Journal of Computer Applications ISSN 1001-9081 2019-07-10
计算机应用,2019,39(7):1869-1882 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2019)07-1869-14 DOI:10.11772/j.issn.1001-9081.2019010174

聚类算法综述

章永来,周耀鉴*

(中北大学 软件学院,太原 030051)

(* 通信作者电子邮箱 zhouyj@nuc.edu.cn)

摘 要:大数据时代,聚类这种无监督学习算法的地位尤为突出。近年来,对聚类算法的研究取得了长足的进步。首先,总结了聚类分析的全过程、相似性度量、聚类算法的新分类及其结果的评价等内容,将聚类算法重新划分为大数据聚类与小数据聚类两个大类,并特别对大数据聚类作了较为系统的分析与总结。此外,概述并分析了各类聚类算法的研究进展及其应用概况,并结合研究课题讨论了算法的发展趋势。

关键词:聚类;相似性度量;大数据聚类;小数据聚类;聚类评价

中图分类号:TP301;TP18 **文献标志码**:A

Review of clustering algorithms

ZHANG Yonglai, ZHOU Yaojian*

(*Software School*, *North University of China*, *Taiyuan Shanxi* 030051, *China*)

Abstract: Clustering is very important as an unsupervised learning algorithm in the age of big data. Recently, considerable progress has been made in the analysis of clustering algorithm. Firstly, the whole process of clustering, similarity measurement, new classification of clustering algorithms and evaluation on their results were summarized. Clustering algorithms were divided into two categories: big data clustering and small data clustering, and the systematic analysis and summary of big data clustering were carried out particularly. Moreover, the research progress and application of various clustering algorithms were summarized and analyzed, and the development trend of clustering algorithms was discussed in combination with the research topics.

Key words: clustering; similarity measurement; big data clustering; small data clustering; clustering evaluation

0 引言

把具有相似特性的实物放到一起是人类最原始的活动之一。这也是聚类的最初目的。早在1984年,Aldenderfer等[1]就已经提出了聚类分析的四大功能:一是数据分类的进一步扩展;二是对实体归类的概念性探索;三是通过数据探索而生成假说;四是一种基于实际数据集归类假说的测试方式。在很多情况下,样本数据集并没有分类,即每一个数据样本都没有分类标签。一般而言,聚类指将没有分类标签的数据集,分为若干个簇的过程,是一种无监督的分类方法[2]。实际上,很难对聚类下一个明确的定义。2001年,Everitt等[3]甚至指出提出聚类的正式定义不仅困难而且也没有必要,因为聚类分析本身是一种建立在主观判断基础上的相对行之有效的方法[4-5]。尽管如此,聚类分析还是表达了一般认为的"类内的相似性与类间的排他性"的目标。Hansen等[6]也已经作了数学上的阐述。给定一个数据样本集:

$$C=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_j,\cdots,\boldsymbol{X}_N;\boldsymbol{X}_j=(x_{j1},x_{j2},\cdots,x_{jd})\in\mathbf{R}^d\} \tag{1}$$

这里,$\boldsymbol{X}_j$ 表示一个向量,称为样本点或者样本;X_{jd} 表示一个变量,通常称为属性、特征、变量或维等。划分聚类将数据集分为 K 个簇,需满足:

$$\begin{cases} C=\{C_1,C_2,\cdots,C_K\};\ K\leqslant N \\ C_i\neq\varnothing;\ i=1,2,\cdots,K \\ \bigcup_{i=1}^{K}C_i=X \\ C_i\cap C_j=\varnothing;\ i,j=1,2,\cdots,K,i\neq j \end{cases} \tag{2}$$

而层次聚类是将数据集构建成一种树状的结构,即:

$$\begin{cases} H=\{H_1,H_2,\cdots,H_Q\};\ Q\leqslant N \\ C_i\in H_m,C_j\in H_l;\ m>l \\ C_i\subset C_j \text{ or } C_i\cap C_j=\varnothing;\ m,l=1,2,\cdots,Q,i\neq j \end{cases} \tag{3}$$

聚类分析是伴随着统计学、计算机学与人工智能等领域科学的发展而逐步发展起来的,为此,这些领域若有较大的研究进展,必然促进聚类分析算法的快速发展。比如机器学习领域的人工神经网络与支持向量机的发展就促生了基于神经网络的聚类方法与核聚类方法。目前,基于人工神经网络的深度学习(如:AlphaGo围棋系统)也必将推动聚类分析方法的进一步发展。到目前为止,聚类研究及其应用领域已经非常广泛,因此,本文主要以聚类分析算法为主要分析对象,兼论聚类分析的全过程。

关于聚类分析,《数据挖掘概念与技术(第二版)》一书中已经有了经典的论述。然而,聚类算法又有了长足的发展与

收稿日期:2019-01-23;修回日期:2019-04-09;录用日期:2019-04-10。 基金项目:国家自然科学基金资助项目(6160051296)。

作者简介:章永来(1978—),男,浙江诸暨人,助理教授,博士,主要研究方向:大数据分析与处理、医疗大数据、海洋大数据;周耀鉴(1987—),男,湖北武穴人,助理教授,博士,主要研究方向:大数据分析与处理、海洋大数据、水下机器人。

进步。本文首先简要介绍了聚类分析的主要过程，然后分析并总结了样本点之间的相似性度量方法，提出了聚类算法的新分类方式，并总结与分析了各种聚类算法，还对如何评价聚类结果作了过程分析。最后，依靠课题组承担的医疗与海洋大数据的聚类分析研究[7-11]，展望了聚类算法的发展趋势，作为本文的结语。

1 聚类分析过程

聚类分析是一个较为严密的数据分析过程。聚类分析的全过程如图 1 所示，从聚类对象数据源开始到得到聚类结果的知识存档为止，其中主要包括四个部分研究内容，即特征选择或变换、聚类算法选择或设计、聚类结果评价与聚类结果物理解析等。

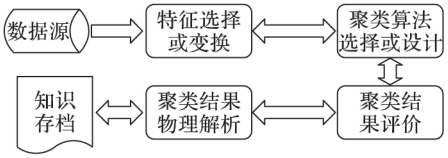

图 1 聚类分析过程

Fig. 1 Clustering analysis process

1.1 特征选择或变换

一般情况下，样本数据是杂乱无章的(特别是大数据时代)，聚类分析首先需要进行数据集的特征选择或变换。实际上，特征选择与特征变换是降维技术的两大分类。特征选择指的是从数据样本集的所有特征(或称属性)中选择更有利于达到某种目标的若干属性，即原始属性集的一个子集，同时也达到了降低维度的目的；而特征变换则是指通过某种变换将原始输入空间的属性映射到一个新的特征空间，然后在特征空间中根据规则选择某些较为重要的变换后的特征。由于特征选择并不改变其原有属性，所以结果只是一个原始属性的优化特征子集，保留了原属性的物理意义，方便用户理解；而特征变换的结果失去了原始特征的物理意义，但能够提取其隐含的特征信息，移除原特征集属性之间的相关性与冗余性。特征选择或变换在聚类分析过程中占据极其重要的地位，结果的优劣将直接影响最后的聚类效果，应该引起足够的重视。有时，特征选择或变换后得到的有效模式(或称子集)的作用甚至超过聚类算法本身的效用。

1.2 聚类算法选择或设计

依据特征选择或变换后的数据集特性，选择或设计聚类算法，是聚类分析的第二部分研究内容。如果样本集数据都是数值型数据，在选择或者设计聚类算法时需要注意量纲不同的问题。一般情况下，样本集数据不一定都是数值型数据，因此，聚类算法需要有处理非数值型数据的能力。各个样本点之间的相似性度量是聚类算法中的首要问题。相似性度量与经常提到的样本间“距离”有着相同的意义，但是，它们的取值却正好相反，即相似性度量值越大，“距离”越近。同样，相似性度量也是聚类分析全过程中的关键问题之一，将在后文进行详细的介绍与分析。

1.3 聚类结果评价与物理解析

聚类簇只能依靠聚类结束准则函数得到[12]，需要特别指出的是，这种准则函数一般由人为设定的终止条件实现，而这些终止条件并没有统一的标准。由此可见聚类分析是一个主观的归类过程，所以在聚类簇生成以后，必须对聚类结果进行综合评价。聚类分析的本来目标是得到特定数据集中隐含的数据结构。更何况，对于同样一个数据集，不同的聚类算法一般会得到不同的聚类簇。然而，对聚类结果作了评价之后，仍然不能改变聚类分析是“通过数据探索而生成假说”的实质，因此，最后需要对聚类结果作物理上的解析。

在聚类结果评价后一段较长的时间内，需要对一种或者几种聚类结果假说，总结出实际的物理意义。聚类簇的物理解析应该与具有实际工作经验的专家作深入的探讨与分析。最后才可以将探讨的结果加入到知识库，作为进一步研究的依据。可见，聚类物理解析并不属于学术研究的范畴，而是一个长期的验证过程。

2 相似性度量

聚类分析是将数据集的相似性样本归为若干类的方法，因此，如何度量样本之间的相似性是聚类算法的关键问题。假设样本间的相似性满足对称性、非负性和反身性，则称样本间的相似性具有可度量性(Metric)。另外，需要注意的是，三角不等式的半度量(SemiMetric)和超度量(UltraMetric)这两种非可度量方式不在本文的探讨范围内。数据集的特征一般分为三种：连续性变量(或称定量型变量)、离散性变量(或称定性型变量)和混合变量。相应的，有三种相似性度量方法。

2.1 连续性变量的相似性度量

1) 欧氏距离(Euclidean Distance)。

这是一种最常用的样本间距离度量方法，计算公式如下：

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sqrt{\sum_{l=1}^{d}(x_{il}-x_{jl})^2} \tag{4}$$

其中：D 表示样本之间的距离；$\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 表示一个向量，或称为样本点或者样本；l 是样本特征的维数；x_{il} 与 x_{jl} 表示一个变量，或称为属性；d 表示样本的总维数，即样本特征的总数量(以下同)。欧氏距离是一种二范数形式，具有在特征空间中转化和旋转的不变性，一般趋向于构建球形聚类簇。然而，属性值相差较大或线性变换都会使相关性产生形变[13-14]。

为了解决这个问题，需要标准化处理目标数据集，使每一个属性对距离的贡献率相同，这也是消除特征之间量纲差异的常规方式。在进行数据分析之前，需要对样本集在均值与方差上作标准化处理[15]。标准化计算公式如下：

$$x_{il}=(x_{il}^{*}-m_l)/S_l \tag{5}$$

其中：m 为均值；S 为方差；* 表示特征的原值(以下同)。另外，为了去掉不同属性值间在量纲上的差别，需要对样本集作正则化处理。例如在 [0,1] 区间内的正则化公式为：

$$x_{il}=\frac{x_{il}^{*}-\min(x_{il}^{*})}{\max(x_{il}^{*})-\min(x_{il}^{*})} \tag{6}$$

