基于密度的散点图去重叠可视化方法

贺怀清, 安楷文*, 刘浩翰, 惠康华

(中国民航大学计算机科学与技术学院 天津 300300)
(2022051005@cauc.edu.cn)

摘 要: 散点图是常用的可视化表征, 但是当数据过多时数据重叠会影响分析的准确性. 为此, 提出基于密度的散点图去重叠可视化方法对散点图进行布局优化. 首先确定待优化区域, 利用数据的密度特征与规则网格划分构建新的采样空间; 然后在采样空间内进行基于密度的采样; 通过采样后处理实现数据关联, 完成采样结果与原数据之间的属性映射; 最后通过数据拆分设定密度阈值以保护散点图中的稀疏点, 维持散点图的局部细节. 在9个数据集上的实验结果表明, 所提方法在保证视觉清晰度、保留原数据分布特征的前提下, 有效地解决了散点图数据重叠的问题, 优化了散点图的布局.

关键词: 散点图; 数据去重叠; 密度; 布局优化; 采样; 可视化
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00814

A Density-Based Visualization for Removing Overlaps of Scatterplots

He Huaiqing, An Kaiwen*, Liu Haohan, and Hui Kanghua

(*Computer Science and Technology College, Civil Aviation University of China, Tianjin* 300300)

Abstract: Scatterplots is a widely used visual analysis method, but when there are excessive points in the scatterplots, data overlap will affect the accuracy of analysis. For this reason, a density-based visualization for removing overlaps of scatterplots is proposed to solve the problem of data overlap. First, determine the area to be optimized. And then use the density-based overlapping removal algorithm. Construct a new sampling space using the density characteristics of the data and regular grid division. Carry out density-based sampling in the sampling space. Then realize data association through post-sampling processing to complete the attribute mapping between the sampling results and the original data. Finally, the density threshold is set through the data splitting step to protect the sparse points in the scatterplots and maintain the local details. The experimental results show that on the premise of ensuring visual clarity and preserving the distribution characteristics of the original data, this method effectively solves the problem of data overlap and optimizes the layout of scatterplots.

Key words: scatterplots; data overlapping removal; density; layout optimization; sampling; visualization

近年来, 随着计算机技术的进步, 数据采集手段逐渐多样, 数据日益丰富, 数据中蕴含的复杂而隐蔽的关系越来越难以观测. 因此, 如何高效地可视化这些数据一直是当前数据可视化中的研究热点.

散点图使用笛卡儿坐标描绘二元点, 是分析变量关系最基本的可视化表示方法之一[1-2]. 散点图的应用极为广泛, 金融分析、医疗健康、市场营

收稿日期: 2023-12-21; 修回日期: 2024-07-17. **基金项目**: 国家自然科学基金(U1333110); 国家重点研发计划(2020YFB1600101); 天津市教委科研项目(2020KJ024). **贺怀清**(1969—), 女, 博士, 教授, 硕士生导师, CCF高级会员, 主要研究方向为图形图像与可视分析; **安楷文**(1999—), 女, 硕士研究生, 论文通信作者, 主要研究方向为图形图像与可视分析; **刘浩翰**(1966—), 男, 硕士, 副教授, 硕士生导师, 主要研究方向为图形图像与可视分析; **惠康华**(1982—), 男, 博士, 副教授, 主要研究方向为图像处理.

销和科学研究等领域常利用散点图探索数据，以协助制定决策. 作为常用的可视化工具，散点图在数据分析和展示中具有多种优势，能够直观地展示数据的分布及变量相关性，便于观察数据趋势、识别异常值和探索异常模式. 但是，散点图也存在数据易重叠和易受异常值干扰的问题[3-5]，阻碍了相关性的分析与模式的比较[6]. 因此，生成能够保留原始散点图特点的去重叠散点图，在数据可视化领域是非常有必要的.

数据的点状分布(简称点数据)是研究空间分布和变量关系的一种常见的数据状态，通过将数据项抽象为二维平面中的二元点，并运用丰富的可视化手段可以直观地反映数据的空间分布密度、区域差异和动态变化等特征. 点数据最重要的特征是空间分布特征，因此，直观、快速、有效地表达点数据在空间中的数量、密度和相互关系，是解决散点图数据重叠问题的重点.

可视化元素由几何标记与视觉通道构成. 从构成角度出发，调整数据点的尺寸和透明度等图元编码是一种常见的减轻数据重叠的方法[7]，但当数据非常密集且规模非常大时收效甚微. 将点要素的空间分布区域按照指定的边界进行分割能够将点要素转换为具有指定边界的面要素[8]，通过统计不同区域内的点数，可以在不受重叠问题的影响下获得区域内数据的数量信息. 在分割边界的过程中，可视化精度与区域划分的精细程度密切相关：区域划分越小，精度越高；区域划分越大，精度也就越低. 因此很难达到合适的区域粒度.

利用采样技术对数据进行抽取能够减少需要绘制的数据点数量，但是采样方法的选取比较困难、处理效率比较低下，不能保证高效地解决散点图的重叠问题. 此外，以改变数据点位置为核心思想的数据移动法也是一种经典的散点图去重叠策略[2-3,9]，这种方法的本质是改变数据本身的分布，能够在根本上解决重叠问题，但其可视化效果往往受到显示空间的大小、数据规模、数据尺寸和数据移动方式的限制. 当显示空间有限且数据点数量较多时，为了使数据点有足够的移动空间，数据点尺寸将被缩小，这往往伴随着可视化清晰度的降低，影响人们对数据对象的视觉感知；当数据点的位移较大时，散点图的整体形状将被扭曲.

虽然上述方法均能在不同程度上减轻散点图的数据重叠问题，但均有较强的局限性. 在对散点图重叠问题与现有解决方法进行充分研究后，本文认为平衡散点图去重叠后的可视化清晰度与布局准确性是十分有必要的. 为此，本文提出基于密度的散点图去重叠可视化方法，通过优化散点图的布局，解决了由数据量大及数据分布不均匀导致的过度绘制问题，提高了可视化的性能和视觉表达；重新设计评价指标的计算方式，对不同方法的可视化效果进行定量评估，并从经典的视觉分析角度对散点图的去重叠效果进行定性评价.

1　相关工作

1.1　密度特征的广泛应用

密度是一个很重要的数据特征，基于密度的可视化方法充分地考虑了数据的聚集性，能够帮助观察者分析数据趋势并识别异常区域[10]. Quadri 等[11]使用合并树构建基于密度的模型来研究人类对聚类的感知；Chen 等[12]提出基于金字塔的散点图采样技术，利用金字塔中的密度值指导不同尺度下的采样；Li 等[13]利用目标的密度信息调整文本标签的位置，以实现更清晰的可视化布局；Wang 等[14]用锚点的局部密度表征簇的差异性.

核密度估计(kernal density estimation, KDE)是点密度评估的常用方法，能够将数据的离散密度特征转换为连续形式[15]，通过连续密度函数反映数据在空间上的聚集性特征. Zinsmaier 等[16]利用 KDE 可视化大型图中重叠的节点和边；Li 等[17]提出基于峰值的 KDE 算法，用于探究时序数据的分布情况. 使用 KDE 时，带宽的确定至关重要，当数据在空间上分布极度不均匀时，离群点会影响带宽参数的设置，导致 KDE 得到的连续密度函数不能很好地反映数据主体的分布特征.

