FANXIQIAN ZHONG DE
KEYI JINRONGJIAOYI SHIBIE 张成虎 著

反洗钱

中的可疑金融交易识别

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

FANXIQIAN ZHONG DE
KEYI JINRONGJIAOYI SHIBIE 张成虎 著

反洗钱中的可疑金融交易识别

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

反洗钱中的可疑金融交易识别/张成虎著．—北京：经济管理出版社，2013.12
ISBN 978-7-5096-2849-2

Ⅰ.①反… Ⅱ.①张… Ⅲ.①洗钱罪—研究—中国 Ⅳ.①D924.334

中国版本图书馆 CIP 数据核字(2013)第 286331 号

组稿编辑：谭 伟
责任编辑：张巧梅
责任印制：黄章平
责任校对：李玉敏

出版发行：经济管理出版社
（北京市海淀区北蜂窝 8 号中雅大厦 A 座 11 层 100038）
网 址：www.E-mp.com.cn
电 话：（010）51915602
印 刷：三河市海波印务有限公司
经 销：新华书店
开 本：720mm×1000mm/16
印 张：13.25
字 数：253 千字
版 次：2013 年 12 月第 1 版 2013 年 12 月第 1 次印刷
书 号：ISBN 978-7-5096-2849-2
定 价：45.00 元

前　言

洗钱活动与贩毒、走私、恐怖活动、贪污腐败和偷漏税等严重犯罪密切联系，它不仅会扰乱市场经济中的有序竞争，影响经济活动中的公平公正，也对国家政治稳定和经济安全构成严重威胁，已被国际社会公认为冷战之后典型的“非传统性安全问题”之一。在经济全球化深入发展和国际恐怖主义甚嚣尘上的时代背景下，打击洗钱犯罪已成为我国大国责任的重要内涵，其工作成效也直接影响着我国的国家形象和国际声誉。反洗钱作为打击洗钱犯罪的直接手段，是我国大国责任的重要体现，也是抑制上游犯罪的有效途径，对维护我国金融体系安全稳健地运行、社会经济活动公正有序地开展，以及打击经济犯罪与惩治腐败有着极其重要的意义。

从我国目前的情况看，我国的洗钱活动涉及通过商业银行等金融机构洗钱的报告数和涉案金额占全部洗钱总量的90%以上，金融机构特别是商业银行，已成为我国反洗钱的前沿阵地和主战场。可疑金融交易识别作为反洗钱工作的核心环节，承担着从日常经济金融活动中发现甄别及追踪与洗钱犯罪有关的可疑交易和相关线索的主要责任，也是反洗钱工作得以顺利展开的重要基础性工作。

我国金融业的迅速发展和金融信息化的不断深入，使得金融活动不仅瞬间完成，而且金融交易数量持续增长，金融机构每天都会有数以亿计的海量交易发生。要在海量金融交易中识别匿藏于其中的可疑交易，仅仅依赖人工识别是比较困难的。加之，可疑金融交易识别不仅涉及相关领域知识、客户背景知识及客户行为模式，还与不断变化的洗钱方式和途径密切相关。由于洗钱活动越来越专业化，特别是现代金融体系越来越复杂，金融产品创新层出不穷，给可疑金融交易识别带来了极大的挑战。此外，能否有效识别已经发生的可疑金融交易，还取决于所采用的识别方法、技术和手段的有效性。

数据挖掘（Data Mining，DM）的发展和广泛应用为识别可疑金融交易和挖掘洗钱线索提供了有效的技术手段。借助于数据挖掘技术，对金融交易、账户群体数据进行分类分析、聚类分析、关联分析和时间序列分析等挖掘处理，可以大大提高可疑金融交易的识别效率和准确性，也能及时发现新的洗钱模式和交易规律。

本书研究团队从2003年开始关注反洗钱问题的研究，西安交通大学也十分重视和肯定本研究团队的工作，批准成立了“西安交通大学金融商务智能与反洗钱研究中心”。团队的研究工作主要分为两个阶段。第一阶段为基于数据挖掘的可疑金融交易识别研究，2005年与国家外汇管理局合作开展了基于数据挖掘的可疑金融交易识别系统研究，2008年得到国际自然科学基金的资助。从2009年开始，鉴于世界范围内反洗钱重点的转移和国际反洗钱规则的扩充与完善，研究团队开始进行反洗钱的理论与机制问题研究，并得到教育部社科基金的资助。本书即为第一阶段研究成果的总结。第二阶段的研究成果正在逐步形成并发表之中，很快也会与读者见面。

