知识发现领域中当今面临的五类重大问题

杨炳儒
（北京科技大学信息工程学院，北京 100083）

［摘要］ 系统地总结与提出知识发现（KD）领域中当今面临的五类重大问题，它们是KD中的两大核心问题、两大猜想问题、主流发展中富有挑战性的问题、应用研究中的相关领域重大问题以及KD技术标准的制定问题，并给出其部分成果或具体分析。这五类问题密切相关，对它们的研究必将KD推向新的发展阶段，并在该领域内外产生深刻的影响。

［关键词］ KD核心问题；KD猜想；KD挑战性问题；蛋白质二级结构预测；KD技术标准
［中图分类号］ TP18 ［文献标识码］ A ［文章编号］ 1009－1742（2009）04－0076－08

1 前言

基于数据库的知识发现（knowledge discovery in databases，KDD）是从海量数据中提取可信的、新颖的、有效的、最终被用户所理解的模式的非平凡提取过程；数据挖掘（data mining，DM）是KDD的关键步骤或处理阶段，此处视为同一[1,2]；KDD（DM）是计算机与人工智能领域的一个新兴、交叉、边缘学科。国际知名调查机构GartnerGroup的一次高级技术调查报告将基于数据库的知识发现和人工智能（artificial intelligent，AI）列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和KDD列为未来五年内投资焦点的十大新兴技术的前两位。METAGroup也曾做出这样的评价“全球重要的企业、组织会发现，到21世纪，数据挖掘技术将是他们商业成功与否的至关重要的影响因素”。这些领域与技术是当代计算机科学家与其他科技人员所面临的极富吸引力的科学发展的前沿，是最富挑战性的布满生长点的“多学科交叉的无人区”。

在知识发现与数据挖掘（KD&DM）的主流发展沿革中，经历了KD&DM概念内涵与外延重要扩展的4个阶段：结构化数据挖掘DM—复杂类型数据挖掘CDM（Web与多媒体数据构成的大型异质异构数据库）—面向系统挖掘（动态—在线—分布式—并行—网络等系统）—基于知识库的知识发现（KDK）；还有KD&DM知识类型、技术方法及应用的扩展。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[3]，这是在相当长的一段时间内保持的主流基调。当前KD&DM研究的趋向主要有：原有理论方法的深化与拓展；复杂类型（系统）数据挖掘成为热点；新技术方法的引入（其他学科领域的渗透）；理论融合交叉性研究；强化基础理论研究等。

就KD&DM面临的机遇与挑战、历史沿革以及主流发展趋向，系统地总结与提出KD领域中当今面临的五类重大问题，即KD进展中的两大核心问题、KD领域中的两大猜想问题、KD主流发展中富有挑战性的问题、KD应用研究中的相关领域重大问题以及KD技术标准的制定问题，并对这五类问题的意义或部分解决方案进行了深入探讨。这五类

［收稿日期］ 2008－06－15；修回日期 2009－02－14
［基金项目］ 国家自然科学基金资助项目（60675030，60875029）；国家自然科学基金重点资助项目（69835001）；教育部科技重点资助项目（［2000］175）；北京市自然科学基金资助项目（4022008）
［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、柔性建模与集成技术；E－mail：bryang_kd@yahoo.com.cn

问题的关系如图 1 所示。

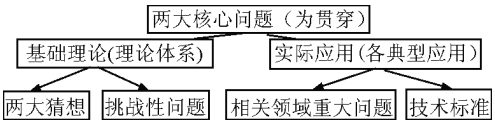

图 1　知识发现领域中五类问题的关系图

Fig. 1　Relational graph between five problems in KD&DM

2　知识发现进展中的两大核心问题

2003 年 8 月 27 日在华盛顿召开了第九届 KD&DM 国际会议，参与讨论的专家一致认为：KD&DM 正面临着巨大的机遇和挑战；并且“最大的绊脚石是基础理论的缺乏”，即提出第一个核心问题。笔者的工作在于将业界已意识到的“缺少杀手锏式应用”这个重要问题与前者联系起来，提到和归结到“两大核心问题”的高度。

2.1　第一大核心问题—基础理论的匮乏

U. Fayyad 认为：“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对所面临的问题和挑战做出清晰明白的阐述”。他认为对于我们要做什么，几乎没有理论甚至工程实践来指导：在今天它仍然是“不为人知的艺术”。我们需要理论来指导我们要做什么以及要如何作。这些理论能够促使工程解决方法的出现，这样我们也可以将我们的“手艺”更有效地教给其他人。而这种形势与从业者以及对应用感兴趣的人们的巨大的热情同时存在，这些人来自不同的领域，但是没有科学根基以及持续的学术发展，该领域不可能得到发展与巩固。R. Uthurusamy 也认为：“ Web 的使用和生产厂家的大肆宣传等都会在短期内影响本领域的发展，它们会使得我们将更多的精力投向数据库营销、CRM 和 OLAP 等方面，而不是致力于使 KDD 从根本上或科学上有大的进步。KDD 的基础研究界必须消除这些干扰而去努力解决 KDD 的真正的根本的问题。”