2) 切比雪夫距离(Chebyshev Distance)。

在二维空间中，切比雪夫距离的典型应用是解决国际象棋中的国王从一个格子走到另一个格子最少需要几步的问题。这种距离在模糊 C-Means 方法[16-17]中得到了有效应用。切比雪夫距离的公式可以表示为：

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j) = \max_l(|x_{xl} - x_{jl}|) \tag{7}$$

此公式的另外一种表示形式为:

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j) = \lim_{p\to\infty}\sqrt[p]{\sum_{l=1}^{d}(x_{xl} - x_{jl})^2} \tag{8}$$

3) 曼哈顿距离(Manhattan Distance)。

在城市中生活,只能沿着街道从一个地方到另一个地方,为此,人们将生活中熟悉的城市街区距离(City Block Distance)形象地称为曼哈顿距离。该距离的计算公式为:

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j) = \sum_{l=1}^{d}|x_{xl} - x_{jl}| \tag{9}$$

曼哈顿距离在基于自适应谐振理论(Adaptive Resonance Theory, ART)的同步聚类(SYnchronization Clustering, SYC)中有较好的应用;但是,需要注意的是这种距离不再符合在特征空间中转化和旋转的不变性。

4) 闵可夫斯基距离(Minkowski Distance)。

闵可夫斯基距离是一种 p 范数的形式,公式可以表示为:

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j) = \sqrt[p]{\sum_{l=1}^{d}(x_{il} - x_{jl})^p} \tag{10}$$

从式(10)可见:若 p 为无穷大时,这种距离可以称为切比雪夫距离;若 $p = 2$ 时就是欧几里得距离;那么当 $p = 1$ 时,就是曼哈顿距离。

5) 马氏距离(Mahalanobis Distance)。

马氏距离是一种关于协方差矩阵的距离度量表示方法,其公式为:

$$D(\boldsymbol{X}_i,\boldsymbol{X}_j) = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_j)} \tag{11}$$

其中:T 表示转置,$\boldsymbol{S}$ 为样本协方差矩阵。马氏距离的优点是距离与属性的量纲无关,并排除了属性之间的相关性干扰。若各个属性之间独立同分布,则协方差矩阵为单位矩阵。这样,平方马氏距离也就转化为了欧氏距离[18-19]。

6) 对称点距离(Point Symmetry Distance)。

当聚类存在对称模式时,就可以使用对称点距离。其表示公式为:

$$D(\boldsymbol{X}_i,\boldsymbol{X}_r) = \max_{j=1,2,\cdots,N,j\neq i}\frac{\|(\boldsymbol{X}_i - \boldsymbol{X}_r) + (\boldsymbol{X}_j - \boldsymbol{X}_r)\|}{\|(\boldsymbol{X}_i - \boldsymbol{X}_r)\| + \|(\boldsymbol{X}_j - \boldsymbol{X}_r)\|} \tag{12}$$

对称点距离是该点到对称点和其他点距离的最小值。

7) 相关系数(Correlation Coefficient)。

距离度量也可以源于相关系数[20],如皮尔逊相关系数的定义为:

$$\rho_{\boldsymbol{X}_i\boldsymbol{X}_j} = \frac{\mathrm{Cov}(\boldsymbol{X}_i,\boldsymbol{X}_j)}{\sqrt{D(\boldsymbol{X}_i)}\sqrt{D(\boldsymbol{X}_j)}} \tag{13}$$

8) 余弦相似度(Cosine Similarity)。

最后一种直接计算相似性的方法是余弦相似度。其表示形式为:

$$S(\boldsymbol{X}_i,\boldsymbol{X}_j) = \cos\alpha = \frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\|\boldsymbol{X}_i\|\|\boldsymbol{X}_j\|} \tag{14}$$

这里,S 表示样本之间的相似性(以下同)。在特征空间中,两个样本越相似,则它们越趋向于平行,那么它们的余弦值也就越大。

在这 8 类聚类相似度测量方法中,需要注意的是最后三类相似性计算方法不再符合对称性、非负性与反身性的要求,即属于非可度量的范畴。连续性变量的相似性度量方法在不同聚类算法中的应用,如表 1 所示。

表 1　连续性变量相似性度量及其应用

Tab. 1　Similarity measurement of continuous variables and its application

度量方法	可度量性	应用领域
欧氏距离	是	K-Means[21-23]
闵可夫斯基距离	是	模糊 C-Means[24-25]
切比雪夫距离	是	模糊 C-Means
曼哈顿距离	是	模糊 ART[26-27]
马氏距离	是	椭球 ART[28]、超椭球[29]
对称点距离	否	对称 K-Means[30]
皮尔逊相关系数	否	微阵列基因测序[31]
余弦相似度	否	文本聚类[32]

2.2　离散变量的相似性度量

依据特征变量的离散取值的不同,其度量方法可以分为二值变量的相似性度量方法和多值变量的相似性度量方法。

2.2.1　二值变量的相似性度量方法

二值型变量(如性别)是一种常见的特征取值类型,在样本集中这种类型的变量较多。一般情况下,可以将二值型变量使用数字"1"和"0"代替(如男性为 1,女性为 0)。在这种假设前提下,样本点 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 之间所有二值型变量相似性度量的计算方法,如表 2 所示。表中 n 表示样本的二值型变量分别取对应数字值的特征的总个数(如:n_{10} 表示两个样本点中所有二值型变量,第一个是 1,第二个是 0。这样的二值型变量有 n_{10} 个)。

表 2　样本之间二值型变量相似性度量数值

Tab. 2　Similarity measurement values of binary variables between samples

$\boldsymbol{X}_j$	$\boldsymbol{X}_i$		合计
	1	0	
1	n_{11}	n_{10}	$n_{11} + n_{10}$
0	n_{01}	n_{00}	$n_{01} + n_{00}$
合计	$n_{11} + n_{01}$	$n_{10} + n_{00}$	—

根据表 2 可以计算出样本间所有二值型变量取不同数字值的总个数,则相应的相似性度量公式为:

$$\begin{cases} S(\boldsymbol{X}_i,\boldsymbol{X}_j) = \dfrac{n_{11} + n_{00}}{n_{11} + n_{00} + w(n_{10} + n_{01})} \\ D(\boldsymbol{X}_i,\boldsymbol{X}_j) = 1 - S(\boldsymbol{X}_i,\boldsymbol{X}_j) \end{cases} \tag{15}$$

其中 w 为匹配系数,表示变量取不同数值时的差异程度。一般的取值为 1、2 或 1/2。特别的,当 w 为 1 时,这样的距离称为海明距离(Hamming Distance)[33-34]。另外,需要指出的是二值型变量的相似性度量公式应该灵活使用。如数值"0"表示"不存在"时,则 n_{00} 就可以从公式中删除掉;又如根据程度的不同,可以为 n 添加不同的权值来进行计算。当然,这些度量公式的灵活运用,也必须以二值型变量的实际物理意义为依据。

2.2.2　多值变量的相似性度量方法

实际上,多值变量的相似性度量方法可以转化为多个二值变量来计算[35],但是,当变量的取值很大时,简单的转化方法会显得力不从心,因此,Everitt 等[36]提出了更有效的单匹配策略来解决多值变量的相似性度量问题,相应的计算公式为:

$$\begin{cases} S(X_i, X_j) = \dfrac{1}{d}\sum_{l=1}^{d} S_{ijl} \\ D(X_i, X_j) = 1 - S(X_i, X_j) \end{cases};$$

$$S_{ijl} = \begin{cases} 0, & \text{第 } l \text{ 维不相同} \\ w, & \text{第 } l \text{ 维相同} \end{cases} \tag{16}$$

其中: d 表示样本中多值变量的个数,匹配系数 w 一般为 1。需要特别注意两点: 一是若样本中多值变量的取值个数相差较大时,取值较多的变量的 w 值应该设定为比 1 更大的值(由于这样的相同更为不易); 二是若样本中多值变量的取值在量纲上存在较大差异时,可以先进行归一化处理。最后,需要指出的是多值变量的相似性度量结果应该是一个程度上的差别,而不是数值上的差值,为此,在计算多值变量的相似性时,应该事先根据实际物理意义,确定程度上的差别等级(如无差别、轻微、中等、严重等)。

2.3 混合变量的相似性度量

现实世界的真实数据集总是既包含连续性特征又包含离散型特征,那么,如何来度量这种具有混合特征样本间的相似性呢? 实际上,混合变量的相似性度量可以转化为如何结合连续性变量与离散性变量的统一度量问题。真实数据集的相似性度量方法有两种: 一是分别度量连续性变量与离散性变量,然后使用一定的权值相加; 二是使用式(17)计算:

$$\begin{cases} S(X_i, X_j) = \dfrac{1}{d}\sum_{l=1}^{d} S_{ijl} \\ D(X_i, X_j) = 1 - S(X_i, X_j) \end{cases};$$

$$S_{ijl} = \begin{cases} \begin{cases} 0, & \text{若第 } l \text{ 维不相同} \\ w, & \text{若第 } l \text{ 维相同} \end{cases}, & \text{离散型变量} \\ 1 - |x_{il} - x_{jl}|/R_l, & \text{连续型变量} \end{cases};$$

$$R_l = \max_m x_{ml} - \min_m x_{ml} \tag{17}$$

这里,m 是数据集中样本的总个数。另外,Chen 等[37]还提出了一种度量混合变量相似性的新方法。

样本间的相似性度量方法是否合理,将直接影响最终的聚类效果,显得尤为重要。对于一个特定的样本集来说,到底哪种相似性计算方法最合适? 如何解释相似度的物理意义? 这些问题,至今也没有一个确切的答案,因此,聚类分析方法不可避免地具有主观性与对问题域的依赖性等特点。

