

平行坐标

肖何 00448247

Parallel Coordinates 平行坐标

平行坐标是一种通常的可视化方法，用于对高维几何和多元数据的可视化。

为了表示在高维空间的一个点集，在 N 条平行的线的背景下，（一般这 N 条线都竖直且等距），一个在高维空间的点被表示为一条拐点在 N 条平行坐标轴的折线，在第 K 个坐标轴上的位置就表示这个点在第 K 个维的值。

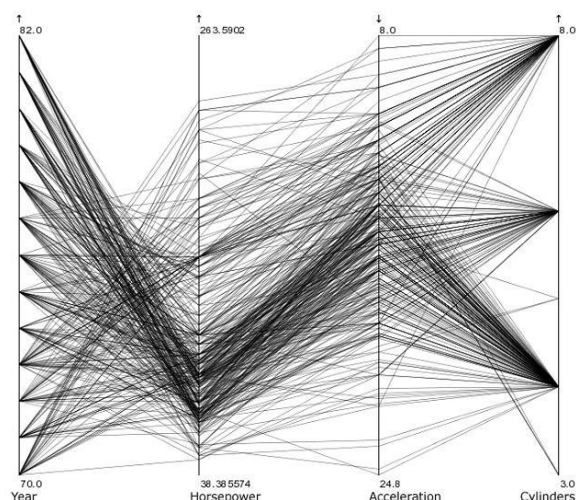


图 1. 一个平行坐标的例子。包含 4 个维度的信息，每根折线代表一个数据集中的数据。

1 平行坐标基础

平行坐标是信息可视化的一种重要技术。为了克服传统的笛卡尔直角坐标系容易耗尽空间、难以表达三维以上数据的问题，平行坐标将高维数据的各个变量用一系列相互平行的坐标轴表示，变量值对应轴上位置。为了反映变化趋势和各个变量间相互关系，往往将描述不同变量的各点连接成折线。所以平行坐标图的实质是将 m 维欧氏空间的一个点 $X_i(x_{i1}, x_{i2}, \dots, x_{im})$ 映射到 2 维平面上的一条曲线。平行坐标图可以表示超高维数据。平行坐标的一个显著优点是它具有良好的数学基础，其射影几何解释和对偶特性使它很适合用于可视化数据分析。^[1]

1.1 平行坐标历史

1.2 平行坐标基本原理

1.3 平行坐标的绘制

见参考文献^{[1][9]}

2 信息混淆处理方法(clutter)

平行坐标已经被证实是强大的可视化工具。但是当大的数据集应用平行坐标的表示方式的时候，我们会遇到线段混乱、平行坐标屏幕限制、折线重叠、不易于发现各维间的隐含关系的问题。大量的折线重叠在背景之上，造成视觉上的信

息混淆，这对我们观察数据的内在模式是很不利的图2。因此，许多处理平行坐标信息混淆的方法被开发了出来。

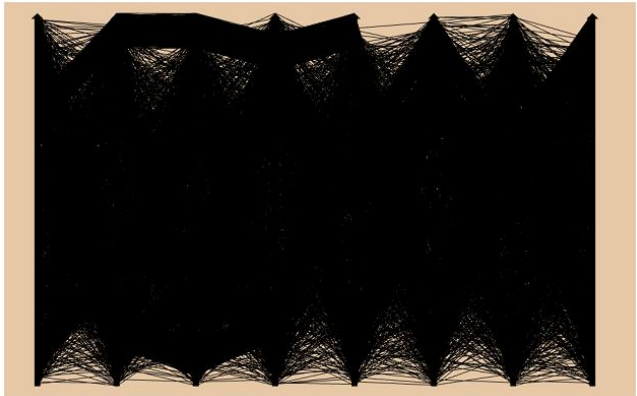


图 2.大量数据用平行坐标表示时造成的信息混淆。

2.1 维度重排^[2]

平行坐标将高维的数据集投影到平面的一系列平行坐标轴上。在实际观察平行坐标的过程中，我们发现相邻两个平行坐标轴所代表的维度在数据集上的模式可以通过观察得到直观的结果。而不相邻的坐标轴表达的信息的直观度就大大降低了。因此，如何决定坐标轴和维度之间的映射关系，也就是维度的顺序，对平行坐标的可视化有重要的影响。图3 图4

在平行坐标的显示中，当坐标轴的顺序变化时，数据点表现出显著不同的形状。参考文献[2]定义了一种衡量平行坐标混淆指数的方法。

我们首先定义在两个相邻维度之间，和周围的数据点孤立，有一定距离的点，称为局外数据点。因为这种局外数据点不遵循数据集的整体模式。有时我们能够通过这种局外数据点得到额外的信息。但这里我们认为局外数据点混淆了平行坐标的表示，我们要排列维度使得局外数据点的影响最小。

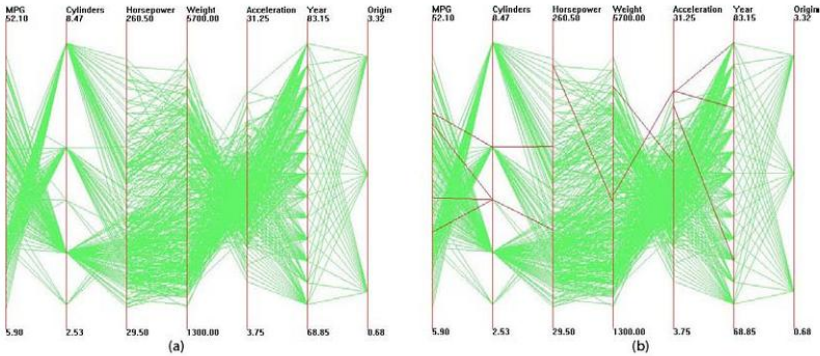


图 3.原始的维度顺序(a) 以及 局外点(红线表示)(b)

