计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(7): 1842−1853, 2026

一种稳定的加速自适应粒球生成方法

张清华[1,2,3] 廖 伟[1,2] 刘 凡[1,2] 高 满[1,2]

[1]（计算智能重庆市重点实验室（重庆邮电大学） 重庆 400065）

[2]（大数据智能计算重点实验室（重庆邮电大学） 重庆 400065）

[3]（网络空间大数据智能安全教育部重点实验室（重庆邮电大学） 重庆 400065）

（zhangqh@cqupt.edu.cn）

A Stable, Accelerated and Adaptive Granular-Ball Generation Method

Zhang Qinghua[1,2,3], Liao Wei[1,2], Liu Fan[1,2], and Gao Man[1,2]

[1]（*Chongqing Key Laboratory of Computational Intelligence* (*Chongqing University of Posts and Telecommunications*), *Chongqing* 400065）

[2]（*Key Laboratory of Big Data Intelligent Computing* (*Chongqing University of Posts and Telecommunications*), *Chongqing* 400065）

[3]（*Key Laboratory of Cyberspace Big Data Intelligent Security* (*Chongqing University of Posts and Telecommunications*), *Ministry of Education*, *Chongqing* 400065）

Abstract Granular-ball generation (GBG) serves as the foundation for various learning tasks in granular-ball computing (GBC), such as classification and clustering. However, existing adaptive GBG methods introduce significant uncertainty due to the strategy of randomly selecting the division center, resulting in a highly unstable granular-ball classification result. Moreover, the *k*-division based adaptive GBG method requires repeated overlap detection and de-overlapping of granular balls, which significantly increases the computational overhead during the iterative process and impairs the efficiency of GBG. To address the above issues, a new method for stable, accelerated and adaptive GBG (SAAGBG) is proposed in this paper, building upon the existing *k*-division strategy. On the one hand, SAAGBG selects the sample point with the smallest difference with the median vector of the data cluster as the division center, which effectively enhances the robustness to outliers and avoids the inconsistency and instability caused random center selection. On the other hand, a new adaptive division condition based on granular-ball coverage is introduced into SAAGBG, which replaces the minimum quality threshold and overlap detection steps in the original *k*-division based GBG method, significantly improving the efficiency of the GBG process. Experimental results on benchmark datasets and datasets with varying noise ratios demonstrate that SAAGBG achieves greater stability and efficiency without loss of classification accuracy compared with the mainstream adaptive GBG methods.

Key words granular-ball computing; granular-ball generation; adaptivity; stability; classification

摘 要 粒球生成是粒球计算中进行分类、聚类等学习任务的基础。然而，现有自适应粒球生成方法因随机选择粒球划分中心，引入了较大不确定性，导致粒球分类结果极不稳定。此外，基于 *k*-division 的自适应粒球生成方法需反复执行粒球重叠检测及去重叠操作，大幅增加了粒球迭代过程的计算复杂度，降低了粒球生成效率。针对上述问题，基于 *k*-division 自适应粒球划分思想，提出了一种稳定、加速的自适应粒球

收稿日期：2025-05-26；**修回日期**：2026-03-17

基金项目：国家重点研发计划项目（2021YFF0704101）；国家自然科学基金项目（62276038, 62221005）；重庆市自然科学基金创新发展联合基金项目（CSTB2023NSCQ-LZX0164）

This work was supported by the National Key Research and Development Program of China (2021YFF0704101), the National Natural Science Foundation of China (62276038, 62221005), and the Chongqing Natural Science Foundation Joint Fund for Innovation and Development Project (CSTB2023NSCQ-LZX0164).

