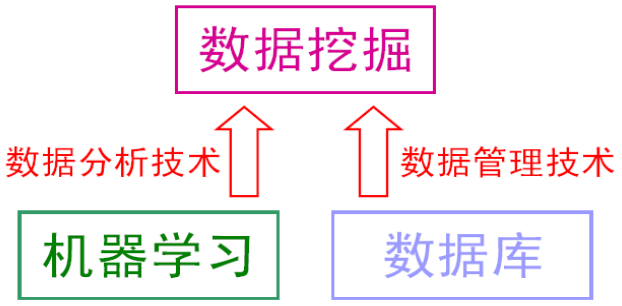
机器学习与数据挖掘

周 志 华

南京大学计算机软件新技术国家重点实验室，南京 210093

“机器学习”是人工智能的核心研究领域之一， 其最初的研究动机是为了让计算机系统具有人的学习能力以便实现人工智能，因为众所周知，没有学习能力的系统很难被认为是具有智能的。目前被广泛采用的机器学习的定义是“利用经验来改善计算机系统自身的性能”[1]。事实上，由于“经验”在计算机系统中主要是以数据的形式存在的，因此机器学习需要设法对数据进行分析，这就使得它逐渐成为智能数据分析技术的创新源之一，并且为此而受到越来越多的关注。

“数据挖掘”和“知识发现”通常被相提并论，并在许多场合被认为是可以相互替代的术语。对数据挖掘有多种文字不同但含义接近的定义，例如“识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”[2]。其实顾名思义，数据挖掘就是试图从海量数据中找出有用的知识。大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据。



因为机器学习和数据挖掘有密切的联系，受主编之邀，本文把它们放在一起做一个粗浅的介绍。

1 无处不在

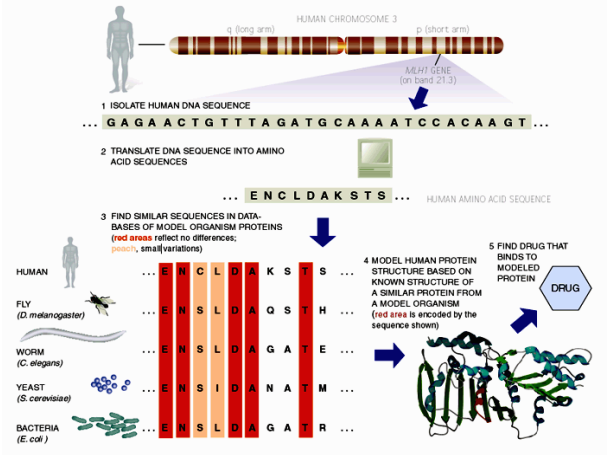
随着计算机技术的飞速发展，人类收集数据、存储数据的能力得到了极大的提高，无论是科学研究还是社会生活的各个领域中都积累了大量的数据，对这些数据进行分析以发掘数据中蕴含的有用信息，成为几乎所有领域的共同需求。正是在这样的大趋势下，机器学习和数据挖掘技术的作用日渐重要，受到了广泛的关注。

例如，网络安全是计算机界的一个热门研究领域，特别是在入侵检测方面，不仅有很多理论成果，还出现了不少实用系统。那么，人们如何进行入侵检测呢？首先，人们可以通过检查服务器日志等手段来收集大量的网络访问数据，这些数据中不仅包含正常访问模式还包含入侵模式。然后，人们就可以利用这些数据建立一个可以很好地把正常访问模式和入侵模式分开的模型。这样，在今后接收到一个新的访问模式时，就可以利用这个模型来判断这个模式是正常模式还是入侵模式，甚至判断出具体是何种类型的入侵。显然，这里的关键问题是如何利用以往的网络访问数据来建立可以对今后的访问模式进行分类的模型，而这正是机器学习

和数据挖掘技术的强项。

实际上，机器学习和数据挖掘技术已经开始在多媒体、计算机图形学、计算机网络乃至操作系统、软件工程等计算机科学的众多领域中发挥作用，特别是在计算机视觉和自然语言处理领域，机器学习和数据挖掘已经成为最流行、最热门的技术，以至于在这些领域的顶级会议上相当多的论文都与机器学习和数据挖掘技术有关。总的来看，引入机器学习和数据挖掘技术在计算机科学的众多分支领域中都是一个重要趋势。

