山东农业大学学报(自然科学版),2015,46(5):765-768 VOL.46 NO.5 2015
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2015.05.023

基于计算机粗糙集的数据挖掘设计与应用分析

朱小刚
南昌大学, 江西 南昌 330096

摘 要: 计算机与网络的日益普及，带来大量繁杂的数据与信息处理。数据挖掘技术能够从大量的数据中提取隐含的信息，而粗糙集是数据挖掘中最常用的方法。本文简单概括了粗糙集的基本属性，并分析了数据挖掘的设计与应用。

关键词: 数据挖掘; 粗糙集; 应用分析
中图法分类号: G202 **文献标识码:** A **文章编号:** 1000-2324(2015)05-0765-04

The Analysis on the Design and Application of Data Mining Based on Computer Rough Set

ZHU Xiao-gang
Nanchang University, Nanchang 330022, *China*

Abstract: There are a lot of data and information treatment on computer with the growing popularity of the computer networks, and these complex data and information can be extracted by the method of the rough set, one of methods in technology of data mining. This paper simply introduced the basic attribute about the rough set, and described the process of the data mining, discussed the most common algorithms , and finally analyzed the application in the rough set technology.

Keywords: Data mining; rough set; application analysis

在过去的几十年中，计算机技术和网络技术日渐普及，随之而来的还有各种各样丰富的数据资源。人们在享受着这些大量的数据所带来的便利的同时，也逐渐认识到一个严重的问题：数据丰富而知识匮乏。现在，数据库系统能够帮助我们对数据进行提取搜索修改等功能，但是我们还需要认识到数据背后所隐藏的规则和有关知识，依靠这些来进行预测[1]。

数据挖掘技术就是在这种情况下应运而生的，它是从众多繁杂的数据中寻找有用的知识的有效的方法[2,3]。数据挖掘涉及到了包括数据库、数学、计算机、人工智能等许多学科的知识[4]。数据挖掘的主要方法是粗糙集。粗糙集最初是由 Pawlak Z 提出来的，最开始是用来对不完善不准确的知识进行整理的数学方法[5]。它的基本理念[6]是将知识进行简化，提炼出问题的解决方法或者分类准则，前提是分类能力没有改变。由于粗糙集理论能够在有限的条件下对知识进行处理和提取，所以已经被应用到了机器、模式、医疗、决策等多个领域，并且在这些领域取得了相当明显的成就。粗糙集理论中主要的一个研究方向就是属性约简[7]，属性约简能够将繁杂的信息简化为规则库，来方便人们进行使用。在这一方面，国内外的许多学者都做了相关研究，但都没有取得重大突破，所以，寻快速的约简算法仍然是众多学者们研究的热点问题。

1 数据挖掘技术简介

1.1 概 念

数据挖掘，简单来说，是一个从数据到知识的转换。数据挖掘能够用比人类更迅速的方法，从大量数据中搜寻到有用的信息，并尝试建立模型，用来简单地叙述复杂的信息。

被挖掘的数据来源可以来自网络，也可以来自数据库。

数据挖掘有两种不同的学习方法：推理，归纳。推理需要结合已有信息去寻找未知信息。归纳则致力于寻找隐藏在不同数据中的模式。对于数据挖掘来讲，归纳学习是一种重要且实用的方法[8]。

数据挖掘方法主要有遗传算法、决策树方法、模糊集合论、粗糙集方法等。

1.2 数据挖掘流程

按照数据挖掘的过程可以分为以下几个主要步骤：

（1）数据准备：对于数据库中的数据，通常是通过长时间积累下来的，用来进行处理很困难。

数据准备过程就是将这些数据进行筛选、净化、推算和简化。这些过程之后会形成数据仓库。

收稿日期: 2013-06-11 **修回日期:** 2013-08-20
作者简介: 朱小刚(1978-),男,硕士研究生,副教授,研究方向:软件工程及网络安全. E-mail:hyxncdx@126.com
数字优先出版: 2015-03-31 http://www.cnki.net

这一步骤做的工作将直接影响到数据挖掘的效率，精确度和模式的效果；

（2）数据挖掘：在前一步骤所整理出来的数据仓库上进行数据挖掘。单独或者综合运用各种数据挖掘方法对数据处理，根据用户需要选择数据挖掘的算法，这些算法（分类，汇总，回归等）将用来提取数据下的模式；

（3）对结果进行分析同化：对上面得到的模型进行评估，看它是不是能够应用于实际，应当确定它是有效的可用的。评估的过程不一致，可以根据用户自己的经验，也可以根据某些数据标准来进行验证。评价并描述数据挖掘这一过程所得到的模式和规则，使得用户能够理解这些知识。

