满足本地差分隐私的混合噪音感知的模糊C均值聚类算法

张朋飞[①②] 程 俊[①] 张治坤*[③] 方贤进[①] 孙 笠[④] 王 杰[⑤] 姜 茸[②]

[①](安徽理工大学计算机科学与工程学院 淮南 232001)

[②](云南省服务计算重点实验室(云南财经大学) 昆明 650221)

[③](浙江大学计算机科学与技术学院 杭州 310058)

[④](华北电力大学控制与计算机工程学院 北京 102206)

[⑤](安徽理工大学安全科学与工程学院 淮南 232001)

摘 要：在大数据和物联网应用中，本地差分隐私(LDP)技术用于保护聚类分析中的用户隐私，但现有方法要么在LDP下交互式地进行聚类，需要消耗大量隐私预算，要么没有同时考虑到聚类数据中蕴含的表示数据质量的高斯噪音以及为满足LDP保护的拉普拉斯噪音，致使聚类精度低下。同时，对于衡量用户提交数据和簇心之间的距离选择较为武断，没有充分利用到用户提交的噪音数据中蕴含的噪音模式。为此，该文创新性地提出一种满足LDP的混合噪音感知的模糊C均值聚类算法(mnFCM)，该算法的主要思想是同时建模用户上传数据中蕴含的表示用户质量的高斯噪音以及为保护用户数据注入的拉普拉斯噪音，进而设计出混合噪音感知的距离替代传统的欧式距离，来衡量样本数据与簇心间的相似性。特别地，在mnFCM中，该文首先设计了混合噪音感知的距离计算方法，在此基础上给出算法新的目标函数，并基于拉格朗日乘子法设计了求解方法，最后理论上分析了求解算法的收敛性。该文进一步理论分析了mnFCM的隐私、效用和复杂度，分析结果表明所提算法严格满足LDP、相对于对比算法更接近非隐私下的簇心以及和非隐私算法具有接近的复杂度。在两个真实数据集上的实验结果表明，mnFCM在满足LDP下，聚类精度提高了10%～15%。

关键词：聚类分析；隐私保护；本地差分隐私；模糊C均值聚类；拉普拉斯机制

中图分类号：TN911; TP391 **文献标识码**：A **文章编号**：1009-5896(2025)03-0739-19

DOI: 10.11999/JEIT241067

1 引言

如今，随着大数据分析的快速发展，个人数字信息的收集和分析得到了极大的发展。新的计算范式的出现，如云计算[1]，增加了从多个来源进行大规模分布式数据分析的可能性。对收集到的数据进行聚类和分析，并将分析结果用于现实生活中，可以改善大数据、物联网应用中的用户体验和服务质量，给人们的日常生活带来了极大的便利。聚类技术是数据挖掘、机器学习、计算机视觉、模式识别、推荐系统等数据驱动应用中最常用的方法之一[2,3]，其目的是将一个数据集依据特定准则划分为多个类或者簇，使得同一簇内的样本数据的相似性较大，不同簇内的样本数据具有较大的差异性。换言之，通过聚类，相同簇内的样本数据在聚类过程后应紧密地聚集在一起，而不同簇内的样本数据则应当被有效地分隔开。由于能够提供产品推荐和行为预测等服务，聚类技术受到了更多的关注。尽管聚类分析为挖掘有价值信息提供了更多的可能性，但由于数据提供者的主观意愿和环境等因素的影响，在聚类数据的上传过程中可能包含噪音。噪音数据的存在可能降低数据的真实性和完整性，从而影响聚类分析的准确性和可靠性。并且聚类数据往往包含用户的隐私信息，若数据拥有者未采取适当的保护措施便收集或公开这些数据，将不可避免地导致大量用户的隐私信息面临泄露的风险[4]。例如，恶意攻击者可能通过分析用户在社交媒体上发布的帖子、照片、评论以及点赞等行为，推断出用户的政治倾向、宗教信仰、社交圈子甚至是家庭状况。这种信息如果被不法分子利用，可能会导致用户遭受骚扰、诈骗或其他形式的攻击。因此，如何在不侵犯用户隐私的情况下收集不同用户的信息并实现聚类建模就显得尤为重要。

在当前的隐私保护技术领域中，本地差分隐私(Local Differential Privacy, LDP)[5,6]是本地设置下

收稿日期：2024-12-04；改回日期：2025-02-25；网络出版：2025-03-07
*通信作者： 张治坤 zhikun@zju.edu.cn
基金项目：安徽理工大学高层次引进人才科研启动基金(2023yjrc92)，云南省服务计算重点实验室开放课题 (YNSC24116)，国家自然科学基金(62202164)

Foundation Items: The Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology (2023yjrc92), The Foundation of Yunnan Key Laboratory of Service Computing (YNSC24116), The National Natural Science Foundation of China (62202164)

标准差分隐私的一种变体，作为在不可信服务器存在的情况下保护用户隐私的理想模型而受到广泛关注。基本上，在LDP的保护下，用户在将数据发送给服务提供商之前，先在本地扰动数据以实现差分隐私。然后，服务提供商只能分析用户数据，而不能收集用户敏感信息。因此，即使服务器拥有大量的设备信息，也无法通过观察设备的输出来推断出敏感数据。换而言之，本地差分隐私技术使设备能够自行处理私有数据。与随机干扰和数据交换等隐私保护技术相比[7]，本地差分隐私在聚类分析中的数据隐私保护方面具有明显的优势。

目前，基于本地差分隐私的聚类研究领域中，关注点和探索方向主要集中于基于划分的聚类算法的实现与优化[8–13]。然而，这种基于划分的聚类算法需要大量的迭代交互，大量的迭代导致隐私预算的分配和个体噪声的增加，从而影响聚类的效用。同时，基于本地差分隐私的聚类工作没有考虑到聚类数据中蕴含的表示数据质量的高斯噪音以及为满足LDP保护注入的拉普拉斯噪音，用户提交数据和簇心之间的距离选择较为武断，从而导致聚类精度低下。

因此，为了在保护聚类数据隐私的同时考虑表示数据质量的高斯噪音和为满足LDP保护注入的拉普拉斯噪音，本文在模糊C均值聚类算法中利用混合噪音机制生成新的噪音感知距离，以之替代传统的欧氏距离来度量样本数据与聚类质心之间的相似性，创新性地提出了一种在本地差分隐私保护框架下采用混合噪音机制的模糊C均值聚类算法。由于没有专门处理噪音的能力，欧式距离容易受到数据集中异常值或极端值的影响，导致聚类边界模糊不清，难以区分不同簇之间的差异。相比之下，本文设计的混合噪音感知距离不仅考虑了数据本身的随机波动(如高斯噪音)，还特别针对人为添加的保护性噪音(拉普拉斯噪音)进行了优化。拉普拉斯噪音因其随机性和无界性，往往会导致用户提交的数据中出现大量离群点。通过改进，即使在存在显著噪音的情况下，所设计的距离度量方法仍能保持较高的稳定性，从而确保聚类结果的可靠性和一致性。

本文的贡献如下：

(1) 提出一种满足LDP的混合噪音感知的模糊C均值聚类算法(mixed noise-aware Fuzzy C-Means, mnFCM)。其主要思想在于同时建模用户上传数据中蕴含的表示用户质量的高斯噪音以及为保护用户数据注入的拉普拉斯噪音，进而设计出混合噪音感知的距离替代传统的欧式距离，来衡量样本数据与簇心间的相似性。

(2) 在mnFCM中，本文首先设计了混合噪音感知的距离计算方法，在此基础上给出算法新的目标函数，并基于拉格朗日乘子法设计了求解方法，最后理论上分析了求解算法的收敛性。

(3) 理论分析了mnFCM的隐私、效用和复杂度，分析结果表明所提算法严格满足LDP、相对于对比算法更接近非隐私下簇心以及和非隐私算法具有接近的复杂度。在两个真实数据集上的实验结果表明，mnFCM在满足LDP下，聚类精度提高了10%～15%。

2 相关工作

2.1 基于差分隐私的聚类方法

Li等人[14]针对差分隐私k均值(k-means)聚类算法面临的聚类可用性和收敛性难题，提出了一种基于遗传算法的隐私预算分配策略。通过优化初始质心的选择和设定最小隐私预算来确保算法的收敛性；同时，它将隐私预算的分配转化为组合优化问题，利用遗传算法寻找最优的隐私预算分配方案，以此提升算法的可用性和效率。Liu等人[15]提出了一种结合机器学习与差分隐私的指纹图像发布算法。此算法通过匹配指纹图像中的特征点并对其进行保护性处理，使用聚类算法对图像进行初步分割，再利用多项式回归算法对分割后的区域进行精确处理，构建回归模型以确定保护区域。为了应对特征点位置不确定性带来的分割难题，研究引入了动态隐私预算分配方法，利用指数机制在差分隐私框架下动态调整隐私预算，以实现指纹图像的局部保护发布。石江南等人[16]提出了一种分布式环境下支持差分隐私的k-means++聚类算法。该算法通过内存计算引擎Spark,创建弹性分布式数据集，利用转换算子及行动算子操作数据进行运算，并在选取初始化中心点及迭代更新中心点的过程中，通过综合利用指数机制和拉普拉斯机制，以解决初始聚类中心敏感及隐私泄露问题，同时减少计算过程中对数据实施的扰动。Wu等人[17]针对现有差分隐私聚类算法在处理非凸数据集时表现不佳的问题，提出了一种新的基于密度的差分隐私聚类算法。该算法通过在密度估计阶段采用拉普拉斯机制向密度添加噪声，以及在聚类扩展阶段实施隐私预算分配方案，有效提高了非凸数据集上聚类的隐私保护能力和聚类质量。Fang等人[18]提出了一种新的差分隐私k-means聚类算法，通过将数据投影到低维空间，私密生成候选中心集，最后在这些候选中心上执行隐私保护聚类，有效解决了传统k-means算法在处理高维稀疏(High-Dimensional Sparse, HiDS)数据

时存在的隐私泄露问题。Diaa等人[19]通过优化差分隐私和安全计算技术，开发出一种既快速又保护隐私、同时保持或提升聚类准确性的方法。特别地，该方法采用计算差分隐私模型，设计了一种轻量级的安全聚合技术，实现了比现有技术快4个数量级的处理速度。此外，通过引入约束聚类技术，进一步提高了聚类的实用性。

2.2 满足本地差分隐私的聚类方法

近年来，本地差分隐私下基于划分的聚类工作引起研究者的持续关注。在基于划分的聚类(如K-means)研究中，许多现有方法依赖于迭代式的交互过程来完成聚类任务，这不仅增加了算法的复杂度，还导致了大量隐私预算的消耗。例如，Luo等人[8]提出了个性化本地差分隐私K-means算法(Personalized Local Differential Privacy K-means, PLDP K-means)，使用了一种扰动机制在用户端扰动用户数据，然后通过本地和服务器端之间的迭代完成聚类。每次迭代都需要分配一定的隐私预算，从而限制了整体的隐私保护水平。一些方法未能同时考虑到聚类数据中固有的高斯噪音以及为满足LDP保护而添加的拉普拉斯噪音。如张少波等人[9]主要关注如何在LDP框架下实现K-means聚类，但忽略了噪音对聚类效果的影响，导致其在处理噪音数据时表现不佳。Xia等人[10]首次设计出一种结合本地差分隐私的K-means聚类机制，他们的机制首先通过扰动用户数据以满足LDP，然后修改了传统的K-means算法，使服务提供商能够通过与高度扰动数据的用户合作获得高质量的聚类结果。他们还提出了一个扩展机制，在每次迭代中都扰动用户的敏感数据和用户聚类的中间结果，以提高基本模型在隐私和效用方面的性能。然而，由于在每次迭代中都要对用户的敏感数据和中间结果进行扰动，这会导致隐私预算的快速消耗。随着迭代次数的增加，可用的隐私预算逐渐减少，从而影响后续迭代的质量。为了确保严格的隐私保护，某些方法不得不牺牲聚类的效用。例如，Lin等人[11]针对高维数据发布和分析中参与者隐私面临的风险，提出了一种基于本地差分隐私的高维数据发布算法(Privacy-Utility Balanced publishing with budgeta-ware local differential privacy, PU_Bpub)，旨在解决现有方法中普遍存在的高计算开销和来自不可信服务器的隐私威胁问题。该算法虽然有效保护了高维数据发布的隐私，但在一定程度上影响了数据的可用性和聚类质量。在寻求聚类分析中数据隐私保护与可用性并重的解决方案时，张国鹏等人[12]设计了一种运用LDP技术来强化聚类过程中的数据隐私保护算法，确保数据在聚类分析时其隐私信息得到充分的屏蔽与防护。此算法显著增加了数据间的区分度，巧妙地规避了局部最优解的困扰，从而显著提升了聚类算法的稳定性。此算法虽然巧妙地规避了局部最优解的困扰，但在某些复杂的优化问题中，仍然可能存在难以完全避免的局部极值问题。此外，这种方法可能依赖于特定的初始化条件或参数设置，缺乏通用性。Li等人[13]提出了一种基于LDP和安全聚合的安全联邦推荐系统方案，且通过基于LDP的K-means聚类算法更好地选择参与每轮训练的用户。旨在实现类似于集中训练的性能，同时平衡数据隐私保护和训练效率。但是其使用的聚类算法主要用于联邦学习环境下的推荐系统，对其他应用场景的通用性有待验证。并且仅依靠LDP机制进行隐私保护，缺乏对多种噪音类型的综合考虑。

