基于粗糙集理论的疾病综合诊断

杨　广[1]，夏　波[1]，王晓龙[2]，张保华[2]，周圣武[1]

（1. 中国矿业大学理学院，江苏徐州　221008；
2. 中国矿业大学机电工程学院，江苏徐州　221008）

摘要：对于医生来说疾病的综合诊断是非常重要的，综合诊断实质上是权系数的确定问题。本文提出了一种基于粗糙集理论的权系数确定方法，将权系数确定问题转化为粗糙集中属性重要程度评价问题，建立了关于组合预测方法的关系数据模型并计算出组合预测模型的权系数。该方法克服了传统权系数确定方法的主观性，避免了线性或非线性极值问题的数值计算，使得组合预测方法更具客观性。本文案例证明了此方法的有效性。

关键词：应用数学；疾病诊断；组合预测；粗糙集
中图分类号：O15　　**文献标识码**：A　　**文章编号**：1674-2850(2008)07-0515-6

Combination diagnosis of the disease based on rough set theory

YANG Guang[1], XIA Bo[1], WANG Xiaolong[2], ZHANG Baohua[2], ZHOU Shengwu[1]

(1. *College of Science, China University of Mining and Technology, Xuzhou, Jiangsu*　221008;
2. *College of Mechanical engineering, China University of Mining and Technology, Xuzhou, Jiangsu*　221008)

Abstract: Combination diagnosis of the disease is very important to the doctor, but the diagnosis is materially equal to the determination of weighting coefficient. A method of determining weighting coefficient based on rough set theory is showed in this paper. Determining weighting coefficient is translated into estimating significance of attributes among rough set. A relation data model about combination forecast is established and the weighting coefficients are computed. The proposed approach overcomes the subjectivity of traditional determination to weighting coefficient, avoids computing linear or nonlinear extremum problem and makes combination forecast more objective. The validity of the proposed approach is verified with a case.

Key words: applied mathematics; disease diagnosis; combination forecast; rough set

0　引言

人们到医院就诊时，通常要化验一些指标来协助医生的诊断。诊断就诊人员是否患肾炎时通常要化验人体内各种元素含量。如附表所示，其中 1～30 号病例是确诊为肾炎病人的化验结果；31～60 号病例是确定为健康人的结果。需要解决的问题是根据附表中的数据特征，确定哪些指标是影响人们患肾炎的关键或主要因素，以便减少化验的指标。

1　基本假设与符号说明

1.1　问题假设

1）假设在各种元素含量中的极少数超高值为偶然性数值，是不合理的，在数值计算中可以不予考虑或者以普通值代替；

作者简介：杨广（1986—），男，本科生，主要研究方向：应用数学
通信联系人：周圣武，教授，主要研究方向：应用数学. E-mail：zswcumt@163.com

2）假设健康人体内的元素含量不会出现大的波动；

3）假设病例只存在患病与不患病两种状态，不存在介于两者之间的状态。

1.2 符号说明

n 为模型中样本总个数；

u_i 为第 i 个样本；

p 为需要观测的指标个数；

c_i 为需要观测的第 i 个指标；

c_{ij} 为第 i 个样本第 j 个指标的数据；

F_i 为每一个主成分的线性加权值；

A 为病症情况判断值：得病（赋值0），健康（赋值1）。

2 问题的分析和模型的建立

2.1 粗糙集理论的相关概念

粗糙集理论（rough set theory）[1~2]是一种处理模糊和不确定知识的数学工具，能处理定性和定量因素。该理论的主要特点是不需提供问题所需处理的数据集合以外的任何先验信息，仅根据观测数据删除冗余信息，分析不完整知识的程度——粗糙度、属性间的依赖性和重要性、生成分类或决策规则等。本文应用该理论来确定患肾炎病人各项病因指标中的权系数，将权系数确定问题转化为粗糙集中属性重要性评价问题，利用粗糙集理论中的知识依赖性和属性重要性的评价方法，计算出各项病因指标权系数。该方法不需要建立解析式的数学模型，完全由数据驱动来确定各优化目标的权系数，克服了传统权系数确定方法的主观性，避免了线性或非线性极值问题的数值计算，使综合目标函数更具有客观性，也更符合具体的规划方案。

1）知识和知识系统。研究对象构成的集合是非空有限集，称为论域U；R 是上一个等价关系称为知识。R 产生等价类 $[x]_R=\{y \mid R: x\rightarrow y, y\in U\}$，等价类集合 $U/R=\{[x]_R \mid x\in U\}$ 称为 U 的知识系统。$\underline{R}(x)=\{x \mid [x]_R\subseteq X, x\in U\}$，称为知识系统 U/R 下集合 X 的下逼近，是 X 中可用知识系统 U/R 的知识精确表示部分；$\overline{R}(x)=\{x \mid [x]_R\cap X\neq\varnothing, x\in U\}$，称为在知识系统 U/R 下集合 X 的上逼近，说明可用知识系统 U/R 的知识 $\overline{R}X$ 粗糙地表示 X。令 $\alpha_R(X)=card[\underline{R}(x)]/card[\overline{R}(x)]$ 反映知识系统 U/R 对集合 X 的可表示程度。

