面向复杂系统的知识发现过程模型 KD(D&K) 及其应用

杨炳儒[1] 李晋宏[2] 宋 威[1] 李 欣[3]

摘 要 为适应复杂系统的知识发现的需要, 在双库协同机制及其诱导的 KDD* 过程模型，双基融合机制及其诱导的 KDK* 过程模型的基础上，借鉴协同原理，提出了将 KDD* 与 KDK* 有机地融合在一起的、双库协同机制与双基融合机制协同工作的知识发现过程模型 KD(D&K)；描述了 KD(D&K) 的总体流程、动态知识库系统及其特征；并在农业施肥和植保领域的应用过程中得到验证.

关键词 知识发现，双库协同机制，双基融合机制，KD(D&K)
中图分类号 TP311

KD(D&K): A New Knowledge Discovery Process Model for Complex Systems

YANG Bing-Ru[1] LI Jin-Hong[2] SONG Wei[1] LI Xin[3]

Abstract Based on double bases cooperating mechanism and double-basis fusion mechanism, as well as process models of KDD* and KDK*, we propose KD (D&K), which is a new KDD(Knowledge discovery in database) process model for complex systems. KD (D&K) integrates the above two mechanisms and the two process models synthetically. This paper discusses the overall flow, the dynamic knowledge base system of KD (D&K), and the application of KD(D&K) to fertilization and plant protection.

Key words Knowledge discovery, double bases cooperating mechanism, double-basis fusion mechanism, KD (D&K)

1 引言

当前知识发现 (Knowledge discovery in database, KDD) 发展的主流是寻求高性能的挖掘算法[1]. 对于较高层次的框架, 乃至理论研究则少有人问津. 然而知识发现存在一些较难克服的问题，如海量数据增加引起的算法失效等，尤其是面对复杂系统时，这些问题尤为突出.

为此我们提出双库协同机制，揭示了数据库与知识库在知识发现过程中，在各自特定构造下的一一对应关系，构建了数据库与知识库的内在联系“通道”. 基于 KDD 与双库协同机制，我们提出并实现了 KDD* 过程模型. 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. 针对基于数据库的知识发现与基于知识库的知识发现这两个发现过程，我们提出知识库中的知识发现 (Knowledge discovery in knowledge base, KDK)，进而又提出了双基融合机制，其本质在于揭示两个发现过程间的内在联系，借助 KDD 来部分地完成 KDK 的发现任务. 基于 KDK 与双基融合机制，我们提出 KDK* 过程模型.

借鉴协同原理，提出融合 KDD* 与 KDK* 的 KD(D&K) 过程模型. 将二者统一在一个知识发现系统中，实现了一种较高的机器智能境界. 本文讨论了 KD(D&K) 过程模型的设计思想、流程、动态知识库和特征，并将 KD(D&K) 过程模型应用于农业施肥和植保领域.

2 双库协同机制及其诱导的 KDD* 过程模型

我们把知识发现系统视为认知系统，从认知心理学的角度来考察知识发现过程，提出双库协同机制[2, 3]，主要内容为：1) 构造两个协调器来模拟认知心理学特征. 用启发型协调器来模拟“创建意向”，实现系统自主发现知识短缺. 用维护型协调器来模

收稿日期 2005-11-11 收修改稿日期 2006-5-15
Received November 11, 2005; in revised form May 15, 2006
国家自然科学基金 (69835001), 北京市属市管高等学校人才强教计划资助项目
Supported by National Natural Science Foundation of P. R. China (69835001), Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality
1. 北京科技大学信息工程学院 北京 100083 2. 北方工业大学信息工程学院 北京 100041 3. 天津师范大学 天津 300074
1. School of Information Engineering, University of Science and Technology Beijing, Beijing 100083 2. College of Information Engineering, North China University of Technology, Beijing 100041 3. Tianjin Normal University, Tianjin 300074
DOI: 10.1360/aas-007-0151

拟“心理信息修复”，实现知识库实时维护. 2) “定向搜索”和“定向挖掘”，等效地缩小搜索空间、降低算法的复杂度. 为此，必须在数据库和知识库的特定构造下，建立二者之间的某种对应关系.