综上所述，已有研究要么需要在聚类过程中进行大量的迭代交互，从而导致隐私预算的消耗，影响聚类算法的性能；要么没有考虑到聚类数据中包含的表示数据质量的高斯噪音和为满足本地差分隐私保护的拉普拉斯噪音，距离的选择较为武断，从而导致聚类精度低下。因此，关于如何在保护聚类数据隐私的同时有效降低噪音数据对聚类结果的影响，目前依旧是一个亟待探索与解决的开放性议题。

3 预备知识

3.1 本地差分隐私

在LDP中，个人(即，数据所有者)在通过扰动私有化数据之后将其数据发送到数据聚合器。因此，这些技术为个人(即，数据所有者)提供了合理的可否认性。数据聚合器收集所有扰动值，并对统计数据进行估计，例如每个值在总体中的频率。形式定义如下：

定义1　ε-本地差分隐私[20,21](ε-LDP):如果对于任意输出v_1, v_2，且对于任意子集$S \subseteq \mathrm{Range}(A)$，则随机化机制$A$满足$\varepsilon$-LDP

$$\Pr[A(v_1) \in S] \leq \mathrm{e}^{\varepsilon} \cdot \Pr[A(v_2) \in S] \tag{1}$$

与差分隐私[22]相比，LDP为数据所有者提供了更多的保护。除了将私有数据直接发送到受信任的聚合器之外，数据所有者可以使用满足ε-LDP的机制来扰动他们的私有数据，然后释放扰动后的数据。

定理1(后处理)[23]如果机制A满足ε-LDP，那么对于任意函数f，$f(A)$也满足ε-LDP。

定理2(并行组合性)[24,25]假设有一组隐私机制算法$A=\{A_1, A_2, \cdots, A_m\}$，在一组不相交数据集上分别满足$\varepsilon_i$-LDP，则组合算法$A$满足($\max\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m\}$)-LDP。

定理3(串行组合性)[24,25] 假设有一组隐私机制算法$A=\{A_1,A_2,\cdots,A_m\}$，在同一个数据集上分别满足ε_i-LDP，则组合算法A满足$\left(\sum_{i=1}^{m}\varepsilon_i\right)$-LDP。

3.2 模糊C均值聚类算法

模糊C均值聚类算法(Fuzzy C-Means, FCM)算法[26,27]在模式识别等领域中被频繁的使用。其糅合了模糊逻辑的核心思想，提供了一种更为灵活且适应性强的聚类结果，允许数据点以不同程度的隶属度同时属于多个聚类。在大部分场景下，数据集中的样本对象难以被清晰地界定并严格分配到互不重叠的类别之中，直接将它们归类于某一特定类别往往显得过于僵化，且可能引发误判的风险。因此，FCM为每个对象和每个类别分配了一个权值，即隶属度，用以表示对象隶属于该类别的程度。诚然，基于概率的方法同样能给出类似的权值，但在某些情况下，确定一个合适的统计模型可能颇具挑战。因此，鉴于FCM算法固有的灵活性和其不依赖于明确概率分布的性质，它成为了一个更加恰当且适用的选择来处理相关问题。

FCM是硬C-均值聚类(Hard C-Means clustering, HCM)的延伸与拓展[28]，它们之间最大的区别就在于隶属度的取值不同，HCM的隶属度只能取0或1中的一个值，而FCM中隶属度的范围是连续的[0,1]区间内的任意实数。除此之外，FCM还需要选取恰当的加权指数m，它的选取对结果有一定的影响，取值范围为$1<m<+\infty$。FCM的目标函数如式(2)所示

$$J_{\mathrm{FCM}}=\sum_{i=1}^{N}\sum_{c=1}^{C}(u_{ci})^{m}\mathrm{Dist}(\boldsymbol{x}_i,\boldsymbol{p}_c) \tag{2}$$

$$\sum_{c=1}^{C}u_{ci}=1,\quad i=1,2,\cdots,N \tag{3}$$

其中，$\mathrm{Dist}(\boldsymbol{x}_i,\boldsymbol{p}_c)$表示样本数据$\boldsymbol{x}_i$和质心$\boldsymbol{p}_c$之间的欧氏距离，用于量化两者间的空间间隔，$u_{ci}$表示样本数据划分隶属度。

3.3 问题定义

假设存在一个数据集$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}\subset\mathbb{R}^{K\times N}$，该数据集包含了$N$个数据样本对象，这些对象在多维空间中分布，并可用于进一步的分析和处理。其中K表示每个数据样本属性的维数。FCM采用了欧氏距离作为样本间的相似性度量，将数据集X划分成C个簇[29]，其中u_{ci}满足$\sum_{c=1}^{C}u_{ci}=1$。在划分过程中，通常含有用户的隐私信息，从而存在泄露敏感信息的风险。ε定义为每个样本数据的隐私预算，用来衡量隐私保护水平，较小的ε提供更好的隐私。服务器希望在保证每个用户的信息遵循ε-LDP的同时得到样本数据与质心之间的距离，然后根据样本数据与质心样本之间的距离对数据进行聚类分析。本文的目标是设计一种基于LDP的模糊C均值聚类算法，能够充分利用用户提交的噪音数据中蕴含的噪音模式，在保护用户隐私的同时提高聚类精度。表1总结了本文中常用的符号。

4 mnFCM算法

4.1 混合噪音感知的距离计算

在大数据和物联网应用场景下，聚类数据中天然存在着表示数据质量的高斯噪音，同时为满足本地差分隐私(LDP)保护要求，需额外注入拉普拉斯噪音。其中高斯噪音源于数据收集过程中的多种因素，如传感器精度限制、环境干扰等，其反映了数据本身的不确定性与质量状况。选择高斯噪音纳入考量，旨在贴合数据实际特性，挖掘数据噪音模式蕴含的潜在信息，提升聚类精度。而拉普拉斯噪音是实现 LDP 的关键要素，LDP 确保用户数据隐私，使服务器无法精准推断原始数据。拉普拉斯机制依据数据敏感度与隐私预算分配噪音，在保证隐私前提下，控制噪音量对数据效用的影响。本文将两者结合的核心在于构建更精准的混合分布模型，设计混合噪音感知距离衡量样本数据与簇心之间的相似性，取代传统欧式距离。这一创新设计能同时兼顾数据固有质量特征与隐私保护需求，避免单一噪音模型缺陷，提升聚类性能与稳定性。

假设存在N个样本数据对象，K是每个数据对象属性的维数，并且加噪后的样本数据表示为$\hat{\boldsymbol{x}}_i$，其位于K维空间$\mathbb{R}^K$中。在本文的模型中，假设每个样本数据$\hat{\boldsymbol{x}}_i$都是一个以$\boldsymbol{p}_c$ (簇c的质心样本)为中心、协方差为$\boldsymbol{\sigma}_i^2\boldsymbol{I}_K$的混合分布，假设$\alpha$是高斯分布在混合分布的比例，1–$\alpha$是拉普拉斯分布在混合分布中所占的比例。每个未知参数$\boldsymbol{\sigma}_i\geq 0$表示样本数据$\hat{\boldsymbol{x}}_i$与其对应簇心$\boldsymbol{p}_c$之间的相似性程度，$\boldsymbol{\sigma}_i$越小，样本数据越接近质心样本。本文强加了一个超参数

表 1 常用符号列表

符号	符号含义
ε	隐私预算
C, N, K	簇、样本数据以及样本属性的个数
u_{ci}	样本数据的划分隶属度
D_{ci}	样本数据与簇心间的混合噪音感知的距离
x_i, $\boldsymbol{p}_c$	样本数据i、簇c的质心
$\hat{\boldsymbol{x}}_i$	加噪后的样本数据i
m, γ	模糊度参数、收敛终止参数
τ	迭代阈值

$\boldsymbol{\sigma}_0 \geq 0$，并且要求每个$i \in [1,N]$时，$\boldsymbol{\sigma}_i \geq \boldsymbol{\sigma}_0$。它合理地被解释为所有样本数据的相似性的上界。

给定该概率模型下的样本数据集$\boldsymbol{X}=\{\hat{\boldsymbol{x}}_i\}_{i=1}^N \subset \mathbb{R}^{K\times N}$和超参数$\boldsymbol{\sigma}_0$，需要找到一个点$\boldsymbol{p}_c$，该点最大化似然函数为

$$\begin{aligned}&\prod_{i=1}^{N}\left[\alpha\mathcal{N}(\hat{\boldsymbol{x}}_i \mid \boldsymbol{p}_c, \boldsymbol{\sigma}_i^2\boldsymbol{I}_K)+(1-\alpha)\mathcal{L}(\hat{\boldsymbol{x}}_i \mid \boldsymbol{p}_c, 2\boldsymbol{\sigma}_i^2\boldsymbol{I}_K)\right]\\&=\prod_{i=1}^{N}\left(\alpha\left(\frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_i}\right)^K\exp\left[-\frac{\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2}{2\boldsymbol{\sigma}_i^2}\right]+(1-\alpha)\right.\\&\left.\cdot\left(\frac{1}{2\boldsymbol{\sigma}_i}\right)^K\exp\left[-\frac{\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|}{\boldsymbol{\sigma}_i}\right]\right)\end{aligned}\tag{4}$$

取负对数并优化所有有效向量上的量$\boldsymbol{\sigma}=(\boldsymbol{\sigma}_i)_{i\in[1,N]}$能得到优化问题为

$$\begin{aligned}\min_{\boldsymbol{p}_c,\boldsymbol{\sigma}}&\left\{\frac{N\cdot K}{2}\ln(2\pi)-N\ln\alpha-N\ln(1-\alpha)\right.\\&+\sum_{i=1}^{n}\left\{K\cdot\ln\boldsymbol{\sigma}_i+\frac{\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2}{2\boldsymbol{\sigma}_i^2}+K\cdot\ln(2\boldsymbol{\sigma}_i)\right.\\&\left.\left.+\frac{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\boldsymbol{\sigma}_i}\right\}\right\}\\&\text{s.t.}\quad \boldsymbol{\sigma}_i\geq\boldsymbol{\sigma}_0\end{aligned}\tag{5}$$

引理1　对于固定的$\boldsymbol{p}_c \in \mathbb{R}^K$，有以下向量$\boldsymbol{\sigma}_i$的取值，使上述目标函数最小化，即

$$\boldsymbol{\sigma}_i=\max\left\{\boldsymbol{\sigma}_0, |\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|/\sqrt{K}\right\}\tag{6}$$

