第5章 机器学习与数据挖掘

5.1 机器学习与数据挖掘综述

5.1.1 机器学习概述

1. 机器学习概念

机器学习(Machine Learning,ML)是计算机模拟或实现人类的学习行为,以获取新的知识或技能,或者重新组织已有的知识结构,使之不断改善自身性能的过程、原理和方法。机器学习也是计算机具有智能的重要标志。

H. A. Simon 认为,学习是系统所做的适应性变化,使得系统在下一次完成同样或类似的任务时更为有效。R. S. Michalski 认为,学习是构造或修改所经历事物的表示。从事专家系统研制的人们则认为学习是知识的获取。这些观点各有侧重,第一种观点强调学习的外部行为效果;第二种则强调学习的内部过程;第三种主要是知识工程的需求。

人类学习有以下几个特点:

(1) 学习是一个缓慢的过程。从上小学到大学毕业通常要花去 16 年时间,要取得博士学位还需要 6 年。而且在实际工作岗位上还要不断地学习。

(2) 人类会"忘记"。人只能记住事物关键的地方,次要的地方会被忘记。对于多次重复的事物,越有特色的事物,记忆越清楚。淡泊的事物总是被忘记。

(3) 人类之间的知识传授很困难。老师讲课要花去很大代价才能教会学生。由于知识传授很困难,也使得人类学习是一个缓慢过程。

(4) 人类能不断地修改知识,使人类逐渐变得聪明。在不断实践过程中,不断地修改知识,使掌握的知识真实地反映事物的规律性。这样,用知识解决实际问题更有效。

机器学习就是让计算机模拟人类的学习,提高获取知识的能力。

2. 机器学习的发展与数据挖掘的兴起

1943 年 McCulloch 与 Pitts 对神经元模型(MP 模型)的研究,第一次揭示了人类神经系统的工作方式。计算机科学与控制理论均从这项研究中受到启发,由于 Pitts 为神经元的工作方式建立了数学模型,正是这个数学模型深刻地影响了机器学习的研究。

机器学习的研究,经历了 5 个发展阶段。

第一阶段始于 20 世纪 50 年代中期,这一阶段的一个重要特点是数值表示和参数调整,代表性工作有 Rosenblatt 在 MP 模型上研究的感知机神经网络;A. L. Samuel 的计算机跳棋

学习程序(曾击败过州级冠军)中采用了判别函数法。

第二阶段始于 20 世纪 60 年代初期,这一阶段主要是概念学习和语言获取,有人称其为符号概念获取阶段。这一时期的代表工作有 E. Hunt 的决策树学习算法 CLS,Winston 的积木世界结构学习系统。另外,在学习计算理论方面,建立了极限辨识理论。

第三阶段始于 20 世纪 70 年代中后期,机器学习逐渐走向兴盛,各学习策略、学习方法相继出现,除了作为主流的归纳学习外,还出现了类比学习、解释学习、观察和发现学习等。这一时期有影响的工作有学习质谱仪预测规则系统 Meta-DENDRAL,利用 AQ11 方法学习大豆疾病诊断规则系统,利用信息论的 ID3 方法,数学概念发现系统 AM,符号积分系统 LEX,及物理化学定律重新发现系统 BACON。在学习计算理论上,L. G. Valiant 提出了概率近似正确 PAC 学习模型,这一成果推动了学习计算理论的发展。

第四阶段始于 20 世纪 80 年代中后期,主要源于神经网络的重新兴起。由于使用隐层神经元的多层神经网络及误差反向传播算法的提出,克服了早期线性感知机的局限性,而使非符号的神经网络的研究得以与符号学习并行发展。同时,机器学习在符号学习的各个方面更加深入和广泛地展开,并形成了较为稳定的几种学习风范,如归纳学习、分析学习(特别是解释学习和类比学习)、遗传学习等。这一时期有影响的工作有多层神经网络反向传播学习算法 BP,基于解释学习,一系列决策树归纳学习方法,J. H. Holland 遗传学习和分类器系统,A. Newell 等的 SOAR 学系统,以及 PRODIGY 学习系统等。

