【17】

上节视频已经介绍了影响大数据质量的各种因素和数据预处理的目的。

我们知道了为什么存在这些问题和处理这些问题的一个理想预期。那么，我们现在就开始着手该如何去处理这些数据质量问题。

怎么处理这些数据质量问题？一般来说有4个流程步骤，分别为数据清理，集成，归约，变换。当然，这些流程步骤中，都有相关的问题描述和应对措施。

接下来，本节视频将介绍数据预处理的第一个流程步骤：数据清理。

【18】

在进入数据清理前，我们首先来看来下数据预处理4大流程步骤中的一个大体流向图，就是数据清理，数据集成，数据变换，数据归约。

整个过程是这样描述的：在原始数据的基础上，我们对有质量问题的数据进行数据清理，然后对干净的数据进行数据的集成，因为在实际的数据分析应用中，数据源是分布式的，存储于不同地方和不同类型的数据库当中，所以在分析之前要将所有数据源进行集成。

集成后的数据量是庞大的，同时这些数据不一定满足数据挖掘的要求，因此要通过数据变换、和数据规约，将数据转为是和适合挖掘的形式。提高挖掘过程的准确性和效率。

这个流程之所以成为大体处理流程，那是说，数据预处理的流程不需要严格按照如图的流程顺序处理，是按实际的工作需要，用相关的几个流程步骤完成数据的处理，甚至，有的步骤是需要反复的做的。比如数据集成后，如果存在脏数据，那么，还需要重新进行数据清理。

【19】

【19.1】首先来看预处理的第一个流程步骤：数据清理。

在数据科学领域，由于操作员重复录入，并发处理等不规范的操作，导致产生不完整，不准确的，无效的数据。

【19.2】这些数据，就是脏数据，如同垃圾一样，不仅没有价值，还会给数据分析带来污染，需要耗费时间精力去清洗，这类脏数据是数据分析应用中最大的障碍。数据有问题，苦心构建的数据库和挖掘算法就失去了价值。正因如此，处理“脏数据”的工作不仅十分必要，而且越早越好。

【19.3】一般在数据录入的时候就会对数据格式等进行约束，在数据库系统设计阶段，会对属性的类型和范围进行设定，比如学生信息中对性别的取值只有男女，对于年龄不可能出现180这样的数值（对应表格红色显示对应部分），有了这些约束条件，可以排除人为自由发挥产生的无效数据和机器故障产生的错误数据。

【20】

尽管有前期对数据类型和范围的约束，但还是避免不了脏数据的存在，比如有些我们需要的值它就是空的，并且数据表中允许这些属性为空。这就是需要对缺失值进行填充（对应表）。

数据清理试图填充空缺的值、识别孤立点、消除噪声，并纠正数据中的不一致性，所对应的两项处理操作，就是缺失值的填充和噪声数据的平滑。

【21】

【21.1】首先来看下空缺值的处理，缺失值是指现有数据集中某个或某些属性的值时不完整的。

【21.2】例如，在分析某公司的客户数据时，你会注意到许多元组的一些属性存在空缺值，例如客户信息表中客户的收入没有记录值，那么怎样才能为该属性填上缺失的值？

【21.3】处理空缺值的基本方法有6种，包括忽略元组、人工填写空缺值、使用全局常量替换空缺值、使用属性的中心度量填充空缺值、使用与给定元组属同一类的所有样本的平均值来填充空缺值、使用最可能的值填充空缺值。

【22】

【22.1】在讲这些处理方法之前我们先来强调几个概念，元组，属性，以及中心度量值。

【22.2】在关系型数据库中，数据库是由许多表组成的，而每张表又是由许多条记录组成，比如客户信息表中每一行是一条客户记录（显示绿色这一行），同时也称为元组，而每一列则为属性，也可称为字段（黄色列）。

【22.3】中心度量是指数据集的平均值或中位数。平均值顾名思义就是数据集的总和除以数据集的个数。中位数是指将一组数据按大小依次排列，把处在最中间位置的一个数（或最中间位置的两个数的平均数）叫做这组数据的中位数。

比如客户信息表中，除去空缺指，所有客户的收入按照从小到大排序，中间的12000是这个数据集的中位数。

【23】

【23.1】再来看下处理方法:

1，首先是忽略元组， 若一条记录中有多个属性值被遗漏了，则将该记录排除在数据挖掘之外；但是，当某类属性的空缺值占整张表的比例很大时，直接忽略元组会使挖掘性能变得非常差。（黑色一行删除）

【23.2】2，第二种是人工填写空缺值的方法

一般来说，该方法很费时，并且当数据集很大，缺失很多值时，该方法可能行不通。

【23.3】3，第三种是采用全局常量替换空缺值，如使用unknown或无穷大来填充。该方法简单，但是并不十分可靠。如果空缺值都用“Unknown”替换，当空缺值较多时，挖掘程序可能误以为它们形成了一个有趣的概念，因为它们都具有相同的值——“Unknown”。