定理 1 (结构对应定理). 论域 X 的推理范畴 $Cr(N)$ 与完全数据子类结构可达范畴 $C_{\propto} < \gamma, \mathfrak{R}^C(\gamma) >$ 等价[2, 3](证略).

该定理给出了知识库中的知识素结点与数据库中“数据子类结构”中的层之间的一一对应关系，从根本上解决了“定向搜索”与“定向挖掘”的问题.

基于双库协同机制，提出 KDD* 新过程模型，简单地说 KDD*=KDD + 双库协同机制 (其中“+”表示融合)[4].

3 双基融合机制及其诱导的 KDK* 过程模型

如何从知识库中进一步发现更深层次的知识，即知识库中的知识发现 KDK，少有人涉足. 针对 KDK，我们提出双基融合机制，用数据库与 KDD 去制约与驱动 KDK 挖掘过程[5].

定理 2 (过程模型逻辑等价定理). 设 KDK 的过程模型为 $M =< W, R, M, c >$，KDD 的过程模型为 $N =< S, F, sup, rel >$. 在依数据子类结构构建数据库，依知识结点网络构建知识库的条件下，M 与 N 各要素间建立了一一对应关系, 即 M 与 N 逻辑等价. 其中，W：知识结点集，R：知识结点间的认知通达关系 (即知识结点间可由归纳推理相连接的关系)，M：正则测度函数，c：正则确信度函数；S：数据子类集，F：数据子类间的可达关系 (即数据库中的挖掘途径)，sup：数据子类的支持度，rel：F 上的挖掘可信度[4] (证略).

双基融合机制是由三个协调器来具体实现的.

1) R 型协调器：通过综合归纳推理来发现新知识. 知识库的组成包括事实和规则两部分.

2) S 型协调器：KDK 发现的规则 (特别是难于决断的知识) 在进行评价前，先将其送入 KDD* 过程中进行定向挖掘，用 KDD* 的发掘结果先行评估，若此条规则在 KDD* 过程中也可被发现，则认为该规则有效的几率较大；反之，则认为此知识缺乏数据支持.

3) T 协调器：在规则已被 KDK 过程确认后，将产生一个定向搜索进程，搜索知识库中对应位置是否有此生成规则的重复、冗余和矛盾. 这样可以对知识库进行实时维护，做到只对那些最有可能成为新知识的假设进行评价，从而最大限度的减少评价量.

基于双基融合机制，我们提出 KDK* 模型，简单地讲 KDK*=KDK + 双基融合机制[6].

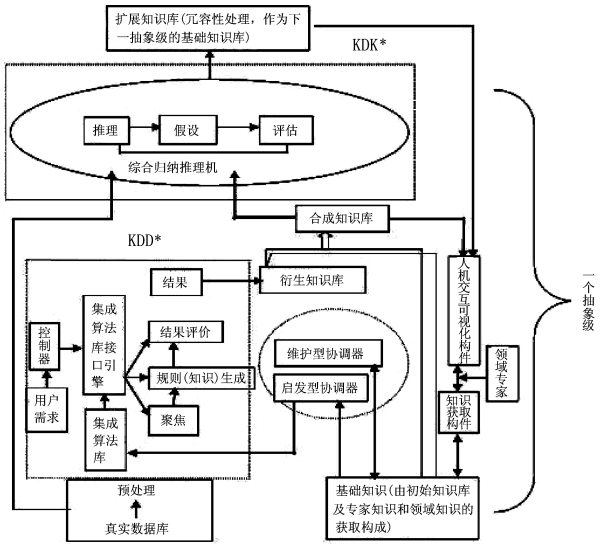


图 1 KD(D&K) 总体结构图

Fig. 1 Overall structure of KD (D&K)

4 KD (D&K) 过程模型

4.1 KD(D&K) 系统的总体结构模型

为适应面向复杂系统的知识发现，借鉴协同原理[7]，我们构造了涵盖 KDD* 与 KDK*，及双库协同机制与双基融合机制的综合型知识发现系统 KD(D&K), 其总体流程如图 1 所示.

说明：这里完备地囊括了现有知识发现的 KDD 与 KDK 系统，按上述构筑的进程可以完成第一个发现 (认识) 阶段，即第一抽象级；以第一抽象级的扩展知识库作为第二抽象级的基础知识库，按类似于第一抽象级的各个发现环节运行，完成第二抽象级. 如此往复，在认识发展与时空环境变迁的不同阶段，不断使知识丰富与升级，不断使认识深化.

