

基于粗糙集流向图的多维定性数据分析方法

范霄文^{1,2}, 朱建平¹

1. 厦门大学经济学院计划统计系, 福建厦门 361005
2. 山西财经大学统计学院, 太原 030006

[摘要] 基于现有多维定性数据分析方法对海量数据分析时存在的局限性, 提出采用流向图理论进行多维定性数据的交叉分类分析。以图形的方式表示多个定性变量间的相互关联, 并从数量上对定性数据间存在的模式进行推理, 克服了高维列联表难以表示的困境, 并进行了实例分析。

[关键词] 流向图; 多维定性数据; 列联表

[中图分类号] O177.99

[文献标识码] A

[文章编号] 1000-7857(2007)16-0034-04

Multidimensional Qualitative Data Analysis Based on the Rough Set Flow Graphs

FAN Xiaowen^{1,2}, ZHU Jianping²

1. Department of Planning and Statistics, College of Economics, Xiamen University, Xiamen 361005, Fujian Province, China
2. College of Statistics, Shanxi University of Finance and Economics, Taiyuan 030006, China

Abstract: In view of the limitations of the existing multidimensional qualitative data analysis methods in dealing with massive data, this paper proposes to use the flow graphs theory for the multi-dimensional qualitative data cross-classification analysis. The inter-relations between qualitative variables are expressed with graphs and the pattern related with the qualitative data is analyzed based on inference from the values of the graphs, which can overcome the difficulty in using the high-dimensional contingency tables. Examples are given.

Key Words: flow graphs; multidimensional qualitative data; contingency table

CLC Number: O177.99

Document Code: A

Article ID: 1000-7857(2007)16-0034-04

0 引言

在现有的统计方法中, 列联表和对数线性模型是多维定性数据交叉分类的主要分析技术。然而, 随着数据采集和存储技术的进步, 所面对的是日益增多的庞大数据库, 大容量的数据带来了传统统计方法难以解决的问题。一方面, 传统的多维定性数据分析的对象是随机抽取的定性数据, 对于定性变量间的关系则是通过独立性

检验来实现的。当数据本身不具有随机性时, 这一方法就无法应用。目前数据库的数据大多都不是随机获取的, 因此, 现有的独立性检验方法已难以满足这种分析需求。另一方面, 数据库中涉及的变量个数越来越多, 传统的列联表的表示方式显然已无法满足海量数据的要求。众所周知, 多于 3 个因素的列联表表示, 既不方便也不便于说明因素间的关系, 高维列联表的表示至今仍是

收稿日期: 2007-06-28

基金项目: 2006 年教育部社科研究项目 (06JA910003)

作者简介: 范霄文, 福建省厦门市厦门大学经济学院计划统计系, 博士研究生, 主要从事数据挖掘和定性数据分析研究;

E-mail: fwx673@sina.com

朱建平 (通讯作者), 福建省厦门市厦门大学经济学院计划统计系, 教授, 主要从事数据挖掘和数理统计研究;

E-mail: jpzhu@163.com

统计研究的一个领域。因此,有必要引入新的方法,改变定性数据传统的表示方法。

流向图 (Flow Graphs) 是粗糙集理论创始人 Z. Pawlak 于 2002 年提出用于分析信息流决策算法的数学工具,本质上也是一种新的分析高维定性数据的推理方法。流向图理论采用图形的方式,对定性数据间存在的模式进行推理,它不仅可以用图同时表示多个定性变量的相互关联,而且可以描述变量间存在的定量关系。基于上述两方面的考虑,本文提出采用粗糙集理论中的信息系统作为原始数据的表示方法,用表示信息流的流向图分析多维定性变量间的关系。

1 粗糙集流向图理论

流向图是一种有向、非循环、有限的信息流向图,定义为 $G=(N, B, \varphi)$ 。其中 N 定义为图中节点的集合, $B \subseteq N \times N$ 定义为有向分支的集合, $\varphi: B \rightarrow R^+$ 定义为信息流函数, R^+ 定义为非负的实数集合。

