

# 基于属性重要性的定性数据聚类分析及应用

朱建平, 曾玉钰

(厦门大学 经济学院 计划统计系, 福建 厦门 361005)

**摘要:**传统的聚类方法大多是基于距离或者是样品间相似度的,这就要求所分析的数据必须是定量的。但是在数据挖掘中,存在着大量的定性数据,传统的聚类分析方法已不再是一个可行的方法,这就需要寻找一个可以有效处理定性数据的聚类方法。粗糙集是处理定性数据的有效方法,在详细阐述粗糙集的相关概念后,利用属性重要性的概念,提出了一种能有效处理定性数据的聚类分析方法,并利用了数据对该方法进行了实证分析,取得了良好的结果。

**关键词:**属性重要性;聚类分析;粗糙集;等价关系

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2007)12-0089-03

## Analysis and Application of Qualitative Data Clustering Approach Based on Attribute Importance

ZHU Jian-ping, ZENG Yu-yu

(Dept. of Planning and Statistics, School of Economics, Xiamen University, Xiamen 361005, China)

**Abstract:** Most of the current clustering approaches are based on the distance among the data or of the similarity of the data, which makes the data analyzed must be quantifiable data. In data mining, there are many qualitative data. That makes the traditional clustering techniques are not useful in tackling the qualitative data as hoped. So need to find an effective clustering method to cope with the qualitative data. Rough set is an useful tool to deal with the qualitative data. After explicating the relative concepts of the rough set, introduced a new clustering approach by using attribute importance concept, which can deal with the high dimensions data effectively. At last, make an empirical analysis of the data and obtain a good clustering result.

**Key words:** attribute importance; clustering analysis; rough set; equivalence relation

### 0 引言

聚类分析(cluster analysis)是数据处理的基本方法,它是研究如何将研究对象按照多个方面的特征进行综合分类的一种统计方法。

传统的聚类方法的共同点是它们都是基于距离或者是样品间相似度的,而且这些方法都要求样品的各个属性必须是可以测度的,即必须是定量的数据,只有这样,才能计算出样品间的距离或相似度。但是,随着信息社会的飞速发展,需要处理的信息量越来越大,而且大多数的数据不再是定量的数据,而是定性的数据,传统的聚类分析方法已经无法直接对定性数据进行处理,而是利用变换将定性数据转化为定量数据后再进行分析,而一旦数据的属性较多,则转化为定性数据就

比较困难了。

为了解决多维定性数据问题,有许多学者进行了这方面的研究,如 Ziarko W. 在文献[1]中提出了变量缩减的粗糙集模型;Barbar'A, D. 和 Chen P. 在文献[2]中利用自反性的概念,提出了 Fractal Clustering (FC)的高维数据聚类方法;李树军、纪宏军在文献[3]中用对应聚类分析方法识别和剔除不必要的或有害的变量,达到了压缩数据的目的;来升强、朱建平在文献[4]中提出了用粗糙集方法选择出最优子空间,并利用等价关系的属性集产生分类的大型数据聚类方法;张文修在文献[5]中全面而详细地阐述了粗糙集的全部理论及其应用。文中也引入了粗糙集理论(roughset theory),提出了通过分析各个属性的相对重要性,略去对于聚类分析作用不大的属性,然后再利用等价关系对定性数据进行聚类的方法。

### 1 粗糙集中的相关概念

粗糙集方法是以不完全信息或知识去处理一些不

收稿日期:2007-03-07

基金项目:国家教育部“新世纪优秀人才支持计划”资助(NCET-04-0608)

作者简介:朱建平(1962-),男,河南浚县人,教授,博士,博士生导师,研究方向为数理统计、数据挖掘。

分明的现象,或者依据观察、量度到的某些不确定结果而进行分类数据的能力。而且粗糙集理论中最核心的概念就是等价关系,这一概念是通过样本在属性集上取值相同或相似来定义的,这和聚类分析中的“物以类聚”的原则是类似的。由此可见,它所处理的属性值可以是定性的数据,而且利用粗糙集的等价关系来对定性数据进行聚类分析是一个可行的办法。

