ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.6, June 2010, pp.1462−1472
doi: 10.3724/SP.J.1001.2010.03460

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

改进的多变元数据可视化方法*

孙 扬+, 唐九阳, 汤大权, 肖卫东

(国防科学技术大学 信息系统与管理学院,湖南 长沙 410073)

Improved Multivariate Data Visualization Method

SUN Yang+, TANG Jiu-Yang, TANG Da-Quan, XIAO Wei-Dong

(School of Information System and Management, National Univisity of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: victor_830514@yahoo.com.cn, http://www.nudt.edu.cn

Sun Y, Tang JY, Tang DQ, Xiao WD. Improved multivariate data visualization method. *Journal of Software*, 2010,21(6):1462−1472. http://www.jos.org.cn/1000-9825/3460.htm

Abstract: Star coordinates (SC), a traditional multivariate data visualization technique, loses much information due to oversimple dimension reduction algorithm. And the SC visualization can't offer the dimension distribution information. Moreover, the manual dimension axis configuration of SC is too complex. To address these problems, the paper proposes the advanced star coordinates (ASC), which uses the diameter instead of the radius as the dimension axis, designs the dimension configuration strategy to optimize the order and the angle of dimension axes, and projects the multi-dimensional object to low dimension visual space through the dimension reduction algorithm. And the dimension reduction process is meaningful to user and the algorithm uses the minimum of the object coordinates variation between the multi-dimensional coordinates and the advanced star coordinates as criterion. Experimental results show that the dimension reduction algorithm is highly efficient and suitable for the aggregation with a great amount of high-dimensional data. The dimension configuration strategy relieves the user's operation burden greatly and helps them detect the connotative characteristics of the multidimensional information aggregation quickly and exactly. The visualization is easy to understand and can express the dimension distribution information effectively, which is helpful for user to view the multi-dimensional information and to discover the implicit information in knowledge discovery process.

Key words: multivariate data visualization; multidimensional data visualization; dimension reduction; dimension configuration strategy; information visualization

摘 要: 针对传统多变元可视化方法——星形坐标法(star coordinates,简称 SC)降维过程信息损失较为严重、可视化结果无法体现维度分布信息及手动配置维度轴十分繁杂的不足,提出一种改进的星形坐标法(advanced star coordinates,简称 ASC),使用沿直径方向的向量作为维度轴,设计维度轴配置策略优化各维度轴之间的夹角及排列顺序,以减小多维信息对象在改进星形坐标系中与在多维坐标系中坐标差别为准则,使用最优化方法实现对用户有意义的降维运算,将多维信息映射到低维可视空间中.实验结果表明,ASC 的可视化结果不仅易于理解,而且能够有效提供维度分布信息,有利于用户发掘隐性知识,基于相关度的维配置策略可以大大减轻用户操作负担,使其能够快速

* Supported by the National Natural Science Foundation of China under Grant No.60172012 (国家自然科学基金)
Received 2008-04-14; Accepted 2008-08-11

而准确地定位多维信息集合中的隐含特征,降维算法高效,适用于数据量较大、维数较高的信息集合.
关键词: 多变元数据可视化;多维数据可视化;降维;维配置策略;信息可视化
中图法分类号: TP391 文献标识码: A

随着信息技术的飞速发展,科学、工程、商业等领域的多维甚至高维信息日益增多,如果依赖数据表格或文字的形式表示多维信息,由于数据量巨大以及人类认知能力固有局限性的存在,用户难以对其进行理解及比较,信息认知活动面临空前的复杂性[1].多变元数据可视化技术通过图形化形式展示多维信息的多属性数据特征,使用户可以有效地在低维可视空间中观察、操纵、研究、浏览、探索及理解抽象的多维信息及其结构,辅助用户在知识发现、信息认知和信息决策过程中快速准确地发掘数据集中隐含的特征、关系、模式、趋势及聚类信息等,因此,该技术被广泛应用于海量多维数据集的探索、理解及分析领域[2].

研究人员已经提出很多有效的多变元数据可视化方法[3−5],如平行坐标法[6,7]、放射坐标法(RadViz)[8]、降维映射技术[9]和星形坐标法(star coordinates,简称 SC)[10,11]等.平行坐标法将 k 维数据属性空间映射为平面上 k 条等距离的竖直平行轴,用一条连接 k 条平行轴上属性值点的折线段表示一个数据项;圆形平行坐标法[12]使用圆形的 k 条半径表示 k 维空间,由于坐标轴内外的几何不对称性,它能够更好地揭示多维信息之间某些特殊关系.(圆形)平行坐标法表达数据关系非常直观,易于理解;但对大数据集进行可视化时,由于折线密度增加产生大量交叠线,难于辨识,而且平行轴的排列次序也是影响发现数据间关系的重要因素.Radviz 使用弹簧模型确定多维对象在坐标系中的位置,其优点是计算复杂度较低 $O(mk)$,相似多维对象的投影点十分接近,容易发现聚类信息;缺点是海量信息投影点的交叠问题及差异较大的多维对象的投影点也可能会很接近,以至于使用户产生误解.降维映射技术是要找到从高维空间到低维可视空间中的一个映射,该映射应尽可能地保持数据间的某种关系不变,虽然它能够较好地展现多维信息集合的整体结构和分布,解决了维数灾难问题[13],但其计算复杂度较高为 $O(m^2)$.更为重要的是,该类方法只是尽可能地保持数据集在低维空间与在高维空间部分统计信息的一致性,而产生的低维可视空间对用户没有直观意义[14],用户无法直接将其与高维空间关联起来.星形坐标法使用从圆心辐射到圆周上的射线(维度轴)表示维,多维对象的各维值对应于相应维度轴上一向量的长度,多维对象由代表各维值向量和的点表示.文献[15,16]分别用不同方法将星形坐标法拓展到三维空间,并使用自组织网络映射法(self-organizing map,简称 SOM)自动配置各维度轴之间的夹角使聚类效果达到最优[16].