机器学习和数据挖掘技术还是很多交叉学科的重要支撑技术。例如，生物信息学是一个新兴的交叉学科，它试图利用信息科学技术来研究从 DNA 到基因、基因表达、蛋白质、基因电路、细胞、生理表现等一系列环节上的现象和规律。随着人类基因组计划的实施，以及基因药物的美好前景，生物信息学得到了蓬勃发展。实际上，从信息科学技术的角度来看，生物信息学的研究是一个从“数据”到“发现”的过程，这中间包括数据获取、数据管理、数据分析、仿真实验等环节，而“数据分析”这个环节正是机器学习和数据挖掘技术的舞台。



正因为机器学习和数据挖掘技术的进展对计算机科学乃至整个科学技术领域都有重要意义，美国NASA-JPL实验室的科学家 2001 年 9 月在《Science》上专门撰文[3]指出，机器学习对科学研究的整个过程正起到越来越大的支持作用，并认为该领域将稳定而快速地发展，并将对科学技术的发展发挥更大的促进作用。NASA-JPL实验室的全名是美国航空航天局喷气推进实验室，位于加州理工学院，是美国尖端技术的一个重要基地，著名的“勇气”号和“机遇”号火星机器人正是在这个实验室完成的。从目前公开的信息来看，机器学习和数据挖掘技术在这两个火星机器人上有大量的应用。

除了在科学研究中发挥重要作用，机器学习和数据挖掘技术和普通人的生活也息息相关。例如，在天气预报、地震预警、环境污染检测等方面，有效地利用机器学习和数据挖掘技术对卫星传递回来的大量数据进行分析，是提高预报、预警、检测准确性的重要途径；在商业营销中，对利用条形码技术获得的销售数据进行分析，不仅可以帮助商家优化进货、库存，还可以对用户行为进行分析以设计有针对性的营销策略；……。下面再举两个例子。

公路交通事故是人类面临的最大杀手之一，全世界每年有上百万人丧生车轮，仅我国每年就有约 10 万人死于车祸。美国一直在对自动驾驶车辆进行研究，因为自动驾驶车辆不仅在军事上有重要意义，还对减少因酒后、疲劳而引起的车祸有重要作用。2004 年 3 月，在美国 DARPA（国防部先进研究计划局）组织的自动驾驶车辆竞赛中，斯坦福大学的参赛车在完全无人控制的情况下，成功地在 6 小时 53 分钟内走完了 132 英里（约 212 公里）的路程，获得了冠军。比赛路段是在内华达州西南部的山区和沙漠中，路况相当复杂，有的地方路面只有几米宽，一边是山岩，另一边是百尺深沟，即使有丰富驾驶经验的司机，在这样的路段上行车也是一个巨大的挑战。这一结果显示出自动

驾驶车辆已经不再是一个梦想，可能在不久的将来就会走进普通人的生活。值得一提的是，斯坦福大学参赛队正是由一位机器学习专家所领导的，而获胜车辆也大量使用了机器学习和数据挖掘技术。

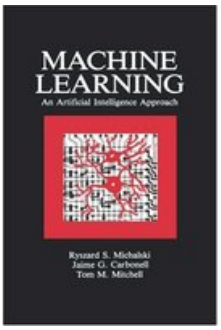
Google、Yahoo、百度等互联网搜索引擎已经开始改变了很多人的生活方式，例如很多人已经习惯于在出行前通过网络搜索来了解旅游景点的背景知识、寻找合适的旅馆、饭店等。美国新闻周刊曾经对 Google 有个“一句话评论”：“它使得任何人离任何问题的答案之间的距离只有点击一下鼠标这么远”。现在很少有人不知道互联网搜索引擎的用处，但可能很多人并不了解，机器学习和数据挖掘技术正在支撑着这些搜索引擎。其实，互联网搜索引擎是通过分析互联网上的数据来找到用户所需要的信息，而这正是一个机器学习和数据挖掘任务。事实上，无论 Google、Yahoo 还是微软，其互联网搜索研究核心团队中都有相当大比例的人是机器学习和数据挖掘专家，而互联网搜索技术也正是机器学习和数据挖掘目前的热门研究话题之一。