有些学者在相关于 KDD 的基础理论的研究中做出一些成果，主要包括从数据库的角度进行研究，它强调 KD 的效率；从机器学习的角度进行研究，它强调 KD 的有效性；从统计分析的角度进行研究，它强调 KD 的正确性；从微观经济学的角度进行研究，它强调的是 KD 的最大效用等。但遗憾的是这些研究或者没有深入探讨其理论基础，或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的性能，也无法解决 KDD 发展过程中一些极富挑战性的问题。

事实上，上述的成果只是提供了 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 过程或系统视为认知过程或系统（不是转化为）；另一个是将 KDD 过程或系统视为非线性动力系统中非平衡态转化的过程或系统。基本构想为如下三点。

2.1.1　构建基于内在认知机理的海量数据挖掘理论体系 I（DMTICM）

笔者于 1997 年跳出主流发展，把 KD 自身本体作为研究对象，在国际上开创了从内在认知机理出发、用认知科学与系统论方法研究知识发现的新路径；其核心思想是把知识发现过程视为认知过程，把知识发现系统视为认知系统，用系统论与认知科学的思想和方法（特别是模型化的方法）来研究复杂的知识发现过程本体。于 2002 年构建了基于内在认知机理的知识发现理论（knowledge discovery theory based on inner cognition mechanism, KDTICM）[3]，KDTICM 是以 4 个机制为贯穿红线与理论支柱，包括 8 个新模型和 17 种新技法，由基础理论层、认知机理层、过程模型层、技术方法层构成的多层递阶、综合集成的 KD 理论，如图 2 所示；并且研发了相应的集成化组合构件式知识发现软件系统 ICCKDSS。

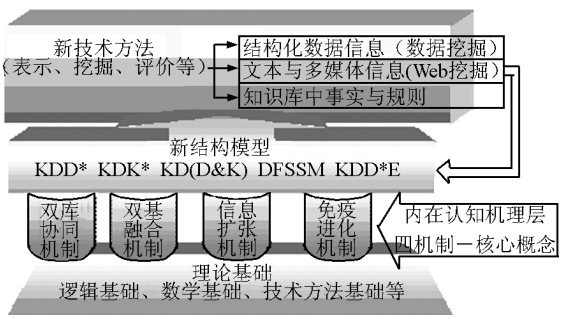

图 2　KDTICM 总体结构图

Fig. 2　Integrated construction graph of KDTICM

基于内在认知机理的海量数据挖掘理论体系（DMTICM）是对 KDTICM 的完善与拓展，也是由理论基础层、内在认知机理层、结构模型层和技术方法层 4 层结构组成。

1）理论基础。从认知科学角度，深入研究构建 DMTICM 的理论基础：逻辑、数学、认知方法等。如扩展语言场理论（知识表示方法），定义语言场梯度、旋度等概念，研究其性质；对知识的熵表示、知识节点网络的构造等提出新方法。

2)认知机理。从双库协同机制[4,5]扩展到复杂类型数据方面,从双基融合机制[6]的高阶逻辑下的KDD与KDK两种发现过程的条件转化方面,从信息扩张机制[7]、免疫进化机制以及动态在线挖掘方面进一步研究与扩展。

3)过程模型。拓展对复杂类型数据的发现特征子空间模型相对于向量空间模型优势的研究,构建复杂的分布式系统和多媒体网络系统中的过程模型、基于大规模复杂结构知识库的知识发现(KDK)过程模型。

4)技术方法:在新的过程模型的规范下,研究基于流数据、代价敏感数据,非平衡态数据,隐私数据等的新算法;对于增强学习、流形学习等新的机器学习方法进行深入的研究;围绕海量问题,研究算法的时空复杂性。

2.1.2 构建基于非平衡态转化的海量数据挖掘理论体系II(DMTBTM)

把非线性系统非平衡态转化与数据挖掘联系起来,即在DM过程中通过一定的流程、运用一定的方法逐步展开,发现用户感兴趣的知识或系统自身判定的短缺知识,便认为挖掘过程达到平衡态。

DM过程的本质是在不断的非平衡态转化中达到稳定态(平衡态);完成挖掘任务即是实现从非平衡态到平衡态的转化。DMTBTM的研究借鉴DMTICM的研究,从以下4个方面构建理论体系:

1)理论基础。从非线性动力学角度,构建基于非平衡态转化的海量数据挖掘理论体系(DMTBTM)特有的基础:对非线性系统非平衡态转化、流形的拓扑不变性、混沌等已有的原理、性质与判定进行梳理,并和DM的认知系统过程相结合,进一步发展在海量数据挖掘条件下解释性的理论基础。

2)内在机理。正象任何事物的转化一样,DM从非平衡态到平衡态的转化也是需要条件的。这些条件有:由非平衡态到平衡态的转化过程中的自适应性、自组织性与自寻优性(和多Agent相关联);涉及到混沌棱理论与同步混沌的研究,提出复杂的DM系统局部活动性机制;在保证挖掘任务精度的前提下,尽量减小挖掘空间与搜索空间,提出压缩挖掘空间与储存空间的机制;提出数据降维与属性约简的条件与机制。

3)过程模型。从非线性系统非平衡态转化的机理出发,构造适用于结构化数据挖掘、复杂类型数据挖掘与动态挖掘进程的全新过程模型,确保挖掘线路流程和信息走向的有效性、便捷性。这种新的过程模型以体系Ⅰ的认知模型为参照,以状态转移的运动视图为特征。

4)技术方法。以机理研究为基础,以过程模型研究为背景,寻求若干高速、可扩展性的挖掘算法,在压缩空间下加速转化;所有新的技术方法,本质上取决于作为非线性系统非平衡态转化过程的一些内在规律。像体系I一样,理论体系II以内在机理层为支柱,形成新的过程模型层与新的技术方法层,其理论基础层是非线性系统非平衡态转化的基本原理与方法论。

将两大理论体系与各类实用智能系统相融合,可引发出各类新型实用智能系统,诸如:基于海量数据挖掘的专家系统、智能决策系统、智能预测系统、计算机辅助创新系统等。

2.1.3 两类理论体系间关系的研究

对于体系I的研究已有一定的基础,通过两类关系的研究,便可由体系I的研究为参照系探求体系II,即以已知探求未知。

基于非平衡态转化的海量数据挖掘是从研究系统状态与系统状态演化过程出发,从非线性动力学的角度来研究海量数据挖掘过程,将海量数据挖掘系统视为非线性动力系统。基于内在认知机理的海量数据挖掘与基于非平衡态转化的海量数据挖掘两者的理论基础不同,出发点不同。但两者研究对象与研究目标一致,因而之间必然存在着一定的关联性。在体系I的挖掘过程中,认知不断演化,对应着不同的认知图;在体系Ⅱ中,由非平衡到平衡的状态变化对应着状态演化图。两类理论体系间关系的研究在某种程度上可归结为认知图与状态图间关系的研究,研究的主要基点是哈肯的"协同论"[8]。

2.2 第二大核心问题——"缺少杀手锏式应用"

在KD&DM的应用中,业界均有其缺少杀手锏式应用的指责。目前国内外KD&DM技术仅在商业的软决策上得到成功的应用,而真正应用于工业生产过程,并产生显著经济效益的DM成果尚属罕见;此外,尚未见到在解决相关学科领域重大问题方面的报道。第二大核心问题"杀手锏式应用"的成功,在很大程度上取决于第一大核心问题基础理论的构建。要实现目标就必须对KD&DM系统自身的认知机理上有所突破,建立科学的理论体系,改变传统的挖掘方法流程,进一步设计高效算法。

笔者利用8年时间做了尝试,即组织实施了上

述 KDTICM 与 ICCKDSS 成功地现场运行于蛋白质结构预测、农业、现代远程教育网、气象、国际商务、铝电解生产、税务、数字资源整合等 8 个领域。a. 有效地验证了 KDTICM；b. 解决了一批领域中的典型问题；c. 在国内外“数据挖掘技术仅在商业的软决策上成功应用，而在工业等领域难获硬效益”的现实面前，后三个领域取得较为显著的直接经济效益与社会效益；如：通过选择全国规模较大、工艺典型的企业与公司为示范（青铜峡铝业股份有限公司、北京华深科技发展有限公司），再在行业内推广，每年带来直接经济效益达 2 050 万元（推广后达 3 亿 3 千万元）；d. 将 KDTICM 与生物信息学领域相结合，主攻结构预测的国际性难题，具有广阔的商业前景。以 KDTICM 为核心技术与方法论应用于蛋白质二级结构预测研究中，达到如下预测结果：在 ILP 相应的数据库 Q3 精度达 93.88 %（国际最高达81 %）；在 RS126 数据库 Q3 精度达 84.1 %（国际最高达 81.65 %）；在 CB513 数据库 Q3 精度达 80.49 %（国际最高达 78.44 %）；上述预测结果均处国际领先水平。

