数据库中知识发现的软件设计

张力峰　张荣肖　(中国科学院北京软件工程研制中心 100083)

摘要:本文讨论了数据库中知识发现 KDD(Knowledge Discovery in Databases)的抽象模型和应用,描述了一种 KDD 方法实例。该方法与领域知识密切结合,借鉴了模式识别的有关知识,实现了友好的人—机界面。

关键词:数据库　知识发现　模式识别　人—机界面

一、引言

当前,信息、物质和能量已并列成为人类社会的三大资源,人类已进入信息社会。计算机、通信和网络三大技术是信息社会的基石,信息社会的需求推动了数据库应用技术的发展。

知识是客观世界规律的反映,知识是由信息综合而成的,信息可从数据库中获取。高层决策者希望把自己的数据库作为知识源,从中提取一些中观的或宏观的知识,他们希望数据库具有推理、类比、联想、预测能力,能主动提供服务,甚至能从中得到意想不到的结果。因此,数据库中的知识发现[1][2]KDD(Knowledge Discovery in Databases) 成为当代数据库应用技术的主要方向之一。

KDD 在数据库的基础上,进行从数据库中发现知识的研究,使得数据库不仅能任意查询存放在库中的数据,而且可以得到对数据库中数据的整体特征的认识,获得一些与数据库中数据吻合的中观或宏观的知识,这不仅有利于数据库自身的增长和管理,而且大大提高了数据库的利用率。KDD 方法的宗旨是分析处理数据库中大量的数据,从中发现有用的知识,给用户提供所需问题的答案。

二、KDD 系统的抽象模型

为了进一步介绍 KDD 系统的结构,首先给出一个系统抽象模型[3],这个模型只是实际 KDD 系统的一个抽象,并不是每个系统都是按照这个模块结构组织的。比如有可能根据需要,系统会把某几个部分联合在一起。但是不管如何安排,一个实际的 KDD 系统模型至少都应该包含如下几个部分及功能,如图 1 所示。其中虚线所包含的部分是发现控制相关模块的具体功能。

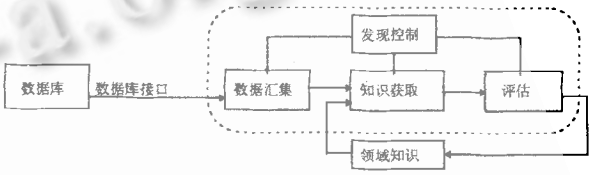

图 1　KDD 系统的抽象模型

1. 数据库和数据库接口

反映现实世界的数据存储在数据库中,用数据库管理系统 DBMS(Database Management System)所提供的查询功能来抽取。一般关系数据库都支持标准数据库查询语言,即结构查询语言 SQL(Structure Query Language),很多 KDD 系统都采用现成的 DBMS 所提供的接口来完成这部分功能。

2. 领域知识

数据库的数据字典对数据库内容涉及的各个域的名称、类型、简单的值约束等进行了描述。另外,关于数据结构、域间约束等信息一般是存放在数据库的说明书、手册以及专家的脑子里,其他关于特定事物或问题的信息则来自终端用户,这些除数据字典以外的信息称为领域知识,在知识发现过程中有着非常重要的作用,它不但能提高系统的发现效率,还可以使发现的知识更加准确。同时,一个理想的 KDD 系统应该能把发现的知识作为领域知识存储起来,以指导或支持以后的发现过程。

3. 发现控制

KDD 系统的知识发现主要来自于控制模块,控制模块是控制器的基本成分。控制器根据领域知识和用户输入信息来控制整个发现过程。如果系统的任务是预先定义好的,而且相对固定,那么控制功能就可以由

一组不变的序列来完成；如果系统的发现任务不是针对某一个问题，而是面向整个数据库，那么控制器就可能比较复杂，一般需要用户参与决策。发现控制模块主要包括数据汇集、知识获取和评估。

(1)数据汇集

数据汇集模块的功能是从大量数据中抽取相关的一部分数据，决定哪些数据记录以及哪部分属性被抽取。用户必须提供详细的数据库结构信息，确定与当前任务相关的属性。如果需要进行抽样，必须采用恰当的随机抽样方法，使所抽取的记录能代表整个数据库。

