【40】

上节视频介绍了数据预处理的数据集成，这节视频介绍预处理的第三个任务，数据规约。

【41】

【41.1】首先要清楚为什么进行数据规约，在现实场景中，数据集是很庞大的，数据是海量的，在整个数据集上进行复杂的数据分析和挖掘需要花费很长的时间。

【41.2】数据规约的目的就是为了**压缩数据量**，帮助从原有庞大数据集中获得一个**精简**的数据集合，并使这一精简数据集**保持原有数据集的完整性**，这样在精简数据集上进行数据挖掘显然效率更高，并且挖掘出来的结果与使用原有数据集所获得结果是**基本相同**。

【41.3】如何进行数据规约，它是有标准的，首先用于规约的时间不应当超过或抵消在规约后的数据上挖掘节省的时间。再者就是规约得到的数据比原数据小得多，但可以产生相同或几乎相同的分析结果。

【42】

数据规约的策略有维归约、数量归约、数据压缩，每种策略下都有一系列相关的处理方法。

【43】

【43.1】我们首先看下维归约策略的第一个方法：主成分分析。

**主成分分析**也称[主分量分析](https://baike.baidu.com/item/%E4%B8%BB%E5%88%86%E9%87%8F%E5%88%86%E6%9E%90/10553278)，旨在利用降维的思想，把多指标转化为少数几个[综合指标](https://baike.baidu.com/item/%E7%BB%BC%E5%90%88%E6%8C%87%E6%A0%87/2849120)，这些综合指标也就是主成分，其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。

【43.2】在这张图中，销售毛利润、销售利润率、资产报酬率，转换为盈利能力的指标。在这张图标中，盈利能力、偿债能力、发展能力、营运能力就是主成分。

【43.3】我们可以看出，这种方法在在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效的数据信息。总而言之，一句话就是简化了原始变量的分析，提高了分析效率。

【44】

【44.1】维归约策略的第二个方法是属性子集选择，主要用于检测并删除不相关、弱相关或冗余的属性。目的是找出最小属性集，使得数据类的概率分布尽可能地接近使用所有属性得到的原分布

【44.2】用于分析的数据集可能包含数以百计的属性，其中大部分属性可能与挖掘任务不相关，或者是冗余的。例如如果分析任务是按顾客听到广告后是否愿意购买新的流行CD时，顾客的分类肯定是与**属性年龄、喜好的音乐类型**相关的，但诸如顾客的**电话号码，住址**等属性多半是不相关的。

【45】

【45.1】接下来，我们来看下数量归约策略。所谓数据归约，是指通过选择替代的、较小的数据表示形式来减少数据量。

【45.2】数值规约也可细分为两种方法，一种是有参数的方法，有参方法一般使用回归和对数的线性模型存储参数。

【45.3】另外一种就是无参数的方法。无参方法一般有直方图、聚类、抽样和数据立方体聚集等方法。

【46】

我们首先看下有参数的方法，它就是使用一个参数模型来评估数据。该方法只需要存储参数，而不是实际数据，能大大减少数据量，但只对数值型数据有效。比如两个属性X和Y支持线性回归函数Y=α+βX，则XY的所有数据可以用这个函数模型来表示，则在实际存储中只需要将参数α，β存储起来即可。

【47】

数量归约的另一种方法是无参的方法，比如用直方图来压缩数据。

直方图是根据属性的数据分布将其分成若干不相交的区间，使用每个区间的高度与其出现的频率成正比，通过一个例子来看下如何使用直方图来压缩数据。有这样一组某商店销售产品的单价数据（显示这组数据），根据单价的取值和出现频率（对应价格清单），绘制了一幅直方图，那么这组数据就可以使用压缩后的直方图来表示。

【48】

聚类属于数量归约中的无参方法的一种子方法，它将原数据集划分成多个群或聚类。一个好的划分的原则是：同类中的数据彼此相似；不同类中的数据彼此不相似。一般上，可通过空间距离来度量相似度，当然，也可以通过别的方法来度量相似度，比如划分法，层次法，密度算法等。聚类的有效性依赖于实际数据的内在规律，也就是不同的数据集具备有不同的特征，由不同的特征决定选取不同的算法来进行聚类。

