粗糙集理论数据处理方法及其研究

张政超[1,2],关 欣[1,3],何 友[1],李应升[2],郭伟峰[2]
(1.海军航空工程学院 信息融合技术研究所,山东 烟台 264001;
2.中国人民解放军63891部队,河南 洛阳 471003;
3.国防科学技术大学 电子科学与工程学院,湖南 长沙 410073)

摘 要:粗糙集理论是一种对数据进行约简,提取规则的数据挖掘的有效工具,在自动控制、电子科学、计算机科学、机器学习、医学、经济学等方面有着广泛应用。根据粗糙集理论处理数据方法的过程,分析和阐述了不完备数据处理、连续数据离散化、属性约简、属性值约简和规则提取、不完备决策系统和不相容决策系统等非标准信息系统的约简、粗糙集理论数据处理方法和其他理论数据处理方法相结合的扩展模型和最新研究进展,及粗糙集理论数据处理的软件实验系统等。

关键词:粗糙集;离散化;属性约简;属性值约简
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2010)04-0012-05

Rough Sets Data Processing Method and Its Research

ZHANG Zheng-chao[1,2], GUAN Xin[1,3], HE You[1], LI Ying-sheng[2], GUO Wei-feng[2]
(1. Research Inst. of Information Fusion, Naval Aeronautical and Astronautical Univ., Yantai 264001, China;
2. PLA Unit 63891, Luoyang 471003, China;
3. Inst. of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Rough sets is a valuable data mining tool in auto control, electrical science, computer science, machine learning, medical science and economics by reducing data and rules extraction. According to the course of rough sets data processing method, incomplete decision information system data processing, data discretization, attribute value reduction, incomplete and inconsistent decision information system reduction, and the model of rough sets combined with other data processing method are analyzed, and their new research development, as well as the software experiment of data processing of rough sets are mentioned in this paper.

Key words: rough sets; discretization; attribute reduction; attribute value reduction

0 引 言

粗糙集理论[1,2]是一种新的处理模糊、不确定信息的数学工具。它的主要思想是在保持分类精度不变的前提下,通过知识约简和规则提取,得到问题的决策或分类规则,目前粗糙集理论已在多个领域得到了广泛的应用。

粗糙集理论的数据处理方法是通过有监督的学习,经过训练数据构成的决策表,通过约简得到最小的决策算法,从而完成对新的数据集合的分类归并。决策表中的数据的来源是多途径的,既可以是通过仪器观察或测量得到的,也可以是专家系统给出的,还可以是通过某种系统模型产生或是仿真获得的。

具体说来,粗糙集理论的数据处理方法可以按以下几个步骤进行:

(1)特征提取。由观测记录的数据及用于进行决策的数据应该能够反映所应用领域的全部属性,通过对数据进行处理、变换,达到能使提取的属性体现系统的全部特征以及降低待处理数据的维数。

(2)决策表的建立。决策表的建立包括三部分:条件属性和决策属性的确定、不完备数据的补齐、连续数据的离散化。

(3)约简。约简包括特征选择(属性约简)、属性值约简、规则提取及最小决策算法的获取。

(4)决策应用。通过以上的数据处理,提取的规则就可以应用于新的数据的分类和决策分析。

正是粗糙集理论数据处理方法的优点,使得粗糙

收稿日期:2009-07-03;**修回日期**:2009-11-08
基金项目:国家自然科学基金资助项目(60572161);国家自然科学基金资助项目(60672140);全国优秀博士论文作者专项资金资助项目(200443);教育部新世纪优秀人才支持计划(NCET-05-0912)
作者简介:张政超(1981-),男,湖北武穴人,助理工程师,硕士研究生,研究方向为粗糙集理论及其应用等;关 欣,博士,副教授,主要研究领域为多传感器信息融合、模式识别。

集理论有着强大的数据约简功能,在没有数据的任何先验分布信息情况下,Let the data speak for themselves[3]。文中对粗糙集理论的数据处理方法各个环节进行了研究。

1　基于粗糙集的数据预处理

从现实世界获得的初始数据,由于各种原因,可能得到的数据值是不完备的。这些缺失的数据导致了所构成的信息系统的不完备。另一方面,各种待处理的数据中,存在着各种连续的、离散的、数值型的、字符型的等各种表现形式。在用粗糙集理论进行数据约简前,需要对这些原始数据进行预处理。

