软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1127−1139 [doi: 10.13328/j.cnki.jos.004951]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于维度扩展的 Radviz 可视化聚类分析方法*

周芳芳[1], 李俊材[2], 黄 伟[2], 王俊韡[1], 赵 颖[1]

[1](中南大学 信息科学与工程学院,湖南 长沙 410075)

[2](中南大学 软件学院,湖南 长沙 410075)

通讯作者: 赵颖, E-mail: zhaoying@csu.edu.cn

摘 要: Radviz 是一种多维数据可视化技术,它通过径向投影机制将多维数据映射到低维空间,使具有相似特征的数据点投影到相近位置,从而形成可视化聚类效果.Radviz 圆周上的维度排列顺序对数据投影结果影响很大,提出将原始维度划分为多个新维度来拓展 Radviz 圆周上的维度排序空间,从而获得比原始维度条件下更好的可视化聚类效果.该维度划分方法首先计算数据在每个原始维度的概率分布直方图,然后使用均值漂移算法对直方图进行划分,最后根据划分结果将原始维度扩展为多个新维度.提出使用 Dunn 指数和正确率来量化评估 Radviz 可视化聚类效果.进行了多组对比实验,结果表明,维度扩展有利于多维数据在 Radviz 投影中获得更好的可视化聚类效果.

关键词: Radviz;可视化数据挖掘;可视化聚类;多维数据;均值漂移

中图法分类号: TP391

中文引用格式: 周芳芳,李俊材,黄伟,王俊韡,赵颖.基于维度扩展的 Radviz 可视化聚类分析方法.软件学报,2016,27(5):1127−1139. http://www.jos.org.cn/1000-9825/4951.htm

英文引用格式: Zhou FF, Li JC, Huang W, Wang JW, Zhao Y. Extending dimensions in Radviz for visual clustering analysis. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1127−1139 (in Chinese). http://www.jos.org.cn/1000-9825/4951.htm

Extending Dimensions in Radviz for Visual Clustering Analysis

ZHOU Fang-Fang[1], LI Jun-Cai[2], HUANG Wei[2], WANG Jun-Wei[1], ZHAO Ying[1]

[1](School of Information Science and Engineering, Central South University, Changsha 410075, China)

[2](School of Software, Central South University, Changsha 410075, China)

Abstract: Radviz is a radial visualization technique that maps data from multi-dimensional space onto a planar picture. The dimensions placed on the circumference of a circle, called dimension anchors, can be reordered to reveal different patterns in the dataset. Extending the number of dimensions can enhance the flexibility in the placement of dimension anchors to explore meaningful visualizations. This paper describes a method that rationally extends a dimension to multiple new dimensions in Radviz. This method first calculates the probability distribution histogram of a dimension. The mean shift algorithm is applied to get centers of probability density to segment the histogram, and then the dimension can be extended according to the number of segments of the histogram. The paper also suggests using Dunn's index and accuracy rate to find the optimal placement of DAs, so the better effect of visual clustering can be achieved and evaluated after the dimension expansion in Radviz. Finally, it demonstrates the effectiveness of the new approach on synthetic and real world datasets.

Key words: Radviz; visual data mining; visual clustering; multi-dimensional data; mean shift

Radviz 是一种径向投影型多维数据可视化方法.它将维度作为维度锚点(dimensional anchors,简称 DA)固定到一个圆环上,数据点在各维度锚点的弹簧拉力作用下被映射到圆内合力为 0 的位置上,各维度锚点发出的拉

* 基金项目: 国家自然科学基金(61103108, 61402540)
Foundation item: National Natural Science Foundation of China (61103108, 61402540)
收稿时间: 2015-05-26; 修改时间: 2015-09-19, 2015-11-09; 采用时间: 2015-12-05

力大小与数据点在对应维度上的取值成正比,具有相似维度取值的数据点将被映射到圆内相近位置,人们可以直观地观察到投影到圆内数据点的簇聚结构,从而实现多维数据在二维空间的可视化聚类效果.Radviz 的圆环上可以布置大量维度锚点,圆内可以容纳海量数据点,其基于弹簧力的径向投影机制能够较好地在低维空间中保持多维数据的原始特性,并且处理过程自动和优雅.自从 Hoffman 等人[1]首次提出 Radviz 原型后,各种基于 Radviz 的多维数据分析方法被广泛应用于生物医疗和商业智能等众多应用领域.

Radviz 圆周上的维度锚点不仅可以任意增加、删除、移动和重排,而且不同维度锚点摆放顺序能够带来不同的投影结果.Radviz 这一特点能够让用户交互地探索同一数据集中可能存在的不同聚类结构,但用户也不得不花费大量精力去寻找理想的维度锚点排序.实际上,由于寻找 Radviz 的最优维度锚点排序是一个 NP 难问题[2],因此一些研究者提出使用启发式策略寻找能够获得较好数据投影结果的维度锚点排序[3,4],然而 Shark 等人[5]则另辟蹊径地设计了向量化的 Radviz(vectorized Radviz,简称 VRV),它通过扩展维度个数(可简称为升维)来获得更大、更灵活的维度锚点排序空间,并从中搜索更好的数据投影结果.VRV 以类别型多维数据为研究对象,首先枚举每个维度的可取值,然后按照枚举值的数量将一个维度划分为多个新维度.例如,某武器类型维度的取值可枚举为(刀、枪、剑、棒),维度扩展后该维度变为刀、枪、剑、棒 4 个新维度,原数据在 4 个新维度上的取值只能为 0 或 1,因此原武器类型维度取值为“刀”的数据在升维后取值变成了向量(1,0,0,0).

VRV 的维度扩展思路为 Radviz 提供了更大、更灵活的维度锚点排序空间,而且可以最大化弹簧力,并将数据点尽可能地映射到靠近圆周的位置,从而拉大不同聚类的间距,实现更好的可视化聚类效果.然而,VRV 仅适合类别型数据集,当应用于数值型数据集时还存在一些问题.首先,类别型维度的划分方法非常简单和直观,但对于可以连续取值的数值型维度如何进行合理的维度划分还有待研究;然后,VRV 在升维时的二值化策略使得数据点只能分布在圆内少量离散位置上,这会破坏数值型数据集的原始特性;最后,维度扩展能否改善数值型数据集在 Radviz 中的可视化聚类效果,还有待进一步的实验和量化评估.

本文聚焦于拓展和完善基于维度扩展的 Radviz 多维数据可视化方法,重点研究数值型多维数据的维度划分方法,以及评估维度扩展对 Radviz 可视化聚类效果的改善程度.本文维度扩展方法的核心思想来自图像处理中的灰度直方图和图像分割技术,我们首先计算出数据在原始维度上的概率密度分布直方图,然后通过均值漂移(mean shift)算法寻找概率密度峰值并实现对原始维度的划分,最后根据划分结果将原始维度扩展为多个新维度.完成维度扩展后,我们利用维度相似性和几何对称性来降低维度锚点排序的复杂度,然后根据 Dunn 指数和正确率两种指标来寻找类内距离小、类间距离大且正确率高的可视化聚类效果.我们在多组数据集上进行了实验,并量化评估和深入探讨了维度扩展对 Radviz 可视化聚类效果的影响.

