doi:10.6041/j.issn.1000-1298.2022.S1.021

基于聚类分析的数据挖掘技术及其农业应用研究进展

冯建英　石　岩　王　博　穆维松
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 基于聚类分析的数据挖掘技术能够推动农业的精准生产、精细管理和精准营销,对于实现农业的智能机械化、精准化,进而促进农业的高效化和现代化具有重要价值。首先对基于聚类分析的数据挖掘技术内涵及方法体系作了阐述,包括特征选择及特征提取、距离度量、聚类算法分类、聚类性能评价指标 4 方面;进而梳理了目前聚类分析在农业领域的动植物遗传繁育数据挖掘、农田分区精准管理、农产品品质评价、农产品市场细分、农户异质性分析与精准服务 5 大方向中的应用研究,最后对农业领域的聚类分析进行了总结与展望。

关键词: 精准农业; 聚类分析; 数据挖掘; 聚类算法; 农业精细管理
中图分类号: S126; TP399　　**文献标识码**: A　　**文章编号**: 1000-1298(2022)S1-0201-12

Cluster Analysis in Data Mining and Its Application in Agriculture

FENG Jianying　SHI Yan　WANG Bo　MU Weisong
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: Data mining technology based on cluster analysis can promote the precision production, fine management, and precise marketing of agriculture, which is of great value to realize the precision of agriculture and then promote the efficiency and modernization of agriculture. The connotation and methodological system of data mining technology based on cluster analysis were reviewed, including feature selection and feature extraction, distance metric, clustering algorithm classification, and clustering performance evaluation index; and then the current research on the application of cluster analysis in five major directions of agriculture—plant and animal genetic breeding data mining, precision management of farmland zoning, agricultural product quality evaluation, market segmentation of agricultural products, and farmer heterogeneity analysis were combed; finally, a summary of cluster analysis in agriculture was presented, and an outlook was given based on the actual needs in agriculture and the development of cluster analysis technology, which provided insight into the theoretical research and in-depth application of clustering technology in agriculture.

Key words: precision agriculture; cluster analysis; data mining; clustering algorithm; agricultural fine management

0　引言

随着大数据时代的到来,各行各业的数据量急剧增长,如何合理利用大数据,推进领域进步成为研究热点之一[1]。数据挖掘作为从大量数据中能够挖掘到潜在且有用的知识和信息的技术,是处理大数据的有效工具。其中,聚类分析是数据挖掘的主要技术之一,能够利用从数据中所发现的描述对象及其相似性的信息,将对象分成不同的簇,进而有针对性地对不同簇开展研究[2-3],具有较高的研究与应用价值。

《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》中强调要加强数字技术与农业农村发展的融合[4],以数字技术推动乡村振兴。基于聚类分析的数据挖掘技术在农业领域的应用,将有助于推动农业生产的精细管理、农

收稿日期: 2022-06-05　修回日期: 2022-08-15
基金项目: 新疆杏产业技术体系专项资金项目(XJCYTX-03)
作者简介: 冯建英(1982—),女,副教授,博士生导师,主要从事农业大数据分析与智能决策研究,E-mail: fjying@cau.edu.cn
通信作者: 穆维松(1967—),女,教授,博士生导师,主要从事农业大数据分析与智能决策研究,E-mail: wsmu@cau.edu.cn

业技术的精准应用[5-7]、农产品的精准营销[8-9]等多方面,对于农业生产节本降耗、提质增效具有重要意义。

因此,本文围绕基于聚类分析的数据挖掘技术及其在农业领域中的应用研究展开综述:首先对聚类分析的核心技术进行阐述,然后综合分析当前聚类分析在农业领域的应用研究进展,最后基于农业角度展望聚类分析的研究趋势,以期为未来基于聚类分析的数据挖掘技术在农业领域的进一步研究与应用、有效挖掘与利用农业数据知识提供支撑。

1 基于聚类分析的数据挖掘技术

数据挖掘(Data mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取出隐含在其中的、人们事先未知的,但潜在有用的信息和知识的过程[10]。聚类分析属于无监督学习,其数据分析过程如图 1 所示。

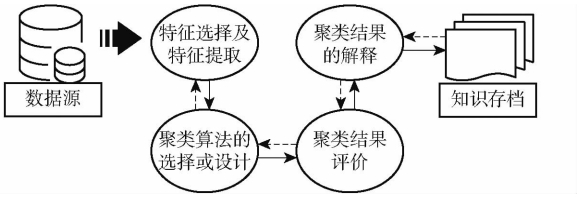

图 1 聚类分析过程

Fig. 1 Cluster analysis process

该过程始于表征聚类对象的数据源,到聚类结果的知识存档为止,其中,特征选择及特征提取、聚类算法的选择或设计、聚类结果评价以及聚类结果的解释为聚类分析的关键环节[11],其中聚类算法的选择或设计环节涵盖了聚类的重要步骤——距离度量。

1.1 特征选择及特征提取

特征选择是从数据样本集的所有属性(特征)中选择出具有较强表达优势能够达到某一目标的若干属性,即初始属性集的一个子集,从而达到降低维度的目的[12]。通过特征选择可以减少原始数据获取的时间,避免数据冗余,缩减数据存储空间,提高模型运行效率和性能。已有的许多特征选择方法属于有监督学习[13-15],但待聚类的数据集一般是缺少数据标签的,针对这一问题,JUNG[16]提出了一种无监督的特征选择办法,通过将特征选择问题转换为估计问题,从而避免组合搜索;此外,马盈仓等[17]提出一种基于流形学习与 $L_{2,1}$ 范数的无监督多标签特征选择方法,谢娟英等[18]提出了基于谱聚类的无监督特征选择思想 FSSC(Feature selectionby spectral clustering)。

特征提取则通过对所选择的特征进行转换形成新的突出特征。主成分分析(Principal component analysis, PCA)为特征提取的典型方法[19-20]。

可以看出,本质上特征选择和特征提取都是降维过程。通过特征选择或提取,能够降低数据维度,舍弃次要特征,凸显关键聚类特征变量的价值,提高聚类分析的效率和性能。

1.2 距离度量

聚类分析的基础是聚类对象之间的相似性和差异性,而对象之间的相似程度或相异程度则需要基于某种距离定义来测量。通常情况下,定义一个距离函数 $d(x,y)$,需满足以下准则[21]:①$d(x,x)=0$,对象与其自身距离为 0。②$d(x,y)\geqslant 0$,非负性。③$d(x,y)=d(y,x)$,对称性。④$d(x,y)\leqslant d(x,k)+d(k,y)$,满足三角不等式。常见的距离度量方法综述如表 1 所示。

然而,对于混合属性数据,表中度量方法往往失效。但是混合类型变量在实际聚类问题中越来越普遍,因此,度量混合属性间的相似性/相异度凸显出研究价值。目前,混合属性数据集的距离度量方法主要是通过融合数值数据与分类数据两部分进行,主流方法有:Goodall 距离[22]、Gower 距离[23]、K-prototypes 距离[24]、MinkoVDM 距离[25]、OCIL 距离[26]等。随着混合属性数据的日益普及和重要,针对混合数据集的距离度量新方法也在不断出现[27-29]。

1.3 聚类算法分类

传统聚类算法主要包括基于划分、基于层次、基于网格、基于密度和基于模型 5 种。同时,随着多学科交叉融合及聚类分析理论的创新,许多新型的聚类算法也应运而生。

1.3.1 传统聚类算法

(1)基于划分的聚类算法

基于划分的聚类算法通过定义或优化一些标准函数,将对象分配到 k 个簇中。此类算法主要包括 K-means[30]、PAM[39]、CLARA[39]、CLARANS[40]、K-means + +、K-modes、ISODATA 等。

