面向金融大数据的若干聚类方法改进与应用研究

王丽敏　韩旭明　著

科学出版社

面向金融大数据的若干聚类方法改进与应用研究

王丽敏　韩旭明　著

科学出版社
北京

内 容 简 介

大数据是一股颠覆性力量，使各行业机遇与挑战并存。大数据时代的来临，使大数据分析成为各行业竞争发展的变革点。麦肯锡全球研究所的研究显示，数据对于企业的重要性正变得与劳动力和资本并驾齐驱。聚类是数据分析的重要手段之一，面对海量数据，提取有价值的信息具有重要意义。本书是作者几年来科研成果的总结，全书共分 6 章，重点是针对吸引子传播聚类等算法进行若干理论改进与应用研究，并将其用于金融领域中，取得了令人满意的结果。

本书可供从事大数据研究的科研人员参考，也可作为高等院校相关专业高年级本科生和研究生的教材。

图书在版编目(CIP)数据

面向金融大数据的若干聚类方法改进与应用研究/王丽敏，韩旭明著. —北京：科学出版社，2017.11

ISBN 978-7-03-052453-9

Ⅰ. ①面… Ⅱ. ①王… ②韩… Ⅲ. ①金融-数据处理-研究 Ⅳ. ①F830.41

中国版本图书馆 CIP 数据核字(2017)第 055699 号

责任编辑：王 哲 霍明亮 / 责任校对：郭瑞芝

责任印制：张 伟 / 封面设计：迷底书装

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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 11 月第 一 版 开本：787×1092 1/16

2017 年 11 月第一次印刷 印张：8 1/2 插页：5

字数：190 000

定价：76.00 元

(如有印装质量问题，我社负责调换)

前　言

大数据时代背景下，数据呈现爆炸式增长。各行各业逐步采用计算机技术管理数据，极大地提高生成、收集、存储和处理数据的能力。面对海量数据，提取有价值的信息变得十分困难，借助聚类技术可以挖掘出有助于提高决策准确度和效率的有用信息。本书是作者几年来科研成果的总结。全书共分 6 章，重点是针对吸引子传播聚类算法进行理论改进和应用研究，并将提出与改进的算法应用到相应领域，取得了令人满意的结果，具体研究内容如下。

（1）鉴于吸引子传播聚类算法的偏向参数对于聚类结果影响巨大。本书提出六种优化参数的算法，分别是基于果蝇优化的吸引子传播（affinity propagation based on fruit fly optimization，FOA-AP）聚类算法、基于果蝇优化的自适应吸引子传播（self-adaptive affinity propagation based on fruit fly optimization，FOA-SAP）聚类算法、基于烟花爆炸优化的半监督吸引子传播（semi-supervised affinity propagation based on fireworks explosion optimization，FEO-SAP）聚类算法、基于布谷鸟优化的半监督吸引子传播（semi-supervised affinity propagation based on cuckoo search，CS-SAP）聚类算法、基于稳定阈值的吸引子传播（affinity propagation based on stability threshold，STAP）聚类算法和基于约束规则的吸引子传播（constraint rules-based affinity propagation，CRAP）聚类算法。与传统的吸引子传播聚类算法相比，本书提到的 FOA-AP、FOA-SAP、FEO-SAP、CS-SAP、STAP 和 CRAP 等聚类算法可以自适应地搜索两个参数空间，快速、准确地定位最优参数位置，强化算法的局部寻优能力，提升算法的全局探索能力，从而获得最佳聚类结果，提高算法的聚类性能。另外，将本书提出的 STAP 聚类算法应用到房地产上市公司财务评价领域，利用 STAP 聚类指数的新概念评价上市公司绩效，为股票投资和上市公司的发展提供一种有效的参考工具，仿真模拟验证表明，具有较好的应用前景。