(2)知识获取

获取知识的算法是KDD系统的核心。所获取的知识一般包括依赖关系、分类识别、抽象描述和异常分析等。

①依赖关系。如果从一个元素A的值能推断出另一个元素B的值，就称为元素B依赖于元素A，也就是说，二者之间存在着依赖关系。一个元素可以是一个域，也可以是域间的一种关系。因此数据依赖关系是一类重要的可被发现的知识。

②分类识别。数据库中，各记录之间的关系对用户而言是没有意义的，然而，KDD能将这些记录按一定的标准划分，分成一系列有一定意义的子类，这些子类的识别对用户可能有直接的意义。

③抽象描述。数据库存放的是一个个记录，但有时人们需要的是数据的整体信息，即量的抽象或内涵的描述，因此从大量数据中归结出抽象级别的信息就尤为重要。

④异常分析。数据库中的数据能反映许多异常的情况，从数据分析中发现这些异常情况也是很重要的，可引起人们对这些特例更多的注意。

(3)评估

数据库中隐含着许多知识，但是用户只对其中的一部分感兴趣，这主要与用户的需求有关。评估模块的功能是给获取的所有知识分别赋予相对数值，用来反映用户感兴趣的程度，然后根据这些值决定提供给用户的知识和排列次序。用户关心的往往是统计意义方面的因素，比如考虑抽象描述是不是能反映整个数据库中隐含的意义。

三、实用的KDD方法

首先，本方法要求与领域知识紧密结合。

任何领域中都有丰富的专门知识，同一领域中，不同的应用需求，涉及的知识内容也不相同，因此，建立一个所有领域都通用的KDD系统是不切实际的。但针对特定领域，建立该领域的KDD系统是可行的。例如，可将领域知识存放在知识文件库中，以供存取，并且随时添加新的领域知识。

图2给出了人口领域中一种实用KDD系统的框图。其中，数据库是我们关心的人口基本信息库，知识库是领域专家提供的领域知识，它以规则的形式存储于知识库中，包括有关的数据字典、常识信息，发现知识过程中要用到的大量属性值之间的概念以及层次关系表等辅助信息，数据基是用户对数据库发出询问时抽取的所有相关数据。从图中可以看出，用户通过界面对数据库发出询问，得到数据基，在一些发现工具的帮助下，利用知识库相关的背景知识，经用户参与，回答用户的询问，同时可以在用户的干预下更新或补充知识库。

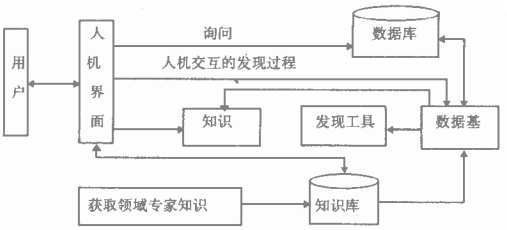

图2　一种实用KDD系统的框架

其次，在获取知识的过程中，可以借鉴模式识别(Pattern Recognition)方法进行。模式识别作为一门技术学科，目的是要研究出能自动进行模式识别和描述的机器系统，以完成人类的模式识别的功能。其系统由预处理、特征或模式基元选择和识别三大部分组成，如图3所示。

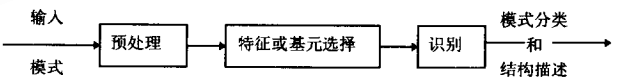

图3　模式识别系统

模式识别中的一个重要理论是统计识别方法，即从被研究的模式中选择能足够代表它的若干特征，每一个模式的特征都组成一个特征向量，于是每一个模式就在特征空间中占有一个位置，可以用各种方法分割特征空间，使得同一类模式大体上都在特征空间的同一区域中。对于待分类的模式，就可根据它的特征向量位于特

征空间中哪一个区域而判定它属于哪一类模式。在人口领域的 KDD 系统中借鉴了模式识别的基本过程。

第一步是对数据进行预处理，得出与用户需求相关的数据。如从某人的生日计算出其人当时的年龄，统计某年龄段的人数等；第二步进行特征选择，由用户根据领域知识或系统缺省参数选定与用户所关心问题密切相关的特征，例如在下文所述的评价人口再生产类型时，其特征是三组特定年龄段的人数。第三步，根据问题数据中的特征值进行识别，得到用户所需结果。例如评价人口再生产类型，只需将问题数据中的三个年龄段与评价标准相比较，利用一定的算法，就能获取所需问题答案。