1 相关工作

1.1 多维数据可视化技术

多维和高维数据普遍存在于我们的日常生活和科学研究中.比如,手机就包括品牌、型号、尺寸、重量、生产日期、屏幕尺寸和电池容量等几十个属性;又如,生物医学领域中的基因表达数据经常会产生成百上千个属性.鉴于人类对于数据的理解集中在 2D 或 3D 的低维空间,因此,采用信息可视化技术将多维数据绘制到低维屏幕空间是实现人与多维数据交互分析的一种解决思路.

多维和高维数据的可视化一直以来都是一个热门的研究领域.以主成分分析(principle component analysis,简称 PCA)和多维尺度分析(multi-dimensional scaling,简称 MDS)为代表的传统高维数据降维方法[6],是通过数学方法将高维数据降维到 2D 或 3D 的低维空间中,并尽量保留高维空间中的原有特性和聚类关系,但这类方法会损失数据在原始维度上的细节信息,而且无法表现维度之间的关系.以散点图矩阵(scatterplot matrix)[7]和平行坐标(parallel coordinates)[8]为代表的经典可视化方法虽然可以保留多维数据在每个维度上的信息,并突出维度间关系,但是当维度数量增多时,有限的屏幕空间难以容纳大量维度,而且还会存在过度绘制问题.以 Radviz[1]和星型坐标(star coordinates)[9,10]为代表的高维数据径向型投影方法,可以大量节约屏幕空间,但维度的排列顺序对数据投影结果影响很大,同时还存在数据重叠问题.此外,还有一部分研究提出先基于用户经验或算法准则

提取高维空间中的部分维度,然后在维度子集中结合上述技术完成高维数据分析,这类方法的代表是特征选择[11]与子空间分析[12].

1.2 Radviz的基本原理与发展现状

Radviz 数据投影原理如图 1 所示,数据集各维度作为维度锚点分布在圆环上,圆内的点代表数据记录,其位置由来自各维度锚点的弹簧拉力共同决定,每个弹簧拉力的大小正比于数据点在相应维度上的取值,这些数据点在所有弹簧拉力的共同作用下稳定在合力为 0 的位置.图 1 中 *A* 点和 *B* 点是一个四维数据集中两个数据点在 Radviz 中的映射,4 个维度被均匀分布在圆环上,记录 *A* 在维度 1 和维度 2 上取值较大,因此受到来自这两个维度锚点的弹簧拉力较大,从而定位在靠近 DA1 和 DA2 的附近;同理,记录 B 在维度 3 和维度 4 上取值较大,所以其位置在这两个维度锚点附近.在 Radviz 的数据投影机制下,具有相似特征的数据点将被映射到圆内相近位置.因此,圆内聚集的点簇可以被人们直观地观察到,从而形成可视化的聚类效果.

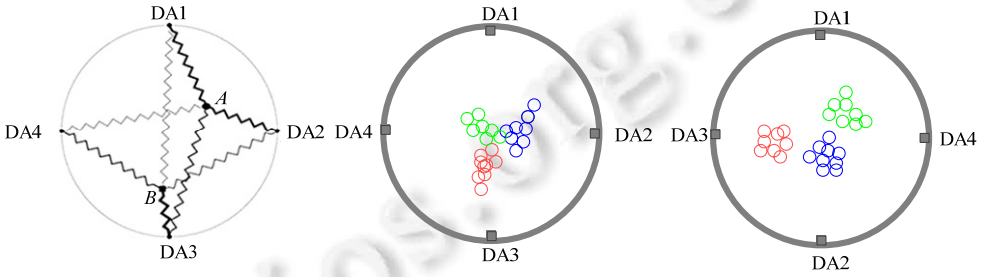


Fig.1 Illustration of Radviz and two different results of data projection generated by different placements of DAs

图 1 Radviz 示意图和两种不同维度锚点排序产生的不同数据投影结果

凭借良好的可交互性、可扩展性和原始特性保留能力,Radviz 被广泛应用于生物医疗、商业智能和故障分析等应用领域.比如,在生物医疗方面,McCarthy 等人[13]利用 Radviz 帮助专业人员从分子层次理解复杂高维生物数据.他们首先将分子化合物按是否对白血病和黑色素瘤产生影响进行分类,编码后将分子化合物按类别分布在 Radviz 圆周,最后根据细胞落在 Radviz 圆内的区位识别癌症细胞.在商业智能方面,Rahul 等人[14]根据 Radviz 的原理设计了一个完整的交互式可视化系统 dotlink360 来帮助用户探索、发现和理解商业生态系统中的合作关系.在故障分析方面,徐永红等人[15]提出了基于 Radviz 的故障诊断方法,并将其应用于化工过程的仿真模拟;Shi 等人[16]将 Radviz 和传感器网络拓扑等多种视图结合起来,设计了一个针对传感器监控的可视化引擎 SAVE,它可以显示一段时间内单一传感器节点在各个属性上取值的变化,以此帮助用户找出短时间内发生显著变化的属性;Zhou 等人[17]将 Radviz 与基于熵度量网络流量异常检测结合起来,帮助网络管理员高效地识别网络异常.

Radviz 方法本身也存在一些缺点:(1) 由于 Radviz 映射为多对一映射,这导致圆内数据点会出现遮挡和重合的现象;(2) Radviz 圆环上维度锚点摆放的位置和顺序对数据投影结果影响很大,如图 1 所示,随机维度锚点布局的可视化聚类效果可能并不理想.针对以上问题,学者们对 Radviz 进行了多方面的改进.

针对 Radviz 方法将多维数据映射到低维空间时会产生数据点互相遮挡和重合的问题,学者们采用的主要改进思路是将 Radviz 扩展到三维空间.Artero 等人[18]设计了 Viz3D.该方法在 Radviz 垂直方向上增加一个坐标轴 *z*,各数据点在 *z* 轴的取值为其有属性的平均值,Viz3D 能够有效保证原始空间内很近的数据在投影空间内也很近,但原始空间中距离较远的记录在投影空间中距离可能变近.随后,Nováková 等人[19]设计了 RadvizS.与 Viz3D 不同的是 *z* 轴坐标记录了原始空间中数据点到坐标原点的距离,从而在三维空间中保留原始空间中数据点间的距离信息.他们进一步采用镜面投影法,让原来重叠的记录点得到有效分离,在一定程度上解决了 Radviz 中数据点重合、遮挡的混乱现象.