本书是由张成虎教授及其博士研究生与西安交通大学金融商务智能与反洗钱研究中心对近年来基于数据挖掘技术与方法的可疑金融交易识别研究与应用的部分成果的系统总结，得到了国家自然科学基金“基于数据挖掘的可疑金融交易识别研究”项目的资助。其中第一章、第二章由张成虎、赵小虎、尹为撰写；第三章至第五章由张成虎、赵小虎、杨彬、陈明哲撰写；第六章由张成虎、李时、周东、陈明哲撰写；第七章由张成虎、尹为、杨彬撰写。张成虎教授负责全书的策划和大纲的拟订，并负责全书的统纂和修改。博士生陈明哲为本书的出版做了大量的资料整理与校正工作。同时，西安交通大学金融商务智能与反洗钱研究中心的孙景、李淑彪、王雪萍等教师及博士研究生王宝运、赵燕、胡啸兵、李霖魁、孙陵霞等同学，在本书的撰写过程中也给予了大量帮助和支持，在此，向他们表示衷心的感谢。感谢国家外汇管理局西安分局和中国人民银行西安分行反洗钱处、中国人民银行济南分行反洗钱处在本书研究过程中给予的支持。

本书在撰写过程中，参考了大量国内外的相关研究成果，在此，对涉及的所有专家和学者表示衷心感谢。

本书的出版得到了经济管理出版社谭伟总编助理的鼓励和支持，没有他多次给予的鼓励和催促，本书不可能按时出版。对此，也向谭伟表示深深的感谢。

由于时间仓促，书中难免有谬误之处，敬请读者指正。本书中出现的错误由作者负责。

目 录

第1章　绪论

随着经济全球化和金融一体化的迅速发展，金融作为现代经济的核心，其交易规模日益庞大，交易手段不断翻新，交易处理的自动化、电子化水平逐步提高。现代金融产品在为参与者提供便利的同时，也为洗钱犯罪带来了可乘之机。洗钱活动日益猖獗，引起国际社会的普遍关注，各国政府也加大了对洗钱犯罪的打击力度。目前各国监管部门普遍建立了金融交易报告制度，为反洗钱工作的开展积累了大量的信息，通过分析金融机构的客户信息及交易数据，采用科学的方法识别可疑金融交易进而发现洗钱线索，已成为反洗钱研究的核心问题。

本章阐述本书的研究背景、国内外研究现状、研究目的和意义、内容与框架安排。

1.1　研究背景

洗钱犯罪是现代社会的毒瘤，洗钱不仅严重破坏社会公平、影响社会稳定、滋生贪腐行为，而且增加犯罪能量（为各类犯罪行为提供资助），甚至会危害国家安全，与洗钱犯罪作斗争是世界各国所共同面临的一个重要问题。

2007年1月1日我国正式实施《中华人民共和国反洗钱法》，这一举措不但从法律层面确立了反洗钱工作的重要性，而且反映出我国所面临的严峻洗钱形势。该法所称的反洗钱，是指为了预防通过各种方式掩饰、隐瞒毒品犯罪、黑社会性质的组织犯罪、恐怖活动犯罪、走私犯罪、贪污贿赂犯罪、破坏金融管理秩序犯罪、金融诈骗犯罪等犯罪所得及其收益来源和性质的洗钱活动。2007年6月28日中国成为金融行动特别工作组（Financial Action Task Force on Money Laundering，FATF）的正式成员，我国反洗钱工作得到国际社会的广泛认可。

随着我国反洗钱工作的深入开展和打击洗钱犯罪活动力度的加强，金融机构被推到了反洗钱工作的前沿，成为反洗钱工作的“主战场”。我国在金融领域反洗钱的主要目的：一是作为一种犯罪威慑手段。洗钱活动帮助犯罪分子逃避法律

制裁，助长了犯罪势力，严重地影响了社会秩序，特别是危害极大的有组织犯罪。反洗钱可以打击各类上游的犯罪，釜底抽薪的制约犯罪能量，对严重的犯罪行为具有抑制作用。二是维护社会经济秩序，保障金融体系稳定运行。洗钱犯罪扰乱了正常的社会经济秩序，妨碍了金融体系的稳定。通过金融机构洗钱对金融机构本身的信誉也是一个损害，金融体系是现代经济的核心，如果经济核心没有诚信，对国家的长期发展不利。三是体现大国责任，维护良好国际形象。洗钱活动泛滥对政府的声誉有严重影响，作为一个负责任的大国，必须预防洗钱活动，遏制、打击洗钱犯罪。

