数据挖掘及其 Python 实现技术研究

刘 班
（武汉职业技术学院，湖北 武汉 430074）

摘要：数据挖掘是“数据库中的知识发现”过程中的一个核心步骤。Python 语言易学易用，且包含了大量与数据挖掘相关的第三方库，在数据挖掘方面的独有优势，使得 Python 成为了数据挖掘的常用工具。论述了数据挖掘的相关知识和用 Python 实现数据挖掘的优势，并用实例说明了 Python 实现数据挖掘核心技术的具体方法。

关键词：数据挖掘；知识发现；机器学习；特征提取；数据清洗；Python；算法
中图分类号：TP311.13 **文献标识码**：A **文章编号**：1673-1131（2020）03-0063-03

1 概述

随着全社会计算机化以及数据收集和存储工具的不断发展，导致各类数据的爆炸式增长。在当前这个大数据时代，人脑已无法深入理解这些海量数据的含义，导致出现了数据丰富但信息匮乏的现象。数据和信息之间的这道越来越宽的鸿沟，促使了数据挖掘工具的出现。本文论述了数据挖掘的相关知识和用 Python 技术实现数据挖掘的优势，并用实例说明了数据挖掘核心技术的 Python 实现方法。

2 数据挖掘的概念

数据挖掘是搜索隐含于某个大数据集中的先前未知的并具有潜在价值信息的过程。数据挖掘技术涉及到机器学习、统计学和数据库系统的交叉方法，是计算机科学和统计学的一个跨学科子领域，其总体目标是通过智能化方法从数据集中提取信息，并将这些信息转换为可理解的结构以供进一步使用。数据挖掘是“数据库中的知识发现”过程（KDD）中的一个核心步骤。数据挖掘技术经常被用于各种形式的大规模数据或信息处理系统以及计算机决策支持系统中。实际的数据挖掘任务主要是对大量数据进行半自动化或全自动化分析，提取以前未知的并包含一定知识内涵的模式。具体来说，数据挖掘包括聚类分析、异常检测、关联规则挖掘、顺序模式挖掘等子任务，完成这些任务通常需要使用空间索引等数据库技术。通过数据挖掘发现的模式可以被视为对输入数据的一种总结，并可以用于后续的进一步数据分析工作中。例如，数据挖掘步骤可以识别数据中的多个组，然后通过决策支持系统使用这些组来获得更准确的预测结果。数据挖掘和数据分析的区别在于，数据分析用于验证数据集上的模型是否正确或者假设是否成立，例如，分析营销活动的可行性。相比之下，数据挖掘使用机器学习和统计模型来发现隐藏于大量数据中的有价值信息或模式。

具体来说，在如下几个领域会大量使用数据挖掘技术：

（1）金融交易领域：在人们日常使用银行卡或支付类 APP 的过程中，会自动产生大量的交易数据，从这些交易数据中可以挖掘出很多有用的信息。

（2）通信领域：当今社会，人和人之间的交流沟通方式很多，但无论采用哪一种方式进行通信，都会产生对应的通信数据，从这些数据中可以提取出大量有用的信息。例如，电话运营商都会经常分析用户的通话记录数据，从中挖掘出用户的相关行为模式。

（3）互联网领域：在互联网上存在有数量巨大的网页，每当用户访问这些网页时，都会在 Web 服务器上生成对应的访问日志，在电商类 Web 服务器上还会生成用户的消费行为数据。另外，互联网上的各网页之间的链接关系也构成了一种特殊的数据集。对这类数据进行挖掘，可以实现 Web 知识发现功能。

（4）物联网领域：物联网中的各种智能设备产生的数据，对于数据挖掘来说，意义也非常巨大。

3 数据挖掘过程

数据挖掘过程是一个线性过程，包含许多阶段，如数据清

收稿日期：2020-07-21
基金项目：本论文为 2018 年度湖北省教育厅科学技术研究计划项目指导性项目“基于云计算平台的高校学生到课率统计与分析系统开发”（编号：B2018453）阶段性成果。
作者简介：刘班（1978-），男，湖北省荆州人，工学硕士，讲师，研究方向：软件工程，Web 计算。

4 结语

本文基于专家系统与远程故障诊断技术实现了指挥通信装备维修支持系统，采用 B/S 架构为通信部（分）队的指挥通信装备故障汇总及上报、故障案例与故障知识的共享提供了方便快捷的手段，设计的故障知识库和案例库为指挥通信装备的故障诊断提供了有力支持，研究的案例检索、模糊推理与交互式推理为不同类型故障的辅助诊断提供了灵活的交互方式。系统经过测试与验证，可以满足部队应用需求。

