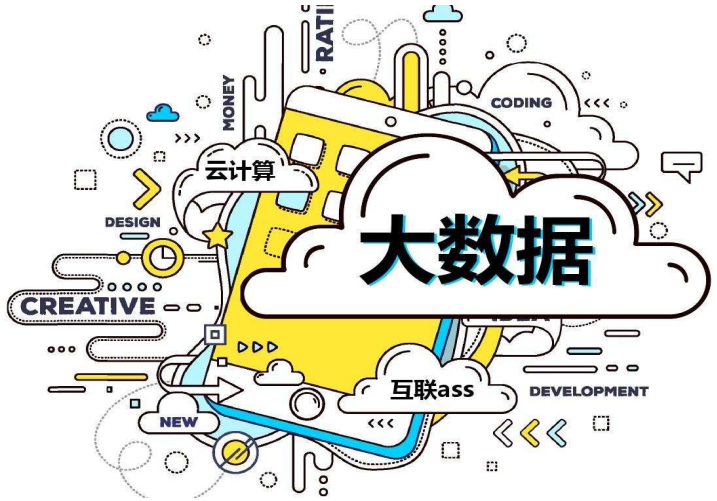
《位置服务网络与大数据》

第三讲-大数据挖掘与算法

主讲教师：姜建武

2022-3-1 · 桂林

本讲内容



- 数据挖掘概述
- 分类算法
- 聚类算法
- 关联规则算法
- 预测规模
- 数据挖掘算法综合应用

回顾

1. 卫星定位系统
2. 蜂窝基站定位
3. 新兴定位系统（AGPS）
4. 无线室内环境定位
5. 传感器网络节点定位技术
6. 传感器网络时间同步技术

数据挖掘概述

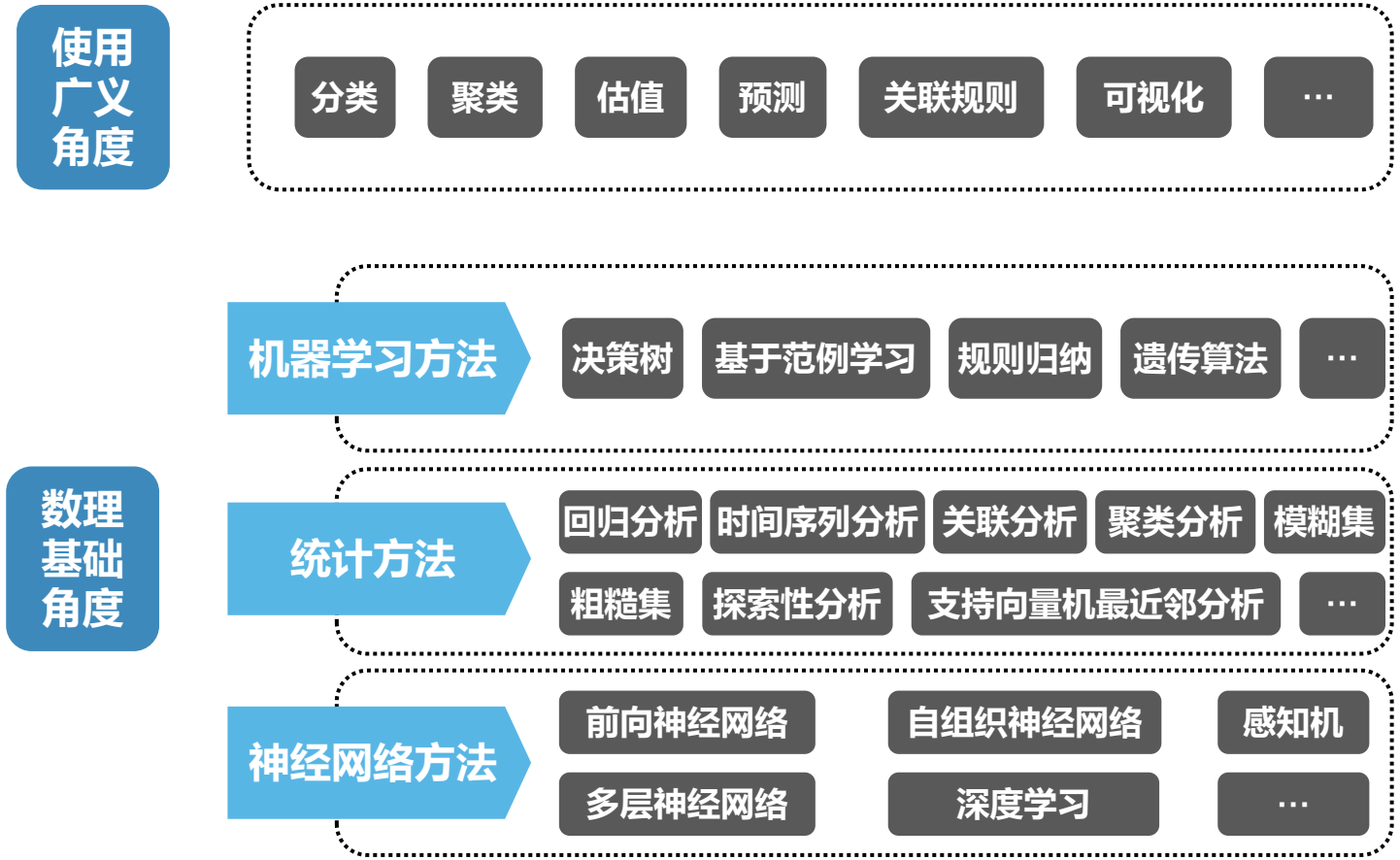
数据挖掘概念

- 20世纪80年代末，数据挖掘（Data Mining，DM）提出。
- 1989年，KDD 这个名词正式开始出现。
- 1995年，“数据挖掘” 流传。
- 从科学定义分析，数据挖掘是从大量的、有噪声的、不完全的、模糊和随机的数据中，提取出隐含在其中的、人们事先不知道的、具有潜在利用价值的信息和知识的过程。
- 从技术角度分析，数据挖掘就是利用一系列的相关算法和技术，从大数据中提取出行业或公司所需要的、有实际应用价值的知识的过程。知识表示形式可以是概念、规律、规则与模式等。准确地说，数据挖掘是整个知识发现流程中的一个具体步骤，也是知识发现过程中最重要的核心步骤。

特征

- 处理大数据的能力更强，且无须太专业的统计背景就可以使用数据挖掘工具
- 从使用与需求的角度上看，数据挖掘工具更符合企业界的需求
- 数据挖掘的最终目的是方便企业终端用户使用，而并非给统计学家检测用的

■ 数据挖掘常用算法


使用广义角度
分类
聚类
估值
预测
关联规则
可视化
…
数理基础角度
机器学习方法
决策树
基于范例学习
规则归纳
遗传算法
…
统计方法
回归分析
时间序列分析
关联分析
聚类分析
模糊集
粗糙集
探索性分析
支持向量机最近邻分析
…
神经网络方法
前向神经网络
自组织神经网络
感知机
多层神经网络
深度学习
…

数据挖掘常用算法

1．分类

- 数据挖掘方法中的一种重要方法就是分类，在给定数据基础上构建分类函数或分类模型，该函数或模型能够把数据归类为给定类别中的某一种类别，这就是分类的概念。

2．聚类

- 聚类也就是将抽象对象的集合分为相似对象组成的多个类的过程，聚类过程生成的簇称为一组数据对象的集合。

3．关联规则

- 关联规则属于数据挖掘算法中的一类重要方法，关联规则就是支持度与信任度分别满足用户给定阈值的规则。

4．时间序列预测

- 时间序列预测法是一种历史引申预测法，也即将时间数列所反映的事件发展过程进行引申外推，预测发展趋势的一种方法。

■ 数据挖掘应用场景

- 按照数据挖掘的应用场景分类，数据挖掘的应用主要涉及通信、股票、金融、银行、交通、商品零售、生物医学、精确营销、地震预测、工业产品设计等领域，在这些领域众多数据挖掘方法均被广泛采用且衍生出各自独特的算法。

1．数据挖掘在电信行业的应用

- 数据挖掘广泛应用在电信行业，可以帮助企业制定合理的服务与资费标准、防止欺诈、优惠政策，为公司决策者提供可靠的决策依据，为市场营销、客户服务、全网业务、经营决策等提供有效的数据支撑，进一步完善了国内电信公司对省、市电信运营的指导，在业务运营中发挥重要的作用，从而为精细化运营提供技术与数据的基础。

2．数据挖掘在商业银行中的应用

- 在美国银行业与金融服务领域数据挖掘技术的应用十分广泛，由于金融业务的分析与评估往往需要大数据的支撑，从中可以发现客户的信用评级与潜在客户等有价值的信息，可成功地预测客户的需求。

数据挖掘应用场景

3．数据挖掘在信息安全中的应用

- 利用机器学习与数据挖掘等前沿技术与处理方法对入侵检测的数据进行自动分析，提取出尽可能多的隐藏安全信息，从中抽象出与安全有关的数据特征，从而能够发现未知的入侵行为。数据挖掘技术可以建立一种具备自适应性、自动的、系统与良好扩展性的入侵检测系统，能够解决传统入侵检测系统适应性与扩展性较差的弱点，大幅度提高入侵检测系统的检测与响应的效能。

