

基于粗糙集的石油安全预警规则提取

单雪红^{1,2}, 吴涛^{1,3}, 徐文婷^{1,3}

(1. 安徽大学 数学科学学院, 安徽 合肥 230039;

2. 宿州学院 数学系, 安徽 宿州 234000;

3. 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:粗糙集理论是分析不确定系统的一种有力的工具。运用粗糙集的理论和方法, 结合我国历年的石油数据, 建立石油安全预警指标模型。利用 Rose 软件, 在保持分类能力不变的前提下, 对该数据的各项指标进行属性约简, 再对该约简的属性值进行约简, 然后提取最小决策规则, 挖掘其中隐含的有用信息, 得出影响我国石油安全预警的重要因素。根据得出的决策规则, 对我国未来几年内的石油预测数据进行分析, 得出我国石油安全属于重警区, 需加强防范的结论。

关键词:粗糙集; 属性约简; 决策规则; 石油安全

中图分类号: TP18

文献标识码: A

文章编号: 1673-629X(2009)01-0251-03

Rules Set of China Oil Security Early Warning Based on Rough Set

SHAN Xue-hong^{1,2}, WU Tao^{1,3}, XU Wen-ting^{1,3}

(1. School of Mathematical Science, Anhui Univ., Hefei 230039, China;

2. Department of Mathematics, Suzhou College, Suzhou 234000, China;

3. Ministry of Education Key Lab. of Intelligent Computing & Signal Processing, Hefei 230039, China)

Abstract: Rough set theory is a powerful tool of uncertainty analysis system. In this paper, using rough set theory and methods, establish a model of early warning of China oil security, according to the data of China oil. By Rose, under the premise of keeping the same classification capacity, the attribute and attribute value of data are reduced, and then, the smallest decision rules are extracted. Get the important factors which affect our country's oil security. The results show that China oil security will be dangerous within the next few years. So must strengthen measure.

Key words: rough set; attribute reduction; decision rule; oil security

0 引言

粗糙集的概念是由波兰数学家 Z. Pawlak^[1] 于 1982 年提出来的, 它的主要思想就是保持分类能力不变的前提下, 通过知识约简找到核值, 导出问题的决策或分类规则, 从中发现隐含的知识, 揭示潜在的规律, 是一种新型的处理模糊和不确定问题的数学工具。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, 已成

为信息分析和数据挖掘的重要方法。

文中利用粗糙集方法, 对我国石油安全的各项指标^[2] 进行分析, 提出决策规则, 并进行预警研究和实例分析。

1 粗糙集的基本概念与方法

1.1 粗糙集的基本概念

定义 1 一个四元组 $S = (U, A, V, f)$ 为一信息系统, 其中 U 是对象的集合, 也称为论域, $A = C \cup D$ 是属性集合, 子集 C 和 D 分别成为条件属性和决策属性, $V = \bigcup_{a \in A} V_a$, V_a 是属性 a 的值域, $f: U \times A \rightarrow V$ 是一信息函数, 当 $D \neq \emptyset$ 时, 称 S 为决策系统, 简称决策表。

定义 2 对于 $R \subseteq A$, $IND(R) = \{(x, y) \mid (x, y) \in U^2, \forall a \in R, f_a(x) = f_a(y)\}$ 称为由 R 决定的不可分辨关系。显然 $IND(R)$ 是一个等价关系。

收稿日期: 2008-05-11

基金项目: 中国博士后基金面上项目 (20070411028); 973 计划 (2004CB318108, 2007BC311003); 国家自然科学基金 (60675031); 安徽省高等学校省级自然科学基金项目 (KJ2008B093, KJ200845ZC); 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: 单雪红 (1980-), 女, 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, 研究方向为机器学习、智能计算及其应用。

记 $U \mid \text{IND}(R) = \{[x]_R \mid x \in U\}$, 简记为 $U \mid R$ 。其中 $[x]_R = \{y \mid xRy, y \in U\}$ 。

