



1.7 数据预处理

数据预处理是数据挖掘中非常重要的一步，通常包括数据清洗、数据集成、数据变换和数据归约。数据清洗主要解决数据中的噪声、缺失值和异常值问题。数据集成是将来自不同数据源的数据合并在一起。数据变换是将数据转换为适合挖掘的形式，如规范化、离散化和聚集。数据归约是减少数据量，提高挖掘效率。

1.8 数据集成

数据集成是指将来自不同数据源的数据合并在一起，形成一个统一的数据视图。数据集成过程中需要考虑数据源的异构性、数据冲突、数据冗余和数据安全问题。常见的数据集成方法包括水平集成、垂直集成和物理集成。

1.9 数据变换

数据变换是指将数据转换为适合挖掘的形式。数据变换包括规范化、离散化和聚集。规范化是将数据映射到一定的范围，如[0,1]。离散化是将连续数据转换为离散数据。聚集是将数据分组，减少数据量。