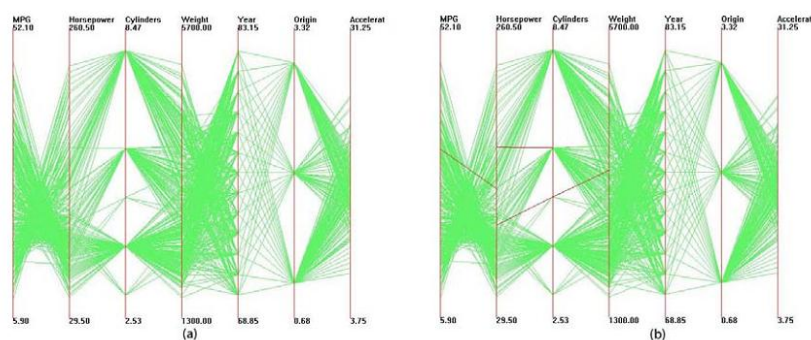


图 4.排列后的的维度顺序(a) 以及 局外点(红线表示)(b)

设每两个相邻之间局外数据点数目之和为 $S_{\text{局外}}$ 。设有 n 维度。那么平均每个维度之间的局外数据点有 $S_{\text{平均}} = S_{\text{局外}} / (n-1)$ 。设 $S_{\text{总}}$ 为数据集的数据总数，我们定义混淆系数 C :

$$C = S_{\text{平均}} / S_{\text{总}} = S_{\text{局外}} / ((n-1) * S_{\text{总}})$$

而重新排列维度的目的就是让混淆系数达到最小。

设数据集的总数为 m ，计算指定两个维度之间的混淆系数的时间复杂度为 $O(m^2)$ ，计算每一对维度之间的混淆系数的时间复杂度为 $O(m^2 n^2)$ 。然后我们需要 $O(n!)$ 的时间来搜索每一种维度排序的情况，每种情况需要 $O(n)$ 的时间来累加。搜索的时间为 $O(n * n!)$ 。整个算法的时间复杂度为 $O(m^2 n^2) + O(n * n!)$ 。

2.2 数据点着色

对数据点着色是重要的平行坐标信息反混淆方法。我们这里提到的对数据集的着色也包括 α 透明值的着色。在特定的颜色空间中，对数据集的每一个数据点 i ，我们都有一个 (r_i, g_i, b_i, a_i) 值跟数据点对应，其中 r, g, b 表示 3 原色红绿蓝， a 表示透明度。

通过应用不同的颜色空间，使得折线之间区别明显，或者突显一部分折线而使其他折线不明显，这样使用户更清晰地了解局部数据的变化规律可以更直观地观察感兴趣的数据。而且颜色空间可以使得我们对相邻较远的坐标轴表示的维度之间的数据模式关系有直观的了解。

有多种构造颜色空间的方法，用以达到不同的效果。

方法一，以某个坐标轴的维度为参照，对数据集的折线进行渐变着色图 5。这样我们就得到了其中一个维度数据在其他维度上分布的直观情况，揭示高纬度数据空间中的数据模式分布。

方法二，和刷方法结合，用户选择的部分集进行着色，未被选择的折线透明化处理或者消掉。刷的方法可以跟别的可视化的方法有机结合，从不同的数据显示角度进行 **brush**，用以观察特定高维数据集在平行坐标轴的趋势和模式表现。

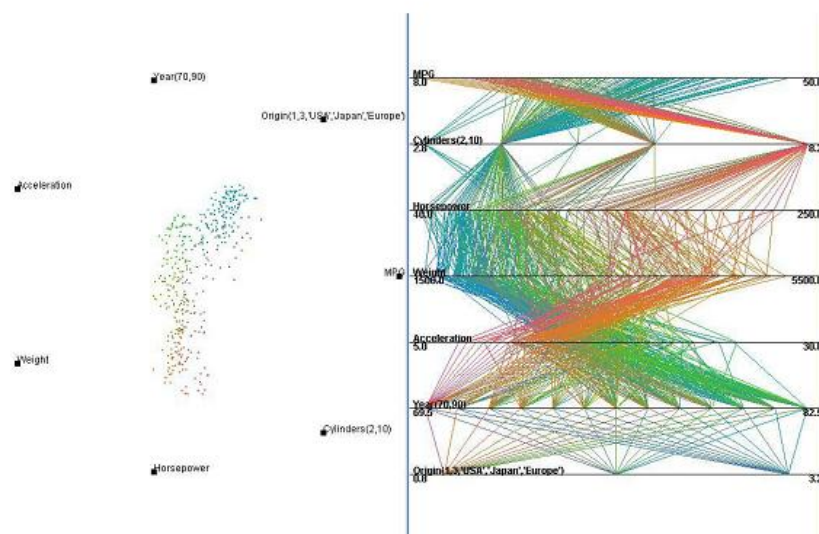


图 5.应用颜色空间的平行坐标表示^[17]