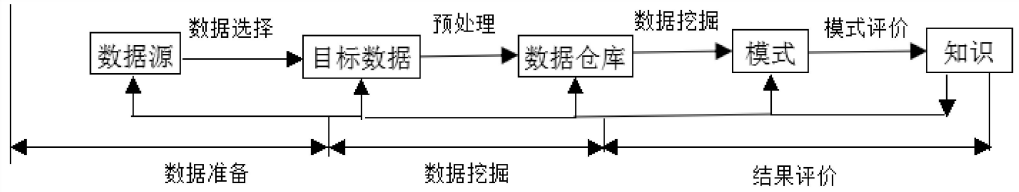

图 1 数据挖掘过程

Fig.1 Process of data mining

2　基于计算机粗糙集的数据挖掘设计

2.1　粗糙集理论

粗糙集是一种从大量数据中导出相关决策和模式的工具，能够对各种不完整的信息分析处理。

在粗糙集理论中，将对对象进行整理的能力称作知识，通常利用决策表来实现这个功能。对象可以是任何能想到的食物，比如抽象概念，数据，时间，状态等。知识必定要与许多分类模式有关，叫做论域[9]。粗糙集理论是建立在集合论上面的，所以知识表示也是基于集合论的。

定义 2.1 信息系统和信息表 将客观世界看成为一个信息系统，也叫做知识库。信息系统 S 可以用四元组来表示：$S=\{U,A,V,f\}$.

在这个四元组中，U 表示对象的有限集合，$U=\{x_1,x_2.....x_n\}$；A 是关于属性的有限集合，$A=\{a_1,a_2.....a_m\}$；V 是关于属性的值域集合，$V=\{v_1,v_2.....v_m\}$，属性 a_i 对应的值域是 v_i，f 表示信息函数，$f:A\times U\to V, f(x_i,a_j)\in v_j$。

定义 2.2 不可分辨关系 不可分辨关也被叫做等价关系，它的定义如下：

$R(B)=\{(X_1,X_2)|f(x_1,b)=f(x_2,b),b\in B\}$。

在上面的定义中，A 的子集是 B，意思是，如果有，都有 $(x_1,x_2)\in R$，对于 $x_1\in X_i, x_2\in X_j, i\neq j$，都有 $(x_i,x_j)\notin R$。$x_1,x_2\in X_i$

U 被 $R(B)$ 分成 k 个等价类，表示为 $R*(B)=\{X_1,X_2.....\mathrm{X_K}\}$。在下面的叙述中，将 $R(B)$ 记作 R，$R*(B)$ 记作 $R*$。就等价类 $\mathrm{X_i}$ 来讲，$\{\mathrm{X_i}\}$ 指的是集合 $\mathrm{X_i}$ 的基数。

在定义中，论域中不能区分的对象组成了等价类。

定义 2.3 上近似，下近似 如果有组对象 $X\subseteq U$，条件属性为 $R\subseteq C$，R 代表等价关系，那么 X 相对于 R 的下近似写成：$\underline{R}X=\{X\in U|R|[X]_R\subseteq X\}$；$X$ 相对于 R 的上近似可以写成 $\overline{R}X=\{X\in U|R|[X]_R\neq\varnothing\}$。下近似说的是，由 R 能够确定 U 中的元素组成的一定是 X 的子集；上近似指的是，根据 R 判断出，U 中的可能属于 X 的集合。有时，我们也将下近似称之为 X 的 R 正域，记作 $pos_R(X)$，$U-\underline{R}(X)$ 叫做 X 的负域，记作 $neg_R(X)$。

定义 2.4 决策系统　若一个信息系统的属性集合中包含决策属性，叫做决策系统。决策系统记作 $T=\langle U,C\quad D\rangle$，属性 A 分成集合 C 和 D，有 $A=C\quad D, C\quad D=\varnothing$，C 代表条件属性的集合，即条件维集，D 代表决策属性的集合，即决策维集。

2.2　基于粗糙集的数据挖掘模型

建立模型的过程：

(1)准备数据。从原始数据开始，选定条件属性及其对应的值域，选定结论属性及其对应的值域，同时转换原始数据，最后会得到符合定义的一个决策系统，表示为 $(U,C\quad\{d\})$。