定义 3^[3,4] 设 $X \subseteq U$, R 是 U 上的一等价关系 $R(X) = \bigcup \{Y_i \mid Y_i \in U \mid R \text{ 且 } Y_i \subseteq X\}$, 称为 X 的下近似集, $\bar{R}(X) = \bigcup \{Y_i \mid Y_i \in U \mid R \text{ 且 } Y_i \cap X \neq \emptyset\}$, 称为 X 的上近似集。 $R(X)$ 是根据知识 R , U 中所有一定能归入集合 X 的元素构成的集合, 又称为 X 的 R 正域, 记为 $\text{POS}_R(X)$ 。 $\bar{R}(X)$ 是根据知识 R , U 中所有一定能和可能归入集合 X 的元素构成的集合。

定义 4^[4] 对于决策表 $S = (U, C \cup D, V, f)$, 决策规则 $A \rightarrow B$ 的可信度 $\text{CF}(A \rightarrow B)$ 定义为 $\text{CF}(A \rightarrow B) = \frac{|X \cap Y|}{|X|}$, 其中 $X = \{x \mid x \in U \wedge A_x\}$, $Y = \{x \mid x \in U \wedge B_x\}$, A_x 表示实例 x 的条件属性值满足公式 A , B_x 表示实例 x 的决策属性值满足公式 B , 即集合 X 是条件属性值满足公式 A 的实例集合, 集合 Y 是决策属性值满足公式 B 的实例集合。

定义 5^[4] 为了表示决策规则的适用性及可信度, 决策规则 $A \rightarrow B$ 的覆盖率 $\text{cover}(A \rightarrow B)$ 定义为 $\text{cover}(A \rightarrow B) = \frac{|X \cap Y|}{|Y|}$, 支持度定义为 $\text{sup}(A \rightarrow B) = \frac{|X \cap Y|}{|U|}$ 。

1.2 属性及属性值约简

在智能数据分析研究中, 原始决策表信息系统中的知识并不是同等重要的, 甚至其中条件属性是冗余的, 在保持知识库中知识不丢失的前提下消除知识库中冗余的属性和属性值, 称为知识的简化。基于粗糙集的数据挖掘, 关键在于决策表的约简过程, 其关键的两个环节就是属性约简和属性值约简。

1.2.1 属性约简

设 U 为一个论域, P 和 Q 为定义在 U 上的两个等价关系, Q 的 P 正域记为 $\text{POS}_P(Q) = \bigcup_{x \in U \mid Q} P(X)$ 。

定义 6^[3] 设 U 为一个论域, P 和 Q 为定义在 U 上的两个等价关系, 若 $\text{POS}_P(Q) = \text{POS}_{P \setminus \{r\}}(Q)$, 则称 r 为 P 中相对于 Q 可省略的, 否则称 r 为 P 中相对于 Q 不可省略的。若 P 中的每一个 r 都是 P 中不可省略的, 则称 P 为 Q 独立的。若 P 的独立子集 $S \subset P$, 有 $\text{POS}_S(Q) = \text{POS}_P(Q)$, 则称 S 为 P 的 Q 约简。

1.2.2 属性值约简

属性约简只是在一定程度上去掉了决策表中冗余属性但是还没有充分去掉决策表中的冗余信息。

决策表 $S = (U, C \cup D, V, f)$, 对 $\forall x \in U$ 用 d_x 表示决策规则, 即 $d_x = \text{des}([x]_C) \Rightarrow \text{des}([x]_D)$, $d_x(a) = a(x)$, $a \in C \cup D$ 。

定义 7^[3] 一个相容知识表达系统 S , 对决策规

则 d_x 有 $[x]_C \subseteq [x]_D$, 若 $\forall r \in C$, 有 $[x]_{C \setminus \{r\}} \not\subseteq [x]_D$, 则 r 为 d_x 的核值属性, r 为 d_x 中不可省略的。若 $[x]_{C \setminus \{r\}} \subseteq [x]_D$, 则 r 不是 d_x 的核值属性, r 为 d_x 中可省略的。

