TVBPS:一种基于 Parallel Sets 的具有度量属性的多变元时态数据可视化方法

孙宁伟[1], 刘海峰[1], 赵 瑜[1], 刘 勇[1], 王 璐[1], 肖卫东[2]
(1. 中国人民解放军 91655 部队, 北京 100036; 2. 国防科学技术大学 信息系统工程重点实验室, 长沙 410073)

摘 要: 针对现有"具有度量属性的多变元时态数据"可视化方法不足,提出 Parallel Sets 分类值排列顺序优化算法 ACLEARCR、基于相关度的 Parallel Sets 变元轴配置算法(VABC)、深度信息 Parallel Sets(DCPS)共同组成基于 Parallel Sets 的具有度量属性的多变元时态数据可视化方法 TVBPS。使用具体数据集对提出的可视化方法进行实验,获得的视图能够挖掘数据集中的隐含知识,证明了该方法的有效性。TVBPS 可视化方法为分析多变元时态数据集提供了有效手段,具有较高的适用性和易用性。

关键词: 度量属性; 多变元; 时态; 信息可视化; Parallel Sets
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2014)05-1591-06
doi:10.3969/j.issn.1001-3695.2014.05.076

TVBPS:time-varying multivariate data visualization method based on Parallel Sets

SUN Ning-wei[1], LIU Hai-feng[1], ZHAO Yu[1], LIU Yong[1], WANG Lu[1], XIAO Wei-dong[2]
(1. *Unit* 91655 *of Chinese People's Liberation Army*, *Beijing* 100036, *China*; 2. *Science & Technology on Information Systems Engineering Laboratory*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: Facing the weakness of the existing visualization method for "time-varying multivariate data with measures", this paper proposed a new visualization method based on Parallel Sets TVBPS, which was composed by a categories layout algorithm ACLEARCR (advanced categories layout based on average heuristic with cardinality reduction), a variate arrangement algorithm based on correlation VABC (variate arrangement based on correlation), and a algorithm called DCPS (depth cue parallel sets). Case studies demonstrate the effectiveness of the method, it shows that the method is conducive to mine the implicit mode in the data sets.

Key words: measures; multivariate data; time-varying data; information visualization; Parallel Sets

在社会科学、环境监测、金融经济、医疗卫生及地理信息等领域中,研究人员通常能够在连续时间段内获得大量多变元数据,既具有多变元特性又具有时态特性,是具有多变元数据和时态数据特点的多类型数据集,称为多变元时态数据集。某些多变元时态数据集中的数值型变元可能是分类型变元的度量,即数据集中的数值型变元的取值表示该数据项分类型变元所确定的类别的具体数量,此时该数值型变元具有了特定的意义,它通常表示数量、比率等,并且必须与相应记录的分类型变元相结合才能表达具体的物理含义。称这种类型的数值型变元(属性)为度量属性,相应的数据类型称为具有度量属性的多变元时态数据。例如,在经济领域的销售数据通常包括多个时间点的销售情况,每个时间点的数据包括多个变元(如产品类型、销售地点、销售数量等),每个变元随时间产生变化又呈现时态特点,销售数量即是度量属性。

多变元时态数据可视化方法研究已经受到国内外学者普遍重视,文献[1~12]针对具体数据集分别提出了众多可视化方法,但均是将多变元可视化视图与时态可视化视图拼凑在一起。多变元可视化视图侧重于构建能够保持原多变元数据拓扑结构的低维展现,以辅助用户在可视空间中分析各数据项多变元属性间的相互关系;时态可视化视图则着重体现数据项之间的演变规律。两视图之间通过联动支持用户对多变元时态数据进行可视化分析。这类方法能够通过对多个视图的综合判断获得感兴趣的知识,但增加了认知负担,且难以分析多个变元随时间的变化趋势情况。

1 相关工作

1.1 Parallel Sets 介绍

Parallel Sets[13,14]是经典的数据可视化方法之一,尤其适合于对具有度量属性的多变元时态数据进行可视化展现。该方法的思想来源于平行坐标系[15~17](parallel coordinates plots, PCP)。PCP 使用竖直的平行坐标轴表示多个变元,同一数据项每个变元上的取值之间用直线相连;而 Parallel Sets 使用水平的平行坐标轴表示数据集中的各变元,使用包围盒(boxes)表现各分类值,boxes 的跨度范围正比于分类值的频率,该频率实际上是该分类值的度量,不同的平行四边形使用不同的颜色

收稿日期: 2013-06-12; **修回日期**: 2013-07-23

作者简介: 孙宁伟(1987-),男,山东临朐人,助理工程师,硕士,主要研究方向为信息可视化、计算机网络(sunningwei@sina.cn);刘海峰(1973-),女,高工,硕士,主要研究方向为计算机网络;赵瑜(1974-),女,工程师,硕士,主要研究方向为计算机网络;刘勇(1976-),男,工程师,硕士,主要研究方向为计算机网络;王璐(1981-),女,工程师,硕士,主要研究方向为计算机网络;肖卫东(1968-),男,教授,博导,博士,主要研究方向为信息资源管理.

