学校编码：10384　　分类号_____密级_____

学　　号：24320071151823　　UDC _____

厦门大学

硕 士 学 位 论 文

高维聚类在银行个人信贷分析中的应用

The Application of High-Dimensional Clustering in Analysis of Bank Personal-Credit

陈 晓 洁

指导教师姓名：董 槐 林 教授

专 业 名 称：计算机软件与理论

论文提交日期：2010 年 5 月

论文答辩时间：2010 年 5 月

学位授予日期：2010 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2010 年　5　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

随着我国金融体制改革的逐渐深入和中国金融业的开放，各银行将面临前所未有的竞争压力，银行如何提高决策能力和速度以适应这种压力是一个挑战。银行业务产生大量数据，利用目前的数据库系统虽然可以高效地进行数据的录入、查询、统计功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。数据挖掘技术可以从数据中抽取具有潜在使用价值的信息，通过数据挖掘技术对银行数据进行分析，可以发现重要的数据模式，将银行的数据“坟墓”转换为知识“金块”,为决策的制定提供有力的支持。因此，在商业银行应用数据挖掘技术具有重要意义。

信贷业务是商业银行的核心业务，个人信贷业务以其风险分散、利息收入稳定、市场潜力大和衍生效益明显等优势成为商业银行的重要战略性业务。本文根据银行数据量大、数据维度高的特点，使用高维聚类分析技术对银行个人信贷业务数据进行具体分析。传统的聚类方法是在聚类前为整个未标记的数据集确定一组单一的特征子集或特征权重。由于忽略了数据集中不同的子结构的存在，不同的子结构要求不同的特征权重子集，聚类过程的性能都被大大地降低。本文实现的局部特征加权聚类算法 SCAD 能够同时执行聚类和特征加权，它的连续的特征权重提供了一种比二元特征选择更富有特征相关性的表示法。其次 SCAD 能够适应存在于数据集中的变化将它分到不同的类，由于使用模糊隶属度，它还允许重叠。

本文首先介绍了数据挖掘、高维聚类技术相关的概念，然后进行高维聚类数据的准备，如数据的获得、数据转换、数据整合及数据清理，并在对 FCM 算法改进的基础上，实现了同步聚类和属性识别算法 SCAD，用该算法对个人信贷偿还信息进行了聚类分析，聚类效果杂乱程度低，精确率高，可以有效的发现影响贷款的主要因素的权重。银行可根据主要因素的权重组合，实现潜在客户预测，对银行高层进行决策提供了科学的依据。

关键词：高维聚类；个人信贷；SCAD

Abstract

With the gradually deepening of China's financial reform and the opening of China's financial industry, banks will face unprecedented competitive pressure. The banks how to improve decision-making ability and speed to adapt to such pressures is a challenge. Using the current database system, banks have a lot of data. Although the system can effectively carry out data entry, query and statistical functions, but it can not find the existence of the relationship and rules in the data. It can not predict the future development of trend based on the data. Data mining technology can potential helpful information form the data. Using data mining techniques to analyze the data of the bank, we can find important data model, convert the data "grave" to knowledge "gold", and provide strong support to the making of decision. Therefore, data mining technology is important to the commercial banks.

Credit business is the core business of commercial banks. Because of these advantages, which are risk diversification, interest income, great market’s potential and derived significant benefits. This dissertation use high-dimensional clustering to analysis the personal-credit information of the banks. The traditional approach in this case is to determine a single subset of feature weights for the entire unlabeled data set prior to clustering. However, by ignoring the existence of different sub-structures in the data set, which requires different subsets of feature weights, the performance of any clustering procedure can be severely degraded. Simultaneous Clustering and Attribute Discrimination performs clustering and feature weighting simultaneously. Its continuous feature weighting provides a much richer feature relevance representation than binary feature selection. Second SCAD can adapt to the changes present in the data set into different categories. With the use of fuzzy membership, it also allows overlapping.

