# 基础篇

# 数据挖掘基础

## 某知名连锁餐饮企业的困惑

国内某餐饮连锁有限公司（以下简称T餐饮）成立于1998年，主要经营粤菜，兼顾湘菜、川菜、中餐等综合菜系。至今已经发展成为在国内具有一定知名度、美誉度，多品牌、立体化的大型餐饮连锁企业。属下员工1000多人，拥有16家直营分店，经营总面积近13000平方米，年营业额近亿元。其旗下各分店均坐落在繁华市区主干道，雅致的装潢，配之以精致的饰品、灯具、器物，出品精美，服务规范。

近年来餐饮行业面临较为复杂的市场环境，与其他行业一样餐饮企业都遇到了原材料[成本](http://www.cnbm.net.cn/add/chengben.asp)升高、人力成本升高、房租成本升高等问题，这也使得整个行业的利润率急剧下降。人力成本和房租成本的上升是必然趋势，如何在保持产品质量同时提高企业效率，成为了T餐饮急需面对的问题。从2000年开始，T餐饮通过加强信息化管理来提高效率， 目前已上线的管理系统包括：

（1）客户关系管理系统

该系统详细记录了每位客人的喜好，为顾客提供个性化服务，满足客户个性化需求。通过客户关怀，提高客户的忠诚度。比如企业能随时查询了解今天哪位客人过生日或其他纪念日，根据客人的价值分类进行相应关怀，如送鲜花、生日蛋糕、寿面等。通过本系统，还可对客户行为进行深入分析，包括客户价值分析、新客户分析与发展，并根据其价值情况提供给管理者，为企业提供决策支持。

（2）前厅管理系统

该系统通过掌上电脑无线点菜方式，改变了传统“饭店点菜、下单、结账一支笔、一张纸，服务员来回跑的局面”，快速完成点菜过程。通过厨房自动送达信息，服务员的写菜速度加快不需要再通过手写，同时传菜部也轻松不少，菜单会通过电脑自动打印出来，差错率降低，也不存在厨房人员看不懂服务员字迹而搞错的问题。

（3）后厨管理系统

信息化技术可实现后厨与前厅沟通无障碍，客人菜单瞬间传到厨房。服务员只需点击掌上电脑的发送键，客人的菜单即被传送到收银管理系统中，由系统的电脑发出指令，设在厨房等处的打印机立即打印出相应的菜单，厨师按单做菜。与此同时，收银台也打印出一张同样的菜单放在客人桌上，以备客人查询以及作结账凭据，使客人明白消费。

（4）财务管理系统

该系统完成销售统计、销售分析、财务审计，实现对日常经营销售的管理。通过报表，企业管理者很容易掌握前台的销售情况，从而达到对财务的控制。通过表格和图形可以显示餐厅的销售情况，如菜品排行榜、日客户流量、日销售收入分析等；统计每天的出菜情况，可以了解哪些是滞销菜，哪些是畅销菜，从而了解顾客的品位，有针对性地制定出一套既适合餐饮企业发展又能迎合顾客品位的菜肴体系和定价策略。

（5）物资管理系统

该系统主要完成对物资的进销存，实际上就是一套融采购管理（入库、供应商管理、账款管理）、销售（通过配菜卡与前台销售联动）、盘存为一体的物流管理系统。对于连锁企业，还涉及到统一配送管理等。

通过以上信息化的建设，T餐饮已经积累了大量的历史数据，有没有一种方法可帮助企业从这些数据中洞察商机，提取价值？在同质化的市场竞争中，找到一些市场以前并不存在的“捡漏”和“补缺”？

## 从餐饮服务到数据挖掘

企业经营最大的目的就是盈利，而餐饮业企业盈利的核心就是其菜品和顾客，也就是其提供的产品和服务对象。企业经营者每天都在想推出什么样的菜系和种类会吸引更多的顾客，究竟各种顾客各自的喜好是什么，在不同的时段是不是有不同的菜品畅销，当把几种不同的菜品组合在一起推出时是不是能够得到更好的效果，未来一段时间菜品原材应该采购多少……

T餐饮的经营者想尽快地解决这些疑问，使自己的企业更加符合现有顾客的口味，吸引更多的新顾客，又能根据不同的情况和环境转换自己的经营策略。T餐饮在经营过程中，通过分析历史数据，总结出一些行之有效的经验：