第五阶段即近期,知识发现和数据挖掘的快速发展,它继承和发展了机器学习方法和技术,从数据库中获取知识。机器学习中的归纳学习、神经网络、遗传算法等都引入数据挖掘中。这一时期有影响的工作主要是粗糙集的属性约简和知识获取、关联规则挖掘以及数据仓库的多维数据分析等。

可见,数据挖掘是机器学习发展的新阶段,它也是机器学习和数据库结合的一门新学科方向。近来,深度学习成为人工智能和机器学习的新潮流。

应该指出的是,数据挖掘中获取知识的方法是采用归纳学习的思想。归纳学习带有或然性(合情性),它受限于所使用的数据库。即所获取的知识能够满足数据库中数据的要求,但不一定满足数据库外数据的要求。可以说,这些知识是较正确的知识。数据库中数据越多,获取的知识就越正确。在平时实践中,这些较正确的知识已经够用了。

3. 机器学习实例

历史上最典型的符号机器学习算法应该是 1980 年 Michalski 提出的 AQ11 与 1986 年 Quinlan 提出的 ID3。AQ11 算法是基于集合论的,而 ID3 算法是基于信息论的,从而形成了两个不同的机器学习家族。

机器学习能自动获取知识,它能解决知识获取中的“瓶颈”现象。例如,从大量的实例中自动归纳,产生描述这些实例的一般规则知识。下面给出两个成功的例子。

1) Michalski 和 R. L. Chilausky 的 PLANT/SS 系统

它是一个大豆病害诊断防治专家系统。该系统用示例学习 AQ11 算法自动产生规则进行诊断。把 631 种有病害的大豆的性状描述(表示为包含 35 种特性的向量)和每种植物的专家诊断一起输入计算机中,选用 290 种作为训练例子(例子间相差很远),利用 AQ11 算法

获得规则知识。再用340个样本作为测试例子,并将专家和计算机的诊断结果进行对比。计算机产生的规则优于专家归纳的规则,专家的正确判断率为71.8%,而计算机的正确判断率高达97.6%。

2) 钟鸣和陈文伟的IBLE算法

利用信息论的信道容量思想,研制出IBLE算法。对已有结论的化学物质的质谱进行学习,得出了质谱规则。然后利用这些规则再去测试未知化学物质的质谱,得出它的种类。对苯、有机磷战剂等8类化学物质共1万5千多种进行分类,IBLE的平均正确判断率高达93.97%。它比基于互信息的ID3算法的平均正确判断率高出10个百分点,而化学专家的正确判断率只在70%左右。

5.1.2 机器学习分类

1. 机器学习分类方法

1) 基本分类方法

机器学习主要有归纳学习、分析学习、遗传学习、连接学习等。

归纳学习从具体实例出发,通过归纳推理,得到的概念或知识。归纳学习的基本操作是泛化和特化。泛化是使规则能匹配应用于更多的情形或实例。特化操作则相反,减少规则适用的范围或事例。

归纳学习是目前研究得最广泛的一种符号学习方法,包括实例学习、概念聚类、发现学习等。实例学习的任务是,给定关于某个概念(或多个概念)一系列已知的实例和反例,要求从中归纳出一般的概念描述,该描述能使这些已知实例可从中再次推导出来,而同时没有任何反例可从中推导出来。概念聚类则是由程序根据实例间的相似度关系自动形成有用的概念描述。发现学习主要是从实验数据、观察实例或数据库中获得知识。

分析学习是利用背景或领域知识,分析很少的典型实例(通常仅一个),然后通过演绎推导形成的知识,使得对领域知识的应用更为有效。分析学习方法的目的在于改进系统的效率性能,而同时不牺牲其准确性和通用性,这不同于归纳学习方法。常见的分析学习方法有解释学习、范例学习、类比学习。

2) 按输入信息分类

根据学习系统的输入信息,机器学习方法分为监督学习、非监督学习和强化学习3种。

监督学习又称有教师学习,所谓"教师"即是对一组给定的输入提供应有的输出结果的训练数据集。监督学习已经产生了许多经典的学习算法,如决策树、人工神经网络、贝叶斯网络、支持向量机等。