【49】

比如，在这种居民的分布图中，通过空间距离的聚类，就可以发现有三个小区，有若干人居住在小区周边，这就是孤立点。

【50】

【50.1】同样的，取样也属于数量归约中的无参方法的一种子方法，它允许用数据的较小随机样本表示大的数据集。

【50.2】取样的方法有以下四种，不放回简单随机取样，放回简单随机取样，聚类取样，分层取样。接下来，我们将一一道来。

【51】

如图所示，不放回简单随机取样，就是随机从数据池中取出数据后，不再将抽出放回资源池。一般常见的应用场景是计算利用摇奖机进行摇奖的概率。

而放回简单随机取样，是从数据池中取出数据后，需要将抽出的数据放回资源池的。常见的应用场景是计算丢骰子时，每个数字出现的概率。

【52】

取样的第三种方式是聚类取样，如图所示，数据集D每100条记录被分成一个聚类，然后再从100条记录中抽取4条记录组成数据子集。

这个取样过程，进行严谨的描述就是：首先将大数据集D划分为M个互不相交的聚类，然后再从M个类中的数据对象分别进行随机抽取，可最终获得聚类采样的数据子集。

【53】

聚类取样常用的场景如图所示，例如我们需要调查人们对产品的建议，由于客户很多，不可能遍历调查每一个客户。为了提升效率，我们先按空间距离方法聚类，得出我们的客户分布在3个小区内，对应图的蓝黑白点。于是，我们在每个小区抽取部分客户调查。这些客户的建议是具有空间区域代表性的。

【54】

取样的第四种方式是分成取样，它是首先将大数据集D划分为互不相交的层，然后对每一层简单随机选样得到D的分层选样。

【55】

不同于的聚类取样的是，我们调查小区客户的对产品的满意度时，如果每个小区的每个客户属性是不一样的，那么就存在很大的偏差。一种比较稳妥的方法就是将客户分层，然后再去调查。

例如根据顾客的年龄组进行分层，然后再在每个年龄组中进行随机选样，从而确保了最终获得分层采样数据子集中的年龄分布具有代表性。这种方法就是分层取样。

【56】

数据立方体聚集也是数量归约策略中的一种无参数方法。数据立方体是指数据的多维建模和表示，由**维度、维度成员和度量值**组成。**维度**是指观察数据的角度，比如图中，可以从**产品类型、时间、地区**三个维度来查看公司产品的销量情况；维度成员指每个维度的取值，比如时间维度它具有一季度，二季度，三季度这三个成员，而地区维度成员有浙江、上海、江苏等地区组成。度量值指具有实际意义的值，比如图中**2300**是**2010年第一季度**，**电子产品**在**浙江**地区的**销售量**

数据立方体聚集是指将n维数据立方体聚集成n-1维的数据立方体的过程。

【57】

比如一家电子公司，它拥有**家庭娱乐、电脑、手机和安全**四种产品，分别在A、B、C、D四个地区有销售点，如图所示的数据立方体（左图）收集了**2002到2004**年期间的销售数据，当公司生产经理想看下各产品在这几年的销量趋势，这时他就不关注各地区的销量，于是可以从**地区**维度做聚集，最终得到表格所示的数据。这个数据可以帮助生产经理去决策下一年每个产品的生产计划。

【58】

数据归约的第三种策略是数据压缩。它利用数据编码或数据转换将原来的数据集合压缩为一个较小规模的数据集合。

数据压缩包括两种方式，一个是无损压缩，指数据经过压缩后，信息不受损失，还能完全恢复到压缩前的原样。比如字符串压缩，常用到的无损压缩的方法有zip或RAR。

另一个是有损压缩，它只能重新构造原数据的近似表示。即将次要的信息数据压缩掉，牺牲一些质量来减少数据量，使压缩比提高。比如音频/视频的压缩。音频能够在没有察觉的质量下降情况下实现 10:1 的压缩比，视频能够在稍微观察质量下降的情况下实现如 300:1 这样非常大的压缩比。

【59】

本节视频介绍了数据规约的各种策略和各种方法，如图所示，数据规约是将原有庞大数据集精简为一个小的数据集，并且这一精简数据集**保持由原来数据集的完整性**，使得数据挖掘的效率更高。