文章编号: 1000-5862(2012)05-0472-05

基于偏最小二乘法的信息粒降维及聚类研究

聂　斌[1], 王　卓[2], 杜建强[1*], 余日跃[3], 徐国良[3], 朱明峰[1]

(1.江西中医学院计算机学院　江西 南昌 330004; 2.南昌大学软件学院　江西 南昌 330047;
3.江西中医学院药学院 江西 南昌 330004)

摘要: 利用偏最小二乘法, 按最大相关性提取出最强解释能力的信息, 实现信息粒降维; 通过欧氏距离聚类, 获取不同层次的信息粒; 根据需求选择合适的粒度进行分析. 实验结果表明: 文中所提方法是可行有效的.

关键词: 偏最小二乘法; 信息粒; 降维; 聚类
中图分类号: TP 39　　文献标志码: A

0　引言

粒计算[1-2]的 3 大基础理论包括: L. A. Zadeh 的模糊集理论[3], E. Pawlak 的粗糙集理论[4]和张钹、张铃的商空间理论[5]. 另外, 在粒计算发展过程中, 梁吉业等[6]提出了一种知识粒度的公理化定义; 苗夺谦等[7]讨论了知识粗糙性与信息熵之间的关系, 并对知识粗糙性的单调性进行了证明; 钱宇华等[8]引入了具有直观知识含量特征的组合熵和组合粒度来度量信息系统中的不确定性和粒度大小, 并讨论了组合熵与组合粒度之间的关系; 王国胤等[9]讨论了分层递阶商空间的信息熵序列随知识粒度变化的规律.

粒计算是人工智能研究领域中的一种新理念和新方法, 它覆盖了所有与粒度相关的理论、方法和技术, 是复杂问题求解、海量数据挖掘、不确定性信息处理的有效工具. 粒计算的主要思想是通过选择合适的粒度来寻找问题的一种较好的、近似的解决方案, 从而降低问题求解的复杂度. 目前, 粒计算模型的种类有模糊集模型、粗糙集模型、商空间理论模型、基于覆盖的粒计算模型、模糊粗糙集模型和粗糙模糊集模型等.

随着社会的发展, 形成了大量庞杂海量的高维数据, 并迫切要求人们有从中获取有效知识能力. 研究者尝试了一些解决方法, 如张燕平等[10]讨论了基于多维数据模型的粒计算方法.

聚类分析是一种常用的信息粒化方法[11], 它将给定数据集中的对象有机地进行聚集, 形成互不相交的划分结果. 聚类分析通常是针对具体问题进行处理. 针对不同的数据类型, 其相应的聚类方法也不同, 根据数据取值值域不同有数值型和区间型; 根据聚类模型有划分型和层次型; 根据聚类算法所基于的策略和特征分有基于密度法、基于图法、基于谱法等.

偏最小二乘法[12]是由 S. Wold 和 C. Albano 等在 1983 年提出, 偏最小二乘回归开辟了一种有效的技术途径, 它利用对系统中的数据信息进行分解和筛选的方式, 提取对因变量的解释性最强的综合变量, 辨识系统中的信息与噪声. 在化学、医药等领域应用十分广泛, 在降维方面[13-14]独具优势.

1　基于偏最小二乘法的信息粒计算

1.1　知识粒度表示

定义 **1**[15]　设 U 是非空有限集, 称为论域. 任何子集 $X \subseteq U$, 称为 U 中的一个概念. U 中的一簇概念, 称为关于 U 的知识. 记 $A=\{X_1, X_2, \cdots, X_n\}$, 若满足

(1) $X_i \subseteq U, X_i \neq \Phi$,

(2) $X_i \bigcap X_j = \Phi, i \neq j, i, j = 1, 2, \cdots, n$,

(3) $\bigcup_{i=1}^{n} X_i = U$,

则称 A 为 U 的划分.