密度聚类也是数据密度特征的应用领域. 张慧杰等[8]利用密度聚类对特征轨迹点进行自适应划分；Kokate 等[18]使用密度聚类对演化中的数据流进行动态聚类. 在密度聚类中，基于密度的噪声应用空间聚类(density- based spatial clustering of applications with noise, DBSCAN)不需要预先指定簇的数量，就能够筛选出数据中的离群点[19]. 本文为空间分布不均匀的数据设计了算法预处理步骤，利用 DBSCAN 筛选并排除原数据中的离群点，使得后续在利用 KDE 对散点图中的数据进行密度评估时，KDE 的带宽设定不受离群点的影响，使 KDE 密度评估结果能够更好地反映数据主体的分布特征. 使用 DBSCAN 时，本文通过估计核心点

邻域内所需最小样本点数目(minPts)和画 *k*-距离图的方法，使聚类参数更符合数据集的特点.

1.2 缓解数据点重叠的方法

目前，已有多种解决散点图数据重叠问题的方法. 对于小规模数据，可以通过绘制中空点或者调整尺寸和透明度[20]来解决. van Onzenoodt 等[21]总结了视觉编码对重叠感知的影响，这类方法简单易用，但不适合重叠度高的场景. 对于大规模点数据，将高密度区域和低密度区域分别处理也能够缓解散点图的过度绘制[22-23]，但缺乏对区域内实体之间的关联性表达.

采样的本质是从完整的数据中挑选出具有代表性的部分数据. 适当的采样技术能够优化数据的选择[24-27]，保留原图的宏观概览. 常用的采样方法有随机采样[28]和蓝噪声采样[29]等，不同的采样策略有其适合的使用场景，因此在面对具有不同特征的数据与不同的分析场景时，采样方法的选择比较困难.

位置移动是解决可视化领域数据重叠的一种常用策略，其核心思想是改变数据点原本的分布位置，在保持一定整体结构的前提下改善可视化效果，如 VPSC[30]，PFS[31]，PRISM[32]等；但是，这些方法的计算效率非常低，并不适用于大规模数据. 网格移动策略是位置移动法的代表性策略，它使用正交的网格线将数据点的分布空间栅格化，并通过网格来限制数据点的位置移动. Keim 等[33]提出的 Gridfit 算法是网格移动策略的经典算法，其主要思想是将被覆盖的数据点放置到最近的未被占用的像素点上，但数据点的移动方向具有不确定性；Hilasaca 等[34]提出一种基于空间划分的散点图去重叠算法 DGrid，其在对散点图进行规则矩形划分时加入一些虚拟点来占据空白网格，从而限制数据点可移动的位置，降低数据点在移动方向上的不确定性，但其不适用于重叠程度较高的数据；赵颖等[2]在使用虚拟点占位的基础上，辅以 Voronoi 网格划分、数据点选择性移动等操作来实现散点图去重，但该方法不适用于数据规模超过 1 万的情况.

上述基于位置移动的方法在本质上改变了数据集的拓扑结构，没有考虑对散点图数据密度和形状的保持，容易造成感知误差. 本文方法充分地利用数据点的密度特征进行建模，结合采样与网格划分的优点，适用于规模大且重叠度高的数据，能够在解决散点图重叠问题的同时，尽可能地拟合原数据的分布特征.

2 本文方法

2.1 框 架

本文方法框架如图 1 所示.

(1) 选定待优化区域作为去重叠算法的输入，如图 1 中①部分所示. 对于空间分布不均匀的散点图，使用 DBSCAN 筛选出离群点，并将其之外的区域定义为待优化区域(case1)，将这个聚类步骤称为算法预处理；对于分布比较均匀的散点图，直接将其整体作为待优化区域(case2). (2) 选定待优化区域后，通过构建采样空间(如图 1 中②部分所示)、基于密度的采样(如图 1 中③部分所示)和采样后处理(如图 1 中④部分所示)，得到带有原数据属性的采样结果点；然后利用构建采样空间时通过 KDE 得到的密度评估结果进行数据拆分(如图 1 中⑤部分所示)，以保护稀疏点；最后将带有原数据属性的采样结果点与稀疏点合并，作为待优化区域的优化结果. 将如图 1 中②~⑤部分封装为基于密度的去重叠(denisty-based overlapping removal, DBOR)算法.

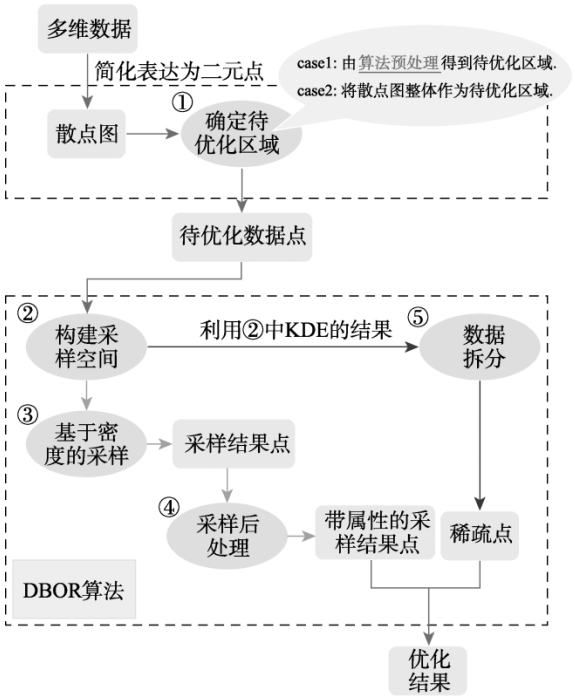


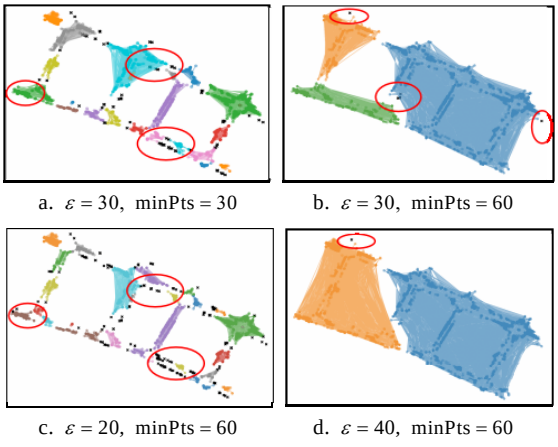
图 1 本文方法框架

2.2 算法预处理

空间分布范围大、密度分布不均匀的散点图通常存在明显的离群点，离群点的存在会影响 KDE 带宽参数的设定，从而影响估计结果的平滑程度. 因此，在使用 DBOR 算法解决这类数据的散点图重

叠问题之前，需要利用 DBSCAN 进行预处理，识别并排除离群点.