4.2 KD(D&K) 的动态知识库系统

动态知识库系统本质上是一个基于数据库和知识库协同机制与融合机制的知识发现系统，作为不同知识层面上知识发现的结果，使得知识库从原有的由专家经验与书本知识为直接源泉的基础知识库不断产生扩充，利用在双库协同机制下形成的 KDD*、在双基融合机制下形成的 KDK* 以及各类推理机制，形成了能处理 Fuzzy 不确定性、随机不确定性及定性信息的具有动态扩展特征的复杂知识库系统. 它包括：1) 基础知识库：即专家经验与书本知识. 2) 衍生知识库：即由 KDD* 发现的新规则. 3) 合成知识库：即经基础知识库和衍生知识库合成后的知识. 即首先把衍生知识库与基础知识库进行

合成，然后利用合成后的知识库去修正基础知识库. 4) 扩展知识库：即通过综合归纳推理机制和基于案例的推理，在合成知识库的基础上发现的新知识；这与知识库中的知识发现 (KDK) 有所不同，我们称之为 KDK*.

总之，KD (D&K) 的知识库系统既包含了书本知识和专家经验 (基础知识库)，又吸纳了通过学习和推理得到的知识 (合成知识库和综合知识库)，还涵盖了在数据库和知识库中发现的新知识 (衍生知识库和扩展知识库). 该知识库系统不仅丰富和提升了经典的知识库结构,而且形成了全新的动态知识库系统. 从而在认识发展与时空环境变迁的不同阶段，使知识不断丰富与升级，认识不断深化.

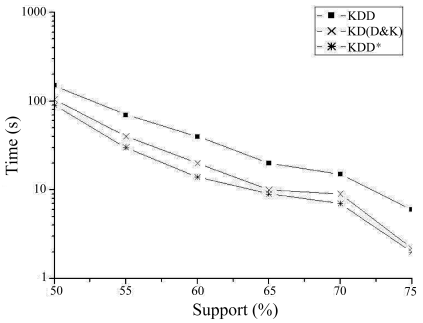
5 KD (D&K)与 KDD 及 KDK 的对比

KD(D&K) 是区别于 KDD 与 KDK 而又包容两者的、融入双库协同机制与双基融合机制的新模型. KD(D&K) 与 KDD 及 KDK 特征上的对比如表 1 所示 (见下页).

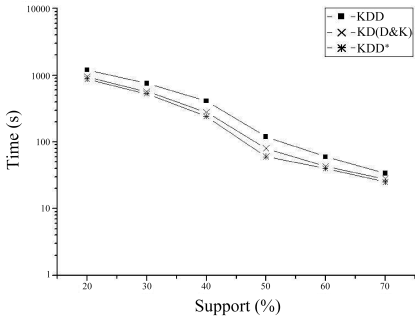
为验证 KD(D&K) 模型的性能，我们分别在 UCI[8]中的 pumsb 数据集和 chess 数据集上进行对比实验. 性能测评的硬件平台是 Pentium IV 2.8GHz CPU, 512 MB Memory, 操作系统是 Windows 2003 Server. 首先利用 KD (D&K)，KDD*，和 KDD 模型进行挖掘，其运行时间对比如图 2 所示. 可以发现，KDD* 的性能最好，KD (D&K) 的挖掘时间比使用 KDD* 模型的挖掘时间略长，但也明显优于传统的 KDD 模型下的挖掘性能. 这是由于 KD (D&K) 和 KDD* 加入了两个协调器，实现了“定向搜索”与“定向挖掘”，较大地减小了搜索空间，提高了挖掘效率. 由于与 KDD* 相比，KD (D&K) 加入了 KDK*，在 KDD* 挖掘的基础上，进一步挖掘深层次的知识；而且引入了动态知识库等部件，这些在一定程度上降低了挖掘的效率.

此外，在以上两个数据集上分别运行 KD(D&K)、KDD 与 KDK 所得到的规则数目对比如图 3 所示.