渲染。在交互性方面,Parallel Sets 提供了选取高亮、交互式查询、过滤、变元轴和分类值重新排序等方式支持多变元分析,如图 1 所示。

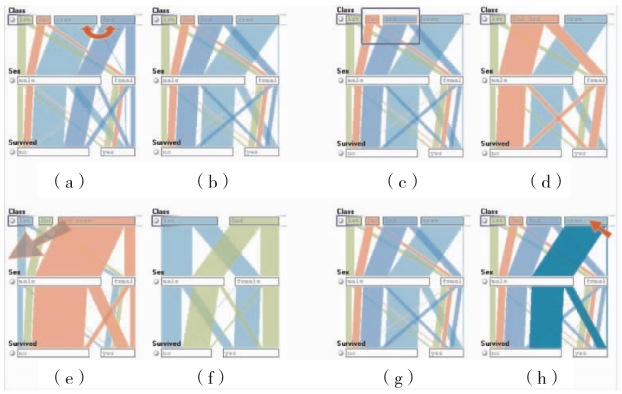

图 1 Parallel Sets 及其基本的交互式操作

Parallel Sets 变元轴的排列顺序由用户选择顺序决定,且每条变元轴上的分类值排列顺序任意排列,会使变元轴间的关联平行四边形产生较多交叠,导致可视化视图出现混乱,数据集较大情况下,难以发现数据集中的隐含模式。虽然 Parallel Sets 允许用户通过拖拉交互的方式排列变元轴和轴上各分类值的排列顺序,但手动操作的方式是一项十分繁杂的任务,且难以保证最终获得满意的视图。

为克服上述缺陷,需要对变元轴的排列顺序及每条轴上各分类值的排列顺序进行自动优化,为进一步的分析提供较好的视图基础。由于 Parallel Sets 中未体现时态特性,为支持多变元时态数据可视化分析,还需要集成时态视图。

1.2 分类值排列顺序方面

文献[18]针对 Parallel Sets 任意排列分类值产生较多交叉的不足,提出了基于中位数的启发式分类值排列算法 CAME(categories arrangement based on median heuristic)。

文献[19]进一步提出基于平均数的带降势启发式分类值布局算法 CLEARCR(categories layout based on average heuristic with cardinality reduction),自动优化各变元轴上分类值的布局,能够有效减轻视图中的可视混乱,适用于数据量较大、分类值较多的数据集,并使用改进的 Parallel Sets 方式对全球恐怖袭击数据进行分析,能够辅助用户获取不同恐怖组织的行为特征等隐性信息。

1.3 变元轴排列顺序方面

Ankerst 等人[20]提出基于相似度的维排列算法,将相似属性排列在相近位置,维排列算法与货郎担问题(travelling salesman problem)具有相同的复杂度,属于 NP 难问题,因此提出了一个启发式算法来解决该问题。

文献[21]为减轻多变元数据可视化中的视图混乱问题,提出一种变元排列算法 SBAA(similarity-based attribute arrangement)应用于可视化视图坐标系排列中,提出降维算法 SBAR(similarity-based attribute reductiont)、mSBAR(modified SBAR)从视图中去除高相关的属性,以减少视图中的混乱。上述方法均使用基于相似度的思想,提出相似度度量指标。

文献[22]同样采用类似的思想,对上述算法进行扩充并应用于星型坐标系,提出了基于相关度的星型坐标系维度轴配置策略,并应用于多变元数据可视化,取得了较为理想的效果。

1.4 集成时态视图方面

文献[23~25]分别对时态数据可视化进行了研究,但是提出的方法均通过增加一个时态维度和通过层次方式展现数据集中的时态特点。文献[26]对相关研究进行了详细综述,指出增加一个时态维度的方法会导致数据集多变元特点和时态特点的分离,不便于进行综合分析。文献[6]在 PCP 中集成数据的时态特点,提出了 TDPC(temporal density parallel coordinates)和 DCPC(depth cue parallel coordinates)技术,基于密度图和转换函数、时态切片(temporal binning)等技术通过层次渲染的方式表现不同的时间段,能够支持对上千个数据项、上千个时间段的多变元数据进行时态分析。

2 基于 Parallel Sets 的具有度量属性的多变元时态数据可视化方法

2.1 定义与说明

参考文献[19],将 Parallel Sets 可视化视图进行抽象,将变元轴上各分类值抽象为节点,节点带有权重,连接不同变元轴的平行四边形抽象为线段。若不同变元轴上分类值之间存在相连线段,说明数据集中存在包含对应两分类值的记录,且连线带有权重信息。Parallel Sets 视图被抽象为带权层次网络图,如图 2 所示。该层次网络图可以形式化描述为 $G(L_1,L_2,\cdots,L_{n-1},L_n,E)$,其中 L_i 表示每一层次的一组有序节点,$E\subseteq(\cup(L_i\times L_{i+1}))$,$i=1,\cdots,n$ 代表节点组 L_{i+1} 与其上层节点 L_i 连线的集合,集合中的任一元素 uv 表示节点 $u(u\in L_i)$ 与节点 $v(v\in L_{i+1})$ 的连线。

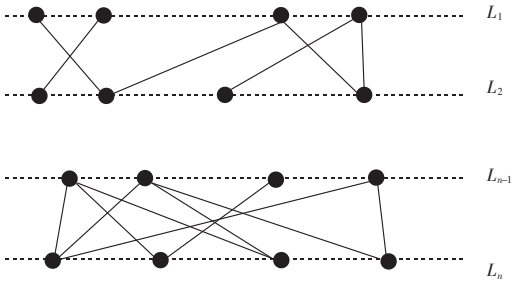

图 2 层次网络图

Parallel Sets 变元轴的排列顺序及每条轴上分类值的排列顺序决定了可视化视图的混乱程度,如图 3 所示,假设数据中有六个变元(d_1,d_2,d_3,d_4,d_5,d_6),每个维度中均有两个分类值。图 3(a)是按照自然顺序展示,交叉数量较多;(b)对分类值的排列顺序进行重新排列,使得视图中不再有交叉出现;(c)通过对变元轴进行重新排列,大大减少了交叉数量。

