

论数据发掘的计算智能方法^{*}

On Computational Intelligence Approaches to Data Mining

童 颖

(上海大学(嘉定校区)计算机科学系 上海市 201800)

摘 要 A general survey on the issues of data mining (DM) and knowledge discovery in databases (KDD) is given first. The machine learning is emphasized as the nature of data mining as KDD. The differences as well as association of DM with respect to on-line analysis processing (OLAP) are discussed. It follows a concise discussion of the essentials of the computational intelligence (CI), characteristics of CI-algorithms as a way of machine learning, and so on. The potentialities and perspectives of its applications are also discussed with a typical case-study: discovering the prediction model of financial market price fluctuation lie hidden in the chaos of the specific data nuggets.

关键词 Datamining, Knowledge discovery in databases, Machine learning, Computational intelligence

一、数据发掘:从数据库发现知识

随着现代科学技术的迅速发展,数据库的规模日益扩大,人们需要有新的、更为有效的手段对各种“数据矿藏”(信息资源)进行开采以发挥其应用潜能。数据发掘与 KDD 正是在这样的应用需求背景下产生并迅速发展起来的,开发信息资源的一整套科学方法、算法及软件工具与环境。

国际上第一次关于数据发掘与知识发现的研讨会于 1989 年 6 月在美国底特律召开。当时仅有数十人参加。此后发展很快。1995 年提升为国际学术大会 (International Conference on Data Mining & Knowledge Discovery in Databases)。此外,还有这一主题的地区性国际大会;相关的学科领域,特别是机器学习、归纳逻辑程序设计 (ILP)、医药数据处理、分布式人工智能、基于实例的推理 (CBR) 等,关于 KDD 的 R&D 也十分活跃。商品化的 KDD 软件工具已开始进入市场,如 IBM 的 Intelligent Miner、SGI 的 MineSet 等。

数据发掘对科学发现的意义,是不言而喻的。事实上,科学家们通过观测、实验发现定理、规律或新事物的过程,就可以说是一种数据发掘过程。在现代计算机问世很早以前,从发现九大行星及开普勒定律到理想气体定律、从发现欧姆定理到高分子材料的分子结构,无不是一个从大量长期积累的观测实

验数据发现新知识的过程。只不过当时科学家们是依靠自身的智慧,直接对大量观测实验数据进行处理而达到科学发现的目的。计算机的应用、尤其是人工智能技术的发展,才为科学发现提供了高科技手段。例如,专家系统 DENDRAL 根据质谱仪给出的数据,能够发现已知或未知的高分子化合物分子结构;机器学习系统 BACON 根据已有实验、观测数据,能够发现物理或天文定律。这也说明,专家系统与机器学习等人工智能技术,以及下面将要论及的计算智能方法,对数据发掘与 KDD 的 R&D 将起到关键性的作用。

数据发掘也是市场竞争的需要,意味着产生新的效益。它将为决策者提供重要的、前所未料的信息或知识,从而产生不可估量的效益。难怪有一位巴黎某公司先进技术研究所所长说:“如果你不抢先在你的竞争对手之前用它(指数据发掘)来预测未来趋向,你就会破产!”

按照数据发掘的深度,大体上可以分为两个层次。在较浅的层次上是利用现有数据库管理系统的查询/检索、报表功能与多维分析、统计分析方法相结合,进行所谓在线分析处理 (OLAP),从而得出可供决策参考的统计分析数据。在较深层上,是要求从数据库或大量数据记录中发现隐含的、前所未有的知识。严格说来,OLAP 在数据发掘这一新概念产生之前及发展初期,不属于数据发掘的范畴。但是就决

^{*} 本文的研究工作得到 NSFC 和国家 863 计划的支持。

策支持的需要而言,二者可起到相辅相成的作用。尽管如此,二者所能解答的问题还是有所不同。例如:OLAP 将能回答这样的问题:去年夏季家用电冰箱在广州和沈阳的销售量各是多少? 哪里的销售量更大? 大多少? 而数据发掘则能回答更深一层、更为广泛的问题,例如:在哪些主要因素会影响家用电冰箱在广州或沈阳两地的销售量? 由此可见,前者主要是在数据库管理系统功能的基础上的 OLAP,就是所谓“验证驱动”(verification-driven)的数据发掘技术途径;后者则是着眼于发现大量数据记录中潜在的有用信息或新的知识,属于所谓“发现驱动”(discovery-driven)的数据发掘技术途径,就其实质而言,与各种机器学习方法密切相关,特别是通过观察/发现学习(Learning by Observation/Discovery)、类比学习(Learning by Analogy)、根据实例学习(CBR)等归纳或分析学习方法以及下面将要提到的计算智能方法。