2）知识的依赖度。设 R、Q 均是 U 上的等价关系，为说明与 Q 之间的不确定关系，定义知识 R 对知识 Q 的依赖程度 $\gamma_Q(R)$ 为

$$\gamma_Q(R)=\sum card[Q([x]_R)]/card(U) \tag{1}$$

显然 $0\leqslant\gamma_Q(R)\leqslant 1$，$\gamma_Q(R)$ 的数值大小反映了知识 R 对知识 Q 的依赖程度。

3）属性的重要性。研究对象集合 U 是论域，条件属性集为 C，决策属性集合为 D，条件属性集 C 中的条件属性 c_i 对于决策属性 D 的重要程度定义为

$$\rho_D(c_i)=\gamma_c(D)-\gamma_{c-[c_i]}(D) \tag{2}$$

显然 $\rho_D(c_i)$ 越大，属性 c_i 的重要性越高。

2.2 粗糙集理论确定多目标综合模型权系数的方法

多目标综合模型为

$$A=\sum_{i=1}^{m} w_i F_i \tag{3}$$

多目标综合模型由 m 个最优化目标函数综合构成，w_i 是第 i 个最优化目标函数的权系数；权系数满足条件 $w_1+w_2+\cdots+w_m=1$。

确定权系数首先建立关系数据模型。将各个最优化目标函数作为条件属性，可以表示为集合 $C=\{F_1,F_2,\cdots,F_m\}$；将综合最优化目标函 A 视为决策属性，表示为集合 $D=\{A\}$，不同的优化方案下得到的综合目标优化值 u_k 为研究对象 A 的一条信息，显然 $u_k=(F_{1k},F_{2k},\cdots,F_{mk};A_k)$，那么就构成了论域 $U=\{u_1,u_2,\cdots,u_n\}$，也称为样本集合。这时研究对象 u_k 的属性为 $F_i(u_k)=c_{ik}$，$A(u_k)=A_k$，其中（$i=1,2,\cdots,m$；$k=1,2,\cdots,n$）。如表 1 所示，由 u_k 构成的二维信息表就是多目表综合最优化模型的关系数据模型。

权系数的确定步骤如下：

1）计算知识 R_D 对知识 R_C 的依赖程度，即计算最优化目标函数集合 C 对预测指标 A 的依赖程度

$$\gamma_{R_C}(R_D)=\sum card[\underline{R_C}([A]_{RD})]/card(U) \quad (4)$$

2）对每个最优化目标函数 F_i，计算知识 R_D 对知识 $R_{C-\{c_i\}}$ 的依赖程度

$$\gamma_{R_{C-\{c_i\}}}(R_D)=\sum card[\underline{R_{C-\{c_i\}}}([A]_{RD})]/card(U) \quad (5)$$

3）计算第 i 最优化目标的重要性

$$\rho_D(c_i)=\gamma_{R_C}(R_D)-\gamma_{R_{C-\{c_i\}}}(R_D) \quad (6)$$

4）第 i 最优化目标的权系数为

$$w_i=\rho_D(c_i)/\sum_{j=1}^{m}\rho_D(c_i),\quad i=1,2,\cdots,m \quad (7)$$

表 1 多目标综合优化模型

Tab.1 Multi-objective optimization model

对象	F_1	F_2	…	F_m	A
u_1	c_{11}	c_{12}	…	c_{1m}	A_1
u_2	c_{12}	c_{22}	…	c_{2m}	A_2
⋮	⋮	⋮	⋮	⋮	⋮
u_k	c_{k2}	c_{k2}	…	c_{km}	A_k

首先，通过极差变化法对 1～60 号病人的数据进行进一步地量化，选用模糊值{含量值 1、2、3、4}来表示。在此考虑划分四个等级，如表 2 所示。

当数值落入区间 [0，0.25）时，认为含量值为 1；

当数值落入区间 [0.25，0.5）时，认为含量值为 2；

当数值落入区间 [0.5，0.75）时，认为含量值为 3；

当数值落入区间 [0.75，1] 时，认为含量值为 4。

$U/A=\{\{1,2,3,\cdots,30\};\{31,32,33,\cdots,60\}\}$，共两大类，得病、不得病。

按 D 分类

$\underline{R_D}([A]_{RD})=\{\{20,21,24,27\},\{17,26,28,29\},\{34,44,50,58\},\{51\},\{31,35,36,37,42,44,46,47,48,51,52,53,56,58\},\{45,50,55,57\},\{40,49\},\{39,45,48,49,50,51,52,57,58\},\{11,20,21,29\},\{11,19\}\}$