非监督学习的输入是没有类别标识的训练数据集,因此非监督学习是没有先验知识的学习,仅凭数据的自然聚类的特性,进行"盲目"的学习。最常用的非监督学习是聚类分析。

强化学习把学习看作试探过程,是一种以环境反馈作为输入的学习方法。强化学习过程是不断尝试错误,从环境中得到相应的奖惩,通过自主学习获得不同状态下哪些动作具有最大的价值,从而发现或逼近能够得到最大奖励的策略。

下面简单介绍几种主要的机器学习方法。

2。通过例子学习(实例学习,Learning from Examples)

对某些概念的正例集合与反例集合,通过归纳推理产生覆盖所有正例并排除所有反例的概念描述。这种概念的描述可以是以规则形式表示或用决策树的方法表示。

例如,给出肺炎与肺结核两种病的一些病例。每个病例都含有 5 种症状:发烧(无、低、高)、咳嗽(轻微、中度、剧烈)、X 光所见阴影(点状、索条状、片状、空洞)、血沉(正常、快)和听诊(正常、干鸣音、水泡音)。

肺炎和肺结核的部分病例集如表 5.1 所示。

表 5.1　肺病实例集

病例 \ 症状 / 病例号		发烧	咳嗽	X 光所见阴影	血沉	听诊
肺炎	1	高	剧烈	片状	正常	水泡音
	2	中度	剧烈	片状	正常	水泡音
	3	低	轻微	点状	正常	干鸣音
	4	高	中度	片状	正常	水泡音
	5	中度	轻微	片状	正常	水泡音
肺结核	1	无	轻微	索条状	正常	正常
	2	高	剧烈	空洞	快	干鸣音
	3	低	轻微	索条状	正常	正常
	4	无	轻微	点状	快	干鸣音
	5	低	中度	片状	快	正常

通过示例学习得到如下诊断:

(1) 血沉=正常 ∧(听诊=干鸣音 ∨ 水泡音)→诊断=肺炎

(2) 血沉=快→诊断=肺结核

这样,就从例子(病例)归纳产生了诊断规则。

实例学习系统较多,其中较有影响的有:

(1) J. R. Quinlan 的 ID3 和 C4. 5;

(2) Michalski 的 AQ11;

(3) 钟鸣和陈文伟的 IBLE。

3. 解释学习

解释学习(Explanation-Based Learning,EBL)是利用领域知识和训练例子,构造对目标概念具有可操作性,能进行推理的规则知识,该知识对领域知识的应用更为有效。

解释学习第一步是演绎,第二步是归纳。它用领域知识指导归纳,增加结果的实用性。

下面用一个例子进行说明。

已知

(1) 目标概念:一对物体(X,Y),使 SAFE_TO_STACK(X,Y),有

$$\text{SAFE\_TO\_STACK}(X,Y) \longleftrightarrow \text{NOT}(\text{FRAGILE}) \vee \text{LIGHTER}(X,Y)$$

(2) 训练例子:

$$\begin{aligned}&\mathrm{ON(OBJ1,OBJ2)}\\&\mathrm{ISA(OBJ1,BOX)}\\&\mathrm{ISA(OBJ2,ENDTABLE)}\\&\mathrm{COLOR(OBJ1,RED)}\\&\mathrm{COLOR(OBJ2,BLUE)}\\&\mathrm{VOLUME(OBJ1,1)}\\&\mathrm{DENSITY(OBJ1,1)}\\&\vdots\end{aligned}$$

(3) 领域知识:

$\mathrm{VOLUME}(P_1,V_1)\wedge\mathrm{DENSITY}(P_1,D_1)\rightarrow\mathrm{WEIGHT}(P_1,V_1\times D_1)$

$\mathrm{WEIGHT}(P_1,W_1)\wedge\mathrm{WEIGHT}(P_2,W_2)\wedge\mathrm{LESS}(W_1,W_2)\rightarrow\mathrm{LIGHTER}(P_1,P_2)$

