基于概念格的数据挖掘理论与方法研究

王璐

（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要： 随着图像数据获取设备和获取手段的迅速发展，我们获取了海量的图像数据，如何充分地利用这些图像数据，从图像数据中挖掘出隐含的、潜在的规律性的知识，是目前迫切需要解决的问题。本文对图像(遥感图像)数据挖掘与知识发现这一新的概念的内涵和外延进行了系统地深入地分析和研究，将这一概念解释为“利用空间数据挖掘的理论和方法(空间聚类分析、空间关联规则分析、空间序列分析等)从图像库(或多幅图像、一幅图像的多个分块)中提取出规律性的潜在的有用的信息、图像数据关系、空间模式等，自动抽取出具有语义意义的信息(知识)，从而为图像的智能化处理服务的过程”，强调这个概念是一个动态的概念，是一个过程，其目的是为图像的智能化处理服务，可以对大量的图像数据库进行挖掘，也可以只对一幅图像进行挖掘，它是在其它相关技术的基础上发展起来的，由于还处于初期阶段，与这些相关技术之间的区别有时候可能还不是很明显。本文认为图像数据挖掘是一个具有自己的独特的研究内容的、具有自己的理论和技术框架的一门新的理论和技术。本文对这一概念与其它相关概念之间的关系进行了分析和对比，对图像数据挖掘的研究内容和研究体系进行了界定。

关键词 数据挖掘 ； 概念格

中图法分类号 TP311 **文献标识码** A

Research on theory and method of data mining based on concept lattice

WangLu

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract With the rapidly development of instruments and methods to obtain image data, great capacity for image data can be obtained. How to utilize these image data, how to get implicit, underlying,disciplinarian knowledge are urgent problems to resolve. This paper systemically, deeply analyzed and researched the intension and extension of the new concept of image(remote sensing image) data mining and knowledge discovery. Image data mining and knowledge discovery is the process of utilizing the theories and methods of spatial data mining, such as spatial clustering analysis, spatial association rule analysis, spatial serial analysis etc., to extract regular implicit useful information, image data relationships, spatial pattern etc. from image databases, or mufti images, or mufti sections of one image”the concept is dynamic concept, is a process, its aim is to promote the intelligentization of image processing, image mining can be done in image databases, or in one image, it developed based on some related technologies, the differences among image mining and these technologies are still not obvious.The paper thought image mining as a new theory and technology, it has its own particular content, own theories and technologies. The paper compared the concept of image mining researching with other related concepts, defined the research content and research system of image mining.

Keywords Data mining;Concept Lattice

1 数据挖掘与概念形成

1.1 数据挖掘与概念形成

数据挖掘首次出现在 1989 年 8 月举行的第十一届国际联合人工智能学术会议上。

收稿日期：2016-10-9
作者简介：王璐（1994-）女，硕士生在读。

数据挖掘被定义为:“从数据中发现隐含的、先前不知道的、潜在有用的信息的非平凡过程”。数据挖掘的过程，即从数据库中发现知识的过程，可以理解为从数据库中形成概念的过程。

概念形成是人脑学习的一个重要特征，从概念形成去探讨人脑学习，从而探讨通过数据库中大量的数据的学习从而产生概念和知识的过程，被认为是一个行之有效的途径。所谓概念，就是在头脑里所形成的反映对象的本质属性的思维形式。把所感知的事物的共同本质特点抽象出来，加以概括，就成为概念，概念都具内涵和外延，并且随着主观、客观世界的发展而变化。如果能够建立一种数学的形式化的数据结构将概念的内涵和外延以及概念与概念之间的不同层次的抽象关系表达出来，将可以对数据挖掘和知识发现的过程进行有效地分析和处理。形式概念分析理论正是提供了这样的一个工具，形式概念分析理论所形成的形式化体系可以很好地利用数学的方法描述概念的形成过程。

