

多元数据的可视化

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 <http://staff.ustc.edu.cn/~zwp/>

论坛 <http://fisher.stat.ustc.edu.cn>

简介

1.2 多元数据的可视化	2
1.3 可视化技术	6
1.3.1 几何投影方法	6
1.3.2 基于像素的可视化技术	29
1.3.3 层次化可视化技术	33
1.3.4 基于图标的可视化技术	41

1.2 多元数据的可视化

Motivations

- 人们相信大量的数据中隐藏着有价值的信息
- 从文字或表格中挖掘有用的信息是非常困难的. 但是人类具有强大的视觉能力, 能够迅速发现图形中的特征, 因此对数据进行有效的图示化是流行的做法
- 信息可视化是通过使用计算机图像表达数据来增强人类的识别能力, 以发现数据中预期或未预期的特征.
- 多元数据可视化中涉及的数据集往往是高维数据, 而且数据的各变量之间具有某种相关性
- 多元数据的广泛性使得人们需要对数据分布进行整体上的认识, 以及了解各属性之间的关系.

Challenges

- 构造一个问题的良好的可视化表示有些时候是非常困难且不明确.
- 映射问题: 将高维数据合适的映射到 2D 上往往比较困难. 一般来说这依赖于数据的本身结构. 而且, 将数据属性和图像元素联系起来也是需要特别注意的, 不合适的联系会对观察者造成误导.
- 维数问题: 多元数据往往数据量很大且维数较大, 因此将这种数据的所有特点在一个图中表达是很困难的, 从而对用户直观的和交互的探索数据空间带来挑战. 维数的顺序也是可视化中的重要因素.
- 设计上的权衡: 多元数据可视化总是在信息量, 简单性和精确性之间进行权衡.

-
- 评价可视化工具发现的结果往往不容易: 我们事先并不知道数据中含有什么样的有用信息. 但是不管怎么样, 如果对数据一无所知, 则我们就根本没有办法评价发现的结果是否具有价值.

维数

- 数据中的属性个数 (number of attributes)
 - 1: 一维 1D/ univariate
 - 2: 二维 2D/ bivariate
 - 3: 三维 3D/ trivariate
 - ≥ 3 : 多维 (multidimensional/hypervariate/multivariate)
- 低维与高维之间的界限是不明确的, 一般高维有 ≥ 4 变量

Dimensions	相互独立的属性 (attribute)
Variables	相互关联的属性
Multidimensional	独立维的维数
multivariate	相关变量的维数

1.3 可视化技术

根据 Keim and Kriegel (1996, 1997), 我们将多元数据的可视化技术分为 4 类: 几何投影方法, 基于像素 (Pixel-oriented) 的可视化技术, 层次化可视化 (Hierarchical display) 和基于图标方法 (Iconography)

1.3.1 几何投影方法

- 寻找将高维数据集在低维空间显示的有效投影和变换方法
- 将数据的属性映射到 2D 平面 (例如散点图) 或者任意的空间 (例如平行坐标图)
- 这种方法有利于发现异常点和不同维数之间的相关性
- 坐标的展示顺序可能会影响我们的视觉发现.

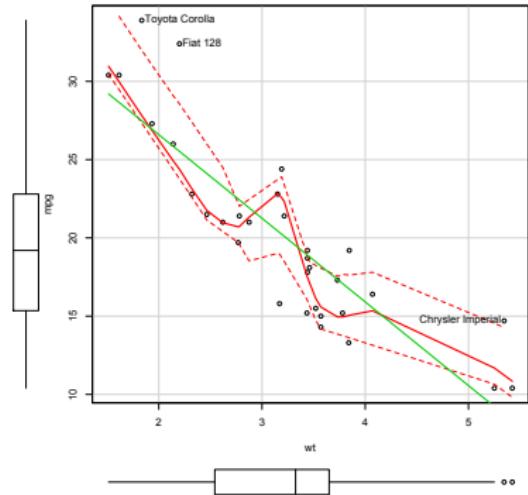
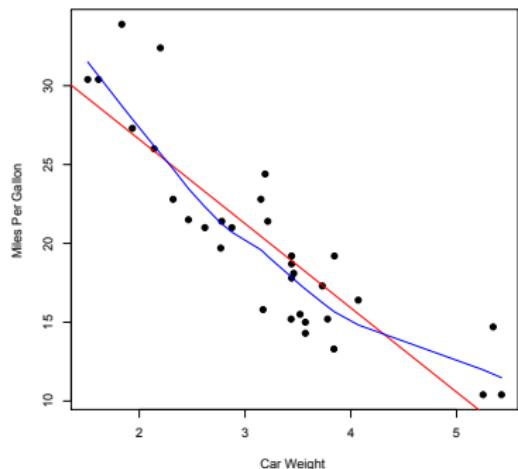
-
- 当数据量或者维数较大时可能会出现观测值或者类的重叠, 从而对视觉发现造成困难