星形坐标法及其改进方法能够较好地体现多维信息集合的聚类、趋势及奇异点信息,并且算法简单,适用于海量高维数据集,但其存在以下缺点:通过各维值向量相加将多维信息映射到低维空间的算法过于简单,信息损失较为严重;可视化结果没有体现多维数据集合在不同维度的分布信息;针对可视化效果对各维度轴排列次序依赖性较强的特点,SC 方法只通过人机交互技术对维度轴进行配置,而且该手动配置过程繁琐、复杂、耗时.针对上述不足,本文提出了改进星形坐标法(advanced star coordinates,简称 ASC),以直径代替半径作为维度轴,清晰展现数据集合的维度分布信息;引入维度轴配置策略优化各维度轴之间的夹角及排列次序,为进一步手动配置维度轴提供较好的初始状态,从而简化维度轴交互配置过程的繁琐性,缩短挖掘隐性知识的时间;以保持多维信息对象在改进星形坐标系中与多维坐标系中的坐标一致性为指标,将多维信息通过对用户有意义的降维运算映射到低维可视空间.通过对海量高维数据集合的整体结构、在各维度上的分布情况及聚类信息的展现,ASC 方法有利于用户全面掌握和深入理解多维数据集合的结构,对基于多维数据集合进行决策和知识发现起到了较好的辅助作用.算法分析和仿真实验结果表明,ASC 的算法效率较高,适用于数据量较大、维度较高的数据集合,可视化结果易于理解;与 SC 方法相比,提高了手动配置维度轴的效率,缩短了发掘有效结论的时间,减少了简单降维过程引入的信息损失,展现了多维数据集合丰富的维度分布信息.

1 改进星形坐标法

1.1 ASC模型

定义 1. 将多属性数据集中相互完全独立属性称作维度(dimension),相关属性称为变元(variate).数据集的记录可以看作定义在 n 维欧式空间 D 中 k 元函数 $F(X)$ 的样本,其中,$F=(f_1,f_2,\ldots,f_k)$ 由 k 维变元组成,$f_i=f_i(x_1,x_2,\ldots,x_n)$,$X=(x_1,x_2,\ldots,x_n)$ 为 D 中一点,$k=0$ 时,$F(X)$ 表示 D 中一点;$n=0$ 时,$F(X)$ 为常数.

维度和变元的概念由来已久[17],通常情况下,多维数据一般针对科学可视化而言,多变元数据更多地与信息可视化相关.其实,也可将多变元数据 $F=(f_1,f_2,\ldots,f_k)$ 看作 k 维空间中(各维之间不存在正交性)的一点,这样就可以将多变元数据和多维数据统一起来.因此,虽然本文涉及的方法主要使用多变元数据集合,但是在此将二者统称为多维数据集合.

定义 2. 设 $G(F)=\{F^1,F^2,\ldots,F^m\}$ 为给定的 $k(k>0)$ 维对象集合,其中,m 为集合基数,即多维对象的数量,$F^i=F^i(f_1^i,f_2^i,\ldots,f_k^i)(f_j^i\in R,f_j^i\geq 0)$ 代表集合中的一个 k 维对象,$f_1^i,f_2^i,\ldots,f_k^i$ 表示每个对象的 k 维属性值.对于多属性数据表 $G(F)$,F^i 表示一条记录,f_j^i 表示其字段值.

定义 3. 在平面直角坐标系中,以 O 为圆心绘制圆形,沿圆形的 k 条直径作射线分别交圆周于 $D_{s1},D_{e1},D_{s2},D_{e2},\ldots,D_{sk},D_{ek}$,称向量 $\overrightarrow{D_{sj}D_{ej}}$ 为改进星形坐标系的维度轴(dimension axis).D_{sj} 是维起始点,D_{ej} 是维终止点.$\overrightarrow{D_{sj}D_{ej}}$ 的方向是维正向(维值增大的方向),相对地,$\overrightarrow{D_{ej}D_{sj}}$ 的方向是维负向.

定义 4. ASC 中任意多维对象由平面直角坐标系中一点 $F^i(x_i,y_i)$ 表示,将该点向改进星形坐标系中的各维度轴投影,称投影点在相应维度轴上的坐标为该多维对象 F^i 的可视坐标(visual coordinates) d_j^i,多维对象的属性值 f_j^i 为在原多维坐标系中的维坐标.

因此,通过在平面直角坐标系中引入改进星形坐标系,可以将代表多维信息对象 $F^i(f_1^i,f_2^i,\ldots,f_k^i)$ 的点 $F^i(x_i,y_i)$ 以其可视坐标表示为 $F^i(d_1^i,d_2^i,\ldots,d_k^i)$,这样就得到了 ASC 方法的多维可视化模型,如图 1 所示.显然,多维信息对象 $F^i(f_1^i,f_2^i,\ldots,f_k^i)$ 只有在多维坐标系中才可以使用一点来准确表示,现在通过降维运算 σ 将其映射到改进星形坐标系(二维空间)中进行表示,即 $F^i(f_1^i,f_2^i,\ldots,f_k^i)\xrightarrow{\sigma}F^i(d_1^i,d_2^i,\ldots,d_k^i)=F^i(x_i,y_i)$,这样必然会带来信息损失.所以,本文给出的算法使用最优化方法实现降维,使 $F^i(x_i,y_i)$ 在改进星形坐标系中可以最大反映多维信息对象 $F^i(f_1^i,f_2^i,\ldots,f_k^i)$ 的各维值,并尽可能地保持各对象之间的空间位置关系.