生成方法。一方面，该方法通过选取与数据簇中值向量差异最小的样本点作为划分中心，降低了对异常值和离群点的敏感性，能够提供更加稳健的集中趋势描述，并有效避免了因随机选择中心而导致的不一致性与不稳定性。另一方面，该方法引入了粒球覆盖率的概念并基于此提出新的自适应划分条件，替代了原始 *k*-division 自适应粒球生成方法中的最低质量下限以及重叠检测和去重叠机制，显著提升了粒球生成过程的效率。在标准数据集及包含不同比例噪声的数据集上的实验结果表明，与主流自适应粒球生成方法相比，所提方法在保持分类准确率的同时，展现了更强的稳定性和更高的效率。

关键词 粒球计算；粒球生成；自适应；稳定性；分类

中图法分类号 TP181

DOI: 10.7544/issn1000-1239.202550360 **CSTR:** 32373.14.issn1000-1239.202550360

随着大数据时代的到来，高效的数据分析与处理成为挖掘数据价值及实现知识发现的关键任务和重要挑战[1-2]。在众多数据处理范式中，粒计算（granular computing, GrC）[3]作为一种高效且可扩展的信息处理方法，通过模拟人脑对信息的多粒度认知机制，能够依据具体任务需求或数据内在特性，从多个层次与维度对相关数据进行有效的抽象或泛化，进而构建具有不同尺度与层次结构的信息粒。基于这些信息粒，GrC 可有效支持知识发现与复杂问题求解，显著提升知识发现的效率，成为处理海量数据的一种有效范式[4-5]。

粒球计算（granular-ball computing, GBC）[6-7]作为粒计算领域近年来发展起来的一种高效学习方法，已引起广泛关注和研究。其核心思想是利用不同粒度的粒球对样本空间进行表示和覆盖，并以粒球为基本信息粒开展后续学习任务。通常，单个粒球包含多个样本点，而粒球的数量远低于样本点的数量。因此，采用粗粒度的粒球替代细粒度的样本点作为模型输入，可以显著降低输入数据的规模，从而提升计算效率。由于其高效性、鲁棒性以及良好的可解释性，GBC 已逐渐拓展至人工智能的多个领域[8-12]。Xia 等人[6]提出粒球的概念，定义了粒球的半径、纯度及标签等属性，并将其应用于分类任务，提出了粒球 *k* 近邻（granular-ball *k*-nearest neighbors, GB*k*NN）分类器和原始的粒球支持向量机。文献 [13] 基于粒子群优化算法，推导了粒球支持向量机的对偶模型，并对原始模型的误差进行了修正。文献 [14−15] 将 GBC 运用于无监督学习，分别提出了球 *k*-means 聚类算法和基于粒球的密度峰值聚类算法，在处理大规模数据时展现出优异性能。Cong 等人[16]将 GBC 与深度学习相结合，提出一种渐进式粒球采样融合框架，通过渐进策略和参数共享机制实现图卷积网络模型的增量训练。此外，GBC 还在粒球采样[17]、粒球粗糙集[18-20]、粒球模糊集[21-22]、粒球三支决策[23-25]、粒球强化学习[26-27]、粒球演化计算[28]等多种模型框架中得到进一步发展和应用。

值得注意的是，GBC 的前提是粒球生成（granular-ball generation, GBG），生成粒球的质量直接影响到后续学习任务的性能。然而，现有自适应的粒球生成方法普遍存在稳定性差和效率低等局限。Xia 等人[6]提出了基于 *k*-means 的粒球生成方法，该方法通过随机选择中心点并设置纯度阈值参数实现粒球的迭代划分，但中心的随机选择导致划分结果不稳定。Liao 等人[29]提出了一种基于最短异质距离的自适应粒球生成方法，尽管该方法能够生成较高质量的粒球，但其需要进行大量样本点间的距离计算，导致模型效率较低。Xia 等人[30]根据粒球的加权纯度和、粒球间的重叠情况以及粒球最低纯度下限 3 个条件，提出了基于 *k*-division 策略的自适应粒球生成方法。然而，该方法在重叠检测时的高复杂度及随机选择中心的不稳定性限制了其在实际应用中的适应性。上述粒球生成方法在迭代过程均采用了随机选择中心的策略，这种随机性导致了粒球划分的稳定性不足，进而影响分类结果的准确性和一致性。此外，频繁的距离计算和重叠检测显著增加了计算开销，降低了粒球生成效率。因此，研究更稳定、高效的粒球生成方法，具有重要的理论价值和实践意义。