3 聚类算法分类

聚类算法一般可以用基于划分、基于密度、基于网格和基于约束等方式来进行分类。然而,在大数据时代背景下,随着数据量的不断增加及其数据形态的日益多样化,聚类算法的应用更加广泛; 同时对算法本身也提出了更高的要求。依据有效数据量 10^{12} 字节为阈值,本文将聚类算法分为小数据聚类和大数据聚类两大类[38]。小数据聚类主要体现的是聚类的基本思想,而大数据聚类的思想主要体现在理念、体系结构与架构等几个方面,至于底层聚类的具体实现算法,其实与小数据聚类算法并没有本质上的差别。换言之,大数据聚类的具体实施算法依然采用小数据聚类技术。本文将迄今为止的聚类算法进行了重新划分,并分别综述了小数据聚类和大数据聚类两种类型的算法,将传统的基于划分、基于密度、基于网格等算法统一归类为基于划分的聚类算法。根据数据对象及其生成聚类簇形式的不同,将小数据聚类算法重新分为传统聚类(Traditional Clustering)与智能聚类(Intelligent Clustering)两大类。其中,传统聚类分为: 划分聚类(Partitional Clustering)和层次聚类(Hierarchical Clustering)两个大类; 智能聚类分为: 人工神经网络聚类(Neural network-based Clustering)、核聚类(Kernel-based Clustering)、序列数据聚类(Sequential Data Clustering)、复杂网络聚类(Complex Network Clustering)与智能搜索聚类(Intelligent Search Clustering)等五个大类。将大数据聚类(Big Data Clustering)分为: 并行聚类(Parallel Clustering)、分布式聚类(Distributed Clustering)和高维聚类(High-dimensional Clustering)等三个大类。聚类算法的新分类方式如图 2 所示。

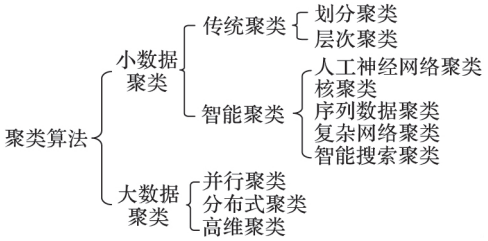

图 2 聚类算法分类

Fig. 2 Classification of clustering algorithms

4 小数据聚类算法

本文将有效数据量在 10^{12} 字节以下的聚类场合,称之为小数据聚类,并将小数据聚类分为传统聚类与智能聚类两个大类。

4.1 传统聚类

本文将传统聚类算法统一分类为划分聚类与层次聚类两大类。

4.1.1 划分聚类

划分聚类算法针对一个包含 n 个样本的数据集,先创建一个初始划分; 然后采用一种迭代的重定位技术,通过样本在类别间移动来改进聚类簇; 最后通过一个聚类准则结束移动并判定结果的好坏。通常情况下的判定准则为平方误差准则:

$$S_E = \sum_{i=1}^{k}\sum_{p\in C_i} |p - m_i|^2 \tag{18}$$

其中: S_E 表示平方误差,p 指样本点,m 指每一个聚类簇的平均值,C 是聚类簇,k 为聚类簇的数目。

划分聚类若从排列组合的角度分析,N 个样本分为 K 类的不同聚类簇[39]为:

$$P(N,K) = \frac{1}{K!}\sum_{m=1}^{K}(-1)^{K-m} C_K^m m^N \tag{19}$$

这种不同聚类簇的数量非常惊人,例如将 30 个样本点分为 3 类的不同划分方式接近 2×10^{14} 种。可见,基于划分的算法不能在整个空间中寻找最优解,必须使用其他方式,其代表算法是 K-Means[40-42]、混合密度聚类、图聚类、模糊聚类等。1967 年提出的 K-Means 算法必须事先为每个类别确定一个聚类中心。该算法是应用较广、较高效的一种聚类方法。然而,算法也存在明显的局限性: 聚类结果的好坏依赖于初始聚类中心的选择; 对异常样本点较为敏感; 只能处理数值型的数据集合等。为克服这些局限性,从 20 世纪 60 年代到现在,许多研究者对 K-Means 算法进行了大量的改进: Bradley 等[43]为克服初始中心的影响提出了一种改进策略; Pelleg 等[44]为了加速迭代过程提出了算法的变体 X-Means 算法; Berkhin

等[45]将 *K*-Means 算法拓展到了分布式聚类领域; Nguyen 等[46]提出的 *K*-MODES 算法,克服了 *K*-Means 算法只能处理数值型数据的缺陷; *K*-MEDOIDS 算法并不是计算求得聚类中心,而是将某个样本直接代表该聚类,这样能有效处理异常数据[47-48]。混合密度聚类算法从概率分布的角度,假设样本集有若干个内在的概率分布,然后利用不同的概率分布来划分聚类簇。这样,聚类过程变成了寻找几个概率分布参数的过程,这也意味着需要人为地构造概率密度函数,因此,这类聚类算法将无监督学习转化成了有监督的分类方法[49]。这些概率分布一般为常用的分布,如高斯分布、t 分布等。最大化似然估计(Maximum Likelihood Estimation)是参数估计中最重要的算法。在这类算法中,最常用的是期望最大化(Expectation Maximization)算法[50-51]。Fraley 等[49]还开发了一个期望最大化算法的软件包。另外,CAST(Cluster Analysis Statistical Test)算法也被认为是一种基于概率模型的聚类算法[52]。图聚类的典型算法为基于最小加权分割的 CLICK(CLustering Identification via Connectivity Kernels)算法[53-54]。该算法正好将样本点与图的顶点、样本点之间的关系(边及边上的权值)建立了图论关联。研究者 Ding 等[55]提出的基于 random walks 的方法也是一种图聚类算法。谱聚类(Spectral Clustering)算法将聚类转化为解二次优化问题,能够识别任意形状的聚类簇并可以收敛于全局最优解[56-57],在图像分析领域有着广泛的应用。Lee 等[58]和 Shi 等[59]又提出了一种具有快捷性和自适应性的 GRC(Graph-based Relaxed Clustering)算法。1981 年模糊聚类方法由 Bezdek 首次实现,该算法就是大名鼎鼎的 FCM(Fuzzy C-Means)算法[60],是图像分割领域应用的较广泛的聚类算法。该算法使用隶属度来确定样本点的相似性,是一种基于目标函数的模糊聚类方法。随后,Bezdek 的研究团队又对模糊目标函数进行了全局性的优化处理[61]。

基于密度的划分聚类方法是一种将数据集看作低密度区域隔开的若干个高密度簇的集合,该方法的主要特点是可以识别任何形状的簇。DBSCAN(Density Based Spatial Clustering of Applications with Noise)算法就是一种基于密度的常用算法,它提出了密度可接近性(Density-reachability)和密度可连接性(Density-connectivity)两个概念[62]。一个簇是基于密度可接近的最大密度可连接的对象集合,除此之外,其他的样本点被认为是异常点。Birant 等[63]在 2007 年又提出了一种新的 ST-DBSCAN(Spatial-Temporal DBSCAN)算法。该算法能在非空间值、空间值和时态值中发现聚类簇[64]。OPTICS(Ordering Points To Identify the Clustering Structure)算法[65]是在一种在参数次序结构上基于密度的聚类算法。数据集中的每个样本点都有两个值: 核心距离(Core-distance)和可接近距离(Reachability-distance)。DENCLUE(DENsity-based CLUstEring)算法[66]在 *K*-Means 和 DBSCAN 的基础上派生出来的一组基于密度分布的聚类算法。STING(STatistical INformation Grid)算法[67]是一种基于网格的多分辨率方法,将空间划分为层次型的矩形单元。WaveCluster 算法[68-69]是一种使用信号处理过程的一种聚类算法。

除了传统的划分聚类算法之外,还出现了一些新的划分聚类算法,如同步聚类、近邻传播聚类、密度峰值快速聚类与大规模数据集聚类等。

同步聚类(SYnchronization CLustering, SYC)是 Shao 等[70]通过对动力学中同步现象的研究提出的基于自适应谐振理论(Adaptive Resonance Theory, ART)的聚类算法。同步聚类算法的主要思路是首先将样本集中的每个样本的每一维分量看作一个相位振子。在最初的振动阶段,各个相位振子(即每一个属性)按照自己的固有本征频率运动; 随着时间的推移,那些比较接近的相位振子会相互影响而产生锁相现象(即以相同的相位作同步运动),最后这种锁相的同步现象影响到了每一个样本点,所有振子(样本点)最终形成了若干个作局部同步运动的聚类簇,在同一个聚类簇的样本点作同步运动的振子相位相同。然而,同步聚类算法的计算量较大,对大规模数据集进行聚类时有相当大的局限性。同步聚类算法基于 Kuramoto 模型:

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \omega_i + \frac{K}{N}\sum_{i=1}^{N}\sin(\theta_i - \theta_j);\ i = 1,2,\cdots,N \tag{20}$$

其中: ω_i 为自然频率(固有频率),θ_i 是第 i 个振荡器的相位,K 是耦合强度。针对 Kuramoto 模型中的耦合强度引入一个分布,认为所有的振荡器相互作用时,耦合是全局的,但每个振荡器的耦合强度 K_i 是不同的,改变了 Kuramoto 模型各个振荡器具有相同的耦合强度的属性,认为耦合强度和固有频率服从一个联合分布,并且二者不相关。该算法巧妙地使用物理模型完成聚类,是结合其他学科理论的重要研究成果,然而其应用并不广泛。

2007 年,近邻传播(Affinity Propagation, AP)聚类(或称仿射传播聚类)是 Frey 等[71]提出的不需要事先设定聚类个数且聚类速度较快的聚类算法。算法通过传递吸引度和归属度两个指标,完成聚类中心的确定[72-73]并且相对比较适合大规模数据集[74]。

2014 年,密度峰值快速聚类(Fast Density Peaks Clustering,FDPC)是意大利研究者 Rodriguez 等[75]提出的一种新的高效聚类算法。实际上,该算法是一种欧氏距离与密度相结合的聚类方法。算法的主要思想是聚类中心具有比邻域高的密度、聚类中心与密度高的点具有相对大的距离。算法的优点是聚类簇数目直观产生、异常点自动发现和剔除、任意形状和任意映射空间维数都可完成聚类。由于该算法需要事先计算任意两个样本之间的距离,故算法的相似度计算开销很大[76]。

大规模数据集聚类是指数据规模在 10^{12} 字节以下,能够在单机上执行的聚类算法。CLARANS(Clustering Large Applications based on RANdomized Sampling)聚类[77]是对 CLARA(Clustering LARge Applications)算法的一次重大的改进。这两种算法本质上都是为了缩小检索空间的一种采样技术。不同的是 CLARA 算法是针对整个搜索过程,而 CLARANS 算法则是针对特定的子图。数据量较大时,上述算法无法将所有样本一次性读入内存。增量聚类,也称在线聚类(Incremental or Online Clustering),是一种不需要将所有样本点都一次性读入内存的方法[78]。

划分聚类算法的应用极其广泛,如无监督的文本聚类、模糊聚类等。

4.1.2　层次聚类

层次聚类具有一个分层的树形结构。按照构建树形结构的方式不同,可以将聚类分为自顶向下和自底向上两种构建

方式，分别称为聚合型层次聚类（Agglomerative Hierarchical Clustering）与分裂型层次聚类（Divisive Hierarchical Clustering）。两种聚类算法都是在聚类过程中构建具有一定亲属关系的系统树图，聚类的大体过程如图 3 所示。