根据**引理1**，优化问题现在只取决于点$\boldsymbol{p}_c$

$$\begin{aligned}\min_{\boldsymbol{p}_c,\boldsymbol{\sigma}}&\left\{\frac{NK}{2}\ln(2\pi)-N\ln\alpha-N\ln(1-\alpha)\right.\\&+\sum_{\frac{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\sqrt{K}}<\boldsymbol{\sigma}_0}\left[K\ln\boldsymbol{\sigma}_0+\frac{\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2}{2{\boldsymbol{\sigma}_0}^2}\right.\\&\left.+K\ln(2\boldsymbol{\sigma}_0)+\frac{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\boldsymbol{\sigma}_0}\right]\\&+\sum_{\frac{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\sqrt{K}}\geq\boldsymbol{\sigma}_0}\left[K\ln\frac{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\sqrt{K}}+\frac{K}{2}\right.\\&\left.\left.+K\ln\frac{2|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|}{\sqrt{K}}+\sqrt{K}\right]\right.\end{aligned}\tag{7}$$

注意，调整$\hat{\boldsymbol{x}}_i$, $\boldsymbol{\sigma}_0$和所有$\boldsymbol{p}_c$，使得只通过一个常数加项改变函数的值。为了简化，本文将对三重体$(\boldsymbol{p}_c, \boldsymbol{\sigma}_0, \{\hat{\boldsymbol{x}}_i\}_{i=1}^N) \mapsto (\boldsymbol{p}_c', \boldsymbol{\sigma}_0'\cdot\{\hat{\boldsymbol{x}}_i'\}_{i=1}^N)$应用一个缩放，使得$\boldsymbol{\sigma}_0=1/\sqrt{K}$，并且去掉撇号。目标函数则变为

$$\begin{aligned}\min_{\boldsymbol{p}_c}&\left\{\frac{NK}{2}\ln(2\pi)-N\ln\alpha-N\ln(1-\alpha)\right.\\&+\sum_{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|<1}\left[-K\ln K+\frac{K\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2}{2}+K\ln 2\right.\\&\left.+\sqrt{K}\,|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|\right]\\&\left.+\sum_{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|\geq1}\left[\frac{K}{2}+K\ln\frac{2\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2}{K}+\sqrt{K}\right]\right.\end{aligned}\tag{8}$$

此外，可以去掉常数项

$$\begin{aligned}&\frac{NK}{2}\ln(2\pi)-N\ln\alpha-N\ln(1-\alpha)+2NK\ln 2\\&\quad-2NK\ln K+\frac{NK}{2}+N\sqrt{K}\end{aligned}\tag{9}$$

然后将整个函数除以$K/2$，目标函数变成

$$\begin{aligned}\min_{\boldsymbol{p}_c}&\left(\sum_{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|<1}\left(\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2+\frac{2\sqrt{K}}{K}|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|\right)\right.\\&\left.+\sum_{|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|\geq1}2\ln\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2\right)\end{aligned}\tag{10}$$

简化上述优化问题得到

$$\min_{\boldsymbol{p}_c}\sum_{i=1}^{N}f(|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|)\tag{11}$$

其中，$f(\ell)=\begin{cases}\ell^2+\dfrac{2\sqrt{K}}{K}\ell, & 0\leq\ell<1\\ 2\ln\ell^2, & \ell\geq1\end{cases}$。

因此，经过混合噪音机制得到混合噪音感知的距离为

$$D_{ci}=\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2+\frac{2\sqrt{K}}{K}|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|+2\ln|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|^2\tag{12}$$

4.2 目标函数设计

mnFCM采用混合噪音感知的距离度量样本间的相似性，故mnFCM的目标函数可以由式(13)表述，样本划分隶属度约束式(14)是其满足的约束条件

$$J_{\text{mnFCM}}=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^m D_{ci}\tag{13}$$

$$\sum_{c=1}^{C}u_{ci}=1,\quad i=1,2,\cdots,N\tag{14}$$

其中，$D_{ci}=\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2+\dfrac{2\sqrt{K}}{K}|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c|+2\ln\|\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\|^2$，是经过混合噪音机制得到的混合噪音感知的距离。$m$为模糊度参数和加权指数，其定义域为$(1,+\infty)$，主要调控聚类结果的模糊程度，即数据点属于各个聚类中心的隶属度分布的模糊性。具体而言，m的

值越大，算法生成的聚类结果就越趋向于模糊，意味着每个数据点以更高的模糊性同时关联于多个聚类中心，从而可能反映出数据内在的更加复杂或重叠的结构特性。相反，m的值越小，聚类结果就越趋于清晰，数据点的隶属度分布更加明确，接近于传统的硬聚类方法。

4.3 算法求解

求解在具有约束条件式(14)下mnFCM的最小值，可以采用拉格朗日乘子法见式(15)

$$J=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^{m}D_{ci}+\sum_{i=1}^{N}\lambda_i\left(\sum_{c=1}^{C}u_{ci}-1\right) \tag{15}$$

其中，λ_i是对应于约束条件式(14)的拉格朗日乘子。

(1)对$\boldsymbol{u}_{ci}$求偏导

$$\left.\begin{aligned}
J&=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^{m}D_{ci}+\sum_{i=1}^{N}\lambda_i\left(\sum_{c=1}^{C}u_{ci}-1\right)\\
\frac{\partial J}{\partial u_{ci}}&=mu_{ci}^{m-1}D_{ci}+\lambda_i=0\\
&\Leftrightarrow mu_{ci}^{m-1}D_{ci}=-\lambda_i\\
&\Leftrightarrow u_{ci}=\left(-\frac{\lambda_i}{mD_{ci}}\right)^{\frac{1}{m-1}}\\
&\Leftrightarrow u_{ci}=\left(-\frac{mD_{ci}}{\lambda_i}\right)^{\frac{-1}{m-1}}=-\frac{(mD_{ci})^{\frac{-1}{m-1}}}{(\lambda_i)^{\frac{-1}{m-1}}}
\end{aligned}\right\} \tag{16}$$

又$\partial J/\partial\lambda_i=\sum_{c=1}^{C}u_{ci}-1=0$，代入式(16)得

$$\left.\begin{aligned}
&\sum_{c=1}^{C}\left[-\frac{(mD_{ci})^{\frac{-1}{m-1}}}{(\lambda_i)^{\frac{-1}{m-1}}}\right]=1\\
&\Leftrightarrow-\frac{\sum\limits_{c=1}^{C}(mD_{ci})^{\frac{-1}{m-1}}}{(\lambda_i)^{\frac{-1}{m-1}}}=1\\
&\Leftrightarrow-\sum_{c=1}^{C}(mD_{ci})^{\frac{-1}{m-1}}=(\lambda_i)^{\frac{-1}{m-1}}
\end{aligned}\right\} \tag{17}$$

所以

$$u_{ci}=\frac{(mD_{ci})^{\frac{-1}{m-1}}}{\sum\limits_{c=1}^{C}(mD_{ci})^{\frac{-1}{m-1}}}=\frac{(D_{ci})^{\frac{-1}{m-1}}}{\sum\limits_{c=1}^{C}(D_{ci})^{\frac{-1}{m-1}}} \tag{18}$$

(2)对$\boldsymbol{p}_c$求偏导

$$\frac{\partial J}{\partial\boldsymbol{p}_c}=\sum_{i=1}^{N}u_{ci}^{m}\left[-2\left(\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\right)-\frac{4}{\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c}-\frac{2\sqrt{K}}{K}\mathrm{sgn}\left(\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\right)\right]=0 \tag{19}$$

(a)当$\hat{\boldsymbol{x}}_i>\boldsymbol{p}_c$时

$$\begin{aligned}
\frac{\partial J}{\partial\boldsymbol{p}_c}&=\sum_{i=1}^{N}u_{ci}^{m}\left[-2\left(\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\right)-\frac{2\sqrt{K}}{K}\mathrm{sgn}\left(\hat{\boldsymbol{x}}_i-\boldsymbol{p}_c\right)\right.\\
&\left.-\frac{4}{\hat{\boldsymbol{x}}_i}-\frac{4}{\hat{\boldsymbol{x}}_i^2}\boldsymbol{p}_c+O\left(\boldsymbol{p}_c\right)\right]=0\\
&\Leftrightarrow\sum_{i=1}^{N}u_{ci}^{m}\left[-2\hat{\boldsymbol{x}}_i-\frac{2\sqrt{K}}{K}-\frac{4}{\hat{\boldsymbol{x}}_i}+2\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]\right.\\
&\left.\cdot\boldsymbol{p}_c+O\left(\boldsymbol{p}_c\right)\right]=0\\
&\Leftrightarrow\boldsymbol{p}_c=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}+\frac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]}
\end{aligned} \tag{20}$$

(b)当$\hat{\boldsymbol{x}}_i<\boldsymbol{p}_c$时，同理可得

$$\boldsymbol{p}_c=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}-\frac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]} \tag{21}$$

综上所述

$$\boldsymbol{p}_c=\begin{cases}\dfrac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}+\frac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]},\hat{\boldsymbol{x}}_i>\boldsymbol{p}_c\\[2ex]\dfrac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}-\frac{2\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]},\hat{\boldsymbol{x}}_i<\boldsymbol{p}_c\end{cases} \tag{22}$$

mnFCM算法实现的具体步骤见算法1。

4.4 收敛性分析

引理2 根据式(18)迭代更新u_{ci}，不会增加式(13)中目标函数J_{mnFCM}的值。

证明 把目标函数J_{mnFCM}看成只关于u_{ci}的函数$J(U)$

$$J\left(U\right)=\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}^{m}D_{ci} \tag{23}$$

经由拉格朗日乘子法，利用式(18)计算出的u^*值是函数$J(U)$的一个驻点。为了证明u^*为极小值点，只需证明海森矩阵$\nabla^2J\left(u^*\right)$在$u^*$处是正定矩阵。

算法 1　mnFCM算法

输入：样本数据集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，模糊度参数$m$，收敛终止参数$\gamma$，最大迭代次数$\tau_{\max}$；

输出：样本划分隶属度u_{ci}，簇心$\boldsymbol{p}_c$

/*本地端*/

for i=1, 2, $\cdots$, N do

　调用Laplace机制对样本数据进行加噪；

　用户将加噪后的样本数据$\hat{\boldsymbol{x}}_i$上传给服务器；

end for

/*服务器端*/

设置初始迭代次数：$\tau=0$；

随机初始化u_{ci}使其满足$0\leq u_{ci}\leq 1$;

迭代开始：

根据式(22)更新$\boldsymbol{p}_c$;

根据式(12)更新D_{ci};

根据式(18)更新u_{ci};

更新迭代次数$\tau=\tau+1$;

迭代终止条件 $\max|u_{ci}(\tau)-u_{ci}(\tau-1)|\leq\gamma$或者$\tau=\tau_{\max}$。

return u_{ci}, $\boldsymbol{p}_c$;

$$\begin{aligned}&\nabla^2 J(u)\\&=\begin{bmatrix}\dfrac{\partial^2 J(u)}{\partial u_{11}\partial u_{11}} & \cdots & \dfrac{\partial^2 J(u)}{\partial u_{11}\partial u_{CN}}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial^2 J(u)}{\partial u_{CN}\partial u_{11}} & \cdots & \dfrac{\partial^2 J(u)}{\partial u_{CN}\partial u_{CN}}\end{bmatrix}\\&=\begin{bmatrix}m(m-1)D_{11}u_{11}^{m-2} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & m(m-1)D_{CN}u_{CN}^{m-2}\end{bmatrix}\end{aligned}\tag{24}$$

对于$\boldsymbol{u}^*$, $u_{ci}\geq 0$, $1<m<+\infty$且$D_{ci}\geq 0$，所以海森矩阵$\nabla^2 J(u^*)$是正定矩阵。由多元函数极值存在的充分必要条件可知$\boldsymbol{u}^*$是目标函数J_{mnFCM}的局部最小值点，所以根据式(18)迭代更新u_{ci}不会增加式(13)中目标函数J_{mnFCM}的值。