收稿日期: 2012-07-14
基金项目: 国家“ 973 ”计划(2010CB530602, 2010CB530603), 国家“ 863 ”计划(2012AA02A609), 国家自然科学基金(81160424), 江西省自然科学基金(2010GZY0174, 20122BAB205083)和江西省教育厅课题(GJJ11541)资助项目.
作者简介: 杜建强(1968-), 男, 江西南昌人, 教授, 主要从事数据挖掘和图像处理的研究.

定义 **2**　称二元序对 $AS=(U,R)$ 是一个近似空间, 其中 U 是非空有限集, 称为论域, R 是 U 上的等价关系, 也称为 U 上的不可区分关系. $U/R=\{[u]_R \mid u\in U\}$ 表示 R 是 U 上的划分, 它由 U 中的每个对象 u 的 R-等价类 $[u]_R$ 组成[15-16].

定义 **3**　知识库可以形式地定义为序对 $K=(U, \Re)$, 其中 U 为论域, R 为 U 上的等价关系簇. 称等价关系 $R\in\Re$ 为知识, 称 R 生成的等价类 $[u]_R$ 为基本知识颗粒, 称商集 $U/R=\{[u]_R \mid u\in U\}$ 为论域 U 的 R-粒划分[15-16].

定义 **4**　设 $K=(U,\Re)$ 是一个知识库, $R\in\Re$ 为论域 U 上的等价关系, 称为知识[17]. 知识 $R\in\Re$ 的粒度, 记为 $GD(R)$, 定义为

$$GD(R)=\frac{|R|}{|U\times U|}=\frac{|R|}{|U|^2}, \tag{1}$$

$|R|$ 表示 $R\subseteq U\times U$ 的基数.

知识 R 的粒度可以表示它的分辨能力, 对 $\forall u,v\in U$, 当 $(u,v)\in R$ 时, 表明对象 u,v 在 R 下不可分辨, 属于 R 的同一个等价关类; 否则, 它们可分辨, 属于不同的等价类. $GD(R)$ 越大, 表明不可分辨的可能性越大, 分辨能力越弱; 否则分辨能力越强.

1.2　偏最小二乘粒降维、聚类及层次转换

设有 q 个决策属性 $\{y_1,\cdots,y_q\}$ 和 p 条件属性 $\{x_1,\cdots,x_p\}$. 原始数据具有 n 条决策规则(成分剂量-药效对), 由此构成了决策属性与条件属性的知识库 $K=(U,\Re)$, 最初生成 n 个等价类 $[u]_R$, $GD(R)$ 粒度最大, 含各原始决策规则信息, 分辨能最差.

令 $X=\{x_1,\cdots,x_p\}$ 和 $Y=\{y_1,\cdots,y_q\}$, 偏最小二乘回归[18]分别在 X 与 Y 中提取出成分 t_1 和 u_1 满足下列 2 个要求: (1) t_1 和 u_1 分别应尽可能大地携带 X 和 Y 中的变异信息; (2) t_1 与 u_1 的相关程度能够达到最大.

如果要 t_1, u_1 能分别很好地代表 X 与 Y 中的数据变异信息, Var 代表方差, r 代表相关关系, Cov 代表协方差, 根据主成分分析原理, 有

$$Var(u_1)\to\max,$$
$$Var(t_1)\to\max.$$

另一方面, 由于回归建模的需要, 又要求 t_1 对 u_1 有很大的解释能力, 有典型相关分析的思路, t_1 与 u_1 的相关度应达到最大值, 即 $r(t_1, u_1)\to\max$, 因此, 综合起来, 在偏最小二乘回归中, 要求 t_1 与 u_1 的协方差达到最大, 即

$$\mathrm{cov}(t_1,u_1)=\sqrt{Var(t_1)Var(u_1)}\times r(t_1,u_1)\to\max.$$

这 2 个要求表明, t_1 和 u_1 应尽可能好的代表, X 和 Y, 同时自变量的成分 t_1 对因变量的成分 u_1 又有最强的解释能力.

