文章编号: 1001-0920 (2012) 03-0321-08

聚类分析研究中的若干问题

王 骏, 王士同, 邓赵红

(江南大学 数字媒体学院，江苏 无锡 214122)

摘 要: 聚类分析是重要的数据挖掘方法, 目的是寻找数据集中所包含的簇结构. 以往研究工作中聚类分析的一些基本问题始终是人们关注的重点, 为此在简要回顾具有代表性的研究成果的基础上, 总结了该研究所面临的若干基本问题及解决方法, 以期能够对相关研究提供有益的参考.

关键词: 聚类分析；聚类方法；无监督学习

中图分类号: TP273 **文献标识码:** A

Survey on challenges in clustering analysis research

WANG Jun, *WANG Shi-tong*, *DENG Zhao-hong*

(School of Digital Media，Jiangnan University，Wuxi 214122，China．Correspondent：WANG Jun，E-mail：wangjun_sytu@hotmail.com)

Abstract: Clustering analysis is an important tool for data mining and its purpose is to find the cluster structures in dataset. However, several fundamental challenges associated with clustering still exist. Therefore, the representative methods are briefly surveyed in this paper. Moreover, the major challenges in current work are summarized and the solutions are also discussed in detail. A valuable reference can be provided to the related research work.

Key words: clustering analysis；clustering method；unsupervise learning

1 引 言

聚类分析是一种重要的无监督学习方法, 作为数据分析的工具, 其重要性在各个领域都得到了广泛的认可. 聚类分析的目的是寻找数据集中的“自然分组”, 即所谓的“簇”. 通俗地讲, 簇是指相似元素的集合, 聚类分析就是一个在数据集中寻找相似元素集合的无监督学习过程. 来自不同应用领域的数据集具有不同的特点, 人们对数据进行聚类分析的目的也不尽相同, 聚类分析的方法因数据集而异, 因使用目的而异. 当前, 聚类分析的新方法层出不穷[1-4], 纵观各种聚类算法, 它们使用的技术互不相同, 其理论背景又彼此交叉、重叠, 很难找到一个统一的标准对其进行归类. 基于文献 [2, 4], 聚类分析的方法可分为基于层次的聚类方法、基于划分的聚类方法、基于图论的聚类方法、基于密度和网格的方法等. 这些方法虽然从不同角度使用不同的理论方法研究聚类分析, 但对于不同的实际问题, 聚类分析中的一些基本内容始终是人们关注的焦点.

本文在对经典聚类分析方法进行简要回顾的基础上, 对聚类分析研究中所面临的若干基本问题及具有代表性的解决思路进行归纳和总结, 这对于进一步研究聚类分析具有重要意义.

2 经典聚类方法回顾

2.1 基于层次的聚类方法

基于层次的聚类算法又称为树聚类算法. 该方法使用数据的联接规则, 通过层次式架构方式反复将数据进行分裂或聚合, 以形成一个层次序列的聚类问题的解[4]. 算法主要有两种策略: 自底向上的聚合式层次聚类和自顶向下的分裂式层次聚类. 近年来, 具有代表性的研究成果有 Hungarian 聚类算法[5]、面向连续数据的粗聚类算法 (RCOSD)[6]和基于 Quartet 树的快速聚类算法[7]等.

层次聚类算法的优点在于不需要用户事先指定聚类数目, 可以灵活控制不同层次的聚类粒度, 并且

收稿日期: 2011-05-08；**修回日期:** 2011-08-20.

基金项目: 国家自然科学基金项目(60903100, 61170122)；江苏省自然科学基金项目(BK2009067)；中央高校基本科研业务费专项资金项目(JUSRP21128, JUSRP111A38).

作者简介: 王骏(1978−), 男, 讲师, 博士, 从事模式识别、智能信息处理等研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、模糊神经网络等研究.

可以清晰地表达簇之间的层次关系. 但是, 层次聚类算法也有其不可避免的缺点: 在层次聚类过程中不能回溯处理已经形成的簇结构; 上一层次的簇形成后, 通常不能在后续的执行过程中对其进行调整. 目前, 大多数层次聚类算法的计算复杂度至少为$O(n^2)$, 其中n为数据集包含的数据点数量. 巨大的计算开销已成为提高层次聚类算法性能的瓶颈, 使其不适用于大规模数据集.

2.2 基于划分的聚类方法

基于划分的聚类方法已在模式识别、数据挖掘等领域得到广泛应用, 至今仍是许多研究工作的思想源头. 假设目标函数是可微的, 首先给出数据集的初始划分; 然后以此为起点, 在迭代过程中不断调整样本点的归属, 从而使目标函数达到最优. 当目标函数收敛时, 便可得到最终聚类结果. k-means 和FCM 是这类算法的典型代表, 近年来的研究成果主要有: 密度加权模糊聚类算法[8], 基于混合距离学习的双指数模糊C均值算法[9]等.

这类方法的优点可归结为收敛速度快且易于扩展, 缺点在于它们通常需要事先指定聚类数目. 此外, 初始簇中心的选择、噪声数据的存在和聚类数目的设置均会对聚类结果产生较大影响.

