作者：宋天龙  
链接：https://www.zhihu.com/question/22077960/answer/223886314  
来源：知乎  
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

数据清洗的主要包括：纠正错误、删除重复项、统一规格、修正逻辑、转换构造、数据压缩、补足残缺/空值、丢弃数据/变量。

**1. 纠正错误**

错误数据是数据源环境中经常出现的一类问题。数据错误的形式包括：

q 数据值错误：数据直接是错误的，例如超过固定域集、超过极值、拼写错误、属性错误、源错误等。

q 数据类型错误：数据的存储类型不符合实际情况，如日期类型的以数值型存储，时间戳存为字符串等。

q 数据编码错误：数据存储的编码错误，例如将UTF-8写成UTF-80。

q 数据格式错误：数据的存储格式问题，如半角全角字符、中英文字符等。

q 数据异常错误：如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期越界、数据前后有不可见字符等。

q 依赖冲突：某些数据字段间存储依赖关系，例如城市与邮政编码应该满足对应关系，但可能存在二者不匹配的问题。

q 多值错误：大多数情况下，每个字段存储的是单个值，但也存在一个字段存储多个值的情况，其中有些可能是不符合实际业务规则的。

这类错误产生的原因是业务系统不够健全，尤其是在数据产生之初的校验和入库规则不规范，导致在接收输入后没有进行判断或无法检测而直接写入后台数据库造成的。

**2. 删除重复项**

由于各种原因，数据中可能存在重复记录或重复字段（列），对于这些重复项目（行和列）需要做去重处理。

对于重复项的判断，基本思想是“排序和合并”，先将数据库中的记录按一定规则排序，然后通过比较邻近记录是否相似来检测记录是否重复。这里面其实包含了两个操作，一是排序，二是计算相似度。

常见的排序算法：

q 插入排序

q 冒泡排序

q 选择排序

q 快速排序

q 堆排序

q 归并排序

q 基数排序

q 希尔排序

常见的判断相似度的算法：

q 基本的字段匹配算法

q 标准化欧氏距离

q 汉明距离

q 夹角余弦

q 杰卡德距离

q 马氏距离

q 曼哈顿距离

q 闵可夫斯基距离

q 欧氏距离

q 切比雪夫距离

q 相关系数

q 信息熵

对于重复的数据项，尽量需要经过业务确认并进行整理提取出规则。在清洗转换阶段，对于重复数据项尽量不要轻易做出删除决策，尤其不能将重要的或有业务意义的数据过滤掉，校验和重复确认的工作必不可少。

**3. 统一规格**

由于数据源系统分散在各个业务线，不同业务线对于数据的要求、理解和规格不同，导致对于同一数据对象描述规格完全不同，因此在清洗过程中需要统一数据规格并将一致性的内容抽象出来。

数据字段的规则大致可以从以下几个方面进行统一：

q 名称，对于同一个数据对象的名称首先应该是一致的。例如对于访问深度这个字段，可能的名称包括访问深度、人均页面浏览量、每访问PV数。

q 类型：同一个数据对象的数据类型必须统一，且表示方法一致。例如普通日期的类型和时间戳的类型需要区分。

q 单位：对于数值型字段，单位需要统一。例如万、十万、百万等单位度量。

q 格式：在同一类型下，不同的表示格式也会产生差异。例如日期中的长日期、短日期、英文、中文、年月日制式和缩写等格式均不一样。

q 长度：同一字段长度必须一致。

q 小数位数：小数位数对于数值型字段尤为重要，尤其当数据量累积较大时会因为位数的不同而产生巨大偏差。

q 计数方法：对于数值型等的千分位、科学计数法等的计数方法的统一。

q 缩写规则：对于常用字段的缩写，例如单位、姓名、日期、月份等的统一。例如将周一表示为Monday还是Mon还是M。

q 值域：对于离散型和连续型的变量都应该根据业务规则进行统一的值域约束。

q 约束：是否允许控制、唯一性、外键约束、主键等的统一。

统一数据规格的过程中，需要重要的一点是确认不同业务线带来数据的规格一致性，这需要业务部门的参与、讨论和确认，以明确不同体系数据的统一标准。

**4. 修正逻辑**

在多数据源的环境下，很可能存在数据异常或冲突的问题。

例如不同的数据源对于订单数量的数据统计冲突问题，结果出现矛盾的记录。通常，这是由于不同系统对于同一个数据对象的统计逻辑不同而造成的，逻辑的不一致会直接导致结果的差异性；除了统计逻辑和口径的差异，也有因为源数据系统基于性能的考虑，放弃了外键约束，从而导致数据不一致的结果；另外，也存在极小的数据丢失的可能性，通常由于并发量和负载过高、服务器延迟甚至宕机等原因导致的数据采集的差异。