3 知识发现领域中的两大猜想问题

3.1 两大猜想问题的提出

笔者的工作在于首次在 KD&DM 领域用较为精确的数学形式提出两大猜想，在所跟踪的知识发现与机器学习等文献中未见同类的表述。

数据挖掘的研究仍处于初级阶段，现有的挖掘方法都或多或少地存在着面对海量数据失效的问题。应当说，方法（包括算法）的改进仍然无法满足数据量不断增加，而信息越来越丰富的现实。因此，能否根据具体问题的需要，找出那些与问题解决最为相关的数据，而不仅是尽可能地提高算法性能来挖掘全部模式；这将对数据挖掘研究进展，及其在解决具体问题中发挥关键性作用有着重要意义。

在大规模数据库中发现知识，其计算和输入输出代价极大。当数据库是静态的，高的计算代价可以接受，因为发现过程是离线的。但是，当在诸如流数据领域（如电子商务、Web 挖掘、股票分析、入侵检测、故障监控等）挖掘的情况下，挖掘过程的时间和存储器空间受到限制；同时，因为数据库被连续更新，用户会交互的修改搜索参数，频繁重新运行发现程序是难以施行的，用户也不能接受漫长的等待时间。因此需要有效的技术，能够挖掘或处理流数据（在线更新、删除）、用户交互（修改、限制搜索空间）下的挖掘问题，我们称之为动态挖掘技术[9]。

动态挖掘领域已经得到科学共同体内部的关注，并且已经取得了一定成果。如 Ganti 等提出的基于数据块演化的动态挖掘模型，该模型解决了数据仓库等情况下数据周期性更新时，模型维护和模型变迁的挖掘问题等[10]；Chi 等提出的用于频繁闭项集增量式挖掘的 Moment 算法[11]；Gao 等提出的挖掘概念偏移数据流的通用框架[12]等。这些成果代表了当前动态数据挖掘的进展；同时，我们也注意到，在动态挖掘领域的成果主要是特殊方法与专门算法的研究，尚缺乏有关基础理论的一般性研究。

图是计算机领域最为复杂的数据结构之一，图挖掘是当前数据挖掘研究的一个热点问题[13]。与一般的数据比较，图能够表达更加丰富的语义，在化学、生物、社会科学等诸多领域有着更为广泛的应用。但这种丰富的语义也增加了数据结构的复杂性，以及挖掘令人感兴趣的子图的难度。尤其是在图挖掘中，图同构性的检验起着至关重要的作用，而这一过程往往又是 NP 完全问题。因此，根据实际问题的需要，挖掘一小部分高质量的频繁子图用于分类是一个富有挑战性的问题。最近提出的 CPAR[14] 和 HARMONY[15] 算法可以直接挖掘一小部分高质量的项集并用于分类，具有比基于频繁项集全集的分类算法如 CBA[16] 和 CMAR[17] 更高的效率和准确率。近年来，从图数据库中挖掘具有 CLIQUE 结构的子图引起了人们的广泛关注，并在诸多领域得到了广泛的应用，如枚举出 Web 中的 CLIQUE、环等结构用于构建 Web 知识库，从电话呼叫记录库中找出 CLIQUE 用以确定用户感兴趣的社区等。而计算一个图所包含的最大 CLIQUE 的大小已被确认为 NP 完全问题；由于频繁闭合模式相对于频繁闭合模式更为简洁且没有信息损失，挖掘频繁闭合子图更加有意义。此外，已发现：真实数据集往往很大（有的超过 100 000 个属性），但往往只有不足 1 % 的属性是真正对分类有用的。

当前主流发展中的若干挑战性问题也与数据的“海量”相关，即困难之所在。如分布式数据挖掘[18]；高维与流数据挖掘[19]；非静态、不平衡、代价敏感数据挖掘[20,21]；异常检测；多种形式的输入数据；知识的维护与更新；网络设计的数据挖掘；安全、隐私数据挖掘[22]等。

3.2 两大猜想的内容

为了解决数据挖掘中海量数据这一难题，我们

提出了旨在化海量为有限量的数据挖掘领域的“逆问题猜想”和“磁铁效应猜想”[23]。

3.2.1 逆问题猜想

数据挖掘就是如何从数据中发现规律，然而现实中人们往往还会遇到相反的问题，即如何寻找支持规律的数据。如验证新方法的科学性，究竟需要多少数据；检验新药的疗效，究竟需要多少病例的临床试验；面对高维的生物数据，究竟使用多少属性，可以满足分类的需要等。尽管人们可以利用统计学的方法对这类问题进行探索，但仍然存在着如下问题：真实数据的分布情况往往是不规律、不确定的；统计方法往往很难真正适用于海量数据。为此，我们提出如下的“逆问题猜想”。