在 Radviz 中寻找最优维度排列顺序是一个 NP 难题[2,20],所以解决该问题的主流方法是应用启发式策略寻找较优解.McCarthy 等人[13]采用把关联度较高的维度放置在邻近位置的策略,并使用概率统计中的抽样分布对特征维度进行分类,将 Radviz 分成多个扇区,每个扇区的弧上放置一类维度锚点.Leban 等人[3]提出了 VizRank,它使用 *K*-NN(*K*-nearest neighbor)分类器对不同的维度顺序组合的投影结果进行评测,并证明了选取评分越高的维度顺序排列于 Radviz 圆环上时,聚类效果越好,但 VizRank 方法在运行时间上具有局限性,寻找最优投影可能需要几分钟的时间.Albuquerque 等人[4]提出使用质量度量(quality measures)的策略对维度顺序进行重排,首先定义度量聚类质量好坏的标准,然后对有类数据集和无分类数据集进行分别度量,并使用贪心法从两个维度开始,每增加一个维度就寻找满足质量度量标准的最优排布,直到所有维度都被添加到 Radviz 圆环上,以此得到较优的维度顺序.经典 Radviz 的维度锚点是均匀地摆放在圆周上,但不均匀地摆放维度锚点也是可行的[21],这无疑进一步增加了求解优化的维度锚点布局的难度.在最新相关研究中,有学者提出了一些新的思路来同时求解优化的维度锚点排序和不均匀的维度锚点摆放位置.Igloo-Plot[22]首先选取一个随机维度并计算其余维度与该维度的相似性距离,然后根据相似性非均匀的线性排布维度锚点,相似性高的维度位于与其相近的位置,同时为了避免最远距离和最近距离在圆周中反而位于临近位置的不合理性,Igloo-Plot 创新性地提出将 Radviz 的圆形布局改为半圆形布局.Cheng 等人[23]则先将维度按照相似性进行 MDS 投影,然后将求解优化的维度锚点排序问题转化为求解旅行商问题,即在 MDS 形成的维度地图上寻找最优的遍历路径.

为了进一步提高 Radviz 的可视化聚类效果,Shark 等人[5]从升维角度出发,提出了 VRV 方法,它通过将类别型的维度切分为多个新维度来增加维度锚点排序空间,并使用优化算法来寻找更好的可视化聚类结果.但是,VRV 的维度划分策略只适合类别型多维数据,而对数值型多维数据并没有进行深入探讨.因此,本文重点研究适用于数值型多维数据的 Radviz 维度扩展策略.

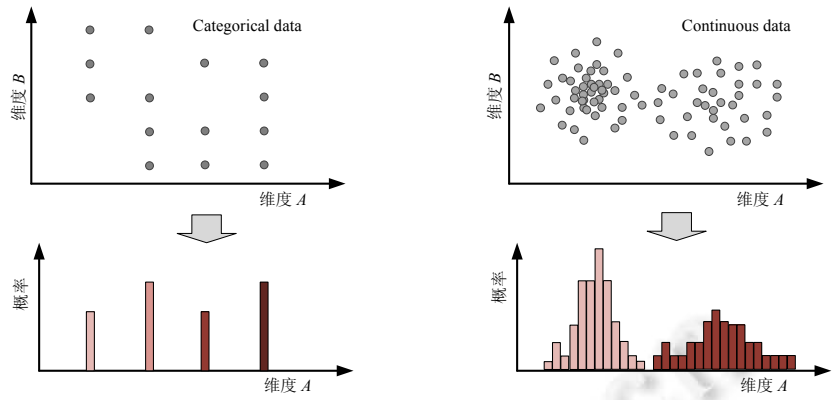
2 Radviz 的维度扩展方法研究

2.1 概率分布直方图的计算

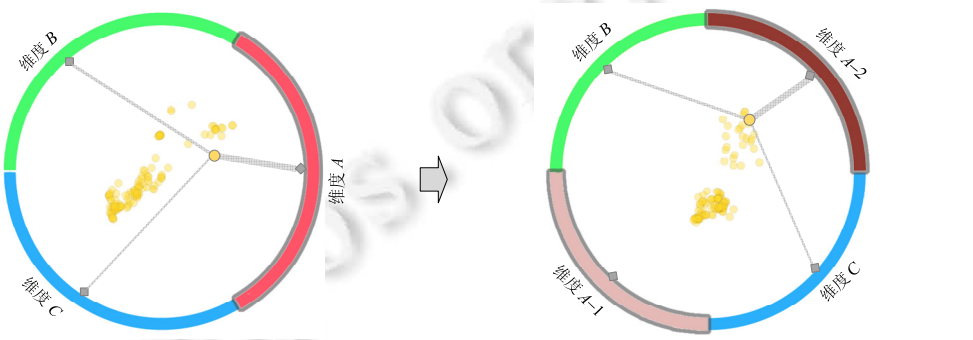
本文提出的维度扩展方法的核心思想来自图像处理中的灰度直方图和图像分割技术.本节首先从计算维度概率分布直方图开始.

以两个维度的类别数据集为例,如图 2(a)所示,数据点只能分布在二维平面的离散位置,所以它们在每个维度取值点上的概率呈离散脉冲状分布,每个脉冲位置都是一个概率密度峰值.类别型数据其实是数值型数据的特例,数值型数据在某个维度的取值上进行分段统计后,也可能呈现多峰的概率密度分布,特别是当数据在这个维度上存在明显聚集效应时,如图 2(b)所示.

对于数值型数据集,数据在每个维度上的取值是连续的,为了对维度进行划分,我们借助图像处理中的灰度直方图的概念计算数据在维度上的概率分布.设有 n 个维度 m 个数据点的数据集合,首先对数据在各维度的取值进行归一化,归一化后维度 $D_i(1\leqslant i\leqslant n)$各数据点的取值范围为[0,1];然后,对维度 D_i 的[0,1]取值范围进行 r 等分;接着,计算在 D_i 维度中,数据点取值落在各 r 分段的中的概率;最后,我们就可以得到维度 D_i 的概率分布直方图 H_i.图 3 显示了鸢尾花(Iris)数据集归一化后,4 个原始维度在 r=50 时的概率分布直方图.



(a) 类别型维度的概率分布直方图示意图 (b) 数值型维度的概率分布直方图示意图



(c) 3 个维度的数值型数据集的 Radviz 数据投影结果 (d) 把维度 *A* 扩展为两个新维度后的 Radviz 数据投影结果

Fig.2 Illustrations of the probability distribution histogram and the dimension expansion in Radviz

图 2 概率分布直方图和 Radviz 升维的示意图

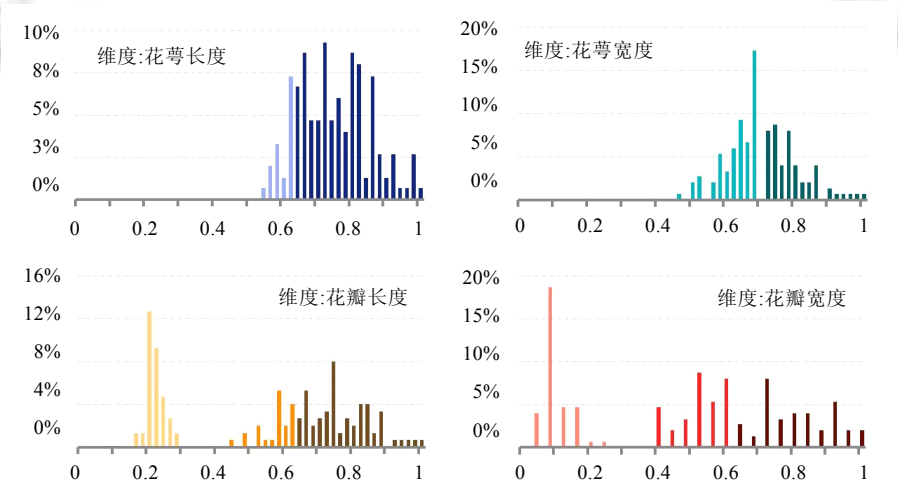


Fig.3 Probability distribution histograms of four dimensions in the iris data and their segmentation results by mean shift (the bandwidth=0.2)

