Data Tricks

2023-10-15

1 参数统计中常用的数据变换

对数变换：即将原始数据 X 的对数值作为新的分析数据：

$$X' = lgX$$

当原始数据中有小于 1 及零的数据时，亦可取

$$X' = lg(X + 1)$$

还可根据需要选用

$$X' = lg(X + k) \quad X' = lg(k - X)$$

对数变化常用于：1. 使服从对数正态的数据正态化。如环境中某些污染物的分布，人体中某些微量元素的分布等，可用对数比变换改善其正态性。2. 使数据达到方差齐性，特别是各样本的标准差与均数成比例或变异系数 CV 接近一个常数时。

平方根变换：即将原始数据 X 的平方根作为新的分析数据：

$$X' = \sqrt{X}$$

当原始数据有小值或零值时，亦可用

$$X' = \sqrt{X + 0.5}$$

平方根变换常用于：1. 使服从 Possion 分布的计数资料或轻度偏态的资料正态化，例如放射性物质在单位时间内的放射次数，某些发病率较低的疾病在时间或地域上的发病例数等，可用平方根变换使其正态化。2. 当样本的方差与均数正相关时，可使资料达到方差齐性。

倒数变换：即将原始数据 X 的倒数作为新的分析数据：

$$X' = \frac{1}{X}$$

倒数变换常用于数据两端波动较大的资料，可使极端值的影响减小。

2 Normalization - 归一化

由以下的定义可以看出 Normalizatin 是一个比 scaling 范围更大的概念。

From here

> In statistics and applications of statistics, normalization can have a range of meanings. In the simplest cases, normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging. In more complicated cases, normalization may refer to more sophisticated adjustments where the intention is to bring the entire probability distributions of adjusted values into alignment. In the case of normalization of scores in educational assessment, there may be an intention to align distributions to a normal distribution. A different approach to normalization of probability distributions is quantile normalization, where the quantiles of the different measures are brought into alignment.

3 Feature scaling

Copyed from here

Rescaling(min-max normalization) 将特征的的范围限制在 [0, 1]

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

如果想把数据限制在任意的 [a, b], 公式如下：

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Mean normalization

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

Standardization(Z-score Normalization)

$$x' = \frac{x - \bar{x}}{\sigma}$$

Scaling to unit length

$$x' = \frac{x}{||x||}$$

真正意义上的 normalization，例如 BOX-BOX 变换会让数据服从于正态分布。

4 缺失值处理

Copyed from here

没有高质量的数据，就没有高质量的数据挖掘结果，数据值缺失是数据分析中经常遇到的问题之一。当缺失比例很小时，可直接对缺失记录进行舍弃或进行手工处理。但在实际数据中，往往缺失数据占有相当的比重。这时如果手工处理非常低效，如果舍弃缺失记录，则会丢失大量信息，使不完全观测数据与完全观测数据间产生系统差异，对这样的数据进行分析，你很可能会得出错误的结论。

造成数据缺失的原因

> 现实世界中的数据异常杂乱，属性值缺失的情况经常发全甚至是不可避免的。造成数据缺失的原因是多方面的：信息暂时无法获取。例如在医疗数据库中，并非所有病人的所有临床检验结果都能在给定的时间内得到，就致使一部分属性值空缺出来。信息被遗漏。可能是因为输入时认为不重要、忘记填写了或对数据理解错误而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失。有些对象的某个或某些属性是不可用的。如一个未婚者的配偶姓名、一个儿童的固定收入状况等。有些信息（被认为）是不重要的。如一个属性的取值与给定语境是无关。获取这些信息的代价太大。系统实时性能要求较高。即要求得到这些信息前迅速做出判断或决策。

对缺失值的处理要具体问题具体分析，为什么要具体问题具体分析呢？因为属性缺失有时并不意味着数据缺失，缺失本身是包含信息的，所以需要根据不同应用场景下缺失值可能包含的信息进行合理填充。下面通过一些例子来说明如何具体问题具体分析，仁者见仁智者见智，仅供参考：“年收入”：商品推荐场景下填充平均值，借贷额度场景下填充最小值；“行为时间点”：填充众数；“价格”：商品推荐场景下填充最小值，商品匹配场景下填充平均值；“人体寿命”：保险费用估计场景下填充最大值，人口估计场景下填充平均值；“驾龄”：没有填写这一项的用户可能是没有车，为它填充为 0 较为合理；” 本科毕业时间”：没有填写这一项的用户可能是没有上大学，为它填充正无穷比较合理；“婚姻状态”：没有填写这一项的用户可能对自己的隐私比较敏感，应单独设为一个分类，如已婚 1、未婚 0、未填-1。