2.3 基于图论的聚类方法

基于图论的聚类方法将待聚类的数据集转化为一个赋权的无向完全图$G = (V, E)$. 其中: 顶点集V为特征空间中的数据点, 边集E及其权重为任意两个数据点之间的联接关系和相似程度. 这样, 便可将聚类问题转化为图划分问题来解决, 所产生的若干个子图对应于数据集包含的簇. 近年来, 代表性的研究成果有GBR 算法[10], 基于最大θ距离子树的聚类算法[11]和基于dominant 集的点对聚类算法[12]等.

基于图论的聚类方法大多使用点对数据来表示数据点之间的相互关系, 与其他方法相比, 这类方法更适于发现数据集中形状不规则的类簇. 但是, 求图的最优划分在数学上可归结为一个NP 难的组合优化问题, 如何面向大规模数据集求图的最优划分仍需要进一步探讨.

2.4 基于密度和网格的聚类方法

基于密度和网格的聚类方法来源于基于密度的聚类方法和基于网格的聚类方法[12]. 前者通常适用于只包含数值属性的数据集, 后者适用于任何属性的数据集. 由于这两类方法在处理数据时都侧重于使用样本点的空间分布信息, 并且经常结合在一起使用, 可将它们归为一类. 该类方法对处理形状复杂的簇具有明显的优势, 近年来具有代表性的研究成果有TFCTMO 算法[13]和ST-DBSCAN 算法[14]等.

3 聚类分析研究面临的基本问题

作为数据挖掘的重要工具, 聚类分析已经得到了广泛关注. 近年来, 随着信息技术的迅猛发展, 具有不同结构特点的数据不断涌现, 为聚类分析的研究提出了新的挑战. 尽管如此, 聚类分析研究中的基本问题始终是人们研究工作的重点内容, 其有效解决对于数据挖掘、模式识别中的许多问题都具有重要的借鉴意义. 这些基本问题包括: 1) 对于不同结构特征的数据, 如何合理计算数据点之间的相异或相似程度; 2) 对于包含噪声或例外点的数据, 如何提高算法的鲁棒性; 3) 对于高维数据, 如何进行特征降维; 4) 对于包含多个类簇的数据, 如何确定数据集包含的聚类数目; 5) 对于大规模数据集, 如何进行高效的聚类. 本节将在前人工作的基础上对解决这些问题的常用方法进行论述.

3.1 如何合理计算数据点之间的相异或相似程度

如何衡量数据点之间的相异或相似程度是设计聚类算法的基础问题, 会直接影响聚类分析的效果, 最直观的方法是使用距离函数或相似性函数. 通常而言, 数据集上的距离函数应该满足对称性和非负性, 如果它还满足三角不等式和自反性, 则该距离函数就是一个度量, 数据集上的相似性函数也有类似的性质.

在模式识别的基础理论中, 很多距离计算方法都可以归结为基于向量p范数的距离, 即Minkowski 距离. 已有研究表明, p的不同取值会对聚类分析的结果产生重要影响[15-16]. 当$p = 2$时, Minkowski 距离退化为欧氏距离, 是最常用的距离计算方法. 但是, 使用欧氏距离的聚类算法大多只能发现低维空间中呈超球状分布的数据, 并且对数据集中的噪声比较敏感. 当$p \to \infty$时, Minkowski 距离演变为Sup 距离; 当$p = 1$时演变为City-block 距离. 研究表明, City-block 距离可以有效提高模糊聚类算法对噪声或例外点的鲁棒性[16].

Mahalanobis 距离为原特征空间中的数据在线性投影空间欧氏距离[17]. 使用Mahalanobis 距离能够使聚类算法成功发现数据集里呈超椭球型分布的类簇, 但Mahalanobis 距离同样会带来较大的计算量[18].

为了克服欧氏距离对于数据集里的噪声比较敏感的缺点, Wu 等人提出了Alternative 距离, 它对数据集里的噪声不敏感, 因此可以用来改造传统聚类算法, 使其对于噪声具有更强的鲁棒性[19-20].

在距离计算过程中, 核技巧的引入可以使原特征空间中线性不可分的数据在核空间中线性可分. 常用的核函数包括Gauss 核函数、多项式核函数和

表 1　常见的相似或相异程度计算方法

方法	表达式	典型应用
Minkowski distance	$d_{ij}=\Big(\sum_{h=1}^{s}\lvert x_{ih}-x_{jh}\rvert\Big)^{1/p}$	基于 Lp 范数的 GFCM[16]
Euclidean distance	$d_{ij}=\Big(\sum_{h=1}^{s}\lvert x_{ih}-x_{jh}\rvert\Big)^{1/2}$	FCM, K-means
City-block distance	$d_{ij}=\lvert x_{ih}-x_{jh}\rvert$	pHCM ($p=1$)[15]; pFCM ($p=1$)[15]
Sup distance	$d_{ij}=\max\lvert x_{ih}-x_{jh}\rvert$	pHCM ($p=\infty$)[15], pFCM ($p=\infty$)[15]
Mahalanobis distance	$d_{ij}=(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}S^{-1}(\mathbf{x}_i-\mathbf{x}_j)$	S 为协方差矩阵; HEC[22], Gustanfson-Kessel[23]
Cosine similarity	$s_{ij}=\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j/(\lVert\mathbf{x}_i\rVert\lVert\mathbf{x}_j\rVert)$	Spherical k-means Algorithm [24]
Alternative distance	$d_{ij}=1-\exp(-\beta\lVert\mathbf{x}_i-\mathbf{x}_j\rVert^2)$	AHCM, AFCM[19-20]
Feature weighted distance	$d_{ij}=\Big(\sum_{h=1}^{s}w_h^{\alpha}\lvert x_{ih}-x_{jh}\rvert\Big)^{1/2}$	DI-FSC [25]

Sigmoid 核函数等. 引入核后, 很多聚类算法都可以改造成相应的核化版本, 从而扩展了其处理非线性数据的能力.

