Hans Journal of Data Mining 数据挖掘, 2019, 9(4), 131-134
Published Online October 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.94016

Hans汉斯

Discussion on the Application of Data Mining Technology in the Information Age

Yilan Wu

Zhongnan University of Economics and Law, Wuhan Hubei
Email: 597385469@qq.com

Received: Sep. 29th, 2019; accepted: Oct. 14th, 2019; published: Oct. 21st, 2019

Abstract

The background and concept of data mining technology are briefly described, and the seven processes of data mining are described in detail. Finally, the main data mining methods of data mining technology and its applicable fields are elaborated.

Keywords

Data Mining, Cluster Analysis, Association Rules

信息时代下数据挖掘技术的应用探讨

吴依兰

中南财经政法大学，湖北 武汉
Email: 597385469@qq.com

收稿日期：2019年9月29日；录用日期：2019年10月14日；发布日期：2019年10月21日

摘　要

本文简述了数据挖掘技术产生的背景及其概念，并进一步详细描述了数据挖掘的七个过程，最后详细阐述了数据挖掘技术的主要数据挖掘方法及其适用领域。

关键词

数据挖掘，聚类分析，关联规则

文章引用: 吴依兰. 信息时代下数据挖掘技术的应用探讨[J]. 数据挖掘, 2019, 9(4): 131-134.
DOI: 10.12677/hjdm.2019.94016

1. 数据挖掘的背景与概念

随着信息时代的到来，在享受信息技术所带来的便利的同时，人们也面临着信息时代所导致的信息爆炸问题，面对着越来越多的信息数据，从中获得有价值的知识变得越来越困难，而从上世纪 80 年代开始，伴随着数据库技术的发展和新应用的提出，数据挖掘技术应运而生。数据挖掘是一种能从海量的、随机的、不完整的、复杂的数据中提取出对人们可能潜在有用的信息和知识的过程，其中涉及到多种技术的内容，包括计算机技术、数据库技术、数据统计技术等。通过数据挖掘技术来处理大量的数据内容极大地提高人们处理信息的效率，从中可以获取大量有价值的信息和知识来帮助人们进行决策工作[1]。

2. 数据挖掘过程

数据挖掘是从海量数据中提取隐含在其中的有用信息和知识的过程，一般来说，我们将数据挖掘过程概括为以下 7 个部分[2]：

1) 定义挖掘目标

清晰明确的挖掘目标是进行数据挖掘的前提，同时也是能够最大限度发挥数据挖掘作用的关键。在进行数据挖掘之前我们必须清楚的知道目标是什么，针对具体的目标，了解的与其相关的应用领域的背景知识，这样有助于从整体把握数据挖掘过程，结合实际问题对数据挖掘结果进行分析。

2) 数据取样

在对数据挖掘目标有了清晰明确的认识之后，接下来需要考虑的就是针对挖掘目标如何选取样本。选取样本时我们需要遵循三大原则，即时效性、可靠性和相关性。必须保证选取的样本数据是最新的、真实可靠的并且与挖掘目标是高度相关的。选取数据是既要考虑数据的完整性与系统性，提示也要考虑到数据的简明性，精选数据，减少数据的计算量，节省资源。

3) 数据探索

获取样本数据之后，我们需要对数据进一步分析探究，数据之间是否存在易被察觉的规律或者趋势，有没有比较明显的类别，数据之间的相关程度如何，这些都是需要进一步分析探究的。

为了保证预测质量，对选取的样本数据进行探索、审核和必要的加工处理是必要的。数据探索就是为了保证用于建模的样本数据的质量，进而为预测质量奠定基础。

数据探索主要包括：相关分析、异常值分析、周期性分析、缺失值分析和样本交叉验证等[3]。

4) 数据预处理

由于用于数据挖掘的样本数据量一般较为庞大，数据结构较为复杂，样本数据可能维度过高，有缺失值，含有噪声，有重复记录，不一致等等，为便于进行数据挖掘，提高预测的准确率和效果，样本数据的预处理是必不可少的。