1967 年 Macqueen 提出的 K-means 算法是这类聚类算法的经典代表,但该算法存在对 k 值和初始聚类中心敏感等问题,很多研究基于此开展了经典算法的改进和优化研究。文献[41]指出最佳 k 值应在 $(1,\sqrt{n})$ 内,n 指数据集大小,这为改进 K-means 算法奠定了一定的理论基础。成卫青等[42]基于 k 与误差平方和(SSE)的关系,指出选择“肘点”所对应的 k 为最优簇数;在此基础上,王建仁等[43]针对簇数 k 与 SSE 的关系图像中“肘点”模糊不明确的问题,提出基于指数函数性质、权重调

表 1　聚类分析中常见的距离度量(相似性度量)方法

Tab. 1　Common distance formula(included similarity)

度量方法	适用变量类型	数学表达式		方法评价
欧氏距离(Euclidean distance)[30-31]	连续型	$D(X_i,X_j)=\sqrt{\sum_{l=1}^{d}(x_{il}-x_{jl})^2}$	(1)	易于理解,具有空间旋转不变性,但该距离没有考虑分量之间的相关性,且受量纲影响(针对其劣势有改进的公式[21])
曼哈顿距离(Manhattan distance)[32]	连续型	$D(X_i,X_j)=\sum_{l=1}^{d}\lvert x_{il}-x_{jl}\rvert$	(2)	相比于欧氏距离,降低了离群点的影响
切比雪夫距离(Chebyshev distance)[33]	连续型	$D(X_i,X_j)=\max_l(\lvert x_{il}-x_{jl}\rvert)$	(3)	受量纲影响;没有考虑各个分量的分布(期望、方差等)可能有所不同
兰氏距离(Canberra distance)[34]	连续型	$D(X_i,X_j)=\sum_{l=1}^{d}\frac{\lvert x_{il}-x_{jl}\rvert}{\lvert x_{il}\rvert+\lvert x_{jl}\rvert}$	(4)	消除了量纲,受异常值影响较小,适合于数据具有高度偏倚的情况,未考虑各维度之间的相关性
马氏距离(Mahalanobis distance)[35-36]	连续型	$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sqrt{(\boldsymbol{X}_i-\boldsymbol{X}_j)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j)}$	(5)	不受量纲影响,考虑了各维度相关性,但该距离通常要求样本总数大于样本维数且计算较复杂
余弦相似度(Cosine similarity)[37]	连续型	$S(X_i,X_j)=\cos\alpha=\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j}{\lVert\boldsymbol{X}_i\rVert\lVert\boldsymbol{X}_j\rVert}$	(6)	适用于高维数据,对数值(向量大小)不敏感
皮尔逊相关系数(Pearson correlation coefficient)[38]	连续型	$\rho_{X_iX_j}=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{D(X_i)}\sqrt{D(X_j)}}$	(7)	一定程度上消除了量纲的影响,具有平移不变性,只适用于连续变量,限制了其应用
简单匹配距离(Simple matching distance)[11]	离散型	$D(X_i,X_j)=\frac{1}{d}\sum_{l=1}^{d}\delta(x_{il},x_{jl})$ 其中 $\delta(x_{il},x_{jl})=\begin{cases}0 & (x_{il}=x_{jl})\\1 & (x_{il}\neq x_{jl})\end{cases}$	(8)	计算过程简单易理解,适用于二值型变量(如性别),具有一定的局限性
杰卡德距离(Jaccard distance)[21]	离散型	$D(X_i,X_j)=\frac{q+r}{p+q+r}$	(9)	可用于二元变量不对称的情况
汉明距离(Hamming distance)[21]	离散型	$H_d(X,Y)=\sum_{i=1}^{n}\lvert X_i-Y_i\rvert$	(10)	可用来表示两个同等长度的字符串,由一个转换为另一个的最小替换次数

注:式(1)X_i、X_j 分别表示第 i 个和第 j 个样本点,l 为样本特征的维数,x_{il}、x_{jl} 表示第 i、j 个样本点的第 l 维,d 为样本总维数即属性个数,$D(X_i,X_j)$ 表示两样本间的距离(其他公式同);$\boldsymbol{S}$ 表示样本的协方差矩阵;式(6)中 S 表示样本间的相似性;式(9)中 X_i、X_j 为两个 d 维变量,所有维度取值 0 或 1,p 表示样本 X_i 与 X_j 都是 1 的维度个数,q 表示 X_i 是 1、X_j 是 0 的维度个数,r 表示 X_i 是 0、X_j 是 1 的维度个数;式(10)中 X、Y 为两个样本,n 指代样本总维数,$X,Y\in\{0,1\}^n$,$n>0$;未作特殊说明的数学符号(如式(7)中的 Cov、D)则具有其通用的数学意义。

节、偏执项和手肘法思想来确定最佳 k 值;WANG 等[44]基于图像分割的思想,结合分水岭算法将数据集划分为多个不同区域以确定最佳 k 值。为了优化初始聚类中心的选择,刘佳鸣等[45]通过 α 狼、β 狼和 δ 狼前 3 种等级靠前的优势狼,不断迭代优化更新候选狼 ω 狼的位置,不断更新优势狼,最终得到最优的聚类中心;ARTHUR 等[46]提出 K-means + + 算法,先考虑样本集内所有样本的分布情况,按照距离尽可能远的原则选取初始中心点,从而避免较大的簇被误划分或者几个较小的簇被误合并的问题;在此基础上,文献[47]结合本地搜索策略对 K-means + + 进行优化。针对 K-means 对噪声和孤立点数据敏感的问题,KAUFMAN 等[39]提出了 K-mediods 算法(包括 PAM 和 CLARA),每个簇用接近该簇的中心对象表示,取得了较好效果;邓斌涛等[48]提出一种基于动态双子种群的差分进化 K 中心点聚类算法,提高了算法精度与运行时间。另外,针对传统 K-means 算法对分类型数据失效的问题,HUANG[49]提出了 K-modes 算法,采用 0 - 1 差异度替代 K-means 算法中的距离。NG 等[40]提出 CLARANS 聚类方法,能有效地处理凸和非凸多边形对象。而 ISODATA 算法则通过动态地进行簇的合并与分裂,能够自动地调节聚类 k 值,通过迭代操作得到较为理想的聚类结果[50]。

(2)基于层次的聚类算法

与基于划分的聚类算法不同,基于层次的聚类算法通常通过对数据集进行层次分解或合并得到聚类结果。根据产生聚类的方式,层次聚类可分为凝聚式和分裂式两种基本方法。凝聚式的层次聚类从点作为个体簇开始,每一步合并两个距离最近的簇;而分裂式的则是从包含所有点的簇开始,每一步进行一次分裂,直至仅剩包含单点的簇。其中,BIRCH[51]、CURE[52]、ROCK[53]、CHAMELEON[54]为 4 种常见的基于层次的聚类算法。

基于层次的聚类算法多从距离度量、合并策略等方面进行改进。针对大多层次聚类算法依赖于对

称距离定义的问题,韩忠明等[55]提出了一种非对称距离下的层次聚类算法,在 Delicious 和 Diggo 数据集上,该算法与传统对称算法在使用不同连接因子的情况下相比,同表象相关系数更接近于 0,即聚类效果更好。李青旭等[56]提出了一种新的基于交点的层次聚类算法,在肝病、甲状腺病、糖尿病数据集上与传统层次聚类算法相比,纯度分别提高了 5、15、3 个百分点。SHI 等[57]提出了基于 HCMS 合并策略的分层聚类算法,在 8 个合成数据集上,准确率均能达到 99% 以上。

(3)基于网格的聚类算法

基于网格的聚类算法通常采用一种多分辨率的网格类型数据结构,将空间量化成为有限数目的单元,聚类过程根据这些单元所形成的网格结构进行,典型的算法有 STING、WaveCluster、CLIQUE 算法等。