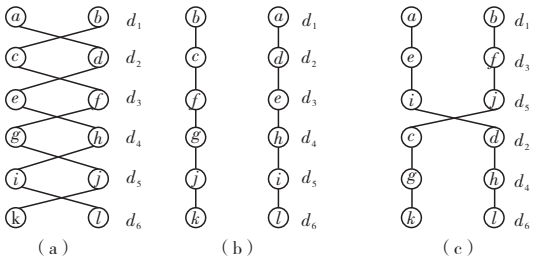

图 3 分类值与变元轴排列示意图

定义 1 具有度量属性的多变元数据集。设 $G(F)=\{F^1,F^2,\cdots,F^m\}$为给定的 $k(k>0)$ 变元对象集合。其中,m 为集合基数,即多变元对象的数量,$F^i=F^i(c_1^i,c_2^i,\cdots,c_{k-1}^i,q^i)$ $(q^i\in\mathrm{R},q^i\geqslant 0)$代表集合中的一个 k 变元对象,$c_1^i,c_2^i,\cdots,c_{k-1}^i$ 表

示每个对象的($k-1$)变元属性,且均为分类型数据,q^i 表示由 $c_1^i, c_2^i, \cdots, c_{k-1}^i$ 确定的数据项的度量值(通常为概率或数量),称 $G(F)$ 为具有度量属性的多变元数据集。对于多变元属性数据表 $G(F)$,F^i 表示一条记录,q^i 称为该条记录的度量。

最常见的具有度量属性的多变元数据集是销售数据,销售数据包括地域、货物类别、时间和销售数量等维度(变元),当固定了地域、货物类别、时间之后,销售数量即成为所确定的数据项的度量。

定义 2　度量变元。对具有度量属性的多变元数据集 $G(F)$,定义其 k 个变元分别表示为 $\{V_1, V_2, \cdots, V_{k-1}, Q\}$,其中 $V_i(1 \leqslant i \leqslant k-1)$ 均为分类型变元,Q 为数值型变元,且变元 Q 任一取值均为非负实数,称 Q 为 $G(F)$ 的度量变元。

定义 3　分类值的度量。考虑具有度量属性的多变元数据集 $G(F)$ 中的变元 $V_l(1 \leqslant l \leqslant k-1)$,设 $\{v_1, v_2, \cdots, v_n\}$ 是 V_l 的 n 个分类值,那么对于任意的分类值 $v_i(1 \leqslant i \leqslant n)$ 在整个数据集中的度量可表示为 $Q(v_i) = \sum_{j=1}^{m} q^j \mid_{V_l = v_i}$,即 $Q(v_i)$ 表示变元 V_l 取分类值 v_i 的所有度量值之和,称 $Q(v_i)$ 为分类值 v_i $(1 \leqslant i \leqslant n)$ 的度量。

定义 4　具有度量属性的多变元时态数据集。对具有度量属性的多变元数据集 $G(F)$,定义其 k 个变元分别表示为 $\{V_1, V_2, \cdots, V_{k-1}, Q\}$,若某一变元 $V_i(1 \leqslant i \leqslant k-1)$ 的取值为若干个时间节点,则整个数据集具有一定的时态特性,称这样的数据集 $G(F)$ 为具有度量属性的多变元时态数据集。

定义 5　具有度量属性的多变元时态数据集。对具有度量属性的多变元数据集 $G(F)$,定义其 k 个变元分别表示为 $\{V_1, V_2, \cdots, V_{k-1}, Q\}$,若某一变元 $V_i(1 \leqslant i \leqslant k-1)$ 的取值为若干个时间节点,则整个数据集具有一定的时态特性,称这样的数据集 $G(F)$ 为具有度量属性的多变元时态数据集。

2.2　Parallel Sets 分类值排列顺序优化算法(ACLEARCR)

分类值排列的顺序是导致 Parallel Sets 显示混乱的重要因素。两条平行变元轴之间边的交叉数量,与对应层节点排列顺序密切相关[19], CLEARCR 算法具有较好的可视化效果,但在分类值顺序排列时未考虑各分类值的度量(见定义 3)的因素,对第一层节点仅仅按照初始顺序赋坐标值;在计算 order_{i+1} 时,如果计算坐标值相等,则按照相邻度进行排序;相邻度相等时,则按照原来的顺序排序。这使得分类值排列顺序在一开始就具有一定的随机性,尤其在相邻度相等时排列顺序的随机性更大。

人的分析认知的习惯往往是首先关注数据中的极端情况,对分类值变元的关注通常集中在度量较大、较小两个极端,因此在对分类值进行排序时参考其度量值可以提供更便于分析的视图。如果较大度量值与较小度量值交叉在一起,则在无形中会导致视图混乱,可能造成对度量较小的分类值的忽略,根据分类值度量顺序对变元进行排序可有效防止出现上述情况。结合定义 3,可以计算出每个变元轴上每个分类值的度量作为分类值排序的参考,基于 CLEARCR 算法,提出 Parallel Sets 分类值排列顺序优化算法 ACLEARCR(advanced CLEARCR[19]),如算法 1 所示。

算法 1　ACLEARCR 算法

a)计算度量值。根据定义 3,计算 Parallel Sets 视图中每一变元的各分类值的度量。

b)节点聚类。

对 PS 视图进行抽象,每个分类值的度量作为节点权值,形成带权层次网络图 G。逐层聚类抽象,聚类方法参考 CLEARCR 算法,形成新的层次网络图 G',在新的视图中,满足一定距离度量的节点被归为一类,同时将归为一类的节点权值相加,作为归类后节点新的权值。

c)按照视图变元轴从上到下的顺序,逐层计算坐标值:

(a)对 G' 初始第一层 L_1 的节点 $\{u_1, u_2, \cdots, u_k\}$,按照各节点权值从大到小对节点进行重新排列,形成顺序 order_1,并为 u_l 分配坐标值 $X_1(u_l) = l$;

(b)根据 L_i 层节点的坐标值计算 L_{i+1} 层全部节点 v 的坐标值 $X_{i+1}(v) = \text{avg}(N_v)$,这里 N_v 表示节点 v 的"上层"相邻节点构成的集合;

