智能聚类分析方法及其应用

李 川　姚行艳　蔡乐才／著

科学出版社

智能聚类分析方法及其应用

李　川　姚行艳　蔡乐才　著

科学出版社

北　京

内 容 简 介

本书主要论述了智能聚类分析的相关理论、方法和典型应用。内容由浅入深，涵盖智能聚类分析的基本概念、基本理论和主要聚类算法，并从基于信息熵粗糙集理论、信息熵自适应并行免疫遗传算法、向量空间模型、有偏观测模糊 C 均值等视角系统阐述了智能聚类分析方法及其典型应用。

本书适合从事智能信息处理、文本分析、故障诊断等领域及相关领域的研究人员参阅，也适合相关学科的高校师生阅读。

图书在版编目(CIP)数据

智能聚类分析方法及其应用 / 李川，姚行艳，蔡乐才著. —北京：科学出版社，2016. 10

ISBN 978-7-03-050226-1

Ⅰ. ①智… Ⅱ. ①李… ②姚… ③蔡… Ⅲ. ①聚类分析-分析方法 Ⅳ. ①O212. 4

中国版本图书馆 CIP 数据核字（2016）第 240910 号

责任编辑：郭勇斌 邓新平 / 责任校对：郑金红

责任印制：张 伟 / 封面设计：众轩企划

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www. sciencep. com

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 10 月第 一 版 开本：720×1000 1/16

2016 年 10 月第一次印刷 印张：9 1/4

字数：124 000

定价：58. 00 元

（如有印装质量问题，我社负责调换）

前　言

通俗地讲，聚类的基本思想就是“物以类聚”，具体来说就是对事物的某些性质根据规则进行归类，把具有相似性质的作为一类，使同一类的数据相似性大，而不同类的数据没有相似点或相似点很少。聚类分析是一种应用数学方法对事物进行分类的研究。面对庞大、高维的数据量，如何对这些数据进行有效的选择，最终达到降维的目的是一个值得研究的问题。随着现代高科技的发展，信息量的高速膨胀，人们开始关注如何从这些数据中获取有效信息，从而做出相应的决策。但是，这些海量数据往往带来数据分析的效率瓶颈，同时，获取数据的多源化导致数据的非结构化问题，使得主要依靠专业知识和经验的经典分类方法已经远远不能满足实际需求。因此，如何从海量数据中提取有效信息在信息处理领域十分重要，如何处理庞大的数据量成为聚类分析研究的一个重要方向。

近几十年来，人工智能技术在国内外受到高度重视，发展迅猛，在自然语言处理、图像识别、语言识别、故障诊断等方面得到广泛应用。遗传算法是在模拟达尔文进化论的基础上发展起来的一种人工智能算法，它是以自然选择和遗传理论为基础的一种启发式局部最优搜索算法。作为最常用的聚类方法的 K 均值聚类算法，对初始聚类中心点敏感，容易陷入局部极小值。免疫遗传算法是一种将生物体内免疫系统的原理应用于基本遗传算法的新型智能算法，在许多地方比基本遗传算法要优越，既保持了基本遗传算法的全局搜索能力，又能克服其局部最优的缺陷，特别是免疫算法的自适应能力及记忆保留策略，使得免疫遗传算法能够更智能地解决一些问题。粗糙集是一种刻

画不完整性和不确定性的数学工具，能有效分析和处理不精确、不一致和不完整等各种不完备的信息，并从中发现隐含的知识，揭示潜在的规律。将这些智能算法与聚类分析相结合产生的智能聚类分析方法，可以有效提升文本分析、故障诊断等领域的聚类分析性能。

本书汇集了作者团队长期以来围绕智能算法和聚类分析等方面的研究成果，介绍了智能聚类分析的方法及其应用。本书主要以文本聚类分析、故障聚类分析等应用案例为背景，对智能聚类分析的相关理论方法进行了相对系统化和完整性的阐述，内容直观，由浅入深且易于理解，不仅包含了智能聚类分析的基础概念、理论和方法，还详细描述了这一领域的最新进展和成果。本书为从事智能信息处理、文本分析、故障诊断等相关研究的科技人员介绍了较为专业化和高层次的有关新理论、新方法和新技术，也为广大教师和学生提供了较为全面而深入的学习素材，对拓展智能聚类分析的应用有着现实的指导意义。

