

面向水产品质量信息的知识发现研究

鄂旭^{1,2}, 方熠东², 胡少华², 金璐璐¹, 林爽¹

(1. 渤海大学 实验管理中心, 辽宁 锦州 121001;

2. 北京交通大学 中国产业安全研究中心, 北京 100084)

摘要:水产品安全知识发现是食品安全监管的一项重要内容,也是食品安全评价与预警的前提和基础。针对水产品安全信息系统中的规则发现问题,文中采用粗糙集理论,从理论、算法及应用三个层次进行了研究,提出了一种面向水产品安全信息系统的规则提取新方法。该方法在保持原水产品安全信息系统分类能力的前提下,通过数据挖掘技术发现数据中蕴涵的知识模式,精简水产品安全决策规则。实验表明该算法是有效可行的。

关键词:知识发现;粗糙集;食品安全;决策规则

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2014)06-0153-03

doi:10.3969/j.issn.1673-629X.2014.06.038

Research on Knowledge Discovery of Aquatic Product Quality Information

E Xu^{1,2}, FANG Yi-dong², HU Shao-hua², JIN Lu-lu¹, LIN Shuang¹

(1. Center of Experiment Management, Bohai University, Jinzhou 121001, China;

2. China Center for Industrial Security Research of Beijing Jiaotong University, Beijing 100084, China)

Abstract: Aquatic product safety knowledge discovery is one of the important contents in food safety supervision and management. And more, it is the premise and foundation of food safety evaluation and early warning. To deal with it, a new method for rule extraction oriented aquatic product safety information system was proposed based on rough set from theory, algorithm and application angles. Under the condition of maintaining the classification capacity, find the potential knowledge mode through data mining technology and reduce the food safety decision rules. Experimental results indicate the algorithm is effective and feasible.

Key words: knowledge discovery; rough set; food safety; decision rules

0 引言

近年来,国内外不断发生“疯牛病”、“三鹿奶粉”等重大食品安全问题,严重影响了人民的身心健康,同时也充分暴露出食品安全管理体系的薄弱。食品安全问题非常复杂,它涉及从“农田”到“餐桌”的整个过程,是一个涉及多个领域、多个环节的动态问题。目前,食品安全已成为世界各国共同关注的问题,已成为我国“十二五”规划的一项重要内容,其中水产品质量安全就是一项重要的研究课题。人们借助信息技术对食品的生产、运输、加工等各个环节进行全方位的监督、管理和控制,利用数据库保存了大量的数据^[1-3]。如何通过数据挖掘和知识发现技术,挖掘隐藏在海量数据中的潜在规则已经成为当前学术界关注的热点和

亟待解决的问题^[4-6]。数据挖掘是从数据库的大量数据中揭示隐含的、人们事先不知道的,但又是潜在有用的信息和知识的非平凡过程。目前最广泛的数据挖掘技术通常主要包括:粗糙集、统计分析方法、聚类算法、决策树算法、遗传算法、人工神经网络、模糊技术等。

粗糙集理论是一种新的数据挖掘工具,是由 Z. Pawlak 于 1982 年提出的一个分析数据的数学理论,它能有效地分析和处理不完备的信息。粗糙集理论能够提供有效的技术用于数据挖掘当中的数据预处理、数据属性约简、规则提取生成、数据依赖关系等方面^[7-10]。它与传统不确定信息处理方法的不同之处如下,例如模糊集理论、证据理论和概率统计理论等由于需要数据的附加信息或先验知识的缺点,该理论减少

收稿日期:2013-06-26

修回日期:2013-10-14

网络出版时间:2014-02-24

基金项目:中国博士后基金项目(2012M520158);辽宁省百千万人才基金择优资助项目(2012921058);辽宁省教育科研项目(L2012397, L2012396, L2012400);辽宁省社科联 2014 年度辽宁经济社会发展立项课题(2014LSLKT DGLX-02);2004 年辽宁省自然科学基金项目