$\gamma_c=\underline{R_D}([A]_{RD})/U=49/60=0.816\ 667$

按 $D-$Ca 分类

$\underline{R_{D-Ca}}([A]_{RD})=\{\{20,21,24,27\},\{17,26,28,29\},$

表 2 病例的优化值

Tab.2 The optimize value of the case

病号	优化值 D							
	Zn	Cu	Fe	Ca	Mg	K	Na	A
1	3	2	1	1	1	4	3	0
2	3	2	1	1	1	4	3	0
3	4	2	1	1	2	4	3	0
4	3	2	1	1	1	4	2	0
5	4	3	1	1	2	4	4	0
6	3	1	1	1	2	4	4	0
7	4	2	1	1	1	4	4	0
8	3	2	1	1	1	4	3	0
9	3	1	1	1	1	4	4	0
10	3	2	1	1	1	4	3	0

注：这里仅给出前十个病例号的优化值

{34,44,50,58},{51},{39,45,48,49,50,51,52,57,58},{11,20,21,29},{11,19}}

$\gamma_{c-Ca} = \underline{R_{D-Ca}}([A]_{RD})/U = 0.466\ 667$

元素 Ca 对结果 A 的重要程度为

$$\sigma(\mathrm{Ca}) = \gamma_c - \gamma_{c-Ca} = 0.816\ 667 - 0.466\ 667 = 0.35$$

同理可得

$$\sigma(\mathrm{Zn}) = 0.133\ 34 \qquad \sigma(\mathrm{Cu}) = 0 \qquad \sigma(\mathrm{Fe}) = 0.083\ 334$$
$$\sigma(\mathrm{Mg}) = 0.15 \qquad \sigma(\mathrm{K}) = 0.066\ 667 \qquad \sigma(\mathrm{Na}) = 0.033\ 34$$

由此可见：$\sigma(\mathrm{Ca}) > \sigma(\mathrm{Mg}) > \sigma(\mathrm{Zn}) > \sigma(\mathrm{Fe}) > \sigma(\mathrm{K}) > \sigma(\mathrm{Na}) > \sigma(\mathrm{Cu})$，即 Ca 的影响性最大，Mg 次之，Cu 和 Na 的影响性最小。将这四种元素在正常人和患者之间的含量用图表示，也说明了这一性质。如图 1 所示。

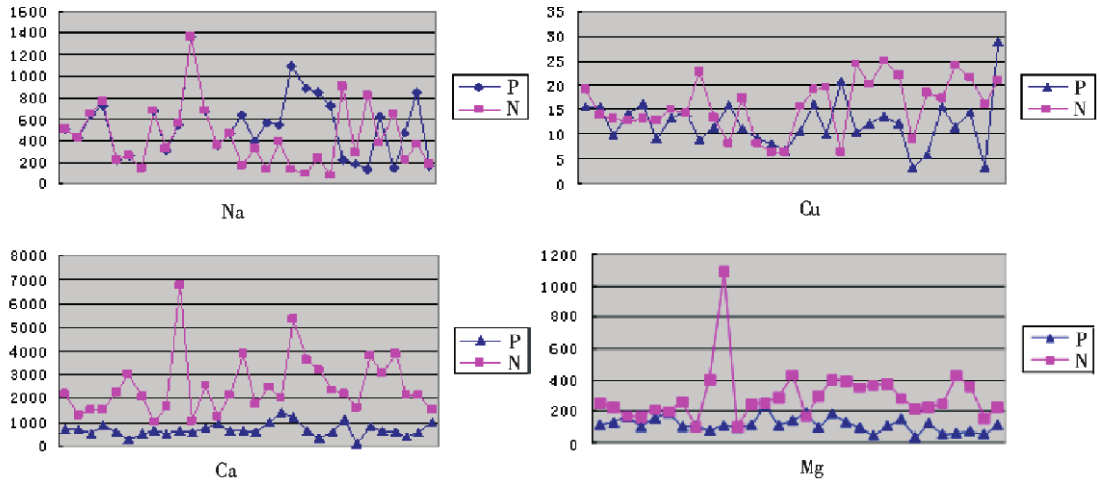

图 1　四种元素在正常人和患者之间的含量

Fig. 1　The content of four elements between normal and patients

同时，利用主成份分析法[3~4]对此问题进行分析，得出的结果为：Ca，Mg，Zn，K，Fe 在健康人中的含量高于病人的含量，为高优指标；Cu，Na 在健康人中的含量低于病人的含量，为低优指标。