1.2 数据挖掘的过程

数据挖掘和知识发现的过程可以根据状态空间理论作一个很好的解释。李德毅教授提出了以发现状态空间理论作为 KDD 的总体框架。

发现状态空间是一个三维立体空间，是发现系统实施多种算法的运作空间。在一个二维的平面基底—知识基上逐步抽象，关系数据库可以抽象地看成一个二维通用大表，纵向为属性，横向为元组。根据知识发现任务，在原始的数据库经过查询、选择(或抽样)、统计和压缩等数据聚焦处理后，形成宏元组，是发现状态空间的基底，也可以认为是初始的知识模板。在发现状态空间进行多种知识汇集和发现操作。模板方向，即面向知识模板的操作，是从微观到宏观的发现知识的操作。由一块知识模板上升到抽象级别更高的另一块模板，是提高知识抽象度的操作，是以归纳为核心的知识发现活动。对于空间数据挖掘对应的状态空间，还增加了一个尺度维，在尺度维上表达了空间数据由细到粗多比例尺或多分辨率的几何变换过程。面向尺度的操作是对空间数据由细到粗的计算、变换、概括、综合过程。

数据挖掘也可以看成一个从不同的视角，从低层到高层，不断地抽象，不断地产生抽象层次更高的概念，从而产生知识的过程。

2 关联规则挖掘

2.1 概述

关联规则是数据挖掘的最重要内容之一，目前的大多数数据挖掘的理论和技术的研究也主要是针对关联规则来进行的。关联规则的概念由 Agrawal, Imielinski, Swami 提出，是数据中一种简单但很实用的规则，关联规则的模式属于描述型的模式，关联规则挖掘的目的是发现大量数据中项集之间有趣的关联或相互关系。

考察一些涉及许多物品的事务，例如，事务 1 中出现了物品甲，事务 2 中出现了物品乙，事务 3 中则同时出现了物品甲和乙。那么，物品甲和乙在事务中的出现相互之间是否有规律可循呢?在数据库的知识发现中，关联规则就是描述这种在一个事务中物品之间同时出现的规律的知识模式。更确切的说，关联规则通过量化的数字描述物品甲的出现对物品乙的出现有多大的影响。

下面给出关联规则的具体的形式化的定义。

关联规则:设 $I=\{i_1,i_2,i_3,\ldots\ldots i_n\}$是项的集合。设任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的集合，使得 $T\subseteq I$。每一个事务有一个标识符，称作 TID。设 A 是一个项集，事物 T 包含 A，当且仅当 $A\subseteq T$。关联规则是形如 $A\rightarrow B$ 的蕴涵式，其中 $A\subseteq I$，$B\subseteq I$,B 之间的交集为空。

可以通过以下四个指标反映关联规则的属性:

(1)支持度(Support)

关联规则 $A\Rightarrow B$ 的支持度就是指:在事务集 D 中包含 AUB(即同时包含 A 和 B)的事务的百分比，它是概率 $P(A\cup B)$，即:$support(A\Rightarrow B)=P(A\cup B)$ 。

支持度描述了项集 A 和项集 B 的并集 C 在事物集 D 中出现的概率有多大。

（2）置信度(Confidence)

关联规则A⇒ B的置信度就是指:在事务集D中包含A的事务同时也包含B的百分比，这是一种条件概率P(BlA)，即:confidence(A⇒ B)=P(BlA) 。

（3）期望置信度(Expected Confidence)在事务集D中包含项集B的事务所占的百分比，即:Expected Confidence(B)=P(B) o 期望置信度描述了在没有任何条件影响下，含有项集 B 的事务在所有的事务中出现的概率有多大。

(4)作用度((Lift)

作用度是置信度与期望置信度的比值。即：Lift (BlA) = confidence(A⇒ B)l ExpectedConfidence(B)。

作用度描述了项集A的出现对项集B的出现有多人的影响。作用度越人，说明项集A对项集B的影响越大。一般情况卜，有用的关联规则的作用度都大于1，说明关联规则的置信度人于期望置信度，说明A的出现对B的出现有促进作用，也说明了它们之间某种程度的相关性，如果作用度不大于1，则此关联规则就没有什么意义了。