4．数据挖掘在科学探索中的应用

- 近年来，数据挖掘技术已经开始逐步应用到科学探索研究中。例如，在生物学领域数据挖掘主要应用在分子生物学与基因工程的研究。
  - 使用概率论模型对蛋白质序列进行多序列联配建模；
  - 特定数据挖掘技术研究基因数据库搜索技术；
  - 在被认为是人类征服顽疾的最有前途的攻关课题“DNA序列分析”过程中，由于DNA序列的构
  - 成多种多样，数据挖掘技术的应用可以为发现疾病蕴藏的基因排列信息提供新方法。

数据挖掘工具

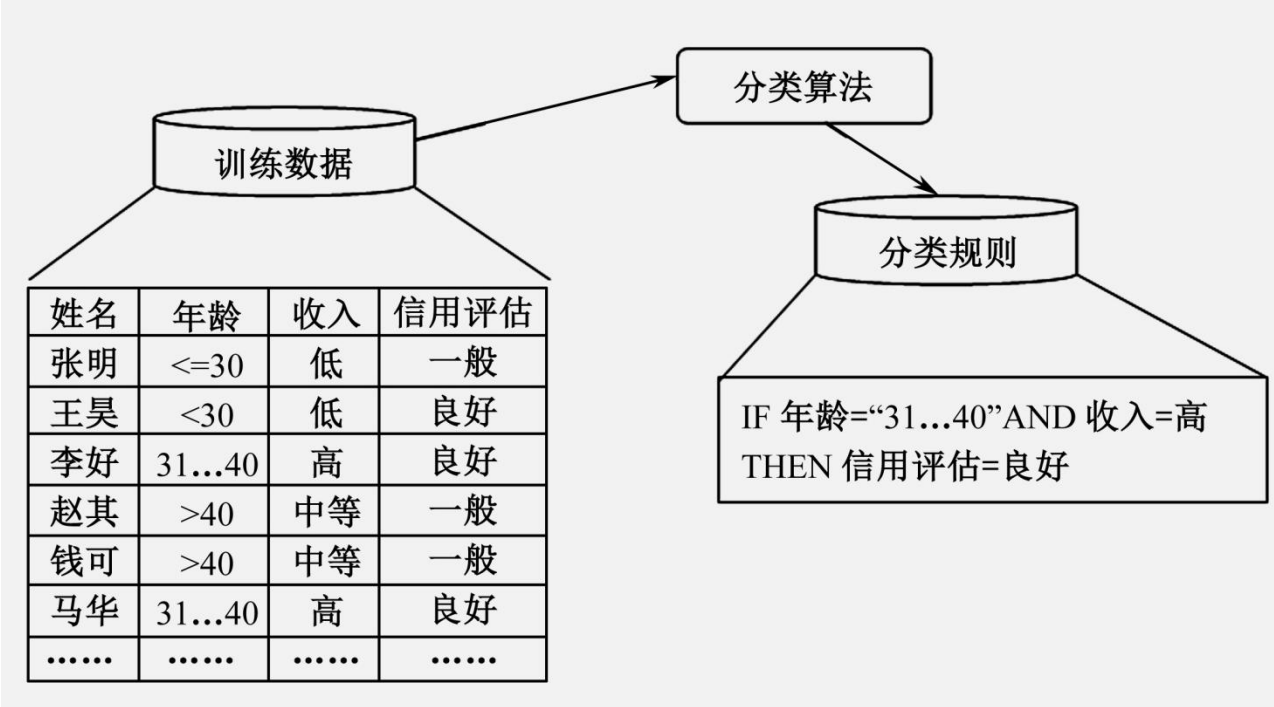
- 根据适用的范围，数据挖掘工具分为两类：专用挖掘工具和通用挖掘工具。专用数据挖掘工具针对某个特定领域的问题提供解决方案，在涉及算法的时候充分考虑数据、需求的特殊性。对任何应用领域，专业的统计研发人员都可以开发特定的数据挖掘工具。

Weka软件	公开的数据挖掘工作平台，集成大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则，以及交互式界面上的可视化。
SPSS软件	SPSS采用类似Excel表格的方式输入与管理数据，数据接口较为通用，能方便地从其他数据库中读入数据。突出的特点是操作界面友好，且输出结果美观。
Clementine软件	Clementine提供出色、广泛的数据挖掘技术，确保用恰当的分析技术来处理相应的商业问题，得到最优的结果以应对随时出现的问题。
RapidMiner软件	RapidMiner并不支持分析流程图方式，当包含的运算符比较多时就不容易查看；具有丰富的数据挖掘分析和算法功能，常用于解决各种商业关键问题。
其他数据挖掘软件	流行的数据挖掘软件还包括Orange、Knime、Keel与Tanagra等

分类算法

- 分类是一种重要的数据分析形式，根据重要数据类的特征向量值及其他约束条件，构造分类函数或分类模型（分类器），目的是根据数据集的特点把未知类别的样本映射到给定类别中。数据分类过程主要包括两个步骤，即学习和分类。

第一步，建立一个模型



姓名	年龄	收入	信用评估
张明	<=30	低	一般
王昊	<30	低	良好
李好	31…40	高	良好
赵其	>40	中等	一般
钱可	>40	中等	一般
马华	31…40	高	良好
……	……	……	……

第二步，使用模型进行分类

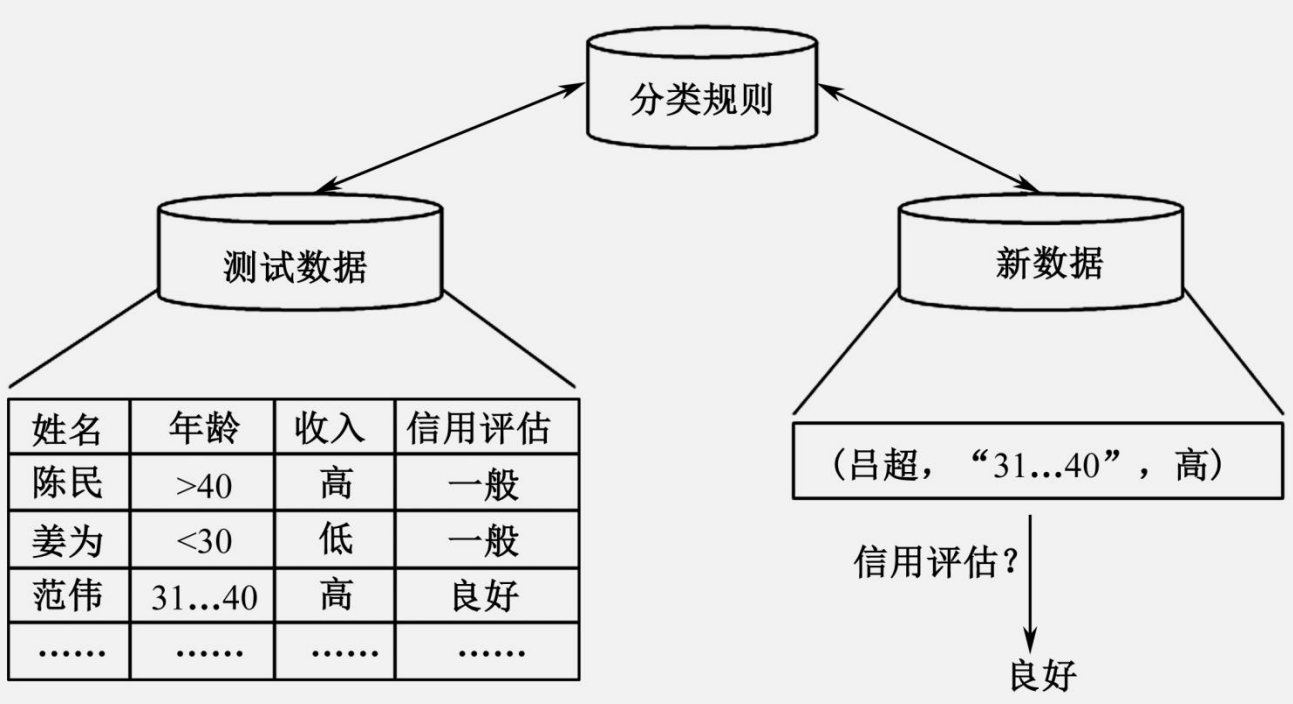


图3-2 使用模型进行分类

分类算法

分类又称为有监督的学习

- 分类分析在数据挖掘中是一项比较重要的任务，目前在商业上应用最多。
- 分类的目的是从历史数据记录中自动推导出对给定数据的推广描述，从而学会一个分类函数或分类模型（也常常称作分类器），该模型能把数据库中的数据项映射到给定类别中的某一个类中。
- 为建立模型而被分析的数据元组形成训练数据集，由一组数据库记录或元组构成，每个元组是一个由有关字段（又称属性或特征）值组成的特征向量，此外，每一个训练样本都有一个预先定义的类别标记，由一个被称为类标签的属性确定。