金融交易活动是洗钱犯罪行为必须利用的一个重要环节，金融机构因此成为防范洗钱的第一道屏障[1]。通过分析金融机构的客户信息和交易数据，识别可疑金融交易进而发现洗钱线索，成为金融机构反洗钱工作面临的核心问题之一。这一问题解决得好坏直接影响全社会反洗钱工作开展的成效。

可疑金融交易（Suspicious Financial Transaction）是指交易的金额、频率、流向、用途、性质等有异常情形或特征的金融交易行为。金融交易是否可疑取决于具体的交易情形，并涉及多个因素。某些交易从单个因素来看似乎正常，是可以忽略的，但从长期的、多个因素综合的情况来看却是可疑的。交易发生时的具体情形是判断交易是否可疑的一个重要因素，而这些情形涉及的内容在不同的商业领域、不同的客户之间存在差异，包括客户商业活动的领域知识、交易历史、客户背景及交易行为。当判断交易是否可疑时，还需要了解正常交易规律，以判断交易是否符合正常的金融活动惯例，进而判断是否异常或可疑。

同时，随着经济全球化、金融电子化和犯罪有组织化的发展，可疑交易藏匿于金融机构成千上万的海量交易数据中，对其甄别的难度很大。尽管各国政府的交易报告制度中，列举了许多可疑交易情形，如美国银行反洗钱工作指南规定了50多种情况。中国人民银行颁布的《金融机构大额交易和可疑交易报告管理办法》（中国人民银行令2006第2号）中规定了48种交易或者行为，作为可疑交易进行报告。但这些指导性的标准只是为可疑交易的识别提供了一个筛选框架，一般是根据已知的洗钱行为和模式来制定的，对于大量未知的洗钱行为采用传统的方法和技术很难识别。洗钱行为的演进和变化会导致这些标准过时，依靠指导性标准来进行可疑交易识别的效率和准确性会大大降低。

因此，可疑金融交易识别是一个比较复杂的过程，它不仅涉及领域知识、客户背景知识及客户行为模式，还与不断变化的洗钱方式和途径密切相关。此外，能否有效识别已经发生的可疑金融交易，还取决于所采用的识别方法、技术和手段的有效性。由于洗钱活动越来越专业化，特别是现在金融体系越来越复杂，金融产品创新层出不穷，洗钱分子常常利用新的金融产品去洗钱，使反洗钱人员越

来越需要新的知识来武装自己。

数据挖掘（Data Mining，DM）是指从大量数据中提取或“挖掘”知识，能够根据分析人员的需要，从数据库、数据仓库或其他信息库存放的海量数据中提取有价值的模式和规律[2]，它的发展和广泛应用为识别可疑金融交易和挖掘洗钱线索提供了有效的技术手段。借助于数据挖掘技术，对金融交易、账户群体数据进行分类分析、聚类分析、关联分析和时间序列分析等挖掘处理，可以大大提高可疑金融交易的识别效率和准确性，也能及时发现新的洗钱模式和交易规律。各类挖掘功能的实现，不仅需要特定的算法，而且需要特定的数据。不同的数据适用于不同的算法，不同的算法适合于不同的挖掘要求。每一类算法也都有应用的约束条件，有其优点和缺陷。因此数据挖掘技术的应用离不开数据的有效准备（数据选择及预处理）和算法的合理选择。

数据挖掘技术是一种基础性技术，包括概念描述、关联分析、分类分析、聚类分析、偏差检测、时间序列分析等主要功能[3]。数据挖掘的众多功能是依靠数据挖掘算法来实现的。当把数据挖掘技术应用于可疑交易识别时，不但要选择适合的挖掘算法，还要结合领域知识对其进行优化和更新。在全球经济一体化环境下，面对金融交易量不断攀升，交易品种层出不穷的形势，将数据挖掘技术应用于可疑金融交易识别，同时面临着应用创新和算法创新的结合，这既是挑战又是机遇。单一算法的简单应用在可疑金融交易识别上存在适用性、效率和约束条件等问题，同时难以对交易的可疑度进行综合判断，只有在深入分析洗钱特征的基础上提出适合于反洗钱监测的新算法，才能在理论研究和应用拓展上取得新突破。本书研究的目标是把数据挖掘技术与金融反洗钱领域知识密切结合，研究能够有效识别我国可疑金融交易的数据挖掘算法。

1.2 本书的目的和意义

可疑金融交易识别是发现和追踪涉嫌洗钱线索的有效途径。本书研究的目的是针对我国反洗钱领域的核心问题——可疑金融交易识别，将数据挖掘技术与金融反洗钱领域知识密切结合，选择适当的洗钱交易识别策略与方法，提出通过“异常模式检测”识别可疑金融交易，建立一套基于数据挖掘的可疑金融交易综合识别方法，在应用创新的基础上实现数据挖掘算法的基础性创新。

