知识发现的过程驱动策略*)

Process Drive Strategies in KDD

刘业政　杨善林　朱卫东

(合肥工业大学计算机网络系统研究所　合肥230009)

Abstract This paper analyses the tasks, problems and relevant factors in the each phase of the KDD process, discusses the KDD's process drive strategies and their advantages and disadvantages, introduces the relevant operators and technologies and proposes an idea of mixed drive strategy.

Keywords KDD, Data mining, Process drive strategy, Mixed drive strategy

1　概述

随着管理信息系统的广泛使用，沉积的数据越来越多，如何有效地使用、分析这些数据，使其对人们的决策起到支持作用，是目前人工智能领域和管理信息系统领域研究的主要内容之一，知识发现(Knowledge Discovery in Databases，KDD)为上述研究提供了有效的工具。本文详细分析了知识发现过程所要完成的任务，并介绍了知识发现的过程驱动策略及相关的操作和技术。

2　KDD处理过程

知识发现又称数据挖掘，是指从大量数据中提取可信的、新颖的、有效的并最终能被人理解的模式的处理过程[1]。知识的概念在这里是有量度的，它必须具有有用、有效、新颖、易懂等特点；知识发现是一般包括实验、反复、用户的交互作用以及许多初步的决策、定制等步骤的处理过程，该过程是复杂的、艰难的。

知识发现有多处理阶段模型和以用户为中心的处理模型。以用户为中心的处理模型认为数据库中的知识发现应该更着重于对用户进行知识发现的整个过程的支持，而不是仅仅限于在数据挖掘的一个阶段上。该模型特别注重对用户与数据库交互的支持，用户根据数据库中的数据，提出一种假设模型，然后选择有关数据进行知识的挖掘，并不断对模型的数据进行调整优化，其代表人物是Brachman和Anand[2]。对于多处理阶段模型，用户将数据库中的知识发现看做是一个多阶段的处理过程，在整个知识发现的过程中包括很多处理阶段，代表人物有Usama M. Fayyad[1,4,5]，George H. John[3]，他们强调由数据挖掘人员和领域专家共同参与KDD的全过程。领域专家对该领域内需要解决的问题非常清楚，在问题的定义阶段由领域专家向数据挖掘人员解释，数据挖掘人员将数据挖掘采用的技术及能解决问题的种类介绍给领域专家。双方经过互相了解，对要解决的问题有一致的处理意见，包括问题的定义及数据的处理方式。在数据挖掘人员得到准确的问题定义和分析后，开始收集需要使用的数据，进行再加工以使得数据更适合后面的挖掘算法使用。根据解决问题的需要选择合适的挖掘算法。提取出来的知识需要向领域专家进行解释，以对知识及整个过程进行评价。图1是典型的多阶段处理模型[4,5]。

1)准备阶段—问题定义。了解应用领域、相关知识及最终用户的目标。使用现有技术(如数据库技术等)，依赖用户和分析师完成此步。需要考虑的因素有：△本领域的瓶颈是什么?哪些任务由机器处理方便，哪些工作留给人做更合适?△KDD的目标是什么?完成的标准是什么?△KDD的最终产品是什么?是分类、可视化、探索、概括吗?△结果可理解吗?怎样权衡所获取知识的简单性和精确性?

2)创建一个目标数据集。利用数据库技术从数据库中选择一个数据集分为样本子集和测试子集[3]。需要考虑的因素：△数据的同构性；△数据的变化；△采样策略(如采用随机策略还是分层策略)；△样本数目；△自由度。

3)数据预处理。本步骤完成的工作有：△剔除"噪音"和与KDD无关的数据"outlier"；△收集与建模有关的必要信息，甄别噪音；△确定处理丢失数据的策略；△检查数据的完整性和一致性。

*)本课题得到国家自然科学基金(编号79970058)和安徽省自然科学基金(编号99043645)资助。刘业政　副教授，博士生，主要从事知识发现及决策支持系统研究。杨善林　教授、博士生导师。朱卫东　副教授、博士生。