2 雄关漫道

机器学习是人工智能研究发展到一定阶段的必然产物。从 20 世纪 50 年代到 70 年代初，人工智能研究处于“推理期”，人们认为只要给机器赋予逻辑推理能力，机器就能具有智能。这一阶段的代表性工作主要有 A. Newell 和 H. Simon 的“逻辑理论家”程序以及此后的“通用问题求解”程序等，这些工作在当时取得了令人振奋的成果。例如，“逻辑理论家”程序在 1952 年证明了著名数学家罗素和怀特海的名著《数学原理》中的 38 条定理，在 1963 年证明了全部的 52 条定理，而且定理 2.85 甚至比罗素和怀特海证明得更巧妙。A. Newell 和 H. Simon 因此获得了 1975 年图灵奖。然而，随着研究向前发展，人们逐渐认识到，仅具有逻辑推理能力是远远实现不了人工智能的。E.A. Feigenbaum 等人认为，要使机器具有智能，就必须设法使机器拥有知识。在他们的倡导下，20 世纪 70 年代中期开始，人工智能进入了“知识期”。在这一时期，大量专家系统问世，在很多领域做出了巨大贡献。E.A. Feigenbaum 作为“知识工程”之父在 1994 年获得了图灵奖。但是，专家系统面临“知识工程瓶颈”，简单地说，就是由人来把知识总结出来再教给计算机是相当困难的。于是，一些学者想到，如果机器自己能够学习知识该多好！

实际上，图灵在 1950 年提出图灵测试的文章中，就已经提到了机器学习的可能，而 20 世纪 50 年代其实已经开始有机器学习相关的研究工作，主要集中在基于神经网络的连接主义学习方面，代表性工作主要有 F. Rosenblatt 的感知机、B. Widrow 的 Adaline 等。在 20 世纪 6、70 年代，多种学习技术得到了初步发展，例如以决策理论为基础的统计学习技术以及强化学习技术等，代表性工作主要有 A.L. Samuel 的跳棋程序以及 N.J. Nilson 的“学习机器”等，20 多年后红极一时的统计学习理论的一些重要结果也是在这个时期取得的。在这一时期，基于逻辑或图结构表示的符号学习技术也开始出现，代表性工作有 P. Winston 的“结构学习系统”、R.S. Michalski 等人的“基于逻辑的归纳学习系统”、E.B. Hunt 等人的“概念学习系统”等。

1980 年夏天，在美国卡内基梅隆大学举行了第一届机器学习研讨会；同年，《策略分析与信息系统》连出三期机器学习专辑；1983 年，Tioga出版社出版了R.S. Michalski、J.G. Carbonell和T.M. Mitchell主编的《机器学习：一种人工智能途径》[4]，书中汇集了 20 位学者撰写的 16 篇文章，对当时的机器学习研究工作进行了总结，产生了很大反响[a]；1986 年，《Machine Learning》创刊；1989 年，《Artificial Intelligence》出版了机器学习专辑，刊发了一些当时比较活跃的研究工作，其内容后来出现在J.G. Carbonell主编、MIT出版社 1990 年出版的《机器学习：风范与方法》[5]一书中。总的来看，20 世纪 80 年代是机器学习成为一个独立的学科领域并开始快速发展、各种机器学习技术百花齐放的时期。



R.S. Michalski等人[4]中把机器学习研究划分成“从例子中学习”、“在问题求解和规划中学习”、“通过观察和发现学习”、“从指令中学习”等范畴；而E.A. Feigenbaum在著名的《人工智能手册》[b] 中[6]，则把机器学习技术划分为四大类，即“机械学习”、“示教学习”、“类比学习”、“归纳学习”。机械学习也称为“死记硬背式学习”，就是把外界输入的信息全部记下来，在需要的时候原封不动地取出来使用，这实际上没有进行真正的学习；示教学习和类比学习实际上类似于R.S. Michalski等人所说的“从指令中学习”和“通过观察和发现学习”；归纳学习类似于“从例子中学习”，即从训练例中归纳出学习结果[c]。20 世纪 80 年代以来，被研究得最多、应用最广的是“从例子中学习”（也就是广义的归纳学习），它涵盖了监督学习（例如分类、回归）、非监督学习（例如聚类）等众多内容。下面我们对这方面主流技术的演进做一个简单的回顾。