数据预处理主要有数据筛选、缺失值处理、数据变量转换、坏数据处理、属性选择、数据标准化和数据规约。

5) 模式发现

数据预处理之后，我们就可以开始构建挖掘模型，在建模之前我们需要考虑本次的挖掘目标是数据挖掘哪方面的应用，也就是上文提出的分类和回归技术、聚类分析、关联规则、时序模式和异常检测，

针对具体的应用类别选取合适的算法。

6) 模型构建

确定了本次数据挖掘应用的具体类别之后，接下来就需要考虑如何构建模型，包括选择什么挖掘算法，模型构建思路，具体的操作过程是怎样等等。通常，我们将样本分为训练样本和测试样本，训练样本用来构建模型，测试样本用来观察模型在新的数据上的表现。

7) 模型评价

对数据挖掘的结果进行评价，并对数据挖掘过程中的不足之处和可取之处进行分析总结，以期在以后的数据挖掘过程中不断改进，最后，结合数据挖掘结果和现实生活中的应用进行分析总结，并对实际应用提出合理的有效的改进意见。

数据挖掘过程可，如图 1 所示。

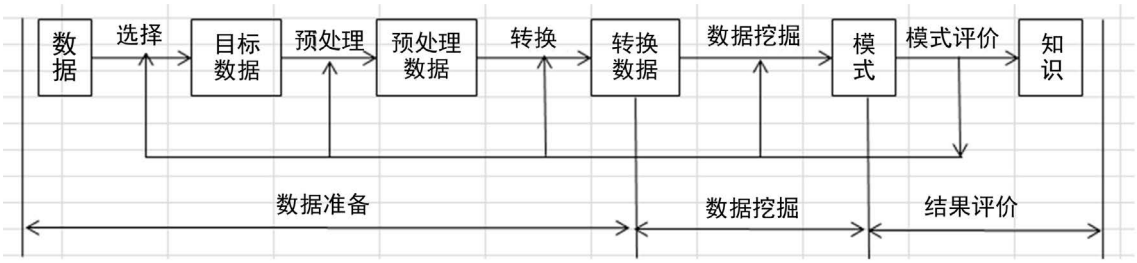

Figure 1. Data mining process
图 1. 数据挖掘过程

3. 数据挖掘方法

1) 分类和回归技术

分类和回归技术是数据挖掘中使用最多最频繁的两种方法。简单地说，分类是将由一系列变量组成数据集合映射到预先定义好的群组或类。分类的前提是这组数据已经有确定了的类别，所以分类又被称为有监督的学习。分类就是构造一个分类模型(分类函数)，用其他的变量值来表示已确定的目标分类变量值。它由模型创建和模型使用两个过程组成，模型创建就是通过对训练数据集的分析学习来挖掘隐含在数据集中可以用来预测的分类模型；模型使用是指利用分类模型预测新的数据所属的类，也就是将未知分类变量值的数据记录能够尽可能准确地被判定到某一类别中去。商业银行可根据客户对于风险的偏好不同对其分类，有针对性的对不同风险偏好者进行理财产品、基金等的推销，这样不仅能提高效率还能获取较大的收益。

回归分析是用属性的历史数据预测未来趋势，找出各个数据之间的相关关系。回归分析通过假设存在可以拟合目标属性的函数，然后利用样本数据进行误差分析，确定最能体现目标属性的函数。简言之，回归分析是处理变量间(包括一对一和一对多)相关关系的一种统计方法。它与分类模型类似，主要区别在于后者采用的是离散预测值，而回归模型采用连续的预测值。利用回归分析，商业银行可对客户的信用风险等级进行评估，为是否发放贷款及贷款额度提供决策依据。分类和回归技术有很多，如决策树、贝叶斯网络、Logistic 回归方法、随机森林算法、遗传算法等[4]。