流向图任一节点 $x \in N$ 的输入定义为 $I(x) = \{y \in N : (y, x) \in B\}$, 输出定义为 $O(x) = \{y \in N : (x, y) \in B\}$; 流向图 G 的输入定义为 $I(G) = \{x \in N : I(x) \in \phi\}$, 输出定义为 $O(G) = \{x \in N : O(x) \in \phi\}$; 流向图 G 的输入和输出节点定义为外节点,其余节点定义为内节点。

如果 $(x, y) \in B$, 则定义 $\varphi(x, y)$ 为节点 x 到节点 y 的信息流量。定义流向图 G 中任一节点 x 的信息流入量为 $\varphi_+(x) = \sum_{x \in I(x)} \varphi(y, x)$, 信息流出量为 $\varphi_-(x) = \sum_{y \in O(x)} \varphi(x, y)$ 。分

别定义 $\varphi_+(G) = \sum_{x \in I(G)} \varphi_+(x)$ 和 $\varphi_-(G) = \sum_{x \in O(G)} \varphi_-(x)$ 为流向图 G 的信息流入量与流出量。

在流向图中,对于任一内节点 x , 其流量 $\varphi(x) = \varphi_+(x) = \varphi_-(x)$; 对于外节点,如果 $x \in I(G)$, 则 $\varphi(x) = \varphi_-(x)$, 如果 $x \in O(G)$, 则 $\varphi(x) = \varphi_+(x)$; 显然,流向图 G 的流量 $\varphi(G) = \varphi_+(G) = \varphi_-(G)$ 。

下面以图 1 为例对上述定义加以说明。

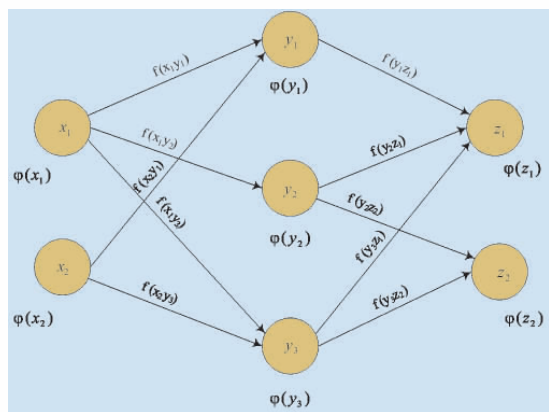


图 1 流向图

Fig. 1 Flow Graph

在图 1 中,圆形表示节点, $N = \{x_1, x_2, y_1, y_2, y_3, z_1, z_2\}$, 有向分支 (x_1, y_1) 用从节点 x_1 到节点 y_1 的有向线段表示。节点 y_1 的输入节点是 x_1 和 x_2 , 输出节点是 z_1 。流向图的输入节点是 x_1 和 x_2 , 输出节点是 z_1 和 z_2 。节点 y_1, y_2, y_3 是流向图的内节点。有向分支 (x_1, y_1) 的流量是 $\varphi(x_1, y_1)$ 。节点 y_1 的流入量 $\varphi_+(y_1) = \varphi(x_1, y_1) + \varphi(x_2, y_1)$, 流出量 $\varphi_-(y_1) = \varphi(y_1, z_1)$, 且 $\varphi_+(y_1) = \varphi_-(y_1)$ 。流向图 G 的流入量 $\varphi_+(G) = \varphi(x_1) + \varphi(x_2)$, 流出量 $\varphi_-(G) = \varphi(z_1) + \varphi(z_2)$, 且有 $\varphi_+(G) = \varphi_-(G) = \varphi(G)$ 。粗糙集流向图中有向分支 (x, y) 的信息流量强度为

$$\sigma(x, y) = \frac{\varphi(x, y)}{\varphi(G)}$$

显然, $0 \leq \sigma(x, y) \leq 1$, 测度的是有向分支 (x, y) 的信息流量占流向图 G 信息流量的比例。相应地我们分别可定义:

$$\sigma_+(x) = \frac{\varphi_+(x)}{\varphi(G)} = \sum_{y \in I(x)} \sigma(y, x)$$

$$\sigma_-(x) = \frac{\varphi_-(x)}{\varphi(G)} = \sum_{y \in O(x)} \sigma(x, y)$$

$$\sigma_+(G) = \frac{\varphi_+(G)}{\varphi(G)} = \sum_{x \in I(G)} \sigma_+(x)$$

$$\sigma_-(G) = \frac{\varphi_-(G)}{\varphi(G)} = \sum_{x \in O(G)} \sigma_-(x)$$

