粗糙集在经济分析中的应用

王 伟[1],高 亮[1],吴 涛[1,2]
(1.安徽大学 数学与计算科学学院,安徽 合肥 230039;
2.安徽大学 智能计算与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:随着计算机技术的快速发展,各种数据急剧增长,如何从这些海量数据中提取有用的信息成为了一个很现实而且重要的问题。在粗糙集理论中,规则的生成是很重要的,从规则中可以得出数据中一些内在规律,这对发现和分析决策表的本质有很大的帮助。该文利用安徽省近几年的经济数据对安徽省经济的运行进行分析,并从中得到一些经济规则,从规则分析中得到了与现实一致的结果。

关键词:粗糙集;约简;规则;经济指标
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)04-0158-03

Application of Rough Set in Economic Analysis

WANG Wei[1], GAO Liang[1], WU Tao[1,2]
(1.School of Mathematics and Computational Science of Anhui University, Hefei 230039, China;
2.Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: As the computer technology rapid development, each kind of data grows suddenly. How to extract useful information from these massive data has become a very real and important issues. It is very important that how to get the rule in the theory of rough set, it may obtain some inherent laws from the rule. It is helpful to analyse the decision table. Based on the economic data of Anhui Province in recent years, analyses the system of Anhui economics by the rule got. The result get from the rules is consistent with the reality.

Key words: rough set; reduct; rule; economic indicators

0 引 言

粗糙集理论是波兰数学家 Z. Pawlak 于 1982 年提出的一种数据分析理论[1],是一种刻画不完整性和不确定的模糊的数学工具。其主要思想就是在保持分类能力的前提下,通过知识约简导出问题的决策或分类规则,从而从中发现隐含的知识,揭示潜在的规律。其最大的优点是无需提供除问题所需处理的数据之外的任何先验信息,完全由数据本身出发来解决问题。由于具有很强的定性分析能力,能够有效地表达不确定或不精确的知识,善于从数据中获取知识,并能利用不确定、不完整的经验知识进行推理等,因此粗糙集理论已成为信息科学最为活跃的研究领域之一。同时该理论还在机器学习、规则生成、决策分析、智能控制、地震预报、语音识别等领域获得了广泛应用[2],特别是在数据挖掘领域,获得了巨大成功,业已成为粒度计算研究领域的主要方向之一。文中结合安徽省经济数据和粗糙集的方法,分析影响安徽省经济发展的关键因素,并从中得到决策规则,以此作为经济决策的依据。

收稿日期:2007-07-04
基金项目:国家自然科学基金(60475017,60675031);安徽省自然科学基金(050420208);安徽省高等学校省级自然科学研究项目(2006KJ244B);安徽大学学术创新团队和安徽大学人才队伍建设经费
作者简介:王 伟(1984-),男,河南信阳人,硕士研究生,研究方向为智能计算与信息处理;吴 涛,博士,副教授,硕士生导师,主要从事机器学习、智能计算及其应用的研究。

1 粗糙集基本概念

粗糙集的基本概念[3]:

定义 1:设一信息系统,即一四元组 $S=(U,A,V,f)$ 是一个知识表达系统,其中:

U:对象的非空有限集合,称为论域;

A:属性的非空有限集合;

$V=\bigcup_{a\in A}V_a$ 其中 V_a 是属性 a 的值域;

$f:U\times A\to V$ 是一信息函数,它为每个对象的每个属性赋予一个信息值,即:

$\forall a\in A, x\in U, f(x,a)\in V_a$

知识表达系统也称为信息系统,通常用 $S=(U,A)$ 来代替 $S=(U,A,V,f)$,其中 $A=C\cup D$,C 称

为条件属性,D 称为决策属性。

定义2:$K=(U,R)$ 为一知识库,对集合 $X\subseteq U$ 和一个等价关系 $R\in \mathrm{ind}(K)$,定义子集:

$$\underline{R}(X)=\bigcup\{Y\in U/R/Y\subseteq X\}=\{x\in U/[x]_R\subseteq X\}$$

$$\overline{R}(X)=\bigcup\{Y\in U/R/Y\cap X\neq\varnothing\}=\{x\in U/[x]_R\cap X\neq\varnothing\}$$

称 $\underline{R}(X)$ 为 X 的下近似集,$\overline{R}(X)$ 为 X 的上近似集,并称集合:

$$\mathrm{bn}_R(X)=\overline{R}(X)-\underline{R}(X)$$

为 X 的 R 边界,

$$\mathrm{pos}_R(X)=\underline{R}(X)$$

为 X 的正域,

$$\mathrm{neg}_R(X)=U-\overline{R}(X)$$

为 X 的负域。

定义3:R 为一等价关系族,$r\in R$,如果

$$\mathrm{ind}(R)=\mathrm{ind}(R-\{r\})$$

称 r 为 R 中不必要的,否则为必要的。若 $\forall r\in R$ 都是 R 中必要的,则称 R 为独立的。设 $Q\subseteq P$,如果 Q 是独立的,且 $\mathrm{ind}(Q)=\mathrm{ind}(P)$,则称 Q 为 P 的一个约简。一个决策表可以有多个约简。