在粗糙集理论中通常称一个数据表为一个由四个要素构成的信息系统,即  $S = \{U, Q, V, f\}$ , 这里,  $U$  是对象(样本)的一个有限集  $\{x_1, x_2, \dots, x_n\}$ ,  $N$  为所研究的对象总量,  $Q$  是描述对象(样本)属性的一个有限集合  $\{q_1, q_2, \dots, q_n\}$ ,  $V$  是属性集  $Q$  中每一个属性的值域,  $f$  是  $U \times Q$  到  $V$  的一个映射, 即:  $f: U \times Q \rightarrow V$ , 且  $f(x, q) \in V_q$ , 它表示对象  $x$  关于属性  $q$  的取值, 指的是  $U$  中每一个对象(样本)的属性值(详见文献[6])。

### (1) 等价关系和等价类。

在一个信息系统  $S = (U, Q, V, f)$  中,  $A \subseteq Q$  是任一子集, 称  $\text{IND}(A) = \{(x_i, x_j) \in U \times U : f(x_i, q) = f(x_j, q), q \in A\}$  为论域  $U$  上的一个等价关系, 记为  $R_A$ 。其中,  $x_i, x_j \in U (i \neq j)$ , 满足  $f(x_i, q) = f(x_j, q), q \in A$ , 则称  $x_i, x_j$  关于属性子集  $A$  是不可识别的, 记为  $x_i \text{IND} x_j$  或  $x_i \tilde{A} x_j$ 。

若  $A \subseteq Q$ , 由等价关系  $R_A$  作为一个准则, 可以将论域  $U$  划分为一个等价类族  $\{X_1, X_2, \dots, X_n\}$ , 这个类族实质上是由等价关系  $R_A$  在  $U$  上导出的一个聚类结果。对于任何一个  $x_i \in U$ , 包含  $x_i$  的一类记为  $[x_i]_A$ , 称其为关于  $A$  的由  $x_i$  生成的等价类。

### (2) 正域和属性重要性。

设  $X \subseteq U$  是任一子集,  $R$  是  $U$  上的等价关系, 即序对  $K = (U, R)$  称为一个近似空间, 则称  $R_- X = \bigcup \{Y \in U/R : Y \subseteq X\}$  为  $X$  的  $R$  正域, 记为  $\text{POS}_R(X)$ 。实际上,  $\text{POS}_R(X)$  是  $K$  中含在  $X$  中的最大可定义集。

属性集  $B' \subseteq B$  相对于由属性集  $C$  引起的分类的重要测度可以表示为:  $\gamma_B(C) - \gamma_{B-B'}(C)$ 。其中,  $\gamma_B(C) = \text{card} \text{POS}_B(C) / \text{card} U$ , 且  $\text{card}(U)$  表示集合  $U$  的基数。这里应该注意到, 测度值的取值范围是  $[0, 1]$ , 而且测度值越大,  $B'$  的重要性越大。

## 2 基于属性重要性的定性数据聚类方法

在许多介绍粗糙集理论的文献中, 大多都有介绍利用等价关系进行聚类的方法, 但是那只是选择一些属性子集对样品进行简单的聚类, 其结果并没有很强的实际意义。况且, 设粗糙集中属性集  $Q$  共有  $n$  个元素, 即有  $n$  个属性, 那么利用等价关系聚类一共可以得

到  $2^n - 1$  个聚类结果, 而在这  $2^n - 1$  个聚类结果中, 并不是所有的聚类结果都是有意义的。在此, 利用基于属性重要性的理论构造定性数据的聚类方法。

首先, 计算每一个属性相对于总属性集  $\theta$  的重要性。某一属性  $q_i$  相对于总属性集  $Q$  的重要性可以由  $\gamma_Q(Q) - \gamma_{Q-\{q_i\}}(Q)$  得到:

其中,  $\gamma_{Q-\{q_i\}}(Q) = \text{card} \text{POS}_{Q-\{q_i\}}(Q) / \text{card} U$ ,  $\gamma_Q(Q) = 1$ 。

