数据流分类

胡学钢 李培培 张玉红 吴信东 著

清华大学出版社

数据流分类

胡学钢 李培培 张玉红 吴信东 著

清华大学出版社
北京

内 容 简 介

本书阐述了数据流分类问题的基础理论、技术方法以及应用实践，为面向实际数据流开展分类数据挖掘任务提供了理论与实践基础。全书共分四篇12章。第一篇是引言篇，本篇首先简介数据挖掘的相关概念，然后介绍数据流挖掘的相关定义、应用背景及理论基础与技术，最后重点总结数据流分类挖掘的主要研究进展并归纳了存在的关键问题；第二篇是基础篇，本篇主要阐述了分类挖掘任务中常用的模型与技术，为后续数据流分类方法提供技术基础；第三篇是专题篇：本篇首先总结分析适宜于数据流环境的几种集成模型，并通过两个示例讲解了基于加权集成模型的数据流分类算法的应用。然后详细介绍若干数据流的概念漂移检测与分类方法、不完全标记数据流分类方法以及面向实际应用数据的特征选择方法，并通过在模拟与实际数据上的大量实验考察了这些方法的分类性能；第四篇是实验资源篇，本篇首先介绍数据流分类算法实验工具ETDSv1.0的功能与用户使用说明，然后归纳总结目前流行的面向数据流环境的实验平台以及在数据流分类任务中常用的数据集。

本书在数据流中概念漂移检测问题、不完全标记问题、特征选择等方面有许多独到见解，总结归纳了近年来在数据流分类任务上的研究成果，并归纳提炼了数据流分类研究任务中存在的重要开放性问题。

本书可作为计算机软件与理论、计算机应用类的研究生教材，也可供对数据流挖掘等领域感兴趣的相关教师、本科生、研究生以及科技工作者参考。另外，本书介绍的相关实验软件平台已开源，可为从事数据流挖掘等方向的科研工作者提供实践与二次开发平台。

图书在版编目（CIP）数据

数据流分类/胡学钢等著. --北京：清华大学出版社，2015
ISBN 978-7-302-40599-3

Ⅰ. ①数…　Ⅱ. ①胡…　Ⅲ. ①数据采集－研究　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2015)第150154号

责任编辑：袁勤勇　徐跃进
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：25　　**字　数**：603千字
版　　次：2015年12月第1版　　**印　次**：2015年12月第1次印刷
印　　数：1～1000
定　　价：49.00元

产品编号：052683-01

序 言

数据挖掘是一个多学科交叉的领域，涉及统计学、信息安全、数据库理论、管理信息学、模式识别、人工智能、机器学习、信息检索、图像与信号处理等多个领域，这些领域内的专家以各自不同的专业视角去关注数据挖掘技术（诸如分类、预测、关联分析、时序分析、聚类等）的进展。虽然数据挖掘已在许多领域开展了应用并取得了成效，然而随着信息技术的高速发展，实际应用领域中的信息获取变得更加容易，生成数据的速率变得越来越快，这种快速的、连续的数据流不断产生必然给计算机系统带来存储空间、计算速度以及通信能力等方面的挑战，同时也使传统的数据挖掘技术面临着新的挑战与机遇。近几年与数据流相关的问题研究已成为数据挖掘领域的研究热点，受到越来越多的研究人员关注。在顶尖国际会议（如 SIGKDD、SIGMOD 和 VLDB 等）上，都为数据流研究开辟了专栏，而 ICDM、KDD、AAAI 中每年也都有大量论文关注这一方向。

经过近十几年的发展，有关数据流的数据挖掘问题在国际上取得了一定的研究成果，分别在 2005—2010 年间陆续出版了 5 部数据流相关专著。相比之下，国内研究工作起步较晚，至今尚未有与数据流相关的中文专著。综合考虑国内外的研究进展，虽然国外出版了几部数据流相关的专著，然而多关注的是数据流的大纲数据与数据管理、面向数据流的聚类、频繁项集挖掘、模式挖掘等数据挖掘任务以及面向具体应用数据（如传感数据流）的数据流相关处理技术研究，尚缺乏一本系统地针对数据流分类问题的专著。从研究现状来看，本书系统地研究了数据流分类问题，是一本面向数据流挖掘领域新的研究成果的学术专著。