（2）特征空间维数越高，聚类分析的复杂性就越高。因此，如何进行特征选取，设计有效特征选取的新方法变得尤为重要。相似性度量的定义会直接影响吸引子传播聚类算法的聚类效果。传统的吸引子传播聚类算法中，以欧氏距离作为相似度量方式的算法并没有考虑数据集的空间特征结构，这样会导致聚类效果不佳。依据数据集的空间特征结构，构造合适的相似度矩阵，研究高效、可扩展、适合复杂结构数据集的吸引子传播聚类算法是一个关键问题。鉴于此，本书提出若干优化相似度矩阵的吸引子传播聚类算法，分别是基于变异赋权的吸引子传播（coefficient of variation affinity propagation，CVAP）聚类算法、基于智能赋权的吸引子传播（intelligent weighting based on affinity propagation，IWAP）聚类算法、基于距离贴近度的吸引子传播（close measures affinity propagation，CM-AP）聚类算法、半监督自适应权重吸引子传播（semi-supervised affinity propagation clustering algorithm based on adaptive feature weighted，AFW-SAP）聚类算法、基于结构相似度的半监督自适应吸引子传播（semi-supervised adaptive affinity propagation based on structural similarity，SAAP-SS）聚

类算法和基于属性分布相似度的吸引子传播（properties distribution similarity-based affinity propagation，PDS-AP）聚类算法，通过优化相似度矩阵，可以有效地消除量纲影响，明显提高聚类效果，同时拓宽算法处理多种数据的能力，具有很好的鲁棒性。

（3）高维数据进行聚类需解决两个重要问题：一是如何消除数据的冗余性，二是如何在稀疏数据点获取数据集的分布。此外，阻尼因子作为决定吸引子传播聚类算法收敛速度的重要参数，在算法运行的全过程中，一经确定，便不再改变。这无疑削弱了算法的收敛性能，使得阻尼因子不能在算法运行的不同阶段动态调整参数以达到最佳聚类性能。为解决上述问题，本书提出三种优化算法，分别是基于熵权法和主成分分析法相结合的吸引子传播（entropy weight method and principal component analysis affinity propagation，EWPCA-AP）聚类算法、基于奇异值分解的自适应吸引子传播（self-adapting affinity propagation clustering algorithm based on singular value decomposition，SVD-SAP）聚类算法和基于最小簇匹配的流形吸引子传播（affinity propagation based on matching min-cluster hierarchical clustering，MMHC）聚类算法。EWPCA-AP 聚类算法利用熵权法的思想对获取样本数据进行加权，消除数据的冗余性，通过主成分分析对数据进行降维；SVD-SAP 聚类算法通过对高维数据进行奇异值分解操作，消除冗余信息，通过奇异值分解逆运算重构数据，进行降维，降低数据规模，提高算法效率；为使算法的收敛更快，本书提出一种非线性函数策略，根据每次能量函数的收敛情况自适应地调整阻尼系数，提高算法的收敛性能；基于最小簇的匹配的流形吸引子传播聚类算法通过建立无向图，构建一种新的基于图的流形聚类，充分利用传统吸引子传播聚类算法优秀的球形和凸集数据聚类能力，不断聚合最小簇，从而完成流形聚类。另外，将 EWPCA-AP 聚类算法用于我国经济领域评价，得到令人满意的结果。因此，EWPCA-AP 聚类算法具有很好的实用性，为各级政府经济决策提供一种新的参考工具，为各省市经济发展提供有效的参考依据。

本书是在国家自然科学基金项目（项目编号：61202306，61472049，61572225，61402193）、国家社会科学基金项目（项目编号：15BGL090）的资助和支持下完成的。本书是作者与金融大数据深度挖掘研究团队几年来的科研成果总结，值此专著完成之际，诚挚地感谢吉林省互联网金融重点实验室全体教师的鼎力支持与帮助。参与本书撰写的还有国薪桐、张利、孙海波、姬强、王依章等；王念博、郑凯月、李明洋、刘美含、郝志远等研究生为本书的完成付出了辛勤劳动，在此一并表示感谢。