Scatter plot

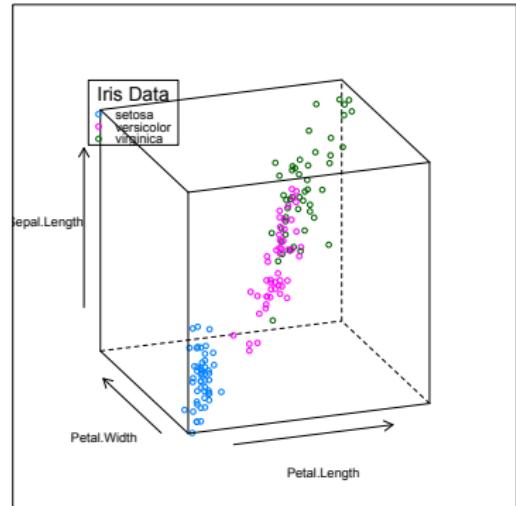
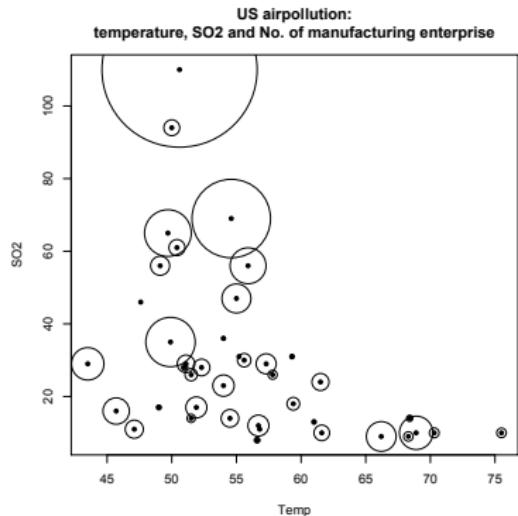
- 散点图 (scatterplot) 用来展示二元离散数据.
- 可以适当的扩展, 以显示更多的信息或者推广到 3 维数据

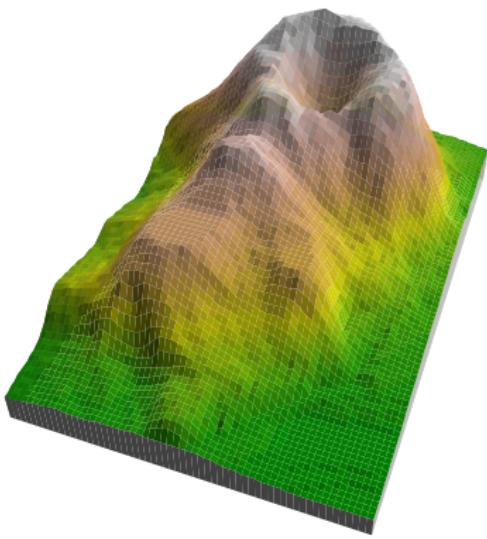
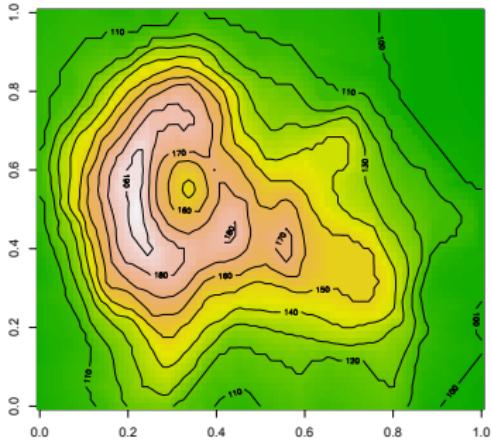
二维数据:

Scatterplot Example



三维数据:





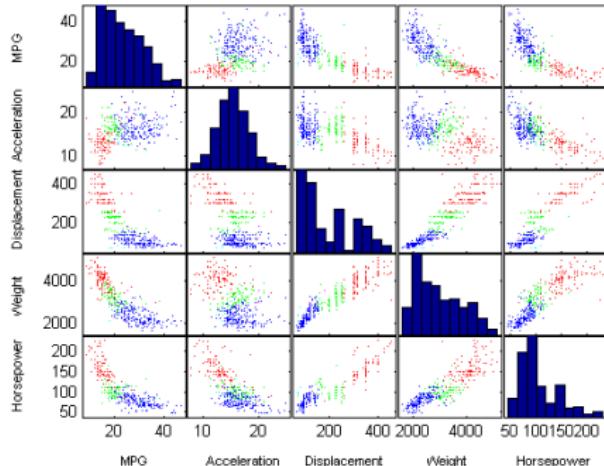
对 3 维以上的数据, 通过两两之间的散点图来发现变量之间的关系, 即:

Scatterplot Matrix

- 对高维数据集, 散点图阵将所有两两变量之间的散点图以矩阵形式排列展示
- 容易发现两两之间的关系, 且只能发现两个维之间的关系, 不能发现多个数据维之间的关系. 此外, 数据点过多时散点图会太过杂乱
- 关联更新 (brushing and linking) 技术可以应用到散点图中以解决上述部分问题: 被刷 (brushed) 的点在系统中所有视图中被高亮显示, 这样对比不同的视图就可以从不同的侧面来看到数据的特征.
- 在每个散点图中, 不同类别 (水平) 的点用不同的颜色来突出显

