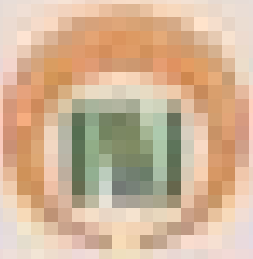
数据挖掘技术及其应用

杨杰 姚莉秀 编著

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS



[redacted]

[redacted]

上海交通大学学术出版基金资助项目

数据挖掘技术及其应用

杨　杰　姚莉秀　编著

上海交通大学出版社

内 容 提 要

本书系统地讲述了数据挖掘的基本概念和基本原理，并列举了在相应领域具有参考价值的算法及其改进和应用，是作者多年来从事教学和科研实践的成果。全书共9章，主要内容有：数据挖掘的基本概念和原理，数据预处理，各种分类、聚类和关联规则提取算法，以及在生物信息学、材料学中的实际应用案例。

本书可用作计算机专业本科高年级学生或研究生的教材或参考书，也可供从事计算机信息处理、数据挖掘、工业优化等有关方面工作的科技人员参考。

图书在版编目(CIP)数据

数据挖掘技术及其应用/杨杰，姚莉秀编著. —上海：上海交通大学出版社，2011

ISBN 978-7-313-06610-7

Ⅰ. 数... Ⅱ. ① 杨... ② 姚... Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2010)第 123090 号

数据挖掘技术及其应用

杨 杰 姚莉秀 **编著**

上海交通大学出版社出版发行

(上海市番禺路 951 号 邮政编码 200030)

电话：64071208 出版人：韩建民

昆山市亭林印刷有限责任公司 印刷 全国新华书店经销

开本：787mm×960mm 1/16 印张：17 字数：319 千字

2011 年 1 月第 1 版 2011 年 1 月第 1 次印刷

ISBN 978-7-313-06610-7/TP 定价：98.00 元

前　言

对于复杂对象(一般而言其特点是:非线性、多因子、高噪声、非高斯分布、各自变量间有强相关)的智能信息处理经常存在着知识获取瓶颈。从“第一原理”出发直接进行数学建模和计算机仿真存在困难。另外为建立数学模型常常在模型和边界条件上进行简化,从书本上或专家那里得到的通用模型和经验规则,并不能保证完全符合对象的特定环境。随着计算机、数据库和传感技术的快速发展,能方便地从复杂对象中获取实时和历史大量传感数据。显然根据研究对象的大量观测数据进行建模和优化更能符合对象的特定环境,因此如何依据大量观测数据进行建模和优化成为关键技术。数据挖掘技术是当今智能系统理论和技术的重要研究内容,它能从大量数据中挖掘和学习有价值的隐含知识,因而近年来得到国内外的极大重视和研究。数据挖掘和知识发现技术已应用于工业、商业、金融、医学、行政管理等领域,如:模糊控制器的建模、故障诊断的建模、DNA 序列分析、金融数据预测、关联特征分析等。美国的信用卡公司和银行通过从客户数据库中挖掘和发现信用卡欺诈或贷款坏账对象的特征,从而调整其相关政策以减少金融风险。

本书介绍了数据挖掘的基础理论和改进算法以及最新研究成果和动态,包括:数据预处理、关联规则提取与粗糙集、数据挖掘分类与回归、聚类分析。数据挖掘已广泛而有效地应用于许多领域,本书重点介绍我们在生物信息学和材料科学的应用实践。

近十年来,著者本着“借鉴—创新—实践应用”的指导思想,在数据挖掘理论和应用研究方面开展了不懈的探索研究。本书综合了著者在数据挖掘理论和应用研究过程中的最新研究成果。在研究过程中先后得到 10 多项国家和省部级科研项目的资助,包括:国家自然科学基金项目“基于数据挖掘和综合模型的脑磁共振图像分析和诊断”、“面向钢铁生产的数据挖掘和数据融合信息处理平台及应用”、国家 863 计划项目“面向建模、优化的数据挖掘技术”、科技部政府间重点国际合作项目“智能机器人的脑功能开发”、教育部跨世纪优秀人才培养计划项目“基于综合策略的智能信息处理软件平台的理论和应用研究”。研究成果先后获上海市自然科学三等奖和教育部科技进步二等奖。没有这些项目的资助,著者不可能取得数据挖掘理论和应用研究的目前成果,也不可能完成本书的撰写。

