

知识的自动发现

Automatic Knowledge Discovery

知识的自动发现是一个横跨机器学习、数据挖掘、认知科学、脑科学和哲学等多学科领域。简单地说，知识的自动发现是用计算机从大量、繁杂的数据中自动地发现其隐含的、有用的并具有一定概括性的知识。例如，从各种行星轨迹的记录中自动发现开普勒定律；从超市大量的客户消费数据中获取哪些物品经常被一起购买的信息。知识的自动发现标志着人类不仅能够对自然界和人类本身有足够的理解(如牛顿定律等)，而且能够重复并扩展这种发现的过程，解释知识发现的本质和机理。同时，在信息化时代，随着海量复杂数据的不断涌现，人们越来越意识到对这些数据进行知识的自动发现可能带来巨大的商机与社会价值。

知识的自动发现与人类对知识的探索是息息相关的。从亚里士多德用逻辑对知识的表达，到达芬奇对各种物理现象的推演和形象描述，直至牛顿等对天文和物理等自然现象利用自然定律的总结，都可以认为是人类早期对知识发现所做出的努力。为了使人类能在更为广泛的领域进行知识发现，学者对知识发现的过程与原理产生了浓厚的兴趣。有学者通过研究认知科学(cognitive science)、综合语言学、心理学、神经科学、哲学和人工智能，来研究大脑如何对知识进行获取、储存、变换与传播，研究结果表明，知识发现中的一个要点是“学习提升智能”^[1]。最近，通过功能磁共振成像(fMRI)等技术，学者可以较精确地测量大脑各区域脑血流的变化及其功能结构，进而研究人的知识发现过程。

为达到同样的目标，人工智能的学者们则探讨如何利用计算机对大量数据进行高通量的积累和处理，来实现知识发现的目的。在给定计算机足够的物理数据、生物数据和商业数据的情况下，计算机能不能自动地发现牛顿定律，推出DNA的双螺旋的结构和自动挖掘出有效的商业信息呢(如图1所示)?

现在，我们得到的知识、发现的成果还远远无法达到“计算机像人类科学家一样进行知识发现”这一目标，但这个目标一直是学者们努力的方向。早在20世纪70年代，Langley教授^[2]就描述了自动发现物理学定律的系统BACON。BACON算法主要是基于专家系统的生成规则对各种变量的子集进行搜索，用穷举法得到如 $F=ma$ 这样的定律。虽然只能发现一些较为规则的物理和化学定律

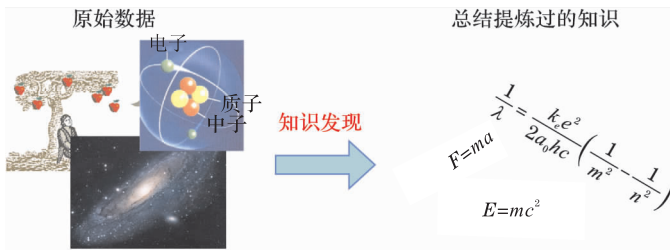


图 1 知识发现的目标

(如开普勒定律等), 并需要非常纯净的少量数据作为输入, 但这些成功的实例为后来的工作提出了两个重要概念。首先, 知识发现需要我们能自动地提出上层的知识模式, 即“学什么”的问题。例如, 在研究物理问题中, 我们有大量的实验变量, 如速度 v 、质量 m 、力 F 、温度 T 等, 那么, 我们应该找 (v, m, T) 之间的关系, 还是应该找 $(\frac{dv}{dt}, \frac{d^2v}{dt^2}, m, F)$ 之间的关系呢? 这涉及如何自动地进行特征构建 (feature generation) 的问题。同时, 我们也要有一整套底层的发现模块, 用来解决“怎么学”的问题。例如, 如果 $F=Wma$, W 是一个常量, 那么, W 应该是多少呢? 至今, 知识发现领域内的主要工作都集中在后者, 即“怎么学”的问题上, 如这个 W 值, 就可以通过回归的方法得到。