一个具体样本的形式可表示为 $\{X_1, \cdots, X_n, C\}$,其中 X_n 表示字段值，C 表示类别。

贝叶斯决策与分类器

数学基础知识

1．条件概率

事件A 在另外一个事件B 已经发生条件下的发生概率，称为在B 条件下A 的概率。表示为 $P(A|B)$

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

2．联合概率

联合概率表示两个事件共同发生的概率。 A 与B 的联合概率表示为 $P(AB)$、 $P(A,B)$或者 $P(A\cap B)$

3．贝叶斯定理

贝叶斯定理用来描述两个条件概率之间的关系，例如，$P(A|B)$ 与 $P(B|A)$。根据乘法法则 $P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$可以推导出贝叶斯公式：

$$P(A|B)=\frac{P(A)P(B|A)}{P(B)}$$

贝叶斯决策与分类器

4．全概率公式

全概率公式为概率论中的重要公式，它将对复杂事件A 的概率求解问题转化为在不同情况下发生的简单事件的概率的求和问题。

设 $B_1,\cdots,B_n$ 构成一个完备事件组，即它们两两互不相容，其和为全集，且 $P(B_i)\geqslant 0(i=1,\cdots,n)$，则事件A的概率为：$P(A)=P(A\mid B_1)P(B_1)+\cdots+P(A\mid B_n),P(B_n)=\sum_{i=1}^{n}P(A\mid B_i)P(B_i)$

贝叶斯分类的工作过程如下：

（1）每个数据样本均是由一个*n* 维特征向量 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$表示，分别描述其*n* 个属性 $A_1,A_2,\cdots,A_n$ 的具体取值。

■ 贝叶斯决策与分类器

4．全概率公式

（2）假设共有m个不同类别，$C_1,C_2,\cdots,C_m$。给定一个未知类别的数据样本X（没有类别号），分类器预测属于X后验概率最大的那个类别。也就是说，朴素贝叶斯分类器将未知类别的样本X归属到类别C_i，当且仅当$P(C_i \mid X) > P(C_j \mid X), 1 \leqslant j \leqslant m, j \neq i$

也就是$P(C_i \mid X)$最大。其中类别C_i就称为最大后验概率的假设。根据贝叶斯公式可得：

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (3\text{-}4)$$

（3）由于$P(X)$对于所有的类别均是相同的，因此，只需要$P(X \mid C_i)P(C_i)$取最大即可。由于类别的先验概率是未知的，则通常假定类别出现概率相同，即$P(C_1) = P(C_2) = \cdots = P(C_m)$。这样对于式（3-4）取最大转换成只需要求$P(X \mid C_i)$最大。而类别的先验概率一般可以通过$P(C_i) = \frac{s_i}{s}$公式进行估算，其中，$s_i$为训练样本集合中类别$C_i$的个数，$s$为整个训练样本集合的大小。

■ 贝叶斯决策与分类器

4．全概率公式

（4）根据所给定包含多个属性的数据集，直接计算 $P(X|C_i)$的运算量非常大。为实现对 $P(X|C_i)$的有效估算，朴素贝叶斯分类器通常都假设各类别是相互独立的，即各属性间不存在依赖关系，其取值是相互独立的。

$$P(X|C_i)=\prod_{k=1}^{n}p(x_k|C_i)$$

可以根据训练数据样本估算 $p(x_1|C_i),p(x_2|C_i),\cdots,p(x_n|C_i)$的值。
如果 A_k 是分类属性，则 $p(x_k|C_i)=\frac{s_{ik}}{s_i}$；其中 s_{ik}是在属性 A_k上具有值 x_k 的类 C_i的训练样本数，而 s_i 是 C_i中的训练样本数。
如果 A_k是连续值属性，则通常假定该属性服从高斯分布。因而

$$p(x_k|C_i)=g(x_k,\mu_{c_i},\sigma_{c_i})\frac{1}{\sqrt{2\pi}\sigma_{c_i}}e^{-\frac{(x-\mu_{c_i})^2}{2\sigma_{c_i}^2}} \quad (3\text{-}6)$$

给定类 C_i 的训练样本属性 A_k 的值，$g(x_k,\mu_{c_i},\sigma_{c_i})$是属性 A_k的高斯密度函数，μ_{c_i} σ_{c_i}分别为均值和方差。

（5）为预测一个未知样本X 的类别，可对每个类别 C_i 估算相应的 $P(X|C_i)P(C_i)$。样本X 归属类别C_i 当且仅当$P(X|C_i)P(C_i)>P(X|C_j)P(C_j)$，$1\leqslant j\leqslant m, j\neq i$即X 属于 $P(X|C_i)P(C_i)$为最大的类 C_i。

■ SVM算法

支持向量机（Support Vector Machine）是建立在统计学习理论的VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（对特定训练样本的学习精度，Accuracy）和学习能力（无错误地识别任意样本的能力）之间寻求最佳折中，以期获得最好的推广能力（或称泛化能力）。

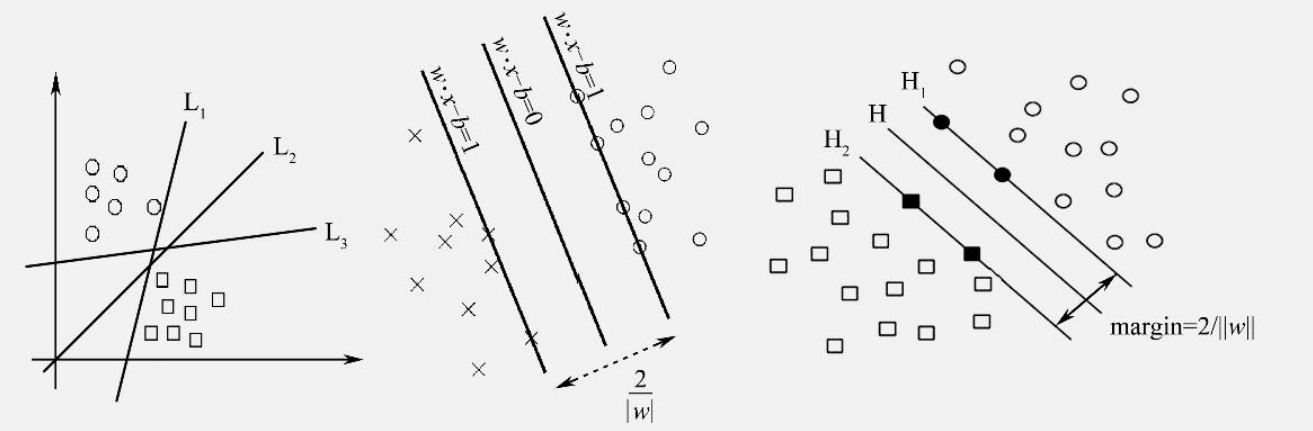


图3-3 超平面

SVM最基本的任务就是找到一个能够让两类数据都离超平面很远的超平面，在分开数据的超平面的两边建有两个互相平行的超平面。分隔超平面使两个平行超平面的距离最大化，平行超平面间的距离或差距越大，分类器的总误差越小。

通常希望分类的过程是一个机器学习的过程。设样本属于两个类，用该样本训练SVM得到的最大间隔超平面。在超平面上的样本点也称为支持向量。

■ SVM算法

线性可分情形SVM

非线性可分情形SVM

支持向量机（SVM）的核函数

■ 案例：在线广告推荐中的分类

互联网的出现和普及，带来的网上信息量的大幅增长，出现信息超载问题。为了解决信息过载的问题，提出了很多解决方案，其中最具有代表性的解决方案是分类目录和搜索引擎。

但是随着互联网规模的不断扩大，分类目录和搜索引擎，不能解决用户的需求。推荐系统就是解决这一矛盾的重要工具。

推荐系统具有用户需求驱动、主动服务和信息个性化程度高等优点，可有效解决信息过载问题。

推荐系统是一种智能个性化信息服务系统，可借助用户建模技术对用户的长期信息需求进行描述，并根据用户模型通过一定的智能推荐策略实现有针对性的个性化信息定制，能够依据用户的历史兴趣偏好，主动为用户提供符合其需求和兴趣的信息资源。