引理3　根据式(22)迭代更新$\boldsymbol{p}_c$，不会增加式(13)中目标函数J_{mnFCM}的值。

证明　与引理2的证明方法相同。

引理4　式(13)中的目标函数J_{mnFCM}有界，即存在常数M，使得$J_{\text{mnFCM}}\geq M$。

证明　因为u_{ci}的最小值为0，且$D_{ci}\geq 0$，所以式(13)中目标函数$J_{\text{mnFCM}}\geq 0$，所以当$M=0$时，$J_{\text{mnFCM}}\geq M$，即目标函数J_{mnFCM}有界。

定理4　mnFCM算法会收敛于满足约束条件式(14)的局部最小值。

证明　引理2和引理3证明了mnFCM目标函数向局部最小值方向更新，引理4证明了mnFCM目标函数存在下限，所以mnFCM目标函数J_{mnFCM}必定收敛于满足约束条件式(14)的局部最小值。

5　算法分析

本节通过对mnFCM的隐私、效用和复杂度进行全面分析，从理论上证明其优越性。

5.1　隐私分析

定理5　mnFCM算法满足ε-LDP。

证明　从理论层面剖析，mnFCM 严格遵循 LDP 准则。在mnFCM算法中，只在算法的初始阶段对样本数据x_i进行扰动，每个样本数据的属性个数为K，那么每个样本数据t的隐私预算为$\varepsilon_t=\varepsilon/K$。根据定理3 LDP的串行组合性，其满足$\varepsilon$-LDP。后续迭代更新阶段仅处理加噪数据，接触不到原始数据，因此不消耗隐私预算，根据定理1 LDP的后处理性质，后续阶段也满足ε-LDP。根据定理2 LDP的并行组合性，mnFCM算法对所有用户的数据满足ε-LDP。证毕。

5.2　效用分析

由于在聚类分析的过程中，簇心的选取有着至关重要的作用。因此，如果mnFCM算法比仅考虑为满足LDP注入的拉普拉斯噪音的模糊聚类算法的簇心更接近非隐私模糊聚类算法的簇心，便可认为根据mnFCM算法求得的聚类结果精度更高。

定理6　mnFCM算法求得的簇心更接近非隐私下的簇心，即$|\boldsymbol{p}_1-\boldsymbol{p}_0|<|\boldsymbol{p}_2-\boldsymbol{p}_0|$。

证明　设$\boldsymbol{p}_1$表示mnFCM算法下的簇心，$\boldsymbol{p}_0$表示非隐私保护下的簇心，$\boldsymbol{p}_2$表示在仅考虑为满足LDP注入的拉普拉斯噪音下的模糊聚类算法的簇心。只需证明$|\boldsymbol{p}_1-\boldsymbol{p}_0|<|\boldsymbol{p}_2-\boldsymbol{p}_0|$，便可说明mnFCM算法求得的簇心更接近非隐私下的簇心

$$\boldsymbol{p}_1=\begin{cases}\dfrac{\sum_{i=1}^{N}u_{ci}^m\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}+\dfrac{\sqrt{K}}{K}\right]}{\sum_{i=1}^{N}u_{ci}^m\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}, \hat{\boldsymbol{x}}_i>\boldsymbol{p}_c\\ \dfrac{\sum_{i=1}^{N}u_{ci}^m\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}-\dfrac{\sqrt{K}}{K}\right]}{\sum_{i=1}^{N}u_{ci}^m\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}, \hat{\boldsymbol{x}}_i<\boldsymbol{p}_c\end{cases}\tag{25}$$

$$\boldsymbol{p}_2=\begin{cases}\dfrac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)},\hat{\boldsymbol{x}}_i>\boldsymbol{p}_c\\ \dfrac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)},\hat{\boldsymbol{x}}_i<\boldsymbol{p}_c\end{cases}\tag{26}$$

$$\boldsymbol{p}_0=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}x_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\tag{27}$$

分别讨论$\hat{\boldsymbol{x}}_i>\boldsymbol{p}_c$和$\hat{\boldsymbol{x}}_i<\boldsymbol{p}_c$两种情况：

(1)当$\hat{\boldsymbol{x}}_i>\boldsymbol{p}_c$时

$$\boldsymbol{p}_1-\boldsymbol{p}_0=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}+\dfrac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}>0\tag{28}$$

$$\begin{aligned}\boldsymbol{p}_2-\boldsymbol{p}_0&=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\\&=-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}<0\end{aligned}\tag{29}$$

所以，根据式(28)和式(29)可知$\boldsymbol{p}_1-\boldsymbol{p}_0>0$，$\boldsymbol{p}_2-\boldsymbol{p}_0<0$，则

$$\begin{aligned}&|\boldsymbol{p}_1-\boldsymbol{p}_0|-|\boldsymbol{p}_2-\boldsymbol{p}_0|\\&=(\boldsymbol{p}_1-\boldsymbol{p}_0)-(\boldsymbol{p}_0-\boldsymbol{p}_2)=\boldsymbol{p}_1+\boldsymbol{p}_2-2\boldsymbol{p}_0\\&=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}+\dfrac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}+\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\\&=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}+\dfrac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\\&=\frac{\left[\begin{array}{l}\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}+\dfrac{\sqrt{K}}{K}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}-\sum\limits_{i=1}^{N}u_{ci}^{m}\left(1+\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\\-2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\end{array}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\dfrac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}}\end{aligned}\tag{30}$$

其中，令$z_1=\sum_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+2/\hat{\boldsymbol{x}}_i+\sqrt{K}/K\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}$，$z_2=\sum_{i=1}^{N}u_{ci}^{m}\left(1+\sqrt{K}/\hat{\boldsymbol{x}}_i\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}$，$z_3=2\sum_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i\cdot\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)$，$z_4=\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}$。

显然$z_1<z_2+z_3$，因为$z_4>0$，则$(z_1-z_2-z_3)/z_4<0$。那么$|\boldsymbol{p}_1-\boldsymbol{p}_0|-|\boldsymbol{p}_2-\boldsymbol{p}_0|<0$，即$|\boldsymbol{p}_1-\boldsymbol{p}_0|<|\boldsymbol{p}_2-\boldsymbol{p}_0|$。

(2)当$\hat{\boldsymbol{x}}_i<\boldsymbol{p}_c$时

$$\boldsymbol{p}_1-\boldsymbol{p}_0=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\dfrac{2}{\hat{\boldsymbol{x}}_i}-\dfrac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\dfrac{2}{\hat{\boldsymbol{x}}_i^2}\right]}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}>0\tag{31}$$

$$\boldsymbol{p}_2-\boldsymbol{p}_0=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}=-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}<0 \tag{32}$$

所以，根据式(31)和式(32)可知$\boldsymbol{p}_1-\boldsymbol{p}_0>0$, $\boldsymbol{p}_2-\boldsymbol{p}_0<0$，则

$$\begin{aligned}
&|\boldsymbol{p}_1-\boldsymbol{p}_0|-|\boldsymbol{p}_2-\boldsymbol{p}_0|\\
&=(\boldsymbol{p}_1-\boldsymbol{p}_0)-(\boldsymbol{p}_0-\boldsymbol{p}_2)=\boldsymbol{p}_1+\boldsymbol{p}_2-2\boldsymbol{p}_0\\
&=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}-\frac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]}+\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(-\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\\
&=\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}-\frac{\sqrt{K}}{K}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]}-\frac{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)}-\frac{2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i}{\sum\limits_{i=1}^{N}u_{ci}^{m}}\\
&=\frac{\left[\begin{array}{l}\sum\limits_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+\frac{2}{\hat{\boldsymbol{x}}_i}-\frac{\sqrt{K}}{K}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}-\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i}-1\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\\ -2\sum\limits_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\end{array}\right]}{\sum\limits_{i=1}^{N}u_{ci}^{m}\left[1-\frac{2}{\hat{\boldsymbol{x}}_i^2}\right]\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}\left(\frac{\sqrt{K}}{\hat{\boldsymbol{x}}_i^2}\right)\cdot\sum\limits_{i=1}^{N}u_{ci}^{m}}
\end{aligned} \tag{33}$$

其中，令$z_5=\sum_{i=1}^{N}u_{ci}^{m}\left[\hat{\boldsymbol{x}}_i+2/\hat{\boldsymbol{x}}_i-\sqrt{K}/K\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}$，$z_6=\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i-1\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}$，$z_3=2\sum_{i=1}^{N}u_{ci}^{m}\boldsymbol{x}_i\cdot\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)$, $z_4=\sum_{i=1}^{N}u_{ci}^{m}\left[1-2/\hat{\boldsymbol{x}}_i^2\right]\cdot\sum_{i=1}^{N}u_{ci}^{m}\left(\sqrt{K}/\hat{\boldsymbol{x}}_i^2\right)\cdot\sum_{i=1}^{N}u_{ci}^{m}$。

显然$z_5<z_6+z_3$，因为$z_4>0$，则$(z_5-z_6-z_3)/z_4<0$。那么$|\boldsymbol{p}_1-\boldsymbol{p}_0|-|\boldsymbol{p}_2-\boldsymbol{p}_0|<0$，即$|\boldsymbol{p}_1-\boldsymbol{p}_0|<|\boldsymbol{p}_2-\boldsymbol{p}_0|$。

综上所述，$|\boldsymbol{p}_1-\boldsymbol{p}_0|<|\boldsymbol{p}_2-\boldsymbol{p}_0|$，证毕。

5.3 复杂度分析

从方案设计可以看出，算法的主要时间消耗在数据传送，迭代以及参数更新上。首先样本数据集X的大小决定了算法的基本操作次数，假设存在一个包含N个数据样本对象的数据集$\boldsymbol{X}$，其中N是数据样本对象的个数。其次是迭代次数，算法直到收敛为止，因此迭代次数取决于收敛速度以及停止准则，假设迭代次数是τ。在每次迭代过程中需要更新样本划分隶属度$\boldsymbol{u}_{ci}$和聚类中心$\boldsymbol{p}_c$，算法的目的是将数据集X划分成C个簇，因此可以得出mnFCM的时间复杂度为$O(CN\tau)$。

6 实验结果与分析

6.1 实验设置

6.1.1 数据集描述

在本文的实验中，使用以下两个真实数据集。

20NewsGroups 语料库[30]：是一个常用的自然语言处理和文本分类任务的基准数据集，其集成了约20 000篇新闻组文章，这些文章均衡地覆盖了20个不同的主题类别。这些主题类别确实在语义上有一定的差异，但有的主题之间可能更为接近或相关，而有的则相对独立。

UW-CAN数据集[31]：是一个包含314个Web页面的集合，这些页面主要来自滑铁卢大学和其他加拿大大学的官方网站，已经被预先分类到10个不同的类别中，适用于各种Web挖掘任务，特别是文本分类和聚类分析。数据集的真实性和代表性使其成为一个理想的基准数据集，用于评估新的Web挖掘算法和技术的性能。

模拟数据集Syn：MNIST 数据集由70 000张28×28像素的灰度图像组成，分为60 000张训练图

像和10 000张测试图像，涵盖10个类别。该数据集主要用于机器学习算法的测试与性能基准比较。

6.1.2 评价指标和实验环境

本文采用F-Measure和Entropy作为评估聚类效果的指标。

(1)F-Measure[32]：精确度和召回率的综合指标，其值越高说明算法聚类效果越好。$F_1 = 2 \cdot P \cdot R/(P+R)$。其中，$R$(召回率)表示结果类与标准类交集的样本对象个数除以标准类的样本对象个数，P(精确度)表示结果类与标准类交集的样本对象个数除以结果类的样本对象个数。

整个聚类算法的F-Measure是对每个标准类i进行加权平均得到的，可以表示成$F = \sum_{i=1}^{C} N_i \times F_1 / \sum_{i=1}^{C} N_i$。其中，$N_i$表示标准类$i$的样本对象个数。

