Business Analytics Software

零售业数据挖掘及客户洞察

分析

分析简介

本文介绍用于零售业市场研究的典型的分析应用；促使我们进行此类分析的动力，是实现更为深入、更可盈利的客户交互方式。直到一段时间以前，理解客户的任务一直是由营销人员来完成的，他们可以使用信息技术执行较浅层面的统计分析。对客户的细分主要基于人口统计信息：例如，将年龄在30到40之间的、有研究生学历的客户划分到一个族群中，将年龄在20到30的研究生归属到另一族群；然后这两个族群还会进一步细分，以采取具体的营销行动。

随着信息技术的进步，新的分析工具和解决方案不断涌现，它们使我们能够越来越轻松地执行越来越复杂的分析，以便在理解客户方面增加更多价值。

近年来，分析流程经历了巨大的变化，数据挖掘积极进军营销活动领域，使对海量数据执行复杂的统计分析成为可能，它可以揭示一般方法不能发现的深层次关系，包括从人口统计变量(年龄、学历、婚姻状况)到交易数据(已购商品、店内消费金额、最喜欢的购物日)中揭示客户特征。

分析的角度发生了显著的变化。例如，客户族群细分分析敞开了一扇通往无人能预见或预测的新锐洞察的大门：例如它有可能发现，年龄在30和40之间的研究生客户群与年龄在20和40之间的研究生客户群拥有共同的行为特征。显然，这两个在静态分析中分属于两个不同族群的亚群体，在统计分析中属于同一个细分人群或族群，如下图所示。

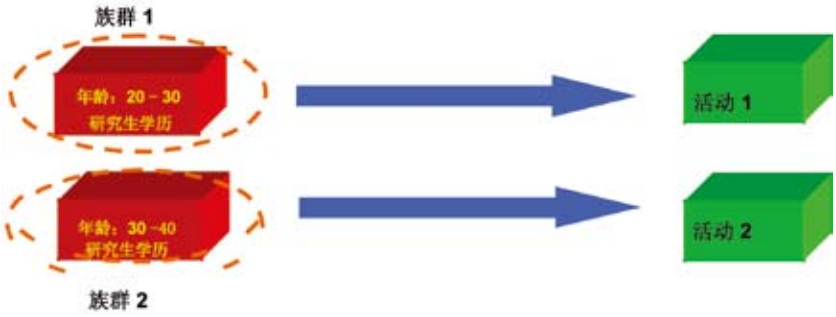


图1 静态分割

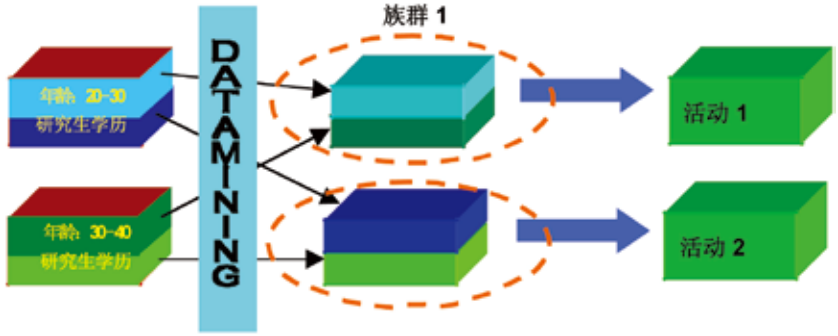


图2 统计分割

另一个典型的用例是预测促销响应。直到不久以前，预测促销活动的有效性和效率一直是极其困难的，主要依靠营销产品的组织者和倡议者的经验。人们曾经试图根据以前促销活动的结果预测这次营销的进展。也曾将或多或少复杂的技术用在选择促销活动和活动的目标客户上。利用数据挖掘可以在过去和未来的行为之间建立一种联系，从而根据客户现在的行为预测他对促销活动的响应。

对客户洞察类型的分析是更为有效和可盈利的方式。它让您能够更深入、更准确地了解您的客户群，而这在不久以前是不可能做到的。本文介绍的分析有:

- 描述性模型: 深化对客户的了解(客户群细分和购物篮分析)
- 预测性模型: 预测客户未来的行为(通过评分预测营销活动的效果)

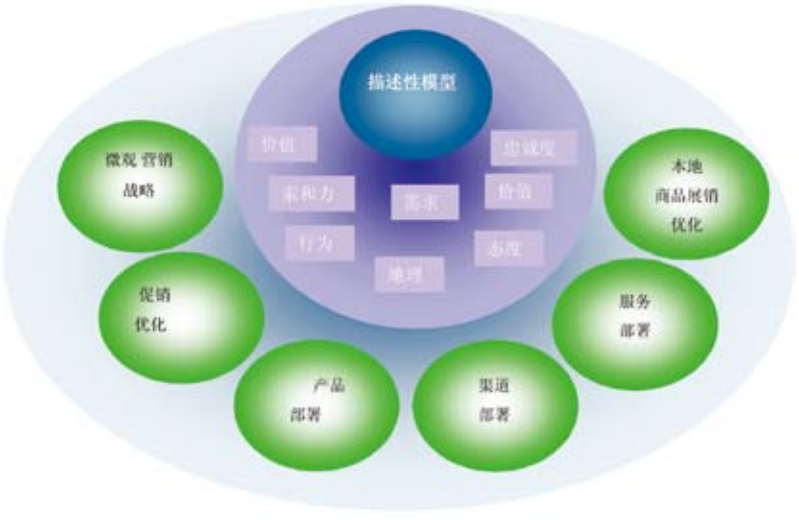


图3 描述性模型

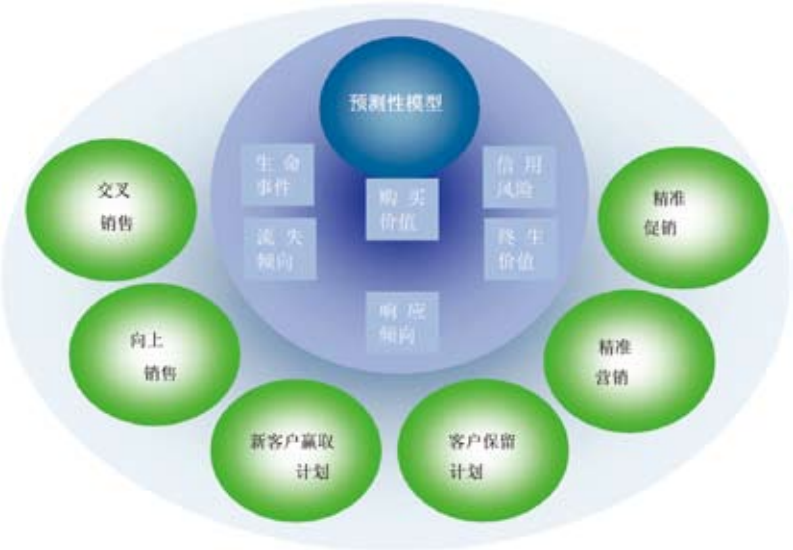


图4 预测性模型

分析最终的结果如下:

1. 客户细分: 如上所述, 根据人口统计数据和行为模式, 每一位客户将被划分到一个特定的族群。一旦识别不同族群的特征和真实价值, 即可针对每个族群实施量身订造的营销行动。典型的案例有:

   针对公司的核心细分客户群体展开客户保留活动, (通常这是一个非常小比例的群体, 但是却能够带来80%的销售额)。因为这些客户对公司的效益做出了较大的贡献, 可以采取行动阻止这些客户流失, 保持较高的购买水平, 加强客户忠诚度。

   "推动" 针对有可能购买产品的客户的营销活动的展开(新客户的赢取)。这些活动旨在吸引极有可能成为客户的潜在客户, 这些客户表现出与现有高价值客户相似的特征。

2. 促销活动的效率: 预测每一位客户对不同的促销活动的响应率, 可以简化营销活动(将促销信息发送给谁, 使用什么渠道发送), 控制营销活动的开支(削减直接邮寄的成本)。