在“怎么学”的问题上, 主要的工作分为分类学习 (classification)、回归算法 (regression) 和聚类算法 (clustering) 等。分类学习的目标是在数据中找到主要的因素, 以对一个特定的标记进行预测。例如, 给定职员的个人信 息, 可以预测职员的收入高低 (如图 2 决策树所示)。回归算法则是对数值的预测, 如预测某种气体的温度有多高。基于有标记的训练数据的学习过程叫做有监督学习, 这样的知识发现任务虽然简单 (预测两类, 多分类或数值标记), 但在科学领域已经非常有用了。例如, 在生物信息学领域, 人类基因组测序计划初步完成和癌症基因组计划的开展积累了大量的实验数据, 机器学习成为生物信息学中的重要工具。在研究疾病, 尤其是癌症的致病基因时, 监督分类学习从病人和正常人的基因序列、SNPs 序列或基因表达数据出发, 通过特征选择得到最能区分病人和正常人的单个基因或者多个基因, 进而帮助设计新的生物学实验以验证致病基因。当知识发现的任务里没有任何监督信息时, 聚类算法也可以从数据之间的相关性出发, 对数据中隐含的聚类知识进行挖掘。学者同时开始研究如何在只有少量有标记数据而存在大量未标记数据的情况下进行学习的问题, 也就是半监督学习问题。半监督学习试图找到未标记数据分布和学习目标之间的联系, 然后利用未标记数据来辅助提高学习性能^[3]。

除了对科学定律进行自动知识发现, 学者们同时对不同应用领域的知识发

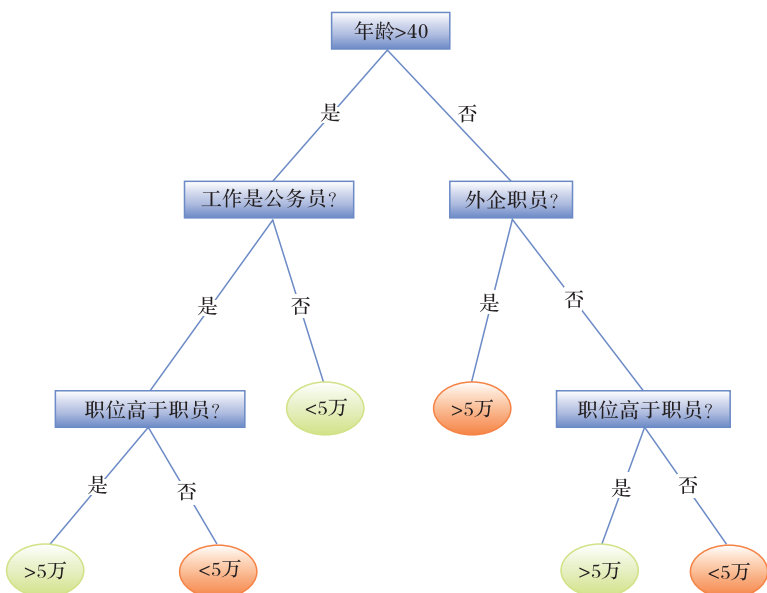


图2 分类算法中的决策树模型

现进行了广泛的研究。在商业应用之中，我们也经常会碰到在大规模数据中找出不同变量之间的联系的问题。众多学者对“关联规则挖掘”（association rule mining）进行了大量研究。早期的工作包括 Agrawal 等在 1994 年提出的 Apriori 算法^[4]。Apriori 算法能够通过广度优先的搜索，发现“买‘知识发现’方面图书的人也很可能买‘数据挖掘’方面的书”这样的关联规则。此后，大批学者开展了后续研究，从时间复杂度与空间复杂度等方面优化关联规则挖掘。在互联网方面，搜索引擎成为最主要的知识发现工具。最成功的实例要数 Page 与 Brin 等在 1999 年提出的 PageRank 算法^[5]。PageRank 算法通过网页之间超链接的关系计算出网页的重要度，并在应用中获得了良好的效果，成为了现在广为人知的 Google 搜索引擎的核心组成部分，每天都被全世界成千上万的用户使用。在电子商务方面，对诸多网络用户而言，互联网上的信息在极大丰富的同时，不可避免地带来了信息过载的困扰。当在售的产品经常数以亿计时，个性化推荐技术应运而生。数据挖掘根据用户买过的商品、用户经常点击的网页等信息，自动去推测会有哪些新的信息跟用户相关，此方面的研究成果已经广泛应用于商业系统，如全球最大的在线书籍销售网站 Amazon 及在线影碟租赁网站 Netflix 都使用了自动知识发现的方法从成千上万的商品中为用户挑选出他们最有可能感兴趣的那些。在社会化网络 (social networks) 应用方面，社会化计算 (如 Web 2.0) 发展迅速，数据挖掘已成为社会网挖掘的重要工具。由此也引出许多

新的研究，如在社会化网络上对信息传播机制的研究、网络动态演变的研究等。继互联网之后，物联网也逐渐流行起来。在这一背景下，传感器及传感器网成为获取人或物体行动信息的重要手段。学者们研究如何用机器学习对传感器数据建模以精确识别人或动物的行为，预测其位置和行为信息。