缺失的类型

> 在对缺失数据进行处理前，了解数据缺失的机制和形式是十分必要的。将数据集中不含缺失值的变量称为完全变量，数据集中含有缺失值的变量称为不完全变量。从缺失的分布来将缺失可以分为完全随机缺失，随机缺失和完全非随机缺失。完全随机缺失（missing completely at random,MCAR）：指的是数据的缺失是完全随机的，不依赖于任

何不完全变量或完全变量，不影响样本的无偏性。如家庭地址缺失。随机缺失 (missing at random,MAR)：指的是数据的缺失不是完全随机的，即该类数据的缺失依赖于其他完全变量。例如财务数据缺失情况与企业的大小有关。非随机缺失 (missing not at random,MNAR)：指的是数据的缺失与不完全变量自身的取值有关。如高收入人群的不原意提供家庭收入。对于随机缺失和非随机缺失，删除记录是不合适的，随机缺失可以通过已知变量对缺失值进行估计；而非随机缺失还没有很好的解决办法。

说明：对于分类问题，可以分析缺失的样本中，类别之间的比例和整体数据集中，类别的比例

缺失值处理的必要性

数据缺失在许多研究领域都是一个复杂的问题。对数据挖掘来说，缺省值的存在，造成了以下影响：系统丢失了大量的有用信息；系统中所表现出的不确定性更加显著，系统中蕴涵的确定性成分更难把握；包含空值的数据会使挖掘过程陷入混乱，导致不可靠的输出。

数据挖掘算法本身更致力于避免数据过分拟合所建的模型，这一特性使得它难以通过自身的算法去很好地处理不完整数据。因此，缺省值需要通过专门的方法进行推导、填充等，以减少数据挖掘算法与实际应用之间的差距。

缺失值处理方法的分析与比较处理不完整数据集的方法主要有三大类：删除元组、数据补齐、不处理。

4.1 删除元组

也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与初始数据集的数据量相比非常小的情况下非常有效，类标号缺失时通常使用该方法。

然而，这种方法却有很大的局限性。它以减少历史数据来换取信息的完备，会丢弃大量隐藏在这些对象中的信息。在初始数据集包含的对象很少的情况下，删除少量对象足以严重影响信息的客观性和结果的正确性；因此，当缺失数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。

说明：删除元组，或者直接删除该列特征，有时候会导致性能下降。

4.2 数据补齐

这类方法是用一定的值去填充空值，从而使信息表完备化。通常基于统计学原理，根据初始数据集中其余对象取值的分布情况来对一个缺失值进行填充。数据挖掘中常用的有以下几种补齐方法：

人工填写（filling manually） 由于最了解数据的还是用户自己，因此这个方法产生数据偏离最小，可能是填充效果最好的一种。然而一般来说，该方法很费时，当数据规模很大、空值很多的时候，该方法是不可行的。

特殊值填充（Treating Missing Attribute values as Special values） 将空值作为一种特殊的属性值来处理，它不同于其他的任何属性值。如所有的空值都用 "unknown" 填充。这样将形成另一个有趣的概念，可能导致严重的数据偏离，一般不推荐使用。

平均值填充（Mean/Mode Completer） 将初始数据集中的属性分为数值属性和非数值属性来分别进行处理。如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值；如果空值是非数值型的，就根据统计学中的众数原理，用该属性在其他所有对象的取值次数最多的值 (即出现频率最高的值) 来补齐该缺失的属性值。与其相似的另一种方法叫条件平均值填充法（Conditional Mean Completer）。在该方法中，用于求平均的值并不是从数据集的所有对象中取，而是从与该对象具有相同决策属性值的对象中取得。这两种数据的补齐方法，其基本的出发点都是一样的，以最大概率可能的取值来补充缺失的属性值，只是在具体方法上有一点不同。与其他方法相比，它是用现存数据的多数信息来推测缺失值。

热卡填充（Hot deck imputation，或就近补齐） 对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。该方法概念上很简单，且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准，主观因素较多。

K 最近距离邻法（K-means clustering） 先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的 K 个样本，将这 K 个值加权平均来估计该样本的缺失数据。

使用所有可能的值填充（Assigning All Possible values of the Attribute） 用空缺属性值的所有可能的属性取值来填充，能够得到较好的补齐效果。但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大，可能的测试方案很多。