3. 购物篮分析: 确定产品的亲和力(哪些产品最有可能在考虑阶段被客户一起购买)。这一数据让我们能够更好地界定一个合适的促销产品范围(促销哪些产品, 如何促销)以及商店的布局(在过道的中间摆放关系相近的、可以一起购买的产品)。

背景

本文所述的分析引用意大利零售市场的一家拥有食品和非食品部门的领先公司(Cash & Carry)的数据。被考虑的数据中包含了每一位客户在12个月内的日常购买情况。

可用的数据的特征如下:

1. 客户群的大小: 分散于全国的30个销售点的所有商店总共拥有800,000名客户。
2. 商店客户的拓扑学: 典型的客户术语食品/餐饮业(旅馆、饭店、披萨饼店、酒吧), 但是还有许多其他类别的客户, 如贸易和自由职业者。
3. 考察的时间段: 2000年7月-2001年7月。
4. 可提供的最详细信息: 在客户-商品水平上的每日购买历史。
5. 此分析仅计算了一个商店的数据, 以作为一个参考示例。

在客户数据集市中收集该数据, 并提取执行数据挖掘分析所需的信息。定义要使用的变量的数目和类型的分析阶段对于数据挖掘项目的成功来说是最基础的, 也是不可或缺的。

在创建分析模型之前, 我们集中精力从业务的角度理解项目的目标和要求并研究所收集到的购物数据。在这种情况下我们将从营销高管的贡献中受益, 他们对于涉及的变量和所研究的市场有着深入的了解。

方法: CRISP-DM

此分析遵循CRISP-DM方法，现在介绍如下。

CRISP-DM的诞生

CRISP-DM是新开发的标准框架，帮助人们进行数据挖掘项目。该标准是由主要位于欧洲的一伙公司开发出来的，命名为跨行业数据挖掘标准过程(Cross-Industry Standard Process for Data Mining，CRISP-DM)。

CRISP-DM是在1996年由当时处于年轻又不成熟的数据挖掘市场的"三剑客"提出来的。

Daimler Chrysler (以前称作Daimler-Benz)、SPSS和NCR。当时，早期市场兴趣也表露出数据挖掘在全球兴起的态势。所有的实践者在发展的过程中都开始研发自己的数据挖掘产品，每一个新的数据挖掘采用者都从尝试和错误中吸取教训。1997年他们已经形成了一个联合体，并获得了欧洲委员会的资助，制定一个行业中立、工具中立、应用中立的标准，以便从尽可能多的实践者和其他对数据挖掘感兴趣的人(如数据仓库供应商和管理顾问)那里获得信息。1999年，一个新的合作伙伴OHRA加入了团队，2000年，CRISP-DM 1.0公开发表。这项标准的

目标是支持来自不同行业的人在数据挖掘活动中使用相同的术语、方法、工具。CRISP-DM不是以理论、学术的方式从技术原则上建立的，也不是由专家级精英团队闭门研制的。这两种制定方法论的方式在过去已经尝试过了，但是几乎没有达成任何实际、成功、广泛采用的标准。CRISP-DM的成功在于它是建立在数据挖掘工程实践和实际经验基础上的合理的标准。

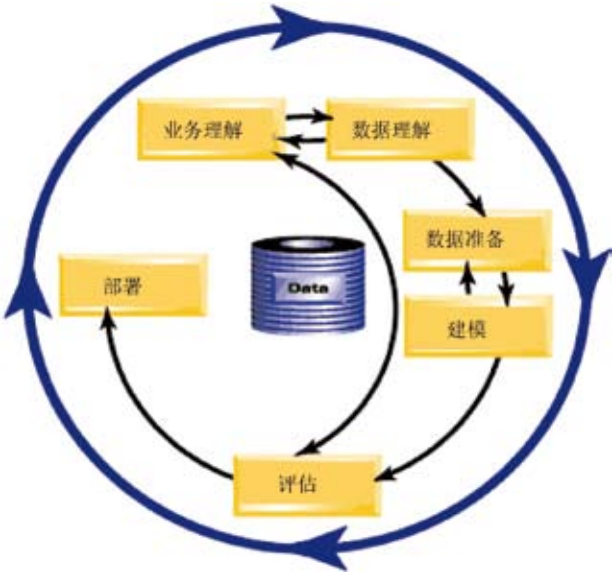
CRISP-DM模型

CRISP-DM的主要特征是，它是一个将业务目标与分析目标结合的开放模型。数据挖掘项目生命周期包含六个阶段。这些阶段的前后顺序并不是固定的。在不同的阶段间向前或向后移动总是必要的。该模型考虑了经常不可避免地返回数据挖掘过程的前一个阶段的因素。接下来要进行哪一步骤或一个阶段的哪项特定任务，取决于每一个阶段的结果。该项目不是一个由起点、一系列预定好的步骤、终点组成的线性过程。而是以循环性为特点。

解决方案部署以后，数据挖掘没有结束。在过程中和部署解决方案时所获得的经验教训可以引发新的、常常是更加有针对性的业务问题。后面的数据挖掘过程将从前面的过程中受益。

下面对每一个阶段进行简要介绍:

- **业务理解:** 这个最开始的阶段专注于从业务的角度理解项目的目标和要求，然后将这个知识转化为一个数据挖掘问题的定义和旨在实现该目标的初步规划。
- **数据理解:** 数据理解阶段以初步数据收集开始，检查数据的可访问性和解决具体的业务问题的充分性，接下来进行一些活动，目的是熟悉数据，识别数据质量的问题，从而获得关于数据的第一手信息，发现有趣的子集，形成对隐含的信息的假设。



- **数据准备:** 数据准备阶段覆盖了所有从初始的原始数据构造最终数据集(将要输入建模工具的数据)的活动。数据准备的任务可能多次执行，并且没有顺序规定。任务包含表格、记录和属性的选择以及为建模工具转换和清洗数据。
- **建模:** 在这一阶段，可以选择并应用不同的建模技术，并且将参数也校准到最优的值。通常用于解决同一类型的数据挖掘问题的几个技术要经过测试，以寻找满足特定需求的最适合的技术。有些技术对数据的格式有具体的要求。因此，经常需要返回数据准备的阶段。
- **评估:** 在项目的这个阶段，你已经构建了一个(或多个)从数据分析的角度来说看上去质量较高的模型。在进行模型的最终部署之前，一定要确定它正确地反映了业务的目标。关键的目标是确定是否有一些重要的业务问题没有充分考虑到，让你必须返回到业务理解阶段。

– **部署:** 在创建模型过程中获得的知识可以被组织起来并以用户能够使用的方式将其呈现。数据挖掘解决方案必须像简单的静态报表一样部署给决策者，或直接写入现有的数据库(数据库计分)。

细分

简介

被考虑的零售商可以处理一系列广泛而精确的客户信息，因为每一位客户都持有一张会员卡，在结账的时候需要出示。这一功能带来了很大的竞争优势，而且还没有得到充分的开发和利用。将客户细分作为数据挖掘过程中的一个部分，让我们能够获得对客户群的更深入的洞察；所获得的结果为我门提供了制定与客户为中心的营销活动(专门针对每一个界定的族群)所需的深度信息。

分析的目标

细分分析在整个全体中的一系列细分族群中向下钻取，每一个族群都以其成员所共有的一些特殊的行为为特征。每一个族群都代表着一个确定类型的客户，可以通过与其他族群不同的行为特征来区分。

客户细分所带来的相关获益是:

1) 更好地理解客户的行为。

数据挖掘发现了在统计分析中出现的变量之间的隐含关系，这在以前的分析方法中是不能发现的。更好地理解客户所能带来的其他价值就是将客户数据转化成了知识，因此超越了传统的分析方法(OLAP和静态报表)而且通过选择和行为让客户具有不同的特征。