在 20 世纪 90 年代中期之前，“从例子中学习”的一大主流技术是归纳逻辑程序设计（Inductive Logic Programming），这实际上是机器学习和逻辑程序设计的交叉。它使用 1 阶逻辑来进行知识表示，通过修改和扩充逻辑表达式（例如Prolog表达式）来完成对数据的归纳。这一技术占据主流地位与整个人工智能领域的发展历程是分不开的。如前所述，人工智能在 20 世纪 50 年代到 80 年代经历了“推理期”和“知识期”，在“推理期”中人们基于逻辑知识表示、通过演绎技术获得了很多成果，而在知识期中人们基于逻辑知识表示、通过领域知识获取来实现专家系统，因此，逻辑知识表示很自然地受到青睐，而归纳逻辑程序设计技术也自然成为机器学习的一大主流。归纳逻辑程序设计技术的一大优点是它具有很强的知识表示能力，可以较容易地表示出复杂数据和复杂的数据关系。尤为重要的是，领域知识通常可以方便地写成逻辑表达式，因此，归纳逻辑程序设计技术不仅可以方便地利用领域知识指导学习，还可以通过学习对领域知识进行精化和增强，甚至可以从数据中学习出领域知识。事实上，机器学习在 20 世纪 80 年代正是被视为“解决知识工程瓶颈问题的关键”而走到人工智能主舞台的聚光灯下的，归纳逻辑程序设计的一些良好特性对此无疑居功至伟[d]。S.H. Muggleton主编的书[7]对 90 年代中期之前归纳逻辑程序设计方面的研

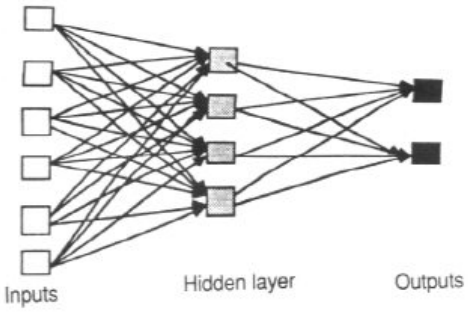
```
plus(0, X, X).
plus(X, 0, X).
plus(A, B, C)  :- dec(A, D),  inc(B, E),  plus(D, E, C).

mult(0, X, 0).
mult(A, B, C)  :- dec(A, D),  mult(D, B, E),  plus(B, E, C).
```

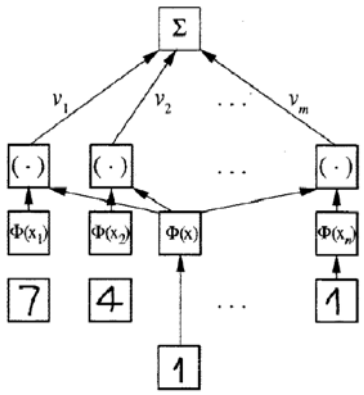
[a] Morgan Kaufmann出版社后来分别于 1986 年和 1990 年出版了该书的续篇，编为第二卷和第三卷。
[b] 该书共 4 卷，分别由E.A. Feigenbaum与不同的学者合作编写而成。
[c] “归纳学习”有狭义的解释和广义的解释。前者要求从训练数据中学得概念，因此也被称为“概念学习”或“概念形成”；后者则对学习结果是否是可理解的概念不做要求。
[d] “归纳逻辑程序设计”这个名字其实是 1991 年S. Muggleton才提出的。

究工作做了总结。然而，归纳逻辑程序设计技术也有其局限，最严重的问题是由于其表示能力很强，学习过程所面临的假设空间太大，对规模稍大的问题就很难进行有效的学习，只能解决一些“玩具问题”。因此，在 90 年代中期后，归纳程序设计技术方面的研究相对陷入了低谷。

20 世纪 90 年代中期之前，“从例子中学习”的另一大主流技术是基于神经网络的连接主义学习。连接主义学习技术在 20 世纪 50 年代曾经历了一个大发展时期，但因为早期的很多人工智能研究者对符号表示有特别的偏爱，例如H. Simon曾说人工智能就是研究“对智能行为的符号化建模”，因此当时连接主义的研究并没有被纳入主流人工智能的范畴。同时，连接主义学习自身也遇到了极大的问题，M. Minsky和S. Papert在 1969 年指出，（当时的）神经网络只能用于线性分类，对哪怕“异或”这么简单的问题都做不了。于是，连接主义学习在此后近 15 年的时间内陷入了停滞期。直到 1983 年，J.J. Hopfield利用神经网络求解TSP问题获得了成功，才使得连接主义重新受到人们的关注。1986 年，D.E. Rumelhart和J.L. McClelland主编了著名的《并行分布处理—认知微结构的探索》[8]一书，对PDP小组的研究工作进行了总结，轰动一时。特别是D.E. Rumelhart、G.E. Hinton和R.J. Williams重新发明了著名的BP算法[e]，产生了非常大的影响。该算法可以说是最成功的神经网络学习算法，在当时迅速成为最流行的算法，并在很多应用中都取得了极大的成功。与归纳逻辑程序设计技术相比，连接主义学习技术基于“属性-值”的表示形式（也就是用一个特征向量来表示一个事物；这实际上是命题逻辑表示形式），学习过程所面临的假设空间远小于归纳逻辑程序设计所面临的空间，而且由于有BP这样有效的学习算法，使得它可以解决很多实际问题。事实上，即使在今天，BP仍然是在实际工程应用中被用得最多、最成功的算法之一。然而，连接主义学习技术也有其局限，一个常被人诟病的问题是其“试错性”。简单地说，在此类技术中有大量的经验参数需要设置，例如神经网络的隐层结点数、学习率等，夸张一点说，参数设置上差之毫厘，学习结果可能谬以千里。在实际工程应用中，人们可以通过调试来确定较好的参数设置，但对机器学习研究者来说，对此显然是难以满意的。