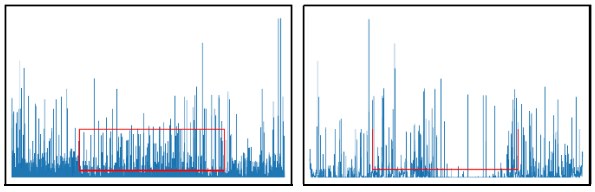
如图 2 所示，在应用 DBSCAN 时，不同的聚类参数会产生不同的聚类结果. 确定聚类参数时，本文使用估计 minPts 和画 *k*-距离图的方法. 在绘制 *k*-距离图时发现，若数据点的索引无序，则生成的 *k*-距离图局部过于密集，可读性较差，如图 3a 所示. 于是，本文重新定义索引序列，将数据点的索引与其相对空间位置对应，生成了如图 3b 所示的有序索引 *k*-距离图.



a. $\varepsilon = 30$, minPts = 30　b. $\varepsilon = 30$, minPts = 60

c. $\varepsilon = 20$, minPts = 60　d. $\varepsilon = 40$, minPts = 60

图 2　不同参数对聚类结果的影响



a. 无序索引　b. 有序索引

图 3　不同索引方式下的 *k*-距离图

k-距离图的拐点表示数据的密度发生了明显的变化，这些拐点的 *k*-距离值对应核心点的邻域半径 ε. 选定某个 ε 值后，不断地调整 minPts 的取值，反复迭代上述步骤形成不同的参数组合.

经过算法预处理后，将不包括离群点在内的区域作为去重叠算法的输入，则由 KDE 得到的连续密度函数将更好地反映数据主体的空间聚集性特征.

2.3　DBOR 算法

2.3.1　DBOR 算法步骤

DBOR 算法结合了采样与规则网格划分的优点. 首先构建采样空间；然后在采样空间内使用基于密度的采样方法，得到少于原数据规模的采样结果点，再通过采样后处理实现与原数据之间的属性映射；最后通过数据拆分设定密度阈值(density threshold, DT)，筛选出待优化区域中的稀疏点.

算法 1. DBOR 算法.

输入. 待优化区域.

输出. DBOR 去重叠散点图.

Step1. 构建采样空间.

Step2. 基于密度的采样. 在采样空间中对待选采样点根据采样概率进行采样，得到采样结果点集 P'.

Step3. 采样后处理. 利用 *k*D 树完成 P' 与待优化区域中数据之间的关联，实现属性映射.

Step4. 数据拆分. 利用 Step1 中 KDE 密度评估的结果确定 DT，筛选出数据中的稀疏点集 D.

Step5. 将 P' 与 D 合并，得到 DBOR 去重叠散点图.

Step6. 算法结束.

2.3.2　采样空间的构建

如果将原始点集作为采样对象，采样所需的计算量与时间的代价较大. 因此，本文不对原数据进行采样，而是使用 KDE 将待优化区域内离散的点密度信息转换为连续密度场，利用原数据的数量、密度和空间分布等特征构建采样空间；然后对采样空间中的待选采样点进行采样，生成无重叠、无覆盖并且拟合原数据分布的采样结果点.

利用 KDE 得到数据的连续密度场后，首先利用网格移动策略的思想，用正交的网格线对数据的分布空间进行规则划分，令网格的宽度等于数据点直径，并且让网格点的尺寸与数据点的尺寸保持一致，保证散点图优化前后具有相同的视觉编码，防止因图元编码不一致造成感知误差；然后将离散的网格点映射到原数据的连续密度场中，赋予网格点密度值，将网格点转化为携带密度信息的待选采样点，由此构建采样空间.

算法 2. 构建采样空间.

输入. 待优化数据点集 P，数据点尺寸 s，数组 M.

输出. 待选采样点.

Step1. 生成连续密度场. 利用 KDE 对 P 进行拟合，将离散的点密度值转换成连续密度场.

Step2. 生成网格点. 用正交的网格线划分 P 的分布空间，使单位网格的宽度等于 s.

Step3. 网格点转化为待选采样点. 将网格点映射到连续密度场中，赋予网格点密度值.

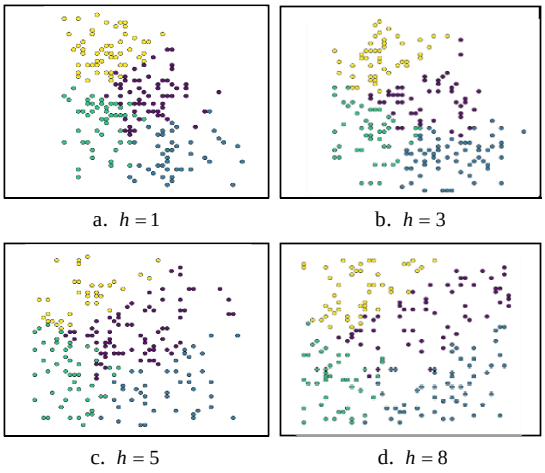
Step4. 将待选采样点序列存储在 M 中.

Step5. 算法结束.

KDE 是一种非参数概率密度估计方法，使用 KDE 对 n 个数据样本拟合，得到其概率密度分布的公式为 $f(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)$. 其中，$f$ 表示概

率密度函数；n 表示数据点的数量；表示核函数，本文采用高斯核；K 表示带宽.

带宽能够控制数据点分布的松散程 h 度，影响散点图去重叠后数据整体所占面积的变化. 如图 4 所示，过小的带宽会使得采样结果点分布过于聚集；过大的带宽会导致采样结果点分布过于松散. 本文使用 Silverman 带宽估计法[35]计算带宽的大小，计算公式为 $h=(4/3n)^{1/5}\sigma$；其中，n 表示数据点的数量，σ 表示数据点的距离标准差.



a. $h=1$ b. $h=3$ c. $h=5$ d. $h=8$

图 4 不同 h 下的采样结果

2.3.3 采样空间中基于密度的采样

在采样空间中，对待选采样点进行基于密度的采样前，首先需要通过归一化将待选采样点的密度值转化为采样概率，归一化公式为 $p(x)=d_x\Big/\sum_{i=1}^{n}d_i$；其中，$p$ 表示采样概率，n 表示数据点的数量，d_x 表示待选采样点 x 的密度值.

对待选采样点进行基于密度的采样后，得到采样结果点. 对每个待选采样点，采样概率越大，被选为采样结果点的概率也越大. 这种基于密度的采样方法充分地利用数据的密度信息保证了采样的质量，使生成的采样结果点更符合原数据的聚集性特征.

2.3.4 采样后处理

采样是在新构建的采样空间中进行的，采样的对象并不是原数据本身，因此采样结果点不包含原数据点的属性信息. 考虑 kD 树常被用于最近邻搜索，所以 DBOR 算法使用 kD 树进行采样后处理. 通过查找每个采样结果点在原始散点图中的最近邻点，将原始散点图中数据点的属性信息赋予采样结果点，实现采样结果与原数据之间的属性关联.

如图 5 所示，采样结果点 A 通过采样后处理查找到其在原始散点图中的最近邻点 B，由此建立了与 B 的关联，获得了 B 的类别属性.