实际应用领域涌现的数据流，如电信通话记录、网上购物交易、网络搜索请求、股票市场交易、交通流量监测数据、卫星探测和天文观测科学研究数据等隐含着丰富的、有价值的知识亟待挖掘。然而，与传统数据形式不同，这些数据流具有连续性、多变性、快速性、无限性等特点，因而如何设计面向数据流环境的分类模型与方法成为极具挑战而重要的研究内容之一。再加上实际应用领域概念漂移现象的客观存在，因而如何检测数据流的时序变化，发现服务质量等概念漂移的规律，发掘有趣的行为模式也成为数据流分类任务中的挑战问题之一。另外，实际应用中数据流的类标签大量缺失、属性高维稀疏、噪音等问题也大大加大了数据流分类任务开展的难度。从本书中，读者可以获得以上问题的解决思路，甚至是具体的解决案例。

从本书的内容上看，一是归纳总结了数十年来针对数据流分类的研究成果，这为从事数据流及其相关方面研究的读者提供了吸收与消化文献的参考思路；二是系统地阐述了数据流分类任务的背景、理论方法与实际应用，并重点以数据流分类中几个重要问题为例，给出了作者自己的观点与方案，为读者们较全面描述了数据流分类的关键科学研究内容，同时也为从事相关数据流挖掘的科技人员提供了理论与实践参考；三是根据作者多年研究经验，归

纳总结了数据流分类中存在的几大开放性问题并提出了自己独到的见解，在一定程度上为该领域的研究者指明了前进的方向，这将有助于推动该领域的研究进展；四是公开了所设计的数据流分类算法的软件实验平台接口与相关代码，这些资源的公开增添了数据流分类的开源资料，为数据流挖掘等方向的科研工作者提供了更多可利用的实验资源。

本书展示了针对若干数据流分类中的开放性问题的探索性成果，也许这些结论并不完美，但本人认为，胡学钢等所著的这本书是一块针对数据流分类研究的敲门砖，相信它的出版能够为该领域的研究者带来理论与实践上的启发，起到抛砖引玉的效果。同时，我也想对每一位有志于从事这方面研究的青年学者说一句：希望大家在以后的研究工作中，继续深入从理论与实践两方面研究数据流相关任务的开放性问题，不断取得新的研究成果，为我国数据挖掘及科学技术的发展做出更多的贡献。

苗夺谦

2015 年 8 月于同济大学

前 言

随着电信、网络等信息技术的迅速发展，实际应用领域涌现了大量的数据流，如电信通话记录、网上购物交易、网络搜索请求、股票市场交易、交通流量监测数据、卫星探测和天文观测科学研究数据等。一方面这些数据流中隐含着丰富的、有价值的知识亟待挖掘。另一方面分类问题研究作为数据挖掘领域的一个重要分支，在预防信用卡欺诈、网络入侵发现等应用中具有重要的研究意义。然而，与传统数据形式不同，这些数据流具有连续性、多变性、快速性、无限性等特点，如果仍采用传统的分类挖掘模型与算法加以处理，大量有用信息将会丢失，因而设计面向数据流环境的分类模型与方法成为极具挑战而重要的研究领域之一。

数据流除具有高速、连续、多变、无限特性外，其最典型的特征则是隐含其中的概念漂移，例如，顾客网上购物偏好随个人兴趣、商品特性、商家信誉、服务类型等因素的改变而变化。而现实世界数据流中这种概念漂移现象是客观存在的，因此研究数据流的时序变化，发现服务质量等概念漂移的规律，发掘有趣的行为模式成为数据流分类任务中的挑战研究内容之一。同时，实际应用中数据流的类标签大量缺失的现象也客观存在，例如：Web 中存在着大量无标签网页；入侵检测领域网络包以及网上产品评论数据的类标签也是大量未标记的。这些仅含少量有标签示例而存在大量无标签示例的不完全标记数据，使得传统的数据挖掘分类算法，甚至已有的数据流分类算法面临严峻挑战。因为一方面，这些算法总是假定训练数据具有完整的类标签，这一假定显然难以满足实际应用数据处理的需求。另一方面，如果简单地选择丢弃无标签示例的信息仅利用少量有标签的示例构建分类器，又会因为有标签训练示例不足难以构建具有泛化能力的分类模型，导致较低的分类精度。因此，如何利用无标签示例辅助少量有标签示例学习以构建具有强泛化能力的在线分类方法也成为现实应用领域数据流分类问题的又一挑战性的研究内容之一。另外，数据流分类任务中还面临着数据流高维稀疏问题、数据分布不平衡问题、噪音问题等众多挑战问题，这些都应该是未来数据流分类任务所应重点关注的极具挑战的研究内容。

