智能信息挖掘与处理

ZHINENG XINXI WAJUE YU CHULI

杨振舰 于彦伟 张运杰 编著

化学工业出版社

智能信息挖掘与处理

杨振舰　于彦伟　张运杰　编著

化学工业出版社
·北京·

本书首先从空间数据流密度聚类、时空轨迹数据流在线聚类及实时查询处理三个方面，分析了现有数据流挖掘算法的挖掘效果、处理效率、伸缩性与参数敏感性等相关问题，提出了一系列适用于海量时空数据流在线分析的方法与处理框架。然后将数据挖掘与可视化技术相结合，对基于可视化数据挖掘技术的城市地下空间 GIS 系统的关键技术和构建方法进行了系统讨论，对城市地下空间超前地质预报展开了深入研究，形成了一整套支持可视化数据挖掘的城市地下空间 GIS 模型及系统。

本书可作为高等学校计算机工程、信息工程、智能机器人学、工业自动化、数据挖掘与模式识别等专业研究生的教材或教学参考书，也可供智能信息挖掘与处理研究人员学习参考。

图书在版编目（CIP）数据

智能信息挖掘与处理/杨振舰，于彦伟，张运杰编著. —北京：化学工业出版社，2014.8

ISBN 978-7-122-20904-7

Ⅰ.①智… Ⅱ.①杨…②于…③张… Ⅲ.①人工智能-信息处理 Ⅳ.①TP18

中国版本图书馆 CIP 数据核字（2014）第 124331 号

责任编辑：董　琳　　装帧设计：刘剑宁

责任校对：李　爽

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）

印　　刷：北京永鑫印刷有限责任公司

装　　订：三河市宇新装订厂

710mm×1000mm　1/16　印张 10　字数 182 千字　　2014 年 9 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899

网　　址：http://www.cip.com.cn

凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：68.00 元

前　言

目前，在需要处理大数据量的科研领域中，数据挖掘受到越来越多的关注，同时，在实际问题中，大量成功运用数据挖掘的实例说明了数据挖掘和智能处理对科学研究具有很大的促进作用。数据挖掘可以帮助人们对大规模数据进行高效地分析处理，以节约时间，将更多的精力投入到更高层的研究中，从而提高科研工作的效率。因此，数据挖掘的研究与应用现已成为各个科学领域中的热点。本著作针对海量时空数据流聚类算法和可视化数据挖掘技术在城市地下空间 GIS 中的应用两个方面进行了相关理论和方法的研究。

无线位置感知技术的快速发展和广泛应用产生了海量的时空数据流大数据，急需有效的海量数据流管理与信息提取的理论和方法。时空数据流聚类分析技术能够发现动态变化的数据对象间的相关模式、结构特征及演化规律等知识，因而具有重要研究意义。本书首先从基于密度的空间数据流聚类、时空轨迹数据流在线聚类及实时查询处理方面，分析了现有数据流挖掘算法的挖掘效果、效率、可伸缩性与参数敏感性等相关问题，提出了一系列适用于海量时空数据流在线挖掘方法与智能处理框架。针对空间数据流聚类分析，提出了一种基于密度的空间数据流在线聚类算法 OLDStream，该算法采用连通性密度概念，在先前聚类结果上聚类新空间数据点。仅对新空间点及其满足核心点条件的邻域数据做局部聚类更新，降低聚类的时间复杂度。针对空间数据流聚类分析，提出了一种基于密度的空间数据流在线快速聚类算法 OLDStream，算法在先前聚类结果上增量更新局部簇结果，降低了聚类的时间复杂度，实现了对空间数据流的在线聚类。针对时空轨迹数据流聚类，提出了一种基于密度的轨迹流在线聚类算法 CTraStream，将轨迹流聚类过程分为轨迹线段流聚类和轨迹簇在线更新两个阶段。最后，设计了一种面向移动目标邻居实时查询处理的时空轨迹数据流挖掘框架 TSMF，能够实时响应用户关于轨迹簇、蜂群模式和 k-NNT 的查询请求。

随着城市地下空间工程的发展，大量的空间和非空间的数据得到采集和存储。如何更有效的利用这些地下复杂空间数据，并服务于城市地下空间工程的超前地质预报和分析，是城市地下空间数据分析和综合利用的重要研究方向。本书进一步将数据挖掘与可视化技术相结合，将数据挖掘和处理结果，转换为人们容易理解的图形、图像的形式，显示在屏幕上，实现数据可视化分析，以提高整个数据挖掘过程的灵活性、有效性与交互性。本书对基于可视化数据挖掘技术的城