(c)以坐标值 $X_{i+1}(v)$ 为依据对 v 进行层内排列,得到 order_{i+1},其中,若两节点坐标值相等,则将相邻度小的节点排列在前面,若相邻度相等时,按照节点权值从大到小进行排列,如果节点权值仍然相等,则这些节点按照原来顺序排列;根据 order_{i+1} 重新为 v_l 分配坐标值 $X_{i+1}(v_l) = l$;依此类推,对所有层的节点顺序进行排列,当 $i = n$ 时,执行 d)。

d)反向迭代。改变迭代方向,即 L_n 变为 L_1,L_{n-1} 变为 L_2,…,L_1 变为 L_n,重复执行 c)。

e)计算终止。

重复步骤 c)d),直至可视化视图中不存在交叉,或者交叉总数不再变化,或者达到初始设定的终止迭代条件(如交叉数量或执行步骤等),则计算终止。

f)还原视图。将由聚类形成的节点逐层替换为原来节点,并恢复原来的相邻关系。

从 ACLEARCR 算法的描述中可以看出,其对 CLEARCR 算法的改进之处在于以下四点:

a)初始阶段,计算每条轴上各分类值的度量,为后续处理提供数据基础。

b)将 Parallel Sets 视图抽象为带权层次网络图。

c)对聚类后形成的层次图的第一层节点按照权值顺序进行排列。

d)在逐层计算阶段,如果两个节点的坐标值和相邻度均相等时,按照权值顺序进行排序。

2.3　基于相关度的 Parallel Sets 变元轴排列算法(VABC)

定义 6　变元相关度。针对具有度量属性的多变元数据集 $G(F)$ 中的变元 $V_1, V_2, \cdots, V_{k-1}$,其度量向量分别为 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{k-1}$,定义任意二变元的相关系数为

$$r(V_i, V_j) = \frac{\alpha}{|n_i - n_j| + 1} + \beta \sum_l w_l$$

其中:$w_l = \begin{cases} 1 & |x_i^l - x_j^l| < \varepsilon \\ 0 & \text{otherwise} \end{cases}$ $(1 \leqslant l \leqslant t)$;$n_i$、$t$ 的定义同定义 4;x_i^l 为度量向量 $\boldsymbol{X}_i$ 的第 l 个分量;α、β 是权值系数,满足 $\alpha \in \mathrm{R}, \alpha > 0, \beta \in \mathrm{R}, \beta > 0$;$\varepsilon$ 是相关度阈值。

相关度的定义有很多种,本文采用两个变元分类值的数量及每个变元各分类值对应度量近似相等的数量来衡量,为此首先通过定义 4 定义了变元的度量向量,以保证整个计算过程中维度一致,然后分别考虑变元分类值的数量、变元的各个分类值的度量(按顺序排列)近似相等的总数量,并按照一定权值求和。这样,两个变元分类值的数量越相近、对应分类值的度量越相近,其相关度越大,极端情况下,两个变元分类值的数量完全相同、一个变元每一分类值的度量与另一变元每一分类值的度量也完全一致,则两个变元的相关度最高。该系数满足相关度的特征,即具有正定性、自反性、对称性。

计算任意二维之间相关度,得到相关度矩阵:

$$R=\begin{bmatrix} r(V_1,V_1) & \cdots & r(V_1,V_{k-1}) \\ \vdots & & \vdots \\ r(V_{k-1},V_1) & \cdots & r(V_{k-1},V_{k-1}) \end{bmatrix}$$

借鉴相似度方法提出基于相关度的 Parallel Sets 变元轴排列算法(variate arrangement based on correlation, VABC),如算法 2 所示。

算法 2 VABC 算法

a) $VR=$"";//VR 为变元轴排列顺序

b)计算相关度矩阵 R;

c)取相关度矩阵中绝对值最大的元素 $r(V_i,V_j)\Rightarrow V_i,V_j(i\neq j)$ 为最相关维度,并将其从矩阵中删除;

d) $EU=V_i$;//变元轴排在最上端

$ED=V_j$;//变元轴排在最下端

$VR=EU+ED$;

e)取 k_1 使得 $r(EU,V_{k_1})$ 最大,取 k_2 使得 $r(ED,V_{k_2})$ 最大,且 $V_{k_1}\notin VR, V_{k_2}\notin VR$;

f) if $r(EU,V_{k_1})>r(ED,V_{k_2})$

$EU=V_{k_1}$;

$VR=EU+VR$;//将 V_{k_1} 排在变元轴最上端

else

$ED=V_{k_2}$;

$VR=VR+ED$;//将 V_{k_2} 排在变元轴最下端

g)重复执行 e)f),直到所有变元都插入到 VR 中,算法结束。

2.4 深度信息 Parallel Sets 算法(DCPS)

定义 7 视图中任意像素点的密度。对于集成时态视图的 Parallel Sets 视图中任意像素点 k,定义其密度值为

$$\rho_k(t_a,t_b)=\sum_j\sum_{i=a}^{b}C_j(k,\Theta(t_i))$$

其中:j 表示活动维的变元序号,j 的值从 1 变化到变元分类值总数;t_a、t_b 表示时间段的开始和结尾;$C_j(k,\Theta(t_i))$ 表示活动维的第 j 个分类值在像素 k 处产生的交叉总数;其余符号的定义参看文献[6]。公式的物理含义是活动维中每个分类值在 k 处产生的交叉数的总和。

定义 8 第 j 个分类值产生的密度。对于集成时态视图的 Parallel Sets 视图中任意像素点 k,单独绘制活动维第 j 个分类值时产生的密度,定义为

$$\rho_{k,j}(t_a,t_b)=\sum_{i=a}^{b}C_j(k,\Theta(t_i))$$

各符号的含义同定义 7。

提出在 Parallel Sets 视图中集成时态视图的深度信息 Parallel Sets(depth cue parallel sets,DCPS)算法,如算法 3 所示。