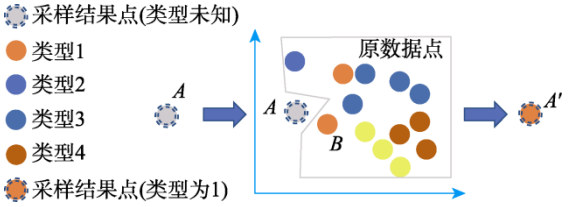
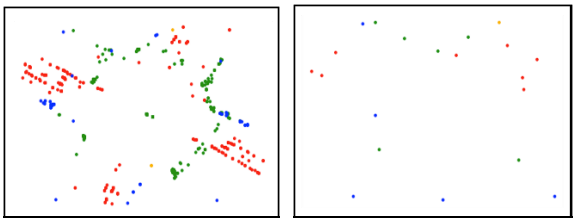


图 5 采样后处理示意图

2.3.5 数据拆分模块

待优化区域中可能存在一些相对稀疏的数据点，它们所处区域的密度相对较低，所以在采样空间对应区域中生成采样结果点的概率也较低，导致丢失散点图的局部细节，影响布局优化的准确性. 针对这一问题，本文在 DBOR 算法中设计了数据拆分模块，其效果如图 6 所示. 利用 KDE 密度评估的结果确定 DT，将密度值小于 DT 的数据点定义为稀疏点并加以保护，将其保留在去重叠后的散点图中. 数据拆分模块涉及 2 个参数：切分阈值(split threshold, ST)和 DT. 这 2 个参数的说明如表 1 所示.



a. 待优化区域 b. 筛选出的稀疏点

图 6 筛选稀疏点

表 1 数据拆分模块的参数说明

参数	确定参数的依据	功能
ST	原数据的密度热力图和散点图	确定 DT
DT	原数据的密度直方图和 ST	筛选稀疏点

(1) 确定 ST. 切分阈值的作用是确定 DT. 首先利用热力图对第 2.3.2 节中得到的连续密度场可视化生成密度热力图，然后将原始散点图与密度热力图进行视图叠加. 如图 7 所示，统计落在热力图颜色最浅区域内的点数量，将其记为 m，再由

$$\mathrm{ST}=\left\lfloor m/10^{\lfloor \lg m\rfloor}\right\rfloor\times 10^{\lfloor \lg m\rfloor} \tag{1}$$

得到切分阈值.

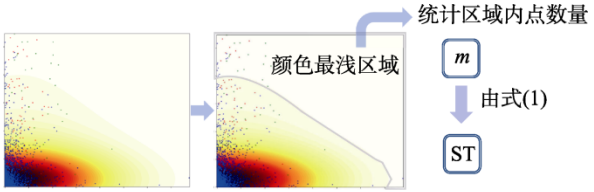


图 7　确定 ST 示意图

(2) 确定 DT. 首先使用直方图对第 2.3.2 节中得到的数据点密度值可视化生成密度直方图，然后对密度直方图进行迭代切分. 将第 x 次切分时直方图第 1 个区间的上限定义为 DT_x，区间内的点数量定义为 n_x. 首先判断 n_1 与 ST 的大小关系，如果 $n_1>\mathrm{ST}$，开启第 2 次切分，即把上一次切分得到的第 1 个区间内的数据作为下一次生成直方图的对象；然后判断 n_2 与 ST 的关系；不断重复上述过程，直到 $n_x\leqslant\mathrm{ST}$ 时停止. 此时，将 DT_x 记作 DT，将密度值属于 $(0,\mathrm{DT})$ 的数据点定义为稀疏点.

算法 3. 确定 DT.

输入. 散点图数据的密度值数组 d，暂存数组 b.

输出. DT.

$h=H(d)$　　//绘制数据的密度直方图

$b=\mathrm{Sort}(h_x)$　//将直方图横坐标按升序存到数组 b 中

$\mathrm{DT}=b_1$　　//将 DT 初始化为数组 b 的第 1 个元素

$d_{\mathrm{low}}=d[d\leqslant\mathrm{DT}]$　　//提取低密度数据

$n=\mathrm{Count}(d_{\mathrm{low}})$　　//统计数据点数目

while $(n>\mathrm{ST})$　// $n>\mathrm{ST}$ 时迭代切分直至满足条件

　$b=\mathrm{Sort}(H(d_{\mathrm{low}})_x)$　　//更新暂存数组的值

　$\mathrm{DT}=b_1$　　//更新 DT

　$d_{\mathrm{low}}=d[d\leqslant\mathrm{DT}]$

　$n=\mathrm{Count}(d_{\mathrm{low}})$

end while

2.4　时间复杂度

设 n 为数据点数量，m 为待优化区域中的数据点数量($m\leqslant n$)，s 为网格点数量，k 为采样数量($k<m$)，d 为数据维度(本文中 $d=2$).

本文方法由算法预处理(非必需步骤)和 DBOR 算法 2 部分组成. 算法预处理的时间复杂度为 $O(n\times\log n)$，在 DBOR 算法中，构建采样空间的时间复杂度为 $O(m\times s)$；基于密度采样的时间复杂度为 $O(m)$；采样后处理分为构建和查找，其中，构建 kD 树的时间复杂度为 $O(d\times m\times\log m)$，查找 k 个采样结果点最近邻的时间复杂度为 $O(k\times\log m)$，因此采样后处理的时间复杂度为 $O(d\times m\times\log m)$；数据拆分的时间复杂度为 $O(m)$. 所以，DBOR 算法的时间复杂度为 $O(m\times s+m+d\times m\times\log m+m)$，简化为 $O\left(\max(m\times s,d\times m\times\log m)\right)$.

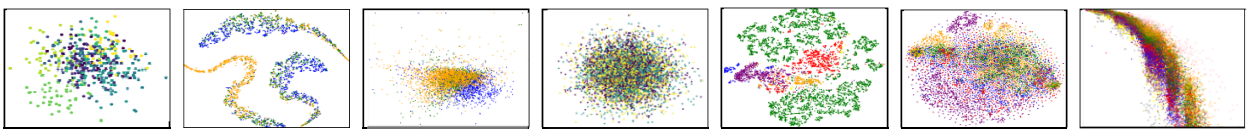
因为算法预处理是可选步骤，所以当需要算法预处理时，本文方法的整体时间复杂度为 $O\left(\max(m\times s,d\times m\times\log m,n\times\log n)\right)$；不需要算法预处理时，本文方法的整体时间复杂度即为 DBOR 算法的时间复杂度.

3　实验结果与分析

本文设计了 3 个实验，对本文方法进行定性评价和定量评估；进一步，通过案例证明其通用性；最后通过用户评估，验证其优越性. 当数据点圆心之间的距离小于数据点半径时，数据重叠；当数据点圆心之间的距离大于等于数据点直径时，数据不重叠.