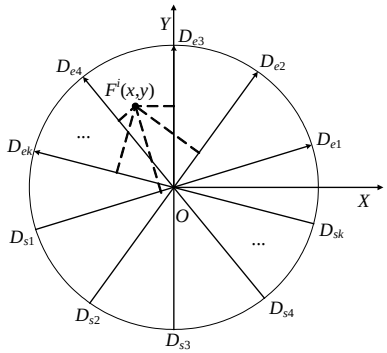


Fig.1 k-Dimensional visualization model of ASC

图 1 ASC 方法的 k 维可视化模型

1.2 算法描述

(1) 在平面直角坐标系中,建立改进星形坐标系,指定各维度轴 $\overrightarrow{D_{sj}D_{ej}}$ 所对应的维,并定义其维正向(各维度轴夹角及长度的初始值相等);

(2) 计算各维度轴的单位向量 $\overrightarrow{u_j}=\dfrac{\overrightarrow{D_{sj}D_{ej}}}{\max_j-\min_j}$,其中,$\min_j=\min\{f_j^i,1\le i\le m\}$,$\max_j=\max\{f_j^i,1\le i\le m\}$;

(3) 求过各维起始点 D_{sj} 并与相应维度轴 $\overrightarrow{D_{sj}D_{ej}}$ 垂直的直线方程 $E_j(x,y)=A_jx+B_jy+C_j=0(1\le j\le k)$;

(4) 建立目标函数:

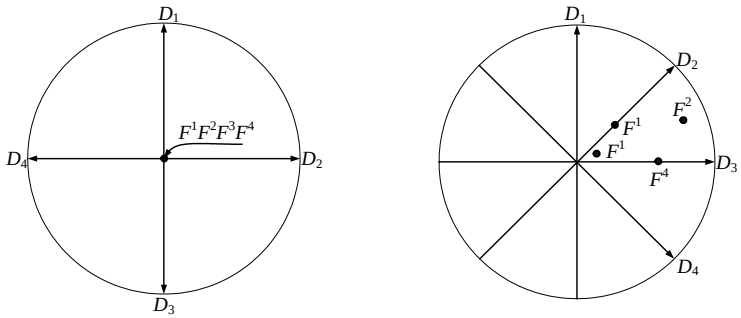
$$\min f(x,y)=\sum_i\sum_j(d_j^i-f_j^i)^2 \tag{1}$$

其中,$x=(x_1,x_2,\ldots,x_m)^T$,$y=(y_1,y_2,\ldots,y_m)^T$,$d_j^i=\dfrac{|A_jx_i+B_jy_i+C_j|}{\sqrt{A_j^2+B_j^2}}\times|\overrightarrow{u_j}|+\min_j$;

(5) 在各维度轴 $\overrightarrow{D_{sj}D_{ej}}$ 上取点 $T_j^i(x_j^i,y_j^i)$,使 $\overrightarrow{D_{sj}T_j^i}=(f_j^i-\min_j)\times\overrightarrow{u_j}$,以点 $\left(\sum_j x_j^i/k,\sum_j y_j^i/k\right)$ 为初始点开始进行迭代,使用步长加速法求解目标函数(1).

在改进星形坐标系中绘制点集 $G(F(x,y))$,并对不同类别的多维对象进行标注.

ASC 使用直径代替半径作为维度轴,并借助最优化算法对多维对象进行降维,解决了多维空间中差异很大的对象在星形坐标系中投影点可能会很接近的问题.如图 2 所示,对于 4 条完全不同的数据(F^1(1,1,1,1),F^2(9,9,9,9),F^3(3,0,3,0),F^4(0,5,0,5)),SC 中的投影交叠于一点;而在 ASC 中,其投影点分散于 4 个位置,体现了多维对象在不同维度的分布情况.



(a) Star coordinates (b) Advanced star coordinates

(a) SC 方法 (b) ASC 方法

Fig.2 Comparison between ASC and SC

图 2 ASC 方法与 SC 方法对比图

1.3 降维算法的优化

传统降维映射技术都是利用线性或非线性方法最小化某优化准则,以尽可能使多维数据间的某种特性在降维前后保持不变.ASC 可以看作是一种降维映射技术,它与传统方法(如多维标度法(MDS))的区别在于,后者产生的低维可视空间对用户是无意义的,而 ASC 通过引入改进星形坐标系,并在此基础上产生算法的降维准则,使多维信息在改进星形坐标系中的可视坐标与在原多维坐标系中的维坐标尽量一致,从而尽可能地保持了各多维对象之间的空间位置关系;高维空间中各维之间的部分语义关系也得以在低维可视空间中继续保持,用户可以通过将改进星形坐标系与高维抽象空间直接关联起来,获取信息集合的整体结构特征及其在各维的分布情况.因此,ASC 产生的低维可视空间对用户是有意义的;而且该算法属于增量式算法,即每一个多维对象在降维过程中都是独立的,只要每一个多维对象达到最优,整个多维信息集合即达到最优取值,当有新的多维对象加入时,只需对其单独进行降维(最优化过程)即可,无须重新计算其余已优化数据.

在 ASC 算法第(4)步中,由于 $d_j^i=\dfrac{|A_jx_i+B_jy_i+C_j|}{\sqrt{A_j^2+B_j^2}}\times|\overrightarrow{u_j}|+\min_j$,使 $f(x,y)$不可微,所以目标函数的求解过程只能采用收敛速度较慢的步长加速法,算法效率较低.