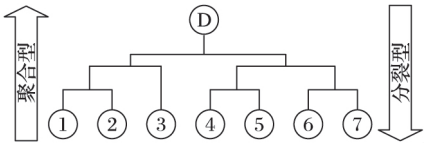

图 3 层次聚类构建示意图

Fig. 3 Schematic diagram of hierarchical clustering construction

聚合型层次聚类，也称自底向上的方法，首先将每一个样本都称为一个聚类簇，然后计算簇间的相似度，分层合并，直到最后只有一个簇为止或满足一定的终止条件；而分裂型层次聚类的迭代过程则正好相反。分裂型层次聚类，也称自顶向下的方法，首先将所有的样本都看作是一个聚类簇，然后在每一步中，上层聚类簇被分裂为下层更小的聚类簇，直到每个簇只包含一个样本，或者满足终止条件为止。由于分裂型算法具有较高的时间复杂度，与聚合型算法相比，分裂型算法并不常见。

最早的层次聚类算法有 AGNES（AGglomerative NESting）算法、DIANA（DIvisive ANAlysis）算法，分别是聚合型与分裂型聚类的早期代表[79]。BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）算法是一种改进的聚合型层次聚类算法[80]。该算法的基本思想是每一个数据点的重要性肯定存在差异性，这样就能够差别化地对待不同的数据样本。BIRCH 算法具有称为 CF-TREE（Clustering Feature Tree）的一种独特的数据结构。算法分为两个步骤：一是扫描所有的样本点，构建内存树；二是利用算法建立树的叶子节点。

一般的聚类算法，都是以单一数据点为聚类的中心，导致最终的聚类簇形状是球形的。为了聚类各种形状的簇，可以采用 CURE（Clustering Using REpresentatives）算法。Guha 等[81]在 1998 年提出了 CURE 算法，该算法的优点是使用具有代表性的一些点来代替聚类簇中的一个中心样本点，这样可以识别各种复杂形状和不同大小的簇，并且具有异常点检测功能。1999 年 Guha 等[82]提出的 ROCK（RObust Clustering using linKs）算法是对 CURE 算法的改进，使之具有识别类别属性的功能。2007 年，Gelbard 等[83]提出了正二进制（Binary-positive）算法，该算法将原始数据转化为二进制位数据，这样样本点之间的相似性度量只在正比特位上进行，直至找到聚类簇为止。

层次聚类算法的主要应用领域包括基因表达谱分析、文本聚类、并行工程组结构等。

4.2 智能聚类

本文将智能聚类算法分为 5 大类：人工神经网络聚类、核聚类、序列数据聚类、复制网络聚类与智能搜索聚类。

4.2.1 人工神经网络聚类

人工神经网络是一种竞争学习方法，由大量神经元（或称处理单元）经相互连接而成的网络，因模拟人类脑神经系统的结构而得名。深度学习也是一种层次较多的人工神经网络。深度学习的飞速发展使人工神经网络研究又上了一个更高的台阶。该聚类方法需要具备三个基本的条件：一是为了使每个输入样本在网络中具有尽量不同的输出，除了一些随机分布的参数外，每一个单元都必须是相同的；二是每个单元具有有限的强度；三是单元间又具有某些竞争性的机制。竞争函数可以表示为：

$$\begin{cases} S_i = \boldsymbol{\omega}_j^{\mathrm{T}} \boldsymbol{X} = \sum_{i=1}^{4} \omega_{ji} x_i \\ Y_i = f(S_i - \theta) \end{cases} \tag{21}$$

其中：$\boldsymbol{\omega}_j$ 是一个权值向量，X 为输入样本，S_i 各个神经元的输入，神经元的输出 Y_i 是输入与此神经元的阈值在传递函数上的映射。人工神经网络通过建立网络模型后，从输入数据中学习知识，来调节神经元的权值向量与阈值，直到网络的输出误差（往往为输出结果与期望结果的范数）达到预期后训练结束。人工神经网络的参数调节方式一般可以分为两种：网格搜索和随机搜索。当然，也可以采用智能算法来搜索参数。

竞争学习方法可分为两种：硬竞争学习和软竞争学习。硬竞争学习机制是用最优的权值向量去匹配输入模式[84]，而软竞争学习是用相似度度量去匹配输入模式。自组织映射（Self-Organizing Map，SOM）[85]就是一种利用人工神经网络进行聚类的算法。该方法将所有的样本点逐一进行处理，并将聚类中心映射到二维空间，从而实现可视化。Yin[86]提出了一种改进的自组织映射算法 VISOM（VIsualization Self-Organizing Map）大幅度提高了传统 SOM 算法的可视化特性。Cao 等[87]提出的基于投影自适应谐振理论的人工神经网络聚类进一步提高了算法的性能。深度学习是人工神经网络的又一次飞跃式的发展。目前，基于深度学习的聚类算法也成为了研究的热点问题[88-90]。

4.2.2 核聚类

伴随着支持向量机[91]的强势推出，从 20 世纪 90 年代以来，基于核函数的方法在机器学习和模式识别领域变得越来越重要[92]。核聚类方法是将样本点从输入空间通过核函数映射到高维空间。这种非线性映射，将不能线性可分的数据集在高维特征空间中变得线性可分，从而在高维空间中利用线性方法完成聚类，这样极大地提高了非线性聚类的性能和可伸缩性；但是这种核聚类方法计算复杂度很高，需要使用 Mecer 理论进行核变换。因为直接寻找非线性映射 Φ 并不容易，所以找到如下的函数：

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \langle \Phi(\boldsymbol{X}_i) \cdot \Phi(\boldsymbol{X}_j) \rangle \tag{22}$$

该式定义的函数 K 就是“核函数（Kernel function）”。核函数的基本作用是通过两个低维空间的向量，计算出经过变换后在高维空间中的向量内积值。支持向量机模型的优化参数有：惩罚因子、核函数的宽度和不敏感参数。一般常常优化前两种参数。支持向量聚类算法是一种典型的无监督学习方法，首先将输入映射到高维空间，然后巧妙地结合高维空间的点在输入空间的位置特性，进行聚类划分。常用的核函数有线性核函数、多项式核函数、高斯核函数与 Sigmoid 核函数。其中，高斯核函数的用途较为广泛。

1995 年，Tax 等[93]提出了支持向量领域描述（Support Vector Domain Description，SVDD）的概念，并初步实现了算法。1999 年，他们进一步改进了支持向量领域描述算法。2004 年，他们又在理论上作了系统的阐述，并将算法的名称改为支持向量数据描述（Support Vector Data Description）算法。该算法是一种全局优化的无监督一分类方法，是一种典型的从数据集中分离出少量异常样本（或称孤立点）的检测

方法。算法将数据集映射到高维空间后,利用一个超球鲁棒地分离了异常样本,并可以通过软间隔动态地控制异常点的数量和映射回输入空间后边界的平滑度。在此基础上,Ben-Hur 等[94-95]提出了支持向量聚类(Support Vector Clustering, SVC)算法。该算法实际上是在分离出异常点后,映射回输入(原始)空间,并且利用任意两个样本点之间连接线上的点在高维空间中的位置特性,进行聚类划分。支持向量聚类算法在手写数字识别等方面的应用中取得了很好的效果。Wang 等[96]将支持向量数据描述称为球形单分类器(Spherical One-Class Classifier, SOCC),并在详细分析了球形单分类器缺点的基础上,提出了一种与聚合型层次聚类相似的结构化单分类器(Structured One-Class Classifier, TOCC)。结构化单分类器首先在特征空间中构建一系列的超椭球,然后用偶内点方法解决这些二次锥规划问题并且检测出孤立点。为了避免求二次锥规划问题,以节省运算时间,2010 年 Rajasegarar 等[97]在结构化单分类器算法的基础上,提出了一种称为中心超椭球支持向量机(Centered-hyperEllipsoidal SVM, CESVM)的单分类算法。该算法只计算一个线性的规划问题,极大地降低了时间复杂度,并应用于传感器网络入侵检测领域取得了较好的效果。Amami 等[98]在增量支持向量机的基础上,又提出了增量支持向量聚类(Incremental Support Vector Clustering)算法,为聚类大规模数据集奠定了基础。

核技巧利用核映射巧妙地解决了非线性问题,应用非常广泛;但是,为了解决二次规划问题或者计算核矩阵,需要大量的计算时间,为此,在处理面向大数据集的核聚类时,也可使用并行计算与云计算等方面的技术。

4.2.3　序列数据聚类

序列数据是指在一定的测度范围内,对某些属性多次测量所得到的数据。序列数据聚类与传统聚类的比较如图 4 所示。

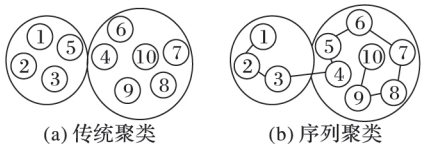

图 4　传统聚类与序列聚类的比较

Fig. 4　Comparison between traditional clustering and sequential clustering

最常用的序列数据是时间序列数据,相应的聚类算法称为时间序列聚类算法[99]。当时间序列数据有先验知识时,可以采用聚类算法直接实现序列划分;当数据没有先验知识时,可以采用聚类算法按数据之间的距离度量划分聚类簇。根据时间序列聚类的执行过程可以将算法大致分成以下三类:基于原始数据的时间序列聚类算法、基于特征的时间序列聚类算法和基于模型的时间序列聚类算法[100]。

轨迹聚类(Trajectory Clustering)也是一种典型的序列数据聚类算法。轨迹聚类算法研究可以追溯到 1994 年[101]。然而直到 1999 年,Gaffney 等[102]提出了轨迹数据的定义及其轨迹聚类的混合回归模型。轨迹聚类将原来孤立点状分布的样本点,扩展到在时间轴上的队列数据。国内的研究者习惯上把轨迹聚类形象地称为时空轨迹聚类。它不仅表示了研究对象的历史模式,同时也隐含了对象的未来模式,具有积极的预测意义。在大数据的背景下,该聚类算法将会是目前及其以后长期的研究热点[103]。

序列数据聚类往往通过子序列的距离或者相似度来生成聚类簇,但它和传统的聚类方法有本质的区别,如在对比基因和蛋白质序列时,聚类过程是通过一系列的替换、插入、删除(或者称编辑)等操作来完成。随着信息技术的发展,近几十年来在基因测序、速度过程检测、文本挖掘、医疗诊断、股票市场分析、顾客交易、网页数据挖掘和机器感知分析等领域出现了大量的序列数据,因此,对这些数据的聚类是非监督学习领域的新挑战。目前,对于序列数据聚类研究主要集中在三个方面:一是对序列数据的距离和相似性度量研究;二是通过特征提取将序列数据转换为熟悉的无序列样本数据;三是直接围绕序列数据的特性建立数学模型。