为减少经典网格聚类算法的计算量,缪裕青等[58]在网格聚类过程中引入动态衰减因子,周期性地移除稀疏网格;而蔡莉等[59]针对因删减稀疏区域而导致原本属于某一簇的数据点被删除的问题,结合网格划分和距离判断提出了一种基于自适应网格划分聚类算法。王玮琪等[60]通过引入维度半径概念进行增量动态网格,提出局部网格动态聚类算法,提高数据处理效率的同时其在 Wine 和 S1 公开数据集上的平均错误率仅为 4.334% 和 1.550%。近年来,许多研究者将密度函数引入网格聚类算法,使得经典的网格聚类算法得到进一步优化[61-62]。

(4)基于密度的聚类算法

通常情况下,基于划分的聚类算法更适用于发现凸形聚类簇,难以识别任意形状的聚类簇。而基于密度的聚类算法则很好地解决了这一问题,该类算法假设聚类结构能够通过样本分布的紧密程度确定,即从对象在空间中分布的密度角度来开展聚类。其中,DBSCAN 算法[63]是最经典的基于密度的聚类算法之一。该算法将处于高密度区域的对象称为核心对象,基于聚类内部任意核心对象不断扩展生成聚类簇。但 DBSCAN 存在聚类质量依赖手动设置的初始参数且不适用于密度变化较大的数据集的问题, OPTICS 算法则在 DBSCAN 的基础上进行了完善,取得了良好的改进效果[64]。RODRIGUEZ 等[65]提出了快速搜索和发现密度峰值的聚类能够自动获得簇的正确个数,在此研究基础之上,BIE 等[66]通过引入模糊规则,提出了一种模糊- CFSFDP 方法,从而有效地自适应选择聚类簇中心。王治和等[67]利用反向最近邻和相似度矩阵找到当前全局密度最大的数据样本,分析此样本周围密度的分布情况,采用自适应方法计算当前簇的邻域阈值,从而提出了一种新的优化初始点和自适应半径的密度聚类算法。

(5)基于模型的聚类算法

基于模型的聚类算法目标是将数据与某个模型达成最佳拟合,主要有基于统计概率模型的聚类方法和基于神经网络模型的聚类方法,前者的代表性算法有高斯混合模型算法(Gaussian mixture models, GMM)、COBWEB 算法,后者中自组织映射网络(Self-organizing maps, SOM)较为经典。高斯混合模型属于一种概率模型,将各个簇服从的分布进行线性组合,线性组合包含每个分布的权重。而自组织映射网络是由全连接神经元阵列所组成的无监督、自组织、自学习网络,能够实现从 n 维输入空间到低维输出平面的映射。余玥等[68]对高斯混合模型聚类算法进行了改进,提出了一种基于分布式聚类的农业环境数据异常检测方法。GRIECO 等[69]利用 SOM 进行文本聚类,有效提高了文本聚类效果。在实际应用中,这类聚类算法通常不稳定,即同一算法多次运行会得到不同的结果。

1.3.2 新型聚类算法

随着机器学习与人工智能理论研究的进一步发展,出现了一些新型聚类算法。其主要可以分为基于样本归属关系(如模糊聚类等)[70-72]、基于样本预处理(如核聚类等)[73-74]、基于样本更新策略(如数据流增量聚类、生物智能的增量聚类等)[75-76]、基于样本高维性等聚类算法(如投影寻踪聚类等)[77-78]以及谱聚类、量子聚类、随机游走聚类等[79-80]。

同时,随着集成学习的发展,集成学习被应用在聚类领域。STREHL 等[81]提出“聚类集成”(Cluster ensembles)的概念,即对于一个数据集的多种划分组合成一个统一聚类结果的方法。与单一聚类算法相比,通常基于集成学习的聚类算法时间成本较高,但其有效提高了聚类的鲁棒性和聚类结果的质量。

新型聚类算法极大地扩展和提升了聚类分析的应用领域和效果。JIA 等[72]提出了一种鲁棒自稀疏模糊聚类算法,将其应用在图像分割领域,克服了传统方法对异常值敏感、过度分割等问题。朱颖雯等[78]基于随机投影和自适应谐振理论提出了一种针对高维数据流的聚类算法,可有效缓解维数灾难导致的问题。诸震宇等[74]基于核聚类方法实现了对不同气味的识别,该模型能够为今后研究更高层次的嗅觉网络水平提供理论基础。张海涛等[82]利用谱聚类算法挖掘虚拟健康社区潜在信息,改善了知识聚合的质量,为知识聚合和知识服务提供了一条新途径。黄飞杰等[80]基于烟草零售数据集成了 CLARA 算法和 K-means 算法构建客户细分模型,实现烟草零售精准营销。

1.3.3 常见聚类算法比较

表 2 对常见聚类算法进行了多维度的比较，对于实际应用中选择恰当的聚类算法有一定的指导意义。

表 2 常见聚类算法比较

Tab. 2 Comparison of common clustering algorithms

聚类算法类别	基本思想	代表性模型	适合大规模数据	适合高维数据	噪声敏感度
基于划分的聚类	将待处理的数据样本依据给定的聚类簇数进行划分	K-means	是	否	高
		PAM	否	否	低
		CLARA	是	否	低
		CLARANS	是	否	低
基于层次的聚类	试图在不同层次对数据集进行划分，从而形成树形的聚类结构	BIRCH	是	否	低
		CURE	是	是	低
		ROCK	否	是	低
		Chameleon	否	否	低
基于网格的聚类	把对象空间量化为有限数目的单元，形成一个网格结构，所有聚类操作均在该网格结构（即量化的空间）上进行	STING	是	是	低
		CLIQUE	否	是	适中
基于密度的聚类	邻近区域的密度超过某个阈值则继续聚类	DBSCAN	是	否	低
		OPTICS	是	否	低
基于模型的聚类	为每个聚类簇假定一个模型，寻找数据对给定模型的最佳拟合	COBWEB	是	否	适中
		SOM	否	是	低
新型聚类算法	基于新的领域理论方法进行聚类	CLICK	是	否	高
		kernel FCM	否	否	低

1.4 聚类性能评价指标

不同聚类方法产生的聚类结果有所差异，因此需要采用聚类评价指标衡量聚类结果的优劣。依据是否利用数据集样本真实标签信息，常见的聚类评价指标可分为外部评价指标和内部评价指标。外部评价指标旨在通过对比聚类结果与真实类别的匹配程度，对聚类结果的有效性进行评价，常见的外部评价指标有 F－measure 值[83－84]、RI（Rand index）[85]、ARI（Adjusted rand index）[79,83]、NMI（Normalized mutual information）[79,83,85]、AMI（Adjusted mutual information）[86]等；而内部评价指标不使用原始数据分布的先验信息（或者数据集无标签信息），而是通过发现数据集内部结构与分布的状态来评价聚类结果的优劣，主要有轮廓系数[83,86]、Calinski－Harabasz 指数[83,87]、DBI（Davies-bouldin index）[83,87]等。

学者们也会结合内部和外部评价指标共同对聚类算法的性能进行评价，使聚类算法更具可靠性。李顺勇等[88]、康雁等[89]、LI 等[83]分别联合利用轮廓系数等内部指标和 ARI、AMI、NMI 等外部指标中的一种或几种评价了其改进算法的有效性。

2 聚类分析在农业领域的应用

聚类分析技术作为数据挖掘的有效工具之一，现已广泛应用到农业生产的全产业链各环节，包括动植物遗传繁育数据挖掘、精准农业管理区划分、农产品品质评价、农产品市场细分、农户异质性分析与精准服务等，对于挖掘农业领域潜在知识以解决农业生产技术问题，促进农业生产提质增效、提高农民经济效益，进而促进农业持续精准高效发展具有重要意义。

