基于聚类感知增强的信息图表数据集可视分析系统

陈嘉澍[1)]，吴子健[1)*]，王立鹏[2)]，赵涓涓[3)]，刘世霞[1)]

[1)](清华大学软件学院 北京 100084)
[2)](复杂系统控制与智能协同全国重点实验室 北京 100074)
[3)](太原理工大学软件学院 太原 030024)
(zijian@mail.tsinghua.edu.cn)

摘 要：为了帮助专家更高效地分析和构建信息图表数据集，提出一种基于聚类感知增强的数据集可视分析系统。系统包括控制视图、概览视图和样本视图 3 个主要模块，支持数据分析人员对数据集进行层次化探索，并对样本进行筛选和检查。首先通过集成类簇形状优化方法和动态颜色分配方法改进用户对类簇的感知，提升分析的准确率；然后引入层次化探索机制，使用户能够系统性地探索、分析大规模数据集。在信息图表数据集上开展的实际案例研究表明，所提系统在排除非目标样本、探索信息图表设计空间和针对性地补充数据方面取得显著效果，采样评测下 100%的样本为目标样本，且总结出的设计空间能够覆盖约 94%的采样样本，验证了该系统能够帮助数据分析人员更直观地理解和分析信息图表数据集，构建高质量的数据集。

关键词：聚类结果可视化；聚类感知；可视分析；信息图表
中图法分类号：TP391.41 **DOI**：10.3724/SP.J.1089.2025-00255

Perception-Enhanced Visual Analysis System for Analyzing Infographic Datasets

Chen Jiashu[1)], Wu Zijian[1)*], Wang Lipeng[2)], Zhao Juanjuan[3)], and Liu Shixia[1)]

[1)](*School of Software, Tsinghua University, Beijing* 100084)
[2)](*National Key Laboratory of Complex System Control and Intelligent Agent Cooperation, Beijing* 100074)
[3)](*School of Software, Taiyuan University of Technology, Taiyuan* 030024)

Abstract: We propose a perception-enhanced visual analysis system for constructing and exploring high-quality infographic datasets. The system consists of three modules: control view, grid view and sample view, supporting hierarchical exploration and sample filtering. The system combines convexity-based layout optimization with dynamic color assignment to enhance cluster perception and improve analytical accuracy. A hierarchical exploration workflow then guides users from overview to detail, allowing them to filter out non-target samples, interrogate design variations, and identify gaps for targeted data augmentation. Case studies show the system effectively excludes irrelevant samples, explores the design space, and augments data. Evaluation results reveal that 100% of the sampled data are target samples, and the design space covers approximately 94% of the samples, demonstrating the system's ability to help analysts better understand and build high-quality infographic datasets.

Key words: visualization methods for clustering results; cluster perception; visual analytics; infographic

收稿日期: 2025-07-05；修回日期：2025-09-03。**基金项目**：国家自然科学基金（U21A20469）。**陈嘉澍**（1999—），男，硕士研究生，主要研究方向为信息可视化、面向机器学习的可视分析；**吴子健**（1987—），男，硕士，工程师，论文通信作者，主要研究方向为计算机图形学、数据质量提升与可视分析；**王立鹏**（1994—），男，硕士，工程师，主要研究方向为智能感知、高质量样本集构建；**赵涓涓**（1975—），女，博士，教授，博士生导师，CCF 会员，主要研究方向为智能信息处理、图像识别；**刘世霞**（1974—），女，博士，教授，博士生导师，CCF 会员，主要研究方向为可解释机器学习、文本可视分析、文本挖掘。

信息图表是可视化技术的重要分支，通过融合图像符号、文字和数据等多模态元素，实现信息的直观化表达和高效传递。相关研究表明，信息图表能够有效地提升信息传递的效率，使阅读者在短时间内获取和理解大量信息[1]。信息图表已经在教育和新闻传媒等多个领域得到广泛应用，其有 3 个重要特点：

（1）视觉引导性。通过精心设计的色彩风格、图形元素和空间布局，信息图表能够建立清晰的视觉层次，引导用户快速地获取关键信息。

（2）简洁性。信息图表通常采用简洁、概括的表达方式，在有限空间内高效、直观地呈现信息的核心内容，降低用户的认知负荷。

（3）跨文化性。作为一种视觉化的交际方式，信息图表具有突破语言和文化障碍的普适性[2]，不同文化的阅读者都能在信息图表中获取信息。

考虑信息图表的重要作用和广泛使用，研究人员开始探索信息图表的自动化生成方法[3]。虽然现有工作在使用深度学习模型理解和生成信息图表上取得了一些进展[4-7]，但是高质量的信息图表非常缺乏，深度学习模型缺少足够的训练数据，严重限制了相关模型性能的提升。在实际中，高质量信息图表数据集的构建十分困难，需要不同领域、不同类型的图表以覆盖各种场景；同时，在色彩风格、图形元素和空间布局等方面需要多样性，避免模型在训练过程中产生偏差。

信息图表具备相对明确的设计空间参数，这些参数能够在聚类结果中被有效地呈现。例如，在信息图表中，图表类型包含柱状图、饼状图和线型图等，布局结构包含并列布局和重叠布局等，颜色风格包含单色、双色和多色等。因此，一种常规的数据集构造方式是首先通过各种途径收集高质量的各类信息图表；然后提取特征进行聚类，利用可视分析系统探索并理解聚类结构，总结对应的设计空间，有针对性地筛选有代表性的样本、剔除重复的样本、补充缺少的类型样本，构建高质量的信息图表数据集。

目前，针对数据集的可视分析系统中，通常使用散点图、网格可视化等聚类结果可视化[8-9]展示数据的整体结构。在使用这类系统探索的过程中，用户需要准确地识别、区分类簇来理解类簇之间的关系，才能高效、正确地完成后续的分析和调整，这种对类簇识别、区分和理解的能力通常被称为聚类感知。然而，现有的聚类结果可视化方法对于聚类感知的考虑并不充分：一是缺少对类簇凸性的考虑，根据格式塔理论[10]，形状凸性较差的类簇容易被误解为多个类簇；二是颜色设计存在局限，特别是在处理大规模层次化数据时会生成接近或不和谐的颜色，干扰用户对类簇的理解与判断。因此，对类簇的形状、颜色等进行改进，一方面能够让各个类簇更易被感知为一个整体；另一方面，能够让不同的类簇更易被区分，并启发类簇间的关系，帮助用户识别、区分和理解类簇，增强用户的聚类感知。