$\mathrm{ISA}(P_2,\mathrm{ENDTABLE})\rightarrow\mathrm{WEIGHT}(P_2,5)$

$\mathrm{LESS}(W_1,5)\wedge\times(V_1,D_1,W_1)\rightarrow\mathrm{LESS}(V_1\times D_1,5)$

$\vdots$

经过解释学习,得到目标概念的实用知识为

$\mathrm{VOLUME}(X,V_1)\wedge\mathrm{DENSITY}(X,D_1)\wedge\mathrm{LESS}(V_1\times D_1,5)\wedge\mathrm{ISA}(Y,\mathrm{ENDTABLE})\rightarrow\mathrm{SAFE\_TO\_STACK}(X,Y)$

该知识的获得是从简单的目标概念开始,利用领域知识展开,再用训练例子实例化,最后用叶结点描述根结点,即用领域知识和训练例子详细地解析目标概念,使该知识更实用。图 5.1 为解释学习推理图。

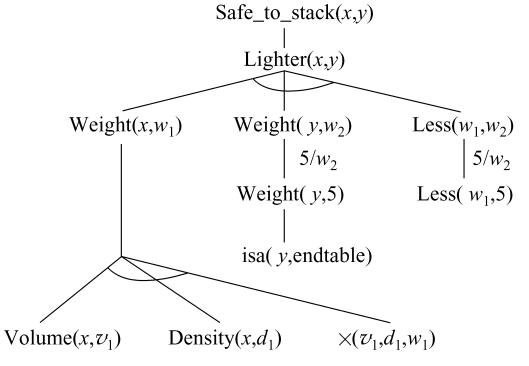

图 5.1 解释学习推理图

基于解释的学习方法,首先由 Mitchell Keller 和 Kadar-Cabell 于 1986 年提出,后来 G. Dejong 和 Mooney 对 Mitchell 的论文进行了全面的讨论、修改和扩充,从而把解释学习推向新高潮。

著名的 EBL 系统有:

(1) T. Mitchell 的 LEX 和 LEAP;

(2) G. Dejong 的 Genesis;

(3) Miton 等的 PRODIGY。

4. 类比学习

有两个不同的领域：源域 S 和目标域 T，S 中的元素 a 和 T 中元素 b 具有相似的性质 P，即 $P(a)\sim P(b)$（$\sim$表示相似），a 还具有性质 Q，即 $Q(a)$。根据类比推理（表示成$|\sim$），b 也具有性质 Q。即

$$P(a)\wedge Q(a), P(a)\sim P(b) \mid\sim Q(b)\sim Q(a)\quad a\in S, b\in T$$

类比学习（Learning by Analogy）在科学技术发展的历史中，起着重要的作用，很多发明和发现是通过类比学习获得的。例如：

（1）卢瑟福将原子结构和太阳系进行类比，发现了原子结构。

（2）水管中的水压计算公式和电路中电压计算公式相似。

类比推理的一般步骤是：

（1）找出源域与目标域的相似性质 P，以及找出源域中另一个性质 Q 和性质 P 对元素 a 的关系：$P(a)\rightarrow Q(a)$。

（2）在源域中推广 P 和 Q 的关系为一般关系，即

$$\forall x(P(x)\rightarrow Q(x))$$

这一步实际是归纳，由个别现象推广成一般规律。

（3）从源域和目标域映射关系，得到目标域的新性质：

$$\forall x(P(x)\rightarrow Q(x))$$

（4）利用假言推理：

$$P(b), P(x)\rightarrow Q(x)\vdash Q(b)$$

最后得出 b 具有性质 Q。

这一步实际是演绎，由一般规律推出个别现象。

类比学习有代表性的工作为：

（1）P. H. Winston 的类比学习和推理系统；

（2）R. G. Reiner 的类比学习系统 NLAG；

（3）J. G. Carbonell 的转换类比学习系统和派生类比学习系统。

5. 发现学习

发现学习（Learning from Discovery）是从大量实验数据中发现规律和定律，即从已知的一组观测结果或数据中求解出能够概括这些数据的一个或多个规律。它主要包括数据驱动法和模型驱动法。