2.1 动植物遗传繁育数据挖掘

动植物遗传繁育数据通常具有数据量大、数据类型多样等特点，聚类分析技术可以辅助挖掘动植物遗传繁育的种群特征及优势。在动物遗传繁育方面，田镇等[90]以江苏黄文蛤原种（SY）、江苏红文蛤原种（SR）及 5 个红壳色文蛤选育群体（SRF1 ~ SR5F5）为研究对象，依据文蛤群体间的遗传距离并采用 UPGMA 聚类方法发现文蛤群体聚类呈两大支。UPGMA 聚类是基于层次聚类的一种方法，多用于生物多样性研究。杨红远等[91]基于云上黑山羊羔羊哺乳阶段体质量日增数据，通过凝聚式层次聚类分析发现，其在哺乳期内生长速度可以分为 3 个阶段，其中，在 0 ~ 10 日龄和 70 ~ 90 日龄生长速度最快。在植物繁育方面，姜存良等[92]根据野生猕猴桃的表型性状采用系统聚类方法在欧氏距离为 20 处将 24 个野生猕猴桃种分为 4 组。程琳等[93]以欧氏距离作为遗传距离，基于杉木无性系 15 个品质性状利用层次聚类法将 40 份杉木无性系分为 4 类群。

2.2 农田分区精准管理

农业分区管理可根据不同的分区特点制定管理策略，为精准农业的实施提供理论基础[5]。通过基于聚类的农田分区管理能够提高作物管理的精确

性,进而达到降低生产成本与增加作物产量的效果。为提高大面积玉米农田作物管理的精确性,陈世超等[5]对甘肃黄羊河农场玉米膜下滴灌示范区地形等特征首先利用主成分分析提取了 3 个主成分作为分区依据,之后利用模糊 c 均值聚类法划分出具有明显差异的 3 个管理区,因而可以依据不同分区特点制定管理策略;针对农田秸秆形态多样、细碎秸秆难以准确识别的问题,安晓飞等[94]提出了一种基于 K-means 聚类和分区寻优结合的秸秆覆盖率计算方法,实现了对不同天气、种植模式、地块条件下的秸秆覆盖率准确计算。另外,在农业分区节水灌溉方面,杨蕊等[6]运用模糊聚类方法将云南省农业节水分区分为水资源缺乏区、季节性缺水区、工程性缺水区、干热河谷缺水区 4 个一级区以及 10 个二级区,为提高节水效率提供了理论基础与新途径。在基本农田规划与土壤环境质量分区方面,吴彦澎[95]针对当前划定基本农田的过程中主观干预过多的问题,提出一种基于网格的基本农田保护规划空间聚类算法,用于指导基本农田保护规划的实施;刘佳斌等[96]基于高斯混合模型的聚类分区方法,充分考虑土壤质地、土壤类型等因素的作用范围及尺度特征,客观反映了土壤环境质量分布的情况。

2.3 农产品品质评价

学者们基于农产品的品质属性特征、利用聚类技术将某种农产品的多个品种细分为若干类,进而比较分析不同品种的品质优劣。为评价陕西省中华和美味 2 种猕猴桃主栽品种与新品种的品质差异,王丹等[97]利用主成分分析对猕猴桃品种的品质指标进行降维,之后利用层次聚类将 8 个猕猴桃品种分为 5 类,针对每一类作出了差异化的品质评价。周琪乐等[98]运用层次聚类分析法建立枸杞营养价值和品质评价方法,分析得出黑枸杞营养品质最好。周双等[99]选择树莓果实的外观品质和营养成分进行分析,通过层次聚类将树莓品种分成 3 个类群,结果表明不同品种的各项指标存在差异。聚类分析作为中间技术用于辅助农产品图像分割,最终提高了分类模型评估农产品新鲜度的准确性。BHARGAVA 等[7]基于苹果图像数据集,在图像处理阶段通过计算像素点间欧氏距离并利用模糊 c 均值聚类来分割 6 种苹果图像,实现苹果图像的有效分割,为利用分类器实现苹果新鲜度准确分类奠定了基础。SHWETA 等[100]利用 K-means 将二值化后的番茄果实图像划分为用红、绿、蓝表示的 3 个簇,划分出的簇有助于提取颜色、图像质量、图像大小等信息,这些信息特征使得后续利用分类模型对果实质量进行实时评估的精度更高。

2.4 农产品市场细分

基于聚类分析的市场细分可以将不同需求、特征的消费者划分成若干不同的消费者群体,根据细分的目标市场开展差异化生产和营销,是农业企业提高经济效益的有效途径。郭燕婷等[101]基于北京城镇居民消费鲜食玉米偏好特征,在使用因子分析提取出物质需求、功能需求和精神需求 3 个公因子后利用 K-means 将消费者市场细分为感情动机型、理智动机型 2 个子市场。陈虹瑶等[8]基于文本挖掘获得葡萄酒在线评论数据,利用 K-means 聚类将消费者细分为追求品质组、性价权衡组、服务偏好组、外观物流组 4 类细分市场。LI 等[83]针对葡萄消费者自身及其偏好特征,提出了一种基于改进 K-means 和自适应粒子群算法的葡萄消费者细分方法,同时证明了所提出的 IKM - ALPSO 算法在应用在消费者细分上的有效性和实用性,细分出的消费者群体在购买行为、偏好方面具有明显差异,可以据此形成细分市场以利于提高葡萄企业竞争力。文献[9]从苹果视觉外观、先前经验、口感/香味、质地 4 种购买动机出发,利用层次聚类的方法将消费者分为 5 类,分别形成相应的细分市场(更好的食用品质追求者、熟悉或美观苹果中具有更好食用品质的追求者、口感追求者、完美产品追求者),使得市场销售更有针对性。

2.5 农户异质性分析与精准服务

聚类分析用于识别农户的异质性类型,把握不同类型农户群体信息需求的差异性及规律性,从而精准有效地提供农业服务。ETUMNU 等[102]利用层次聚类法将农户进行聚类,发现农民群体在其优先考虑战略管理因素方面是异质的,相关研究结果为农民、农场投入提供商和农场服务提供商提供了有价值的参考。JESSE 等[103]基于农户属性及其农业活动数据相关变量对小农户奶农进行 K-means 聚类分析以表征差异化的农户群体。GRASKEMPER 等[104]根据德国农民个人特征及农场相关特征使用 PAM 算法将德国农民聚类成传统种植者、多才多艺的年轻人、基于家庭的农户 3 个群体,群体之间在个人及农场特征方面明显不同,因而对于未来农业生产活动政策的制定以及进行相应的农业信息服务等具有指导意义。

2.6 农业领域应用比较分析

表 3 总结了前文所述的聚类分析在农业领域的研究与应用。可以看出,在实际农业领域的应用中,经典聚类方法的使用仍是主流,近年来也有诸多学者探索性使用一些新的聚类方法,如基于机器学习改进的聚类方法、模糊聚类等[6,83],提高了聚类分析

表 3　聚类分析在农业领域应用比较

Tab. 3　Comparison of cluster analysis applied in agriculture

应用领域	功能描述	常用聚类方法类型
动植物遗传繁育数据挖掘	挖掘动植物遗传繁育数据中潜在的有价值信息，探讨遗传性状、繁育相关特征的差异，对育种或动植物个体研究等工作有一定的指导意义	基于层次[90-93]
农田分区精准管理	体现不同划分区之间的差异，提高农田区域管理的精确性，进而达到提高农田利用效率、降低生产成本、提高作物产量，节水灌溉、农田可持续利用等效果	基于划分[94]、基于样本归属关系[5-6]、基于网格[95]、基于模型[96]
农产品品质评价	探索品种间的品质差异或作为辅助技术以评估果实品质优劣	基于划分[100]、基于层次[97-99]
农产品市场细分	将消费者按照人口统计特征、心理、行为等指标细分为不同的客户群体，对不同群体进行针对性的产品研发及营销措施，提高企业绩效	基于划分[8,101]、基于层次[9]
农户异质性分析与精准服务	构建群体用户画像，挖掘分析农户群体信息需求差异	基于划分[104]、基于层次[102]