本文在平面直角坐标系中引入极坐标,以点 $F^i(r_i,\varphi_i)$代替 $F^i(x_i,y_i)$表示多维对象 $F^i(f_1^i, f_2^i,..., f_k^i)$.如图 3 所示,目标函数转换为

$$\min f(r,\varphi)=\sum_i\sum_j (d_j^i - f_j^i)^2,$$

其中, $d_j^i=\left(r_i\cos(\varphi_i-\psi_j)+\frac{\max_j-\min_j}{2}\right)\times|\overrightarrow{u_j}|+\min_j, r=(r_1,r_2,\ldots,r_m)^T, \varphi=(\varphi_1,\varphi_2,\ldots,\varphi_m)^T, \psi_j$ 为各维度轴的维正向到极轴的顺时针夹角.这里的 $f(r,\varphi)$可微,所以能够采用收敛速度较快的拟牛顿法或变尺度法进行计算.

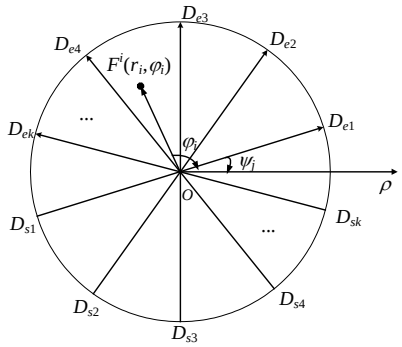


Fig.3 Improved ASC algorithm based on polar coordinates

图 3 基于极坐标的 ASC 优化算法示意图

2 基于相关度的维度轴配置策略

ASC 方法与 SC 方法都存在可视化效果对维度轴排列次序依赖性较强的缺陷,维度轴的随机排列极易引起可视混乱和对象交叠,使用户很难发掘多维信息集合潜在的组织结构特征及各多维对象之间预期的隐含关系,不利于用户从中获取隐性知识.SC 方法通过人机交互操作对维度轴进行配置,寻找最优的维度轴排列顺序.但是,随着维数不断升高,用户手动操作的繁琐性越来越高,可行性越来越差.因此,为了减轻用户的负担,改进可视化效果,提高隐性知识的发现效率,我们引入维度轴配置策略对 ASC 方法维度轴初始状态进行优化.维度轴配置策略的基本思想是,根据维相关度配置维度轴,将高相关维度轴排列在一起并使其具有尽可能小的维正向夹角,以避免因将它们配置在相反方向引起效用相互抵消而产生可视混乱和对象交叠,从而使多维信息集合内在的组织结构特征及各多维信息对象之间的隐含关系更加清晰.因此,本文借鉴 Ankerst 等人[18]基于相似度的维排列算法提出了适用于 ASC 方法的基于相关度的维度轴配置策略.

2.1 相关度的定义

定义 5. 针对多维信息集合 $G(F)=\{F^1,F^2,\ldots,F^m\}$中的维度 $D_1,D_2,\ldots,D_k$,定义任意二维的相关系数为

$$r(D_i,D_j)=\sum_l w_l,$$

其中, $w_l=\begin{cases}1, & \left|\frac{f_i^l-\mathrm{Min}_i}{\mathrm{Max}_i-\mathrm{Min}_i}-\frac{f_j^l-\mathrm{Min}_j}{\mathrm{Max}_j-\mathrm{Min}_j}\right|<\varepsilon \\ 0, & \text{otherwise}\end{cases}$ $(1\le l\le m, 1\le i\le k, 1\le j\le k)$, $\mathrm{Max}_i,\mathrm{Min}_i$ 分别为维度 D_i 的最大、最小值, ε 为用户设定的相关度阈值.

定义 6. 领域专家通过分析各维间的语义相关性,使用 AHP 方法构造各维的语义相关系数矩阵:

$$S=\begin{bmatrix} S(D_1,D_1) & \ldots & S(D_1,D_k) \\ \ldots & \ldots & \ldots \\ S(D_k,D_1) & \ldots & S(D_k,D_k)\end{bmatrix},$$

其中,$-9\le S(D_i,D_j)\le 9$ 且 $S(D_i,D_j)$为自然数,则任意二维的相关度 $R(D_i,D_j)=S(D_i,D_j)r(D_i,D_j)$.

相关度有很多不同的定义方法,我们以多维数据集中所有数据任意两维对应数值近似相等的数量作为二者的相关系数.由定义 5 可知,该系数满足相关度的特征——正定性、自反性及对称性,而且该系数对于维度的缩放及平移具有恒定性,即对诸如 $D(0,1,1,0,0)$,$D'(10,11,11,10,10)$,$D''(20,22,22,20,20)$此类维度值与其他任一维的相关系数都相等,这一点对于不同维间存在缩放及平移关系的多维信息集合比较重要.相关系数没有考虑维度负相关的问题,因此定义 6 中引入领域专家给出的维语义相关系数矩阵(其中的负数即表示二维度在语义层次上负相关.如文献[10]城市指标多因素分析中的教育、娱乐及犯罪率维,前两者是语义正相关的,而它们与第三者之间存在明显的负相关性),然后将二者结合得到任意两维的相关度 $R(D_i,D_j)$,而且语义相关系数矩阵对于由噪声数据带来的统计误差也有一定的修正作用.但是,这里要求 S 具有较好的一致性,对于一致性较差的 S 需要进行调整,否则会引起维度轴配置算法的谬误.

2.2 维度轴配置问题的描述

为了形式化描述维度轴配置问题,首先构造相关度矩阵:

$$R=\begin{bmatrix} S(D_1,D_1)r(D_1,D_1) & \dots & S(D_1,D_k)r(D_1,D_k) \\ \dots & \dots & \dots \\ S(D_k,D_1)r(D_k,D_1) & \dots & S(D_k,D_k)r(D_k,D_k) \end{bmatrix},$$

其中,$S(D_i,D_j)r(D_i,D_j)=S(D_j,D_i)r(D_j,D_i)$,$S(D_i,D_i)r(D_i,D_i)=9m$.然后引入维度邻接矩阵:

$$N=\begin{bmatrix} n(D_1,D_1) & \dots & n(D_1,D_k) \\ \dots & \dots & \dots \\ n(D_k,D_1) & \dots & n(D_k,D_k) \end{bmatrix},$$

其中, $n(D_i,D_j)=\begin{cases}1, & \text{if } D_i \text{ and } D_j \text{ are neighbors}\\ 0, & \text{otherwise}\end{cases}$.

