国家科技支撑计划重点课题
工业和信息产业科技与教育专著出版资金资助出版

中国少数民族特需用品数字化工程丛书

动态数据流分类方法及其在民族信息数据挖掘中的应用

◎ 姚　远　张俊星　徐国凯　著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

国家科技支撑计划重点课题
工业和信息产业科技与教育专著出版资金资助出版

中国少数民族特需用品数字化工程丛书

动态数据流分类方法及其在民族信息数据挖掘中的应用

◎ 姚　远　张俊星　徐国凯　著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

动态数据流挖掘是近年来学术界研究的热点问题之一，由于其本身具有海量性、实时性及动态变化性等特点，因此在利用传统数据挖掘方法对数据流进行处理时，很难得到令人满意的结果。

本书以动态数据流分类问题为切入点，着重介绍了近些年的动态数据流分类问题研究进展，提出了相关技术方法，以民族信息处理为背景，通过案例对相关技术方法进行深入说明。本书内容新颖，融入了近年来学术界与工商业界提出的新方法和新技术，并给出了进一步扩展。

本书可作为信息与工程、计算机应用技术等专科、本科，以及研究生教学用书，也可供相关研究人员参考。

图书在版编目（CIP）数据

动态数据流分类方法及其在民族信息数据挖掘中的应用/姚远，张俊星，徐国凯著. —北京：电子工业出版社，2014.11
（中国少数民族特需用品数字化工程丛书）
ISBN 978-7-121-24652-4

Ⅰ. ①动… Ⅱ. ①姚… ②张… ③徐… Ⅲ. ①统计分析—应用—少数民族—民族文化—数据采集—研究—中国 Ⅳ. ①K28-39

中国版本图书馆 CIP 数据核字（2014）第 249230 号

策划编辑：曲 昕
责任编辑：康 霞
印 刷：三河市双峰印刷装订有限公司
装 订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：720×1 000 1/16 印张：18.75 字数：327 千字
版 次：2014 年 11 月第 1 版
印 次：2014 年 11 月第 1 次印刷
定 价：48.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

丛书序

56个民族56枝花。在中华文明发展的历史长河中，勤劳勇敢的各少数民族群众在生产和生活实践中创造了各具特色的民族特需用品。它们是人类智慧与文明的结晶，是中华民族宝贵的物质与精神财富，是连接民族情感的纽带和维系国家统一的基础。

我国的少数民族特需用品种类繁杂、内容丰富，载体形式多样，蕴含着丰富的民族文化信息，一旦失传，必将带来无法挽回的损失。现实情况恰恰是，相当一部分基于传统工艺的少数民族特需用品因年代久远或受现代大工业生产的冲击已经接近失传，急需保护与发扬。

《中国少数民族特需用品数字化工程丛书》是“十一五”国家科技支撑计划课题“民族特需品数字化关键技术研究与示范应用”（课题编号2009BAH41B05）成果的体现，是我国第一部针对少数民族特需用品进行系统挖掘、整理、研究和展示的学术著作。

本丛书内容涵盖了少数民族特需用品的发展历史与国家相关政策、系统化的评价规范、各类数字化技术（信息系统体系建设、多媒体数据库建设、数字媒体技术、虚拟现实技术）、多语种词汇库（汉语、少数民族语、英语）、先进的数字化软件及其相关应用实例。

在本丛书的完成过程中，创作团队多次深入少数民族地区进行数据的挖掘、采集与整理，对纷繁浩瀚的民族特需用品资料进行了系统与全面的归纳、分析和整理，进行了挑战性极强的后期研究与数字化工作。丛书的完成历时

三年之久。希望丛书的出版能够对我国少数民族特需用品及其所承载文化的传承、保护和发扬尽绵薄之力。

非常感谢国家民族事务委员会、科学技术部、大连民族学院，以及工业和信息化部“工业和信息产业科技与教育专著出版资金”评审委员会对本丛书的出版给予的精心指导与大力支持。

徐国凯　张俊星

2014 年 5 月

前言

分类问题作为数据挖掘领域一个经典而重要的课题，一直受到学术界的关注。然而，随着物联网的推广，以及“大数据”时代的到来，传统数据分类方法正面临新的挑战，首当其冲的就是数据形式的变化，即从传统的静态数据向动态的数据流形式转变。与静态数据相比，动态数据流具有三个特点，即海量性、实时性和动态变化性，这些特点大大增加了数据流分类的难度。因此，如何设计一种数据流分类模型，既能够满足数据流的特点，又能够对数据流进行有效分类，成为当前学术界研究的热点问题。