本书由杨杰、姚莉秀负责执笔,杨杰负责审定所有书稿内容,书稿内容整合了著者负责的实验室在数据挖掘理论和应用研究方面的最新研究成果,整合了所指

导的研究生(博士生:叶晨洲、全勇、王猛、刘惠、王向阳、沈红斌;硕士生:黄欣、薛莉、刘国平)的学位论文内容。因此本书是一项热情与和谐气氛的集体创作。

著者在数据挖掘理论和应用研究过程中与中科院上海冶金所(现微系统所)研究员及上海交通大学兼职教授陈念贻先生开展多年合作,共同承担了国家863计划项目并合著了一本专著。陈先生最早将数据挖掘技术应用于我国的冶金和材料领域,获得国家自然科学奖等多项成果奖。陈先生勇于创新、诲人不倦、鞠躬尽瘁的品质是著者的楷模,他的逝世是我国学术界的一大损失。本书的出版是对他的纪念。

著者在数据挖掘理论和应用研究过程以及写作过程中得到了国内模式识别与智能系统专业的元老李介谷教授、著名学者施鹏飞教授的支持和指导,著者杨杰在德国汉堡大学攻读博士学位和开展数据挖掘理论和应用研究时,得到导师国际人工智能领域的权威专家 Bernd Neumann 教授(曾担任欧洲人工智能学会主席,全球计算机大学主席)的指导和帮助。在本书的写作过程中,博士生朱林和硕士生秦蓓蓓、王强、李娟、王小念、常青青帮助编辑排版。在此一并表示感谢!最后,感谢上海交通大学学术专著出版基金的资助以及上海交通大学出版社编辑的出色工作,使得本书得以顺利出版。

杨杰　姚莉秀

2010年1月18日

目　录

第1章 导 论

1.1 数据挖掘技术的源起与发展

近十几年，随着科学技术飞速的发展，经济和社会都取得了极大的进步，与此同时，在各个领域产生了大量的数据，如人类对太空的探索、银行每天的巨额交易数据等。显然在这些数据中蕴涵着大量的信息，如何处理这些数据得到有益的信息，人们进行了很多的研究探索。计算机技术的迅速发展使得处理数据成为可能，这就推动了数据库技术的极大发展，但是面对如潮水般不断增加的数据，人们不再满足于数据库的查询功能，而是提出了深层次问题：能不能从数据中提取信息或者知识为决策服务。也就是说如何从底层的数据转变成一种知识(图 1-1)。就数据库技术而言已经显得无能为力了。

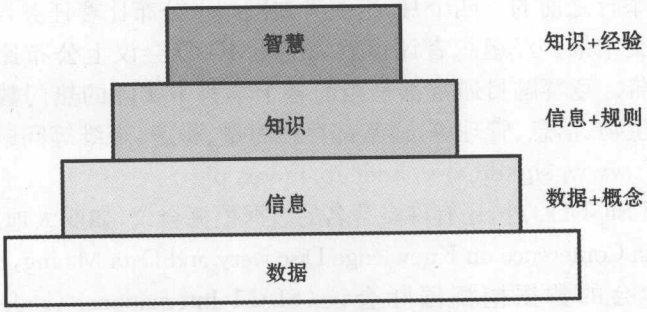


图 1-1 数据金字塔

其实，数据库本身也是一个发展的过程，从开始的原始数据文件处理到数据库系统再到具有索引、查询等复杂功能的数据库管理系统，再发展到更高级的数据库，如数据仓库、基于 Internet 的全球化信息等。数据越来越多，数据库越来越复杂，功能越来越强，但还是偏向于管理，要想发现数据中的关联和规则，或是根据现有的复杂数据预测未来的发展趋势还是不可能。所以迫切需要一种能自动地把数据转换成有用信息和知识的技术和工具。

为研究和解决上述问题，1989 年在美国底特律召开的 11 届国际人工智能联合会议的专题讨论会上，首次提出了知识发现(Knowledge Discovery in Database，KDD)的概念，内容涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据