参考文献：

[1] 马也. 军用通信电台故障诊断专家系统的设计与实现[D]. 东北大学,2008
[2] 陈维，陈永革，赵强. 基于 BP 神经网络的装备故障诊断专家系统研究[J].指挥控制与仿真,2008,Vol.30 No.4:103-105.
[3] 陈学文.野战地域网通信装备故障诊断专家系统研究[D]. 重庆:重庆大学, 2011.
[4] 赵佳，吕弘，刘宝，於拯威. 基于模糊 Petri 网的卫星导航接收系统故障诊断[J].测试技术学报,2017.Vol.31 No.5:438-442.
[5] 胡学东, 侯燕. 聚类分析方法在基于案例推理系统中的应用[J]. 微机发展,2004,Vol12: 32-35.
[6] 邓兴宇, 胡双演, 李钊, 等. 基于 SVSM 的装备故障案例相似度匹配算法[J]. 无线电工程,2016, Vol.46 No. 02: 31-35
[7] 赵磊磊. 基于专家系统的船舶电力系统故障诊断研究[J]. 舰船电子工程, 2018, Vol.38 No.9: 132-134

理、特征提取和算法设计等。典型的数据挖掘应用程序的工作流通常包括以下几个阶段：

(1) 数据采集：该阶段可能需要使用专门的硬件(如传感器网络)、手工劳动(如用户问卷调查)或软件工具(如 Web 爬虫引擎)来采集相关数据。虽然这个阶段是与应用类型密切相关的，并且通常超出了数据挖掘的范围，但这个阶段的工作非常重要，因为在这个阶段的工作完成情况可能会对后续的数据挖掘过程产生重大影响。在数据采集阶段之后，数据通常存储在数据库或数据仓库中，等待进行下一阶段的处理。

(2) 特征提取和数据清洗：当数据被采集到数据库或数据仓库中之后，它们的形式通常并不适合于数据挖掘工具进行直接处理。例如，采集到的部分数据可能缺少某个特征值，或者由于数据采集过程中的某方面原因，导致部分数据产生了错误。为了使数据更适合处理，必须将其转换为对数据挖掘算法友好的格式，如多维、时间序列或半结构化格式。多维格式是最常见的一种格式，其中数据的不同字段对应于被称为特性、特征或维的不同度量属性。挖掘过程中相关特征的提取是至关重要的，特征提取操作通常与数据清理操作同时进行，在操作过程中发现的数据缺失和数据错误部分会被补充或纠正。在很多情况下，数据可以从多个来源进行提取，然后使用计算机程序集成到统一的格式中进行处理，这一阶段的最终结果是一个具有良好结构的数据集。特征提取和数据清洗阶段结束之后，数据会再次存储到数据库中，为下一阶段的处理做好准备。

(3) 分析处理和结果输出：数据挖掘过程的最后一阶段是针对数据库中被处理的数据设计有效的分析方法和算法，或者选择合适的分析工具，应用统计方法、事例推理、决策树、规则推理、模糊集，甚至神经网络、遗传算法的方法处理信息，获得有用的分析结果。然后，从商业角度，由行业专家来验证数据挖掘结果的正确性。最后，将数据挖掘的分析结果以可视化的形式呈现给用户，或作为新的知识存入知识库中，供其它应用程序使用。

完整的数据挖掘流程如图 1 所示。

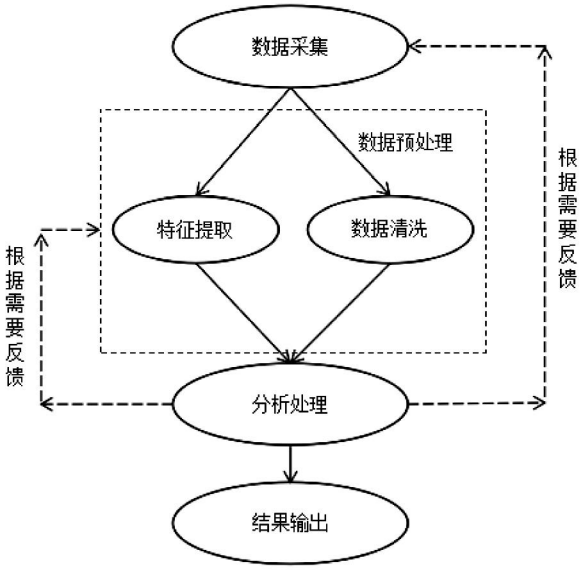

图 1 数据挖掘流程

4 Python 语言实现数据挖掘的优势

Python 是一种易于学习、功能强大的编程语言。它具有高效的数据结构和简单有效的面向对象编程方法。Python 优雅的语法和动态类型，加上其解释性质，使其成为在大多数平台上进行多领域脚本编制和快速应用程序开发的理想语言。另外，Python 提供了大量功能强大的第三方库，用以解决计算机各领域内的实际工作问题。在 Python 中提供的与数据挖掘相关的第三方库有：