本书以数据挖掘技术为手段，应用金融领域知识和反洗钱领域知识，进行可疑金融交易的识别研究，探讨适合我国金融交易特征的可疑金融交易识别数据挖

掘方法，并在反洗钱工作建立的有效的技术和方法基础上，提高可疑金融交易识别方法的基础创新能力，进而提高我国反洗钱工作的科学化和有效性，推进我国反洗钱战略的实施，这具有重要的理论意义和实用价值，为我国可疑金融交易识别提供理论和方法上的参考与借鉴。

1.3 国内外研究进展

1.3.1 基于传统统计分析和静态数据挖掘的可疑金融交易识别研究

国外关于反洗钱的研究起步比较早。20 世纪 70 年代就开始了反洗钱立法方面的研究。而将信息技术应用到反洗钱领域的研究也在 90 年代就提出了。

Senator（1995）[4]较为系统地介绍了 FinCEN（Financial Crimes Enforcement Network，金融犯罪执法网）的 FAIS（FinCEN Artificial Intelligence System，FinCEN 人工智能系统）的系统结构、监测识别关键技术及其应用。FAIS 采用人工智能技术，通过分析被提交的交易报告，发现可疑金融交易行为。Kingdon J. 和 Feldman K. S.（2002）[5]设计了银行交易数据监测和分析系统，可以检测到支付欺诈和洗钱活动。Kingdon J.（2004）[6]提出了利用人工智能自动识别客户行为模式，从而可以识别异常交易行为（Unusual Behavior）的方法。Shijia Gao 等（2006）[7]提出了将智能代理技术应用于反洗钱领域，以适应不断变化的洗钱风险和方式，介绍了一个新颖的、开放式的多代理的反洗钱结构体系。

Office of Technology Assessment（U. S. Congress）（1995）[8]发表了如何利用信息技术控制洗钱犯罪的报告，文中阐述了基于统计学和人工智能技术的聚类方法在发掘潜在交易特征方面的应用。Sara Reese Hedberg（1995）[9]介绍了并行技术应用于数据挖掘，并说明了其在 FinCEN 系统中的应用。Ronald J. Brachman（1996）[10]对数据挖掘技术在金融投资、反洗钱等多个金融领域的应用进行了探索。Aussem A. et al.（1998）[11]将小波分析与动态递归神经网络结合，对股票市场收盘价格做出了有效预测。Goldberg H. G. 和 Senator T. E.（2000）[12]阐述了基于链接分析的聚类方法在反洗钱中的应用。Zbigniew R. 和 Struzik（2001）[13]通过多重分形标度分析的方法研究了 S&P（标准普尔）指数时间序列的多重分形标度特征，揭示出 S&P 基于奇异性指数的局部相关性。Bolton R. J. 和 Hand D. J.（2002）[14]介绍了基于统计技术的欺骗检测和洗钱识别的方法和工具。Laurence Jacobs（2003）[15]介绍了大量监测洗钱的智能方法。Zhongfei（Mark）Zhang

(2003)[16]提出一种基于相关性分析的链接分析方法，用于分析文本性文档以发现洗钱犯罪团伙及内在的关系。Sangbae Kim 和 Francis In (2003)[17]借助小波分析技术发掘各类金融变量与真实经济活动之间的联系，得出美国工业生产与金融指标存在超前—滞后现象的结论。John S. Zanowicz (2004)[18]提出采用数据挖掘技术识别通过国际间贸易洗钱的活动，该方法以美国商品贸易数据库 (U. S. Merchandise Trade Database) 作为数据源，针对进出口价格异常，来发现可疑金融交易活动。