2) 深入探索具体的客户细分。

在对客户群进行族群细分之后，可以执行具体而深入的分析进一步描述每一个族群展示的典型习惯。例如研究一个以购买食品为主的群体，你很可能将注意力集中在客户所选择的更详细的商品或产品类别上。如果分析“核心客户”细分族群和潜在“核心客户”细分族群的行为，你会尝试查看这两个群体购买的商品，寻找能够通过向上销售活动来促销的产品组。

3) 修改客户沟通。

了解细分族群的组成和每一族群的具体特征和购买习惯之后，如果不能有区别的对待整个客户群，那么这些分析就变得徒劳无益了。必须以有针对性的营销活动来对待每一个族群，例如针对核心客户进行客户保留活动以提高客户忠诚度、通过促销活动让客户转变成为更高价值的客户、针对专卖便宜货的客户进行疯狂促销活动等等。

实施后的模型

对于族群细分分析来说，通过使用会员卡(卡上带有一个代码，在结账时会链接到所有买的商品)，可以获得日常的购买数据。这些卡的使用为我们提供了每一位购买数据的长期记录，可供我们分析。

数据准备

通常在分析和开发分析模型的过程中，要用到一个样本，而不是整个数据集，这个样本由数量有限的观测值组成。样本的选取有多种方法，最常见的一种是随机取样(只表示想要的记录的数目或百分率)和分层取样。此处一定要定义用于分层的算法和变量。

在我们的例子中，不必抽取样本。此次分析已经专门针对一个全体的子集进行，因为它只考虑了一个特定的销售点。

准备变量

在数据准备的阶段，从原始变量中获得的一组高价值属性，是可用于进一步分析的。(图5中的阶段1):

– 每位客户在销售点的总访问次数

– 距离客户最后一次访问商店的天数

– 根据销售额和购买价格之间的差异计算的销售利润。这个指标可以按售出的全部商品来计算，也可以按食品和非食品以及促销产品进行分别计算

– 在发放会员卡的商店(home store)购物的倾向

– 在另一个销售点(foreign store)购物的倾向

– 购买促销商品的倾向

– 购买促销/非促销食品类/非食品类商品的倾向

– 花在促销商品上的销售价值的比例

– 花在促销/非促销的食品类/非食品类商品上的销售价值的比例

倾向变量根据所占客户购买的商品总数的百分比计算。例如，若要确定购买促销商品的倾向，只需计算客户购买的促销商品数与购买的商品总数之比。销售价值的比例根据占店内消费总金额的百分比计算。例如，促销商品的销售价值的比例，是根据用于购买促销商品的销售价值占客户全部销售价值的百分比计算的。

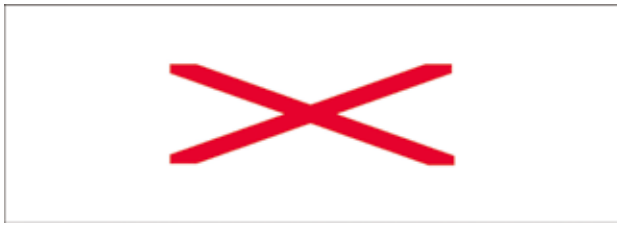


图5 数据准备的流程

全体(population)的定义

定义了所有有用的变量之后，接下来的步骤就是对客户进行分析(图5中的阶段2)，从而将所有观测值从数据集中删除，因为它们会因各种原因损害模型的质量。在本例中，分析已将所有的穿行的客户、零售商雇佣的员工以及消费额极为庞大的(超过100万意大利里拉)的客户排除在外。否则极端的值会影响分析，使生成的模型带有很大的误导性。

这一阶段对于成功部署模型来说是基本的，借助之前对客户的了解，随着他们的宏观行为和有效的变量都被纳入考虑之中，你可以从一个干净一致的数据库开始，构造一个模型。这样，统计模型可以充分利用数据构建精确的模型并输出可靠的结果。很明显，在这个阶段，对于业务的深入透彻的理解是成功进行分析所不可缺少的。

数据理解

下列步骤用于探索数据和可用的变量。为了在这个阶段取得成功，一定要仔细检查分类变量的分布状态和/或针对连续变量获得一些描述性统计数据和直方图。

这样做，你才可以开始准确了解数据库的内容。这是发现发现有缺陷的数据和错误信息的时刻。极端的、异常的或者超出了范围的值以及某些类别可疑的频度分布，可以降低模型的质量和准确度。在这个阶段，还可以检测用于编码缺失值或有一个以上编码的类别的非标准值(图6显示了一些超出范围的值/极端值，后来将其删除了)。

在有争议的情况下，删除了一些大于100%的观测值，和所有销售变量为负值的观测值。

(例如：促销中的商品价格或非食品类的销售价值)这些观测值可能是前一阶段的商品展销在这一阶段所产生的结果。

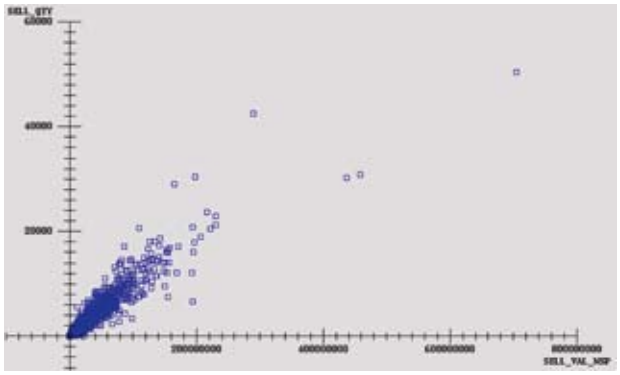


图6 显示极端值存在的图

建模

对客户进行细分最常见的方法是称为 "K-means" 的聚类算法和称为 "Kohonen networks" 的无监督的神经网络。第一个模型使用非层级算法，将观测值分配到不同的族群中，而第二个模型使用在神经网络十分典型的学习技法。在神经网络中，神经元(相当于族群)以纯二维网格分布(Kohonen地图)。

在本例中，非层级聚类算法(K-means)用于对数据库进行细分。该算法通过将具有相似行为的客户放在一起，努力实现聚类后所有族群之间的差异最大化。族群的数目必须在建模前被分析师定义好。因此，针对不同的族群数目，因建立不同的模型。在仔细评估族群和族群中的变量值后，应选择最佳的族群数目。

选择用来构造模型的客户行为驱动因素包括：

- 根据客户需为商品支付的价格而确定的销售价值
- 每位客户和每月平均销售价值

- 促销商品的销售价值
- 已购买商品的总数
- 已购买的促销商品的总数
- 食品类和非食品类商品的总数
- 对此商店的访问的总次数
- 距离客户最后一次访问商店的时间
- 购买食品类或非食品类商品以及购买促销的食品类或非食品类商品的趋势
- 花在食品类和非食品类商品(促销或非促销)上的销售价值比例

在分析过程中并非所有可用的变量都被使用。例如，所计算的利润和所有相关的变量都被排除，因为它们与客户行为无关，而与商店的管理层的决策有关，但这对于客户来说是不可见的。

SELL_QTY	SELL_VAL_NSP_AVG	SELL_VAL_NNBP_AVG	SELL_QTY_AVG	SELL_VAL_NSP_PRM	SELL_VAL_NNBP_PRM	SELL_QTY_PRM	VI
62.823	207702	165313.827	15.706	18280	17298.0	2.0	3
114.419	162091	139961.254	19.07	141415	133471.368	10.476	6
30.68	88657	74653.126	10.227	91527	89769.311	3.68	2
175.0	636262	517989.622	21.875	1000590	794916.85	16.0	6
61.175	86266	73004.375	6.118	183492	152275.921	9.175	10
18.0	144082	114940.285	6.0	228000	202807.0	2.0	3
142.43	84715	73561.887	8.902	80200	81211.494	12.0	15
1325.738	304404	251402.106	33.143	515010	495579.706	59.898	30
317.349	363286	313773.923	26.446	318250	328438.771	36.0	11
870.334	230116	180455.927	14.506	1069077	913436.938	63.751	48