从图中可以发现：使用 KD(D&K) 模型挖掘得到的规则数目明显要比使用 KDD 模型挖掘得到的规则数目少，且置信度越低这种差距越明显. 这是由于 KD(D&K) 模型引入了两个协调器，部分实现了“定向搜索”与“定向挖掘”，使得规则数目明显减少. 另一方面，由于 KD(D&K) 得到的规则包括使用 KDD* 从数据中挖掘得到的规则，也包括使用 KDK* 从事实和规则中发掘的更深层次的知识，故其挖掘得到的规则数目比使用 KDK 模型得到的规



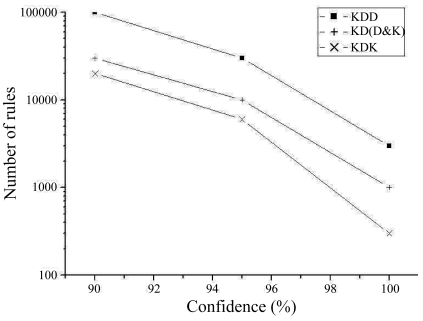
(a) pumsb 数据集 (pumsb dataset)



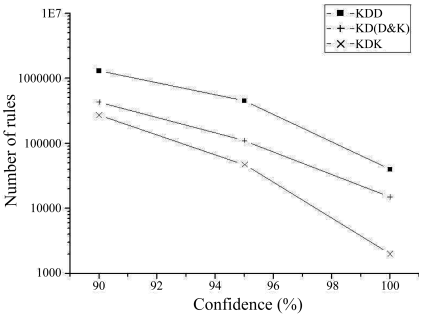
(b) chess 数据集 (chess dataset)

图 2　使用 KD (D&K) 与 KDD 模型在不同数据集上挖掘的时间对比

Fig. 2　Comparisons of execution times between KD (D&K) and KDD



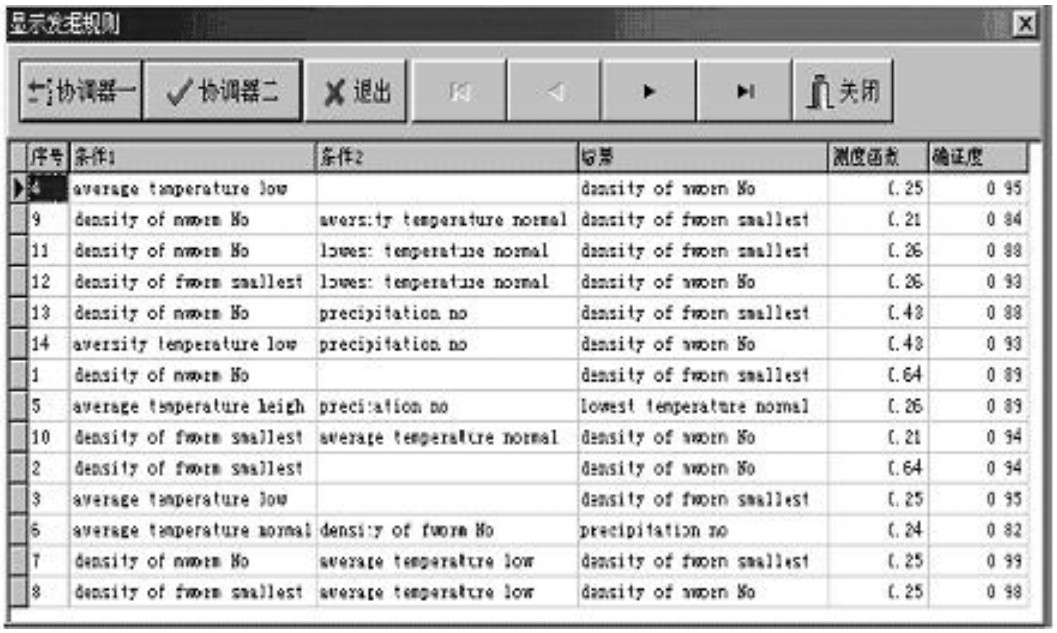
(a) pumsb 数据集 (pumsb dataset)



(b) chess 数据集 (chess dataset)