* 在点餐过程中，由有经验的服务员根据顾客特点进行菜品推荐，一方面可提高菜品的销量，另外一方面可减少客户点餐的时间和频率，提高用户体验；
* 根据菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行预测，以便餐饮企业提前准备原材料；
* 定期对菜品销售情况进行统计，分类统计出好评菜和差评菜，为促销活动和新菜品推出提供支持；
* 根据就餐频率和金额对顾客的就餐行为进行评分，筛选出优质客户，定期回访和送去关怀。

上述措施的实施都依赖于企业已有业务系统中保存的数据，但是目前从这些数据中获得有关产品和客户的特点以及能够产生价值的规律更多依赖于管理人员的个人经验。如果有一套工具或系统，能够从业务数据中自动或半自动地发现相关的知识和解决方案，这将极大地提高企业的决策水平和竞争能力。这种从数据中“淘金”，从大量数据（包括文本）中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程，就是**数据挖掘**；它是利用各种分析工具在大量数据中寻找其规律和发现模型与数据之间关系的过程，是统计学、数据库技术和人工智能技术的综合。

这种分析方法可避免“人治”的随意性，避免企业管理仅依赖个人领导力的风险和不确定性，实现精细化营销与经营管理。

## 数据挖掘的基本任务

数据挖掘的基本任务包括利用分类与预测、聚类分析、关联规则、时序模式、偏差检测、智能推荐等方法，帮助企业提取数据中蕴含的商业价值，提高企业的竞争力。

对餐饮企业而言，数据挖掘的基本任务是从餐饮企业采集各类菜品销量、成本单价、会员消费、促销活动等内部数据，以及天气、节假日、竞争对手以及周边商业氛围等外部数据；之后利用数据分析手段，实现菜品智能推荐、促销效果分析、客户价值分析、新店选点优化、热销/滞销菜品分析和销量趋势预测；最后将这些分析结果推送给餐饮企业管理者及有关服务人员，为餐饮企业降低运营成本，增加盈利能力，实现精准营销，策划促销活动等提供智能服务支持。