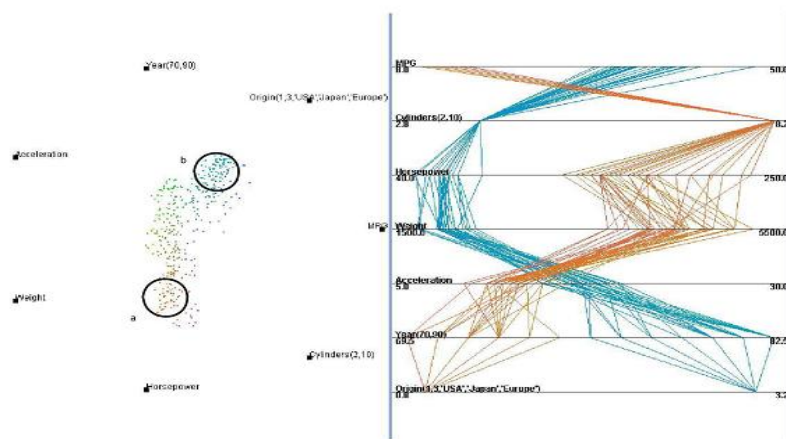


图 6.平行坐标表示,对用户选择的数据点进行着色, 其他数据点隐去^[17]

在实际的平行坐标的可视化过程中, 对大的数据集的可视化, 折线的透明化是必然的选择, 大量的折线重叠在背景之上, 很容易就达到了屏幕显示的极限。图 2 进行折线的透明化调整, 也是类似用密度参数来调整折线在屏幕上显示的亮度范围。把折线密度最高的地方调整为屏幕显示的最高极限, 而把折线密度最低的位置(一般是单根折线)调整为肉眼识别的下限。由此来达到更大的折线区分效果。

2.3 聚类

聚类把数据集按某种特征划分为少量的群组, 从而发现数据中的重要规则。进行聚类可以减少显示的数据集的数目, 也就是说我们把大数目的数据集映射到数目少的数据集之上, 再用平行坐标进行可视化, 以达到减少信息混淆, 展示信息模式的目的。

首先我们要定义数据之间的距离。

设数据集有 n 个数据点, 有 m 个数据维度。 X_{ij} 为第 i 个数据点在第 j 个维度上的投影。任意两个样品 i 与 j 之间的距离可定义为:

$$d_{ij} = \{ [\sum (X_{ik} - X_{jk})^2] / m \}^{1/2}$$

定义好了距离系数，我们就可以对数据集进行聚类了。聚类的具体算法我们这里不多做讨论。

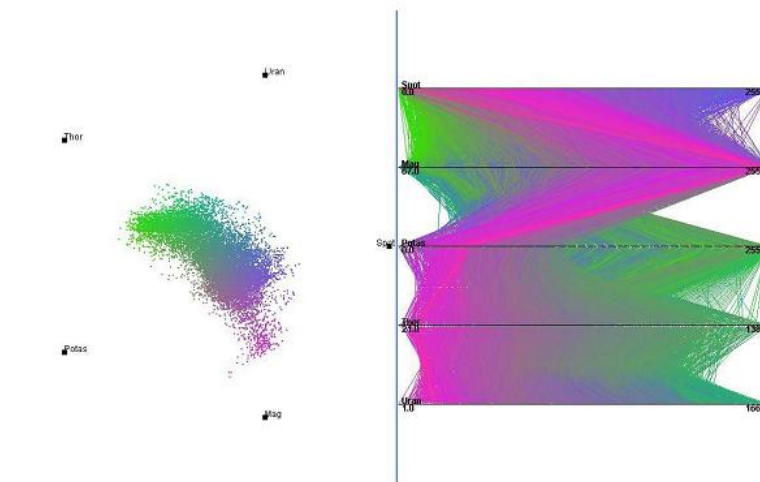


图 7.应用聚类之前的平行坐标(大数据集高混淆)