基于以上思路，本文提出一种基于聚类感知增强的信息图表数据集可视分析系统。该系统集成类簇的形状优化方法和动态颜色分配方法，对网格可视化进行优化，增强用户的聚类感知，提高分析的效率和准确率；基于该系统，用户可以对数据集进行高效的探索和分析，并在样本视图中检查对应的数据，总结数据集的设计空间。最后通过案例研究，验证了该系统在分析和构建信息图表数据集的场景中，能够有效地帮助专家排除非研究目标样本，并探索数据集的设计空间构建高质量的信息图表数据集。

1　相关工作

本文提出的可视分析系统集成了基于形状和颜色的聚类感知增强方法，下面从基于形状的聚类感知增强方法、基于颜色的聚类感知增强方法和数据集分析系统 3 个方面论述相关工作。

1.1　基于形状的聚类感知增强方法

基于形状的聚类感知增强的目的是改进聚类结果可视化中类簇的凸性，减少用户对类簇的误解，核心是确定合适的凸性度量方法。基于“凸集”的数学定义[11]，研究人员提出多种凸性度量方法，根据不同的计算方式，可以分为基于面积的凸性度量方法和基于轮廓的凸性度量方法。

（1）基于面积的凸性度量方法依靠几何图形的面积衡量凸性。最常用的方法是计算图形面积与其凸包面积的比值[12-13]。一些方法使用图形内部最大凸多边形（称为凸骨架）的面积与图形面积或凸包面积之间的比值作为凸性的度量[14-15]。Rosin 等[16]使用最大重叠凸图形替代凸骨架，解决了基于面积的方法对细长形状尤为敏感的问题。还有的方法从凸性的统计定义出发，通过估计连线落于子集内部的概率度量凸性[17-18]。Held 等[19]使用形状内任取 2 点连线完整落入区域内部的概率作为凸性度量，可以对 2 点及其连线上任意一点进行遍历计算，称为三元组比例。

（2）基于轮廓的凸性度量方法通过几何图形的边界轮廓特征计算凸性指标。一种基础方法是基于欧几里得空间下凸包周长与图形实际周长的比值衡量凸性，称为周长比例[20]。Žunić 等[14]使用包围盒周长与 L_1 周长之间的比值衡量凸性，选取任意直角坐标系下最小的比值作为凸性得分。

Zhou 等[21]通过分析发现，虽然不存在一种凸性度量符合所有用户的感知，但是用户对凸性度量的偏好可以分为 2 类：约 70%的用户偏好基于面积的方法，三元组比例[18]更符合他们的感知；约 30%的用户偏好基于轮廓的方法，周长比例[20]更符合他们的感知。基于此，Zhou 等[21]提出新的类簇形状优化方法，充分地考虑了类簇的近邻性、紧凑性和凸性。本文使用该方法改善网格可视化类簇的凸性，通过形状增强用户的聚类感知。

1.2 基于颜色的聚类感知增强方法

基于颜色的聚类感知增强方法的目的是通过色彩编码提升信息传输的有效性[22]。针对非层次化结构，现有方法在颜色的可区分性[23-25]、颜色的和谐性[26-29]，以及数据的空间分布[30-32]上进行了大量的工作；但是，随着数据量的增大，类别的数量也随之提升。为了满足更多类别需求，现有的聚类结果可视化方法采用层次化技术处理大规模数据[33-34]，进一步要求颜色分配方法能够适应层次化数据。

为了适应层次化数据，早期的研究为层次结构中的每个类别分配静态的颜色[35-36]。Tree colors[35]将色环分为若干个范围，并为树的不同分支分配颜色；随着层级的增加，每个分支的色相范围逐渐缩小，而饱和度增加。这种方法在整个层次结构中生成了多样但一致的颜色分配结果。然而，随着层级的增加，类别数量呈指数级增长，这些方法在颜色可区分性方面仍然面临可扩展性问题。为此，后续研究采用动态的颜色分配策略，仅在探索过程中为可见数据分配颜色[37-38]，更好地利用了色彩空间，提高了颜色分配结果的整体质量。

尽管这些方法在一定程度上改进了颜色分配的结果，但仍然存在 2 个问题：（1）由于它们未同时考虑颜色可区分性、和谐性和空间分布，因此难以在每层生成高质量的颜色分配结果；（2）其父类和子类颜色之间的严格约束导致颜色的可区分性不足。针对这些问题，Chen 等[39]提出一种动态颜色分配方法，在进行多目标优化时，还要考虑颜色的可区分性、和谐性和对数据空间分布的适应程度，以生成高质量的颜色分配结果；同时，确保不同层次间颜色的一致性，使得子类别的颜色与对应的父类别在视觉上接近，保持用户在层次化数据探索过程中思维的连贯性。本文使用该方法改善网格可视化颜色分配，通过颜色增强用户的聚类感知。

1.3 基于聚类结果可视化的数据集分析方法

在数据集分析中，散点图和网格可视化是最常用的 2 类聚类结果可视化方法。

在散点图中，每个样本被表示为一个点，点的位置根据样本的属性或降维结果确定，点的颜色编码样本类别标签。散点图能够帮助用户高效、直观地检查样本及样本间的关系，被广泛使用于各类数据集的分析评估中。Chen 等[40]采用散点图结合点线图、柱状图，帮助用户分析有标和无标样本，基于半监督学习方法提高模型的性能；Yang 等[41]在散点图中加入样本图片，帮助用户选择合适的样本子集和学习器，提升少样本学习模型的效果；为了更方便地比较模型，Zhang 等[42]在系统中使用多个小型散点图对展示产生不同预测的原因。为了解决散点图中存在元素重叠和空间浪费的问题，一些研究人员提出网格可视化方法。该方法采用网格布局将样本内容放置在格点内，避免样本间的重叠并高效地利用空间，能够有效地帮助专家发现异常样本并改进异常检测模型。但是，随着数据量的不断增大，网格中每个样本的空间不断变小，阻碍了用户的探索和分析。TargetVue[43]使用三角形网格布局展示社交网络中的用户样本。Chen 等[33]提出一种分层网格可视化支持对大量样本的探索，解决网格布局可扩展性的问题：首先采样出关键的样本生成第 1 层布局，并将剩余的样本关联在其近邻下方；然后允许用户在探索过程中不断地选择感兴趣的部分深入检查，准确地发现训练数据中的分布外样本。基于该方法，Uni-Evaluator[44]在网格可视化中展示筛选出的样本子集，并在网格中标记目标位置以支持目标检测和分割任务。