的有效性。

3　总结与展望

本文围绕聚类分析的核心技术，分析了聚类分析过程中的特征选择与提取、距离度量、聚类算法分类以及聚类结果评价的主流方法，阐述了聚类算法在农业领域中的应用研究进展。文献显示基于聚类分析的数据挖掘技术可以利用农业数据特征变量，将数据对象分为有着明显差异的不同群体，获得农业生产等过程中的潜在知识，使研究进一步细化或为后续研究奠定一定的理论与技术基础，为农业技术改革创新，农业生产提质增效、转型升级，发展高效精准农业提供支撑。基于农业发展的实际需求，聚类分析技术在农业领域中的研究与应用呈现以下几方面趋势：

(1)高维度、大规模农业数据聚类

农业生产现实环境具有复杂多变的特点，土壤、水分、品种、病虫害、杂草等对于作物产量具有重要影响，精准农业更有利于保证粮食安全，而在发展精准农业过程中对多种现代信息技术的需求增大，所产生的数据如遥感、光谱学数据等，呈现出数据量大、维度高的趋势，这为利用聚类分析实现农田分区精准管理等带来了挑战。根据具体情况选取有效的特征选择与特征提取方法以降低数据处理维度、提高数据效率在精准农业等方面表现出了良好效果[105-106]，因此，针对特征选择与特征提取方法的研究与应用将成为聚类分析的重要发展方向之一。

(2)基于农业多源异构混合属性数据的聚类分析

随着农业领域信息化技术发展，农业生产各环节产生的数据通常不仅为单一类型的数据，越来越多的非数值类型的数据也具有重要的意义，如农户和农产品市场中消费者特征属性的多样性要求聚类分析能够多角度挖掘农户及消费者市场信息，因此针对多源异构混合属性数据的聚类分析对于农业领域将更具应用价值。现有的聚类算法大多基于欧氏距离等传统距离度量方式，往往能够有效处理数值型数据而对于分类型数据或混合属性数据则不适用，因此，根据领域知识研究适用于多源异构混合属性数据的距离度量方式从研究和应用角度都具有较高的现实意义。

(3)基于群智能进化算法优化现有聚类方法

现有的聚类算法存在参数敏感的问题，这导致其在农业研究中结果不理想。作为有效的寻优算法，群智能进化算法(如粒子群、遗传算法等)应用在对常见聚类算法的聚类中心等寻优方面表现出了强大的优越性，能够明显改善聚类结果。新群智能进化算法在展现优越性能的同时也能为聚类算法优化拓展思路[107]，这对于农业数据有效聚类分析并获得更可靠、可解释性高的聚类结果有着较大的研究与应用价值。

(4)新型聚类分析技术

农业的可持续发展需要与时俱进的机械化、智能化技术与方法，目前，农业领域应用传统型聚类算法进行聚类分析仍占主流，新型的聚类分析技术在提高聚类有效性方面有较为明显的优势，同时也存在发展空间。其中，谱聚类算法在农产品图像分割上具有一定的适用性[108]，由于其相似性矩阵是误差来源的第一环节，现有的构造方法众多，许多并未使用其他方面的先验信息，如密度先验信息，且该方法时间复杂度较高。因而在已有算法基础上深入研究，结合多方面特征构造相似性矩阵，进而提高谱聚类算法的精确性并降低时耗是未来谱聚类算法改进的重要研究方向。模糊聚类则适用于具有模糊隶属性的农业数据，算法参数对聚类结果的质量有着明显影响，可结合聚类结果评价性指标与智能优化算法确定相关参数，提高聚类效果。针对农业数据聚类结果质量与数据处理效率等问题，新型聚类分析技术或将克服传统聚类方法的弊端，发挥自身优势，以适应应用场景，达到农业数据价值最大化以及使研究成果更具科学有效性的目的。

参 考 文 献

[1] RABHI L, FALIH N, AFRAITES A, et al. Big data approach and its applications in various fields[J]. Procedia Computer Science, 2019, 155: 599 - 605.

[2] GHOSAL A, NANDY A, DAS A K, et al. A short review on different clustering techniques and their applications[M]// MANDAL J, BHATTACHARYA D. Emerging technology in modelling and graphics: advances in intelligent systems and computing. Singapore: Springer Singapore, 2020: 69 - 83.

[3] HIGUCHI A, MAEHARA R. A factor-cluster analysis profile of consumers[J]. Journal of Business Research, 2021, 123: 70 - 78.

[4] 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要[N]. 人民日报, 2021-03-13(001).

[5] 陈世超, 杜太生, 王素芬. 基于模糊c均值聚类法的玉米农田管理分区研究[J]. 农业机械学报, 2019, 50(11): 293 - 300.
CHEN Shichao, DU Taisheng, WANG Sufen. Delineating management zones in maize field based on fuzzy c-means algorithm [J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11): 293 - 300. (in Chinese)

[6] 杨蕊, 胡贤群, 王龙, 等. 基于主成分分析和模糊聚类的云南省农业节水分区[J]. 节水灌溉, 2021(4): 92 - 97.
YANG Rui, HU Xianqun, WANG Long, et al. Zoning of agricultural water-saving in Yunnan Province based on principal component analysis and fuzzy cluster[J]. Water Saving Irrigation, 2021(4): 92 - 97. (in Chinese)

[7] BHARGAVA A, BANSAL A. Classification and grading of multiple varieties of apple fruit[J]. Food Analytical Methods, 2021, 14(7): 1359 - 1368.

[8] 陈虹瑶, 李淑琦, 王春盈, 等. 葡萄酒网购消费者购买决策影响因素与客户细分——基于在线评论数据[J]. 酿酒科技, 2021(11): 127 - 132.
CHEN Hongyao, LI Shuqi, WANG Chunying, et al. Influencing factors of consumers' purchase decision in wine online shopping and customer segmentation based on online review data[J]. Liquor-Making Science & Technology, 2021(11): 127 - 132. (in Chinese)

[9] BEJAEI M, CLIFF M A, SINGH A. Multiple correspondence and hierarchical cluster analyses for the profiling of fresh apple customers using data from two marketplaces[J]. Foods, 2020, 9(7): 873.

[10] HAN J, PEI J, KAMBER M. Data mining: concepts and techniques[M]. Elsevier, 2011.

[11] 章永来, 周耀鉴. 聚类算法综述[J]. 计算机应用, 2019, 39(7): 1869 - 1882.
ZHANG Yonglai, ZHOU Yaojian. Review of clustering algorithms[J]. Journal of Computer Applications, 2019, 39(7): 1869 - 1882. (in Chinese)

[12] 蒋盛益, 郑琪, 张倩生. 基于聚类的特征选择方法[J]. 电子学报, 2008, 36(增刊1): 157 - 160.
JIANG Shengyi, ZHENG Qi, ZHANG Qiansheng. Clustering-based feature selection[J]. Acta Electronica Sinica, 2008, 36 (Supp.1): 157 - 160. (in Chinese)

[13] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014, 40 (1): 16 - 28.

[14] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226 - 1238.

[15] 戴建国. 一种新的有监督特征选择方法[J]. 陕西理工学院学报(自然科学版), 2017, 33(4): 89 - 92.
DAI Jianguo. A novel method for supervised feature selection[J]. Journal of Shaanxi Institute of Technology (Natural Science Edition), 2017, 33(4): 89 - 92. (in Chinese)

[16] JUNG S H. Queen-bee evolution for genetic algorithms[J]. Electronics Letters, 2003, 39(6): 575 - 576.