针对现有方法的不足，本文提出了一种稳定的加速自适应粒球生成方法（SAAGBG）。该方法基于 *k*-division 自适应粒球生成策略，通过改善粒球划分中心选择策略，并引入新的自适应粒球划分条件，有效克服了现有方法稳定性差和低效率问题。具体而言，SAAGBG 通过计算数据簇的中值向量，并选择与中值向量差异最小的样本点作为划分中心，不仅增强了对离群点的鲁棒性，还有效避免了随机中心选取导致的不确定性。同时，SAAGBG 引入粒球覆盖

率的概念并基于此提出新的自适应划分条件，显著减少了重叠检测和去重叠过程的计算开销，从而提高了粒球生成效率。此外，SAAGBG 在粒球生成过程中引入异常值检测机制，通过剔除仅包含单个样本点的孤立粒球，进一步增强了算法的鲁棒性。

本文的主要贡献如下：

1)提出了一种稳定的加速自适应粒球生成方法，设计基于中值向量的划分中心选择策略，有效避免了随机选取中心带来的不确定性，保证了粒球分类结果的一致性和稳定性。

2)引入粒球覆盖率的概念并基于此提出新的自适应划分准则，取代了现有基于 k-division 自适应粒球生成方法的粒球重叠检测与去重叠过程，大幅降低了粒球迭代过程的计算复杂度，从而显著提升粒球生成效率。

3)在标准数据集及包含不同噪声水平数据集上的对比实验表明，所提方法在基本保持分类精度的前提下，具有更优的稳定性和迭代效率。

1 基础知识

1.1 基本定义

定义 1[6]. 粒球中心和半径。给定一个包含 m 个样本的粒球 $gb=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^m$，$gb$ 的中心 $\boldsymbol{o}$ 和半径 r 分别定义为

$$\boldsymbol{o}=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i, \tag{1}$$

$$r=\frac{1}{m}\sum_{\boldsymbol{x}_i\in gb}d(\boldsymbol{x}_i,\boldsymbol{o}), \tag{2}$$

其中 $d(\boldsymbol{x}_i,\boldsymbol{o})$ 表示样本 $\boldsymbol{x}_i$ 和中心 $\boldsymbol{o}$ 之间的欧氏距离。根据定义 1，粒球的中心是粒球中所有样本的均值向量，即粒球的质心。而粒球的半径则定义为粒球中所有样本点到质心的距离的平均值。

定义 2[6]. 粒球纯度和标签。给定一个包含 m 个样本的粒球 $gb=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^m$，$gb$ 中样本所对应的标签集合为 $L=\{l_i|l_i\in\{w_1,w_2,\cdots,w_k\}\}_{i=1}^m$，$gb$ 的纯度 $p(gb)$ 和标签 $l(gb)$ 如式(3)(4)所示。

$$p(gb)=\frac{\max|w_i|}{m},\ i\in\{1,2,\cdots,k\}, \tag{3}$$

$$l(gb)=\arg\max_{w_i}|w_i|,\ i\in\{1,2,\cdots,k\}, \tag{4}$$

其中 $|w_i|$ 表示 gb 中属于 w_i 类别的样本数量。由定义 2 可知粒球的纯度表示粒球中多数类样本占粒球中总样本的比率，粒球的标签由粒球中多数类样本决定。

1.2 相关研究

为了突出贡献，本节首先介绍了基于 k-division 的自适应粒球生成方法[30]。随后介绍了粒球 k-近邻(granular-ball k-nearest neighbors, GBkNN)[6] 分类器。