Petrus C. Van Duyne (1999)[19]通过分析荷兰 1994～1999 年的可疑数据，指出可疑交易监测系统和反洗钱策略方面存在的问题，并提出改进建议。Sara Reese Hedberg (1995)[20]提出了数据挖掘在反洗钱领域的应用。Ronald J. Brachman (1996)[21]对数据挖掘技术在金融投资、反洗钱等多个金融领域的应用进行了探索。Jun Tang 和 Jian Yin (2005)[22]提出了基于 SVM (Support Vector Machine，支持向量机) 的异常交易行为探测算法，用于替代传统的基于预先设定规则的可疑交易过滤系统，通过实验数据验证了异常交易识别在误报率和检测率方面性能的改进。Bolton R. J. 和 Hand D. J. (2003)[23]提出采用统计方法和机器学习辅助反洗钱活动。Jason Kingdon (2004)[24]提出了为每个银行客户产生一个多维的自适应概率矩阵 (Multidimensional Adaptive Probabilistic Matrix)，通过使用概率加权汇总来评估个人行为，进而识别可疑交易。Henry G. Goldberg 和 Raphael W. H. Wong (2005)[25]描述了如何将 FinCEN 数据库重组，支持使用链接分析方法在现金交易数据库中发现和分析洗钱迹象的全过程。David J.，Hannah B. 和 Matthew R. (2003)[26]提出了一种数据挖掘技术方案，用于解决在欺诈交易识别中准确定位和隐私保护之间的矛盾，其中聚类方法是设计方案的重要组成部分。Tang J. (2006)[27]提出了一个基于金融交易数据特征的交叉孤立点检测模型，同时给出了优化的计算偏离度的近似算法，根据真实银行交易数据和人造孤立点的混合数据表明此模型在显著提高检出率的同时能有效降低假阳性比率 (False Positive Rate)。

我国对反洗钱方面的研究在 20 世纪 90 年代后逐渐增多。目前在我国对于人工智能技术、数据仓库技术、数据挖掘技术等信息技术在反洗钱中的应用的研究仅仅处于起步阶段。

贺时忠 (2003)[28]根据洗钱活动的特点，总结了经验观察法 (还原分析法)、经典分析法 (经济分析法)、技术分析法 (数学分析方法) 三类方法可用于对大额和可疑外汇资金进行交易分析，来发现洗钱线索。徐志春、肖伟平、何宏 (2003)[29]介绍了反洗钱系统中用到的几个关键数据开采技术，包括数据集成、数据分类、关联分析、聚类分析和可视化技术。谭德彬、陈藻 (2003)[30]描

述了基于数据挖掘技术的反洗钱系统的实现框架和数据挖掘的关键技术，包括数据分类、关联和聚类等技术。徐加根（2004）[31]从法律、行为、技术、经济等角度综合阐述了洗钱活动的特征。刘丽、史奇中（2004）[32]提出应用网格技术建立反洗钱数据监控分析平台，实现各部门、各行业相关反洗钱数据的集成和统一，为可疑交易识别提供数据基础和应用平台。

张焱（2004）[33]分析了数据挖掘技术的应用特点，提出了一个应用系统原型，并举例说明了基于统计和决策树等挖掘方法。孙小林、卢正鼎（2004）[34]把中心点的思想应用到 BIRCH 算法的聚类特征计算过程，改进了增量聚类算法，通过判断距离是否大于阈值来决定是否为正常交易或是可疑交易，并应用金融交易电汇数据进行测试，验证了算法的有效性。杨胜刚、王鹏（2005）[37]分析了将数据挖掘技术应用于上报交易数据中的必要性和可行性，综述了各类挖掘算法，描述了一个反洗钱系统框架。张成虎、李时（2005）[36]综述了基于 AI 技术的反洗钱系统框架及其相关的数据集成及数据挖掘等技术。胡秋灵（2005）[37]通过实验综合对比分析了各类聚类方法应用于金融交易数据挖掘时异常点的查找性能。汤俊、熊前兴（2006）[38]分析了我国反洗钱工作中对判别规则含混，可操作性差，可疑行为的定义包含大量诸如“频繁、突然、异常、与客户背景不符”之类的原则性的术语等现状，对基于单数据集和多数据集的离群点算法进行研究，给出了一个用于可疑金融交易数据判别的跨数据集对比离群点检测模型，该模型通过数学定义清晰描述了参照集和对比集之间离群点模式的判别检测关系，为深入研究切合金融数据挖掘特点的算法建立形式化描述体系。邓凯旭、宋宝瑞（2006）[39]以预测股票收盘价格为例说明了小波变换对金融数据进行分析和预测的有效性。张成虎、高薇（2006）[40]探讨了朴素贝叶斯分类算法在反洗钱中的应用，设计出了适合于可疑交易识别的朴素贝叶斯分类算法及模型，用数据对该算法进行了实验验证，并提出了与聚类算法相结合的综合运用设想。侯守国、张世英（2006）[41]利用最大重复离散小波变换对沪深股市的交叉互相关性做了小波分析，识别出沪深股市高频收益序列的互相关性。李时、张成虎（2007）[42]提出一种快速的基于 FP－tree 的约束最大频繁项目集挖掘算法，把发现约束频繁项目集的问题转化为发现约束最大频繁项目集的问题，提高运行效率，同时可确定挖掘主属性，确定生成的关联规则范围，为有效开展可疑金融交易识别提供了有益的参考。刘芳、伏峰（2007）[43]基于约束非指导性链接发现技术，通过对目标节点进行约束性判断，再计算“兴趣度”和路径“作用值”来对外汇资金交易数据库进行自动分析，挖掘出用户感兴趣的交易信息。傅强、彭选华和毛一波（2007）[44]研究了小波变换方法在金融时序分析中模型变点探测的应用，提出了金融时间序列变点探测的小波模极大值线方法。陈云开、马君华（2007）[45]从技