图 3 鸢尾花数据集 4 个原始维度的概率分布直方图以及基于均值漂移(带宽设置为 0.2)的维度划分结果

2.2 基于均值漂移的维度划分

求得概率分布直方图后,需要在概率分布直方图中选取合理分界点,并利用这些分界点对维度取值区间进行划分.在 VRV 中,类别型数据集在维度上的概率分布直方图仅在离散取值点上呈现出明显的脉冲特性,可以根据冲激位置直接进行划分,但可以连续取值的数值型维度的划分策略则更为复杂.

本文使用图像分割技术中的均值漂移算法对概率密度直方图进行分段.均值漂移算法是一种统计迭代算法[24],在一定条件下,它可以收敛到最近的一个概率密度分布的稳态点,因此均值漂移算法可以用来检测概率密度函数中存在的模态.由于均值漂移算法完全依靠特征空间中的样本点进行分析,并不需要先验知识,而且收敛速度快,近年来被广泛应用于图像分割和运动跟踪的相关领域[25,26].

均值漂移算法的基本思想是通过反复地迭代搜索特征空间中的样本点最密集的区域,如图 4 所示,搜索点沿着样本点密度增加的方向"漂移"到局部密度极大点,漂移到同一极大值点的数据点将被归为一类.设有 r 个概率密度点,$P=\{p_1,p_2,\ldots,p_r\}$,对每一个点 $p_i,i=1,\ldots,r$,迭代计算搜索区域内的密度重心点 $m(p_i)$.

$$m(p_i)=\frac{\sum_{j=1}^{k} g\left(\left\|\frac{p_i-p_j}{h}\right\|^2\right)p_j}{\sum_{j=1}^{k} g\left(\left\|\frac{p_i-p_j}{h}\right\|^2\right)} \tag{1}$$

其中,h 是搜索区域的带宽,k 是搜索区域内点的个数,g 是控制算法收敛的核函数,它可以是高斯函数或均匀核函数.$m(p_i)$是均值漂移算法的基础,表示采样点的加权平均值,类似于"重心"的概念.一般 $m(p_i)$处的密度大于 p_i 处的密度,因此均值漂移总是漂向密度大的方向,即密度梯度增加的方向,均值漂移算法的收敛点为局部密度极大值点.如图 4 所示,均值漂移算法的计算过程如下:

Setp 1. 在特征空间中任意选择初始点 p_i,以 p_i 为圆心,带宽 h 为半径确定一个圆形的搜索区域.

Setp 2. 根据公式,计算圆中采样点的重心 $m(p_i)$,它是圆形区域中的密度最大值点.

Setp 3. 将搜索圆的圆心从 p_i 处漂移到 $m(p_i)$处,并计算 p_i 与 $m(p_i)$之间的距离,记为漂移向量.

Setp 4. 如果均值漂移向量的模小于允许误差ε,则迭代算法结束;否则,执行 Setp 2.

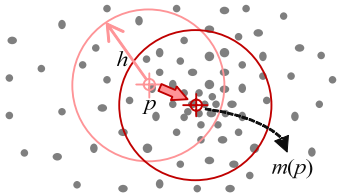


Fig.4 Illustration of mean shift algorithm

图 4 均值漂移算法示意图

对概率分布直方图的划分实际上可以看做对一个二维数据进行均值漂移,其中一个维度是直方图的 y 轴,即概率值,另外一个维度是直方图的 x 轴,即取值.考虑到这两个维度的取值范围为[0,1],本文在两个维度上使用相同的带宽和核函数,核函数选取的标准高斯型核函数.图 3 显示了 Iris 数据集 4 个维度的概率密度直方图在带宽 h=0.2 的情况下均值漂移的维度划分结果,其中,花萼长度和花萼宽度两个维度的直方图都被划分为两段,花瓣长度和花瓣宽度两个维度的直方图则都被划分为 3 段.

完成对维度概率分布直方图的分割后,就可以根据分割结果将一个维度扩展为多个新的维度.对于数据在新维度上的取值,本文没有采用 VRV 的二值化方法,而是在新维度上保持数据点在原始维度对应区段的取值.例如,6 个多维数据点在维度 A 上的取值分别为 0.1,0.2,0.2,0.5,0.5 和 0.6,如果其概率分布直方图在 0.5 处被划分为 $A1<0.5$ 和 $A2\geqslant 0.5$ 两段,维度 A 则可以扩展为两个新维度 $A1$ 和 $A2$,那么 6 个数据点在新维度 $A1$ 上的取值分别为 0.1,0.2,0.2,0,0 和 0,在新维度 $A2$ 上的取值分别为 0,0,0,0.5,0.5 和 0.6.

3 Radviz 的投影结果评价与维度排序化简

3.1 基于Dunn指数和正确率的聚类结果评估

维度扩展为 Radviz 提供了更大、更灵活的维度排序空间,在拓展的排序空间中我们可能会得到一个更好的数据投影结果.那到底什么样的数据投影结果是好结果呢?为此,我们需要建立一种量化评估机制来衡量多维数据在 Radviz 二维平面投影中位置分布的好坏.

从聚类的角度看来,好的数据投影能够使簇内的数据点尽可能地内聚(簇的内聚性和紧致性),不同簇之间的数据点尽可能地分离(簇的分离度和孤立性).Dunn 指数[27]将数据的相似性定义为欧拉距离,用类内和类间距离的非线性组合来评价聚类效果,因此本文选取它作为对 Radviz 投影结果的评价函数.Dunn 指数的公式如下:

$$D_{nc}=\min_{i=1,\ldots,nc}\left\{\min_{j=i+1,\ldots,nc}\left(\frac{\min\limits_{x\in C_i,y\in C_j}d(x,y)}{\max\limits_{k=1,\ldots,nc}\left\{\max\limits_{x,y\in C_k}d(x,y)\right\}}\right)\right\} \tag{2}$$

其中,$d(x,y)$是二维空间中两个点之间的欧式距离,$\min\limits_{x\in C_i,y\in C_j}d(x,y)$ 是两个簇 C_i 和 C_j 之间的最小距离,用来表示簇间分离程度,$\max\limits_{x,y\in C_k}d(x,y)$ 是簇 C_k 的直径,用来表示一个簇的内聚程度.$\max\limits_{k=1,\ldots,nc}\left\{\max\limits_{x,y\in C_k}d(x,y)\right\}$ 是所有簇的直径最大值.很明显,如果簇内越紧凑,簇间越分离,那么簇内最大距离越小,簇间最小距离越大,因此 Dunn 指数取值越高,聚类效果越好.