(1) NumPy：一个功能强大的 Python 第三方库，允许更高级的数据操作和数学计算，主要用于对大型、多维数组和矩阵执行计算，广泛应用于数据挖掘的相关开发和研究工作中。NumPy 的前身 Numeric 最初是由 Jim Hugunin 和其他几个开发人员共同创建的。2005 年，Travis Oliphant 在 Numeric 中集成了另一个程序库 Numarray 的特色，并加入了其它扩展而开发了 NumPy。NumPy 目前由许多协作者以开源形式共同维护开发。

(2) SciPy：是 Python 中的一个用于数学、科学以及工程领域的开源第三方库，SciPy 依赖于 NumPy，用于快速计算 Numpy 矩阵。同时，SciPy 还提供了许多易于使用且能高效运行的数值例程，可以方便地处理插值、积分、优化、图像处理、常微分方程数值求解、信号处理等问题。SciPy 安装速度快，可以跨平台运行，功能强大，而且它的使用是完全免费的。

(3) Pandas：是一个基于 Numpy 的第三方数据分析库，为用户提供了快速、灵活和富有表现力的数据结构，从而可以简单、直观地处理关系型、标记型数据。Pandas 是专门为处理表格和混杂数据设计的，能够将数据预处理、清洗、分析工作变得更快更简单。经过多年的不断改进，Pandas 已经成为 Python 数据挖掘实践与实战的必备高级工具。

(4) Matplotlib：是用于 Python 和 Numpy 的绘图库。利用 Matplotlib 提供的一个面向对象的 API，可以将 Matplotlib 生成的图形嵌入到基于通用 GUI 框架(比如 wxPython、Qt 或 GTK+)开发的应用程序中。Matplotlib 中有一个基于状态机的过程式接口 pylab，该接口的设计非常类似于 MATLAB 的状态机，包含了 NumPy 和 Pyplot 模块中的很多常用函数，方便用户进行快速计算和绘图操作，十分适合在 IPython 交互式环境中使用。

(5) IPython：是 Python 的原生交互式 Shell 的增强版，支持变量自动补全和代码自动缩进，内置了许多很有用的功能和函数，同时还包含了强大的并行计算工具，是利用 Python 进行科学计算和可视化交互的一个强大解释器。

(6) Scikit-Learn：原名 scikits.learn，是 David Cournapeau 在 2007 年的 GoogleSummer of Code 中发起的一个项目。Scikit-Learn 的名称源于它是“SciKit”(SciPy 工具包)的概念。SciKit 是独立开发和分发的 SciPy 第三方扩展，内部实现了各种成熟的算法，需要 NumPy 和 SciPy 等其它包的支持，是 Python 中专门针对机器学习应用而发展起来的一款开源第三方库。

5 数据挖掘核心技术的 Python 实现

下面将通过几个简单的实例来论述基于 Python 语言及其第三方库来实现数据挖掘核心技术的方法。

5.1 数据分类

数据分类是监督学习的一种类型，能够实现根据训练数据集来识别被观察数据属于哪一组类别的功能。用于分类的最常用 Python 库是 Scikit-Learn。以水果分类为例，设计一个数据分类方案，能够根据水果的大小、颜色和形状特征数据，识别出水果的类型(比如：苹果，橘子，西瓜)。在正式分类前，可以先建立一个包含各种水果特征数据的训练数据集，然后将训练集中的水果特征数据映射到对应的水果类别下面。建立了这些映射关系之后，就可以通过分类算法建立相应的模

型。这时候如果出现了一个新的数据，此模型可以对该数据代表的水果进行预测，判断它到底属于哪一种水果。采用决策树分类器方法的 Python 实现代码如下所示：

```
fromsklearn import datasets
fromsklearn.metrics import confusion_matrix
fromsklearn.model_selection import train_test_split
iris = datasets.load_iris()
a = iris.data
b = iris.target
a_train, a_test, b_train, b_test = train_test_split(a, b, random_state = 0)
fromsklearn.tree import DecisionTreeClassifier
dtree_model = DecisionTreeClassifier (max_depth = 2) .fit (a_train, b_train)
dtree_predictions = dtree_model.predict(a_test)
c = confusion_matrix(b_test, dtree_predictions)
```

以上代码的执行时，首先加载数据集并将数据集分割成训练数据和测试数据两个部分，然后使用分类方法在训练数据上完成决策树分类器的生成，接着通过决策树分类器对测试数据对应的水果类型进行预测，最后计算出预测结果的精确度。