可视化、高性能计算、专家系统等领域。但由于内涵太广，理论和技术难度很大，使KDD技术一时难以满足需要。1995年在Canada蒙特利尔召开的第一届知识发现学术会议上，人们开始重新认识数据、认识存储、认识数据统计和分析。美国计算机学会(ACM)于当年提出了数据挖掘(Data Mining，DM)的概念，把大型数据库看成是存放有价值信息的矿藏。引发了知识发现和数据挖掘理论及应用研究的热潮。随后，各类KDD会议、研讨会纷纷涌现。IEEE的Knowledge and Data Engineering会刊率先在1993年出版了KDD技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论，到了脍炙人口的程度。在生物、化学化工、材料、纳米等领域的会议、期刊也都把数据挖掘、知识发现等列为重点的专题和专刊讨论。

从1995年开始，知识发现与数据挖掘国际性会议(ACM Special Interest Group on Knowledge Discovery and Data Mining，ACM SIGKDD)每年一次在欧美地区召开(http://www.sigkdd.org/)，每次都会有许多国家的相关领域研究人员参与，互相交流最新进展。此外，1997年起，美国计算机协会(ACM)和会议举办方一起，举办“KDD cup”的知识发现和数据挖掘国际竞赛，向数据挖掘领域的学术界和工业界开放，以期找出最有创新性和最有效的技术与方法。在每年的ACM SIGKDD会议举行之前的三四个月，组织者在网站上公布比赛任务，参赛者必须在指定日期前提交结果。经组织者评审后，在SIGKDD会议上公布比赛结果，并邀请优胜者作报告。竞赛题目通常都是当前各个学科中实际的热门数据挖掘问题，如经济、制药、生物、信息、管理等，通常都存在海量、噪声、高维等问题，具有较高的挑战性(http://www.sigkdd.org/kddcup/index.php)。

除了ACM SIGKDD外，还有许多著名的数据挖掘会议，如亚太地区数据挖掘会议(Pacific—Asia Conference on Knowledge Discovery and Data Mining，PAKDD)，工业和应用数学学会的数据挖掘国际会议(SIAM International Conference on Data Mining)(http://www.siam.org/meetings/)等。中国计算机学会人工智能与模式识别专委会也于2009年在原中国分类技术与应用研讨会基础上扩展为中国数据挖掘会议(CCDM，China Conference on Data Mining)，为学术界和工业界的广大研究人员提供一个交流、合作平台，使得研究人员之间可以更广泛地分享数据挖掘与知识发现领域的初创性研究成果、创新思想、最新研究进展以及系统开发经验。

1.2 数据挖掘的概念

1) 数据挖掘的定义

数据挖掘的历史虽然不长，但从20世纪90年代以来，它的发展速度很快，加

之它是多学科综合的产物，目前还没有一个完整的定义，人们多是根据研究方法与应用对象提出数据挖掘的定义，例如：

SAS研究所(1997)认为数据挖掘是“在大量相关数据基础之上进行数据探索和建立相关模型的先进方法”。

Bhavani(1999)认为数据挖掘是“使用模式识别技术、统计和数学技术，在大量的数据中发现有意义的新关系、模式和趋势的过程”。

韩家炜等人(2000)认为“数据挖掘就是在大型数据库中寻找有意义、有价值信息的过程”。

现今资料流通量之巨大已到了令人咂舌的地步，就实际限制而言，就遇到了诸如巨量的纪录，高维的资料增加了传统分析技术上的困难，且搜集到的资料仅有5%至10%用来分析，以及资料搜集过程中并不探讨特性、存在冗余或噪声等问题，这就让我们不得不利用数据挖掘技术。

所以，我们更倾向于韩家炜先生的关于数据挖掘的定义[1]，即数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，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。这个定义包括了好几层含义：① 数据源必须是大量的、真实的，真实的数据往往含有噪声或缺失；② 发现的是用户感兴趣的知识；③ 发现的知识要可接受、可理解、可运用，能支持特定的发现问题，能够支持决策，可以为企业带来利益，或者为科学研究寻找突破口。

