# 统计学方法在数据挖掘中的应用探究

来源：网络 作者：九曲桥畔 更新时间：2024-01-25

*数据挖掘就是指从众多实际应用数据中获取批量大、有噪声、且随机性强的数据，将潜在的信息与数据提取出来，就是从数据中挖掘有价值的知识，而大多数原始数据具有一定的结构化特征，比如，关系数据库中的数据;也可以通过文本、图形、图像等半结构化发掘有用知...*

数据挖掘就是指从众多实际应用数据中获取批量大、有噪声、且随机性强的数据，将潜在的信息与数据提取出来，就是从数据中挖掘有价值的知识，而大多数原始数据具有一定的结构化特征，比如，关系数据库中的数据;也可以通过文本、图形、图像等半结构化发掘有用知识，这些知识可以是数学的也可以是非数学形式的;数据挖掘能以归纳形式存在，能够被广泛应用到信息查询、信息管理、信息决策控制中，方便数据的维护与管理。由此可见，数据挖掘是一门交叉性强的学科，加强对其的研究非常有意义，下面将对统计方法在数据挖掘中的具体应用进行分析。

一、数据挖掘与统计学的关系

(一)数据挖掘的内涵

通常来说，数据挖掘的定义较为模糊，没有明确界定，大部分对其的定义只是停留在其背景与观点的内容上。通过对不同观点的统一整理，人们最终将其描述为：从大量多样化的信息中发现隐晦性、规律性等潜在信息，并对这些信息进行创造、加工的过程。数据挖掘作为一门重要的交叉学科，能够将数据库、人工智能、机器学习、统计学等众多的科学融入到一起，从而实现技术与理论的创新与发展。其中，数据库、人工智能与统计学是数据挖掘当中的三大支柱理论。数据挖掘的目的是从数据库当中发掘各种隐含的知识与信息，此过程的方法非常多，有统计学知识、遗传算法、粗集方法、决策法、模糊逻辑法等，还可以应用向邻近的可视技术、模式识别技术等，在以上所有技术的支持上能够使数据挖掘更为科学、有序。

(二)数据挖掘与统计学间的关系

通常来说，统计学的主要功能是对统计原理与统计方法进行研究的科学。具体来说就是指对数字资料进行的收集、整理、排序、分析、利用的过程，数字资料是各种信息的归纳与总结，可以将其作为特性原理的认知、推理方法。而统计学则表示的是使用专业的统计学、概率理论原理等对各种属性关系的统计与分析过程，通过分析成功找到属性间的关联与发展的规律。在此过程中，统计分析方法是数据挖掘最为重要的手段之一。

在数据挖掘这一课题被提出来之前，统计分析技术对于人们来说更熟悉，也是人们日常开展工作、寻找数据间规律最常使用的方法。但是不能简单的将数据挖掘作为统计学的延伸与替代工具，而是要将两者的区别认识到位，再结合两者间的不同特点分析其应用特点。大部分的统计学分析技术都是建立在数学理论与技巧上的，预测通常较为准确，效果能够让大部分人满意。数据挖掘能够充分借鉴并吸收统计学技术，在融入到自身特点以后成为一种数据挖掘技术。

统计学与数据挖掘存在的目标都是一致的，就是不断对数据结构进行发掘。鉴于统计学与数据挖掘在目标上的一致性，致使很多研究学者与专家将数据挖掘作为了统计学的一个分支机构。但是这种认知非常不正确，因为数据挖掘不仅体现在与统计学的关系上还体现在思想、工具与方法上，尤其是在计算机科学领域对数据挖掘起到的作用非常大。比如，通过借助数据库技术与人工智能的学习，能够关注到更多统计学与数据挖掘上的共通点，但是两者存在的差异依然非常大。数据挖掘就是指对大量的数据信息不断挖掘的过程，DM能够对数据模式内的数据关系进行充分挖掘，并对观测到的数据库处理有着极高的关注度。

二、数据挖掘的主要过程

从数据本身出发探讨数据挖掘过程，数据挖掘的过程分为信息的收集、数据集成、数据处理、数据变换、数据挖掘实施等过程。

首先，要将业务对象确定下来，明确不同业务定义，并认清数据挖掘的目的，这是做好数据挖掘最关键的一步，也是最重要的一步，虽然挖掘的结果不能被准确预测到，但却需要对问题的可预见性进行探索。其次，还要做好数据准备工作，包含数据清理、数据变换等工作，数据清理的实际意义是将噪声与空缺值补全，针对这一问题，可以使用平滑技术，而空缺值的处理则是属性中最常见的，可以将统计中最可能出现的值作为一个空缺值。

