计算机科学1997 Vol. 24 №. 5

知识发现的若干问题及应用研究*)

王清毅 陈恩红 蔡庆生

(中国科技大学计算机系 合肥230027)

摘 要 This paper first discusses several issues on knowledge discovery, such as the knowledge discovery process, the methods and algorithms of data mining and discovering association rules. Then some practical development tools and systems of knowledge discovery are given. The paper concludes with the future research work of knowledge discovery.

关键词 Knowledge discovery, KDD, Data mining, Association rule

知识发现是一个众多学科诸如人工智能、机器学习、模式识别、统计学、数据库和知识库、数据可视化等相互交叉、融合所形成的一个新兴的且具有广阔应用前景的领域。目前国际上对知识发现的研究与开发进展很快。第一和第二届数据库中知识发现的国际会议(KDD95)和(KDD96)已经召开,第三届也于97年8月在美国的加州举行。第一本关于知识发现的国际学术杂志《Data Mining and Knowledge Discovery》于97年3月创刊。众多领域的知识发现系统和工具也不断投入市场。

知识发现的研究始于从数据库中发现有用的模式这一概念并先后有着不同的术语,如数据挖掘,知识提取,信息发现,数据模式处理,数据考古学以及数据库中的知识发现。其中数据挖掘术语多为统计学家、数据分析学家及管理信息系统团体采用;而数据库中的知识发现术语则是在于89年的第一届KDD专题讨论会上被首次采用。这一术语强调了知识是数据发现的最终产品并很快在人工智能和机器学习领域得到广泛应用。

长期以来,在知识发现领域这两个术语的范畴和使用界限一直不很清晰。直到KDD96国际会议上,知识发现研究领域的知名学者Fayyad, Piatetsky-Shapiro和Smyth就这两个术语的关系作了如下阐述:KDD是指从数据库中发现知识的全部过程,而Data Mining则是此全部过程中的一个特定步骤。

一、知识发现的一般过程

在KDD96国际会议上,Fayyad, Piatetsky-Shapiro和Smyth对KDD下了最新定义:

定义1 KDD是识别出存在于数据库中有效的、新颖的、具有潜在效用的乃至最终可理解的模式的非平凡的过程。

一般说来,知识发现的过程可描述如下:

(1)熟悉应用领域的数据、背景知识,明确所要完成的KDD发现任务性质,例如是要验证系统的假设还是要发现新的数据模式。

(2)确定与发现任务相关的数据集合。

(3)从与发现任务相关的数据集合中除去明显错误的数据和冗余的数据。

(4)数据预处理,即创建知识发现算法所需要的数据组织形式,如概念层次。如果发现任务属于验证,则还需提出相应的假设。

(5)针对所要发现任务的所属类别,例如归类、回归分析、简约、聚类、发现关联规则等等,设计或选择有效的数据挖掘算法并加以实现。

(6)测试和评价所发现的知识。例如对知识一致性、新颖性和效用性的处理。

(7)解释和运用指的是对所发现的知识进行解释以及在实际系统中的应用。

以上各个过程都涉及数据可视化技术。整个发现过程也不是简单的线性流程,步骤之间包含了循环和反复。这样可以对所发现的知识不断求精、深化

*)得到国家自然科学基金和国家教委博士点基金的支持。 王清毅 博士研究生,研究方向为机器学习、知识发现。陈恩红 博士,研究方向为机器学习,知识发现,约束满足问题。蔡庆生 博士生导师,研究方向为人工智能、机器学习、知识发现。

并使其易于理解。

二、知识发现的方法和算法

长期以来，对知识发现方法和算法的研究主要集中在数据挖掘阶段，但近两年对关联规则的发现研究正逐渐增多。本节首先讨论数据挖掘的方法和算法，然后再讨论关联规则的发现方法和算法。

2.1 数据挖掘的方法

大部分的数据挖掘方法都是基于机器学习、模式识别以及统计学三个领域中的方法。具体说来，它们是：

·归类。对数据库中的每一类数据，挖掘出关于该类数据的描述或模型。已有许多数据分类方法，如决策树(ID3和C4.5)方法、统计方法、神经网络方法和粗集方法等。最近，一些面向数据库的分类方法已经开始由 Mehta，Agrawal，& Rissanen 等人开始研究。J. Han 等人在他们开发的知识发现系统 DBMiner 中采用了基于概括的决策树方法，该方法集成了面向属性的归纳和决策树归纳技术。

