计算机科学2000Vol. 27№. 2

关于KDD的一类开放系统KDD*的研究*)

Research of the Open KDD System(KDD*)

杨炳儒 申江涛
(北京科技大学信息工程学院 北京100083)

Abstract On the basis of KDD(Knowledge Discovery based on Database), the paper proposes the general framework of the open KDD system——KDD*, discusses its theoretical foundation and realization of its key technology——double base cooperating mechanism. The result of initial illustration shows that the structure of KDD* is effective and available.

Keywords Knowledge discovery, Cooperating mechanism, General framework

1 引言

目前国际上KDD的研究主要是以知识发现的任务描述、知识评价与知识表示为主线,以有效的知识发现算法为中心。具体而论,在各类真实数据库(关系、演绎、时序、空间、分布式、面向对象等)中,利用统计学、证据理论、归纳学习、近似推理、人工神经网络、遗传算法、概念树提升算法、关联算法、分类算法、Rough集理论、现代数学分析方法及其集成方法等技术,发掘诸如关联规则、分类规则、数据聚类、序贯模式、相似模式、混沌模式等知识;开发其原型系统与实用系统(如Regian大学的KDD-R系统,Kansas大学的LERS系统,Lock Head Martin公司的Recon系统等);研究与开发基于KDD的通用工具(如S. S. Anand等开发的MKS)。几次有关KDD的国际会议基本上围绕着KDD的基础理论、发现算法、数据仓库、可视化技术、知识表示方法、发现知识的再利用、网络环境下的数据发掘等几个专题进行研讨。

然而,对于KDD自身的发掘过程、本体结构与运行机制却很少有人问津,1996年,Sarabjot S. Anand等提出的基于证据理论的数据发掘一般框架ESD中,提及了"用户的先验知识与先前发现的知识可以耦合到发现过程中"[1];1992年,在G. Piatesky-Shapiro等开发的知识发现平台KDW中提出过"采用领域知识辅助初始发现的聚焦,限制性的搜索"的思想[2];1993年,Jong P. Yoon与Larry Kerschberg提出一个数据库中知识发现与进化的概念,提出使用正反两个方面的例子来发现新旧知识的协调一致,以及知识与数据库同步进化的思想[3]。但他们都没有提供理论基础与具体实现方法,更没有从提高发掘效率的角度对KDD固有的结构与运行过程加以改造与优化。

虽然当今学术界对KDD的研究已取得了一定的成就,但从信息融合的角度考虑问题,将知识库与数据库所提供的信息有机地融合起来,从而更有效地发掘知识还是一个崭新的课题。本文从知识发现、认知科学与智能系统交叉结合的角度,提出了基于双库协同机制的KDD*新系统。粗略地讲:

KDD* △ KDD* 双库协同机制

上式中的符合*表示在KDD技术的基础上融入双库协同机制,即构建数据库与基础知识库的内在联系"通道",从而用基础知识库去制约与驱动KDD的发掘过程,改变KDD固有的运行机制,在结构与功能上形成了相对于KDD而言的一个开放的、优化的扩体。

本文的工作主要有:(1)给出KDD*系统总体结构(与KDD对比可见其特征);(2)给出双库协同机制的部分理论基础与具体实现方法(本文不是对发掘算法的研究);(3)通过实例,验证KDD*在提高知识发现效率与扩展KDD功能等方面的有效性、可行性与创新性。(4)给出了基于知识发现的广义诊断型专家系统(ESKD)的总体结构。

2 KDD*系统总体结构图及其特征

2.1 系统总体结构图

图1说明了系统的逻辑结构和各部分之间的关系,以下分述之:

*)国家自然科学基金重点项目(69835001)资助。杨炳儒 教授,博士导师。

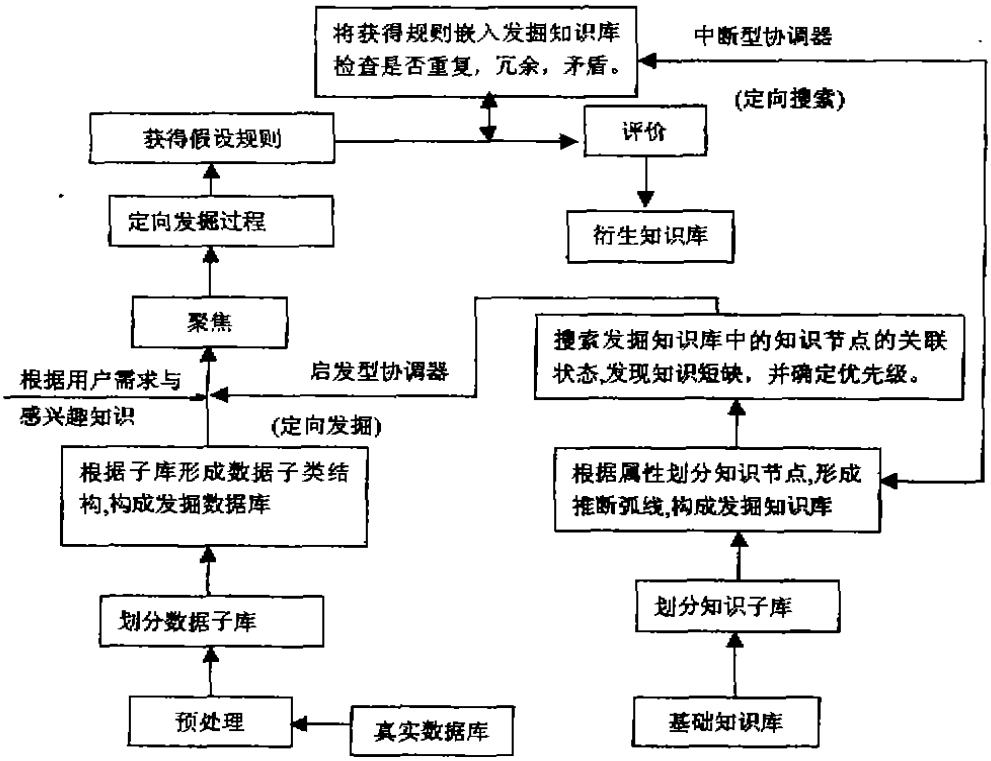

图1 KDD*系统总体结构图

2.1.1 预处理 对原始数据进行包括数据净化、数值化与特定转换等在内的处理，形成发掘数据库DMDB，以供数据发掘过程使用。

2.1.2 聚焦 即从发掘数据库里进行数据的选择，进行聚焦的方法主要是利用聚类分析和判别分析。指导数据聚焦的方式有：(i)通过人机交互由专家提出感兴趣的内容，让专家来指导数据发掘的方向，(ii)利用启发式协调器进行定向的数据发掘。

2.1.3 求取假设规则 这是KDD的核心，它是针对真实数据库(具有大数据量、不完全性、不确定性、结构性、稀疏性等特点)中数据所隐藏的、先前未知的及具有潜在应用价值的信息进行非平凡抽取。在本系统中主要是抽取因果关联规则，从而进一步丰富基础知识库。使用的发掘方法是统计归纳推理法与因果关系定性推理法，对于前者将在下面第3节中加以介绍。

2.1.4 双库协同机制 即采用中断型协调器、启发型协调器，分别对所获得的假设规则进行处理和利用关联强度激发数据聚焦进行数据发掘。这是我们的主要创新点，将在下面第4节中加以介绍。

2.1.5 评价 这一环节主要用于对所获得的假设规则进行评价，以决定所得的规则是否存入知识库。使用的方法主要有：(i)由规则的关联强度，通过设定一定的阈值，由计算机来实现。(ii)通过人机交互界面由专家来评价，也可利用可视化工具所提供的各类图形和分析资料进行评价。将经评价认可的规则作为新知识存入衍生知识库中。

2.2 KDD*的特征

KDD*相对于KDD而言，是KDD与双库协同机制相融合的一种知识发现的新结构，它具有以下特征：

1)KDD*有机地沟通与融合了KDD*新发现的知识与基础知识库中固有的知识，使它们成为一个有机的整体；即实现了"用户的先验知识与先前发现的知识可以耦合到发现过程中"。