定义4:设 $Y\subseteq D$ 是 $(U,C\cup D)$ 中的一些决策属性子集,则 Y 关于 X 的支持度为:

$$\mathrm{spt}_X(Y)=/S_X(Y)/\,/U/$$

其中 $S_X(Y)$ 是 Y 关于 X 的支持子集。

定义5:设 $S=(U,A)$ 为一决策表,令 X_i 和 Y_j 分别代表 U/C 与 U/D 中的各个等价类,$\mathrm{des}(X_i)$ 表示对 X_i 的描述,$\mathrm{des}(Y_j)$ 表示对 Y_j 的描述。规则定义为:

$$r_{ij}:\mathrm{des}(X_i)\rightarrow\mathrm{des}(Y_j),X_i\cap Y_j\neq\varnothing$$

规则的确定性因子 $\mu(X_i,Y_j)=/X_i\cap Y_j/\,/\,/X_i/$,$0<\mu(X_i,Y_j)\leqslant 1$,当 $\mu(X_i,Y_j)=1$ 时,r_{ij} 是确定的,否则为不确定的。

在决策表中,最重要的就是决策规则的产生,因为可以根据决策规则从决策表中得到一些重要的信息,并可以由决策规则来指导未来的一些经济决策[4]。

在决策规则产生之前还要对决策表进行一些预处理,比如数据中有缺失,要对缺失数据进行处理,有些属性是连续的,还要对其离散化等。然后对决策表进行约简,最后得到决策规则。

2　基于 Rough 集的经济决策分析

在经济数据中,通常以国内生产总值(GDP)作为衡量经济发展快慢的因素,故将其作为决策属性,其他的一些指标作为条件属性。文中采用安徽省 1980～1998 年全省一些经济指标进行分析[5],找出各经济指标在安徽省经济发展中所扮演的角色,并从中得到一些关键规则。影响国内生产总值增长的因素有很多,限于篇幅,主要分析以下经济指标 $C=\{$第一产业 $C1$(亿元),第二产业 $C2$(亿元),第三产业 $C3$(亿元),粮食总产量 $C4$(万吨),固定资产投资 $C5$(亿元),进出口总额 $C6$(万美元),消费零售总额 $C7$(亿元),旅游总收入 $C8$(万美元),能源生产总量 $C9$(万吨标准煤),科研教育总投入 $C10$(万元)$\}$ 对安徽省国内生产总值 D(亿元) 的影响。具体数据见表1。

表1　年间增幅(%)

序号	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	D
1980	64.68	50.06	26.14	1453.9	17.3	3982	62.5	104.3	1733.03	1274	140.88
1981	88.44	52.48	29.59	1818.5	15.7	11201	68.04	115.2	1712.18	1608	170.51
1982	89.97	59.66	37.39	1933	34	16617	75.77	107.9	1725.19	1942	187.2
1983	96.55	73.2	45.93	2010.5	46.2	19416	83.2	110	1817.11	2276	215.6
1984	116.26	92.83	56.65	2202.5	62.2	29113	97.6	116.3	1979.27	2610	265.74
1985	140.97	117.84	72.43	2168	80.7	43013	119.9	117	2086.98	2944	331.24
1986	155.91	137.57	89.28	2371.9	103.5	48960	142.5	128	2169.64	3278	382.76
1987	176.35	157.74	108.26	2428.7	117.2	62603	165.6	116.1	2072.69	4080	442.73
1988	210.53	200.95	135.46	2310.3	137.9	70373	208.3	167	2194.12	8812	546.97
1989	225.39	227.85	163.01	2424.67	114.4	70082	224.4	209	2239.73	8924	616.26
1990	246.17	251.48	160.37	2520.13	123	73668	226.5	620	2306	8729	658.03
1991	190.48	280.28	192.84	1749.15	137.3	85420	247.5	1035.31	2221.68	6956	663.67
1992	230.51	333	237.65	2341.92	214.9	110439	285.04	1418.51	2419.74	14265	801.23
1993	284.06	494.85	290.93	2595.9	321	128776	343.8	1779.76	2572.99	18453	1069.92
1994	336.72	744.16	407.59	2361.24	399.5	184853	453.2	2481.3	2910.86	25393	1488.56
1995	581.24	908.87	513.47	2652.74	532.5	230769	586.5	4435.9	3188.35	31852	2003.66
1996	665.44	1067.25	606.56	2700.26	614.3	275111	727.1	5800.4	3646.93	51336	2339.34
1997	732.37	1214.82	722.76	2802.7	687.3	311605	859.8	8205.7	3509.49	72039	2670.08
1998	739.7	1253.53	812.22	2590.5	729	312012	924.8	7036.05	3278.11	70960	2805.56