This dissertation introduces the relevant concept of data mining and high-

dimensional clustering, then prepare the personal-credit data, such as data acquisition, data transform, data integration and data cleaning, then on the basis of FCM algorithm, achieve Simultaneous Clustering and Attribute Discrimination algorithm. Using SCAD in analysis of the data of the bank personal-credit, the effect of clustering is low clutter and high precision. It can grasp the situation of personal-credit customers and find the weight association of the main factors in personal-credit information. On this basis, the bank can adjust the policy of releasing and predict potential customer group.

Keywords: High-Dimensional Clustering; Personal-Credit; SCAD

目录

Contents

第1章 绪论

1.1 选题的依据及意义

数据挖掘(Data Mining，DM)是当今信息科学领域中十分活跃的研究热点。自 20 世纪 80 年代中期以来，随着数据库技术和信息技术的不断发展，大量数据库和信息存储技术被用于事务管理、科学研究、工程开发、信息检索和数据分析等。随着我国金融体制改革的逐渐深入和中国金融业的开放，各银行将面临前所未有的竞争压力，银行如何提高决策能力和速度以适应这种压力是一个挑战。银行每天的业务都会产生大量数据，利用目前的数据库系统虽然可以高效地进行数据的录入、查询和统计功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。这样，使得银行很多重要的决定不是基于数据库信息丰富的数据，而是基于决策者的直觉。而利用数据挖掘聚类分析技术进行数据分析，不但可以从海量数据中发现隐藏在其后的客观规律，将银行的数据“坟墓”转换为知识“金块”，而且可以很好地降低金融机构存在的风险。因此，在商业银行应用数据挖掘技术具有重要意义。

对于金融这个高风险和高回报的行业，关键是要能够在把风险控制到自己能承受的限度的同时，获得最大的利润。由于银行业是典型的以客户为导向的服务行业，对客户的理解程度很大程度上决定了企业的成功与否[1]。信贷业务是银行的核心业务，个人信贷业务以其风险分散、利息收入稳定、市场潜力大和衍生效益明显等优势成为商业银行的重要战略性业务[2]。大力发展个人信贷业务已成为各家银行的普遍共识。个人信贷是指商业银行将资金借贷给个人或家庭使用和消费[3]，近年来，商业银行个人信贷业务迅猛发展。为达到风险控制，获得最大利润，必须对客户信息进行科学的分析，以及时发现问题，化解风险。在这个方面，数据挖掘应用的效果将是突出的。本课题旨在运用数据挖

掘聚类分析技术，对银行个人信贷业务进行分析，找出其内在的规律，提高管理效率和决策的科学性，降低运营风险，提高银行的竞争力。

1.2 国内外研究现状及发展趋势

数据挖掘是 20 世纪 80 年代后期兴起的新技术，是多门学科和多种技术相结合的产物，也是一个非常年轻而又活跃的研究领域。20 世纪 90 年代中后期，国外数据挖掘已经形成高潮，国内研究数据挖掘的学者数量也在迅速增长。

经过十几年的研究，数据挖掘技术得到了飞速的发展，数据挖掘的方法层出不穷，其中主要有：关联规则、分类分析、聚类分析、统计分析等等[4]，其中每类方法又有多个不同的具体算法，大量的学者对各种算法进行研究，使得算法的性能不断提高。在研究过程中产生了许多数据挖掘工具，通过数据挖掘工具来实现数据挖掘的目的，显然可以达到事半功倍的效果。

数据挖掘在银行领域的应用主要有四个方面[5]:

(1)银行客户关系管理(Customer Relation Management, CRM)

