聚类算法中的优化方法应用

JULEISUANFA
ZHONGDE YOUHUA FANGFA YINGYONG

陈新泉 著

电子科技大学出版社

聚类算法中的优化方法应用

JULEI SUANFA ZHONGDE YOUHUA FANGFA YINGYONG

陈新泉 著

电子科技大学出版社

图书在版编目（CIP）数据

聚类算法中的优化方法应用 / 陈新泉著. —成都：
电子科技大学出版社，2014.7
ISBN 978-7-5647-2409-2

Ⅰ. ①聚… Ⅱ. ①陈… Ⅲ. ①聚类分析－最优化算法
Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字（2014）第 128449 号

内 容 提 要

数据挖掘是从统计学、机器学习、最优化方法等学科中发展起来的一门新兴交叉学科，目前已被广泛应用到电子商务、医学、科学研究以及工程技术等领域中，它具有重要的理论与应用价值。当前，海量数据和混合属性数据集的数据挖掘应用越来越多，面对如此复杂的数据挖掘类型，现有的许多数据挖掘算法力不从心。如何充分利用优化方法来提高数据挖掘算法的效率，改善挖掘的结果，是众多研究者关心的热点。

本书将优化方法充分应用到聚类分析领域，从多个角度研究将特征权重优化嵌入到混合属性数据集的聚类算法中，以期优化后的特征权重有助于构造出更简洁、更精确的分类器。

本书可作为聚类分析领域研究生的教学和科研参考书，也可成为智能数据分析与处理技术人员的自学研究参考书。

重庆三峡学院科学研究项目计划资助（NO.12RC01）

聚类算法中的优化方法应用

陈新泉 著

出　　版：电子科技大学出版社（成都市一环路东一段 159 号电子信息产业大厦
邮编：610051）
策划编辑：曾　艺
责任编辑：曾　艺
主　　页：www.uestcp.com.cn
电子邮箱：uestcp@uestcp.com.cn
发　　行：新华书店经销
印　　刷：成都蜀通印务有限责任公司
成品尺寸：170 mm×240mm　　印张 8.75　　字数 200 千字
版　　次：2014 年 7 月第一版
印　　次：2014 年 7 月第一次印刷
书　　号：ISBN 978-7-5647-2409-2
定　　价：30.00 元

作 者 简 介

陈新泉，男，1974 年 7 月生，湖南安仁人，理学学士，工学硕士，工学博士，电子科技大学计算机科学与技术博士后流动站在站博士后，副教授（2009 年由讲师晋升为副教授，目前受聘于重庆三峡学院），CCF 与 ACM 会员。曾被多个国际、国内 EI 会议及若干期刊邀请担任审稿人，也曾为 SCI 期刊《Computers & Electrical Engineering》《Computational Intelligence》和《Expert Systems With Applications》评审过多篇论文。

陈新泉已在数据挖掘领域从事了十多年的研究，有着较为丰富的科学研究和工作经历，在混和型数据集的加权聚类分析及特征权重优化方面有着较为深入的研究，取得了一些成果。到目前为止，主持完成省级科研项目两项和校级科研项目一项，主持校级科研项目一项，以第一作者在 JCR 二区 SCI 期刊《Journal of Intelligent Information Systems》，国际 EI 期刊《Journal of Computers》和《Journal of Software》，国内 CSCD 核心库期刊《数值计算与计算机应用》《计算机工程与应用》《计算机工程与科学》等共发表 40 余篇学术论文。

前　言

数据挖掘是从统计学、机器学习、最优化方法等学科中发展起来的一门新兴交叉学科，目前已被广泛应用到电子商务、医学、科学研究以及工程技术等领域中，它具有重要的理论与应用价值。当前，海量数据和混合属性数据集的数据挖掘应用越来越多，面对如此复杂的数据挖掘类型，现有的许多数据挖掘算法力不从心。如何充分利用优化方法来提高数据挖掘算法的效率，改善挖掘的结果，是众多研究者关心的热点。