若  $\gamma_Q(Q) - \gamma_{Q-\{q_i\}}(Q)$  的值越大, 说明  $q_i$  在整个属性集  $Q$  的重要性越大, 则它对于样品的聚类结果的影响作用越大, 是决定样本的重要属性之一。反复进行这一步计算, 直到计算出所有属性的重要性。设将每一个属性按重要性大小排列为  $q_1, q_2, \dots, q_n$ 。

其次, 在聚类分析中, 在数据繁多时, 往往可以利用其主要信息进行分析, 而忽略影响样本的不重要的信息, 这样做既能减少聚类的复杂度, 也不会影响聚类的结果。因此, 可以按属性顺序为  $q_1, q_2, \dots, q_n$  的顺序画出新的数据表, 这样从左到右, 属性的重要性依次减小。这里根据实际问题分析的需要可以设定一个阈值  $\epsilon$ , 去掉那些重要性小于等于阈值  $\epsilon$  的属性, 留下那些重要性大于阈值  $\epsilon$  的属性。通常来说阈值  $\epsilon$  可以设为零, 因为阈值  $\epsilon$  取零, 意味着这个属性对于聚类结果没有影响作用, 可以舍弃; 也可以根据所有属性的属性重要性大小的分布图, 在相对大的属性重要性和相对小的属性重要性之间, 选择一个值作为阈值  $\epsilon$  的取值, 这样能舍弃那些相对于所有属性来说, 对聚类结果影响作用相对小的属性, 而保留了对聚类结果影响作用相对大的属性。

这样, 通过了阈值判断的  $m$  个属性, 可以将其按属性重要性从大到小排列为  $q_1^*, q_2^*, \dots, q_m^*$ , 可以说这些属性都是在属性集  $Q$  中必需且相对重要的属性, 是决定样本的关键属性。它们对于样本聚类分析起着决定作用, 这样就可以以  $\{q_1^*, q_2^*, \dots, q_m^*\}$  作为属性集  $X$ , 并利用等价关系的概念对数据进行聚类分析, 即若样本  $i$  和样本  $j$  在属性集  $\{q_1^*, q_2^*, \dots, q_m^*\}$  上取值相同, 则样本  $i$  和样本  $j$  可以归为一类。

由于  $\{q_1^*, q_2^*, \dots, q_m^*\}$  是根据阈值判断后属性集  $Q$  中的相对重要的属性, 因此这样得到的聚类结果将具有良好的分类效果, 且意义明确。得到的这些影响样品的几个重要因素  $\{q_1^*, q_2^*, \dots, q_m^*\}$  也有利于对样品进行其他的分析。

## 3 实证分析

为了进一步地说明该方法的基本思路, 这里选取

了某大学计算机系某班 34 名同学的综合测评的成绩进行分析,这里的综合测评是根据 5 个指标来评分的,每个指标的取值集合为{好、中、差}。这样,这些数据构成了一个信息系统,共有 34 个样本,每个样本都有 5 个属性,每个属性的值域都是{好、中、差}。样本数据如表 1 所示,每个属性的含义如表 2 所示。

表 1 34 名同学的综合测评成绩

学号	属性 I	属性 II	属性 III	属性 IV	属性 V
1	中	差	差	差	差
2	好	中	好	好	中
3	中	好	差	差	差
4	中	差	好	好	差
5	好	好	中	中	好
6	中	差	中	中	差
7	好	好	差	差	中
8	好	好	差	差	中
9	中	中	中	中	中
10	中	中	中	中	中
11	中	中	中	中	中
12	中	中	差	中	差
13	好	好	好	好	好
14	中	中	中	好	中
15	中	差	好	好	差
16	中	差	中	中	差
17	中	中	差	中	差
18	中	中	差	差	差
19	中	中	中	好	中
20	中	中	差	差	差
21	好	好	中	中	好
22	中	中	中	中	中
23	中	差	差	差	差
24	中	好	中	中	中
25	中	中	差	中	差
26	中	差	中	好	差
27	好	好	好	好	好
28	好	中	好	好	中
29	中	差	好	好	差
30	中	差	差	差	差
31	中	差	差	差	差
32	中	中	中	中	中
33	中	差	好	好	差
34	中	中	差	差	差