(2)Entropy(熵指数)[33]：基于各类别中样本分布的均匀性来计算。当各类别中的样本分布越趋于均衡时，Entropy值会相应减小，这标志着聚类效果更为理想；反之，若Entropy值变大，则说明聚类结果愈发混乱，各类别样本分布不均衡。结果类c的Entropy可以表示为：$e_c = -\sum_{c=1}^{C} P\log_2 P$。

聚类算法的整体熵指数是通过对所有结果类别c进行加权求和得到的结果。$E = \sum_{c=1}^{C} N_c/N e_c$。其中，$N_c$表示结果类$c$的样本对象个数。

本文搭建的所有实验环境和所有算法的实现均在一台Inter i9-14900HX, 32 G内存的24核32线程笔记本电脑上进行，采用Python 3.7来编写所有算法。特别地，迭代停止阈值γ=0.001。

6.1.3 对比方法

本文将mnFCM算法与非隐私(None Privacy, NoPriv)算法、隐私K-Means(Privacy K-Means, PrivKM)算法[14]、隐私距离(Privacy Distance, PrivDis)算法[15]和隐私强化(Privacy Professional, PrivPro)算法[16]进行了比较。其中，NoPriv算法是非隐私下的FCM算法，PrivKM算法是K-means算法的隐私保护版本，PrivDis算法则是对FCM算法中的样本数据与质心间的距离进行隐私保护，PrivPro算法是普通隐私保护下的FCM算法的强化版本。

6.2 实验结果

6.2.1 参数的影响

实验1 隐私预算ε的影响。

在本实验中，设置固定的模糊度参数m=2，以此来判断不同隐私预算ε对各算法性能的影响。

实验结果如图1所示，其中的图1(a)、图1(b)、图1(c)和图1(d)分别展示了在数据集NG和UW上，不同隐私预算(ε=0.1, 0.3, 0.5, 0.7, 1)下的聚类效果对比。从中可以明显看出，无论在哪个数据集中，所有算法的性能均低于非隐私算法NoPriv，这是由于隐私保护机制中注入了噪音，从而降低了聚类算法的效用。此外，随着隐私预算ε的增加，各聚类算法的性能普遍呈现出逐渐提升的趋势。这是因为更高的隐私预算允许在保护隐私的同时引入更少的噪音，从而降低对聚类过程的干扰，使得聚类结果更加贴近真实情况，进而提高聚类的准确性和效果。特别地，本文提出的mnFCM的性能都优于PrivKM, PrivDis和PrivPro。具体而言，在第1个数据集NG中，当ε变化时，mnFCM的F-Measure仅比非隐私版本NoPriv降低了11.5%～15.3%，而PrivKM, PrivDis和PrivPro相比NoPriv则分别降低了14.2%～19.2%, 21.1%～25%和16.9%～21.1%。同样，mnFCM的Entropy仅比NoPriv提高了9%～17.6%，而PrivKM, PrivDis和PrivPro则分别提高了12.7%～20.9%, 23.4%～32.5%和16.7%～23.2%。这是因为，mnFCM创新性地考虑了聚类数据中表示数据质量的高斯噪音和为满足LDP保护注入的拉普拉斯噪音，设计出一种混合噪音感知的距离来改进样本间的相似性度量方式，从而在有效保护隐私的同时，更好地兼顾了聚类性能。在UW数据集上的实验结果与数据集NG上的实验结果相似，进一步验证了mnFCM算法的有效性和优越性。

实验2 模糊度参数m的影响。

文献[34]通过聚类有效性的实验发现m的最佳选取区间为[1.5, 2.5]。因此在本实验中，设置模糊度参数m的值分别为1.5, 1.75, 2, 2.25和2.5，并设置固定的隐私预算ε=0.5，以此来判断不同模糊度参数m对各算法性能的影响。

实验结果如图2所示，图2(a)、图2(b)、图2(c)和图2(d)分别呈现了数据集NG和UW在不同模糊度参数m下的聚类效果对比。从中可以看出，在所有数据集中，相较于非隐私保护算法NoPriv，其他算法的性能均有所降低，这是因为其他的算法都为了实现隐私保护注入了噪音，降低了聚类算法的效用。此外可以发现，当m=2时，各算法的F-Measure值达到峰值，而Entropy值则降至最低点。这一结果强有力地验证了m=2是所有算法在聚类效果上的最佳平衡点，既不过于模糊也不过于确定，确保了聚类结果的准确性和稳定性。进一步观察发现，随着m值的减小，所有算法的实验结果逐渐趋近于PrivKM算法。这是因为m值越接近于1，其聚类划分越趋向于硬划分，即模糊C均值聚类算法越接近于传统的硬聚类算法。

尤为突出的是，本文提出的mnFCM算法在不同m值下均展现出卓越的性能优势。其F-Measure

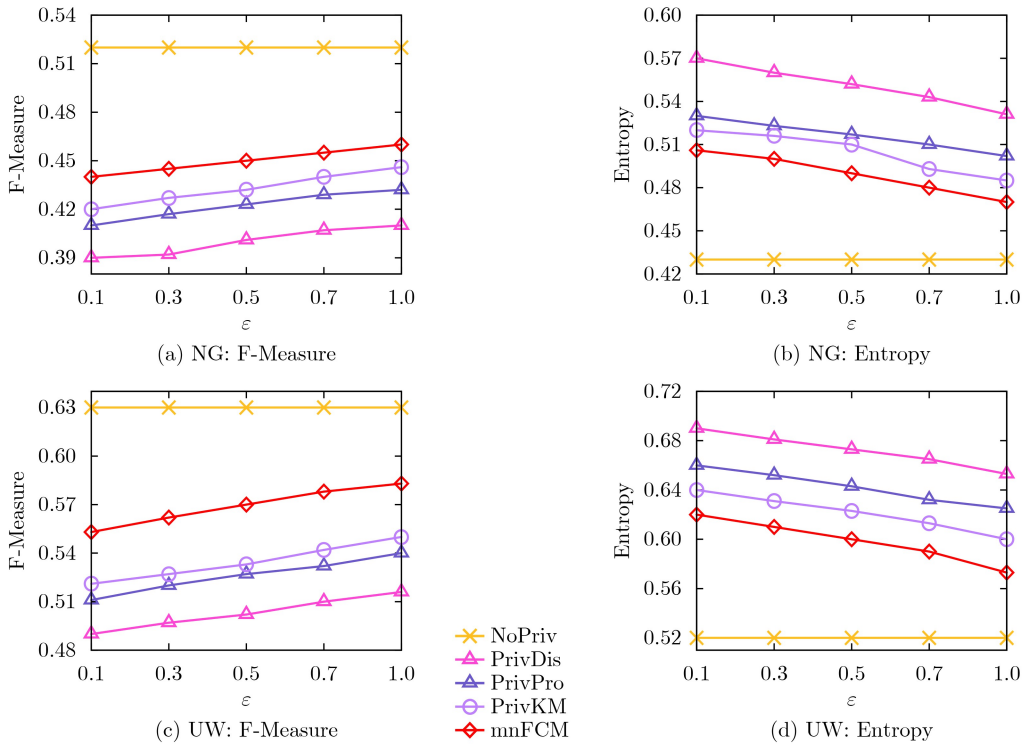


图 1 NG和UW数据集上不同隐私预算ε对各算法性能的影响

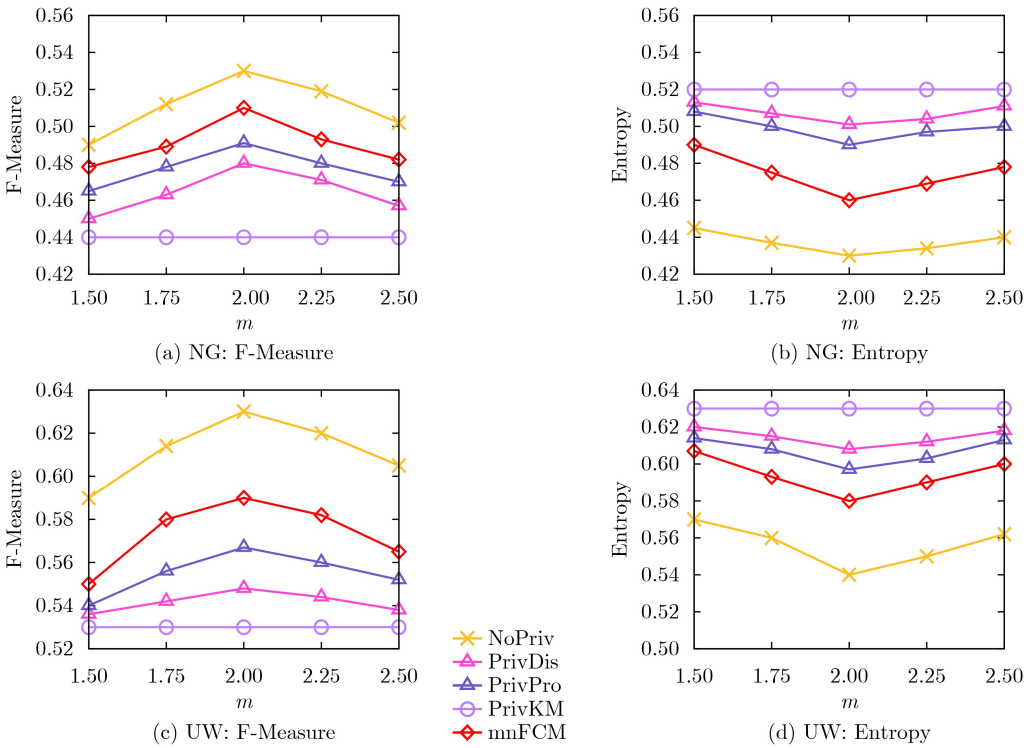


图 2 NG和UW数据集上不同模糊度参数m对各算法性能的影响

值显著高于PrivKM, PrivDis和PrivPro算法，而Entropy值则明显低于这3种算法，充分证明了mnFCM算法在聚类效果上的优越性。这一结果不仅体现了mnFCM算法在隐私保护聚类领域的创新价值，也为后续研究提供了坚实的基础和广阔的展望。

6.2.2 迭代次数

本文分别在数据集NG和UW上对所提算法进行迭代，来观察mnFCM在不同迭代次数下的收敛情况。从图3可以看出，不论在数据集NG还是数据集UW上，目标函数值随着迭代次数的增加均呈现

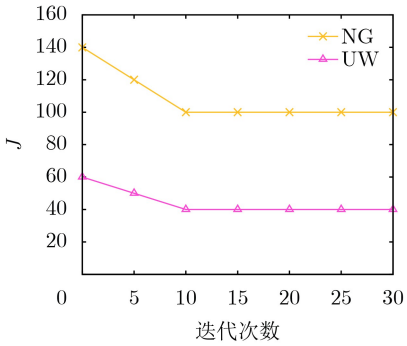


图 3 NG和UW数据集上目标函数值随迭代次数的变化情况

逐渐下降的趋势。目标函数值越小，算法的聚类性能越强。尤为显著的是，当迭代次数大于10时，目标函数值逐渐趋于稳定。这是因为迭代次数超过10次后，算法达到了预设的迭代终止条件，即算法收敛。

以上实验结果与本文的理论分析紧密契合，不仅证实了mnFCM算法的确具有收敛性，而且凸显了其卓越的收敛速度，这对于提升聚类任务的执行效率至关重要。综上所述，mnFCM算法在保持隐私保护的同时，还通过其高效的收敛机制，为聚类分析领域带来了新的活力与优势。

6.2.3 运行时间

聚类算法的运行时间(即聚类时间)同样是衡量其性能优劣的一个重要指标。它反映了算法处理数据集所需的时间成本，对于实际应用中的效率考量至关重要。NoPriv, PrivDis, PrivPro, PrivKM和mnFCM算法的时间复杂度都可以表示成$O(CN\tau)$，其中τ为迭代次数。