上面提到 Langley 的 BACON 系统提出知识发现的两个层次，其下层模块“怎么学”的问题已经有了很多的研究和应用。但是，对于上层的“学什么”的问题，至今尚未得到很好的解答。可见，对这样一种自动学习过程，需要从全局角度将不同的学习问题有机的结合起来。下面，我们将展望知识发现领域中三个重要且相互联系的热点问题。概括起来说，这三个问题是如何能够自动地进行“借鉴和类比”、“复杂问题的简约”和“产生灵感”。

首先，知识的自动发现需要计算机有联想、借鉴和类比的功能，使其能够举一反三。就这一点，我们来讨论迁移学习。在现实问题中，我们可能难以得到同种类或同分布的训练数据。此时，学者们开始尝试借用其他任务的数据或其他种类的数据辅助学习，这种学习称为迁移学习^[6]。迁移学习类似于我们人类利用象棋的思想与套路学习机器人足球(如图 3 所示)，它从很大程度上量化了人工智能早期的类比学习(learning by analogy)的目标^[7]，使得我们能够从其他相关领域中得到的启发，利用现代机器学习的手段来解决一个新出现的较难问题。现阶段的迁移学习研究可以分为两个方面：同构空间下的迁移学习和异构空间下的迁移学习。在同构空间下，尽管辅助训练数据和源训练数据或多或少会有些不同，但辅助训练数据中还是应该存在一部分数据比较适合用来训练一个有效的分类模型，并且适应测试数据，这是基于实例的迁移学习的基本思想。基于特征的迁移学习的基本思想是同时考虑源数据与辅助数据，以得到一个更好的共同特征空间，通过在这个新空间表示源数据实现迁移学习。

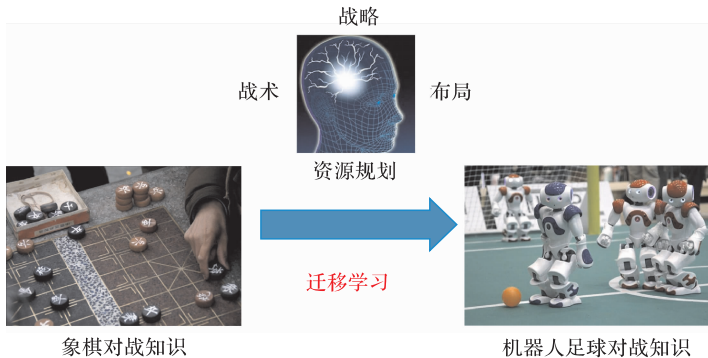


图 3 迁移学习

在异构空间下，即非常不同的领域之间进行迁移学习，是实现借鉴和类比功能的另一个方面。在非常不同的领域之间进行迁移学习难度在于如何找到能迁移的知识。例如，我们会问：学习绘画创作与学习音乐创作之间是否有公共知识可以迁移呢？人类可以做到这一点，因为在创作意图、情感及表达上，作为艺术的两种形式，美术与音乐有很多相通之处。而对于计算机来说，做到这一点并不容易。在已有的工作中，学者已经在文字和图像之间的转换问题上做了很好的尝试^[8]。一般来说，异构空间下的迁移学习致力于解决训练数据与测试数据分别属于两个不同特征空间的情况，其基本思想是：首先获得一个连接不同特征空间的桥梁，然后通过这个桥梁将辅助数据迁移到源数据特征空间里去，用一个统一的模型进行学习及预测。能够在更加广泛、缺少直接联系的任务间做到举一反三和触类旁通是当前知识自动发现这一方向的挑战与热点之一。

其次，知识的自动发现需要计算机能够把不同的知识联系起来，同时，把一个复杂的问题化解或简约(reduction)为几个简单的问题来解决。结合领域的背景知识把不同的学习子任务联系起来，是关系学习(relational learning)的一个主要目的。其中的一个重要手段是将逻辑引入统计学习，以达到把一个全局的学习问题分解与简化的目的。马尔可夫逻辑网(Markov logic networks, MLNs)^[9]就是这样的一个方法，它的基本思想可以通过一个例子来表达。例如，假设我们要通过网页间的联系来发现一个学校里老师和学生之间诸如“教育与被教育”之类的关系，这种关系可以通过研究大量的数据中存在的共同现象被发现与证实。例如，他们经常在同一门课的网页里出现或常在同一篇文章中为合作者，这些都可以作为他们师生关系的依据。在马尔可夫逻辑网里，这些关系可以用带权值的逻辑公式来描述。如果我们对某些人不是很确定他们有这样的关系，那么，可以赋予这个逻辑公式一个较小的权值，反之，也可以增加权值，而这些权值可以从数据中通过优化及近似等方法得到。这样一个有权值的逻辑网络架构能够更好地表达很多仅用统计所不能表达的特征。今天，马尔可夫逻辑网已被应用于文本分析、生物信息学及机器人的动作模型学习中。