定义 **7**. 对于给定的维相关度矩阵 R,维度轴最优配置问题可以定义为

$$\sum_{i}^{k}\sum_{j}^{k} n(D_i,D_j)\,|\,S(D_i,D_j)r(D_i,D_j)\,| \text{的值最大},\theta(D_i,D_j)\propto|S(D_i,D_j)r(D_i,D_j)| \tag{2}$$

其中, $\sum_{j}^{k} n(D_i,D_j)=2$,$\forall i=1,\dots,k$,$\theta(D_i,D_j)$表示 D_i,D_j 间的顺时针夹角.

定义 7 描述了维度轴最优圆形排列与相邻维夹角配置的问题,其中没有考虑维相关度的正负性,这是因为两条维度轴是否相邻以及相邻维间夹角的大小只依赖于维相关度的绝对值,绝对值越大,二维越相邻夹角越小,反之亦然;但维正向的选取是与相关度的正负有关联的.下面给出维度轴配置算法,其中包括维度轴排列顺序,相互之间夹角的计算及维正向的标定.

2.3 维度轴配置算法

Ankerst 证明了维度轴最优排列问题与货郎担问题(travelling salesman problem)具有相同的计算复杂度,都属于 NP 问题[19].因此,Ankerst 基于蚁群算法提出了启发式计算方法,文献[20]使用原理相同但较简单的近邻算法(nearest neighbor heuristic)解决这类问题.但是,上述算法只考虑了维度轴的排列次序,没有涉及各维度轴之间的夹角,而且在引入维正向概念之后,配置算法也需进行相应改变.本文对该算法进行扩充,提出了基于相关度的维度轴配置算法,见算法 1.

算法以相邻维度轴间夹角的余弦值表示相关度绝对值(归一化处理后)的大小,而且考虑了相邻维度轴夹角之和应该小于π的约束(否则会破坏维度轴的排列顺序).如果违反约束,即将各夹角缩小相同的倍数,夹角计算完成之后,再根据二维度之间相关度的正负性标定其维正向,结束维度轴配置.

算法 **1**. 维度轴配置算法.

Step 1. DR="", A=""; //DR 为维度轴的排列顺序,A 为相邻维度轴的夹角

Step 2. 计算相关度矩阵 R;

Step 3. 取相关度矩阵中绝对值最大元素$|S(D_i,D_j)r(D_i,D_j)|$⇒最相关的维度 D_i,D_j,并将其从矩阵中删除;

Step 4. $EL=D_i$; //维度轴排列 DR 的最左端

$ER=D_j$; //维度轴排列 DR 的最右端

$DR=EL+ER$

Step 5. 选 D_p,使$|S(D_i,D_p)r(D_i,D_p)|$最大,选 D_q,使得$|S(D_j,D_q)r(D_j,D_q)|$最大,且 D_p,D_q 不在 DR 中;

Step 6. if $|S(D_i,D_p)r(D_i,D_p)|\geq|S(D_j,D_q)r(D_j,D_q)|$

$EL=D_p$;

$DR=EL+DR$; //将 D_p 添加到 DR 的最左端

else

$ER=D_q$;

$DR=DR+ER$; //将 D_q 添加到 DR 的最右端

Step 7. 重复执行 Step 5~Step 6 直到将所有维度插入到 DR;

Step 8. 依次计算 DR 中相邻维度轴之间的夹角 $\theta_{u,u+1}=\begin{cases}\dfrac{\pi}{\omega_{1,k}+\sum\limits_{u}\omega_{u,u+1}}\omega_{u,u+1}, & \omega_{1,k}+\sum\limits_{u}\omega_{u,u+1}>\pi\\ \omega_{u,u+1}, & \text{otherwise}\end{cases}$,

其中, $\omega_{u,u+1}=\arccos\dfrac{|S(DR_u,DR_{u+1})r(DR_u,DR_{u+1})|}{9m}$,$u$=1,…,($k$–1),然后将其记录到 A 中;

Step 9. 由 $S(DR_u,DR_{u+1})r(DR_u,DR_{u+1})$正负性标定 DR_u,DR_{u+1} 的维度正向,如果为正,则为同向,否则为反向.

3 实验结果

3.1 实验设置

我们通过实验验证 ASC 方法的有效性,通过对比维度轴配置策略引入前后的可视化结果验证维度轴配置策略的正确性;对整个多维信息数据集在改进星形坐标系中的分布进行观察,获取较 SC 方法更为丰富的隐性知识并发掘各属性间的关系;在不同数据量和维数的情况下,将算法性能与经典降维算法性能进行比较.