动态数据流分类的核心问题是如何设计分类器，满足数据流的三个特点。数据流的海量性要求分类模型具有处理“大数据”的能力，这种数据至少是 GB 级别的，通常使用动态数据库与实时数据库作为数据源；数据流的实时性特点要求分类模型具有快速处理数据的能力，对分类结果应在数据流不断产生中同步给出；数据流的动态变化性特点要求分类模型能够自我更新以适应新数据环境的分类需求。因此，对于数据流分类问题的研究，需要在传统分类方法的基础上改进或提出新颖的数据分类技术。

本书介绍了近年来应用于动态数据流分类领域中的几种较新颖的分类方法，包括集成学习数据流分类方法、增量式学习数据流分类方法、数据流概念漂移检测及学习，以及数据流分类技术在民族信息数据挖掘中的应用等内容，研究了它们的原理、特点、性能及应用情况。全书共分为 6 章：第 1 章概述传统数据挖掘技术和动态数据挖掘

技术，重点介绍动态数据流挖掘技术产生的背景、发展现状和未来发展方向；第 2 章探讨数据流挖掘技术及其应用、数据流挖掘算法，以及网络数据流实时监测系统；第 3 章介绍集成学习数据流分类方法、不同分类器集成方法及分类结果融合技术等内容；第 4 章研究基于增量式学习的数据流分类模型的构建，着重研究了动态数据流环境下，基于学习方式的分类模型增量式学习技术，以及不同分类方法对于增量式学习的改进和优化；第 5 章介绍动态数据流概念漂移的产生原因，检测方法及概念漂移学习策略；第 6 章将前面动态数据流分类方法进行融合，以民族信息数据为背景，针对民族信息数据流分类问题展开研究，提出多种具有民族特点的数据流分类方法。总体来说，各章内容相对独立又相互联系，较为系统地阐述了动态数据流分类方法的研究现状。

本书是作者在博士期间针对数据流挖掘所开展工作的系统而全面的总结，通过将已获得的研究成果与最新国内外研究进展相结合，扩充了数据流分类方法的内容，使得本书内容与当前最新动态数据流分类问题研究的前沿接轨，给读者提供最新的研究现状。

在撰写本书过程中，考虑到民族信息分析的复杂性，这部分内容由徐国凯教授和张俊星教授执笔，使得本书理论内容与少数民族信息挖掘具体问题相结合，这也是本书的最大特色之一。此外，在本书的撰写过程中，参考了国内外的相关研究成果和著作，在此感谢所涉及的所有专家和研究人员。

最后，本书的顺利完成还需要感谢默默在背后支持我的父母和妻子李纯，没有你们给我创造的写作环境，这本书也不会那么快完成。另外，我的儿子姚斐，尽管你现在还那么小（6 个月），但希望这本书作为爸爸送给你的礼物，祝愿你茁壮成长，开开心心每一天，爸爸永远爱你。

由于作者水平有限，不妥之处在所难免，恳请同行与读者批评指正。

著　者

目录

第1章

绪　论

数据挖掘（Data Mining）是一种与数据打交道的学科，由于不同领域产生的数据不同，使得数据挖掘成为一门融合多学科、跨领域的交叉学科。早期（20 世纪后期）数据挖掘研究主要针对静态数据展开。所谓静态数据，直观来说就是能够存储在数据库中，且数据量有限的数据。这种数据往往规模有限，且使用过程中可以对数据的总体进行宏观把握，因此，对于提出的模型要求并不高，但是针对静态数据的数据挖掘仍然产生了巨大的影响力。例如，最为经典的“啤酒与尿布”的故事就是早期数据挖掘的典型代表，但随着信息技术的发展，以及“大数据”时代的到来，使得传统数据发生了本质的变化，即数据由静态形式转变为动态数据流（Data Stream）形式。所谓动态数据就是数据是海量的、实时的、动态变化的，这对于传统针对静态数据设计的数据挖掘模型来说提出了全新的挑战。在数据流环境下，简单套用传统数据挖掘方法所得结果令人无法满意，甚至有些时候挖掘模型会完全失效，因此，针对数据流特点设计相应数据挖掘方法与模型成为当前学术界及工商业界研究的热点问题[1]。本章通过循序渐进的方式介绍数据挖掘的相关内容、常用技术及基本概念，让读者对数据挖掘领域有一个宏观的认识，然后阐述动态数据流挖掘的背景及国内外研究现状，最后对一些重要的动态数据流挖掘技术进行详细介绍。