如表2所示是mnFCM算法与NoPriv, PrivDis, PrivPro和PrivKM算法的聚类时间对比，可以看到在NG和UW两个数据集上，mnFCM算法的运行时间略高于非隐私算法NoPriv。这是因为相对于NoPriv，在本文的方法中增加了噪音类型的考量。与NoPriv算法直接采用欧式距离度量样本间相似性的简洁方法不同，mnFCM融入了更精细的噪音处理机制，从而增加了更多的距离计算上的开销。然而，值得注意的是，尽管mnFCM在处理噪音上付出了更多计算资源，但从表2中可以发现，mnFCM算法的运行时间远比PrivDis, PrivPro和PrivKM算法的运行时间要短。原因在于mnFCM算法同时考虑了聚类数据中蕴含的表示数据质量的高斯噪音以及为满足LDP保护注入的拉普拉斯噪音，设计了一种混合噪音感知的距离来度量样本间的相似性，充分利用了用户提交的噪音数据中蕴含的噪音模式。这种双重噪音处理的精妙平衡，不仅保障了数据的隐私安全，还极大提升了聚类性能，使得mnFCM在效率与效果上均展现出对其他对比算法的显著优势。

表 2 算法运行时间对比

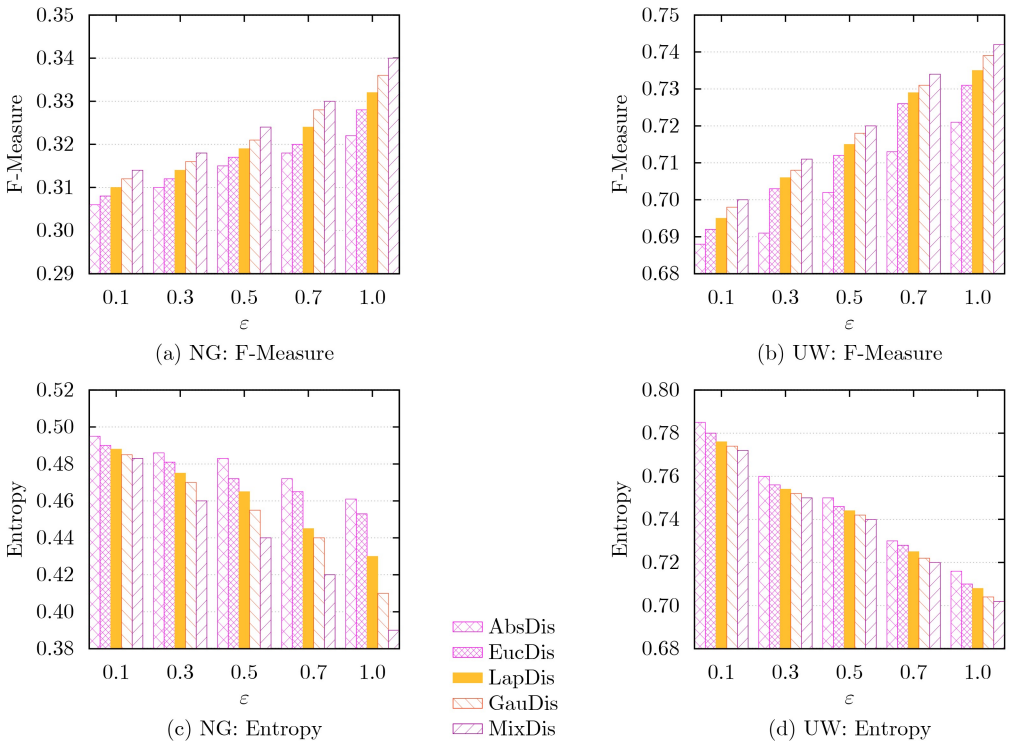
对比算法	NG (min)	UW(s)
NoPriv	5	8.33
PrivDis	19	23.67
PrivPro	26	44.36
PrivKM	24	34.63
mnFCM	7	8.42

6.2.4 消融实验

为了凸显本文所提出算法的显著优势，特此设计并实施了以下消融实验作为验证手段。

为了全面评估mnFCM算法的有效性，本文精心挑选了4个具有代表性的对比方案进行深入研究，它们是AbsDis, EucDis, LapDis和GauDis。具体而言，AbsDis方案用样本数据与质心样本之间的绝对值距离来表示样本间的相似性度量；EucDis方案则进一步引入了更为精确的欧氏距离来表示样本间的相似性度量；LapDis方案仅考虑为满足LDP注入的拉普拉斯噪音，用拉普拉斯距离来评估样本间的相似性度量；而GauDis方案则聚焦于聚类数据中蕴含的表示数据质量的高斯噪音，通过高斯距离来度量样本间的相似性。而本文的MixDis方案同时建模用户上传数据中蕴含的表示用户质量的高斯噪音以及为保护用户数据注入的拉普拉斯噪音，进而设计出混合噪音感知的距离替代传统的欧式距离，来衡量样本数据与簇心间的相似性。通过将MixDis方案与这四个精心设计的对比方案进行对比分析，本文能够全面、客观地评估mnFCM算法在隐私保护聚类任务中的表现，从而验证其有效性和优越性。

图4展示了MixDis的有效性，有以下观察结果。首先，比较了5个方案的F-Measure值，如图4(a)和图4(b)所示。结果表明，当隐私预算ε增加时，所有方案的聚类效果都会提高。因为当隐私预算ε增加时，LDP机制允许的隐私损失量增加，从而聚类精度提高。正如预期所示，考虑噪音的方法在两个数据集上的表现都明显优于AbsDis和EucDis，这表明考虑噪音可以提高聚类的准确性。其次，GauDis的表现优于LapDis。原因在于，高斯噪音的连续性和平滑性有助于减少噪音对数据本身特性的影响，从而更准确地度量样本间的相似性。同时，高斯机制在隐私保护和噪音量控制方面也具有一定的优势。图4(c)和图4(d)展示了在NG和UW两个数据集上AbsDis, EucDis, LapDis, GauDis和MixDis的



(a) NG: F-Measure　(b) UW: F-Measure　(c) NG: Entropy　(d) UW: Entropy

图 4 NG和UW数据集上不同隐私预算下聚类效果对比

Entropy值。与F-Measure值不同的是，Entropy值随隐私预算ε增加而减少。由图4可以看出，与其他4个对比方案相比，在数据集NG上，MixDis方案的F-Measure值分别提高了3.4%(相对于GauDis)到28.5%(相对于AbsDis)，并且Entropy值分别降低了2.7%(GauDis)～18.3%(AbsDis)。而在数据集UW上，MixDis方案的F-Measure值分别提高了5.5%(相对于GauDis)～18.7%(相对于AbsDis)，并且Entropy值分别降低了1.9%(GauDis)～16.6%(AbsDis)。由以上的对比分析可知，MixDis方案的聚类效果不论在NG数据集上还是UW数据集上均表现最优。这是因为上传数据中同时存在为保护数据隐私注入的拉普拉斯噪音和固有的高斯噪音，同时考虑它们可以使经过聚类后的数据对噪音更加鲁棒。综上所述，本实验验证了MixDis方案在mn-FCM算法中的有效性，其通过综合考虑多种噪音因素，不仅提高了聚类算法的性能，还确保了数据的隐私安全。

6.3 模拟数据集Syn

为了更全面地评估算法的普适性，我们进一步扩展了实验范围，新增了模拟数据集Syn。这部分实验设置了4个可调参数，分别是隐私预算ε、样本量N、维度K和分簇数C，我们通过变换可调参数的值来观察算法的性能。

6.3.1 密度聚类算法对比

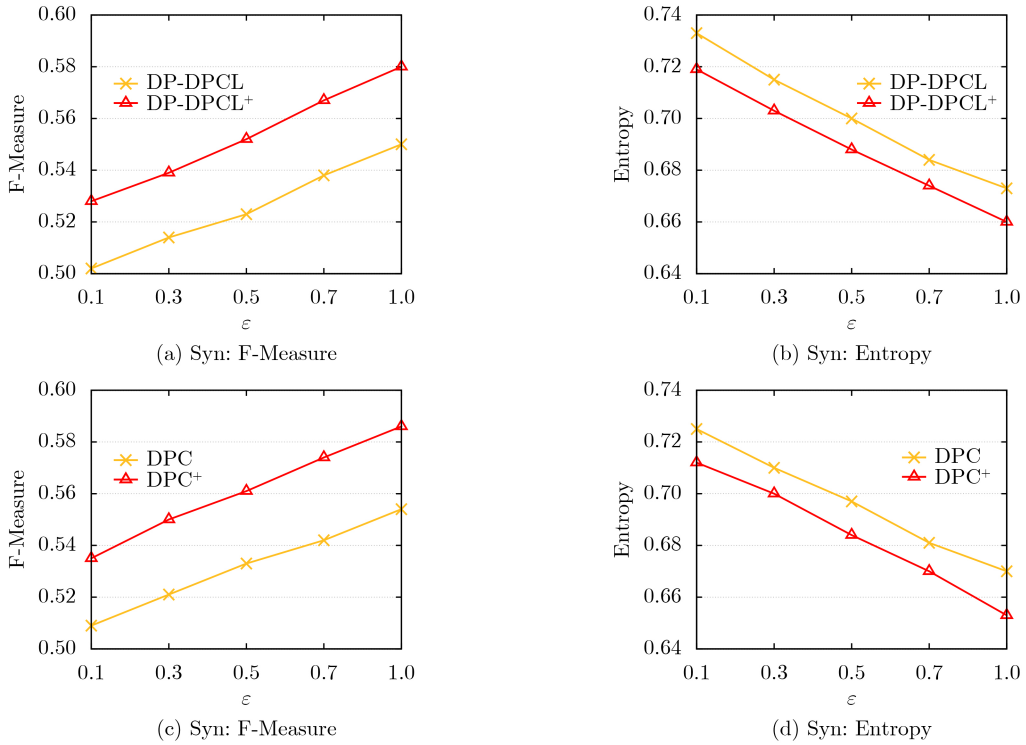
为了进一步全面地评估算法的普适性，本文在Syn数据集上与其他致力于基于隐私保护的聚类算法(如基于密度的聚类算法)进行了对比分析，包括基于局部密度加噪的密度峰值聚类差分隐私保护算法(Differentially Private Density Peaks Clustering with Local-density adding noise, DP-DPCL)算法[35]、密度峰值聚类(Density Peaks Clustering, DPC)算法[36]。其中，DP-DPCL算法是一种改进的密度峰值聚类算法的差分隐私保护方案，DPC算法则是一种基于差分隐私的自适应聚类中心密度峰值聚类算法。而DP-DPCL$^+$和DPC$^+$算法则是在DP-DPCL算法和DPC算法上分别应用本文提出的混合噪音感知距离方案。

从图5(a)和图5(b)中可以看出，随着隐私预算ε的增加，DP-DPCL算法和DP-DPCL$^+$算法的F-Measure值都逐渐提高。同时，Entropy都逐渐下降。但DP-DPCL$^+$算法效果始终好于 DP-DPCL算法。同样的，图5(c)和图5(d)的实验结果表明DPC$^+$的算法效果也始终好于DPC算法。原因如下：(1)更精确的距离度量：混合噪音感知距离方案通过综合考虑数据质量的高斯噪音和为保护隐私而注入的拉普拉斯噪音，提供了更精确的距离度量，减少了聚类误差。(2)更好的噪声处理：该方案能够有效识别和处理不同类型的噪音，即使在较低的隐私预算下也能保持较高的聚类性能。(3)更高的稳定性：混合噪音感知距离方案增强了算法的稳定性，确保聚类结果更加准确和可靠。

6.3.2 参数变化结果

图6(a)和图6(b)展示了在改变隐私预算ε时，所有方法的F-Measure和Entropy情况。从中可以观察到随着ε的增加，所有方法的F-Measure值都有所提高，这表明随着隐私保护程度的降低(即ε值增大)，模型的性能有所提升。同时，所有方法的Entropy值都随着ε的增加而减小，这意味着模型的不确定性减少，分类更加明确。其中，mnFCM方法在所有ε值下都表现出较高的F-Measure值和较低的Entropy值，显示出较好的聚类效果和稳定性。总的来说，随着隐私预算的增加，模型的性能普遍提高，而mnFCM方法在不同隐私预算下的表现均优于其他方法。