·回归分析。利用回归分析的方法产生一个将数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系。

·聚类。识别出一组聚类规则，将数据分成若干类。主要采用可能性稠密度估计法。

·简约。给出一个数据子集，寻求对它的紧致描述。例如可以采用关联规则、可视化技术以及商用图表等来描述。

·构造依赖关系，构造一个描述变量之间函数依赖关系或相关关系的模型。例如信念网络。

·变化和偏差分析，偏差包括很大一类潜在有趣的知识，如分类中的反常实例，模式的例外，观测结果对期望的偏离以及量值随时间的变化等等。该方法的基本思想是寻找观察结果与参照量之间有意义的差别。观察可以是一组变项值的某个模式，参照可以是给定模型的预测、外界提供的标准量或另一个观察。

2.2 数据挖掘的算法

数据挖掘算法是对上述数据挖掘方法的具体实现。所有数据挖掘算法都含有以下三个构件(下文中的模型是指从数据库中所发现的模型)：

(1)模型表示。用于描述要发现的模型的语言。如果语言的描述能力较强，就有助于发现精确的数据模型。但要注意的是，能力过强的描述语言却有可能导致所发现的模型的过分一般化，降低了预测的精确度。常用的模型表示方法有决策树、非线性回归、基于事例的推理、贝叶斯网络和归纳逻辑程序设计等方法。

(2)模型评价标准。对一个所发现的模型在多大程度上符合发现目的的要求作出定量的评价。对预测类的模型，可以利用一些测试数据集来评价其精确度。对描述类的模型，可以在精确度、新颖性、实用性以及可理解性等多个方面进行评价。

(3)发现方法。分为参量发现和模型发现。在上述模型表示和模型评价标准被确定之后，数据挖掘就完全变成了一个优化任务，即从数据的描述中发现最适合评价标准的参量或模型。具体说来，参量发现就是在确定数据集合和模型表示之后，寻找最适合模型评价标准的参量；模型发现是一个循环的试探过程，需要不断更改模型表示，最后确定出恰当数量的模型。

一般说来，不存在一个普遍适用的算法。一个算法在某个领域非常有效，但在另一个领域却可能不太适合。例如，决策树分类法在问题维数高的领域可以得到很好的分类结果，但对数据类之间的决策分界采用二次多项式描述的分类问题却不太适用。因此在实际应用中，要针对特定的领域，精心选择有效的数据挖掘算法。换句话说，实际的开发工作往往转变成对领域问题、领域知识和发现任务的形式化，而不是对所选择的数据挖掘算法进行细节上的优化。

2.3 关联规则的发现方法

从数据库中发现关联规则的方法和算法近几年研究较多。所谓关联规则就是描述数据库中数据项(属性，变量)之间所存在的(潜在)关系的规则。例如，“百分之八十的顾客如果购买了牛奶，那就一定会再购买面包”，用关联规则表示成“milk→bread (80%)”，其中80%是规则的信念因子。发现关联规则的意义如下例所述：①发现所有以 bread 作为后件的关联规则。这将有助于商场决策者采取相应措施来促进 bread 的销售。②发现前件中含有 milk 的关联规则。这将使得商场决策者了解如果中止销售 milk 将会影响其它什么商品的销售。③发现商场里货架 A 上和货架 B 上的商品之间的关联规则。这会帮助确定两个货架上商品之间的销售关系。

目前已经从单一概念层次关联规则的发现发展到了多个概念层次的关联规则的发现。例如在上述关联规则的层次下还存在一条事实，即“百分之八十的顾客如果买了小麦面包，就会再买巧克力牛奶”，用关联规则表示为：wheat_bread→chocalate_milk (80%)。这样在概念层次上一层层往下，从一般到具

体，发现的关联规则所提供的信息就越具体，越明了。这实际上也是个逐步深化所发现的知识的过程。下面我们以一个交易数据库为例，讨论一下文[4]给出的关联规则的发现方法。假设该交易数据库由交易数据集 T 和数据项描述集 D 组成，其中 T 是交易 $(T_i,\{A_p,\cdots,A_q\})$ 的集合，T_i 是交易标志；D 是 T 中每一个数据项的描述的集合，形式为 $(A_i,\text{description}_i)$。下面给出一些定义。