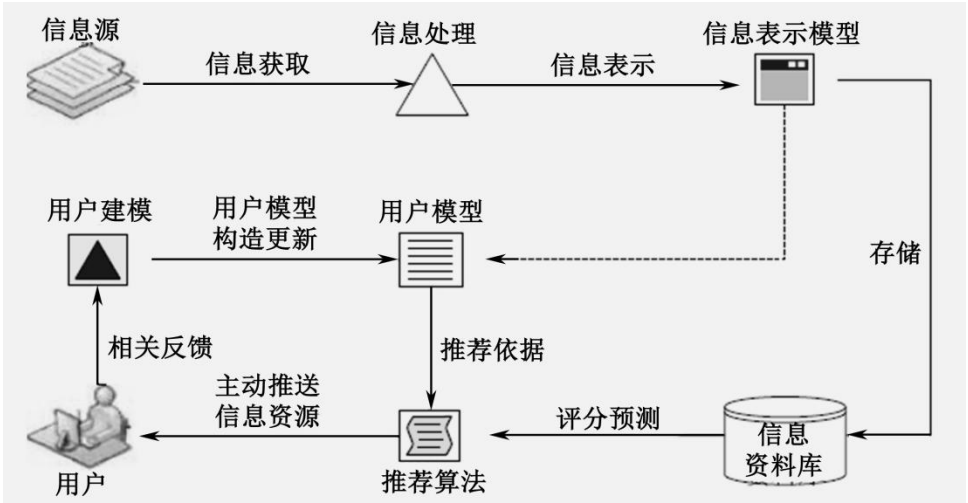


图3-6 推荐系统的工作原理

案例：在线广告推荐中的分类

推荐系统利用推荐算法将用户和物品联系起来，能够在信息过载的环境中帮助用户发现令他们感兴趣的信息，也能将信息推送给对他们感兴趣的用户。

根据已有用户注册信息和购买信息，使用朴素贝叶斯分类预测一个新注册用户购买计算机的可能性，从而向该用户推荐计算机类广告。训练样本如表3-1所示。

序号 ID	年龄 Age（岁）	收入等级 Income_level	是否学生 student	信用等级 Credit rate	类别：是否购买计算机 Class:buy computer
1	30以下	高	否	良	否
2	30以下	高	否	优	否
3	31到40	高	否	良	是
4	40以上	中	否	良	是
5	40以上	低	是	良	是
6	40以上	低	是	优	否
7	31到40	低	是	优	是
8	30以下	中	否	良	否
9	30以下	低	是	良	是
10	40以上	中	是	良	是
11	30以下	中	是	优	是
12	31到40	中	否	优	是
13	31到40	高	是	良	是
14	40以上	中	否	优	否

表3-1 训练课本

■ 非监督机器学习方法与聚类

聚类（clustering）就是将具体或抽象对象的集合分组成由相似对象组成的为多个类或簇的过程。由聚类生成的簇是一组数据对象的集合，簇必须同时满足以下两个条件：每个簇至少包含一个数据对象；每个数据对象必须属于且唯一地属于一个簇。

聚类分析是指用数学的方法来研究与处理给定对象的分类，主要是从数据集中寻找数据间的相似性，并以此对数据进行分类，使得同一个簇中的数据对象尽可能相似，不同簇中的数据对象尽可能相异，从而发现数据中隐含的、有用的信息。

聚类过程

聚类算法的要求

■ 常用聚类算法

1．层次聚类算法

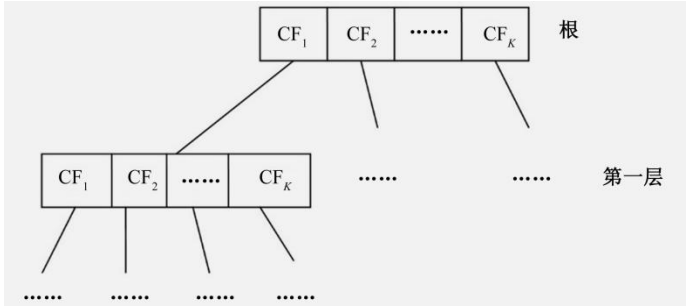
层次聚类算法的指导思想是对给定待聚类数据集合进行层次化分解。此算法又称为数据类算法，此算法根据一定的链接规则将数据以层次架构分裂或聚合，最终形成聚类结果。

从算法的选择上看，层次聚类分为自顶而下的分裂聚类和自下而上的聚合聚类。

分裂聚类初始将所有待聚类项看成同一类，然后找出其中与该类中其他项最不相似的类分裂出去形成两类。如此反复执行，直到所有项自成一类。

聚合聚类初始将所有待聚类项都视为独立的一类，通过连接规则，包括单连接、全连接、类间平均连接，以及采用欧氏距离作为相似度计算的算法，将相似度最高的两个类合并成一个类。如此反复执行，直到所有项并入同一个类。

典型代表算法，BIRCH（Balanced Iterative Reducing and Clustering Using Hierarchies，利用层次方法的平衡迭代规约和聚类）



常用聚类算法

2．划分聚类算法

划分法属于硬聚类，指导思想是将给定的数据集初始分裂为*K*个簇，每个簇至少包含一条数据记录，然后通过反复迭代至每个簇不再改变即得出聚类结果。

K-Means算法也称作K-平均值算法或者K均值算法，是一种得到广泛使用的聚类分析算法。

常用距离算法

1）欧氏距离

$$d\left(x_i,x_j\right)=\left|\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2\right|^{\frac{1}{2}}$$

2）曼哈顿距离

$$d\left(x_i,x_j\right)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$$

3）闵可夫斯基距离

$$d\left(x_i,x_j\right)=\left|\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^r\right|^{\frac{1}{r}}$$

4）切比雪夫距离

$$d\left(x_i,x_j\right)=\max_{k\in\{1,2,\cdots,p\}}\left\{\left|x_{ik}-x_{jk}\right|\right\}$$

■ 常用聚类算法

2．划分聚类算法

K-Means算法是解决聚类问题的一种经典算法，简单快速，对于处理大数据集，该算法是相对可伸缩的和高效的

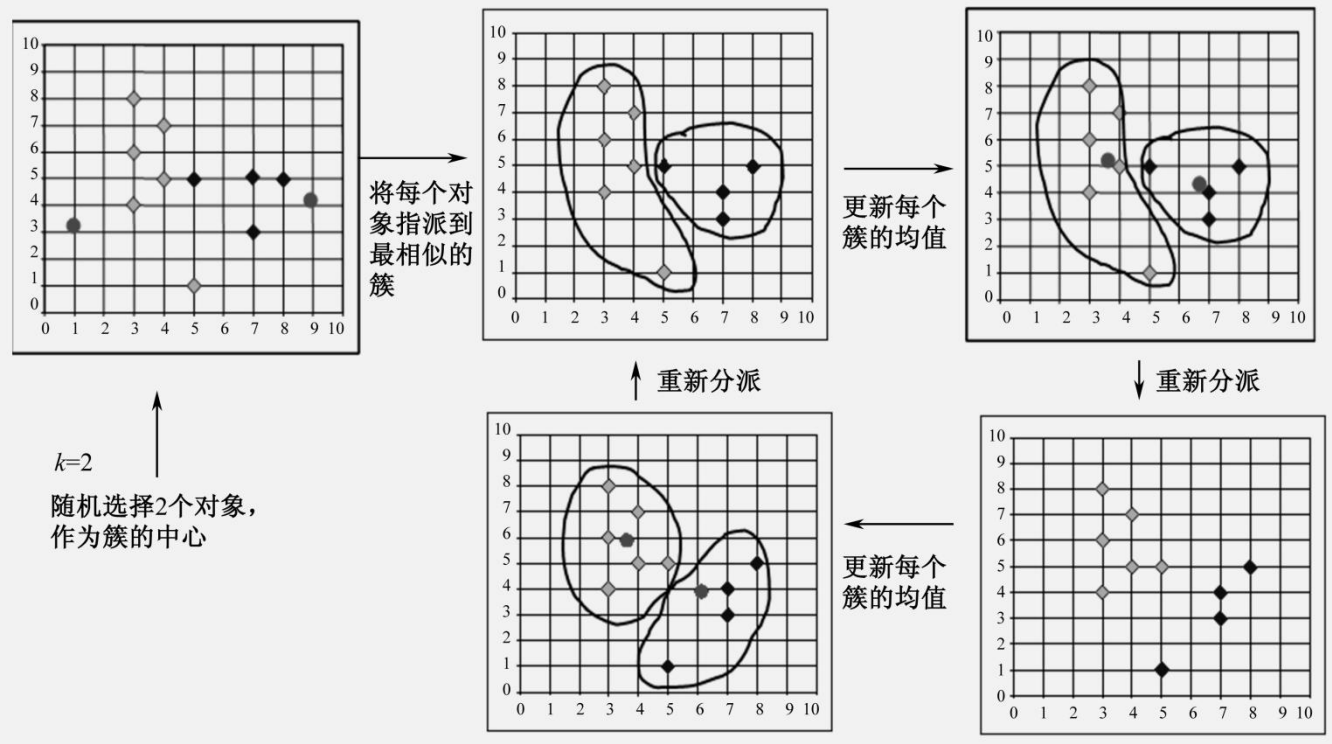


图3-8 K-Means算法流程

常用聚类算法

3．基于密度的聚类算法

基于密度聚类的经典算法DBSCAN（Density-Based Spatial Clustering of Application with Noise，具有噪声的基于密度的空间聚类应用）是一种基于高密度连接区域的密度聚类算法。

DBSCAN的基本算法流程如下：从任意对象*P*开始根据阈值和参数通过广度优先搜索提取从*P*密度可达的所有对象，得到一个聚类。若*P*是核心对象，则可以一次标记相应对象为当前类并以此为基础进行扩展。得到一个完整的聚类后，再选择一个新的对象重复上述过程。若*P*是边界对象，则将其标记为噪声并舍弃

缺陷

- 如聚类的结果与参数关系较大
  - 阈值过大容易将同一聚类分割
  - 阈值过小容易将不同聚类合并
- 固定的阈值参数对于稀疏程度不同的数据不具适应性
  - 密度小的区域同一聚类易被分割
  - 密度大的区域不同聚类易被合并