4.2.4　复杂网络聚类

聚类分析是机器学习领域发展最快的研究方向之一。复杂网络聚类(Complex Network Clustering)与智能方法聚类的发展日新月异,已经成为聚类算法分类方式的新成员。复杂网络聚类是伴随着互联网的飞速发展及其巨大影响力而提出的一类领域性较强的专用聚类方法[104-105]。复杂网络自身具有无标度、小世界和分形的特点[106-107]。复杂网络聚类可以分为基于优化的方法(optimization based method)和启发式方法(heuristic method)两种[69],例如,利用启发式规则的 Girvan-Newman 算法[108]、郭玉泉等[109]提出的面向社区结构的分形聚类检测算法。

复杂网络聚类的主要应用领域是流行病传播网、互联网上的社交网络、现实世界的社会系统、生物系统联系网等。

4.2.5　智能搜索聚类

智能搜索聚类(Intelligent search clustering),是指运用智能方法搜索解空间的启发式聚类算法。基于划分的聚类方法本身是一个 NP 难的问题,而且搜索空间大小随着样本点的增加以指数级的方式增长,以智能搜索解空间为研究路径,研究者提出了复杂的智能方法聚类算法。如为了弥补 K-Means 聚类算法容易陷入局部最优解的问题,可以结合进化算法,模拟出多个种群,并可以根据环境的不同,动态改变其交叉变异概率,增强种群的多样性。这样既解决了容易陷入局部最优的问题,又保留了进化算法全局最优的收敛性。

一般情况是,研究者提出一种新型智能搜索算法,就会被很快应用于某种聚类算法的应用领域,并对应产生一种新的智能方法聚类算法。代表性的智能搜索算法有:进化算法[110](包括遗传算法及结合量子理论的进化算法)、退火算法[111]、Tabu 搜索[112]、粒子群算法[113]及其他生物觅食算法(如蜂群算法、鱼群算法、蛙跳算法、萤火虫算法)等。2014 年 Mirjalili 等[114]又提出了一种模仿狼群狩猎的新智能搜索算法,称为灰狼算法。该算法在多参数的组合优化问题上表现出良好的性能[115]。

5　大数据聚类算法

IBM 和国际数据公司(International Data Corporation, IDC)分别提出了大数据的 4V 特点,综合来看大数据具有 5V 特性,即数据量庞大(Volume)、多样性或称异构数据(Variety)、实时性(Velocity)、真实性(Veracity)与大价值(Value)。本文将有效数据量在 10^{12} 字节以上的聚类技术,称之为大数据聚类。大数据聚类的核心思想是处理计算复杂度和计算成本,与可扩展性和速度之间的关系问题,因此,大数

据聚类算法关注的焦点是:以最小化地降低聚类质量为代价,提高算法的可扩展性与执行速度。本文将大数据聚类分为分布式聚类(Distributed Clustering)、并行聚类(Parallel Clustering)和高维聚类(High-dimensional Clustering)等三个类别。其中并行聚类与分布式聚类算法,需要在计算机集群中执行,因此,这两种算法合称为多机聚类(Multi-machine Clustering)。多机聚类的硬件架构如图5所示。

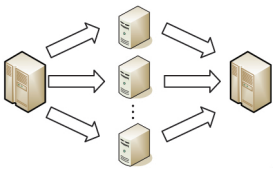

图5 多机聚类的硬件架构

Fig. 5 Hardware architecture of multi-machine clustering

从图5可知,多机聚类将数据划分到多台机器分开执行聚类的过程。若划分与处理过程是人为干预下执行的聚类,称为并行聚类;若划分与处理过程是由分布式框架自动执行的聚类,称为分布式聚类。因此,分布式聚类的执行框架对用户来说,既隐藏了负载均衡、出错控制与计算资源的分配等网络问题,又自动执行数据划分、信息交换等数据处理问题,体现了"计算向数据靠拢"的执行理念;而并行聚类则在处理网络与处理数据两个方面都需要人工干预,需要消耗大量的时间与精力,执行聚类的难度很大,体现的是"数据向计算靠拢"的执行理念。多机聚类算法的核心问题在于,在多台机器之间尽可能少地交换信息的情况下,获得较好的聚类效果。在多机聚类过程中,影响聚类效果的因素主要体现在聚类标准和数据划分两个方面:其一,不同的机器可能使用不同的基本聚类算法;其二,即使所有机器强制使用同一种基本聚类算法,由于数据划分的不用,聚类效果可能也会大相径庭。多机聚类算法的执行过程如图6所示。首先,划分数据到不同的机器,然后执行分组聚类;第二,综合并分析分组聚类的结果;第三,依据分析的结果,自动改进聚类过程;第四,重新进行分组聚类;依次循环执行,直到符合判定准则或者满足终止条件。可见,多机聚类算法是波浪式、循环、不断前进地构造聚类簇的过程。

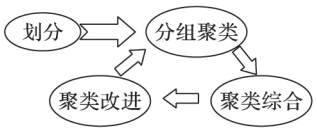

图6 多机聚类算法的执行过程

Fig. 6 Run process of multi-machine clustering

5.1 分布式聚类

负载均衡与数据交换等问题,由分布式框架自动执行的聚类方式,称为分布式聚类。对于分布式聚类算法的评价,一般可以从三个方面入手:一是执行速度,即数据量不变的情况下,随着机器数量的增加,执行时间的变化率;二是可伸缩性,即执行时间不变的情况下,机器数量的增加能够处理数据量的容忍度;三是数据吞吐量,即机器数据不变的情况下,随着执行时间的增加,能够处理的数据规模。目前,从分布式聚类的研究文献来看,主要是使用MapReduce框架执行聚类。MapReduce框架的工作流程如图7所示。对于编程人员来说,只需要编写Map函数与Reduce函数,该框架会自动执行比较复杂的信息交换(Shuffle)过程。

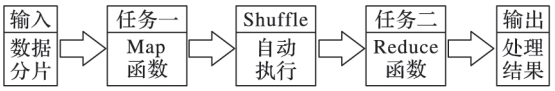

图7 MapReduce的工作流程

Fig. 7 Work flow of MapReduce

PK-Means(Parallel K-Means based on MapReduce)算法是较简单也是经典的K-Means算法在MapReduce大数据平台上的应用。PK-Means聚类算法在执行时间与可伸缩性两个方面都有线性的提高[116]。MR-DBSCAN(DBSCAN based on MapReduce)算法在可伸缩性与数据吞吐量两个方面的性能也有了长足的进步[117]。另外,为了提高大数据集的执行效率,基于GPU的MapReduce聚类[118]与基于SPARK框架的聚类方法[119],在较新文献中也已经崭露头角。

5.2 并行聚类

相对于分布式聚类来说,并行聚类算法的实现较为困难,但是,并行聚类最大的优势是执行过程尽在编程人员的掌握之中。DBDC(Density Based Distributed Clustering)算法是DBSCAN算法在大数据聚类中的典型应用,但是在聚类效果相同的前提下,执行效率比DBSCAN提高了30倍[2]。图划分方法是一种常用的聚类技术。类似的,ParMETIS(Parallel METIS)算法是METIS算法的大数据并行执行版本[120],是一种面向大数据的层次型并行聚类技术。该并行聚类算法也能在不损失聚类效果的情况下,极大地提高执行效率。近几年,聚类研究领域出现了基于GPU(Graphic Processing Unit)的并行聚类算法。例如,G-DBSCAN(Graphic DBSCAN)算法与G-OPTICS(Graphic OPTICS)算法。G-DBSCAN聚类算法,首先以距离为标准,构建样本集图,然后在图上执行聚类,结果比基于CPU的方法聚类速度提高了112倍[121]。基于GPU的G-OPTICS算法显示了更好的聚类效果,比基于CPU的方法聚类速度提高了200倍[122]。可见,基于GPU的聚类技术将是未来较为热门的一种大数据聚类方法。

5.3 高维聚类

样本点维度极高的聚类算法,称为高维聚类算法。当聚类高维数据时,传统的方法是先降低维度(或称为维度约简(Dimension Reduction),简称降维)。降维算法种类很多,不在这里作过多介绍。需要注意的是该方法会不可避免地带来数据信息损失,可能也会降低聚类的有效性;并且,降维算法不适合处理甚高维的情况。直接对甚高维数据集进行传统聚类,几乎很难找到聚类簇,因此可以将数据集划分为若干个子空间,这样,原高维空间就是子空间的并集,为此,处理高维数据集的方法可以命名为子空间聚类(Subspace Clustering)算法[123-124]。该种聚类算法可以分为硬子空间聚类(Hard Subspace Clustering)和软子空间聚类(Soft Subspace Clustering)。硬子空间聚类算法又分为自底向上和自顶向下两种,如CLIQUE(CLustering In QUEst)[125]、ENCLUS(ENtropy-based CLUStering)[126]、MAFIA(Merging of Adaptive Finite IntervAls)[127]等算法为自底向上的类型;ORCLUS(arbitrarily ORiented projected CLUster generation)[128]、FINDIT(a Fast and INtelligent subspace clustering algorithm using DImention voTing)[129]等算法为自顶向下的类型。针对硬子

空间聚类算法的研究相对比较成熟,Sim 等[130]已经作了相关的详细阐述。软子空间算法主要是研究特征加权聚类[131-132],Deng 等[133]已经作了深入的阐述。

对于高维聚类,还有一种聚类算法值得注意。那就是双聚类(Biclustering)算法。随着对基因数据的深入分析,出现了众多的基因表达数据的情况,产生了双聚类算法。这些数据大多以矩阵形式表示和存储,并且维度极高(有时可能达到几千万维)。基因芯片数据中隐藏了大量有用的局部模式,为寻找这些信息,2000 年 Cheng 等[134]提出了双聚类的概念: 双聚类的目的是在基因表达数据矩阵中寻找满足条件的子矩阵,使子矩阵中基因集在对应的条件集上一致表达。目前主要的几类双聚类模型为矩阵等值模型、矩阵加法模型、矩阵乘法模型和信息共演变模型。值得注意的是类似的子空间聚类方法的发展速度很快,也是目前极高维聚类的一个研究热点。

6　算法对比分析

划分聚类的应用较为广泛,收敛速度也很快,且能够比较容易地扩展,以用于大规模的数据; 缺点在于它倾向于聚集样本数量大小比较接近、以密度作为主要的区别手段来进行聚类分析,并且初始聚类中心的选择和孤立点(或噪声)数据会对聚类分析产生较为明显的影响。由于划分聚类一般从总体上评判样本间的相似性,故不适合聚类高维数据集。

