数据挖掘及其在商业银行中的应用

上海交通大学管理学院　杨　辉

摘要　本文对数据挖掘的基本原理作了阐述,分析了数据挖掘的流程及主要功能,介绍了数据挖掘工具的算法和特点,并结合具体实例说明了数据挖掘在商业银行的应用。

关键词　数据挖掘　预测　决策　商业银行

Abstract　The basic principle of data mining is discussed in this paper. The process and main functions of data mining are analysed. The algorithms and characteristics of data mining tools are introduced. The examples of data mining in commercial bank are illustrated.

Keywords　data mining, predict, decision making, commercial bank

随着互联网的出现，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率？数据挖掘技术应运而生，越来越显示出其强大的生命力。金融事务需要搜集和处理大量的数据，由于银行在金融领域的地位、工作性质、业务特点以及激烈的市场竞争决定了它对信息化、电子化比其它领域有更迫切的要求。利用数据挖掘技术可以帮助银行产品开发部门描述客户以往的需求趋势，并预测未来。美国商业银行是发达国家商业银行的典范，许多地方值得我国学习和借鉴。

一、数据挖掘的基本原理

1. 数据挖掘概述

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的新技术。它能开采出潜在的模式，找出最有价值的信息，指导商业行为或辅助科学研究。还有很多和数据挖掘(DM)这一术语相近的术语，如从数据库中发现知识(KDD)、数据分析、数据融合(Data Fusion)等。原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。已有的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门广义的交叉学科，它汇聚了不同领域的研究者尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

2. 数据挖掘的流程

(1)数据取样(Sample)。当进行数据挖掘时，首先要从企业大量数据中取出一个与要搜索的问题相关的样板数据子集，而不是动用全部企业数据。通过对数据样本的精选，不仅能减少数据处理量，节省系统资源，而且能通过对数据的筛选，使数据

更加具有规律性。

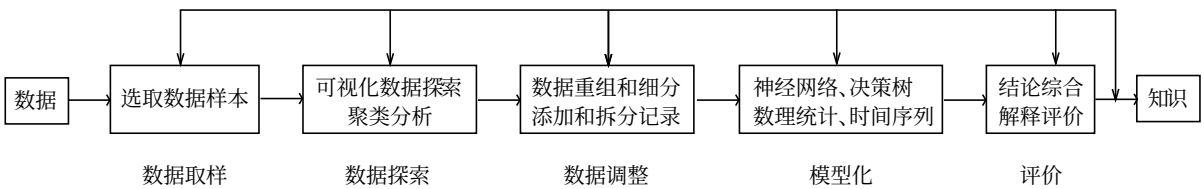
(2) 数据探索(Explore)。数据探索就是通常所进行的对数据深入调查的过程，从样本数据集中找出规律和趋势，用聚类分析区分类别，最终要达到的目的就是搞清楚多因素相互影响的、十分复杂的关系，发现因素之间的相关性。

(3) 数据调整(Modify)。通过上述两个步骤的操作，对数据的状态和趋势有了进一步的了解，这时要尽可能对问题解决的要求能进一步明确化、进一步量化。针对问题的需求要对数据进行增删，按照对整个数据挖掘过程的新认识组合或生成一个新的变量，以体现对状态的有效描述。

(4) 模型化(Model)。在问题进一步明确，数据结构和内容进一步调整的基础上，就可以建立模型。这一步是数据挖掘的核心环节，运用神经网络、决策树、数理统计、时间序列分析等方法来建立模型。

(5) 评价(Assess)。从上述过程中将会得出一系列的分析结果、模式和模型，多数情况会得出对目标问题多侧面的描述，这时就要综合它们的规律性，提供合理的决策支持信息。评价的一种办法是直接使用原先建立模型样本和样本数据来进行检验；另一种办法是另找一批数据并对其进行检验，已知这些数据能反映客观实践的规律性；再一种办法是在实际运行的环境中取出新鲜数据进行检验。

以上叙述的是数据挖掘的基本流程如下图所示。这一过程要反复进行，在反复过程中，不断地趋近事物的本质，不断地优先问题的解决方案。



3. 数据挖掘的功能

数据挖掘通过预测未来趋势及行为，做出前摄的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识，主要有以下五类功能。

(1) 自动预测趋势和行为。数据挖掘自动在大型数据库中寻找预测性信息，以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题，数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户，其它可预测的问题包括预报破产以及认定对指定事件最可能作出反应的群体。

(2) 关联分析。数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有可信度。

(3) 聚类。数据库中的记录可被化分为一系列有意义的子集，即聚类。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类学。80年代初，Michalski 提出了概念聚类技术，其要点是，在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。

