

深入浅出谈数据挖掘

段 勇

编者的话：本文对数据挖掘概念的产生，数据挖掘与常规数据分析的主要区别，所能解决的几大类问题和所应用的领域都有着非常清晰的论述。作者在此篇文章中认为数据挖掘最重要的要素是分析人员的相关业务知识和思维模式。丰富的业务知识是设计有效的相关变量的必要条件，而分析人员的思维模式从另外一个方面也保障了设计变量的结构化和完整性。所以我们在掌握丰富的业务知识同时，如果能够按照正确的思维模式去思考问题，将会发现解决问题并不是很困难的。

一、 数据挖掘的本质

一般来说，比较狭义的观点认为数据挖掘区别于常规数据分析的关键点在于：数据挖掘主要侧重解决四类问题：分类、聚类、关联、预测（关于这四类问题后文会详细阐述），而常规数据分析则侧重于解决除此之外的其他数据分析问题：如描述性统计、交叉报表、假设检验等。

让我们来看一个例子：某移动运营商想了解目前彩铃业务的发展现状如何？解决这个问题的方法就是常规的数据分析，通过描述性统计和交叉报表，可以知道目前彩铃业务的用户数、普及率、收入情况？不同品牌用户间的情况和差异？不同消费水平用户间的情况和差异……。这样的分析主要解决了企业过去发生了什么以及存在什么问题；如果该运营商希望建立一个模型（或者规则），从没有使用彩铃的用户群中找出一部分用户作为彩铃营销活动的目标用户，如通过短信或者外呼的方式告知用户可以免费试用彩铃一个月。解决这个问题则需要使用数据挖掘的方法，如通过决策树方法可以找出使用彩铃业务可能性较高的用户的一系列特征规则，然后根据这些规则去筛选目标用户。当然数据挖掘也并不是解决这个问题唯一办法，因为在没有数据挖掘这个概念之前（1990年以前），这样的问题在商业中也是普遍存在的。通过常规的数据分析依然能解决这个问题，例如研究不同品牌、不同消费水平、不同年龄、不同……的用户使用彩铃的情况，也可以总结出一套比较实用的规则来作为筛选彩铃目标用户的规则。当然，这样的方法跟数据挖掘方法相比存在一定的不足，由于篇幅的限制，这个问题留给大家

去思考。

个人的观点：数据挖掘很大程度上来说更像一个框架概念。它所使用的各种方法在这个概念形成之前已经普遍存在，例如统计学中的多元回归、Logistic 回归，人工智能中的神经网络等。在上个世纪 90 年代，由于数据库的高速发展，企业对精确化营销的迫切需求，导致了数据挖掘这个概念和新名词的诞生。当然我们也不能简单的认为数据挖掘就是一个“新瓶装老酒”，毕竟，数据挖掘根据所解决的不同类型的问题，把包含统计学在内的各种方法进行了整合和重新设计，形成了一套新的数据分析方法论和框架，在这个框架内，源源不断的很多人投入进来，这其中主要包含两类人：一类人是在更新设计新的算法；一类人是在不断的探索既有的方法在商业中的各种应用。

二、 数据挖掘主要解决的四类问题

数据挖掘非常清晰的界定了它所能解决的几类问题。这是一个高度的归纳，数据挖掘的应用就是把这几类问题演绎的一个过程。下面让我们来看看它所解决的四类问题是如何界定的：

■ 分类问题

分类问题属于预测性的问题，但是它跟普通预测问题的区别在于其预测的结果是类别（如 A、B、C 三类）而不是一个具体的数值（如 55、65、75……）。

举个例子，你和朋友在路上走着，迎面走来一个人，你对朋友说：我猜这个人是个上海人，那么这个问题就属于分类问题；如果你对朋友说：我猜这个人的年龄在 30 岁左右，那么这个问题就属于后面要说到的预测问题。

商业案例中，分类问题可谓是最多的：给你一个客户的相关信息，预测一下他未来一段时间是否会离网？信用度是好/一般/差？是否会使用你的某个产品？将来会成为你的高/中/低价值的客户？是否会响应你的某个促销活动？……。

有一种很特殊的分类问题，那就是“二分”问题，显而易见，“二分”问题意味着预测的分类结果只有两个类：如是/否；好/坏；高/低……。这类问题也称为 0/1 问题。之所以说它很特殊，主要是因为解决这类问题时，我们只需关注预测属于其中一类的概率即可，因为两个类的概率可以互相推导。如预测 $X=1$ 的