显然有

$$\sigma_-(G) = \sigma_+(G) = \sigma(G) = 1$$

$$\sigma(y|x) = \frac{\sigma(x, y)}{\sigma_-(x)}$$

$$\sigma(x|y) = \frac{\sigma(x, y)}{\sigma_+(y)}$$

节点序列 $[x_1, x_2, \dots, x_n]$ 形成了 G 中的路径, $(x_i, x_{i+1}) \in B, 1 \leq i \leq n-1$ 。路径 $[x_1, x_2, \dots, x_n]$ 的 $\sigma(x_n|x_1)$ 定义为

$$\sigma[x_n|x_1] = \prod_{i=1}^{n-1} \sigma(x_{i+1}|x_i) \quad \sigma[x_1|x_n] = \prod_{i=1}^{n-1} \sigma(x_i|x_{i+1})$$

从统计学的角度看,流向图中的每个节点对应定性变量的一个取值,有向分支表示多个定性变量不同取值之间的联系,不同变量对应流向图中不同的“层”,同一变量的不同取值对应流向图同一层的不同节点。节点 x 到节点 y 的信息流量 $\varphi(x, y)$ 对应于多维定性变量交叉分类的频数,而信息流强度则可对应交叉分类的频率,流向图中的数量关系与多维定性变量交叉分类后的数量关系具有对等关系。且流向图是基于数据推理而非推断来进行分析的,虽然其数量关系满足贝叶斯理论,但流向图是从信息流的分布比例出发来进行分析,因此,可以应用流向图进行多维定性数据的交叉分类分析。

2 基于流向图的多维定性数据交叉分类实证分析

流向图是建立在对多维定性数据进行交叉分类的基础上,下面以来自数据库的 11 254 名客户的经营类

型 Y 、市场类型 S 、经营规模 G 和守法情况 D 为例,就流向图应用于多维定性数据交叉分类进行实证分析。

表 1 给出了 (Y, S) , (S, G) 和 (G, D) 的交叉分类频数表,每一个频数表可以定义一个二元的流向图。如 (Y, S) 形成的频数表, Y 和 S 的取值集合构成了流向图的节点集, $N=\{\text{便利店, 超市, 商场, 食杂店, 烟酒店, 娱乐服务, 其他, 城市, 农村}\}$, 频数表中的每一行是流向图中的一个有向分支的节点及流量(交叉分类频数)。将 3 个二元流向图组合起来就形成了如图 2 所示的反映多维定性变量关系的流向图,图中的每一个有向分支分别标示 3 个数值: $\sigma(X, Y)$, $\sigma(X|Y)$ 和 $\sigma(Y|X)$, 具体数值见表 2。这些数值是在表 1 基础上按照前述公式计算得到的,反映了变量间的数量联系。如在表 2 中, $\sigma(\text{便利店, 城市})$ 为 0.083 1, 表示在客户中经营类型属于便利店而市场类型属于城市的客户所占比例是 8.31%; $\sigma(\text{城市, 便利店})$ 为 0.876 3, 表示经营类型是便利店的客户中, 市场类型

表 1 (Y, S) , (S, G) , (G, D) 交叉分类频数表
Table 1 (Y, S) , (S, G) , (G, D) cross-classification frequency table

Y	S	$\varphi(Y, S)$
便利店	城市	935
便利店	农村	132
超市	城市	97
超市	农村	10
商场	城市	23
商场	农村	2
食杂店	城市	5 006
食杂店	农村	3 423
烟酒店	城市	963
烟酒店	农村	50
娱乐服务	城市	160
娱乐服务	农村	4
其他	城市	374
其他	农村	75
S	G	$\varphi(S, G)$
城市	大	1 507
城市	中	4 559
城市	小	1 492
农村	大	746
农村	中	2 193
农村	小	757
G	D	$\varphi(G, D)$
大	合法	2 248
大	不合法	5
中	合法	6 727
中	不合法	25
小	合法	2 245
小	不合法	4