■ 常用聚类算法

4．基于网格的聚类算法

基于网格的聚类算法是采用一个多分辨率的网格数据结构，即将空间量化为有限数目的单元，这些单元形成了网格结构，所有的聚类操作都在网格上进行。

STING（STatistical INformation Grid，统计信息网格）算法将空间区域划分为矩形单元

针对不同级别的分辨率，通常存在多个级别的矩形单元，这些单元形成了一个层次结构——高层的每个单元被划分为多个低一层的单元

WaveCluster（Clustering using wavelet transformation，采用小波变换聚类）是一种多分辨率的聚类算法

先通过在数据空间上加一个多维网格结构来汇总数据，然后采用一种小波变换来变换原特征空间，在变换后的空间中找到密集区域

常用聚类算法

5．基于模型的聚类算法

基于模型的聚类算法是为每一个聚类假定了一个模型，寻找数据对给定模型的最佳拟合。

概念聚类是机器学习中的一种聚类方法，给出一组未标记的数据对象，它产生一个分类模式。概念聚类除了确定相似对象的分组外，还为每组对象发现了特征描述，即每组对象代表了一个概念或类。

概念聚类过程主要有两个步骤：首先，完成聚类；其次，进行特征描述。

统计学方法(EM和COBWEB算法)

神经网络方法将每个簇描述成一个模型。模型作为聚类的一个“原型”，不一定对应一个特定的数据实例或对象。

神经网络聚类的两种方法：竞争学习方法与自组织特征图映射方法。神经网络聚类方法存在较长处理时间和复杂数据中复杂关系问题，还不适合处理大数据库。

神经网络方法(SOM算法)

■ 案例：海量视频检索中的聚类

图像分割是图像处理到图像分析的关键步骤，也是一种基本的计算机视觉技术，一般来说，图像分割是把图像分成每个区域并提取感兴趣目标的技术和过程。颜色、灰度、纹理是比较常见和主要的特性，目标可以对应多个区域，也可以对应单个区域，主要与实际应用和目标有关。

K-Means聚类算法简捷，具有很强的搜索能力，适合处理数据量大的应用场景，在数据挖掘和图像领域中得到了广泛的应用。

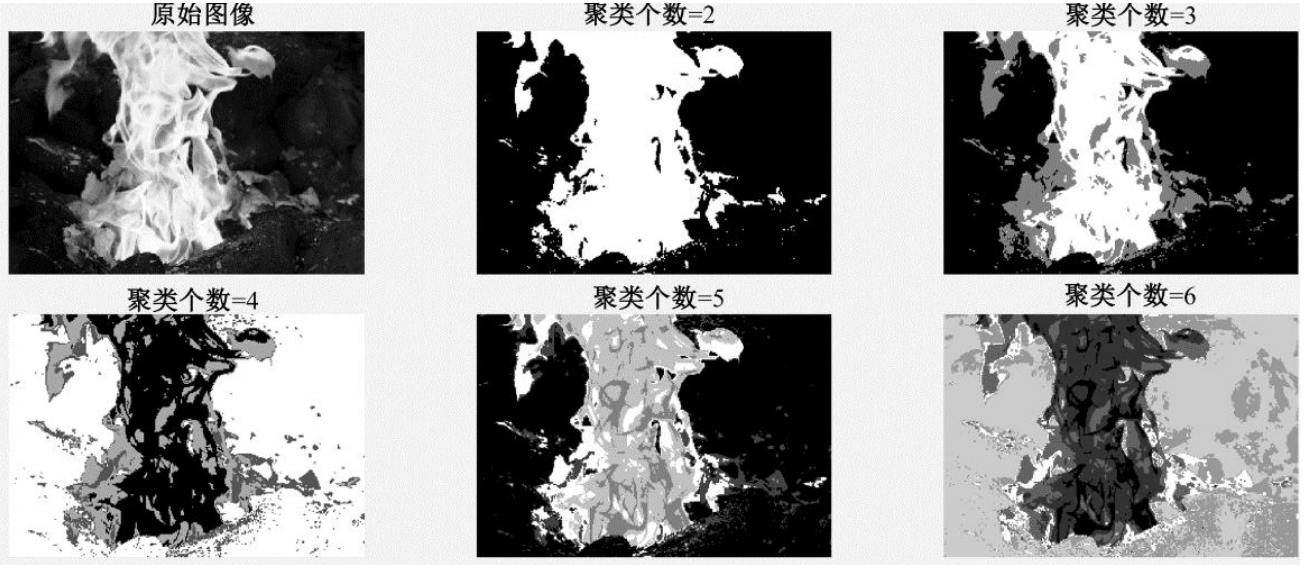


图3-9 K-Means聚类算法进行图像分割示意图

关联规则算法

关联规则的概念

关联规则是数据挖掘中最活跃的研究方法之一，是指搜索业务系统中的所有细节或事务，找出所有能把一组事件或数据项与另一组事件或数据项联系起来的规则，以获得存在于数据库中的不为人知的或不能确定的信息，它侧重于确定数据中不同领域之间的联系，也是在无指导学习系统中挖掘本地模式的最普通形式。

关联规则的概念

一般来说，关联规则挖掘是指从一个大型的数据集（Dataset）发现有趣的关联（Association）或相关关系（Correlation），即从数据集中识别出频繁出现的属性值集（Sets of Attribute Values），也称为频繁项集（Frequent Itemsets，频繁集），然后利用这些频繁项集创建描述关联关系的规则的过程。

关联规则挖掘问题：

发现频繁项集	生成关联规则
发现所有的频繁项集是形成关联规则的基础。通过用户给定的最小支持度，寻找所有支持度大于或等于Minsupport的频繁项集。	通过用户给定的最小可信度，在每个最大频繁项集中，寻找可信度不小于Minconfidence的关联规则。