除了利用 Dunn 指数作为评估指标外,我们还引入了正确率来评价聚类结果的准确性,即以真实的分类结果为基准,计算 Radviz 投影后数据点被正确聚类的百分比.需要特别说明的是,Radviz 产生的聚类结果只是一种可视化的聚类呈现,它是人通过观察并基于大脑认知加工而成.为了能够计算 Dunn 指数值和正确率,我们需要用自动化聚类算法或交互标定手段对 Radviz 数据投影结果进行处理,从而获得计算机可使用的聚类结果描述.本文实验数据都采用具有真实分类标签的多维数据集,对 Radviz 投影结果,我们使用 *K*-means 算法进行聚类处理,其中,k 值设定为数据集的真实分类数目.根据 *K*-means 算法的聚类结果既可以计算 Dunn 值,又可以通过比较真实分类情况来计算聚类正确率.

3.2 基于维度相似性的维度排序化简

在维度数不变的情况下有两个因素影响 Radviz 的投影结果:维度权重与维度排序.本文需要对比升维前和升维后的可视化聚类效果,因此要尽量消除这两大因素对投影结果对比实验的影响.为了简化实验设计,突出升维对投影质量的影响,本文的实验对维度权重和维度布局进行如下设定:首先,设定所有维度具有相同的权重,因为维度权重的设置在一定程度上依赖用户对数据或维度的先验知识,该设定可以消除用户经验对实验的影响.然后,我们将圆周上维度锚点的放置策略设定为均匀摆放,这样可以简化维度排序的复杂度.在两个假设的条件下,我们只需要考虑升维前和升维后的维度均匀排列顺序.

为了综合评价升维对 Radviz 投影结果的整体影响,我们没有使用优化算法来自动寻找优化的维度排序,而是尽量穷举所有维度排序,并分别计算每种排序情况下投影结果的 Dunn 指数值和正确率,这样我们就可以分别得到两个评价指标在多种排序情况下的最小值、最大值以及平均值.

但穷举维度排序还是会严重影响实验效率.由于 n 个维度存在 $n!$种排列组合,考虑到圆形排列中有些维度顺序是等价的,即对 Radviz 而言应该有 $n!/n$ 种排列.因此,对于维度不高的数据集,可以用穷举法来进行;而对于维度较高的数据集,则不宜采用穷举法,比如当维度数大于 13 时,普通计算条件下很难快速完成实验.因此,我们还是需要引入一些规则来减少计算量.本文采用了 Radviz 维度排序优化策略中最常用的维度相似性规则,即将相似性高的维度尽量放在靠近的位置[13,20],其中衡量维度间相似性的指标采用皮尔逊指数.

4 实验结果与分析

在实验中,我们选择了 1 个人工合成数据集和 3 个真实世界数据集[28].人工合成数据集简单易懂,主要用来说明我们的实验过程;3 个真实世界数据集各有特点,主要用来比较我们的方法在不同数据集上的应用效果.

4.1 第1组实验结果与分析

第 1 组实验使用一个人工合成的数据集.该数据集共有 4 个维度,350 个数据点,3 个真实分类分别包含 139 个、137 个和 74 个数据点.由于具有高斯分布特征的概率分布直方图比较适合均值漂移对其进行划分,因此,我们在合成数据时,让所有数据记录在第 1 个和第 2 个维度分别形成两个高斯分布,并在第 3 个和第 4 个维度分别形成 3 个高斯分布.图 5 中显示了 4 个原始维度的概率分布直方图的划分结果,其中第 1 个和第 2 个维度的均值漂移带宽设置为 0.35,第 3 个和第 4 个维度的均值漂移带宽设置为 0.2.

实验从 4 个原始维度开始,我们首先获得了原始维度的所有 6 种可能的维度锚点排序情况下的 Radviz 投影结果,并分别计算每个投影结果的 Dunn 值和正确率,然后统计 6 组 Dunn 值和正确率的最小值、最大值和平均值.图 6(a)是上述 6 个投影结果中最好的结果(Dunn 指数:0.037,正确率:95.4%),其聚类正确率比较高,但红色和黄色数据点的边界并不清晰,所以聚类效果还不太理想.表 1 的第 1 行是 4 个原始维度的 6 种可能维度锚点排序情况下的 Dunn 值和正确率的最小值、最大值和平均值.接下来的实验都按照上述类似流程进行.根据图 5 的维度划分结果,我们先后测试了单维度升维和多维度升维等多种组合,并根据 Dunn 指数和正确率寻找优于图 6(a)的可视化聚类效果.

图 6(b)显示了将原始维度 1 扩展为两个新维度后,可以得到的最好 Radviz 可视化聚类效果,由于维度 1 只有两个高斯分布,而真实结果是 3 个分类,因此扩展维度 1 对改善聚类结果很有限.单列升维效果最好的是将维度 4 扩展为 3 个新维度,图 6(c)是这种情况下的最佳可视化聚类效果,其中 Dunn 指数值明显变大,而且正确率也达到了 100%.当然,扩展维度 4 后,并不是所有维度锚点排序都可以获得如此好的效果,图 6(d)就显示了一个不好的投影结果.总的来说,扩展维度 4 能够大幅度提升平均 Dunn 值,这说明将维度 4 扩展为 3 个新维度可总体改善该数据集的 Radviz 可视化聚类效果.我们又实验了两个维度同时升维的各种组合,图 6(e)显示了两个维度同时升维可以得到的最好结果,其中,维度 2 被扩展为两个新维度,维度 4 被扩展为 3 个新维度.值得注意的是,虽然图 6(e)的 Dunn 值比图 6(c)进一步提高了,但此时最小 Dunn 值却明显下降,这说明当更多原始维度同时升维时,虽然能够获得更好的可视化聚类效果,但拓展后的排序空间也可能产生更糟的数据投影结果.