首先对数据预处理:由于1980～1984年间的科研教育总投入的数据缺失,为此用最小二乘法对科研教育总投入($C10$)的数据进行拟合,并将数据补齐,完备后的决策表见表1。用每年比上年的增幅作为研究对象,由于每年的增幅为连续属性,用平均增长幅度作为每个属性的间断点,以反映各属性增长的快慢,1为快,0为慢。离散结果见表2。

表2 离散后的各指标

序号	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$	$C9$	$C10$	D
1	1	0	0	1	0	1	0	0	0	0	1
2	0	0	1	1	1	1	0	0	0	0	0
3	0	1	1	0	1	0	0	0	1	0	0
4	1	1	1	1	1	1	1	0	1	0	1
5	1	1	1	0	1	1	1	0	1	0	1
6	0	0	1	1	1	0	1	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	0	0	0	1	1	1	1	1
9	0	0	0	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	1	0	0	0
12	1	0	1	1	1	0	0	1	1	1	1
13	1	1	1	1	1	0	1	0	1	1	1
14	1	1	1	0	0	1	1	1	1	1	1
15	1	1	1	1	1	0	1	1	1	0	1
16	0	0	0	0	0	0	1	0	1	1	0
17	0	0	0	0	0	0	1	1	0	1	0
18	0	0	0	0	0	0	0	0	0	0	0

对离散决策表进行计算:

可以算出 $\mathrm{spt}_C(D)=1$,即条件属性 C 对决策属性 D 的支持度是1,这说明该10个经济指标(条件属性)能够很好地反映国内生产总值(决策属性)的状况,这也说明了指标选取的合理性和可行性。

文中用Rose2软件对决策表进行约简和提取最小规则。共得到12个约简,其中最小约简为{$C1$},在最小相对支持度为75% 的情况下,共得到3个最小规则:

(1) $(C1=0)\Rightarrow(D=0)$;

(2) $(C1=1)\Rightarrow(D=1)$;

(3) $(C2=1)\wedge(C7=1)\Rightarrow(D=1)$;

因为:

$U/D=\{(1,4,5,8,12,13,14,15),(2,3,6,7,9,10,11,16,17,18)\}$;

$U/C1=\{(1,4,5,8,12,13,14,15),(2,3,6,7,9,10,11,16,17,18)\}$;

$U/(C2,C7)=\{(1,2,7,9,10,11,12,18),(3),(4,5,8,13,14,15),(6,16,17)\}$;

按照定义5算出各规则的确定性因子:

$\mu(1)=10/10=1;\mu(2)=8/8=1;\mu(3)=6/6=1$;

即该3个规则都是确定的。

从以上规则可以看出:

(1) 第一产业($C1$)对安徽省的国内生产总值的影响是比较大的。从历史上看,安徽省是一农业大省,地处中部农业发展区,其第一产业作为该省的基础,因而在GDP中占很大的分量,尤其在"六五"和"七五"时期(1981～1990),工业开始发展,更需农业的支持。

(2) 从第三个规则可以看出第二产业($C2$)和消费零售($C7$)增长快,则GDP增长也快。因为在20世纪80年代,改革开放开始,外资的引进加快了工业发展的进程,其对国内生产总值的影响开始增加。但在九十年代中后期,经济开始"软着陆",人民消费疲软,急需扩大内需,这也使消费零售对GDP的影响显著增加。

由以上可以看出,由粗糙集得到的规则所说明的问题与现实基本吻合,这也说明了该方法的有效性,并且由得到的规则还可以对未来几年的经济政策做出相应的调整,比如,消费疲软,如何从宏观调控中扩大内需等。

3 结 语

经济发展需要一个衡量标准,这对经济发展具有规范性和引导的作用,量化的指标具有直观性且容易操作,成为促进社会经济发展的有效手段和途径[6]。文中主要是利用粗糙集这一工具对经济数据进行加工处理,从中得到一些有用信息,得到的规则对经济发展有个参考作用,从文中的实例也可看出所得到的结果与现实情况基本保持一致,说明了该方法的可行性。

参考文献:

[1] Pawlak Z. Rough sets[J]. International journal of computer and information science, 1982, 11(5): 342－356.

[2] Pawlak Z. Rough Set—Theoretical Aspects of Reasoning about data [M]. Dordrecht: Kluwer Academic Publishers, 1991.

[3] 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: 1－40.

[4] 林 毅, 梁家荣. 基于粗糙集的规则的挖掘[J]. 微机发展, 2004, 14(9): 92－93.

[5] 安徽省政府. 安徽五十年[M]. 北京:中国统计出版社, 1999.

[6] 顾爱华, 陈玉林. 地方政府促进知识经济发展的指标模型及职能分析[J]. 东北大学学报, 2006, 8(1): 38－41.