层次聚类方法较好地克服了划分聚类的一些缺陷,适用于识别各种形状的聚类簇; 由于聚类的层次性,也可以控制不同层次水平的聚类簇,以适应于不同粒度的应用场合,同时以牺牲算法的时间复杂度作为代价。

人工神经网络聚类是一种基于机器学习的聚类算法,主要优点: 一是聚类的灵活性,可以通过调节参数,理论上可以无限度地减少误差; 二是聚类运算的并行性,可以根据数据集维度的不同,调节输入节点的个数,并行计算; 三是算法易于实现。同时算法主要缺点是容易出现过学习问题和聚类结果不稳定两个方面。

核聚类算法主要用于非线性可分的数据集。这种数据集无法使用线性可分的聚类算法解决。该算法也巧妙地解决了 Bellman 教授提出的处理高维数据集时的“维度诅咒”问题。另外,核聚类算法能有效地处理噪声与离群点的问题,并不需要先验知识的指导。然而,核函数的选择、特征空间与原始空间的关系、大数据样本集的核矩阵存储与计算等问题也亟待解决。

序列数据聚类是针对特定的数据特征而提出的领域聚类分析算法,是在时间与空间的可测度范围内,对某些属性多次测量所得到的数据,进行聚类的过程。

大数据聚类主要着眼于超大规模数据集(10^{12}字节以上)与高维数据集两个角度。超大规模数据集聚类算法与其他算法的主要区别是在理念、体系结构与架构方面。至于聚类的具体实施过程,一般更倾向于采用较为简单的基本算法(如 K-Means)。

表 3 选取了聚类算法的时间与空间计算复杂度、能否处理纵向的大规模数据集和横向的高维数据集等最受关注的三个方面,总结比较了一些常用的典型算法的性能。

表 3　典型聚类算法的性能对比

Tab. 3　Performance comparison of typical clustering algorithms

算法	计算复杂度	大规模	高维
K-Means	$O(NKd)$(时间),$O(N+K)$(空间)	是	否
CURE	$O(N^2 \log N)$(时间),$O(N)$(空间)	是	是
DBSCAN	$O(N \log N)$(时间)	是	否
WaveCluster	$O(N)$(时间)	是	否
CLARANS	$O(N^2)$(时间与空间)	是	否
FCM	接近 $O(N)$(时间)	否	否
BIRCH	$O(N^2 \log N)$(时间)	否	否
DENCLUE	$O(N \log N)$(时间)	是	是
SVC	$O((N-n_{bsv})\,n_{sv}{}^2)$(时间)	是	是
SYC	$O(TN^2)$(时间)	否	否
FDPC	$O(N)$(时间)	是	是
G-OPTICS	$O(E \log N)$(时间)	是	否

注: N 表示数据的规模; K 表示聚类簇的数目; n_{sv} 表示支持向量的个数; n_{bsv} 表示边界支持向量的个数; E 表示样本图中边的数量。

7　聚类结果评价

聚类算法是一种无监督的主观的分类方法。聚类结果的客观分析、聚类簇数目的合理性、聚类结构的物理意义等内容,都属于聚类评价的范畴[135]。聚类结果的评价在很多文献中已经从有效性函数方面作了大量的研究,本文主要是从聚类分析的三大类准则的角度作简要分析,这些准则分别是外部准则(external criteria)、内部准则(internal criteria)与相对准则(relative criteria)[136]。

外部准则是一种利用聚类结构先验知识的聚类结果评价方法。例如,在聚类评估时可以借助专家的指导(或者采用人工的标准方法)来开展聚类结果分析。最后,通过假设检验的方法比较算法结果与人工的经验结果的一致性,从而确定聚类算法的有效性。

内部准则也是一种假设检验方法,只是并不是利用先验知识,而是利用样本集的内部特性来评估聚类结果的有效性。例如,内部准则通过样本集的相似矩阵、CPCC(CoPhenetic Correlation Coefficient)指标等来评估聚类结果。相对准则通过不同算法(或者相同算法选择不同输入参数)的聚类结果对比来综合评估结果的一种准则,与外部准则和内部准则有很大的不同之处。实际上,相对准则依据聚类分析四大功能之一的“实际数据集上的其他技术归类假说的测试方式”而产生。该准则并不使用假设检验方法,故其计算量相对较小。

外部准则较客观,相对准则较不客观,而内部准则居于两者之间。可见,在聚类结果的评价过程中,能使用外部准则最好,其次是内部准则,最后才选择相对准则。

这里需要特别说明的是: 由于确定聚类簇的数量比较困难(有关聚类数目的确定可参见 José-García 等[137]的深入分析),而且聚类数量评估是结果评估的中心问题[138],相对准则主要用于聚类簇数目的评估; 相对准则也有很多有效性指标、停止规则、可视化方法、启发式方法等[135]来评估聚类簇的数目。

聚类结果的评价是聚类分析中的一个重要步骤,也是最

困难、最无所适从的一步。聚类分析全过程中的最后一步是聚类结果的物理解析。该步骤是在结果评价后的长期实践中逐步探索形成知识的过程，这里不再详细叙述。

8 结语

课题组在聚类分析医疗与海洋大数据的需求背景下，本文是聚类项目的研究成果。现将聚类算法的展望如下：

1）聚类分析不仅仅是选择或者设计聚类算法的过程，数据预处理与特征提取其实非常重要。针对实际的数据源，数据预处理时有大量细致入微的工作需要去完成，特征提取的数据集质量的优劣也会直接影响最后的聚类结果。课题组建议预处理与特征提取尽量有领域专家的指导，而且这一过程可能需要占用大量的时间与精力。

2）在实际应用中，数据集会存在复杂性和多样性的特点，选择任何一种聚类算法可能都不一定适用，因此，需要在了解基本聚类算法的优缺点基础上，研究多种算法的融合问题。

3）在聚类实际数据集时，能够产生任何形状聚类簇的算法将会是聚类算法研究的发展方向，因此，先使用核变换将数据集映射到高维空间，再利用传统聚类算法的核聚类算法是一个重要的研究方向。

4）在大数据时代背景下，大数据聚类算法研究会有较好的发展前景，例如基于 GPU 的聚类、基于 SPARK 的聚类、基于图计算框架（如 Pregel）的聚类等新的聚类理念与技术。

5）随着云计算、物联网与大数据技术的应用日趋成熟，在这种情况下，待聚类的数据复杂性前所未有，因此，在复杂数据背景下，如何探讨聚类的有效性也是一个重要的、有难度的研究热点方向。

6）另外，在查阅外文文献的过程中，课题组研究人员发现，对于同一个概念，在不同的文献中，可能会出现不同的称谓。如果将数据集看作一个由行和列构成的二维表，不同的外文文献会将行称为 Samples、Data Objects、Patterns、Entities、Cases、Instances、Observances、Units 等；将列称为 Dimensions、Variables、Features、Properties 等；将聚类簇称为 Clusters、Subsets、Groups、Categories 等。

参考文献（References）

[1] ALDENDERFER M S, BLASHFIELD R K. Cluster Analysis [M]. Los Angeles: Sage Publications, 1984: 2 - 12.

[2] AGGARWAL C C, REDDY C K. Data Clustering: Algorithms and Applications [M]. London: Taylor and Francis Group, 2014: 4 - 7.

[3] EVERITT B, LANDAU S, LEESE M. Cluster Analysis [M]. 4th ed. London: Arnold, 2001: 144 - 201.

[4] BARALDI A, ALPAYDIN E. Constructive feedforward ART clustering networks—Part I and II [J]. IEEE Transactions on Neural Networks, 2002, 13(3): 645 - 677.

[5] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review [J]. ACM Computing Surveys, 1999, 31(3): 264 - 323.

[6] HANSEN P, JAUMARD B. Cluster analysis and mathematical programming [J]. Mathematical Programming, 1997, 79: 191 - 215.

[7] 章永来. 基于聚类的社区居民健康指数预测模型研究 [D]. 北京: 中国科学院大学, 2015. (ZHANG Y L. Research on prediction model of health index for community residents based on clustering [D]. Beijing: University of Chinese Academy of Sciences, 2015.)

[8] ZHOU X, ZHANG Y, SHI M, et al. Early detection of liver disease using data visualisation and classification method [J]. Biomedical Signal Processing and Control, 2014, 11: 27 - 35.

[9] ZHANG Y, ZHOU X, SHI H, et al. Corrosion pitting damage detection of rolling bearings using data mining techniques [J]. International Journal of Modeling, Identification and Control, 2015, 24(3): 235 - 243.

[10] ZHOU Y, YU J, WANG X. Time series prediction methods for depth-averaged current velocities of underwater gliders [J]. IEEE Access, 2017, 5: 5773 - 5784.

[11] 章永来，史海波，尚文利，等. 面向乳腺癌辅助诊断的改进支持向量机方法 [J]. 计算机应用研究, 2013, 30(8): 2373 - 2376. (ZHANG Y L, SHI H B, SHANG W L, et al. Improved method for computer-aided diagnosis of breast cancer based on support vector machines [J]. Application Research of Computers, 2013, 30(8): 2373 - 2376.)

[12] KLEINBERG J. An impossibility theorem for clustering [C] // Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2002: 463 - 470.

[13] DUDA R O, HART P E, STORK D G. Pattern Classification [M]. 2nd ed. New York: John Wiley and Sons, 2001: 47 - 56.

[14] GAO J, WANG Y, LI J. Bounds on covering radius of linear codes with Chinese Euclidean distance over the finite non chain ring F - 2 + vF(2) [J]. Information Processing Letters, 2018, 138: 22 - 26.

[15] HOGG R, TANIS E. Probability and Statistical Inference [M]. 7th ed. Upper Saddle River: Prentice Hall, 2005: 120 - 145.

[16] BOBROWSKIL, BEZDEK J C. C-means clustering with the L_1 and L_∞ norms [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 545 - 554.

[17] ANTER A, HASSENIAN A E, OLIVA D. An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural [J]. Expert Systems with Applications, 2019, 118: 340 - 354.

[18] MAO J, JAIN A K. A self-organizing network for HyperEllipsoidal Clustering (HEC) [J]. IEEE Transactions on Neural Networks, 1996, 7(1): 16 - 29.

[19] ZHAN J, WANG R, YI L. Health assessment methods for wind turbines based on power prediction and Mahalanobis distance [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(2): 1951001.

[20] KAUFMAN L, ROUSSEEUW P J. Finding Groups in Data: An Introduction to Cluster Analysis [M]. New York: John Wiley and Sons, 2009: 82 - 85.

[21] XU R, DONALD C W. Clustering [M]. New York: John Wiley and Sons, 2009: 12 - 95.