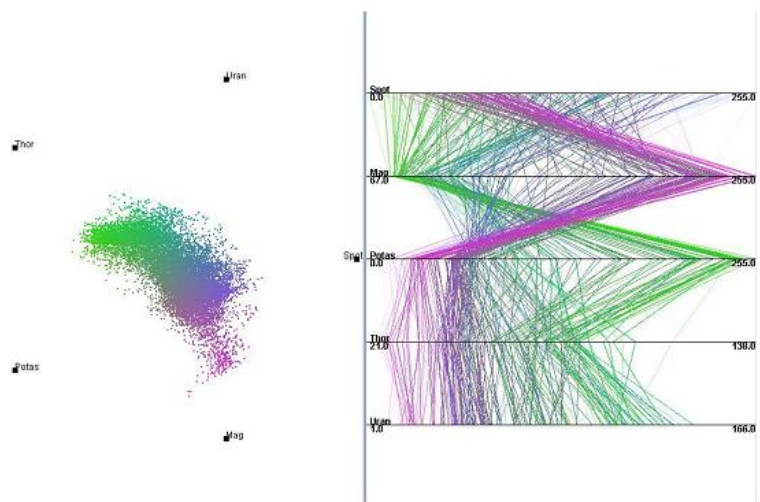


图 8.应用聚类之后的平行坐标(大数据集高混淆)，数据集的模式得到的直观的展示。

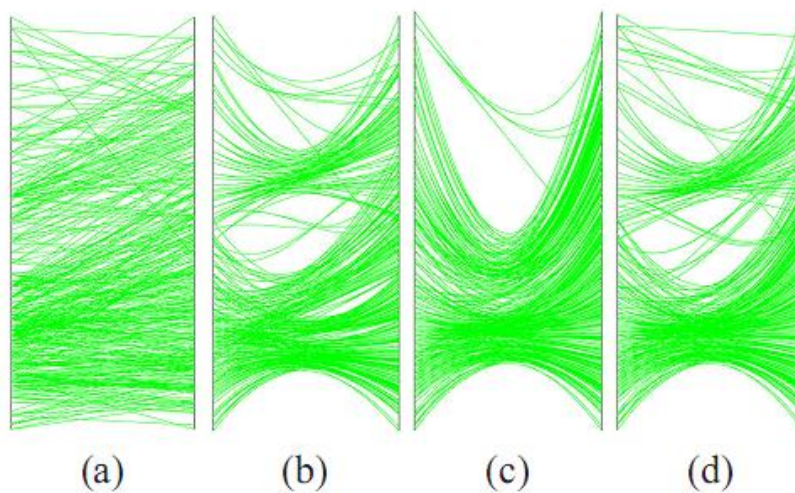


图 9.应用引力扭曲的聚类平行坐标的显示方法。[14]

图 7 图 8 是直接合并相同类的例子。图 7 是一个搞混淆的平行坐标的效果图，应用了颜色空间，但是折线的重叠情况严重，很难看到内部的数据模式。图 8 是合并了相同聚类后的效果图，每个类用他们的平均数据点来代表。可以看到聚类大大较少的数据集的大小，对平行坐标的重叠情况的改善非常有效果。我们可以从图 8 中得到更多信息。

图 9 显示了一种特殊的平行坐标聚类方法。这种方法应用了引力聚类的概念，将距离相近的数据点吸引到一起，以达到减少重叠的效果。通过调整引力的参数，（引力衰减系数方程）来获得不同的聚类效果。

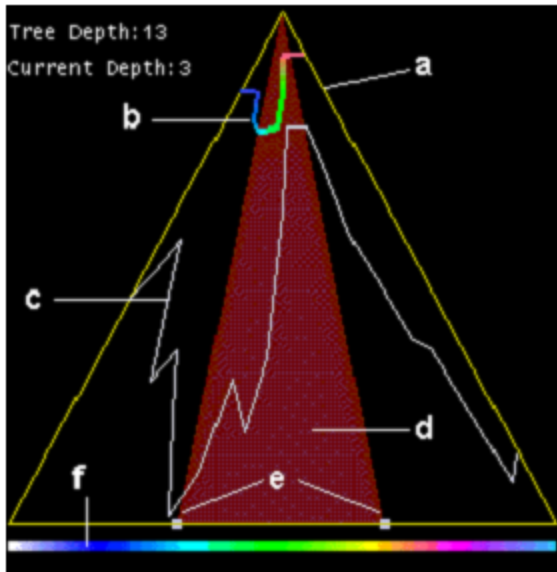


图 10. 分层聚簇树。 一条横贯树的线就是一个对树的分层。

分层平行坐标(Hierarchical Parallel Coordinates)^[4]也是一种特殊的聚类方法。采用分层显示模式对数据集进行多种层次的显示,克服平行坐标中数据显示混乱的缺点。分层显示模式中利用分层聚类算法构造分层聚簇树(Hierarchical Cluster Tree)^[4],分层聚簇树可以从不同抽象层次上构造和表达大型数据集。分层聚簇树上的每个节点都表达一个聚簇。图10^[30]

分层平行坐标图 11 是由普通平行坐标演变而来的,其中显示的是聚簇而不是像普通平行坐标一样显示单独的数据项。聚簇的平均值是一条穿过所有坐标轴的折线。折线周围的带表示聚簇覆盖的范围,带与坐标轴相交的上边缘是簇中数据在该维上的最大值,下边缘为最小值。平均值突现为红色的折线是要着重显示的部分。^[30]