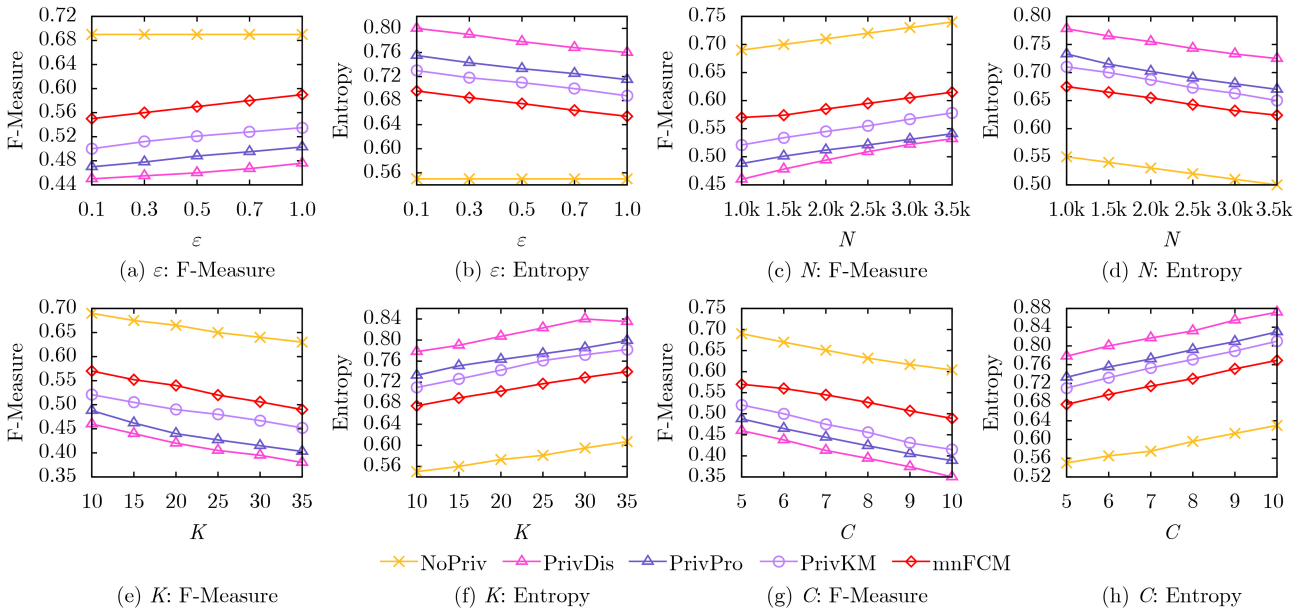
图6(c)和图6(d)展示了在改变样本量N时，所



(a) Syn: F-Measure (b) Syn: Entropy

(c) Syn: F-Measure (d) Syn: Entropy

图 5 Syn数据集上不同隐私预算下密度聚类算法效果对比



(a) ε: F-Measure (b) ε: Entropy (c) N: F-Measure (d) N: Entropy

(e) K: F-Measure (f) K: Entropy (g) C: F-Measure (h) C: Entropy

图 6 Syn数据集上各算法在不同参数下的聚类效果对比

有方法的F-Measure和Entropy情况。从中我们观察到mnFCM的F-Measure值至少比PrivDis高出69.6%，且Entropy值至少比PrivDis低10.2%。此外，随着样本量N的增大，所有方法的F-Measure也变大，同时所有方法的Entropy都变小。这是因为当样本量较小时，模型可能会过度适应训练数据，即过拟合。更大的样本量可以帮助模型学习到更普遍的模式，而不是特定于训练集的特征，从而提高泛化能力，算法的效果越好。

图6(e)和图6(f)显示了当改变维度K时，不同方法的F-Measure和Entropy的变化情况。从图6(e)和图6(f)中，观察到mnFCM的F-Measure值在所有维度下都比PrivDis至少高29.4%。此外，本文还观察到，随着维度K的增大，所有方法的F-Measure都变小，而Entropy则变大。原因在于，维度K的值越大，每个样本的数据复杂度就越高，从而导致算法的性能越差。因此，在样本量N不变的情况下，F-Measure会在一定程度上减小，而Entropy则会增加。

图6(g)和图6(h)展示了在改变分簇数C时，所有方法的F-Measure和Entropy情况。通过图6中可以注意到，随着C值的增加，所有方法的F-Measure值都有所下降。这表明随着分簇数的增加，模型可能导致过拟合现象，从而导致模型的性能有所降低。同时，所有方法的Entropy值都随着C值的增加而增加。这意味着模型的不确定性增加，分类更加模糊，每个簇的纯度可能会降低(即每个簇内部的同质性减少)。其中，mnFCM方法在所有C值下相较于其他对比算法都表现出较高的F-Measure值和较低的Entropy值，显示出较好的聚类效果和稳定性。

7　结束语

本文针对模糊聚类算法中聚类数据的隐私保护问题，同时考虑了表示数据质量的高斯噪音和为满足LDP保护注入的拉普拉斯噪音，创新性地提出了一种满足本地差分隐私的混合噪音感知的模糊C均值聚类算法mnFCM，实现了隐私安全与聚簇质量的兼顾。本文首先设计了混合噪音感知的距离计算方法，在此基础上给出算法新的目标函数，并基于拉格朗日乘子法设计了求解方法，最后理论上分析了求解算法的收敛性。进一步理论分析了算法的隐私、效用和复杂度，分析结果表明所提算法严格满足LDP、与对比算法相比更接近非隐私下的簇心以及和非隐私算法具有相近的复杂度。在公开的标准数据集上进行实验，实验结果表明本文算法在真实数据集上的效用均优于隐私下的对比算法，聚类精度提高了10%～15%。但本文的不足之处在于，在混合噪声中拉普拉斯噪声的隐私预算计算可能受高斯噪声影响。在未来的研究中，本文将进一步探索自适应噪声比例分配策略(如动态调整高斯/拉普拉斯噪声权重)，以优化隐私-效用权衡。

参 考 文 献

[1] MISTRY H K, MAVANI C, GOSWAMI A, *et al*. The impact of cloud computing and AI on industry dynamics and competition[J]. *Educational Administration: Theory and Practice*, 2024, 30(7): 797–804. doi: 10.53555/kuey.v30i7.6851.

[2] 常璐瑶, 牛新征, 罗涛, 等. 基于子博弈完美均衡的启发式聚类算法[J]. 电子学报, 2024, 52(3): 740–750. doi: 10.12263/DZXB.20221206.
CHANG Luyao, NIU Xinzheng, LUO Tao, *et al*. Heuristic clustering algorithm based on sub-game perfect equilibrium[J]. *Acta Electronica Sinica*, 2024, 52(3): 740–750. doi: 10.12263/DZXB.20221206.

[3] 黄鹤, 李文龙, 杨澜, 等. 跳跃跟踪SSA交叉迭代AP聚类算法[J]. 电子学报, 2024, 52(3): 977–990. doi: 10.12263/DZXB.20220209.
HUANG He, LI Wenlong, YANG Lan, *et al*. Jump tracking SSA hybrid iterative AP clustering algorithm[J]. *Acta Electronica Sinica*, 2024, 52(3): 977–990. doi: 10.12263/DZXB.20220209.

[4] 张强, 叶阿勇, 叶帼华, 等. 最优聚类的k-匿名数据隐私保护机制[J]. 计算机研究与发展, 2022, 59(7): 1625–1635. doi: 10.7544/issn1000-1239.20210117.
ZHANG Qiang, YE Ayong, YE Guohua, *et al*. k-Anonymous data privacy protection mechanism based on optimal clustering[J]. *Journal of Computer Research and Development*, 2022, 59(7): 1625–1635. doi: 10.7544/issn1000-1239.20210117.

[5] LIN Wanyu, LI Baochun, and WANG Cong. Towards private learning on decentralized graphs with local differential privacy[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2936–2946. doi: 10.1109/TIFS.2022.3198283.

[6] 傅培旺, 丁红发, 刘海, 等. 基于本地差分隐私的分布式图统计采集算法[J]. 计算机研究与发展, 2024, 61(7): 1643–1669. doi: 10.7544/issn1000-1239.202330628.
FU Peiwang, DING Hongfa, LIU Hai, *et al*. Statistics collecting algorithms of distributed graph via local differential privacy[J]. *Journal of Computer Research and*

Development, 2024, 61(7): 1643–1669. doi: 10.7544/issn1000-1239.202330628.

[7] 李宗维, 孔德潮, 牛媛争, 等. 基于人工智能和区块链融合的隐私保护技术研究综述[J]. 信息安全研究, 2023, 9(6): 557–565. doi: 10.12379/j.issn.2096-1057.2023.06.08.
LI Zongwei, KONG Dechao, NIU Yuanzheng, *et al*. Towards a privacy-preserving research for AI and blockchain integration[J]. *Journal of Information Security Research*, 2023, 9(6): 557–565. doi: 10.12379/j.issn.2096-1057.2023.06.08.

[8] LUO Yuling, WANG Zhangrui, ZHANG Shunsheng, *et al*. Efficient-secure *k*-means clustering guaranteeing personalized local differential privacy[C]. 22nd International Conference on Algorithms and Architectures for Parallel Processing, Copenhagen, Denmark, 2023: 660–675. doi: 10.1007/978-3-031-22677-9_35.

[9] 张少波, 原刘杰, 毛新军, 等. 基于本地差分隐私的K-modes聚类数据隐私保护方法[J]. 电子学报, 2022, 50(9): 2181–2188. doi: 10.12263/DZXB.20201374.
ZHANG Shaobo, YUAN Liujie, MAO Xinjun, *et al*. Privacy protection method for K-modes clustering data with local differential privacy[J]. *Acta Electronica Sinica*, 2022, 50(9): 2181–2188. doi: 10.12263/DZXB.20201374.

[10] XIA Chang, HUA Jingyu, TONG Wei, *et al*. Distributed *K*-Means clustering guaranteeing local differential privacy[J]. *Computers & Security*, 2020, 90: 101699. doi: 10.1016/j.cose.2019.101699.

[11] LIN Aixin and MA Xuebin. PU_Bpub: High-dimensional data release mechanism based on spectral clustering with local differential privacy[C]. 17th International Conference on Wireless Algorithms, Systems, and Applications, Dalian, China, 2022: 572–581. doi: 10.1007/978-3-031-19214-2_48.

[12] 张国鹏, 陈学斌, 王豪石, 等. 面向本地差分隐私的K-Prototypes聚类方法[J]. 计算机应用, 2022, 42(12): 3813–3821. doi: 10.11772/j.issn.1001-9081.2021101724.
ZHANG Guopeng, CHEN Xuebin, WANG Haoshi, *et al*. K-Prototypes clustering method for local differential privacy[J]. *Journal of Computer Applications*, 2022, 42(12): 3813–3821. doi: 10.11772/j.issn.1001-9081.2021101724.

[13] LI Weiqing, CHEN Hongyu, ZHAO Ruifeng, *et al*. A federated recommendation system based on local differential privacy clustering[C]. 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), Atlanta, USA, 2021: 364–369. doi: 10.1109/SWC50871.2021.00056.

[14] LI Yong, SONG Xiao, TU Yuchun, *et al*. GAPBAS: Genetic algorithm-based privacy budg*et al*location strategy in differential privacy K-means clustering algorithm[J]. *Computers & Security*, 2024, 139: 103697. doi: 10.1016/j.cose.2023.103697.

[15] LIU Chao, ZHI Zhaolong, ZHAO Weinan, *et al*. Research on local fingerprint image differential privacy protection method based on clustering algorithm and regression algorithm segmentation image[J]. *IEEE Access*, 2024, 12: 27127–27146. doi: 10.1109/ACCESS.2024.3363494.