术角度探讨了外汇领域的洗钱侦测系统及其关键算法的实现，提出了一种以语义核心树为基础的增量概念聚类算法。刘芳、薛蕾（2006）[46]提出利用数据挖掘技术监测外汇公有账户资金流失，首先将频繁交易的个人账户聚类，然后对每个簇与公有账户之间的交易行为进行链接分析，从而发现可疑个人账户簇以及与之关联的公有账户。陈云开（2006）[47]提出分布式异构计算环境下基于数据挖掘技术的洗钱侦测系统体系结构，并从逻辑层次结构，系统基本框架和系统基本流程三个方面对洗钱侦测系统的体系结构进行了阐述。

孙景、李峰（2008）[48]根据数据挖掘技术中的多层关联规则方法，提出了建立金融交易中的行业间关联规则的思想，并给出基于多层行业分类表，银行企业客户信息和大额历史交易数据的行业间关联规则的建立过程。通过建立行业间关联规则发现那些原本关联度很小的可疑金融交易，为金融反洗钱识别提供参考。张成虎、赵小虎（2009）[49]根据小波分析在数学上具有严格意义上的突变点诊断能力，不依赖于经验模型，适合检测金融交易序列中的可疑成分，针对可疑金融交易特征，探讨借助小波分析方法在信号奇异性检测方面具有的独特优势，从金融交易序列中识别出具有异常交易行为特征的账户。李玉华、李栋才和毕威（2011）[50]提出一种结合数据挖掘中聚类、关联规则和低序马尔科夫模型的混合马尔科夫模型，并基于置信度进行剪枝以降低时间复杂度，将该模型用于预测反洗钱领域中账户之间的交易。

1.3.2 数据流挖掘在可疑金融交易识别中的应用研究

1.3.2.1 *数据流挖掘方法研究*

自从 Henzinger M. R.、Raghavan P. 和 Rajagopalan S.（1999）[51]在论文“Computing on Data Stream”中首次将数据流作为一种数据处理模型提出之后，数据流挖掘已经成为数据挖掘领域一个非常活跃的研究方向。Babcock B.、Babu S. 和 Datar M.（2002）[52]在比较了传统的静态数据和数据流基础上，概括出数据流的四大特征：①数据流实时非匀速到达；②数据到达次序独立，不受系统控制；③数据持续到达，不可预知数据量，无法对其全部保存；④数据流原则上只能被访问一次。由于数据流是无限、连续到达的大量数据，不可能存储所有的数据，因此，许多传统的数据挖掘算法不适合于数据流的挖掘。Hulton G.、Spencer L. 和 Domingos P.（2001）[53]引入窗口滑动技术以子树替代方式处理“概念漂移”问题，构筑了 CVFDT（Concept - adaptive Very Fast Decision Tree）算法。Hulten Geoff（2001）[54]提出了 VFDT 决策树学习器，针对大型数据库中连续变化的数据流，提出了动态决策树学习器 CVFDT（Concept - adapting Very Fast Decision Tree Learner），该算法能比较有效地结合历史信息和更新信息，根据更新数据动

