对偶刻度法及其在群体研讨中应用

刘怡君[1,2] 唐锡晋[1]

(1.中国科学院数学与系统科学研究院系统科学研究所,北京 100080;
2. 中国科学院研究生院,北京 100049)

摘要: 群体研讨中对参与成员意见的有效综合能够为人们深入分析研讨进程、理解并合理使用相关结果提供帮助。本文介绍一种处理定性数据的多元统计方法——对偶刻度法,结合它在群体研讨中的具体应用,描述其对专家发言(定性数据)的实时分析和处理(算法)、可视化显示(结果)等。

引 言

统计分析中经常考虑分类数据(categorical data)。分类数据一般有两种基本度量,一种为有序(ordered),一种是无序(unordered)。分析中前者称为序数型变量(ordinal variable),后者称为标称变量(nominal variable)。分类变量常常被认为是定性的,以区分数值型的定量变量[1]。故实际中我们讨论的定性数据,是指那些反映类别(categories)和水平(level),而不是区间刻度(interval scale)和定量的数据。此类数据多源自社会和经济统计问题(如问卷调查数据等),且一般以表格(tabular table)的形式表示。定性数据一般有五种表现形式:(1) 列联表(contingency table),(2) 应答频数表 (response-frequency table),(3)应答模式表(response-pattern table),(4)排序表(rank order table),(5)多维表(multidimensional table)[2]。

以表格形式表示的数据,其行与列所包含的实际意义难以直观的理解,本文介绍了一种多元统计方法——对偶刻度法(dual scaling method)用于分析这些分类数据。我们将该方法译为"对偶刻度法",也可译成"对偶标度法"。该法最早应用于18世纪,当时叫做代数特征值理论(Algebraic Eigenvalue Theory)。S.Nishisato在《Analysis of Categorical Data:dual scaling and its applications》一书中提到,Fisher和Guttman被认为是它的发明者[2]。

对偶刻度法又被称作Guttman加权法(Guttman weighting)、相关分析法(Correspondence Analysis)、定性数据的主成分分析法(Principal Components Analysis of Qualitative Data)、优化刻度法(Optimal Scaling)等等。

下面结合对偶刻度法在群体研讨中的应用,详细介绍其基本特性,具体算法等,并给出处理专家意

收稿日期:2004-07-05
基金项目:国家自然科学基金项目(79990580;70221001)
作者简介:刘怡君,中国科学院数学与系统科学研究院博士研究生;唐锡晋,博士,中国科学院数学与系统科学研究院副研究员。
致 谢:感谢清华大学智能技术与系统国家实验室郭崇慧博士和中科院系统科学研究所顾基发研究员对本文的支持。

见的可视化结果图。

对偶刻度法的统计描述和具体算法

本节将从对偶刻度法在群体研讨中的具体实例入手,介绍其基本的统计描述及特性,为下一步介绍对偶刻度法的基本算法做准备。

群体研讨中,一般有多个专家共同探讨某个问题,对专家发言进行有效综合能够提取发言过程所蕴涵的丰富信息,促进人们对相关结果的理解与合理使用。我们采用了对偶刻度法进行这样的尝试,期望从专家发言这类定性数据中提取一些定量的信息,如一些关于专家意见收敛或发散的程度等。其中,专家的发言(utterance)构成对象集(object set),而与发言相关的关键词(keyword)作为对象的属性集(attribute set),这样就形成了以专家发言和相应关键词为列和行的应答频数表[3,4]。

设:各专家发言中的关键词(keyword)的权值为,$x_1,x_2,\cdots,x_m$;各专家发言(utterance)的权值为,$y_1,y_2,\cdots,y_n$。见表1:

表1 专家意见与相应关键词集的应答频数表

x / y	keyword_1	keyword_2	$\cdots$	keyword_m	
	x_1	x_2	$\cdots$	x_m	
utterance_1 y_1	a_{11}	a_{12}	$\cdots$	a_{1m}	$y_1=\sum_{i=1}^{m} a_{1i}x_i$
utterance_2 y_2	a_{21}	a_{22}	$\cdots$	a_{2m}	$y_2\sum_{i=1}^{m} a_{2i}x_i$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
utterance_n y_n	a_{n1}	a_{n2}	$\cdots$	a_{nm}	$y_n\sum_{i=1}^{m} a_{ni}x_i$

其中,$a_{ij}=\begin{cases}0,\text{当 } keyword_i \text{没有在 } utterance_j \text{中出现}\\ 1,\text{当 } keyword_i \text{在 } utterance_j \text{中出现}\end{cases}$