由于作者水平有限，加之金融大数据研究领域纵深宽广，书中难免有不足之处，敬请广大读者批评指正。

王丽敏

2017 年 6 月于长春

目　录

第1章 绪 论

1.1 研究背景与意义

得益于云计算的推动，大数据乘“云”而上成为时代的特征，中国各行业已步入大数据时代。据统计 Facebook 上每天都有超过 1000 万张的照片更新和 30 亿次的点击率；谷歌公司每天要处理超过 24PT 的数据，这个数据是美国国家图书馆所有纸质出版物所包含数据量的上千倍；同样，Twitter 上几乎每天有超过 4 亿条推文出现。早在 2008 年，*Nature* 就推出 *Big Data* 专刊，而后 2011 年又推出 *Dealing with Data* 专刊。与此同时，世界各国也纷纷围绕大数据问题展开深入研究。2012 年 3 月，美国奥巴马政府斥资 2 亿美元专项启动“Big Data Research and Development Plan”（大数据研究与发展规划），计划在科学研究、环境、生物、金融、医学等领域利用大数据技术实现新的突破。麦肯锡全球研究所调查数据显示，数据对于企业正变得日趋重要，与劳动力和资本的变化并驾齐驱。大数据是一股颠覆性力量，对于国民经济各个部门既是机遇又是挑战。面对日益增多的大数据，数据挖掘技术的更新应用也日益受到人们的关注。模拟人工智能，数字、信号进行智能化处理，从具有随机特性的数据中提取和挖掘潜在的、有价值的信息，探索各行业在纷繁复杂、瞬息万变的表象中所隐含的内在规律，是数据分析、管理与决策的核心工作。面对每时每刻都在不断激增的海量数据，人类真正能从中获取有价值的信息却寥寥无几，我们逐渐发现，海量信息在给人们带来便利的同时，也带来前所未有的困难与挑战。鉴于此，如何对海量的原始数据进行有效的分析与利用，挖掘出有价值的信息，成为众多研究者共同关注的热点与课题。

数据挖掘的本质就是从大规模的原始数据中，通过一系列科学的分析与处理，发现其中潜在的规律和有价值的信息。聚类是数据挖掘领域中的重要研究方向。针对一组未知数据群，聚类分析是将数据看成对应的点，根据点之间的关系将数据进行分类，最终实现类内相似、类间疏远。聚类是在无监督状态下寻找最优划分的过程，聚类结果将影响人们对数据的利用。因此，聚类结果的精确性尤为重要。而有监督学习是指对新来的数据对象进行划分的一个过程，也就是事先知道各个类的属性，根据每一类确定的属性来划分新的数据对象。目前，聚类分析技术已经成功应用于各个领域，例如，在生物信息方面，经常利用聚类技术对生物的基因进行聚类，对蛋白质结构进行预测，对植物动物进行分类等；在图像处理方面，聚类可以直接用于初步图像分割，提取视频中的关键帧；在模式识别中，聚类分析技术被应用于手写体识别、信号识别、文本识别等方面；另外，聚类技术能用于文本信息的提取选择，并对大量文档进行分类；聚类算法也可以作为分类算法或者其他算法的预处理步骤，经聚类得到未标记样本的类别信息，然后利

用这些标记样本作为训练样本创建分类器等；除此之外，聚类分析还经常被应用在选址布局、信息安全、计算机视觉、模糊控制、信息检索、机器学习、数据挖掘、基因选择、最优航空路径选择以及其他社会科学等领域。历经几十年的变化，聚类算法由当初经典的 K-means 算法、FCM 算法、K-medoids 算法、谱聚类等，到现在新算法的不断涌现。本书选择一种比较新的聚类算法——吸引子传播聚类作为研究对象。该算法是 2007 年 Frey 和 Dueck 在 *Science* 上提出的一种基于信念传播的聚类算法。与其他聚类算法相比，具有明显优势。

吸引子传播（affinity propagation，AP）聚类算法将所有数据点视为网络图的节点，通过节点之间的信息传递和竞争，建立数据点之间的依存关系，并且通过不断地竞争，得到最终的聚类中心。该算法与常用的基于划分的聚类算法不同，它事先不指定初始聚类中心，而是将所有的样本点作为潜在的类中心，从而有效地避免初始点选择不佳导致聚类结果陷入局部最优的情况[1]。AP 聚类算法的输入值是样本点之间的相似度矩阵而非原始数据样本，使得其具有较高的运行效率和可伸缩性，能够胜任大规模的数据处理。目前，AP 聚类算法已经成功应用于基因识别[2]、最优航空路线确定[3]、人脸图像聚类[4]等诸多领域，但该算法还处在刚刚发展的阶段。尽管 AP 聚类算法具有很多其他算法无法比拟的优势，并在实际应用中有着良好而稳健的效果，但是它依然要面临一些难题和挑战。鉴于此，本书在保留传统 AP 聚类算法优势的前提下，对其不足之处进行合理有效地改善，提高算法的聚类性能，从而使其更广泛地应用于实际工作中，为政府和企业提供更为有效的决策依据，具有非常重要的意义。