## 数据挖掘建模过程

从本节开始，将以餐饮行业的数据挖掘应用为例，详细介绍数据挖掘的建模过程，如图1‑1所示。

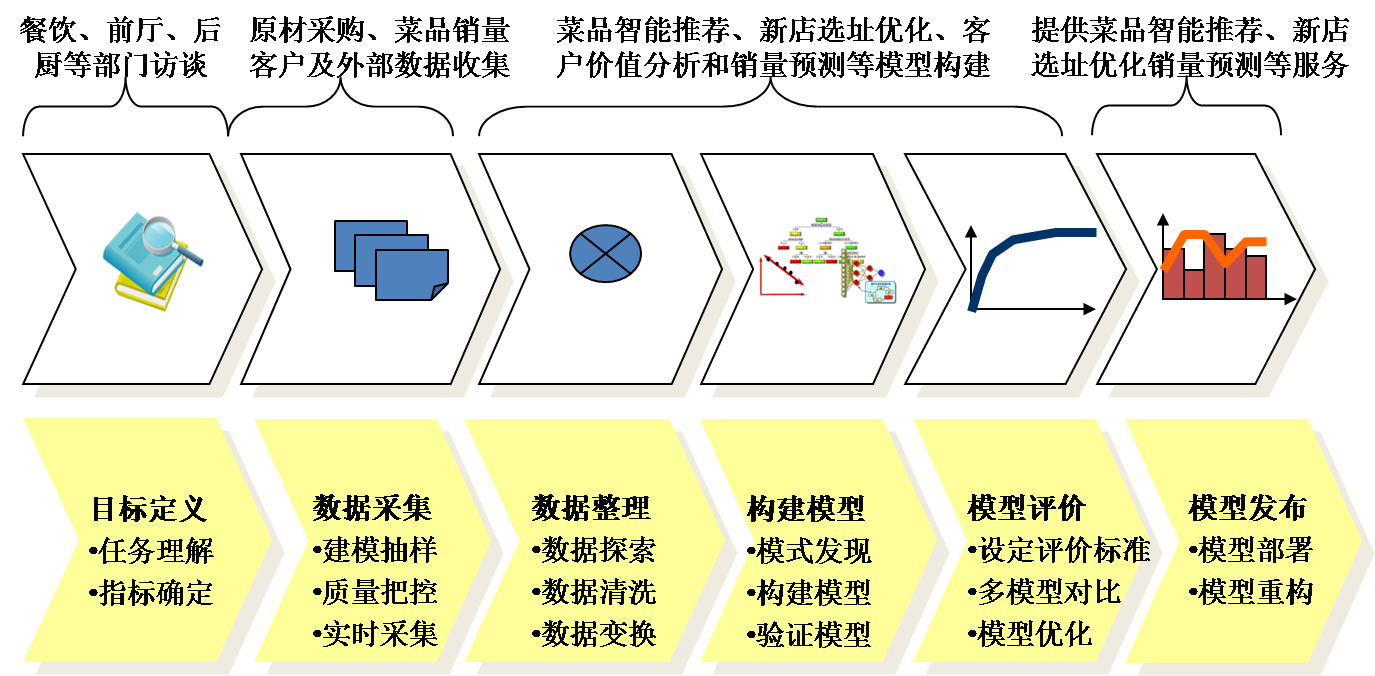


图1‑1 餐饮行业数据挖掘建模过程

### 定义挖掘目标

针对具体的数据挖掘应用需求，首先要明确本次的挖掘目标是什么？系统完成后能达到什么样的效果？因此我们必须分析应用领域，包括应用中的各种知识和应用目标，了解相关领域的有关情况，熟悉背景知识，弄清用户需求。要想充分发挥数据挖掘的价值，必须要对目标有一个清晰明确的定义，即决定到底想干什么。

针对餐饮行业的数据挖掘应用，可定义如下挖掘目标：

* 实现动态菜品智能推荐，帮助顾客快速发现自己感兴趣的菜品，同时确保推荐给顾客的菜品也是餐饮企业所期望的，实现餐饮消费者和餐饮企业的双赢；
* 对餐饮客户进行细分，了解不同客户的贡献度和消费特征，分析哪些客户是最有价值的，哪些是最需要关注的，对不同价值的客户采取不同的营销策略，将有限的资源投放到最有价值的客户身上，实现精准化营销；
* 基于菜品历史销售情况，综合考虑节假日、气候和竞争对手等影响因素，对菜品销量进行趋势预测，方便餐饮企业准备原材料；
* 基于餐饮大数据，优化新店选址，并对新店位置的潜在顾客口味偏好进行分析，以便及时进行菜式调整。

### 数据取样

在明确了需要进行数据挖掘的目标后，接下来就需要从业务系统中抽取出一个与挖掘目标相关的样本数据子集。抽取数据的标准，一是相关性，二是可靠性，三是有效性，而不是动用全部企业数据。通过数据样本的精选，不仅能减少数据处理量，节省系统资源，而且使我们想要寻找的规律性更加突显出来。

进行数据取样，一定要严把质量关。在任何时候都不能忽视数据的质量，即使是从一个数据仓库中进行数据取样，也不要忘记检查其质量如何。因为数据挖掘是要探索企业运作的内在规律性，原始数据有误，就很难从中探索规律性。若真的从中还探索出来了什么“规律性”，再依此去指导工作，则很可能会造成误导。若从正在运行的系统中进行数据取样，更要注意数据的完整性和有效性。

衡量取样数据质量的标准包括：

1. 资料完整无缺，各类指标项齐全。
2. 数据准确无误，反映的都是正常（而不是异常）状态下的水平。

对获取的数据，可再从中作抽样操作。抽样的方式是多种多样的，常见的有：

* 随机抽样：在采用随机抽样方式时，数据集中的每一组观测值都有相同的被抽样的概率。如按10%的比例对一个数据集进行随机抽样，则每一组观测值都有10%的机会被取到。
* 等距抽样：如按 5%的比例对一个有100 组观测值的数据集进行等距抽样，则有：100 / 5 = 20，等距抽样方式是取第20、40、60、80 和第100 五组观测值。
* 分层抽样：在这种抽样操作时，首先将样本总体分成若干层次（或者说分成若干个子集）。在每个层次中的观测值都具有相同的被选用的概率，但对不同的层次可设定不同的概率。这样的抽样结果通常具有更好的代表性，进而使模型具有更好的拟合精度。
* 从起始顺序抽样：这种抽样方式是从输入数据集的起始处开始抽样。抽样的数量可以给定一个百分比，或者直接给定选取观测值的组数。
* 分类抽样：在前述几种抽样方式中，并不考虑抽取样本的具体取值。分类抽样则依据某种属性的取值来选择数据子集。，如按客户名称分类、按地址区域分类等。分类抽样的选取方式就是前面所述的几种方式，只是抽样以类为单位。