图7 客户表格的部分展示

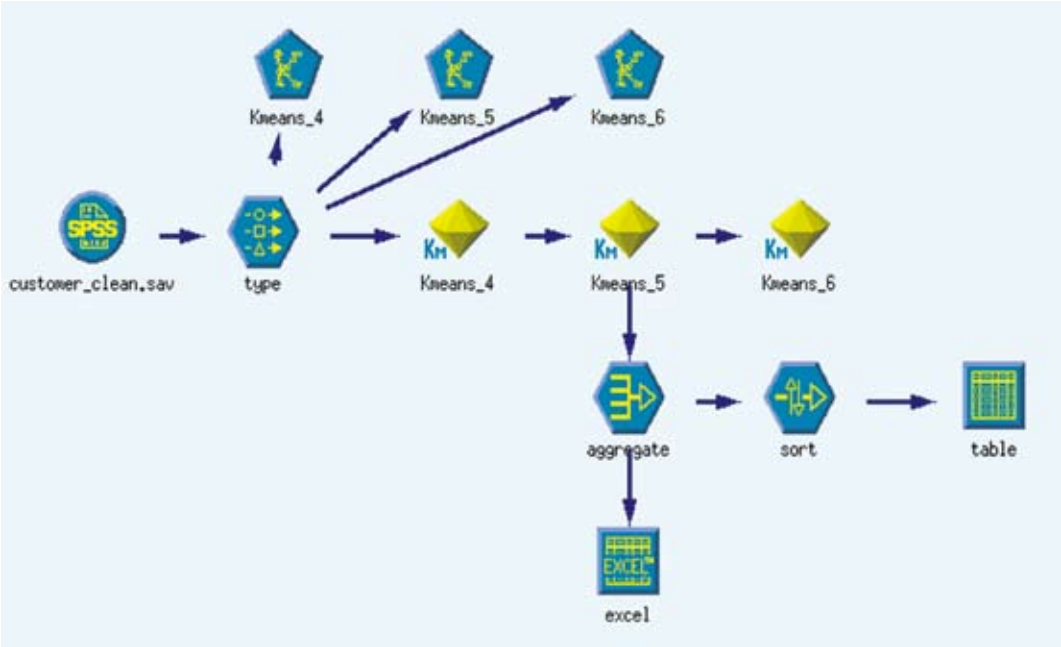


图8 创建三个聚类模型(带有4个、5个和6个族群)的建模流程

如图8所示，三个不同的模型(有4个、5个、6个族群的)经过了测试。出于下面的原因，5族群的解决方案被认为是最满意的细分方案：6族群的细分中，有两个族群只包含很少的几个观测值(可能只代表极端的行为)；4族群解决方案没有很好地描述数据库中存在的客户类型的差别，生成的细分族群不具备较显著和比较有特色的行为。

细分为5个族群，可以很好地划分客户：每一个族群都是由具有相似购买行为的个体组成，在数量上具有一致性，并且与其他族群差异很大。

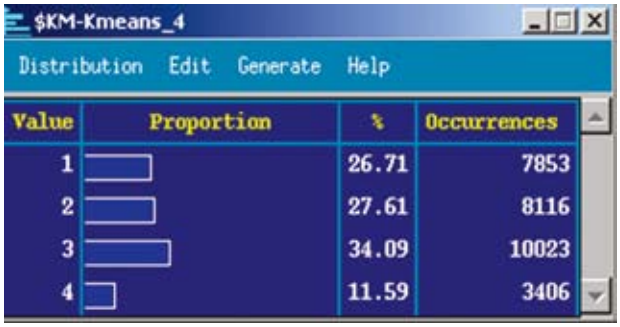


图9 在4族群解决方案中可以看到一个很大的族群(3)它可以被进一步分解，还有一个相对较小的族群(4)

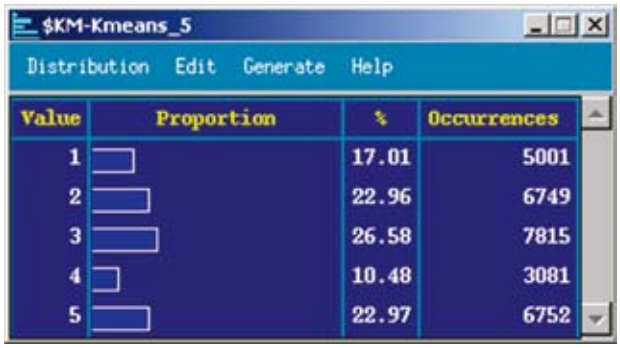


图10 5族群解决方案很好地取得了平衡，包含3个相对较大的族群(2、3、5)，但是并不含有超大族群

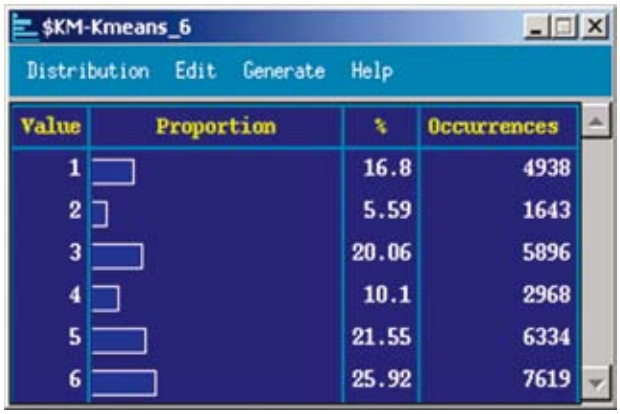


图11 6族群的解决方案有两个非常小的族群(2、4)，而且与5族群的解决方案相比，没有什么改进。

族群识别与分析。

这是细分的最后一个部分。此阶段的目的是识别已定义的族群的名称，确定它们对公司的实际价值。从统计学的观点看，聚类分析有重要意义，但是它不受业务评估的支持，不具有实际价值，因为对营销活动不能产生利润。

图12 (下一页)总结了对有29398名客户的数据库细分为5个族群的结果。

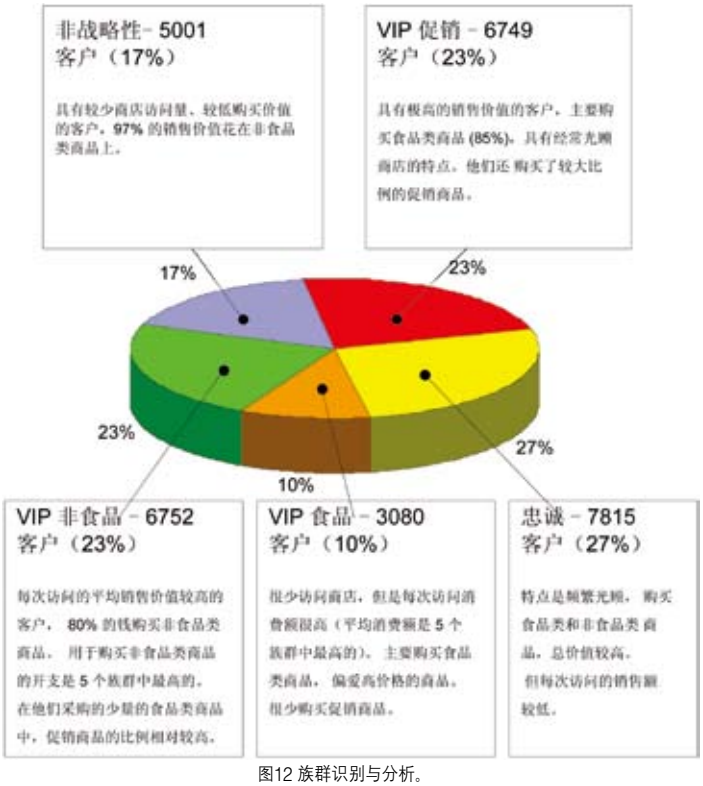


图12 族群识别与分析。

促销评分(活动优化)