高维空间中, 使用传统的距离函数会出现不稳定现象, 即随着数据维数的增加, 数据点到最近邻点的距离将趋于它到最远点的距离[21]. 文献 [22] 通过研究表明, 该现象是由距离分布的变化引起的, 基于相关的理论分析, 进一步设计了新的距离计算方法来克服不稳定现象, 有效提高了高维空间中机器学习算法的性能.

表 1 对常见的距离或相似性计算方法进行归纳, 并给出了典型应用.

3.2　如何提高聚类算法对噪声、例外点的鲁棒性

噪声和例外点在各种类型的数据集中普遍存在. 为了减少聚类过程中噪声、例外点对正常数据的影响, 提高聚类算法的抗噪性能具有重要意义.

文献 [26] 为噪声数据引入了单独的簇, 提出了 NC(noise clustering) 算法. 在 NC 算法中, 到各簇中心的距离大于噪声距离 δ 的数据点被归入到这一单独的簇中. NC 算法成功地减小了聚类过程中噪声或例外点对正常簇的影响, 但不具有标识例外点的功能. 文献 [27] 将 NC 算法进行扩展, 给出了一种计算噪声距离 δ 的新方法, 并将其用于例外点检测.

在聚类过程中, 使用对噪声鲁棒的距离度量来计算数据点之间的距离成为提高算法鲁棒性的另一种方法. 如, 文献 [18-19] 利用噪声鲁棒的距离度量来代替 k-means 和 FCM等算法中的欧氏距离, 提出了对噪声更加鲁棒的 AHCM 和 AFCM 等算法. [28] 采用类似方法来增强算法对噪声和例外点的鲁棒性. [29-30] 使用 Vapnik's ϵ- 不敏感损失函数作为距离度量, 分别将其用于极大熵聚类和模糊聚类神经网络, 从而得到了更加鲁棒的聚类效果, 并且成功地标识了例外点.

为参与聚类的样本赋予权值是提高算法对噪声和例外点鲁棒性的另一种有效方法. 基于 FCM 算法, 文献 [31] 为每一个样本点分配一个动态权值, 以此来衡量各样本点对于聚类的贡献大小, 通过动态迭代过程在进行模糊聚类的同时发现了数据集包含的例外点. 在这一过程中, 动态权值的引入增强了算法的鲁棒性. 采用类似的思路, [30] 基于极大熵聚类框架研究了鲁棒的极大熵聚类算法 RMEC, [32] 研究了基于图像直方图加权的 FCM 图像分割算法, [33] 将样本加权与特征加权相结合提出了新的模糊聚类算法.

在 FCM 算法中, 样本点对于某一个簇的隶属度通常取值在 [0,1] 之间. 文献 [34] 讨论了一类基于"簇核"概念的模糊聚类算法. 通过引入"簇核", 样本对某一个簇的隶属度可以等于 1. 文献 [35] 将此类算法进行改进, 使其适用于非球状分布的数据. 研究表明, 基于"簇核"概念的模糊聚类算法对噪声或例外点具有较好的鲁棒性.

可能性聚类是软聚类的一种重要形式[36], FCM 要求数据点属于不同类别的隶属度之和为 1. 从本质上讲, FCM 中的模糊隶属度体现了数据点被不同簇的共享程度. 但是, 同一个簇中具有相同隶属度值的两个数据点到这个簇中心的距离可能会大不相同, 因此, FCM 无法根据模糊隶属度来判断并处理数据集包含的噪声和例外点. 可能性聚类放宽了这一约束, 并采用新的方法来计算数据点对各个簇模糊隶属度. 由于模糊隶属度包含了样本点到聚类中心的距离信息, 远离聚类中心的例外点或噪声对聚类中心的影响有限, 从而使得算法对例外点或噪声有更好的鲁棒性.

3.3　如何对高维数据进行聚类

许多聚类算法对高维数据显得无能为力, 这是因为在高维空间中, 传统的距离函数会出现不稳定现象, 数据点之间的距离变得几乎相等[21]. 在实际应用中, 文本数据、基因数据、时间序列、基因表达数据、生物特征数据等都是典型的高维数据, 如何对这些高维数据进行有效的聚类成为当前研究的热点和难点.

