知识发现系统框架及其理论体系的构造方法论

杨炳儒
(北京科技大学计算机与通信工程学院,北京 100083)

[摘要] 当前知识发现的主流发展是围绕着寻求在各类数据库和应用背景下高性能、高扩展性的挖掘算法这一主题而展开的。事实上有比算法重要的决定挖掘流程的过程模型的研究,还有更为重要的决定模型和算法的内在机理(反映知识发现系统或过程本身规律)的研究,尚未得到应有的重视。笔者另辟蹊径,将所论三者有机地融合与集成,构造了一类基于认知心理特征的自主知识发现"系统框架",通过对类似的几类"系统框架"的交叉融合、综合集成,构造出基于内在认知机理的知识发现理论体系 KDTICM。研究与实验结果表明:这种高起点、高层次的构造方法论研究,有可能形成高效能挖掘系统与新的研究方向;这种构造方法论的研究,可使长时间得不到解决的"领域知识实质性地介入到知识发现过程中"对知识库进行"动态实时维护"等重要问题得以解决;通过揭示知识发现的潜在规律与复杂性,可反作用于主流发展。最后,给出了此种构造方法有效性的有力佐证。

[关键词] 双库协同机制;过程模型 KDD*;Maradbcm 算法;系统架构 MMAKDS;理论体系 KDTICM
[中图分类号] TP182 [文献标识码] A [文章编号] 1009-1742(2011)09-0083-09

1 前言

数据库中的知识发现(KDD, knowledge discovery in database)是一门新兴的交叉学科。通过对现今各种 KDD 系统十几年的跟踪可见:不同领域学者对其研究的视角不同,主要包括从数据库的角度进行研究,它强调知识发现的效率(efficiency)[1,2];从机器学习的角度进行研究,它强调知识发现的有效性(effectiveness)[3,4];从统计分析的角度进行研究,它强调知识发现的正确性(valid)[5,6];从微观经济学的角度进行研究[7],它强调的是知识发现的最大效用。2003 年 8 月 27 日在华盛顿召开了第九届知识发现与数据挖掘国际会议,参与讨论的专家一致认为:"数据挖掘正面临着巨大的机遇和挑战";"从科学发展的长远来看,最大的绊脚石是基础理论的缺乏以及所面临的问题和挑战的清晰明白的阐述"[8]。

目前,许多有关知识发现的研究或者没有深入探讨其理论基础,或者没有给出具体的实现方法。因此,无法从根本上明显提高现有知识发现过程的性能,也无法解决 KDD 发展过程中极富挑战性的一些问题。事实上,有关知识发现的研究成果只是提供了 KDD 的方法论基础,而要真正构建其理论体系,必须抓住 KDD 的本质,形成与其本质相适应的理论基础。KDD 的本质何在?至少有两个可信的路径:一个是将 KDD 过程(系统)视为认知过程(系统),不是转化为认知系统中;另一个是将 KDD 过程(系统)视为非线性动力系统中非平衡态转化的过程(系统)。笔者从前者出发,经近十余年的研究得到如下结果。

2 基于认知沿机理—模型—算法线路构造的自主知识发现"系统框架"

现代研究表明:分层递阶结构是降低系统复杂

[收稿日期] 2010-07-30
[基金项目] 国家自然科学基金项目(60675030、60875029、61175048)
[作者简介] 杨炳儒(1943—),男,天津市人,北京科技大学教授、博士生导师,主要研究方向为知识发现与智能系统、柔性建模语集成技术;E-mail: bryang_kd@yahoo.com.cn

度的最有效的处理手段,而有序的粒度空间理论是建立复杂系统的分层递阶结构最有效的手段之一。笔者构造了一类以内在机理为理论支柱、以过程模型与挖掘算法为上层建筑的多层递阶的知识发现“系统框架”MMAKDS—沿机理—模型—算法线路构造的自主知识发现“系统框架”,其根源是将认知科学(认知心理学等)的基本原理嫁接到知识发现领域中的结果;形成了以认知自主性为贯穿主线的带有普遍意义的“系统框架”,其结构见图1所示。