3.1　去重叠效果对比

由于本文方法结合了采样与网格移动策略中规则网格划分的思想，因此将原始散点图、随机采样散点图、蓝噪声采样散点图、Gridfit 去重叠散点图、DGrid 去重叠散点图与通过本文方法得到的去重叠散点图(简称本文去重叠散点图)在视觉效果上进行对比，分别从相对区域密度、离群点保持、长宽比保持和视觉清晰度上进行描述. 本文使用 7 个数据集进行实验，其中，2 个由 Python 合成，分别是 Dpython1(简称 Dp1)和 Dpython2(简称 Dp2)；另外 5 个来自 Li 等[36]，分别是 Abalone(简称 Ab), Avila(简称 Av), Crowded_sourced(简称 Cr), Epileptic_seizure(简称 Ep)和 Diamonds(简称 Di). 数据集的规模为 500~53 931，具有不同的分布特征和重叠程度；使用不同的颜色为每个数据集中数据的类别进行编码. 7 个数据集如图 8 所示，它们的相关信息如表 2 所示.



a. Dp1　b. Ab　c. Av　d. Dp2　e. Cr　f. Ep　g. Di

图 8　7 个数据集

图 9 所示为 Di 数据集上整体与局部去重叠效果对比. 其中, 图 9a~图 9f 所示为整体去重叠效果, 使用黑色虚线框圈出了局部区域, 用编号①~编号⑥表示; 图 9g 所示为局部去重叠效果, 编号①~编号⑥分别与图 9a~图 9f 中的 6 个黑色虚线框对应.

表 2 数据集相关信息

名称	规模	数据类别	距离标准差	重叠率/%
Dp1	500	5	10	20.6
Ab	4 177	3	175	35.2
Av	5 672	3	89	40.3
Dp2	6 000	3	60	87.2
Cr	10 845	6	162	94.5
Ep	11 500	5	152	95.8
Di	53 931	8	192	99.2

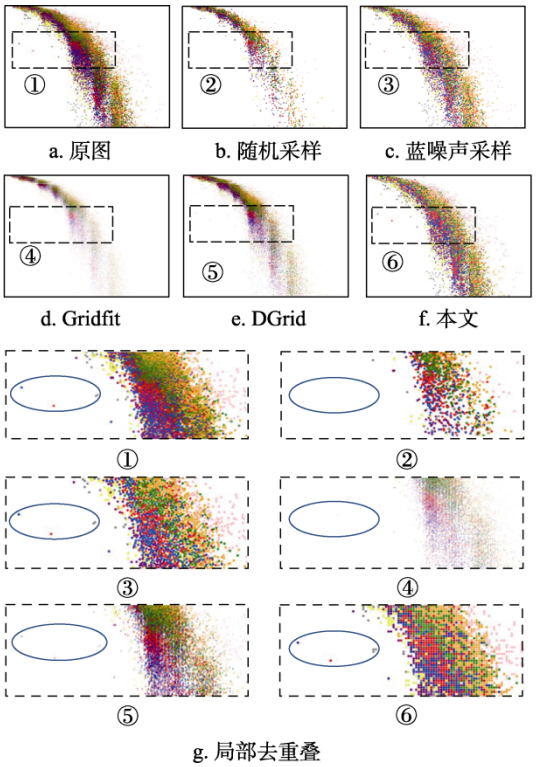


图 9 Di 数据集整体与局部去重叠效果对比

从图 9 可以看出: (1) 相对区域密度. 散点图的过度绘制会掩盖不同区域间的相对密度差异, 图 9g 中, ①③⑥的右下区域不能明显地看出密度差异, 而②④⑤的相应位置则能明显地看到白色空隙; 在反映相对区域密度差异上, 随机采样、Gridfit 与 DGrid 的表现较好. (2) 离群点保持. 图 9g 中, ①的蓝色椭圆框内有少量稀疏点, ③和⑥中也有, ②和④中则完全没有; 虽然能勉强感知到⑤的蓝色椭圆框内也有少量稀疏点, 但不是非常明显; 只有③和⑥对局部细节的保持比较好, 所呈现的信息空间更完整. (3) 长宽比保持. 长宽比指物体或图像的宽度与高度之间的比例关系, 散点图的长宽比反映了数据的空间关系, 对长宽比的保持能够体现散点图去重叠前后的扭曲程度. 图 9a~图 9f 中, 随机采样、Gridfit 与 DGrid 在宽度上的变化较大, 与原图相比更狭窄, 这分别是由随机采样的随机性、数据移动策略对大规模高重叠率的数据集的不适用性所导致的; 蓝噪声采样与本文方法对长宽比的维持则更好, 在视觉上较好地保留了原数据的空间关系. (4) 视觉清晰度. 主体与背景原则[37]认为, 图形与背景的对比越大, 图形的轮廓越明显, 则图形越容易被发觉. 图 9g 中, ④⑤对应 Gridfit 和 DGrid 虽然在原理上实现了去重叠, 但为了有足够的移动空间, 它们都缩小了数据点的尺寸, 降低了视觉清晰度, 需要不断地放大视图才能更清楚地看到单个数据点, 这无疑增加了可视分析的烦琐性; ⑥则在去重叠的基础上保证了与①中数据点尺寸相同, 视觉清晰度良好, 能够减少视觉识别过程心理任务的工作量[38]; ②和③中仍有重叠现象.

数据点完全平铺对于规模大且重叠度高的数据会占用更多的显示空间, 在有限的显示空间内, 为了让数据点有足够的移动空间, 不得不缩小数据点的尺寸. 因此, 经过 Gridfit 和 DGrid 去重叠后得到的数据点会比原数据点小得多; 而数据点越小, 人们对其感知程度就越弱, 并且当显示设备的性能较差时, 微小的数据点甚至会完全被人眼忽略. 本文方法可通过设定少于原数据个数的采样结果点数, 以提高算法的效率和显示空间的利用率, 在解决重叠问题的基础上保证视觉编码的一致性. 面对密度不均匀的数据时, 本文方法能够根据数据的密度信息调整采样概率, 使数据点数量与区域密度一致, 防止在高密度区域进行过多采样而导致信息冗余, 同时也避免在低密度区域欠采样导致信息丢失, 很好地捕捉到了数据的聚集性规律. 此外, KDE 的带宽能够限制散点图整体面积变化, 使散点图更符合数据的原始结构.

除了 Di 数据集外, 图 10 所示为本文方法在其余 6 个数据集上的整体与局部去重叠效果. 实验结果再次表明, 即便在未放大视图时, 本文方法仍具有良好的视觉清晰度, 很好地维持了原散点图的长宽比和视觉纹理, 能够避免具有分组特征的数据簇之间相互融合, 保证了清晰的边界, 较好地保留了原数据集的全局和局部特征.