支持度是对关联规则的重要性的衡量，置信度是对关联规则的准确度的衡量。支持度说明了这条规则在所有事务中有多大的代表性，支持度越大，表示这种关联规则越重要。有些关联规则的置信度虽然很高，但是支持度很小，说明该关联规则出现的机会很小，实用价值不大，因此，也不重要。

有趣规则:什么样的规则是有趣的、有意义呢?根据关联规则的定义可以看出，任意两个项集之间都存在关联规则，如果不考虑关联规则的属性指标值的大小，那么，在事务集D中可以发现很多很多的规则，但是并不是所有的规则都是有意义的。

规则的理解:利用各种关联规则的挖掘算法可以挖掘出相应的关联规则，但是如果需要了解这些规则的具体的意义，还需要结合具体的领域，根据领域相关知识进行解释。下面介绍关联规则挖掘的具体步骤，首先介绍两个相关的概念:

(1) k 项集:包含 k 个项的项集称为 k-项集，所谓项集就是项的集合。

(2)频繁项集:如果在事务集D中项集出现的频率大于或等于 min sup 与 D 中事务总数的乘积，则称它为频繁项集(frequent itemset)。

关联规则挖掘的主要过程包括以下几个步骤:

(l)准备数据;

(2)设定最小支持度值和最大置信度值:

(3)根据数据挖掘的算法找出所有支持度大于或等于最小支持度闽值的所有频繁项集;

(4)根据频繁项集生成所有置信度大于或等于置信度闽值的有趣规则(强规则);

(5)如果生成的规则过多或者是过少，则需要对支持度阂值和置信度阂值进行调整，并重新生成强关联规则;

(6)关联规则的理解，挖掘出关联规则以后，还需要结合领域相关知识对规则的意义进行解释、进行理解，这样才能体现出数据挖掘概念的挖掘出有意义的规则的含义。

在这几个步骤中，最繁杂、最耗时的工作是第三步即生成频繁项集的工作;第四步根据频繁项集生成关联规则的工作相对简单，但是如何避免过多的、冗余的规则的生成也是需要认真考虑的。其它的步骤可以认为是一些相关的辅助性的步骤。

2.2 Apriori 算法

1 算法概述

Apriori 算法是一种使用频繁项集的先验知识从而生成关联规则的一种算法，也是最有影响的关联规则挖掘算法，最早由Agrawal 等人提出。

该算法的基本过程如下:

(1)首先计算所有的 C_l;

(2)扫描数据库，删除其中的非频繁子集，生成 L_1(1_频繁项集);

(3)将 L，与自己连接生成 C_2 候选 2 一项集);

(4)扫描数据库，删除 C:中的非频繁子集，生成 L_2(2_频繁项集);

(5)依此类推，通过从 L_{k-1} (k-1 一频繁项集)与白己连接生成 C_k(候选 k-项集)，然后扫描数据库，生成 L_K 繁 k-项集)，直到不再有频繁项集产生为止。

在连接过程中，为了连接的方便，将项

集(项目的集合)中的项按照辞典序排列。执行连接 Lk 与 L、的连接时，如果某两个元素的前 k-1 项相同，则认为二者是可连接的，否则，认为二者是不可连接的，不作处理。

3 概念格理论研究

概念格理论，也称形式概念分析理论，首先由德国的数学家于 1982 年提出。形式概念分析理论是一种基于概念和概念层次的数学化的表达的应用数学的一个分支。因此，在应川形式概念分析理论时，需要用数学的思维方式进行概念数据分析和知识的处理。形式概念分析中的“形式”一词表示我们正在处理数学领域的工作，通过与这些工作相联系的结构化的概念的联系，发现可理解的、有意义的知识。

形式概念分析理论用来进行数据表达的基本形式就是交义表(cross table)，用来表示形式背景(formal context)。形式概念分析理论可以用来表示单值属性背景，也可用来表示更为复杂的数据类型，如多值属性背景，对于多值属性背景，一种处理方式是通过概念定标(conceptual scaling)的方法将其转换为基本的类型(即单值属性背景)，另外一种处理方式就是不转换，直接根据低层概念之间的关系抽象出高层的概念。形式概念分析理论的基本概念是形式背景(formal context)和形式概念(formal concept)。