(2)银行风险管理

(3)银行信用等级评估

(4)银行服务分析和预测

数据挖掘技术在国外银行业有很多成功的应用案例:美国 Firstar 银行使用 Masksman 数据挖掘工具，根据客户的消费模式预测何时为客户提供何种产品。美国 Bankone 银行用各种集中起来建立数据仓库，从建立的数据仓库中挖掘出为银行创造利润的这部分客户，从复杂的客户信息中建立模型，对客户记录信息进行动态跟踪和监测，计算客户价值，锁定特定客户群，制定不同市场需求、不同客户群的市场战略，根据客户的价值选定服务产品配置，从而与创造利润的优良客户建立长期关系[7][8]。这些模式帮助提高了客户忠诚度。爱尔兰最大的银行 AIB 采用 IBM 的 Intelligent Miner 数据挖掘工具，成功地从大量的

账户和交易数据中挖出宝贵的商业信息，使该银行能够较准确地预测诸如客户欺诈贷款的可能性、不同客户接受一个特殊银行产品的倾向等事件，指导银行及时推出针对性的营销活动，帮助银行不断扩大市场份额[9]。Mellon 银行使用 Intelligent Miner 数据挖掘软件对银行数据进行分析，发现其数据模式及特征，然后可以发现某个客户、消费群体或组织的金融和商业兴趣，并可观察金融市场的变化趋势[10]。

中国商业银行的信息化在近 10 年内有了长足的发展，从最初的业务处理电子化，到后来各银行内部网络和垂直业务体系的建成，直至数据大集中工程的实施，达到了前所未有的高度。但是，信息化本质是保证银行具备核心竞争力的一系列重要工具，而在信息化工具组合中，更为锐利、高效和复杂的数据挖掘工具，还没有被中国银行业所广泛掌握。目前中国银行业数据管理应用的普遍现状是，银行汇集了大量数据，但缺乏挖掘数据底层隐藏知识的手段和工具，往往导致“数据爆炸但知识贫乏”。如何游出“数据海洋”，把海量数据用于提升客户关系、挖掘客户价值、掌握业务规律——这一切难题，在没有掌握数据挖掘能力的银行，目前还都处于“瓶颈”阶段，有待解决。在银行业，由于银行产品具有相当的同质性，因此银行之间的差别，往往在于谁掌握了客户关系，以及海量的业务和客户信息背后的独特业务规律，谁就可以科学地制定决策。现在银行实施的大多数系统所基于的数据库只能实现数据的录入、查询、统计等较低层次的功能，但却无法发现数据中存在的关联关系和业务规律，更难以根据现有的数据预测未来业务的发展趋势。目前看来，在银行管理客户生命周期的各个阶段都会用到数据挖掘技术：数据挖掘能够帮助银行确定客户的特点，从而可以为客户提供有针对性的服务；通过数据挖掘，可以发现购买某类金融产品的客户特征，从而可以扩大业务；如果找到了流失客户的特征，就可以在具有相似特征的客户还未流失之前，采取针对性的措施——银行的客户获得、交叉销售（Cross-selling）、客户关怀与保持等方面，数据挖掘技术都将发挥重要作用。

1.3 研究目标与内容

1.3.1 研究目标

对数据挖掘技术进行深入研究，主要是探讨数据挖掘的过程、数据预处理的方法和高维数据聚类方法。通过本人已有的研究——三种局部特征加权聚类算法的性能比较和分析，确定本文应用的高维数据聚类算法SCAD(Simultaneous Clustering and Attribute Discrimination)。使用聚类分析技术对银行个人信贷业务数据进行具体分析，掌握个人信贷客户的情况，发现贷款信息中主要因素的权重组合，为银行高层防范风险、开发新产品、推出新的服务提供决策依据，吸引更多的客户，帮助银行不断扩大市场份额，提高银行的竞争力。

1.3.2 研究内容

（1）数据挖掘技术分析。了解当前国内外数据挖掘的研究现状及发展趋势。

（2）高维数据聚类算法研究。掌握高维数据的特点及聚类算法，并通过已有的研究，确定本文中需应用的聚类算法 SCAD，并实现了 SCAD 算法。该算法可以同时地搜索最优聚类中心和最优特征权重集。