在第一个成分 t_1 和 u_1 被提取后, 偏最小二乘回归分别实施 X 对 t_1 的回归以及 Y 对 u_1 的回归. 如果回归方程已经达到满意的精度, 则算法终止; 否则, 将利用 X 被 t_1 解释后的残余信息以及 Y 被 t_1 解释后的残余信息进行第 2 轮的成分提取. 如此往复, 直到能达到一个较满意的精度为止. 若最终对 X 共提取了 m 个成分 t_1, …, t_m, 偏最小二乘回归将通过实施 y_k 对 t_1, …, t_m, 的回归, 然后再表达成 y_k 关于条件属性 x_1, …, x_m 的回归方程, k=1, 2, …, q.

通常情况下, $m\leqslant p$, 可实现降维效果. 至此, $GD(R)$ 粒经降维后, 保留原有用信息"浓缩粒". $GD(R)$ 粒度最大, 含各原始决策规则信息.

根据欧氏距离对各决策属性 y_k 聚类. 按需要, 取不同层次粒. 首先计算所有 y_k 之间欧氏距离, 对距离值排序, 取最短的 2 个点的中点; 然后, 以此中点作为起点, 求其与其它 $k-2$ 个点之间距离, 取最近的点加入, 并算出 3 个点的质心; 依此类推, 就是聚类过程以及可以求出聚类结果.

Step1: 降维后的 y_k, 拥有 m 个条件属性;

Step2: $\min(D_{a,b})=\sqrt{\sum\left(a[i]-b[i]\right)^2}$ 其中 a, b 属于 y_k 中的 2 个点, (i = 1, 2, …, m);

Step3: $[u]_j=\{a,b\}$, $j\leqslant m$;

Step4: 重复 Step2, Step3, 将新的点加入等价类, 直到所有点计算完为止.

2　实验

在药效研究中, 给定了 12 个成分指标, 作为条件属性信息, 即总蒽醌芦荟大黄素、总蒽醌大黄素、总蒽醌大黄酸、总蒽醌大黄酚、结合蒽醌大黄素甲醚、结合蒽醌芦荟大黄素、结合蒽醌大黄素、结合蒽醌大黄酸、结合蒽醌大黄酚、厚朴酚酸大黄素甲醚、厚朴酚酸和厚朴酚、厚朴酚; 2 个药效指标, 作为决策属性信息, 即通便时间和肠梗阻时间. 共 10 组不同配方得到的不同药效数据, 课题目的是研究

这些药效物质在不同配伍配比和药效之间的关系. 在此, 将每组配方及其得到的药效数据作为一个信息粒, 通过粒降维和分层聚类, 分析各组方之间关系及量效关系. 在图 1 中, R_2Y 表示主成分对决策属性 Y 的解释能力, Q_2 表示主成分对条件属性 X 的解释能力, 图 1 反映 X 从 12 维降到了 7 维, 根据实际需要, 还可继续降维. 下文各图中的 Group 和 G 代表的一个等价类, 或称信息粒. 图 2 表明聚类的层次和过程, 图 3、图 6、图 9 表明 2 个等价类, $[u]_1=\{1,2,3,4,5,6\}$, $[u]_2=\{7,8,9,10\}$; 图 4、图 7、图 10 表明 3 个等价类, $[u]_1=\{1,2,3,4,5,6\}$, $[u]_2=\{7,10\}$, $[u]_3=\{8,9\}$, 粒度变小, 分辨度增加; 图 5、图 8、图 11 表明 4 个等价类, $[u]_1=\{1,2,6\}$, $[u]_2=\{3,4,5\}$, $[u]_3=\{7,10\}$, $[u]_4=\{8,9\}$, 粒度进一步减小, 分辨度进一步增加. 同理, 可向更小粒转化. 并可实现粒度大小之间相互转换. 经医药学专家确认, 并且此结果对于药效物质的配伍及配比同药效间的关系研究有重要意义.