本书将优化方法与数据挖掘结合起来进行研究，分析了数据挖掘的两个重要分支——聚类和分类，将特征权重优化与聚类和分类联系起来并进行相互融合，形成一条贯穿全文的主线。从多个角度研究将特征加权嵌入混合属性数据集的聚类和分类中，以期优化后的特征权重能有助于构造出更简洁、更精确的分类器。

本书的创新点主要表现在以下六个方面：

（1）为克服 k-means 聚类算法对初始化过于敏感的缺点，提出了一种具有单纯形思想的 k-中心点轮换法。仿真实验及分析表明，该方法在应用于那些具有一定聚类结构、各个簇大小相差不太大的数据点集时，具有良好、稳定的（对初始中心点集的选取不敏感）聚类效果，但其缺点是时间复杂度较高。从仿真实验结果中还归纳出一个具有直观性的实验结论。为在聚类质量与时间复杂度之间取得良好均衡，提出了一种基于近似类抽样的组合聚类算法。仿真实验表明，该方法效果良好，并具有一定的实用性。

（2）将求解单点优化解的 Rosenbrock 搜索法应用到具有 k-代表点优化解特征的聚类分析中，给出了一种适合于数值型数据集的新的聚类分析算法。

（3）为使特征加权后的数据点集具有更好的聚类分布性质，提出了一个可体现“聚类之内的数据点最大限度的相近，聚类之间的数据点最大限度的相离”（相近相离原则）的混合目标函数。为求解该混合目标函数，提出了一种基于负投影梯度的特征权重的自适应优化方法。仿真实验表明，该方法在优化连续有序数据集的特征权重时是有效的。

（4）利用核映射将原始样本空间中的分类问题与特征空间中的聚类问题联系起来，提出了一个可体现核空间中数据点像集相近相离原则的带线性约束条件的非线性混合目标函数。为解决该非线性优化问题，提出了一种基

于核映射的属性权重的自适应优化方法。仿真实验表明，该方法在属性选择、确定属性权重方面是有效的。

（5）参照 Joshua Zhexue Huang 等将 k-means 聚类算法与特征权重优化相结合的方法，推导出 FCM 聚类算法与特征权重优化相结合的优化迭代公式，形成加权 FCM 算法。将加权 FCM 算法中计算聚类均值项的公式代入到计算隶属度的更新公式和特征权重的更新公式中，得到加权 FCM 扩展算法。由于这个扩展算法消去了均值项，它对于有序属性和无序类别属性的隶属度和特征权重的更新公式具有统一的形式，因此可以很方便地应用到混合属性数据集的加权聚类分析中来。该算法的收敛性分析与 FCM 类似，算法迭代结束后能给出一组优化的特征权重值。仿真实验结果与 WKMeans 算法的结果基本一致，说明该方法在优化混合属性数据集的特征权重时是有效的。

（6）应用决策树方法来获取混合属性数据集的“规则聚类区域”，利用“异类子聚类相离，同类子聚类相近”的原则来交替优化有序属性和无序类别属性的权重，提出了基于决策树划分的特征权重优化方法。该方法在一定程度上解决了有效获取数据子集的子聚类问题和混合属性数据集的特征权重优化难题。仿真实验表明，该方法在优化混合属性数据集的特征权重时是有效的。

最后，列出了一些与本书相关的可行研究方向。

著　者

2014 年 1 月

目　录

第1章 绪 论

第一节 引 言

数据的分析与处理早已存在于科学研究中了，如 17 世纪的开普勒将其老师第谷花费几十年记录的天文观测数据资料进行数据建模分析，从而得到行星运动三大定律；19 世纪物理学的一些规律就是对实验数据进行分析和拟合而得到的。到 20 世纪 90 年代为止，由于数据库技术在科学、工业、商业等领域中的长期广泛应用而积累了大量的数据资源，但这些数据资源对决策者并未起到多大作用，从而出现了“海量数据，知识贫乏”的状况，数据挖掘技术就是为解决这种困境而于 20 世纪 80 年代诞生、90 年代得到蓬勃发展的新的交叉研究领域。