2) 数据的分类

数据是指一个有关事实的集合，用来描述事物有关方面的信息，是我们进一步发现知识的原材料。数据类型除了一般的静态的数值型的数据外，还包括时间序列数据、空间数据、文本数据、多媒体数据等。

(1) 时间序列数据，是与时间有关的一系列数据，可以进一步分为时间相关数据和序列相关数据。时间相关数据与数据产生的绝对时间有关，如股票价格、银行账务、设备运行日志、某地区每日不同时间段的自来水用水量等；序列相关数据与数据产生的绝对时间关系不大，而注重数据间的先后次序。典型的序列相关数据有传感器输出数据、生物信息中的蛋白质、DNA序列数据等。

(2) 空间数据，是与空间位置或地理信息有关的数据，如二维或三维图像数据、地理信息系统GIS数据、人口普查数据等。

(3) 文本数据，就是我们应用的一般文字，如报刊杂志、设备维护手册、故障描述等的内容。对文本数据的挖掘主要是发现某些文字出现的规律以及文字与语义、语法间的联系，用于自然语言处理，如机器翻译、语音识别、信息检索等。当前一个十分活跃的研究方向是Web日志(Web log)的挖掘，目的是有效发现Internet

用户访问站点的模式，从而提高服务的针对性。

（4）多媒体数据，是随着多媒体技术而日益涌现的声音、图形、图像、超文本等数据。

1.3 数据挖掘的过程

图 1-2 显示了从海量的数据中通过数据挖掘等技术形成知识的过程。包含以下几个步骤：

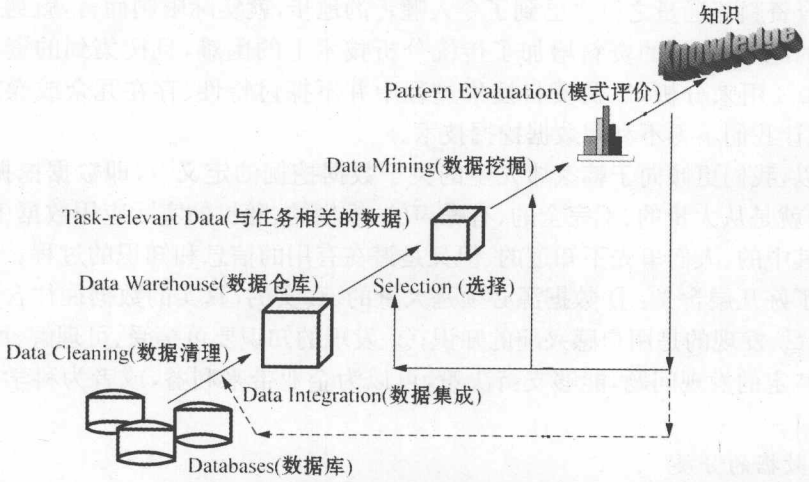


图 1-2 数据转化成知识过程

步骤 1 数据的准备：这是问题的提出和数据的选取部分。必须根据所要解决的问题，确定待挖掘的目标，并收集与解决该问题有关的数据供后续处理。例如有一个挖掘目标是提高某产品的产量或质量，我们就有必要收集整个生产流程工艺数据、原材料配比数据等，用于数据挖掘以获取规则，指导生产。

步骤 2 数据预处理：在数据挖掘的定义中已经说明，数据挖掘的数据来源于实际记录，由于客观或主观原因，经常存在数据缺失、数据噪声、高维、数据来源不一致导致的数据重复冗余等问题。为了后续的数据挖掘工作能够顺利有效进行，获得的知识更具代表性，更具指导作用，我们有必要对原始数据记录进行清理、去噪、降维等预处理工作。预处理的成功与否很大程度上决定了后面的知识获取的有效程度。

步骤 3 狭义的数据挖掘：广义的数据挖掘包括了从数据收集到指导决策的过程，但在整个过程中，人们习惯把用各种数据挖掘方法对预处理后的数据提取有用信息的步骤也称为数据挖掘，我们称其为狭义的数据挖掘。