为方便统计描述,引进一些记号:

y_{ji}:keyword_i 相对于 utterance_j 的权值;y_t:权值的和;f_t:权值数目的和;y_j:utterance_j 的权值的和;f_j:utterance_j 的权值数目的和;$\bar{y}_j=y_j/f_j=\text{utterance}_j$ 的平均权重;$\bar{y}_t=y_t/f_t=$总的 utterance_j 的平均权重;$c=y_t^2/f_t$:c 为修正项(correction term);

F:F 为(f_{ji})的矩阵,即

$$F=\begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1m}\\ x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nm}\end{pmatrix}$$

f:矩阵 F 中的列数据的和形成的 m×1 的向量,即

$$f=\begin{pmatrix}x_{11}+x_{21}+\cdots+x_{n1}\\ x_{12}+x_{22}+\cdots+x_{n2}\\ \vdots\\ x_{1m}+x_{2m}+\cdots+x_{nm}\end{pmatrix}$$

D:矩阵 F 中的列数据的和形成的 m×m 的对角阵,即

$$D=\begin{pmatrix}x_{11}+x_{21}+\cdots+x_{n1} & 0 & \cdots & 0\\ 0 & x_{12}+x_{22}+\cdots+x_{n2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & x_{1m}+x_{2m}+\cdots+x_{nm}\end{pmatrix}$$

D_n:矩阵 F 中的行数据的和形成的 n×n 的对角阵,即

$$D_n=\begin{pmatrix}x_{11}+x_{12}+\cdots+x_{1m} & 0 & \cdots & 0\\ 0 & x_{21}+x_{22}+\cdots+x_{2m} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & x_{n1}+x_{n2}+\cdots+x_{nm}\end{pmatrix}$$

下面是关于对偶刻度法的一些特性定义及其基本算法。

1、对偶刻度法遵循原则及其两点之间的距离度量和相关度的定义[2,5,6]

对偶刻度法遵循的原则是:Guttman 的内部一致性原则(Guttman's internal consistency),也就是说,根据平方和分解定理,有如下恒等式:

$$SS_t=SS_b+SS_w \qquad (1)$$

其中,SS_t 表示总偏差平方和,SS_b 表示组间平方和,SS_w 表示组内平方和。

总偏差平方和的大小反映了全部数据的波动大小;组内平方和反映了因随机误差的作用而在数据中引起的波动,应该越小越好;组间平方和反映了各列向量的不同作用在数据中引起的波动,应该越大越好。因此,定义特征值为:

$$\eta^2=SS_b/SS_t\ (0\leqslant\eta^2\leqslant1) \qquad (2)$$

对偶刻度法应用所求最大及次大特征值作为二维平面图的坐标轴,最大程度展示表格中行与列数据的特性和关系。

对偶刻度法中两点之间的距离度量不是传统的欧氏距离(Euclidean distance),而是 x^2 距离。其相关度定义如下:

$$\text{相关度(correlation)}=\frac{\text{对象与其属性间的协方差}}{\sqrt{\text{对象的方差}}\sqrt{\text{属性的方差}}}$$

2、对偶刻度法的基本算法[2]

根据上述已定义的统计描述及其变量，对偶刻度法的具体算法如下：

$SS_t=\sum\sum(y_{ji}-\bar{y}_t)^2$，经过运算并整理得：

$$SS_t=\sum\sum y_{ji}^2-c \tag{3}$$

同理：$SS_b=\sum f_j(\bar{y}_j-\bar{y}_t)^2$，整理得：

$$SS_b=\sum\left(y_j^2/f_j\right)-c \tag{4}$$

$SS_w=\sum\sum(y_{ji}-\bar{y}_j)^2$，整理得：

$$SS_w=\sum\sum y_{ji}^2-\sum\left(y_j^2/f_j\right) \tag{5}$$

中心化后得简化表达式：$SS_t=x'Dx$ (6)

$$SS_b=x'F'D_n^{-1}Fx \tag{7}$$

由(2),(6),(7)式得：

$$\eta^2=x'F'D_n^{-1}Fx/x'Dx \tag{8}$$

因此，可以通过标准的过程求解最大化的 η^2。下面介绍对 η^2 的优化：

不妨设 $x'Dx=f_t$ 为常量，根据(8)式，最大化 η^2，就是最大化 $x'F'D_n^{-1}Fx$。

在 $x'Dx=f_t$ 的条件下，设：$\omega=D^{1/2}x$，则 $x=D^{-1/2}\omega$，即在 $\omega'\omega=f_t$ 的条件下，最大化 $\omega'D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega$。应用拉各朗日乘子法，有：

$$Q(\omega)=\omega'D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega-\lambda(\omega'\omega-f_t) \tag{9}$$