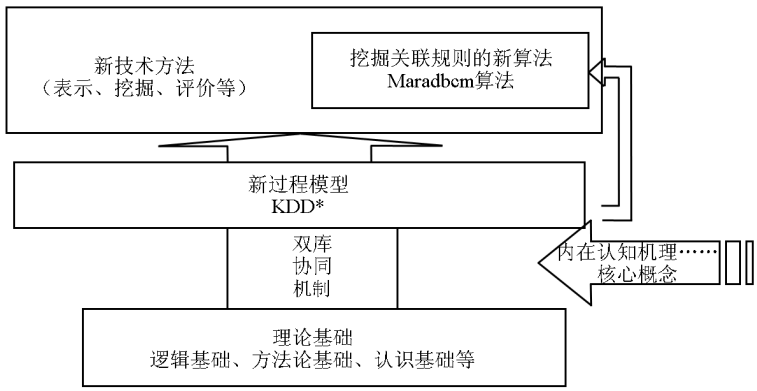

图1 MMAKDS系统框架图

Fig.1 MMAKDS system framework

2.1 内在认知机理

认知心理学兴起于20世纪50年代中期,它是以信息加工观点为核心的心理学,其核心是揭示认知过程的内部心理机制,即信息是如何获取、贮存、加工和使用的。在知识发现系统中,模拟“创建意象”和“心理信息修复”这两项认知心理特征进而提高系统的认知自主性,正是研究的出发点。

利用认知心理学的两个重要特征(即“创建意象”与“心理信息修复”)来研究知识发现的两个重要主题,从而对知识发现的过程模型进行创新。具体而言:a.通过模拟“创建意象”来实现系统自主发现知识短缺,实施启发式的聚焦(除用户感兴趣式的聚焦外)。为此,笔者构造了启发型协调器来模拟“创建意象”,从而实现系统自主地发现知识短缺;该协调器由启发型协调算法来实现[9]。b.通过模拟“心理信息修复”来实现知识库的实时维护。为此,构造了维护型协调器来模拟“心理信息修复”,从而实现知识库的实时维护;该协调器由维护型协调算法来实现[9]。

实现上述两个协调器(算法)的核心技术是要采取“定向搜索”和“定向挖掘”;从而,等效地缩小搜索空间、降低算法的复杂度。为此,在几类布尔代数及其关系的理论基础之上,在数据库和知识库的特定构造下,构建了挖掘数据库中数据子类结构的层与挖掘知识库中知识素结点间的一一对应关系(见图2),称之为“双库协同机制”[9]。“双库协同机制”从一个特定角度揭示知识发现的潜在规律与复杂性。至今这种深入到其系统内部探索规律(内在机理)的研究,实属罕见。

2.2 新过程模型——KDD*

将双库协同机制及其支持下的两个协调器的构造,融入经典的KDD过程中,形成笔者独立提出的KDD*新过程模型,从根本上改变了原有的知识发现进程与运行机制。KDD*过程模型见图3所示。

1) 原有的知识发现过程模型KDD在如下技术与功能方面存在着不足之处:a.领域知识不能实质性地介入到数据挖掘(知识发现)过程中。b.系统不能自主地实现对短缺知识需求和挖掘。c.仅根据用户的兴趣度产生聚焦,确立挖掘方向,会导致大量重复、冗余规则的产生;与系统自身挖掘的短缺知识不能较好地吻合。d.不能对知识库进行动态实时维护。e.模型的实现是基于语义层面的。

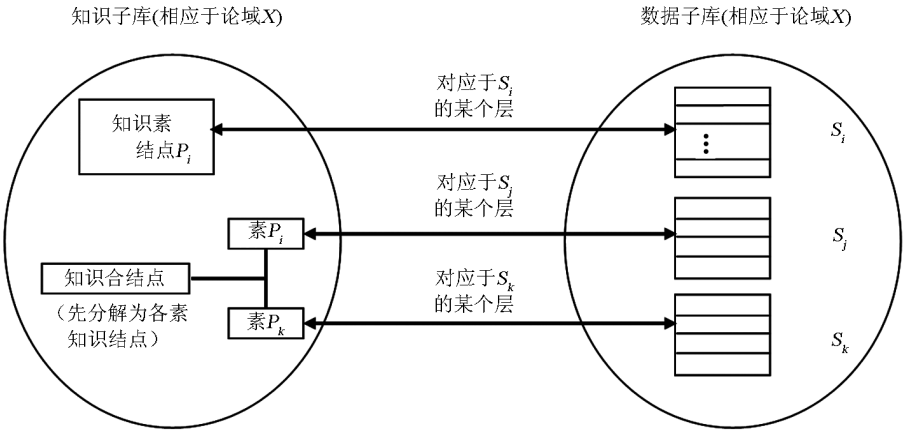

图2　数据子类结构的层与知识素结点之间的对应关系

Fig. 2　Mapping between the layer of data sub elass structure of the database and primitive knowledge node of knowledge base