P. Langley 等人的 BACON 系统是数据驱动发现学习系统，该系统重新发现欧姆定律、牛顿万有引力定律和开普勒行星运动定律等物理化学定律。

D. B. Lenat 的 AM 系统是典型的模型驱动发现学习系统，是一个用于模拟初等数学研究的程序，它在大量启发式集合的引导下产生新的概念。它包括各种各样的搜索法（242 个启发式规则）指导在数据领域中的搜索，从集合、表、项等 100 多个基本数学概念出发，使用具体化、一般化、类比、复合等操作去产生新的数学概念，然后把这些操作再应用于所得到的新的数学概念，最终产生出像相乘、自然数、质数等重要的数学概念。这个系统还找到了与这些概念有关的定性规律，如唯一因子分解定理等。

5.1.3 知识发现与数据挖掘综述

1. 知识发现过程

知识发现(Knowledge Discovery in Database,KDD)被认为是从数据中发现有用知识的整个过程。数据挖掘被认为是 KDD 过程中的一个特定步骤,它用专门算法从数据中抽取模式(pattern)。

KDD 过程定义为(Fayyad、Piatetsky-Shapiror 和 Smyth,1996):

KDD 是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的高级处理过程。

其中,数据集:事实 F(数据库元组)的集合;模式:用语言 L 表示的表达式 E,它所描述的数据是集合 F 的一个子集 F_E,它比枚举所有 F_E 中元素更简单,我们称 E 为模式;有效的、新颖的、潜在有用的,以及最终可被人理解:表示发现的模式有一定的可信度,应该是新的,将来有实用价值,能被用户所理解。

KDD 过程图如图 5.2 所示。

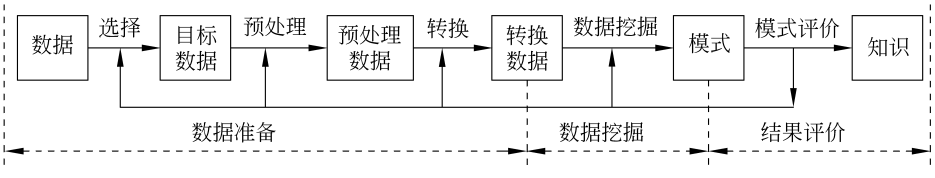

图 5.2 KDD 过程图

KDD 过程可以概括为 3 部分,即数据准备(data preparation)、数据挖掘(data mining)及结果的解释和评价(interpretation & evaluation)。

1) 数据准备

数据准备又可分为 3 个子步骤:数据选择(data selection)、数据预处理(data preprocessing)和数据转换(data transformation)。

数据选择的目的是确定发现任务的操作对象,即目标数据(target data),是根据用户的需要从原始数据库中选取的一组数据。数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录等。数据转换的主要目的是完成数据类型转换(如把连续值数据转换为离散型数据,以便于符号归纳,或是把离散型数据转换为连续值型数据,以便于神经网络计算),尽量消减数据维数或降维(dimension reduction),即从初始属性中找出真正有用的属性以减少数据挖掘时要考虑的属性个数。

2) 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的,如数据分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后,就要决定使用什么样的挖掘算法。选择实现算法有两个考虑因素:一是不同的数据有不同的特点,因此需要用与之相关的算法来挖掘;二是用户或实际运行系统的要求,有的用户可能希望获取描述型的(descriptive)、容易理解的知识(采用规则表示的挖掘方法显然要好于神经网络之类的方法),而有的用户只是希望获取预测准确度尽可能高的预测型(predictive)知识。选择了挖掘算法后,就可以实施数据挖掘操作,

获取有用的模式。

3）结果的解释和评价

数据挖掘阶段发现出来的模式，经过评价，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要回退到发现过程的前面阶段，如重新选取数据、采用新的数据变换方法、设定新的参数值，甚至换一种挖掘算法等。另外，KDD 由于最终是面向人类用户的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的另一种表示，如把分类决策树转换为 if…then…规则。