简介

公司将广告目录用作于客户沟通的主要方式。目录中有不同的宣传册，分别针对食品类和非食品类产品，每两周向客户发送一次，或者在有重大活动的时候发送。处于降低邮寄成本的需要，分析的目标是寻找一种方式，确定可以将哪些客户从收件人列表中删除，集中争取最有可能响应的客户，优化促销活动的效率。

分析的目标

这类分析基于预测模型，旨在“预测”客户对于促销活动的反应，观察期其目前的行为：在一个时间段内对每一个客户进行监视，定义一个能够与客户“促销活动反应评分”相关联的标准或规则并构建一个模型，将此规则与客户过去的行为相连接。

分析的获益是：

1) 界定一个促销活动收益率最高的时间段

在启动促销活动之前预测多少人以及谁会响应：可以从广告的产品中估计响应率和销售量。

2) 识别对促销活动不感兴趣的客户

对始终对促销不感兴趣的客户进行调查，能让我们将他们与其他客户相区分。因为不同的客户群体的结构和习惯不同，可以采取精准的营销活动，或者极端地决定在有些客户身上干脆不做任何投资，因为他们只有极低的销售价值。

3) 简化营销行动

一定要不仅仅确定促销的目标群体，而且确定促销哪些产品。

实施后的模型

数据准备

就像在细分分析的数据准备的阶段中讲过的那样，这里跳过抽样阶段，专注于分析可用客户数据的特定子集。标准还与前面描述的分析的相同。

变量的定义

此模型是用于预测的，使用4个月时间段的数据，从下列4个月中计算的一个变量(目标变量)。数据准备阶段是十分复杂的并且包含了一系列步骤。每月购买情况的数据加入到了一个独特的数据集中，它包含每客户和每月一个记录，用“MESE”(月)变量表示。下面的步骤用于收集一个客户级别的数据，将所有的购买情况的数据聚集在一起，计算下列变量的每月平均值：

- 根据客户需为商品支付的价格而确定的销售价值
- 成本价计算的销售价值(为供应商提供的价格)
- 向客户销售的量

在数据准备阶段要执行的操作步骤与细分分析中的操作步骤相同。因此，创建了下列变量：

- 每位客户在销售点的总访问次数
- 距离客户最后一次访问商店的时间
- 根据销售额和购买价格(向供应商支付的价格)之间的差异计算的销售利润。该指标根据销售的总价值计算，也可按食品类和非食品类商品分别计算
- 在发放会员卡的商店(home store)购物的倾向
- 在另一个销售点(foreign store)购物的倾向
- 购买促销商品的倾向
- 购买促销/非促销食品类/非食品类商品的倾向
- 花在促销商品上的销售价值的比例
- 花在促销/非促销的食品类/非食品类商品上的销售价值的比例

倾向变量根据所占客户购买的商品总数的百分比计算。例如，要确定购买促销商品的倾向，只需计算客户购买的促销商品数与购买的商品总数之比。销售价值的比例根据占店内消费总金额的百分比计算。例如，花在促销商品上的销售价值的比例，是根据用于购买促销商品的销售价值占客户全部销售价值的百分比计算的。

目标变量的定义

在选择变量构建模型后，必须定义目标变量。这可以利用从试点活动中获得的数据进行，如本例，使用在数据集中呈现的变量构造一个变量。

例如，在第一个案例中，目标变量可以假定为值：“积极响应”-”消极响应“，这些是模型要预测的值。

如果没有可以用来衡量响应的方法(或者当不可能定义测试活动或缺乏历史数据时)，可以使用第二个标准，要求从数据集中呈现的变

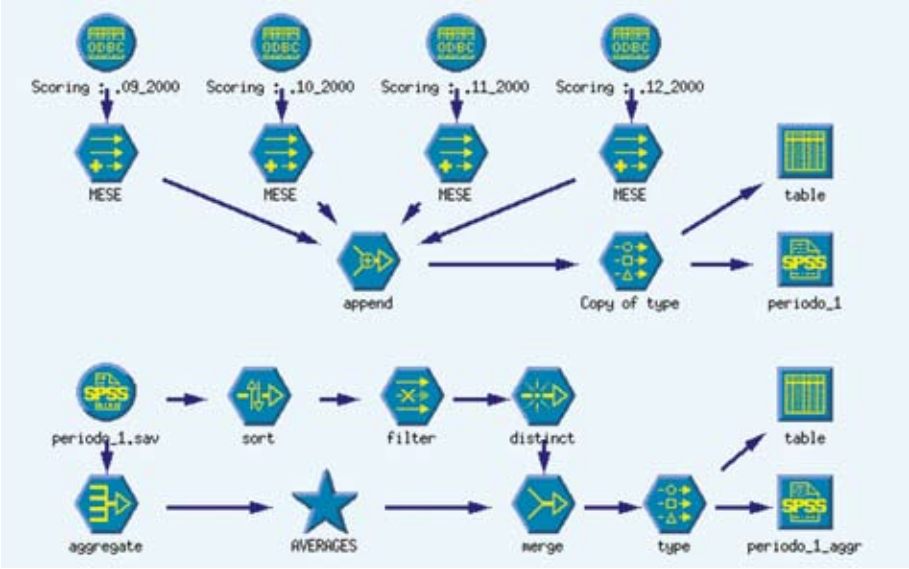


图13 数据准备的流程

量直接计算目标变量。一定要正确定义模型以便并入与业务紧密相关的人员的知识，定义区分"好"和"坏"客户的标准因为模型的质量严重依赖对于分类标准的正确定义。

对于有争议的观测值，没有从营销活动获得响应变量，选择第二个标准：目标变量的两个结果0/1是根据一个叫做效率的变量值计算的，定义为：

$$\log\left(PCT\_NSP\_P * PCT\_AP * \left(\frac{VISIT\_TOT}{120} \right) \right)$$

- PCT_NSP_P表示花在促销商品上的销售价值的比例它是所购促销商品的价值占总购买价值的比例；
- PCT_AP表示购买促销商品的倾向，定义为所购的促销商品的数量与购买商品总数量之比
- VISIT_TOT表示每位客户在调查阶段访问销售点的总次数。

根据变量的效率是否超过了预定义的阈值，将客户分为"好客户"和"坏客户"。阈值的定义用作确定促销产品销售价值的标准：以覆盖邮寄成本，一个客户至少花费£12,000在促销商品上，这是单一目录(£1,500)定价和在调查期间促销数量(4个月内8个目录)的所得的结果。

变量效率的阈值目前已定义，分析它的分布状态并且识别它的价值，至少全体的50%已经在促销商品上花费超过12,000里拉的价值。(图14)

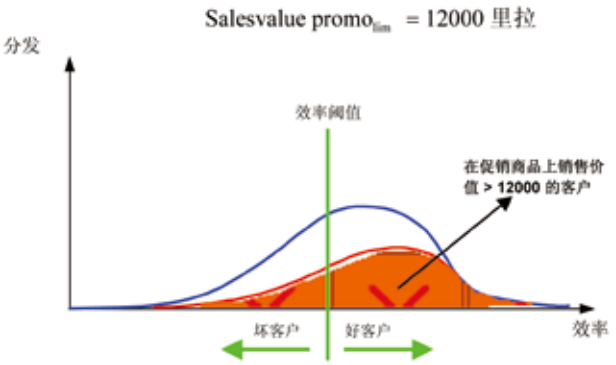


图14 定义效率的标准

建模

从目标变量获得的信息被添加到数据集，它包含预测变量(驱动因素)并随后被细分为两个不同的数据集：一个用于操练(training)模型和其他内容，另一个用于测试(test)所构造的模型的质量。客户群被随机分成两个数据集。当构造预测模型时，这种战略值得推荐。操练的数据集可帮助构造和操练模型，直到构建了尽可能最佳的模型，同时通过测试集评估模型的性能。模型测试阶段是在新数据上进行的，以免因被操练过度而生产过于精确的模型，由于过于精确