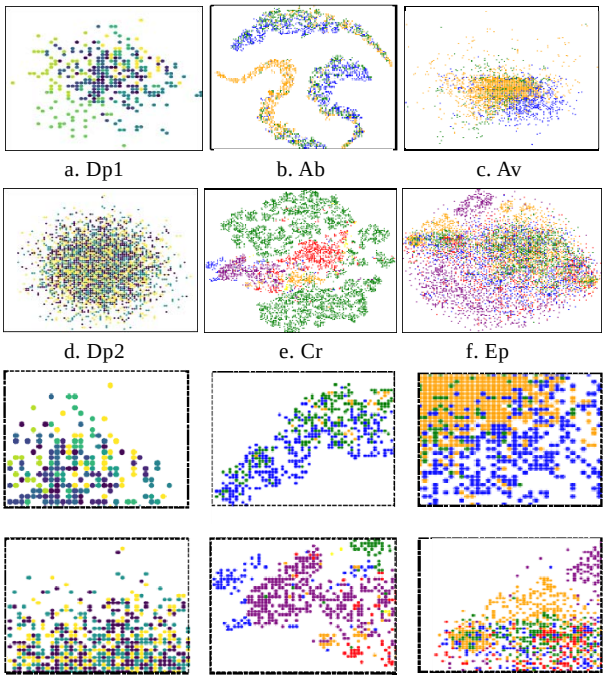


图 10　整体与局部去重叠效果

3.2　定量评价

Gridfit, DGrid 和本文方法均改变了原始散点图的布局，因此，结合指标对散点图变化前后的布局进行定量分析是非常有必要的. 受到 Li 等[36]工作的启发，本文选择了几个具有代表性的评估指标，并针对本文方法的特点相应地调整了计算方式. 设原始散点图为 S_{original}；由 Gridfit, DGrid 和本文方法处理得到的散点图分别为 S_{Gridfit}，S_{DGrid} 和 S_{ours}；S_{ours} 中的数据为 p_i，其在 S_{original} 中对应的最近邻点为 q_i；p_i 的集合为 $\{p_i\}_N$，点数为 N；q_i 的集合为 $\{q_i\}_N$，点数为 N；S_{original}，S_{Gridfit}，S_{DGrid} 和 S_{ours} 中由 q_i 围成的区域面积分别为 A_{original}，A_{Gridfit}，A_{DGrid} 和 A_{ours}.

评估指标如下:

(1) 位移距离(displacement distance, DD). 对于 S_{Gridfit} 和 S_{DGrid}，直接计算数据点移动前后圆心间的平均欧几里得距离作为 DD; 本文方法中，由于每个 p_i 都能够在 S_{original} 中找到其最近邻 q_i，因此计算 p_i 与 q_i 之间的平均欧几里得距离作为 DD. 其数值越小，表示对散点图全局结构的维持越好.

(2) 局部面积变化(local shape variation, LSV). 首先在 S_{ours} 中选取 x 个($x \geqslant 3$)数据点，得到 $\{p_i\}_x$ 并计算其围成的面积 A_{ours}；然后在 S_{original} 中找到 $\{q_i\}_x$ 并得到其围成的面积 A_{original}，将 $\left|A_{\text{ours}}-A_{\text{original}}\right|$ 作为 S_{ours} 的 LSV. 对于 Gridfit 与 DGrid，由 $\{q_i\}_x$ 得到 A_{Gridfit} 与 A_{DGrid}，将 $\left|A_{\text{Gridfit}}-A_{\text{original}}\right|$ 与 $\left|A_{\text{DGrid}}-A_{\text{original}}\right|$ 作为 S_{Gridfit} 与 S_{DGrid} 的 LSV 值; 不断地改变 $\{p_i\}$ 的选取方式，取多次 LSV 的平均值. LSV 数值越小，表示局部面积变化越小.

(3) 密度维持(density preservation, DP). 该指标体现方法对局部密度的维持情况. S_{Gridfit} 与 S_{DGrid} 的计算方法与 Li 等[36]的方法相同，此处不赘述. 对于 S_{ours}，Δq 的计算则略有不同: 假设 S_{ours} 共有 y 个数据点，首先计算 S_{ours} 中所有数据点 $\{p_i\}_y$ 的 KNN 平均距离的倒数作为局部密度，并将 p_i 按照 KNN 平均距离进行排序，统计不同 KNN 平均距离值对应的 p_i 的个数; 然后在 S_{original} 找到相应的 $\{q_i\}_y$，对其进行上述操作，最终得到 S_{ours} 的 Δq.

(4) 时间消耗(time cost, TC). 该指标统计不同方法在 7 个数据集上运行 30 次的平均时间消耗，单位为 s.

在 7 个数据集上, Gridfit, DGrid 和本文方法的定量分析结果如表 3 所示. 可以看出，对于 Dp1, Ab 和 Av 这 3 个数据集，本文方法在除 TC 外的指

表 3　各数据集上不同方法定量分析结果

数据集	方法	DD	LSV	DP	TC/s
Dp1	Gridfit	9.412	9.411	0.698	4.74
	DGrid	0.913	7.283	0.642	**1.79**
	本文	**0.416**	**5.436**	**0.615**	4.61
Ab	Gridfit	1.965	649.037	0.757	6.42
	DGrid	3.418	639.985	0.691	**4.32**
	本文	**0.813**	**638.484**	**0.632**	6.96
Av	Gridfit	2.392	196.093	0.815	7.64
	DGrid	1.255	194.221	0.726	**5.42**
	本文	**0.835**	**192.876**	**0.708**	12.13
Dp2	Gridfit	26.191	89.325	0.793	9.83
	DGrid	13.967	75.499	0.775	8.54
	本文	**1.453**	**56.156**	**0.651**	**7.19**
Cr	Gridfit	10.087	881.124	0.797	27.81
	DGrid	3.825	746.254	0.792	23.47
	本文	**2.478**	**413.773**	**0.683**	**12.11**
Ep	Gridfit	3.213	172.325	0.792	13.25
	DGrid	2.359	143.586	0.801	26.35
	本文	**1.641**	**123.119**	**0.768**	**12.17**
Di	Gridfit	12.762	168.872	0.882	45.21
	DGrid	10.767	110.452	0.816	51.32
	本文	**1.434**	**55.851**	**0.651**	**14.93**

注. 粗体表示最优值.

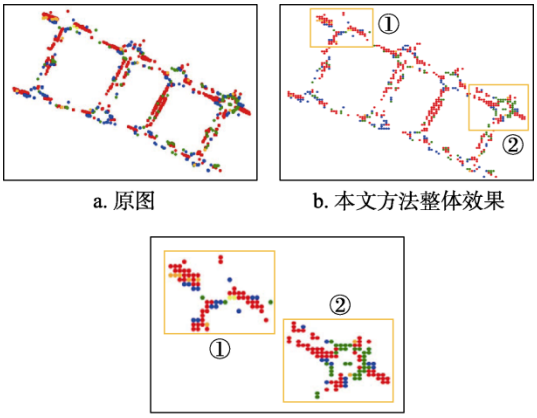
标上表现最好, DGrid 在 TC 指标上表现最好; 随着数据规模的增大, 本文方法的优势越来越明显, 在 Dp2, Cr, Ep 和 Di 这 4 个数据集的所有指标上都取得了最好的结果, 尤其是在 Di 数据集上, 本文方法的指标结果与其他 2 种方法的差异非常显著, 再次验证了本文方法的通用性. 这是因为当数据规模较小且重叠率不高时, Gridfit 与 DGrid 都有足够的空间来移动数据, 并且 DGrid 对数据的选择性移动也能够保证一定的效率; 随着重叠程度与分布范围的增加, 对可移动位置的判断与移动方式的确定, 影响了数据移动类算法在 TC 指标上的表现, 当邻近可用位置被占用时, 不得不将数据移动到更远的位置, 导致 Gridfit 与 DGrid 在 DD 指标上表现不佳; 对于 DP 指标, Gridfit 与 DGrid 没有考虑原数据的密度信息, 在移动数据点时破坏了数据原有的聚集性; 而本文方法利用原数据的密度特征构建采样空间, 基于密度的采样考虑了密集程度对空间分布特征的影响.