“逆问题猜想”：若给定挖掘任务 T 以及精度 δ，则存在“最小”数据子集 $K \subseteq D$，（D 为真实海量数据库），其势记作 $\Omega(\Omega << |D|)$，使得在数据子集 K 中实施挖掘任务 T 至少具有精度 δ，且 Ω 是可定量估计的；我们称 K 为数据库 D 的“核集”。

简单说来，数据挖掘逆问题猜想的基本思想是化海量为有限量，这个有限的数据库的大小远小于原始数据库，且挖掘这个有限的数据库在精度 δ 的前提下，与挖掘原始真实数据库等效。数据挖掘逆问题猜想的本质是最终得以确认挖掘数据集的绝对规模是确定的，且其基数存在上确界。一般地，K 的“势”Ω 是可计算的，或者其计算代价是可估计的。

3.2.2 磁铁效应猜想

找出真实数据库的“核集”对数据挖掘的学术研究和应用都具有重要的意义。但“核集”的势仅宏观地给出了“标准（健壮）样本空间”总体量的估计，进而我们要问：哪些具体的样本有资格可充当 K 中的元素呢？即 K 集如何能行地构造？为此，我们提出了“磁铁效应猜想”。

“磁铁效应猜想”：若给定挖掘任务 T、真实海量数据库 D 以及精度 δ，则 D 的“核集”K，可以通过对若干初始数据样本（称为“核吸引子”）作有限步扩展直到其势达到 Ω 来构造。简单说来，数据挖掘“磁铁效应”猜想的基本思想是首先找出若干初始的“核吸引子”，然后以每个核吸引子为中心，吸纳“近测度”者，逐步进行扩展，将更多符合条件的数据不断纳入到数据库的“核集”中，直至其势达到 Ω。应当指出：数据挖掘“磁铁效应”猜想与中心聚类思想有些类似；其区别在于：数据挖掘“磁铁效应”猜想的目的是在海量真实数据库中找到可近似代替原始数据集，又远小于原始数据集的“核集”，并且是在“逆问题猜想”的规范下构造的；而中心聚类则是要对全体数据进行分割并且无前提规范。

显然，这两个猜想构成相互关联的统一的有机体，本质上是围绕着要解决海量数据这一难题。

3.3 两大猜想的意义

1）两大猜想对数据集合与知识集合关系的理解，将深化知识发现内在认知机理的内涵，这对知识发现理论研究具有重要的推动作用。如果上述猜想得以理论证明和实验确认，那么我们可以得到性能优良的挖掘方法：当确认数据集合达到确定的规模时，对知识的确认过程即可停止，无须再对后继数据展开分析；对于一个确定的动态数据流而言，达到确定规模的确认数据集合可以看作为需要验证整个知识库的数据集合的“不动点”，显然对于海量动态数据挖掘过程，这一基于上述假说的挖掘方法具有高效性和高扩展性。

2）两大猜想对促进 KD 主流发展和当前进展中两大核心问题的解决，均具有极大的带动性。

3）从哲学的角度来看：一个学科真正成熟的标志之一在于其基础理论的奠定和发展；基础理论在解释已有特殊经验规律的同时，能够扩展出解释新现象的新的经验规律和解决新问题的方法；基础理论的确立，表征着学科发展的深度和广度。知识发现作为一个学科，刚刚走过它的初步发展阶段，其研究目前尚缺乏统一性的理论作为整个学科的基础，这种状况限制了知识发现的长远发展。从这个意义上讲，两大猜想的提出可能会派生出新的研究方向。

3.4 两大猜想初探

1）KDTICM 内涵的许多具体的结果，特别是双库协同机制，给出了知识库中的“知识素结点”与数据库中“数据子类结构”中层之间的一一对应关系，为两大猜想的实现奠定了理论基础。

2）目前最新的知识发现方法，如最大频繁模式的挖掘、频繁闭合模式的挖掘、多神经网络集成等均在不同的程度上避免了处理全体数据集。这些方法的成功，为两大猜想的实现提供了借鉴。

3）最近提出的 CPAR 和 HARMONY 算法可以直接挖掘一小部分高质量的项集并用于分类，具有比基于频繁项集全集的分类算法如 CBA 和 CMAR 更高的效率和准确率。