所以只适用于它所操练的数据。

图15 显示了一个高度准确的模型但是它不适合做预测。实际上，只要添加一个与操练集合不同的新的观测值，模型就会失败。图16 显示了适度精确的模型，它更适合处理和使用新的观测值进行预测

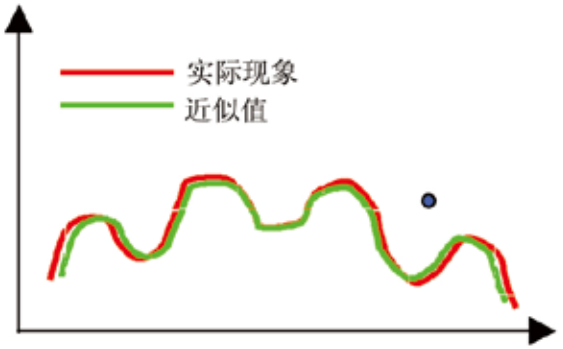


图15 过度拟合的演示

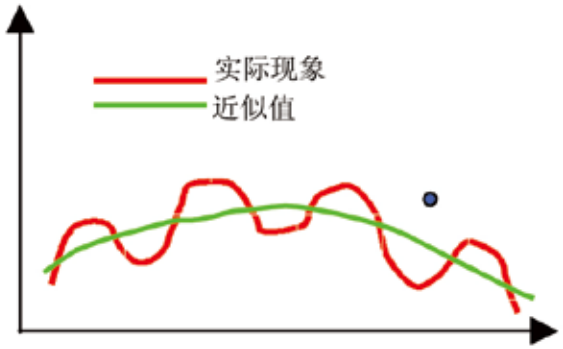


图16 适合预测的模型的演示

在此应用中，测试了两个不同的模型: 神经网络和决策树。

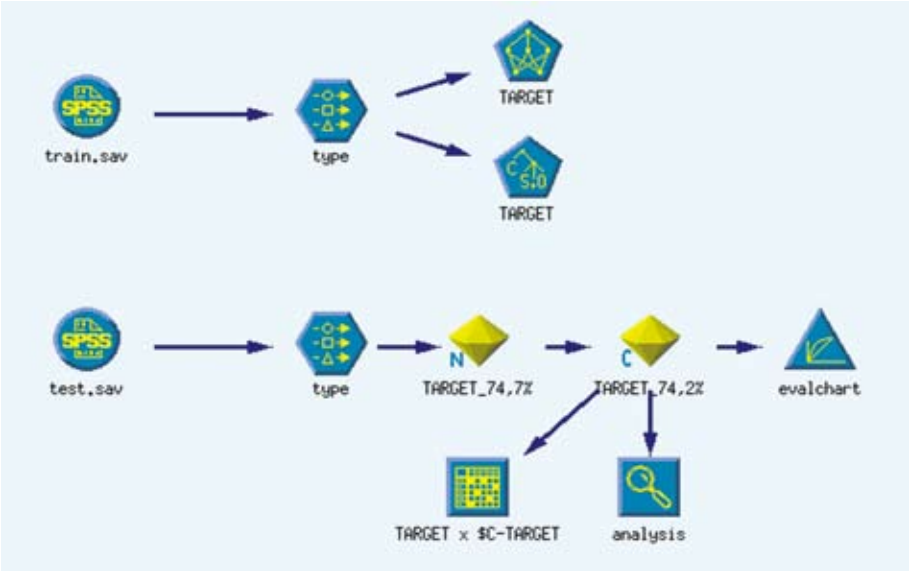


图17 建模和测试模型的流程

- 受监督的神经网络

神经网络让我们能够模拟输入和输出变量之间的复杂关系。神经网络是关于神经系统运行方式的一个简单模型。它的基本单位是神经元，神经元通常成层分布。输入数据呈献给第一层后，信息由每一个神经元向下一层的每一个神经元传递。在传递过程中该值因权重而被修改。结果在输出层得出。起初，所有的权重都是随机的。网络通过训练得到了学习。输出被获知的示例反复不断地呈现给网络，它所给出的结果被与已知的结果进行比较(受监督的学习)。从比较中获得的信息通过网络传递回去，逐渐改变了权重。一旦训练完毕，网络可以被应用到结果未知的未来观测值。

为了阻止算法过度拟合，在模型创建过程中使用70%的输入数据(随机选取)来构建模型，其余的30%用于测试所构造的模型。

- 规则归纳/决策树

规则归纳创建了一个表示如何将数据分类成不同的结果或节点的规则的决策树显示了数据中的属性如何将全体分割成与问题相关的子集。

通过一个示例可以解释规则归纳算法的工作原理。最初的一个包含全体的节点，被分为两个子集(或者超过两个子集，取决于所使用的算法)。根据第一次分类时使用的属性的值，将客户放在第一个节点或第二个节点上。每建立一个子节点，就重复这个过程，因此形成了分支，故名决策树。树的每一分支都以一个叶而结束，叶可以链接到所做出的决策并形成一个规则。该算法的目标是实现一个解决方案，其中得到的叶都互不相同，每个叶中的个案都尽可能是同类的。

在构造决策树的过程中，需要应用修剪技术来防止树过度分支，要求叶最少同时包含20个记录。

评估

在评分应用中，评估包含比较所获得的模型的结果。尤其是要使用模型评估的下列方法:

- 重合矩阵
- 模型的精确度指数
- 处理时间
- 累积反应曲线
- 捕获反应曲线

利用这些标准可以选择最佳的最适合解决这个问题的模型。两个模型的结果非常相似，决策树的准确性为74,2%，神经网络的准确性为74,7%; 这两个模型在89%的观测值中对目标变量得出的预测结果(“好客户”和“坏客户”)具有关联性。模型的评估没有到此为止，并且继续深入研究重合矩阵，该矩阵显示象征性目标的目标字段与生成的(预测的)字段之间的匹配模式。在此矩阵中，行是根据实际值定义的，列包含的是预测值。这对于在预测中识别系统错误很有用。因此，可以针对每一类别的目标变量控制正确预测的百分比。分析这两个矩阵，发现与神经网络相比，规则归纳模型预测的“坏”类别更坏一些，预测的”好“类别更好一些，划分总群体的准确度是76,26%。在仔细评估结果后，神经网络模型被选择，因为它似乎能够产生更稳定的结果。