以上方法均是利用聚类结果可视化，有效地帮助用户对数据集进行分析。然而，这些方法未充分地考虑聚类感知，用户在分析的过程中可能产生错误，影响解决问题的效果。因此，本文提出一种聚类感知增强的分析系统，结合凸性和颜色上对聚类感知的增强，灵活地整合现有的可视化方法，帮助用户更好地分析信息图表数据集中的问题、构建高质量的数据集。

2 需求分析

在开始设计系统前，本文与 2 位数据分析领域

的专家合作，通过访谈整理了他们的 4 条需求。

R1：支持对数据集的层次化探索。在处理大规模数据集时，数据分析人员需要同时把握数据的整体分布和局部细节。考虑到大规模数据集通常被组织为层次结构，因此系统应该支持自顶向下的层次化探索，允许用户从全局视图逐步深入到感兴趣的局部区域，并在不同层次间自由切换，实现对数据集的全面理解。

R2：提供样本筛选和检查的功能。在数据集分析和构建的过程中，数据分析人员需要筛选并检查特定位置的数据样本。一方面，检查类簇的样本有助于分析人员提取数据的模式；另一方面，检查边界样本和异常样本有助于分析人员发现数据集中存在的数据问题。系统应该提供灵活的样本筛选功能，并允许用户详细地查看单个样本的信息，确保数据集分析的准确性和可靠性。

R3：改进类簇的形状和颜色，增强用户的聚类感知。聚类感知指用户辨别和感知类簇的能力，形状和颜色都会影响数据分析人员对于类簇的判断。系统应当集成聚类感知增强方法，基于形状改进类簇的凸性，帮助数据分析人员准确地判断整体的结构；同时，基于颜色生成可区分、和谐的颜色分配结果，并适配层次化数据，保持分析人员探索中思维的连贯性。

R4：实现简洁有效的数据分析交互。系统的交互应该简洁、直观，能够让数据分析人员快速上手，有效地提高数据分析的效率。例如，系统应该支持数据集的切换、选择区域深入探索、返回上一层、筛选样本批量展示、查看选中样本信息等基本的交互功能。

3　本文系统

综合考虑以上需求，本文以迭代开发的方式进行系统研发，并在此过程中与专家保持联系和沟通，确保系统功能能够满足需求。本文提出基于聚类感知增强的信息图表数据集分析系统，通过集成类簇的形状优化方法和动态颜色分配方法增强用户的聚类感知，提高探索、分析信息图表数据集的准确性和效率。

3.1　系统数据与方法

本文系统的底层支撑包括数据和方法 2 部分。（1）数据部分。使用包含文本/图片格式的原始数据及其标签和特征信息，特征信息为可选项。通常，特征可以通过相关领域的专家设计并提取，也可以通过预训练模型进行提取。数据进入数据处理模块后，采用广泛使用的多模态模型 CLIP[45]提取图片的高维特征；然后进行层次聚类和数据降维，采用通用的层次化 *K*-means 算法[46-47]作为默认的层次聚类方法，可增量式 *t*-SNE 算法[34]作为默认的降维算法。（2）方法部分。集成了类簇形状优化方法[21]和动态颜色分配方法[39]，同时改进网格可视化类簇的形状和颜色，增强用户的聚类感知，提高用户探索、分析数据集的准确率和效率。

3.2　可视化设计

本文系统的界面如图 1 所示，包含控制视图、概览视图和样本视图。

（1）控制视图。为用户提供数据集选择的接口并展示基本信息。该视图中，从上到下分为上方、中间和下方 3 块，分别对应不同的功能需求。

a. 上方进行数据集的选择，数据分析人员可以根据自己的需求选择对应的数据集，并在完成 1 个阶段的分析任务后进行切换。该部分同样提供了数据集相关的信息，包括数据集的样本数量、类别数量和基本的描述，使用户能够初步地了解切换出的数据集。

b. 中间展示当前界面数据的类别分布，主要展示层次化数据的 2 层类别。对于展示的数据，首先检查其对应的类簇，并在层次化聚类获得的类簇树上逐步展开，直到类簇数目达到预期的范围；随着用户对数据的深入探索，类簇也会逐渐深入，直到抵达类簇树的叶子节点。该过程与用户的探索思路一致，能够逐步深入并最终了解数据的全貌。例如，第 1 层可能展示 2 个类簇：柱状图构成 1 个类簇，饼状图构成一个类簇；而当选择柱状图簇并深入探索到第 2 层时，用户可以进一步发现并探索分组柱状图、堆叠柱状图等不同的类簇。

采用 2 层的层次化类别，是希望帮助用户更全面地了解数据在层次结构中的多级关系，以更有效地分析出数据中潜在的模式和问题。控制视图中数据类别分布呈现如图 2 所示，表现为 2 条带状图，色带的长度分别对应该类簇下样本的数量。其中，第 1 层的类簇数目更少，颜色之间用白色的空隙隔开；第 2 层则是将第 1 层的类簇依次展开的结果，大类间同样用白色空隙隔开。基于这样的可视化设计，用户能够准确地理解数据的类别分布和类簇之间的层级关系。