1.2 国内外研究现状

AP 聚类算法建立一个数据点网络，每个数据点是网络中的一个节点，不断沿着网络递归传输真实信息来获取一个较好的聚类划分。信息的迭代依据是最小化能量函数，这些信息反映数据点的归属关系。

AP 聚类算法高效快速实用，但是仍旧面临着一些缺陷。

（1）无法获取真实类数。因为该算法内含一个重要参数，即偏向参数（Preference），对聚类结果影响巨大[1]。图 1.1 是 Frey 等在其论文中给出的聚类结果和偏向参数的变化关系图，大概遵循着偏向参数的绝对值越大，类数越少的规律。但是无法获取哪个偏向参数值对应的类数才是真实类数，需要我们采用一些方法搜索到这个合适的偏向参数。王开军等自适应扫描偏向参数空间来搜索聚类数空间以寻找最优聚类结果，在 11 个数据集上都获得真实类数[2]。储岳中等将 AP 聚类算法与核匹配追踪相结合，能够识别 2 个标准 UCI 数据集和遥感图像数据，提高了聚类结果准确率[3]。肖宇等利用半监督技术与 AP 聚类算法结合，大幅度提高聚类精度[4]。董俊等利用指数函数构建了新的可变相似度，将流形数据结构重新构建，增加类间可区分性，在多个数据集上得到真实类数[5]。

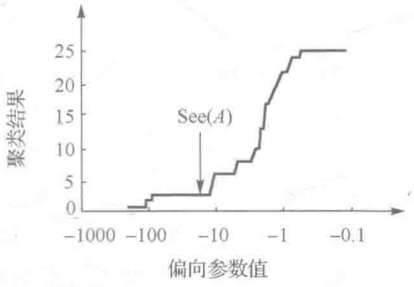

图 1.1 聚类结果和偏向参数的变化关系图（*A* 是中值）

（2）AP 聚类算法的另一个主要问题就是处理复杂数据，该算法本质上是一个基于距离的算法，因为其中两个重要信息矩阵都是基于两点间相似度的，所以该算法对于满足欧氏空间约束的数据集聚类效果更好。而面对复杂数据，例如，不同密度类簇、多流形等数据聚类效果不佳。冯晓磊针对不同密度簇提出高斯核共享最吸引子相似度，扩展 SNN（spiking neuron networks）在非欧氏空间的扩展性，使距离近的点更加靠近，离较远的点距离更远。针对流形数据提出局部保持的核流形距离，利用高斯核函数构建了新的流形相似度矩阵[6]。Wang 等[7]认为单一类代表模型不足，如场景分析和特征识别模型，提出一种多样本吸引子传播算法，不需要预先指定聚类数目，利用稀疏数据大大减少计算时间和存储，仿真实验使用手写体数字，成果显著。Guan 等[8]提出一种新的半监督文本聚类算法，即基于种子的吸引子传播算法，其主要贡献有两个：①一种新的相似性度量方法，捕捉到文本的结构信息；②一种半监督聚类过程种子的工作方法。将其应用到 Reuters-21578 数据集，比较分析 K-means 的算法、传统 AP 聚类算法和改进算法的聚类性能，结果表明，提出的相似性度量效果明显，比原 AP 聚类算法的 F 值高约 21%，提出的半监督策略迭代次数只用原算法的 76%，明显提高了聚类执行时间（是 K-means 算法的 20 倍），并与其他方法相比提高了鲁棒性。Kumar 等[9]提出了一种新的基于图形的方法，用来提取手写文本行的单色阿拉伯语文档图像。第一步利用主成分定义线和变音符号，每个主成分建立一个稀疏相似性图，然后使用最短路径算法计算非相邻组件之间的相似性。第二步将次要组件分配给每一个文本行，并使用像素匹配的标准评估该方法。他们还提出了一种人为减少文本行之间间距检测方法，证明其方法的鲁棒性。Akl 等[10]提出了一种手势识别系统，该系统采用 AP 聚类算法规整动态时间，实现几乎完美的用户依赖性识别。Dueck 等[11]认为寻找一个治疗组合是 NP-hard 问题，但各种贪婪算法可以应用，引入 AP 聚类算法用于处理组合设计问题。Givoni 等[12]扩展了 AP 聚类算法模型来解释半监督聚类，提出一个半监督算法，可以使用实例约束来指导聚类，相比其他的半监督学习方法，其可以实现更好的性能。