3.3.1 特征约简技术

为了对高维数据进行聚类, 通常可以先进行特征约简, 将高维特征空间中的数据转换到低维特征空间, 然后使用 k-means 等传统的聚类算法在低维特征空间中进行聚类[37]. 特征约简技术可分为特征选择和特征提取两种策略: 前者是指从一组特征中选取一些最有代表性的特征以达到降低特征空间维数的目的; 后者是指将高维空间中的数据通过线性或非线性变换映射到低维空间中. 典型的线性映射方法有 PCA[38], ICA[39], LPP[40]等, 非线性映射方法有 ISOMAP[41], LLE[42], NPE[43]等. 特征约简技术是一个相对独立的研究领域, 本文不再展开.

特征约简技术可以有效降低计算开销, 并且使用户对感兴趣的数据有更清晰的理解. 但是, 在特征约简过程中会不可避免地发生信息损失, 从而使聚类结果产生失真.

3.3.2 特征加权技术

对于高维数据而言, 每个特征在聚类过程中所起的作用是不同的, 部分特征在聚类过程中起了主导作用, 它们对簇的形成起到积极作用; 而另一部分特征在聚类过程中起的作用通常不大, 有时甚至会引入噪声而为簇的生成带来负面影响. 以此为出发点, 对特征进行加权成为处理高维数据的有效方法, 这相当于在欧氏空间中拉长或缩短不同特征所对应的轴.

文献 [44] 首次提出了特征加权算法 SYNCLUS. 在该算法中, k-means 聚类和计算特征权重交替进行, 直到算法收敛, 但是 SYNCLUS 巨大的计算开销使之不适用于大规模数据集. 文献 [45] 提出了基于特征加权的 Convex k-means. 为了在预先指定的若干组特征权重中选取最适合数据集的权重, 引入广义 Fisher 比 Q 作为评价指标, 从而得到理想的聚类结果. 但是, 这种方法需要预先得到特征权重的候选集合, 无法保证最优的特征权重出现在这个候选集合中. 此外, 对于高维数据, 为每一维特征指定权重在实际应用中也不尽可行.

在聚类前首先对特征权重进行学习, 文献 [46-48] 提出了几种面向 k-means 或 FCM 的特征权重学习算法. 这些方法的共同点在于, 先通过学习算法对特征权重进行学习, 在此基础上形成特征加权的距离函数; 然后使用 k-means 或 FCM 算法对目标函数进行迭代优化, 从而得到数据集的划分.

近年来, 自动特征加权技术得到了充分研究, 这类方法将特征权重的学习融合在聚类分析的过程中. 例如, 文献 [49] 提出了 W-k-means 算法, 该算法基于传统 k-means 算法在迭代过程中增加了新的步骤来计算特征权重, 通过迭代优化实现了在聚类分析的同时进行特征权重的计算. 与某些固定权重的聚类算法相比, 这种方法在一些数据集上可以得到更好的聚类结果. [50] 对其进行了扩展, 将这一思想用于模糊聚类. [33] 将样本加权与特征加权相结合, 提出了具有特征排序功能的模糊聚类算法. 从距离学习的角度看, 这类算法在进行聚类分析的同时也成功地学习了适合于数据集的最佳特征加权距离.

3.3.3 子空间聚类技术

在高维空间中, 属于不同类簇的样本点通常分布在由不同特征子集构成的子空间中. 子空间聚类技术就是针对类簇的这一分布特点而设计的聚类方法. 根据各个特征对于不同类簇的从属关系, 子空间聚类算法可以分为硬子空间聚类和软子空间聚类两大类.

在硬子空间聚类中, 数据集不同的特征子集张成不同的子空间, 硬子空间聚类则在这些不同的子空间中搜索类簇. 对于硬子空间聚类而言, 某个特征或者属于某个类簇, 或者不属于某个类簇. 与数据集“硬划分”概念不同的是, 一个特征可以同时从属于多个类簇. 研究表明, 硬子空间聚类算法能够成功地发现高维数据集不同子空间中任意形状的类簇, 但它们对参数的选取比较敏感, 如何合理地选取参数仍是进一步需要研究的问题.

在特征子集的选择上引入模糊概念, 学术界提出了“软子空间”的概念[51]. 相应地, 软子空间聚类技术也成为近年来的研究热点, 其基本思想是, 数据集中的各类别赋予不同的特征权重向量, 以此来表示聚类过程中各维特征对此类别贡献的大小. 在聚类过程中, 每一维特征对于各个类别都有不同的贡献, 因此每一类都有不同的特征权重向量, 从而在整个特征空间中形成了若干个“软子空间”, 聚类过程就是在各个“软子空间”中进行的, 典型算法有 EWKM[51], FWKM[52], FSC[53], LAC[54]等. 这类方法的收敛速度较快, 其计算复杂度与数据集中包含的样本数量和特征数量均成线性关系, 因此也适用于大规模数据集和高维数据集, 但这些算法均基于数据集的硬划分. 文献 [25,55] 先后将模糊划分的概念引入软子空间聚类. [25]研究了目标函数中带两个模糊矩阵的适用于高维稀疏数据集的软子空间聚类算法; [55] 进一步将模糊划分、类内和类间信息融入软子空间聚类技术中, 提出了增强的软子空间模糊聚类算法 ESSC. 研究表明, 软子空间聚类技术能够广泛应用于基因表达数据、文本数据等高维数据的聚类和数字图像的纹理分割等任务. 与基于整个数据集采用单一特征权重的特征加权聚类算法相比, 这些算法可以取得更好的聚类效果.