本书共分为6章。第1章为绪论，给出了聚类分析的基本概念，概要介绍了应用到聚类分析过程的遗传算法、免疫遗传算法、粗糙集理论等智能技术的现状和研究进展。第2章在概述数据挖掘基础知识之上，介绍了聚类分析的定义及聚类分析过程，总结、比较了目前的主要聚类方法、主要聚类结果评估方法，以及聚类分析方法面临的问题。第3章介绍了一种基于信息熵的粗糙集理论用于聚类分析数据预处理的算法。第4章在改进的免疫遗传算法的基础上，介绍了基于信息熵的自适应并行免疫遗传算法的K均值聚类算法。第5章针对免疫遗传算法的相似性，介绍了一种基于向量空间模型的路径相似度聚类分析及其对文本的聚类分析。第6章介绍了有偏观测模糊C均值智能聚类分析算法，并将其应用到轴承故障诊断的故障识别与聚类分析。

本书由李川、姚行艳和蔡乐才等共同撰写完成，其中第1章由蔡

乐才撰写，第2～5章由姚行艳撰写，第6章由李川撰写。作者团队的研究生李勇、成志伟、李雪娇等参与了本书相关内容整理和校对，在此向他们一并表示感谢。

由于作者水平有限，书中难免存在一些不妥之处，敬请广大读者批评指正。

作　者

2016年4月

目　　录

第1章 绪 论

1.1 引言

随着互联网的迅速普及，企业信息量的急速膨胀，如何从众多纷繁的数据中按照某种规则获得一些有用的数据，在一定程度上对于企业的存活起着至关重要的作用。数据挖掘（Data Mining，DM）就是从大量的数据库、数据仓库或其他信息储存库中获取新颖的、有效的、潜在有用的、最终可理解模式的过程。

由于各种信息资源呈指数形式增长，面对如此庞大的数据量，人们的需求已经不是简单的数据查询统计，而是需要从大量信息中挖掘可以得到决策的模式、规则或规律等。因此，如何从中得到自己需要的信息显得尤为重要，由此，数据挖掘技术应运而生。数据挖掘一般是指从大量数据中通过相关算法得到隐藏的信息的过程[1]。

数据挖掘这一概念最早由美国计算机协会（Association for Computing Machinery，ACM）于1995年提出。在提出数据挖掘概念之前，国际联合人工智能学术会议上提出了数据库知识发现这一概念。知识发现的过程一般包括3个步骤，即数据准备、数据挖掘及对结果的评价解释。其中，数据准备包括数据选择、数据预处理和数据转换3个步骤；数据挖掘是知识发现的核心，在得到良好的挖掘效果之前，需要事先对各种数据挖掘技术进行全面了解[2]。

聚类分析是数据挖掘[3,4]的一个重要研究内容，它涉及诸如数据挖掘、统计学、经济学、机器学习及生物工程等研究领域[5]。“所谓聚类分析就是根据各样本自身的不同，将数据集划分为不同的簇，使

数据源之间用相似性来衡量，即一些基本相似的个体尽可能划分在同一簇中，而一些相差较大的个体划分在不同簇，从而整个数据集就可以用少数的几个簇来描述（当然，尽管数据集中的一些细节信息可能会丢失，但它却将数据集进行了概化，节省了数据集的内存）。”[6]正因为聚类分析具有如此强大的功能，通过聚类分析，人们可以或可能会发现数据集中所蕴涵的某种信息或知识，并为人们所用。从孩提时代开始，人类就从未停止过进行聚类分析。通过对所见、所闻的一切事物经过某种下意识的分析后，随着知识的积累和不断发现，不断改进聚类模式而对事物进行某种聚类，从而达到分类的目的。目前，聚类分析已广泛应用于商业、生物、地理、保险业、电子商务及互联网等很多方面。常见的聚类分析方法有：K 均值聚类算法、模糊 C 均值智能聚类分析算法、最大似然估计算法和基于图论的算法。

K 均值聚类算法是基于规则的聚类算法中的一种简单常用算法。首先，该算法选择一个特定距离度量作为模式间的相似度，然后由所选择的聚类准则函数来评价聚类划分结果。在给定初始聚类中心点后，采用迭代的方法找出取决于聚类准则函数的最佳聚类分区。这种算法的一个缺点就是初始聚类中心点的选择不当可能导致早收敛的问题。在 K 均值聚类算法的基础上，模糊 C 均值智能聚类分析算法有效集成了模糊技术进行聚类分析。最大似然估计算法是以概率论为基础的一种聚类算法，它根据事先所假设的某种先验概率分布计算出后验概率来实现数据分类。基于图论的算法主要是根据所估计的每个点的密度梯度值生成方向树，然后通过求出的谷点密度函数对数据进行分类[7]。

为了提高聚类分析的效果，可以将遗传算法、进化算法、粗糙集理论、模糊理论等智能技术与聚类分析结合起来，形成智能聚类分析方法。本书通过对智能聚类分析方法的介绍，将其应用到文本分析、故障诊断等典型案例中。