概率为 $P(X=1)$ ，那么 $X=0$ 的概率 $P(X=0) = 1 - P(X=1)$ 。这一点是非常重要的。

可能很多人已经在关心数据挖掘方法是怎么预测 $P(X=1)$ 这个问题的了，其实并不难。解决这类问题的一个大前提就是通过历史数据的收集，已经明确知道了某些用户的分类结果，如已经收集到了 10000 个用户的分类结果，其中 7000 个是属于“1”这类；3000 个属于“0”这类。伴随着收集到分类结果的同时，还收集了这 10000 个用户的若干特征（指标、变量）。这样的数据集一般在数据挖掘中被称为训练集，顾名思义，分类预测的规则就是通过这个数据集训练出来的。训练的大概思路是这样的：对所有已经收集到的特征/变量分别进行分析，寻找与目标 0/1 变量相关的特征/变量，然后归纳出 $P(X=1)$ 与筛选出来的相关特征/变量之间的关系（不同方法归纳出来的关系的表达方式是各不相同的，如回归的方法是通过函数关系式，决策树方法是通过规则集）。

如需了解细节，请查阅：决策树、Logistic 回归、判别分析、神经网络、Inpurity、Entropy、Chi-square、Gini、Odds、Odds Ratio……等相关知识。

■ 聚类问题

聚类问题不属于预测性的问题，它主要解决的是把一群对象划分成若干个组的问题。划分的依据是聚类问题的核心。所谓“物以类聚，人以群分”，故得名聚类。

聚类问题容易与分类问题混淆，主要是语言表达的原因，因为我们常说这样的话：“根据客户的消费行为，我们把客户分成三个类，第一个类的主要特征是……”，实际上这是一个聚类问题，但是在表达上容易让我们误解为这是个分类问题。分类问题与聚类问题是有本质区别的：分类问题是预测一个未知类别的用户属于哪个类别（相当于做单选题），而聚类问题是根据选定的指标，对一群用户进行划分（相当于做开放式的论述题），它不属于预测问题。

聚类问题在商业案例中也是一个非常常见的，例如需要选择若干个指标（如价值、成本、使用的产品等）对已有的用户群进行划分：特征相似的用户聚为一类，特征不同的用户分属于不同的类。

聚类的方法层出不穷，基于用户间彼此距离的长短来对用户进行聚类划分的方法依然是当前最流行的方法。大致的思路是这样的：首先确定选择哪些指标对用户进行聚类；然后在选择的指标上计算用户彼此间的距离，距离的计算公式很多，最常用的就是直线距离（把选择的指标当作维度、用户在每个指标下都有相应的取值，可以看作多维空间中的一个点，用户彼此间的距离就可理解为两者之间的直线距离。）；最后聚类方法把彼此距离比较短的用户聚为一类，类与类之间的距离相对比较长。

如需了解细节，请查阅：聚类分析、系统聚类、K-means 聚类、欧氏距离、闵氏距离、马氏距离等知识。

■ 关联问题

说起关联问题，可能要从“啤酒和尿布”说起了。有人说啤酒和尿布是沃尔玛超市的一个经典案例，也有人说，是为了宣传数据挖掘/数据仓库而编造出来的虚构的“托”。不管如何，“啤酒和尿布”给了我们一个启示：世界上的万事万物都有着千丝万缕的联系，我们要善于发现这种关联。

关联分析要解决的主要问题是：一群用户购买了很多产品之后，哪些产品同时购买的几率比较高？买了 A 产品的同时买哪个产品的几率比较高？可能是由于最初关联分析主要是在超市应用比较广泛，所以又叫“购物篮分析”，英文简称为 MBA，当然此 MBA 非彼 MBA，意为 Market Basket Analysis。

如果在研究的问题中，一个用户购买的所有产品假定是同时一次性购买的，分析的重点就是所有用户购买的产品之间关联性；如果假定一个用户购买的产品的时间是不同的，而且分析时需要突出时间先后上的关联，如先买了什么，然后后买什么？那么这类问题称之为序列问题，它是关联问题的一种特殊情况。从某种意义上来说，序列问题也可以按照关联问题来操作。

关联分析有三个非常重要的概念，那就是“三度”：支持度、可信度、提升度。假设有 10000 个人购买了产品，其中购买 A 产品的人是 1000 个，购买 B 产品的人是 2000 个，AB 同时购买的人是 800 个。支持度指的是关联的产品（假定 A 产品和 B 产品关联）同时购买的人数占总人数的比例，即 $800/10000=8\%$ ，

有 8% 的用户同时购买了 A 和 B 两个产品；可信度指的是在购买了一个产品之后购买另外一个产品的可能性，例如购买了 A 产品之后购买 B 产品的可信度 $=800/1000=80\%$ ，即 80% 的用户在购买了 A 产品之后会购买 B 产品；提升度就是在购买 A 产品这个条件下购买 B 产品的可能性与没有这个条件下购买 B 产品的可能性之比，没有任何条件下购买 B 产品可能性 $=2000/10000=20\%$ ，那么提升度 $=80\%/20\%=4$ 。