数据挖掘仅仅是整个过程中的一个步骤。数据挖掘质量的好坏有两个影响要素：一是所采用的数据挖掘技术的有效性；二是用于挖掘的数据的质量和数量（数据量的大小）。如果选择了错误的数据或不适当的属性，或对数据进行了不适当的转换，则挖掘的结果不会好。

整个挖掘过程是一个不断反馈的过程。例如，用户在挖掘途中发现选择的数据不太好，或使用的挖掘技术产生不了期望的结果。这时，用户需要重复先前的过程，甚至从头重新开始。

可视化技术在数据挖掘的各个阶段都扮演着重要的作用。特别是在数据准备阶段，用户可能要使用散点图、直方图等统计、可视化技术来显示有关数据，以期对数据有一个初步的了解，从而为更好地选取数据打下基础。在挖掘阶段，用户则要使用与领域问题有关的可视化工具。在表示结果阶段，则可能要用到可视化技术以使得发现的知识更易于理解。

2. 数据挖掘任务

数据挖掘任务有关联分析、时序模式、聚类、分类、偏差检测和预测 6 项。

1）关联分析

关联分析是从数据库中发现知识的一类重要方法。若两个或多个数据项的取值之间重复出现且概率很高时，它就存在某种关联，可以建立起这些数据项的关联规则。

例如，买面包的顾客有 90％的人还买牛奶，这是一条关联规则。若商店中将面包和牛奶放在一起销售，将会提高它们的销量。

在大型数据库中，这种关联规则是很多的，需要进行筛选，一般用“支持度”和“可信度”两个阈值来淘汰那些无用的关联规则。

“支持度”表示该规则所代表的事例（元组）占全部事例（元组）的百分比，如买面包又买牛奶的顾客占全部顾客的百分比。

“可信度”表示该规则所代表事例占满足前提条件事例的百分比，如买面包又买牛奶的顾客占买面包顾客中的 90％，称可信度为 90％。

2）时序模式

通过时间序列搜索出重复发生概率较高的模式。这里强调时间序列的影响。例如，在所有购买了激光打印机的人中，半年后 80％的人再购买新硒鼓，20％的人用旧硒鼓装碳粉；在所有购买了彩色电视机的人中，有 60％的人再购买 VCD 产品。

在时序模式中，需要找出在某个最小时间内出现比率一直高于某一最小百分比（阈值）的规则。这些规则会随着形式的变化做适当的调整。

时序模式中，一个有重要影响的方法是“相似时序”。用“相似时序”的方法，要按时间顺

序查看时间事件数据库,从中找出另一个或多个相似的时序事件。例如,在零售市场上,找到另一个有相似销售的部门,在股市中找到有相似波动的股票。

3) 聚类

数据库中的数据可以划分为一系列有意义的子集,即类。简单地说,在没有类的数据中,按"距离"概念聚集成若干类。在同一类别中,个体之间的距离较小,而不同类别上的个体之间的距离偏大。聚类增强了人们对客观现实的认识,即通过聚类建立宏观概念,如将鸡、鸭、鹅等都聚类为家禽。

聚类方法包括统计分析方法、机器学习方法和神经网络方法等。

在统计分析方法中,聚类分析是基于距离的聚类,如欧氏距离、海明距离等。这种聚类分析方法是一种基于全局比较的聚类,它需要考察所有的个体才能决定类的划分。

在机器学习方法中,聚类是无导师的学习。在这里距离是根据概念的描述来确定的,故聚类也称概念聚类,当聚类对象动态增加时,概念聚类则称谓概念形成。

在神经网络中,自组织神经网络方法用于聚类,如 ART 模型、Kohonen 模型等,这是一种无监督学习方法。当给定距离阈值后,各样本按阈值进行聚类。

4) 分类

分类是数据挖掘中应用的最多的任务。分类是在聚类的基础上,对已确定的类找出该类别的概念描述,它代表了这类数据的整体信息,即该类的内涵描述。一般用规则或决策树模式表示,该模式能把数据库中的元组影射到给定类别中的某一个。