分别对 ω,λ 求偏导：

$$\frac{\partial Q(w)}{\partial w}=2D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega-2\lambda I\omega=0 \tag{10}$$

$$\frac{\partial Q(w)}{\partial \lambda}=\omega'\omega-f_t=0 \tag{11}$$

由式(10)和(11)得：

$$D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega=\lambda\omega \tag{12}$$

对式(12)两端前乘 ω'，有：

$$\omega'D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega=\lambda\omega'\omega \tag{13}$$

由式(13),(8)得：

$$\lambda=\frac{\omega'D^{-1/2}F'D_n^{-1}FD^{-1/2}\omega}{\omega'\omega}=\frac{x'F'D_n^{-1}Fx}{x'Dx}=\eta^2 \tag{14}$$

此时，得到了对偶刻度法的一个很好的性质，即拉各朗日乘子中的对偶变量 λ 就是所求的 η^2，使以后的运算更加简洁方便。对偶变量 λ 的引入也是我们将对偶刻度法中的"dual"翻译为"对偶"的依据。

最后，所求的问题变成解下面的式子：

$$\left(D^{-1/2}F'D_n^{-1}FD^{-1/2}-\eta^2 I\right)\omega=0 \tag{15}$$

s.t. $\omega'\omega=f_t$

显然，问题转化为求解对称矩阵的特征向量。因此，采用经典主成分分析（principal component analysis- PCA）中的迭代算法即可求得最大的特征值 η^2。PCA 求取最大特征值 η^2 的实质是对原坐标系进行平移变换和旋转变换，使得新坐标系的原点与数据群点的重心相重合。

对偶刻度法的可视化显示实例

上述对偶刻度法的算法介绍是对其理性上的理解，为更好体会这个方法，下面我们以一个群体研讨的实例来说明。香山科学会议是我国科技界以探索科学前沿、促进创新为主要目标的高层次、跨学科、小规模的学术会议，从香山科学会议网站上所公布的历次香山会议的有关资料，我们主要抽取了 7 次有关"脑、意识和智力"前沿研究讨论的有关信息，以"脑与意识"为议题，以专家在会上的报告题目和主要的问答记录为蓝本进行了一次试验。发言次序按照会议的时间顺序和每个会议中报告顺序。专家们各抒己见，其发言和关键词形成如表 1 所示的应答频数表，(X,Y)构成了一个随着专家发言动态变化的矩阵，我们所研制的支持群体研讨的计算机环境(GAE)设定了一个定时器(2 分钟)，这样系统会定时动态地对当前已有的发言内容分析处理。

每一次动态的处理专家发言及其关键词时，智暴研讨室都会根据当前的最大特征值 η_1^2 和次大特征值 η_2^2，以 (x_{1i},x_{2i}) 为 keyword$_i$ 的坐标，(y_{1j},y_{2j})为 utterance$_j$ 的坐标，形成二维可视化图。其中，以最大特征值 η_1^2所表示的轴与数据变异的最大方向对应，次大特征值 η_2^2 决定的轴与前一轴标准正交，并且对应于数据变异的第二大方向[7]。两个正交的轴构成了一个平面坐标系，在此平面图中可反映专家发言及关键词间的关联程度。图 1,2 分别是试验过程中所捕获的有关发言关联的过程图和结果图。

对偶刻度法是一种探索性数据分析方法(exploratory data analysis)。用对偶刻度法生成的图和拓扑结构(topological structures)有很大的区别：拓扑图更大程度上是设计好后形成的结构化的图，如树型结构，网络的拓扑结构等；而生成的图则体现了数据自身存在的关联，是自然存在的。因此，这种生成的图被叫做解释图(interpretable graph)或感性图(perceptual map)，也就是说，对可视化结果图的理解