4）信息熵、各种演化算法等新技法的成功，为两大猜想的实现提供了方法论基础。由统计学原理

可知:一定数量的小样本数据在容差范围内可以体现总体数据集合的特征,这为两大猜想的探索提供了参考性依据。

5)两大猜想得以实现的一个关键在于“面向知识”的距离测度的定义。一般聚类方法的距离测度研究可以借鉴,但是在我们的假说中,距离测度本质在于反映样例间确证知识能力的相似程度,所以必须作进一步的探索。

6)两大猜想的研究,应涉及到模式确认所需的“有趣性测度”定义的基础研究问题。“有趣性测度”定义在不同的具体挖掘任务中有不同的特定内涵,同时历史上部分具体研究已经给出了主、客观测度的区分。因此在本猜想的研究中,我们将提出涵盖不同任务定义和主客观测度的“有趣性测度”的一般性定义。

7)我们可以首先面向动态数据关联规则发现任务展开研究,继而展开描述性数据挖掘任务(涵盖 clustering,subgroup 等任务)的一般性研究。

上述表明:两大猜想的提出有其客观基础,从目前发展动态、研究方法和研究路线而论呈现了实现的可行性。

4 知识发现主流发展中的富有挑战性的问题[24]

挑战性问题指的是利用当前可公开获得的海量数据与数据挖掘技术(作为主要技术)非常难于解决的问题,问题要目标明确、感兴趣、易理解,并且其解决具有重要的科学价值和显著效益。列举如下:内在机理研究;领域知识参与和用户参与挖掘过程;挖掘过程的自动化;知识的感兴趣度的评价;知识集成的评价;知识库的实时维护与更新;挖掘系统与其他系统的集成与融合;高维与高速流数据挖掘(遥感、生物、环境);动态、非平衡、代价敏感数据挖掘;安全、隐私数据挖掘(被控数据含隐私,保用户安全);人文数据挖掘;文本挖掘;多媒体及其网络挖掘;图挖掘;时序与序贯模式挖掘;隐涵因果关联规则挖掘;多关系挖掘;Link 挖掘;多形式输入的数据挖掘;网络设计数据挖掘;分布式数据挖掘(传感器网络);异常检测(处理海量以太网、多镜像站点等);大量预测模型的评测;融入混沌过程的量子遗传算法;数据挖掘中的微观经济学方法;流形学习中拓扑不变性与流形分类;数据挖掘中的语音识别方法与视觉感知方法等。

5 知识发现应用研究中的相关领域的重大问题

从复杂系统的开放性与系统间集成融合的角度,从自然科学发展的专与博的统一性规律而论,将 KD&DM 理论、方法、技术应用于相关领域中重大问题的研究,是值得探索的重要方向。笔者的主要工作是在分子生物学(生物信息学)领域,选择了国际公认的、大难度的蛋白质二级结构预测问题;经两年多努力,取得了如前所述的预测精度的突破。在微观经济学、医学信息学等其他相关学科领域中,都存在着体现知识发现理论体系科学价值和学科带动性,并有可能形成新的分支学科或研究方向的研究课题。

蛋白质二级结构预测工作的原创性体现:

1)我们没有步国际主流发展的仅研究预测方法的后尘,而是作为预测系统加以研究,它涵盖了系统模型、系统方法、系统优化等核心构件。这是迄今为止,在解决蛋白质二级结构预测这一国际性难题中,我们力图实现的学术思想、技术路线和方法论上的突破。

2)就系统模型而论:实现了形如“合成金字塔”式的多层递阶、综合集成、逐步求精的模型构造;在每一层中,均采用物化属性判定与结构序列判定相结合的构建线路;特别是其核心判定层的构造,以高纯度、高起点的训练数据与精化规则为基础,并采用了作者独立提出的原创性理论 KDTICM,对构造高准确度的关联分类器形成有效支撑。

3)就系统方法而论:是彼此联系、无缝对接、走向有序、相互融合的若干预测方法构成的方法群,故而称之为系统方法。所涵盖的方法中,有的是原创性的(如基于 KDTICM 理论的 KDD * 模型与 M 算法、基于动态规划与神经网络的同源性分析等),有的是对现有方法做出重要改进(如优化的 SVM 类化分析),故而形成了一个高内聚、低耦合、紧密协同的预测方法体系。特别是利用我们独立提出的基于知识发现的因果自动机理论,寻求与优化影响蛋白质空间构象的物化因素,这样深入到机理层面为其物化属性预测法奠定了坚实的基础。

4)就系统优化而论:在系统模型自下而上的演绎中,各层的粒度空间由粗到细,最后到顶层——优化层。这种分层结构的设置,体现了优化的迭代、体现了渐近细化的粒度空间构建思想、也体现了领域知识与背景知识的贯穿性,从而确保了预测精度的