表 2 属性的含义

	属性含义
属性 I	思想要求上进,为人正直友善
属性 II	学习勤奋刻苦,成绩优异
属性 III	社会活动能力强,具有很强的领导才能和组织才能
属性 IV	积极参与集体活动,具有很强的集体荣誉感和社会责任感
属性 V	科技创新,勇于探索,具有很强的求知欲,能灵活应用学科知识,并有所拓展

下面可以计算出这 5 个属性的属性重要性:  
首先,利用这 5 个属性对样本数据进行聚类分析,

得到聚类结果为: {1,23,30,31}、{2,28}、{3}、{4,15,29,33}、{5,21}、{6,16}、{7,8}、{9,10,11,22,32}、{12,17,25}、{13,27}、{14,29}、{18,20,34}、{24}、{26}。

接着,计算属性 I 的属性重要性:以属性 II、III、IV、V 对样本数据进行聚类分析,得到聚类结果为: {1,23,30,31}、{2,28}、{3}、{4,15,29,33}、{5,21}、{6,16}、{7,8}、{9,10,11,22,32}、{12,17,25}、{13,27}、{14,29}、{18,20,34}、{24}、{26}。由于  $\text{card}U = 34$ , 且  $\text{cardPOS}_{Q-\{I\}}(Q) = 34$ , 根据属性重要性计算公式可以得到: 属性 I 的属性重要性 =  $\gamma_Q(Q) - \gamma_{Q-\{I\}}(Q) = 1 - \text{cardPOS}_{Q-\{I\}}(Q)/\text{card}U = 0$ 。

利用同样的方法,就可以算出属性 II、III、IV、V 的属性重要性分别为 0.412,0.147,0.471,0。根据 5 个属性的属性重要性大小的分布图,可以看到,属性 II、IV 的重要性相对于属性 I、III、V 是比较大的,且按属性重要性的大小从大到小可以将属性排列为:IV、II、III、I 和 V,因此在属性 II 和属性 III 的属性重要性的值之间取一个值作为阈值  $\epsilon$  的值,这里可以取  $\epsilon$  为 0.15。

这样,利用“阈值  $\epsilon$  为 0.15”对属性重要性进行判断,可以在聚类分析中不考虑属性 I、属性 III 和属性 V,而仅仅利用属性 II、IV 对样本进行聚类分析。得到结果如表 3 所示。

表 3 考虑属性重要性的聚类结果

类别	含义
{1,23,30,31}	各方面表现都比较差的同学,属于班上的后进生
{2,14,28,29}	学习成绩优良,积极参与并组织集体活动,是班上的班干
{3,7,8}	学习成绩优秀,但对班级事务不太关心
{4,15,26,29,33}	学习成绩差,但是集体活动的骨干和积极参与者
{5,21,24}	学习成绩优秀,对集体活动的积极性一般,有一定的组织能力
{6,16}	学习成绩差,对集体活动的积极性一般,有一定的组织能力
{9,10,11,12,17,22,25,32}	各方面表现都比较一般,属于班上的中游分子
{13,27}	各方面表现都很优秀,是班级乃至学校的佼佼者
{18,20,34}	学习成绩一般,对班级事务不太关心

由综合测评成绩和上述聚类结果可以看到,该聚类结果良好,意义明确,符合该班同学的实际情况。

4 结束语

提出了一种基于属性重要性的定性数据聚类的方法。它通过求解出属性集中各个属性的相对于整个属性集的重要性,并设定了阈值  $\epsilon$  对属性重要性进行判断区分,得到影响样品的决定因素,最终在利用粗糙集的等价关系这一核心概念的基础上对样品进行聚类分析。在利用某班同学综合测评的数据进行实际分析中,也得到了较好的聚类结果,进一步地说明了该方法

(下转第 95 页)