3.4 如何确定数据集包含的聚类数目

聚类过程将数据集划分为若干个子集, 虽然在某些情况下, 用户根据自身经验可以为数据集选择较为合理的聚类数目, 但大多数情况下, 数据集包含的聚类数目对用户而言是未知的. 许多聚类算法将聚类数目作为一个需要预先设定的输入参数, 对这类聚类算法而言, 聚类结果的质量与此参数的设置密切相关. 如果用户设置的聚类数目过大, 则会使聚类结果过于复杂而难以解释; 相反, 如果数目过少, 则聚类结果中会丢失许多有价值的信息. 可见, 为数据集确定合理的聚类数目, 无论对实际应用本身还是对聚类算法的有效运行都具有十分重要的意义.

估计聚类数目最简单的方法是将数据可视化. 对于可以有效地投影到二维欧氏空间中的数据集而言, 通过数据点在二维空间中的分布图可以直观地获取数据集包含的聚类数目信息. 但是, 对于高维数据和结构复杂的数据而言, 这种方法往往不适用. 下面就如何估计数据集包含的最佳聚类数目这一问题对相关技术进行总结.

1) 基于聚类有效性指标的方法

聚类有效性指标用来评价聚类算法在数据集上生成划分的质量. 通过构造合适的聚类有效性指标来完成聚类数目的估计是一种有效的方法, 其思路是: 在一定范围内设置不同的聚类数目值 k, 并在数据集上进行聚类, 使用聚类有效性指标来评价其结果, 有效性指标的值最大、最小或出现明显拐点时所对应的 k 值即为最佳聚类数目 k_{opt}. 文献 [56] 对常见的有效性指标作了归纳, 这些指标大多从簇内紧凑性、簇间分离性出发来考虑问题, 并综合考虑平方误差、数据的统计和几何特性、参与聚类的数据集大小、簇的相似或非相似程度、簇的数量等因素.

为了提高估计最佳聚类数目 k_{opt} 的效率, 必须确定 k_{opt} 的搜索范围, 即确定 $k_{\max}$ 以满足 $k_{\mathrm{opt}} \leqslant k_{\max}$. 多数研究者使用经验规则$k_{\max} \leqslant \sqrt{n}$, 文献 [57] 就这一问题给出了理论探讨, 并指出其在理论上的合理性.

基于聚类有效性指标通过穷举搜索的策略来确定最佳聚类数目有以下缺点: 1) 该方法必须依次尝试每一个 k 值, 这将导致巨大的运算量; 2) 对于每一个 k 值, 不能保证聚类结果为全局最优解; 3) 当数据集存在噪声时, 聚类有效性指标可靠性不强.

2) 启发式方法

为了克服以上缺点, 近年来一些新方法相继提出, 其主要思路是基于某些准则来指导聚类的过程, 在聚类过程中对聚类数目进行调整, 这样在完成聚类任务的同时也可得到数据集合适的聚类数目. 代表性成果有基于分裂式层次聚类思想的 X-means 算法[58]. 在聚类过程中, X-means 使用 BIC 准则来决定哪个簇应该进行分裂, 随后在这个簇上运行 k-means, 从而为数据集找到最佳聚类数目.

与上述过程相反, RCA[59]通过一个竞争聚合过程来完成对数据集的聚类, 并且得到了合理的聚类数目. 在 RCA 中, 失去竞争力的簇被舍弃, 并被其他簇吸收. 文献 [60] 对这一过程进行推广, 在确定聚类数目的过程中考虑了算法复杂度和聚类准确性之间的平衡, 并将其用于图像分割. 此外, 从数据集概率密度分布的角度出发, 基于单点迭代技术的均值漂移算法为自动确定聚类数目提供了新的思路, 其基本思想是: 通过均值漂移, 收敛于同一点的数据点应该属于同一类. 文献 [61] 指出, 均值漂移算法在满足一定条件下必然可以收敛到与起始点最近的一个概率密度函数的稳态点, [62] 给出了理论证明, 为利用均值漂移算法确定聚类数目提供了理论基础.

将单点迭代技术与层次聚类技术相结合, 文献 [28] 提出了一种基于相似度聚类的聚类数目估计方法. 其思路是: 将相似度聚类转化为核密度估计问题, 随后引入单点迭代技术, 使算法收敛于核密度最大的点. 在此基础上进行层次聚类, 从而得到数据集包含的聚类数目及相应的簇. 此外, [63] 提出了基于谱分解的聚类数目确定方法, [11] 提出的 GBR 算法可以从优化问题的解析解中得到聚类数目, 这些工作为相关研究拓宽了思路.

3.5 如何对大规模数据进行高效聚类

近年来, 大规模数据集在各领域的频繁出现对聚类分析研究提出了新的挑战. 对大规模数据集进行聚类分析通常从两方面进行考虑：一方面, 开发算法复杂度较低的算法, 通常认为, 当算法的时间与空间复杂度与数据集大小接近线性关系时, 该算法适合于处理大规模数据集; 另一方面对原数据集进行采样或压缩, 在不影响聚类效果的前提下得到原数据集的子集. 下面就这两个方面对现有的相关工作进行回顾.