c. 下方通过信息图表的图片展示用户鼠标位置样本的详细信息。用户可以随机地检查大量的样本，总结数据中潜在的模式。系统提供了图片展示

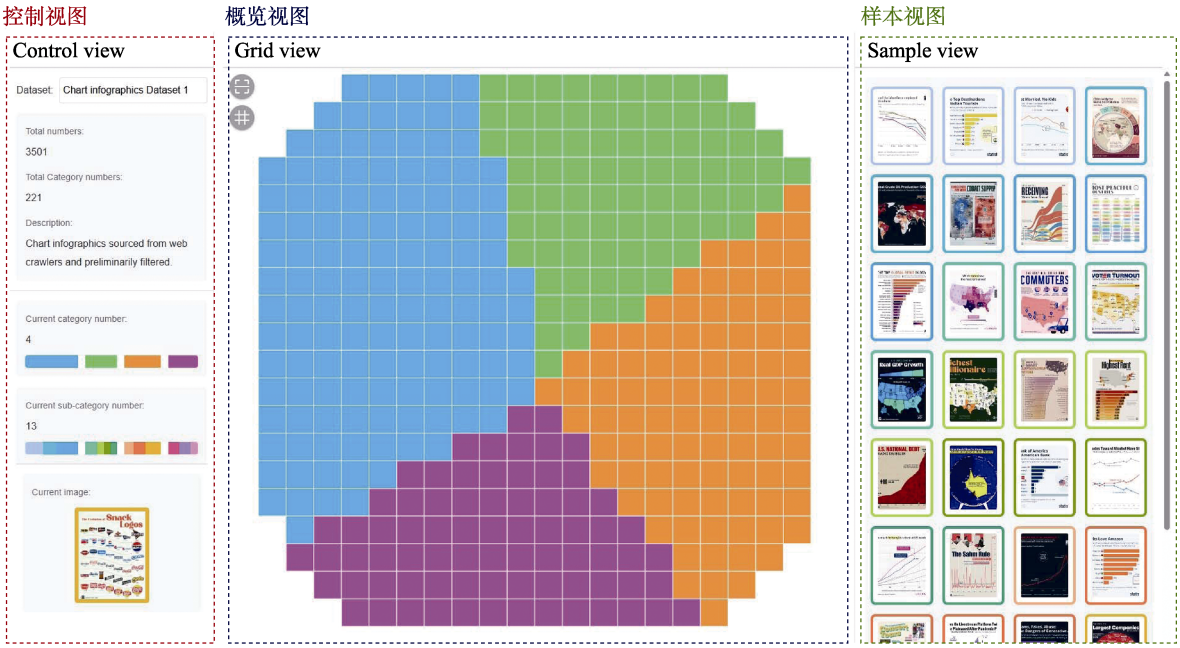


图 1 本文系统界面

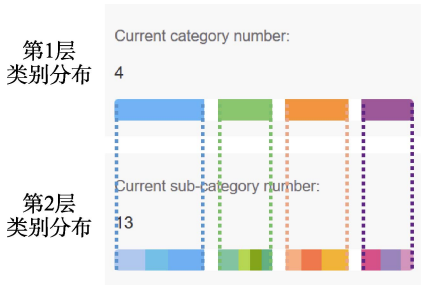


图 2 控制视图中数据类别分布呈现

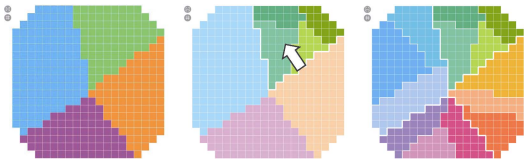


图 3 网格可视化的不同状态

的交互，用户可以点击图片查看具体细节。

（2）概览视图。为用户展示数据的具体分布，并支持数据分析人员进行层次化的探索。概览视图的核心是一个多功能的网格可视化，其规模可以根据实际需求进行调整，通常使用 20×20 或 30×30，能够在界面上清晰地展示丰富的数据样本。网格可视化与左侧控制视图中的类别分布部分一致，展示数据样本多层类别信息。

如图 3 所示，网格可视化分为 3 种状态。图 3a 所示为通常状态，展示第 1 层类别的结果，不同的颜色表示不同的类簇，其中的颜色与类别分布展示的颜色一致。图 3b 所示为当鼠标悬停在某一块区域上时，系统会展示该区域的第 2 层类别情况，并对该区域进行如下设计：为了增强用户对该区域的专注，其他的区域被灰色的蒙版覆盖，以更突出选中区域。图 3c 展示第 2 层类别的分布情况，其中的颜色同样与类别分布中一致。用户可以点击左上角的井型按钮在 2 种模式下进行切换。考虑包含的子类簇数量较多，系统在子类簇、类簇之间均增加了白色的轮廓线，以帮助用户定位类簇和子类簇边界。

概览视图还为用户的层次化探索提供支持。系统采用现有的层次化聚类结果可视化方法[34]，首先从处理好的数据集中采样顶层数据，然后将剩余的数据依据所计算的相似度关联在最相似的样本下；当用户在系统中选择某一块区域深入探索时，系统提取出选中的数据样本和关联在这些样本之下的数据样本，对关联的样本进行重采样，展示下一层的数据。其中，系统支持使用多种不同的采样方法[48]灵活地应对各类场景。例如，随机采样能够最好地保持数据的分布不变，蓝噪声采样能够保证数据的完整性，分布外数据采样方法更容易发现数据集中的离群点。本文默认采用随机采样，以保证分布一致。网格可视化中进行层次化探索的流程如图 4 所示。用户通过点击左上角的“检查”按钮进入选择状态，其中，整个网格被覆盖在一层灰色的蒙版下，表示可以拖动的区域，用户在区域中随意拖动矩形选择自己感兴趣的区域；点击矩形框右

下角的“放大镜”按钮进入选中区域的下一层，其中的数据按照同样的方式生成网格、分配颜色并进行展示；如果用户对选中的数据不满意或者完成了局部的分析，可以点击左上角的返回按钮，由于系统在栈中存储探索的历史记录，因此返回时直接复原出上一层的网格可视化结果及对应的颜色分配。重复以上流程，用户可以灵活地在层次化数据中进行探索和分析。

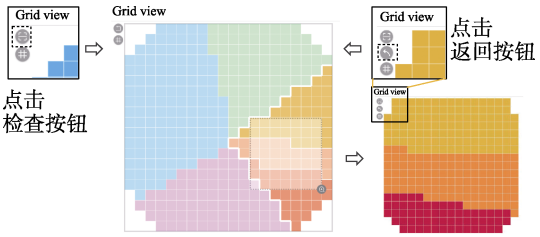


图 4　网格可视化中的层次化探索

（3）样本视图。为用户批量展示整体/选中局部区域的样本信息。考虑到加载速度和展示的空间限制，系统自动设置一个最大样本数阈值，当展示区域对应的样本数超过该阈值时，系统调用随机采样算法挑选该数目的样本进行展示。如图 1 所示，样本视图展示筛选出样本的图片，并使用该样本类别对应的颜色作为图片的边框。系统根据类别的颜色进行排序：一方面，可以让相同类别的样本聚集在一起方便用户分析；另一方面，排序的顺序参考控制视图中类别分布的顺序，使系统的整体顺序保持一致，保证用户探索中思维的连贯性。随着用户进行层次化探索，样本视图也会进行对应的更新。

3.3　聚类感知增强的网格可视化方法

本文集成了聚类感知增强的类簇形状优化方法[21]和动态颜色分配方法[39]，优化网格可视化中的网格生成和颜色分配。基于该网格可视化，系统实现了核心的概览视图，允许用户对聚类结果进行层次化探索和分析。