市地下空间 GIS 系统的关键技术和构建方法进行了系统讨论，优化了机器学习算法、空间和非空间的聚类算法，研究结合挖掘算法的相关可视化技术，对城市地下空间超前地质预报展开了深入研究工作，形成了一整套支持可视化数据挖掘的城市地下空间 GIS 原型系统。最后，在空间数据挖掘算法设计基础上，设计了城市地下空间 GIS 系统架构，并实现了基于插件形式进行城市地下空间数据挖掘的 GIS 原型系统，应用于天津市城市地下空间超前地质预报。

本书研究内容旨在提出支持大规模空间数据挖掘和智能处理的高效解决方法，有助于提高海量信息挖掘和智能处理的水平，为城市管理应用邻域内的各类空间数据处理提供重要的理论基础，加快城市信息化建设、提高城市管理效率和水平，也可以为大数据的挖掘和智能处理技术研究提供有益参考。

本书第 1～4 章由烟台大学于彦伟老师和天津城建大学张运杰老师共同编著，第 5～9 章由天津城建大学杨振舰老师负责编著，张运杰老师负责全书的统稿，天津城建大学胡建平教授、河北工业大学夏克文教授审阅了全书并提出了许多宝贵意见。

本书在编著过程时参考或引用了部分单位、专家学者的相关资料，得到了许多业内人士的大力支持，在此表示衷心的感谢。限于编著者水平有限和时间紧迫，书中疏漏及不当之处在所难免，敬请广大读者批评指正。

编著者

2014 年 5 月

目 录

1 概 论

随着无线通信技术的快速发展，各种空间数据采集设备得到了广泛应用，如GPS、智能手机、视频监控、RFID等已普及到了我们日常生活。基于这些位置感知技术的时空数据应用也越来越多，如各类人员和移动设备实时监控系统、基于车辆GPS的智能交通管理系统、移动社交网络、移动计算、位置服务等。基于这类空间数据的挖掘和处理结果不仅是实现智能交通管理、基于位置的服务、安全防卫等重要应用的基础，也可为城市规划、社会服务网络、疾病传播、人类社会行为分析、自然环境变化等研究提供重要参考。

目前，在需要处理大数据量的科研领域中，数据挖掘受到越来越多的关注，同时，在实际问题中，大量成功运用数据挖掘的实例说明了数据挖掘和智能处理对科学研究具有很大的促进作用。数据挖掘可以帮助人们对大规模数据进行高效的分析处理，以节约时间，将更多的精力投入到更高层的研究中，从而提高科研工作的效率。因此，数据挖掘的研究与应用现已成为各个科学领域中的热点。本著作针对海量时空数据流聚类算法和可视化数据挖掘技术在城市地下空间GIS中的应用两个方面进行了相关理论和方法的研究。

无线位置感知技术的快速发展和广泛应用产生了海量的时空数据流大数据，需要有效的海量数据流管理与信息提取的理论和方法。时空数据流聚类分析技术能够发现动态变化的数据对象间的相关模式、结构特征及演化规律等知识，因而具有重要研究意义。本书首先从基于密度的空间数据流聚类、时空轨迹数据流在线聚类及实时查询处理方面，分析了现有数据流挖掘算法的挖掘效果、效率、可伸缩性与参数敏感性等相关问题，提出了一系列适用于海量时空数据流在线挖掘方法与智能处理框架。

针对空间数据流聚类分析，提出了一种基于密度的空间数据流在线聚类算法——OLDStream。该算法采用连通性密度概念，在先前聚类结果上聚类新空间数据点，仅对新空间点及其满足核心点条件的邻域数据做局部聚类更新，降低

聚类的时间复杂度。同时设计一种空间 Eps-网格索引和聚类簇核心点列表索引结构优化区域查询搜索、簇扩展和簇合并等操作，实现了对空间数据流的在线聚类。实验结果显示仅有 4%的数据点消耗最坏运行时间，平均聚类速度约为 0.033ms/空间点。OLDStream 算法能够快速处理大规模空间数据流，实时获取任意形状的聚类簇结果，对数据流的输入顺序不敏感，具有较高的效率和伸缩性。

