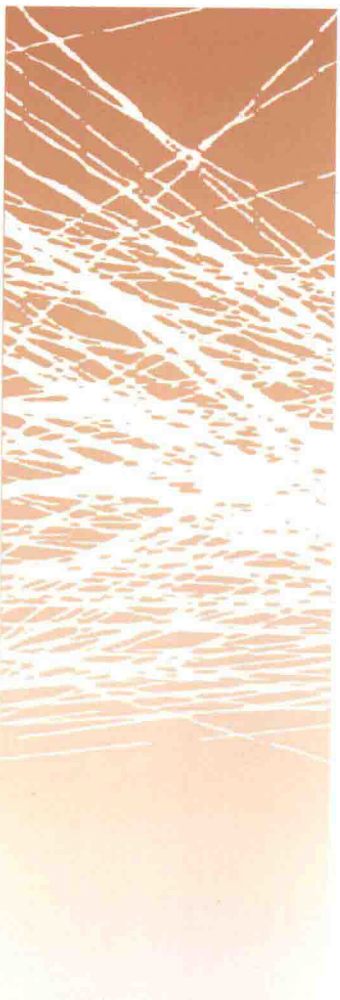
祝琴 著

Subspace Recognization for
Object-Attribute Space with High-Dimension
Sparse Feature

高维数据分析预处理技术



社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

祝琴 著

高维数据分析预处理技术

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

图书在版编目(CIP)数据

高维数据分析预处理技术/祝琴著. —北京：社会科学文献出版社，2015.12

ISBN 978-7-5097-8569-0

Ⅰ.①高… Ⅱ.①祝… Ⅲ.①统计数据-统计分析 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字（2015）第 312302 号

高维数据分析预处理技术

著　　者／祝　琴

出 版 人／谢寿光
项目统筹／王玉敏
责任编辑／王玉敏　张文静

出　　版／社会科学文献出版社·国际出版分社（010）59367243
地址：北京市北三环中路甲 29 号院华龙大厦　邮编：100029
网址：www.ssap.com.cn
发　　行／市场营销中心（010）59367081　59367090
读者服务中心（010）59367028
印　　装／三河市尚艺印装有限公司

规　　格／开 本：787mm×1092mm　1/16
印 张：11.25　字 数：156 千字
版　　次／2015 年 12 月第 1 版　2015 年 12 月第 1 次印刷
书　　号／ISBN 978-7-5097-8569-0
定　　价／49.00 元

本书如有破损、缺页、装订错误，请与本社读者服务中心联系更换

序　言

数据挖掘中有一类典型的数据分析问题，是对所分析的对象集合进行聚类、分类分析。但是，在数据挖掘的实际应用中，由于对象属性的高维特征，导致数据挖掘问题的规模巨大，数据挖掘变得异常地困难，甚至导致传统、经典的数据挖掘算法由于计算量大而丧失实用价值。

高属性维数据是比较常见的一种数据形式，对高属性维数据的处理能力是数据挖掘研究与应用中的重要内容。大量的生产管理实践表明，数据挖掘的实际应用问题面对的数据具有高维特性，同时，这些属性的取值却具有稀疏的特征，这类问题称为高维稀疏数据挖掘问题，其本质是数据分析的对象数据具有高属性维，即描述每个对象的属性有很多，但这些属性有很大一部分取值为零。

对于高维稀疏数据挖掘问题，大部分研究工作都集中在数据对象间相似度的度量方法及挖掘算法方面，如高属性维稀疏数据聚类的稀疏特征聚类法（sparse feature clustering，SFC）、基于稀

疏特征向量的聚类算法（clustering algorithm based on sparse feature vector，CABOSFV）等。

本书中，作者针对高维稀疏数据挖掘问题，从数据预处理的角度，研究对象—属性空间的划分问题，其目的是把所研究的数据挖掘空间分解为若干规模较小的对象—属性空间，从而降低实际数据挖掘的难度。