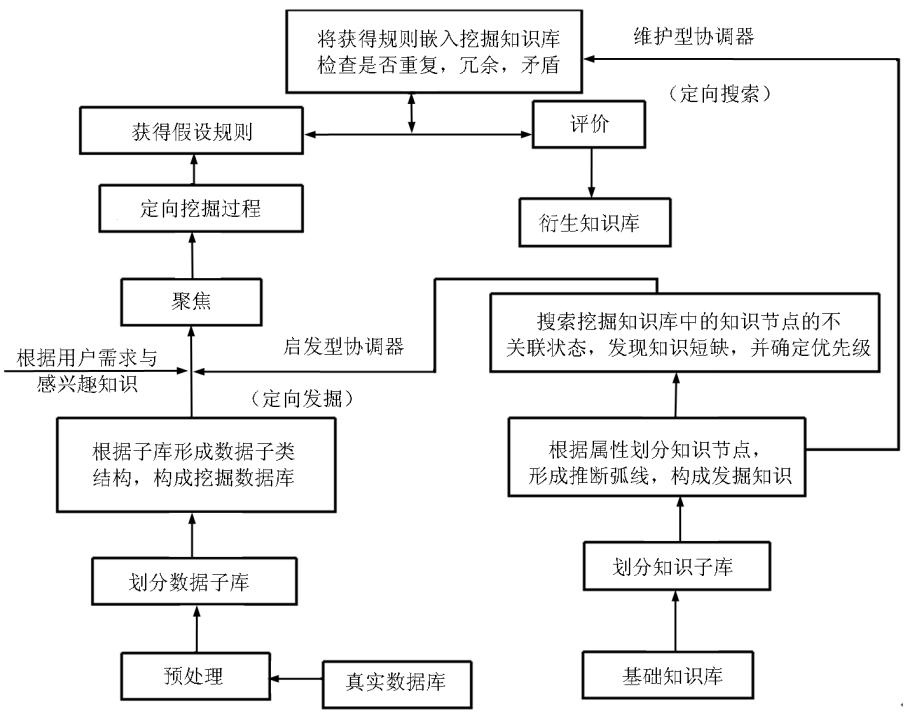

图3　KDD* 过程模型

Fig. 3　KDD* process model

2) KDD* 过程模型针对上述各种不足给出了具体的创新方法和实现技术,具体列下: a. 在挖掘的过程中,领域知识通过两个协调器直接地、具体地介入到挖掘过程中,其主要思想来源是借用同步进化和协同计算的思想。b. 系统能通过有向超图[10]的邻接矩阵产生定向聚焦,自主地实现对短缺知识需求和挖掘。c. 聚焦问题: 定向挖掘的方向与进程当且仅当在用户“感兴趣点”与系统自主发现的“短缺知识点”相吻合的情况下才能产生。这样,不至于挖掘出大量重复、冗余的知识,大大减少规则评价量。这样做的主要目的是减少搜索空间,提高了算法的效率,为算法能通过处理少量数据而达到挖掘效率提供了必要的技术支持。d. 随着知识的积累,知识库的知识也会越来越多,为了能快速地对应用问题做出反映,在新模型中加入了维护协调器,有效地、动态地、实时地处理了知识的重复、冗余、矛盾、

循环与从属。e. 新模型是基于认识心理学的“创建意象”与“心理信息修复”两个认知特征的，故新的模型有坚实的理论基础；模型的实现是基于理论层面的。

3) 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：a. 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”。b. 在知识发现过程中，KDD* 对于冗余性的、重复性的、不相容的信息做出了实时处理，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提供了先决条件，实现了“知识与数据库同步进化”[11]。c. 在数据库的数据积累过程中，虽然知识库结构具有一定的稳定性，但它也是随着数据的积累而不断进化的，并且这种进化的能力是双库协同机制本身所具有的，无须领域专家的干预。d. 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。e. 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”[12]。f. 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、提高挖掘效率提供了有效的技术保证[13,14]。g. 双库协同机制与其诱导的新结构模型 KDD*，对知识发现的主流发展有着重要的作用，由此派生出新的关联规则与数据聚类规则的挖掘算法，与目前流行的算法对比，具有更好的可扩展性与有效性。

在 KDD* 的基础上，还诱导出 KD（D&K）过程模型[15] 以及针对复杂类型数据挖掘的 DFSSM 过程模型[16] 等，此不赘述。

2.3 新挖掘算法——Maradbcm

在双库协同机制与 KDD* 的基础之上，提出了挖掘关联规则的 Maradbcm 算法（以下简称 M 算法）[17]，具体流程与步骤不再赘述。现仅将 M 算法与挖掘关联规则的权威算法 Apriori 算法及其改进型在理论上作一典型的对比分析：

1) 基于的学术思想不同：M 算法是基于双库协同机制的内在认知机理研究，具体而论是基于“知识短缺”（利用有向超图）进行“定向挖掘”以及知识库的实时维护；而 Apriori 算法及其改进型是基于组合论的数据库全局搜索。