示

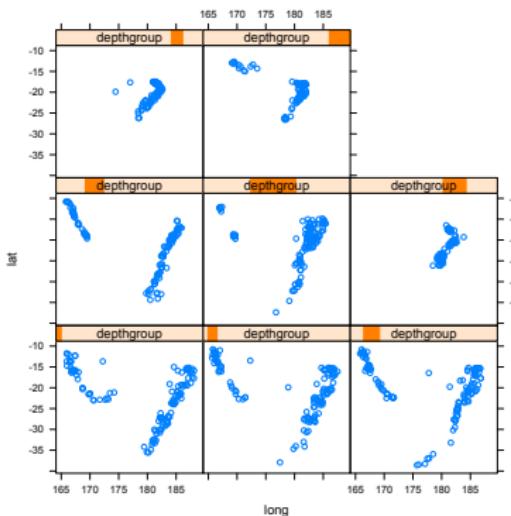
- 下图中，车辆数据按照缸体个数用不同的颜色突出显示.



displayed in R via the pairs() function

Trellis Displays(Becker and Cleveland, 1996)

- 一类基于条件化一个或者多个变量对复杂多元数据进行可视化
的技术, 常表现为一些图形构成的矩阵阵列, 因此称为 *Trellis*.
- 对变量的不同水平或者不同区间下的数据集绘制相同的图形
(比如直方图, 散点图等)
- 许多统计软件使用 *trellis plots*或者 *crossplots*这样的名称来进
行多面板条件化作图
- 在 R 中使用 **lattice**包来载入 *trellis*系统. 可以参考 Sarkar,
Deepayan (2008) Lattice: Multivariate Data Visualization
with R. Springer, New York.

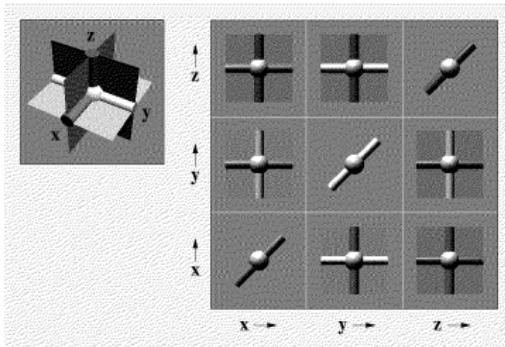


- **quakes**数据集, 感兴趣是三个变量 lat,long 和 depth
- 图形从左到右, 从下到上, 深度 (depth) 逐渐增加
- 8 个不同的深度区间, 每个包含约相同的地震次数
- 较浅的地震在两个倾斜的断层板块上均有发生; 西边的板块主要是浅地震, 而东边的板块深浅地震都有.

Hyperslice(Jarke and Robert, 1993)

- 一种可视化多维标量函数的方法, 本质想法是将一个多维函数表示为一些正交二维切片 (slice) 图构成的矩阵
- 对 p 维标量函数 $f(\mathbf{x}) = f(x_1, \dots, x_p)$, 记 $\mathbf{c} = (c_1, \dots, c_p)$ 为感兴趣的 (当前) 点, ω_i 表示第 i 维的宽度, 即 $R_i = [c_i - \omega_i/2, c_i + \omega_i/2]$ 为感兴趣的区间
- 二维切片 $S_{k,l}(k \neq l)$ 为 $f(\mathbf{x})$, $x_k \in R_k, x_l \in R_l$; 其他 $x_i = c_i, i \neq k, l$ 的表示
- 一维图 G_k 为 $f(\mathbf{x})$, $x_k \in R_k$, 且其他 $x_i = c_i, i \neq k$ 的图 (横轴为 x_k , 纵轴为 $f(\mathbf{x})$) 的表示
- HyperSlice 图就是一个 $p \times p$ 矩阵图, 对角为一维图 G_i , 非对角元 (k, l) 为二维切片图 $S_{k,l}$, 用相应于函数值大小的灰度值表示函数的值

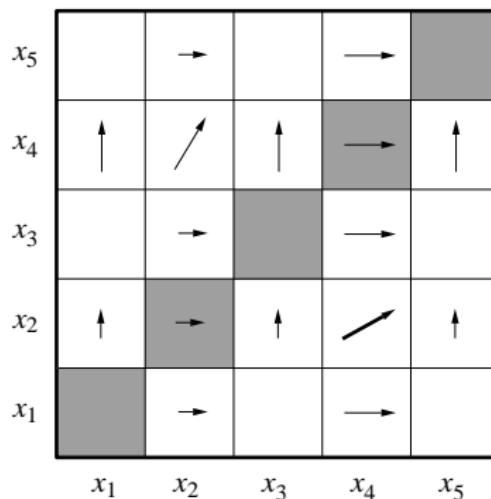
- $p = 3$ 的 HyperSlice 示意图: 左边为当前点及其邻域, 右边为相应的 HyperSlice 图