形式概念分析理论是基于数学中的偏序理论(partial ordering theory)的，特别是基于完备格的理论，因此，在介绍形式概念分析理论之前，首先简单介绍一下格理论的相关知识。

4 概念格的化简

一般的概念格的复杂性为 $O(2^n)$，在具体的建格过程中应设法减少复杂性，在此对这个问题进行简单的介绍。一般可以从以下几个方面着手实现概念格的化简。

(1)通过选择那些能够反映兴趣焦点的属性集合，将属性集合 M 减少为 M'。这也可以叫做概念的内涵缩减。

(2)减少对象的数量，当数据对象的数量十分巨人时，尽可能地从大量的数据集合中选择有代表性的样本，将这些样本作为对象，从而构建概念格。

(3)通过对属性的分布的分析，寻找有效的编码技巧，从而减少空间复杂性。

对于复杂的和大型的应用，建立概念格时可能会产生大量的数对，涉及大量的处理和存储资源，因此，必须采取剪枝策略，减少格节点的数量。例如，只生成出现的频率大于支持度 I}I 值的频繁概念格 T}点，这时所生成的格称为半格或者部分格。

当数据很复杂时，所画出的 Hasse 图可能非常复杂，以至于难以观察。一个可行的方法是通过嵌套的 Hasse 图来减少视觉上的复杂性。其基本方法就是对概念格本身进行分析，寻找一些相对关系密切的概念子格，将概念子格作为一个整体，利用一个更宏观的概念节点来表示，构建嵌套的概念格.从而实现嵌套的从 Hasse 图的绘制。

5 本章小结

概念格理论是一个利用数学的形式化的方法描述概念形成过程的形式化的方法，可以作为规则表达的自然基础。本章详细分析讨论了概念格的基本理论;分析了基于概念格的知识的表达与处理方法.章对基于概念格的数据挖掘的原理进行了分析，研究讨论了相关的数据挖掘算法，研究出两种集概念格的构建和 Hasse。图的绘制为一体的快速算法，其中第二种算法通过建立基于辞典序的索引树的方法建立格节点的索引，并将概念格节点根据内涵基数进行分层存放，设计了一个统一的数据结构表示概念格节点，同时确定概念格节点的坐标及其父子关系，有效地解决了在构建概念格的同时进行 Hasse 图的绘制的问题。

参考文献

[1]李德仁，程涛，从 GIS 数据库中发现知识，测绘学报，1995, 22 (4) :37-430

[2]李德仁，关泽群，空间信息系统的集成与实现。武汉:武汉测绘科技大学出版社，2000 年:39-590

[3]李德仁，王树良，史文中，王新洲，论空间数据挖掘和知识发现，武汉大学学报(信

息科学版)，2001 年 06 期:491-4990
[4]李德仁，王树良，李德毅，王新洲，论空间数据挖掘和知识发现的理论与方法，武汉大学学报(信息科学版)，2002 年 03 期:221-2330
[5]李德毅，发现状态空间理论，小型微型计算机系统，1994, 15 (11) :1-60
[6]李德毅，邸凯昌，李德仁，史雪梅，用语言云模型发掘关联规则(英文)，软件学报，2000 年 02 期:143-1580
[7]邸凯昌，李德仁，KDD 技术及其在 GIS 中的应用与扩展，中国地理信息系统协会第二届年会论文集，北京，1996 年 9 月:1-90
[8]邸凯昌，李德仁，李德毅，空间数据发掘和知识发现的框架，第十届全国遥感技术学术交流会，青岛，1997 年 7 月，武汉测绘科技大学学报，Vol. 22, No. 4, 1997: 328-3320
[9]邸凯昌，李德仁，李德毅，从空间数据库发现聚类:一种基于数学形态学的算法，中国图象图形学报，1998 年 03 期:173-1780
[10]邸凯昌，空间数据发掘和知识发现的理论和方法:「博士论文」，武汉:武汉测绘科技大学，1999 年:1-1190