5.2 聚类

聚类是根据一定的标准，把一组数据对象分成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大于不在同一个簇内的数据对象的相似性，也就是要最大化簇内的相似性，并最小化簇与簇之间的相似性。与数据分类不同的是，聚类是一种无监督学习，最流行的聚类技术是 K-Means。以随机数数据集分类为例，使用 K-Means 算法根据特征对一个包含 50 个随机点的数据集进行分组，将这些随机点分为两组，具体代码如下所示：

```
importmatplotlib.pyplot as plot
fromsklearn.datasets import make_blobs
a, b = make_blobs(
n_samples=50, n_features=2,
centers=2)
km = KMeans(
n_clusters=2, init='random ',
n_init=5, max_iter=50,
tol=1e-04, random_state=0)
y_km = km.fit_predict(a)
```

以上代码运行时，先从样本点中选取 2 个质心作为聚类中心，然后将样本点分配给距离它们最近的质心，接着将质心移向分配给质心的样本点的中心位置，再重复上述步骤直到质心位置不再改变，最后将 Scikit-Learn 库中的 KMeans 类作用于测试数据集，并对测试数据集的分组情况进行预测。

5.3 回归

回归是一种基于自变量 (b) 来预测因变量 (a) 的值的监督学习算法，它计算输入和输出变量之间的线性关系，并在图上画一条直线。下面通过一个例子来演示如何使用 Python 第三方库 NumPy 来实现线性回归。

```
fromnumpy.random import rand
a = rand(10,1) # explanatory variable
b = a*a+rand(10,1)/5 # dependent variable
fromsklearn.linear_model import LinearRegression
linreg = LinearRegression()
linreg.fit(a,b)
fromnumpy import linspace, matrix
c = linspace(0,1,10)
plot(a,b, 'o ',c,linreg.predict(matrix(c).T), '- r ')
show()
```

6 结语

数据挖掘是一项强大的新技术，如同从大山中开采有价值的矿石一样，具有极大的潜力和价值。依靠数据挖掘技术可以预测行为和未来趋势，回答传统上过于耗时而无法解决的业务问题，找出专家可能会错过的预测信息，指导企业做出前瞻性的、知识驱动的决策。Python 语言的语法简单，易学易用，并且还包含了大量与数据挖掘相关的第三方库，因此，是实现数据挖掘技术的强有力的工具。可以预见，在 Python 的支持下，数据挖掘技术的实现难度将变得越来越低，从而会有更多的程序员进入数据挖掘领域进行软件开发，使数据挖掘相关软件的功能变得越来越强大。

参考文献：

[1] NongYe. Data Mining Theories, Algorithms, and Examples [M].USA:CRC Press,2014.

[2] Charu C. Aggarwal. DataMining: The Textbook[M]. USA: Springer Press,2015.

[3] Robert Layton. Learning Data Mining with Python [M]. USA:Packt Press,2015.

[4] Megan Squire. Mastering Data Mining with Python-Find patterns hidden in your data[M].USA:Packt Press,2016.

[5] 王岚, 王萍. 数据挖掘在远程教育系统中个性化教育的应用研究[J].计算机工程与科学,2008,30(10):93-95,124.

[6] 谭磊.大数据挖掘[M].北京:电子工业出版社,2013.

[7] 王蕾,戎杰.数据挖掘在金融风险管理中的应用[J].产业与科技论坛,2013,12(10):233-234.

[8] 沈立鸿.基于数据挖掘的中小地税机构税收预测系统分析与设计[D].厦门:厦门大学,2013.

[9] 吴超.浅谈电力企业电子档案数据挖掘与利用管理[J].中国高新技术企业,2015,(16):172-173.

[10] 顾洁,吕毅,沈炜,王怡平.小议大数据在安全播出上的应用[J].现代电视技术,2015,(3):118-119,140.

[11] 冯嘉关.基于数据挖掘技术的电信网络故障管理模型的研究和应用[D].北京:北京邮电大学,2016.

[12] 黄晓珊.湖南省高速公路道路养护管理系统的设计与实现[D].长沙:湖南大学,2018.

[13] 宋君.数据驱动的产业集群广义制造系统的产品综合评价方法研究[D].宁波:宁波大学,2018.

[14] 电子发烧友网.结合代码实例带你上手 python 数据挖掘和机器学习技术[EB/OL].http://www.elecfans.com/d/659869.html, 2018.

[15] 郭有庆.基于MapReduce的频繁项集挖掘算法研究[D].重庆:重庆邮电大学,2019.

[16] 李生.基于 Storm 的分布式流数据关联规则挖掘[D].镇江:江苏大学,2019.

[17] 高风.基于数据挖掘的中医药治疗慢性心力衰竭用药规律研究[D].北京:中国中医药科学院,2019.

[18] 黄妙红,何胜,王珏,肖嘉丽.移动技术在审计中的应用[J].电子技术与软件工程,2020,(6):192-194.