[17] 马盈仓, 张要, 张宁, 等. 基于流形学习与$L_{2,1}$范数的无监督多标签特征选择[J]. 纺织高校基础科学学报, 2021, 34 (3): 102 - 111,120.
MA Yingcang, ZHANG Yao, ZHANG Ning, et al. Unsupervised multi-label feature selection based on manifold learning and $L_{2,1}$ norm[J]. Basic Sciences Journal of Textile Universities,2021,34(3):102 - 111,120. (in Chinese)

[18] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009 - 1024.
XIE Juanying, DING Lijuan, WANG Mingzhao. Spectral clustering based unsupervised feature selection algorithms[J]. Journal of Software, 2020, 31(4): 1009 - 1024. (in Chinese)

[19] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1 - 3): 37 - 52.

[20] 李静, 郝耀军, 杨瑜. 一种改进DBSCAN特征聚类的学习者类画像方法[J]. 计算机与数字工程, 2022, 50(4): 703 - 708,756.
LI Jing, HAO Yaojun, YANG Yu. A learner class profile method based on improved DBSCAN feature clustering[J]. Computer and Digital Engineering, 2022, 50(4): 703 - 708, 756. (in Chinese)

[21] 石胜飞. 大数据分析与挖掘[M]. 北京: 人民邮电出版社, 2018.

[22] GOODALL D W. A new similarity index based on probability[J]. Biometrics, 1966, 22(4): 882 - 907.

[23] GOWER J C. A general coefficient of similarity and some of its properties[J]. Biometrics, 1971, 27(4): 857 - 871.

[24] HUANG Z. Clustering large data sets with mixed numeric and categorical values[C]// Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD), 1997: 21 - 34.

[25] ZHOU Z H, YU Y. Ensembling local learners throughmultimodal perturbation[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2005, 35(4): 725 - 735.

[26] CHEUNG Y M, JIA H. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number[J]. Pattern Recognition, 2013, 46(8): 2228 - 2238.

[27] 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.

[28] 邱保志, 张瑞霖, 李向丽. 基于残差分析的混合属性数据聚类算法[J]. 自动化学报, 2020, 46(7): 1420 - 1432.
QIU Baozhi, ZHANG Ruilin, LI Xiangli. Clustering algorithm for mixed data based on residual analysis[J]. Acta Automatica Sinica, 2020, 46 (7): 1420 - 1432. (in Chinese)

[29] FOSS A H, MARKATOU M, RAY B. Distance metrics and clustering methods for mixed-type data [J]. International Statistical Review, 2019, 87(1): 80 - 109.

[30] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967: 281 - 297.

[31] HOSSAIN M Z, AKHTAR M N, AHMAD R B, et al. A dynamic K-means clustering for data mining[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2019, 13(2): 521 - 526.

[32] FAISAL M, ZAMZAMI E. Comparative analysis of inter-centroid K-means performance using euclidean distance, canberra distance and manhattan distance[C]// Journal of Physics: Conference Series, 2020: 012112.

[33] GULTOM S, SRIADHI S, MARTIANO M, et al. Comparison analysis of K-means and K-medoid with Ecluidience distance algorithm, Chanberra distance, and Chebyshev distance for big data clustering[C]// IOP Conference Series: Materials Science and Engineering, 2018: 012092.

[34] YANG Y M, JIA R, XUN H, et al. Determining the number of instars in *Simulium quinquestriatum* (*Diptera*: *Simuliidae*) using K-means clustering via the Canberra distance[J]. Journal of Medical Entomology, 2018, 55(4): 808 - 816.

[35] MAHALANOBIS P C. On the generalized distance in statistics[J]. Proceedings of the National Institute of Sciences of India, 1936, 2(1): 49 - 55

[36] LAHAV A, TALMON R, KLUGER Y. Mahalanobis distance informed by clustering[J]. Information and Inference, 2018, 8 (2): 377 - 406.

[37] SREE K, MURTHY D V R. Clustering based on cosine similarity measure[J]. International Journal of Engineering Science & Advanced Technology, 2012, 2(3): 508 - 512.

[38] BENESTY J, CHEN J, HUANG Y, et al. Pearson correlation coefficient[M]// Noise reduction in speech processing. Berlin: Springer, 2009: 1 - 4.

[39] KAUFMAN L, ROUSSEEUW P. Finding groups in data: an introduction to cluster analysis[M]. New Jersey: John Wiley & Sons, 1990.

[40] NG R T, HAN J. CLARANS: a method for clustering objects for spatial data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(5): 1003 - 1016.

[41] REZAEE M R, LELIEVELDT B P, REIBER J H. A new cluster validity index for the fuzzy c-mean[J]. Pattern Recognition Letters, 1998, 19(3 - 4): 237 - 246.

[42] 成卫青, 卢艳红. 一种基于最大最小距离和 SSE 的自适应聚类算法[J]. 南京邮电大学学报(自然科学版), 2015, 35 (2): 102 - 107.
CHENG Weiqing, LU Yanhong. Adaptive clustering algorithm based on maximum and minimum distances, and SSE[J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science), 2015, 35(2): 102 - 107. (in Chinese)

[43] 王建仁, 马鑫, 段刚龙. 改进的 K-means 聚类 k 值选择算法[J]. 计算机工程与应用, 2019, 55(8): 27 - 33.
WANG Jianren, MA Xin, DUAN Ganglong. Improved K-means clustering k-value selection algorithm [J]. Computer Engineering and Applications, 2019, 55 (8): 27 - 33. (in Chinese)

[44] WANG X, JIAO Y, FEI S. Estimation of clusters number and initial centers of K-means algorithm using watershed method[C]// 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2015: 505 - 508.

[45] 刘佳鸣, 况立群, 尹洪红, 等. 灰狼优化的 k 均值聚类算法[J]. 中国科技论文, 2019, 14(7): 778 - 782,807.
LIU Jiaming, KUANG Liqun, YIN Honghong, et al. k - means clustering algorithm based on grey wolf optimization[J]. China Sciencepaper, 2019, 14(7): 778 - 782,807. (in Chinese)

[46] ARTHUR D, VASSILVITSKII S. K - means + + : the advantages of careful seeding[C]// 18th ACM - SIAM Symposium on Discrete Algorithms, 2007: 1027 - 1035.

[47] LATTANZI S, SOHLER C. A better K - means + + algorithm via local search[C]// Proceedings of the 36th International Conference on Machine Learning, 2019: 3662 - 3671.

[48] 邓斌涛, 徐胜超. 基于动态双子种群的差分进化 K 中心点聚类算法[J]. 计算机与现代化, 2021(7): 54 - 59,70.

DENG Bintao, XU Shengchao. A differential evolution K-mediods clustering algorithm based on dynamic gemini population [J]. Computer and Modernization, 2021(7): 54 - 59,70. (in Chinese)

[49] HUANG Z. A fast clustering algorithm to cluster very large categorical data sets in data mining[J]. DMKD, 1997, 3(8): 34 - 39.

[50] DIK A, EL MOUJAHID A, JEBARI K, et al. A new dynamic algorithm for unsupervised learning[J]. Int. J. Innov. Comput. Inf. Control, 2015, 11: 1325 - 1339.

[51] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases[C]// SIGMOD'96, 1996.

[52] GUHA S, RASTOGI R, SHIM K. CURE: an efficient clustering algorithm for large databases[C]// SIGMOD'98, 1998.

[53] GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes[C]// Proceedings 15th International Conference on Data Engineering, 1999: 512 - 521.

[54] KARYPIS G, HAN E H, KUMAR V. Chameleon: hierarchical clustering using dynamic modeling[J]. Computer, 1999, 32: 68 - 75.

