3-5-cn
1
本节我们讨论数据预处理中的数据规约相关问题
2
数据规约（数据消减）
大型数据库内容的复杂数据分析通常需要大量时间。
使用数据规约（减法）技术，帮助从原始庞大的数据集中得到一个精简的数据集，并使这个精简的数据集保持原始数据集的完整性，这样对精简数据集的数据分析显然效率更高，分析结果与使用原始数据集得到的结果基本一致。
3
数据规约标准，进行数据规约要满足两个条件，分别是：
1）用于数据归约的时间不应当超过或“抵消”在归约后的数据上挖掘节省的时间
2）归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
4
数据降维技术包括降低维数和减少数量。
降维方法包括属性子集选择、主成分分析和小波变换
数量约简可以分为参数化方法和非参数化方法。
非参数方法包括数据立方体聚合、聚类、抽样直方图等。
5
首先让我们看看降维
删除不相关的属性，减少数据分析处理的数据量。
6
属性子集选择：**维归约——选择相关属性子集**
逐步向前选择
从一个空属性集（作为属性子集初始值）开始，
每次从原来属性集合中选择一个当前最优的属性添加到当前属性子集中。
直到无法选择出最优属性或满足一定阈值约束为止。
7
属性子集选择：**维归约——选择相关属性子集**
逐步向后删除
从一个全属性集（作为属性子集初始值）开始，每次从当前属性子集中选择一个当前最差的属性并将其从当前属性子集中消去。直到无法选择出最差属性为止或满足一定阈值约束为止。
还可以使用向前选择和向后删除结合的方法
8
属性子集选择的其他方法
1）判定树（决策树）归纳
利用决策树的归纳方法对初始数据进行分类归纳学习，获得一个初始决策树，所有没有出现这个决策树上的属性均认为是无关属性，因此将这些属性从初始属性集合删除掉，就可以获得一个较优的属性子集。
2）基于统计分析的归约

9 现在让我们看看数据压缩
10
数据压缩技术分类：可以分为无损(loseless)压缩和有损(lossy)压缩
无损(loseless)压缩：可以不丢失任何信息地还原压缩数据。
例如：字符串压缩
有广泛的理论基础和精妙的算法

有损(lossy)压缩：只能重新构造原数据的近似表示。
例如：音频/视频压缩
有时可以在不解压整体数据的情况下，重构某个片断

11

数据归约——数据压缩

数据压缩——用数据编码或者变换，得到原始数据的压缩表示。
在数据挖掘领域通常使用的两种数据压缩方法均是有损的：
主成分分析法（PCA）
假定待压缩的数据由 n 个取自 k 个维的元组或数据向量组成。主要成分分析并搜索得到 c 个最能代表数据的 k 维正交向量，这里 $c \leqslant k$。这样就可以把原数据投影到一个较小的空间，实现数据压缩
另外还有小波变换的数据压缩方法

12

减少数据量的方法可以分为无参数的方法和有参数的方法；
其中无参数的方法包括，数据立方体聚合，聚类，采样，直方图等方法

13

让我们看一下数据立方体聚合的数据规约方法
数据立方体是数据的多维建模和表示，由维度和事实组成。
维度：属性
事实：数据
数据立方体聚合定义——将 n 维数据立方体聚集成 n-1 维数据立方体。
使用数据立方体聚合进行近似查询在图中，3 维数据立方体聚合成 2 维，男性和女性的销售额聚合在一起。

14

数据归约——离散化与概念分层生成

现在让我们看看数据规约方法的离散化方法。
属性值可以是 Name 类型。一个无序集合中的值。值和连续值。实数。
离散化技术
通过将属性(连续值)域值的范围划分为几个区间来减少连续(值)属性的值的数量。

15

第三个数据规约方法是使用概念分层生成进行数据规约
概念层次定义了一组从低级概念集到高级概念集的映射。
它允许在各种抽象层次上处理数据，从而在多个抽象层次上发现知识。
使用更高级的概念替换更低级的概念（例如本例中的年龄值可以抽象化为青年，中年和壮年），以减少值的数量。
虽然一些细节在数据泛化的过程中消失了，但是这样得到的广义数据可能更容易理解，也更有意义。
对规约数据集的数据分析显然更有效。概念层次可以用一棵树来表示，树的每个节点代表一个概念。

16

本节课我们介绍了数据规约和相关技术，今天的课就到这里，谢谢大家