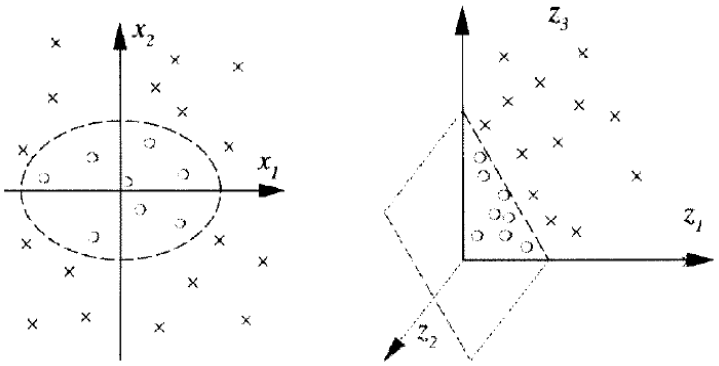
20 世纪 90 年代中期，统计学习粉墨登场并迅速独占鳌头。其实早在 20 世纪 6、70 年代就已经有统计学习方面的研究工作，统计学习理论[9]在那个时期也已经打下了基础，例如V.N. Vapnik早在 1963 年就提出了“支持向量”的概念，他和A.J. Chervonenkis在 1968 年提出了VC维，在 1974 年提出了结构风险最小化原则等，但直到 90 年代中期统计学习才开始成为机器学习的主流技术。这一方面是由于有效的支持向量机算法在 90 年代才由B.E. Boser、I. Guyon和V.N. Vapnik提出，而其优越的性能也是到 90 年代中期才在T. Joachims等人对文本分类的研究中显现出来；另一方面，正是在连接主义学习技术的局限性凸显出来之后，人们才把目光转向了统计学习。事实上，



[e] 实际上，P. Werbos在他 1974 年哈佛大学的博士学位论文中曾经发明了这个算法，但由于当时正处于连接主义的“冰河期”，因此没有得到应有的重视。

统计学习与连接主义学习有着密切的联系，例如RBF神经网络其实就是一种很常用的支持向量机。

在支持向量机被普遍接受后，支持向量机中用到的核（kernel）技巧被人们用到了机器学习的几乎每一个角落中，“核方法”也逐渐成为机器学习的一种基本技巧。但其实这并不是一种新技术，例如Mercer定理是在 1909 年发表的，核技巧也早已被很多人使用过，即使只考虑机器学习领域，至少T. Poggio在 1975 年就使用过多项式核。如果仔细审视统计学习理论，就可以发现其中的绝大多数想法在以往机器学习的研究中都出现过，例如结构风险最小化原则实际上就是对以往机器学习研究中经常用到的最小描述长度原则的另一个说法。但是，统计学习理论把这些有用的片段整合在同一个理论框架之下，从而为人们研制出泛化能力[f]有理论保证的算法奠定了基础，与连接主义学习的“试错法”相比，这是一个极大的进步。然而，统计学习也有其局限，例如，虽然理论上来说，通过把原始空间利用核技巧转化到一个新的特征空间，再困难的问题也可以容易地得到解决，但如何选择合适的核映射，却仍然有浓重的经验色彩。另一方面，统计学习技术与连接主义学习技术一样是基于“属性-值”表示形式，难以有效地表示出复杂数据和复杂的数据关系，不仅难以利用领域知识，而且学习结果还具有“黑箱性”。此外，传统的统计学习技术往往因为要确保统计性质或简化问题而做出一些假设，但很多假设在真实世界其实是难以成立的。如何克服上述缺陷，正是很多学者正在关注的问题。