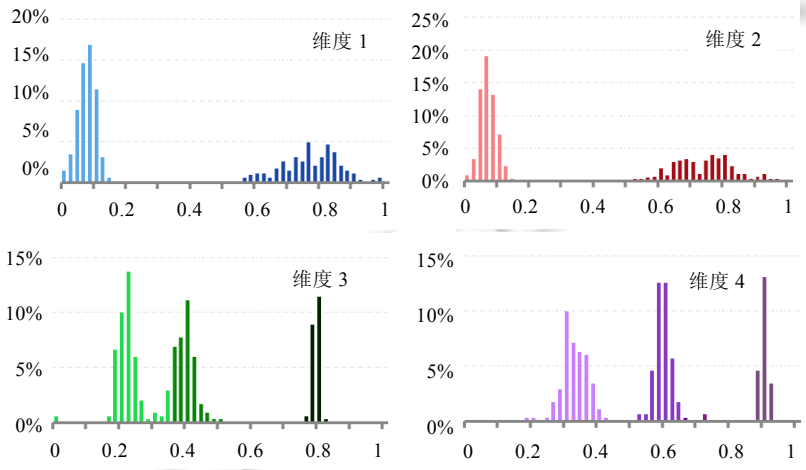
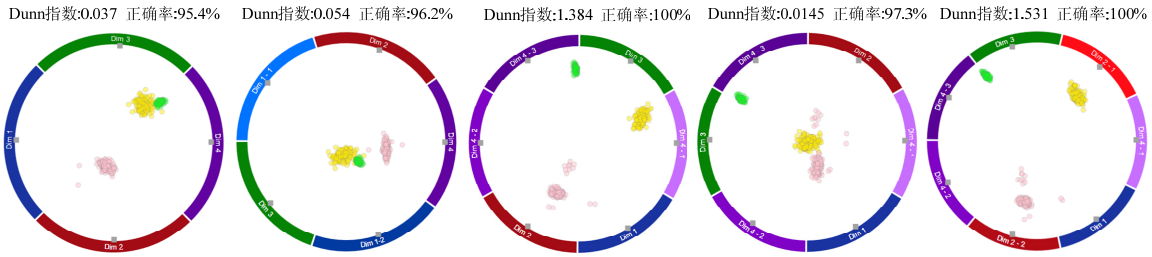


Fig.5 Probability distribution histograms of four dimensions in the synthetic data set and their segmentation results by mean shift

图 5 人工合成数据集的 4 个原始维度的概率分布直方图以及均值漂移的维度划分结果



(a) 原始维度可以得到的最好可视化聚类效果 (b) 维度 1 扩展为两个新维度后得到的最好可视化聚类效果 (c) 维度 4 扩展为 3 个新维度后得到的最好可视化聚类效果 (d) 维度 4 扩展为 3 个新维度后得到的最差可视化聚类效果 (e) 两个维度同时升维可以得到的最好可视化聚类效果

Fig.6 Some experimental results of using Radviz on the synthetic data

图 6 人工合成数据集的一些 Radviz 实验结果

Table 1 Dunn index and accuracy of dimension expansion experiments on synthetic data set

表 1 人工合成数据集升维实验的 Dunn 指数与正确率

	Dunn 指数			正确率(%)		
	最小值	最大值	平均值	最小值	最大值	平均值
图 6(a)对应的升维实验	0.020 4	0.037	0.029 4	93.71	96.28	95.14
图 6(b)对应的升维实验	0.014 1	0.054	0.025 9	69.14	97.14	91.36
图 6(c)对应的升维实验	0.014 5	1.384	0.342 5	96.28	100.00	98.47
图 6(e)对应的升维实验	0.004 2	1.531	0.398 3	81.42	100.00	98.58

4.2 第2组实验结果与分析

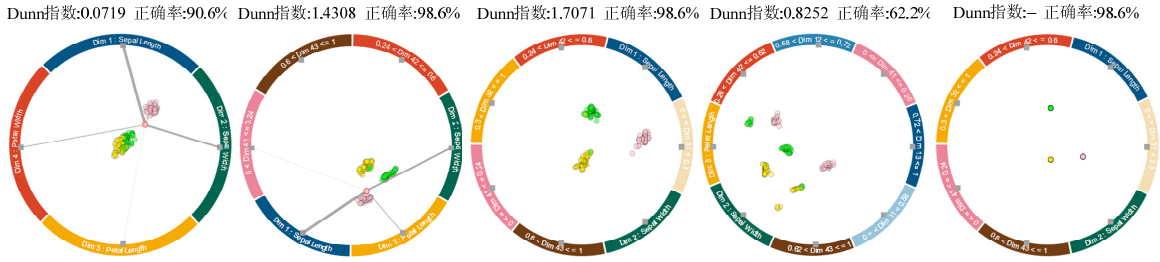
第 2 组实验使用的数据是著名的鸢尾花(iris)数据集,该数据集包含 150 个数据点,共 4 个维度,分别是花萼长度、花萼宽度、花瓣长度和花瓣宽度.该数据集的真实分类结果是 3 类,每一类都有 50 个数据点,我们分别用粉红、绿色和黄色的数据点来表示.图 7(a)是 Iris 数据集 4 个原始维度在 Radviz 中最好的可视化聚类效果,其中,粉红色数据点被很好地映射到一起,而绿色和黄色数据点之间却没有明显的边界.

Iris 数据集是真实世界数据集,其 4 个维度的概率分布直方图的高斯峰值特性明显弱于上组实验中的合成数据集,如图 3 所示,因此均值漂移对该数据的维度扩展更容易受带宽参数 h 的影响.由于带宽设置得太大或太小,会造成划分不足或过度划分情况的发生,因此我们对 Iris 数据集的 4 个原始维度设置了多个带宽(包括 0.15、0.2、0.25 和 0.4).同时,我们还测试了单维度升维和多维度升维等多种组合,并利用 Dunn 指数和正确率寻找优于图 7(a)的可视化聚类效果.

图 7(b)显示了单一维度升维时 Iris 数据集的最佳可视化聚类效果,图中花瓣宽度维度被均值漂移算法划分为 3 个新维度(均值漂移的带宽为 0.2),3 类数据点被明显地分成 3 个簇,只有非常少量的绿色数据点被错聚到黄色数据点区域.在花瓣宽度维度被划分为 3 个新维度的基础上,我们进一步将花瓣长度维度划分为两个新维度(均值漂移的带宽为 0.4),图 7(c)显示了这两个原始维度升维后的最佳可视化聚类效果,与图 7(b)相比,3 个簇的簇内距离更小,簇间距离更大了.然而,维度划分并不总能获得更好的投影结果,因为更多的维度可能会产生过度聚类,图 7(d)是花瓣长度和花瓣宽度两个维度都被划分为 3 个新维度后的一个过度聚类结果.这 3 个实验都找到了明显优于图 7(a)的可视化聚类结果,并且升维后的多种排序情况下的平均 Dunn 值也明显升高,见表 2.但更小的 Dunn 值和正确率的出现再次说明,升维后如果不合理排序维度,则可能获得更糟的投影结果.

在本组的最后一个实验中,我们采用 VRV 的二值化方法对升维后的 Iris 数据进行二值化.图 7(e)是对图 7(c)的数据二值化,并设置相同维度锚点布局后获得的可视化聚类效果.我们发现两图的聚类中心点位置相似,但图 7(e)中同簇数据点完全重叠在一起,这种过度离散化的投影结果说明 VRV 的二值化升维策略并不适合数值型数据集.

上述两组实验初步验证了升维能够较大幅度地提高 Radviz 可视化聚类效果.我们进一步分析其原因,发现这两个数据集都有一个相同特点,即它们的真实分类结果与某些维度相关度非常高.对这类维度合理升维后,不但能够找到更好的可视化聚类结果,而且可以总体改善不同维度排序情况下的投影结果.因此,我们可以得到一个结论,对与实际分类结果相关性高的维度进行升维,能够明显提高 Radviz 的可视化聚类质量.