- 当我们拖动切片 $S_{k,l}$ 位移 $[d_k, d_l]$ 时, 当前点从 c 变为

$$c_k - d_k \rightarrow c_k, c_l - d_l \rightarrow c_l$$

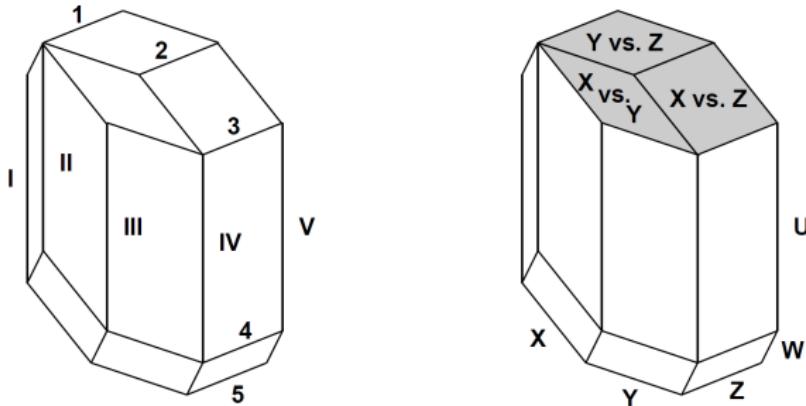
此时的效果图如下 ($S_{4,2}$ 被拖动, 同一列的切片图水平移动 d_k , 同一行的垂直移动 d_l)



Hyperbox (Alpern and Carter, 1991)

- 类似于前面的散点图阵和 HyperSlice, 不同之处在于使用 p 维盒子来建立图形, 而不是散点矩

- Hyperbox 是 p 维盒子的二维描述, 下面是一个 5 维 hyperbox 图



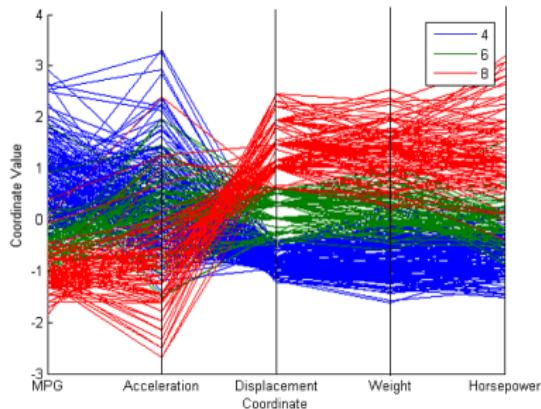
- 一个 p 维 hyperbox 由 p^2 个线和 $p(p - 1)/2$ 个面构成, 线的长度和斜率可以是任意值
- 同样长度和方向的线构成方向集 (direction set). 上图中, 线 1,2,3,4,5 构成一个方向集, 线 I, II, III, IV, V 构成另外一个方

向集.

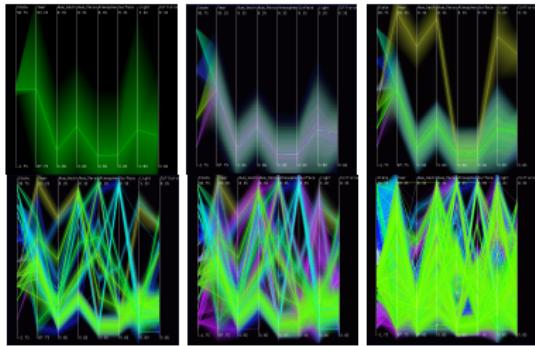
- 给定 5 个变量 x, y, z, ω 和 u , 每个变量映射到一个方向, 如上右图所示
- 每个面可以用来绘制两个变量之间的散点图或者其他线图
- Hyperbox 比散点图阵功能要强大的多, 它可以将变量映射到面的大小和形状, 可以强调或者不强调某些变量
- p 维数据映射到 2 维, 长度和方向的任意性可能会导致错误的信息

Parallel Coordinates(Inselberg and Dimsdale, 1990)

- 平行坐标图 (Parallel Coordinates) 是最早提出的以二维形式表示 p 维空间数据的可视化技术之一
- 基本思想是将 p 维数据用 p 条等距离的平行轴映射到二维平面上, 每条轴线代表一个属性维



- 优点是表达数据关系非常直观，易于理解
- 缺点是维数非常大时垂直轴非常靠近。数据个数非常大时的重叠会掩盖数据属性的内容关系。此外，垂直平行轴之间的安排顺序对发现数据之间关系有着重要影响
- 层次平行坐标图可以解决数据重叠所带来的问题

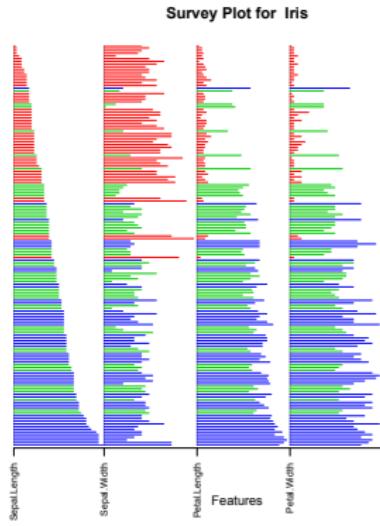


230,000 个严重事故数据。

第一个图为聚类根节点水平下，最后一个为所有数据点下。中间为不同聚类层次下。

Survey plot

- 使用等距平行线表示坐标轴, 所有数据点按照某个属性排序后, 将每个数据点的每个属性值通过从相应坐标上延伸出的线表示, 线长与该属性值成比例.
- 有助于发现两个变量之间的相关性
- 对不同的类使用不同的颜色有助于寻找能够最佳分类数据的坐标