基于上节定义的针对餐饮行业的挖掘目标，需从客户关系管理系统、前厅管理系统、后厨管理系统、财务管理系统和物资管理系统抽取用于建模和分析的餐饮数据，主要包括：

1. 餐饮企业信息：名称、位置、规模、联系方式；部门、人员、角色等；
2. 餐饮客户信息：姓名、联系方式、消费时间、消费金额等；
3. 餐饮企业菜品信息：菜品名称、菜品单价、菜品成本、所属部门等；
4. 菜品销量数据：菜品名称、销售日期、销售金额、销售份数；
5. 原材料供应商资料及商品数据：供应商姓名、联系方式、商品名称；客户评价信息；
6. 促销活动数据：促销日期、促销内容、促销描述；
7. 外部数据，如天气、节假日、竞争对手以及周边商业氛围等数据。

### 数据探索

前面所叙述的数据取样，多少是带着人们对如何实现数据挖掘目的的先验认识进行操作的。当我们拿到了一个样本数据集后，它是否达到我们原来设想的要求；其中有没有什么明显的规律和趋势；有没有出现从未设想过的数据状态；属性之间有什么相关性；它们可区分成怎样一些类别……，这都是要首先探索的内容。

对所抽取的样本数据进行探索、审核和必要的加工处理，是保证最终的挖掘模型的质量所必需的。可以说，挖掘模型的质量不会超过抽取样本的质量。数据探索和预处理的目的是为了保证样本数据的质量，从而为保证模型质量打下基础。

针对1.4.2节采集的餐饮数据，数据探索主要包括：异常值分析、缺失值分析、相关分析、周期性分析等，有关介绍详见第3章。

### 数据预处理

当采样数据维度过大时，如何进行降维处理、缺失值处理等都是数据预处理要解决的问题。

由于采样数据中常常包含许多含有噪声、不完整、甚至不一致的数据，对数据挖掘所涉及的数据对象必须进行预处理。那么如何对数据进行预处理以改善数据质量，并最终达到完善最终的数据挖掘结果的目的呢？

针对采集的餐饮数据，数据预处理主要包括：数据筛选、数据变量转换、缺失值处理、坏数据处理、数据标准化、主成分分析、属性选择、数据规约等，有关介绍详见第3章。

### 挖掘建模

样本抽取完成并经预处理后，接下来要考虑的问题是：本次建模属于数据挖掘应用中的哪类问题（分类、聚类、关联规则、时序模式或是智能推荐），选用哪种算法进行模型构建？

这一步是数据挖掘工作的核心环节。针对餐饮行业的数据挖掘应用，挖掘建模主要包括基于关联规则算法的动态菜品智能推荐、基于聚类算法的餐饮客户价值分析、基于分类与预测算法的菜品销量预测、基于整体优化的新店选址。

以菜品销量预测为例，模型构建是对菜品历史销量，综合考虑节假日、气候和竞争对手等采样数据轨迹的概括，它反映的是采样数据内部结构的一般特征，并与该采样数据的具体结构基本吻合。模型的具体化就是菜品销量预测公式，公式可以产生与观察值有相似结构的输出，这就是预测值。

### 模型评价

从1.4.5节的建模过程中会得出一系列的分析结果，模型评价的目的之一就是从这些模型中自动找出一个最好的模型出来，另外就是要根据业务对模型进行解释和应用。

对分类与预测模型和聚类分析模型的评价方法是不同的，具体评价方法详见第5章相关章节介绍。

## 常用数据挖掘建模工具

数据挖掘是一个反复探索的过程，只有将数据挖掘工具提供的技术和实施经验与企业的业务逻辑和需求紧密结合，并在实施过程中不断地磨合，才能取得好的效果。下面简单介绍几种常用的数据挖掘建模工具：