二、数据发掘的计算智能方法

1. 计算智能:模仿生命的自适应优化/进化过程

1994 年,关于神经网络、演化程序设计、模糊系统的三个 IEEE 国际学术会议在美国佛罗里达州奥兰多市联合举行了“首届计算智能世界大会”(The First IEEE World Congress on Computational Intelligence, WCCI'94),把本来是不同学科领域的专家们聚在一起,进行了题为“计算智能:模仿生命(Computational Intelligence: Imitating the Life)”的主题讨论会,取得了关于计算智能的共识。继人工智能之后,计算智能犹如异军突起,吸引着众多研究开发者投身于这一新领域的开拓。尽管关于模糊逻辑、神经网络、演化程序设计的研究开发历史都可以追溯到五、六十年代,但它们却在计算智能共识的启示下获得了新的内涵。所谓计算智能方法,大体上包括以下几类:

- 神经计算(NC)。这是在微观层次上模仿脑神经网络的功能。其实质在于对外界刺激/信号的自适应反应能力。但是必需先要对神经网络进行训练,使它“学会”对输入刺激模式做出期望反应而具备这种自适应反应能力,如图 1 所示。神经网络的学习过程,就是一种自适应调节过程。主要是根据实际输出反应与期望反应之间的偏差 δ ,按照给定的学习/训练算法,对神经网络的参数 w (联接“权重”)和/或 θ (阈值)进行反复的自适应调节,直到对给定输入刺激模式集中每一个模式的输出反应偏差 δ 都在允许范围以内。只有通过这样训练的神经网络,才能作为系统中的一个功能模块发挥“神经计算”的作用。

- 演化计算(EC)。这是在微观或宏观两个不同层次上模仿生物的演(进)化过程。即:□遗传算法(GA)。模仿生物通过染色体的交配及其基因的遗传

变异机制来达到自适应优化的过程。□演化程序设计(EP)或演化策略(ES)。是受到达尔文进化论的启示,模仿自然界“物竞天择”的物种自适应进化过程。EP 和 ES 在算法上可谓大同小异,只是前者发源于美国,着眼于不同种群间的竞争;而后者则发源于德国,着眼于种群内个体间的竞争。

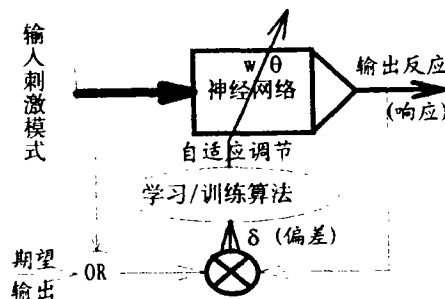


图 1 神经网络的学习/训练过程示意图

演化计算所对应的生态学仿生原型如图 2 所示。其中组成群体的“染色体”、“种群”或“个体”,实际上是一些代表给定问题可能解的数据结构;演化计算的一般算法,就是关于如何按优胜法则选配“染色体”、“种群”或“个体”进行“繁殖”、如何进行自适应变异、如何评价个体的适应性、如何保持整体优势、如何最终求得可接受优化解等等的策略及其实施步骤的描述。

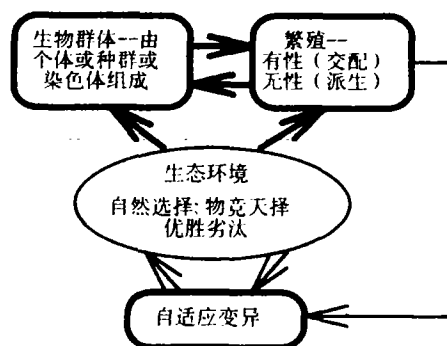


图 2 演化计算的生态学仿生原型

- 模糊计算或模糊推理。可以说是在计算语义变量的隶属度函数值的基础上,进行概念聚类,是对人类在日常生活中进行近似或非精确推断、决策能力的模拟。

2. 计算智能方法的性质与特点

从理论上讲,与传统人工智能(AI)相比,CI 的最大特点、也是它的潜力所在是:它不需要建立问题本身的精确(数学或逻辑)模型,也不依赖于知识表示,而是直接对输入数据进行处理得出结果。以神经网络为例,只要把数据输入到一个已经训练好的神经网络输入端(相当于刺激模式),就可从输出端直接得到预期的结果(反应)。这个神经网络本身是一

个“黑箱”。问题是如何训练神经网络。但这类训练算法并非求解问题本身的算法,而是使神经网络“学会”怎样解决问题的算法,可以说是一种准元算法;类似地,演化算法,无论是模仿生物通过遗传、还是“物竞天择”的自然进化机制来达到优化的目的,都是模仿生物进化方式的算法而不是根据问题本身的数学/逻辑模型来制订的算法,所以也是一类准元算法。