（3）数据挖掘应用研究。这方面主要进行银行个人信贷业务分析，即聚类分析技术在银行个人信贷方面应用。其中涉及到多个核心问题的解决及数据挖掘任务的实施和完成。根据具体情况采用多种方法对数据进行预处理；使用MATLAB7 实现系统；对确定的分析主题进行了具体的挖掘实验，并对结果进行评估和解释。

1.4 主要特色及内容安排

1.4.1 主要特色

本课题主要对银行个人信贷数据进行聚类分析，以发现贷款信息中主要因

素的权重组合，为银行高层防范风险、开发新产品、推出新的服务提供决策依据。传统的聚类方法是在聚类前为整个未标记的数据集确定一组单一的特征子集或特征权重。由于忽略了数据集中不同的子结构的存在，不同的子结构要求不同的特征权重子集，聚类过程的性能都被大大地降低。而本文应用的局部特征加权聚类算法 SCAD 能够同时地执行聚类和特征加权，它的连续的特征权重提供了一种比二元特征选择更富有特征相关性的表示法。其次 SCAD 能够适应存在于数据集中的变化将它分到不同的类，由于使用模糊隶属度，它还允许重叠。该算法可以同时地搜索最优聚类中心和最优特征权重集，大大的提高了聚类过程的性能。

1.4.2 内容安排

本文将以如下的组织结构展开描述：

第一章 绪论。介绍论文研究背景，相关技术在国内外银行业的应用现状，论文的研究目标和内容，以及论文的主要特色及内容安排。

第二章 高维数据聚类分析。简要描述高维数据聚类相关的基本理论，包括数据挖掘的定义及其过程，聚类问题的描述、聚类中常用的概念及聚类技术分类，高维数据聚类的概念及其分类。

第三章 个人信贷数据准备。主要是确定数据挖掘所需的信息，并进行相应的数据预处理，包括数据清理，数据变换，数据规约等。

第四章 高维聚类在银行个人信贷分析中的应用。主要阐述同步聚类和属性识别算法 SCAD，并介绍本文应用的两种评价标准。应用 MATLAB7 实现聚类系统，并结合第三章准备好的数据，进行具体的挖掘实验，并对挖掘结果进行分析。

第五章 结论与展望。

第2章 高维数据聚类分析

2.1 数据挖掘简介

2.1.1 数据挖掘的定义

随着数据库信息系统和计算机网络的发展与应用，各种数据源中都积攒了大量的数据，造成了“数据丰富，但信息贫乏”的尴尬局面。数据挖掘作为一个面向这个问题的技术就应运而生了。数据挖掘,顾名思义就是从大量的数据中挖掘出有用的信息，即从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中发现隐含的、规律性的、人们事先未知的，但又潜在有用的并且最终可理解的信息和知识的非平凡过程[12]。数据挖掘又称作数据库中的知识发现（Knowledge Discovery in Database，KDD），在 KDD 96 国际会议上，KDD 被定义为：对数据库中蕴涵的、未知的、有潜在应用价值的、非平凡的模式的提取。它包括数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估和知识表示等多个步骤，数据挖掘是对经过预处理的数据进行处理抽取知识的过程[13]。

数据挖掘的对象往往是大规模的高维数据，这些数据可能来自于数据库、数据仓库或其它数据源。同时，数据挖掘的结果是准确的、有用的、未知的、可解释的，知识可能以各种形式存在：概念、规则、模式、约束等。另外，数据挖掘的目的是支持决策分析，由于决策分析往往是有时间要求的，所以数据挖掘过程必须高效。

数据挖掘是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程，这些模型和关系可以进行预测，它帮助决策者寻找数据间潜在的关联，发现被忽略的模式，因而被认为是解决当今时代所面临的数据爆炸而信息贫乏问题的一种有效方法。数据挖掘是一门交叉学科，融合了数据库[13]、人工智能[14]、机器学习[15]、统计学[16]等多个领域论和技术。数据库、人工智能和数理统