1.2　聚类分析的研究进展

1.2.1　聚类分析的基本方法

聚类是数据挖掘的一个重要方法，也是人类一种基本的认知活动。聚类分析是指将未知分布的一组数据，利用数据对象之间的关系，尽可能将具有相似性质的数据聚集成一类，使类间相似性尽可能小，而同类中数据的相似性尽可能大，这种方式实际上是一种无标签分类，因此，聚类也属于无监督学习方法。同时，聚类和分类之间又存在明显的区别。聚类的最终目的是找到数据的特征及潜在的数据类别的分布情况；而分类则是对已经标记好的数据集进行训练，并通过学习预先获得数据的特征以建立一个分类模型，进而利用该分类模型对数据的类别进行预测。聚类算法作为一种有效的数据分析方法，目前已在数据挖掘、语音识别、机器学习及生物信息处理等领域广泛应用。同时，聚类分析还可以将聚类算法应用于商业分析，区分消费者数据库中的不同消费人群，以帮助市场决策人员归纳总结出每一类消费者的消费习惯或者消费模式。目前聚类算法主要有以下几种：基于谱的聚类算法，基于支持向量机的聚类算法，基于密度的聚类算法，基于遗传算法的聚类算法，等等。

国外学术研究中比较著名的具有聚类分析功能的系统主要有WEKA、CLUTO 等。WEKA 是来自新西兰怀卡托大学的一款开源软件，是到目前为止功能最为完备的数据挖掘工具之一，被誉为数据挖掘学习史上的里程碑[8]。WEKA 中集成了多种数据挖掘算法，不仅包括数据的预处理，而且还包括数据的分类和回归、聚类及关联规则等可视化界面。用户还可以通过 JAVA 语言进行二次开发。

CLUTO 是由美国明尼苏达大学的 Karypis 教授团队开发的一款聚类工具，该工具不仅可以处理低维数据，还能够处理高维数据，而

且，针对不同聚类的结果可以对结果的类簇进行分析[9]。CLUTO 软件包中包括多个独立可执行的程序和库文件，它可以应用于多种领域，如信息检索、生物学及商业等。CLUTO 软件包含多种聚类算法及聚类准则函数，不仅可以辨别出各类别的特征属性，还能够根据所识别的特征属性对类别中的对象提供总结。

由于聚类分析强大的功能，其潜在的应用也对聚类算法提出了更高的要求，主要要求如下[10]。

（1）可伸缩性。一般来说，常用的聚类算法在处理较小数据集时效果较好，但面对海量数据处理对象的时候，效果则没那么好。虽然可以通过海量数据进行抽样聚类，但总体来说，这种抽样聚类的效果并不理想，往往会与实际值存在很大偏差。因此，这就要求聚类算法在处理不同特征数据集时，具有一定的可伸缩性。

（2）能够聚类任意形状的类簇。目前，常用的确定类簇的方法主要是基于欧氏距离等相似性度量方法，但这类方法只能发现具有类似大小和密度的圆形或球状的类簇。事实上，每一个类簇的形状可能是任意的，那么，如何设计有效的聚类算法来处理任意形状的类簇就显得尤为重要。

（3）多类型数据的处理能力。聚类算法需要对多种类型的数据进行聚类，而不仅限于某一类型的数据，如枚举型数据、二值型数据等。

（4）对异常数据处理的能力。真实数据中往往存在很多孤立点、缺失的数据甚至错误的数据等，而这些异常数据对聚类结果的影响较大，聚类算法如何处理这些异常数据也是需要考虑的问题。

（5）对高维数据处理的能力。大多数聚类算法能够较好地处理低维数据，而对于如文本数据等高维数据的聚类效果则并不是很理想，这也是聚类算法研究中面临的一项巨大挑战。

1.2.2　聚类分析的典型应用

随着科技的发展，聚类分析已经在各种领域得到广泛应用，如文本分析、语言识别、图像处理、故障诊断等方面。

以文本分析为例，统计表明，一个组织中约有80%的信息存储都以文本形式存在，主要有新闻报纸、学术论文和专著、历史资料存档、门户网站、论坛、博客、电子邮件和Office文档等。由于文本数据固有的特点，大多数是结构化或者半结构化的数据，并且数据又存在维度高和稀疏的特点，因此，基于传统的结构化数据挖掘技术常常不能够直接应用于文本挖掘，从而，如何从大量的文本信息中发掘出有用信息受到越来越多学者的关注，具体内容包括对文本信息的分析和组织、如何提取文档中所隐含的规则和模式等。文本挖掘需要多种技术相结合来实现，如机器学习、信息处理、信息检索及数据挖掘等。文本挖掘的主要目标在于文本的结构分析、信息提取、关联和预测分析、文本的分类和聚类等。文本挖掘这一概念于1995年由Feldman正式提出[11]，自此之后，国内外很多学者就文本挖掘的理论及应用进行了许多研究。据调查发现，文本挖掘技术已经成为数据挖掘分支中一个日益重要的领域。文本聚类的流程图如图1.1所示。