针对时空轨迹数据流聚类问题，提出了一种基于密度的轨迹流在线聚类算法 CTraStream，该算法将轨迹流聚类分为两个阶段：轨迹线段流聚类和轨迹簇在线更新。首先，将移动目标的轨迹看成不断增加的线段流，提出了一种线段距离测量方法。对于轨迹线段流，设计了一种增量密度聚类算法，实现了对当前采样时间区间内的线段流进行在线密度聚类。其次，对于轨迹簇更新，设计了一种轨迹簇树 TC-Tree 结构存储所有关闭轨迹簇。根据线段簇结果，在线更新 TC-Tree，实现对轨迹数据流的在线聚类。相比已有算法，CTraStream 能实时发现更接近于移动目标真实运动轨迹的聚类簇结果，同时实验结果显示算法的平均处速度可达 2.4K 轨迹段/S。针对时空轨迹数据流上的实时查询问题，设计了一种面向邻居查询的轨迹数据流挖掘框架 TSMF，该框架包括两个部分：在线的轨迹数据流挖掘和离线的面向移动目标的实时查询处理。在线部分集成了 OLDStream 和 CTraStream 算法，实现对轨迹数据流的在线聚类，此外，通过建蜂群模式哈希表存储索引（Swarm-HT）实现对蜂群模式的在线挖掘。离线部分实现了当前关闭轨迹簇 *CCTC*、当前关闭蜂群模式 CCSwarm 和邻居轨迹 *k*-NNT 三种面向移动目标邻居的实时查询处理，当有用户请求查询时，从实时更新的轨迹簇和蜂群模式中快速搜索结果以响应用户的实时查询处理。随着城市地下空间工程的发展，大量的空间和非空间的数据得到采集和存储。如何更有效地利用这些地下复杂空间数据，并服务于城市地下空间工程的超前地质预报和分析，是城市地下空间数据分析和综合利用的重要研究方向。本书进一步将数据挖掘与可视化技术相结合，将数据挖掘和处理结果转换为人们容易理解的图形、图像的形式，显示在屏幕上，实现数据可视化分析，以提高整个数据挖掘过程的灵活性、有效性与交互性。本书对基于可视化数据挖掘技术的城市地下空间 GIS 系统的关键技术和构建方法进行了系统讨论，优化了机器学习算法、空间和非空间的聚类算法，研究结合挖掘算法的相关可视化技术，对城市地下空间超前地质预报展开了深入的研究工作，形成了一整套支持可视化数据挖掘的城市地下空间 GIS 原型系统。

在可视化空间数据挖掘技术研究方面，从数据挖掘技术特点、海量数据特征以及多维、多源数据集成的角度进行综合分析，采用可视化数据挖掘和 GIS 技术相结合的集成应用。在空间数据挖掘技术上，主要采用基于空间关联规则、支持向量机和聚类分析的空间数据挖掘方法。在空间数据挖掘的可视化技术上，提

出了一种基于平行坐标理论的多维多时相空间数据可视化方法，能较好地处理海量空间数据可视化问题，使用 Java 3D 技术实现了复杂地质体的建模显示，以及空间插值结果的三维展示功能。

结合空间关联规则和基于案例推理（CBR）学习思想，对基于支持向量机的空间数据挖掘方法进行了深入分析，以 GIS 技术以及空间数据模型为切入点，提出了进一步提高分类精度和缩短训练时间的两种改进方法，即 CBR 初选训练子集和基于空间区域划分的 SVM 算法。与常规方法进行对比实验，结果表明两种改进算法能够缩短训练时间，在大数据量情况下提高进行空间数据挖掘的效率；其中基于空间区域划分的 SVM 算法还可以在一定程度上缩短训练时间。此外，对于空间数据挖掘中基于距离测度的空间分类方法做了改进，即以统计距离代替欧氏距离可以消除数据自身相关性带来的错误分类影响。