态地建立新枝或删除旧枝。Aggarwal C. C. （2003）[55]提出能够诊断出进化数据流变化的一个算法框架。Ordonez C. （2003）[56]使用 K - means 算法对二进制数据流聚类进行了改进。Gama J. 、Rocha R. 和 Medas （2003）[57]利用精确决策树对高速数据流进行了挖掘实验。Rushing J. 、Graves S. 和 Criswell E. 等 （2004）[58]对多种分类模型结合而成的集成分类器进行了理论探索，为集成分类器的数据流挖掘应用作了铺垫性理论探索。Daniel Kifer、Shai Ben David 和 Johannes Gehrke （2004）[59]提出了一种新的数据流变化检测和估计的非参数方法，这种方法不需要假设数据流的概率分布，仅需数据流中的数据独立。Beringer J. 和 Hullermeier E. （2006）[60]提出了一种关于并行数据流在线实时聚类方法，可以快速地计算数据流之间的近似距离。Heinz C. 和 Seeger B. （2007）[61]设计了一种新颖的基于小波密度估计的方法来估计数据流的分布密度。Lior Cohen、Gil Avrahami 和 Mark Last 等 （2008）[62]认为，数据流分类挖掘需要解决的核心问题是动态数据流在具体分类模型中的实时表达和由于其动态实时性引起的“概念漂移” （Concept Drift） 问题。Chowdhury Farhan Ahmed、Syed Khairuzzaman Tanbeer 和 Byeong - Soo Jeong 等 （2008）[63]为频繁模式挖掘中的每一项引入了动态权重概念，并在此基础上提出了一个动态权重频繁模式挖掘算法，该算法能够动态处理每个单独项的权重变化情况，并运用模式增长挖掘技术来避免产生候选集，而且只需要扫描数据库一次，所以非常适合于用于对数据流的挖掘。Ludmila I. Kuncheva （2009）[64]通过对数据流挖掘中“概念漂移问题”的研究，提出了 STAGGER 系统，运用内置式属性—节点架构和阻缓机制，构建了一种解决数据流挖掘中分类漂移的算法思路。Qinghua Hu、Maozu Guo 和 Daren Yu 等 （2010）[65]将“信息熵”概念引入到数据流分类挖掘中，利用信息熵对数据流分类判定树算法进行基于概率分布的优化，对基于 Hoeffding 不等式理论增量滑动窗数据流分类模型进行了改进。Cioara Tudor、Anghel Ionut 和 Salomie Ioan 等 （2010）[66]在对数据流分类挖掘自适应算法 （Self - adaptive Algorithm） 研究后认为，利用分类器分类精度作为反馈信号对分类器技能型自适应系统调整对分类器分类效率提高具有重要作用。Thanaa M. Ghanem、Ahmed K. Elmagarmid 和 Per Ake Larson 等 （2010）[67]提出了一种同步 SQL （Structured Query Language） 查询语言，将数据流定义为一个根据关系变化而修正参数的数据序列，其中同步 SQL 语言可以通过统一的查询、输入和输出解释来集成查询功能，从而显著降低了对数据流进行实时查询的成本，并且能针对相同查询在不同的执行方法中找到最优方案。Wei Xia 和 Zhang Wei （2010）[68]通过异相矩阵提取数据流中的可疑数据，调整滑动时间窗口，建立回归方程和回归分析模型，对网络文本的可疑数据流进行了识别和预测。Hanady Abdulsalam、David B. Skillicorn 和 Patrick Martin （2011）[69]针对传统分类算法只能一次性

地扫描数据的问题，设计了一种数据流随机森林分类算法，该算法不仅能够多次扫描数据流，还能处理间歇到来的标示数据包，通过调整数据包的参数，回应不断变化的概念漂移。Hua－Fu Li（2011）[70]在基于数据结构的新的字典树框架基础上，提出了一个高实用项集挖掘算法 MHUI－max（Mining High－Utility Itemsets based on LexTree－maxHTU），用于挖掘对交易敏感的滑动窗口数据流。Qing Ling Mei 和 Ling Chen（2011）[71]提出了一种新的挖掘数据流频繁项的算法，一方面采用时间衰减因子算法突出动态数据流的时间序列，另一方面采用散列算法在 Hash 表连续记录数据项的密度，在密度阈值 s 和整数 k 条件下挖掘顶层 k 频繁项，每个数据项的平均处理时间为 O（1），提高了挖掘精度和运行速度。Jun Shan Tan、Zhu Fang Kuang 和 Guo Gui Yang（2012）[72]提出了基于有向图和数据流概要结构的频繁模式挖掘算法 FPD－Graph，通过数据流滑动窗口对 FPD－Graph 列表节点进行插入和删除操作，在有效提高频繁模式挖掘效率的同时，降低查询时间。

