Atitit 数据算法 数据规约 压缩

数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量（完成该任务的必要前提是理解挖掘任务和熟悉数据

数据归约主要有两个途径：属性选择和数据采样，分别针对原始数据集中的属性和记录。

假定在[公司](https://baike.baidu.com/item/%E5%85%AC%E5%8F%B8/251430" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)的[数据仓库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93/381916" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)选择了[数据](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE/5947370" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)，用于分析。这样数据集将非常大。在海量数据上进行复杂的[数据分析](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90/6577123" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)扣挖掘将[需要](https://baike.baidu.com/item/%E9%9C%80%E8%A6%81/9109492" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)很长时间，使得这种分析不现实或不可行。

1、特征归约

特征归约是从原有的特征中删除不重要或[不相关](https://baike.baidu.com/item/%E4%B8%8D%E7%9B%B8%E5%85%B3" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)的特征，或者通过对特征进行[重组](https://baike.baidu.com/item/%E9%87%8D%E7%BB%84" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)来减少特征的个数。其原则是在保留、甚至提高原有判别能力的

同时减少特征向量的维度。特征归约算法的输入是一组特征，输出是它的一个子集。在领域知识缺乏的情况下进行特征归约时一般包括3个步骤:

(1)搜索过程:在特征空间中搜索特征子集，每个子集称为一个状态由选中的特征构成。

(2)评估过程:输入一个状态，通过评估函数或预先设定的阈值输出一个评估值搜索算法的目的是使评估值达到最优。

(3)分类过程:使用最终的特征集完成最后的算法。

特征归约处理的效果：

（1）更少的数据，提高挖掘效率

（2）更高的数据挖掘处理精度

（3）简单的数据挖掘处理结果

（4）更少的特征。

2、样本归约

样本都是已知的，通常数目很大，[质量](https://baike.baidu.com/item/%E8%B4%A8%E9%87%8F" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)或高或低，或者有或者没有关于实际问题的先验知识。

样本归约就是从数据集中选出一个有代表性的[样本](https://baike.baidu.com/item/%E6%A0%B7%E6%9C%AC" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)的子集。子集大小的确定要考虑计算成本、存储要求、[估计量](https://baike.baidu.com/item/%E4%BC%B0%E8%AE%A1%E9%87%8F" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)的精度以及其它一些与算法和数据特性有关的因素。

初始数据集中最大和最关键的维度数就是样本的数目，也就是数据表中的记录数。数据挖掘处理的初始数据集描述了一个极大的总体，对数据的分析只基于样本的一个子集。获得数据的子集后，用它来提供整个数据集的一些[信息](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)，这个子集通常叫做估计量，它的质量依赖于所选子集中的元素。取样过程总会造成取样误差，取样误差对所有的方法和策略来讲都是固有的、不可避免的，当子集的规模变大时，取样误差一般会降低。一个完整的数据集在理论上是不存在取样误差的。与针对整个数据集的[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)比较起来，样本归约具有以下一个或多个优点:减少[成本](https://baike.baidu.com/item/%E6%88%90%E6%9C%AC" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)、速度更快、范围更广，有时甚至能获得更高的精度。

3、特征值归约

特征值归约是特征值离散化技术，它将连续型特征的值离散化，使之成为少量的区间，每个区间映射到一个离散符号。这种[技术](https://baike.baidu.com/item/%E6%8A%80%E6%9C%AF" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)的好处在于简化了数据描述，并易于理解数据和最终的挖掘结果。

特征值归约可以是有参的，也可以是无参的。有参方法使用一个[模型](https://baike.baidu.com/item/%E6%A8%A1%E5%9E%8B" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)来评估数据，只需存放参数，而不需要存放实际数据;有参的特征值归约有以下两种:

(1)回归:[线性回归](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)和多元回归;

(2)对数线性模型:近似离散多维概率分布。

无参的特征值归约有3种:

(1)[直方图](https://baike.baidu.com/item/%E7%9B%B4%E6%96%B9%E5%9B%BE" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank):采用分箱近似数据分布，其中V-最优和MaxDiff直方图是最精确和最实用的;

(2)[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank):将数据元组视为对象，将对象划分为群或聚类，使得在一个聚类中的对象“类似”而与其他[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB" \t "https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E5%BD%92%E7%BA%A6/_blank)中的对象“不类似”在数据归约时用数据的聚类代替实际数据;

(3)选样:用数据的较小随机样本表示大的数据集，如简单选择n个样本(类似样本归约)、聚类选样和分层选样等。