2) 基本流程（或基于的过程模型）不同：M 算法是一条一条短缺知识的挖掘；而 Apriori 算法及其改进型是所有的规则一并挖掘。

3) 基础不同：M 算法是基于规则强度，它考虑了主观和客观两个方面；涵盖了 Apriori 算法及其改进型的支持度阈值。

4) 发现知识的量不同：在 M 算法中知识库直接参与挖掘过程，从而能真正发现新颖的、用户感兴趣的知识，这正是符合了 KDD 定义；而 Apriori 算法及其改进型是把满足条件的规则全部挖掘出来；另外，由于 M 算法中的支持度可以设置得比较小（因为该算法主要是由规则强度来聚焦的），即对短缺知识的删除是比较谨慎的，因此 M 算法部分地克服了 Apriori 算法及其改进型的一个缺陷——遗漏重要规则。

5) M 算法可融入 KDD 中形成新的开放型的过程模型——KDD*，整个算法实现的运算背景是 KDD* 结构；而 Apriori 算法及其改进型是原有的 KDD 过程模型。

3 基于内在认知机理的知识发现理论体系——KDTICM 的构造

这种“系统框架”的研究是构造知识发现理论体系的有效路径。事实上，从纵向研究：笔者从认知心理学、认知物理学、认知生物学等不同的角度出发，先后发现了 4 条机制，即双库协同机制（如前）、双基融合机制（揭示了基于数据库的知识发现模型与基于知识库的知识发现模型的逻辑等价）、信息扩张机制（揭示了动态挖掘进程中规则参数的演变规律）、免疫进化机制（揭示了动态挖掘进程中人工免疫与进化演算的协同性），从而对应于每个机制构建了相互独立的 4 类”系统框架”。再从横向研究：对 4 类“系统框架”进行整合集成，交叉融合，在此过程中诱导出 8 个新过程模型，派生出 17 种新技术方法，最终构建了见图 4 所示的一类基于内在认知机理的知识发现理论体系 KDTICM。

4 理论体系有效性与先进性的佐证

以上述的 KDTICM 理论体系及其构造为指导，

笔者针对蛋白质二级结构预测——生物信息学领域中的国际性难题,提出了具有普适性的智能预测系统模型——复合金字塔模型。它采取了逐步求精、多层递阶的4层架构,各个层次各有侧重、功能相对独立且通过智能接口无缝对接,其模型架构见图5所示。

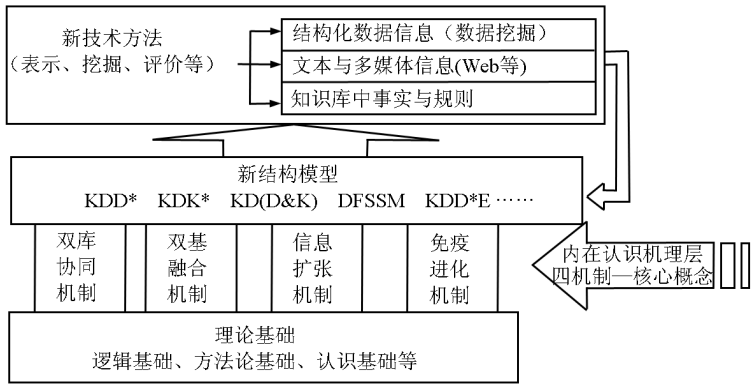

图4 KDTICM 理论体系图

Fig. 4 KDTICM theoretical system

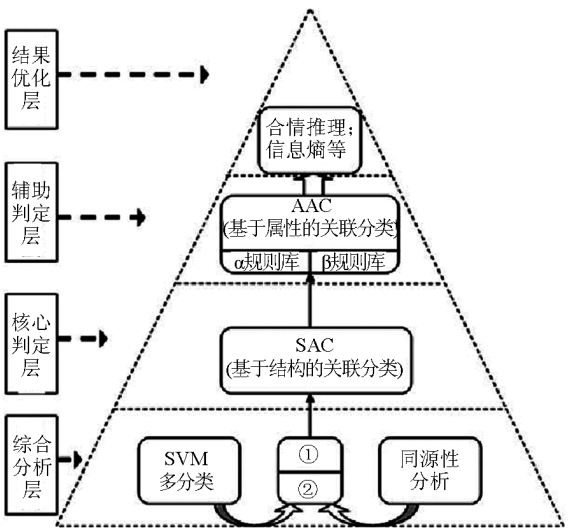

图5 复合金字塔模型

Fig. 5 Compound pyramid model(CPM)