[55] 韩忠明, 陈妮, 张慧, 等. 一种非对称距离下的层次聚类算法[J]. 模式识别与人工智能, 2014(5): 410 - 416.
HAN Zhongming, CHEN Ni, ZHANG Hui, et al. A hierarchical clustering algorithm based on asymmetric distance[J]. Pattern Recognition and Artificial Intelligence, 2014(5): 410 - 416. (in Chinese)

[56] 李青旭, 陈天鹰, 胡波. 基于交点的新层次聚类算法[J]. 信息技术与网络安全, 2020, 39(10): 18 - 22.
LI Qingxu, CHEN Tianying, HU Bo. New hierarchical clustering algorithm based on intersection[J]. Information Techology and Network Security, 2020, 39(10): 18 - 22. (in Chinese)

[57] SHI J, ZHU Q, LI J. A novel hierarchical clustering algorithm with merging strategy based on shared subordinates[J]. Applied Intelligence, 2022, 52(8): 8635 - 8650.

[58] 缪裕青, 高韩, 刘同来, 等. 基于网格聚类的情感分析研究[J]. 中国科学技术大学学报, 2016, 46(10): 874 - 882.
MIAO Yuqing, GAO Han, LIU Tonglai, et al. Sentiment analysis based on grid clustering[J]. Journal of University of Science and Technology of China, 2016, 46(10): 874 - 882. (in Chinese)

[59] 蔡莉, 江芳, 许卫霞, 等. 一种基于自适应网格划分和决策图的聚类算法研究[J]. 小型微型计算机系统, 2019, 40(10): 2033 - 2038.
CAI Li, JIANG Fang, XU Weixia, et al. Research on a clustering algorithm based on adaptive grid partition and decision graph[J]. Journal of Chinese Computer Systems, 2019, 40(10): 2033 - 2038. (in Chinese)

[60] 王玮琪, 万仁霞, 周方祥. 局部网格动态聚类算法[J]. 现代电子技术, 2020,43(1):102 - 106.
WANG Weiqi, WAN Renxia, ZHOU Fangxiang. Dynamic clustering algorithm based on local grid[J]. Modern Electronics Technique, 2020, 43(1): 102 - 106. (in Chinese)

[61] XU X, DING S, DU M, et al. DPCG: an efficient density peaks clustering algorithm based on grid[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(5): 743 - 754.

[62] BROWN D, SHI Y. A distributed density-grid clustering algorithm for multi-dimensional data[C]// 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020: 1 - 8.

[63] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// KDD, 1996: 226 - 231.

[64] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure[J]. ACM Sigmod Record, 1999, 28(2): 49 - 60.

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

[66] BIE R, MEHMOOD R, RUAN S, et al. Adaptive fuzzy clustering by fast search and find of density peaks[J]. Personal and Ubiquitous Computing, 2016, 20(5): 785 - 793.

[67] 王治和, 曹旭琰, 杜辉. 一种优化初始点与自适应半径的密度聚类算法[J]. 计算机工程, 2022, 48(01): 51 - 59.
WANG Zhihe, CAO Xuyan, DU Hui. A density clustering algorithm with optimized initial points and adaptive radius[J]. Computer Engineering, 2022, 48(1): 51 - 59. (in Chinese)

[68] 余玥, 邓丽, 庞洪霖, 等. 基于分布式聚类的作物生长环境异常检测方法[J]. 应用科学学报, 2018, 36(6): 1010 - 1021.
YU Yue, DENG Li, PANG Honglin, et al. Environmental anomaly detection method during crop growth based on distributed clustering[J]. Journal of Applied Sciences, 2018, 36(6): 1010 - 1021. (in Chinese)

[69] GRIECO A, PACELLA M, BLACO M. On the application of text clustering in engineering change process[J]. Procedia CIRP, 2017, 62: 187 - 192.

[70] 李中衡, 杨奔, 张劲节, 等. 基于相关熵的快速聚类算法[J]. 西安交通大学学报, 2021, 55(6): 121 - 130.
LI Zhongheng, YANG Ben, ZHANG Jinjie, et al. Fast correntropy-based clustering algorithm[J]. Journal of Xi'an Jiaotong University, 2021, 55(6): 121 - 130. (in Chinese)

[71] 武斌, 周树斌, 武小红, 等. Gath - Geva 联合模糊聚类的生菜近红外光谱聚类分析[J]. 光谱学与光谱分析, 2021, 41(3): 932 - 935.
WU Bin, ZHOU Shubin, WU Xiaohong, et al. Gath - Geva allied fuzzy C-means clustering analysis of NIR spectra of lettuce[J]. Spectroscopy and Spectral Analysis, 2021, 41(3): 932 - 935. (in Chinese)

[72] JIA X, LEI T, DU X, et al. Robust self-sparse fuzzy clustering for image segmentation[J]. IEEE Access, 2020, 8: 146182 – 146195.
[73] KUMAGAI M, KOMATSU K, SATO M, et al. Ising-based combinatorial clustering using the kernel method[C]//2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 2021: 197 – 203.
[74] 诸震宇, 王如彬. 基于核聚类的嗅觉神经网络对气味模式的识别[J]. 动力学与控制学报, 2020, 18(1): 93 – 101.
ZHU Zhenyu, WANG Rubin. Odor pattern recognition of the olfactory neural network based on kernel clustering[J]. Journal of Dynamics and Control, 2020, 18(1): 93 – 101. (in Chinese)
[75] 舒月. 一种基于 PDStream 的增量聚类算法研究[D]. 武汉:武汉科技大学, 2013.
[76] CAI J, WEI H, YANG H, et al. A novel clustering algorithm based on DPC and PSO[J]. IEEE Access, 2020, 8: 88200 – 88214.
[77] CUI Z, JING X, ZHAO P, et al. A new subspace clustering strategy for AI-based data analysis in IoT system[J]. IEEE Internet of Things Journal, 2021, 8(16): 12540 – 12549.
[78] 朱颖雯, 陈松灿. 基于随机投影的高维数据流聚类[J]. 计算机研究与发展, 2020, 57(8): 1683 – 1696.
ZHU Yingwen, CHEN Songcan. High dimensional data stream clustering algorithm based on random projection[J]. Journal of Computer Research and Development, 2020, 57(8): 1683 – 1696. (in Chinese)
[79] HUANG D, WANG C D, PENG H, et al. Enhanced ensemble clustering via fast propagation of cluster-wise similarities[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 51(1): 508 – 520.
[80] 黄飞杰, 郎旭明, 尹健康, 等. 聚类集成算法在客户细分模型中的应用[J]. 西南科技大学学报, 2020, 35(1): 75 – 80.
HUANG Feijie, LANG Xuming, YIN Jiankang, et al. Application of clustering integration algorithm in customer segmentation model[J]. Journal of Southwest University of Science and Technology, 2020, 35(1): 75 – 80. (in Chinese)
[81] STREHL A, GHOSH J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3(Dec): 583 – 617.
[82] 张海涛, 宋拓, 周红磊, 等. 基于谱聚类的虚拟健康社区知识聚合方法研究[J]. 图书情报工作, 2020, 64(8): 134 – 140.
ZHANG Haitao, SONG Tuo, ZHOU Honglei, et al. Research on knowledge aggregation method of virtual healthy community based on spectral clustering[J]. Library and Information Service, 2020, 64(8): 134 – 140. (in Chinese)
[83] LI Y, CHU X, TIAN D, et al. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm[J]. Applied Soft Computing, 2021, 113: 107924.
[84] KUMAR S, SINGH M. A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem[J]. Big Data Mining and Analytics, 2019, 2(4): 240 – 247.
[85] 李斌, 狄岚, 王少华, 等. 基于改进核模糊 C 均值类间极大化聚类算法[J]. 计算机应用, 2016, 36(7): 1981 – 1987.
LI Bin, DI Lan, WANG Shaohua, et al. Clustering algorithm with maximum distance between clusters based on improved kernel fuzzy C-means[J]. Journal of Computer Applications, 2016, 36(7): 1981 – 1987. (in Chinese)
[86] 尹世庄, 王韬, 谢方方, 等. 基于互信息和轮廓系数的聚类结果评估方法[J]. 兵器装备工程学报, 2020, 41(8): 207 – 213.
YIN Shizhuang, WANG Tao, XIE Fangfang, et al. Protocol clustering evaluation method based on mutual information and contour coefficient[J]. Journal of Sichuan Ordnance, 2020, 41(8): 207 – 213. (in Chinese)
[87] RENJITH S, SREEKUMAR A, JATHAVEDAN M. Pragmatic evaluation of the impact of dimensionality reduction in the performance of clustering algorithms[M]// SENGODAN T, MURUGAPPAN M, MISRA S. Advances in electrical and computer technologies. Singapore: Springer Singapore, 2020: 499 – 512.
[88] 李顺勇, 崔文秀, 荆鹏霏. 一种基于信息熵的加权聚类算法[J]. 云南民族大学学报(自然科学版), 2020, 29(6): 568 – 576.
LI Shunyong, CUI Wenxiu, JING Pengfei. A weighted clustering algorithm based on information entropy[J]. Journal of Yunnan University of Nationalities(Natural Sciences Edition), 2020, 29(6): 568 – 576. (in Chinese)
[89] 康雁, 崔国荣, 李浩, 等. 融合自注意力机制和多路金字塔卷积的软件需求聚类算法[J]. 计算机科学, 2020, 47(3): 48 – 53.
KANG Yan, CUI Guorong, LI Hao, et al. Software requirements clustering algorithm based on self-attention mechanism and multi-channel pyramid convolution[J]. Computer Science, 2020, 47(3): 48 – 53. (in Chinese)
[90] 田镇, 陈爱华, 曹奕, 等. 红壳色文蛤选育群体遗传多样性的微卫星分析[J]. 南方农业学报, 2021, 52(9): 2582 – 2589.
TIAN Zhen, CHEN Aihua, CAO Yi, et al. Microsatellite analysis on genetic diversity of breeding populations of red shell color Meretrix meretrix[J]. Journal of Southern Agriculture, 2021, 52(9): 2582 – 2589. (in Chinese)
[91] 杨红远, 王鹏, 江炎庭, 等. 哺乳期云上黑山羊羔羊生长规律及影响因素研究[J]. 中国畜牧杂志, 2021, 57(5): 127 – 130.
[92] 姜存良, 吴勇, 邓浪, 等. 云南猕猴桃资源的收集及表型多样性分析[J]. 西南林业大学学报, 2021, 41(2): 38 – 45.
JIANG Cunliang, WU Yong, DENG Lang, et al. Kiwifruit resources collection in Yunnan Province and phenotypic diversity analysis[J]. Journal of Southwest Forestry University, 2021, 41(2): 38 – 45. (in Chinese)
[93] 程琳, 戴俊, 贺锦锋, 等. 基于球果和种子杉木品质遗传多样性分析及综合评价[J]. 南方农业学报, 2021, 52(10): 2805 – 2813.
CHENG Lin, DAI Jun, HE Jinfeng, et al. Genetic diversity analysis and comprehensive evaluation of cone and seed quality of