层次式聚类算法大多采用单联接、全联接、类内平均联接、类间平均联接等联接规则, 其时间复杂度和空间复杂度至少为 $O(n^2)$(其中 n 为数据集包含样本点的数量), 因此它们不适用于大规模数据集的聚类问题.

以 k-means 为代表的划分式聚类算法在算法复杂度方面有突出的优势. 对于 k-means 聚类算法而言, 算法的时间复杂度为 $O(ndk)$. 其中: n 为数据集大小, d 为数据集的维数, k 为聚类数目. 显然这类算法的复杂度与数据集的大小和维数均成线性关系, 因此适合大规模数据集. 但这类划分式聚类算法通常只适用于呈超球状、超椭球状分布的数据. 此外, 当类簇之间的

大小差异明显时, k-means 所使用的平方误差准则不能有效地将不同类簇的数据点分开. 因此, 对于形状不规则或类簇大小差异明显的大规模数据集而言, 如何以较低的算法复杂度实现聚类仍然是一个亟待解决的问题.

对于高时间复杂度的聚类算法, 可以在聚类前对原数据集进行采样. 在保证采样方法不破坏数据集中类簇分布形状的前提下, 这种方法可以有效提高聚类算法的效率. 随机采样是最简单最有效的采样策略, 文献 [64] 深入研究了大规模数据集聚类分析问题的随机采样技术, 并通过 Chernoff 界来保证在丢失类簇概率较低的前提下, 所需随机采样的样本点数目最小.

机器学习问题大多可以归结为优化问题. 对于优化问题而言, 如果它能够在数据集的某一子集上获得与它在整个数据集上相近的解, 则称这一子集为核心集[65]. 核心集技术可以看作是在某种启发式准则下一种特殊的采样方法. 近年来, 相关理论得到了充分的研究: 文献 [66-67] 先后指出了在一维空间和高维空间中存在适用于 k-center 聚类的 ϵ-coreset; 文献 [68] 研究了 k-means 和 k-median 聚类的核心集问题.

最近, 文献 [69] 提出了一种面向最小包含球问题的核心集求解算法, 其最大特点在于该算法所得核心集的大小与原数据集的维数和大小均无关. 文献 [70] 将该技术用于谱聚类, 提出了新的适用于大规模数据集的谱聚类算法. 该算法将聚类问题转化为特殊形式的二次规划问题, 然后将其与中心约束最小包含球 CCMEB 问题建立起等价关系, 设计了相应的 CCMEB 核心集选取策略, 并结合 k 最近邻技术, 实现了对大规模数据集的聚类.

4 结 论

聚类分析作为无监督学习的一种重要形式, 具有广泛的应用前景. 本文重点就聚类分析研究的基本问题以及常见的解决方法进行了归纳和总结, 这些问题在今后相当长的时间内仍将是学术研究的热点.

当前, 随着信息技术的飞速发展, 各行各业出现了各种复杂结构的数据集. 下一步的工作是在已有成果的基础上, 针对这些复杂结构的数据集的聚类分析问题开展有针对性的研究, 以进一步丰富聚类分析的研究内容和理论方法.

参考文献(References)

[1] Xu R, Wunsch D. Survey of clustering algorithms[J]. IEEE Trans on Neural Networks, 2005, 16(3): 645-678.

[2] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[3] Jain A K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[4] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
(Sun J G, Liu J, Zhao L Y. Clustering algorithms research[J]. J of Software, 2008, 19(1): 48-61.)

[5] Goldberger J, Tassa T. A hierarchical clustering algorithm based on the Hungarian method[J]. Pattern Recognition Letters, 2008, 29(1): 1632-1638.

[6] Kumar P, Krishna P R, Bapi R S, et al. Rough clustering of sequential data[J]. Data & Knowledge Engineering, 2007, 3(2): 183-199.

[7] Cilibrasi R L, Vitányi P MB. A fast quartet tree heuristic for hierarchical clustering[J]. Pattern Recognition, 2011, 44(3): 662-677.

[8] Hathaway R J, Hu Y. Density-weighted fuzzy c-means clustering[J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 243-252.

[9] 王骏, 王士同. 基于混合距离学习的双指数模糊 C 均值算法[J]. 软件学报, 2010, 21(8): 1878-1888.
(Wang J, Wang S T. A dsouble-indexed FCM algorithm based on Hybrid distance metric learning[J]. J of Software, 2010, 21(8): 1878-1888.)

[10] Lee C H, Zaiane O R, Park H-H, et al. Clustering high dimensional data: A graph-based relaxed optimization approach[J]. Information Sciences, 2008, 178(23): 4501-4511.

[11] Li Y J. A clustering algorithm based on maximal θ-distant subtrees[J]. Pattern Recognition, 2007, 40(5): 1425-1431.

[12] Pavan M, Pelillo M. Dominant sets and pairwise clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 167-172.

[13] Nanni M, Pedreschi D. Time-focused clustering of trajectories of moving objects[J]. J of Intelligent Information Systems, 2006, 27(3): 267-289.