表 2 图 2 流向图的数值表
Table 2 Value table of flow graphs

Y	S	$\sigma(Y, S)$	$\sigma(S Y)$	$\sigma(Y S)$
便利店	城市	0.083 1	0.876 3	0.123 7
便利店	农村	0.011 7	0.123 7	0.035 7
超市	城市	0.008 6	0.906 5	0.012 8
超市	农村	0.000 9	0.093 5	0.002 7
商场	城市	0.002	0.92	0.003
商场	农村	0.000 2	0.08	0.000 5
食杂店	城市	0.444 8	0.593 9	0.662 3
食杂店	农村	0.304 2	0.406 1	0.926 1
烟酒店	城市	0.085 6	0.950 6	0.127 4
烟酒店	农村	0.004 4	0.049 4	0.013 5
娱乐服务	城市	0.014 2	0.975 6	0.021 2
娱乐服务	农村	0.000 4	0.024 4	0.001 1
其他	城市	0.033 2	0.833	0.049 5
其他	农村	0.006 7	0.167	0.020 3
S	G	$\sigma(S, G)$	$\sigma(G S)$	$\sigma(S G)$
城市	大	0.133 9	0.199 4	0.668 9
城市	中	0.405 1	0.603 2	0.675 2
城市	小	0.132 6	0.197 4	0.663 4
农村	大	0.066 3	0.201 8	0.331 1
农村	中	0.194 9	0.593 3	0.324 8
农村	小	0.067 3	0.204 8	0.336 6
G	D	$\sigma(G, D)$	$\sigma(D G)$	$\sigma(G D)$
大	合法	0.199 8	0.997 8	0.200 4
大	不合法	0.000 4	0.002 2	0.147 1
中	合法	0.597 7	0.996 3	0.599 6
中	不合法	0.002 2	0.003 7	0.735 3
小	合法	0.199 5	0.998 2	0.200 1
小	不合法	0.000 4	0.001 8	0.117 6

属于城市的客户所占比例为 87.63%; $\sigma(\text{便利店, 城市})$ 为 0.123 7, 表示在市场类型属于城市的客户中, 经营类型属于便利店的客户所占比例为 12.37%。通过这 3 个数值不同变量间的联系可见一斑。

为了进一步分析经营类型与守法情况之间的关系, 可以将 $\{Y, S, G, D\}$ 流向图合并处理为如图 3 所示的 $\{Y, D\}$ 流向图。其中:

$$\sigma\langle D|Y \rangle = \sum_{Y, S, G} \sigma(S|Y) \cdot \sigma(G|S) \cdot \sigma(D|G)$$

$$\sigma\langle Y|D \rangle = \sum_{Y, S, G} \sigma(Y|S) \cdot \sigma(S|G) \cdot \sigma(S|D)$$

$$\sigma\langle Y|D \rangle = \sigma(Y) \cdot \sigma(D|Y) \cdot \sigma(D) \cdot \sigma(Y|D)$$

由图 3 可以看出, 客户经营类型主要集中在食杂店、便利店和烟酒店, 这 3 类客户占总客户数的 93%, 其中 75% 的客户是食杂店。在所有客户中, 只有 0.3% 的客户属于不合法, 而不合法的客户主要是食杂店、烟酒

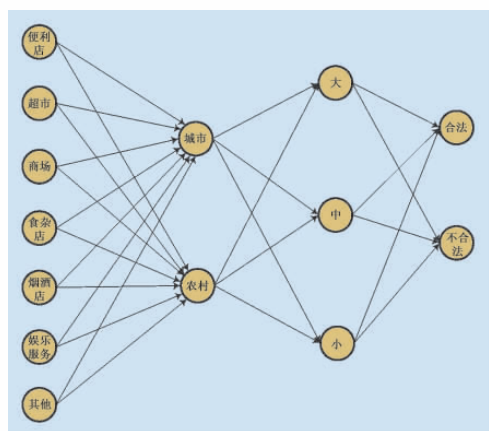


图 2 {Y,S,G,D}流向图

Fig. 2 {Y,S,G,D} flow graph

店和其他 3 种经营类型的客户。在不合法的客户中,食杂店占到了 53%,烟酒店占 37%。可见,管理的重点客户是食杂店和烟酒店。在管理中,更关心的是哪些经营类型的客户可能会出现不合法的情况,因此,需要测度经营类型与不合法之间的依存关系。在图 3 中,每一个向分支节点的依存关系可以用依存系数 $\eta(x, y)$ 进行测度,公式如下:

$$\eta(x, y) = \frac{\sigma(y|x) - \sigma(y)}{\sigma(y|x) + \sigma(y)}$$

当 $\eta(x, y) > 0$ 时,表明两节点间存在正的依存关系;当 $\eta(x, y) < 0$ 时,表明两节点间存在负的依存关系;当 $\eta(x, y) = 1$ 时,表明两节点间存在完全的正依存关系;当 $\eta(x, y) = -1$ 时,表明两节点间存在完全的负依存关系。

表 3 图 2 的部分依存系数

Table 3 Dependency coefficient for Fig.2

Y	D	$\eta(Y, D)$
便利店	不合法	-1
超市	不合法	-1
商场	不合法	-1
食杂店	不合法	-0.168 18
烟酒店	不合法	0.605 839
娱乐服务	不合法	-1
其他	不合法	0.251 173

从表 3 给出的计算结果可以看出,便利店、超市、商场和娱乐服务与不合法之间存在完全的负依存关系,表明这些经营类型的客户不合法的可能性很小。烟酒店与不合法存在较强的正依存,其他与不合法也存在一定的正依存,表明经营类型为烟酒店的客户不合法的可能性较大,而经营类型是其他的客户也具有一定的可能性。食杂店与不合法存在负依存关系,表明经营类型属于食杂店的客户为不合法的可能性较小。虽

然不合法中食杂店的所占比例较大,但考虑到食杂店占客户总数的比例为 75%,这一结果还是较为合理的。

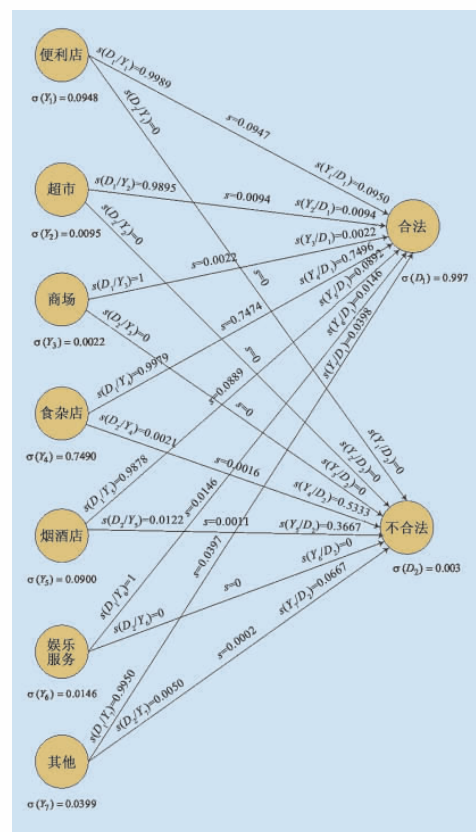


图 3 {Y, D}流向图

Fig. 3 {Y, D} flow graph

3 结论

多维定性数据存在于各个研究领域,采用流向图进行多维定性数据的交叉分类分析,不同于传统的高维关联分析。其结果表明:① 可以直观清晰地表示高维定性变量间的关系;② 对于多维定性变量间的数量关系是基于推理而非推断。因此,它解决了传统方法在对非随机获取多维定性数据进行分析时存在的问题。将流向图引入多维定性数据的分析仅仅是研究的开始,今后还需要对流向图各有向分支间的数量关系做进一步深入的理论研究。

参考文献 (References)

- [1] PAWLAK Z. Flow graphs and data mining [M]// PETERS J F, SKOWRON A Eds. Transaction on rough sets III, vol 3400. Heidelberg: Springer, 2005: 1-36.
- [2] BUTZ C J, YAN W, YANG B. An efficient algorithm for inference in rough set flow graphs [M]// PETERS J F, SKOWRON A Eds. Transacting on rough sets V, Vol 4100. Heidelberg: Springer, 2006: 102-122.

(责任编辑 李慧政)