2)在知识发现过程中，KDD*对于冗余性的、重复性的、不相容的信息作出了实时处理，有效地减少了由于过程积累而造成的问题的复杂性，同时为新旧知识的融合与合成提供了先决条件；实现了"知识与数据库同步进化"。

3)KDD*改变与优化了知识发现的过程与运行机制，实现了"多源头"聚焦与减少评价量。

4)从认知科学的角度看，KDD*强化并提供了知识发现的智能化程度，提高了认知自主性(这将是今后

相当长的一阶段内保持的研究基调),较有效地克服领域专家的自身局限性、实现了"采用领域知识辅助初始发现的聚焦"。

5)作为 KDD* 的核心技术—双库协同机制的研究,揭示了在一定的建库原则下,知识子库与数据子类结构之间的对应关系,为实现"限制性的搜索"而减小搜索空间、提高发掘效率提供了有效的技术方法。

3 KDD* 系统的理论研究和技术实现

3.1 双库协同机制的理论基础

KDD* 的核心是双库协同机制,这一技术的实现是要构造中断型协调器与启发型协调器。为此,首先在大型(基础)知识库中,根据各个具体的论域划分为若干个相关的知识子库;同时在真实数据库中,也相应地抽取与各个具体论域相关的数据子库。这样可以建立知识子库中"知识结点"与数据子库中"数据子类(结构)"的层之间的一一对应关系,其理论基础是我们提出的泛同伦概念和下列的结构对应定理(详见文[4,5])。

结构对应定理 对于论域 X,在相应的知识子库与数据子库中,关于知识结点的拓扑空间〈E,Ж〉与关于数据子类(结构)的拓扑空间〈F,Б〉是同一泛同伦型的空间。

此定理给出了知识子库中"知识结点"与相应数据子库的"数据子类结构"中的层之间的一一对应关系。

从以上讨论可得到:在知识发现系统中,数据库与知识库的数学结构本质上都可以归结为泛同伦范畴,即数据库是数据子类(结构)集合连同"发掘线路"构成的泛同伦范畴,称为数据发掘范畴;知识库是知识节点集合与"推理弧线"构成的泛同伦范畴,称为知识推理范畴,并进一步得到:在〈E,Ж〉中的知识推理范畴 $C_R(E)$ 与在〈F,Б〉中的数据发掘范畴 $C_D(F)$ 的同构性与制约机制的一些结果,从而从根本上解决了"定向搜索"与"定向发掘进程"的问题。

3.2 双库协同机制的技术实现

3.2.1 *中断型协调器* 其主要功能是当从真实数据库的大量数据中经聚焦而生成感兴趣的与具有一定可信度的规则(知识)后,使 KDD 进程产生"中断",而去定向搜索知识库中对应位置有无此生成规则的重复、矛盾。若有重复,则取消该生成规则而返回 KDD 的"始端";若无,则继续 KDD 进程。对于矛盾的处理,采用约束规则的条件与根据其可信度或关联强度来裁决等方法。

由于中断型协调器对 KDD 过程的介入,可以实时地将重复、矛盾知识淘汰掉,作到只对那些有可能成为新知识的假设进行评价,从而最大限度地减少评价工作量、提高 KDD 的效率。

3.2.2 *启发型协调器* 其功能是在以属性为基础的知识库建库原则下,通过搜索知识库中"知识结点"的不关联态,以发现"知识短缺",产生"创见意象"、从而启发与激活真实数据库中相应的"数据子类",以产生"定向发掘进程"。为了防止"海量定向发掘"现象的产生,必须规定优先级,以定向发掘较可信与关联性强的待定规则。方法之一是计算整个因果网络中相应知识结点间的因果关联规则强度,它由一个三元组构成,记为:

$$\pi(H,E)=\langle\alpha,\beta,\gamma\rangle$$

其中:$\alpha=CF(E)*P(E)$;$\beta=CF(H,E)$;$\gamma=CF(H)*P(H)$。而 CF(E)为前提的可信度,P(E)为前提的先验概率,CF(H,E)为规则的可信度,CF(H)为结论的可信度,P(H)为结论的先验概率。它包含了关于这条规则的随机不确定和模糊不确定的全部信息,为研究规则之间的关系和实现对规则的评价提供了很好的依据。

