分析方法为基础的粗糙集容差关系和PCA不完备信息系统

王峻慧
(重庆师范大学经济与管理学院 重庆 400047)

摘 要 如何在不完备信息系统中利用已有的数据进行属性约简、获取属性权重一直是粗糙集理论中的一个研究重点。提出了一种在不完备信息系统中基于相容关系和PCA(主成份分析)相结合的粗糙集分析方法,这种方法属于启发式的数据分析方法。给出了这种分析方法的算法和一个算例。结果表明,该方法能够较准确地确定决策表的核,并用定量分析的方法获取属性权重。

关键词 相容关系,主成份分析(PCA),粗糙集,属性约简
中图法分类号 TP393 **文献标识码** A

Rough Sets Tolerance Relation Based on Analyzing Method and PCA Incomplete Information System

WANG Jun-hui
(School of Economics & Management, Chongqing Normal University, Chongqing 400047, China)

Abstract How to acquire the optimal attributes reduction and weighs of attributes in incomplete information system is a hot topic for researching in rough sets theory recently. A rough set analysis method based on tolerance matrix and principle component analyze was provided in incomplete information system, which is one of heuristic data analyzing methods. Corresponding algorithm and an example were also given in this article. The method provided in this paper was compared with other methods. The result shows the method provided in this paper can find efficiently the core of an incomplete information system and acquire weighs of attributes through quantitative analyses.

Keywords Tolerance matrix, Principle component analyze, Rough sets, Attributes reduction

1 引言

不完备信息系统中的知识获取是智能信息处理的重要环节。粗糙集理论RS(Rough Sets)是由波兰华沙理工大学的Z. Pawlak教授等一批科学家提出的一种研究不完整、不精确或者是模糊知识的一种组织和分析方法。利用粗糙集理论进行数据挖掘,最重要的内容之一就是基于粗糙集的属性约简。根据区分矩阵和区分函数及布尔运算可以求出属性集的所有最小约简,但是由于求所有最小约简是NP问题[1],因此一般只适合小数据集。实际应用中大都采用启发式算法。

目前常见的属性约简方法是启发式算法。很多学者从不同角度提出了多种处理方法。黄兵等人在相容关系的基础上通过定义属性的信息量来度量信息系统中属性的重要性和属性的相对重要性,以此来设计决策表的属性约简算法[2]。Kryszkiewicz M提出相容关系后[3],相关的研究不断深入。王国胤等人在相容关系基础上提出了限制相容关系[4],放宽了相似关系的条件而严格控制了相容关系的条件,解决了相似和相容关系的一些局限。文献[5]在相容关系的基础上定义了广义决策矩阵来求出粗糙集的属性约简。还有学者提出了概率差别矩阵的概率来获得属性约简[6]。信息熵也被用于不完备信息系统的属性约简中。付昂、王国胤等人阐明了不完备信息系统中的知识约简在信息观和代数观下的差异,在此基础上提出了一种基于信息熵的不完备信息系统中的知识约简算法[7]。本文提出一种在不完备信息系统中基于相容关系和PCA(主成份分析)方法的属性约简方法。这种分析方法也属于启发式的属性约简方法,它以主成份分析的结果作为计算属性重要度的依据。

2 相关理论

2.1 相关定义

定义1 信息系统 $S=(U,A,V,f)$,U 表示全体对象的集合,是一个非空有限集;A 是非空有限集,表示全体属性的集合;V 表示属性取值的集合,f 定义一个信息函数,即 $f:U\times AT\to V$。称 U 上的二元关系 T 是相容的,若 T 满足:

(1)自反性,即 $\forall u\in U, uTu$ 成立;

(2)对称性,即若 uTv,则 $vTu, v,u\in U$。

定义2 设 T 是论域 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 上的相容关系,则 T 对应的相容矩阵定义如下:

$$M_T=(r_{ij})_{|U|\times|U|},\text{其中}(r_{ij})=\begin{cases}1 & u_iTu_j\\ 0 & \text{其它}\end{cases}$$