3.3 交通数据案例验证

采用 ChinaVis2023 挑战赛[39]提供的交通轨迹数据集, 验证本文方法对空间分布范围较大且密度不均匀的散点图的适用性. 在交通轨迹数据集中, 每项数据记录了不同时间戳下道路传感器接收到的车辆信息, 不同颜色代表不同类别的交通实体. 在删除了重复和异常的数据项并将其按时间维度分组后, 本文提取出 8 147 条数据, 并根据位置属性与类型属性构造了交通实体散点图. 如图 11 所示, 使用本文方法对原始交通散点图进行布局优化, 能够在解决重叠问题的同时尽可能地拟合原数据的空间分布特征, 有利于对散点图做定量分析.

3.4 用户评估

本文招募了 26 位参与者进行用户评估, 包括 13 位男性和 13 位女性, 年龄分布在 23~28 岁. 所有参与者都没有包括色盲、色弱在内的视觉障碍问题.



a. 原图 b. 本文方法整体效果

c. 本文方法局部效果

图 11 散点图效果对比

(1) 数据和实验描述

用户评估使用的数据集包含 kaggle 网站 3 类等级的用户信息共 3 000 条, 每条用户信息有 19 个属性维度. 经数据清洗后, 本文筛选出 2 694 条有效数据, 并基于自定义语义形成 4 种二元属性组合, 分别绘制不同属性组合下的原始散点图、Gridfit 去重叠散点图、DGrid 去重叠散点图和本文去重叠散点图(分别简称 A 类、B 类、C 类和 D 类).

本文设定 3 个可视分析任务, 通过对比用户在 4 类散点图上的任务完成情况, 证明本文方法的优越性.

(2) 任务设计

该数据集的可视分析任务设计如下:

任务 1. 相对数量排序. 要求用户在每种散点图上对 3 类数据点进行相对数量的排序.

任务 2. 识别相对密度. 要求用户依据每种不同的散点图对 3 类数据点按照可视感知结果进行相对密度的识别, 用户需识别出相对密度最大的数据点种类.

任务 3. 描述空间分布特征. 要求用户根据每种不同的散点图分析数据点的空间分布特征.

识图时, 用户在每种散点图上都有 20 s 的识图时间, 采用赋分制评估可视分析任务的完成情况. 正确完成任务赋 1 分, 反之赋 0 分.

(3) 评估结果与分析

表 4 所示为 26 名用户在 4 类散点图上的得分情况. 可以看出, 26 名用户在 D 类散点图下的单项任务得分和总得分均最高; 通过调查用户的认知

表 4 在不同散点图上可视化任务得分表

任务	类型	组合 1	组合 2	组合 3	组合 4	总分
1	A	10	8	5	11	34
	B	16	15	10	11	52
	C	14	15	16	12	57
	D	**24**	**23**	**24**	**26**	**97**
2	A	5	7	9	11	32
	B	10	11	13	12	46
	C	12	17	15	12	56
	D	**19**	**22**	**26**	**22**	**89**
3	A	0	0	0	0	0
	B	5	7	4	8	24
	C	10	7	9	11	37
	D	**26**	**26**	**24**	**25**	**101**

注. 粗体表示最优值.

过程得知, A 类散点图中存在严重的散点遮盖问题, 且通过放大视图的方法也无法解决，导致用户无法感知到不同种类数据点的相对数量、相对密度以及空间分布特征; B 类散点图与 C 类散点图虽然在表现上好于 A 类散点图，但是由于在处理过程中没有充分地利用原数据自身的密度信息，使得生成的布局破坏了数据的聚集性，因此多类数据点之间的相对数量差异在视觉上被减弱，对用户造成了干扰；使用本文方法处理前后的散点图对比如图 12 所示，用户评估结果再次验证了本文方法在去重叠方面的优越性.

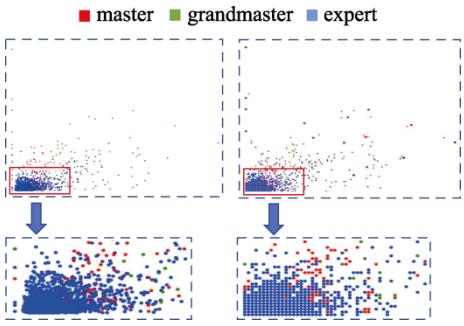


图 12　本文方法处理前后的散点图对比

4　总　结

本文针对散点图的数据重叠问题提出基于密度的散点图去重叠可视化方法，其充分地利用数据密度特征，并结合采样与网格划分思想提出去重叠的 DBOR 算法；并优化 DBOR 算法的输入，为空间分布不均匀的数据设计算法预处理——利用 DBSCAN 筛选出离群点，并将其之外的区域作为DBOR算法的输入. 实验结果表明，本文方法对于不同领域的数据集均有较强的适用性；在规模大、重叠率高的数据集上，与已有的散点图去重叠方法相比，该方法能够更为恰当地拟合原数据的空间分布特征，得到的散点图优化结果具有更好的视觉清晰度，且在布局优化的过程中具有更高的空间利用率和处理效率. 相比已有方法，本文方法更好地平衡了散点图去重叠后的可视化清晰度和布局准确性.

在合适的参数组合下，DBSCAN 聚类得到的离群点之间重叠程度不高甚至是完全不重叠的，并且在获得去重叠散点图后，可以利用多视图协同的方法将其与原始散点图进行对比观察，获得更全面的可视化效果. 虽然对于离群点的处理并不是本文的工作重心，但是在极端情况下，离群点也可能非常致密甚至重叠，本文缺少对这类情景的研究. 未来将提高本文方法在不同数据结构下的扩展性. 由数据拆分筛选出的稀疏点虽然很好地保留了散点图的局部特征，但其所在位置的附近可能会形成新的采样结果点，所以稀疏点可能会与采样结果点之间出现轻微重叠. 但是，稀疏点的个数可以通过 DT 控制，且在其附近产生采样结果点的概率也很低，综合看来，这种情况对全局可视化效果的影响是可以忽略不计的. 此外，本文在构建采样空间时使用规则网格划分的方式，未来将尝试使用非规则划分方式[2]进一步优化散点图的去重叠效果.