【24】

【24.1】然后比较人性化的方法是使用属性的中心度量填充空缺值，对于对称分布的数据可以采用均值的方法，而倾斜的数据分布采用中位数。比如所有顾客的平均收入为$15200，则使用该值替换收入中的空缺值（对应只有收入一列的表）。

【24.2】使用与给定元组属同一类的所有样本的平均值，这种方法适用于分类数据挖掘，

如将顾客按信用度分类，则用具有相同信用风险的顾客的平均收入替换收入中的空缺值。比如客户信息表中，与0005号客户风险等级相同的是0002和0003，那么使用这两个用户的平均值来替代0005客户的收入。

【24.3】最后一种方法也是在数据挖掘中用的最多的方法，利用回归、贝叶斯计算公式或判定树归纳确定，推断出该条记录特定属性最大可能的取值，比如利用数据集中其他顾客的属性，构造一棵判定树，预测收入的空缺值。

【25】

数据清理需要处理的第二种类型数据是噪声数据，

【25.1】噪声数据是指一个测量变量中的随机错误或偏差，它包括错误的值或偏离期望的数据。噪声数据是无意义的数据，在真实世界中也是永远存在的，它会影响数据挖掘分析的结果。比如在做各种生物实验时，会由于仪器的脏污或环境温度影响等，都会对得到实验结果造成偏差。

【25.2】这里有一类比较特殊的值，叫孤立点，孤立点的定义是不符合数据模型的数据。它通常是真实存在的且准确的，但是与正常数据偏差比较大，这一类数据的用途是做异常值检测，比如在欺诈和网络攻击检测中，目标就是从大量正常对象或事件（大数据）中发现不正常的对象和事件。

【25.3】如果不是做特殊用途，孤立点也属于噪声数据的一种，在数据分析过程中尽量要消除噪声数据，降低对最终结果的影响。

【25.4】引起噪声数据的原因可能有数据收集工具的问题、数据输入错误、数据传输错误、技术的限制或命名规则不一致。

【25.5】针对这些原因，通常采用分箱法、回归、聚类等数据平滑方法来消除噪声数据。

【26】

【26.1】第一种噪声处理的方法是分箱，分箱法是通过考察数据的“近邻”即周围的值来光滑有序数据值。这些有序的值被分布到一些“桶”或箱中。由于分箱方法考察近邻的值，因此它进行的是局部的光滑。

【26.2】举个例子说明，有这样一组数据8、24、15、41、6、10、18、67、25，对它使用分箱的方法进行数据的平滑。

首先对数据按大小进行排序

然后将它们分到等深的箱中；比如这里将9个数据分到箱深为3的三个箱中。在这里箱深也就是箱子里数据的个数，这里为3。

【26.3】然后可以按箱的平均值、按箱中值或者按箱的边界等方法分别对三个箱中的数据进行平滑。

【26.4】按箱的平均值平滑：箱中每一个值被箱中的平均值替换

【26.5】按箱的中位数平滑：箱中的每一个值被箱中位数来替换

【26.6】按箱的边界平滑：箱中的最大和最小值被视为箱边界，箱中的每一个值被最近的边界值替换。

【27】

【27.1】第二种方法是回归方法，它采用一个函数拟合数据来光滑数据。

【27.2】其中线性回归涉及找出拟合两个属性（或变量）的最佳直线，使得一个属性能够预测另一个，比如将表格（最左表格）中的人均GDP和城市化水平做线性回归，拟合出中间图中的一条直线，通过这条直线，可以预测出具有更高的人均GDP的地区它的城市化水平是多少（中间图）。

【27.3】多元线性回归是线性回归的扩展，它涉及多个属性，并且数据拟合到一个多维面，如（最右边的）图所示。

使用回归方法，找出适合数据的数学方程式，能够帮助消除噪声。

【28】

【28.1】第三种噪声处理的方法是通过各种聚类算法找出数据集中的离群点并删除过程。

【28.2】聚类操作会将类似的值组织成群或簇，落在簇集合之外的点视为离群点，一般这种离群点是异常数据，最终会影响整体数据的分析结果，因此对离群点处理操作是删除。

比如图中对所有的点集进行聚类分析，形成了c1和c2两个簇，而o1，o2，o3是离群点。

【28.3】聚类的算法有多种，包括K均值法，层次聚类方法和基于密度的聚类法等。

【29】

本节视频介绍了数据清理主要完成的任务，包括缺失值的处理和噪声数据的处理，对于缺失值我们提供了六种不同的方法（显示图中的六个方法）来进行填充。对于噪声数据则使用到了分箱，回归和聚类的技术。下节视频将介绍数据集成。