通过和用户交互的分层聚簇树的选择

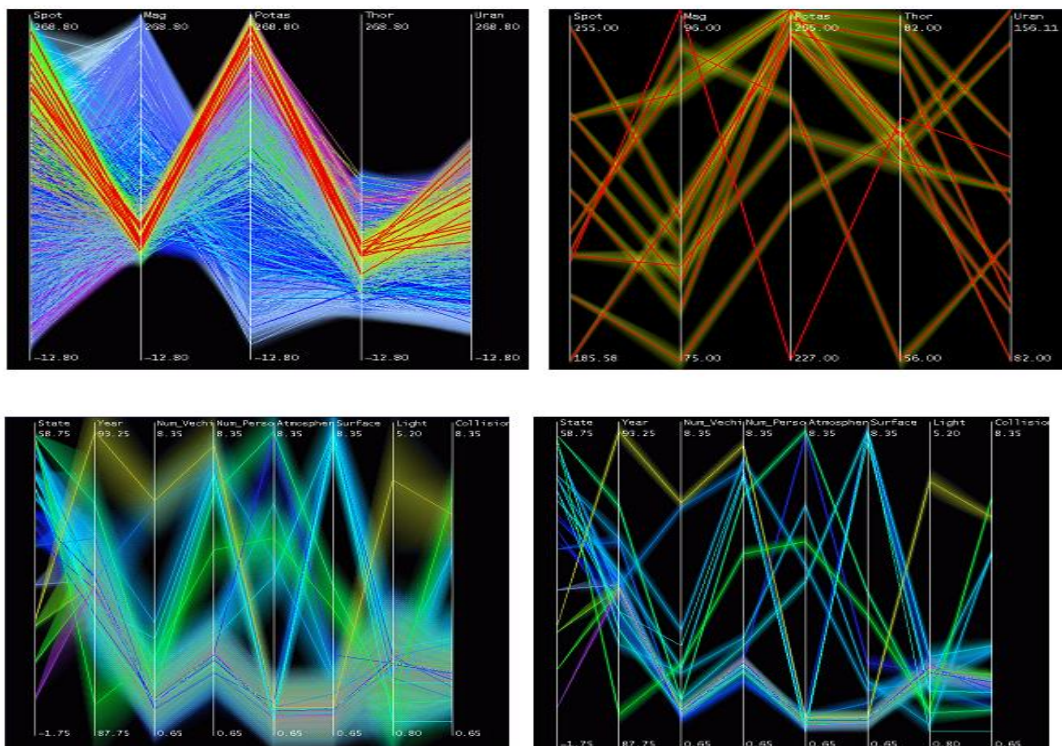


图 11.分层平行坐标。聚簇和聚簇的平均值。每个数据集的分层聚簇树对应一种分层平行坐标。

2.4 抽样^[20]

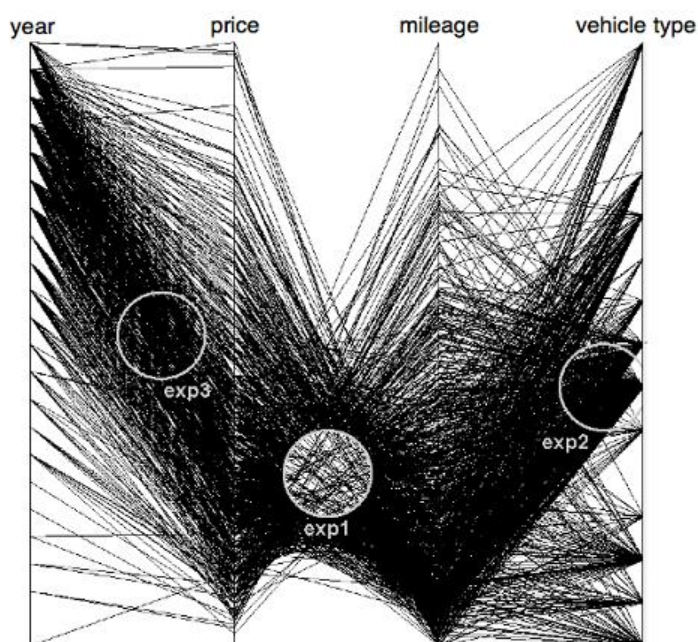


图 12.抽样透镜和 focus+context 的方法。

抽样的方法是很有效的信息反混淆的方法。如图 12 所示，对高度重叠的折线，在透镜的范围内应用抽样，使得一部分的数据模式信息能够直接被观察到。

对平行坐标应用抽样的方法来进行反混淆，优点是能解决折线重叠的问题，能给用户更清晰的观察视角。缺点在于抽样带来的信息的丢失。通过改进的算法，我们能够抽取更具代表性的样本。这样的抽样有些类似于聚类的算法，用更精简的簇的平均值来代表簇的信息。

2.5 动画^{[6][13]}

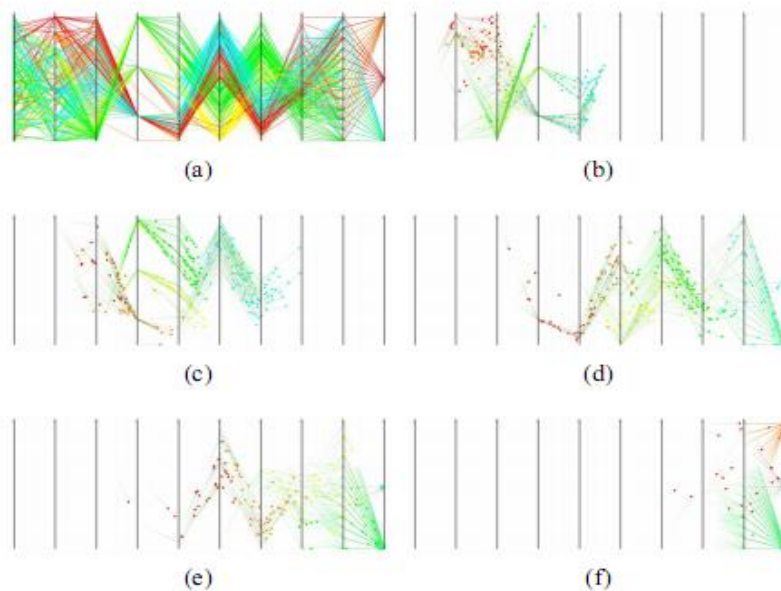


图 12.对平行坐标应用时间轴的变化，达到动画的效果。

在平行坐标的展示中加入时间维度的平行坐标变化，那么在同一时间在平行坐标的区域内折线的密度就会降低。图 12 展示的是在垂直于坐标轴的方向上应用时间维度的变化。在同一时刻，数据集用颜色空间中的颜色标识的点表示折线的当前位置，用渐隐的半透明的线来提示折线的变化历史记录。

在平行于坐标轴的方向上也可以加入时间维度的变化。以选定的一个坐标轴做参照，在这个维度上的数据依次扫描过整个数据空间，同样也能达到反混淆的效果，同时能展示足够多的信息。