如何迅速高效地发现所有频繁项集，是关联规则挖掘的核心问题，也是衡量关联规则挖掘算法效率的重要标准。

■ 频繁项集的产生及其经典算法

格结构（Lattice Structure）常常被用来枚举所有可能的项集。

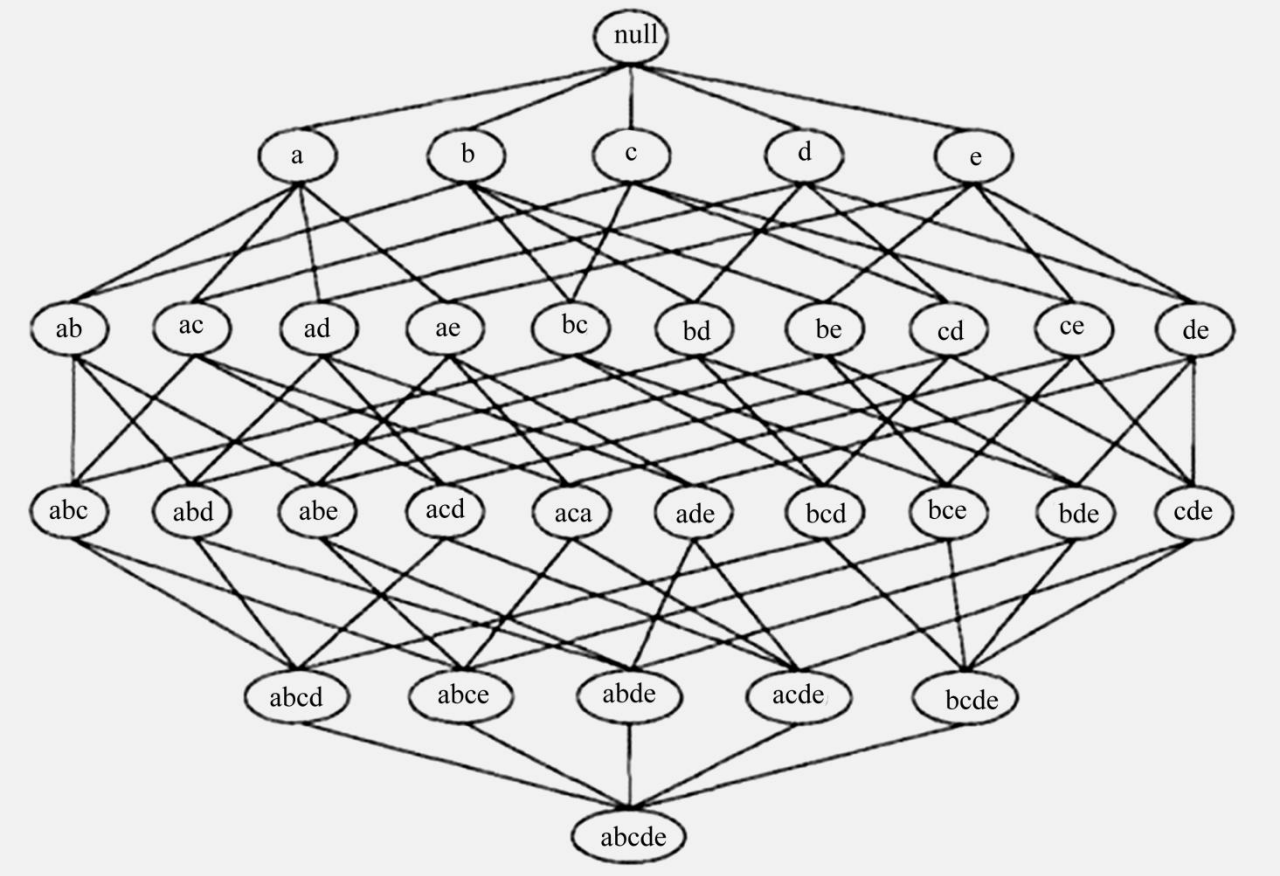


图3-10 项集的格

关联规则算法

■ 频繁项集的产生及其经典算法

格结构（Lattice Structure）常常被用来枚举所有可能的项集。

查找频繁项目集 —— 按照挖掘的策略不同：

- 经典的查找策略
- 基于精简集的查找策略
- 基于最大频繁项集的查找策略

经典的挖掘完全频繁项集方法

- 基于广度优先搜索策略的关联规则算法：Apriori算法、DHP算法
- 基于深度优先搜索策略的算法：FP-Growth算法、ECLAT算法 COFI算法

与经典查找不同方法

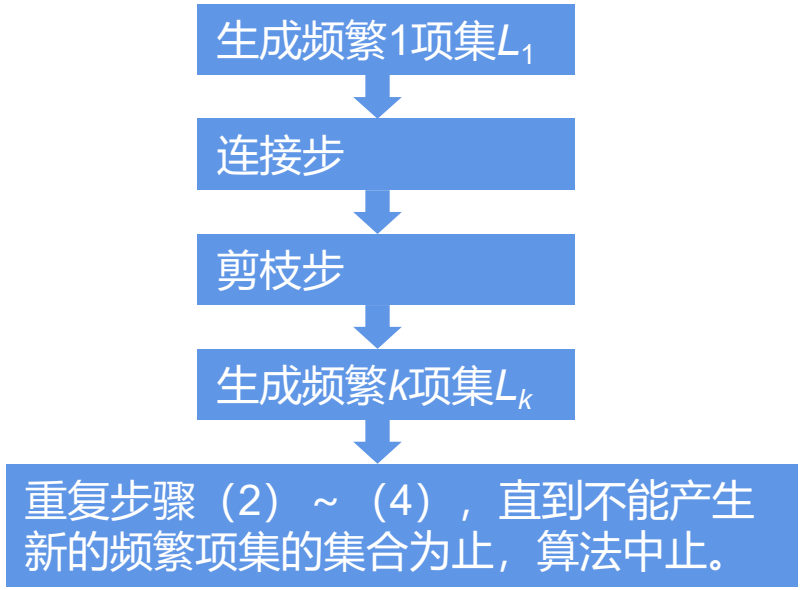
- 基于精简集的方法：A-close算法
- 基于最大频繁项目集的方法：MAFIA算法、GenMax算法 DepthProject算法

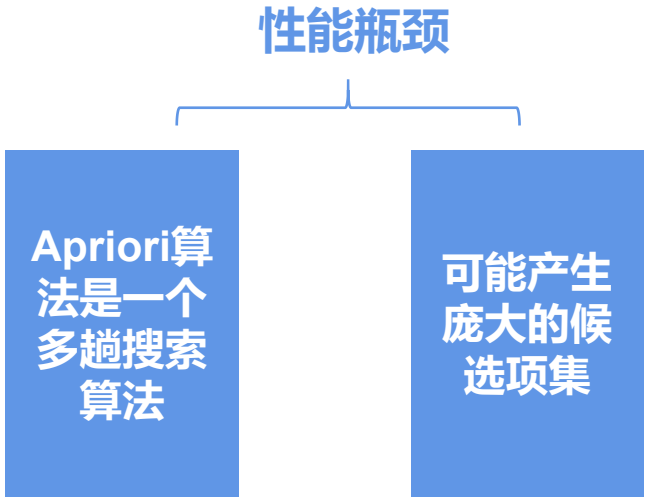
频繁项集的产生及其经典算法

1．Apriori算法

Apriori算法基于频繁项集性质的先验知识，使用由下至上逐层搜索的迭代方法，即从频繁1项集开始，采用频繁k项集搜索频繁k+1项集，直到不能找到包含更多项的频繁项集为止。

Apriori算法由以下步骤组成，其中的核心步骤是连接步和剪枝步：





■ 频繁项集的产生及其经典算法

2．FP-Growth算法

频繁模式树增长算法（Frequent Pattern Tree Growth）采用分而治之的基本思想，将数据库中的频繁项集压缩到一棵频繁模式树中，同时保持项集之间的关联关系。然后将这棵压缩后的频繁模式树分成一些条件子树，每个条件子树对应一个频繁项，从而获得频繁项集，最后进行关联规则挖掘。

FP-Growth算法由以下步骤组成：

1. 扫描事务数据库D，生成频繁1项集L_1
2. 将频繁1项集L_1按照支持度递减顺序排序，得到排序后的项集L_1
3. 构造FP树
4. 通过后缀模式与条件FP树产生的频繁模式连接实现模式增长

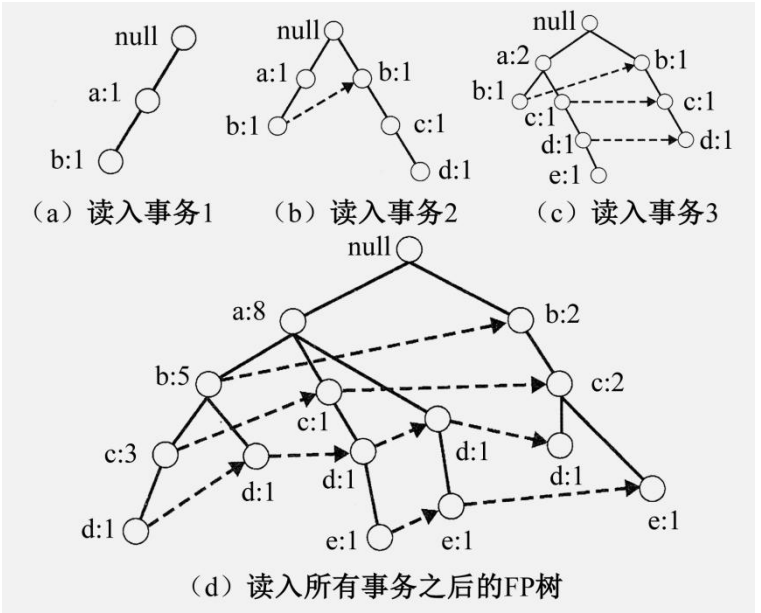


图3-11 FP树的构造

■ 频繁项集的产生及其经典算法

3．辛普森悖论

虽然关联规则挖掘可以发现项目之间的有趣关系，在某些情况下，隐藏的变量可能会导致观察到的一对变量之间的联系消失或逆转方向，这种现象就是所谓的辛普森悖论（Simpson’s Paradox）。

为了避免辛普森悖论的出现，就需要斟酌各个分组的权重，并以一定的系数去消除以分组数据基数差异所造成的影响。同时必须了解清楚情况，是否存在潜在因素，综合考虑。

■ 分类技术

分类技术或分类法（Classification）是一种根据输入样本集建立类别模型，并按照类别模型对未知样本类标号进行标记的方法。

1．决策树

决策树就是通过一系列规则对数据进行分类的过程。

决策树分类算法通常分为两个步骤：构造决策树和修剪决策树。

关联规则算法

分类技术

构造决策树

根据实际需求及所处理数据的特性，选择类别标识属性和决策树的决策属性集

↓

在决策属性集中选择最有分类标识能力的属性作为决策树的当前决策节点

↓

根据当前决策节点属性取值的不同，将训练样本数据集划分为若干子集

↓

针对上一步中得到的每一个子集，重复进行以上两个步骤，直到最后的子集符合约束的3个条件之一

① 子集中的所有元组都属于同一类。
② 该子集是已遍历了所有决策属性后得到的。
③ 子集中的所有剩余决策属性取值完全相同，已不能根据这些决策属性进一步划分子集。

↓

根据符合条件不同生成叶子节点

修剪决策树

对决策树进行修剪，除去不必要的分枝，同时也能使决策树得到简化。

常用的决策树修剪策略

- 基于代价复杂度的修剪
- 悲观修剪
- 最小描述长度修剪

按照修剪的先后顺序

- 先剪枝（Pre-pruning）
- 后剪枝（Post-pruning）

■ 分类技术

2．k-最近邻

最临近分类基于类比学习，是一种基于实例的学习，它使用具体的训练实例进行预测，而不必维护源自数据的抽象（或模型）。它采用*n* 维数值属性描述训练样本，每个样本代表*n* 维空间的一个点，即所有的训练样本都存放在*n* 维空间中。若给定一个未知样本，k-最近邻分类法搜索模式空间，计算该测试样本与训练集中其他样本的邻近度，找出最接近未知样本的*k* 个训练样本，这*k* 个训练样本就是未知样本的*k* 个“近邻”。其中的“邻近度”一般采用欧几里得距离定义：两个点 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$ 的Euclid距离是 $d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$。

最近邻分类是基于要求的或懒散的学习法，即它存放所有的训练样本，并且直到新的（未标记的）样本需要分类时才建立分类。其优点是可以生成任意形状的决策边界，能提供更加灵活的模型表示。