对不完整数据的挖掘、剔除噪音数据是知识发现任务中的难点，一些新的理论和方法[8,21,23,24]的提出将有助于提高数据预处理的效果。

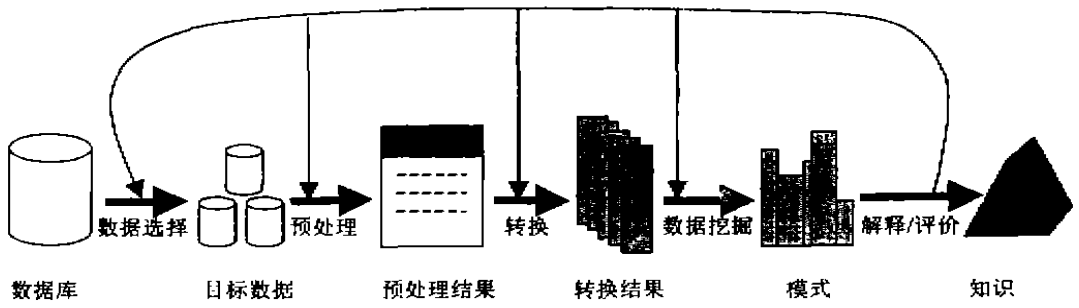

图1 知识发现过程

4)数据缩减和变换。本步骤要完成的工作有：Δ特征抽取。根据KDD的目标，找到能有效描述对象的特征；Δ使用降维或变换方法减少变量数目，发现数据中的常量；Δ将数据映射到一个易求解的空间。

该步骤是KDD的关键一步。特征抽取需要较多的领域知识。对于领域专家来说，定义数据的特征并非困难，专家合理地使用分类技术可直接从样本中得到数据所属的类别，关键问题是数据工程师如何与领域专家合作。粗集理论是目前最行之有效的方法[21,23-25,26]之一。

5)数据挖掘。本步骤是知识发现的核心，也是KDD过程的难点。为了获得满意结果，需要考虑以下问题：

Δ确定数据挖掘类型。是让数据挖掘系统为用户产生假设(发现型)，还是用户自己对于数据库中可能包含的知识提出假设(验证型)[6]。

Δ算法选择。根据数据挖掘的任务即KDD的目标是不是分类、回归、聚类、概括、相关性建模或检测数据的变化和偏离等选择适当的算法。运行算法寻找数据中的模式或建立能恰当描述数据的模型。挖掘方法要与KDD目标相一致，最终用户感兴趣的也许是对模型的理解而不是模型的预测能力。

Δ数据挖掘。用某种特殊的形式或用一组如下的描述：分类规则或树、回归、聚类等寻找感兴趣的模式。

6)结果评估。确定什么是可信的知识，常用方法有：

Δ通过对"兴趣度(interestingness)"[7]和其它内容的测量定义一个自动过滤知识的方案，测量是统计性的，应有良好的适应性、简洁性；

Δ依靠可视化技术帮助分析师决定所抽取知识的可用性或获得一些隐含在数据/现象背后的结论；

Δ完全依靠用户对所抽取的模式进行筛选，以期得到感兴趣的条目。

7)整合所发现的知识。把所发现的知识溶合到系统中，或以文件和报表的形式呈现给用户。这期间也包含对知识的一致性的检查，以确信本次发现的知识不与以前发现的知识相抵触。

上述仅仅是一个过程，至于在每一步选用什么样的技术，这些技术需要什么样的参数以及如何表述这些问题都是非常复杂的，而且整个过程往往要重复多遍。

3 数据挖掘的过程驱动策略及其相关操作和技术

数据挖掘根据任务的不同，常常采用两种形式：验证驱动形式和发现驱动形式[6-8,9]。验证驱动(自顶向下)：抽取信息以验证用户提出的假设的有效性的过程，常使用统计、多维分析等技术。发现驱动(自底向上)：使用工具如符号和神经元聚类、关联发现，有导师归纳法等自动抽取信息。两种方法所抽取的信息为以下形式之一：事实、趋势、回归或分类模型，数据库记录之间的关系和偏离准则的数据即outliers。