2) 聚类分析

聚类分析是根据数据之间的相似度进行数据分类的一种方法，它是在没有划分数据类的前提下进行的。所以，聚类又被称为无指导的学习。聚类的输入对象是一组事先未被分类的数据，通过确定数据之间在原本的属性上的相似性来完成聚类任务。不管研究对象中是否真的有不同的类别，运用聚类分析都

能将样本数据分成若干个类别，但其结果并不是唯一的，选择哪一个分类结果最终是由研究者的主观判断和分析总结决定的。聚类的原则就是使同一聚类中的数据尽可能的相似，不同聚类中的数据尽可能的相异，也就是尽量保证组内相似性最大，组间相似性最小。利用聚类分析，商业银行可以对大量客户进行分类，找到他们的共同点，比如风险爱好者、风险中立者和风险规避者，然后可以有针对性的介绍相关业务，推销理财产品，从而达到事半功倍的效果。

3) 关联规则

关联规则是应用较为广泛的数据挖掘技术，它在商业销售、保险业、银行业和制造业等方面都具有很强的实用性。关联规则，顾名思义，它是从错综复杂的数据中发现事物之间可能存在的关联或者联系，这种关联是隐含在数据中没有直接显现出来的。关联规则又可表示为：如果 X 发生，那么 Y 发生的可能性是百分之 Z。Z 又称为关联规则的置信度，即条件概率。关联规则挖掘大致可以分为两步：第一步是从原始数据中找出频繁项目集；第二步是从频繁项目集中挖掘出满足最低置信度的关联规则。利用关联规则挖掘技术，商业银行可以分析客户购买理财产品 A 的同时又购买理财产品 B 的可能性，还能用于发现客户基本信息中与优质客户存在关联关系的属性和它们之间的关联程度以及客户违约的前后因果关系等。关联规则挖掘算法有很多，其中最出名的就是 Apriori 算法[5]。

4) 时序模式

时序模式与回归分析类似，是用已知的数据对未来的趋势进行预测。区别是这些数据的属性值是随时间不断变化的，因此，时序模式主要考虑的是大量复杂多变的数据在时间维度上的关系。时序模式包含序列分析和序列发现。商业银行可以利用时序模式对每个月信用卡、基金等的销售额量进行预测，找到影响销量的因素，从而及时采取有效的措施提高销售额。

5) 异常检测

异常检测又被称为偏差检测，是用来发现与其他大部分对象不同的异常或变化，并进一步分析这种变化是合理的变化，还是恶意的欺诈行为。异常检测分为两个阶段：第一个阶段是对所研究的数据中的异常对象给出清晰的定义，第二个阶段是结合具体研究对象找到合适的方法挖掘出这些异常对象。异常检测对于金融领域中的欺诈交易的防范与侦破具有重大意义，对我国经济正常有序的发展具有推动作用。

4. 结束语

数据挖掘技术从概念的提出到理论的完善、算法的成熟一步步成为了一套完整的体系，并成功应用在许多领域，例如金融、电子商务、医疗、机械工业、网络等领域，这也表明了数据挖掘技术有着广泛的应用前景和研究价值。面对信息时代所产生的大量数据，根据实际的业务需求我们可以按照本文介绍的数据挖掘过程一步步分析研究，并结合适当的数据挖掘方法提取出所需要的信息和知识，挖掘出潜在的模式关系。

参考文献

[1] 程陈. 大数据挖掘分析[J]. 软件, 2014(4): 130-131.

[2] 范明, 孟小峰, 译. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2012.

[3] 翟峰. 数据挖掘语言分析及应用探析[J]. 通讯世界, 2015(11): 273.

[4] 丁秀玲. 数据挖掘算法和研究方向[J]. 办公自动化(学术版), 2014(8): 33-34, 56.

[5] 刘羿, 陈宝钢. 数据挖掘的应用及优化浅析[J]. 电子商务, 2014(3): 52-53.