算法 3 DCPS 算法

a)获取活动维分类值总数 J,令 j 表示分类值的序号,初始情况下 $j=1$。

b)根据定义 7 计算 $\rho_{k,j}$,并将 $\rho_{k,j}$ 划分成 B 个块,从而形成分层,分别记为 $\rho_{k,j}^1,\rho_{k,j}^2,\cdots,\rho_{k,j}^B$,并使得在划分过程中每个块所包含的时间步长的个数尽量平衡,从而将时间段 $[t_a,t_b]$ 划分为 B 个子段,记为 $[t_a^t,t_b^t]$,每一段的密度值 $\rho_{k,j}^t$ 的计算方法如定义 7 所示,$t=1,\cdots,B$,记 ζ_t 等于第 t 个子块中所包含的数据项的总数与其中包含的时间步长个数的乘积。

c)给定 λ 为控制系数,分别计算 B 个块的透明度为 $\alpha_{j,t}=1-e^{-\rho_{k,j}^t/\zeta_t\lambda}$,$t=1,\cdots,B$。

d)根据给定的活动维第 j 个分类值的颜色的色调 h_j,给出 B 个颜色值,并用 HSV 色彩空间表示为

$$\{c_{j,1},c_{j,2},\cdots,c_{j,B}\}=\{(h_j,s_{j,1},v_{j,1}),(h_j,s_{j,2},v_{j,2}),\cdots,(h_j,s_{j,B},v_{j,B})\}$$

其中:$s_{j,1}<s_{j,2}<\cdots<s_{j,B}$,$v_{j,1}<v_{j,2}<\cdots<v_{j,B}$,以使得在时间顺序上,越靠前的颜色饱和度越低、亮度越低,越靠后的颜色饱和度越高、亮度越高。

e)给定初始值 $c_{j,\text{tot}}=0$,$\alpha_{j,\text{tot}}=0$,对 $t=1,\cdots,B$,分别计算合成颜色和合成透明度为

$$c_{j,\text{tot}}\leftarrow c_{j,\text{tot}}+(1-\alpha_{j,\text{tot}})\alpha_{j,t}c_{j,t}$$

$$\alpha_{j,\text{tot}}\leftarrow\alpha_{j,\text{tot}}+(1-\alpha_{j,\text{tot}})\alpha_{j,t}$$

f) $j=j+1$;

if $j<J+1$ then

转 b)

else

转 g)

end if

g)对于视图中任意像素 k,计算其最终的合成颜色和合成透明度为

$$c_{\text{tot}}=\sum_{j=1}^{J}c_{j,\text{tot}}$$

$$\alpha_{\text{tot}}=\sum_{j=1}^{J}\alpha_{j,\text{tot}}$$

进行视图渲染,完成绘制过程。

2.5 TVBPS 可视化方法

将算法 1(ACLEARCR)、算法 2(VABC)、算法 3(DCPS)依次进行整合,形成一种具有度量属性的多变元时态数据可视化方法 TVBPS(time-varying multivariate data visualization based on parallel sets)。

a)列举出各变元轴各个时间内所有可能出现的分类值,使用 ACLEARCR 算法对每条变元轴上的各分类值进行排列。

b)使用基于相关度的 Parallel Sets 变元轴配置算法(VABC),对多变元时态数据的变元轴进行排列。

c)使用 DCPS 算法,对 Parallel Sets 视图进行时态展现,以表现出多变元数据集的时态特点。

d)参照经典的 Parallel Sets 视图,对形成的视图提供合理的人机交互策略。

通过前两个步骤确保了形成的视图具有较少的混乱,所形成的排列方式便于进一步分析,减少了排列中的随机性,降低了进一步手工调整的难度,易于发现更多的隐含知识。特别注意步骤 a)中,由于需要考虑进行时态分析,考虑到随着时间的变化变元的取值可能出现增加或减少的情况,因此必须首先综合分析,罗列出变元在整个时间跨度内所有的可能取值;步骤 c)着重体现数据集的时态特点,层次化的视图能够较好地展现时态性;步骤 d)为形成的视图提供合适的人机交互策略。

3 实验结果

3.1 数据集描述与实验环境说明

本节以院校招生数据集为例对提出的可视化方法进行案例研究,该数据集积累了近年来大量的招生计划数据,包括年度、招生院校、招生专业、学制、学历、性别、文/理科、划线类别、学生生源地等 11 个分类型变元(维度),每一条记录还对应一个具体的度量值变元即招生数量,该变元取值类型为数值型,整个数据集记录约有 40 000 条,包含的时间节点(年度)为 8 个。

本章开展案例研究的硬件环境为 Intel Core2 Duo CPU P8600 2.40 GHz,2.93 GB 内存物理地址,ATI Mobility Radeon

HD 3400 Series 显卡,320 GB 硬盘;软件环境为 Microsoft Windows XP Professional SP3,Eclipse,MyEclipse 6.5,JDK1.6,Oracle 9i;相关的开源软件包为 Parallel Sets[27]。

3.2　ACLEARCR 与 VABC 算法结合的效果分析

图 4 能够保证度量值较大的分类值之间的平行四边形尽量少地与其他平行四边形交叉,有利于迅速定位出数据集中的关键元素,同时借助视图提供的交互式方法,也可以进一步查看度量值较小的分类值,因而效果更好。

图 5 中,变元轴的随机排列,导致数据集中一些较为明显的模式被分隔开形成不同的平行四边形,这些平行四边形与度量值较小的分类值的平行四边形相互交叠,可读性较差。图 6 完全采用随机方式排列分类值和变元轴,视图中混乱程度最为严重。