该书的研究成果，针对高维稀疏数据挖掘问题，降低数据挖掘规模，建立了体系完整的数据预处理理论和方法，具有很强的理论意义和实践应用前景。

北京科技大学经济管理学院

致　谢

衷心感谢恩师高学东教授给予我跨学科在北京科技大学攻读博士学位的机会，让我能在自己感兴趣的数据挖掘领域进行学习和研究。本书从选题、课题研究、书稿撰写到书稿完善都是在高教授的悉心指导下完成的。高教授不仅学识渊博、治学严谨，而且思想开明、实事求是。他开阔的视野和敏锐的思维给了我深深的启迪。师从高教授不仅让我学到了知识，更重要的是学到了思想，领悟到了很多做人的道理，这将让我受益一生。在此，再次表示最衷心的感谢和最崇高的敬意！

衷心感谢北京科技大学经济管理学院张群教授、高俊山教授、李铁克教授、王道平教授、张晓冬教授、马风才教授、鲍新中教授对我的帮助，感谢经济管理学院全体教师！

衷心感谢武森教授在学习和生活上给予我的关心和帮助！

衷心感谢喻斌老师、王莹老师、周晓光老师、魏桂英老师、崔巍博士、国宏伟博士、徐章艳博士、王阅博士、陈敏博士、杨珺博士、戴爱明博士、孟陶然博士、吴玲玉博士，感谢课题组全

体老师和同学！

衷心感谢南昌大学管理科学与工程系贾仁安教授、涂国平主任、邓群钊主任的关怀和支持，感谢同事们的帮助和鼓励！

最后，特别感谢我的先生赖平红博士，感谢他一直以来给予我的理解、支持与鼓励！深情感谢我最亲爱的父母以及所有爱我的亲人和朋友们！

目录

第 1 章　引言 …… 1

第 2 章　文献综述 …… 5

2.1　知识发现与数据挖掘 …… 5

2.2　聚类分析 …… 13

2.3　数据挖掘所面临的挑战 …… 24

2.4　高维数据 …… 27

2.5　维度约简 …… 31

2.6　高维数据聚类 …… 38

2.7　本章小结 …… 43

第 3 章　基于排序的高属性维稀疏数据聚类方法 …… 44

3.1　高维稀疏数据 …… 44

3.2　高属性维聚类问题描述 …… 47

3.3　经典高属性维稀疏数据聚类 CABOSFV 方法分析 …… 54

3.4　基于排序的 CABOSFV 方法——CABOSFVABS 方法 …… 59

3.5　本章小结 …… 68

第 4 章　对象—属性空间分割的两阶段联合聚类方法 …… 70
4.1　具有高维稀疏特征的对象—属性空间分割问题的提法 …… 70
4.2　传统对象—属性空间分割方法基于内聚度方法 …… 71
4.3　联合聚类方法 …… 75
4.4　两阶段联合聚类方法（MTPCCA） …… 86
4.5　本章小结 …… 96

第 5 章　对象—属性子空间重叠区域的归属问题 …… 98
5.1　问题描述及相关研究工作 …… 98
5.2　对象—属性子空间的边缘重叠区域归属方法——OASEDA 方法 …… 108
5.3　本章小结 …… 126

第 6 章　对象—属性子空间优化 …… 128
6.1　高维稀疏特征的对象—属性非关联子空间 …… 130
6.2　剔除非关联子空间 RNASAUBSC 方法 …… 131
6.3　RNASAUBSC 方法算例 …… 136
6.4　RNASAUBSC 方法应用 …… 138
6.5　本章小结 …… 141

第 7 章　结论 …… 142

参考文献 …… 145

后记 …… 165

图目录

图 1 – 1　本书结构图 …… 4
图 2 – 1　数据库知识发现的过程图 …… 6
图 2 – 2　聚结型层次聚类和分解型层次聚类法的比较 …… 17
图 2 – 3　高维数据聚类方法的分类图 …… 39
图 3 – 1　CABOSFV 聚类方法的两层结构图 …… 58
图 4 – 1　传统聚类与联合聚类 …… 76
图 4 – 2　行和列均独立的联合聚类 …… 78
图 4 – 3　格子结构的独立联合聚类 …… 79
图 4 – 4　独立行的联合聚类 …… 79
图 4 – 5　独立列的联合聚类 …… 80
图 4 – 6　树型的没有重叠独立的联合聚类 …… 80
图 4 – 7　没有独立、没有重叠的联合聚类 …… 81
图 4 – 8　层次结构重叠的联合聚类 …… 81
图 4 – 9　6 个对象 10 种属性的对象—属性空间图 …… 90
图 4 – 10　第一阶段聚类分割后的对象—属性空间图 …… 92