上述神经网络的训练算法或演化算法,作为一类准元算法,有一个重要的特点,就是它们的计算复杂性不是 NP-完备的,尽管它们要解决的问题可能是 NP 难的。具体说来,多层前馈神经网络的学习算法,其复杂性上界为: $O(n^2)$, 其中 n 为输入单元数或某一隐单元层神经元数中的较大者;对于演化算法,则为 $O(m \times n)$, 其中 m 为组成群体的个体数, n 为个体中可参与变异的变量数。

正是 CI 的上述特点,使它适用于解决那些用传统 AI 技术难以有效地处理、甚至无法处理的问题。特别是对高维非线性随机、动态或混沌系统行为的分析、预测,例如,从反映市场价格变动及其相关诸因素的大量数据记录中发现预测模型。

3. 计算智能方法在金融市场数据发掘中的应用

市场价格波动作为一种社会现象,不但受到供求关系这一普遍规律的制约,而且还受参加交易的人群中流行的主观意念的影响。而参加者的主观意念又受到市场价格波动的影响。设 y 代表市场价格的波动; x 代表参加者的主观意念,便可用以下两个函数表示二者的相互影响:

$$y=f(x); x=g(y)$$

由此可导出以下两个递归函数:

$$y=f(g(y)); x=g(f(x))$$

这就是关于市场价格波动的“反射性”理论(Theory of Reflexivity)^[7]。这一理论揭示了市场、特别是金融市场价格波动的不稳定性和混沌性的本质:永远是混沌性的涨、跌交替序列(Boom/Bust Sequence)和良性、恶性相继循环(Benign/Vicious Circles)。这也说明了为什么经典的数学方法或传统的 AI 技术难以用来对具有这种性质的大量历史数据记录进行发掘、建立实用的预测模型。然而计算智能方法本身就有的自适应、自组织学习机制,正好可以在这方面发挥作用。因此,早已引起了以华尔街为代表的金融界的重视,进行了应用开发。例如,伦敦花旗银行早在 90 年代初期就已开发应用了一个预测外汇市场兑换率变化趋势的系统,它的主体是一个神经网络。用遗传算法对种种外汇市场技术指标进行优化预处理后,再汇同有关市场基础数据输入神经网络。神经网络

将按先经过训练而建立(发现)的预测模型输出关于应当买进还是卖出的建议^[8]。我们也研究开发了一个分析、预测股票市场的实验系统原型,其主体是一个神经网络(多层感知器),不同的是,不是用遗传算法来对输入数据进行预处理,而是配合神经网络的训练过程,对它的结构、性能参数(隐单元数目、学习率、动量系数、初始权值范围、输出偏差允许范围等)进行优化,从而自动实现一个训练时间较短、准确性较高的神经网络预测模型。此外,为进一步提高预报的准确率,还用了一个模糊决策子系统来综合处理神经网络的输出和影响市场心理的那些难以预料或量化的因素,如天灾人祸、社会动荡等等^[9,10]。我们还试验了直接用 GA 代替 BP 算法训练神经网络,取得了明显的效果(将另文发表)。

总之,数据发掘技术的研究开发,将涉及数据库系统、决策支持系统、人工智能与计算智能、知识工程、分布式并行处理、多媒体网络技术、计算可视化等多种学科分支的理论与技术。特别是数据发掘的目标是要从数据库发现隐藏在大量数据记录中的未知知识,这种知识发现的实质是机器学习,而机器学习又是人工智能所面临的难题之一。所以,数据发掘作为一项新兴的高新技术,为解决所面临的种种理论上或实现技术上的难题、难点,已形成一个方兴未艾的国际前沿研究开发新领域。

主要参考文献

- [1] Special Issue of "IEEE EXPERT" on Data Mining, Oct. 1996
- [2] KDD Nuggets electronic newsletter, <http://info.gte.com/~kdd/nuggets/>, Feb./Mar. 1997
- [3] Database Research: Achievements and Opportunities into the 21st century, by Avi Silberschatz, et al., 1995, 中译文见《计算机科学》1996 第 4 期
- [4] Software Datamining, by Cheyl Gerber, DATAMATION, May 1, 1996
- [5] Zurada, J. M., et al. ed., 《Computational Intelligence: Imitating Life》, IEEE Press, 1994
- [6] Soros, George, 《The Alchemy of Finance-Reading the Mind of the Market》, John Wiley & Sons, 1984-1985
- [7] Special Issue on "AI in Finance", AI EXPERT, 1992
- [8] Colin, A. M., Neural Network and Genetic Algorithm for Exchange Rate Forecasting, Proc. of IJCNN' 92, Beijing, Nov. 15, 1992
- [9] 费良俊, 童 颖, 计算智能在金融市场分析中的应用, 第四届中国人工智能联合学术会议(CJCAI' 96)论文集, 清华大学出版社, 1996
- [10] 费良俊, 童 颖, 发现金融市场预测模型的计算智能方法, (待发表)