2 基于 Rough Set 的石油安全预警数据的规则提取步骤

第一步: 建立指标集, 进行数据收集, 进行数据处理。影响中国石油安全的因素很多, 结合目前我国石油安全的状况, 选取了可靠性较强, 同时兼顾灵敏性与完整性的 1993-2005 年 10 个石油安全指标进行分析: 石油储采比(a1), 储量替代率(a2), 石油消费增长率(a3), 石油消费弹性(a4), 石油增长速度与消费增长速度比(a5), 石油占总能源的比重(a6), 石油对外依存度(a7), 石油进口集中度(a8), 石油价格(a9), 油价波动率(a10), 实际安全程度(d)。

粗糙集理论主要研究离散系统, 但石油数据多以连续数据的形式存在, 因此在运用粗糙集的理论与方法进行数据挖掘^[5], 需要先将数据离散化。这里, 为了消除量纲的影响首先对数据进行归一化, 然后进行离散。

第二步: 建立安全预警知识表达系统。

定义安全预警决策表 $S = (U, C \cup D, V, f)$, 其中 $C = \{a1, a2, a3, a4, a5, a6, a7, a8, a9, a10\}$, $D = \{d\}$ 。

第三步: 决策表的知识约简。

对建立的决策表进行属性和属性值约简。

第四步: 提取最小决策规则。

第五步: 分析决策规则。

3 实例分析

收集我国 1993-2005 年石油安全数据, 建立知识表达系统, 如表 1 对数据进行归一化处理, 使其在 $[-1, 1]$ 区间内。

文中应用 Rose2 软件对决策表进行约简和提取最小规则, 得到最小约简为 $\{a7, a9\}$, 对 $\{a7, a9\}$ 进行属性值约简, 得规则见表 2。

选取规则可信度、覆盖度和支持度较高的规则作为决策的依据, 求取最小规则如下:

T1: $(a7=0) \& (a9=0) \Rightarrow (d=0)$,

$(a7=1) \Rightarrow (d=1)$,

$(a7=2) \Rightarrow (d=2)$,

$(a9=2) \Rightarrow (d=3)$ 。

T2: $(a7=0) \& (a9=0) \Rightarrow (d=0)$,

$(a7=1) \Rightarrow (d=1)$,
 $(a7=2) \Rightarrow (d=2)$,
 $(a9=3) \Rightarrow (d=3)$.

以上规则可以看出:

(1)对外依存度对我国石油安全的影响比较大。从1993年我国成为石油净进口国,对外依存度不断加大,到2004已达到45%^[6],同时中东地区是世界石油供应的中心也一直是主要进口来源地,进口单一,也使得我国石油安全存在很大隐患。

表1 原始数据

序号	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	d
1993	15.39	0.88	10.12	0.75	0.14	18.2	6.19	51.66	23.09	-14.4	0
1994	15.22	0.79	1.58	0.13	0.92	17.4	1.74	87.17	21.07	-8.7	0
1995	14.85	0.49	8.17	0.78	0.24	17.5	5.09	78.03	22.03	4.6	0
1996	14.29	0.82	8.49	0.88	0.75	18	7.97	75.12	25.94	17.7	1
1997	14.48	1.14	12.77	1.45	0.08	20.4	18.33	74.05	23.51	-9.4	1
1998	14.89	0.99	0.51	0.07	0.12	21.5	14.69	68.26	15.71	-33.2	1
1999	15.35	1.02	6.4	0.9	0	23.2	21.92	54.44	21.41	36.3	1
2000	15.07	0.95	6.48	0.8	0.29	24.6	31.13	59.21	32.88	53.6	2
2001	14.69	0.67	1.79	0.22	0.73	24.3	28.46	60.64	27.34	-16.8	2
2002	14.24	0.81	8.33	0.92	0.22	24	30	60.8	27.36	0.1	2
2003	14.08	1.02	9.72	0.97	0.19	22.7	36.45	59.15	30.62	11.9	2
2004	13.97	1.27	6.27	0.62	0.46	22.7	45.15	60.07	39.57	29.2	3
2005	13	1.16	6.6	0.67	0.59	22.7	42.9	61.01	54.52	37.8	3