组合完整化方法（Combinatorial Completer） 用空缺属性值的所有可能的属性取值来试，并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法，能够得到好的约简结果；但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大。

回归（Regression）基于完整的数据集，建立回归方程。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。当变量不是线性相关时会导致有偏差的估计。

期望值最大化方法（Expectation maximization，EM）EM 算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E 步（Excepctaion step, 期望步），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M 步（Maximzation step，极大化步），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在 E 步和 M 步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

多重填补（Multiple Imputation，MI）多重填补方法分为三个步骤：为每个空值产生一套可能的填补值，这些值反映了无响应模型的不确定性；每个值都被用来填补数据集中的缺失值，产生若干个完整数据集合。每个填补数据集合都用针对完整数据集的统计方法进行统计分析。对来自各个填补数据集的结果进行综合，产生最终的统计推断，这一推断考虑到了由于数据填补而产生的不确定性。该方法将空缺值视为随机样本，这样计算出来的统计推断可能受到空缺值的不确定性的影响。该方法的计算也很复杂。

C4.5 方法通过寻找属性间的关系来对遗失值填充。它寻找之间具有最大相关性的两个属性，其中没有遗失值的一个称为代理属性，另一个称为原始属性，用代理属性决定原始属性中的遗失值。这种基于规则归纳的方法只能处理基数较小的名词型属性。

就几种基于统计的方法而言，删除元组法和平均值法差于热卡填充法、期望值最大化方法和多重填充法；回归是比较好的一种方法，但仍比不上 hot deck 和 EM；EM 缺少 MI 包含的不确定成分。值得注意的是，这些方法直接处理的是模型参数的估计而不是空缺值预测本身。它们合适于处理无监督学习的问题，而对有监督学习来说，情况就不尽相同了。譬如，你可以删除包含空值的对象用完整的数据集来进行训练，但预测时你却不能忽略包含空值的对象。另外，C4.5 和使用所有可能的值填充方法也有较好的补齐效果，人工填写和特殊值填充则是一般不推荐使用的。

4.3 不处理

补齐处理只是将未知值补以我们的主观估计值，不一定完全符合客观事实，在对不完备信息进行补齐处理的同时，我们或多或少地改变了原始的信息系统。而且，对空值不正确的填充往往将新的噪声引入数据中，使挖掘任务产生错误的结果。因此，在许多情况下，我们还是希望在保持原始信息不发生变化的前提下对信息系统进行处理。

不处理缺失值，直接在包含空值的数据上进行数据挖掘的方法包括贝叶斯网络和人工神经网络等。

贝叶斯网络提供了一种自然的表示变量间因果信息的方法，用来发现数据间的潜在关系。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系。贝叶斯网络仅适合于对领域知识具有一定了解的情况，至少对变量间的依赖关系较清楚的情况。否则直接从数据中学习贝叶斯网的结构不但复杂性较高（随着变量的增加，指数级增加），网络维护代价昂贵，而且它的估计参数较多，为系统带来了高方差，影响了它的预测精度。

人工神经网络可以有效的对付缺失值，但人工神经网络在这方面的研究还有待进一步深入展开。

知乎上的一种方案：

4. 把变量映射到高维空间。比如性别，有男、女、缺失三种情况，则映射成 3 个变量：是否男、是否女、是否缺失。连续型变量也可以这样处理。比如 Google、百度的 CTR 预估模型，预处理时会把所有变量都这样处理，达到几亿维。这样做的好处是完整保留了原始数据的全部信息、不用考虑缺失值、不用考虑线性不可分之类的问题。缺点是计算量大大提升。而且只有在样本量非常大的时候效果才好，否则会因为过于稀疏，效果很差。

4.4 总结

大多数数据挖掘系统都是在数据挖掘之前的数据预处理阶段采用第一、第二类方法来对空缺数据进行处理。并不存在一种处理空值的方法可以适合于任何问题。无论哪种方式填充，都无法避免主观因素对原系统的影响，并且在空值过多的情形下将系统完备化是不可行的。从理论上来说，贝叶斯考虑了一切，但是只有当数据集较小或满足某些条件（如多元正态分布）时完全贝叶斯分析才是可行的。而现阶段人工神经网络方法在数据挖掘中的应用仍很有限。值得一提的是，采用不精确信息处理数据的不完备性已得到了广泛的研究。不完备数据的表达方法所依据的理论主要有可信度理论、概率论、模糊集合论、可能性理论，D-S 的证据理论等。

5 参考：

1. 卫生统计学教程 - 王燕等