注:综合分析层综合了SVM多分类方法与同源性分析方法,其结果分为两个集合,其中集合①基本确定为H、E、C结构,直接输送结果库;②为尚待SAC、AAC模块进一步判定的部分

综合分析层综合了改进的同源性分析方法与优化的SVM类化方法,即综合了物化属性分析与结构序列分析结果,是整个模型的基础层,可以完成50%以上的特征明显的待测氨基酸二级构象的预测,其中同源序列[18]分析采用Apssp2方法,是一种较为成熟的基于多序列匹配的蛋白质二级结构预测方法;SVM多分类(SVM multiple classification)[19]模块是将惩罚性因子与随机采样引入到SVM方法中,采用属性与结构相结合"轮换式"多分类方法。

核心判定层的原理基于二级结构间的关联影响,也即二级结构之间构象影响信息,其核心理论基础为前述基于内在认知机理的知识发现理论KDTICM,工具为该理论下的KDD*过程模型及其Maradbcm算法。通过与相关关联规则方法的比较,可以发现Maradbcm算法在相同的支持度与可信度下,通常可以获得更多的规则。SAC方法依据蛋白质二级结构的临近信息以及KDD*挖掘得到的蛋白质知识库,利用CMAR(classification based on multiple association rules)算法进行蛋白质二级结构的多分类预测。CMAR[20]作为经典关联分类预测方法CBA的改进类型,突破了使用单一规则进行预测的方法,而是使用了多条满足条件的关联规则进行联合预测。

辅助判定层的核心同SAC一样是笔者独立提出的AAC(attribute association classifier)[21]模块。通过对氨基酸物化属性的关联分析,建立精化规则库,然后利用改进的CBA算法,对下两层无法判断数据进行预测。

优化层主要设计倾向性因子、位能函数及合情推理3类方法。前两类方法属于生物信息学的固有方法,其主要利用生物信息背景知识进行结构预测。合情推理方法是建立在各种二级结构具备的不同物化属性规律基础上的。3种方法从不同角度对其下3层的结果加以优化,以最大程度提高整体预测精度。

笔者使用了3种不同的数据集,以开发和测试

CPM 及其新方法。所选择的测试集为 RS126[22] 数据集、CB513[23] 数据集和 CASP8 [24] 数据集。同时采用 Q3 标准作为评价指标，其定义为预测正确的氨基酸数与氨基酸总数的比值，见式(1)。这个评分只依赖三个状态(螺旋/折叠片/卷)，因此它的名字取为 Q3。在整个的三态预测正确的残基准确度的定义(Q3)是测量预测性能的。

$$Q_3 = \frac{\sum_{i\in\{H,E,C\}} \#\text{ of residues correctly predicted}_i}{8_{i\in\{H,E,C\}} \#\text{ of residues in class i}} \times 100 \quad (1)$$

二级结构的每个类型的残基的准确性(Q^H, Q^E, Q^C, Q_H^{pre}, Q_E^{pre}, Q_C^{pre})的计算公式为(2)、(3)：

$$Q_i(\%) = \frac{\#\text{ of residues correctly predicted}_i}{\#\text{ of residues in class i}} \times 100 \quad (2)$$

$$Q_i^{pre}(\%) = \frac{\#\text{ of residues correctly predicted}_i}{\#\text{ of residues predicted}_i} \times 100 \quad (3)$$

这里的 i 分别代表 H, E 或 C。

CPM 的每一层的结果显示在表 1 和表 2 中。同时，笔者与最好的 6 个二级结构预测的方法(包括 PSIPRED[19], SSPRO[25], SAM - T02[26], PHD Expert[27], PROF[28], JPRED[29]) 在 RS126 和 CB513 数据集上进行对比实验。实验结果显示见图 6。对于数据集 CASP8，笔者选择预测结果最好的 4 个方法作为对比对象，实验结果显示见图 7。

表 1　每个层预测的准确性和在 RS126 数据集上的 CPM 的范围

Table 1　Each layer pridiction accuracy and scale of CPM on the RS126 data set

模型层次	准确率	范围
Comprehensive analysis layer	16 937/18 646 = 90.83 %	18 646/24 806 = 75.17 %
Kernel judgment layer	3 885/6 053 = 64.18 %	6 053 24 806 = 24.40 %
Assistant judgment layer	15/107 = 14.02 %	107/24 806 = 0.43 %
Total	20 837/24 806 = 84.00 %	

表 2　每个层预测的准确性和在 CB513 数据集上的 CPM 的范围

Table 2　Each layer pridiction accuracy and scale of CPM on the CB513 data set

模型层次	准确率	范围
Comprehensive analysis layer	105 106/113 638 = 92.49 %	113 638/146 233 = 77.71 %
Kernel judgment layer	19 961/32 194 = 62.00 %	32 194/146 233 = 22.02 %
Assistant judgment layer	86/ 401 = 21.45 %	401/146 233 = 0.27 %
Total	125 153/146 233 = 85.58 %	

