数据挖掘到底能挖到什么

文 / 章贵军 谢邦昌

有人说我们生活的时代是信息爆炸的时代，我们生活周围充斥着各种各样的信息，这些信息丰富了我们的生活，使我们生活变得便捷。然而，也有人认为，这些信息来源渠道五花八门，有图像的，有音频的，有文字的，有数字的，让我们有些应接不暇。因此，对后者来说，这些信息打乱了他们的生活节奏，令他们无所适从；对前者来说，这些信息让他们的生活更加丰富多彩，令他们无所不能。

在大统计时代，数据处理内容不仅仅局限于数字，数据处理的对象包罗万象，包含自然社会提供给我们的一切信息，不仅包括我们能直接感觉到的信息，譬如上面提到的图像、音频、文字和数字；还应包括我们感觉不到的信息，如天文、地理、微生物、反物质……显然，对于身处信息数据海洋中的我们而言，如果我们不能很好地整理、分析这些数据信息，我们必将迷失前进的方向。换句话说，数据挖掘技术就是告知我们如何整理、分析海量数据信息的工具，就是数据大海中的指南针，掌握了它，我们便知道了前进方向。

可能有人会觉得这种比方太抽象，仍然会问数据挖掘到底能挖到什么？

数据挖掘能挖到什么？从消费者的角度而言，如果我们很好地掌握了数据挖掘技术，那么这门技术会让我们受益无穷；它会帮助我们甄别商品真伪，区分我们交往人群的善恶，分析购买的商品是否物有所值，预测理财产品是否收益丰厚。从生产者角度而言，如果我们很好地掌握了数据挖掘技术，那么这门技术会让我们无往不利，它会帮助我们判断生产过程是否正常，告诉我们仓储货运是否合理，产品销售是否顺畅，市场的需求倾向和产品改进的方向是否一致。可能，我们在生活中在潜意识的状态下或多或少地运用了数据挖掘技术，只是运用的方法、目的还没有系统化和明确化。

就目前而言，数据挖掘技术主要有六种模式：分类、回归、时间序列、聚类、关联及序列发现。各种技术模式既可以有针对性的应用在各个专业领域，也可以综合应用解决某一类问题，为了更清楚地展现数据挖掘技术的强大生命力，本文接下来的部分将具体阐述这些技术模式和应用实例。

1. 分类。数据分类是根据一组数据对象的特征给出数据对象数学划分的过程，是数据挖掘的重要内容之一。简而言之，数据挖掘技术的分类过程就是根据一堆样品，然后通过样品的指标找到一个分类规则，在这个规则的基础上对新的样品进行归类。目前，数据分类的方法主要有决策树方法、逻辑斯回归法、类神经网络法、卡方自动影响探测器（Chi-Square Automatic Interaction Detector）法等。从消费者角度看，分类技术应用的一个具体例子是，假设某个消费者想网购一款手机，买到价廉物美的商品是很多理性消费者的目的，懂得数据挖掘技术的消费者会首先搜集网上关于手机的相关信息，比方说使用寿命、价格、售后服务、通话质量、图像质量、待机时间和购买者评价等一系列数据，然后可能会对手机性价比做一个基本分类，譬如按性价比分为低中高三类。从厂商角度看，分类技术应用的具体例子是，根据某大型国有银行提供的数以千万计的使用信用卡客户的基本信息如（户籍所在地、性别、年龄、学历、职业、婚姻等），经济状况（家庭月收入、个人月收入等），信用卡消费状况（信用卡张数、使用频率、月刷卡金额等）。根据这些基本信息，应用数据挖掘的分类方法就可以对使用信用卡的消费者进行分类。对于商业银行而言，可以以大概率事件轻松找到潜在的优质客户和违约客户。在此基础上，商业银行便可以制定相应的服务措施，一方面可以为潜在的优质客户提供针对性地服务，另一方面通过对潜在的违约客户采取防范未然的服务措施，从而避免或降低违约风险。

2. 回归。回归是使用一系列的现有数值来分析自变量对因变量的影响并预测因变量将来出现的可能值。目前，回归分析与分类方法在很多方面已不分彼此，常用的二元分类方法如决策树方法、逻辑斯回归法、类神经网络法也可用来进行预测类别变量。数据挖掘技术案例中应用到回归分析方法的例子不胜枚举。现举一某品牌服装厂商成功应用数据挖掘的例子。该品牌厂商通过对大量数据研究发现，在服装质量，服装价格，款式相同的情况下，新款产品推出的时间与服装销售量有很大关系，于是该厂商改进生产工艺，缩短从设计到生产的时间，这种改进为该厂商赢得了大量市场

份额。

3. 聚类。聚类规则是识别一组数据对象的内在规则，从而将对象分组，构成相似对象类，以导出数据的分布规律，也就是力图去发现隐含在一组混杂的数据对象的分类规则。一个常见的具体的例子是保险公司通过分析大量投保人投保前行为和投保后行为以及投保人关于性别、年龄、学历、职业、婚姻、收入、身体状况、家庭财产、犯罪记录等方面的数据信息后会得出一个分类规则，然后根据该规则对准投保人进行判别从而有效侦测保险欺诈行为。