在城市地下空间 GIS 分类技术分析与数据质量控制方面，针对城市地下空间点、线、面数据，可以采用基于距离、数学形态、拓扑关系和空间关联规则的空间聚类分析方法来进行分类；对于文本分类，可以经过文本预处理、特征选择、特征项权重确定和具体分类等过程来实现。另外，针对空间分析过程中的抽样布点问题，采用基于三明治空间抽样模型的空间抽样方法对城市地下空间数据采集过程中的抽样布点问题进行模拟和改进，最终达到在不损失可信度和精度的前提下降低地质数据采集成本的目的。最后，在空间数据挖掘算法设计基础上，设计了城市地下空间 GIS 系统架构，并实现了基于插件形式进行城市地下空间数据挖掘的 GIS 原型系统，应用于天津市城市地下空间超前地质预报。

本章将首先概述时空数据挖掘基本概念、数据流聚类方法、轨迹数据挖掘、实时查询处理方法、GIS 可视化技术和城市超前地质预报的研究现状。内容如下：概述时空数据挖掘现状；介绍空间数据流聚类算法研究现状；介绍轨迹数据挖掘及实时查询处理方面的研究现状；概述 GIS 可视化空间数据挖掘技术；介绍城市超前地质预报的发展现状。

1.1 时空数据挖掘研究概述

随着雷达、红外、GPS 技术、无线射频、GIS 技术、电子跟踪及定位设备的普及，大量的时间空间数据呈指数增长，时空数据的积累已经远远超出人们的分析能力。如何从这些海量时空数据中提取有趣的空间知识呢？人们开始越来越关注空间数据挖掘和知识发现（Spatial Data Mining and Knowledge Discovery，SDMKD）的研究应用。目前 SDMKD 已经成为国际研究热点，并且取得了相当的理论和技术成就。Han 和 Kamber 在专著中，系统讲述了时空数据挖掘及流数据挖掘的概念和技术。李德仁等较早开始关注空间数据挖掘和知识发现，研究从 GIS 数据库中发现知识，构筑了空间数据挖掘和知识发现的理论框架，系统研究

了聚类分析、统计分析、空间在线数据挖掘等理论和技术，其著作总结了他们多年在空间数据挖掘领域的研究成果和应用。Murray 和 Estivill-Castro 在空间数据分析的聚类发现技术方面，分析了基于统计学、数据挖掘和地理信息系统的空间模式识别和知识发现方法。

从狭义角度来看，可以将时空数据流看作是变化更新速度快、数量无限增长的数据集合。鉴于这样的数据通常具有明显的实效性，因而将这些源源不断产生的数据全部存储起来是不现实，也是不必要的。从广义角度来看，数据流是针对静态传统数据库而言的，是只能进行线性扫描等操作的大规模海量连续数据集合。如电信公司的电话通话记录、客户点击流、交易数据、证券交易所的股票形势分析以及移动目标管理领域中实时时空数据等数据集合。通常情况下，这些超大规模的海量数据是无法完全存放在主存中，只能存放在硬盘、磁盘等类的外存中。在这些应用中，基于时效性要求及数据特性，采用在线方式实时处理数据流是非常必要的。

（1）在线挖掘需求

时空数据流挖掘技术研究的目的是解决随时间快速增长的时空数据的实时分析和应用关键问题。目前，时空数据流挖掘在智能家庭、实时监管、天气预报、智能交通导航、安全监控和位置服务等领域有广泛的应用需求。

（2）在线监测应用需求

在时空数据采集系统中，大量应用要求对获取的时空数据进行在线监测，如在自然灾害预警中，需要实时发现飓风的可能路径；在交通管理系统中，需要在线监测热点交通路段、交通趋势、异常的交通状态；在目标跟踪系统中，需要实时跟踪目标的运动轨迹、分析目标间的邻近关系、预测目标的运动趋势等；在应急安防系统中，需要在线预防异常事件的发生。这些应用都要求对时空数据流进行在线处理。

（3）大数据挖掘需求

时空数据流随时间不断积累和更新，形成大数据（Big Data），如果静态地对这些大数据进行挖掘，不仅要消耗大量时间，还需要借助大量存储设备和计算量能力巨大的计算设备，而在线数据流处理方案降低了对系统计算能力的要求，能够处理源源不断的数据，是一种重要的大数据处理模式。对一般计算系统而言，对大数据的挖掘需要在线处理数据流。