Cunninghamia lanceolata[J]. Journal of Southern Agriculture, 2021, 52(10): 2805 - 2813. (in Chinese)

[94] 安晓飞, 王培, 罗长海, 等. 基于 K-means 聚类和分区寻优的秸秆覆盖率计算方法[J]. 农业机械学报, 2021, 52(10): 84 - 89.
AN Xiaofei, WANG Pei, LUO Changhai et al. Corn straw coverage calculation algorithm based on K-means clustering and zoning optimization method[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(10): 84 - 89. (in Chinese)

[95] 吴彦澎. 一种基于网格的空间聚类算法及在基本农田规划中的应用[J]. 江苏农业科学, 2020, 48(16): 250 - 254.

[96] 刘佳斌, 郜允兵, 李永涛, 等. 基于高斯混合模型的土壤环境质量分区研究[J]. 农业环境科学学报, 2021, 40(8): 1746 - 1757.
LIU Jiabin, GAO Yunbing, LI Yongtao, et al. Soil environmental quality zoning based on the Gaussian mixture model[J]. Journal of Agro-Environment Science, 2021, 40(8): 1746 - 1757. (in Chinese)

[97] 王丹, 梁锦, 黄天姿, 等. 基于主成分和聚类分析的不同品种猕猴桃鲜食品质评价[J]. 食品工业科技, 2021, 42(7): 1 - 8.
WANG Dan, LIANG Jin, HUANG Tianzi, et al. Fresh food quality evaluation of kiwifruit based on principal component analysis and cluster analysis[J]. Science and Technology of Food Industry, 2021, 42(7): 1 - 8. (in Chinese)

[98] 周琪乐, 龚凌慧, 纪凤娣, 等. 红枸杞、黄枸杞和黑枸杞营养成分比较[J]. 中国酿造, 2021, 40(10): 43 - 49.
ZHOU Qile, GONG Linghui, JI Fengdi, et al. Comparison of nutritional constituents of red wolfberry, yellow wolfberry and black wolfberry[J]. China Brewing, 2021, 40(10): 43 - 49. (in Chinese)

[99] 周双, 孙兰英, 杨光, 等. 20 个品种树莓果实品质评价[J]. 食品研究与开发, 2022, 43(2): 171 - 176.
ZHOU Shuang, SUN Lanying, YANG Guang, et al. Evaluation of fruit quality of twenty varieties of raspberry[J]. Food Research and Development, 2022, 43(2): 171 - 176. (in Chinese)

[100] SHWETA S D, BARVE S S. External feature based quality evaluation of tomato using K-means clustering and support vector classification[C]// 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021: 192 - 200.

[101] 郭燕婷, 刘瑞涵, 姜修胜. 北京城镇居民鲜食玉米市场细分研究[J]. 中国农学通报, 2019, 35(32): 153 - 157.
GUO Yanting, LIU Ruihan, JIANG Xiusheng. Fresh corn consumption by Beijing urban residents: market segmentation[J]. Chinese Agricultural Science Bulletin, 2019, 35(32): 153 - 157. (in Chinese)

[102] ETUMNU C, GRAY A W. A clustering approach to understanding farmers' success strategies[J]. Journal of Agricultural and Applied Economics, 2020, 52(3): 335 - 351.

[103] JESSE O O, DANIEL O, SHEM O W, et al. A cluster analysis of variables essential for climate change adaptation of smallholder dairy farmers of Nandi County, Kenya[J]. African Journal of Agricultural Research, 2020, 16(7): 1007 - 1014.

[104] GRASKEMPER V, YU X, FEIL J H. Farmer typology and implications for policy design—an unsupervised machine learning approach[J]. Land Use Policy, 2021, 103: 105328.

[105] 许新华, 吴海燕. 面向精准农业中的高维复杂数据的降维方法[J]. 农业与技术, 2021, 41(15): 30 - 33.

[106] 王轩慧, 陈建毅, 郑西来, 等. 基于 SGA - RF 算法的农业土壤镉浓度反演研究[J]. 农业机械学报, 2018, 49(10): 261 - 269.
WANG Xuanhui, CHEN Jianyi, ZHENG Xilai, et al. Inversion of cadmium content in agriculture soil based on SGA - RF algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(10): 261 - 269. (in Chinese)

[107] WANG Y R, DING S F, WANG L J, et al. A manifold p-spectral clustering with sparrow search algorithm[J]. Soft Computing, 2022, 26(4): 1765 - 1777.

[108] 李大华, 赵辉, 于晓. 基于改进谱聚类的重叠绿苹果识别方法[J]. 光谱学与光谱分析, 2019, 39(9): 2973 - 2980.
LI Dahua, ZHAO Hui, YU Xiao. Overlapping green apple recognition based on improved spectral clustering[J]. Spectroscopy and Spectral Analysis, 2019, 39(9): 2973 - 2980. (in Chinese)