（3）除了理论改进之外，大量学者将 AP 聚类算法应用各个领域：管仁初等[13]提出了一种能够包含文本结构信息的非欧氏空间相似性度量算法，选用标准数据集 Reuters-21578（文本数据集）进行了验证，明显优于其他算法。Yang 等[14]构建了一个关于样本点邻域及其覆盖点数的 Preference 计算方式，使可以成为代表点的样本点代表性放大，并构建一个二元相似性函数，应用在图像分割和彩色图像聚类中，比传统算法获得的效果更好。何晏成[15]在文

本聚类中，利用 AP 聚类算法的聚类结果，采用凝聚层次聚类算法二次聚类，克服了文本数据高维稀疏的问题。鲁伟明等[16]将 MapReduce 技术引入到 AP 聚类算法中，实现了对大规模高维数据的高效聚类，采用了 MSRA 数据集、JAFFE 数据集和人工数据集验证该机制的有效性。许晓丽等[17]将 MS（mean shift）算法和 AP 聚类算法结合，用于彩色图像分割，MS 先对目标图像进行初步分割，计算分割后的区域所有像素的彩色向量平均值作为 AP 聚类算法的输入数据，运行时间和分割效果都有显著性的提高。Hassanabadi 等[18]将 AP 聚类算法用于车载 Ad Hoc 网络。由于网络是高度动态的，有恶劣的信道条件，所以一个合适的聚类算法必须对信道错误具有鲁棒性并考虑节点移动性。他提出了一种新的流动性车辆自组织网络集群计划，形成集群过程中使用 AP 聚类算法。该算法考虑节点在集群形成中的移动性，产生高稳定性的集群。仿真结果证实其相比其他流动性为基础的聚类算法性能更加优越。Givoni 等[19]为了将 AP 聚类算法应用在真正 HIV（human immunodeficiency virus）序列中，提出层次吸引子传播算法，设计了一个高阶图形化模型。在人工数据集仿真模拟实验中，模仿艾滋病毒的应变突变动力学，该算法优于相关的算法。在进行神经元分类研究中，Santana 等[20]发现定义一个神经元细胞类，需要一个高效的无监督聚类算法聚类分析神经元细胞类的形态、生理或分子特征，选择 AP 聚类算法完成这部分工作，使用 337 类四亚型测试数据集，结果证明该算法能够正确分类神经元。Zhang 等[21]把 AP 聚类算法引入到数据流聚类中，解决变化检测问题，在标准数据集和实际应用中验证提出的算法。Shang 等[22]提出了一种新的快速吸引子传播聚类方法，同时考虑局部和全局数据结构中包含信息，是一种优质的多级图划分方法，内含一种新的快速采样算法，即一种密度加权谱聚类方法。所有数据点聚类任务可以通过其相应的有代表性的范例实现。实验结果表明，该算法优于 AP 聚类算法和谱聚类算法。Sun 等[23]为了识别弱图案和减少局部最优，提出了一种新的算法 AP-Motif，应用 AP 聚类算法在 DNA 序列产生良好的候选图案，然后使用 EM 算法从候选图案中获得最佳图案。仿真模拟实验表明，AP-Motif 优于其他四种广泛使用的算法[24]。

1.3 本书主要研究内容

本书针对智能聚类算法进行系统梳理，重点是在 AP 聚类算法基础上，进行理论改进研究和应用研究，并将提出和改进后的算法应用到不同领域，其主要研究内容及创新之处概括如下。

（1）鉴于 AP 聚类算法的偏向参数对于聚类结果影响巨大，本书提出五种优化参数的算法，并与传统的 AP 聚类算法相比发现，本书提出的五种优化参数的算法可以自适应地搜索两个参数空间，快速、准确地定位最优参数位置，强化算法的局部寻优能力，提升算法的全局探索能力，从而获得最佳聚类结果，提高算法的聚类性能。