[22] FORGY E W. Cluster analysis of multivariate data: efficiency vs. interpretability of classification [J]. International Journal of Environmental Studies, 1965, 21(3): 41 - 52.

[23] ANTOINE G B, CATHY M R, ANDREA R. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data [J]. Journal of Applied Statistics, 2019, 46(1): 47 - 65.

[24] HATHAWAY R J, BEZDEK J C, HU Y. Generalized fuzzy c-means clustering strategies using L_p norm distances [J]. IEEE Transactions on Fuzzy Systems, 2000, 8(5): 576 - 582.

[25] 耿宗科,王长宾,张振国. 基于模糊 c-means 与自适应粒子群优化的模糊聚类算法 [J]. 计算机科学,2016,43(8):267 - 272. (GENG Z K, WANG C B, ZHANG Z G. Fuzzy c-means and adaptive PSO based fuzzy clustering algorithm [J]. Computer Science, 2016, 43(8): 267 - 272.)

[26] CARPENTER G A, GROSSBERG S, ROSEN D B. Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system [J]. Neural Networks, 1991, 4(6): 759 - 771.

[27] CHANDRAPRABHA K, GEETHA B G. Wireless network confidence level improvement via fusion adaptive resonance theory [J]. Cluster Computing, 2018(2): 1 - 11.

[28] ANAGNOSTOPOULOS G C, GEORGIOPOULOS M. Ellipsoid ART and ARTMAP for incremental unsupervised and supervised learning [C] // Proceedings of the 2001 International Society of Optical Engineering. Bellingham: SPIE Publications, 2001: 1 - 6.

[29] MOSHTAGHI M, RAJASEGARAR S, LECKIE C, et al. An efficient hyperellipsoidal clustering algorithm for resource-constrained environments [J]. Pattern Recognition, 2011, 44(9): 2197 - 2209.

[30] SU M C, CHOU C H. A modified version of the K-Means algorithm with a distance based on cluster symmetry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 674 - 680.

[31] EISEN M B, SPELLMAN P T, BROWN P O, et al. Cluster analysis and display of genome-wide expression patterns [J]. Proceedings of the National Academy of Sciences of the United States of America, 1998, 95(25): 14863 - 14868.

[32] STEINBACH M, KARYPIS G, KUMAR V. A Comparison of Document Clustering Techniques [M]. New York: John Wiley and Sons, 2000: 50 - 56.

[33] GERSHO A, GRAY R M. Vector quantization and signal compression [J]. Springer International, 1992, 159(1): 407 - 485.

[34] ABDEL-GHAFFAR K A S. Sets of binary sequences with small total Hamming distances [J]. Information Processing Letters, 2019, 142: 27 - 29.

[35] GETZ G, LEVINE E, DOMANY E. Coupled two-way clustering analysis of gene microarray data [J]. Proceedings of the National Academy of Sciences, 2000, 97(22): 12079 - 12084.

[36] EVERITT B, HOTHORN T. Cluster Analysis [M]. New York: John Wiley and Sons, 2011: 111 - 134.

[37] CHEN J Y, HE H H. A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data [J]. Information Sciences, 2016, 345: 271 - 293.

[38] SHIRKHORSHIDI A S, AGHABOZORGI S, WAH T Y, et al. Big data clustering: a review [C] // Proceedings of the 14th International Conference on Computational Science and Its Applications. Berlin: Springer, 2014: 707 - 720.

[39] KALYANI P. Approaches to partition medical data using clustering algorithms [J]. International Journal of Computer Applications, 2013, 49(23): 7 - 10.

[40] LIU G. Introduction to Combinatorial Mathematics [M]. New York: McGraw Hill, 1968: 12 - 16.

[41] KRISHNA K, MURTY M N. Genetic K-means algorithm [J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 1999, 29(3): 433 - 439.

[42] LU Y, LU S, FOTOUHI F. FGKA: a fast genetic K-means clustering algorithm [C] // Proceedings of the 2004 ACM Symposium on Applied Computing. New York: ACM, 2004: 622 - 623.

[43] BRADLEY P, FAYYAD U. Refining initial points for K-means clustering [C] // Proceedings of the 15th International Conference on Machine Learning. New York: ACM, 1998: 91 - 99.

[44] PELLEG D, MOORE A. X-means: extending K-means with efficient estimation of the number of the clusters [C] // Proceedings of the 17th International Conference on Machine Learning. New York: ACM, 2000: 111 - 117.

[45] BERKHIN P, BECHER J. Learning simple relations: theory and applications [C] // Proceedings of the 2nd International Conference on Data Mining, Washington, DC: IEEE Computer Society, 2002: 333 - 349.

[46] NGUYEN H H. Privacy-preserving mechanisms for K-modes clustering [J]. Computers and Security, 2018, 78: 60 - 75.

[47] LACKO D, HUYSMANST, VLEUGELS J, et al. Product sizing with 3D anthropometry and k-medoids clustering [J]. Computer-Aided Design, 2017, 91: 60 - 74.

[48] NAKAGAWA K, IMAMURA M, YOSHIDA K. Stock price prediction using k-medoids clustering with indexing dynamic time warping [J]. Electronics and Communications in Japan, 2019, 102(2): 3 - 8.

[49] FRALEY C, RAFTERY A. Model-based clustering, discriminant analysis, and density estimation [J]. Journal of the American Statistical Association, 2002, 97: 611 - 631.

[50] McLACHLAN G, KRISHNAN T. The EM Algorithm and Extensions [M]. New York: John Wiley and Sons, 1997: 6 - 9.

[51] ZHOU Y, XU S, JIN C, et al. Multiple point sets registration based on expectation maximization algorithm [J]. Computers and Electrical Engineering, 2018, 70: 1 - 11.

[52] BEN-DOR A, SHAMIR R, YAKHINI Z. Clustering gene expression patterns [J]. Journal of Computational Biology, 1999, 6: 281 - 297.

[53] SHARAN R, SHAMIR R. CLICK: A clustering algorithm with applications to gene expression analysis [C] // Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. San Diego: [s. n.], 2000: 307 - 316.

[54] NGUYEN M N, SIM A Y L, WAN Y, et al. Topology independent comparison of RNA 3D structures using the CLICK algorithm [J]. Nucleic Acids Research, 2017, 45(1): e5.

[55] DING C F, LI K. Centrality ranking in multiplex networks using topologically biased random walks [J]. Neuro-computing, 2018,

312: 263 - 275.

[56] SHANG R, ZHANG Z, JIAO L, et al. Global discriminative-based nonnegative spectral clustering [J]. Pattern Recognition, 2016, 55: 172 - 182.

[57] 宋健,许国艳,夭荣朋. 基于差分隐私的数据匿名化隐私保护方法 [J]. 计算机应用,2016,36(10):2753 - 2757. (SONG J, XU G Y, YAO R P. Spectral clustering algorithm based on differential privacy protection [J]. Journal of Computer Applications, 2016, 36(10): 2753 - 2757.)

[58] 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.

[59] SHI D, WANG J, CHENG D, et al. A global-local affinity matrix model via EigenGap for graph-based subspace clustering [J]. Pattern Recognition Letters, 2017, 89: 67 - 72.

[60] BEZDEK J. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. New York: Plenum Press, 1981: 37 - 89.

[61] HATHAWAY R, BEZDEK J. Fuzzy c-means clustering of incomplete data [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2001, 31(5): 735 - 744.

[62] ESTER M, KRIEGEL H, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1996: 56 - 69.

[63] BIRANT D, KUT A. ST-DBSCAN: an algorithm for clustering spatial-temporal data [J]. Data and Knowledge Engineering, 2007, 60(1): 208 - 221.

[64] TRISMININGSIH R, SHAZTIKA S S. ST-DBSCAN clustering module in SpagoBI for hotspots distribution in Indonesia [C] // Proceedings of the 3rd International Conference on Information Technology. New York: ACM, 2017: 60 - 67.

[65] ANKERST M, BREUNIG M, KRIEGEL H, et al. OPTICS: ordering points to identify the clustering structure [C] // Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1999: 49 - 60.

[66] HINNEBURG A, KEIM D. An efficient approach to clustering in large multimedia databases with noise [C] // Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1998: 58 - 65.

[67] WANG W, YANG J, MUNTZ R. STING: a statistical information grid approach to spatial data mining [C] // Proceedings of the 23rd Conference on Very Large Data Bases. New York: ACM, 1997: 186 - 195.

[68] SHEIKHOIESIAMI G, CHATTERJEE S, ZHANG A. WaveCluster: a multi-resolution clustering approach for very large spatial databases [C] // Proceedings of the 24th Conference on Very Large Data Bases. New York: ACM, 1998: 428 - 439.

[69] YILDIRIM A A, WATSON D. A comparative study of the parallel wavelet-based clustering algorithm on three-dimensional dataset [J]. Journal of Supercomputing, 2015, 71(7): 2365 - 2380.

[70] SHAO J, HE X, BOHM C, et al. Synchronization-inspired partitioning and hierarchical clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 893 - 905.

[71] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972 - 976.

[72] 刘晓楠,尹美娟,李明涛,等. 面向大规模数据的分层近邻传播聚类算法 [J]. 计算机科学,2014,41(3):185 - 188. (LIU X N, YIN M J, LI M T, et al. Hierarchical affinity propagation clustering for large-scale data set [J]. Computer Science, 2014, 41(3): 185 - 188.)

[73] AKASH O M, AZMI M S B. A new similarity measure based affinity propagation for data clustering [J]. Advanced Science Letters, 2018, 24(2): 1130 - 1133.

[74] ZHAO J L, QU H, ZHAO J H. Towards controller placement problem for software-defined network using affinity propagation [J]. Electronics Letters, 2017, 53(14): 928 - 929.

[75] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492 - 1496.

[76] XIAO X, DING S, SUN T. A fast density peaks clustering algorithm based on pre-screening [C] // Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing. Piscataway, NJ: IEEE, 2018: 456 - 462.

[77] GHOSH S, MITRA S. Clustering large data with uncertainty [J]. Applied Soft Computing, 2013, 13(4): 1639 - 1645.

[78] LE H S, NGUYEN D T. Tune up fuzzy c-means for big data: some novel hybrid clustering algorithms based on initial selection and incremental clustering [J]. International Journal of Fuzzy Systems, 2017, 19(5): 1585 - 1602.

[79] LIU A, SU Y, NIE W, et al. Hierarchical clustering multi-task learning for joint human action grouping and recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 102 - 114.

[80] MADAN S, DANA K J. Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering [J]. Pattern Analysis and Applications, 2016, 19(4): 1023 - 1040.

[81] GUHA S, RASTOGI R, SHIM K, et al. CURE: an efficient clustering algorithm for large databases [J]. Information Systems, 1998, 26(1): 35 - 58.