(a) 原始维度可以得到的最好可视化聚类效果

(b) 单维度升维可以得到的最好可视化聚类效果

(c) 两个维度同时升维可以得到的最好可视化聚类效果

(d) 一个过度聚类的可视化聚类效果

(e) 使用二值化编码(c)中数据得到的过度离散化的投影结果

Fig.7 Some experimental results of using Radviz on the iris data

图 7 鸢尾花数据集的一些 Radviz 实验结果

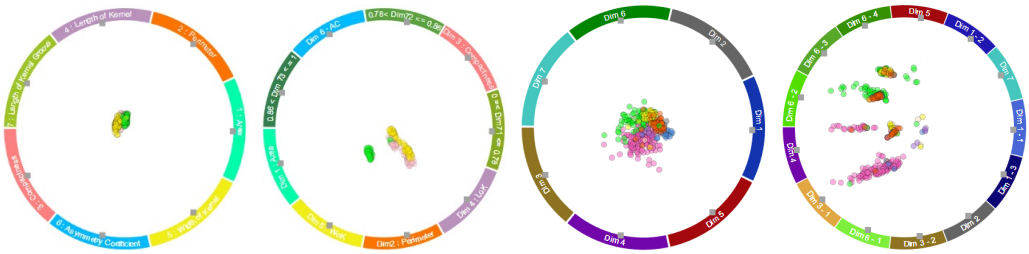
Table 2 Dunn index and accuracy of dimension expansion experiments on iris data set

表 2 鸢尾花数据集升维实验的 Dunn 指数与正确率

	Dunn 指数			正确率(%)		
	最小值	最大值	平均值	最小值	最大值	平均值
图 7(a)对应的升维实验	0.058 2	0.071 9	0.063 1	73.30	90.60	84.60
图 7(b)对应的升维实验	0.018 3	1.430 8	0.413 3	78.10	98.60	96.60
图 7(c)对应的升维实验	0.005 4	1.707 1	0.504 2	66.20	98.60	96.50
图 7(d)对应的升维实验	0.000 9	0.825 2	0.300 3	57.30	99.30	69.90

4.3 第3组实验结果与分析

在第 3 组实验中,我们使用了另外两个真实世界数据集:Seeds 和 Ecoli 数据集,这两个数据集的各维度与真实分类结果的相关度都不高.首先,我们对 Seeds 数据集进行实验,它有 7 个维度和 3 个分类.图 8(a)是 Seeds 数据集 7 个原始维度在不同排序情况下可以获得的最好可视化聚类效果,3 种不同颜色的数据点都集中在圆心附近,彼此根本无法区分开来,这也反映出 7 个原始维度与真实分类结果确实不太相关的事实.图 8(b)是对 Seeds 数据集进行单列升维后可以取得的较好效果,大部分的绿色数据点已经从另外两类中分离出来,而粉色和黄色数据点仍然无法区分开.接下来,我们对 Ecoli 数据集进行实验,它也有 7 个维度,但它有更多的真实分类数.图 8(c)是 Ecoli 数据集 7 个原始维度在 Radviz 中最佳可视化聚类效果,图中 8 种颜色数据点都混合分布在 Radviz 的中心区域.图 8(d)是对 Ecoli 的某 3 个原始维度进行升维后的较好结果,可以发现,可视化聚类效果还是有明显的改善.结合前两组实验,我们可以得到另一个结论,在原始维度与真实分类结果的相关度不高的情况下,通过合理的升维操作仍然能够带来更好的 Radviz 可视化聚类效果,并且这种改进在维度数和目标分类数适当增加时依然有效.



(a) Seeds 数据集原始维度可以得到的最好可视化聚类效果

(b) Seeds 数据集单维度升维可以得到的最好可视化聚类效果

(c) Ecoli 数据集原始维度可以得到的最好可视化聚类效果

(d) Ecoli 数据集三个维度升维后得到的一个改进可视化聚类效果

Fig.8 Some experimental results on the Seeds and Ecoli data sets

图 8 Seeds 和 Ecoli 数据集的一些实验结果

5 总结与展望

本文围绕 Radviz 的维度扩展问题,着重研究了针对数值型数据集的维度扩展策略,并提出了基于维度概率直方图和均值漂移算法的维度线性划分方法.本文同时还提出了使用 Dunn 指数和正确率来量化评估 Radviz 的可视化聚类效果.通过多组对比实验我们发现,对与真实分类结果相关性高的维度进行合理升维处理,可以较大幅度地提高 Radviz 的最佳可视化聚类效果,并且能够从总体上改善不同维度锚点排序情况下的数据投影结果.另外我们还发现,对与真实分类结果相关性不高的维度进行合理升维,也能在一定程度上改善 Radviz 的可视化聚类效果.

本文的方法还存在一些局限性,这也是我们在下一步工作中要考虑解决的问题.首先是维度扩展算法,一方面均值漂移很难自动获得每个维度的最优带宽,过大和过小的带宽设置都会影响维度划分结果,另一方面,本文目前的维度扩展方法属于线性方法,这会影响升维的适用范围,在未来工作中我们考虑为用户设计一个可视化工具,允许用户根据领域知识交互地扩展维度.其次,虽然本文的重点不是讨论维度锚点排序问题,但智能和高效的维度锚点排序方法有利于推动维度扩展的应用,特别是当我们需要对高维数据进行分析时,快速获得 Radviz 维度锚点优化布局尤为重要.最后,我们希望通过更多的实验来归纳和总结一套维度扩展的规则体系,并尝试将维度扩展策略应用到其他放射型可视化[29,30]和数据投影方法中,如星型坐标[31]等.另外,维度扩展还可以考虑与特征选择等技术结合起来,因为在高维数据分析中各维度的重要性可以不相同,也经常不需要一次性使用全部维度,用户可以在一个或者一组特定的维度子空间中实现数据聚类和其他数据分析任务.

References:

[1] Hoffman PE, Grinstein GG, Marx K, Grose I, Stanley E. DNA visual and analytic data mining. In: Proc. of the IEEE Visualization. 1997. 437−441. [doi: 10.1109/VISUAL.1997.663916]

[2] Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: Proc. of the IEEE Symp. on Information Visualization. 1998. 52−60. [doi: 10.1109/INFVIS.1998.729559]

[3] Leban G, Zupan B, Vidmar G, Bratko I. Vizrank: Data visualization guided by machine learning. Data Mining and Knowledge Discovery, 2006,13(2):119−136. [doi: 10.1007/s10618-005-0031-5]

[4] Albuquerque G, Eisemann M, Lehmann DJ, Theisel H, Magnor M. Improving the visual analysis of high-dimensional datasets using quality measures. In: Proc. of the IEEE Symp. on Visual Analytics Science and Technology. 2010. 19−26. [doi: 10.1109/VAST.2010.5652433]

[5] Sharko J, Grinstein G, Marx KA. Vectorized Radviz and its application to multiple cluster datasets. IEEE Trans. on Visualization and Computer Graphics, 2008,14(6):1444−1427. [doi: 10.1109/TVCG.2008.173]

[6] Ingram S, Munzner T, Irvine V, Tory M, Bergner S, Moller T. Dimstiller: Workflows for dimensional analysis and reduction. In: Proc. of the IEEE Symp. on Visual Analytics Science and Technology. 2010. 3−10. [doi: 10.1109/VAST.2010.5652392]

[7] Sedlmair M, Munzner T, Tory M. Empirical guidance on scatterplot and dimension reduction technique choices. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2634−2643. [doi: 10.1109/TVCG.2013.153]

[8] Guo P, Xiao H, Wang Z, Yuan X. Interactive local clustering operations for high dimensional data in parallel coordinates. In: Proc. of the IEEE Pacific Visualization Symp. 2010. 97−104. [doi: 10.1109/PACIFICVIS.2010.5429608]

[9] Kandogan E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In: Proc. of the IEEE Information Visualization Symp. 2000. 9−12.