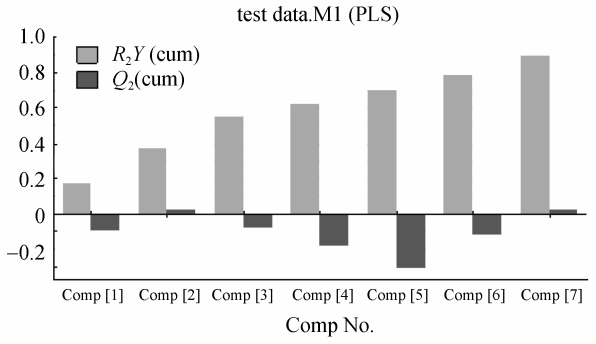


图 1 主成分图

	1	2	3	4	5
1	7				
2	10	3.328 83			
3	8		24.794 5		
4	9	3.233 71			
5	2				
6	1		2.207 06		34.322 9
7	6	1.675 59			
8	4			7.952 1	
9	3		1.487 63		
10	5	0.693 656			

图 2 粒聚类过程及层次

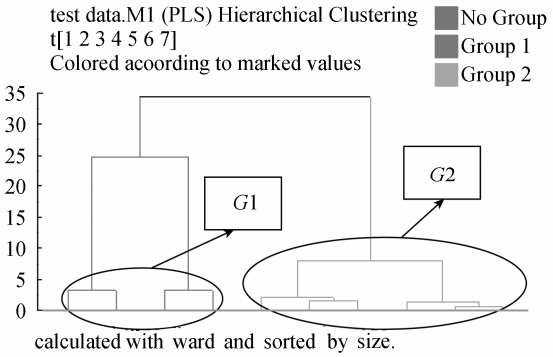


图 3 粒聚类 2 个等价类树形图

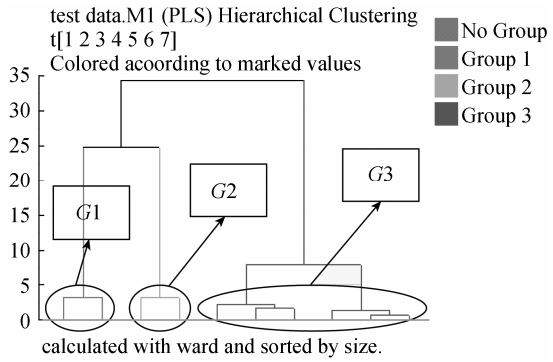


图 4 粒聚类 3 个等价类树形图

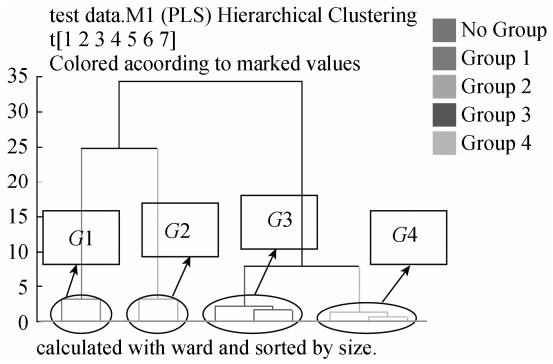


图 5 粒聚类 4 个等价类树形图