（1）系统集成了类簇形状优化方法[21]。在用户研究中发现：虽然没有单一的凸性度量能够完美地符合所有用户的感知，但是用户对凸性的感知可以分为偏好基于面积的度量和基于轮廓的度量 2 类。对于这 2 类用户，三元组比例[18]和周长比例[20]最符合他们的感知。因此，本文选择这 2 种方法作为代表性的凸性度量，将其应用于网格可视化中以满足不同用户的需求。基于所选的凸性度量对现有的网格可视化结果进行改进，包括整体优化和局部调整。整体优化中，基于二分图匹配改进类簇的近邻性和紧凑性；局部调整中，基于随机交换改进类簇的凸性。相关实验证明，该方法具备较好的优化凸性目标的能力。因此，本文对该方法进行集成，改进网格中类簇的凸性，以增强用户的聚类感知。

（2）系统集成了动态颜色分配方法[39]。该方法包括颜色分配方法和动态范围选择算法 2 部分。颜色分配方法基于连续性方法和模拟退火算法逐级优化，生成具有较好可区分性、和谐性并适应数据空间分布的颜色结果。动态范围选择算法则进一步在优化中增加约束，确保层次化探索中颜色的一致性。颜色生成模块的运行流程与用户的探索流程一致。用户选择数据集后，首先生成顶层的网格可视化，颜色分配方法使用完整的颜色空间为顶层的类簇分配颜色；然后基于这组颜色再次调用颜色分配方法，为子类簇分配合适的颜色。参考图 1 和图 3 的例子，顶层包含 4 个类簇，分别分配为蓝色、绿色、橙色和紫色。这 4 个类簇又可以进一步展开为 13 个子类簇。当用户选中一些区域时，选中部分的颜色将作为下一层的父类颜色，调用动态范围选择算法对这些颜色进行微调，并确认每个父类对应的子类的颜色范围。基于该动态颜色分配方法，系统能够生成可区分且和谐的高质量颜色分配结果，增强用户的聚类感知，帮助用户更高效地探索和分析数据集。

4　案例研究

本文通过 2 个应用案例，介绍数据分析人员使用本文系统分析、构建信息图表数据集的流程，分为研究配置、案例分析 2 部分。研究配置部分介绍现有信息图表数据集的来源和基本信息；案例分析部分介绍 2 组分析人员探索的真实案例，以及分析人员总结出的部分信息图表设计空间。案例研究结果证明，本文系统可以有效地帮助数据分析人员分析、构建完备的信息图表数据集。

4.1　研究配置

本文邀请 5 位从事信息图表研究的数据分析人员，对一个包含 3 501 张高质量信息图表的数据集进行后续的分析。该数据集通过相关领域专家在 Pinterest、Voronoi 和 Statista 等主流设计社区上收集获得。为了确保数据的质量，数据集中采用多阶段的数据处理流程：首先进行重复样本的识别与剔除，然后利用大语言模型对图表进行质量评估和筛选，保留高质量的信息图表样本。从图表类型看，

柱状图是数据集中最主要的可视化形式，共有 1 666 条的样本，约占总样本量的 47.6%；其次是折线图，包含 714 条样本，约占总样本量的 20.4%。这种分布特征与实际应用中的使用频率基本吻合。

本文对该数据集进行特征提取、层次聚类、数据降维等处理。（1）采用 CLIP[45]获取信息图表的视觉语义特征，使用在尺寸较大的图片上表现较好的 clip-vit-large-patch14-336 模型。（2）基于提取的特征向量，通过层次化 K-means 算法[46-47]对数据集进行层次聚类。限制每层的类簇数量范围 $[K_{\min}, K_{\max}]$，在该区间逐次实验以选择最合适的类簇数量。其中，系统使用轮廓系数作为度量聚类效果的标准，因为该系数能够有效地度量各类簇之间的分离程度和各类簇内的紧凑程度[49]；自顶向下地进行聚类，直到所有的类簇均小于预设的最大规模时终止；聚类结果形成一个 6 层的层次结构，包含 221 个叶子节点类簇，各层次间的样本分布相对均匀，以有效地捕捉数据集的内在结构特征。

需要注意的是，尽管采用大模型辅助筛选，但数据集中仍然存在一定比例的非研究目标样本，包括不含有图表的信息图，以及现有方法难以实现的复杂信息图表。通常，后者具有较强的设计性，难以被自动化方法复现。对于以上噪声，数据分析人员使用数据集分析系统筛选出这部分样本。

对于剩余的优质样本，数据分析人员则关注聚类产生类簇的特性。基于这些特性，分析人员希望能够总结出信息图表数据集的设计空间，可以更加系统地构建高质量的信息图表数据集。

在实际探索中，基于聚类感知增强方法，系统能够为不同的类簇赋予更加易识别、易区分的颜色和形状，帮助分析人员更高效地筛选非研究目标样本，并总结出信息图表数据集的设计空间。

4.2 案例分析

根据以上研究配置，本文整理了 2 个数据分析人员探索中发现的研究案例，以说明系统对非研究目标样本排除、设计空间探索的有效性。

4.2.1 案例 1：非研究目标样本排除

案例 1 中，数据分析人员从全局角度检查最顶层的样本，发掘其中的非研究目标样本。然后，开展进一步探索，以批量地找出不合理的样本。

图 5 所示为数据分析人员的一组发现。图 5a 中，数据分析人员在概览视图的顶层进行探索，基于聚类感知的增强，在湖蓝色类簇和浅紫色类簇的交界处发现了 2 个意外的样本 A_1 和 B_1。A_1 是一个美国地图上的热图，尽管热图是一种常见的图表，但是基于地图的热图十分复杂，难以简单地进行复现。分析人员认为，这类样本不适合目前阶段的数据集构建，但应该在设计空间中记录，作为未来的工作。B_1 则是一个不包含图表的信息图，图中展示了一批运动员和他们的年龄，但并未组织为任何图表。这类样本不属于信息图表，传递的信息有限，因此分析人员认为需要排除此类样本。