4 KDD* 的应用示例

有了以上的理论研究成果,我们可以对很多领域的问题采用 KDD* 系统来实现。下面是我们的一些研究实例。

1)**故障诊断问题**

在气轮发电机组振动的故障诊断问题中,通过对数据库的发掘,我们发现了以下两组规则并存入知识库,即:

油粘度下降→油膜破坏→烧瓦→机组强烈振动;

冷油器水侧结垢→冷油器故障→油温很高;

然而,在专家看来,机组强烈振动的根本原因并不在于油粘度下降。如何发现实际系统中问题发生的根本原因?利用双库协同机制可发现知识短缺,产生定向发掘,结果产生了新规则"油温很高→油粘度下降"(这是计算机自动发掘之结果,并非人为的"有意注意"与参与的结果)。这样,就形成了一个因果链:

冷油器水侧结垢→冷油器故障→油温很高→油粘度下降→油膜破坏→烧瓦→机组强烈振动

从而找到了冷油器水侧结垢这一根本原因。这是非常有意义的,因为故障的传播是一个由低层到高层的逐层传播过程,诊断过程必须找到根本原因,才能真正解决问题。

2)**KDD* 在农业规划中的应用**

在农业系统中,贮藏着非常丰富的数据,这些数据组成各种各样的数据库,如关系数据库、时空数据库、面向对象数据库、多媒体数据库等。如何有效发掘这些数据库中的有用知识呢?我们建立起相应的基础知识