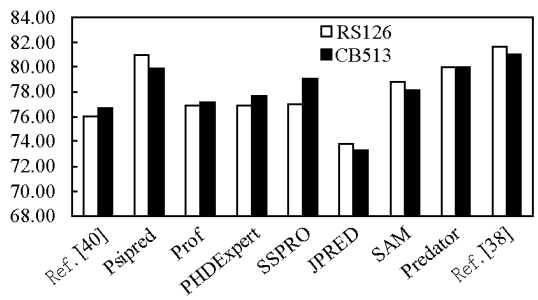

图 6　在 RS126 和 CB513 数据集上 CPM 与其他研究结果的 Q3 准确度比较

Fig. 6　Q3 accuracy comparison with other research results seperately on the RS126 and CB513 data set

注：Ref. [40]: 76. 10 % 和 76. 60 %, Psipred: 81. 01 % 和 79.95 %, Prof: 76. 95 % 和 77. 13 %, PHDExpert: 76.92 % 和 77.61 %, SSPRO: 77. 01 % 和 79.07 %, JPRED: 73. 82 % 和 73.37 %, SAM: 78.81 % 和 78. 17 %, Predator: 80. 06 % 和 80.04 %, Ref. [38]: 81.65 % 和 80.99 %, CPM: 84.31 % 和 86.78 %

对典型的研究文献回顾：Hu 使用 SVM 方法[31] 使准确度达到 78.8 %(在 RS126 数据集上)；Xie[32] 使用神经网络得到了 79.65 % 和 69.11 %(分别在 RS126 和 CB513 数据集上)；Chen[33] 使用层次的神经网络得到 74. 38 % 的准确度(在 RS126 数据集上)；Chopra[34] 使用细胞自动机方法得到了 58.21 % 和 56.51 % 的准确度(分别在 RS126 和 CB513 上)；Liu[35] 使用文章分析的方法分别在 RS126 和 CB513 数据集上得到了 69.8 % 和 69.6 % 的准确度，Guo[36] 使用了双层 SVM 方法在 CB513 数据集上得到了 75.2 % 准确度；Wang[37] 使用了优化的 SVM 方法在 CB513 数据集上得到了 78.44 % 的准确度。实验结果见图 8 和图 9。

可以看到：CPM 的预测精度比其他方法都要

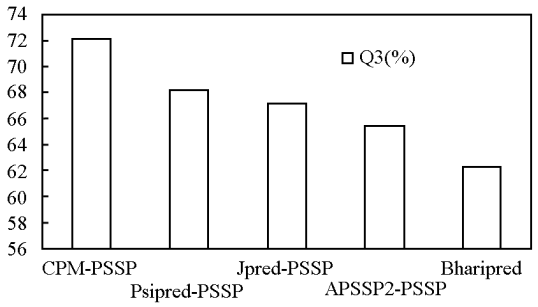

图7　在 CASP8 数据集上 CPM 与其他4种方法的预测结果对比

Fig. 7　Comparison with the results of 4 methods on the CASP8 data set

注: Psipred PSSP: 68.21 %, Jpred PSSP: 67.15 %, ASSP2 PSSP: 65.50 %, Bharipred: 62.26 %, CPM PSSP: 72.13 %

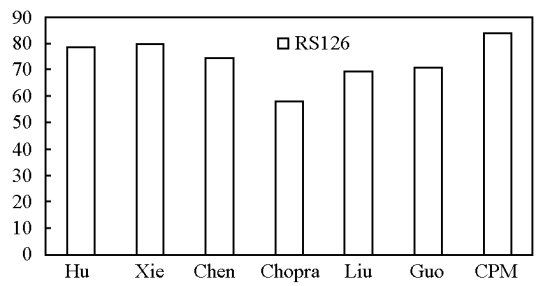

图8　RS126 数据集上 Q3 准确率比较

Fig. 8　Q3 accuracy comparison with typical literature on RS126

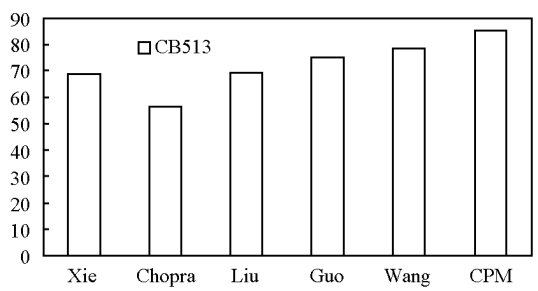

图9　CB513 数据集上 Q3 准确率比较

Fig. 9　Q3 accuracy comparison with typical literature on CB513

高。需要指出的是,笔者可以进一步优化和改善 CPM 方法,使得预测结果更准确。此国际性难题的典型实例充分佐证了 KDTICM 理论体系及其构造方法的有效性,并体现了它的科学价值与实用价值。