最后，知识的表达运用了多种形式，不仅有表格和准自然语言，而且包括饼图、棒图等。界面友好，明白易懂。

根据所述的实用的 KDD 方法，数据库中的知识发现过程可以分为以下三个相对独立的阶段，即预处理阶段、知识获取阶段和后处理阶段，如图 4 所示。

预处理部分负责理解用户的发现意图，为进一步的知识发现活动准备好相关的数据。它包括接收并理解用户的发现要求，描述用户的发现任务，利用数据库和系统字典提取出所有有关的数据并加以整理，形成初始知识模板。

知识获取部分是实现从微观数据到宏观知识这一转化过程的核心，包含了多种用于知识发现的工具，负责完成与发现任务相关的一系列抽象归纳和对比分析操作，使初始知识模板中的数据逐步浓缩，直至抽象成为所需要的知识。

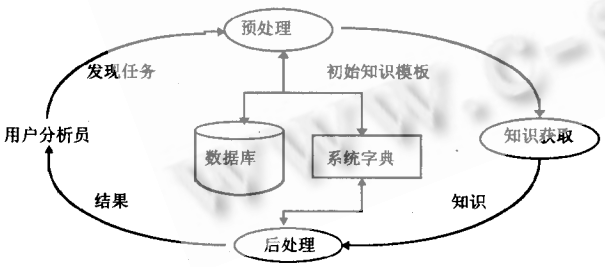

图 4　KDD 系统工作的逻辑过程图

后处理部分负责以多种面向人类的形式输出知识发现的结果，包括数据表格、各种统计图形和准自然语言的发现报告等。

四、应用实例

现具体介绍预测评价人口再生产类型的应用实例。

瑞典人口学家桑德巴氏根据现有人口年龄构成与未来的人口出生率、自然增长率的关系，提出了人口再生产类型评价的国际通用标准，他指出，从人口年龄构成情况可表明人口再生产类型是增加型，还是稳定型或减少型。标准见表 1 所示。

表 1　人口再生产类型国际标准

国际标准	0－14 岁(%)	15－49 岁(%)	50 岁以上(%)
增加型	40	50	10
稳定型	26.5	50.5	23
减少型	20	50	30

为了预测将来某个时刻的人口再生产类型，首先必须进行该时刻的人口年龄构成情况预测，并且提供能按年龄段进行组合的计算方法和衡量人口再生产类型标准的知识文件，在此基础上设计供用户操纵使用系统的界面程序和计算程序。

实现上述具体应用的功能模块关系见图 5。

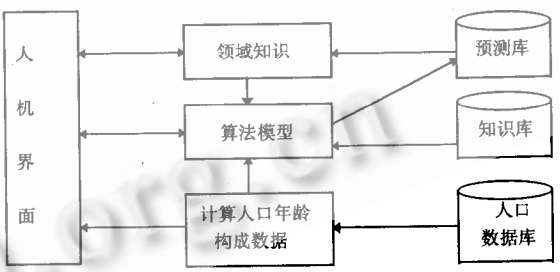

图 5　功能模块关系

进行人口年龄构成情况预测，要求以人口数据库中的当前数据和历史上的数据(如 1990 年的人口普查数据)为基础，预测某年后的人口数据，将预测结果存入预测库中。

预测计算采用直接推算法，如已知 1990 年人口普查的某年龄段的总人数，求出当前的相应的总人数，并以此为基础，计算 X 年(或 X 月)后的相应的总人数。算法模型存放在知识库中。根据人口领域知识，该实例选用了以下三种，用户可随时根据需要添加。各算法模型公式如下：

算法模型 1(线性增量法)：Pt = P0 * (1 + r * t)