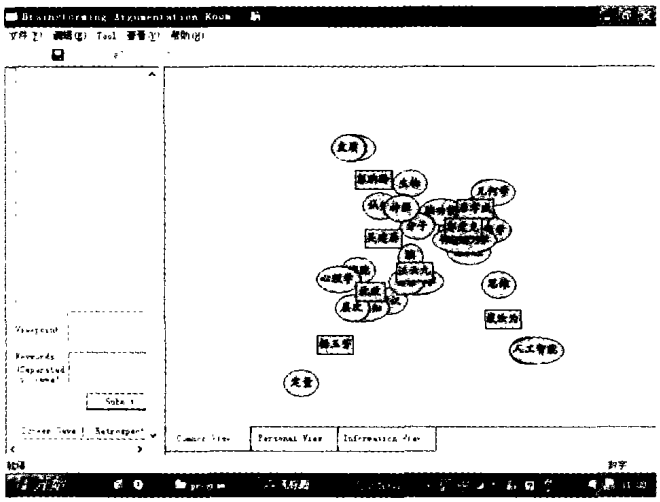

图 1 应用对偶刻度法处理研讨信息的可视化过程图

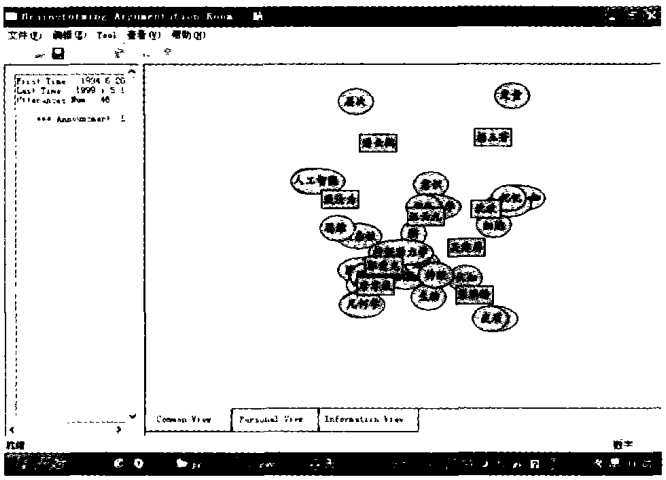

图 2 应用对偶刻度法处理研讨信息的可视化结果图

是基于解释的,而不是精确计算的,解释则需要领域专家的直觉和经验[8]。

一般情况下,在得到优化解后,需进行显著性检验。我们的研究主要目的是通过可视化视图给出一种实时的发言关联,以此激发专家的思考,激发其知识联想,拓展其思考空间,故没有进行检验。我们的工作借鉴了日本 ATR 的研究人员的研究成果[9]。

结束语

本文主要介绍了一种多元统计方法——对偶刻度法,并将其在群体研讨中加以具体应用,分析和处理了专家智暴过程中的定性研讨内容,得到文本聚类结果,并用二维图加以可视化显示,便于专家的直观理解和进一步的深入思考。

S.Nishisato 在 20 世纪 70 年代中期正式提出对偶刻度法的名字。近期,他又指出该算法的一些需继续研究的方面[10]。其中引起我们关注的一点是:算法本身在求其最大特征值时采用的是主成分分析方法,传统的主成分分析法最终是将所求特征向量投影到正交坐标轴中显示,但在很多实际数据中,正交轴并不能很好的表示形成的簇(clusters),即定性数据间的相关性,可能在多维空间中簇的意义更好的被解释。针对将输入数据映射到高维空间这一点,可以利用 Mercer 核(kernel)的原理,将核函数应用于主成分分析中,形成核主成分分析(kernel PCA)[11],这也是我们下一步要进行的研究。

参考文献:

[1] Agresti A. An Introduction to Categorical Data Analysis. New York: John Wiley & Sons, 1996: 1-3

[2] Nishisato S. Analysis of Categorical Data:dual scaling and its applications. University of Toronto Press, 1980:1-53

[3] 唐锡晋,刘怡君. 群思考的计算机支持工具研究. 西部开发与系统工程(顾基发主编). 中国系统工程学会第 12 届年会论文集. 北京:海洋出版社,2002, pp351-356

[4] 刘怡君,唐锡晋. 群思考中对偶刻度法的应用. 研究报告 No.AMSS-2002-25,中国科学院数学与系统科学研究院,2002. 11

[5] 庄楚强,吴亚森. 应用数理统计基础. 广州:华南理工大学出版社,1999

[6] Michael J. Greenacre. Correspondence Analysis in Practice. Academic Press, 1993

[7] 王惠文. 时序立体表数据分析的理论研究及其应用. 博士学位论文,北京航空航天大学研究生院,1992

[8] http://www2.chass.ncsu.edu/garson/pa765/correspondence.htm.

[9] Mase K, Sumi Y, Nishimoto K. Informal Conversation Environment for Collaborative Concept Formation. In: Ishida T eds. Community Computing: Collaboration over Global Information Networks, New York: John Wiley & Sons, Inc, 1998, pp165-205

[10] http://fcis.oise.utoronto.ca/~snishisato/

[11] Schölkopf B, Smola A, Müller K. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10: 1299-1319