

粗糙集理论分析及其在农业数据中的应用

郭红波¹,周明全^{1,2},耿国华¹,王小凤¹

(1. 西北大学信息科学与技术学院, 陕西 西安 710069;

2. 北京师范大学信息学院, 北京 100875)

摘要:从数据库中挖掘有用信息,将难理解的纯数据变为容易利用的规则,从而为以后的决策提供依据。以粗糙集理论和规则提取算法为基础,将基于信息量的粗糙集属性约简算法和规则提取算法集成起来提出一种集成算法,应用粗糙集约简掉冗余属性,然后利用规则提取算法得出有效规则。将此集成算法应用于农业领域,得出规则,并且效果良好,理论分析和应用都表明了本算法的有效性和实用性。此集成算法可以应用于各种大型数据库中,从中得出有效规则,让历史数据为以后的决策服务。

关键词:粗糙集;信息量;属性约简;值约简

中图分类号: TP391.4

文献标识码: A

文章编号: 1673-629X(2007)10-0165-03

Rough Set Theory Analysis and Its Application in Agricultural Field

GUO Hong-bo¹, ZHOU Ming-quan^{1,2}, GENG Guo-hua¹, WANG Xiao-feng¹

(1. School of Information Science and Technology in Northwest University, Xi'an 710069, China;

2. School of Information, Beijing Normal University, Beijing 100875, China)

Abstract: To dig useful information from database, and get the easy using rules from difficult understand pure data, then offer future decision information according to the rules obtained in the system. Based on the rough set theory and rules extraction algorithm, an integrated algorithm is proposed. Using rough set theory the redundant attribute can be reduced, and then the efficient rules are obtained according to the rules extraction algorithm. The data in agricultural fields is analyzed by this algorithm and effective rules are obtained. The analysis in theory and the applications show that it is not only effective but also feasible. The algorithm can be widely used in all kinds of large databases, from which the effective rules can be obtained. Then the history's data can serve the future decision-making.

Key words: rough set theory; information quantity; reduction of attribute; value reduction

0 引言

粗糙集(Rough Set)理论是由 Z. Pawlak 于 1982 年提出的^[1],它是一种刻画不完整性和不确定性的数学工具,能有效地分析和处理不精确、不一致、不完整等各种不完备信息,并从中发现隐含的知识,揭示潜在的规律;它的主要思想是在保持信息系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。它的核心内容是属性约简和规则提取。属性约简是根据属性的重要程度约简的,在属性重要性的度量

上可以采用基于属性依赖度、信息熵、互信息、加权平均和信息量度量等^[2-8]方法,在比较了上述几种方法后采用基于信息量度量^[8]的方法;在规则提取过程中,必须先进行值约简,然后导出规则。文中将用此算法分析农业数据,得到有效的规则。

1 粗糙集理论基本概念

在粗糙集理论中,对象的知识是通过指定对象的基本特征(属性)和它们的特征值(属性值)来描述的。一个知识表达系统定义^[9]为

$$S = \langle U, C, D, V, F \rangle$$

式中, U 是对象的集合, $C \cup D = A$ 是属性集合(等价关系集合),子集 C 和 D 分别称为条件属性和结果属性, $V = \bigcup_{a \in A} V_a$ 是属性值的集合, V_a 表示了属性 $a \in A$ 的范围, $f: U \rightarrow V_a$ 为一单射,使论域 U 中任一元素取属性 a 在 V_a 中的某一个唯一值。这种定义方式使对

收稿日期:2006-12-24

基金项目:国家自然科学基金资助项目(60372072);中国博士后科学基金资助项目(2003033519)

作者简介:郭红波(1978-),男,陕西人,硕士研究生,研究方向为智能挖掘和音频处理等;周明全,教授,博士生导师,研究方向为图形图像处理、数据挖掘等;耿国华,教授,博士生导师,研究方向为智能信息处理、数据库、数据挖掘等。