[14] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[15] Bobrowski L, Bezdek J C. C-means clustering with the L_1 and L_∞ norms[J]. IEEE Trans on System, Man and Cybernetics, 1991, 21(3): 545-554.

[16] Hathaway R J, bezdek J C, Hu Y. Generalized fuzzy c-means clustering strategies using Lp norm distances[J]. IEEE Trans on Fuzzy Systems, 2000, 8(5): 576-582.

[17] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. J of Machine Learning Research, 2009, 10(2): 207-244.

[18] Mao J, Jain A K. A self-organizing network for hyperellipsoidal clustering(HEC)[C]. IEEE Int Conf on Neural Networks. Orlando: IEEE, 1994, 5: 2967-2972.

[19] Wu K L, Yang M S. Alternative c-means clustering algorithms[J]. Pattern Recognition, 2002, 35(10): 2267-2278.

[20] Zhang D, Chen S: A comment on alternative c-means clustering algorithms[J]. Pattern Recognition, 2004, 37(2): 173-174.

[21] Beyer K, Goldstein J, Ramakrishnan R, et al. When is "nearest neighbor" meaningful[J]. Lecture Notes in Computer Science, 1999, 1540: 217-235.

[22] Hsu C M, Chen M S. On the design and applicability of distance functions in high-dimensional data space[J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(4): 523-536.

[23] Gustafson D E, Kessel W C. Fuzzy clustering with a fuzzy covariance matrix[C]. Proc of the IEEE CDC. San Diego: IEEE, 1979: 761-766.

[24] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering[J]. Machine Learning, 2001, 42(1): 143-175.

[25] 王骏, 王士同, 王晓明. 基于特征加权距离的双指数模糊子空间聚类算法[J]. 控制与决策, 2010, 25(8): 1207-1210.
(Wang J, Wang S T, Wang X M. Double-indices fuzzy subspace clustering algorithm based on feature weighted distance[J]. Control and Decision, 2010, 25(8): 1207-1210.)

[26] Dave R N. Characterization and detection of noise in clustering[J]. Pattern Recognition Letters, 1991, 12(11): 657-664.

[27] Rehm F, Klawonn F, Kruse R. A novel approach to noise clustering for outlier detection[J]. Soft Computing, 2007, 11(5): 489-494.

[28] Yang M S, Wu K L. A similarity-based robust clustering method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(4): 434-448.

[29] 邓赵红, 王士同. 鲁棒性的模糊聚类神经网络[J]. 软件学报, 2005, 16(8): 1415-1422.
(Deng Z H, Wang S T. Robust fuzzy clustering neural networks[J]. J of Software, 2005, 16(8): 1415-1422.)

[30] Wang Shi-tong, Chung Fu-lai, Deng Zhao-hong. Robust maximum entropy clustering algorithm with its labeling for outliers[J]. Soft Computing, 2006, 10(7): 555-563.

[31] Keller A. Fuzzy clustering with outliers[C]. Proc of the 19th Int Conf of the North American Fuzzy Information Processing Society, Atlanta. USA: IEEE, 2000: 143-147.

[32] 高新波, 李洁, 姬红兵. 基于加权模糊 c 均值聚类与统计检验指导的多阈值图像自动分割算法[J]. 电子学报, 2004, 32(4): 661-664.
(Gao X B, Li J, Ji H B. A multi-threshold image segmentation algorithm based on weighting fuzzy c-means clustering and statistical test[J]. Acta Electronica Sinica, 2004, 32(4): 661-664.)

[33] 皋军, 王士同. 具有特征排序功能的鲁棒性模糊聚类方法[J]. 自动化学报, 2009, 35(2): 145-153.
(Gao J, Wang S T. Fuzzy clustering algorithm with ranking features and identifying noise simultaneously[J]. Acta Automatica Sinica, 2009, 35(2): 145-153.)

[34] Wu K L, Yu J, Yang M S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests[J]. Pattern Recognition Letters, 2005, 26(5): 639-652.

[35] Yang M S, Wu K L, Hsieh J N, et al. Alpha-cut implemented fuzzy clustering algorithms and switching regressions[J]. IEEE Trans on Systems, Man and Cybernetics, 2008, 38(3): 588-603.

[36] Zhang J S, Leung Y W. Improved possibilistic c-means clustering algorithms[J]. IEEE Trans on Fuzzy Systems, 2004, 12(2): 209-217.

[37] Nasser A, Hamad D. K-means clustering algorithm in projected spaces[C]. Proc of the 9th Int Conf on Information Fusion. Florence: ISIF, 2006: 1-6.

[38] Jolliffe I T. Principal component analysis[M]. New York: Springer-Verlag, 1989.

[39] Jutten C, Herault J. Independent component analysis versus PCA[C]. Proc of European Signal Processing Conf. Elsevier, 1988: 287-314.

[40] He X, Niyogi P. Locality preserving projections[C]. Advances in Neural Information Processing Systems 16. Cambridge MA: MIT Press, 2004: 585-591.

[41] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for Nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[42] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[43] He X, Cai D, Yan S, et al. Neighborhood preserving embedding[C]. Proc of the Tenth IEEE Int Conf on Computer Vision. Washington DC: IEEE Computer Society, 2005: 1208-1213.