对于这类的数据矛盾，首先需要明确各个源系统的逻辑、条件、口径，然后定义一套符合各个系统采集逻辑的规则，并对异常源系统的采集逻辑进行修正。

某些情况下，也可能存在业务规则的错误导致的数据采集的错误，此时需要从源头纠正错误的采集逻辑，然后再进行数据清洗和转换。

**5. 转换构造**

数据变换是数据清理过程的重要步骤，是对数据的一个的标准的处理，几乎所有的数据处理过程都会涉及该步骤。数据转换常见的内容包括：数据类型转换、数据语义转换、数据值域转换、数据粒度转换、表/数据拆分、行列转换、数据离散化、数据离散化、提炼新字段、属性构造、数据压缩等。

**数据类型转换**

当数据来自不同数据源时，不同类型的数据源数据类型不兼容可能导致系统报错。这时需要将不同数据源的数据类型进行统一转换为一种兼容的数据类型。

**数据语义转换**

传统数据仓库中基于第三范式可能存在维度表、事实表等，此时在事实表中会有很多字段需要结合维度表才能进行语义上的解析。例如，假如字段M的业务含义是浏览器类型，其取值分为是1/2/3/4/5，这5个数字如果不加转换则很难理解为业务语言，更无法在后期被解读和应用。

**数据粒度转换**

业务系统一般存储的是明细数据，有些系统甚至存储的是基于时间戳的数据，而数据仓库中的数据是用来分析的，不需要非常明细的数据，一般情况下，会将业务系统数据按照数据仓库中不同的粒度需求进行聚合。

**表/数据拆分**

某些字段可能存储多中数据信息，例如时间戳中包含了年、月、日、小时、分、秒等信息，有些规则中需要将其中部分或者全部时间属性进行拆分，以此来满足多粒度下的数据聚合需求。同样的，一个表内的多个字段，也可能存在表字段拆分的情况。

**行列转换**

某些情况下，表内的行列数据会需要进行转换（又称为转置），例如协同过滤的计算之前，user和term之间的关系即互为行列并且可相互转换，可用来满足基于项目和基于用户的相似度推荐计算。

**数据离散化**

将连续取值的属性离散化成若干区间，来帮助消减一个连续属性的取值个数。例如对于收入这个字段，为了便于做统计，根据业务经验可能分为几个不同的区间：0～3000、3001～5000、5001～10000、10001～30000、大于30000，或者在此基础上分别用1、2、3、4、5来表示。

**数据标准化**

不同字段间由于字段本身的业务含义不同，有些时间需要消除变量之间不同数量级造成的数值之间的悬殊差异。例如将销售额进行离散化处理，以消除不同销售额之间由于量级关系导致的无法进行多列的复合计算。数据标准化过程还可以用来解决个别数值较高的属性对聚类结果的影响。

**提炼新字段**

很多情况下，需要基于业务规则提取新的字段，这些字段也称为复合字段。这些字段通常都是基于单一字段产生，但需要进行复合运算甚至复杂算法模型才能得到新的指标。

**属性构造**

有些建模过程中，也会需要根据已有的属性集构造新的属性。例如，几乎所有的机器学习都会讲样本分为训练集、测试集、验证集三类，那么数据集的分类（或者叫分区）就属于需要新构建的属性，用户做机器学习不同阶段的样本使用。

提示 在某些场景中，也存在一些特殊转换方法。例如在机器学习中，有些值是离散型的数据但存在一定意义，例如最高学历这个字段中包含博士、研究生、大学、高中这4个值，某些算法不支持直接对文本进行计算，此时需要将学历这个字段进行转换。常见的方法是将值域集中的每个值拆解为一个字段，每个字段取值为0或1（布尔型或数值型）。这时，就会出现4个新的字段，对于一条记录来看（通常是一个人），其最高学历只能满足一个，例如字段博士为1，那么其余的字段（研究生、大学、高中）则为0。因此这个过程实际上是将1个字段根据值域（4个值的集合）拆解为4个字段。