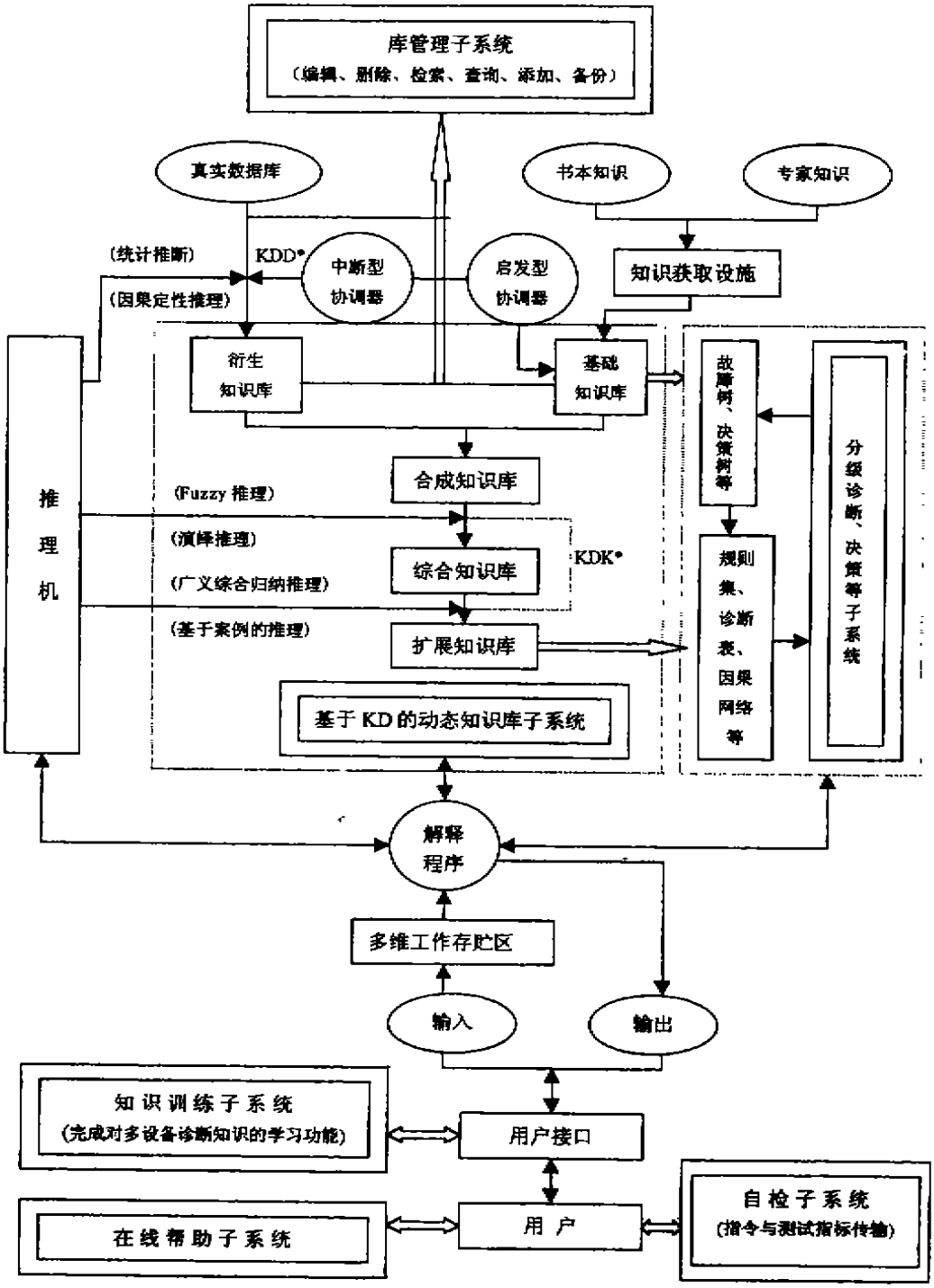

图2 ESKD 总体结构图

库，然后进行知识发掘。例如：硒是人和动物必须的微量生命元素，具有多重生物学功能，缺硒是人克山病、大骨病、动物白肌病等的主要原因，且与多种癌症、白内障及衰老、乳房炎等疾病有密切关系。大米是世界上各国人民的主食之一，其含硒量高低与人体硒营养状况密切相关，但广大的水稻产区多属缺硒或低硒地区，因此发现水稻对硒的生物富集作用的动态变化规律，对指导农业生产和促进人体健康具有非常重要的意义。利用 KDD* 对相关数据的分析，可以发现水稻干物质累积和对硒的累积不同步，前者高峰在生长中期，后

者以生长后期为主，形成一条规则后，存于知识库中。根据此规则，在农艺措施上应在稻灌浆充实前增施一次硒肥。另外水稻对硒有一定的生物富集作用，在缺硒和低硒地区施用硒肥，能显著提高水稻含硒量，改善其营养品质。这样就为我们的决策提供了合理的依据，可避免资源不必要的浪费。它一方面可以指导我们合理地施加肥料，另一方面也可以指导肥料生产厂家在不同阶段添加不同的微量元素，以适应农业生产的需要。对于农作物其它的数据，也可以这样处理。

以上只是两个简单的事例，事实上，凡拥有大量数据但尚不能很好地利用其支持生产、决策的部门，都可以应用 KDD* 解决实际问题，提高工作效率。

5 ESKD 简介

ESKD(基于知识发现的广义诊断型专家系统)是在基于数据库与知识库协同机制的综合型知识发现系统 KD(D&K)的基础上，提出的基于知识发现系统的一类新型专家系统，其理论基础是我们提出的基于数据库与知识库协同机制的综合型知识发现系统 KD(D&K)，它以多个知识源、多种知识融合、多抽象级与不同知识层次结构形成了极其丰富的动态知识库系统与相应的集成推理机制，它为解决专家系统构造中的核心技术提供了一条有效的途径，也从根本上提高了专家系统的实用化功能。ESKD 的总体结构图如图2所示。