需要说明的是，机器学习目前已经是一个很大的学科领域，而本节只是管中窥豹，很多重要的内容都没有谈及。T.G. Dietterich曾发表过一篇题为《机器学习研究：当前的四个方向》[10]的很有影响的文章，在文章中他讨论了集成学习、可扩展机器学习（例如对大数据集、高维数据的学习等）、强化学习、概率网络等四个方面的研究进展，有兴趣的读者不妨一读。

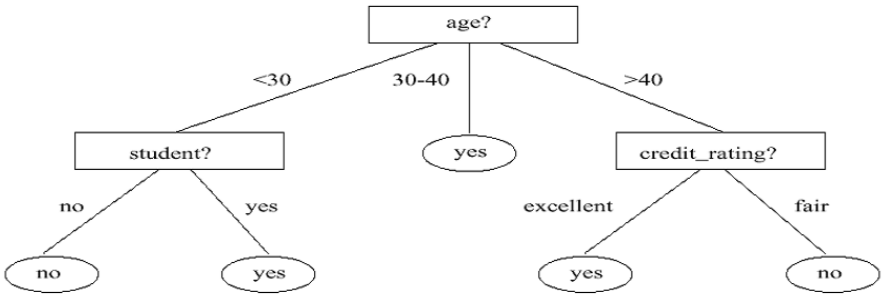
如前所述，机器学习之所以备受瞩目，主要是因为它已成为智能数据分析技术的创新源之一。但是机器学习还有一个不可忽视的功能，就是通过建立一些关于学习的计算模型来帮助人们了解“人类如何学习”。例如，P. Kanerva在 20 世纪 80 年代中期提出SDM（Sparse Distributed Memory）模型时并没有刻意模仿人脑生理结构，但后来的研究发现，SDM的工作机制非常接近于人类小脑，这为理解小脑的某些功能提供了帮助。自然科学研究的驱动力归结起来无非是人类对宇宙本源、物质本性、生命本质、自我本识的好奇，而“人类如何学习”无疑是一个有关自我本识的重大问题。从这个意义上说，机器学习不仅在信息科学中占有重要地位，还有一定的自然科学色彩。与此不同，数据挖掘[11]则是一个直接为实际应用而生的学科领域。20 世纪 60 年代，早期的数据库问世，人们开始利用计算机对数据进行管理；到了 70 年代之后，随着关系数据库的出现和发展，人们管理数据的能力越来越强，收集存储的数据也越来越多。如果只利用数据库进行一些简单的事务处理，显然没有对数据进行充分的利用，从数据中挖掘出有用的知识，才可以更好地实现数据的价值。

[f] 提高泛化能力（generalization ability）是机器学习中最重要的问题之一。泛化能力表征了机器学习系统对新事件的适用性，简单地说，泛化能力越强，系统对新事件的适用能力（例如做出正确预测的能力）就越强。

1989 年 8 月，第 11 届国际人工智能联合会议（IJCAI'89）在美国底特律举行，GTE 实验室的G. Piatetsky-Shapiro在J.G. Carbonell、W. Frawley、K. Parsaye、J.R. Quinlan、M. Siegel、R. Uthurusamy等人的支持下，组织了一个名为“在数据库中发现知识”的研讨会，这个研讨会后来被认为是数据挖掘成为一个领域的标志。早期人们一直称其为“数据挖掘与知识发现”，但随着该领域的发展壮大，越来越多的人直接称其为数据挖掘[g]。值得注意的是，数据挖掘的对象早就不限于数据库，而可以是存放在任何地方的数据，甚至包括Internet上的数据。

数据挖掘受到了很多学科领域的影响，其中数据库、机器学习、统计学无疑影响最大[12]。粗糙地说，数据库提供数据管理技术，机器学习和统计学提供数据分析技术。由于统计学界往往醉心于理论的优美而忽视实际的效用，因此，统计学界提供的很多技术通常都要在机器学习界进一步研究，变成有效的机器学习算法之后才能再进入数据挖掘领域。从这个意义上说，统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术。