在发现驱动方法中，由训练集驱动搜索，每次处理一个输入样本，逐步泛化最新的概念集直到算出最终的合取泛化。其优点是支持增量学习，学习过程不仅从指定的训练样本开始，也可以从已经发现到的规则开始，学习系统在考虑每个新样本后能够更新已有的假设，但由此也可看出该方法抗干扰能力差，对噪音数据缺少免疫能力，该方法的典型例子是删除候选概念算法[10,11]。在验证驱动方法中，用一个预置模型(假设)控制搜索，找到一个可能泛化集以试图发现几个满足一定要求的"最佳"假设。该方法的优点是有良好的噪音免疫能力，当一个假设集被有噪音的数据验证时，出现一到二个反例并不能否定一个假设，而是使用统计方法判断解释这些假设的优劣，其缺点是由于验证或否定假设是基于对所有数据的检验，很难用在增量学

习环境中，当新样本可用时，必须重新从最开始启动学习过程，

3.1 验证驱动的相关操作

验证驱动的数据挖掘的操作包括：查询和报表、多维分析以及统计分析。

△查询和报表[12,13]。构成了决策支持和数据挖掘的最基本形式。其目标是验证用户所描述的假设的有效性。向数据库发出一个或一组查询(该查询能最有效描述用户所定义的假设)，对返回的结果进行解释和分析，确定是与假设一致还是矛盾，从而进一步改善对假设的表示。一般用图形、表格或文本形式来报告查询结果。

△多维分析[14]。虽然查询和报表能满足几种验证驱动的数据挖掘，然而在某些领域，有效的数据挖掘需要建立非常复杂的查询。这些查询通常包含一个嵌入的临时的维且能表述两种指定事件间的变化。如某百货连锁店的地区经理要求"按部门呈报1994年和1995年第一季度中西部地区的周销售量"。多维数据库的特点：按照预先定义的维(如时间、部门)组织数据，拥有利用多维结构中稀疏部分的工具，提供专门化语言使用户在处理查询任务时能迅速、方便地按维完成查询。多维数据库同时也允许按每维实行数据的层次组织即摘要置于高层、实际数据置于较低层。如季度销售量置于第一层，月销售量置于第二层，而日销售量置于底层。

△统计分析[15]。简单的统计分析通常仅在多维分析中使用，在查询和报表中也使用，然而为了验证更复杂的假设，需要将统计操作如主元分析回归建模与数据可视化工具结合使用。

3.2 发现驱动的相关操作

发现驱动的数据挖掘的操作包括预测建模、数据库分段、连接分析以及偏离探测。

△预测建模。这些包括符号归纳法如 CART、C4.5[16,17]和神经网络技术如 BP 法[18]。该操作利用反映历史(过去行为的信息)的数据库记录自动产生一个能预测未来行为的模型。如一个保险公司的保险商希望预测其政策对一名客户的影响；市场行政人员希望预测某个特殊消费者是否要更换某个专用产品的品牌等。发现驱动建模技术特别是符号归纳技术的结果是一组被描述成 if-then 规则集且可理解和可解释的模型。

△数据库分段。该操作通常必须自动地将一个数据库分割成若干相关记录的集合，这些记录集使用户能总结每个数据库(产生一个摘要)或选择数据库的较小的一部分数据进行不同的数据挖掘操作如建模、连接分析等。例如，通过分割一个百货销售点的数据，用户可自动地创建包括特殊时期如"开学日"或"春节后"事务的段。

△链接分析(link analysis)。建模和分段操作的目标是创建一个能刻画数据库内容的概括性描述，而链接分析是为了寻求在数据库的记录之间建立联系。例如，销售商必须确定哪些商品能在一起销售——如男式衬衫和领带、男士香水一起销售——以便商场进哪些商品(衬衫、领带、香水)以及如何放置这些商品(如将领带和香水靠近衬衫摆放)。

△偏离探测。这种操作试图找到不适合某个段的那些数据(outliers)，然后确定它们是不是噪音或者作进一步的验证。该操作常常与数据库分段操作相关联，同时由于"outliers"表达了对期望值的偏离，往往能引起一些真实的发现。

3.3 发现驱动的数据挖掘技术[19~22]