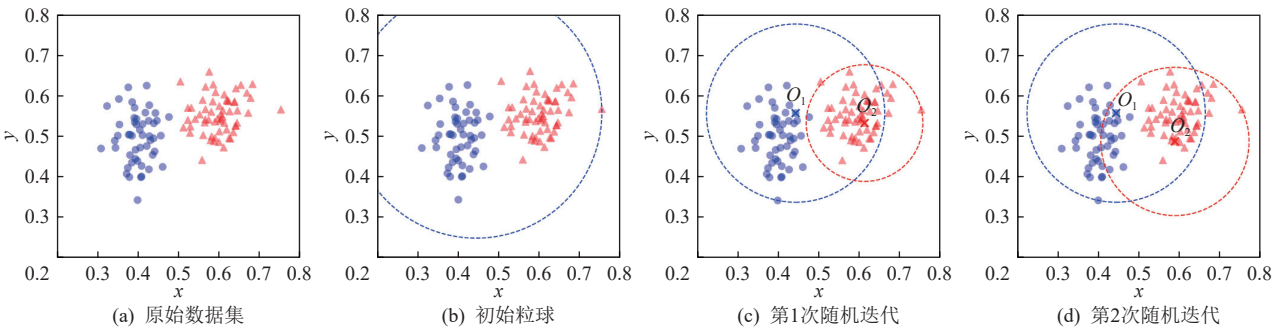
1.2.1 基于 k-division 的自适应粒球生成方法

k-division 的核心思想是在原始数据集中随机选择一个样本作为初始划分中心，并将整个数据集初始化为一个粒球。随后，基于当前粒球，随机选择 $k-1$ 个异质样本点作为子粒球划分中心，根据样本与各中心的距离将当前粒球划分成 k 个子粒球，其中 k 表示当前粒球中样本的类别数。该过程不断迭代，直至所有粒球的纯度达到预设的阈值。k-division 方法的迭代过程如图 1 所示。从图 1 (c)(d)可以看出，随机选择的中心具有较大的不确定性，导致生成的粒球也不同。

基于 k-division 的自适应粒球生成方法包括一个预划分步骤，在此过程中，Xia 等人[6] 提出了 3 个自适应条件，用以判断一个粒球是否需要进一步划分。

条件 1. 粒球纯度不小于其子粒球的加权纯度和。

定义 3[30]. 加权纯度和。给定一个粒球 gb 及其子粒球 $gb_i(i\in\{1,2,\cdots,k\})$，其中 $(\cup_{i=1}^k gb_i)=gb$，$(\cup_{i=1}^k gb_i)=\varnothing$，$gb$ 的子粒球的加权纯度和 $W(gb)$ 被定义为



(a) 原始数据集 (b) 初始粒球 (c) 第1次随机迭代 (d) 第2次随机迭代

Fig. 1 Example of k-division method

图 1 k-division 方法示例

$$W(gb)=\frac{|gb_1|}{|gb|}\times\frac{|gb_1^*|}{|gb_1|}+\cdots+\frac{|gb_k|}{|gb|}\times\frac{|gb_k^*|}{|gb_k|}=\frac{\sum_{i=1}^{k}|gb_i^*|}{|gb|},\quad(5)$$

其中gb_i^*表示粒球gb_i中的多数类样本集合。在定义 3 的基础上，Xia 等人[30]进一步提出了定理 1 和定理 2。

定理 1[30]. 如果对于任意一个子粒球gb_i，都有$l(gb_i)=l(gb)$，那么$W(gb)=p(gb)$。

定理 2[30]. 如果存在一个子粒球gb_i，使得$l(gb_i)\neq l(gb)$，那么$W(gb)>p(gb)$。

根据定理 1 和定理 2，Xia 等人[30]提出，如果父粒球中的少数类样本在部分子粒球中成为多数类样本，则子粒球的加权纯度和将大于父粒球的纯度。换言之，当一个粒球的子粒球加权纯度和超过其自身纯度时，该粒球应被进一步细分。