针对原始数据的数值缺失和不全是离散值的特点,粗糙集理论数据预处理方法主要有两方面:不完备数据处理和数据离散化。

1.1　不完备数据处理

数据的缺失的原因是多方面的:由于技术、人员、设备等原因无法及时获取;依赖数据间的联系需要推理获得;无意义的信息;获取的代价太高等。数据集中不含缺失值的变量(属性)称为完全变量,数据集中含有缺失值的变量称为不完全变量。数据缺失机制有以下三种[4]:

(1)数据的缺失与不完全变量以及完全变量都是无关的完全随机缺失。

(2)数据的缺失仅仅依赖于完全变量的随机缺失。

(3)不完全变量中数据的缺失依赖于不完全变量本身,且这种缺失是不可忽略的非随机、不可忽略缺失。

缺失值对于数据挖掘造成了丢失信息、增加不确定性、导致系统输出不可靠的结果。数据挖掘算法与实际应用之间的差距可以通过用推导、填充空缺的数据得到减少。

对不完备数据处理的方法主要有两种:删除对象和数据补齐。

1.2　数据离散化

离散化经典的粗糙集理论不能处理具有连续属性值的信息系统,在对数据处理之前,除了必要的不完备数据处理,还要进行离散化处理。连续数据的离散化应尽可能满足以下两点:

1)连续属性离散化后的空间维数应尽可能小,也就是经过离散化后的每一个属性都应包含尽量少的属性值的种类;

2)连续属性值离散化处理后丢失的信息应尽量少。

最优离散化问题已经被证明是一个 NP－hard 问题。

连续属性的离散化问题被广泛研究,并取得了大量成果,研究人员从不同领域提出了多种离散化方法,可以从以下三个方面进行分类[5]:

a. 局部离散化和全局离散化。

b. 监督离散化和非监督离散化。

c. 静态离散化和动态离散化。

从粗糙集的观点看,离散化的实质是在保持决策表分类能力不变,即条件属性和决策属性相对关系不变的条件下,寻找合适的分割点集,对条件属性构成的空间进行划分。评价属性离散化的质量,主要看分割点的选择和多少,并确保信息系统所表达的样本之间的"不可分辨关系",即既能准确地对论域中实例进行分类,又不引入新的噪声。当然离散化的结果可能不唯一,同一个决策表可能有多种离散化结果。

连续属性离散化方法除了传统的等宽度离散化方法、等频率离散化方法、Naïve Scaler 离散化方法、Semi Naïve Scaler 离散化方法、粗糙集与布尔逻辑相结合的离散化方法、S H Nguyen 和 H S Nguyen 改进的离散化方法、超平面离散化方法、超曲面的离散化方法[6]外,近年来又涌现出了很多新的离散化方法:

(1)聚类离散化方法。

苗夺谦利用决策表的相容性的反馈信息,提出了一种领域独立的基于动态层次聚类的离散化方法[7],韩中华等人提出了基于谱系聚类的离散化方法,并经实验证明得到较好的应用效果[8],文献[9]给出了基于新聚类学习算法对决策表属性值进行离散化的算法,王伟等人[10]还提出了一种基于模糊聚类的离散化方法。

(2)基于信息论与熵的离散化方法。

胡逢彬在文献[11]中提出了一种基于相对熵的决策表连续属性离散化算法,徐如燕等人[12]则使用信息论的方法进行连续属性的离散化,引入 Hellinger 偏差作为每个区间对决策的信息度量,从而定义切分点的信息熵,最终的结果是使各区间的信息量尽可能平分。李春贵等人在一致性假设前提下,以数据集的统计性质作为启发式知识,从候选离散点集中选择离散点,根据数据集的期望值和方差来确定搜索最优离散点的区域,提出一种新的基于信息熵粗集数值属性离散化算法,并采用 UCI 国际标准数据集来验证新算法[13]。

(3)基于属性重要度的离散化方法。

刘凌霞[14]研究了粗糙集理论在决策表离散化中的应用,提出了一种基于粗糙集理论属性重要性的决策表离散化算法。该算法首先使用依赖度来定义属性的重要性,并据此对条件属性按照重要性由小到大排