参考文献(References):

[1] Sarikaya A, Gleicher M. Scatterplots: tasks, data, and designs[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 402-412

[2] Zhao Ying, Xiu Yuhong, Tang Tao, *et al*. Incomplete overlapping removal algorithm for scatterplots[J]. Journal of Software, 2023, 34(2): 945-963(in Chinese)
(赵颖, 秀昱宏, 唐涛, 等. 一种非完全的散点图去重叠算法[J]. 软件学报, 2023, 34(2): 945-963)

[3] Chen H D, Chen W, Mei H H, *et al*. Visual abstraction and exploration of multi-class scatterplots[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1683-1692

[4] Yuan J, Xiang S X, Xia J Z, *et al*. Evaluation of sampling methods for scatterplots[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1720-1730

[5] Tao W B, Hou X L, Sah A, *et al*. Kyrix-S: authoring scalable scatterplot visualizations of big data[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 401-411

[6] Chen F T, Piccinini L, Poncelet P, *et al*. Node overlap removal algorithms: an extended comparative study[J]. Journal of Graph Algorithms and Applications, 2020, 24(4): 683-706

[7] Ellis G P, Bertini E, Dix A. The sampling lens: making sense of saturated visualisations[C]. //Proceedings of The ACM Conference on Human Factors in Computing Systems, New York: ACM Press, 2005: 1351-1354

[8] Zhang Huijie, Wang Rong, Chen Bin, *et al*. Dynamic identification of urban functional areas and visual analysis of time-varying patterns based on trajectory data and POIs[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(9): 1728-1740(in Chinese)
(张慧杰, 王蓉, 陈斌, 等. 基于轨迹和兴趣点数据的城市功能区动态识别与时变规律可视分析[J]. 计算机辅助设计与图形学学报, 2018, 30(9): 1728-1740)

[9] Raidou R G, Gröller M E, Eisemann M. Relaxing dense scatter plots with pixel-based mappings[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(6): 2205-2216

[10] Yin Jingjing. Research and realization on visualization method of multidimensional data[D]. Shenyang: Northeastern Univer-

sity, 2019(in Chinese)
(殷晶晶. 多维数据可视化方法研究与实现[D]. 沈阳: 东北大学, 2019)

[11] Quadri G J, Rosen P. Modeling the influence of visual density on cluster perception in scatterplots using topology[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1829-1839

[12] Chen X, Zhang J, Fu C W, *et al*. Pyramid-based scatterplots sampling for progressive and streaming data visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 593-603

[13] Li Z Y, Wang T, Wang M, *et al*. Construct boundaries and place labels for multi-class scatterplots[J]. Journal of Visualization, 2022, 25(2): 407-426

[14] Wang Y Z, Wang D, Pang W, *et al*. A systematic density-based clustering method using anchor points[J]. Neurocomputing, 2020, 400: 352-370

[15] Song Zhao. Density field based large-scale time series data visual analysis system[D]. Jinan: Shandong University, 2022(in Chinese)
(宋钊. 基于密度场的大规模时序数据可视分析系统[D]. 济南: 山东大学, 2022)

[16] Zinsmaier M, Brandes U, Deussen O, *et al*. Interactive level-of-detail rendering of large graphs[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2486-2495

[17] Li C H, Baciu G, Wang Y Z, *et al*. DDLVis: real-time visual query of spatiotemporal data distribution via density dictionary learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 1062-1072

[18] Kokate U, Deshpande A V, Mahalle P N. Density-based clustering method for trends analysis using evolving data stream[J]. International Journal of Synthetic Emotions, 2020, 11(2): 19-36

[19] Ester M, Kriegel H P, Sander J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[C] //Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI Press, 1996: 226-231

[20] Matejka J, Anderson F, Fitzmaurice G. Dynamic opacity optimization for scatter plots[C] //Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York: ACM Press, 2015: 2707-2710

[21] van Onzenoodt C, Huckauf A, Ropinski T. On the perceptual influence of shape overlap on data-comparison using scatterplots[J]. Computers & Graphics, 2020, 90: 169-181

[22] Mayorga A, Gleicher M. Splatterplots: overcoming overdraw in scatter plots[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(9): 1526-1538

[23] Trautner T, Bolte F, Stoppel S, *et al*. Sunspot plots: model-based structure enhancement for dense scatter plots[J]. Computer Graphics Forum, 2020, 39(3): 551-563

[24] Guo T, Feng K Y, Cong G, *et al.* Efficient selection of geospatial data on maps for interactive and visualized exploration[C] // Proceedings of the International Conference on Management of Data. New York: ACM Press, 2018: 567-582

[25] Zhao J H, Liu X, Guo C, *et al*. Phoenixmap: an abstract approach to visualize 2D spatial distributions[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(3): 2000-2014

[26] Bertini E, Santucci G. By chance is not enough: preserving relative density through nonuniform sampling[C] //Proceedings of the 8th International Conference on Information Visualisation. Los Alamitos: IEEE Computer Society Press, 2004: 622-629

[27] Hu R Z, Sha T K, van Kaick O, *et al*. Data sampling in multi-view and multi-class scatterplots via set cover optimization[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 739-748

[28] Keim D A, Hao M C, Dayal U, *et al.* Generalized scatter plots[J]. Information Visualization, 2010, 9(4): 301-311

[29] Yan D M, Guo J W, Wang B, *et al*. A survey of blue-noise sampling and its applications[J]. Journal of Computer Science and Technology, 2015, 30(3): 439-452

[30] Dwyer T, Marriott K, Stuckey P J. Fast node overlap removal[C] //Proceedings of the 13th International Symposium on Graph Drawing. Heidelberg: Springer, 2006: 153-164

[31] Misue K, Eades P, Lai W, *et al.* Layout adjustment and the mental map[J]. Journal of Visual Languages & Computing, 1995, 6(2): 183-210

[32] Gansner E R, Hu Y F. Efficient, proximity-preserving node overlap removal[J]. Journal of Graph Algorithms and Applications, 2010, 14(1): 53-74

[33] Keim D A, Herrmann A. The Gridfit algorithm: an efficient and effective approach to visualizing large amounts of spatial data[C] //Proceedings of the Visualization. Los Alamitos: IEEE Computer Society Press, 1998: 181-188

[34] Hilasaca G M, Marcílio-Jr W E, Eler D M, *et al*. A grid-based method for removing overlaps of dimensionality reduction scatterplot layouts[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(1): 1-14

[35] Silverman B W. Density estimation for statistics and data analysis[M]. London: Chapman and Hall, 1986

[36] Li Z Y, Shi R Z, Liu Y, *et al*. Dual space coupling model guided overlap-free scatterplot[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 657-667

[37] Chen Wei, Shen Zeqian, Tao Yubo, *et al*. Data visualization[M]. 2nd ed. Beijing: Publishing House of Electronics Industry, 2019: 49-53(in Chinese)
(陈为, 沈则潜, 陶煜波, 等. 数据可视化[M]. 北京: 电子工业出版社, 2019: 49-53)

[38] Zhang Huijun. Research on the design model of visual analytics systems based on complex cognition theory[D]. Taiyuan: Taiyuan University of Technology, 2022(in Chinese)
(张慧军. 基于复杂认知理论的可视分析系统设计模型研究[D]. 太原: 太原理工大学, 2022)

[39] ChinaVis. The 10th China visualization and visual analytics conference[EB/OL]. [2023-12-21]. http://www.chinavis.org/ 2023/challenge.html(in Chinese)
(ChinaVis. 第十届中国可视化与可视分析大会[EB/OL]. [2023-12-21]. http://www.chinavis.org/2023/challenge.html)