[16] 石江南, 彭长根, 谭伟杰. Spark框架下支持差分隐私保护的*K*-means++聚类方法[J]. 信息安全研究, 2024, 10(8): 712–718. doi: 10.12379/j.issn.2096-1057.2024.08.04.
SHI Jiangnan, PENG Changgen, and TAN Weijie. *K*-means++ clustering method supporting differential privacy protection in spark framework[J]. *Journal of Information Security Research*, 2024, 10(8): 712–718. doi: 10.12379/j.issn.2096-1057.2024.08.04.

[17] WU Fuyu, DU Mingjing, and ZHI Qiang. Density-based clustering with differential privacy[J]. *Information Sciences*, 2024, 681: 121211. doi: 10.1016/j.ins.2024.121211.

[18] FANG Shuhui, WAN Xuejun, WANG Jun, *et al*. HiDS Data clustering algorithm based on differential privacy[C]. 2024 International Conference on Networking and Network Applications (NaNA), Yinchuan, China, 2024: 131–136. doi: 10.1109/NaNA63151.2024.00029.

[19] DIAA A, HUMPHRIES T, and KERSCHBAUM F. FastLloyd: Federated, accurate, secure, and tunable *k*-means clustering with differential privacy[EB/OL]. https://arxiv.org/abs/2405.02437, 2024.

[20] SONG Haina, HAN Xinyu, LV Jie, *et al*. MPLDS: An integration of CP-ABE and local differential privacy for achieving multiple privacy levels data sharing[J]. *Peer-to-Peer Networking and Applications*, 2022, 15(1): 369–385. doi: 10.1007/s12083-021-01238-8.

[21] YANG Wenjun and AL-MASRI E. ULDP: A user-centric local differential privacy optimization method[C]. 2024 IEEE World AI IoT Congress (AIIoT), Seattle, USA, 2024: 0316–0322. doi: 10.1109/AIIoT61789.2024.10579023.

[22] 曾卓, 汪成亮, 马飞. 基于差分隐私的活动模式保护与时空轨迹发布方法[J]. 电子学报, 2023, 51(3): 552–563. doi: 10.12263/DZXB.20210631.
ZENG Zhuo, WANG Chengliang, MA Fei. Differentially private activity pattern and spatial-temporal trajectory publication[J]. *Acta Electronica Sinica*, 2023, 51(3):

552–563. doi: 10.12263/DZXB.20210631.

[23] HERNANDEZ-MATAMOROS A and KIKUCHI H. Comparative analysis of local differential privacy schemes in healthcare datasets[J]. *Applied Sciences*, 2024, 14(7): 2864. doi: 10.3390/app14072864.

[24] DU Minxin, YUE Xiang, CHOW S S M, *et al*. Sanitizing sentence embeddings (and labels) for local differential privacy[C]. Proceedings of the ACM Web Conference 2023, New York, USA, 2023: 2349–2359. doi: 10.1145/3543507.3583512.

[25] YANG Mengmeng, GUO Taolin, ZHU Tianqing, *et al*. Local differential privacy and its applications: A comprehensive survey[J]. *Computer Standards & Interfaces*, 2024, 89: 103827. doi: 10.1016/j.csi.2023.103827.

[26] KRASNOV D, DAVIS D, MALOTT K, *et al*. Fuzzy C-means clustering: A review of applications in breast cancer detection[J]. *Entropy*, 2023, 25(7): 1021. doi: 10.3390/e25071021.

[27] ALI N A, EL ABBASSI A, and BOUATTANE O. Performance evaluation of spatial fuzzy C-means clustering algorithm on GPU for image segmentation[J]. *Multimedia Tools and Applications*, 2023, 82(5): 6787–6805. doi: 10.1007/s11042-022-13635-z.

[28] 徐久成, 侯钦臣, 瞿康林, 等. 面向时间序列的鲁棒性半监督模糊C均值聚类[J]. 计算机工程与应用, 2023, 59(8): 73–80. doi: 10.3778/j.issn.1002-8331.2207-0445.

XU Jiucheng, HOU Qinchen, QU Kanglin, *et al*. Robust semi-supervised fuzzy C-means clustering for time series[J]. *Computer Engineering and Applications*, 2023, 59(8): 73–80. doi: 10.3778/j.issn.1002-8331.2207-0445.

[29] ARORA J, TUSHIR M, and DADHWAL S K. A new suppression-based possibilistic fuzzy c-means clustering algorithm[J]. *EAI Endorsed Transactions on Scalable Information Systems*, 2023, 10(3): e3. doi: 10.4108/eetsis.v10i3.2057.

[30] FANG Siguo, HUANG Dong, CAI Xiaosha, *et al*. Efficient multi-view clustering via unified and discrete bipartite graph learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(8): 11436–11447. doi: 10.1109/TNNLS.2023.3261460.

[31] JAEGER A and BANKS D. Cluster analysis: A modern statistical review[J]. *WIREs Computational Statistics*, 2023, 15(3): e1597. doi: 10.1002/wics.1597.

[32] BESHARATNIA F, TALEBPOUR A, and ALIAKBARY S. An improved grey wolves optimization algorithm for dynamic community detection and data clustering[J]. *Applied Artificial Intelligence*, 2022, 36(1): 2012000. doi: 10.1080/08839514.2021.2012000.

[33] ALEMAZKOOR N, TOOTKABONI M, NATEGHI R, *et al*. Smart-meter big data for load forecasting: An alternative approach to clustering[J]. *IEEE Access*, 2022, 10: 8377–8387. doi: 10.1109/ACCESS.2022.3142680.

[34] ZHAO Wenhao, MA Jin, LIU Qiyuan, *et al*. Comparison and application of SOFM, fuzzy c-means and k-means clustering algorithms for natural soil environment regionalization in China[J]. *Environmental Research*, 2023, 216: 114519. doi: 10.1016/j.envres.2022.114519.

[35] 葛丽娜, 陈圆圆, 王捷, 等. 改进的密度峰值聚类算法的差分隐私保护方案[J]. 郑州大学学报: 工学版, 2023, 44(6): 19–24. doi: 10.13705/j.issn.1671-6833.2023.03.010.

GE Lina, CHEN Yuanyuan, WANG Jie, *et al*. Differential privacy protection scheme of adaptive clustering by fast search and find of density peaks[J]. *Journal of Zhengzhou University: Engineering Science*, 2023, 44(6): 19–24. doi: 10.13705/j.issn.1671-6833.2023.03.010.

[36] CHEN Hua, ZHOU Yuan, MEI Kehui, *et al*. A new density peak clustering algorithm with adaptive clustering center based on differential privacy[J]. *IEEE Access*, 2023, 11: 1418–1431. doi: 10.1109/ACCESS.2022.3233196.

张朋飞：男，讲师，研究方向为数据安全与隐私保护、数据挖掘.

程　俊：男，硕士生，研究方向为数据安全与隐私保护.

张治坤：男，副教授，研究方向为隐私计算、数据隐私保护、机器学习隐私与安全.

方贤进：男，教授，研究方向为数据安全与隐私保护，智能计算.

孙　笠：男，讲师，研究方向为数据挖掘和机器学习.

王　杰：男，教授，研究方向为智能检测与智能仪表、粉尘防治技术研究.

姜　茸：男，教授，研究方向为数据安全与隐私保护，智能计算.

责任编辑：余　蓉

Fuzzy C-Means Clustering Algorithm Based on Mixed Noise-aware under Local Differential Privacy

ZHANG Pengfei[①②] CHENG Jun[①] ZHANG Zhikun[③] FANG Xianjin[①]
SUN Li[④] WANG Jie[⑤] JIANG Rong[②]

[①](*School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan* 232001, *China*)

[②](*Key Laboratory of Service Computing, Yunnan University of Finance and Economics, Kunming* 650221, *China*)

[③](*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310058, *China*)

[④](*School of Control and Computer Engineering, North China Electric Power University, Beijing* 102206, *China*)

[⑤](*School of Safety Science and Engineering, Anhui University of Science and Technology, Huainan* 232001, *China*)

Abstract:

Objective In big data and Internet of Things (IoT) applications, clustering analysis of collected data is crucial for enhancing user experience. To mitigate privacy risks from using raw data directly, Local Differential Privacy (LDP) techniques are often employed. However, existing LDP clustering studies either require interactive execution, consuming significant privacy budgets, or fail to balance Gaussian noise in clustering data with Laplacian noise for LDP protection, resulting in low clustering accuracy. Moreover, distance metrics for similarity measurement are chosen arbitrarily without fully utilizing the noise characteristics of user-submitted noisy data. This study designs a hybrid noise-aware distance calculation method integrated into the fuzzy C-means clustering algorithm, effectively reducing noise impact on clustering results while protecting data privacy, ensuring both privacy security and clustering quality. It provides a robust solution for sensitive information processing in high-dimensional data environments.

Methods This paper innovatively proposes a mixed noise-aware Fuzzy C-Means clustering algorithm (mnFCM) under LDP. The core idea is to model both Gaussian noise (representing data quality) and Laplacian noise (for data protection) in uploaded user data by constructing a more accurate mixed distribution model, and design a mixed noise-aware distance to replace Euclidean distance for measuring similarity between samples and cluster centers. Specifically, in mnFCM, this paper first designs a mixed noise-aware distance calculation method. On this basis, a new objective function for the algorithm is proposed, and a solution method is designed based on the Lagrange multiplier method. Finally, the convergence of the solution algorithm is theoretically analyzed.

Results and Discussions The experimental results show that as the privacy budget ε increases, the performance of various clustering algorithms generally improves. Notably, mnFCM achieves at least a 8.5% improvement in accuracy compared to the state-of-the-art PrivPro algorithm (Fig.1). This is because mnFCM innovatively considers both Gaussian noise (reflecting data quality) and Laplacian noise (for LDP protection), designing a hybrid noise-aware distance metric to enhance sample similarity measurement, thereby effectively protecting privacy while balancing clustering performance. Experiments on the fuzziness parameter m reveal that when m=2, all algorithms reach peak F-Measure values and lowest Entropy values (Fig.2), strongly validating m=2 as the optimal balance point for clustering effectiveness. Additionally, running time of mnFCM is 1.0 to 1.4 times that of the non-privacy-preserving Nopriv algorithm (Table 2), due to its refined noise processing mechanism. Ablation experiments demonstrate that the MixDis scheme achieves the best clustering performance on both NG and UW datasets (Fig.4), as it considers both Laplacian and Gaussian noise, making the clustered data more robust. Comparative analysis on the synthetic dataset Syn with other privacy-preserving clustering algorithms shows that DP-DPCL+ consistently outperforms DP-DPCL, and DPC+

consistently outperforms DPC (Fig.5). In addition, by varying the values of the four adjustable parameters—privacy budget ε, sample size N, dimension K, and cluster number C—it is evident that the mnFCM method outperforms other privacy protection schemes (Fig. 6).

Conclusions　This paper addresses the privacy protection issue in fuzzy clustering algorithms by simultaneously considering Gaussian noise (reflecting data quality) and Laplacian noise (for LDP protection), and innovatively proposes a mixed noise-aware fuzzy C-means clustering algorithm, mnFCM, satisfying LDP to balance privacy security and clustering quality. It designs a mixed noise-aware distance calculation method, formulates a new objective function, and solves it using the Lagrange multiplier method, while theoretically analyzing the algorithm's convergence. Theoretical analysis shows that the algorithm strictly satisfies LDP, is closer to non-private cluster centroids compared to baseline algorithms, and has similar complexity to non-private algorithms. Experiments demonstrate that the algorithm improves clustering accuracy by 10%～15% on real datasets compared to baseline privacy-preserving algorithms. However, a limitation of this study is that the privacy budget calculation for Laplacian noise in the mixed noise setting may be influenced by Gaussian noise. In future research, the adaptive noise proportion allocation strategies, such as dynamically adjusting the weights of Gaussian/Laplacian noise, will be further explored to optimize the privacy-utility trade-off.

Key words: Clustering analysis; Privacy protection; Local Differential Privacy (LDP); Fuzzy C-Means (FCM) clustering; Laplace mechanism