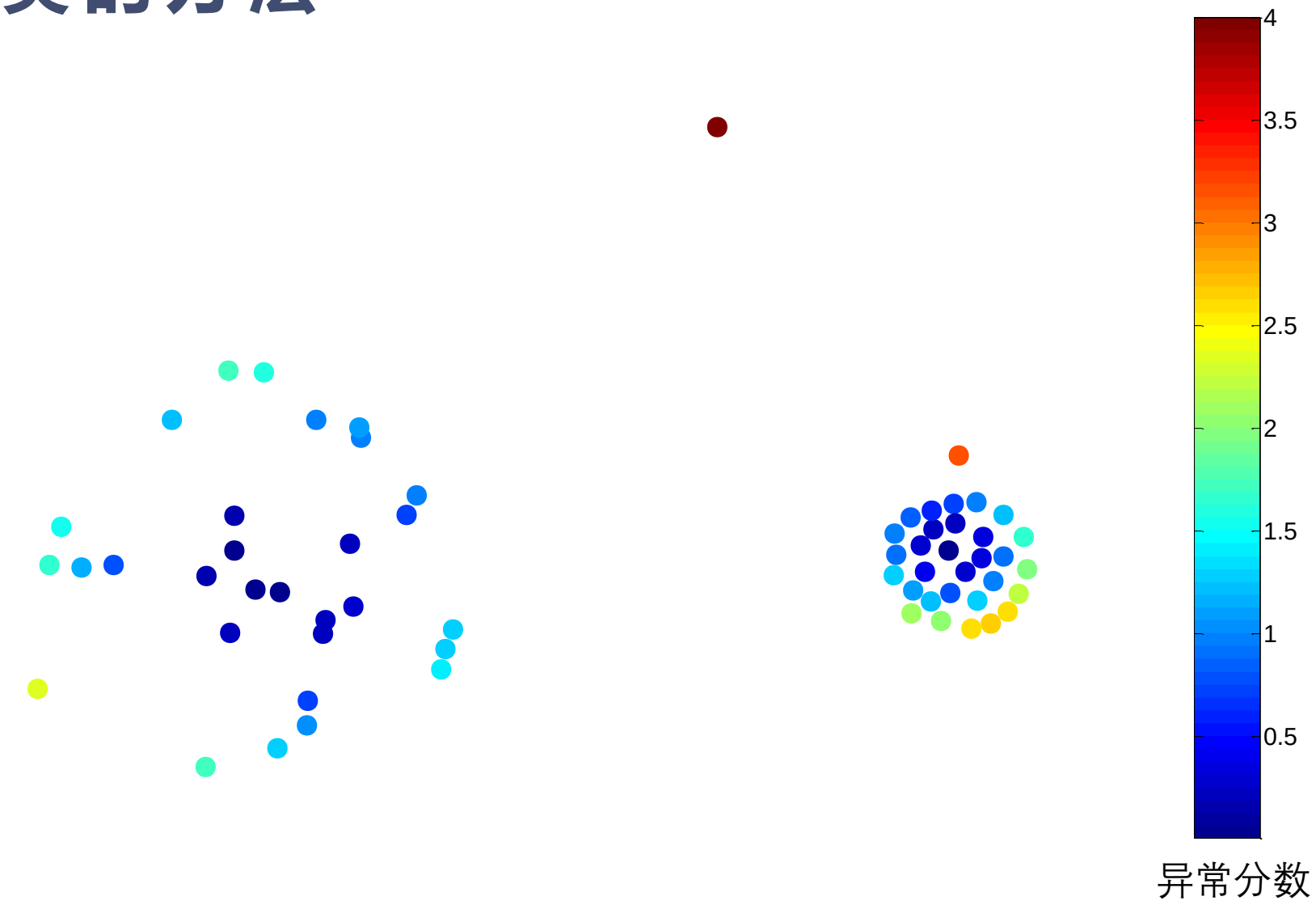
$$M(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}$$

5.基于聚类的方法



点到最近质心的距离

5.基于聚类的方法



相对距离是该点与质心的距离跟该簇中所有点与质心之间的中位距离之比

点到最近质心的相对距离