# 大数据挖掘是什么，数据挖掘的方法主要有哪些？

人工智能4 月 2 日


数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

**数据挖掘对象**

根据信息存储格式，用于挖掘的对象有关系数据库、面向对象数据库、数据仓库、文本数据源、多媒体数据库、空间数据库、时态数据库、异质数据库以及 Internet 等。

**数据挖掘流程**

定义问题：清晰地定义出业务问题，确定数据挖掘的目的。

数据准备：数据准备包括：选择数据–在大型数据库和数据仓库目标中 提取数据挖掘的目标数据集;数据预处理–进行数据再加工，包括检查数据的完整性及数据的一致性、去噪声，填补丢失的域，删除无效数据等。

数据挖掘：根据数据功能的类型和和数据的特点选择相应的算法，在净化和转换过的数据集上进行数据挖掘。

结果分析 :对数据挖掘的结果进行解释和评价 ,转换成为能够最终被用户理解的知识。

**数据挖掘分类**

直接数据挖掘：目标是利用可用的数据建立一个模型，这个模型对剩余的数据，对一个特定的变量(可以理解成数据库中表的属性，即列)进行描述。

间接数据挖掘：目标中没有选出某一具体的变量，用模型进行描述;而是在所有的变量中建立起某种关系。

**数据挖掘的方法**

**神经网络方法**

神经网络由于本身良好的鲁棒性、自组织自适应性、并行处理、分布存储和高度容错等特性非常适合解决数据挖掘的问题，因此近年来越来越受到人们的关注。

**遗传算法**

遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化方法。遗传算法具有的隐含并行性、易于和其它模型结合等性质使得它在数据挖掘中被加以应用。

**决策树方法**

决策树是一种常用于预测模型的算法，它通过将大量数据有目的分类，从中找到一些有价值的，潜在的信息。它的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。

**粗集方法**

粗集理论是一种研究不精确、不确定知识的数学工具。粗集方法有几个优点：不需要给出额外信息;简化输入信息的表达空间;算法简单，易于操作。粗集处理的对象是类似二维关系表的信息表。

**覆盖正例排斥反例方法**

它是利用覆盖所有正例、排斥所有反例的思想来寻找规则。首先在正例集合中任选一个种子，到反例集合中逐个比较。与字段取值构成的选择子相容则舍去，相反则保留。按此思想循环所有正例种子，将得到正例的规则(选择子的合取式)。

**统计分析方法**

在数据库字段项之间存在两种关系：函数关系和相关关系，对它们的分析可采用统计学方法，即利用统计学原理对数据库中的信息进行分析。可进行常用统计、回归分析、相关分析、差异分析等。

**模糊集方法**

即利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。系统的复杂性越高，模糊性越强，一般模糊集合理论是用隶属度来刻画模糊事物的亦此亦彼性的。

**数据挖掘任务**

**关联分析**

两个或两个以上变量的取值之间存在某种规律性，就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。一般用支持度和可信度两个阀值来度量关联规则的相关性，还不断引入兴趣度、相关性等参数，使得所挖掘的规则更符合需求。

**聚类分析**

聚类是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

**分类**

分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则。分类可被用于规则描述和预测。

**预测**

预测是利用历史数据找出变化规律，建立模型，并由此模型对未来数据的种类及特征进行预测。预测关心的是精度和不确定性，通常用预测方差来度量。

**时序模式**

时序模式是指通过时间序列搜索出的重复发生概率较高的模式。与回归一样，它也是用己知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。

**偏差分析**

在偏差中包括很多有用的知识，数据库中的数据存在很多异常情况，发现数据库中数据存在的异常情况是非常重要的。偏差检验的基本方法就是寻找观察结果与参照之间的差别。