（2）传统的 AP 聚类算法中，以欧氏距离作为相似度量方式的算法并没有考虑数据集的空间特征结构，这样会导致聚类效果不佳，设计有效特征选取的新方法变得尤为重要。依据数据集的空间特征结构，构造合适的相似度矩阵，研究高效、可扩展、适合复杂结构的数据集的吸引子传播聚类算法是一个关键问题。鉴于此，本书提出几种优化相似度矩阵的 AP 聚类算法，通过优化相似度矩阵，可以有效地消除量纲影响，聚类效果明显提高，同时，拓宽算法处理多种数据的能力，具有很好的鲁棒性。

（3）针对高维数据进行聚类和阻尼因子对 AP 聚类算法聚类性能的影响，本书提出三种优化算法，通过主成分分析对数据进行降维，奇异值分解操作，消除冗余信息，进行降维，降低数据规模，提高算法效率；为使算法的收敛更快，提出一种非线性函数策略，根据每次能量函数的收敛情况自适应地调整阻尼系数，提高算法的收敛性能；基于最小簇的匹配的流形 AP 聚类算法通过建立无向图，构建一种新的基于图的流形聚类，充分利用传统 AP 聚类算法优秀的球形和凸集数据聚类能力，不断聚合最小簇，从而完成流形聚类。

1.4 本书结构安排

本书共 6 章，各章内容概括如下。

第 1 章，绪论部分，主要介绍撰写本书的背景和意义，AP 聚类算法的国内外研究现状，以及本书的主要研究内容。

第 2 章，聚类算法的理论基础，介绍相似度计算方式，以及对目前较为流行的聚类算法归纳、分类，重点介绍 AP 聚类算法的具体内容。

第 3 章，提出优化参数的算法，分别是基于果蝇优化的吸引子传播（FOA-AP）聚类算法、基于果蝇优化的自适应吸引子传播（FOA-SAP）聚类算法、基于烟花爆炸优化的半监督吸引子传播（FEO-SAP）聚类算法、基于布谷鸟优化的半监督吸引子传播（CS-SAP）聚类算法、基于稳定阈值的吸引子传播（STAP）聚类算法和基于约束规则的吸引子传播（CRAP）聚类算法。

第 4 章，提出若干优化相似度矩阵的 AP 算法，分别是基于变异赋权的吸引子传播（CVAP)聚类算法、基于智能赋权的吸引子传播（IWAP）聚类算法、基于距离贴近度的吸引子传播（CM-AP）聚类算法、半监督自适应权重吸引子传播(AFW-SAP)聚类算法、基于结构相似度的半监督自适应吸引子传播（SAAP-SS）聚类算法。

第 5 章，提出三种优化算法，分别是基于熵权法和主成分分析法相结合的吸引子传播（EWPCA-AP）聚类算法、基于奇异值分解的自适应吸引子传播（SVD-SAP）聚类算法和基于最小簇匹配的流形吸引子传播（MMHC）聚类算法。

第 6 章，结论与展望。

参 考 文 献

[1] 张震, 汪斌强, 伊鹏. 一种分层组合的半监督吸引子传播聚类算法. 电子与信息学报, 2013, (3): 645-651.

[2] 王开军, 张军英, 李丹. 自适应仿射传播聚类. 自动化学报, 2007, 33(12): 1242-1246.

[3] 储岳中, 徐波, 高有涛. 基于吸引子传播聚类与核匹配追踪的遥感图像目标识别方法. 电子与信息学报, 2014, (12): 2923-2928.

[4] 肖宇, 于剑. 基于吸引子传播算法的半监督聚类. 软件学报, 2008, 19(11): 2803-2813.

[5] 董俊, 王锁萍, 熊范纶. 可变相似性度量的吸引子传播聚类. 电子与信息学报, 2010, 32(3): 509-514.

[6] 冯晓磊. 近邻传播聚类算法研究. 郑州: 解放军信息工程大学, 2011: 1-84.

[7] Wang C D, Lai J H, Suen C, et al. Multi-exemplar affinity propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2223-2237.

[8] Guan R C, Shi X H, Maurizio M, et al. Text clustering with seeds affinity propagation. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(4): 627-637.

[9] Kumar J, Wael A A, Kang L, et al. Handwritten arabic text line segmentation using affinity propagation// Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, New York, 2010: 135-142.