定义2 一个数据项集合 A 是一个数据项的集合 $\{A_i|A_i\in T\}$。集合 S 中数据项集合 A 的支持 $\sigma(A/S)$ 是包含 A 的交易的数目与 S 中的交易的数目之比。集合 S 中规则 A→B 的信念 $\varphi(A\to B)$ 是 $\sigma(A\cup B/S)$ 与 $\sigma(A/S)$ 之比，即指当数据项集合 A 在 S 中出现时数据项集合 B 也在 S 中出现的可能性。

定义3 集合 S 中的第 l 概念层的数据项集合 A 是大的（出现频率高），如果 A 的支持不小于该层的最小支持阈值 σ_l。第 l 概念层次的规则“A→B/S”是高的，如果规则的信念因子不小于该层的最小信念阈值 φ_l。

此定义中的最小支持阈值和最小信念阈值可由用户或领域专家给出。

定义4 一条规则“A→B/S”是强的，如果1）数据项集合 A 和 B 中的每一项的每一个祖先在各自相应的概念层次上是大的；2）“A∪B/S”在当前层次上是大的；3）“A→B/S”在当前层次上是高的。

定义表明关联规则的发现过程是沿着每一概念层次上的大的数据项集合由上至下一层层进行的，从而避免了对统计意义不大的数据项集的检测。

假定交易数据库由两种关系组成，一是销售商品的属性描述表；二是销售交易表。

表1 销售商品的属性描述

bar_code	category	trade-mark	content	price(rmb)	storage_period(days)
19971	milk	shang_hai	chocalate	4.00	90
…	…	…	…	…	…
…	…	…	…	…	…

表2 销售交易表

transaction_id	bar_code_set
345345	{19971,56723,12364,88973,…}
981245	{87612,34907,67134,90654,…}
…	{…,…}

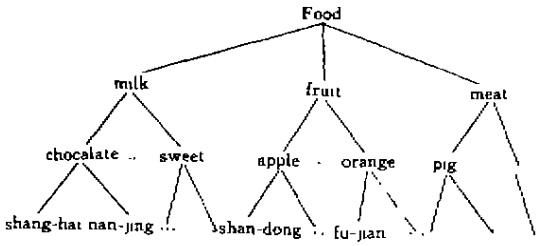

图1 概念层次树

再假设要发现的关联规则是关于保质期不超过180天的食品的销售情况，并且只观察食品的类别（牛奶，面包，肉，蔬菜，水果等）、食品成份和商标三个属性，则可以通过下面的 DMQL 查询语句首先进行检索：

```
find association rules from sales_transactions
    T,sales_item I
where T.bar_code=I.bar_code and I.category
    ="food" and I.storage_period<180
with interested attributes category, content,
    trade-mark
```

由检索结果，可以得到表3所示的一般化的销售商品属性描述表，其中每一个元组代表了具有相同的 category，content 和 trade-mark 属性值的元组的合并，在条码集属性中表明了被合并的元组的条码值。

表3 一般化的销售商品属性描述

TID	bar_code_set	category	content	trade-mark
112	{19971,53245,67532}	milk	chocalate	shang_hai
131	{52134,23145,99331}	milk	sweet	hang_zhou
211	{23157,98345,86354}	fruit	apple	shan_dong
311	{54328,43218,65432}	meat	pig	he_fei
…	{…,…}	…	…	…

这样，低的概层次上的关联规则的发现就只和这三个属性相关。如果以 category 作为第一概念层次，以 content 作为第二概念层次，以 trade-mark 作为第三概念层次，表3实际上就蕴含了一个如图1所示的概念层次树。

首先在第一概念层次上发现大的单一数据项，设该层的最小数据项支持为10%，例如可以发现大的单一数据项：milk(20%)，bread(30%)，…；再发现大的双数据项：milk，bread (15%)，milk meat (18%)，…；最后再发现关联规则：milk→bread (20%)，bread→meat(17%)，…。

在第二概念层次上，只有那些在第一概念层次上双亲是大的数据项的数据项才被检测，利用同样的方法可以在第二概念层次上继续发现大的单一数据项：chocalate(15%)，sweet(20%)，…；大的双数据

项：chocalate sweet(20%)，chocalate apple(18%)，…；和关联规则：chocalate→sweet(15%)，sweet→apple(23%)，…。