基于在第 1 层样本中的发现，分析人员认为有必要进行更深入的探索，以发掘类似的非研究目标样本，进行批量的排除。图 5b 中，分析人员选择指定区域后展开网格可视化。可以看出，聚类感知增强后的网格可视化更利于分析人员的识别。一方面，网格可视化的布局与上一层非常接近，并且类簇都具有相对较好的凸性，有助于分析人员快速辨别不同类簇。另一方面，网格可视化的颜色在上一层的基础上进行了调整：（1）由于父类颜色（浅紫色和浅蓝色）比较接近，因此采用动态颜色范围选择方法对这 2 种颜色进行微调，调整为深紫色和浅蓝色，扩大了子类可用的颜色范围。（2）采用颜色分配方法在选择出的颜色范围中进行优化，获得相对可区分且和谐的颜色分配。图 5b 中，分析人员进一步检查样本发现，由于网格可视化方法考虑了样本的近邻性，因此相关的样本类簇聚集在一起：A_2 部分的样本来自网格可视化右下角的浅蓝色区域，同样是基于美国/非洲地图的热图信息图表；A_3 部分的样本来自网格可视化右下角的深蓝色区域，紧邻 A_2 的样本，也是一批非洲地图热图的信息图表。尽管属于相近的类别，但由于整体的颜色风格等差异，这些样本分散在不同类簇中，但聚类结果可视化方法将这些类簇一起展示，基于其相似性，分析人员能够快速地发现所需要的样本。

事实上，随着分析人员更加全面地探索，他们发现这类样本并不全都集中在这 2 个类簇。如图 6 所示，分析人员在绿色的区域同样发现了地图上热图的信息图表，这批图表以美国地图为主。从图 6a 可以看出，A_1 部分的样本正是这类热图样本。因此分析人员选择对应的区域进行深入探索，并在图 6b 的网格可视化及其对应的样本视图中发现了大量类似的样本，这类样本也分为 2 个类簇；其中，大部分背景为白色的热图样本集中在 A_2 对应的浅蓝色类簇，而背景偏浅黄色的热图样本则集中在 A_3 对应的黄褐色类簇。

基于这些发现，分析人员以这批样本作为种子，通过样本特征进行搜索；并结合大模型再一次



图 5　案例 1：热图样本与非信息图表样本探索

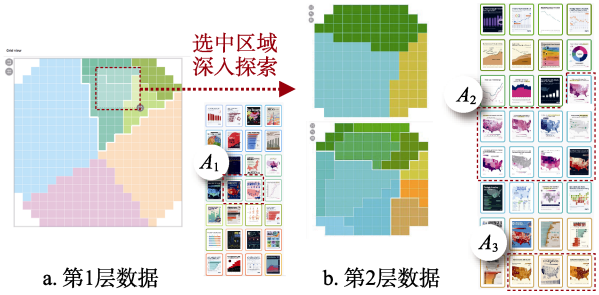


图 6　案例 1：对热图样本的层次化探索



图 7　信息图表设计空间的 2 个维度

筛选，筛选出了约 220 张地图上的热图信息图表。考虑到这类图表的广泛存在（占总数的 6.3%），分析人员记录了这部分数据，计划作为未来工作的方向。关于 B_1 样本，同样检查出少量的类似样本，这部分样本不包含任何图表，是依附于类簇的离群点，分析人员对其进行了有针对性的排除。

通过类似的操作，分析人员排除了大部分非研究目标样本。在探索操作后，分析人员对数据集进行采样检测，在随机采样的 200 个样本中，100.0% 的样本为符合预期的研究样本。

4.2.2　案例 2：数据集设计空间探索

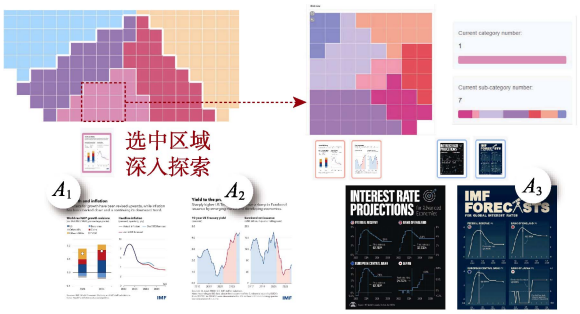
案例 2 中，分析人员进行进一步的探索，目标是发现现有设计空间中不足的地方以进行完善。

分析人员将信息图表的设计空间总结为信息图表布局与信息图表风格 2 个维度，如图 7 所示。

信息图表布局用于确定包含的元素，以及元素的位置和大小。图 7a 中，企鹅主题的信息图表包含文本、图表和图片 3 种元素，这些元素按照特定大小和位置分布。其中，图表元素可以根据表的类型和内部结构进一步拆分，数据分析人员总结更多的布局模板，在设计时通过推荐模型为输入的素材选择最合适的信息图表布局模板。在实践中，包含单一图表的信息图表最普遍，因此总结出的布局模板主要为此类模板。

信息图表风格用于为信息图表赋予一致和谐的风格，风格可以进一步被划分为颜色风格和形状风格 2 个部分。如图 7b 所示，常见的信息图表颜色风格包括单色系、双色系、渐变色系和彩色系；形状风格则包括扁平和立体等。在实践中，风格样式的模板更难总结，相关分析人员通过整理常用的样式库，可以在生成信息图表时灵活地选择统一的样式风格，增强信息图表的生动性。

图 8 所示为数据分析人员发现的一组样本。在顶层的概览视图中进行探索时，分析人员发现浅粉色的类簇中包含一个特殊的样本 A_1，其整体布局是常见的垂直布局，但通过双栏展示了 2 个并列相关的图表。分析人员认为，这类数据对于布局模板的完善是有价值的，应该在数据集生成中考虑这种多个子图的情况。为了探索更多类似布局结构的案例，分析人员选中浅粉色区域中的一块，将其展开进行深入探索，展开后的网格可视化结果如图 8b 所示。可以看出，该结果体现了本文集成的方法在实现细节上的完整性；该部分的父类为上一层的浅红色的类簇，没有其他的颜色可以构成范围选择的约束，是动态颜色范围选择方法的一种边界情况。在实践中，本文集成的方法根据子类的数量估计颜色的大概范围，确保颜色整体上的差异不会过大。可以看出，浅粉色的类簇展开为 7 个不同的子类簇，颜色分配方法在周围的颜色范围进行分配，获取可区分但父类颜色都相对接近的分配结果，保持了探索中思维的连贯性。



a. 第1层选中粉色的1个类簇　b. 第2层展开为7个不同的类簇

图 8　案例 2：包含多个子图的样本探索

在图 8b 所示的第 2 层样本中，数据分析人员在进一步的检查中发现，A_2 具有与 A_1 一致的布局结构，这 2 个样本位于同一个类簇（浅橙色）；此外，在左上角的浅蓝色类簇中，存在特殊的样本，如 A_3 所示。这些样本采用四格划分的子图表展示相关的数据信息，分析人员认为这种布局模板在生成时，可以进一步扩展出更多包含子图表的信息图表布局模板。