[10] Akl A, Feng C, Valaee S. A novel accelerometer-based gesture recognition system. IEEE Transactions on Signal Processing, 2011, 59(12): 6197-6205.

[11] Dueck D, Frey B J, Jojic N, et al. Constructing Treatment Portfolios Using Affinity Propagation. Berlin: Springer, 2008: 360-371.

[12] Givoni I E, Frey B J. Semi-supervised affinity propagation with instance-level constraints//Proceedings of the International Conference on Artificial Intelligence and Statistics, Clearwater Beach, 2009: 161-168.

[13] 管仁初, 裴志利, 时小虎. 权吸引子传播算法及其在文本聚类中的应用. 计算机研究与发展, 2010, 47(10): 1733-1740.

[14] Yang C, Bruzzone L, Guan R C, et al. Incremental and decremental affinity propagation for semisupervised clustering in multispectral images. IEEE Transactions on Geoscience and Remote Sensing, 2013, 51(3): 1666-1679.

[15] 何晏成. 基于吸引子传播和凝聚层次的文本聚类方法. 哈尔滨: 哈尔滨工业大学, 2010: 1-60.

[16] 鲁伟明, 杜晨阳, 魏宝刚. 基于 Mapreduce 的分布式吸引子传播聚类算法. 计算机研究与发展, 2012, 49(8): 1762-1772.

[17] 许晓丽, 卢志茂, 张格森. 改进吸引子传播聚类的彩色图像分割. 计算机辅助设计与图形学学报, 2012, 24(4): 514-519.

[18] Hassanabadi B, Shea C, Zhang L, et al. Clustering in vehicular ad hoc networks using affinity propagation. Ad Hoc Networks, 2014, 13: 535-548.

[19] Givoni I, Chung C, Frey B J. Hierarchical affinity propagation. Preprint, arXiv: 1202. 3722, 2012: 1-9.

[20] Santana H R, McGarry L M, Bielza C, et al. Classification of neocortical interneurons using affinity propagation. Front Neural Circuits, 2013, 7(185): 185.

[21] Zhang X L, Furtlehner C, Cecile G R, et al. Data stream clustering with affinity propagation. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1644-1656.

[22] Shang F H, Jiao L C, Shi J R, et al. Fast affinity propagation clustering: A multilevel approach. Pattern Recognition, 2012, 45(1): 474-486.

[23] Sun C X, Huo H W, Yu Q, et al. An affinity propagation-based DNA motif discovery algorithm. BioMed Research International, 2015: 853-461.

[24] Gelbard R, Goldman O, Spiegler I. Investigating diversity of clustering methods: An empirical comparison. Data & Knowledge Engineering, 2007, 63(1): 155-166.

第 2 章　聚类算法的理论基础

2.1　相似性度量方式

聚类算法本质上是基于样本间相似度进行分组的，所以定义一个合理的相似性度量方式是至关重要的，这在很大程度上决定了算法的聚类性能。对象之间的相似性通常采用距离函数或相似性度量函数表示，这里的距离不仅仅是欧氏空间上的距离，还包括语义、时间、状态、密度等产生的差距。同样，在处理不同的问题和数据时，应该充分考虑数据自身的特点来采用合适的距离函数。给定两个数据点 $x_i=(x_{i1},x_{i2},\cdots,x_{id})^{\mathrm{T}}$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jd})^{\mathrm{T}}$，本节将对几种常用的相似性度量方式进行详细介绍。

（1）欧氏距离。欧氏距离是指在一个 d 维空间内两个样本点之间直线距离，是数据挖掘中最常用的距离函数，其定义为

$$d_{ij}=\sqrt{\sum_{k=1}^{d}(x_{ik}-x_{jk})^2} \tag{2.1}$$

（2）曼哈顿距离（Manhattan distance）。曼哈顿距离为 L1-距离或城市区块距离，即在欧氏空间的固定直角坐标系上两点所形成的线段对轴产生投影的距离总和，其定义为