全局优化。

5)研究成果具有自主知识产权,全部模型、算法与程序均是自行设计研发的。

6 知识发现技术标准的制定

在此项研究中,国际上有部分较为成功且具可借鉴性的工作经验,也有未成功的教训。笔者的主要工作在于明确提出涵盖 KD&DM 本体技术、基于 KD&DM 智能系统、行业实施的 3 类技术标准,并指出从规范逐步走向标准是一条可行之路。

6.1 KD&DM 技术标准制定的意义

针对 KD 发展过程中极富挑战性的一些问题,从数据采集、数据预处理等步骤开始直到知识表示、结果展示、分析评价为止的 KD 全过程中,针对每一个步骤形成一整套数据挖掘的技术规范,有助于形成一个可以有效记录工作经验的统一体系,有助于进行项目计划和项目管理,还有助于使新手顺利地完成数据挖掘的整个工作流程。

从 KD 过程模型与技术方法、实用智能系统、行业实施等方面相应技术标准的制定和应用推广,能够让人们反复使用现成的模块、程序或系统部件,有助于降低产品开发成本,提高数据挖掘技术各过程数据或结果的共享和各步骤间相互连接的能力,促进数据挖掘系统在企业和社会中使用和推广,必将会产生巨大的经济效益与社会效益。

KD&DM 技术的标准化将有助于数据挖掘系统的研发工作,促进数据挖掘系统在企业和社会中使用和推广,并将极大程度地推动数据挖掘领域的主流发展。

6.2 KD&DM 技术标准的特点与内容

数据挖掘与具体应用问题密切相关,每一种数据挖掘方法在算法与技术要求上都有自身的特点和实现步骤,因此人们从系统化和方法学的角度,提出了一些数据挖掘过程的参考模型或标准。如 SPSS 提出 5A(Assess – Access – Analyze – Act – Automate)过程模型,5A 强调的是支持数据挖掘过程的工具应具有的功能;美国 SAS 研究所在多年的数据处理研究工作中积累的一套行之有效的数据挖掘方法——SEMMA(Sample – Explore – Modify – Model – Assess)过程模型,SEMMA 强调的是结合其工具的应用方法;数据挖掘特别兴趣小组提出的“数据挖掘交叉行业标准过程”(Cross – Industry Standard Process for Data Mining, CRISP – DM)。CRISP – DM 从方法学的角度强调实施数据挖掘项目的方法和步骤,并独立于每种具体数据挖掘算法和数据挖掘系统。

构建 KD&DM 技术标准的理论基础是系统机理与方法论。即将 KD 过程(系统)视为认知过程(系统),或将其视为非平衡态转化的过程(系统)进行研究。用统计分析、机器学习、微观经济学、数据库技术等构建方法论基础,确保标准制订的深刻性与可持续性。

KD&DM 技术标准主要包括以下 3 种类型,由其可形成 KD&DM 总体技术标准:

1) 由内在认知机理的研究成果诱导出决定信息流程、挖掘线路与技术方法的 KD&DM 本体技术标准。具体而论:包含过程模型、技术方法与软件工具,其中技术方法包含 DM 原语、DM 语言、算法;知识表示、评价、优化、预处理、后处理、结果表示与可视化;DM 与各类数据库的集成等。

2) 基于 KD&DM 智能系统构造技术标准,其涵盖总体架构、典型构件、应用流程、评价模型、开发工具等。

3) 行业实施中的重要技术标准(针对典型示范的行业领域)。

考核指标是设计出以上 3 种标准、示范架构及其评价体系,制定 KD&DM 国家技术的标准文本。将其 KD&DM 技术标准正式文本提交作为国家技术标准审定;并向国际相关组织提供以作 KD&DM 国际技术标准的主流参考源。

7 结语

在对知识发现历史沿革、现实发展与发展趋向剖析的基础上,系统地总结与提出知识发现领域中当今面临的五类重大问题。全文以知识发现进展中的两大核心问题为贯穿主线,沿理论层面提出“两大猜想”与总结补充了“挑战性问题”;沿应用层面提出与总结了“相关领域重大问题”与“技术标准问题”;此外,对这五类问题的意义或部分解决方案进行了深入探讨;部分地阐述了笔者在长时间的实际运行与实验中得到有效验证的成果。

参考文献

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco: Morgan Kaufmann, 2000

[2] Fayyad U, Piatetsky – Shapiro G, Smyth P. From data mining to knowledge discovery in databases[J]. AI Magazine, 1996, 17(3): 37 – 54