定义3 有不完备决策表 $S=(U,C\cup D,V,f)$,其中 C 为条件属性,$D\{d\}$ 为决策属性,$C\cap D=\emptyset$。对任意的 $u\in U$,$f(u,d)\neq *$,定义 u 关于 $a\in C$ 的广义决策函数如下:

$$\partial_a(u_i)=\{f(u_i,d)\mid f(u_i,a)=f(v,a)\vee f(u_i,a)=*\vee f$$

到稿日期:2009-02-27 返修日期:2009-05-22

王峻慧(1966-),女,学士,副教授,高级会计师,主要研究方向为会计、粗糙集分析等。

$(v,a)=*, v\in U\}$

定义 4 有不完备信息系统 S,A 为属性集，定义其条件属性 $a\in A$ 确定的相容关系 T_a 如下：

$uT_av\Leftrightarrow f(v,d)\in\partial_a(u)\quad u,v\in U$

2.2 主成份分析

主成份分析方法(PCA)是一种评价属性贡献率的统计分析方法，它设法将原来指标重新组合成一组新的互相无关的几个综合指标来代替原来指标，同时根据实际需要从中可取几个较少的综合指标，尽可能多地反映原来指标的信息。因此可以将主成份分析方法应用于属性排序，将排序结果用作属性选择的一个参考，以此作为属性选择的一种启发式规则。

在决策问题的研究中，常常会遇到影响此问题的很多属性，这些属性可能有很多并且有一定的相关性，因此从中综合出一些主要的指标是非常必要的。若有指标系列$(x_1,x_2,\cdots,x_p)$，取它们的某些线性组合 F，要使这些线性组合 F 包含尽可能多的信息，即 $var(F)$ 最大，这样得到的 F 记为 F_1，然后再找 F_2，F_1 与 F_2 无关，以此类推，便找到了一组综合变量 $F_1,F_2,\cdots,F_k$，这组变量基本包含了原来变量的所有信息。

设指标集$(x_1,x_2,\cdots,x_p)$的协方差矩阵 Σ 的特征根为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$，分别为对应的主成份的特征值，则 $var(F_i)=\lambda_i$，称 F_1 为第一主成份，F_2 为第二主成份，……，F_p 为第 p 主成份。

定义 5 称 $\lambda_1/\sum_{i=1}^{p}\lambda_i$ 为第一主成份的贡献率。称 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$ 为前 m 个主成份的累计贡献率。

这个值越大，表明第一主成分综合 $x_1,\cdots,x_p$ 信息的能力越强。前 k 个主成分的累计贡献率达到 85%，表明取前 k 个主成分基本包含了全部测量指标所具有的信息。

主成份确定属性重要性的一般过程是：先按累积方差贡献率不低于某一阈值的原则确定主成份的个数，即前 k 个主成份，然后以每个主成份的贡献率作为权数，最后各属性的重要度以各主成份贡献率和各属性的主成分系数加权综合来确定。第一主成份能够最大限度地反映评价对象之间的差异，是概括评价指标差异信息的最佳线性函数，尤其是当第一成份的方差贡献率大于 60%时。

3 基于相容关系和 PCA 的不完备信息系统属性约简

不完备信息系统中，基于相容关系和 PCA 的粗糙集分析方法的基本思想是：通过相容关系和相容矩阵确定不完备决策表的核，再通过某种数据补全算法将决策表中不完备数据补全，使之成为一个信息完备的决策表。接着利用 PCA 方法来确定属性的重要度，以此作为启发式算法中非核属性加入的依据。这种属性约简的方法属于启发式算法。

设有不完备决策表 $S=(U,A,V,f)$，$A=C\cup D$。其中 C 为条件属性集，D 为决策属性集。具体算法如下：

(1)计算 S 的相容矩阵 M_A 和去掉某个属性 a 后的相容矩阵 $M_{A-\{a\}}$，$a\in A$；

(2)根据相容关系和相容矩阵计算出 S 的核 $Core_D(C)$；

(3)用 ConditionedMeanCompleter 算法对决策表 S 进行完备化处理，得到 S'；

(4)用 PCA 方法进行主成份分析，计算出主成份 $F_1,F_2,\cdots,F_m$ 及各主成份的贡献率；

(5) 给定一个阈值 ω，$0\leqslant\omega\leqslant1$，根据前 m 个主成份的累计贡献率 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i\geqslant\omega$ 来确定需要的主成份的个数；