(2)衍生节点。对 $S_{RED}(C,\{d\})$ 进行计算，将计算结果作为模型建立的初始节点，对于属性相同的

节点，放在模型的一层里。其次，将每个节点中的一个属性删掉，会得到后继节点，重复这一过程，直至遇到空节点，也就是没有条件属性的节点，如下图 2 所示。

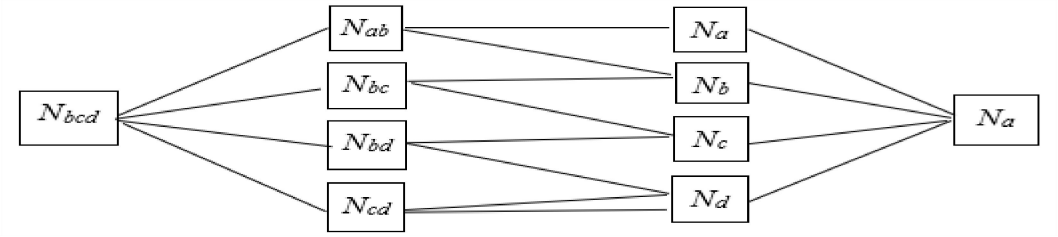

图 2 模型中各节点关系图

Fig.2 The relationship of each node in model

图中，$N_i^{\{q\}}$ 指的是第 q 层中 i 个节点，与之相对应的决策系统是 $\{U, C_i^{\{q\}} \quad \{d\}\}$

(1)节点计算，得到规则集。用 $ind(C_i^{\{q\}})$ 把 U 分成不可分辨类 $X_1, X_2, \ldots, X_i$，计算任意不可分辨类 X_i 的相对简约，可以获取简化之后的一列规则；之后，对规则粗糙所属函数 μ 进行进行，按照 $\mu > \mu_0$ 的条件，纳入节点的规则集中集中。

(2)停止条件。若某个节点中有前向节点，该前向节点的置信度小于 μ_0，那么计算在此时停止。

2.3 模型的运行

当模型创建好之后，就可以用来处理新的数据源了，运行模型的过程如下：

(1)用数据源的条件属性和模型中的节点（如图 2）相配合，若有匹配节点，那么选择第一个节点进行配合；

(2)将已经匹配上的属性值去跟节点的规则集进行匹配，如果有匹配规则，那么之后重新回到这些规则，不然，跳到步骤 3；

(3)对于该节点中全部的后续节点，都返回到步骤 2 中。

根据上面的分析，结果有下面几种形式：1）返回只有一条规则；2）返回有多条规则；3）没有任何返回规则。

第一种情况较好处理，而第三种情况表示无法给出相应解，第二种情况多有下面几种处理办法：

A.若两种规则分别归于两个节点，那么将后续节点的规则去掉；

B.根据评判算法，选取优先级比较高的规则进行返回；

C.而对于优先级的评判，可以用粗糙隶属函数或者综合评判算法[10]来判断。

如果所有的规则包括 d_1，d_2，…，d_r 个结论，得到的 d_i 的规则有 r_1，r_2，…，r_m，我们定义 $v(d_i) = \sum_{j=1}^{m} s_j / \sum_{j=1}^{m} (s_j / \mu_j)$。在式子中，规则 r_j 的支持数为 s_j，所谓的支持书，也就是在 U 中，支持 r_j 的元素个数，粗糙集隶属函数为 μ_j。

2.4 属性约简

2.4.1 基本思想 属性是否能够进行约简取决于各属性间关系的紧密程度。众多属性约简的算法的基本思想大体是一致的，都把属性重要性看成是启发式信息，最佳约简往往是从信息系统和决策系统中提取的，被叫做启发式算法。基于操作方法，可以分为两种不同的想法：前向选择法，后向删除法。目前常用的算法是前向选择法。

2.4.2 常见算法 区分矩阵 区分矩阵又被叫做可辨识矩阵，差别矩阵，是数学家 Skowron 提出的。利用区分矩阵，我们可以描述出繁杂信息系统中的所有不可区分的关系。

定义 2.5 区分矩阵 若决策表 $S = \langle U, A, V, f \rangle$，$|U| = n$，属于 S 的则是一个 $n \times n$ 的矩阵，其中的元素 $a(x,y) = \{a \in A \mid f(x,a) \neq f(y,a)\}, a \in A$。而 $a(x,y)$ 被用来对 x 和 y 的全部属性集。

定义 2.6 区分函数　运用布尔函数作为区分函数，记作 Δ，针对属性 $a \in A$，都有一个相应的布尔变量 a，如果 $a(x,y) = \{a_1, a_2, \ldots, a_k\} \neq \varnothing$，对应的为 $a_1 \vee a_2 \vee \ldots a_k$，记作 $\sum a(x,y)$；如果 $a(x,y) = \varnothing$，则对应的布尔变量是 1.将布尔函数或者区分函数 Δ 定义为：$\Delta = a(x,y)$。