（4）效率需求

单独对大规模海量数据进行数据挖掘是非常低效的，而数据流在线挖掘在已有挖掘基础上更新挖掘结果，处理时间短，挖掘效率较高。

目前，时空数据流挖掘的研究内容已涉及聚类分析、时空规则挖掘、频繁轨迹模式、周期模式、异常检测、轨迹预测、语义地理信息挖掘、语义时空数据库

研究、空间索引结构、邻域查询处理等多个方面。近年来，基于时空数据流的实时系统应用越来越广泛，相应的数据分析与挖掘技术也成了研究热点并取得了一定成绩。但是，我们在相关部门的调研中发现，目前针对海量时空数据流分析与挖掘等处理能力还远不能满足要求。例如，北京市公安局、交通管理局每时每刻都会有大量人、车、物等相关目标的轨迹数据输入，但基本上都以原始数据的形式被存放起来，而没有为管理决策提供实时支持。

1.2 空间数据流聚类算法研究

聚类分析是一种非常重要的数据分析方法，已经广泛应用在许多领域，包括模式识别、数据分析、图像视频处理以及商业市场分析。通过聚类，可以使人们容易地识别出密集和稀疏的区域，进而发现全局的分布模式以及数据属性之间有趣的相互关系。

在商业上，聚类分析能帮助市场分析人员从客户信息库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。在生物学上，聚类能用于生物计算领域，有助于对基因分类与划分、物种分布及分类等方面的研究。聚类还可以对地理信息数据库中有相似特性的区域分类、相似习惯和爱好的人员分组、房屋和传染性疾病的分布以及蔓延趋势预测等方面发挥作用。聚类分析还能应用在海量网络文本信息提取、文档分类等领域。

在数据挖掘领域，聚类分析已经被广泛研究了许多年，主要集中在基于距离的聚类分析。如基于 k-means、k-medoids 的聚类算法已经被加入到许多统计分析软件或系统。主要研究都关注在聚类算法的可伸缩性，对复杂形状的聚类效果，以及高维聚类分析技术。近年来，数据流聚类也得到了巨大关注，在移动网络管理、网络监控、商业交易管理、金融分析及移动数据管理等领域得以广泛应用研究。

数据流聚类相对于传统的数据库操作模式下的聚类分析而言，主要区别在于，其处理对象由静态的数据集变成了动态流数据，以及由于流数据自身特点所带来的一些特殊处理要求。

数据流聚类分析是移动目标管理系统常用的分析方法，对移动目标的位置数据分析起到了重要作用。在轨迹挖掘领域，聚类分析即可作为一个独立的工具来获得移动目标的分布情况等知识；也可以作为轨迹数据的预处理步骤，其他挖掘算法在其获取的聚类结果上做进一步的挖掘。

1.2.1 基于密度的聚类算法

聚类算法一般可以划分为如下几类。

(1) 基于划分的方法 (Partitioning Methods)

划分方法首先创建给定数据的一个初始 k 划分。然后采用一种迭代的重定位

技术，尝试通过对象在划分间移动来改进划分，最终将所有数据对象划分成 k 个分组。划分的一般准则是：在同一个组中的对象之间的距离尽可能小，而不同组中的对象之间的距离尽可能大。典型的划分算法主要有 k-means、k-medoids、k-modes、CLARANS 等方法。

(2) 基于层次的方法 (Hierarchical Methods)

层次的方法对给定数据集合进行层次分解，将数据对象构成一颗聚类树。根据层次的形成过程，层次聚类方法可以被分为凝聚方法和分裂方法。

① 凝聚方法。也称自底向上的方法，将每个对象作为一个簇，然后将相互邻近的簇合并成一个更大的簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。

② 分裂方法。也称自顶向下的方法，将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达到了设定的终结条件。

层次聚类方法的缺陷在于，一旦某个步骤完成，它就不能被撤消，将影响到最终结果。典型的层次聚类算法有 BIRCH、CURE 和 ROCK。

(3) 基于网格的方法 (Grid-based Methods)

基于网格的方法采用一个多分辨率的网格数据结构，把对象空间量化为有限数目的网格单元，将网络操作代替在数据上的聚类操作。这种方法的主要优点是处理速度快，处理时间独立于数据对象的数量，只与划分的网格数目有关。但是，这类算法效率的提高是以降低聚类结果精度为代价的。典型算法包括 STING、CLIQUE。

(4) 基于模型的方法 (Model-based Methods)