4. 时间序列预测。时间序列预测类似于回归预测，与回归预测不同的是，时间序列预测是用现有自变量预测将来自变量可能出现的数值。我们日常中碰到的很多数据资料都是在连续时间段观察到的有顺序的数据的集合，如证券市场交易数据、失业率、工厂生产流水线每日的产能及某种产品每月的销售量等。在证券市场中，我们用过去时间某只股市价格作为自变量来对该股票未来价格进行预测，这便是时间序列预测。

5. 关联规则。目前，很多学者认为关联规则就是相关分析，或者认为关联规则分析就是要找出某一事件或是资料中会同时出现的东西。上述理解都具有片面性，前者混淆了一般统计学概念中的相关分析和海量数据挖掘中关联规则的概念；后者只是说了关联规则分析的目的，属于以偏概全。由于关联规则的定义要涉及专业的数学术语，此处以举例形式介绍关联分析。关联规则挖掘的主要对象是事务数据库。在进行事务数据库分析时，经常会考察到一些涉及许多属性项的事务：譬如事务 A 中出现了属性项甲，事务 B 中出现了属性项乙，事务 C 中同时出现了属性甲和乙，研究属性甲乙之间的相互影响性便可称之为关联规则分析。关联规则由于其商业用途广，相关软件便于操作，目前是大型数据挖掘时应用最多的一种方法。现举一例说明目前关联规则的商业用途：某知名饮料公司欲向市场推销一种新的饮料，已知市场现有饮料品种为，冰糖红茶、雪梨、苹果汁、水晶葡萄、冰糖绿茶、菠萝汁、鲜橙汁、冰糖石榴、水蜜桃汁等，通过对调查各种饮料的购买数据并进行关联规则分析，该公司决定向市场推出主打产品——冰糖雪梨。后来的事实表明，冰糖雪梨获得了广大消费者喜爱，成功占领饮料市场一定份额。

6. 时态规则。时态规则与关联规则有很多相似之处，二者最大的不同是时态规则中相关的项目是通过时间区别开来的（例如：某股票 A 在某一天上涨 8%，并且当天股市加权指数下降，则 B 股票在未来两天上涨概率是 54%）。传统的数据挖掘着重于研究某一时间点上的静态数据并关注总体信息，往往忽视了个体差异性。随着专业分工细化、市场细分需要以及个性化服务的需要，现在越来越多的行业要求数据挖掘能提供关于个体行为特征和规律的信息，从而便于生产和销售管理。因此，时态规则分析是建立在对个体进行足够多次重复观察数据的基础上的。一个具体例子是分析手机用户的使用情况，现有一个按月记录的手机消费数据库，数据库中包含 48748 个手机用户连续 7 个月关于用户类型、营业收入、本地话费、长途话费、漫游费、国际话费、信息费、月租费和服务费等变量数据。然后采用简单随机抽样的方法从原始数据中依次抽取了 10%，20%，……，60% 的样本。然后分析十点规则的支持度平均值、时点规则的可信度平均值和时点规则的作用度平均值等数值在不同抽样比例下的变化情况。分析结果反映了手机用户在月度消费相关项目上的滞后效应以及手机消费行为特征及各时段的支持度和可信度情况，并借此分析手机用户的消费习惯特征。

除了上述介绍的单一的数据挖掘方法针对性地在生产生活中的应用外，在信息爆炸和人们追求多样化生活的时代，我们经常面对的是纷繁芜杂的数据资料，因此，我们更多地是运用多种数据挖掘方法分析来自生产生活中的数据资料。其他综合运用数据挖掘方法分析的经典案例有：信用卡公司使用数据挖掘的方法分析持卡人性别、职业、年龄、受教育程度、工作年限、犯罪记录、刷卡地点、刷卡数额等指标从而推销信用卡、设定担保额、分析持卡人购买行为及侦测信用卡欺诈行为等；大型超市通过数据挖掘方法分析会员年龄、收入、每次购买商品情况等数据资料，从而了解消费者商品偏好以便及时向其会员推荐其可能有购买倾向的商品、决定商品存储内容和存储数额和以使其存储成品最小以及决定商品架商品的摆放等 ；零售商使用数据挖掘方法分析消费者购物篮和电子销售点资料从而决定促销手段、广告投入、甚至侦测收银员的欺诈行为；航空公司利用数据挖掘方法分析各类乘客的收入、职业、公务活动以及交通支付行为等以便为各类潜在购买者量身打造服务产品从而在激烈竞争的航空业求得立足之地；数据挖掘技术还被用在制造业的生产控制过程中以侦测不合格产品以及用在医疗业中预测手术、用药、诊断过程的疗效。

还有一点必须强调的是，数据挖掘技术对于统计工作的意义重大。数据挖掘与统计工作向来是联系紧密的，统计学理论为数据挖掘方法提供理论支持，离开了统计学，数据挖掘方法就是无源之水无本之木，数据挖掘方法是统计方法在新时期遇到新问题时发展起来的新方法，是统计学强大生命力的体现。一个具体的应用是数据挖掘可以为统计抽样工作提供辅助信息从而指导抽样工作，提高抽样估计精度。

在信息爆炸的时代，在数据资料满世界的今天，数据挖掘技术几乎无处不在，无所不能，它在人们的生产生活中将发挥越来越重要的作用。因此，可以毫不夸张的说，对于深谙数据挖掘的专业人员来说，满世界的数据资料意味着满世界的机遇，也意味着满地的黄金，他们每一次的数据挖掘都将挖到货真价实的真金白银。

作者单位：厦门大学经济学院