5　结语

经过十余年的研究,在对双库协同机制的内涵、知识库及其结构、数据库及其结构、两库间在本质上的对应关系研究的基础上,提出了在知识发现系统与过程中,两个用于模拟认知心理特征从而实现系统自主地发现知识短缺和进行知识库的实时维护的两个协调器构造的理论基础与技术实现方法;作为前者的逻辑必然,诱导出 KDD* 新过程模型;由双库协同机制与 KDD* 派生出新的 M 算法,由此提出了一类由机理—模型—算法构造线路经融合与集成而构造的自主知识发现系统框架 MMAKDS,进而构建出基于内在认知机理的知识发现理论体系 KDTICM。最后,在蛋白质二级结构预测实践中,验证了自主知识发现系统框架 MMAKDS 及知识发现理论体系 KDTICM 的有效性与先进性。

理论分析与实验证实:基于内在认知机理的自主知识发现系统框架及知识发现理论体系的研究,对 KDD 的主流发展将起到重要的推动作用,对"从科学发展的长远来看,最大的绊脚石是基础理论的缺乏以及所面临的问题和挑战的清晰明白的阐述"问题的解决产生深刻影响;基于内在认知机理自主知识发现系统框架及知识发现理论体系具有一般性。笔者的研究成果已在国家级重点科研项目的资助下,有效地应用于农业、现代远程教育网、气象、国际商务、铝电解生产、税务、数字资源整合、医学信息学与生物信息学 9 个领域。特别是对新兴交叉学科,已深刻地验证了理论体系 KDTICM 的有效性与先进性,并解决了一批领域中的典型问题,这类理论体系的构造方法论对其他学科领域具有重要的示范作用。

参考文献

[1] Chen M S, Han J, Yu P S. Data mining: an overview from a database perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866 – 883.

[2] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann, 2001.

[3] Indranil Bose, Mahapatra R K. Business data mining – a machine learning perspective[J]. Information & Management, 2001, 39: 211 – 225.

[4] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations[M]. San Francisco: Morgan Kaufmann, 2000.

[5] Friedman H. Data mining and statistics: what is the connection? [C]// Keynote Speech of the 29th Symposium of the Interface: Computing Sciences and Statistics, Houston, TX, 1997.

[6] Hand D, Mannila H, Smyth P. Principles of Data Mining[M]. Cambridge: MIT Press, 2001.

[7] Kleinberg J, Papadimitriou C, Raghavan P. A microeconomic view of data mining[J]. Data Mining and Knowledge Discovery, 1998, 2(4): 311 – 324.

[8] 杨炳儒. 知识发现进展中的两大核心问题[J]. 中国科学技术前沿: 中国工程科学版, 2006(9): 205 – 269.

[9] 杨炳儒. 基于内在认知机理的知识发现[M]. 北京: 国防工业出版社, 2009.

[10] Wang J F, Lee T T. An invariant for hypergraphs[J]. Chinese AC - TA Mathematical Application Sinica, 1996, 2(2): 113 – 120.

[11] Piatetsky – shapiro G, Matheus C J. Knowledge discovery work – bench for exploring business databases[J]. International Journal of Intelligent Systems, 1992, 7: 668 – 675.

[12] Yang Bingru. KDK based double – basis fusion mechanism and its structural model[J]. International Journal of Artificial Intelligence Tools, 2005, 14(3): 399 – 423.

[13] 杨炳儒, 李晋宏, 宋 威, 等. 面向复杂系统的知识发现过程模型 KD(D&K)及其应用[J]. 自动化学报, 2007, 33(2): 151 – 155.

[14] 杨炳儒, 宋 威, , 徐章艳. 基于知识发现创新技术的专家系统新构造[J]. 中国科学(E 辑), 2007, 37(6): 738 – 747.

[15] Yang Bingru, Xiong Fanlun. KD(D&K) and double – bases co - operating mechanism[J]. Journal of Systems Engineering and Electronics, 1999, 10(2): 48 – 54.

[16] Yang Bingru, Tang Jing. Research of discovery feature sub space model (DFSSM) based on complex type data[C]//Proceedings of 2002 International Conference on Machine Learning and Cybernetics, 2002, 1: 256 – 260.