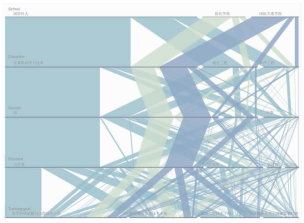

图 4　ACLAERCR + VABC 算法效果图

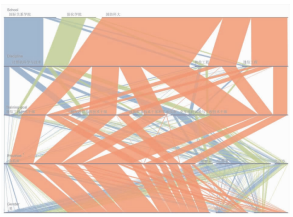

图 5　分类值 CLEARCR 算法排列、变元轴随机排列效果图

交叉数量是度量视图混乱程度的一种有效方式。按照不同的记录条数规模、变元个数、分类值数量抽取本文数据集,分别记录形成的可视化视图中的交叉数量,得到的结果如表 1 所示。

表 1　ACLEARCR + VABC 算法效果对比分析表

记录条数	变元个数	分类值数量	分类值变元轴均任意排列产生的交叉数	分类值采用 CLEARCR 算法,变元轴任意排列产生的交叉数	ACLEARCR + VABC 算法产生的交叉数
470	5	21	148	49	20
700	5	21	154	54	32
1000	5	21	161	58	53
1000	3	21	143	47	21
1000	4	21	151	52	40
1000	5	21	161	58	42
1000	5	10	11	3	3
1000	5	15	38	12	12

从表 1 中可以看出,ACLEARCR + VABC 算法在大部分情况下都能有效地减少视图中出现的交叉,尤其是在初始视图变元轴的排列顺序不合理的情况下,通过 VABC 对变元轴进行排列之后交叉数明显减少。在某些变元轴的排列顺序已经达到合理的状态情况下,ACLEARCR + VABC 算法与分类值采用 CLEARCR 算法排列、变元轴任意排列的交叉数相当或相同,这说明了 ACLEARCR + VABC 算法能够更好地克服变元轴任意排列、分类值采用 CLEARCR 算法排列方法中存在的随机性。

从上述可视化效果和交叉数量分析中可以看出,ACLEARCR 与 VABC 算法共同为基于 Parallel Sets 进行具有度量属性的多变元时态数据的分析提供了较好的视图基础,用户可以在现有视图基础上经过较少的调整获得混乱程度较小、既直观又美观的可视化效果图。

3.3　DCPS 算法的案例验证

由于 DCPS 算法视图渲染过程中,第一变元轴上分类值的数量直接决定了整个视图中所使用的颜色色调的数量,需要展示的时间段(时间节点)的数量直接决定了视图中出现的层次效果的数量,因此在进行 DCPS 有效性时需要对第一变元轴上分类值的个数和时间节点的个数作一限定。本节以第一变元轴上有两个分类值、三个时间节点为例进行说明。

从图 7 中可以较为明显地查看出数据集中的时态信息,不同时间节点的数据集的情况展示非常明显,越旧的时间点使用饱和度、亮度越低的颜色层表示,越新的时间点使用饱和度、亮度越高的颜色层表示,这与大家的常规认知(即越旧的东西往往色彩越暗淡)相吻合。符合可视化技术将抽象数据进行直观视图展现的原则,便于用户接受,DCPS 算法是对时态相关数据进行可视化的一种有效方法。

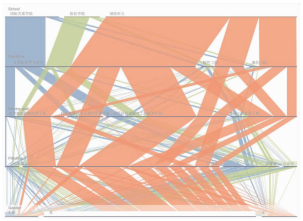

图 6　分类值与变元轴均随机排列效果图

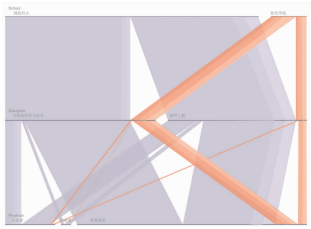

图 7　DCPS 方法可视化效果图

3.4　TVBPS 可视化方法的案例验证

对 2010—2012 年的招生院校、招生专业、学生生源地三个维度进行可视化展现和分析,重点研究军队指挥类、指挥信息系统与工程、飞行器系统工程三个专业(discipline);具有上述三个专业、且招生数量较多的院校(school)选取国防科技大学、信息工程大学、理工大学、海军工程大学、空军工程大学、第二炮兵工程大学;在学生生源地(province)方面为综合考虑全国的情况,分别选取我国东北的吉林省、华东的山东省、中部的湖北省、西北的甘肃省和西南的云南省,上述省份教育教学质量在该地区范围内相对较好,历年的招生计划数也较多。从而形成具有度量属性的多变元时态数据集:

$$G(F) = \{\text{year}, \text{school}, \text{discipline}, \text{province}, \text{number}\}$$

各个变元的取值如上所述。根据本文提出的可视化方法 TVBPS,获得视图如图 8 所示。第一变元 school 决定了整个视图中颜色的数目,每个学校的招生数被各个 province 划分为不同的子块,各 province 进一步将招生数按照专业划分,同一色调不同的颜色饱和度和亮度表现了不同的时间点,饱和度越高、亮度越高表示的时间点越新。如果对应的 school、province 在 2010—2012 年之间招生数有变化,则相应的平行四边形被划分为变化次数相同的不同层次,即如果三年内变化三次则平行四边形有三层,如果三年内只有两年之间有变化则平行四边形有两层,如果不变化则平行四边形只有一层。从图 8 中至少可以得出以下结论:

a)国防科大在山东省、湖北省的招生数量逐年增加,其他各省的招生数量基本稳定不变,国防科大在山东省招生数量中,军队指挥类专业在逐年减少,指挥信息系统与工程专业招生数量在增加。