在该层探索中，数据分析人员也发现了有趣的现象。如图 9 所示，玫红色类簇下包含直接可见的 4 个相似的样本，经检查后发现它们仅在数据上和文字存在差异，整体布局和风格都保持完全一致。因此分析人员指出，在进行信息图表数据集的构建时，也可以考虑以类似的方式对数据集进行扩充，进一步丰富数据集。



图 9　案例 2：相似度极高的样本

图 10 所示为数据分析人员发现汽车主题的样本。在探索中分析人员发现，某一层中的浅绿色和浅黄色类簇覆盖了一批汽车主题相关的信息图表样本，它们的特点是在信息图表中大量插入汽车的图片以增强表达能力。可以看出，聚类结果可视化将相近的样本排布在一起，有效地提升了分析人员检查的效率。这些样本引起了分析人员的关注，因为其中部分样本仅仅放置了汽车主题的图片和介绍而没有任何图表元素，所以它们可能不是信息图表，应该被归类为非研究目标样本。数据分析人员认为，图 10a 中展示的 3 个样本属于信息图表，并且应该作为设计空间的一种丰富；在这些样本中，汽车作为辅助元素呈现，而数据通过车上的圆形图案进行编码，因此这类图表可以被视为布局较为特殊的一种柱状图或面积图。这种新奇的布局方式并不复杂，结合更多的图像元素，能够更生动地传递信息，分析人员认为这是一类需要在设计空间中考虑的布局模板。与图 10a 中的样本相比，数据分析人员认为，图 10b 中的 2 个样本不应该视为信息图表，这些样本仅使用图片作为装饰，本质上是信息的枚举，缺少图表元素的呈现，因此不应该将这类



a. 应该视为信息图表　　b. 不应该视为信息图表

图 10　案例 2：汽车主题的样本分析

样本包含在信息图表数据集中。

在设计空间探索的最后，分析人员同样对数据集进行采样检查，在随机采样的 200 个样本中，94% 的样本可以被分析人员总结的设计空间覆盖，未覆盖的样本主要是设计上非常特殊的样本，不具备一般性。因此基于本文系统，分析人员能够不断地完善信息图表数据的设计空间，有针对性地补充缺少的类别和样本数量较少的类别数据，构建高质量的信息图表数据集。

4.3　专家反馈与思考

案例研究后，专家也根据系统的使用情况进行反馈，这些反馈可以分为 2 类：关于系统的界面设计和系统的交互设计。

在界面设计方面，专家称赞了本文系统的简洁性，能够快速地理解并上手使用。专家指出，网格可视化中类簇的形状完整且凸性良好，类簇的颜色可区分且美观，帮助用户更容易地分辨出不同的类簇，有效地提高了工作的效率。然而，形状和颜色的设计也带来了一些顾虑。在形状方面，聚类感知增强可能导致一些离群点和类簇绑定在一起，需要在检查中进行进一步的判断。专家建议，可以将这些离群点尽可能地放置在类簇的边缘，有效地减少多余的检查。在颜色方面，与现有方法相比，尽管动态颜色分配方法更容易产生可区分及和谐的颜色，但是随着颜色范围的不断变小，仍然面临视觉上的困难。在这种情况下，如果要保证颜色的可区分性，则需要大量使用上一层其他部分的颜色范围，如图 8 中的粉色类簇展开后使用了上一层其他部分的紫色和红色，可能给用户带来误解；但如果强制要求避免使用这些颜色，那么颜色的可区分性无法保证。在实践中，可以通过调整本文在动态颜色分配方法中设计的参数进行权衡，但仍然难以做到两全其美。

在交互设计方面，专家指出本文系统的交互简洁、直接且容易使用，能够为用户提供便利的层次化探索方式。在层次化数据中，能够灵活地选择数据并深入检查。专家还指出，界面和交互中类别的一致性保持得很好，在探索和分析时能够通过多个视图的联动更快地理解数据。然而，交互也存在一些不足，专家提出了 2 点建议：（1）在网格可视化部分，内部的空间利用率可以进一步提高。专家希望能够在网格局部展示代表性的数据图片，并通过交互调整展示样本的密度，对可视化进行进一步的丰富。（2）现有系统交互时的响应偏慢，这是因为系统往往需要向前端加载大量的高清图片数据。专家建议，可以在本地预留更多缓存的空间，提高交互的响应速度，进一步提升用户的分析效率。

专家认为，本文系统有效地帮助用户对数据集进行探索和分析，并进一步构建完备的数据集。

5　结　语

本文提出基于聚类感知增强的信息图表数据集分析系统。该系统支持数据分析人员对数据集进行层次化探索，提供样本筛选和检查的功能，并实现了简洁、有效的数据分析交互；集成聚类感知增强方法，改进了网格可视化的类簇的形状和颜色，可帮助用户更高效地探索分析数据集。通过与信息图表相关专家合作开展了案例研究，验证所开发系统的有效性。一方面，专家使用本文系统排查非研究目标的样本，提升了数据集的质量；另一方面，专家在分析中总结、完善了现有的信息图表设计空间，在实践中构建更完备的信息图表数据集。案例研究结果表明，本文系统能够有效地帮助专家探

索、分析、构建信息图表数据集。

参考文献（References）：

[1] Ware C. Information visualization: perception for design[M]. 3rd ed. Amsterdam: Elsevier, 2012

[2] Lu Ting. Research on design method of infographic in big data era——taking the design of dynamic distribution data chart for college students as an example[D]. Suzhou: Soochow University, 2018(in Chinese)
(陆庭. 大数据时代背景下信息图表设计方法研究——以高校在校生校内动态分布数据图表设计为例[D]. 苏州: 苏州大学, 2018)

[3] Chen M, Mao S W, Liu Y H. Big data: a survey[J]. Mobile Networks and Applications, 2014, 19(2): 171-209

[4] Masry A, Islam M S, Ahmed M, *et al*. ChartQAPro: a more diverse and challenging benchmark for chart question answering[OL]. [2025-07-05]. https://arxiv.org/abs/2504.05506

[5] Cui W W, Zhang X Y, Wang Y, *et al*. Text-to-Viz: automatic generation of infographics from proportion-related natural language statements[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 906-916

[6] Qian C Y, Sun S Z, Cui W W, *et al.* Retrieve-then-adapt: example-based automatic generation for proportion-related infographics[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 443-452

[7] Wang Y, Sun Z D, Zhang H D, *et al*. DataShot: automatic generation of fact sheets from tabular data[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 895-905