虽然发现驱动的操作仅有四种，但支持的技术很多。如有导师归纳法支持预测模型的建立，关联发现和截列发现技术支持连接分析，统计技术支持偏离探测等。

△有导师归纳法。该过程包括从一个被称为训练集(training set)的一组记录(样本)中自动地创建一个分类模型。训练集可能是被挖掘的数据库或数据仓库的一个样本集，也可能是整个数据库或数据仓库。训练集中的记录必须属于分析师预先定义的类，归纳出的模型由模式(本质上讲是对记录的概括)组成，不同的模式描述不同的类。模型一旦被归纳出来，就可用来自动预测那些未被分类的记录属于哪个类。有导师归纳法可以是神经元法也可以是符号法，神经元法如 BP 方法，用结点和带权重的连接的架构图来描述模型；符号法所创建的模型常用决策树或一组 if-then 规则来描述。如果数据挖掘具备下述三个特点则特别适合使用有导师归纳法：训练集中的数据即使是有噪声或不完整也能产生高质量的模型；结果模型是可理解和可解释的以便用户能理解系统是如何作出决策的；能接受领域知识以便在迅速完成归纳任务的同时改善归纳模型的质量，

△关联发现。给定一组项(即关系中的字段)以及每条记录均包含有该组项的若干项的一个记录集，关联发现就是对记录集进行操作返回各项存在的关系。这些关系用一组规则表示如"含有 A、B、C 项的所有记录中有72%含有 D 和 E。"百分数(72)被称为关联度(confidence factor)，A、B、C 关联边与 D、E 关联边相对，每个关联边的项数任意。

△截列发现(sequence discovery)。在事务日志中，购买商品的客户本人信息往往是不知道的，若这些信息存在，则可根据该客户重复购买的情况分析购买记录之间的关系。

△聚类。将数据库分段成子集或簇，每个簇中的成员共享许多感兴趣的属性。聚类操作的结果将完成下

列二项功能之一：分析每个簇而不是数据库中的每条记录的特点总结目标数据库的内容；或作为其它方法如有导师归纳法的数据输入。簇对于有导师归纳学习元而言是一个更小且更易管理的数据集。

簇使用神经和符号无导师归纳法。不同的神经和符号法由目标数据库中允许记录所取的属性值的类型（如数值型、名词型、结构型）、它们表示每个簇的方法以及它们组织一组簇的方法（分层或平行列表）加以区别。对数据库聚类后，分析师利用可视化组件对所形成的簇进行检测以发现有用的或感兴趣的簇。

神经聚类法，如特征映象，将一个簇表示成一个与被聚类的数据集中样本子集相关联的原型。符号聚类法主要操作具有名词型属性的记录，该方法考虑所有刻画每个样本的属性，使用基于人工智能的搜索方法建立描述已建立的每个簇的属性子集。

Δ 粗集方法（Rough set）。波兰人 Pawlak 提出的粗集理论，近年来得到了迅速发展，其应用遍及多个领域，尤其是为数据挖掘和知识发现领域提供了不同于常规数据库方法的一种有效而新颖的理论。粗集理论的实际应用与理论探讨是当前计算机科学中的一个热点问题。粗集方法在数据分析中能够解决的基本问题包括：根据属性值表征对象集；发现属性间的（完全或部分）依赖；冗余属性（数据）的简化；发现最重要的属性（核）；生成决策规则。

结论 目前有80～90%的挖掘工具只采用了一种操作方法。首先，工具开发者通常只理解一或二种操作方法，且这些工具往往只基于一种技术；其次，最终用户所提出的需要解决的问题比较狭窄，往往用一种操作就能解决。

在70%的应用中用户使用验证驱动操作完成数据挖掘。数据分析师和应用领域专家可能完全理解查询和报表、多维分析或统计分析，因此自然就习惯于使用该方法解决决策支持问题，而神经归纳法和符号归纳法对应用领域专家来讲是比较陌生的。另外有两个因素直接制约了发现驱动挖掘方法的应用：研制每一种挖掘工具所必须付出的有效的努力和被挖掘数据的不恰当的表示，数据经常不正确、不完全，不同数据库的数据模型和格式不兼容，命名难以理解等。由于这些问题的存在且缺乏良好的数据集成和净化工具，据测算仅有10%的数据能用来进行分析。数据仓库技术能够缓解这方面的问题，它能保证分布的数据以一种数据模型集成且在集成过程中进行净化，同时保证元数据编码明确。