序,然后按排序后的顺序,考察每个条件属性的所有断点,将冗余的断点去掉,从而将条件属性离散化。

2 基于粗糙集的约简

粗糙集理论的数据处理方法的主要方式表现在对数据的约简上。主要包括属性约简和属性值约简两大方面。属性约简是通过粗糙集合理论的数据处理方法,去除信息系统中多余的属性,达到提取信息系统中不可缺少的信息。属性值约简则是在属性约简的基础上,对属性的取值进一步地分析,删除多余的取值信息,为提取规则打下基础。

2.1 属性约简

属性约简是粗糙集理论的核心内容之一。在信息系统中各个属性并不是同等重要的,甚至有些是冗余的。所谓属性约简,就是在保持数据分类能力不变的条件下,约简冗余的属性,即消去决策表中一些不必要的列。每约简一个属性后查看新表是否出现了不一致,若无不一致,则此属性可以约去,否则不能约去。这种方法也被称为属性约简的数据分析。

属性约简的方法有很多,大致可分为以下几类:

1)简单删除法。

依次从数据表中删除属性,看删除属性后的决策表的不可分辨关系是否改变,如果没有改变,则删除此属性,继续比较;如果不可分辨关系发生改变,则恢复到前一个信息表删除另一属性。这种算法在属性较少和数据量较小的情况下比较有效,但当条件属性较多或含有大量的记录时,计算的复杂度比较高。

2)基于属性重要性的约简算法。

基于属性重要性的属性约简算法的主要思想是在决策表的核集的基础上,依次添加属性,直到满足添加后的属性集合的属性依赖度与所有条件属性的决策依赖度相等。在此基础上,依次删除非核属性,直到所有属性都满足是不可缺少为止。

路松峰等人在文献[15]中针对现有属性约简算法存在的问题,利用信息论和粗糙集理论,提出了基于属性依赖的属性约简算法,该算法不用求核。首先利用单个条件属性与决策属性的依赖度来选择条件属性,取与决策属性依赖度大的属性,计算完毕后,将得到的条件属性两两之间进行依赖度计算,删除冗余属性,最后得到条件属性的约简。江敬之在文献[16]中为解决多约简决策表的约简选取问题,在综合考虑约简中属性的平均重要性以及属性个数的基础上,提出了约简重要性的概念,以此概念为准则对多个约简进行比较时,可选出一个最佳约简。

3)基于差别矩阵的约简算法[17~19]。

基于差别矩阵的约简算法通过构造一个差别矩阵,求出差别函数,在此基础上,得到决策表的属性约简结果。

4)基于集合近似质量的决策系统的属性约简算法。

基于集合近似质量的决策系统的属性约简算法与基于属性重要性的约简算法类似,只不过度量的函数由属性重要性、属性依赖度换成了近似质量函数。

除了上述四种基本的属性约简方法之外,近年来又涌现出了许多新的属性约简算法:

(1)基于信息熵、互信息的属性约简算法[20,21]。

(2)基于遗传算法的属性约简算法[22,23]。

(3)基于概念格的属性约简算法[24,25]。

(4)基于免疫的属性约简算法[26,27]。

(5)基于排序的属性约简算法[28,29]。

(6)增量式属性约简算法[30,31]。

2.2 属性值约简和规则提取

决策信息系统经过属性约简后,若要达到去除冗余属性值的目的,还必须进行属性值约简。约简的主要目的是获取规则。

2.2.1 属性值约简

现有的属性值约简算法有一般属性值约简算法、归纳值属性约简算法、启发式属性值约简算法、基于决策矩阵的属性值约简算法等[4]。

宋旭东等人在文献[32]中提出了一种改进的基于属性值重要性的 Rough 集值约简算法,该算法在执行效率上有很大的提高,通过实例分析验证了该算法的可行性和有效性。

2.2.2 规则提取

决策信息系统经过属性约简和属性值约简后便可以提取规则。

规则提取的方法也有很多[33~35]。一般来说,为了考察提取出的规则的分类性能,还需要用未知对象对其进行分类测试和评估。

规则应用的方法一般如下:

(1)将分类规则作为新的分类器识别新对象时,首先在规则集中寻找匹配的规则;

(2)如果没有匹配的规则,则将此对象作为新类别对待;