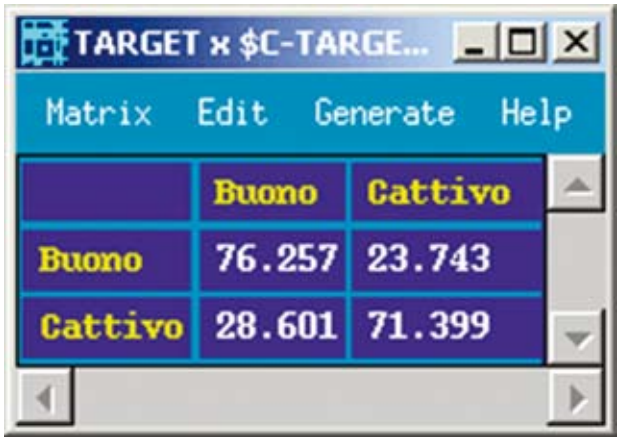


图18 行显示观测的类别，列显示决策树预测的类别

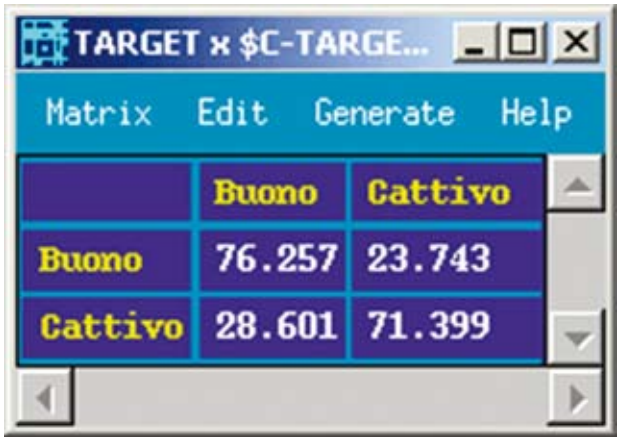


图19 行显示观测的类别，列显示神经网络模型预测的类别

因此模型的选择并不是立即进行的，而是随着彻底的评估进程，，显示了模型的效果，找出了满足分析需求的模型。

购物篮分析

分析的目标

这种类型的分析有分析单一客户或一群同类客户的“购物篮”的目标，以发现常常在一起购买的产品。通过对常常一起购买的产品和服务的类型和顺序进行研究，可以得出有用的相关规则，帮助执行向上销售或交叉销售活动。

购物篮分析基于对于自然发生的事件的频度的计算或与其他相关

分析相结合。在零售业，每一个产品的购买率(单一或结合)都可以被计算出来，从而找出具有最高亲和力的一组产品。

进行深入的购物篮分析的一个卓越的战略是在不同的水平上操作，首先定义产品类别之间的关联性(因此必须知道产品层次结构信息)，然后深入了解显示出最高亲和力的产品类别，只对属于这一范围内的产品进行深入的分析。

为了让购物篮分析更加有效，应该先对全体进行细分：这样你就可以将不同的相关产品关联在一起，从而针对每一个发现的客户细分族群有针对性地开展不同的营销活动。采用哪一种分析方法，可以从分析中获得什么样的收益，主要取决于输入模型中的数据的类型。

专注于产品间的最佳组合：这种类型的分析考虑了在所有客户细分族群以及所有现有的产品组之间发生的所有交易。识别具有高度关联性的产品可以导致关联的规则对于整个全体有效。此方法的目标是：

- 重新定义商店的布局并对商店内部的展销商品进行重新分组；
- 选择在大众市场促销的产品。

专注于特定产品：这种类型的分析的目标是促进某些特定产品的销售。注意力的焦点转移到只分析所关注的产品涉及的交易。此分析考虑到整个全体。此方法的目标是：

- 采取交叉销售等营销行动以便向与被分析的产品具有强大关联性的客户提供该产品；
- 创建一个旨在促进特定产品销售的类别。

专注于客户群：这种类型的分析是在专门准对特定客户族群规划营销活动时进行的。目的是进一步详细理解被分析客户组的购买类型。

此方法的目标是：

- 定义针对特定客户族群的向上销售和交叉销售活动，改善客户忠诚度或增加盈利能力。
- 为不同的客户细分族群创建不同的广告手册

实施后的模型

此类分析方法集中关注单一商品的购买记录和同时购买的两个或更多产品之间的关系

数据准备

可用于分析的初始数据集为考察期间内所购买的每个产品提供了一个行。每个项目包含产品ID和购买日期。第一步是为每个产品创建一个标记字段，表示产品是否在访问商店的过程中被购买。结果是一个数据集，每一行显示一次访问，而访问所涉及的所有产品或产品组都在列中显示。根据产品是否与当前行所显示的其他商品一同购买，每一列显示值1或0。

TRANSACTION_ID	3_ALIMENTARI	3_BEVANDE	3_CARNE FRESCA	3_CONSERVE	3_DETERGENZA	3_DOLCIUMCI
4459420000717	1	1	0	0	1	0
4249920000129	0	1	0	0	0	0
4440220010419	0	1	0	0	0	0
8201820010320	0	1	0	0	1	0
4733720000731	0	1	0	0	1	0
1861420010508	0	1	0	0	0	0
6348320000704	0	1	0	0	0	0
1267920000229	0	1	0	0	1	0
5832820000310	0	1	0	0	1	0
1085220000729	0	1	0	0	1	0

图20 用于购物篮分析的交易数据集的格式

需要注意的另一个方面就是分析选择的详细水平。通常，在一个商店有数千种商品(定义为品牌-模型)。这些产品被分为产品家族，还可以被进一步划分。

因此可以进行不同类型的分析，研究在不同产品组之间关系的示例，或下钻到更详细的水平，寻找特定产品之间的关系。

在将数据转换进一个包含每一商品的所有分类层次的文件后，从单个商品到产品组或家族(例如“乳制品”类别包含了奶酪、半硬质奶酪、软奶酪以及牛奶等子类)，都可以进行两个不同类型的分析。一个是宏观分析，寻找宏观类别之间的关系，一个是微观分析，寻找更细微的关系。下面的示例参照第一种类型的分析。

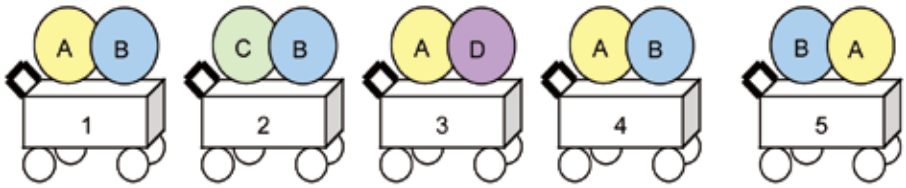
关联规则

在研究产品关联之前，不妨进行一些初步的分析，以获得关于要研究的关联度强弱的第一印象，以便更好理解可用的数据集。

与所有的数据挖掘分析一样，发现产品之间的关联度，需要进行一系列不同的尝试，调整参数并对于要解析的信息进行评估。在分析后续部分，需要定义关联规则。关联规则将一个特定的结论(特定商品的购买)与一组条件(其他商品的购买)相关联。例如：

如果条件1、条件2、条件3同时发生，这常常得出结论X

我们考虑产品A和B:



$$\text{支持 } (A \& B) = \frac{\text{购买次数 } (A \& B)}{\text{总交易数量}} = 3/5$$

规则：A <= B (如果条件 = B，那么结论 = A)

$$\text{置信度} = \frac{\text{购买次数 } (A \& B)}{\text{购买次数 } B} = 3/4$$

由下列规则得出的一个具体分析示例

乳制品<=熟食&香肠(580:12.1%, 0.784)

它显示:

覆盖度12,1%
准确度78,4%。

这意味着在12%的总交易中，熟食和香肠是同时购买的。满足条件的78%的观测值中，乳制品也被购买。

如果一个规则有两个条件，例如，条件1是购买熟食产品，条件2是购买香肠。结论X就是购买面包。

可以使用不同的方法评估所发现的关联规则:

- 先验概率: 说明一个特定产品被购买的可能性，忽略同时购买的其他产品，它就是购买产品X这一事件的概率。
- 覆盖度(或支持): 覆盖度是指数据中满足规则的先行部分(antecedent)的记录的比例-也就是说，当规则的“if”部分为真时，观测值的数量。对于特定产品来说，支持被定义为产品被购买的交易数量除以总的购买次数。

  因此，覆盖度可表示规则的通用性，或者是规则对百分之多少的数据适用。
- 准确性(或置信度): 准确性是既满足(先行)条件又符合结论的记录的比例。因此它表示关联性的强弱。

数字示例:

为了定义一个良好的关联规则，所有参数都必须被考虑。特别是先验概率不应被低估: 例如如果我们找到了一个有70%准确度但是先验概率为68%的规则，这并不能给我们多少新的信息。可以用来评估所发现的关联的重要性的度量方法有:

- **规则置信度:** 可以指定一个准确度的标准以便让规则保留在规则集中。凡是准确度小于指定标准的规则将被弃用。有最高置信度/准确度的规则是最令人感兴趣的
- **置信度与先验概率的绝对差异:** 此项评估衡量在规则置信度及其先验概率之间的绝对差异此

方法防止了结果不均匀分布的偏斜，帮助阻止"明显的(obvious)"规则被保留。

- **置信度商数(confidence quotient)与1之间的差异:** 这个指标衡量规则置信度与先验置

  信度的比值被1减所得的差值。

  像置信度与先验概率之间的绝对差异一样，这种方法把不均匀的分布考虑在内。它特别擅长寻找能够预测稀有事件的规则。
- **与先验概率之间的信息差异:** 这项评估方法基于信息增益(information gain)方法。

  信息差异是被给予先行条件(antecedent)的信息增益与被给予结果(consequent)的先验置信度的信息增益之间的差异。它考虑了规则的支持。
- **正态卡方度量方法:** 这种度量方法使用经典统计学中众所周知的卡方检验来检测先行条件(antecedent)和结果(consequent)之间的依赖关系。这种度量方法是正态的，采用从0到1的值，非常依赖支持。

一个示例将会帮助你理解为什么规则选取的方法不可忽视。下面的规则

洗涤剂<=饮料&糖果(1241:25.9%, 0.753)

准确度75.3%，但洗涤剂的先验概率已经是64.82%。因此，准确度75.3%，绝对来看是非常高的，但没有提供任何新的价值。

为了找出那些为我们提供新的有用信息的规则，你可以为每一个上面解释的标准指定一个最小值，例如置信度与先验概率或第三方度量方法显示的概率之间的绝对差异:

$$\left| 1 - \frac{\text{准确度}}{\text{先验概率:}} \right|.$$

这表示(以百分率)当规则的条件被满足时，与其一般的概率相比，结论的概率增加的可能性有多大。为了更好地理解这一概念，我们使用以前描述过的两个关联规则作为例子。

上文提到的第二个规则将洗涤剂作为结论，显示了规则的准确性为75.3%。这表示如果我们想要基于类别饮料和糖果的购买预测这种类别的产品的购买，我们必须让所有观测值的75%都符合。洗涤剂的先验概率为64.8%，也就是说，在所有交易中，有65%的购买洗涤剂，而且没有考虑同时购买的其他产品。

这个规则给我们提供了一些新的价值，但是相对来说较低；原先定义的标准假定了一个值为0.162。应用这一规则让我们能够比不考虑关联性的情况下解释16.2%更多的信息。

前面描述过的第一个规则的结论是乳制品，准确度为78.4%。购买这些产品的先验概率

是36.5%。如你所见，关于这一关联所获得的信息是很重要的。这次，该标准假定了一个等于1.148的值。因此，该规则为我们提供了115%的更多的信息。

为这个规则选取方式指定一个最小值57%，条件数量最大值等于2，覆盖度最小值为10%，准确性至少20%，可以得到下列规则集:

香肠<=乳制品&罐头食品(767:16.0%, 0.573)熟食<=乳制品&鲜肉(494:10.3%, 0.621)鲜肉<=熟食&冷冻产品(489:10.2%, 0.466)香肠<=罐头食品&熟食(556:11,6%, 0.579)

可见，第一个规则准确性为57%，明显比之前解释的规则(洗涤剂<=饮料&糖果(1241:25.9%, 0.753))低，但是"香肠"这个产品的先验概率为23.91%，因此所获得的信息是值得参考的。

可使用许多不同的选择方法识别关联规则，没有哪一个方法是最好的。它取决于你面对的业务问题的类型，并且一直需要由熟悉业务目标的人引导分析按照正确的方向进行，并且淘汰无关紧要的规则，保留那些能给我们提供新鲜、重要的信息的规则。

模型部署

模型的创建通常不是项目的最终结果，它只能解决单一的问题。这类项目的目标是总是拥有关于感兴趣的现象的最新的数据集，以便能够尽可能以最佳的方式遵循营销经理或CRM项目主管的目标。

要实现这一点，必须将模型应用到组织决策流程中去。需要经常对结果进行更新，以便用户可以一直使用必要的信息。

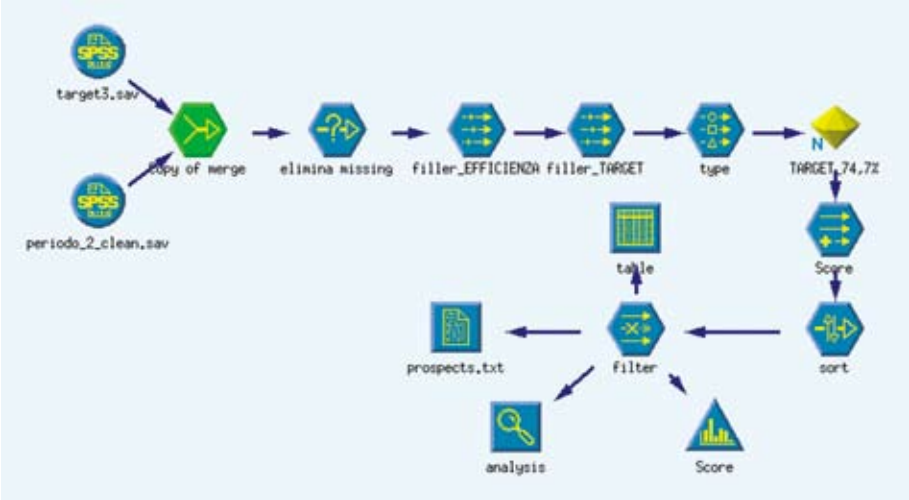


图21 部署流程

分析的整个流程转换成一系列执行规则的指令，规则可应用于数据库或文件中的一个或多个表中呈现的数据。

这些指令可以被保存在一个单独的文件中然后以批处理的模型执行，例如，由原始数据所驻留的计算机在夜间进行。该流程终止后，信息就准备好可以使用了。

通过一些例子你可以更容易理解模型的好处，它超过了斑点分析，可以转化为制定决策的一个连续固定的起始点。在分类时，模型的有用性不仅限于不同客户细分的存在和特点。所开发的模型让您能够获取规则，使其与整体中每个客户的族群标签建立因果关系。周期性地重复对客户进行分类，你能够一直了解族群结构的最新状态，及时观察特征，在客户从一个族群跨越到另一个族群时监视其状态，创建流动的矩阵，从而为公司已有的信息增加价值。

在评分模型的案例中，一旦规则得到制定并且模型得到评估，可以将分数与所有的客户建立因果关系，不仅仅是在模型创建阶段的客户。

因此，例如，可以知道什么“类型”的客户在结账或跟运营人员通电话，应该在交叉销售的活动中联系什么样的客户，总是能够妥善利用最佳潜在客户的名单。作为公司客户关系管理的一部分，此信息对于有效管理促销活动和营销活动(一对一或一对多)，是十分关键的。了解最可能在一起售出的产品组合，让你能够制定重要的战略决策，例如促销活动、交叉销售活动以及商店布局或销售点的一个部门的产品安排。此外，它还能优化存货和订单管理。从此，将模型投入生产的重要性就是更新数据以在尽可能短的时间内将企业的信息系统整合的可行性。要实现这一点，必须要下决心充分利用数据挖掘分析所提供的暗示来“激活”知识。

IBM Canada
3755 Riverside Drive
Ottawa,ON,Canada K1G

在中国印刷
2012年1月

请回收利用