一个类的内涵描述分为特征描述和辨别性描述。

特征描述是对类中对象的共同特征的描述。辨别性描述是对两个或多个类之间的区别的描述。特征描述允许不同类中具有共同特征,而辨别性描述对不同类不能有相同特征。辨别性描述用的更多。

分类是利用训练样本集(已知数据库元组和类别所组成的样本)通过有关算法而求得。

建立分类决策树的方法,典型的有 ID3、C4.5、IBLE 等方法。建立分类规则的方法,典型的有 AQ 方法、粗集方法、遗传分类器等。

目前,分类方法的研究成果较多,判别方法的好坏,可从 3 个方面进行:

(1) 预测准确度(对非样本数据的判别准确度);

(2) 计算复杂度(方法实现时对时间和空间的复杂度);

(3) 模式的简洁度(在同样效果情况下,希望决策树小或规则少)。

在数据库中,往往存在噪声数据(错误数据)、缺损值、疏密不均匀等问题。它们对分类算法获取的知识将产生坏的影响。

5) 偏差检测

数据库中的数据存在很多异常情况,从数据分析中发现这些异常情况也是很重要的,以引起人们对它更多的注意。

偏差包括很多有用的知识,大体有以下内容:

(1) 分类中的反常实例;

(2) 模式的例外;

(3) 观察结果对模型预测的偏差;

(4) 量值随时间的变化。

偏差检测的基本方法是寻找观察结果与参照之间的差别。观察常常是某一个域的值或多个域值的汇总。参照是给定模型的预测、外界提供的标准或另一个观察。

6）预测

预测是利用历史数据找出变化规律，建立模型，并用此模型来预测未来数据的种类、特征等。

典型的方法是统计机器学习中的回归分析，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。预测时，只要输入任意的时间值，通过回归方程就可求出该时间的预测值。

近年来，发展起来的神经网络方法，如 BP 模型，实现了非线性样本的学习，能进行非线性函数的判别。

分类也能进行预测，但分类一般用于离散数值。回归预测用于连续数值。神经网络方法预测既可用于连续数值，也可以用于离散数值。

5.1.4　数据浓缩与知识表示

1. 数据浓缩

数据浓缩就是在满足某种等价条件下，将复杂的难以理解的数据库，变换成简洁的、容易理解的高度浓缩的数据库。

数据浓缩包括属性约简和元组（记录）压缩两方面。

1）属性约简

属性约简一般用于分类问题。属性约简的原则是保持数据库中分类关系不变。目前，属性约简一般采用粗糙集（rough set）方法，也可以采用信息论方法。

在数据库（S）的分类问题中，属性分为条件属性（C）和决策属性（D）。属性约简是在条件属性中删除那些不影响对决策属性进行分类的多余的属性。经过研究对条件属性一般分为可省略属性和不可省略属性。不可省略属性实质是对决策属性进行分类的核心属性（Core(S)）。而可省略属性（Choice(S)）并不是全部都可省略的属性，需要在可省略属性中挑选出部分属性与核心属性组合成等价原数据库的分类效果。

例如，有如表 5.2 所示的汽车数据库（CTR），有 9 个条件属性，1 个决策属性（里程）。

表 5.2　汽车数据库（CTR）

序号	类型 a	汽缸 b	涡轮式 c	燃料 d	排气量 e	压缩率 f	功率 g	换挡 h	重量 i	里程 D
1	小型	6	Y	1 型	中	高	高	自动	中	中
2	小型	6	N	1 型	中	中	高	手动	中	中
3	小型	6	N	1 型	中	高	高	手动	中	中
4	小型	4	Y	1 型	中	高	高	手动	轻	高
5	小型	6	N	1 型	中	中	中	手动	中	中
6	小型	6	N	2 型	中	中	中	自动	重	低
7	小型	6	N	1 型	中	中	高	手动	重	低
8	微型	4	N	2 型	小	高	低	手动	轻	高