作者简介:鄂旭(1971-),男,教授,博士,硕士生导师,研究方向为数据挖掘与食品安全物联网。

网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0857.002.html

了除问题所需处理的数据集合之外的任何先验信息^[8-14]。多年来,随着数据库知识发现和数据挖掘研究的兴起,由于粗糙集在分类分析、聚类分析等算法中的突出表现,已经成为这些领域研究中的一种有效分析方法,受到了全世界众多学者的高度重视。

文中基于粗糙集理论提出一种水产品质量信息中的知识发现策略。

1 粗糙集主要相关概念

由一个多值属性集合描述的一个对象集合是粗糙集的研究对象,每个对象及其属性都有一个值作为其描述符。这些信息在粗糙集中用决策表来表示,对象、属性和描述符是决策表的基本组成要素。决策表可以被看成是一个二维表格,表格的行与对象对应,列与对象的属性相对应;各行包含了相应对象信息的描述符以及类别信息。粗糙集及其在应用中所涉及的一些基本概念如下^[8-12]。

定义 1:在粗糙集中,信息系统可以用一个四元组来表示: $S = (U, A, V, f)$ 。

其中, U 是一个非空有限的对象集合,被称为论域; A 是一个非空有限的属性集合; V 是属性值的集合, $V = \cup (V_a)$, $a \in A$, V_a 是属性 a 的值域; f 是 $U \times A \rightarrow V$ 的映射函数,即 $\forall a \in A, x_i \in U, f(x_i, a) \in V_a$ 。

定义 2:信息系统 $S = (U, A, V, f)$, 设 $P \subseteq A$ 而且 $x, y \in U$, 称 x, y 关于 P 是不可分辨的,如果满足: $f(x, a) = f(y, a)$, $\forall a \in P$, 由属性 P 产生的不可分辨关系也称等价关系,它将 U 划分为若干个等价类,记为 $U/IND(P)$ 。

定义 3:令 $X \subseteq U$, 其中 R 是一等价关系。当 X 是某些等价关系 R 基本范畴的并集时,就称 X 为 R 可定义的,否则 X 是 R 不可定义的。 R 的可定义集是论域 U 的子集,在知识库 K 中可被精确的定义,而在知识库中 R 的不可定义集则不能被定义。 R 的可定义集被称作 R 精确集,而 R 的不可定义集被称作非精确集或 R 粗集。

定义 4:假定 $K = (U, R)$ 代表知识库,对于每个子集 $X \in U$ 和一个等价关系 $R \in IND(K)$,可以根据等价关系 R 的基本集合的描述对集合 X 进行划分。为了判定 $\{des(Y_i), Y_i \in R\}$ 精确地说明 X 中对象的隶属度情况,采用两个子集:

$$R_-(X) = \cup \{Y \in U/R; Y \subseteq X\}$$

$$R^-(X) = \cup \{Y \in U/R; Y \cap X \neq \emptyset\}$$

分别表示 X 的 R 下近似和 R 上近似。

它们也可以表征如下:

$$R_-(X) = \{x \in U; [X]R \subseteq X\}$$

$$R^-(X) = \{x \in U; [X]R \cap X \neq \emptyset\}$$

定义 5:设 R 是一个等价关系族, $r \in R$, 如果 $IND(R) = IND(R - \{r\})$, 则称 r 在 R 中是可被约去的知识;如果 $P = R - \{r\}$ 是独立的,则 P 是 R 中一个约简。

定义 6:假定 $Q \subseteq P$ 独立,且存在 $IND(Q) = IND(P)$, 则可以把 Q 看作是关系簇集 P 的一个约简。在 P 中所有绝对必要的关系集合称为 P 的核,记为 $core(P)$, 其结果为 P 的所有约简集合的交集,即: $core(P) = \cap red(P)$ 。