原型采用 Visual C# 2005 基于 DirectX3d 开发包编写实现,软件平台 Microsoft Windows XP,机器配置为 Intel P4 2.6,1GB,120GB 硬盘.实验数据采用真实的汽车信息数据集和 AAUP(American Association of University Professors)数据集,汽车数据集包含 398 种汽车,涉及 9 个属性(mpg(每加仑汽油可行驶英里数,油耗倒数),cylinders(气缸大小),displacement(排量),horsepower(马力),weight(车重),acceleration(加速度),model year(样车出厂时间(1970~1982)),origin(产地,分为美国、欧洲、日本),name(唯一标识)).为了便于观察和对比,我们对上面列出的前 8 个属性进行可视化并以 name 属性作为信息点的标识;AAUP 数据集包含 1994 年 3 月美国 1 161 所高等院校的各类教师数量及其工资、奖金数据,涉及 15 个属性(prof.(s)(教授的平均工资(salary)),assoc.prof.(s)(副教授的平均工资),assist.prof.(s)(教授助理的平均工资),all profs.(s)(所有教授的平均工资),prof.(c)(教授的平均奖金(compensation)),assoc.prof.(c)(副教授的平均奖金),assist.prof.(c)(教授助理的平均奖金),all profs.(c)(所有教授的平均奖金),prof.(教授的数量),assoc.prof.(副教授的数量),assist.prof.(教授助理的数量),instructors(讲师的数量),faculty(全体教员的数量),college name(院校名称),Type(院校类别,分为 I,IIA,IIB)).我们可视化该数据集的前 13 个属性并以 college name 属性作为信息点标识,以 Type 作为标注依据.汽车数据集来源于 http://www.ics.uci.edu/~mlearn/MLRepository.html,AAUP 数据集来源于 http://lib.stat.cmu.edu/datasets/colleges/.

3.2 整体效果

图 4 给出了 ASC 方法针对汽车数据集的可视化结果,并且分别对各种汽车的产地进行了形状标注.其中,正三角形表示美国车,倒三角形表示欧洲车,方块表示日本车.从中可以直观分析得到 398 条汽车数据在各维度的分布情况及在改进星形坐标系中的整体结构:首先可以看到,origin 维、year 维以及相关度较高的

cylinders,horsepower,weight 和 displacement 维、acceleration 和 mpg 维分别聚集在一起;其次,沿着 origin 维数据集以条带形状聚集为 3 大类,说明 3 大地区生产的汽车相互之间差别较大;然后,沿着 year 维观察,美国车又分为 3 类,而且欧洲车和日本车基本出现在 year 维度轴的后半部分,说明汽车业在日本和欧洲进入繁荣阶段的时间比较晚,符合汽车产业在美国发源的历史;再从 cylinders,horsepower,weight 和 displacement 维分析,发现美国车的这几个指标都比较大,而日本车相同的指标比较小,这也反映了各类汽车的特点;最后对 acceleration,mpg 维进行研究,了解到早期的美国车油耗很高、加速度较小,而日本车和欧洲车油耗小、加速度高的特点比较明显.并且在图中可以看出,随着时间的推移,美国车的各类指标在不断向欧洲车、日本车靠近,说明欧洲车和日本车虽然发展比较晚但是代表了汽车工业发展的趋势.用户可以通过旋转、缩放维度轴进一步获取集合中的信息,借助分析可视化结果得到的隐性知识.我们可以通过分析给定的各类指标判断某种汽车的产地,根据汽车产地预测未知指标的数值.政府部门可以通过分析汽车工业发展趋势,制定国家汽车产业发展的方针;普通用户可以通过对各类汽车特点的了解,指导其进行私家车的购买.

图 5 为使用 ASC 方法可视化 AAUP 数据集的效果图,并按照院校类别对信息点形状进行了标注,其中,正三角形表示 I 类院校,倒三角形表示 IIA 类院校,方块表示 IIB 类院校.从图中可以看出,prof.(c),prof(s)并没有与相关度较高的 assoc.prof.(s),assist.prof.(s),all profs.(s),assoc.prof.(c),assist.prof.(c),all profs.(c)维聚在一起(维聚类 1),prof.也没有与 assoc.prof.,assist.prof.,instructors,faculty 维聚在一起(维聚类 2),而是 prof(s)聚入维聚类 2,prof.聚入维聚类 1,这说明美国院校的教授与其余教职员工差别较大,地位比较特殊;3 类院校虽然互有交叠,但是近似形成了 3 大聚类,这说明美国院校的分类标准并不是完全按照 13 种属性制定,并且 3 类院校之间在一般教职员工数量、教授工资、教授奖金等方面差别较大,而在各类院校内部一般教职员工的薪资水平和教授数量也有一定差别;效果图还直观地表现出类别高的院校师资力量比较强,教职员工的薪资水平也比较高.借助分析可视化结果得到的隐性知识,我们可以通过分析给定的各指标判断某院校所属类别,根据院校类别预测未知指标的数值.政府决策部门可以制定更为科学的院校分类标准及相应的发展策略,求职人员可以根据自身条件及需求申请不同院校工作.