图 3　使用 KDD、KD (D&K) 与 KDK 模型在不同数据集上挖掘得到的规则数目对比

Fig. 3　Comparisons of execution times between KD (D&K) and KDD



序号	条件1	条件2	结果	测度函数	确证度
4	average temperature low		density of worm No	(.25	0 95
9	density of worm No	aversity temperature normal	density of worm smallest	(.21	0 84
11	density of worm No	lower temperature normal	density of worm smallest	(.26	0 88
12	density of worm smallest	lower temperature normal	density of worm No	(.26	0 93
13	density of worm No	precipitation no	density of worm smallest	(.43	0 88
14	aversity temperature low	precipitation no	density of worm No	(.43	0 93
1	density of worm No		density of worm smallest	(.64	0 89
5	average temperature heigh	precipitation no	lowest temperature normal	(.26	0 89
10	density of worm smallest	average temperature normal	density of worm No	(.21	0 94
2	density of worm smallest		density of worm No	(.64	0 94
3	average temperature low		density of worm smallest	(.25	0 95
6	average temperature normal	density of worm No	precipitation no	(.24	0 82
7	density of worm No	average temperature low	density of worm smallest	(.25	0 99
8	density of worm smallest	average temperature low	density of worm No	(.25	0 98

图 5 使用 KD(D&K) 挖掘庐江螟虫害数据库的部分结果

Fig. 5 Some mining results on pest database of Lujiang region using KD (D&K)

表 1 KD(D&K) 与 KDD/KDK 的特征对比表

Table 1 Comparisons of features between model KD(D&K) and KDD/KDK

模型	知识源	KDD 过程	KDK 过程	总体结构	方法技术	核心概念
KDD 或 KDK	数据库或知识库	基本上形成封闭系统	仅从知识库中利用综合归纳推理机制等发现新知识 (规则)	单级、单层 KDD (聚焦－规则生成－结果评价－结果)	KDD 相关的数据挖掘方法与技术	认知自主性
KD(D&K)	数据库、知识库、两库合成, 推理机制导出等	通过协调器使知识库制约与驱动 KDD, 形成 KDD*, 即形成 KDD 的开放系统.	在 KDK 中融入双基融合机制, 形成 KDK* 新结构的新技法	多个抽象级、不同知识层面的多层递阶、综合集成结构	在知识挖掘、表示、评价、优化、可视化、冗余性与相容性处理等方法与技术上均有相应的改进	认知自主性＋创见意象等 (赋予"感兴趣度"、"进化知识"以新的意义)

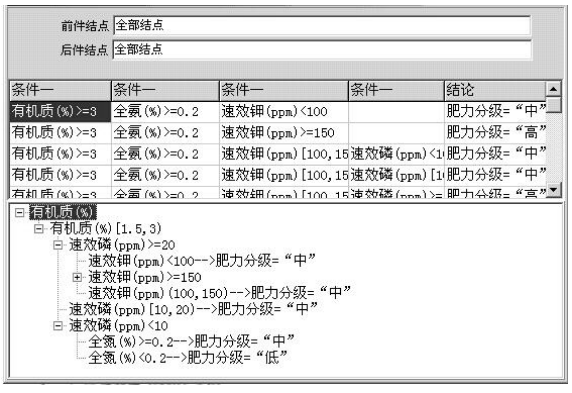


图 4 利用 KD(D&K) 模型挖掘得到的施肥规则

Fig. 4 Some mining results on fertilization using KD(D&K)

则数目要多.

6 KD(D&K)的应用

我们以安徽省合肥农业示范区为基地, 将 KD(D&K) 模型应用于农业施肥、植保领域.

6.1 面向施肥的知识发现系统

在农业生产中, 施肥不科学、肥料投入不足、投肥结构及比例失调等肥料施用误区, 严重影响着农作物的优质高产. 构造指导农民科学、合理施用肥料的施肥专家系统是解决施肥问题的一种有效方法. 在施肥专家系统中, 缺乏有效的知识获取工具一直是其应用过程中的瓶颈, 而建造面向施肥的、以 KD(D&K) 模型为核心的知识发现系统是解决这一瓶颈的有效途径. 图 4 展示了利用 KD(D&K) 模型

挖掘施肥数据得到的一些结果.

6.2 面向植保的知识发现系统

面向植保的知识发现系统可实现：1) 发现病虫害诊断知识；2) 发现气象与病虫害发生规律的关联关系；3) 预测虫害发生数量；4) 发现虫害各代发生程度的序贯模式.