3 平行坐标的变种

在基础的平行坐标的显示方式上，加入其新的特性，得到新的平行坐标的显示方式。

3.1 三维平行坐标（3-Dimensional Display）

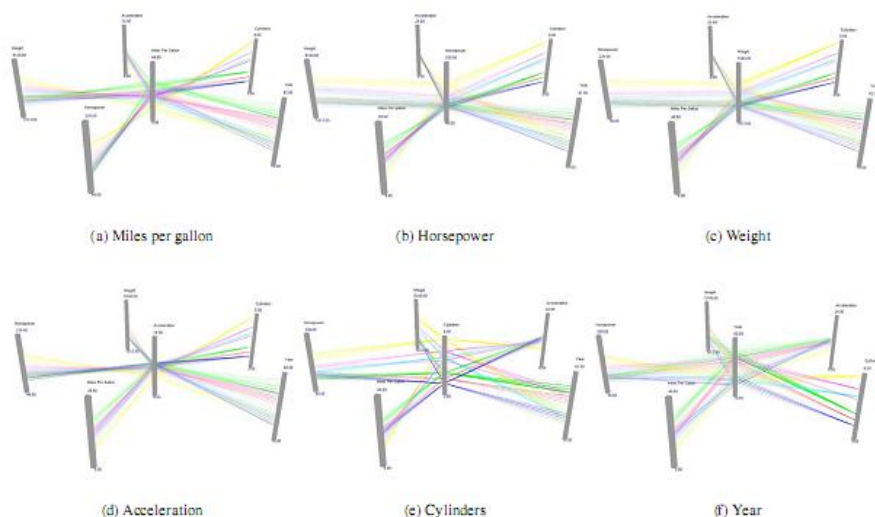


图 13.3 维的平行坐标的形式。 a-f 以不同的维度为中心轴。

将平行坐标的坐标轴投影在三维空间，就得到了三维平行坐标图 13。在三维空间中我们能得到更多的相邻坐标轴以及维度。然而在三维空间中折线重叠情况也更为复杂。

其中 5 个坐标轴和中心轴的数据关系都能直观的展现出来。这样的平行坐标相比普通的平行坐标有更高的表示效率。但在数据点密集的时候，三维平行坐标的缺点也很明显，那就是前排的折线会遮挡后排的折线。这在数据密度低的时候不明显，而在数据集很大的时候，三维坐标的显示效果就很糟糕了。前排和后排在视觉上都被混乱，完全表达不出任何的信息。

3.2 平行坐标和星形图 (Parallel Coordinates and Star Glyphs) ^[11] 图 14

三维平行坐标是坐标轴扩展到空间中的结果，而这里提到的平行坐标和星形图是对坐标轴本身的扩展。我们把坐标轴扩展为一个类似坐标平面的概念。因为如果在整个平面上都分布有点，那么最后展示出来的效果会非常的混乱。所以我们将坐标平面替换为了星形图。通过星形图，我们可以消除折线在圆柱面投影的交叉。(虽然在用户的视线中仍然有交叉，不过混乱程度已经大大降低了)图 15 演示了将普通的平行坐标扩展为星形图的步骤。一步步展开，消除折线的交叠，就是平行坐标和星形图结合的意义。