算法模型 2(几何增长法)：Pt = P0 * (1 + r)t

算法模型 3(指数增长法)：Pt = P0 * e r * t

其中，P0 为已知时刻 t0 时的人数，t 为从时刻 t0 到预测时刻 t01 时的时间间隔，Pt 为 t01 时刻的人数，r 为该时间间隔 t 内的增长比率。

假设从历史上某一时刻(如 1990 年)到未来的某一时刻，人数增长规律都遵照某一算法模型，则可以先用人口库中的当前数据(如 1995 年数据)和历史数据(如 1990 人口普查数据)计算这段时间间隔内的人数增长比率 r，再以此计算出未来的某一时刻所求人数。

如已知人口库中当前的人数 P01，从历史数据中得到的人数 P02，相差时刻 t12，那么：

对于线性增量法，则 r = (P01 - P02) / (t12 * P02)，从而，Pt = P0 * (1 + r * t)；

对于几何增长法，则 r = exp((ln P01 - ln P02) / t12) - 1，从而，Pt = P0 * (1 + r)t；

对于指数增长法，则 r = (ln P01 - ln P02) / t12，从而，Pt = P0 * e r * t。

评价人口再生产类型，可从预测库中使用计算程序求得年龄在 0 到 14 岁之间、15 到 49 岁之间和 50 岁以上人数百分比，分别记为 x、y 和 z，且有 x + y + z = 1。根据表 1 的判定法则，三种类型的 y 值均在 50% (0.5)附近，选用如下存放于知识库中的判定法则：

当 x-z > = 0.1 时，所求类型为增加型；

当 x-z < 0.1 且 z-x > = 0.1 时，所求类型为减少型；

当 x-z < 0.1 且 z-x < 0.1，即 | x-z | < 0.1 时，所求类型为稳定型。

人口再生产类型预测评价的工作过程如下：

(1) 扫描人口数据库中的当前(1995 年)数据，求出当前的 0 到 14 岁之间、15 到 49 岁之间和 50 岁以上这三个年龄段的人数，记为 P01i(i = 1,2,3)；

(2) 从人口数据库中的历史数据得到 1990 人口普查的 0 到 14 岁之间、15 到 49 岁之间和 50 岁以上这三个年龄段的人数，记为 P02i(i = 1,2,3)；

(3) 由用户在知识库里的算法模型 1，或算法模型 2，或算法模型 3 中进行选择(当然，也可由用户自己提供合适的算法)；

(4) 用户输入需预测的时间间隔 t12；

(5) 系统调用算法进行计算，得到预测结果，记为 Pti(i = 1,2,3)；

(6) 将预测结果数据存入人口预测数据库；

(7) 利用预测数据值，计算表 1 中相应的 x、y、z 值，根据知识库里的人口再生产类型判定法则，得到用户要求时刻的人口再生产类型评价结果，通过人机界面提供给用户。

五、使用 KDD 方法的不利因素

现实数据库存在一些对知识发现不利的情况，主要有以下几点：

·数据库中常常有噪声，即存在一些并不反映事实的记录，从而影响了抽取模式的准确性；

·数据库中的数据不完全，有些记录的属性域存在空值现象，也会影响模式的准确性；

·数据库中信息冗余，导致用户对抽取出来的模式大多不感兴趣；

·数据库中数据稀疏，或者抽取出来的模式不能反映整体情况，或者抽取模式犹如大海捞针，难度很大。

在实际数据库中有时存在一种或几种不利的因素，在设计一个应用系统时，必须面向数据库的实际情况，有的放矢，针对主要存在的不利因素设计系统结构和算法，使得系统能解决实际问题。

随着数据量成指数增长，数据库中的知识发现 KDD 必将和七八十年代的 DBMS 一样影响和支撑着整个信息社会，并广泛地应用于自然科学和社会科学的各个领域，给人类带来不可估量的益处。

参考文献

[1] S. Manivannan, "A Knowledge - Based Fatal Incident Decision Model", IEEE Transactions on Knowledge and Data Engineering Vol. 6, No. 4, Aug. 1994

[2] Beat Wuthrich, "Probabilistic Knowledge Bases", IEEE Transactions on Knowledge and Data Engineering Vol.7, No.5, Oct. 1995

[3] 范建华，"KDD - 基于数据库的知识获取技术"，计算机世界，1995,3,22