[82] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes [J]. Information Systems, 1999, 25(5): 345 - 366.

[83] GELBARD R, GOLDMAN O, SPIEGLER I. Investigating diversity of clustering methods: an empirical comparison [J]. Data and Knowledge Engineering, 2007, 63(1): 155 - 166.

[84] BARALDI A, BLONDA P. A survey of fuzzy clustering algorithms for pattern recognition—Part I and II [J]. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 1999, 29(6): 778 - 801.

[85] ALAHAKOON D, HALGAMUGE S K, SRINIVASAN B. Dynamic self-organizing maps with controlled growth for knowledge discovery [J]. IEEE Transactions on Neural Networks, 2000, 11(3): 601 - 614.

[86] YIN H. VISOM: a novel method for multivariate data projection and structure [J]. IEEE Transactions on Neural Networks, 2002, 13(1): 237 - 243.

[87] CAO Y, WU J. Dynamics of projective adaptive resonance theory model: the foundation of PART algorithm [J]. IEEE Transactions on Neural Network, 2004, 15(2): 245 - 260.

[88] ZHANG Y, LU J, LIU F. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding [J]. Journal of Informetrics, 2018, 12(4): 1099 - 1117.

[89] HAN J, TAO J, WANG C. FlowNet: a deep learning framework for clustering and selection of streamlines and stream surfaces [J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 11: 678 - 689.

[90] ZHAO Z, BARIJOUGH K, GERSTLAUSER A. DeepThings: distributed adaptive deep learning inference on resource-constrained IoT edge clusters [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(11): 2348 - 2359.

[91] CORTES C, VAPNIK V. Support vector networks [J]. Machine Learning, 1995, 20: 273 - 297.

[92] SCHÖLKOPF B, BURGES C, VAPNIK V. Incorporating invariances in support vector learning machines [C] // Proceedings of the 1996 International Conference on Artificial Neural Networks. Berlin: Springer, 1996: 47 - 52.

[93] TAX D M J, DUIN R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11): 1191 - 1199.

[94] BEN-HUR A, HORN D, SIEGELMANN H T, et al. A support vector clustering method [C] // Proceedings of the 2000 International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2000: 2724 - 2727.

[95] BEN-HUR A, HORN D, SIEGELMANN H T, et al. Support vector clustering [J]. Journal Machine Learning Research, 2001, 2: 125 - 137.

[96] WANG D, YEUNG D S, TSANG T C C. Structured one-class classification [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2006, 36(6): 1283 - 1295.

[97] RAJASEGARAR S, LECKIE C, BEZDEK J C. Centered hyperspherical and hyperellipsoidal one-class support vector machines for anomaly detection in sensor networks [J]. IEEE Transactions on Information Forensics and Security, 2010, 5(3): 518 - 533.

[98] AMAMI R, SMITI A. An incremental method combining density clustering and support vector machines for voice pathology detection [J]. Computers and Electrical Engineering, 2017, 57: 257 - 265.

[99] 李海林,梁叶. 基于中心度的标签传播时间序列聚类方法 [J]. 控制与决策,2018,33(11):33 - 41. (LI H L, LIANG Y. Time series clustering method with label propagation based on centrality [J]. Control and Decision, 2018, 33(11): 33 - 41.)

[100] 熊英志. 时间序列的特征表示与聚类方法研究 [D]. 镇江: 江苏大学,2018:36 - 61. (XIONG Y Z. Research on feature representation and clustering algorithm for time series [D]. Zhengjiang: Jiangsu University, 2018:36 - 61.)

[101] JONES R H. Longitudinal data with serial correlation: a state space approach [J]. Journal of the Royal Statistical Society, 1994, 36(2): 231 - 239.

[102] GAFFNEY S, SMYTH P. Trajectory clustering with mixtures of regression models [C] // Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1999: 63 - 72.

[103] DU F, ZHU A, QI F. Interactive visual cluster detection in large geospatial datasets based on dynamic density volume visualization [J]. Geocarto International, 2016, 31(6): 597 - 611.

[104] HU A, CAO J, HU M, et al. Distributed control of cluster synchronisation in networks with randomly occurring non-linearities [J]. International Journal of Systems Science, 2015, 65(4): 1 - 10.

[105] GONG M, CAI Q, CHEN X, et al. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition [J]. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 82 - 97.

[106] GUIMERA R, AMARAL L A N. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028): 895 - 900.

[107] 杨博,刘大有, LIU J-M,等. 复杂网络聚类方法 [J]. 软件学报,2009,20(1):54 - 66. (YANG B, LIU D Y, LIU J-M, et al. Complex network clustering algorithms [J]. Journal of Software, 2009, 20(1): 54 - 66.)

[108] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Science, 2002, 99(12): 7821 - 7826.

[109] 郭玉泉,李雄飞. 复杂网络社区的分形聚类检测方法 [J]. 吉林大学学报(工学版),2016,46(5):1633 - 1638. (GUO Y Q, LI X F. Fractal clustering method for uncovering community of complex network [J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(5):1633 - 1638.)

[110] 赵凤,刘汉强,范九伦. 基于互补空间信息的多目标进化聚类图像分割 [J]. 电子与信息学报,2015,37(3):672 - 678. (ZHAO F, LIU H Q, FAN J L. Multi-objective evolutionary clustering with complementary spatial information for image segmentation [J]. Journal of Electronics and Information Technology, 2015, 37(3): 672 - 678.)

[111] 张引,潘云鹤. 基于模拟退火的最大似然聚类图像分割算法 [J]. 软件学报,2001,12(2):212 - 218. (ZHANG Y, PAN Y H. Simulated annealing based maximum likelihood clustering algorithm for image segmentation [J]. Journal of Software, 2001, 12(2): 212 - 218.)

[112] GLOVER F. Tabu search, Part I [J]. ORSA Journal of Computing, 1989, 1(3): 190 - 206.

[113] BOUYER A, HATAMLOU A. An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms [J]. Applied Soft Computing, 2018, 67: 172 - 182.

[114] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer [J]. Advances in Engineering Software, 2014, 69(3): 46 - 61.

[115] 李麟玮,吴益平,苗发盛. 基于灰狼支持向量机的非等时距滑坡位移预测 [J]. 浙江大学学报(工学版),2018,52(10):167 - 175. (LI L W, WU Y P, MIAO F S. Prediction of non-equidistant landslide displacement time series based on grey wolf support vector machine [J]. Journal of Zhejiang University (Engi-

neering Science), 2018, 52(10): 167 - 175.)

[16] DENG C, LIU Y, XU L, et al. A MapReduce-based parallel K-means clustering for large-scale CIM data verification [J]. Concurrency and Computation Practice and Experience, 2016, 28(11): 3096 - 3114.

[17] HE Y, TAN H, LUO W, et al. MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data [J]. Frontiers of Computer Science, 2014, 8(1): 83 - 99.

[18] LI J, CHEN Q, LIU B. Classification and disease probability prediction via machine learning programming based on multi-GPU cluster MapReduce system [J]. Journal of Supercomputing, 2017, 73(5): 1782 - 1809.

[19] LU Y, CAO B, REGO C. A tabu search based clustering algorithm and its parallel implementation on Spark [J]. Applied Soft Computing, 2018, 63: 97 - 109.

[20] ZHOU A, WANG H, SONG P. Experiments on light vertex matching algorithm for multilevel partitioning of network topology [J]. Procedia Engineering, 2012, 29: 2715 - 2720.

[21] ANDRADE G, RAMOS G, MADEIRA D, et al. G-DBSCAN: a GPU accelerated algorithm for density-based clustering [J]. Procedia Computer Science, 2013, 18(1): 369 - 378.

[22] MELO D, TOLEDO S, MOURÃO F, et al. Hierarchical density-based clustering based on GPU accelerated data indexing strategy [J]. Procedia Computer Science, 2016, 80: 951 - 961.

[23] PARSONS L, HAQUE E, LIU H. Subspace clustering for high dimensional data: a review [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 90 - 105.

[24] YIN M, XIE S, WU Z, et al. Subspace clustering via learning an adaptive low-rank graph. [J]. IEEE Transactions on Image Processing, 2018, 27(8): 3716 - 3728.

[25] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications [C] // Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1998: 94 - 105.

[26] CHENG C H, FU A W, ZHANG Y. Entropy-based subspace clustering for mining numerical data [C] // Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1999: 84 - 93

[27] GOIL S, NAGESH H, CHOUDHARY A. MAFIA: efficient and scalable subspace clustering for very large data sets [C] // Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1999: 443 - 452.

[28] AGGARWAL C C, YU P S. Finding generalized projected clusters in high dimensional spaces [C] // Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 70 - 81.

[29] WOO K G, LEE J H, KIM M H, et al. FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting [J]. Information and Software Technology, 2004, 46(4): 255 - 271.

[30] SIM K, GOPALKRISHNAN V, ZIMEK A, et al. A survey on enhanced subspace clustering [J]. Data Mining and Knowledge Discovery, 2013, 26(2): 332 - 397.

[31] BOUGUILA N. A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(12): 1649 - 1664.

[32] CHEN L, WANG S, WANG K, et al. Soft subspace clustering of categorical data with probabilistic distance [J]. Pattern Recognition, 2016, 51(C): 322 - 332.

[33] DENG Z, CHOI K-S, WANG J, et al. A survey on soft subspace clustering [J]. Information Sciences, 2014, 348: 84 - 106.

[34] CHENG Y, CHURCH G M. Biclustering of expression data [C] // Proceedings of the 2000 International Conference of Intelligent Systems for Molecular Biology. New York: ACM, 2000: 93 - 103.

[35] HALKIDI M, BATISTAKIS Y, VAZIRGIANNIS M. Cluster validity methods [C] // Proceedings of the 2002 International Conference on Special Interest Group on Management of Data. New York: ACM, 2002: 127 - 131.

[36] THEODORIDIS S, KOUTROUMBAS K. Pattern Recognition [M]. 3rd ed. San Diego: Academic Press, 2006: 56 - 63.

[37] JOSÉ-GARCÍA A, GÓMEZ-FLORES W. Automatic clustering using nature-inspired metaheuristics: a survey [J]. Applied Soft Computing, 2015, 41: 192 - 213.

[38] LIAN C, RUAN S, DENOEUX T, et al. Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions [J]. IEEE Transactions on Image Processing, 2019, 28(2): 755 - 766.

This work is partially supported by the National Natural Science Foundation of China (6160051296).

ZHANG Yonglai, born in 1978, Ph. D., assistant professor. His research interests include big data analysis and processing, medical big data, ocean big data.

ZHOU Yaojian, born in 1987, Ph. D., assistant professor. His research interests include big data analysis and processing, ocean big data, underwater robots.