b)信息工程大学招生总数也呈上升趋势,但该学校在军队指挥类专业的招生总数基本保持稳定,指挥信息系统与工程专业、飞行器系统工程专业的招生数量在逐年增加。

c)理工大学在湖北省的招生数量呈现增加趋势,在军队指挥类专业的视图仅包括两层,可以看出理工大学从 2011 年

才开始招收军队指挥类专业学生,但 2012 年招生数量即呈现出减小的趋势,与之相反,在指挥信息系统与工程专业的招生数量呈现出增加趋势。

d)海军工程大学在湖北省、吉林省、甘肃省、云南省的招生数量维持平稳,在山东省的招生总数在减少,但在山东省招收的指挥信息系统与工程专业的数量却呈现出增加趋势。

e)近三年来,空军工程大学、二炮工程大学在上述省份和专业方面的招生情况基本维持平稳。

f)上述六所院校在上述省份、专业的招生总数差别不大,山东省的招生总数最多,军队指挥类专业的招生数量比其他两个专业的招生数量都大。

以上完成了可视化效果实验验证,证明了可视化方法的有效性。

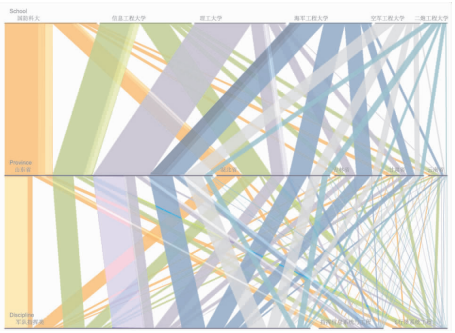

图 8 TVBPS 可视化方法的案例验证效果图

4 TVBPS 可视化方法分析

通过对可视化效果图进行分析,总结 TVBPS 可视化方法具有如下特点:

a)支持的分析任务方面,重点展示数据集中各个变元之间的相互关系,能够较为明显地展示出任意变元之间、各分类值之间度量值的相互关系;但层次分析效果不够明显。

b)支持的数据特点方面,使用层次方式展示数据集的时态特性,适用于时间段更多的情况;但未将不同变元或不同分类值的度量值在同一坐标范围内进行排列,因此只能进行较为粗略的对比展示。

c)视图特点方面,视图绘制过程完全由算法实现,不需要对数据集本身有更多的理解,反映数据特点更为客观;但视图调整的灵活度不够,关注对象不能调整到视图中心,展示效果不太明显,对度量值较小的分类值的时态特点展现效果不明显。

5 结束语

通过对典型的分类型数据可视化方法 Parallel Sets 进行改进,提出 Parallel Sets 分类值排列顺序优化算法(ACLEARCR)、基于相关度的 Parallel Sets 变元轴配置算法(VABC)、深度信息 Parallel Sets(DCPS),共同组成基于 Parallel Sets 的具有度量属性的多变元时态数据可视化方法 TVBPS,案例研究表明该方法能够有效满足本文研究目标,有利于挖掘数据集中的隐含模式。如何对变元轴上大量的度量值较小的分类值进行排列以减少视图中的混乱,有效发现数据集中的隐含模式,是值得进一步研究的课题。

参考文献:

[1] FOUT N, MA K L, AHRENS J. Time-varying multivariate volume data reduction[C]//Proc of ACM Symposium on Applied Computing. 2005:1224-1230.

[2] GUO Dian-sheng, CHEN Jin, MACEACHRENn A M, *et al*. A visualization system for space-time and multivariate patterns (VIS-STAMP)[J]. IEEE Trans on Visualization and Computer Graphics, 2006, 12(6):1461-1474.

[3] GUO Dian-sheng, LIAO Ke, MORGAN M. Visualizing patterns in a global terrorism incident database[J]. Environment and Planning B: Planning and Design, 2007, 34(5):767-784.

[4] AKIBA H, MA K L. An interactive interface for visualizing time-varying multivariate volume data[C]//Proc of APVIS. 2007.

[5] AKIBA H, MA K L. A tri-space visualization interface for analyzing time-varying multivariate volume data[C]//Proc of IEEE-VGTC Symposium on Visualization. 2007:115-122.

[6] JOHANSSON J, LJUNG P, COOPER M. Depth cues and density in temporal parallel coordinates[C]//Proc of Eurographics/IEEE VGTC Symposium on Visualization. 2007:35-42.

[7] JOHANSSON J. Efficient Information visualization of multivariate and time-varing data[D]. Linköping Studies in Science and Technology Dissertations. NorrRöping: Linkoping University, 2008.

[8] LEE T Y, SHEN Han-wei. Visualization and exploration of temporal trend relationships in multivariate time-varying data[J]. IEEE Trans on Visualization and Computer Graphics, 2009, 15(6):1359-1366.

[9] CHEN Cheng-kai, WANG C, MA K L, *et al*. Static correlation visualization for large time-varying volume data[C]//Proc of IEEE Pacific Visualization Symposium. 2011.

[10] FERMSTAD S J, JOHANSSON J. A task based performance evaluation of visualization approaches for categorical data analysis[C]//Proc of the 15th International Conference on Information Visualisation. 2011.

[11] 贾澎涛,何华灿,刘丽,等. 时间序列数据挖掘综述[J]. 计算机应用研究,2007,24(11):15-29.

[12] 郝智勇,贺明科,谭文堂,等. 基于多维标度法的专利文本可视化聚类研究[J]. 计算机应用研究,2010,27(12):4608-4611.

[13] BENDIX F, KOSARA R, HAUSER H. Parallel Sets: visual analysis of categorical data[C]//Proc of IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society, 2005:133-140.

[14] KOSARA R, BENDIX F, HAUSER H. Parallel Sets: interactive exploration and visual analysis of categorical data[J]. IEEE Trans on Visualization and Computer Graphics, 2006, 12(4):558-568.

[15] 孙扬,封孝生,唐九阳,等. 多维可视化技术综述[J]. 计算机科学,2008,35(11):1-7.

[16] CAAT M, MAURITS N M, ROERDINK J B. Design and evaluation of tiled parallel coordinate visualization of multichannel EEG data[J]. IEEE Trans on Visualization and Computer Graphics, 2007, 13(1):70-79.