最经典的数据流分类方法是由 Domingos 与 Hulten 于 2000 年在 KDD 国际会议上提出的基于增量式决策树模型的数据流分类算法 VFDT(Very-Fast-Decision-Tree learner)。经过十几年的发展，数据流分类问题在国际上取得了一定的研究成果，如 Muthukrishnan 于 2005 年出版了关于“数据流算法与应用(Data Streams：Algorithms and Applications)”的专著，Aggarwak 于 2007 年出版了关于“数据流模型与算法(Data Streams：Models and Algorithms)”的专著，Gama 等于 2007 年、2010 年分别出版了“面向传感器网络处理技术的数据流学习问题研究(Learning from Data Stream：Processing Techniques in Sensor Networks)”与“面向数据流的知识发现问题研究(Knowledge Discovery from Data Streams)”的专著，Bifet 于 2010 年发表了“针对进化数据流的模式识别与数据挖掘的适应

性流挖掘问题研究(Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams)"的专著。相比之下,国内的研究工作起步较晚,研究工作也仅限于西北工业大学、中科院等若干高校与研究机构,而至今尚未有数据流相关的中文专著。综合考虑国内外的研究进展,虽然国外出版了几部与数据流分类相关的专著,然而多关注的是数据流的大纲数据与数据管理、面向数据流的聚类、频繁项集挖掘、模式挖掘等数据挖掘任务以及面向具体应用数据(如传感数据流)的数据流相关处理技术研究,尚缺乏一本系统地针对数据流分类问题的专著。本书系统阐述了数据流分类任务的背景、理论方法与实际应用,归纳总结了十几年来针对数据流分类任务的研究成果,重点介绍了数据流分类任务中几个重要挑战问题的研究进展,并探讨了数据流分类任务中的开放性研究问题。

本书在总结归纳过去研究工作的基础上阐述了数据流分类问题的基础理论、技术方法以及应用实践,为面向实际数据流开展分类挖掘任务提供了理论与实践基础。全书共分四篇 12 章,第一篇是对数据挖掘的大背景、数据流挖掘的背景以及数据流分类问题研究的挑战性问题的描述与归纳,包括第 1、2、3 章;第二篇是对分类模型与方法以及特征选择方法的介绍,包括第 4、5 章;第三篇详述了数据流分类任务中的若干挑战问题的研究成果,包括第 6、7、8、9 章;第四篇介绍了作者设计的数据流分类实验工具,并总结了已有的若干实验平台与常用的数据集,包括第 10、11、12 章。本书由胡学钢统一规划设计,由胡学钢、李培培、张玉红、吴信东共同撰写。全书最后由胡学钢统稿,李培培校稿。

第 1 章是数据挖掘导论,是对全书将要描述的问题领域的铺垫。介绍了数据挖掘的相关概念、任务及其应用热点。

第 2 章是数据流挖掘相关知识介绍。阐述了数据流挖掘的背景、相关定义、应用领域、数据流处理的理论基础与技术以及数据流挖掘的任务与挑战。

第 3 章是数据流分类的关键研究问题。层层铺垫,引出了本书的主题,同时总结了数据流分类问题中的七大关键研究问题,包括合理的概念描述模型及数据流大纲数据提取机制、数据流中概念漂移检测问题研究、噪音数据流问题研究、数据分布不平衡问题研究、不完全标记数据流、高维数据流。

第 4 章是分类模型与方法。阐述了作为数据挖掘的任务之一的分类的基本知识、常用的分类模型以及评估方法等。

第 5 章是特征选择。介绍了特征选择的背景、相关定义、经典的特征选择方法以及所面临的挑战问题。

第 6 章是数据流的集成分类方法研究。概述了数据流分类中常用的集成分类模型,即 WE 集成方法、AP 集成方法、WE 与 AP 的混合方法以及基于 WE 的混合集成方法,同时,重点介绍了两个基于 WE 集成方法的典型算法。