1.1 引言

随着信息技术的发展，传感器技术（物联网）、网络技术（社交网络）和通信技术（移动互联网）正改变着人们的生活方式和社会发展。面对数据海洋，人们往往呈现出两种态势：一种是将数据完全存储起来，这种方式对数据仅保存而不利用，最终结果是在数据的海洋中“沉没”；另一种是将收集到的数据利用起来，从中发掘出有价值的知识和规律，并反过来指导具体领域的行为，创造更大的价值。在此背景下数据挖掘应运而生，通过使用数据挖掘技术，将沉睡在数据库中的数据进行分析，重新焕发了历史数据的价值，解决了信息时代“数据爆炸，知识匮乏”的问题。

所谓数据挖掘（Data Mining）就是将数据中所隐含的知识与信息进行

提取，即从大量的、有噪声的、不确定的、随机的数据中挖掘出隐含于内的、不为人知的、具有实际作用与价值的信息与知识。

传统数据挖掘技术的发展过程可以追溯到 20 世纪 80 年代，其里程碑事件是 1989 年在美国底特律召开的第 11 届人工智能联合会。此次会议首次将数据库知识挖掘的内容进行专题讨论（KDD Workshop）。随后，关于数据挖掘的 KDD 会议成为数据挖掘领域的顶级会议，该会议每两年举办一次。KDD 会议每次讨论的议题均不同，例如，KDD1999 年会议的议题是网络攻击检测，KDD2001 年的议题是生物制药数据方面的挖掘等，总体来说，每次 KDD 会议的议题都会根据近期最新的技术或趋势进行设定，最近一次 KDD 会议于 2011 年在北京举办，主要讨论的内容是社交网络数据下的挖掘，这也是当前比较热门的问题之一。除了 KDD 会议之外，还有很多国际顶级会议涉及数据挖掘相关领域，如 VLDB、SIGMOD、PKDD、ICDM、SDM、PODS 等，针对不同的数据挖掘问题进行专门讨论，在这些会议的推动下，数据挖掘相关理论和技术得到了极大丰富和长足进步。

如何从数据中提取出有价值的知识，其中蕴含了两个基本问题：第一是需要设计数据挖掘方法，这部分主要涉及理论研究，利用统计等知识在数据层面进行研究。第二要考虑数据所产生的背景与环境，这需要跟具体领域背景结合，从中寻找出真实的、需要挖掘的内容。只有将上述两点有机结合起来，才能既保证数据挖掘的可执行性，又保证挖掘结果的有效性。

目前，支撑数据挖掘的相关理论基础主要有数学理论、机器学习理论、数据库理论和可视化理论等。其中，数学理论包括统计学理论、模糊集理论、粗糙集理论、概率论等内容。机器学习理论包括支持向量机（SVM，Support Vector Machine）理论、贝叶斯学习理论、归纳学习、决策树、类比学习与基于样本的学习和计算智能等。数据库理论包括关系数据库理论、数据库事务理论、逻辑与数据库、面向对象数据库理论等。可视化理论涉及计算机图形学、图像处理、计算机辅助设计、计算机视觉，以及人机交互等。

随着数据挖掘研究的深入，针对不同领域背景下的问题基本上都有解决方案被提出。这些方法无论如何变化都离不开数据挖掘三大理论支柱的支持，即数据库技术、人工智能技术、概率与数理统计。数据库技术面向

数据本身，着重对数据存储与读取进行研究，通过改善系统输入/输出效率，提高数据的处理速度。人工智能技术力求对数据进行升华，在更高层面对数据进行理解与分析，其分析方法采用了模拟人类智慧的方式，并利用计算机快速处理的能力大大提高数据分析效率。概率与数理统计则面向数据本身，从数学属性出发对数据进行分析与处理，通过使用数学方法获得数据内部隐含的关联关系。当前，数据挖掘的发展方向是所构建挖掘模型趋向复杂性，其原因一方面是由于所要面对的问题更为复杂，另一方面是随着挖掘技术的深入，用户对数据挖掘技术及结果要求也随之提高。因此，一个完整的数据挖掘模型往往是对数据库技术、人工智能及数理统计相关工具的集成[2]。