信息收集指的是按照特定的数据分析对象，可以将分析中需要的特征信息抽象出来，并在此基础上选择出较为科学、适合的信息收集方法，将全部的信息全部录入到特定的数据库中。如果数据量较大，则可以选择一个专门的管理数据的仓库，实现对信息的有效保护与管理;数据集成就是指将来源不同、格式不同、性质不同、特点不同的数据集成到一起，进而为企业提供更为全面、系统的数据共享平台;数据变换就是通过聚集、概化、规范化等方式对数据进行挖掘，对于一些实用数据，则可以通过分层与分离方式实现对数据的转换;数据挖掘就是结合数据仓库中的数据信息点，并选择正确的分析方法实现对有价值数据的挖掘，事例推理、规则推理、遗传算法等都是应用较多的方法。

三、统计学方法中的聚类分析

在统计学聚类方法基础上能够构建出潜在的概率分布假设，可以使用试图优化的方法构建数据与统计模型的拟合效果。基于统计学聚类方法当中，Cobweb方法是在1987年由Fisher提出的，能够以分类树作为层次聚类创建的方法，在分类树上，每一个节点都能代表着一个概念，该方法就是对节点概率描述的过程。Cobweb方法还使用了启发式估算方式，使用分类效用对分类树的构建进行指导，从而实现对最高分类的划分目的，能够将不同分类对象全部归类到一个类别中，并依据这些内容创建出一个新的类别。但是这种方法也存在一定局限性，局限性在于假设的属性概率分布都是独立的，并不能始终处于成立状态中。只有在掌握了Cobweb算法以后才能对概念聚类算法的特点进行探究。Cobweb算法能够以分类树方式创建层次聚类，可以将概率表现为p(Ai=Vii/Ck)条件概率，其中，Ai=Vij是一个类别下的，同属于一个值对，Ck是概念类中的一种。在给出一个特定的对象以后，Cobweb能够将全部对象整合到一个节点上，从而计算出分类效应，分数最高的效用就是对象所在的节点位置。如果对象构建失去节点，则Cobweb能够给出一个新的节点，并对其进行分类使用，这种节点计算方法起步较晚，能够对现有的节点与计算相互对比，从而划分出最高的分类指标，将全部对象统一到已有的分类中，从而构建出一个新的类别。

Classitci是Cobw eb方法的一种延伸与发展，能够使用其完成聚类数据的处理，在该方法下，节点中的每一个存储属性都是处于连续分布状态中，能够将其作为分类效果修正的方法，并以度量的形式表现出来，这种度量基础上能够实现连续性的积分，从而降低分散发生率，该方法是积分过程而不是对属性的求和过程。

Auto Class方法也是一种应用较为普遍的聚类方法，该方法主要采用统计分析对结果类的数目进行估算，还可以通过模型搜索方式分析空间中各种分类的可能性，还能够自动对模型数量与模型形态进行描述。在一定类别空间中，不同的类别内属性存在关联性，不同的类别间具有相互继承性，在层次结构当中，共享模型参数是非常重要的。

还有一种使用较为普遍的模型是混合模型，混合模型在统计学聚类方法上使用也非常普遍。该方法最为基本的思想就是概率分布决定着每一种聚类状态，并且模型中的每一个数据都是由多个概率在分布状态下产生的。混合模型还能够作为一种半参数密度评估方法，其能够将参数估计与非参数估计的优点全部集中到一起，并将参数估计法与非参数估价法的诸多优点融合到一起，因为模型具有一定复杂性，为此，不能将其限制在概率密度函数表达形式上，这种复杂性决定了模型与求解存在关联，与样本集合的联系非常少。通过以上的研究可以了解到，数据发掘中应用聚类方法非常有效，并且较为常见。比如，构建出Cobweb模型与混合模型，采用Clara与Clarans方法中的抽样技术，将Denclue方法用在概率密度函数中。

结束语

统计学方法自产生开始已经有非常久远的历史，将严谨的数学逻辑作为基础，将分类算法假定作为独立条件，属性值之前能够相互保持独立，对假定进行计算，当假定成立时，可以再与其他分类算法进行对比，这种分类算法准确性非常高。为此，其不仅能够对连续值进行预测，还可以通过线性回归方程对系数进行比较，从而归纳出结果。

本DOCX文档由 www.zciku.com/中词库网 生成，海量范文文档任你选，，为你的工作锦上添花,祝你一臂之力！