(6)根据各属性的得分确定属性的重要度，属性 a_i 得分由式(1)得出：

$$Score(a_i)=r(F_1)*x_{a_i}+\cdots+r(F_m)*x_{a_i}\qquad(1)$$

其中，$r(F_i)$表示主成份 F_i 的方差贡献率，x_{a_i} 表示对象 a_i 在主成份 F_1 中对应的系数；

(7)在 $Core_D(C)$中按照属性重要性将所有非核属性按属性重要性依次加入，直至得出需要的属性约简结果。

4 算例分析

有不完备决策表 S[3](如表 1 所列)，论域 $U=\{1,2,3,4,5,6\}$，属性集 $A=C\cup D$，其中条件属性集 $C=\{P,M,S,X\}$，决策属性集为 $D=\{A\}$。

表 1 不完备决策表 S

U	P	M	S	X	A
1	High	High	Full	Low	Good
2	Low	*	Full	Low	Good
3	*	*	Compact	High	Poor
4	High	*	Full	High	Good
5	*	*	Full	High	Excel
6	Low	High	Full	*	Good

根据广义决策函数的定义计算出不完备信息表中每个对象的广义决策值：

$\partial_A=\{\{Good\},\{Good\},\{Poor\},\{Good,Excel\},\{Good,Excel\},\{Good,Excel\}\}$。

根据∂_A 可以计算出不完备决策表 S 的相容矩阵 M_A。

$$M_A=\begin{bmatrix}1&1&0&1&0&1\\1&1&0&1&0&1\\0&0&1&0&0&0\\1&1&0&1&1&1\\1&1&0&1&1&1\\1&1&0&1&1&1\end{bmatrix}$$

当从属性集 AT 中单独去掉各个属性时，可以依次得到对应的相容矩阵 $M_{A-\{P\}}$，$M_{A-\{M\}}$，$M_{A-\{S\}}$ 和 $M_{A-\{X\}}$。由 $M_{A-\{P\}}$，$M_{A-\{M\}}$，$M_{A-\{S\}}$ 和 $M_{A-\{X\}}$ 的结果可以得出：

$M_A=M_{A-\{P\}}=M_{A-\{M\}}\neq M_{A-\{S\}}$

$M_A=M_{A-\{P\}}=M_{A-\{M\}}\neq M_{A-\{X\}}$

因此，属性 S 和属性 X 即为决策表 S 的核。

对决策表 S 中各对象的属性值进行数字化处理，再利用有条件平均化填充算法(ConditionedMeanCompleter 算法)对决策表做完备化处理，得到决策表 S'，如表 2 所列。

表 2 ConditionedMeanCompleter 算法处理后的决策表 S'

U	P	M	S	X	A
1	0	0	3	2	1
2	0	0	4	2	2
3	2	2	4	1	3
4	1	2	4	1	3
5	2	2	4	2	3
6	1	2	4	1	3

对于决策表 S'，利用 PCA 方法计算出各属性的重要度。再和前面计算出的决策表的核$\{P,X\}$相结合，利用 PCA 的主

(下转第 287 页)

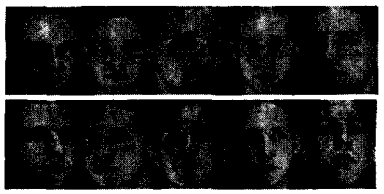

图1 ORL人脸图像的样本

图2 YALE人脸图像的样本

表1 算法在ORL人脸库中的识别结果(%)

训练样本数	PCA	LPP	FELG
3	78.2	82.1	85.2
4	83.7	87.6	91.3
5	86.8	89.8	93.9
6	89.1	93.3	96.7
7	92.4	94.1	97.9

表2 算法在Yale人脸库中的识别结果(%)

训练样本数	PCA	LPP	FELG
3	70.9	73.1	79.6
4	73.4	75.6	82.8
5	73.9	76.2	84.5
6	75.3	78.5	86.9
7	76.8	79.7	87.6