中国几千年的历史中不时闪烁着优化方法的智慧，中国古兵书中就有应用优化策略的影子，例如，田忌赛马的故事中就提到了如何运用优化策略来获取最终的胜利。在国民经济和军事上，优化方法和技术被广泛地应用和研究，优化这个概念更是与生产、生活息息相关。

数据挖掘是从数据中挖掘有价值的知识和规律，在整个挖掘过程中，时时、处处都可以利用优化方法来提高算法的挖掘效率，改善挖掘结果。甚至有些数据挖掘问题还可以采用数学的形式化描述，将它转换为一个数学规划问题（即数学建模），然后利用优化方法进行求解，将求解得到的最优解或满意解置入具体问题中来验证其有效性与合理性。从数学规划的角度研究数据挖掘问题，是数据挖掘与最优化方法结合研究的一个重点。

本书主要从优化方法与数据挖掘的结合点处展开研究，采用相关或改进的优化方法来改善挖掘结果。着重研究如何采用优化方法来提升聚类算法的聚类质量，从多个角度研究特征权重优化问题，提出了适应于混合属性数据点集的数据分析与处理算法，以及如何应用特征加权的聚类方法来构造出更简洁、更高效的分类器。

第二节 本书的研究背景与意义

到目前为止，数据挖掘在继承与发展机器学习、统计学等学科已有成果的基础上取得了很大的进步，构造出一个独具特色的理论体系。随着应用需求的推动以及数据挖掘技术本身和相关技术的发展，又促进了新的数据挖掘

理论和技术的出现。新的数据挖掘技术的出现，反过来又会对特定的应用产生推动作用。因此，对数据挖掘理论和算法的探讨将是长期而艰巨的任务。

随着数据挖掘应用的日益广泛，寻找可伸缩的、高效的、能处理复杂数据类型的数据挖掘方法成为当前的迫切需要。目前在高维数据挖掘、地理空间数据挖掘、多媒体数据挖掘、时间序列数据挖掘以及 Web 挖掘方面已取得一些进展，但仍不能满足实际应用的需求，所以有必要充分应用优化方法来开发出满足实际需要的数据分析方法和数据挖掘算法。

2005 年国家自然科学基金提出“最优化与数据挖掘（G0110）”这个支持项目，提到对这两个领域进行相结合研究的一点思路：以用最优化及其相关方法探寻并发展高水平的、具有实际应用价值的数据挖掘技术，旨在高效率、高精度地从海量的数据中发现潜在、新颖及有用的知识。主要研究内容应涉及基于最优化及其相关方法的分类、聚类、预测、模式发现等数据挖掘技术的探讨，从建模、特征、算法、有效性、实用性等方面寻求众多最优化及其相关方法与各种数据挖掘技术的结合；研究应考虑问题及求解中的非结构化、非线性、近似性、不确定性等特点，并应结合实际管理决策的应用问题展开研究[1]。

聚类与分类是数据挖掘领域两个重要的研究分支，在许多行业中都有其应用。本书主要研究既可处理有序属性，又可处理无序类别属性的数据挖掘算法，特别是将特征加权与聚类和分类融合起来，从多个角度研究了特征权重优化方法，并比较评估了优化后的特征权重对于提高基于抽样的加权最近邻分类器的分类精度是否有效。而确定合适的相似性度量和混合属性数据集的聚类分析与分类建模在数据挖掘理论研究和工程应用中都具有重要的地位，因此，本书的研究具有一定的理论价值和实际意义。

第三节　本书的研究思路与研究主线

一、本书的研究思路

从数据中挖掘出知识的过程可以看成是一个优化过程，目的是为了得到最优知识。许多数据挖掘应用都是目标驱动的，即针对不同的应用目的，采用适合实际问题的数据挖掘算法及相应的优化方法来进行挖掘。

图 1-1 可以大致说明最优化方法与数据挖掘这两个研究领域能互相结合起来研究，具体的结合往往是根据实际问题的需要来开发高效的数据挖掘算法，寻找最优挖掘过程和发现最优参数配置的规律。如果整个挖掘过程是一个最优过程，根据组合优化原理，最优挖掘过程的每个子过程也都是一个最优的子过程。将最优化方法应用到这个数据挖掘框架的具体挖掘算法中，以获得最优挖掘结果。