基于模型的方法为每个簇设定一个模型，将数据分配给最佳匹配的模型。主要有基于统计和基于神经网络两类模型。典型算法主要有基于统计的 AutoClass 和基于神经网络的 SOFM 等。

上述大多数聚类方法只能发现类球状的聚类簇。为了发现空间中任意形状的聚类簇，基于密度的聚类方法 (Density-based Methods) 开始被关注研究。密度聚类方法将数据簇看作是数据空间中由低密度区域分割开的高密度对象区域。主要思想是：对于每个数据点，只要其临近区域的密度超过某个阈值，就被聚在簇中。然后，对其邻域内的数据点继续执行相同聚类操作。这类方法可以过滤“噪声”数据，有效地发现任意形状的聚类簇。

基于密度的算法一般可分为基于局部连通性和基于密度函数两类。基于局部连通性的典型算法有 DBSCAN、GDBSCAN、OPTICS 和 DBCLASD 等，基于密度函数的算法有 DENCLUE。此外，基于密度的算法还包括很多在这些典型算法上的改进算法。

DBSCAN 是一种典型的基于密度的聚类算法，通过检查数据库中的每个对象

邻域内的邻居点个数来聚类。下面给出一些基于密度的基本概念。*Eps*（Epsilon neighborhood）是邻居距离阈值，*MinPts*（Minimum number of Points）是邻居点数量阈值。

① 核心点（Core Point）。如果一个点 p 的 *Eps* 邻域包含多于 *MinPts* 个邻居点，则点 p 是某个簇的核心点。

② 边界点（Border Point）。如果一个点 q 的 *Eps* 邻域包含少于 *MinPts* 个邻居点，但 q 是核心点 p 的 *Eps* 邻域内的邻居点，则 q 是包含点 p 的簇的边界点。

③ 直接密度可达（Directly Density-reachable）。如果一个点 p 是核心点 q 的 *Eps* 邻域点，则称点 p 直接密度可达点 q。

④ 密度可达（Density-reachable）。给定一系列的点 p_1，p_2，…，p_n，且 p_i 直接密度可达 p_{i+1}（$i=1,2,\cdots,n-1$），若 $p_1=p$，$p_n=q$，则称点 p 密度可达点 q。

⑤ 密度相连（Density-connected）。如果点 p、q 同时密度可达点 o，则称点 p 密度相连点 q。由定义可知，密度相连是对称的。

DBSCAN 算法是将密度相连的数据点聚集在同一个簇内。实现方法是：初始任意选择一数据点 p，检测点 p 是否是核心点，若是，则创建一个以 p 为核心点的簇，然后将直接密度可达到点 p 的邻域点加入到簇内。然后对点 p 邻域内点做相同操作，这样按递归的方式将到核心点密度可达的数据点加入簇，直到没有新的点被处理时算法结束。DBSCAN 需要用户输入 *Eps* 和 *MinPts* 两个重要参数，即领域范围和最小点个数。聚类结果对输入参数的选择非常敏感，参数不易选择。算法采用欧式距离，实现简单，聚类效果较好，可以直接处理整个数据集。但数据量比较大时，需要大内存的支持，I/O 消耗也比较大，采用空间索引时时间复杂度为 $O(nlogn)$，否则为 $O(n^2)$。

为了克服 DBSCAN 参数选择困难和全局参数的缺点，Mihael Ankerst 等提出一种排序数据点的簇结构识别方法 OPTICS（Ordering Points to Identify the Clusering Structrue）。该算法并不显式地产生一个数据聚类簇，而是为自动和交互的聚类分析计算一个数据点排序，这个排序代表了数据集合在一系列参数下的基于密度的聚类结构。每个数据点需要额外存储两个属性：核心距离和可达距离。核心距离和可达距离定义如下。

① 核心距离（Core-distance）。一个对象 p 的核心距离是点 p 到它的第 *MinPts* 个邻居点的距离。如果 p 不是核心对象，则 p 的核心距离没有定义。

② 可达距离（Reach-distance）。一个对象 q 关于另一个对象 p 的可达距离是 p 的核心距离和 q 到 p 的距离两者中的较大值。如果 p 不是一个核心对象，p 和 q 之间的可达距离没有定义。

OPTICS 算法采用螺旋式方式从某核心对象开始，按照可达距离排序数据点，之后，根据数据点排序列表的可达距离信息选择某 *subEps* 产生基于密度的