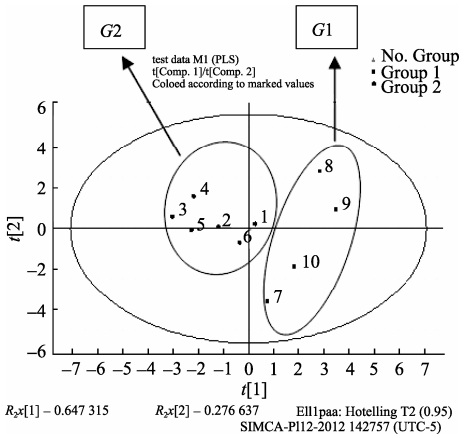


图 6 粒聚类 2 个等价类 2 维图

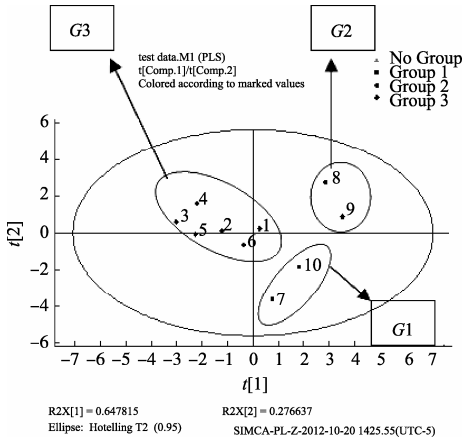


图 7　粒聚类 3 个等价类 2 维图

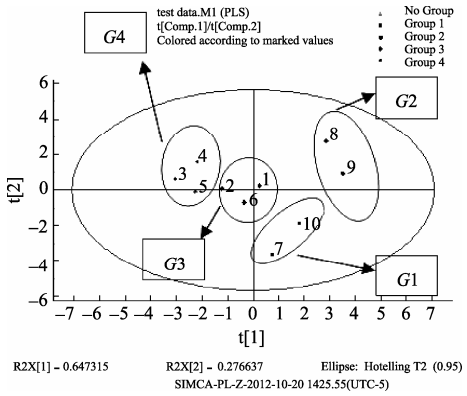
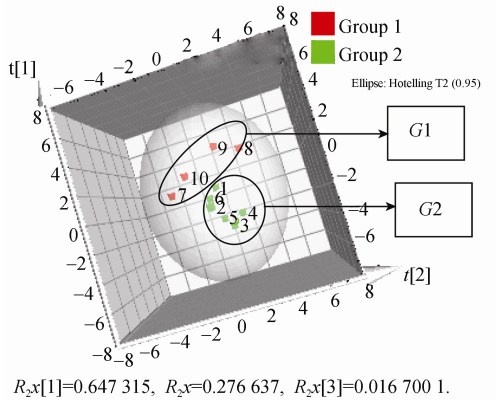
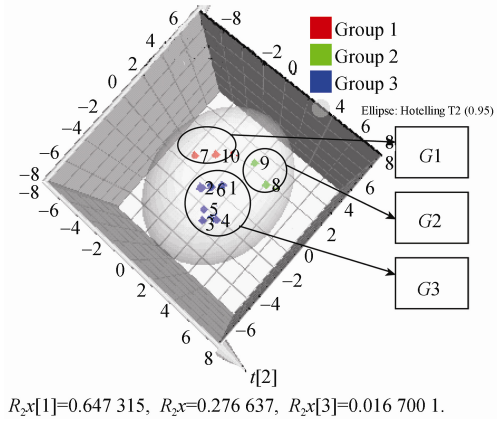


图 8　粒聚类 4 个等价类 2 维图



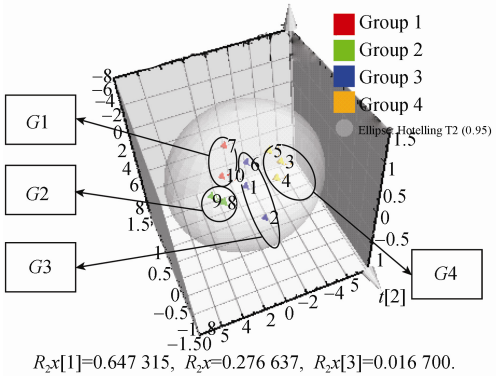
$R_2x[1]$=0.647 315, R_2x=0.276 637, $R_2x[3]$=0.016 700 1.

图 9　粒聚类 2 个等价类三维图



$R_2x[1]$=0.647 315, R_2x=0.276 637, $R_2x[3]$=0.016 700 1.