从实验结果中可以看出，各类算法在ORL人脸库中的结果比较理想，这是因为其人脸的表情、光照等因素变化较少，适合提取关键特征。从实验数据可以看出，LPP算法的识别率高于PCA，这是因为PCA只是较好地在低维空间中展现了原数据，而LPP考虑了人脸的局部信息，人脸空间更可能存在于非线性子空间中，LPP在一定程度上能解决这类问题。GELG不仅从全局和局部角度出发，而且也对数据的类别信息进行了综合，因此识别效果较理想。

结束语 本文提出了一种基于局部特性和全局特性的新的特征提取算法，既考虑了人脸的全局信息又兼顾了局部信息；构造出了更加准确、鲁棒的识别系统，并且考虑了类别信息，因而具有良好的识别性能。实验结果表明，算法可以有效提取出特征，并使不同类尽可能分离，因而可以提高识别率。

参考文献

[1] Turk M, Pentland A. Eigenfaces for recognition[J]. Cognitive Neurosci, 1991, 3(1): 71-86

[2] Belhumeur P N, Kriegman D J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19: 711-720

[3] Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290: 2319-2323

[4] Rowies S, Saul L. Nonliear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290: 2323-2326

[5] Belkin M, Niyogo P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396

[6] He X, Niyogi P. Locality preserving projections[M]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2004

[7] Zhao Haitao. Combining labeled and unlabeled data with graph embedding[J]. Neurocomputing, 2006, 69: 2385-2389

(上接第269页)

成份特征根、方差贡献率以及方差累计贡献率(如表3所列)和主成分因子得分系数表(如表4所列)，最终获得所需要的属性约简。

表3 主成分特征根及方差贡献率

成份	特征根λ	方差贡献率的比重(%)	累积方差贡献率(%)
1	2.829	70.728	70.728
2	0.613	15.331	86.059
3	0.500	12.500	98.559
4	0.058	1.441	100.000

表4 主成份因子得分系数矩阵

主成份	1	2	3	4
P	0.303	0.616	0.655	−2.258
M	0.341	0.012	−0.378	3.226
S	0.274	0.341	1.195	−0.217
X	−0.266	1.065	0.000	1.339

当选择特征值大于1的主成份时，累计方差贡献率只有70%，因此以85%作为阈值选出前两个主成份 F_1，F_2，4个主成份各自的方差贡献率为权重，根据式(1)计算出各属性的得分分别为0.309，0.243，0.246，0.025。在非核属性中，属性 P 的得分最高，因此选择属性 P 加入到核属性集合中，可以得到决策表的一个约简$\{S,X,P\}$，可以求出 $M_A=M_{\{S,X,P\}}$。

结束语 当数据量很大时，应用粗糙集算法进行数据的分析处理时间复杂度较高，因此如何能在有限的时间内求出最佳约简一直是科研工作者的一个研究重点。启发式方法是近年来研究的热点，它将粗糙集方法和现有的一些比较成熟的、效率较高的统计分析方法相结合，从而获得属性约简和相应的规则。

在本文提出的基于相容关系和主成份分析方法相结合的粗糙集分析方法中，如何进一步降低求解相容矩阵的时间复杂度是下一步要继续研究的问题。

参考文献

[1] Wong S K, Ziarko W. On optimal decision rules in decision tables[J]. Bulletin of Polish Academy of Science, 1985, 33: 693-696

[2] 黄兵，周献中，张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践，2005(4)：55-60

[3] Kryszkiewicz M. Rough set approach to incomplete information system[J]. Information Sciences, 1998(112): 39-49

[4] 王国胤. Rough集理论在不完备信息系统中的扩充[J]. 计算机研究与发展，2002，39(10)：1238-1243

[5] 张腾飞，王锡淮，肖建梅. 不完备信息系统的一种属性相对约简算法[J]. 计算机工程，2007(5)：184-186

[6] 闫德勤. 概率差别矩阵与不完备信息系统属性约简[J]. 计算机科学，2005(8)：164-166

[7] 付昂，王国胤，胡军. 基于信息熵的不完备信息系统属性约简算法[J]. 重庆邮电大学学报：自然科学版，2008，32(5)：586-592