二、本书的研究主线

本书研究分析了数据挖掘中的两个重要分支——聚类和分类，通过优化方法在数据挖掘中的应用研究，将特征权重优化与聚类和分类联系起来，形成一条贯穿全文的研究主线。以特征权重优化与聚类和分类这三个研究方向的相互融合来展开研究，将特征加权嵌入到混合属性数据集的聚类与分类中，从多个角度展开分析和研究特征加权在聚类与分类中的应用，以期应用优化后的特征权重能构造出更简洁、更精确的分类器。如图 1-2 所示。

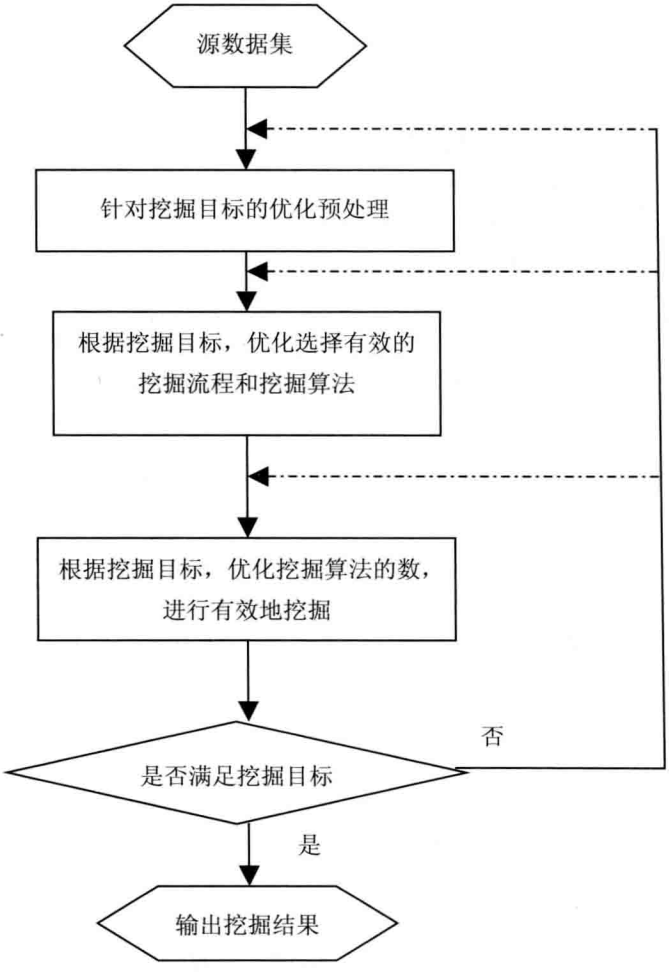

图 1-1　最优化方法应用到数据挖掘中的框架图

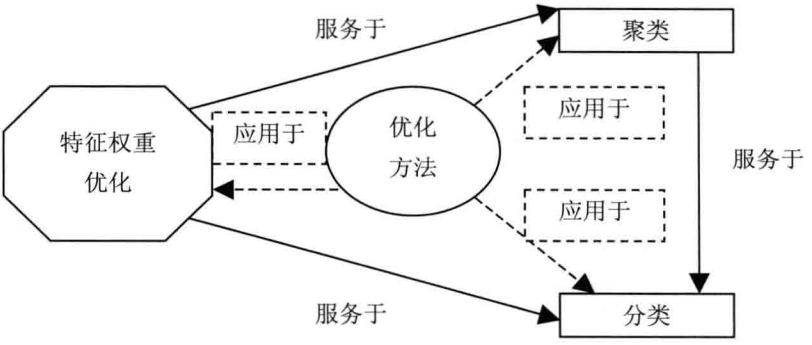

图 1-2　本书的研究主线框图

第四节　与本课题相关的国内外研究进展

最优化与数据挖掘相结合的研究最早可以追溯到聚类和分类的研究。从最优化的角度看，FCM 聚类算法和 SOFM 网络（自组织特征映射网络）是类似的，甚至 BP 算法也可看成是一种类似的优化问题。SVM（支撑向量机）是数学规划方法在分类上的一个典型应用。Kleinberg 等[2]提出一种数据挖掘的微观经济学理论，把数据挖掘看成一种优化问题。