Andrews Curve(Andrews, 1972)

- 使用一条光滑的曲线表示每个 p 元观测点 $\mathbf{x} = (x_1, \dots, x_p)$:
 p 为奇数时

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + \dots + x_{p-1} \sin\left(\frac{p-1}{2}t\right) + x_p \cos\left(\frac{p-1}{2}t\right)$$

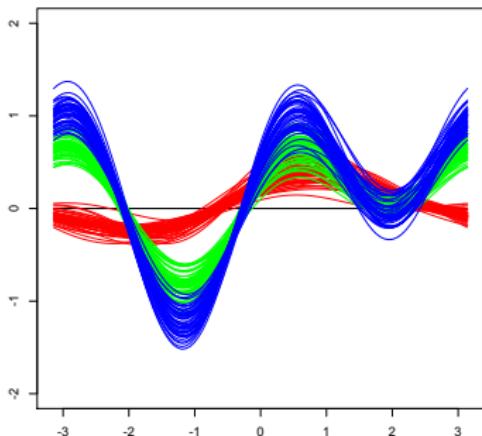
p 为偶数时

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + \dots + x_p \sin\left(\frac{p}{2}t\right)$$

其中 $-\pi < t < \pi$

- 相近点的曲线表示也类似, 不同点的曲线表示也不同. 因而方便用于发现类和异常点.

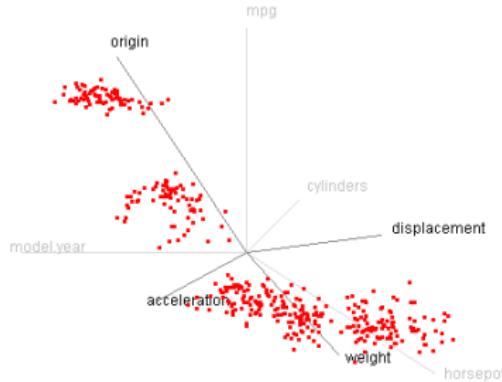
Andrew's curve for Iris data



Star Coordinates(Kandogan, 2000)

- 散点图往高维的一种推广
- 对 p 维数据, 使用 p 条半径将圆均分, p 条半径作为坐标轴, 再

将原数据点映射到以 p 条半径为坐标轴的圆上. 进而还可以对坐标轴进行旋转和变换. 例如下图所示



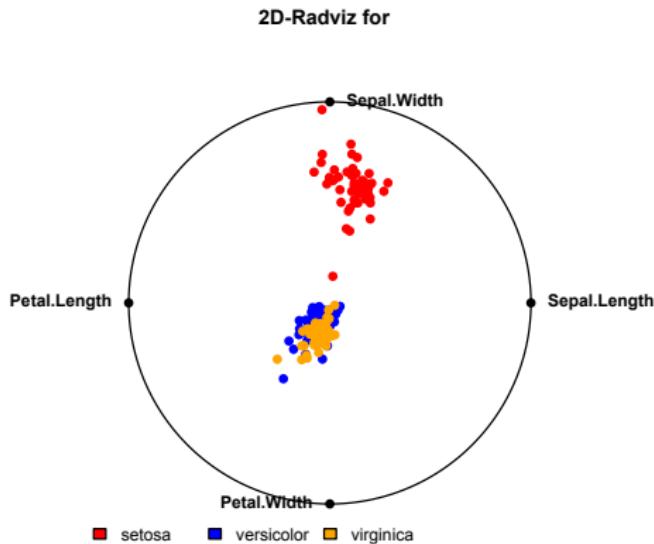
- 数据中相近的点变换到圆上也很相近, 因此利用发现数据中的类

Radiz or Radial Visualization Plots(Hoffman, 1999)

- 径向可视化图的想法是利用弹簧常数来表示数据点之间的相关值
- 对 p 维数据, 选择 p 个点等距散布在圆周上 (这些点称为固定点), 每个固定点连接一个弹簧的一头, 弹簧的另一个头链接到一个点 (该点为映射后的点), 该点的 p 个弹簧力和为 0
- 弹簧常数 K_i 等于数据点的第 i 个坐标值
- 所有的数据点一般需要通过正则化映射到 0 到 1 之间. 当一个数据点的所有 p 个坐标值都相同时, 映射点将为圆心; 如果数据点为单位向量, 则恰好映射到圆边上的固定点. 多个数据点可以映射到同一个点.
- 这种方法是数据的一个非线性映射, 其保持了某种对称性. 各坐标近似相同的数据点被映射后靠近圆心; 具有类似值但是坐

标方向相互相反的点映射后也靠近圆心

- p 维直线映射成一条线; 一个球映射为一个椭圆; p 维平面映射为一个有界多边形



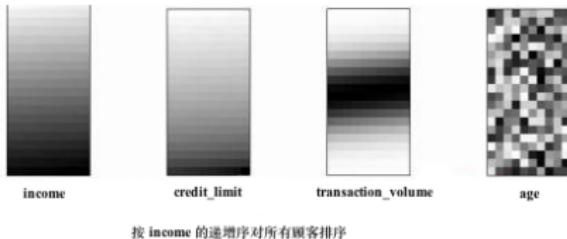
几何投影方法总结

- 可以处理大的或者非常大的数据集，但是对大的数据集可能会出现点或者类的重叠
- 可以合理的处理中高维数据集
- 所有变量同等对待，轴的顺序对图形的形状进而可视化发现结果具有影响
- 能够有效发现异常点和不同变量之间的相关性等