从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域。但能否认为数据挖掘只不过就是机器学习的简单应用呢？答案是否定的。一个重要的区别是，传统的机器学习研究并不把海量数据作为处理对象，很多技术是为处理中小规模数据设计的，如果直接把这些技术用于海量数据，效果可能很差，甚至可能用不起来。因此，数据挖掘界必须对这些技术进行专门的、不简单的改造。例如，决策树是一种很好的机器学习技术，不仅有很强的泛化能力，而且学得结果具有一定的可理解性，很适合数据挖掘任务的需求。但传统的决策树算法需要把所有的数据都读到内存中，在面对海量数据时这显然是无法实现的。为了使决策树能够处理海量数据，数据挖掘界做了很多工作，例如通过引入高效的数据结构和数据调度策略等来改造决策树学习过程，而这其实正是在利用数据库界所擅长的数据管理技术。实际上，在传统机器学习算法的研究中，在很多问题上如果能找到多项式时间的算法可能就已经很好了，但在面对海量数据时，可能连$O(n^3)$的算法都是难以接受的，这就给算法的设计带来了巨大的挑战。



另一方面，作为一个独立的学科领域，必然会有一些相对“独特”的东西。对数据挖掘来说，这就是关联分析。简单地说，关联分析就是希望从数据中找出“买尿布的人很可能会买啤酒”这样看起来匪夷所思但可能很有意义的模式[h]。如果在 100 位顾客中有 20 位购买了尿布，购买尿布的 20 位顾客中有 16 位购买了啤酒，那么就可以写成“尿布→啤酒 [支持度=20%，置信度=80%]”这样的

[g] “数据挖掘”这个词其实很久以前就在统计学界出现并略带贬义，但由于数据挖掘领域的发展壮大，这个词目前已经没有贬义了。

[h] “尿布和啤酒”的故事可能是对数据挖掘最好的宣传策划。对“买尿布的人很可能会买啤酒”的一个解释是说，婴儿出世后母亲在家照管孩子，父亲在下班回家的路上买尿布，会顺手捎几瓶啤酒回家。

一条关联规则。挖掘出这样的规则可以有很多用处，例如商家可以考虑把尿布展柜和啤酒展柜放到一起以促进销售。实际上，在面对少量数据时关联分析并不难，可以直接使用统计学中有关相关性的知识，这也正是机器学习界没有研究关联分析的一个重要原因。关联分析的困难其实完全是由海量数据造成的，因为数据量的增加会直接造成挖掘效率的下降，当数据量增加到一定程度，问题的难度就会产生质变，例如，在关联分析中必须考虑因数据太大而无法承受多次扫描数据库的开销、可能产生在存储和计算上都无法接受的大量中间结果等，而关联分析技术正是围绕着“提高效率”这条主线发展起来的。在R. Agrawal等人首先对关联规则挖掘进行研究之后，大批学者投身到这方面的研究中并产生了很多成果，代表性工作有R. Agrawal和R. Srikant的Apriori算法以及J. Han等人的FP-Growth算法等，有兴趣的读者可以参考一些相关书籍[11][13]。

3 坐看云起

机器学习和数据挖掘在过去 10 年经历了飞速发展，目前已经成为子领域众多、内涵非常丰富的学科领域。“更多、更好地解决实际问题”成为机器学习和数据挖掘发展的驱动力。事实上，过去若干年中出现的很多新的研究方向，例如半监督学习、代价敏感学习、流数据挖掘、社会网络分析等，都起源于实际应用中抽象出来的问题，而机器学习和数据挖掘领域的研究进展，也很快就在众多应用领域中发挥作用。值得指出的是，在计算机科学的很多领域中，成功的标志往往是产生了某种看得见、摸得着的系统，而机器学习和数据挖掘则恰恰相反，它们正在逐渐成为基础性、透明化、无处不在的支持技术、服务技术，在它们真正成功的时候，可能人们已经感受不到它们的存在，人们感受到的只是更健壮的防火墙、更灵活的机器人、更安全的自动汽车、更好用的搜索引擎……