表2 简化的决策规则表

编号	a7	a9	d	CF (A→B)	cover (A→B)	sup (A→B)
93,94,95	0	0	0	1	1	3/13
96	0	1	1	1	0.25	1/13
97,98,99	1	*	1	1	0.75	3/13
00,01,02	2	*	2	1	0.75	3/13
03	3	1	2	1	0.25	1/13
04	*	2	3	1	0.5	1/13
05	*	3	3	1	0.5	1/13

(2)从第四个规则可以看出石油价格的频繁波动

(上接第250页)

的物流配送系统的核心和出发点。在实际的作业过程中还涉及其他的作业管理,如库存盘点、入库管理、人事管理、设备管理、客户管理、系统维护等等。

(3)第三方物流管理信息系统功能结构。第三方物流管理信息系统功能结构可以划分为四个层次:数据管理层、业务处理层、决策管理层及战略管理层。

总之,物流软件和物流软件公司由于自身的不足,没有统一标准、缺乏分工合作、没有集成先进物流技术,还没有形成拥有拳头产品和明显优势的物流软件产业。因此急需制定统一的物流软件标准,特别是物流软件评测标准、建立物流软件质量保证平台来促进

使得我国石油安全状况日益恶化。1999年以来,国际油价增长了近7倍,据亚太经合组织估计,石油价格每上升10美元/桶,就会使通货膨胀上升0.5个百分点,经济增长率下降0.25个百分点。只有石油价格趋于稳定,世界各国的经济才能向健康稳定增长的方向发展。

根据未来某一年份的各项原始预测指标^[2],如2010年,2015年和2020年的原始预测指标见表3(略),这里的预测值有的是采自一些权威研究机构公开发表的有关数据,有些是根据其他数据测算而得。将2010,2015,2020年的预测值归一化,离散化得2010年 $a7=3,a9=3$,2015年 $a7=3,a9=3$,2020年 $a7=3,a9=3$ 。可见这三年我国石油安全属于重警区,需加强防范。

4 结束语

用粗糙集的方法对石油数据进行分析,得出影响石油安全的重要因素。根据不同年份的预警指标数据,建立相应的预警系统模型,实例也可看出所得到的结果与现实情况基本一致,说明了该方法的可行性。

参考文献:

[1] Pawlak Z. Rough set[J]. International Journal of Computer Information Sciences,1982,11(5):342-356.
[2] 范秋芳. 中国石油安全预警与对策研究[D]. 合肥:中国科学技术大学,2007.
[3] 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.
[4] 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.
[5] 李 剑,范小军,黄 沛. 基于粗糙集的知识理论及其应用[J]. 系统工程理论方法应用,2001,10(3):184-188.
[6] 李仕婷,王 欣. 中国石油安全现状及战略分析[J]. 西北工业大学学报,2007,27(1):42-44.

物流软件产业的发展和形成。

参考文献:

[1] 李孟刚. 中国物流产业安全问题研究[J]. 中国流通经济, 2007(12):7-10.
[2] 王庆智,王喜富. 基于供应链管理的物流信息平台设计研究[J]. 物流技术,2007(8):203-205.
[3] 孙宗虎,李世宗. 物流管理流程设计与工作标准[M]. 北京:人民邮电出版社,2007.
[4] 王坚红,王 京. 国外救灾物流的运作方式及启示[J]. 中国物流与采购,2006(6):56-58.
[5] 杨 巍,葛 星. 信息化如何提升物流利润[J]. 物流时代, 2006(10):51-53.