(3)如果多于一个规则与新对象匹配,则需要加入人工干预。

3 非标准信息系统的约简

非标准信息系统是针对相容完备的信息系统而言的,主要有不完备信息系统和不相容完备系统。在许

多情况下,面临的信息系统是非标准的。

3.1 不完备决策系统的约简

不完备的信息系统主要是指属性值有空值的信息系统。这样的信息系统可以通过删除对象和数据补齐的方式达到完备的目的,但是人们还是希望在保持原始信息不发生变化的前提下对信息系统进行处理,直接在包含空值的数据上进行数据挖掘。

对于不完备信息系统的约处理方法主要有以下几种[36]:

(1)基于容差关系的粗糙集处理。

(2)基于非对称相似关系的处理。

(3)基于限制容差关系的处理。

近年来,出现了许多新的不完备信息系统的约简方法:

钟波在文献[37]中为了对不完备信息进行有效识别,引入了粗糙集贝叶斯定理,结合粗集约简识别方法,建立基于粗糙集的最小错误率贝叶斯决策准则,归纳出不完备信息模式的一种统计意义上的识别方法。邢化玲等人在文献[38]中根据分层递阶约简算法,提出了一种直接在不完备信息系统上进行数据挖掘的方法。该方法首先将信息系统中由所有属性构成的单层知识表示转变成由部分属性所构成的多层知识表示,即由完备属性和不完备属性表示;然后建立了两个不同层次的子系统,并推导出各个子系统的规则集。

3.2 不相容决策系统的约简

不相容信息系统也称为不协调信息系统。对于不相容信息系统的数据约简用得最多的是用变精度粗糙集模型解决的[39~41]。近年来又有许多新的关于不相容决策系统约简方法的报道:汪小燕[42]针对目前求核方法存在的问题,提出一种基于分布函数的用于计算核属性的改进的二进制可辨矩阵。改进的二进制可辨矩阵不仅规模小,而且适用于任何决策表求核。在获取核属性的基础上,提出一种新的不一致决策表的属性约简算法,只要在用于计算核属性的改进的二进制可辨矩阵中简单增加相应的行,就可以利用逻辑运算来获取属性约简。

陈鑫影、邱占芝[43]定义了决策包含度约简和最大决策包含度约简的概念,讨论了决策包含度约简和最大决策包含度约简的关系,即最大决策包含度约简弱于决策包含度约简,为解决不协调决策信息系统的知识约简问题提供了新方法。

4 关于粗糙集的其它问题

粗糙集理论在数据处理方面有着强大的约简功能,可以提取出规则,去除多余的属性、属性值。将粗糙集理论的数据处理方法和其它的数据处理方法结合起来,可以发挥各自的优点,更为有效地解决现实世界的数据处理问题。

目前针对不同的应用背景和选取的约简方法的不同,世界各地科学家开发了多种粗糙集数据处理方法软件实验系统。

4.1 粗糙集与其它方法结合的扩展模型

迄今为止,粗糙集合已成功地和各种方法结合,应用于各个领域:

(1)与 Agent 结合[44]。

(2)与贝叶斯网络、概率论结合[45,46]。

(3)与聚类分析方法相结合[47,48]。

(4)与模糊集理论相结合的[49,50]。

(5)与小波分析相结合的[51,52]。

(6)与蚁群算法相结合的[53,54]。

除此之外,粗糙集理论数据处理方法还和云模型、遗传算法、主成分分析有着良好的互补性质。

4.2 粗糙集软件实验系统

世界各地的科学家建立了不少基于粗糙集的知识数据挖掘系统,其中最有代表性的有 KDD-R、ROSE、LERS 等。

(1)KDD-R。

KDD-R 是由加拿大的 Regina 大学开发的基于可变精度粗糙集模型,采用知识发现的决策矩阵方法开发了 KDD-R 系统,这个系统被用来对医学数据分析,以此产生症状与病证之间新的联系,另外它还支持电信工业的市场研究。

(2)ROSE。

波兰 Poznan 科技大学基于粗糙集开发了 ROSE (Rough Set data Explorer),用于决策分析。它是 Rough Das & Rough Class 系统的新版,其中 Rough Das 执行信息系统数据分析任务,Rough Class 支持新对象的分类,这两个系统已经在许多实际领域中得到应用。目前 ROSE 已有新的版本 ROSE2。

(3)LERS。

LERS(Learning from examples based on Rough Set)系统是美国 Kansas 大学开发的基于粗糙集的实例学习系统。它是用 Common Lisp 在 VAX9000 上实现的。LERS 已经在 NASA 的 Johnson 空间中心应用了两年。此外,LERS 还被广泛地用于环境保护、气候研究和医疗研究。

5 结束语

粗糙集理论以其强大的数据约简功能,用属性约简、属性值约简、规则提取的数据处理方法,在保持分

类精度不变的前提下,完成对决策信息系统的识别、归类。随着研究的深入以及粗糙集数据处理方法与其他数据处理方法的结合,必将开发出更多更为有效的软件实验系统应用于各领域。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science,1982,11(5):341-356.