图 4－11　两阶段联合聚类识别的对象—属性子空间图 …… 92
图 4－12　30 个对象、45 种属性的对象—属性空间图 …… 95
图 4－13　基于内聚度分割方法识别的对象—属性
子空间图 …… 95
图 4－14　基于 MTPCCA 方法识别的对象—属性
子空间图 …… 96
图 5－1　子空间中的交叉重叠区域图 …… 99
图 5－2　交叉重叠区域中零属性值现象 …… 99
图 5－3　聚类边界不准现象 …… 100
图 5－4　扩展 1/2 网格图 …… 104
图 5－5　同位置点距离计算情况图 …… 104
图 5－6　聚类边界点、噪声、孤立点图 …… 104
图 5－7　边界效应引起聚类效果不好图 …… 107
图 5－8　受力分析图 …… 109
图 5－9　子空间的交叉重叠区域分块图 …… 116
图 5－10　8 个对象、10 种属性的对象—属性空间图 …… 121
图 5－11　MTPCCA 方法识别的对象—属性
子空间图（一） …… 121
图 5－12　8 个对象、10 种属性的对象—属性子空间图 …… 122
图 5－13　26 个客户订购 45 种产品的对象—属性空间图 …… 124
图 5－14　MTPCCA 方法识别的对象—属性
子空间图（二） …… 125
图 5－15　根据 OASEDA 方法得出的对象—属性子空间图
…… 126

图 6 - 1　8 个对象、10 种属性对应的对象—属性子空间图 …… 129
图 6 - 2　对象—属性稀疏子空间图 …… 130
图 6 - 3　对象—属性非关联子空间图 …… 132
图 6 - 4　RNASAUBSC 方法运算过程图 …… 133
图 6 - 5　4 个对象、5 种属性对象—属性空间的优化过程图 …… 135
图 6 - 6　30 个对象、45 种属性的对象—属性子空间图 …… 137
图 6 - 7　30 个对象、45 种属性优化后的对象—属性子空间图 …… 137
图 6 - 8　8 个客户订购 10 种产品的对象—属性子空间图 …… 140
图 6 - 9　对象—属性子空间 C 的优化过程图 …… 140

表目录

表 3 - 1　高维稀疏数据 …… 46
表 3 - 2　高维稀疏的数据归一化 …… 46
表 3 - 3　高维稀疏二态数据表 …… 47
表 3 - 4　二态变量取值统计 …… 49
表 3 - 5　对象数据例表 …… 51
表 3 - 6　分类变量转化为不对称二态变量 …… 52
表 3 - 7　6 个客户订购 8 种产品的稀疏特征表 …… 55
表 3 - 8　6 个客户订购 8 种产品情况的压缩存储 …… 57
表 3 - 9　15 个客户对 48 种产品的订购情况 …… 64
表 3 - 10　CABOSFV 方法聚类结果 …… 65
表 3 - 11　CABOSFVABS 方法聚类结果 …… 67
表 4 - 1　所包含的数值都相等的联合聚类 …… 76
表 4 - 2　同列包含数值相等的联合聚类 …… 77
表 4 - 3　同行包含数值相等的联合聚类 …… 77
表 4 - 4　加法模型 …… 77

表4－5　乘法模型 …… 77
表4－6　演变趋势一致的联合聚类 …… 78
表4－7　8个客户订购10种产品的稀疏特征表（一） …… 89
表4－8　8个客户订购10种产品的压缩存储表 …… 89
表4－9　30个对象、45种属性取值的情况表 …… 93
表5－1　8个客户订购10种产品的情况表 …… 119
表5－2　8个客户订购10种产品的归一化结果表 …… 120
表5－3　8个客户订购10种产品的稀疏特征值表 …… 120
表5－4　26个客户订购45种产品的情况 …… 123
表6－1　8个客户订购10种产品的稀疏特征表（二） …… 129
表6－2　2个客户订购7种产品的稀疏特征表 …… 133
表6－3　4个客户订购5种产品的稀疏特征表 …… 134
表6－4　8个客户订购10种产品的统计表 …… 138
表6－5　8个客户订购10种产品数量归一化的数据表 …… 139
表6－6　8个客户订购10种产品的稀疏特征表（三） …… 139