定义 7:属性 a 的重要性可以利用两个集合间的依赖性 $r_R(P)$ 来表示。设属性集合 $P, R \subseteq Q$, 有 $K = r_R(P) = card(POS_R(P))/card(U)$, $POS_R(P) = x \in U/IND(P)_{R-(X)}$ 。式子中的 $card$ 表示集合元素的个数, $POS_R(P)$ 为 R 在 $U/IND(P)$ 中的正区域。 P 与 R 的关系分以下几种情况:

- (1) $K=0$, 表明 P 与 R 没有关系,完全独立。
- (2) $K=1$, 表明 $P \rightarrow R$, P 与 R 是完全依赖关系。
- (3) $0 < K < 1$, 表明 P 与 R 是粗糙依赖或部分依赖。

2 决策知识算法描述

文中提出的算法包含如下基本步骤:

(1) 数据预处理:对采集到的用户原始数据进行转换,生成二维表形式,并划分确定条件、决策属性集合;

(2) 信息约简:对二维表进行概念抽象,生成面向对象的分辨矩阵,并应用约简算法简化信息表,抽取重要属性,生成约简属性集合;

(3) 获取知识:运用决策规则提取策略,获取相应的知识。

具体描述该算法如下:

输入: $S = (U, A, V, f)$ 为信息表, $R = C \cup D$, C 为条件属性集, D 为决策属性集;

输出:决策表 S 的决策知识。

Step1:首先扫描二维数据表,当两个对象具有相同属性时,消去重复对象;

Step2:比较不同对象属性值,消除相容属性列;

Step3:针对二维表,抽取可分辨矩阵 M :

for ($i = 1; i \leq n; i++$)

for ($j = 1; j < i; j++$)

$M = (C_{ij})$;

Step4:应用公式计算核属性 $Core_d(C)$, 初始化 $C_0 = Core_d(C)$;

Step5:设 $B = C_0$, $AR = C - B$ 为其他件属性;计算每一属性 $a_i \in AR$ 的重要性 $f(a_i)$, 并按降序方式对属性进行排序;

Step6:利用公式求取 $r_B(D)$ 和 $r_C(D)$;

Step7:while($r_B(D) \neq r_C(D)$)and($AR \neq \emptyset$)

$\{ f(a_k) = \max(f(a_i)) ;$

$B = B \cup \{ a_k \} ;$

$AR = AR - \{ a_k \} ; \}$

Step8:抽取、生成属性相对约简表;

Step9:合并重复规则,通过规则提取手段抽取决策知识。

3 应用举例

给定一个水产品质量信息表,如表1所示。其中, a, b, c, d 为条件属性, e 为决策属性,即 $C = \{ a, b, c, d \}, D = \{ e \}$ 。

表1 原始信息表

U	a	b	c	d	e
1	L	P	B	G	0
2	L	P	G	M	0
3	M	P	G	M	0
4	M	G	G	M	1
5	M	V	V	B	0
6	H	G	B	M	0
7	H	G	G	M	1
8	H	G	B	G	1
9	E	V	G	M	0
10	E	P	B	V	0

针对表1进行转换,生成分辨矩阵 M ,如表2所示。

表2 分辨矩阵

	1	2	3	4	5	6	7	8	9
1									
2	\emptyset								
3	\emptyset	\emptyset							
4	$abcd$	ab	b						
5	\emptyset	\emptyset	\emptyset	bcd					
6	\emptyset	\emptyset	\emptyset	ac	\emptyset				
7	$abcd$	ab	ab	\emptyset	$abcd$	c			
8	ab	$abcd$	$abcd$	\emptyset	$abcd$	ad	\emptyset		
9	\emptyset	\emptyset	\emptyset	ab	\emptyset	\emptyset	ab	bcd	
10	\emptyset	\emptyset	\emptyset	$abcd$	\emptyset	\emptyset	$abcd$	b	\emptyset