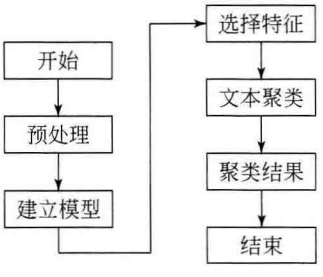

图1.1　文本聚类流程图

文本聚类技术是一种无监督的学习方法，是对文本信息进行分析、组织和分类的重要手段。如前所述，文本聚类就是在对文本信息

没有标记任何类别的情况下，自动识别出文本类别的过程。通常的聚类方法是采用明确的定量方法处理结构化数据，而文本聚类处理的是非结构化的文本信息，对此，文本聚类就需要采用一系列文本分析的处理技术，如文本分词、特征选择、降维及文本表示等。

文本聚类的应用主要在以下几个方面。

（1）自然语言的预处理。通过聚类分析技术可以加快用户在文本浏览系统中寻找有效信息的速度，为用户提供了很大方便。聚类分析技术还可以用于多文档摘要的自动生成，可以从互联网上搜集许多当天重要的文本新闻来聚类，然后对每个聚类后的文本集的主要内容聚集成简单的摘要以供用户浏览。

（2）对搜索引擎结果聚类。为方便用户及时、迅速定位到所需的有效信息，需要采用聚类分析技术对搜索引擎的结果进行聚集分类。

（3）发现并追踪热点主题。如何从每天海量的互联网信息中获得有效的热点主题并进行追踪，对于研究热点和维护社会的稳定都具有重要的意义。通过聚类分析及聚类相关算法不仅可以找出目前已经关注的主题信息，而且还能发现新热点。

（4）改善文本分类的性能。通过文本聚类技术可以从海量数据中选择出特征空间，从而使文本分类的性能得以改善。

（5）优化网站结构和挖掘用户感兴趣的模式。利用文本聚类技术可以从互联网中大型数据中聚集用户感兴趣的模式，以实现对信息的自动过滤和推荐。

国内外许多研究机构和公司对文本聚类和挖掘技术进行了研究，并取得一定的成果。例如，IBM 公司针对文本聚类技术开发了一款数据挖掘软件 Text Miner，其主要功能是实现对文本信息的特征提取、文档聚类和分类、检索。Text Miner 支持十几种语言，采用深层次的文本分析和索引实现对多种文本格式的数据检索。Bow 是一个专门用

来处理文本信息处理的开源包，由卡内基梅隆大学开发，主要包括3个部分：文本分类、文本检索及文本聚类。文本检索系统（Text Retrieval System，TRS）是国内外第一个实用的中文文本数据挖掘产品，采用智能化的信息处理方式代替传统的机械化的文本处理模式，可以自动地对文本进行归类生成主题词。FudanNLP是由复旦大学开发的开源中文自然语言处理工具包，同时还包括实现机器学习的算法和数据集。它的主要功能包括信息检索、中文处理及结构化学习等。

1.2.3　聚类分析方法面临的挑战

聚类分析是在没有任何预知信息的前提下，对所要划分的分类是未知的，并且由于聚类数据自身的特殊性而加大了聚类的难度。这种特殊性主要表现在两个方面：一方面是由于聚类信息的高维性和稀疏性导致的非结构性；另一方面则体现在因聚类对象自身的多样性导致的聚类困难。因此，研究聚类分析是一项极具挑战性的课题。常见的聚类分析应用困难主要包括以下几个方面。

（1）聚类对象的高维性。例如，一篇文本信息经过处理后最少包括几百个甚至成千上万个属性。传统文本数据中包括的属性比较少，而随着信息技术的发展，现代聚类分析过程与传统的文本信息处理大不相同。无论是文本信息维度的增加，还是聚类算法的可伸缩性、处理任意形状及处理孤立点的能力等方面，都面临巨大的挑战。传统的处理低维数据的聚类算法，在面对高维数据集时，其聚类效果大大下降。

（2）聚类对象的稀疏性。经过预处理的聚类文本信息大多数采用词袋（Bag of Words）向量表示。词袋向量表示都是以特征词的形式从该集合中选择出来，并且往往属于某一特定对象的特征词，可能仅占有整个集合的一小部分，而文本空间上的文本特征向量在大部分