该系统的核心技术是双库协同机制，它解决问题时所用的知识库并非基础知识库，而是已经过了一系列提升的过程形成的扩展知识库。其提升过程大致如下：

基础知识库→衍生知识库→合成知识库→综合知识库→扩展知识库

在知识库的不断提升过程中，综合利用了多种推理机制以及 KDD* 的成果，限于篇幅，在这里就不再详细描述了。

对于 ESKD 的 KDD* 模块我们完成了在网络与 Windows95/98环境下，基于 Oracle 数据库、VC++5.0 平台的程序设计与实例运行，其结果良好。

结论 (1)与传统的 KDD 相比，可明显地看到如上所述的 KDD* 的功能特征与创新点。(2)两类协调器可独成系统，可以形成"接口"装在任何已有的 KDD 软件系统中，沟通与固有知识库的联系，大大提高发掘效率。(3)所开发的软件系统支持常用的数据库如 Oracle，Access，FoxPro 等，并采用了数据字典；对于不同领域的数据库，只需变换此数据字典，就可以适用于不同的领域，具有通用性。另外，适于在 Internet 网络环境下工作。

参 考 文 献

1 Anand S S, et al. EDM. A General Framework for Data Mining Based on Evidence Theory. Data & Knowledge Eng., 1996, 18: 189～223
2 Piatetsky-shapiro G, Matheus C J. Knowledge Discovery Work-bench for Exploring Business Databases. Intl. J. of Intelligent Systems, 1992, 17: 675～686
3 Yoon J P, Kerschberg L. A Framework for Knowledge Discovery and Evolution in Databases. IEEE Trans. on Knowledge and Data Eng., 1993, 5: 973～979
4 Yang Bingru. KD(D&K) and Double-Bases Cooperating Mechanism. J. of System Engineering and Electronics, 1999, 10(1)
5 Yang Bingru. Double-Base Cooperating Mechanism in KDD. Intl. Symposium on Computer, 1998. 149～152
6 Yang Bingru. FIM and CASE for Evaluation of Hazavd Level Based on Fuzzy Language Field. Fuzzy Sets and System, 1997, 95(2): 83～99
7 Lchiang R H, et al. A Framework for the Design and Evaluation of Reverse Engineering Methods for Relational Databases. Data & Knowledge Engineering, 1997, 21: 57～77
8 Ong H.-L., Lee H.-Y. A New Visualization Technique for Knowledge Discovery in OLAP. In: Proc. of the First Pacific-Asia Conf. on Knowledge Discovery and Data Mining, 1997. 279～286
9 Li C, Biswas G. Unsupervised Clustering With Mixed Numeric and Nominal Data-A New Similarity Based Agglomerative System. In: Same to[8]. 35～48

(上接第78页)

合交通控制子模块时应用了面向对象的设计思想，同时为了封装模块和弥合两种设计方法之间的差异，设计了一个外壳对象 AIModule，作为模块的外部接口。

面向对象的设计思想认为现实世界中的系统由对象组成，各个对象有自身的属性和方法。系统的功能通过各个对象相互通信(即调用其他对象的方法)来合作完成。基于这个思想，可以把混合交通控制系统抽象为五个主要的对象。AIObject：FA 的软件实体，代表了活动物体；Traffic：FA 的输入决策产生式系统；Router：路由选择产生式系统；AIObjAdmin：AIObject 的管理者，负责 AIObject 的生成、初始化和撤销后的资源回收；Navigator：负责实现 AIObject 以当前的速度和方向在道路上移动一个 Step。五个对象的关系如图6所示。

参 考 文 献

1 Seidel R J. Chatelier P R. Perspective on Virtual Reality and Related Emerging Technologies—— Virtual Reality Training's Future?1997, 6
2 Oliver D, Anderson S. Zigon B, Mclord J. Tricks of The Game Programming Gurus 1996, 10
3 汪成为，高文，王行仁. 灵境(虚拟现实)技术的理论、实现及应用. 清华大学出版社，1996, 9
4 陈世福，陈兆乾，等. 人工智能与知识工程. 南京大学出版社，1997, 12
5 廖阳. 三维汽车驾驶仿真系统中的人工智能技术.[硕士学位论文]. 南京大学，1998, 4