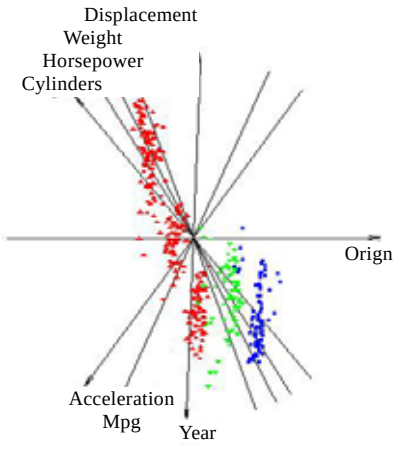


Fig.4 Visualization of cars using ASC

图 4 ASC 汽车数据集效果图

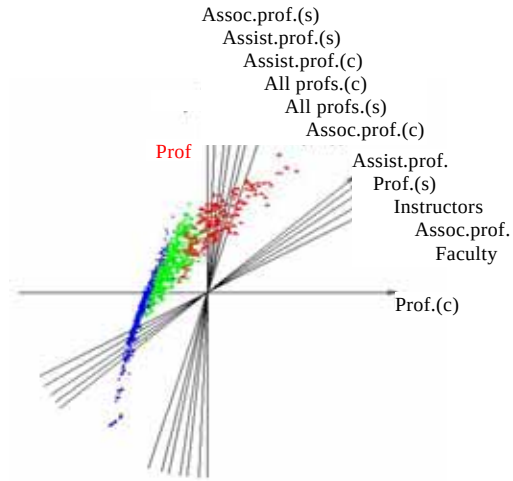


Fig.5 Visualization of AAUP using ASC

图 5 ASC AAUP 数据集效果图

为体现 ASC 方法的优越性,我们将 ASC 的效果图与无维度轴配置策略 ASC,SC 的效果图进行对比,由于篇幅所限,在此只选取了汽车数据集进行了对比实验及说明.图 6 是未使用维度轴配置策略的 ASC 效果图,图中聚类信息比较混乱,各聚类的交叠现象严重.虽然我们勉强能发现日本车、欧洲车、美国车在 mpg 维的分布特点,但是非常模糊,而且也丧失了各聚类在其他维的分布特点,这在一定程度上说明了维配置策略对于 ASC 的重要性及必要性.图 7 是使用 SC 方法可视化汽车数据集的效果图,从中可以得到 4 簇聚类,但是无法得到上述丰富的

维聚类信息、各聚类结构特点以及趋势信息等,也不能明确说明相对维度(如 origin,weight)在聚类产生过程中所起的作用.而且,虽然能够通过用户手动操作对维度轴进行优化配置使多维数据集的内部特点更加明显,但这一过程较为繁琐、缓慢,具有极大的不确定性.对比图 4 可知,ASC 方法较 SC 方法的聚类效果更好,不仅可以产生更细粒度的聚类(ASC 中美国车按照时间可进一步划分为 3 类(代)),而且聚类内部也比较紧密(ASC 中聚类中心与各信息对象在条带横向上的距离较小);能够揭示更多的隐性知识和信息集合的结构特征,尤其是其中隐含的维聚类信息及维度分布信息(在 SC 中无法观察得到上面阐述的部分 ASC 提供的知识);维配置策略能够使用户更加快速、准确地定位信息集合中的隐性知识(SC 经过多步旋转、缩放等操作才能获得较好的可视化结果(如图 7 所示)).

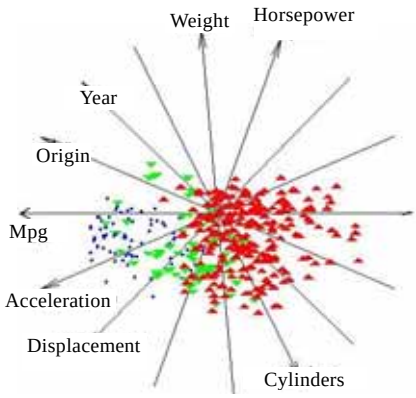


Fig.6 Visualization of cars using ASC without dimension configuration strategy

图 6 无维配置策略的 ASC 汽车数据集效果

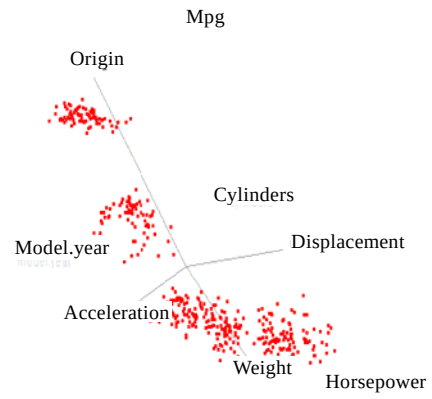
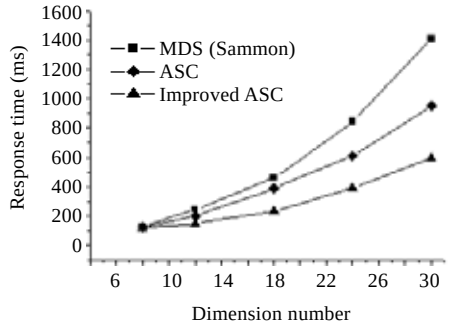


Fig.7 Visualization of cars using SC with manual dimension configuration[10]

图 7 手动配置维度轴的 SC 汽车数据集效果[10]

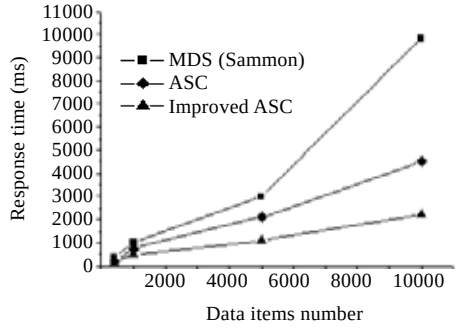
3.3 性能比较

本文选取经典降维映射技术 MDS 算法(使用 Sammon 映射实现,目标函数为 $\min f(x,y,z)=\frac{1}{2}\sum_{i<j}\frac{(d_{ij}-\delta_{ij})^2}{\delta_{ij}}$,其中,$d_{ij}$表示两多维信息在低维空间中的距离,$\delta_{ij}$为两多维信息在高维空间中的距离)与 ASC 算法及其优化形式进行了性能对比分析,我们分别以汽车信息数据集的数据量和维度数量为标准各随机生成了 4 组数据集,对 3 种算法在相同数据量、不同维数和相同维数、不同数据量的情况下进行了仿真实验,结果如图 8 所示.



(a) Relationship between the response time and dimension number

(a) 响应时间与维数的关系



(b) Relationship between the response time and data items number

(b) 响应时间与数据量的关系

Fig.8 Comparison of the MDS, ASC and improved ASC

图 8 MDS 算法、ASC 算法及其优化算法的性能比较

从图中可以看到,在不同维数和数据量的情况下,ASC 优化算法的效率最高,MDS 算法的效率最低,而且 ASC 优化算法的效率与多维信息集合的维数和数据量之间基本上呈线性关系.相对于 MDS 算法,数据量越大、维数越高,算法性能的优势越明显.这是因为,相对于 MDS 算法,ASC 算法属于增量式算法,最优化过程对每一个多维信息是相互独立的,而 MDS 算法不属于增量式算法,每一个多维信息之间在最优化过程中是相关的.因此,随着维数和数据量的增高,优化变量的数量增多,MDS 算法的效率下降较快;而相对于 ASC 算法,ASC 优化算法使用拟牛顿法或变尺度法求解最优化过程,收敛速度比较快,所以效率较高.