当折线在圆柱面上的投影完全展开平行后，数据集的所有信息其实都蕴含在了星形图之中，折线已经是多余的了，我们通过观察星形图的形状来获得数据的模式信息。

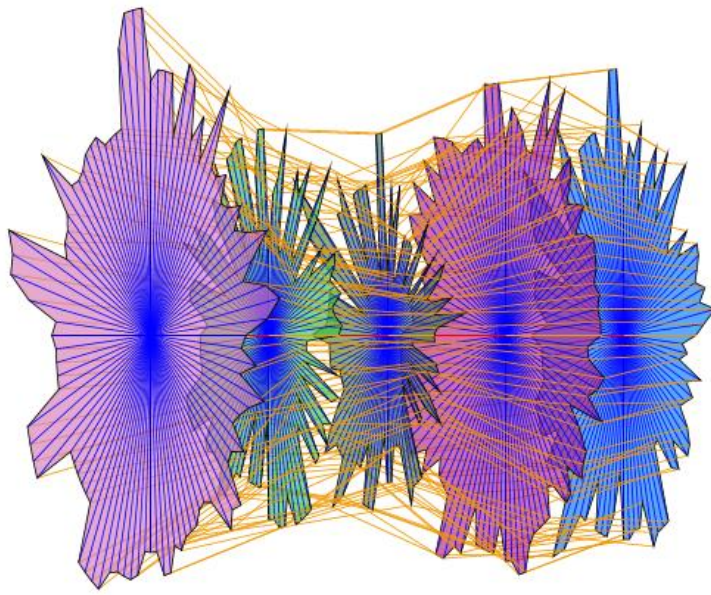


图 14.平行坐标和星形图的结合、

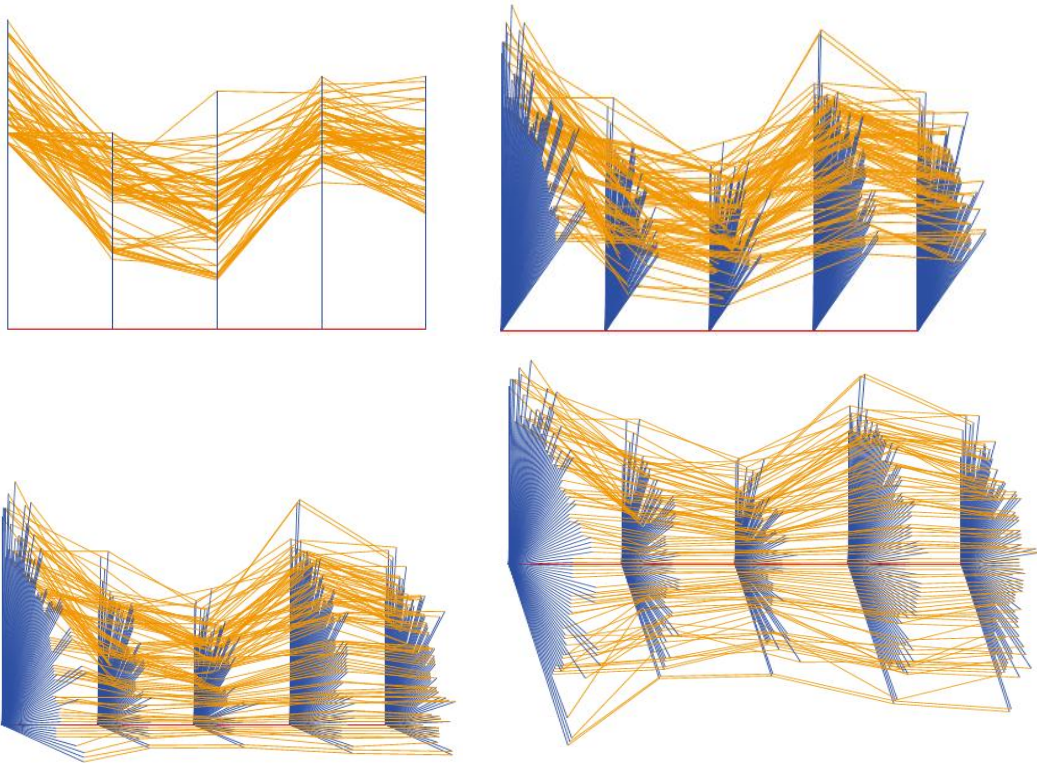


图 15.星形图的展开

3.3 平行集合(Parallel Sets) [22][23]

当平行坐标在每个维度的取值都非常离散时，我们可以应用一种特殊的平行坐标——平行集合。图 16

平行集合的每个维度的每个取值在坐标轴上有一个范围，这个范围由数据集在这个维度取这个值的点的数目比上数据集的大小来决定。相当于每个坐标轴都是一个集合的划分。而每个集合中的元素通过一条有宽度的折线连接到相邻坐

标轴的集合。这样的平行集合有许多不确定因素影响到可视化的效果，类似于集合在坐标轴上的排布顺序，不同的折线在连接相邻坐标轴集合中元素的具体位置。这些因素都是需要谨慎决定的。

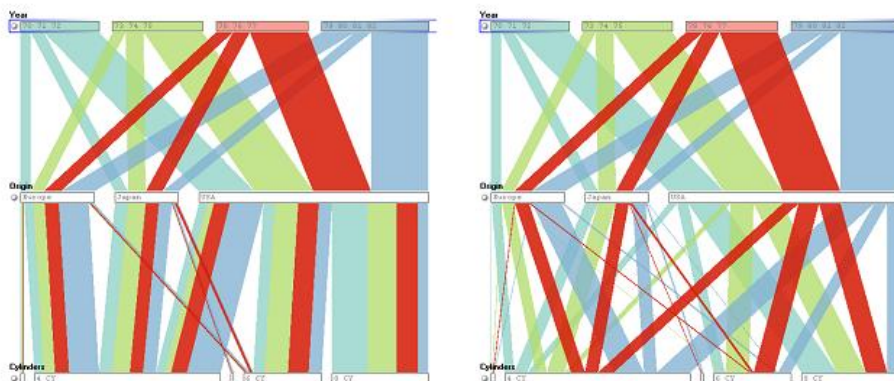


图 16.平行集合的例子

4 结论

平行坐标对于信息可视化有着相当重要的作用，是目前对高维数据的可视化必不可少的应用工具。目前的研究工作的重点都集中在平行坐标工具的拓展，以及平行坐标信息混乱混淆的情况应对措施，发展出了很多有效的消除混淆，显示信息的方法。这些方法有的非常有效，有的是可以普遍应用的方法。不过目前平行坐标的信息混淆仍然是困扰用户的首要问题。平行坐标的高维数据显示能力，与其他可视化工具的结合仍然有很大的研究潜力。目前在平行坐标与其他方法的组合上面研究成果相对比较少。现在的平行坐标的工具，在和数据本身的特性结合上做的不够，研究方向主要在平行坐标工具本身。平行坐标在显示数据模式时，需要很强的交互来让用户自己发掘数据内在的规律，不够智能化。平行坐标向三维空间的扩展仍然有发展潜力。这都是我猜测的可能的研究方向。关于平行坐标的可视化工作，仍然有待更多的进展。

参考文献：

- [1]徐永红，高直，金海龙等.平行坐标原理与研究现状综述. 燕山大学学报.2008年9月第32卷第5期.289-392页.
- [2]Peng W, Ward MO, and Rundensteiner EA. **Clutter reduction in multi-dimensional data visualization using dimensional reordering.** *Proceedings of the IEEE Symposium on Information Visualization 2004*, edited by Keahey A, Austin TX, 2004, p. 89–96.
- [3]H. Hauser, F. Ledermann, and H. Doleisch. **Angular brushing of extended parallel coordinates.** In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*,2002.
- [4]Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. **Hierarchical parallel coordinates for exploration of large datasets.** Technical report, Worcester Polytechnic Institute, 1999.
- [5] A. Inselberg and B. Dimsdale. **Parallel coordinates: A tool for visualizing multidimensional geometry.** In *Proc. Of IEEE Conference on vis'90*, pages 361–378,

1990.