1.3.2 基于像素的可视化技术

- 用一个像素的颜色来表示一个数值的大小
- 对 n 个 p 维数据点集, 在屏幕上创建 p 个窗口, 每一维一个. 数据集的每一维值大小通过一个窗口的 n 个像素颜色来表示
- 在窗口内, 数据值按所有窗口共用的某种全局序安排. 据此可分为查询相关 (query-dependent) 和查询无关 (query-independent) 两种
- 查询相关就是给定一个点, 按照所有数据点与该点的相似度排序; 查询无关则是按照数据的某个维自然顺序排列.
- 适用于数据量巨大场合, 利用发现变量之间的关系, 趋势等.
- 对理解数据在多维空间里的分布帮助不大, 例如并不能显示多维子空间是否存在稠密区域.

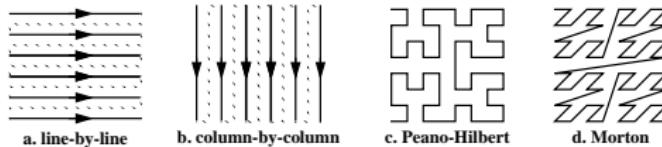
- 一家电子商务网站的顾客信息表，包含 4 个维：income, credit_limit, transaction_volume 和 age，则可通过基于像素的可视化技术发现 income 和其他变量之间的相关性（值越小，颜色越淡）：



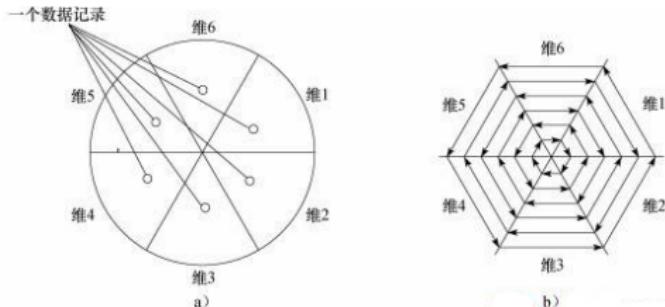
- credit_limit 随 income 增加而增加
- 收入处于中部区间的顾客更可能从该电子商购物
- income 和 age 没有明显相关性
- 当窗口过宽时，以线性方法安排数据记录填充窗口的效果可能不好。每行的第一个像素与前一行的最后一个像素离得太远，

尽管它们的对象在全局序下是彼此贴近的. 此外, 像素与它上面的像素相邻, 但这两个像素的对象在全局序下并非彼此相邻.

- 为此, 使用空间填充曲线 (space-filling curve) 来安排数据记录填充窗口 (图 c 和 d)



- 窗口也不必是矩形的. 例如:



基于像素的可视化技术总结

- 在高解析度的显示器上可以处理大规模以及非常大规模的数据集
- 可以合理的处理中高维数据集
- 每个数据点被唯一的映射到一个像素，因此不会出现点或者类的重叠现象

1.3.3 层次化可视化技术

对大维数据集, 很难同时对所有维可视化. 层次化技术将所有维划分为子集 (即子空间), 这些子空间按层次可视化, 从而将所有维数展示出来.

Dimensional Stacking(Leblance et al., 1990)

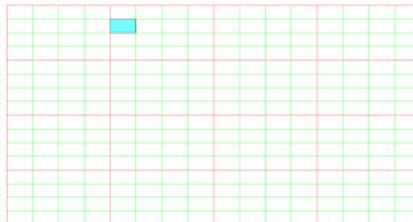
- 将 p 维数据空间划分为嵌套堆栈的 2 维子空间
- 适用于每一维具有有序低势 (不同可能值的个数) 的低维数据集 (一般 $p \leq 20$)
- 过程:
 - 选择两个最重要的变量 x_i 和 x_j , 根据它们的势定义 2 维格子
 - 在剩余的变量中再选择两个最重要的变量, 根据其势将每个格子划分为更小的格子, 重复直至所有变量被表示 (奇

数个变量时候引入哑变量)

- 使用颜色标出数据点所处的格子 (颜色可以依据变量的值, 或者格子内数据的頻数)

例如对四个变量:

变量	最小值	最大值	势
研究生	0	10	4
教师	0	20	4
助教	0	10	4
助研	0	10	4

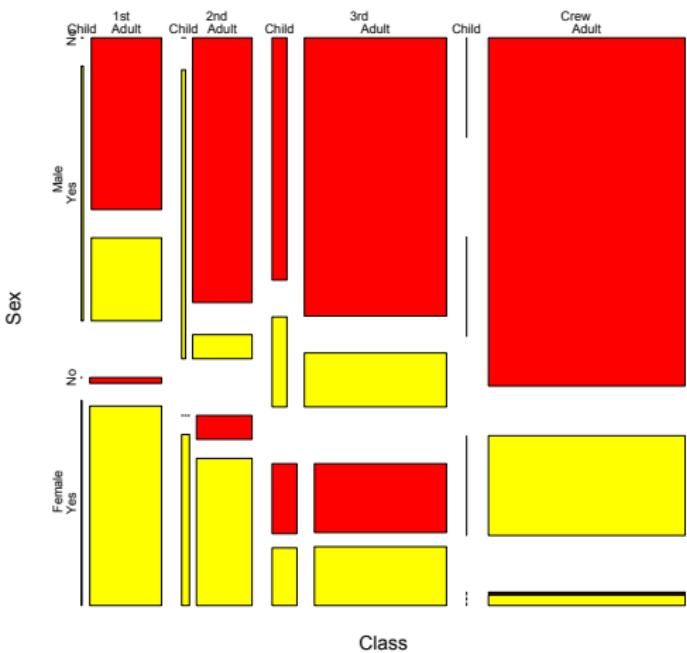


颜色强调的格子表示: 研究生 (2.5-5.0, 最外层, 水平方向); 教师 (15.0-20.0, 最外层, 垂直方向); 助教 (0.0-2.5, 内层, 水平方向); 助研 (5.0-7.5, 内层, 垂直方向).

Mosaic Plot(Friendly, 1994)

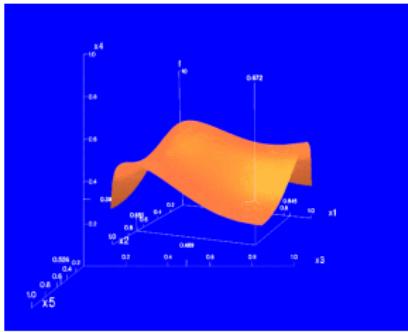
- 一种公认的用于属性数据的可视化方法
- 在 m 维列联表中通过嵌套的矩形表示数据的频数
 - 矩形的面积与数据的频数成比例
- 过程:
 - 首先将一个正方形在水平方向上按照变量 X_1 的边际总数成比例的分割成矩形
 - 对 X_1 的每个类别, 在垂直方向上按照 X_2 的条件频率成比例分割该矩形
 - 然后, X_1 和 X_2 的每个类别组合的矩形按照 X_3 的条件频率成比例分割
 - 重复以上过程, 直至所有变量都已经包含在图中

Survival on the Titanic



Worlds-within-worlds(Feiner and Beshers, 1990)

- “世界中的世界 (Worlds-within-worlds)”又称 n-Vision, 是一种代表性的层次 (动态) 可视化方法.



例 如 对 6 维 点
 $(f, x_1, x_2, x_3, x_4, x_5)$, 感兴趣是 f 维如何随其他维变化. 我们先将 x_3, x_4, x_5 固定在某选定的值, 比如 c_3, c_4, c_5 . 然后使用 3D 图对 (f, x_1, x_2) 可视化 (内世界). 如左图所示.

- 内世界的原点位于外世界的点 (c_3, c_4, c_5) 处. 外世界是另外一个三维图, 使用维 x_3, x_4, x_5 .
- 用户可以在外世界中交互地改变内世界原点的位置, 然后观察

内世界的变化结果. 此外, 也可以改变内外世界使用的维.

- 给定更多维, 可以使用更多的世界层

Treemap (Shneiderman, 1992)

- 树图也是一种流行的层次化可视化技术, 它把层次数据显示成嵌套矩形的集合.
- 下图是一个 Google 新闻的树图, 所有的新闻报道被分为 7 个类, 每个显示在一个唯一颜色的矩形中, 每个矩形的大小和不同媒体的新闻报道文章数成正比.
- 在每个类别内, 新闻报道进一步分成较小的子类.
- 用户能够快速识别那类新闻是“最热”的(报道最多的).

<http://newsmap.jp/>

Plushenko on thin ice in Russia after Olympic dropout		Another blast of snow makes its way into Northeast	
TJ Oshie leads United States to win in shootout over Russia	Sochi 2014: USA men edge Russia in ice hockey classic	US official tasked to task for it's Syrian conflict	
USA defeats Russia in rivalry game decided in shootout	Jermaine Marshall comes up huge in Arizona State's upset win over No. 2 Arizona	Duke survives close call against Maryland	College basketball: Top-ranked Syracuse remains unbeaten
Danica Patrick excited to be in Sprint Unlimited with boyfriend	American's Skeleton Run Goes Away	Florida emerges as national title favorite	Arsenal v Bayern Munich: Champions League ticket dancing to coach Pochettino
NBA All-Star Saturday 2014: Live event results from New Orleans	Yankees big-bucks pitcher Tanaka wobbled after 1-mile run	NBA players begin time-picking new union term	UConn tops Memphis in OT
Holiday Inn Express & Suites	Red Sox agree to terms with LHP Miller	Victor of Bullying	Wall Street Journal's Best Places to Work
Photo: AP Photo/Mark Humphrey	Photo: AP Photo/Matt Slocum	Photo: AP Photo/Elise Amendola	Photo: AP Photo/Jessica Hill
Sun News Staff Sun February 18, 2014 14:48:22	Sun News Staff Sun February 18, 2014 14:48:22	Sun News Staff Sun February 18, 2014 14:48:22	Sun News Staff Sun February 18, 2014 14:48:22

层次化可视化技术总结

- 可以处理小规模以及中等规模的数据集
- 更适宜处理低维至中维的数据集
- 变量被不同对待，不同的映射得到数据的不同可视化
- 可视化结果的解释需要经过训练

1.3.4 基于图标的可视化技术

- 基于图标 (icon-based) 可视化技术将多维数据点映射到一个图标 (icon/glyph).
- 一个图标总是由多个图形参数决定, 因此可以用来表示多维数据.
- 数据点的所有维没有等同对待, 因为在图标中图形元素不是同等重要突出的. 人对图形的认知差异在解释结果时导致偏差.
- 类似的数据点映射成图标后性状也类似, 因此利于发现数据中的类和趋势.

