浅说数据挖掘

Herb Edelstein （美国 Two Crows 公司）

一、发掘数据中的隐藏趋势

数据挖掘是数据仓库中最重要的应用之一，它是被称为“知识发现”大处理过程中的一部分。由于描述了保证获得有意义结果所必须采取的步骤，知识发现处理过程是成功地实现数据挖掘的基础。一般来说，数据挖掘能够揭示隐藏的模式和关系。在认识到数据挖掘技术的重要性之后，IBM 公司在世界上率先开发了一整套被称为“智能挖掘机(Intelligent Miner)”的应用工具。

需要说明的是，数据挖掘并非一些人想象中的魔法，——数据挖掘工具不会驻扎在数据库中，监视数据库里发生的一举一动，当发现有趣的趋势时，便自动发出一封电子邮件来提醒用户注意。并非有了数据挖掘工具就不需要用户亲自去了解销售情况，或者说就不需要统计方法了，它所发现的趋势还需要用户的验证。除此之外，数据挖掘所进行的分析需要建立在一些基本假设的基础上，因此，它既不能验证假设，也不能估计或测定其价值。

数据处理专业人士提出的最普遍问题之一是：数据挖掘和 OLAP 的区别是什么？对这个问题的答案是：OLAP 是由用户驱动的，分析家设定一些假设，并用 OLAP 工具去验证这些假设；相反地，数据挖掘应用在数据上，并产生一个假设。类似地，当使用者使用 OLAP 和一些其他的查询工具来发掘数据时，是用户来指导发掘过程；然而，当用户使用数据挖掘工具发掘数据时，是挖掘工具来进行开发。

例如，分析家假设那些高负债和低收入的人有信用风险，他们可以用 OLAP 以各种方式验证或反验证这个假设；然而，数据挖掘工具可以用来发现给予信用的风险因素，比如，可能会发现具有高债务和低收入的人有信用风险，它可能还会发现一种分析家们难以置信的模式，如负债/收入比和年龄所预示的风险。——这就是数据挖掘和 OLAP 互补的地方：在使用一种发现模式操作之前，分析家需要知识通过这种模式来控制谁将获得信用的财政含义；此外，通过帮助用户更好地理解数据(如集中注意那些重要的变量、识别异常情况或发现相互作用等)，OLAP 能够加强知识发现过程的早期阶段。这些操作是相当重要的，用户对自己的数据理解的越充分，知识发现过程也就越有效。由于使用这种引导用户的 OLAP 方式及数据发掘可以补充数据挖掘，IBM 已经集成了智能决策服务软件和智能数据挖掘软件，以期利用 IDS 图形开发环境来调用数据挖掘操作并分析操作结果。

二、数据挖掘的独特之处

数据挖掘使用数据建立现实世界的模型，建模的结果是一种数据中的模式和关系的描述。我们可以用两种基本方式利用这些模型：第一种方式描述数据库中的模式和关系，可以提供指导经营活动的知识，例如，食品杂货店的关联模型对上货计划是非常有用的；第二种方式，模式可以用来进行预测，例如，用户能够在邮寄名单数据库中发现一种模式，它将帮助使用者预测哪些人最可能对一种邀请作出反应，因此，用户不必再向数据库名单中的所有人发信，而只需要对模式所识别出的子集中的人发信，这样可以通过减少信函数量节约大量的邮寄费用。

数据挖掘可以用 6 种模型解决商业问题：分类、回归、时间序列、分类归并、相关分析和顺序发现模型，分类和回归模型主要用于预测，相关分析和顺序发现模型主要用于描述或说明从用户数据库中捕获的行为，分类归并模型主要用于预测或说明。

分类模型通过计算范畴变量值将实例进行分组或分类(范畴数据与一小组离散目录如“可能回信”或“不可能回信”相适应)，分类模型被广泛用于解决诸如上述的邮寄名单等问题。这时，分类模型检验一个实例收集，这个实例收集的归属是已知的。分类模型应用这些数据来确定属性的模式，该模式识别这些实例的属组，——这个模式可以用来理解已有数据及预测新的实例如何进行分类。

回归模型使用已有值和它们的属性来预报后续值(后续数据可具有实数区间中的任意值)，与此类似，时序预测模型使用一系列现存值和它们的属性预测将来值，区别于时序分析的是，这些值依赖于时间。工具可以用来发掘时间的特性，特别是时期的分层结构(包括 5 天或 7 天工作周、包含 13 个月的年等等)，季节、日历的影响(如假期)、日期的算法和某些特殊的考虑(如多少数据是和将来相关的)等。

分类归并模型将一个数据库分割成许多组，目的是找出相互区别的组及相似的组。和分类模型不同，分类归并开始时，用户并不知道是什么样子的，或依据什么属性对数据进行分类归并，因此，商业分析家必须对这些分类归并进

行解释。

相关分析模型对在一个事件或记录中同时发生的事项进行分析，相关分析工具用来发现形式的规则：如果事项 A 是一个事件的组成部分，而 x% 的时间(置信度)中事项 B 是事件的一部分，例如：如果购买低脂肪的农家乳酪和脱脂酸奶，85% 的时间也购买脱脂牛奶。