(4) 概念描述。概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。生成区别性描述的方法很多，如决策树方法、遗传算法等。

(5) 偏差检测。数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。

二、数据挖掘工具

1. 基于神经网络的工具

由于对非线性数据的快速建模能力,基于神经网络的数据挖掘工具现在越来越流行。其开采过程基本上是将数据聚类,然后分类计算权值。神经网络很适合非线性数据和含噪声数据,所以在市场数据库的分析和建模方面应用广泛。典型的产品有 HNC Software 公司开发的 Marksman。

2. 基于规则和决策树的工具

大部分数据挖掘工具采用规则发现或决策树分类技术来发现数据模式和规则,其核心是某种归纳算法。这类工具通常是对数据库的数据进行开采,生产规则和决策树,然后对新数据进行分析和预测。这类工具的主要优点是,规则和决策树都是可读的。典型产品有 Angoss Software 公司开发的 Knowledge Seeker,广泛应用于市场和金融分析。

3. 基于模糊逻辑的工具

其发现方法是应用模糊逻辑进行数据查询、排序等。该工具使用模糊概念和"最近"搜索技术的数据查询工具,它可以让用户指定目标,然后对数据库进行搜索,找出接近目标的所有记录,并对结果进行评估。典型的产品有 Information Builders Inc. 开发的 Level5 Quest。

4. 综合多方法工具

不少数据挖掘工具采用了多种开采方法,这类工具一般规模较大,适于大型数据库(包括并行数据库)。这类工具开采能力很强,但价格昂贵,并要花很长时间进行学习。典型产品有 IBM 公司开发的 Intelligent Miner。

三、数据挖掘技术在商业银行中的应用

数据挖掘技术在美国银行金融领域应用广泛。金融事务需要搜集和处理大量数据,对这些数据进行分析,发现其数据模式及特征,然后可能发现某个客户、消费群体或组织的金融和商业兴趣,并可观察金融市场的变化趋势。商业银行业务的利润和风险是共存的。为了保证最大的利润和最小的风险,必须对帐户进行科学的分析和归类,并进行信用评估。

Mellon 银行使用 Intelligent Miner 数据挖掘软件提高销售和定价金融产品的精确度,如家庭普通贷款。零售信贷客户主要有两类,一类很少使用信贷限额(低循环者),另一类能够保持较高的未清余额(高循环者)。每一类都代表着销售的挑战。低循环者代表缺省和支出注销费用的危险性较低,但会带来极少的净收入或负收入,因为他们的服务费用几乎与高循环者的相同。银行常常为他们提供项目,鼓励他们更多地使用信贷限额或找到交叉销售高利润产品的机会。高循环者由高和中等危险元件构成。高危险分段具有支付缺省和注销费用的潜力。对于中等危险分段,销售项目的重点是留住可获利的客户并争取能带来相同利润的新客户。但根据新观点,用户的行为会随时间而变化。分析客户整个生命周期的费用和收入就可以看出谁是最具创利潜能的。Mellon 银行认为"根据市场的某一部分进行定制"能够发现最终用户并将市场定位于这些用户。但是,要这么做就必须了解关于最终用户特点的信息。数据挖掘工具为 Mellon 银行提供了获取此类信息的途径。Mellon 银行销售部在先期数据挖掘项目上使用 Intelligent Miner 寻找信息,主要目的是确定现有 Mellon 用户购买特定附加产品:家庭普通信贷限额的倾向,利用该工具可生成用于检测的模型。据银行官员称:Intelligent Miner 可帮助用户增强其商业智能,如交往、分类或回归分析,依赖这些能力,可对那些有较高倾向购买银行产品、服务产品和服务的客户进行有目的的推销。该官员认为,该软件可反馈用于分析和决策的高质量信息,然后将信息输入产品的算法。Intelligent Miner 还有可定制能力。

美国 Firstar 银行使用 Marksman 数据挖掘工具,根据客户的消费模式预测何时为客户提供何种产品。Firstar 银行市场调查和数据库营销部经理 Ted Bratanow 发现:公共数据库中存储着关于每位消费者的大量信息,关键是要透彻分析消费者投入到新产品中的原因,在数据库中找到一种模式,从而能够为每种新产品找到最合适的消费者。Marksman 能读取 800 到 1000 个变量并且给它们赋值,根据消费者是否有家庭财产贷款、赊帐卡、存款证或其它储蓄、投资产品,将它们分成若干组,然后使用数据挖掘工具预测何时向每位消费者提供哪种产品。预测准客户的需要是美国商业银行的竞争优势。