[6] L. J. S. N. Barlow. **Animator: A tool for the animation of parallel coordinates.** In Proceedings of the Eighth International Conference on Information Visualization (IV04), 2004.

[7] H. Siirtola. **Direct manipulation of parallel coordinates.** In IEEE International Conference on Information Visualization, pages 373–378, 2000.

[8] K. Zhao, B. Liu, T. M. Tirpak, and A. Schaller. **Detecting patterns of change using enhanced parallel coordinates visualization.** In Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), 2003.

[9]. A. Inselberg, “**The plane with parallel coordinates,**” The Visual Computer 1, pp. 69–91, 1985.

[10]. J. Johansson, M. Cooper, and M. Jern, “**3-dimensional display for clustered multi-relational parallel coordinates**” in 9th IEEE International Conference on Information Visualization, pp. 188–193, 2005.

[11]. E. Fanea, S. Carpendale, and T. Isenberg, “**An interactive 3D integration of parallel coordinates and star glyphs**” in IEEE Symposium on Information Visualization, pp. 149–156, 2005.

[12]. E. J. Wegman, “**Hyperdimensional data analysis using parallel coordinates**” Journal of the American Statistical Association 85(411), pp. 664–675, 1990.

[13] **Animation for Clutter Reduction in Parallel Coordinates**

[14] Hong Zhou, Xiaoru Yuan, Huamin Qu1, Weiwei Cui, Baoquan Chen. **Visual Clustering in Parallel Coordinates.** Eurographics IEEE-VGTC Symposium on Visualization 2008 *Volume 27 (2008), Number 3*

[15] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz. **Uncovering clusters in crowded parallel coordinates visualizations.** In 10:th IEEE Symposium on Information Visualization, pages 81–88, 2004.

[16] MR Berthold, LO Hall, T Inc, CA San Francisco. **Visualizing fuzzy points in parallel coordinates.** IEEE Transactions on, 2003

[17] E.D. Bertini, L. Aquila, and G. Santucci. “**Springview: Cooperation of Radviz and Parallel Coordinates for View Optimization and Clutter Reduction**”. in Proceedings, the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2005 (CMV 2005). p. 22- 29, 2005.

[18]M. ten Caat, N. M. Maurits, and J. B. T. M. Roerdink. **Tiled parallel coordinates for the visualization of time-varying multichannel EEG data.** In Proceedings Euro-graphics/IEEE VGTC Symposium on Visualization, pages 61–68, 2005.

[19]G. Ellis and A. Dix. **A taxonomy of clutter reduction for information visualisation.** IEEE Transactions on Visualization and Computer Graphics, 13(6):1216–1223, 2007.

[20] GP. Ellis and A. Dix, “**Enabling Automatic Clutter Reduction in Parallel Coordinate Plots**”, IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis'06), 12(5), pp. 717-723, Sept 2006

[21] Wegman, E. J., and Solka, J. L. (2002), **On some mathematics for visualizing high dimensional data,**” Sanhkyā (A), 64 (2), 429{452.

[22] BENDIX F. , KOSARA R., HAUSER H.: **Parallel sets: visual analysis of**

categorical data. In Proc. of IEEE Symp. on Information Visualization (2005), pp. 133–140.

[23] KOSARA R., BENDIX F., HAUSER H.: **Parallel sets: Interactive exploration and visual analysis of categorical data.** IEEE Trans. on Vis. and Comp. Graph. 12,4 (2006), 558–568.

[24] Yang, L. (2005). **Pruning and visualizing generalized association rules in parallel coordinates.** IEEE Transactions on Knowledge and Data Engineering, 17, 60–70.

[25]. J. Johansson, P. Ljung, M. Jern, and M. Cooper, “**Revealing structure within clustered parallel coordinates displays,**”in IEEE Symposium on Information Visualization, pp. 125–132, 2005.

[26]C Forsell, J Johansson. **Task-based evaluation of multi-relational 3D and standard 2D parallel coordinates.** SPIE's International Symposium on Electronic Imaging

[27] A. Inselberg and T. Avidan, “**The automated multidimensional detective,**” presented at the IEEE Symp. Information Visualization, San Francisco, CA, 1999.

[28] GRAHAM M., KENNEDY J.: **Using curves to enhance parallel coordinate visualizations.** In Proc. of Intl. Conf. on Information Visualization (2003), pp. 10–16.

[29]K Daniel Ericson, Jimmy Johansson and Matthew Cooper. **Visual Data Analysis using Tracked Statistical Measures within Parallel Coordinate Representations.** In Proceedings of IEEE International Conference on Coordinated & Multiple Views in Exploratory Visualization, pages 42–53. London, UK, 2005.

[29]Brodbeck, D. and Girardin, L. (2003). “**Visualization of Large-Scale Customer Satisfaction Surveys Using a Parallel Coordinate Tree**”, 9th IEEE Symposium on Information Visualization (InfoVis 2003). Seattle, USA: IEEE Computer Society.

[30]翟旭君,李春平.平行坐标及其在聚类分析中的应用. 计算机应用研究 2005 年 42 1