目前，数据挖掘技术已经应用于各个领域，例如，在医学领域，使用数据挖掘技术对脑电波数据进行检测，预防老年痴呆症的发生；在金融领域，使用数据挖掘相关方法对个人消费行为进行分析，杜绝洗钱及恶意贷款的发生；在航天领域，同样使用数据挖掘技术对航天器外形及飞行控制等方面做出优化。可以说，数据挖掘技术已经潜移默化地渗透到各个领域中，通过对数据进行挖掘，从数据中要知识、要规律，大大提高服务质量和人们的生活水平。综上所述，目前大部分学科和行业都投入到数据挖掘的理论和应用研究中，并取得了良好的结果，衍生出很多高质量的产品。

然而，在数据挖掘技术大发展的过程中，依然存在很多亟待解决的问题。这些问题中，除了新背景产生的问题外，例如，社交网络中物联网等可以归纳为数据形式的改变，即数据从传统静态数据转变为动态数据流形式，因此，对传统数据挖掘方法也提出了新的挑战。由于传统数据挖掘方法是基于静态数据类型而设计的，因此，当面对动态数据流数据时就会显得力不从心或者完全失效，这也是当前数据挖掘领域需要面对的问题。

1.2 数据挖掘概述

数据挖掘又称知识发现，是从数据库中提取有价值规则的过程。近些年众多研究者投入到数据挖掘领域中，大量新概念、新技术被提出，本节

将对数据挖掘相关概念与技术进行介绍，从宏观上为读者讲述数据挖掘发展现状。

1.2.1　数据挖掘基本概念介绍

1. 数据挖掘定义

从本质上说，数据挖掘是一种从原始数据中提取知识的工具。与具体应用背景相结合，数据挖掘就是按照既定目标，对大量实际数据进行探索与分析，以揭示隐藏的、未知的规律性并将其模式化，从而为决策活动提供支持。因此，数据挖掘只有与具体应用背景相结合才能真正发挥它的作用，也使得相关数据挖掘理论更加经得住实际应用的考验，提高理论的完备性。此外，在数据挖掘技术中，很难有一种方法能够适应多种不同数据环境与背景的挖掘需要，因此，所有方法都是相对的、有局限的，更加突出问题背景的重要性。

从技术角度看，数据挖掘的核心内容是从海量的、不完全的、有噪声的、模糊的、不确定的数据中，透过数据的纷繁复杂提取出数据中蕴含的有价值的但人们不知道的知识的过程。这里原始数据可以是结构化的标准数据，也可以是非结构化的数据，甚至是异构数据等。挖掘知识的方法可以是数学的方法，也可以是非数学的方法，可以是演绎推理方法，也可以是归纳方法，而最后获取的知识可以进行信息检索、管理、决策支持及过程优化等。因此，数据挖掘既可以看成一种全新的领域，又可以看成多个传统领域相互融合的结果。由于数据挖掘的出现，使得人们对数据有了全新的认识，也让数据从存储价值变为发掘价值，甚至当前数据对于一个企业的价值远超人才价值。

在数据挖掘中，所谓知识就是信息经过加工和改造形成的结果。这里的信息以数据为载体，一般信息量使用著名的 Shannon 信息熵来描述。知识是人类在实践基础上产生又经过实践检验的对客观实际的、可靠的反应。知识是人脑创新的成果，是人类智慧的结晶。而知识的最高层面智慧，是人类文明的源泉，是推动历史发展的永恒动力，是生产力诸多要素的核心。因此，对于数据挖掘得到的知识，从另一个方面来说，若成为智慧，将对

整个人类文明进程产生巨大影响。

从数据挖掘流程来说，一个完整的数据挖掘过程包含以下 5 部分。

（1）对原始数据进行选择。此部分的主要内容是将需要进行分析和挖掘的数据从数据库（源）中提出为后续工作做准备，并且所需提取的数据量由实际问题和计算机软硬件环境所决定，并不是越多越好。

（2）对数据进行初始化。此过程相当于对数据进行初始调整，因为原始数据存在很多不适应后续挖掘的情况，如数据中存在噪声、数据值缺失、数据类型不统一等问题，因此初始化的目的是将数据进行规整，以便后续操作。值得注意的是，当数据初始化无法满足条件时，从原始数据库中重新选择数据进行初始化。