由于机器学习和数据挖掘技术的重要性，各国都对这方面的研究非常关注。例如，美国计算机科学研究的重镇——卡内基梅隆大学 2006 年宣布成立“机器学习系”。而美国DARPA从 2003 年开始启动 5 年期的PAL（Perceptive Assistant that Learns）计划[14]，首期 1-1.5 年投资即达 2 千 9 百万美元，总投资超过 1 亿美元。从名字就可以看出，这是一个以机器学习为核心的计划。具体来说，该计划包含两个子计划，一个称为RADAR，由卡内基梅隆大学单独承担，其目标为研制出一种软件，它“通过与其人类主人的交互，并且通过接收明晰的建议和指令来学习”、“将帮助繁忙的管理人员处理耗时的任务”。另一个子计划称为CALO，牵头单位为斯坦福国际研究院，参加单位包括麻省理工学院、斯坦福大学、卡内基梅隆大学、加州大学伯克利分校、华盛顿大学、密歇根大学、德克萨斯大学奥斯汀分校、波音公司等 20 家单位，首期投资即达 2 千 2 百万美元。显然，CALO是整个PAL计划的核心，因为其参加单位不仅包含了美国在计算机科学和人工智能方面具有强大力量的主要高校以及波音公司这样的企业界巨头，其经费还占据了PAL计划整个首期投资的 76%。DARPA没有明确公布CALO的目标，，但从其描述[15]可见端倪：“CALO软件将通过与为其提供指令的用户一起工作来进行学习 …… 它将能够处理常规任务，还能够在突发事件发生时提供协助”，考虑到 911 之后美国对突发事件处理能力的重视，以及波音公司对该计划的参与，该计划的（部分）成果很可能会用于反恐任务。DARPA还说[15]，“CALO的名字源于拉丁文calonis，含义是‘战士的助手’”，而且

DARPA 曾在网站上放置了这样一幅军官与虚拟参谋人员讨论战局的画面，可以预料，该计划的（部分）成果会直接用于军方。从上述情况来看，美国已经把对机器学习的研究上升到国家安全的角度来考虑。

如果要列出目前计算机科学中最活跃的研究分支，那么机器学习和数据挖掘必然位列其中。随着机器学习和数据挖掘技术被应用到越来越多的领域，可以预见，机器学习和数据挖掘不仅将为研究者提供越来越大的研究空间，还将给应用者带来越来越多的回报。

对发展如此迅速的机器学习和数据挖掘领域，要概述其研究进展或发展动向是相当困难的，感兴趣的读者不妨参考近年来机器学习和数据挖掘方面一些重要会议和期刊发表的论文。在机器学习方面，最重要的学术会议是 NIPS、ICML、ECML 和 COLT，最重要的学术期刊是《Machine Learning》和《Journal of Machine Learning Research》；在数据挖掘方面，最重要的学术会议是 SIGKDD、ICDM、SDM、PKDD 和 PAKDD，最重要的学术期刊是《Data Mining and Knowledge Discovery》和《IEEE Transactions on Knowledge and Data Engineering》。此外，人工智能领域的顶级会议如 IJCAI 和 AAAI、数据库领域的顶级会议如 SIGMOD、VLDB、ICDE，以及一些顶级期刊如《Artificial Intelligence》、《Journal of Artificial Intelligence Research》、《IEEE Transactions on Pattern Analysis and Machine Intelligence》、《Neural Computation》等也经常发表机器学习和数据挖掘方面的论文。

参 考 文 献

[1] T. M. Mitchell. *Machine Learning*, New York: McGraw-Hill, 1997.

[2] U. Fayyad, G. Piatetsky-Shapiro, R. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In: *Proc. KDD'96*, Portland, OR, 82-88.

[3] E. Mjolsness, D. DeCoste. Machine learning for science: State of the art and future prospects. *Science*, 2001, 293(5537): 2051-2055.

[4] R. S. Michalski, J. G. Carbonell, T. M. Mitchell, eds. *Machine Learning: An Artificial Intelligence Approach*, Palo Alto, CA: Tioga Publishing Co., 1983.

[5] J. G. Carbonell, ed. *Machine Learning: Paradigms and Methods*, Cambridge, MA: MIT Press, 1990.

[6] P. R. Cohen, E. A. Feigenbaum, eds. *The Handbook of Artificial Intelligence*, vol.3, New York: William Kaufmann, 1983.

[7] S. H. Muggleton, ed. *Inductive Logic Programming*, London: Academic Press, 1992.

[8] D. E. Rumelhart, J. L. McClelland, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press, 1986.

[9] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[10] T. G. Dietterich. Machine learning research: Four current directions. *AI Magazine*, 1997, 18(4): 97-136.

[11] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd edition, Singapore: Elsevier, 2006.

[12] Z.-H. Zhou. Three perspectives of data mining. *Artificial Intelligence*, 2003, 143(1): 139-146.

[13] P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Reading, MA: Addison-Wesley, 2006.

[14] DARPA News Release. DARPA, Jul. 2003.

[15] CALO Overview. DARPA, 2003.

作者介绍：

周志华，南京大学计算机科学与技术系教授，博士生导师，教育部长江学者特聘教授。2000 年于南京大学计算机科学与技术系获博士学位。中国计算机学会人工智能与模式识别专业委员会副主任。主要研究领域为人工智能，机器学习，数据挖掘等。