[17] Yang Bingru, Sun Haihong, Xiong Fanlun. Mining quantitative association rules with standard SQL queries and its evaluation [J]. Journal of Computer Research and Development, 2002, 39 (3): 307 – 312.

[18] Kevin Karplus, Barrett C, Hughey R, et al. Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods[J]. Journal of Molecular Biology, 1998, 284: 1201 – 1210.

[19] David T, Jones. Protein secondary structure prediction based on position – specific scoring matrices[J]. J Mol Biol, 1999, 292: 195 – 202.

[20] Li Wenmin, Han Jiawei, Pei Jian. CMAR: accurate and efficient classification based on multiple class association rules[C]//Proc of the 2001 IEEE International Conference on Data Mining, San Jose, California, 2001: 369 – 376.

[21] Yang Bingru, Hou Wei, et al. KAAPRO: an approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model [J]. Expert Systems with Applications, 2009, 36(5): 9000 – 9006.

[22] Rost B, Sander C. Prediction of secondary structure at better than 70 % accuracy[J]. J Mol Biol, 1993, 232(2): 584 – 599.

[23] Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction[J]. Proteins: Structure, Function and Genet, 1999, 34: 508 – 519.

[24] Protein Structure Prediction Center. http://predictioncenter org/.

[25] Baldi P, Brunak S, Frasconi P, et al. Bidirectional dynamics for protein secondary structure prediction[C]//Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden, 1999.

[26] Karplus K, Karchin R, Draper J, et al. Combining local structure, fold recognition, and new fold methods for protein structure prediction[J]. Proteins, 2003, 53: 491 – 496.

[27] Rost B, Sander C, Schneider R. PHD an automatic mail server for protein secondary structure prediction[J]. Comput Appl Bios - ci, 1994, 10: 1153 – 1160.

[28] Ouali M, King R. Cascaded multiple classifiers for secondary structure prediction[J]. Protein Science, 2000, 9: 1162 – 1176.

[29] Cuff J, Clamp M, Siddiqui A, et al. JPRED: a consensus secondary structure prediction server[J]. Bioinformatics, 1998, 14: 892 – 893.

[30] Hyunsoo Kim, Haesun Park. Protein secondary structure prediction based on an improved support vector machines approach[J]. Protein Engineering, 2003, 16(8): 553 – 560.

[31] Hu H J, Pan Yi, Robert Harrison, et al. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier[J]. IEEE Transactions on NanoBioscience, 2004, 3(4): 265 – 271.

[32] Xie Xiao, Yang Bo, Chen Yuehui. Protein secondary structure prediction based on nerve network[J]. Journal of University of Jinan (Science and Technology), 2008, (2): 111 – 115.

[33] Chen Jfinmiao, Narendra S Chaudhari. Cascaded bidirectional recurrent neural networks for protein secondary structure prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4(4): 572 – 582.

[34] Paras Chopra, Andreas Bender. Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature[J]. Silico Biol, 2007, 7(1): 87 – 93.

[35] Liu Yan, Jaime Carbonel, Judith Klein Seetharaman, et al. Context sensitive vocabulary and its application in protein secondary structure prediction[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, ACM, 2004, 538 – 539.

[36] Guo Jian, Chen Hu, Sun Zhirong, et al. A novel method for protein secondary structure prediction using dual layer SVM and profiles[J]. Proteins, 2004, 54(4): 738 – 743.

[37] Wang Longhui, Liu Juan, Li Yanfu, et al. Predicting protein secondary structure by a support vector machine based on a new coding scheme[J]. Genome Informatics, 2004, 15(2): 181 – 190.

The construction methodology of knowledge discovery system framework and theoretical system

Yang Bingru

(School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

[Abstract] The mainstream of development in knowledge discovery is researching on new high performance and high scalability mining algorithm. In fact, the research of process model and inner mechanism reflecting the law of knowledge discovery system or process and determining model and algorithm is more important, which has not got enough attention. This paper proposed a new independent knowledge discovery system framework, which combines those three elements: mechanism, model and algorithm. Through the cross integration and comprehensive integration of several "systems framework", a kind of knowledge discovery theory based on inner cognitive mecha - nism(KDTICM) has been constructed. Researches and experimental results show that this high starting point and high level researches on the construction methodology are likely to form high performance mining system methodology and new research direction; this researches on the KD(knowledge discovery) construction methodology can substantively involve domain knowledge long unresolved into "the solution to these important issues such as knowledge discovery process" and "dynamic real time maintenance" of knowledge base; by exposing the essence, the regularity and complexity of KD would react on the mainstream development. Finally, the paper gave strong evi - dence of effectiveness of such construction methodology.

[Key words] double collaborative mechanism; process model KDD*; Maradbcm; system framework MMAKDS; theoretical system KDTICM