由于数据预处理能力的增强、数据约简更有效、低复杂度的挖掘算法增多，受 T. G. Dietterich 和 R. S. Michalski 等人的启迪[23]，在 KDD 中使用混合驱动策略，结合两者的优点，会有助于解决上述现象，这是本课题目前研究的主要内容之一。

参 考 文 献

1 Fayyad U M. Data mining and knowledge discovery：making sense out of data, Microsoft Research, IEEE Expert, Oct. 1996 20～25

2 Brachman R, Avand T. The Process of Knowledge Discovery in Databases: A Human Centered Approach in AKDDM, AAAI/MIT, 1996. 37～58

3 John G H. ENHANCEMENTS TO THE DATA MINING PROCESS, PhD. dissertation, Stanford University, 1997. Chapter 1～2

4 Fayyad U M, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery: An Overview. in Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, (AKDDM), AAAI/MIT Press, 1996

5 Fayyad U M, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proc. of the 2nd Int Conf. on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, Aug 1996

6 Reality cnieck for data mining. Evangelos Simoudis, IBM Almadan Research Center. IEEE Expert, Oct. 1996. 26～33

7 周欣，沙朝锋，朱扬勇，等 兴趣度—关联规则的又一个阈值. 计算机研究与发展，2000，37(5)：628～633

8 Hu Xiaohua. Knowledge Discovery in Databases: An Attribute-Oriented Rough Set Approach: [PhD-dissertation]. University of Regina, 1995. 10～11

9 史忠植. 高级人工智能. 科学出版社，1998. 180～190

10 Mitchell T M. An Analysis of Generalization as a Search Problem. In: Proc. of the 6th IJCAI Conf. Tokyo, Japan, 1997. 577～582

11 Han J, Cai Y, Cercone N. Data-Driven Discovery of Quantitative Rules in Relational Databases. IEEE Trans. Knowledge and Data Engineering, 1992, 5(2)

12 Graefe G. Query Evaluation Techniques for Latge Databases ACM Computing Surveys, 1993, 25(2)

13 Gupta H, et al. Index Selection for OLAP, Working papre, 1996

14 Shukla A, Deshpande P M, et al. Storage Estimation for Muitidimensional Aggregates in the Presence of Hierarchies. In: Proc. of the 22nd VLDB Conf. 1996

15 Mannila H. Data Mining: Machine Learning, Statistics and Database. In: Proc. of the 8th Int Conf. on Scientific and Statistical Database Management, Stockholm, June 1996

16 Quinlan J R. Induction of Decision Trees. Machine Learning, 1986(1): 81～106

17 Quinlan J R. C4.5. Programs for Machine Learning, Morgan Kaufmann, 1993

18 Lu Hongjun, Rudy Setiono, Liu Huan. Effective Data Mining Using Neural Networks. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 957～961

19 Frawley W J G. Piatetsky-Shapiro, &Matheus C. J., Knowledge Discovery in Databases: An Overview. AI Magazine, 1992, 13(3): 57～70

20 Bramer M A. Knowledge Discovery and Data Mining, Publised by the Institution of Electrical Engineers, London, 1999

21 Pawlak Z. Rough set—Theoretical Aspects of Reasoning about Data. Dordrecht, Boston, London: Kluwer Academic Publishers, 1991

22 马溪骏，刘业政，等. 基于粗集理论的决策分析. 合肥工业大学学报（自然科学版）（已录用）

23 Dietterich T G, Michalski R S. A Comparative Review of Selected Methods for Learning from Examples, Machine Learning: An Artificial Intelligence Approach, 1983, 1: 42～82

24 Ramoni M, Sebastiani P. Robust Learning with Missing Data, 1996. Available at: http: // kmi. open. ac. uk/techreports/KMi-TR-28

25 王珏，王任，等. 基于 Rough set 理论的"数据浓缩". 计算机学报，1998，21(5)：393～400

26 苗夺谦，胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展，1999，36(6)：681～684