性质：属性集 A 的全部约简是布尔函数 Δ 的最小析取式中的全部合取式。

分类质量 这种方法是求属性约简的时候，忽略核属性，把核看成是空集，按照属性重要度，即

分类质量来进行排列，排列的时候把重要度高的属性纳入最简集，直到最简集的分类能力和原表分类能力一样，此时终止算法。

遗传算法 遗传算法是根据达尔文遗传学理论和自然选择学说而逐渐发展起来的一种优化算法，组成算子主要有三个：选择，交叉，变异。该算法用遗传算子来处理初始群体，将初始群体引导向最优解。

3 应用领域

在数据挖掘方面，粗糙集显示出了其独特的作用，因此，近些年粗糙集的应用得到了人们的深入研究。粗糙集的应用主要在以下几个领域：

（1）多方法结合。Jelonek 学者对于粗糙集理论（Rough Set，以下简称 RS）进行了研究，将它用作对神经网络训练数据进行了属性及其支予的缩减，使得在近似分类差错率很低的情况下，提高了效率。研究人员 Hu 还发现了 RS 可以与一种归纳的思路相结合，这种方法运用概念树爬升技术泛化属性，用 RS 理论进行缩减生成所需的知识。

（2）缩减数据。生成规则 Lenarcik 等人对基于所有不相同的对象值的粗糙分类器进行了研究，思路是对每个对象重新赋予一个新的值属性。Grzymala-Busse 等人将同时对可能和确定规则以及只用确定规则进行了研究，发现后者的错误率要高于前者。

（3）粗糙逻辑。很多学者都致力于基于 RS 的 Rough 逻辑。Lin，Liu 等人根据拓扑学原理给出了上近似算子 H 和下近似算子 L 的概念，前者与模态逻辑中的必然算子十分接近，后者与模态逻辑中的必然算子十分接近。含有 L 和 H 两个算子的相应公式被叫做 Rough 逻辑公式。同时，与模态逻辑类似的理论化的 Rough 集的逻辑演绎系统和规则也被建立。Yao，Lin 对模态逻辑和 RS 进行了讨论，得到一些关于扩展粗糙集的特性。

（4）大数据集。基于关联规则挖掘算法理论，有些研究学者将其用来确认粗糙集和生成相关的规则，将粗糙集在数据挖掘中的繁杂计算进行简化。Nfuyen 等人提出了一种新的决策表分解办法。重复递归这个过程，直到得到我们所需要的决策表为止。关于最后生成的小决策表，生成相应的规则。有新对象出现的时候，从顶部一直匹配到叶子。

4 小 结

(1)数据挖掘技术中运用最多的方法就是粗糙集。粗糙集是一种从大量数据中找出潜在规则的工具。粗糙集中比较重要的内容是知识约简，知识约简的过程就是对信息系统中的相关信息进行分析，删除冗余信息的过程；

(2)数据挖掘模型的建立过程包括数据的准备、节点的衍生和计算、停止条件几个步骤。模型建立好之后，就可以进行运行了；

(3)粗糙集的应用十分广泛，目前，学者们发现可以运用粗糙集进行多方法结合、大数据集、缩减数据和形成规则、粗糙逻辑等操作处理。

参考文献

[1] Shortland R, Scarfe R. Digging for Gold[J]. IEE Review, 199541(5):213-217
[2] 曾黄麟.粗集理论及其应用[D].重庆:重庆大学出版社,1998
[3] 韩家伟.数据挖掘中的知识分类[C]//1999 年数据挖掘论文集.上海:复旦大学出版社,1999:16-18
[4] Pauray S M Tsai, Chien-Ming Chen. Discovering Knowledge from Large Databases Using Prestored Information [J], Information systems, 2001,26(1):1-14
[5] Pawlak Z. Rough Sets: theoretical aspects of reasoning about data[M]. Netherlands:Kluwer Academic Publishers,1991
[6] Wojciech P Ziarko. Rough Sets, Fuzzy Sets and Knowledge Discovery: Proceedings International Workshop on Rough Set and Knowledge Discovery[C]. Berlin: Springer-Verlag Berlin and Heidelberg GmbH &Co.K,1993
[7] Bautista R, Millan M, Diaz J F. An Efficient Implementation to Calculate Relative Core and Reducts[C]. New York:18th International Conference of the North American on Fuzzy Information Processing Society, 1999
[8] Pete C, Julian C, Randy K, *et al*. The CRISP-DM 1.0[M]. USA:SPSS,2000
[9] 张文修,吴伟志,李德玉.粗糙集理论与方法[D].北京:科学出版社,2001
[10] Aasheim O T, Solheim H G. Rough sets as a framework for data mining[R]. Trondheim:Norwegian University of Science and Technology, 1996