在国外，将数学规划方法应用到数据挖掘的研究有：P.Narendra 等[3]于 1977 年采用分支定界法来解决数据挖掘中的特征选择问题；P.S.Bradley 等[4]于 1998 年总结了当前的应用对数据挖掘提出的一些挑战，也列举了为了将数学规划方法更有效地应用到数据挖掘中需要做出哪些新的突破；F.J.Iannatilli 等[5]于 2003 年采用混合整数线性规划方法来解决特征选择；R.Shioda 等[6]于 2003 年采用整数优化模型来解决分类和回归问题；Ioannis P. 等[7]于 2004 年提及了可以将数学规划的最优化方法应用到分类、聚类、SVM、多类 SVM 中去。

在国外，将最优化方法应用到聚类分析领域的研究有：P.Hansen 的两篇论文[8,9]提出了一种可变近邻元启发式搜索(Variable Neighborhood Search metaheuristic)框架；P.Hansen 等[10]提出一种新的局部启发式搜索算法 J-Means，这是一种采用中心点到数据点的分配策略，说是可以更有效地找到更小的目标函数值；N.Belacel 等[11]将一种局部启发式搜索方法 Fuzzy J-Means 嵌入到可变近邻元启发式搜索框架中，称其实验效果比 FCM 聚类算法要好。

在国内，邓乃扬于 2003 年出版的专著《最优化与数据挖掘中的核方法》总结了数据挖掘中广泛应用到的最优化方法及相关技术；同一年，袁玉波[12]研究了最优化技术在数据挖掘中的应用，通过分析几类数据挖掘问题，建立相应问题的优化模型，研究求解这些问题的优化方法，从多个角度进行算法

研究和数值实验，取得了若干的创新成果；丁世飞[13]将信息熵应用到模式识别领域，也充分利用了优化方法。

目前，数学规划已经成功地应用于许多数据挖掘问题。例如，数学规划的思想在解决特征提取、聚类、分类等问题已经显示了它们的影响力。从数值角度上说，数学规划方法，尤其是线性和二次规划方法是可靠的、有效的。但是一些适应性很强的数学规划算法对于特定的数据挖掘问题却表现出很多缺陷，例如时间和空间复杂度高。所以对于特定问题，需要使用更具有针对性的算法来提高速度和准确率[14]。

第五节 本书相关的技术与方法

本书涉及的基础知识主要有数据挖掘和优化方法等，下面对一些相关的技术与方法作简要的介绍。

一、数据挖掘

Fayyad et al.[15]于 1996 年给出一个公认的知识发现过程定义：数据挖掘是指从数据库中发现潜在有用的、新颖的、可理解模式的高级处理过程，它是利用确定的算法从准备好的数据中挖掘或提取有用知识的过程。

在 Fayyad[15]给出的知识发现过程模型中，为了能得到更贴近用户需要的挖掘结果，可以在知识发现过程中的全局过程以及具体每一步都根据用户的要求作优化或次优化处理。这里一般把优化目标选择为用户的要求，所以需要对用户的要求形式化为能被计算机接受的优化目标函数。事实上，对总体过程的优化以及具体每一步的优化可以选择具体不同的优化目标函数，但其目的都是为了满足用户的要求。

数据挖掘将传统的数据分析方法与处理大量数据的复杂算法相结合[16]。它涉及最优化方法、统计学习、机器学习、模式识别以及人工智能等方面的理论。在数据挖掘中，应用较广泛的技术来自机器学习和统计学习领域。最优化方法经常融入数据挖掘中的许多算法中，这是因为许多数据挖掘问题最终都可以转化为一个优化问题来进行求解，如 Web 挖掘中的优化搜索、对具有复杂结构数据集的聚类分析、目前流行的 SVM 方法等。