**6. 数据压缩**

数据压缩是指在保持原有数据集的完整性和准确性，不丢失有用信息的前提下，按照一定的算法和方式对数据进行重新组织的一种技术方法。

对大规模的数据进行复杂的数据分析与数据计算通常需要耗费大量时间，所以在这之前需要进行数据的约减和压缩，减小数据规模，而且还可能面临交互式的数据挖掘，根据数据挖掘前后对比对数据进行信息反馈。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基本相同。

数据压缩的意义不止体现在数据计算过程中，还有利于减少存储空间，提高其传输、存储和处理效率，减少数据的冗余和存储的空间，这对于底层大数据平台具有非常重要的意义。

数据压缩有多种方式可供选择：

q 数据聚合：将数据聚合后使用，例如如果汇总全部数据，那么基于更粗粒度的数据更加便利。

q 维度约减：通过相关分析手动消除多余属性，使得参与计算的维度（字段）减少；也可以使用主成分分析、因子分析等进行维度聚合，得到的同样是更少的参与计算的数据维度。

q 数据块消减：利用聚类或参数模型替代原有数据，这种方式常见于多个模型综合进行机器学习和数据挖掘。

q 数据压缩：数据压缩包括无损压缩和有损压缩两种类型。数据压缩常用于磁盘文件、视频、音频、图像等。

**7. 补足残缺/空值**

由于各种主客观原因，很多系统存在残缺数据，残缺数据包含行缺失、列缺失、字段缺失三种情况。行缺失指的是丢失了一整条数据记录，列缺失指的是丢失一整列数据，字段缺失指的是字段中的值为空值。其中空值也分两种情况：

q 缺失值。缺失值指的是的数据原本是必须存在的，但实际上没有数据。例如年龄这个字段每个人都会有，所以如果系统强制验证是不应该为空。

q 空值。空值指的是实际存在可能为空的情况，所以空值不一定是数据问题。例如身份证号这个字段，只有成人之后才有这个字符串，因此也可能存在非成人的用户，所以可能为空。

对于缺失值和空值的填充处理主要包含两种方式：

一是手工填入可能的值；

二是利用规则填充可能的值：某些缺失值可以从本数据源或其它数据源推导出来，这就可以用数据分布的状态和特征，使用众数、中位数、平均值、最大值、最小值填充，或者使用近邻分析甚至更为复杂的概率估计代替缺失的值，从而达到填充的目的，某些情况下也可以直接以未知或unknown填充，这是一种先期不处理而等到后期业务在处理数据的时候再处理的方法。

提示 对缺失数据进行填补后，填入的值可能不正确，数据可能会存在偏置，导致数据并不是十分可靠的。除了明显的可以确定的规则来填充值以外，基于已有属性来预测缺失值是一种流行的方法。假如性别字段部分记录为空，可以将性别字段作为目标变量进行建模分析，对完整样本建模后得出缺失数据性别为男、女的概率，然后进行填充。对于更多的不确定值的数据样本，如果不影响整体计算逻辑的，建议先保持原样；如果会成为计算和建模噪音的数据，则可以采取上述方法进行处理，尽量使其在建模过程的作用消减。

**8. 丢弃数据/变量**

对于数据中的异常数据，包括缺失值、空值、错误值、不完整的数据记录等，除了使用上面讲的方法进行清洗、转换、提升外，还有另外一种方法——丢弃。丢弃也是提升数据质量的一种方法。丢弃数据的类型包含两种：

q 整条删除，指的是删除含有缺失值的样本。某些情况下，由于各种原因可能存在大量的有某些字段缺失的数据记录，这会导致完整的数据很少，此时需要慎重使用。因此，这只适合关键变量缺失，或者含有无效值或缺失值的样本比重很小的情况。

q 变量删除。如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要，则可以考虑将该变量删除。这种做法减少了供分析用的变量数目，但没有改变样本量。

提示 数据丢弃或删除操作要慎重执行。一方面，被丢弃的数据很可能存在业务实际意义，而这些意义作为开发人员是不清楚的；另一方面，后期可能会需要针对异常数据进行处理，并成为重要的研究课题。例如，营销领域存在流量欺诈，电商领域存在黄牛订单，银行保险领域存在高风险业务，这些课题对应的底层数据可能都是异常数据。