3　模型的评价

本文就疾病的综合诊断问题进行了实证性的分析，建立了基于权重分析的粗糙集模型，确定了在多种因素影响下疾病的主导因素和次要因素，减少了多种因素的化验指标，节约了医疗资源。同时本文还采用了主成份分析法对本例进行分析，得出的结果比较相近，证明了该方法的有效性。

[参考文献] (References)

[1] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001.
WANG G Y. Rough set theory and knowledge gain[M]. Xi-an: Xi-an Jiaotong University Press, 2001. (in Chinese)

[2] 钟波. 组合预测中基于粗糙集理论的权值确定方法[J]. 重庆大学学报，2002 (7)：127～129.
ZHONG B. Determination to weighting coefficient of combination forecast based on rough set theory[J]. Chongqing University Journal, 2002(7): 127～129. (in Chinese)

[3] 李艳双. 主成分分析法在多指标综合评价方法中的应用[J]. 河北工业大学学报，1999 (1)：28～51.

LI Y S. Application of primary component analysis in the methods of comprehensive evaluation for many indexes[J]. Journal of Hebei University of Technology, 1999(1): 28～51. (in Chinese)

[4] 胡永宏，贺思辉. 综合评价方法[M]. 北京：科学出版社，2000.
HU Y H, HE S H. Quality synthetic evaluation method[M]. Beijing: Science Press, 2000. (in Chinese)

附表

确诊病例的化验结果
The diagnosed case's lab results

病例号	Zn	Cu	Fe	Ca	Mg	K	Na
1	166	15.8	24.5	700	112	179	513
2	185	15.7	31.5	701	125	184	427
3	193	9.80	25.9	541	163	128	642
4	159	14.2	39.7	896	99.2	239	726
5	226	16.2	23.8	606	152	70.3	218
6	171	9.29	9.29	307	187	45.5	257
7	201	13.3	26.6	551	101	49.4	141
8	147	14.5	30.0	659	102	154	680
9	172	8.85	7.86	551	75.7	98.4	318
10	156	11.5	32.5	639	107	103	552
11	132	15.9	17.7	578	92.4	1 314	1 372
12	182	11.3	11.3	767	111	264	672
13	186	9.26	37.1	958	233	73.0	347
14	162	8.23	27.1	625	108	62.4	465
15	150	6.63	21.0	627	140	179	639
16	159	10.7	11.7	612	190	98.5	390
17	117	16.1	7.04	988	95.5	136	572
18	181	10.1	4.04	1 437	184	101	542
19	146	20.7	23.8	1 232	128	150	1 092
20	42.3	10.3	9.70	629	93.7	439	888
21	28.2	12.4	53.1	370	44.1	454	852
22	154	13.8	53.3	621	105	160	723
23	179	12.2	17.9	1 139	150	45.2	218
24	13.5	3.36	16.8	135	32.6	51.6	182
25	175	5.84	24.9	807	123	55.6	126
26	113	15.8	47.3	626	53.6	168	627
27	50.5	11.6	6.30	608	58.9	58.9	139
28	78.6	14.6	9.70	421	70.8	133	464
29	90.0	3.27	8.17	622	52.3	770	852
30	178	28.8	32.4	992	112	70.2	169
31	213	19.1	36.2	2 220	249	40.0	168
32	170	13.9	29.8	1 285	226	47.9	330
33	162	13.2	19.8	1 521	166	36.2	133
34	203	13.0	90.8	1 544	162	98.90	394
35	167	13.1	14.1	2 278	212	46.3	134
36	164	12.9	18.6	2 993	197	36.3	94.5
37	167	15.0	27.0	2 056	260	64.6	237
38	158	14.4	37.0	1 025	101	44.6	72.5

续表

39	133	22.8	31.0	1 633	401	180	899
40	156	135	322	6 747	1 090	228	810
41	169	8.00	308	1 068	99.1	53.0	289
42	247	17.3	8.65	2 554	241	77.9	373
43	166	8.10	62.8	1 233	252	134	649
44	209	6.43	86.9	2 157	288	74.0	219
45	182	6.49	61.7	3 870	432	143	367
46	235	15.6	23.4	1 806	166	68.8	188
47	173	19.1	17.0	2 497	295	65.8	287
48	151	19.7	64.2	2 031	403	182	874
49	191	65.4	35.0	5 361	392	137	688
50	223	24.4	86.0	3 603	353	97.7	479
51	221	20.1	155	3 172	368	150	739
52	217	25.0	28.2	2 343	373	110	494
53	164	22.2	35.5	2 212	281	153	549
54	173	8.99	36.0	1 624	216	103	257
55	202	18.6	17.7	3 785	225	31.0	67.3
56	182	17.3	24.8	3 073	246	50.7	109
57	211	24.0	17.0	3 836	428	73.5	351
58	246	21.5	93.2	2 112	354	71.7	195
59	164	16.1	38.0	2 135	152	64.3	240
60	179	21.0	35.0	1 560	226	47.9	330