Chen 等[17]从数据库的角度来解释数据挖掘。Ramakrishnan 和 Grama[18]给出了数据挖掘的一些讨论，提出了若干观点。Lambert[19]考察统计学在挖掘大型数据集上的应用，并对数据挖掘与统计学各自的角色提出了一些看法。Smyth[20]介绍数据挖掘研究所面临的一些挑战，有些难题至今仍需要研究人员去努力加以解决。Wu 等[21]讨论了如何将数据挖掘的研究成果转化为工具，为实际生活创造价值。

二、最优化理论和方法

Lagrange 极值法是近代著名的一种经典优化方法，但最优化方法成为一门学科（通常又称为运筹学），则是在第二次世界大战之后。第二次世界大战以后，线性规划方法得到蓬勃发展，特别是 G.B.Dantzig 于 1947 年提出用单纯形法来解决线性规划问题，这是优化理论上的一个里程碑突破，接着又对这种方法进行了发展和推广[22,23,24]。Operations Research 50th Anniversary[25]中列出了最优化理论上的一些重大的突破，从中可以看出优化理论的发展轨迹。

最优化是确定某些数学上能定义的客观现实问题（数学模型）的最优解的科学。具体的研究内容包括各种问题的最优性条件的研究，数值求解方法的确定，优化求解方法结构的研究，以及在试验性条件下对实际问题的计算机仿真实现。最优化方法几乎应用到所有具有数值信息的活动，如科学、工程、数学、经济、商业、贸易等[26]。

可以说，最优化理论和方法是从实际问题中归纳产生出方法和理论的，为了体现这门学科的价值，它又必须应用到实际问题的解决当中去，这才是这门学科的研究方法和研究动力。近年来，随着海量数据挖掘的兴起，为了更好地处理海量数据，需要充分应用最优化理论和方法。同时，在新的应用推动下，又对最优化理论提出了新的研究问题。

三、特征选择和特征加权

对于高维数据，许多分类和聚类算法（以及其他数据分析算法）都有维数灾难的麻烦，经常出现分类准确率下降，聚类质量降低，这时就迫切需要减少维数。降低维度的一种经典方法是主充分分析（Principal Component Analysis, PCA）[27]，这是一种对原属性进行线性组合，希望得到数据最大变差的一种正交方法。

降低维度的另一种方法就是特征子集选择方法，这是一个搜索所有可能的特征子集的过程。有三种标准的特征选择方法：嵌入、过滤和包装。Molina 等[28]的综述提供了该主题的广泛材料。

特征加权是一种保留或删除特征的可供选择的方法。特征越重要，所赋予的权值就越大，而不太重要的特征就赋予较小的权值。在 SVM 分类法中，可以产生出将每个特征赋予一个权值的分类模型，这样保证具有较大权值的特征在模型中所起的作用更加重要。如果最近邻分类器能选择合适的距离度量，则可以更好地判断当前待预测数据点应该归属到哪个最近邻子聚类，从而可以提高预测的准确度。

四、聚类分析

1. 聚类分析介绍

在机器学习中，聚类又称为无监督学习，指的是试图发现无标号数据集中内在的分布结构。如果是划分为有意义的组，则划分后的簇应当能获取数据的自然结构。从直觉上来说，同一个聚类里的数据点应该比不同聚类的数据点更相似（接近）一点。聚类在许多领域中都得到应用，如：心理学和其他社会科学、生物学、统计学、模式识别、信息检索、机器学习和数据挖掘[16,99]。

在多元统计分析中，聚类被认为是一种分析实值数据集分布的方法。随着数据挖掘的兴起，聚类方法开始被应用到实际问题中经常出现的具有混合属性（有序属性和无序类别属性）数据集的分析中来。聚类方法大致可分为模糊聚类方法、神经网络聚类方法、基于优化目标函数的聚类方法、层次聚类方法等。

多元统计分析学科中的聚类分析是作为一种连续型数据集的数据分析方法来进行介绍的，这可看成是聚类分析发展的早期阶段。Duda 等[29]的模式识别、Mitchell[30]的机器学习和 Hastie 等[31]的统计学专著都是从统计学的角度来对聚类方法进行介绍的。