（3）数据转换。这部分内容是在数据初始化的基础上对数据做进一步整理，使其能够完全适应数据挖掘的要求。由于不同数据挖掘问题所采用的数据挖掘方法和模型不同，对数据形式的要求也不同（例如，有的模型要求是字符型数据，而有的模型要求是纯数值型数据），因此，在进行挖掘之前有必要将数据进一步转化为模型能够采用的形式。此外，数据规整度较好，可以对挖掘效果产生积极影响，有利于得到所需的知识和模式。

（4）数据挖掘。此部分是整个数据挖掘过程中的核心内容，也是大部分学者研究的热点环节。针对不同问题所提出的方法也不同，例如，数据分类、聚类、降维、压缩等，并且挖掘方法还需具体问题具体分析。对每一个小问题，目前都有很多种方法进行挖掘，这也是数据挖掘领域近些年发展成果比较集中的部分。

（5）评估和解释。这部分内容主要是对挖掘方法输出的结果进行评估和评价，并且对其中所蕴含的潜在知识进行分析和解释。评估的作用是对所使用的数据挖掘模型是否满足解决问题的要求给出一种度量。一般来说会使用一种标准来衡量误差，并且通过误差来调整数据挖掘模型的参数，以达到模型优化的目的。此外，尽管数据挖掘模型已经给出挖掘结果，但模型毕竟只是一个工具，如果想获得知识，还需要人类进行后续分析才能最终得到想要的知识。当从海量数据中通过数据挖掘方法获取到有价值、有用的知识后才算完成了数据挖掘的全过程，整体挖掘过程如图 1.1 所示。

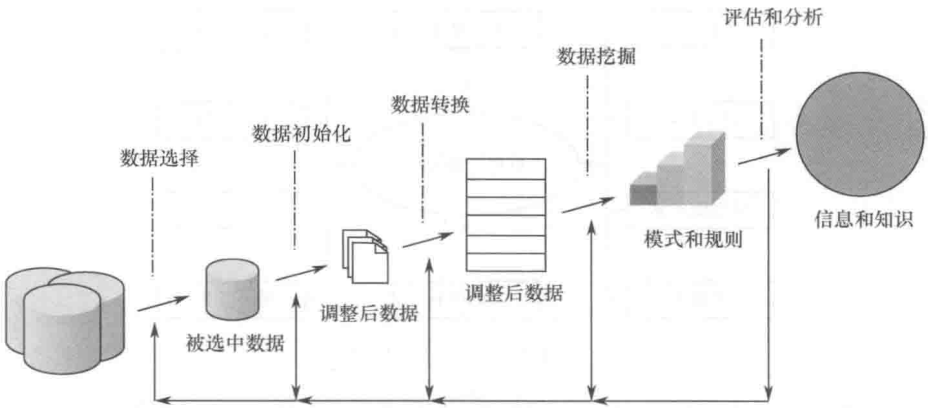

图 1.1　数据挖掘过程

对数据挖掘相关内容的研究我国起步相对较晚。从 1993 年国家自然科学基金对数据挖掘领域项目进行支持开始，国内众多高校和研究所加入到数据挖掘研究的热潮中，随着人力、物力投入的增加，在数据挖掘领域也做出了很多优异的成果，如大连理工大学、北京大学、清华大学、国防科技大学、南京大学、上海交通大学、中国科学院计算所和数学所等[3]。在这些高校和研究所的共同努力下，国内已经形成一股研究数据挖掘的热潮，并且也开展了很多研讨会供学者们进行交流，如高级数据挖掘与应用国际会议（ADMA）目前已经举办了 8 届；模糊系统与知识发现国际学术会议（FSKD）目前已经举办了 9 届。除此之外，民间也自发组织了一些数据挖掘相关的兴趣小组，例如，豆瓣网中的数据挖掘群，每天讨论都很热烈，在群中大家分享自己的发现和研究成果，互相答疑解惑，这也为数据挖掘技术在我国的研究和推广产生了积极作用。

在学者们的共同努力下，数据挖掘技术已经形成一种多学科交叉的研究领域，所涉及领域如图 1.2 所示，并且随着研究的深入，越来越多的新学科也逐步融入数据挖据领域中。

通过上述对数据挖掘的介绍，可以总结出数据挖掘具有以下 5 个特点。

（1）数据挖掘所使用的数据量往往是巨大的，因此，在挖掘过程中如何提高效率、如何针对具体应用背景提出有效的挖掘算法，以及对所用数据进行筛选等，都成为研究数据挖掘学者需要考虑的问题。