国内针对数据流的动态挖掘研究起步较晚。黄磊（2007）[73]和孙玉芬、卢炎生（2007）[74]对国内外数据流挖掘的研究进展，从数据流模型、数据流技术、数据流挖掘算法等多个方面进行了系统的综述。常建龙、曹锋和周傲英（2007）[75]等提出了受伪（False Positive）和拒真（False Negative）两种聚类特征指数直方图，分别支持受伪误差和拒真误差窗口的聚类分析，并提出了一种基于滑动窗口的数据流聚类方法。该方法在占用窗口大小的次线性内存空间前提下，能够及时保存最近数据记录的分布状况，从而实现对滑动窗口内的数据进行聚类。李人和、宫学庆和常建龙等（2007）[76]指出，除了有限内存和扫描一次的限制之外，数据流聚类算法还要求事先不假设聚类个数、能发现任意形状类、能够处理异常点等，提出了满足上述三个要求的一种新的聚类算法：基于密度的带噪声的演变数据流聚类算法 DenStream，并且对实际数据和人工合成数据进行了聚类分析检验，证明了该算法的有效性。王涛、李舟军和颜跃进等（2007）[77]从数据流平稳分布和带概念漂移两个方面对国内外数据流分类挖掘进行了研究综述，对数据流分类挖掘中存在的数据连续性处理、概念漂移、样本抽取、分类精度以及数据流预处理问题等进行了详细的文献述评，并提出了若干该领域需要解决的难点问题。陈安龙、唐常杰和元昌安等（2007）[78]提出了基于 Haar 小波技术和耦合特征的多数据流压缩方法，设计了多尺度能量分解压缩算法以及多尺度重构算法，该算法揭示了数据流的耦合度与变化趋势的相关性、耦合度的平移不变性及等价规律，采用特征流序列的小波系数和流能量近似表示流的趋势。胡彧、闫巧梅（2008）[79]改进了基于滑动窗口的数据流聚类算法，采用聚类特征指数直方图来支持数据处理，减少了直方图结构的维护数，与传统基于界标模型的聚类算法相比，在复杂度、聚类效果上得到了进一步改善。朱小栋、黄志球和陈圣青等

(2008)[80]提出了一种基于 Web 服务的数据流挖掘过程模型算法管理框架 PMAMF - DSM，在 Eclipse 上基于该框架实现了一个数据流挖掘算法管理系统，实验结果表明了该框架的灵活性与自适应性。李光、赵虎和代春明等（2009）[81]运用流形理论对分类数据挖掘进行了区别于传统全局性静态算法的局部性分类算法动态优化，为数据流挖掘提供了一种有效思路。毛国君、宗东军（2009）[82]提出了基于多维数据流挖掘技术的入侵检测模型与算法。欧阳震诤（2009）[83]在其博士学位论文中对不平稳数据流分类技术进行了全面的梳理和讨论，对不平稳数据流分类挖掘的框架、系统和具体算法进行了迄今为止较为详尽的理论总结。王锡文、贾银山（2010）[84]在对数据流动态到达的实时性、到达顺序的不可控性以及海量数据流存档困难等特点进行考察和对经典的处理概念漂移的动态裁剪集成分类器（SEA）算法改进的基础上提出了 ECRRC 算法，有效兼顾了分类器自动更新和保持被抛弃概念模式等数据流挖掘要求，从而提高了分类器模式更新的效率和分类精度。谭军、卜英勇和杨勃（2010）[85]针对频繁模式增长算法无法适应数据流的无限性和流动性的特点，提出一种新颖的 FP - tree 的变形结构 - SP - tree，只需单次扫描便能容纳全部数据库信息，从而有效解决了针对数据流的频繁模式挖掘算法中无法多次扫描的问题。袁正午、程宇翔和梁均军等（2010）[86]针对关系型数据流，提出一种基于流立方体框架的频繁模式挖掘算法，通过数据流的不断到达动态地创建流立方体来保存近期数据流信息，当用户提出查询请求时在以创建的流立方体基础上进行频繁模式的挖掘计算，可以快速地挖掘数据流备维之间存在的所有频繁模式。罗聪、任广伟（2011）[87]提出了一个基于改进的距离和密度的孤立点挖掘算法，该算法在数据流上利用聚类的方式，快速过滤掉数据域中比较稠密的数据，在稀疏区域找到可能成为孤立点的候选集合，然后利用聚类特有的统计信息对孤立点的偏离程度进行估计，以便于用户查询。屠莉、陈峻（2011）[88]提出了一种数据流上的频繁项挖掘算法 SW - COUNT，该算法通过数据采样技术挖掘滑动窗口下的数据流频繁项，在 O（ε^{-1}）空间复杂度下，检测误差在 εn 内的数据流频繁项，对每个数据项的平均处理时间为 O（1），具有较好的精度质量以及时间和空间效率。刘畅（2011）[89]提出了一种改进的加权随机抽样算法 IWRS，该算法根据数据流变化的快慢程度，动态的对数据流加权，将权值作为数据项的键值，根据键值大小、skipping 因子和退避因子对数据流进行抽样，解决了现有的抽样算法生成的概要数据与原始数据偏离大小不确定以及数据稳定度低时生成概要数据效率不高的问题。戴奇波、倪志伟和王超等（2011）[90]提出了一种基于动态数据流挖掘的案例推理模型，其中动态数据流挖掘算法采用改进的数据流聚类算法，以解决案例推理中知识获取、知识库更新等应用的“瓶颈”问题。