4 结论与进一步工作

本文在对星形坐标法及维排列算法进行研究的基础上提出了改进星形坐标法(ASC),产生了对用户有意义的低维可视空间,并通过引入极坐标对其降维算法进行了优化,最后借鉴基于相似度的维排列算法设计了维度轴配置策略.实验结果表明,ASC 方法可以高效展现海量高维信息集合的整体结构、在各维度上的分布及聚类情况等信息,能够辅助用户在知识发现过程初期,建立对多维信息集合的整体直观感受,能够辅助其在整个知识发现工作的过程中对多维信息集合进行更加深入的研究,而且维度轴配置策略可以极大地简化用户手动操作的繁杂性,使用户可以快速、准确发掘隐性知识,进而作出决策.

对于 ASC 方法,以后还需要在加强对分类数据(categorical data)的处理能力、提高算法效率、完善人机交互界面及多维信息语义的引入等方面进行研究.

致谢 非常感谢贺明科副教授在基于相关度的维度轴配置算法设计过程中所给予的指导和帮助.

References:

[1] Chen CM. Top 10 unsolved information visualization problems. IEEE Computer Graphics and Applications, 2005,25(4):12−16. [doi: 10.1109/MCG.2005.91]

[2] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999,31(3):264−323. [doi: 10.1145/331499.331504]

[3] Keim DA. Information visualization and visual data mining. IEEE Trans. on Visualization and Computers Graphics, 2002,8(1):1−8. [doi: 10.1109/2945.981847]

[4] de Oliveira MCF, Levkowitz H. From visualization to visual data mining: A survey. IEEE Trans. on Visualization and Computer Graphics, 2003,9(3):378−394. [doi: 10.1109/TVCG.2003.1207445]

[5] Wilkinson L, Anand A, Grossman R. High-Dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Trans. on Visualization and Computers Graphics, 2006,12(6):1363−1372. [doi: 10.1109/TVCG.2006.94]

[6] Inselberg A. The plane with parallel coordinates. The Visual Computer, 1985,1(2):69−91. [doi: 10.1007/BF01898350]

[7] Inselberg A, Dimsdale B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proc. of the IEEE Conf. on Visualization. Los Alamitos: IEEE Computer Society, 1990. 361−378. http://portal.acm.org/citation.cfm?id=949588

[8] Hoffman PE. Table visualizations: A formal model and its applications [Ph.D. Thesis]. Lowell: University of Massachusetts, 1999.

[9] Shao C, Huang HK. A new data visualization algorithm based on SOM. Computer Research and Development, 2006,43(3):429−435 (in Chinese with English abstract).

[10] Kandogan E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In: Proc. of the IEEE Information Visualization Symp. Los Alamitos: IEEE Computer Society, 2000. 4−8. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.8909

[11] Kandogan E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2001. 107−116. http://portal.acm.org/citation.cfm?id=502512.502530

[12] Hoffman PE, Grinstein GG, Marx K, Grosse I, Stanley E. DNA visual and analytic data mining. In: Proc. of the 8th Conf. on Visualization. Los Alamitos: IEEE Computer Society, 1997. 437−441. http://portal.acm.org/citation.cfm?id=266989.267116

[13] Carreira-Perpinan MA. A review of dimension reduction techniques. Technical Report, Report No.CS-96-09, Department of Computer Science, University of Sheffield, 1997.

[14] Yang J, Ward MO, Rundensteiner EA, Huang S. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In: Proc. of the Symp. on Data Visualization. Aire-la-Ville: Eurographics Association. 2003. 19−28.

[15] Nathan DC, Robert PB. Extension of star coordinates into three dimensions. In: Proc. of the Conf. on Visualization and Data Analysis. Bellingham: SPIE, 2007. 137−147. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.684

[16] Shaik JS, Yeasin M. Visualization of high dimensional data using an automated 3D star co-ordinate system. In: Proc. of the 2006 Int'l Joint Conf. on Neural Networks. Los Alamitos: IEEE Computer Society, 2006. 1339−1346.

[17] dos Santos SR, Brodlie KW. Gaining understanding of multivariate and multidimensional data through visualization. Computers & Graphics, 2004,28:311−325.

[18] Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: Proc. of the 1998 IEEE Symp. on Information Visualization. Washington: IEEE Computer Society, 1998. 52−60. http://portal.acm.org/citation.cfm?id=721216

[19] Wang LL, Zhang L. A study on traveling salesman problem. Computer Science, 2002,29(1):103−105 (in Chinese with English abstract).

[20] Artero AO, Oliveira MCF, Levkowitz H. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement. In: Proc. of the Conf. on Information Visualization. Washington: IEEE Computer Society, 2006. 707−712.

附中文参考文献:

[9] 邵超,黄厚宽.一种新的基于 SOM 的数据可视化算法.计算机研究与发展,2006,43(3):429−435.

[19] 汪林林,张林.对"货郎担问题"的深入解析.计算机科学,2002,29(1):103−105.

孙扬(1983－),男,山东济南人,博士生,主要研究领域为信息可视化,人机交互技术,语义 Web.

汤大权(1971－),男,博士生,副教授,主要研究领域为信息资源管理,信息可视化,信息检索.

唐九阳(1978－),男,博士,讲师,主要研究领域为对等网,信息集成,知识管理.

肖卫东(1968－),男,博士,教授,博士生导师,主要研究领域为信息管理,信息可视化,智能决策技术.