[8] Yang Weikai, Chen Changjian, Zhu Jiangning, *et al*. A survey of visual analytics research for improving training data quality[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(11): 1629-1642(in Chinese)
(杨维铠, 陈长建, 朱江宁, 等. 基于可视分析的训练数据质量提升综述[J]. 计算机辅助设计与图形学学报, 2023, 35(11): 1629-1642)

[9] Lyu Fei, Chen Changjian, Zhang Jiapeng, *et al*. Visualization for supercomputer system: a survey[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(3): 321-335(in Chinese)
(吕斐, 陈长建, 张嘉鹏, 等. 面向超级计算机系统的可视化综述[J]. 计算机辅助设计与图形学学报, 2024, 36(3): 321-335)

[10] Todorovic D. Gestalt principles[EB/OL]. [2025-07-05]. http://www.scholarpedia.org/article/Gestalt_principles

[11] Rosin P L, Žunić J. Probabilistic convexity measure[J]. IET Image Processing, 2007, 1(2): 182-188

[12] Efrat A, Hu Y F, Kobourov S G, *et al*. MapSets: visualizing embedded and clustered graphs[C] //Proceedings of the 22nd International Symposium on Graph Drawing. Heidelberg: Springer, 2014: 452-463

[13] Sonka M, Hlavac V, Boyle R. Image processing, analysis, and machine vision[M]. 4th ed. Stamford: Cengage Learning, 2014

[14] Žunić J, Rosin P L. A new convexity measure for polygons[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 923-934

[15] Bozeman J R, Pilling M. The convexity ratio and applications[J]. Scientiae Mathematicae Japonicae, 2013, 76(1): 47-53

[16] Rosin P L, Mumford C L. A symmetric convexity measure[J]. Computer Vision and Image Understanding, 2006, 103(2): 101-111

[17] Rahtu E, Salo M, Heikkila J. A new convexity measure based on a probabilistic interpretation of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1501-1512

[18] Žunić J, Rosin P L. Measuring shapes with desired convex polygons[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(6): 1394-1407

[19] Held A, Abe K. On approximate convexity[J]. Pattern Recognition Letters, 1994, 15(6): 611-618

[20] Peura M, Iivarinen J. Efficiency of simple shape descriptors[C] //Proceedings of the 3rd International Workshop on Visual Form. Singapore: World Scientific, 1997: 443-451

[21] Zhou Y X, Yang W K, Chen J S, *et al*. Cluster-aware grid layout[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(1): 240-250

[22] Liu S Q, Tao M T, Huang Y F, *et al*. Image-driven harmonious color palette generation for diverse information visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(7): 3089-3103

[23] Healey C G. Choosing effective colours for data visualization[C] //Proceedings of the 7th Annual IEEE Visualization. Los Alamitos: IEEE Computer Society Press, 1996: 263-270

[24] Maxwell B A. Visualizing geographic classifications using color[J]. The Cartographic Journal, 2000, 37(2): 93-99

[25] Lu K C, Feng M, Chen X, *et al*. Palettailor: discriminable colorization for categorical data[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 475-484

[26] Schloss K B, Palmer S E. Aesthetic response to color combinations: preference, harmony, and similarity[J]. Attention, Perception, & Psychophysics, 2011, 73(1): 551-571

[27] Ou L C, Luo M R. A colour harmony model for two-colour combinations[J]. Color Research and Application, 2006, 31(3): 191-204

[28] Matsuda Y. Color design[J]. Asakura Shoten, 1995, 2(4): Article No.10

[29] Lara-Alvarez C, Reyes T. A geometric approach to harmonic color palette design[J]. Color Research & Application, 2019, 44(1): 106-114

[30] Smart S, Wu K K, Szafir D A. Color crafting: automating the construction of designer quality color ramps[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1215-1225

[31] Zeng Q, Wang Y Q, Zhang J, *et al*. Data-driven colormap optimization for 2D scalar field visualization[C] //Proceedings of the IEEE Visualization Conference. Los Alamitos: IEEE Computer Society Press, 2019: 266-270

[32] Lee S, Sips M, Seidel H P. Perceptually driven visibility optimization for categorical data visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(10): 1746-1757

[33] Chen C J, Yuan J, Lu Y F, *et al*. OoDAnalyzer: interactive analysis of out-of-distribution samples[J]. IEEE Transactions

on Visualization and Computer Graphics, 2021, 27(7): 3335-3349

[34] Xiang S X, Ye X, Xia J Z, *et al*. Interactive correction of mislabeled training data[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2019: 57-68

[35] Tennekes M, de Jonge E. Tree colors: color schemes for tree-structured data[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2072-2081

[36] Delon J, Desolneux A, Lisani J L, *et al*. Automatic color palette[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2005: Ⅱ-706

[37] Waldin N, Le Muzic M, Waldner M, *et al*. Chameleon: dynamic color mapping for multi-scale structural biology models[C] //Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine. Aire-la-Ville: Eurographics Association Press, 2016: 11-20

[38] Waldin N, Waldner M, Le Muzic M, *et al*. Cuttlefish: color mapping for dynamic multi-scale visualizations[J]. Computer Graphics Forum, 2019, 38(6): 150-164

[39] Chen J S, Yang W K, Jia Z L, *et al*. Dynamic color assignment for hierarchical data[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(1): 338-348

[40] Chen C J, Wang Z W, Wu J, *et al*. Interactive graph construction for graph-based semi-supervised learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(9): 3701-3716

[41] Yang W K, Ye X, Zhang X X, *et al*. Diagnosing ensemble few-shot classifiers[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(9): 3292-3306

[42] Zhang J W, Wang Y, Molino P, *et al*. Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 364-373

[43] Cao N, Shi C L, Lin S, *et al*. TargetVue: visual analysis of anomalous user behaviors in online communication systems[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 280-289

[44] Chen C J, Guo Y K, Tian F Y, *et al*. A unified interactive model evaluation for classification, object detection, and instance segmentation in computer vision[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(1): 76-86

[45] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the 38th International Conference on Machine Learning. Cambridge: PMLR, 2021: 8748-8763

[46] Chen B, Tai P C, Harrison R, *et al*. Novel hybrid hierarchical-k-means clustering method (H-*K*-means) for microarray analysis[C] //Proceedings of the IEEE Computational Systems Bioinformatics Conference - Workshops. Los Alamitos: IEEE Computer Society Press, 2005: 105-108

[47] Qi J P, Yu Y W, Wang L H, *et al*. An effective and efficient hierarchical *k*-means clustering algorithm[OL]. [2025-07-05]. https://journals.sagepub.com/doi/10.1177/155014771772862 7

[48] 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

[49] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65