条件 2. 无重叠的异质粒球。

定义 4[30]. 异质粒球重叠问题。给定 2 个粒球gb_1和gb_2，它们的中心分别为$\boldsymbol{o}_1$和$\boldsymbol{o}_2$，半径分别为r_1和r_2。异质粒球重叠问题可定义为

$$d(\boldsymbol{o}_1,\boldsymbol{o}_2)\leqslant r_1+r_2\ \text{且}\ l(gb_1)\neq l(gb_2)\ 。$$

根据条件 2，基于 *k*-division 的自适应粒球生成方法在每一轮迭代中都需要进行粒球重叠检测及去重叠。具体而言，需要将相互重叠的异质粒球中的样本合并为一个新的数据簇，并对该数据簇应用 *k*-division 方法重新划分。由于该过程需要反复迭代，且每次再划分可能无法完全消除粒球重叠问题，因此计算开销较大，导致粒球生成方法的效率较低。

条件 3. 粒球纯度高于最低质量下限。

Xia 等人[30]将初始粒球的纯度设定为最低纯度下限，即任意粒球的纯度不能低于初始粒球的纯度，否则该粒球需被进一步细分。假设初始粒球为*initGB*，其纯度为$p(initGB)$，对于任意一个粒球gb，其纯度为$p(gb)$，都需满足条件$p(gb)\geqslant p(initGB)$。初始粒球的纯度定义为原始数据集中多数类样本占总样本数的比例，该比值通常较低。因此，将初始粒球的纯度作为自适应划分的条件，可能无法保证生成粒球的质量。

如上所述，粒球的划分仅在同时满足上述 3 个条件时才会停止，否则需要一直迭代。由于划分中心是随机选择的，不同的中心选择使得生成的粒球有所不同，进而导致分类结果的不一致。此外，频繁的重叠检测显著影响了粒球生成过程的效率。

1.2.2 GB*k*NN

k 近邻（*k* nearest neighbors, *k*NN）算法是一种典型的基于实例的监督学习算法，被广泛应用于分类和回归任务。其核心思想可概括为“物以类聚”，即一个样本的类别或数值可以由其最近邻的 *k* 个样本共同决定。具体而言，对于待分类样本，若其 *k* 个最相邻样本中的多数样本属于某一个类别，则该样本也被归类为这个类别，并具有该类别的特征。给定一个待预测的样本，*k*NN 算法首先计算该样本与训练集中所有样本的距离（通常使用欧氏距离），然后选取距离最小的 *k* 个样本，并根据这些邻居的类别或数值进行最终预测。对于分类任务，*k*NN 采用多数投票机制，即选择 *k* 个邻居中出现频率最高的类别作为预测结果。尽管 *k*NN 算法具有直观易懂的优点，但其仍存在计算开销大以及对 *k* 值选择敏感等局限性。为了克服这些固有缺陷和不足，研究者们持续对 *k*NN 算法进行改进与拓展，以提升其性能与实用性[31-32]。

Xia 等人[6]将粒球与 *k*NN 算法相结合，提出了 GB*k*NN 分类器。GB*k*NN 的基本思想是将待分类样本的标签预测为与其距离最近的粒球的标签，其中样本与粒球之间的距离度量如定义 5 所示。

定义 5[6]. 样本与粒球的距离。给定一个样本$\boldsymbol{x}$和一个粒球gb，粒球的中心为$\boldsymbol{o}$，半径为r，$\boldsymbol{x}$与gb的距离如式（6）所示。

$$dist(\boldsymbol{x},gb)=\begin{cases}d(\boldsymbol{x},\boldsymbol{o})-r, & d(\boldsymbol{x},\boldsymbol{o})>r\\ 0, & \text{其他}\end{cases},\quad(6)$$