号量的值。

参数 cmd 所能指定的操作:IPC\_STAT 获取信号量信息,信息由 arg.buf 返回;IPC\_SET 设置信号量信息,待设置信息保存在 arg.buf 中(在 manpage 中给出了可以设置哪些信息);GETALL 返回所有信号量的值,结果保存在 arg.array 中,参数 semnum 被忽略;GETNCNT 返回等待 semnum 所代表信号量的值增加的进程数,相当于目前有多少进程在等待 semnum 代表的信号量所代表的共享资源;GETPID 返回最后一个对 semnum 所代表信号量执行 semop 操作的进程 ID;GETVAL 返回 semnum 所代表信号量的值;GETZCNT 返回等待 semnum 所代表信号量的值变成 0 的进程数;SETALL 通过 arg.array 更新所有信号量的值;同时更新与本信号集相关的 semid\_ds 结构的 sem\_ctime 成员;SETVAL 设置 semnum 所代表信号量的值为 arg.val;调用返回:调用失败返回 -1,成功返回与 cmd 相关。

## 2 结 论

在 Linux 信号量机制中,内核信号量与用户信号量之间的区别在于内核信号量由系统定义,且只有一

个值,只能实现内核进程简单的同步,使用方法简便。而用户信号量则可以由用户定义一组信号量,用于进程间共享多个资源的同步,并使用 3 个系统调用进行操作,要求操作必须具有整体性,而且各数据结构之间的关系较复杂,通过对内核信号量与用户信号量的深入研究,为进一步应用信号量机制打下了坚实的基础。

### 参考文献:

- [1] Molloy S, Man P H. Scalable Linux Scheduling[R]. CITI Technical Report. [s.l.]:[s.n.],2001:285-295.
- [2] LI Chun-guang, WEN Tao, XU Qiang. Analysis of Linux Kernel Service Mechanism[J]. Fushun Shiyu Xuebao, 1998, 9(18):65-68.
- [3] 王文义,武华北. Linux 中进程间信号通信机制的分析及其应用[J]. 计算机工程与应用,2005(3):108-115.
- [4] 王社国. Linux 信号量通信机制分析与实践[J]. 微机发展, 2002,12(6): 63-66.
- [5] 李 晋,葛敬国. Linux 下互斥机制及其分析[J]. 计算机应用研究,2005(8):72-77.
- [6] 胡希明,毛德操. Linux 内核源代码情景分析[M]. 杭州:浙江大学出版社,2001.
- [7] 倪继利. Linux 内核分析及编程[M]. 北京:电子工业出版社,2005.

(上接第 88 页)

- Machine Vision[M]. 2nd ed. New York: PWS Publishing, 1999.
- [8] Reddy B S, Chatterji B N. An FFT - Based Technique for Translation, Rotation, and Scale Invariant Image Registration [J]. IEEE Trans Image Procession, 1996, 5 (7): 1215 - 1220.
  - [9] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509 - 522.
  - [10] 章夏芬,庄越挺,鲁伟明,等. 根据形状相似性的书法内容

检索[J]. 计算机辅助设计与图形学学报, 2005, 17(11): 2565-2569.

(上接第 91 页)

的可行性,及其对简单利用等价类的概念进行聚类分析方法的改进。这种聚类方法解决了传统聚类方法中对定性数据处理的不足,它无需把定性数据变换为定量数据,并且该方法又不是简单地利用等价关系进行聚类,而是通过分析属性的重要性,并对数据进行约简,得到了意义较为明确的聚类结果,避免了简单利用等价关系聚类而产生结果的不明确性。

### 参考文献:

- [1] Ziarko W. Variable Precision Rough Set Model[J]. Journal of

Computer and System Science, 1993(40):39-59.

- [2] Barbar'A D, Chen P. Using Self-Similarity to Cluster Large Data Sets[J]. Data Mining and Knowledge Discovery, 2003, 7:123-152.
- [3] 李树军,纪宏军. 对应聚类分析与变量选择[J]. 地球物理学进展,2005,20(3):694-697.
- [4] 来升强,朱建平. 数据挖掘中高维定性数据的粗糙集聚类[J]. 统计研究,2005(8):56-60.
- [5] 张文修. 粗糙集理论与方法[M]. 北京:科学出版社,2001.
- [6] 朱建平. 数据挖掘的统计方法及实践[M]. 北京:中国统计出版社,2005.