由于聚类分析属于一个交叉研究领域，它融合了多个学科的方法和技术，所以可以从多角度、多层次来分析现有的聚类分析算法。Qlan Wei-ning 等[32]从多个角度分析了现有的许多聚类算法。Johannes Grabmeier 等[33]从数据挖掘的角度（如相似度的定义、相关的优化标准等）分析了许多聚类算法。Arabie 和 Hubert 的论文[34]是一个关于聚类方面的很好的参考文献。Jain 等[35,36]对聚类分析领域作了一个较好的综述。

因为商业中经常遇到的是事务数据，目前事务数据聚类是一个比较重要的研究领域，Ganti 等[37]，Gibson 等[38]，Han 等[39]，Peters 和 Zaki[40]的研究是这方面的代表性成果。随着通信产业的兴起，流数据也越来越普遍，越来越重要，因此流数据聚类也是一个重要的研究方向，Barbara[41]和 Guha 等[42]做了这方面的引导性工作。概念聚类是一个还需进一步研究的课题，虽然这个领域已有一些研究成果，如 Fisher 和 Langley[43]，Jonyer 等[44]，Mishra 等[45]，Michalski 和 Stepp[46]以及 Stepp 和 Michalski[47]。

2. 聚类算法的分类

聚类算法一般可以分为五类：

（1）划分方法

这是从最优化的角度来研究聚类分析的，最经典的有 k-均值算法[48]，k-

中心点算法 PAM 和 CLARA[49]，CLARANS[50,100]，k-模（聚类分类数据）和 k-原型（聚类混合数据）算法[51]，EM(Expectation Maximization，最大期望)算法[52]以及一些改进算法。

（2）层次方法

如果需要在不同层次上获取数据的内部结构，层次聚类方法是个很好的选择。Kaufman 等[53]提出一种凝聚的层次聚类和分裂的层次聚类方法。Zhang 等提出的 BIRCH[54]首先采用 CF 树来进行层次聚类以达到改进的目的。Guha 等提出的 CURE[55]和 ROCK[56]，以及 Karypis 等提出的 Chameleon[57]都对原始的层次聚类方法进行了一定的改进。

（3）基于密度的方法

OPTICS(Ordering Points to Identify the Clustering Structure)[58] 和 DBSCAN(Density- Based Spatial Clustering of Applications with Noise)[59]是这种类型的比较著名的聚类算法。

（4）基于网格的方法

STING(Statistical Information Grid)[60]是基于网格的一种多分辨率的聚类方法，它将空间量化为网格结构，后续的所有聚类操作都是在网格上进行的。CLIQUE(Clustering In QUEst)[61]和 WaveCluster[62]是混合基于网格和基于密度的聚类方法。

（5）基于模型的方法

COBWEB[63]，CLASSIT[64]和 AutoClass[65]是这种类型的比较著名的聚类算法，自组织特征映射网络[66]也可归属到这一类中。

五、分类

分类是一种根据输入数据集建立分类模型的系统方法，它是一个热门的研究课题[29,67]。主要的分类方法有决策树分类法、基于规则的分类法、神经网络分类法、支撑向量机分类法和朴素贝叶斯分类法等。分类是神经网络、统计学习和机器学习领域的一个主要研究课题[16]。分类技术一般都使用一种学习算法来建立一个能很好地拟合输入数据中属性集和类标号之间的联系的分类模型，这就是分类模型与训练数据集的贴近度要求；同时还需要该分类模型对待预测的未知样本具有较高的准确率，这实际上就是模型的贴近度与泛化能力之间的一种合适折中。

在经典统计学中，一种典型的分类方法是 Fisher 线性判别分析[68]，它是一种寻求产生不同类对象之间最大区分能力的数据的线性投影方法。

Mitchell[30]从机器学习的角度介绍了许多分类技术。Duda 等[29]，Webb[69]，Fukunaga[70]，Bishop[71]，Hastie 等[72]，Cherkassky 和 Mulier[73]，Witten 和 Frank[74]，Hand 等[75]，Han 和 Kamber[76]以及 Dunham[77]对分类技术进行了广