以上过程反复进行，直到概念层次的最底层，这样就一步步发现了越来越具体的关联规则。

2.4 关联规则的发现算法

基于上节讨论的思想和方法，研究者们已经提出了许多关联规则的发现算法。Agrawal 等人几年前提出了关联规则的发现算法 AIS 和 SETM，94年又提出了改进的算法 Apriori 和 AprioriTid，后两个算法与前两个算法的不同在于：在对数据库的一次遍历中，哪些候选数据项集被计数以及产生候选数据项集的方法。AIS 和 SETM 算法都是在将交易数据读入数据库的过程中迅速生成候选数据项集。在读入新的交易数据之后，就要决定前次过程中的大的数据项集中的哪些应该和这些读入的交易数据中的数据组合，以产生新的候选数据项集。这种方法的缺点是会导致许多不必要的数据项集的生成和计数。而 Apriori 和 AprioriTid 算法只利用前次过程中生成的大的数据项集来生成新的候选数据项集。因此，具有 K 个数据项的候选数据项集可以通过组合具有 K－1个数据项的大数据项集来生成，删除了那些不大的数据项集。所产生的候选数据项集要小得多，提高了算法的效率。文[4]中给出了多层关联规则的发现算法 ML_T2L1，ML_T1LA，ML_TML1 和 ML_T2LA，这几个算法与 Agrawal 等人的算法不同在于采用了新的优化技术。本人所在的研究小组也在文[12]中提出了对 Agrawal 等人的算法的改进算法。

三、实际的知识发现工具和应用系统

目前国外已在市场、金融、工业、医疗保健和科学研究等领域开发了众多的知识发现工具和应用系统，并投入市场。下面我们介绍几个典型实例。

3.1 知识发现工具

目前的知识发现工具一般分为三类：(1)通用单任务类。这类工具已开发出许多，主要采用了决策树，神经网络，基于例子和基于规则的方法，所用的发现任务大多属于归类范畴。在具体应用中，主要用于知识发现的数据挖掘步骤，而且需要相当工作量的预处理和后处理。(2)通用多任务类。这类工具可以执行多个领域的知识发现任务。一般集成了归类、可视化、聚类和简约等多项发现策略。例如 Clementine，IBM Intelligent Miner，SGI MineSet 等系统。(3)面向专门领域类。即用于专门领域的知识发现，这类工具包括用来产生反映市场零售额变化情况分析报告的 Opportunity Explore(Anand 1995)和分析篮球比赛统计数字寻求最佳阵容的 IBM Advanced Scout 等等。

3.2 应用

下面是一些知识发现的典型应用领域。

·市场 Integral Solutions 为 BBC 开发了一个采用神经网络和规则归纳方法来预测收视率的发现系统。其中规则归纳用于确定在收视率和节目编排的关系中，哪些因素起着最重要的作用。最终开发的系统的预测能力已等价于人类专家。再一个就是 Spotlight 系统，该系统采用自然语言和商用图表来分析超级市场销售额数据，确定销售渠道和价格变化之间的因果关系。该系统在1995年被扩充到了 Opportunit Explorer 系统中，其它的市场应用领域是市场购物分析，目的是发现顾客所购的不同商品之间的关联。许多公司包括 IBM 和 SGI 都提供了这样的发现系统。

·工业 KDD 在工业界的应用目标是要发现最佳生产条件。一个典型例子就是正在欧洲一家化学公司运行的 KDD 系统。该系统分析一家聚合塑料工厂的生产过程，所分析的数据包括控制变量、过程变量和质量变量。再一个就是 CASSIOPEE 故障发现系统，该系统被欧洲三大航空公司用来诊断和预测波音737飞机的故障问题，其应用获得了"欧洲创新应用"的一等奖。

·金融 这类应用系统大都采用统计回归或神经网络的方法来构造预测模型。例如 Morgan Stanloy 等人已经开发了 AI(Automated Investor)系统。该系统通过采用聚类、可视化和预测技术来寻求最佳投资时机。Daiwa Securities 利用 MATLAB 工具包建立了一个有价证券管理系统，旨在分析大量的证券数据。

·科学研究 在生物界，开发了 HMMs 和 SAM 两个发现系统，已被用于基因发现和构造核糖核酸模型。在地学领域也开发了 Quakfinder 系统，通过利用卫星采集的数据来监测地壳的构造活动。系统采用了统计推理、大规模并行计算以及全局优化的技术。微软公司的 Fayyad 等人已经开发了 SKICAT—宇宙图象分类和分析系统。该系统采用决策树学习算法来确定每一个宇宙客体的所属类别(例如是否属银河系)。