[44] Desarbo W S, Carroll J D, Clark L A, et al. Synthesized clustering: A method for amalgamating clustering bases with differential weighting variables[J]. Psychometrika, 1984, 49: 57-78.

[45] Modha D S, Spangler W S. Feature weighting in k-means clustering[J]. Machine Learning, 2003, 52(3): 217-237.

[46] 王熙照, 王亚东, 湛燕, 等. 学习特征权值对 K-均值聚类算法的优化[J]. 计算机研究与发展, 2003, 40(6): 869-873.
(Wang X Z, Wang Y D, Zhan Y, et al. Optimization of k-means clustering by feature weight learning[J]. J of Computer Research and Development, 2003, 40(6): 869-873.)

[47] 王丽娟, 关守义, 王晓龙, 等. 基于属性权重的Fuzzy C Mean算法[J]. 计算机学报, 2006, 29(10): 1797-1803.
(Wang L J, Guan S Y, Wang X L, et al. Fuzzy C mean algorithm based on feature weights[J]. Chinese J of Computers, 2006, 29(10): 1797-1803.)

[48] 李洁, 高新波, 焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1): 89-92.
(Li J, Gao X B, Jiao L C. A new feature weighted fuzzy clustering algorithm[J]. Acta Electronica Sinica, 2006, 34(1): 89-92.)

[49] Huang Z, Ng M K, Rong H, et al. Automated variable weighting in k-means type clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[50] Wang Q, Ye Y, Huang J Z. Fuzzy k-means with variable weighting in high dimensional data analysis[C]. Proc of the 6th Int Conf on Web-Age Information Management. Zhangjiajie: IEEE, 2008: 365-372.

[51] Jing L, Ng M K, Huang J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[52] Jing L, Ng M K, Xu J, et al. Subspace clustering of text documents with feature weighting k-means algorithm[C]. Proc of the 9th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Hanoi, 2005: 802-812.

[53] Gan G, Wu J. A convergence theorem for the fuzzy subspace clustering(FSC) algorithm[J]. Pattern Recognition, 2008, 41(6): 1939-1947.

[54] Domeniconi C, Gunopulos D, Ma S, et al. Locally adaptive metrics for clustering high dimensional data[J]. Data Mining and Knowledge Discovery, 2007, 14(1): 63-97.

[55] Deng Zhao-hong, Choi Kup-sze, Chung Fu-lai, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. Pattern Recognition, 2010. 43(3): 767-781.

[56] Bezdek J C, Pal N R. Some new indexes of cluster validity[J]. IEEE Trans on Systems, Man and Cybernetics, 1998, 28(3): 301-315.

[57] 于剑, 程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学 (E辑), 2002 32(2): 274-280.
(Yu J, Chen Q S. Search range of the optimal cluster number in fuzzy clustering[J]. Science in China(Series E), 2002, 32(2): 274-280.)

[58] Pelleg D, Moore A W. X-means: Extending k-means with efficient estimation of the number of clusters[C]. Proc of the 17th Int Conf on Machine Learning. Stanford, 2000: 727-734.

[59] Frigui H, Krishnapuram R. A robust competitive clustering algorithm with applications in computer vision[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1999, 21(5): 450-465.

[60] Boujemaa N. Generalized competitive clustering for image segmentation[C]. Proc of 19th Int Conf of the North American Fuzzy Information Processing Society (NAFIPS'00). Atlanta, 2000: 133-137.

[61] Cheng Y. Mean shift, mode seeking and clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790-799.

[62] Comaniciu D, Meer P. Mean Shift: A robust approach toward feature space analysis[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619.

[63] Girolami M. Mercer kernel-based clustering in feature space[J]. IEEE Trans on Neural Networks, 2002, 13(3): 780-784.

[64] Guha S, Rastogi R, shim K. CURE: An efficient clustering algorithm for large databases[C]. Proc of the ACM SIGMOD Conf. Washington: ACM Press, 1998: 73-84.

[65] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. J of Machine Learning Research, 2005, 6(6): 363-392.

[66] Agarwal P K, Procopiuc C M, Varadarajan K R. Approximation algorithms for k-line center[C]. Proc of the 10th Annu. European Sympos. Algorithms, 2002: 54-63.

[67] Har-Peled S. No coreset, no cry[C]. Proc of the 24th Foundations of Software Technology and Theoretical Computer Science. Chennai: Springer, 2004.

[68] Har-Peled S, Mazumdar S. Coresets for k-means and k-median clustering and their applications[C]. Proc of the 36th Annual ACM Symposium on Theory of Computing. Chicago, 2004: 291-300.

[69] Badoiu M, Clarkson K L. Optimal core sets for balls[C]. DIMACS Workshop on Computational Geometry. Piscataway, 2002.

[70] 钱鹏江, 王士同, 邓赵红. 基于最小包含球的大样本数据集快速谱聚类算法[J]. 电子学报, 2010, 38(9): 2035-2041.
(Qian P J, Wang S T, Deng Z H. Fast spectral clustering for large data sets using minimal enclosing ball[J]. Acta Electronica Sinica, 2010, 38(9): 2035-2041.)