第 7 章是数据流中概念漂移检测与分类问题研究。概述了基于增量式决策树模型的概念漂移检测与分类方法,同时详细描述了若干基于漂移检测机制与决策树改进模型的数据流分类方法。

第 8 章是不完全标记数据流分类问题研究。总结了不完全标记数据流的处理技术、不完全标记数据流中的概念漂移检测包括重现概念漂移检测方法,同时详述了基于不完全标记数据流处理技术与概念漂移检测方法设计而成的分类算法,最后探讨了不完全标记数据流中的开放性问题。

第 9 章是面向应用数据的特征降维方法研究。重点介绍了面向实际应用数据如 Web 文本数据、基因数据、入侵检测数据流、商业交易数据流开展的特征降维研究工作。

第 10 章是数据流分类算法实验工具包 ETDSv1.0。详细介绍了数据流分类算法实验工具包 ETDSv1.0 的功能与用户使用手册。

第 11 章是经典的数据流分类算法实验工具。主要介绍了若干经典的数据流分类算法实验工具的主要功能与使用方法

第 12 章是数据流分类算法常用的实验数据集。归纳了数据流分类领域常用的模拟数据流数据集的分类并描述了数据集的特点。

本书在总结归纳了已有数据流分类方法基础上，介绍了作者针对数据流分类问题在理论上与实践上的研究成果，并阐述了作者对数据流分类问题的思考。希望通过本书的“抛砖引玉”，能进一步加强数据流分类的理论与应用研究的发展。

本书的研究工作得到了教育部创新团队“多源海量动态信息处理”(No. IRT13059)、973 课题(No. 2013CB329604)、国家自然科学基金面上项目(No. 60975034，No. 61273292，No. 61229301)、国家自然科学基金青年项目(No. 61305063，No. 61503112)、教育部博士点博导基金项目(No. 20130111110011)与中国博士后面上基金项目(No. 2014M551801)的资助，特此向支持和关心作者研究工作的所有单位与个人表示衷心的感谢，感谢清华大学出版社的广大员工为本书出版付出的辛苦劳动。书中部分内容参考了国内外有关单位或个人的研究成果，均已在参考文献中列出，在此一并致谢。

另外，由于作者水平所限，虽几经改稿，书中不足和缺点在所难免，欢迎广大读者不吝赐教。

作　者

2015 年 12 月

目 录

引 言 篇

专 题 篇

实验资源篇

引 言 篇

本篇首先简单介绍数据挖掘的相关概念以及代表性十大算法，然后介绍数据流挖掘的相关定义、应用背景及理论基础与技术，最后重点介绍数据流分类挖掘的主要研究进展与存在的关键问题。

第 1 章
数 据 挖 掘

随着信息技术的高速发展，数据库应用的规模、范围和深度不断扩大，人们面临的主要问题不再是缺乏足够的信息可以使用，而是面对浩瀚的数据海洋如何有效地利用这些数据。面对这一挑战，数据挖掘(Data Mining，DM)和数据库知识发现(Knowledge Discovery in Databases，KDD)应运而生，并显示出强大的生命力。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。本章主要介绍数据挖掘的相关概念、任务十大算法及其应用热点。

1.1 KDD 定义和过程

1. KDD 的定义

众多的学者根据自己对 KDD 的认识和理解，给出了很多定义，而其中被公认为比较完整、深刻和全面的是由 W. J. Frawley 和 U. Fayyad 分别在 1991 年和 1996 年的会议论文中对 KDD 的定义[1,2]：

The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

即 KDD 是从大量数据中提取出有效的、新颖的、有潜在作用的、可信的，并能最终被人理解的、模式的、非平凡的处理过程。由此可知，KDD 是从数据库中提取有价值知识的过程[2]，进行 KDD 的研究是为了将知识发现的研究成果应用于实际数据处理中，为科学的决策提供支持。

2. KDD 处理模型

KDD 过程是交互的、重复的，包括大量由用户决策的阶段，它是一个多步骤的处理过程。图 1.1 是 Fayyad 等给出的九阶段处理模型，多步骤之间相互影响，反复调整，形成一种螺旋式上升过程[1,4]。

KDD 基本步骤如下。

步骤 1，数据准备：对应用领域和相关先验知识的理解以及从客户的角度确定 KDD 过程的目标。

步骤 2，数据选择(Data Selection)：根据用户的要求从数据库中提取与 KDD 目标相关