图 10　粒聚类 3 个等价类 3 维图



$R_2x[1]$=0.647 315, R_2x=0.276 637, $R_2x[3]$=0.016 700.

图 11　粒聚类 4 个等价类 3 维图

3　结束语

粒计算的研究涉及到其基础研究和应用研究，文中探索性地使用偏最小二乘法引入粒计算，实现信息粒降维、聚类及粒层转换．经实验表明，此法是可行有效的．

4　参考文献

[1] 苗夺谦，王国胤，刘清，等．粒计算：过去、现在与展望 [M]．北京：科学出版社，2007.

[2] 王国胤，张清华，马希骜，等．知识不确定性问题的粒计算模型 [J]．软件学报，2011，22(4)：676-694.

[3] Zadeh L A. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338-353.

[4] Pawlak Z. Rough sets [J]. Int Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[5] 张铃，张钹．问题求解理论及应用：商空间粒度计算理论及应用 [M]．2 版．北京：清华大学出版社，2007.

[6] Liang Jiye, Qian Yuhua. Axiomatic approach of knowledge granulation in information system [J]. Lecture Notes in Artificial Intelligence, 2006, 4304: 1074-1078.

[7] Miao Dtioqian, Wang Jue. On the relationships between information entropy and roughness of knowledge in rough set theory [J]. PR & AI, 1998, 11(1): 34-40.

[8] Qian Yuhua, Liang Jiye. Combination entropy & combination granulation in rough set theory [J]. Int Journal of Uncertainty Fuzzinessand Knowledge-Based Systems, 2008, 16(2): 179-193.

[9] Wang Guoyin, Zhang Qinghua. Uncertainty of rough sets in different knowledge granularities [J]. Chinese Journal of Computers, 2008, 31(9): 1588-1598.

[10] 张燕平．商空间与粒计算：结构化问题求解理论与方法 [M]．北京：科学出版社，2010.

[11] 钱宇华．复杂数据的粒化机理与数据建模 [D]．太原：山西大

学, 2011.

[12] Esbensen K H, Wold S, Simca, et al. Space and Unfold: the ways towards regionalized principal components analysis and subconstrained N-way decomposition—with geological illustrations [C]. In: O. J. Christie(Ed.), Proc Nord Symp Appl Statist. Stavanger, 1983.

[13] 郭建校. 改进的高维非线性偏最小二乘回归模型及应用 [M]. 北京: 中国物质出版社, 2010.

[14] 曾雪强, 李国正. 基于偏最小二乘降维的分类模型比较 [J]. 山东大学学报: 工学版, 2010, 40(5): 41-47.

[15] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法 [M]. 北京: 科学出版社, 2001.

[16] Pawlak Z. Rough sets: theoretical aspects of reasoning about data [M]. Dordrecht: Kluwer Academc Publishers, 1991.

[17] 苗夺谦, 范世栋. 知识的粒度计算及其应用 [J]. 系统工程理论与实践, 2002, 22(1): 48-56.

[18] 王惠文. 偏最小二乘回归方法及其应用 [M]. 北京: 国防工业出版社, 1999.

The Research for Information Granule Reduction and Cluster Based on the Partial Least Squares

NIE Bin[1], WANG Zhuo[2], DU Jian-qiang[1*], YU Ri-yue[3], XU Guo-liang[3], ZHU Ming-feng[1]

(1. School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang Jiangxi 330004, China;
2. Software School of Nanchang University, Nanchang Jiangxi 330047, China;
3. College of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang Jiangxi 330004, China)

Abstract: According to the biggest correlation and the best interpret ability to achieve reducing dimensions based on the partial least squares. To cluster use Euclidean distance, and to obtain different information granular. Analyze the suitable granular in the actual field. It was proved to be feasible and effective after tested with a database.

Key words: partial least squares; information granular; reducing dimensions; cluster

(责任编辑: 冉小晓)