[17] BLAAS J, BOTHA C, POST F. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets[J]. IEEE Trans on Visualization and Computer Graphics, 2008, 14(6):1436-1451.

[18] 孙扬,周城,汤大权,等. 带分类值排列算法的改进的 Parallel Sets[J]. 武汉大学学报:信息科学版,2011,36(2):144-147.

[19] 肖卫东,周城,孙扬,等. Parallel Sets 的改进及其在全球恐怖袭击数据分析中的应用[J]. 国防科技大学学报,2011,33(1):115-119.

(下转第 1600 页)

3 结束语

本文基于 SPM 框架提出了一种基于 Gist 特征信息检测的码本区分性增强方法。通过对图像 Gist 特征信息量的检测，对目标和背景进行划分，使用非均匀密集采样有效提取目标特征，过滤与类别无关特征，从而使得 K-means 聚类生成的码本更加具有类别区分性，提高了分类效果。实验结果表明，本文方法进行样本筛选得到的码本在类别区分性上优于没有经过筛选的码本，分类效果优于 LSC、ScSPM、KSPM 等方法。本文方法和 LLC 分类效果基本一致，但是本文方法实现更加简单、快速、易于理解，码本生成方法与后续编码方法完全独立，因此生成的码本更具有通用性。在将来的工作中，笔者将对信息检测的方法进行改进，尝试用其他更加简洁、更准确的方法来检测信息，降低检测所需要的时间，同时更加精确地获得与类别相关的特征，这样将进一步提高分类效率和准确率。

参考文献：

[1] LI Fei-fei, FERGUS R, PERONA P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2004: 178-186.

[2] NOWAK E, JURIE F, TRIGGS B. Sampling strategies for bag-of-features image classification[C]//Proc of European Conference on Computer Vision. 2006: 490-503.

[3] FERGUS R, PERONA P, ZISSERMAN A. Object class recognition by unsupervised scale-invariant learning[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2003: 264-271.

[4] BERG A, BERG T, MALIK J. Shape match and object recognition using low distortion correspondences[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2005: 509-522.

[5] LAZEBNIK S, SCHMID C, PONCE A J. A maximum entropy framework for part-based texture and object recognition[C]//Proc of International Conference on Computer Vision. 2005: 832-838.

[6] GRAUMAN K, DARRELL T. Pyramid match kernels: discriminative classification with sets of image features[C]//Proc of International Conference on Computer Vision. 2005: 1458-1465.

[7] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]// Proc of International Conference on Computer Vision and Pattern Recognition. 2006: 2169-2178.

[8] YANG Jian-chao, YU Kai, GONG Yi-hong, *et al*. Linear spatial pyramid matching using sparse coding for image classification[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2009: 1794-1801.

[9] WANG Jin-jun, YANG Jian-chao, YU Kai, *et al*. Locality-constrainted linear coding for image classification[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2010: 3360-3367.

[10] LIU Ling-qiao, WANG Lei, LIU Xin-wang. In defense of soft- assignment coding[C]//Proc of International Conference on Computer Vision. 2011: 2486-2493.

[11] RUSSAKOVSKY O, LIN Yuan-qing, YU Kai, *et al*. Object-centric spatial pooling for image classification[C]//Proc of European Conference on Computer Vision. 2012: 1-15.

[12] MAIRAL J, BACH F, PONCE J, *et al*. Supervised dictionary learning [C]//Advanceds in Neural Information Processing System. 2008: 1033-1040.

[13] OLIVA A, TORRALBA A. Modeling the shape of the scene: a holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[14] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[15] LI Fei-fei, ERONA P P. A Bayesian hierarchical model for learning natural scene categories[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2005: 524-531.

[16] FAN R E, CHANG K W, HSIEH C J, *et al*. LIBLINEAR: a library for large linear classification[J]. Journal of Machine Learning Research, 2008, 9(1): 1871-1874.

[17] JURIE F, TRIGGS B. Creating efficient codebooks for visual recognition[C]//Proc of European Conference on Computer Vision. 2005: 604-610.

[18] YANG Jian-chao, YU Kai, GONG Yi-hang, *et al*. Linear spatial pyramid matching using sparse coding for image classification[C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2009: 1794-1801.

[19] HOLUB G G, PERONA P A D. Caltech-256 object category dataset, TR7694[R]. [S. l.]: California Institute of Technology, 2007.

(上接第 1596 页)

[20] ANKERST M, BERCHTOLD S, KEIM D A. Similarity clustering of dimensions for an enhanced visualization of multidimensional data [C]//Proc of IEEE Symposium on Information Visualization. Washington: IEEE Computer Society, 1998: 52-60.

[21] ARTERO A O, de OLIVEIRA M C F, LEVKOWITZ H. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement[C]//Proc of the 10th International Conference on Information Visualization. 2006: 5-7.

[22] 孙扬，唐九阳，汤大权，等. 改进的多变元数据可视化方法[J]. 软件学报，2010，21(6)：1462-1472.

[23] SHI Cong-lei, CUI Wei-wei, LIU Shi-xia, *et al*. RankExplorer: visualization of ranking changes in large time series data[J]. IEEE Trans on Visualization and Computer Graphics, 2012, 18(12): 2669-2678.

[24] LOUDCHER S, BOUSSAID O. OLAP on complex data visualization operator based on correspondence analysis[C]//Lecture Notes in Business Information Processing. Berlin: Springer, 2012: 172-185.

[25] LIVINGSTON M A, DECKE J W, AI Zhu-ming. Evaluating multivariate visualizations on time-varying data[C]//Proc of SPIE 8654 on Visualization and Data Analysis. 2013.

[26] KEHRER J, HAUSER H. Visualization and visual analysis of multifaceted scientific data: a survey[J]. IEEE Trans on Visualization and Computer Graphics, 2013, 19(3): 495-513.

[27] http://eagereyes.org/parallel-sets[EB/OL].