[10] Sun Y, Tang JY, Tang DQ, Xiao WD. An improved multivariate data visualization method. Ruan Jian Xue Bao/Journal of Software, 2010,21(6):1462−1472 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3460.htm [doi: 10.3724/SP.J.1001.2010.03460]

[11] Bertini E, Tatu A, Keim DA. Quality metrics in high-dimensional data visualization: An overview and systematization. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12):2203−2212. [doi: 10.1109/TVCG.2011.229]

[12] Yuan X, Ren D, Wang Z, Guo C. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2625−2633. [doi: 10.1109/TVCG.2013.150]

[13] Mccarthy JF, Marx KA, Hoffman PE, Gee AG, Oneil P, Ujwal ML, Hotchkiss J. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. Annals of the New York Academy of Sciences, 2004,1020:239−262. [doi: 10.1196/annals. 1310.020]

[14] Basole RC, Clear T, Hu M, Hu MD, Mehrotra H, Stako J. Understanding interfirm relationships in business ecosystems with interactive visualization. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2526−2535. [doi: 10.1109/TVCG.2013.209]

[15] Xu YH, Hong WX, Chen MM. Visual fault diagnosis method based on Radviz and its optimization. Application Research of Computers, 2009,26(3):840−842 (in Chinese with English abstract).

[16] Shi L, Liao Q, He Y, Li R, Striegel A, Su Z. SAVE: Sensor anomaly visualization engine. In: Proc. of the IEEE Conf. on Visual Analytics Science and Technology. 2011. 201−210. [doi: 10.1109/VAST.2011.6102458]

[17] Zhou F, Huang W, Zhao Y, Shi Y, Liang X, Fan X. ENTVis: A visual analytic tool for entropy-based network traffic anomaly detection. IEEE Computer Graphs and Applications, 2015,35(6):42−50. [doi: 10.1109/MCG.2015.97]

[18] Artero AO, De Oliveira MCF. Viz3D: Effective exploratory visualization of large multidimensional data sets. In: Proc. of the XVII Brazilian Symp. on Computer Graphics and Image. 2004. 340−347. [doi: 10.1109/SIBGRA.2004.1352979]

[19] Nováková L, Stepankova O. Radviz and identification of clusters in multidimensional data. In: Proc. of the IEEE 17th Int'l Conf. on Information Visualization. 2009. 104−109. [doi: 10.1109/IV.2009.103]

[20] Caro LD, Frias-Martinez V, Frias-Martinez E. Analyzing the role of dimension arrangement for data visualization in Radviz. In: Advances in Knowledge Discovery and Data Mining. 2010. 125−132. [doi: 10.1007/978-3-642-13672-6_13]

[21] Gee AG, Yu M, Grinstein GG. Dynamic and interactive dimensional anchors for spring-based visualizations. Technical Report, Computer Science, University of Massachussetts Lowell, 2005.

[22] Kuntal BK, Ghosh TS, Mande SS. Igloo-Plot: A tool for visualization of multidimensional datasets. Genomics, 2014,103(1):11−20. [doi: 10.1016/j.ygeno.2014.01.004]

[23] Cheng SH, Mueller K. Improving the fidelity of contextual data layouts using a generalized barycentric coordinates framework. In: Proc. of the IEEE Conf. on Pacific Visualization. 2015. [doi: 10.1109/PACIFICVIS.2015.7156390]

[24] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002,24(5):603−619. [doi: 10.1109/34.1000236]

[25] Zhou FF, Zhao Y, Ma KL. Parallel mean shift for interactive volume segmentation. In: Proc. of the MLMI 2010. LNCS 6357, Berlin, Heidelberg: Springer-Verlag, 2010, 67−75. [doi: 10.1007/978-3-642-15948-0_9]

[26] Peng NS, Yang J, Liu Z, Zhang CF. Automatic selection of kernel function window width in mean-shift tracking algorithm. Ruan Jian Xue Bao/Journal of Software, 2005,16(9):1542−1550 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20050903&journal_id=jos [doi: 10.1360/jos161542]

[27] Dunn JC. Well-Separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 1974,4(1):95−104. [doi: 10.1080/01969727408546059]

[28] Machine Learning Repository. 2014. http://archive.ics.uci.edu/ml/

[29] Chen Y, Zhang XY, Feng YH, Liang J, Chen HQ. Sunburst with ordered nodes based on hierarchical clustering: A visual analyzing method for associated hierarchical pesticide residue data. Journal of Visualization, 2015,18(2):237−254. [doi: 10.1007/s12650-014-0269-3]

[30] Wu YD, Wang S, Wang HY, Li QS, Jiang HY, Zou YG. A total variation-based hierarchical radial video visualization method. Journal of Visualization, 2015,18(2):255−267. [doi: 10.1007/s12650-014-0266-6]

[31] Manuel R, Laura R, Francisco D, Alberto S. A comparative study between Radviz and Star Coordinates. IEEE Trans on Visualization and Computer Graphics, 2016,22(1):619−628. [doi: 10.1109/TVCG.2015.2467324]

附中文参考文献:

[10] 孙扬,唐九阳,汤大权,肖卫东.改进的多变元数据可视化方法.软件学报,2010,21(6):1462−1472. http://www.jos.org.cn/1000-9825/3460.htm [doi: 10.3724/SP.J.1001.2010.03460]

[15] 徐永红,洪文学,陈铭明.基于 Radviz 及其优化的可视化故障诊断方法.计算机应用研究,2009,26(3):840−842.

[26] 彭宁嵩,杨杰,刘志,张风超.Mean-Shift 跟踪算法中核函数窗宽的自动选取.软件学报,2005,16(9):1542−1550. http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20050903&journal_id=jos [doi: 10.1360/jos161542]

周芳芳(1980－),女,湖南株洲人,博士,副教授,CCF 专业会员,主要研究领域为科学可视化,信息可视化.

王俊韡(1992－),男,硕士生,主要研究领域为机器学习与可视化.

李俊材(1990－),男,硕士生,主要研究领域为高维数据可视化.

赵颖(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为可视化与可视分析.

黄伟(1991－),女,硕士生,主要研究领域为可视化数据挖掘.