[2] Pawlak Z. Rough Sets - Theoretical Aspects of Reasoning about Data[M]. Boston: Kluwer Academic Publishers,1991:72-80.

[3] Düntsch I, Gediga G. Rough set data analysis:A road to non-invasive knowledge discovery[M]. UK:Methodos Publishers,2000.

[4] 王国胤. Rough集理论与知识获取[M]. 西安:西安交通大学出版社,2001:168-217.

[5] 赵晓伟. 数据挖掘技术探讨:基于错误率的全局离散化方法[EB/OL]. 2007-08-06[2009-03-14]. http://www.eNet 硅谷动力.com.

[6] 胡寿松,何亚群. 粗糙决策理论与应用[M]. 北京:北京航空航天大学出版社,2006:27-44.

[7] 苗夺谦. Rough Set 理论中连续属性的离散化方法[J]. 自动化学报,2001,27(3):296-302.

[8] 韩中华,马 斌,许 可,等. 基于谱系聚类的粗糙集数据挖掘预处理方法[J]. 计算机工程与应用,2008,44(2):194-196.

[9] 赵晓霞. 基于新聚类学习的离散化方法[J]. 现代电子技术,2007(22):197-199.

[10] 王 伟,高 亮,吴 涛. 一种基于模糊聚类的离散化方法[J]. 计算机技术与发展,2008,18(3):53-55.

[11] 胡逢彬,桂现才. 基于相对熵的决策表连续属性离散化算法[EB/OL]. 2008-05-06[2009-03-16]. http://www.studa.net.

[12] 徐如燕,鲁汉榕,郭齐胜. 基于信息论的连续属性离散化[J]. 空军雷达学院学报,2001,15(2):20-23.

[13] 李春贵,王 萌,原庆能. 基于启发式信息熵的粗集数值属性离散化算法[J]. 广西科学院学报,2007,23(4):235-237.

[14] 刘凌霞. 基于粗糙集理论属性重要性的离散化算法[J]. 广西轻工业,2007(10):75-76.

[15] 路松峰,刘 芳,胡 波. 一种基于属性依赖的属性约简算法[J]. 华中科技大学学报:自然科学版,2008,36(2):39-41.

[16] 江敬之. 基于约简重要性的最佳约简求解算法[J]. 计算机应用,2008,28(6):1435-1437.

[17] 吕 静,陈 炼. 基于分明矩阵方法的属性约简方法[J]. 微计算机信息,2008,24(2-3):236-238.

[18] 薛安荣,韩红霞,潘雨青. 基于可辨识矩阵的快速粗糙集属性约简算法[J]. 计算机工程与设计,2007,28(20):4987-4989.

[19] 王加阳,高 灿. 基于分辨矩阵的快速完备约简算法[J]. 计算机工程与应用,2008,44(8):92-94.

[20] 丁守祯,桑琳,朱全英,等. 基于信息熵的粗糙集属性约简及其应用[J]. 计算机工程与应用,2007,43(35):245-248.

[21] 颜 艳,杨慧中. 一种基于互信息的粗糙集知识约简算法[J]. 清华大学学报:自然科学版,2007,47(S2):1903-1906.

[22] 肖厚国,桑琳,丁守珍,等. 基于遗传算法的粗糙集属性约简及其应用[J]. 计算机工程与应用,2008,44(15):228-230.

[23] 颜 艳,杨慧中. 基于遗传算法的粗糙集属性约简算法[J]. 计算机工程与应用,2007,43(31):156-158.

[24] 胡学钢,王昕娅,张玉红. 基于概念格模型的约简表示及求解[J]. 合肥工业大学学报:自然科学版,2007,30(3):278-281.

[25] 王 霞,张文修. 概念格的属性约简与属性特征[J]. 计算机工程与应用,2008,44(12):1-4.