顺序发现模型和相关分析模型紧密相关，除非相关事项跨越时间界限。为发现这些顺序，必须捕获每一笔交易及交易者识别的细节。例如：如果进行手术过程 X，那么 45% 的时间感染 Y 将会发生；或者，如果在特定的一天中股票 A 上涨超过 12%，并且 NASDAQ 指数下滑，那么 68% 的时间股票 B 将在两天之内上涨。

IBM 的智能挖掘机是能够建立所有这些模型的少数几个数据挖掘工具之一，建立相连接的模型来发现这些不同的模式，将大大增加数据挖掘的效率。

三、数据挖掘工具和技术

需要记住的最重要的事情在于：没有一个或一套工具是万能的。对任何给定的问题，数据的性质将影响所选择的工具，因此，需要使用各种不同的工具和技术来发现最好的模型。分类模型就是这种模型之一，所以我们将在这里说明建立这些模型的最流行方法。

分类模型至少包含两个数据仓库统计技术——对数回归(线性回归的一般化处理)和判别式分析中的一个，然而，当数据挖掘变得越来越普遍时，神经网络和决策树也受到越来越多的重视。尽管这些方法很复杂，但它们不要求用户具有熟练的统计学知识。

神经网络使用许多参数(隐藏层中的节点)来建立模型，此模型采用并结合一套输入来预测一个后续或范畴变量，每一个隐藏节点的值都是所有这些前述节点值的加权和函数，例如节点 4 的值是：

节点 4 的值 = f(权$_{1-4}$ × 节点 1 的值 × 权$_{2-4}$节点 2 的值)

建立模型的过程包括找出连接权的值，通过使用数据“训练”神经网络，可以产生最精确的结果。最常用的训练方法是逆向传播，使输出结果和已知的正确值进行比较，每次比较之后，对权数进行调整并计算新的结果。经过足够的训练之后，神经网络变成一个非常好的预测机。然而，神经网络存在两个问题，首先，反对神经网络的观点之一是它的不透明性，从而导致预测值的因素不明显；其次，神经网络容易过适应，它们在预测实验值时非常完美，但这是以牺牲新数据的精确性为代价的。现在，在仔细操作的情况下，我们已经有了几项技术可以用来避免这个问题。

与神经网络不同，决策树代表一系列的规则以连接类或值，例如，用户可能希望对贷款申请进行信用风险高低的分类，一个简单的决策树就可以解决这个问题。使用决策树和贷款申请，贷款官员可以确定一个申请的信用风险高低，“年收入大于 4 万美元”和“高债务”的人可被列入“高风险”类，而“年收入小于 4 万美元”和“工作时间大于 5 年”的人可被列入“低风险”类。

由于决策树的精确性可以接受，而且不象神经网络那样难以理解，现在这项技术已非常流行。建立决策树花费的时间比神经网络少，然而，标准的决策树运算法则确也有其缺点：它不能发现基于变量合并的规则，分支间的拆分不够平滑，进行拆分时不考虑其对将来拆分的影响。神经网络和决策树可以用来进行回归，一些类型的神经网络甚至可用来进行分类归并，IBM 的智能挖掘机就提供了神经网络和决策树算法规则。

四、数据挖掘的应用

数据挖掘是如此重要的策略应用，以至于许多公司不会透露他们的计划。根据美国 Two Grows 公司所做的一项调查表明，数据挖掘的 3 个最重要终途应用是在销售领域：客户概况、目标市场和购买方式分析。

在客户概况中，通过预测谁将会帮助商家发现新的前景来确定好顾客的特点。数据挖掘可以在客户数据库中发现一种模式，将它应用到一个期望的数据库中，就能够使获得客户的目标得以实现。例如，通过鉴别可以提供邮寄和目录的候选人，邮购商可以降低成本和增加销售，针对已有客户和潜在客户的促销活动可以获得相应的效益。

购买方式分析可以帮助零售商了解顾客会同时购买哪些商品，使用数据挖掘，零售商能够确定哪些商品应放在哪些商店，甚至在商店中如何摆放这些商品，数据挖掘还可以用来评估促销和优惠券的效果。

数据挖掘在许多机构中的另一通常用途是帮助管理顾客关系，通过分析确定那些可能离开并走向竞争对手的顾客特点，公司可以采取行动留住这些顾客，因为这样做比招徕新顾客的花费要小的多。

企业由于诈骗而遭受的损失是非常巨大的，因此，通信公司、信用卡公司、保险公司、股票交易所、政府机构等对诈骗侦测有极高的兴趣和热情，使用数据挖掘技术，这些机构可以识别潜在的诈骗交易，进而控制可能发生的伤害。金融公司可以使用数据挖掘技术来确定市场和工业特征，并预测某一公司和股票的运作情况。数据挖掘技术另一个有趣的应用是医药行业：它能够帮助预测手术过程的效率、诊断测试、药物治疗、服务管理和过程控制。

(来稿时间：1998 年 1 月)