·医疗保健 1996 年，Matheus，Piatetsky-Shapiro 和 McNeill 等人开发了该领域的第一个知

识发现系统—KEFIR。该系统在多个方向上自动进行数据挖掘，确定特性的定量测量值和原先值以及期望值的偏差，从中提取有用的信息。KEFIR 区别于其它系统的一个先进之处是它可以通过 Web 浏览器采用自然语言和商用图表以超文本的形式报告它的发现。

四、进一步的研究方向

在知识发现的研究和开发已经取得了令人瞩目的进展的同时，一个亟待解决的课题也摆在了研究者的面前。

·**效率和可扩放性** 海量数据库中存有成百个属性和表，成百万个元组。GB 数量级的数据库已不鲜见，TB 数量级的数据库也开始出现。这就必然导致海量数据库中问题的维数很大，不仅增大了发现算法的搜索空间，也增加了盲目发现的可能性。因此，必须利用领域知识除去与发现任务无关的数据，有效地降低问题的维数，设计出更加有效的知识发现算法。

·**数据的时序性** 应用领域的数据库中的数据大都是动态随时间变化的，可能使得原先发现的知识失去效用。但从另一角度来看，数据的时序性又为开发强有力的知识发现系统提供了潜在的舞台，因为重新训练一个系统要比重新训练一个人(即改变他的思维、观点等)要容易得多。Matheus 等人已经提出采用对所发现的模式随时间逐步修正和数据的变化来指导以后的发现过程。

·**和其它系统的集成** 一个方法、功能单一的发现系统其适用范围必然受到限制，而且开发的知识发现系统仅局限于数据库领域。然而要在更广阔的领域发现知识，知识发现系统就应该是数据库、知识库、专家系统、决策支持系统、可视化工具、网络等多项技术集成的系统。

·**交互性** 许多目前的知识发现系统和工具缺乏和用户的交互性，在知识的发现过程中，难以充分有效地利用领域知识。对此可以利用贝叶斯方法确定数据可能性和分布来利用先前的知识。再就是利用演绎数据库本身的演绎能力发现知识，并用于指导知识发现过程。

·**发现模式的精练** 当发现搜索空间很大时，就会获得许多发现结果。其中有些是偶然、盲目的，这时可以利用领域知识进一步精练所发现的模式，从中提取有用的知识。

·**互联网络上的知识发现** WWW 正日益普及，在这信息的海洋中可以发现大量的新知识。已有一些资源发现工具可用来发现含有关键字的文本。例如，我们可以在 WWW 上通过搜索机寻找到有关"data mining"的文本。至今对 WWW 上发现知识的研究仍然不多。加拿大的 Han 等人提出利用多层次结构化的方法，通过对原始基本数据的一般化，构造出多层次的数据库。例如可以将 WWW 上的图象描述而不是图象本身存储在高层数据库中。目前的问题是，如何从复杂的数据例如多媒体结构化的数据中提取有用的信息，对多层次数据库的维护，以及如何处理数据的异类性(heterogeneity)和自主性等等。

参 考 文 献

[1] Fayyad, U. et al., Knowledge Discovery and Data Mining Towards a Unifying Framework, KDD-96 Proc. 2nd Intl. Conf. on Knowledge Discovery & Data Mining, AAAI press, 1996

[2] Piatetsky-Shapiro, G. et al., An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, Same to[1]

[3] Agrawal, R., and Srikant, R., Fast Algorithms for Mining Association Rules, In Proc. 1994 Int. Conf. VLDB

[4] Han, J. and FuY., Discovery of Multiple-Level Association Rules from Large Databases. In Proc. 1995 Int. Conf. VLDB

[5] David W. Cheung et al., Maintenance of Discovered Knowledge: A Case in Multi-level Association Rules, Same to[1]

[6] Fayyad, U. M. et al., KDD for Science Data Analysis; Issues and Examples. Same to[1]

[7] Han, J. etc., DBMiner: A System for Mining Knowledge in Large Relational Databases, Same to[1]

[8] R. Agrawal et al., Database mining: A performance perspective, IEEE Trans. on Knowledge and Data Eng., 5(6)1993

[9] Jiawei Han et al., Intelligent Query Answering by Knowledge Discovery Techniques, 同[8], 8(3)1996

[10] Vasant Dhar, Alexander Tuzhilin, Abstract-Driven Pattern Discovery in Databases, Same to[8]

[11] 蔡庆生、陈小平，机器发现的进展与趋势，国防科学大学学报，1995年6月

[12] 叶煊、陈恩红、蔡庆生，关联规则的发现算法研究，技术报告1996