1. R

R 是一种为统计计算和图形显示而设计的语言环境，是贝尔实验室的Rick Becker，John Chambers和Allan Wilks开发的S语言的一种实现。在S语言源代码的基础上，1995年Auckland 大学的Robert Gentleman和Ross Ihaka 编写了一套能执行S语言的软件，并将该软件的源代码全部公开，这就是R软件的雏形，其命令被统称为R语言。用户可以自己设计相应的程序，并且可以做成拓展包发布。其他的使用者可以根据需要下载并加载软件包，从而非常方便的拓展R的内容。

1. Python

Python是一门简单易学且功能强大的编程语言。它拥有高效的高级数据结构，并且能够用简单而又高效的方式进行面向对象编程。Python优雅的语法和动态类型，再结合它的解释性，使其在大多数平台的许多领域成为编写脚本或开发应用程序的理想语言。

1. SAS Enterprise Miner

Enterprise Miner（EM）是SAS推出的一个集成的数据挖掘系统，允许使用和比较不同的技术，同时还集成了复杂的数据库管理软件。它的运行方式是通过在一个工作空间（workspace）中按照一定的顺序添加各种可以实现不同功能的节点，然后对不同节点进行相应的设置，最后运行整个工作流程(workflow)，便可以得到相应的结果。

1. IBM SPSS Modeler

IBM SPSS Modeler原名Clementine，2009年被IBM收购后对产品的性能和功能进行了大幅度改进和提升。它封装了最先进的统计学和数据挖掘技术，来获得预测知识并将相应的决策方案部署到现有的业务系统和业务过程中，从而提高企业的效益。IBM SPSS Modeler拥有直观的操作界面、自动化的数据准备和成熟的预测分析模型，结合商业技术可以快速建立预测性模型。

1. SQL Server

Microsoft的SQL Server中集成了数据挖掘组件——Analysis Servers，借助SQL Server的数据库管理功能，可以无缝地集成在SQL Server数据库中。在SQL Server2008中提供了决策树算法、聚类分析算法、Naive Bayes 算法、关联规则算法、时序算法、神经网络算法、线性回归算法等9种常用的数据挖掘算法。但是其预测建模的实现是基于SQL Server平台的，平台移植性相对较差。

1. MATLAB

MATLAB（Matrix Laboratory，矩阵实验室）是美国Mathworks公司开发的应用软件，具备强大的科学及工程计算能力，它不但具有以矩阵计算为基础的强大数学计算能力和分析功能，而且还具有丰富的可视化图形表现功能和方便的程序设计能力。MATLAB并不提供一个专门的数据挖掘环境，但它提供非常多的相关算法的实现函数，是学习和开发数据挖掘算法的很好选择。

1. WEKA

WEKA (Waikato Environment for Knowledge Analysis)是一款知名度较高的开源机器学习和数据挖掘软件。高级用户可以通过Java编程和命令行来调用其分析组件。同时，WEKA也为普通用户提供了图形化界面，称为WEKA Knowledge Flow Environment和WEKA Explorer，可以实现预处理、分类、聚类、关联规则、文本挖掘、可视化等。

1. RapidMiner

RapidMiner也称为YALE（Yet Another Learning Environment，https://rapidminer.com），提供图形化界面，采用类似Windows资源管理器中的树状结构来组织分析组件，树上每个节点表示不同的运算符（operator）。YALE中提供了大量的运算符，包括数据处理、变换、探索、建模、评估等各个环节。YALE是用Java开发的，基于Weka来构建，可以调用Weka中的各种分析组件。RapidMinner有拓展的套件Rhadoop，可以和Hadoop集成起来，在Hadoop集群上运行任务。

## 小结

本章从一个知名餐饮企业经营过程中存在的困惑出发，引出数据挖掘的概念、基本任务、建模过程及常用工具。

如何帮助企业从数据中洞察商机，提取价值，这是现阶段几乎所有企业都关心的问题。通过发生在身边案例，由浅入深引出深奥的数据挖掘理论，让读者在不知不觉中感悟到数据挖掘的非凡魅力！本案例同时也贯穿到后续第3章至第5章的理论介绍中。