■ 案例：保险客户风险分析

1．挖掘目标

由过去大量的经验数据发现机动车辆事故率与驾驶者及所驾驶的车辆有着密切的关系，影响驾驶人员安全驾驶的主要因素有年龄、性别、驾龄、职业、婚姻状况、车辆车型、车辆用途、车龄等。因此，客户风险分析的挖掘目标就是上述各主要因素与客户风险之间的关系，等等。

2．数据预处理

数据准备与预处理是数据挖掘中的首要步骤，高质量的数据是获得高质量决策的先决条件。在实施数据挖掘之前，及时有效的数据预处理可以解决噪声问题和处理缺失的信息，将有助于提高数据挖掘的精度和性能。

数据清洗

去除数据集之中的噪声数据和无关数据，处理遗漏数据和清洗“脏”数据等。

数据清洗处理通常包括处理噪声数据、填补遗漏数据值/除去异常值、纠正数据不一致的问题，等等。

数据转化

在处理完噪声数据后，就可以对数据进行转化，主要的方法有：

- 聚集
- 忽略无关属性
- 连续型属性离散化等。

■ 案例：保险客户风险分析

3．关联规则挖掘

序号	关联规则	支持度	置信度
1	驾龄（X，A）∧被保车辆的价值（X，A） 年赔付金额（X，B）	0.1825	0.2965
2	投保人年龄（X，A）∧驾龄（X，A） 年赔付次数（X，B）	0.1679	0.2571
3	驾龄（X，B）∧车辆用途（X，A） 年赔付金额（X，B）	0.1663	0.3337
4	驾龄（X，B）∧车辆用途（X，B） 年赔付次数（X，A）	0.1789	0.4851
5	驾龄（X，B）∧被保车辆的价值（X，C） 年赔付金额（X，C）	0.1809	0.3003
6	驾龄（X，C）∧车辆用途（X，C） 年赔付次数（X，A）	0.1994	0.5864
7	驾龄（X，C）∧被保车辆的价值（X，C）∧车辆用途（X，C） 年赔付次数（X，A）	0.1031	0.6639
8	驾龄（X，A）∧被保车辆的价值（X，A）∧车辆用途（X，B） 年赔付金额（X，B）	0.1025	0.3654
9	投保人年龄（X，B）∧驾龄（X，A）∧被保车辆的价值（X，D） 年赔付金额（X，D）	0.0934	0.4546
10	驾龄（X，B）∧被保车辆的价值（X，A）∧车辆用途（X，A） 年赔付金额（X，B）	0.0968	0.4487
11	投保人年龄（X，C）∧被保车辆的价值（X，C）∧车辆用途（X，C） 年赔付金额（X，B）	0.0909	0.3531
12	投保人年龄（X，C）∧驾龄（X，B）∧被保车辆的价值（X，C） 年赔付次数（X，A）	0.0827	0.6094

详细分析所得数据，可以为公司业务提供数据支撑，针对不同客户提供偏好服务，既能确保公司收益，又能给予用户更多的实惠。

表3-7 客户风险关联规则

预测与预测模型

预测分析是一种统计或数据挖掘解决方案，包含可在结构化与非结构化数据中使用以确定未来结果的算法和技术，可为预测、优化、预报和模拟等许多其他相关用途而使用。

时间序列预测是一种历史资料延伸预测，以时间序列所能反映的社会经济现象的发展过程和规律性，进行引申外推预测发展趋势的方法。

时间序列预测及数据挖掘分类

- 依据研究的方式分类
  - 将时间序列数据作为一种特殊的挖掘对象，找寻对应的数据挖掘算法进行专门研究
  - 从时间序列数据中提取并组建特征，仍用原有的数据挖掘框架与算法进行数据挖掘
- 依据研究的内容分类
  - 时态模式挖掘
  - 相似性问题挖掘
- 依据研究的对象分类
  - 事件序列的数据挖掘
  - 事务序列的数据挖掘
  - 数值序列的数据挖掘

预测与预测模型

预测方案分类

依据预测方法的性质

- 定性预测方法
- 时间序列预测
- 因果关系预测

时间序列的统计特征

1）均值函数

$$\mu_t = E[X_t] \triangleq \int_{-\infty}^{+\infty} x f_t(x) dx$$

2）自协方差函数

$$\gamma_{t,s} = Cov(x_t, x_s) \triangleq E[(x_t - Ex_t)(x_s - Ex_s)]$$

3）自相关函数

$$\rho_{t,s} \triangleq \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$$

预测与预测模型

预测方案分类

时间序列模型

1）自回归模型

$$x_t = \varnothing_1 x_{t-1} + \varnothing_2 x_{t-2} + \cdots + \varnothing_p x_{t-p} + \varepsilon_i$$

2）移动平均模型

$$x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

3）自回归移动平均模型

$$x_t = \varnothing_1 x_{t-1} + \varnothing_2 x_{t-2} + \cdots + \varnothing_p x_{t-p} + \varepsilon_i + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

■ 时间序列预测

时间序列：对按时间顺序排列而成的观测值集合，进行数据的预测或预估。

典型的算法：序贯模式挖掘SPMGC算法

序贯模式挖掘算法SPMGC（Sequential Pattern Mining Based on General Constrains）SPMGC算法可以有效地发现有价值的数据序列模式，提供给大数据专家们进行各类时间序列的相似性与预测研究。

时间序列领域约束规则

■ 时间序列预测

SPMGC算法的基本处理流程

扫描时间序列数据库，获取满足约束条件且长度为1的序列模式L_1，以序列模式L_1作为初始种子集

↓

根据长度为i-1的种子集L_{i-1}，通过连接与剪切运算生成长度为i 并且满足约束条件的候选序列模式C_i，基于此扫描序列数据库，并计算每个候选序列模式C_i的支持数，从而产生长度为I 的序列模式L_i，将L_i作为新种子集

↓

在此重复上一步，直至没有新的候选序列模式或新的序列模式产生

SPBGC算法首先对约束条件按照优先级进行排序，然后依据约束条件产生候选序列。SPBGC算法说明了怎样使用约束条件来挖掘序贯模式，然而，由于应用领域的不同，具体的约束条件也不尽相同，同时产生频繁序列的过程也可采用其他序贯模式算法。

■ 案例：地震预警

1．地震波形数据存储和计算平台

南京云创大数据有限公司为山东省地震局研发了一套可以处理海量数据的高性能地震波形数据存储和计算平台，将从现有的光盘中导入地震波形数据并加以管理，以提供集中式的地震波形数据分析与地震预测功能，为开展各种地震波形数据应用提供海量数据存储管理和计算服务能力。

图3-12山东省地震波测数据云平台的显示界面

■ 案例：地震预警

2．地震波形数据存储和计算平台的主要性能指标

数据存储和处理指标	每年的原始地震波形数据及相关辅助信息约为15TB，为保证数据存储的可靠性，要求采用3倍副本方式保存数据，云平台每年需要提供约45TB的总存储量，同时系统必须能实时接收和处理高达10MB/s的入库数据
系统响应时间指标	千兆网络环境下，局域网客户端从分布式文件存储系统中读取4096B存储内容的响应时间不高于50毫秒
地震波形数据存储性能指标	采用HDFS格式进行数据读取，读取性能为40～80MB/s节点，数据规模10PB，数据负载均衡时间可依据流量配置而确定，集群重新启动时间按10PB规模计算达到分钟级别