如需了解细节，请查阅：关联规则、aprior 算法中等相关知识。

■ 预测问题

此处说的预测问题指的是狭义的预测，并不包含前面阐述的分类问题，因为分类问题也属于预测。一般来说我们谈预测问题主要指预测变量的取值为连续数值型的情况。

例如天气预报预测明天的气温、国家预测下一年度的 GDP 增长率、电信运营商预测下一年的收入、用户数等？

预测问题的解决更多的是采用统计学的技术，例如回归分析和时间序列分析。回归分析是一种非常古典而且影响深远的统计方法，最早是由达尔文的表弟高尔顿在研究生物统计中提出来的方法，它的主要目的是研究目标变量与影响它的若干相关变量之间的关系，通过拟和类似 $Y=aX_1+bX_2+\dots$ 的关系式来揭示变量之间的关系。通过这个关系式，在给定一组 $X_1、X_2\dots$ 的取值之后就可以预测未知的 Y 值。

相对来说，用于预测问题的回归分析在商业中的应用要远远少于在医学、心理学、自然科学中的应用。最主要的原因是后者是更偏向于自然科学的理论研究，需要有理论支持的实证分析，而在商业统计分析中，更多的使用描述性统计和报表去揭示过去发生了什么，或者是应用性更强的分类、聚类问题。

如需了解细节，请查阅：一元线性回归分析、多元线性回归分析、最小二乘法等相关知识。

三、 数据挖掘的应用领域

数据挖掘一开始就是面向应用而诞生的，前面说到数据挖掘主要解决四大类的问题，如果把这些问题演绎到不同的行业，我们将看到数据挖掘的应用是非常广泛的。

以我们经常接触的移动通信行业来说，结合前面提到的四大类问题，我们看看数据挖掘在通信行业都有哪些应用。

分类问题：

- 离网预测：预测用户在未来一段时间内离网的风险。
- 信用申请评分：根据用户资料评估用户是否可以授信（如预付费用户可以透支、后付费用户可以延长帐期）。
- 信用行为评分：根据用户过去的消费行为特征评估信用得分高低，便于调整话费透支额度或者付费帐期。
- 定位产品（如彩铃、WAP、增值数据业务等）目标用户：构建模型筛选产品营销的目标用户群。

聚类问题：

- 用户细分：选择若干指标把用户群聚为若干个组，组内特征相似、组间特征差异明显。当然用户细分的方法很多，不一定是采用聚类方法。聚类的优点是可以综合处理多维变量，缺点是随之带来的不易解释性。一种便于解释的细分方法是结合业务对用户群进行人为的划分，习惯上称为 Pre-Define 的方法。这种方法的优点是便于解释且应用性强，缺点是对业务要求比较高，划分边界比较难定，对多维变量处理有难度。

关联问题：

- 交叉销售：针对用户已经使用的产品和业务，向其推荐他没有使用的，但可能有兴趣的产品。交叉销售的问题从某种角度上来也可以理解为分类问题，与定位产品目标用户这个问题比较相似。

预测问题：

比较成型的应用不多，一般多为用户数预测、收入预测等。

四、 什么是数据挖掘最重要的要素？

回到文章一开始举的那个案例来说，如果某运营商需要建立一个模型来筛选一部分目前还没有用彩铃的用户作为推广彩铃业务的目标用户，那么这样一个任务要取得成功的关键要素是什么呢？是分析人员的思维模式、分析采用的方法、相关业务知识还是分析采用的工具？

从技术的角度来看这个问题，能不能得出精准的答案主要取决于是否寻找到与目标（是否使用彩铃）相关的变量。而影响变量选择的关键并不是选择了不同分析方法，而是是否提供了足够和有效的变量的去供分析方法选择。也就是说不同的分析方法选择相关变量的能力是相差不大的，关键是是否提供了足够的变量供选择。

变量的提供取决于变量的收集和设计，影响它最关键的两个因素是：相关业务知识和分析人员的思维模式。丰富的业务知识是设计有效的相关变量的必要条件。分析人员的思维模式从另外一个方面保障了设计变量的结构化和完整性。麦肯锡公司一个重要的思维模式就是 MECE，即不重叠、不遗漏。这是一个非常要命的观点，如果都能按照这个模式去思考问题，你会发现解决问题原来也并不是那么困难。

分析人员的业务知识和思维模式不仅仅简单的影响着变量的设计，还包括整个数据挖掘任务的方案框架设计以及后续的结果应用，在这里以终为始的思维模式又显得尤为重要。

纵观其他要素，分析方法对结果的影响主要体现在结果的解释性和稳定性上：例如在信用评分应用中，Logistic 回归的结果就更便于解释和应用；而决策树方法对极值、非线性关系的处理就比其他方法更稳健。

此外，分析工具对结果的影响较小，但是在功能、操作的便利性和效率方面差别也是相当大的。SAS 软件相比 SPSS、SPLUS 等软件来说在效率和功能方面有较大的优势。