步骤4 模式的评价：主要指对上述步骤中提取的信息知识模式的有效性、可靠性、泛化能力等进行评价。任何获取的知识模式，要经过评价检验才能判定是否有效。不具指导性的知识是没有意义的。

在用多种数据挖掘方法解决问题时，应该同时考虑其可靠程度、繁简程度、可理解性等。

1.4 数据挖掘的功能

数据挖掘通过对现有数据记录的分析，预测未来趋势及行为，做出基于知识的决策。数据挖掘的目标是从大量数据中发现隐含的、有意义的知识，主要有以下几类功能。

1）概念描述

被分析的数据称为目标数据集，像程序设计语言C++里面的对象与类一样，对于一个数据集，我们可以通过其共有的属性和行为来描述它，最终获得简明、准确的描述，使之更具代表性。

2）相关分析（关联分析）

就是从给定的数据集发现频繁出现的项集模式知识，即发现各属性之间的关联关系并用关联规则描述出来。

3）分类和回归预报

根据一系列已知数据，训练产生一套能描述或区别数据的类别或概念的模型，并能够根据这个模型来预测未知数据的结果。人脸识别、指纹识别、商业中的客户识别分类、工业上故障诊断等都是分类问题。

4）聚类

根据物以类聚原则，利用属性特征将数据集合分成为由类似的数据组成的多个类的过程称为聚类。聚类后，同一类之间的数据具有很强的相似性而非同类之间的数据具有很强的非相似性。

5）趋势分析

实现前面提到的四种功能时，事件产生的顺序信息都被忽略，被简单地作为一条静态的记录来对待。而趋势分析是对随时间变化的数据对象的变化规律和趋势进行建模描述，根据前一段时间的运动预测下一个时间点的状态。解决的问题一般可分为两类：

（1）总结数据的序列或者变化趋势，如预测股票/期货交易，网页点击顺序记录等。

（2）检测数据随时间变化的变化；如自来水厂用水量的日、周、月、年等周期

变化。

6）离群点的分析

数据挖掘的功能大多致力于建立模型，根据模型对未知数据进行预测。除此之外，数据挖掘还可用于模式探测，即用统计、距离测量等方法来寻找明显不同于其他点的数据，判断是否是离群点。

离群点的检测对于调查商业欺诈、偷税漏税等行为尤其有用，如，税务稽查机关在选案过程中，先根据企业的规模、行业、地区等特性找出若干相关企业的数据，对于税额明显偏低的企业，就可以通过数据挖掘方法以离群点分析的方式检测出来。

7）复杂类型数据的挖掘

随着数据处理工具、先进数据库技术以及万维网技术的迅速发展，大量的形式各异的复杂类型数据（如超文本、视频、多媒体、时间序列数据等）不断涌现。因此复杂数据的挖掘也是数据挖掘面临的一个重要课题。

1.5 数据挖掘的典型应用领域

与数据挖掘功能相对应的，数据挖掘技术可以应用到各个有数据记录的领域，本书只描述几个重要的应用领域。

1）销售，市场营销

数据挖掘的最早应用领域之一是市场营销，“啤酒/尿布”问题至今被人们津津乐道。为分析顾客最有可能一起购买什么商品，美国加州某公司利用数据挖掘工具，对数据库中的大量数据进行分析，竟然发现跟尿布一起购买最多的商品是啤酒。获得如此关联规则，我们可以给出合理的解释：在家带孩子的太太们会叮嘱她们的丈夫，下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了两瓶啤酒。既然尿布与啤酒一起购买的机会很多，公司给出的决策就是将他们摆放在一起，通过之后啤酒与尿布销量的双双增长可验证该规律的有效性。

2）生物信息处理

生物信息学是21世纪生物学的必然[2,3,4]。人类为了更深入地了解和认识自身，制定了宏伟的人类基因组计划，产生了大量的生物分子数据。充分利用这些数据，通过数据分析、处理，揭示这些数据的内涵，从而得到对人类有用的信息，在指导试验、精心设计试验方面发挥重要作用，是生物学家、数学家和计算机科学家所面临的一个严峻挑战，是数据挖掘应用的新领域。