- (1)利用核属性公式求得 $C_0 = \{ b, c \}$;
- (2)根据公式计算得 $AR = C - B = C - C_0 = \{ a, d \}$;
- (3)分别计算 AR 中属性 a 和 d 的重要性,得5和2.47;
- (4)降序排列属性 a 和 d ,根据属性重要性得 $\{ a, d \}$;
- (5)计算得 $r_B(D) = 0.8, r_C(D) = 1$;
- (6)由于 $r_C(D) \neq r_B(D)$,故 $B = B \cup \{ a \} = \{ a, b, c \}$;
- (7)计算 $r_B(D) = 1$,又由于 $r_C(D) = r_B(D)$,所以 $B = \{ a, b, c \}$ 为一个最小相对约简集;
- (8)整合规则信息表,获取决策知识;

$$a_0b_0 \vee a_1 \vee a_2b_1c_0 \vee a_3 \Rightarrow e_0;$$
$$a_1b_1c_1 \vee a_2b_1c_1 \vee a_3b_1c_0 \Rightarrow e_1.$$

4 结束语

在粗集理论中,决策规则是知识库的一种表示形式。知识发现的本质概括抽取决策信息中的决策规则,所以信息表中属性的多少或简单与否将直接影响获取知识的质量。然而,属性相对约简却是一个 NP-hard 问题,通常采取的启发式搜索方法虽然可以提取重要属性,但通常效率很差。为了解决这个问题,文中基于粗糙集中分辨矩阵的概念,以属性重要性作为迭代启发信息,提出了一个新的知识发现算法,大大缩减了求取水产品质量表征属性的搜索空间,有利于高效获取水产品质量的相关知识。

参考文献:

[1] 邓聪文,朱雪冬,王俊能. 食品安全评价及其方法简述[J]. 食品安全,2009(6):8-10.

[2] 郭旭强,王大建,王秀霞,等. 影响水产食品安全因素的分析[J]. 齐鲁渔业,2009,26(12):49-51.

[3] 鄂旭,韩芳,侯建,等. 面向食品安全评价的属性约简方法研究[J]. 吉林大学学报(信息科学版),2013,31(3):314-319.

[4] Pawlak Z. Rough sets and fuzzy sets[J]. Fuzzy Sets and Systems,1985,17(1):99-102.

[5] Krysikiewicz M. Rough set approach to incomplete information system[J]. Information Sciences,1998,112(1-4):39-49.

[6] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2005.

[7] 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,2006.

[8] 曾黄麟. 粗糙集理论及其应用[M]. 重庆:重庆大学出版社,1996.

[9] 武森,高学东,Bastian M. 数据仓库与数据挖掘[M]. 北京:冶金工业出版社,2003.

[10] Hu X H, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338.

[11] 常犁云,王国胤,吴渝. 一种基于 Rough set 理论的属性约简及规则提取方法[J]. 软件学报,1999,10(11):1206-1211.

[12] 鄂旭,高学东,喻斌. 基于扫描向量的属性约简方法[J]. 北京科技大学学报,2006,28(6):604-608.

[13] 李仁璞,黄道. 基于 RS 理论的不完备信息系统处理方法[J]. 华东理工大学学报(自然科学版),2005,31(2):227-231.

[14] E Xu, Yang Yuqiang, Ren Yongchang. A new method of attribute reduction based on information quantity in an incomplete system[J]. Journal of Software,2012,7(8):1881-1888.

面向水产品质量信息的知识发现研究

作者：鄂旭，方熠东，胡少华，金璐璐，林爽，E Xu，FANG Yi-dong，HU Shao-hua，JIN Lu-lu，LIN Shuang

作者单位：鄂旭, E Xu(渤海大学 实验管理中心, 辽宁 锦州 121001; 北京交通大学 中国产业安全研究中心, 北京 100084)，方熠东, 胡少华, FANG Yi-dong, HU Shao-hua(北京交通大学 中国产业安全研究中心, 北京, 100084)，金璐璐, 林爽, JIN Lu-lu, LIN Shuang(渤海大学 实验管理中心, 辽宁 锦州, 121001)

刊名：计算机技术与发展

英文刊名：Computer Technology and Development

年，卷(期)：2014(6)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201406038.aspx