[26] 张 旭,郭 晨. 基于免疫原理的粗糙集属性约简[J]. 计算机工程,2007,33(23):51-53.

[27] 向长城,黄席樾,杨祖元,等. 基于免疫算法的粗糙集知识约简[J]. 计算机仿真,2007,24(11):155-158.

[28] 徐伟华,张文修. 基于优势关系下信息系统分配约简的矩阵算法[J]. 计算机工程,2007,33(14):4-7.

[29] 戴 毓,周德群. 决策分析中属性约简的择优算法[J]. 系统工程,2007,25(8):89-93.

[30] 杨 明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报,2007,30(5):815-822.

[31] 王 杨,闫德勤,张凤梅. 基于粗糙集和决策树的增量式规则约简算法[J]. 计算机工程与应用,2007,43(1):170-172.

[32] 宋旭东,朱伟红,宁 涛. 基于属性值重要性的 Rough 集值约简算法[J]. 计算机技术与发展,2007,17(6):77-79.

[33] 徐凤生. 一种属性与值约简及规则提取算法[J]. 计算机工程与科学,2008,30(2):61-63.

[34] 邓九英,毛宗源,杜启亮,等. 基于粗糙集的决策规则设计算法研究[J]. 计算机工程与应用,2007,43(30):209-212.

[35] 周春来,李志刚,孟跃进. 决策规则获取算法及规则表示[J]. 计算机工程与应用,2007,43(4):102-105.

[36] 汤路金,魏大宽,汤路生. 基于对称相似关系的不完备信息系统粗糙集拓展模型[J]. 湖南农业大学学报:自然科学版,2007,33(2):239-243.

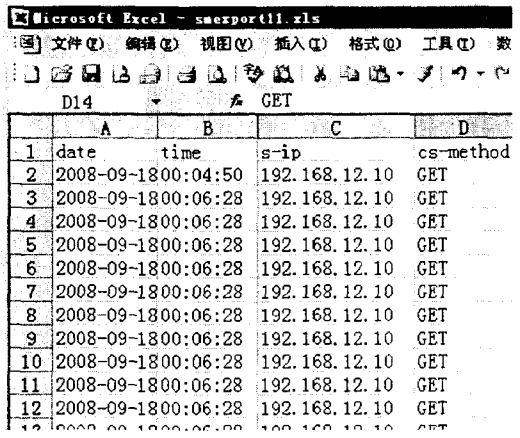
[37] 钟 波,罗会亮. 不完备信息的粗糙集-贝叶斯识别方法[J]. 重庆大学学报,2008,31(1):74-76.

[38] 邢化玲,刘思伟,高社生,等. 不完备信息系统的数据挖掘方法研究[J]. 计算机应用研究,2008,25(1):90-92.

[39] Ziarko W. Variable precision Rough Set model[J]. Journal of

(下转第 20 页)

了日志挖掘中数据预处理的基本概念和基本过程，并用开发工具制作了专门软件来完成日志文本信息到关系型数据格式的转换。但由于 Web 日志的多样性和不确定性，还有很多问题亟待解决，有待于进一步去研究和探索。


Microsoft Excel - smexport11.xls

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数

D14 GET

	A	B	C	D
1	date	time	s-ip	cs-method
2	2008-09-18	00:04:50	192.168.12.10	GET
3	2008-09-18	00:06:28	192.168.12.10	GET
4	2008-09-18	00:06:28	192.168.12.10	GET
5	2008-09-18	00:06:28	192.168.12.10	GET
6	2008-09-18	00:06:28	192.168.12.10	GET
7	2008-09-18	00:06:28	192.168.12.10	GET
8	2008-09-18	00:06:28	192.168.12.10	GET
9	2008-09-18	00:06:28	192.168.12.10	GET
10	2008-09-18	00:06:28	192.168.12.10	GET
11	2008-09-18	00:06:28	192.168.12.10	GET
12	2008-09-18	00:06:28	192.168.12.10	GET

图 3 预处理产生的 Xls 文件

```
<?xml version="1.0" encoding="UTF-8" ?>
- <dataroot xmlns:od="urn:schemas-micros
  xsi:noNamespaceSchemaLocation="smex
  - <smexport>
    <ID>1</ID>
    <data>2008-09-18</data>
    <time>00:04:50</time>
    <sip>192.168.12.10</sip>
    <method>GET</method>
    <cs-uri-query>/rsc/zsbbook/leavew
    <cs-username>8001</cs-username>
    <cip>74.6.22.183</cip>
    <ie>Mozilla/5.0+(compatible</ie>
  </smexport>
  - <smexport>
```

图 4 预处理产生的 XML 文件

注：本程序在 http://www.hftc.edu.cn/xjzx/datamining.rar 有下载。

参考文献：

[1] Cooley R, Srivastava J. Grouping Web page references into transactions for mining world wide Web browsing patterns [C]//Proceedings of KDEX'97. Newport Beach, CAUSA: [s.n.], 1997:2-7.