3）工业控制

随着工业上工艺调控、数据采集等过程的自动化技术不断提高，利用数据挖掘

技术对相关数据进行分析，获取规则，指导工艺调节，实现最优的过程控制也成了可能。例如研究者们可通过对有缺陷和无缺陷样本的学习，提取相关规则和识别规则，鉴别产品制造过程中的缺陷，管理由异常行为引起的不良结果等，实现故障诊断或过程优化。

4）保险、银行、证券等金融领域

金融领域存有大量的客户信息记录、自身服务记录等，可用数据挖掘技术分析客户需求和兴趣，银行方面可以预测存、贷款趋势，优化存、贷款利率，或者推出更多的有意义的服务等。保险方面可以通过分析保险客户的要求、信誉，防范风险。

1.6 目前国际上流行的数据挖掘软件

Knowledge and Data Engineering，Pattern Analysis and Machine Intelligence 是目前国际上最有影响的数据挖掘期刊。此外，在 Internet 上还有一些 KDD 电子出版物，http://www.kdnuggets.com 上会提供一些进展报告，也可以下载各种各样的数据挖掘工具软件和典型的样本数据，供人们测试和评价。国内也有一些数据挖掘技术交流网站，如 http://www.dmgroup.org.cn 和 http://www.dmresearch.net。

目前，世界上比较有影响的商业数据挖掘系统有 SAS 公司的 Enterprise Miner（http://www.sas.com）、SPSS 公司的 Clementine（http://www.spss.com）、IBM 公司的 Intelligent Miner、SGI 公司的 SetMiner、Sybase 公司的 Warehouse Studio、RuleQuest Research 公司的 See5 等。主要的实验数据挖掘系统有加拿大 Simon Fraser 大学的 DBMiner、新加坡国立大学的 CBA 和 IAS、德国 Dortmund 大学的 MiningMart 等等。人们还可访问 http://www.datamininglab.com 网站，该网站提供了许多数据挖掘系统和工具的性能测试报告。

参考文献

[1] 韩家炜，坎伯. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2007.

[2] 李衍达. 与信息科学的结合为生命科学的研究开辟新的前景[J]. 中国科学基金，1999，13(5)：307-308.

[3] Chou K C. Review：Structural bioinformatics and its impact to biomedical science [J]. Current Medicinal Chemistry，2004，11：2105-2134.

[4] 孙啸，陆祖宏，谢建明. 生物信息学基础[M]. 北京：清华大学出版社，2005.

第2章　数据预处理

数据挖掘技术处理的是大量的日常业务数据。原始业务数据是知识和信息提取的源泉，但同时实际应用系统中收集到的原始业务数据通常存在噪声、数据缺失、记录不一致等现象。这些现象的存在势必对数据挖掘的信息规则提取产生干扰。所以为提高数据的质量，便于得到更好、更有效的数据挖掘结果，有必要对其进行预处理。

预处理的目标就是接受并理解客户的挖掘需求，确定挖掘任务，抽取与任务相关的知识源，根据背景知识中的约束性规则对原始数据进行检查，通过清理和归纳等操作，生成供挖掘核心算法使用的目标数据。目标数据汇集了原始数据库中与发现任务相关的所有数据的总体特征，在数据挖掘处理这些数据时即可用如下矩阵形式表示：

$$\boldsymbol{X} = (x_{ij})_{N\times M} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

矩阵中每一行表示一个样本(或元组)，每一列表示一个特征(或属性)，其中 N 为样本数，M 为特征数，故 $\boldsymbol{X}$ 为 $N\times M$ 阶矩阵，x_{ij} 为第 i 个样本的第 j 个特征参数。

根据面向的要处理的问题，预处理可以分为数据清理、数据集成、数据转换和数据约简四个部分。

2.1　数据清理

数据清理主要解决样本的不完整、噪声和不一致的问题，以优化样本，提高其后的挖掘过程的精度和性能，主要处理如下两类情况。

2.1.1　缺失数据的处理

数据挖掘面向的是实际应用数据，由于实际生产或记录过程中，总会有一些意想不到的特殊情况发生，如仪器故障、工人疏忽等，导致一些记录的缺失。基于此，有必要对缺失数据进行预处理。