[3] 杨炳儒,宋 威, 徐章艳. 基于内在认知机理的知识发现理论及其应用[J]. 自然科学进展, 2005, 15(12):107 – 115

[4] 杨炳儒, 王建新. KDD 中内在机理研究(I)[J]. 中国工程科学, 2002, 4(4): 41 - 51
[5] 杨炳儒, 王建新, 孙海洪. KDD 中双库协同机制的研究(Ⅱ)[J]. 中国工程科学, 2002,4(5): 34 - 43
[6] Yang Bingru, Shen Jiangtao, Song Wei. KDK * based double ~ basis fusion mechanism and its process model[J]. International Journal on Artificial Intelligence Tools, 2005, 14 (3): 399 - 423
[7] Yang Bingru, Song Wei. Research on post - processing of association rules during dynamic knowledge discovery process[A]. Arabnia H R, Mun Y eds. Proceedings of International Conference on Artificial Intelligence[C]. Las Vegas USA, 2004. Bogart USA: CSREA Press, 2004. 664 - 670
[8] [德]赫尔曼· 哈肯. 协同学—大自然构成的奥秘[M]. 凌复华译. 上海: 上海译文出版社, 2005
[9] Yang Bingru, Song Wei, Li Linna, et al. Research overview of regulations in the process of dynamic mining[J]. Journal of Computational Information Systems, 2006, 2 (3): 973 - 980
[10] Ganti V, Gehrke J, Ramakrishnan R. Mining data streams under block evolution[J]. SIGKDD Explorations, 2002, 3(2): 1 - 10
[11] Chi Yun, Wang Haixun, Yu Philip, et al. Catch the Moment: Maintaining closed frequent itemsets over a data stream sliding window[J]. Knowledge And Information Systems, 2006, 10 (3): 265 - 294
[12] Gao Jing, Fan Wei, Han Jiawei. A general framework for mining concept - drifting data streams with skewed distributions[J]. SDM07: 3 - 14
[13] Cook D J, Holder L B. Mining Graph Data[M]. New Jersey: John Wiley & Sons, 2007
[14] Yin Xiaoxin, Han Jiawei. CPAR: Classification based on predictive association rules[J]. SDM'03: 369 - 376
[15] Wang Jianyong, Karypis George. HARMONY: Efficiently mining the best rules for classification [J]. SDM'05: 205 - 216
[16] Liu Bing, Hsu Wynne, Ma Yiming. Integrating classification and association rule mining[J]. KDD'98: 80 - 86
[17] Li Wenmin, Han Jiawei, Pei Jian. CMAR: Accurate and efficient classification based on multiple class - association rules [J]. ICDM'01: 369 - 376
[18] Silvestri C, Orlando S. Distributed approximate mining of frequent patterns[J]. SAC 05: 529 - 536
[19] Wang Wei, Yang Jiong. Mining High Dimensional Data[A]. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers[C]. Kluwer Academic Publishers, 2005
[20] Bernstein A, Provost F J, Hill S. Toward intelligent assistance for a data mining process: An ontology - based approach for cost - sensitive classification[J]. IEEE Transaction on Knowledge and Data Engineering, 2005, 17(4): 503 - 518
[21] Zhang Shichao, Qin Zhenxing, Ling Charles X, et al. "Missing Is Useful": Missing Values in Cost ~ Sensitive Decision Trees [J]. IEEE Transaction on Knowledge and Data Engineering, 2005, 17(12): 1689 - 1693
[22] Zhong Sheng. Privacy - preserving algorithms for distributed mining of frequent itemsets[J]. Information Sciences, 2007, 177 (2): 490 - 503.
[23] Yang Bingru. Two conjectures of knowledge discovery field[J]. Journal of communication and computer, 2007,29(4) :1 - 4
[24] Yang Qiang, Wu Xindong. 10 challenging problems in data mining research[J]. International Journal of Information Technology & Decision Making, 2006, 5 (4): 597 - 604

Five important issues in the present field of knowledge discovery

Yang Bingru

(*School of Information Engineering, University Science and Technology Beijing, Beijing 100083, China*)

[**Abstract**] In this paper, the author summarized and proposed five important issues in the present field of knowledge discovery and data mining, which comprise two important core issues, two conjectures, challenge issues in mainstream development, important issues of relevant fields in appliance research and technique standard in knowledge discovery (data mining). Some achievements and special analysis was also presented. Those five issues correlate closely, and these researches will push the development of knowledge discovery (data mining) to a higher stage, and bring great influence to the field inside and outside.

[**Key words**] important core issues in KD; conjectures in KD; challenge issues in KD; protein secondary structure prediction; technique standard in KD