第1章　引言

面对“信息爆炸”，如何迅速从海量数据中获得所需的知识，成为一个迫切需要解决的问题。在这种背景下诞生了数据挖掘（data mining，DM）技术[1]。

随着信息技术的迅猛发展，数据挖掘技术面临的不仅是数据量越来越大的问题，更重要的还是数据的高维度问题。受“维度效应”影响，许多在低维数据空间表现良好的数据挖掘方法，在处理高维数据时，从中发现有价值的知识比较困难，甚至出现错误的结果[2~6]。

具有高维稀疏特征的对象—属性空间中的对象维和属性维的数据都是高维数据，如上所述，不能将传统的数据挖掘方法直接运用到高维稀疏数据的处理中。如果能对具有高维稀疏特征的对象—属性空间进行分割以获得其相应的子空间，那么高维稀疏数据的数据挖掘问题就能转化为维数较低的稀疏特征的对象—属性子空间的数据挖掘问题，高维稀疏数据的数据挖掘问题就会大大简化。

本书重点研究高维稀疏数据问题对象—属性空间识别技术，并针对该领域的若干相关问题，提出一些解决问题的新方法和新思路，并通过实验证明其合理性。

针对具有高维稀疏特征的对象—属性空间识别问题，本书开展如下研究工作。

（1）研究已有的高维数据聚类方法。

研究者用不同的思路设计了不同的高维数据聚类方法，本书将分析这些方法的优点与不足，为进一步提出更合理的方法奠定理论基础。

（2）研究已有的高维稀疏数据常用的数据预处理方法——维数约简方法。

在已有的高维稀疏数据维数约简方法研究工作中，研究者一般从选维（特征选择）和降维两个方面设计维数约简方法。本书将研究分析这些方法的实质，为提出更适合高维稀疏数据的数据预处理方法提供理论参考。

（3）改进和提出高效的高属性维数据聚类方法。

研究分析经典的高属性维数据聚类 CABOSFV 方法，针对该方法的局限性，提出一种改进的 CABOSFV 方法，这是本书的一个重要内容。

（4）提出高效的具有高维稀疏特征的对象—属性空间分割方法。

针对高维稀疏数据具有高维度和稀疏性的特点，对具有高维稀疏特征的对象—属性空间直接分割识别其对应的子空间，从而实现高维稀疏数据的预处理。本书将研究具有高维稀疏特征的对

象—属性空间分割技术及其子空间进一步的优化问题，通过该技术可以获得具有高维稀疏特征的对象—属性的子空间，这是本书的核心研究内容。

针对以上优化问题和研究内容，本书分为七章。

第 1 章：论述本书的目的与意义和主要研究内容，最后给出全书的组织结构。

第 2 章：对本书所涉及的数据挖掘与知识发现理论做了较为基础的概述，重点介绍聚类分析内容、高维数据的形态和特点，分析高维数据常用的预处理方法——维数约简，最后系统概述目前几种主要的高维数据聚类分析方法。

第 3 章：提出一种改进的 CABOSFV 高属性维稀疏数据聚类方法。研究分析经典的高属性维稀疏数据聚类 CABOSFV 方法的不足，提出融合排序思想的高属性维稀疏数据聚类方法。

第 4 章：给出具有高维稀疏特征的对象—属性空间的定义，提出对具有高维稀疏特征的对象—属性空间分割的方法识别其子空间的思想，并提出一种新型的两阶段联合聚类的方法，实现对高维稀疏数据的对象维和属性维进行聚类分割以识别其子空间。

第 5 章：提出对象—属性边缘重叠区域的归属判断方法。研究发现了具有高维稀疏特征的对象—属性子空间边缘可能存在交叉重叠区域现象，设计了对象—属性子空间交叉重叠区域的归属判断目标函数。

第 6 章：提出高维稀疏对象—属性子空间优化方法。通过对对象—属性子空间识别过程的分析，发现对象属性取值全为零的子空间，在此基础上给出了非关联子空间的定义，揭示了非关联