■ 案例：地震预警

3．地震波形数据存储和计算平台的功能设计

■ 案例：地震预警

4．平台的组成、总体构架与功能模块

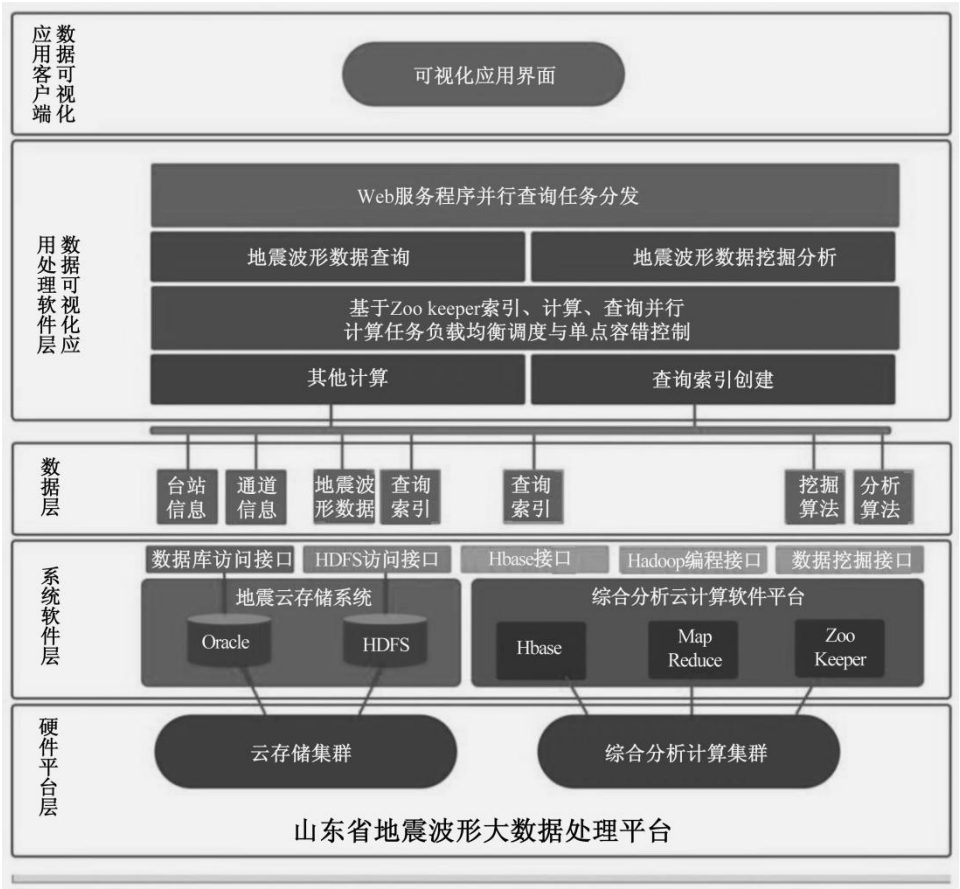


图3-13 地震波形数据云平台总体构架与功能模块

■ 案例：地震预警

5．地震中的时间序列预测

地震预测的主要手段也就是对地震序列进行特征研究。通过对地震序列的特征研究，可以帮助判断某大地震发生后地质活动的规律，掌握一定区域内地震前后震级次序间的某种内在关联性，有利于判断次地震发生后，震区地质活动的客观趋势

1）地震数据收集和预处理

采用SPBGC算法，预处理的流程步骤具体如下：

- 设定地震序列的空间跨度，并划分震级标准M
- 依据地震目录数据库，将震级大于或等于震级标准M的地震信息存入大地震文件
- 获取大地震文件中的每一条记录E，并取得震级M与震中所在位置G
- 扫描地震目录数据，对每一地震记录E，均判断当前地震位置与震中G的距离是否满足设定的空间跨度。如果满足空间跨度，则将该记录标注为与震中等同的序列号，同时将震中为圆心的区域范围内地震的次数加I；否则继续处理下一条地震记录
- 大地震文件处理完毕后，该阶段地震数据收集和预处理阶段结束

案例分析：精确营销中的关联规则应用

数据挖掘在各领域的应用非常广泛，只要该产业拥有具备分析价值与需求的数据仓储或数据库，都可以利用挖掘工具进行有目的的挖掘分析。一般较常见的应用案例多发生在零售业、制造业、财务金融保险、通信业及医疗服务等。

- 如何通过交叉销售，得到更大的收入？
- 如何在销售数据中发掘顾客的消费习性，并由交易记录找出顾客偏好的产品组合？
- 如何找出流失顾客的特征与推出新产品的时机点？

通过关联规则挖掘来发现和捕捉数据间隐藏的重要关联，从而为产品营销提供技术支撑。

数据挖掘算法综合应用

■ 挖掘目标的提出

电子商务网站中的商品推荐为例

客户忠诚度

影响因素

- 客户自身原因
- 企业原因
- 其他因素：如社会文化、国家政策等

数据挖掘技术可以建立客户忠诚度分析模型，了解哪些因素对客户的忠诚度有较大的影响，从而采取相应措施。因此，基于数据挖掘技术的客户忠诚度分析具有重要的应用价值。

■ 分析方法与过程

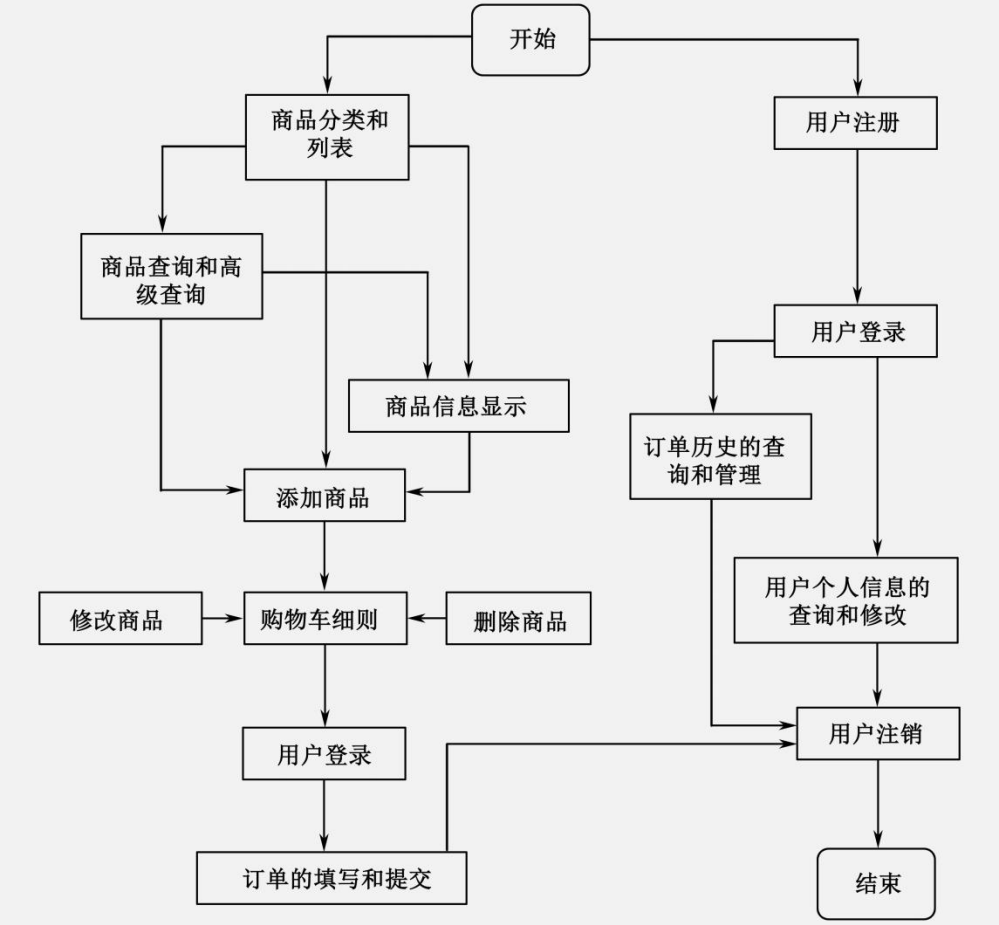


图3-14 电子商务网站操作流程

■ 分析方法与过程

在电子商务系统中，忠诚度分析所需要的客户信息和交易信息分别存放在网站数据库的客户表、订单表及订单明细表中。

将客户的忠诚度分为4个等级：0——忠诚；1——由忠诚变为不忠诚；2——由不忠诚变为忠诚；3——不忠诚。

客户编号	性别	年龄（岁）	教育程度	……	距最近一次购买时间（天）	月均购买频率	已消费金额	忠诚度级别
20120001	男	40	大专	……	5	3.4	801.6	0
20120002	女	28	本科	……	11	1.9	246.3	1
……	……	……	……	……	……	……	……	……

表3-9 经抽取而成的客户信息表

所得到的用户数据很难做到完整全面，用户在注册时可能选择不填注册信息的几项，造成数据项空缺。对于空缺的数据项，要视情况排除或填入默认值。
按照一般的统计划分经验来对属性值进行分段，实现离散化。

■ 分析方法与过程

客户编号	性别	年龄（岁）	教育程度	……	距最近一次购买时间（天）	月均购买频率	已消费金额（元）	忠诚度级别
20120001	男	30～40	大专	……	0～10	2～4	800～1000	0
20120002	女	20～30	本科	……	10～20	0～2	0～500	1
……	……	……	……	……	……	……	……	……

表3-10 经离散变换后的客户信息表

本案例采用基于信息论的ID3决策树分类算法进行客户忠诚度分析。

客户群细分使得公司可以更好地识别不同的客户群体，区别对待不同客户，采取不同的客户策略，达到最优化配置客户资源的目的。

使用聚类算法进行客户群，数据项处理过程主要将这些表内反映客户身份背景、购买兴趣度等相关信息提取出来，并加以清理，除去噪声数据，对信息不完全的数据填入默认值或舍去，进行必要的离散化变换。

■ 分析方法与过程

客户编号	性别	年龄（岁）	教育程度	类别1购买量	类别2购买量	……	类别49购买量
20120001	男	30～40	大专	0	17	……	61
20120002	女	20～30	本科	23	1	……	0

表3-11 客户兴趣度表

商品推荐是电子商务网站用来向访问网站的顾客提供商品信息和建议，并模拟销售人员帮助顾客完成购买过程。它是利用数据挖掘技术在电子商务网站中来帮助顾客访问有兴趣的产品信息。推荐可以是根据其他客户的信息或此客户的信息，参照该顾客以往的购买行为预测未来的购买行为，帮助用户从庞大的商品目录中挑选真正适合自己需要的商品。推荐技术在帮助了客户的同时也提高了顾客对网站的满意度，换来对商务网站的进一步支持。

总结

1. 数据挖掘概述
2. 分类算法
3. 聚类算法
4. 关联规则算法
5. 预测规模
6. 数据挖掘算法综合应用

谢谢聆听

Email： fbx2020@163.com
WebSite： http://www.coohelp.vip/