Chernoff faces(Chernoff,1973)

- 统计学家 Herman Chernoff 于 1973 年引入. 有助于发现数据中的趋势和类.

- 能够将多达 18 个变量 (维) 的多维数据以卡通人脸显示.

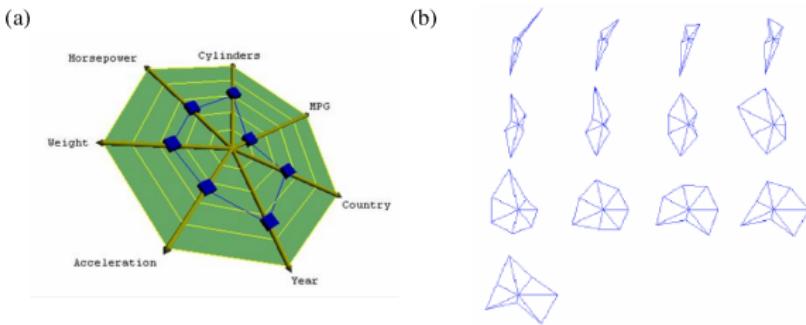
脸的要素, 如眼、耳、口、鼻等用其形状、大小、位置和方向表示维的值。例如, 维可以映射到如下面部特征: 眼的大小、两眼的距离、鼻子长度、鼻子宽度、嘴巴曲度、嘴巴宽度、嘴巴阔度、眼球大小、眉毛倾斜、眼睛偏离程度和头部偏离程度。



- 已经发现, 眼睛大小和眉毛的歪斜是重要的.
- 脸具有垂直 (关于 y 轴) 对称性, 因此脸的左右两边是相同的. 非对称的切尔诺夫脸使面部特征加倍, 这样允许显示多达 36 维.

Star glyph(Chambers et al. 1983)

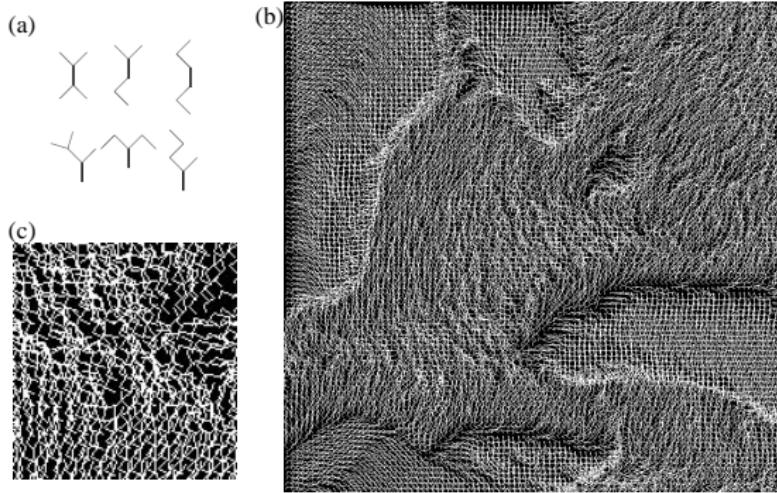
- 星形图是许多使用符号 (glyph) 可视化技术中最流行的一种.
- 对 p 维数据, 通过从圆心引出的 p 条等距射线来表示, 这些射线的长度代表变量的数值, 以直线链接射线的端点即构成一个星图.



- 每个数据点由一个星形图表示. 这样, 当数据点个数很大时, 同时可视化所有数据点来发现模式将变的不可能.

Stick figure(Pickett and Grinstein, 1988)

- 人物线条画 (stick figure) 可视化技术把多维数据映射到 5-段人物线条画, 其中每个画都有四肢和一个躯体. 两个维被映射到显示轴 (x 和 y 轴), 而其余的维映射到四肢角度和 (或) 长度 (下图 a)
- 如果数据项关于两个显示维相对稠密, 则结果可视化显示纹理模式, 反映数据趋势.
- 下图 b 为用人物线条画表示的人口统计数据.



基于图标的可视化技术总结

- 能够处理小规模到中等规模的数据集
- 可以用于高维数据集, 但是结果解释不直接, 需要一定的训练
- 变量被不同对待, 因为图标的一些特征比其他的更令人注意
 - 数据变量被映射到图标特征的方式严重决定了可视化的表达力以及感知力
- 定义合适的映射可能比较困难, 特别是更高维的数据
- 当一些变量映射到显示位置时, 数据点可能会重叠

最后

A picture is worth a thousand words