另一个新的备受关注的方向是如何把一个困难的学习问题转化为一系列的相对容易或已有答案的学习问题。例如，学习一个英文句子的语法时，可以化解为一系列的子问题，如“学习第 t 个词的词意”和“学习第 t 个词和第 $t+1$ 个词之间的联系”等，我们对其中每一个子问题都已有一些可行的算法。这类问题叫做“简约学习”(machine learning reductions)^[10]。学者们现在的做法大都是把一个有结构的问题(structured learning problem)化简为许多子学习问题。

我们看到，目前的知识发现系统仅局限于建立模型、完成简单的任务，如分类等。今天的计算机还不能做到自动地发现普遍存在于宏观和微观社会中的科学定律和自然规律的程度，如从大量的物理文献、理论与实验观察里得出像爱因斯坦的 $E = mc^2$ 那样深奥的理论。其中一个很重要的原因是我们对人类知识发现过程中“灵感”的产生机理还不太了解。这是我们要提的第三个问题，也是一个非常有趣的问题：知识发现过程中的灵感到底是从哪里来呢？

可以想象，当数据量和复杂程度达到一定规模的时候，所能发现的知识很可能发生某种“相变”，就像水在某种温度下变成冰一样，而这种相变很可能就相当于人的灵感的产生。从计算的角度，在数据量骤增的情况下，知识发现是不是也有一个从量变到质变的过程，可以让计算机像人类一样地产生灵感呢？这种研究现在，已经有了一定的可能性，其基础之一就是大规模分布式的机器学习算法。现在，这种算法的研究已经成为科研与在线工业应用的一个新的热点。从特征的角度，我们要解决的实际问题中所包含的特征数目通常非常巨大，例如，从 2010 年的知识发现大赛的数据中我们可提取多达三千万个特征，显然，这些特征是互相联系的，并且含有冗余。人类可以从大量的特征中综合、提炼出最重要的数个重点进行知识发现，并自动地产生新的特征，计算机是否也能做到这一点呢？现在，许多学者开始关注这个问题，思考如何自动地进行特征构建，如建立层次化的特征空间或深度信念网络(deep belief networks)。当前，自动地进行特征构建仍然属于知识发现领域中的薄弱环节，但渐渐成为了一个研究热点。

灵感来源于实践。我们过去的经验很可能为我们将来的学习问题提供答案，这就需要我们所说的终生学习。“活到老，学到老”是中国人耳熟能详的名言，我们同样希望计算机能够终身学习。当然，对于计算机而言，“终生学习”指的是永不休止的学习。现有的机器学习技术主要是针对某个特定的任务与数据集设计，一旦算法结束，学习过程也将中止，而这个任务学到的知识也不会用到以后的学习过程中。因此，为了更好地利用各种来源的数据与知识，我们希望机器学习也能终生学习。现在，这类研究已经开始，如卡内基梅隆大学的“永不终止的语言学习”(never ending language learning)系统。

知识的自动发现这一领域从探索之初至今仅五十多年的时间，但却已渗透到了生活中的方方面面。虽然当前知识的自动发现还面临诸多挑战，与人类相比仍显得十分简单，但是，考虑到人类进化、发展了二十万年才获得的当前知识与文明，假以时日，知识的自动发现一定也能有更大的发展。

致谢 感谢陈雨强、刘楠、金鸥、徐倩、胡昊等同学及朱小燕教授的讨论和校对。

参 考 文 献

- [1] 史忠植,余志华. 认知科学和计算机. 北京:科普出版社,1990.
- [2] Langley P. Rediscovering physics with BACON. 3. Proceedings of the 6th International Joint Conference on Artificial Intelligence,1979,1:505—507.
- [3] Zhu X. Semi-supervised learning literature survey. Madison; University of Wisconsin-Madsion,2006.
- [4] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, 1994: 487—499.
- [5] Page L, Brin S, Motwani R, et al. The page Rank citation ranking: Bringing order to the web. Proceedings of ASIS98,1998.
- [6] Pan S J, Yang Q. A Survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345—1359.
- [7] Falkenhainer B, Forbus K D, Gentner D. The structure-mapping engine. Proceedings of the American Association of Artificial Intelligence, 1986: 272—277.
- [8] Lu R, Zhang S. Automatic Generation of Computer Animation: Using AI for Movie Animation. New York: Springer, 2002.
- [9] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006, 62(1—2): 107—136.
- [10] Beygelzimer A, Langford J, Zadrozny B. Reductions in machine learning. Proceedings of International Conference on Machine Learning, 2009.

撰稿人: 杨 强¹ 薛贵荣²

1 香港科技大学计算机科学与工程系

2 上海交通大学计算机科学与工程系