[2] Wong J S K, Nayar R. A framework for a world wide web based data mining system[J]. Journal at Network and Computer Applications, 2000, 21:163-185.

[3] Pei J, Han J. Mining access patterns efficiently from Web logs [C]//Sun Liping, Zhang Xiuzhen. PAKDD' 00, Kyoto, Japan2000. Efficient Frequent Pattern Mining on Web Logs. APWeb 2004. [s.l.]: [s.n.], 2004:533-542.

[4] Ezeife, Lu Yi. Mining Web Log Sequential Patterns with Position Coded Pre - Order Linked WAP - Tree[J]. Data Mining and Knowledge Discovery, 2005(10):5-38.

[5] 马瑞民，李向云. Web 日志挖掘中数据预处理技术的研究[J]. 计算机工程与设计，2007(10):2358-2359.

[6] 刘造新. 基于本体的 XML 关联规则挖掘方法[J]. 计算机应用，2008(9):2319-2320.

[7] 李雪竹. 一种基于 XML 的 Web 数据抽取的实现[J]. 科学技术与工程，2008(9):2473-2474.

[8] Delphi[EB/OL]. 2009-03-21. http://baike.baidu.com/view/3297.htm. baidu, Linking-2009-03-21.

(上接第 16 页)

Computer and System Sciences, 1993, 46(1):39-59.

[40] Quafatou M. a-RST: a generalization of rough set theory [J]. Information Sciences, 2000, 124(1-4):301-306.

[41] Beyond M. Reducts within the variable precision rough sets model :a further investigation[J]. European Journal of Operational Research, 2001, 134(3):592-605.

[42] 汪小燕，杨思春. 一种新的不一致决策表属性约简算法[J]. 计算机应用，2008，28(2):525-527.

[43] 陈鑫影，邱占芝. 不协调决策信息系统的约简[J]. 计算机工程与应用，2008，44(7):193-195.

[44] 贺 鹏，王庆林. 可重构制造系统故障诊断多 Agent 自学习模型[J]. 计算机工程与设计，2007，28(8):1741-1743.

[45] 朱永利，吴立增，李雪玉. 贝叶斯分类器与粗糙集相结合的变压器综合故障诊断[J]. 中国电机工程学报，2005，25(10):159-165.

[46] 张秋娜，董双勤. 粗糙集模型和概率粗糙集模型的若干研究[J]. 重庆文理学院学报：自然科学版，2007，26(4):13-14.

[47] 刘高峰，王 飞. 基于聚类分析的粗糙集模型及其应用[J]. 内江师范学院学报，2008，23(8):28-31.

[48] 印 勇，孙如英. 基于聚类有效性分析的模糊粗糙集归纳学习方法[J]. 计算机工程，2008，34(10):86-88.

[49] 宋云雪，张传超，史永胜. 基于模糊粗糙集的飞机远程故障诊断模型研究[J]. 中国民航大学学报，2007，25(6):15-19.

[50] 庄曰平，李 伟. 基于粗糙集和模糊理论的变电站电压无功控制策略[J]. 武汉大学学报：工学版，2007，40(5):112-115.

[51] 张 明，方 敏. 基于粗糙集和小波矩的车牌字符识别[J]. 安徽建筑工业学院学报：自然科学版，2007，15(3):95-98.

[52] 车志宇，夏明革，何 友. 基于粗糙集与小波分析的图像融合算法[J]. 电光与控制，2005，12(1):18-21.

[53] 郑小霞，钱 锋. 基于粗糙决策模型和蚁群算法的故障诊断[J]. 系统工程理论与实践，2007(3):140-144.

[54] 马 昕，林丽清. 蚁群算法在面向属性的数据约简中的应用[J]. 计算机仿真，2007，24(9):158-160.