我们收集了有关庐江地区三化螟的数据作为挖掘对象. 庐江螟虫害数据库中，记录的属性分别为雌虫密度，雄虫密度，最低温度，平均温度，降雨量，日照时数等，利用 KD(D&K) 模型对该数据库进行挖掘，部分结果如图 5 所示.

以上两个系统在国家 863 智能农业信息技术应用安徽示范区进行了应用，取得了满意的效果和领域专家的认同.

7 结论

本文提出了将 KDD* 与 KDK* 有机地融合在一起、双库协同机制与双基融合机制协同工作的知识发现过程模型 KD(D&K)；给出了 KD(D&K) 的总体流程、动态知识库系统及其特征；面向农业领域，开发了施肥和植保知识发现应用系统. KD(D&K) 的研究从本质上扩展了现有 KDD 的研究思路，为复杂系统知识发现提供了全新路径和有力工具. 实践证明：基于 KD(D&K) 模型的知识发现较好地解决了植保、施肥及相关领域的一些先前尚未解决或解决不好的问题.

References

1 Mannila H. Theoretical frameworks for data mining. *SIGKDD Explorations*, 2000, **1**(2): 30∼32

2 Yang Bing-Ru, Wang Jian-Xin. A study on double bases cooperating mechanism in KDD (I). *Engineering Science*, 2002, **4**(4): 41∼51
(杨炳儒，王建新. KDD 中双库协同机制的研究(I). 中国工程科学, 2002, **4**(4): 41∼51)

3 Yang Bing-Ru, Wang Jian-Xin, Sun Hai-Hong. A study on double bases cooperating mechanism in KDD (II). *Engineering Science*, 2002, **4**(5): 34∼43
(杨炳儒，王建新, 孙海洪. KDD 中双库协同机制的研究(II). 中国工程科学, 2002, **4**(5): 34∼43)

4 Yang B R. *Knowledge Discovery Based on Inner Mechanism: Construction, Realization and Application*. USA: Elliott & Fitzpatrick Inc, 2004. 144∼154

5 Yang Bing-Ru, Shen Jiang-Tao, Chen Hong-Jie. Research on the structure model and mining algorithm for knowledge discovery based on knowledge base (KDK). *Engineering Science*,2003, **5**(6): 49∼54
(杨炳儒，申江涛, 陈泓婕. 基于知识库的知识发现 (KDK) 的结构模型与挖掘算法的研究. 中国工程科学, 2003, **5**(6): 49∼54)

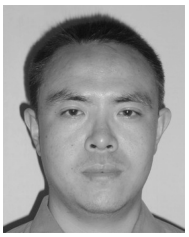
6 Yang B R, Shen J T, Song W. KDK based double-basis fusion mechanism and its process model. *International Journal on Artificial Intelligence Tools*, 2005, **14**(3): 399∼423

7 Harken H. *Synergetics: An Introduction*. Berlin: Springer Verlag, 1997

8 Newman D J, Hettich S, Blake C L, Merz, C J. UCI Repository of Machine Learning Databases[Online], available: http://www.ics.uci.edu/˜mlearn/MLRepository.html, 1998

杨炳儒 北京科技大学信息工程学院教授，博士生导师，研究方向为知识发现与智能系统，柔性建模与集成技术. 本文通信作者. E-mail: bryang_kd@yahoo.com.cn
(**YANG Bing-Ru** Professor in School of Information Engineering at University of Science and Technology Beijing. His research interests include knowledge discovery and intelligent system, flexible modelling, and integration technology. Corresponding author of this paper.)

李晋宏 北方工业大学教授, 研究方向为数据挖掘与专家系统. E-mail: ljh121@263.net
(**LI Jin-Hong** Professor in North China University of Technology. His research interests include data mining and expert systems.)



宋威 北京科技大学信息工程学院博士研究生，研究方向为知识发现. E-mail: sgyzfr@yahoo.com.cn
(**Song Wei** Ph.D. candidate in School of Information Engineering at University of Science and Technology Beijing. His research interests include knowledge discovery, *etc.*)

李欣 博士，天津师范大学副研究员，研究方向为知识发现. E-mail: tj_lixin@163.com
(**Li Xin** Ph.D.，associate professor in Tianjin Normal University. His research interests include knowledge discovery, *etc.*)