$$d_{ij}=\sum_{k=1}^{d}\left|x_{ik}-x_{jk}\right| \tag{2.2}$$

（3）切比雪夫距离。切比雪夫距离是指在向量空间中，两个向量对应维度相减后，绝对值最大的数，其定义为

$$d_{ij}=\max_{k}\left(\left|x_{ik}-x_{jk}\right|\right) \tag{2.3}$$

（4）闵可夫斯基距离（Minkowski distance）。闵可夫斯基距离定义的不只是一个距离函数，而是一组距离函数的集合，其定义为

$$d_{ij}=\sqrt[p]{\sum_{k=1}^{d}\left|x_{ik}-x_{jk}\right|^p} \tag{2.4}$$

其中，$p>0$ 是一个参数，不难看出，当 $p=1$ 时，该式即为曼哈顿距离；当 $p=2$ 时，即为欧氏距离；当 p 趋向于无穷大时，即为切比雪夫距离。

（5）马氏距离（Mahalanobis distance）。马氏距离表示数据集的协方差距离，它充分考虑每个特性之间的关联，即独立于测量尺度，其定义为

$$d_{ij}=\sqrt{(x_i-x_j)^{\mathrm{T}}S^{-1}(x_i-x_j)} \tag{2.5}$$

其中，S 为协方差矩阵。若 S 为单位矩阵，马氏距离即为欧氏距离。

（6）余弦相似度。余弦相似度是用向量空间中两个向量夹角的余弦值作为衡量两个样本间差异程度。该值越大，表明夹角越小，即两者的方向越一致，两个向量越相似，其定义为

$$d_{ij}=\cos\theta=\frac{x_i\cdot x_j}{\|x_i\|\|x_j\|}=\frac{\sum_{k=1}^{d}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{d}x_{ik}^2}\sqrt{\sum_{k=1}^{d}x_{jk}^2}} \tag{2.6}$$

（7）皮尔逊相关系数。皮尔逊相关系数是用于反映变量之间相关关系密切程度的统计指标。皮尔逊相关系数是按积差方法计算，同样以两变量与各自平均值的离差为基础，通过两个离差相乘来反映两变量之间相关程度，其定义为

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E((X-E(CX))(Y-E(Y)))}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{2.7}$$

皮尔逊相关系数是衡量随机变量 X 与 Y 相关程度的一种方法。相关系数的绝对值越大，则表明 X 与 Y 相关度越高。当 X 与 Y 线性相关时，相关系数取值为 1（正线性相关）或–1（负线性相关）。

2.2　聚类算法分类

由于数据的复杂性和多样性，往往并不存在某种单一的聚类算法使其能够适用所有情况，鉴于此，众多国内外学者根据不同领域的特殊问题，提出大量表现出色的聚类算法。在聚类算法的选择上，需要综合地分析数据的规模与结构特点，并根据算法自身的特性与优势，选择适合的方法。本节对目前较为流行的聚类算法进行归纳、分类，如图 2.1 所示。

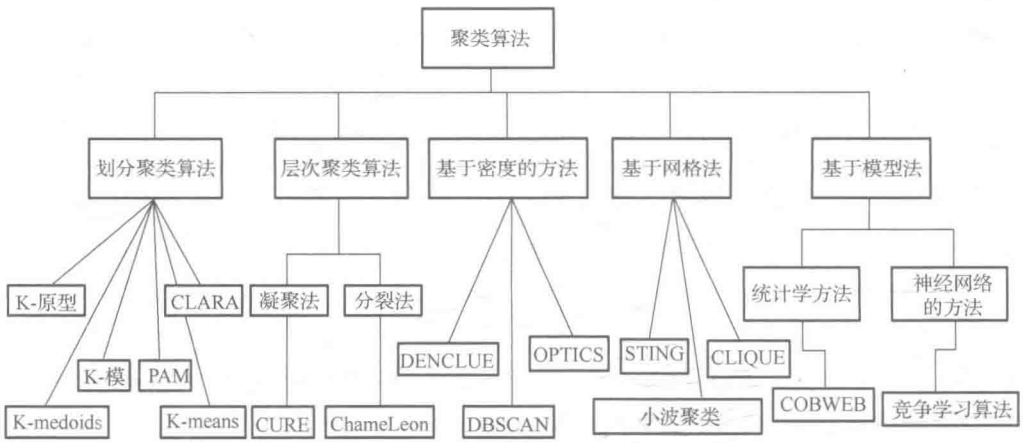

图 2.1　聚类算法分类

2.2.1　基于划分的方法

基于划分的聚类算法是应用最为广泛的聚类算法之一，其核心思想是：对于包含 n 个