本质上，*k*NN 和 GB*k*NN 中待分类样本的标签都由其近邻样本的标签决定。不同之处在于，*k*NN 算法需要预设一个参数 *k*，*k* 值决定了在预测阶段所考虑的最近邻样本数。若 *k* 值过小，可能导致对噪声数据的过拟合；而若 *k* 值过大，则可能引起欠拟合。与此不同，GB*k*NN 无需预设 *k* 值，因为粒球内包含的样本已经在粒球生成过程中根据不同粒度自适应确定了。此外，粒球的数量通常远小于原始数据集中的样本数，GB*k*NN 查找最近粒球的计算量显著低于 *k*NN 查找 *k* 近邻样本的计算量。因此，相比传统的 *k*NN 算法，GB*k*NN 更加高效。

2 SAAGBG 方法

2.1 基于中值向量的划分中心选择

在 *k*-division 策略中，数据簇的划分中心通过从每个标签类别中随机选取样本来确定，这一过程引入了较大的不确定性。随机参数不同，选取出的划分中心也就不同，导致生成的粒球存在差异，进而使粒球分类结果不一致。换言之，原始基于 *k*-division

的粒球生成方法表现出较为明显的不稳定性问题。

在SAAGBG方法中，划分中心不再通过随机初始化选择，而是选取与数据簇中值向量差异最小的样本作为划分中心。

假设给定一个数据集$D=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^n$，其中每个样本$\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^f)$包含$f$个特征。则$D$的中值向量可表示为

$$\boldsymbol{m}=(m_1,m_2,\cdots,m_f)\,, \tag{7}$$

其中$m_j(j\in\{1,2,\cdots,f\})$表示数据集$D$中所有样本在第$j$个特征上的中位数。由于中值向量$\boldsymbol{m}$是一个统计量，它可能不对应于数据集$D$中的任何真实样本(例如当样本数量为偶数时)。因此，为确保划分中心始终来源于原始数据，SAAGBG方法通过计算单个样本与中值向量之间的差异，选择与中值向量差异最小的样本作为划分中心。当存在多个样本与中值向量的差异同时达到最小时，为确保划分中心的确定性，选择其中样本索引i最小的样本作为最终结果。

定义 6. 样本与中值向量的差异。给定一个数据集$D=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^n$，其中值向量为$\boldsymbol{m}$，单个样本$\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^f)$与$\boldsymbol{m}=(m_1,m_2,\cdots,m_f)$的差异$diff(\boldsymbol{x}_i,\boldsymbol{m})$由两者间的曼哈顿距离表示：

$$diff(\boldsymbol{x}_i,\ \boldsymbol{m})=\sum_{j=1}^{f}|x_i^j-m_j|,i\in\{1,2,\cdots,n\}\,, \tag{8}$$

其中$|x_i^j-m_j|$表示x_i^j与m_j的差的绝对值。

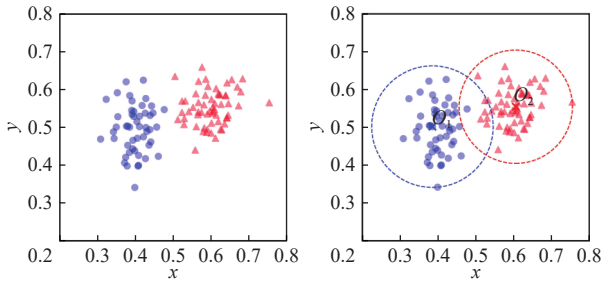
根据定义6，与中值向量差异最小的单个样本索引可表示为

$$index=\min\{i|i\in\arg\min_j diff(\boldsymbol{x}_j,\boldsymbol{m})\},\ \ j\in\{1,2,\cdots,n\}\,, \tag{9}$$

因此，数据集D的划分中心为

$$\boldsymbol{c}=\boldsymbol{x}_{\text{index}}\,。 \tag{10}$$

如图2所示，对于给定的数据集，其中值向量是唯一且确定的，通过计算每个样本与中值向量之间的差异，选择与中值向量差异最小的样本作为划分中心。当多个样本与中值向量的差异相同时，选择其中索引最小的样本作为划分中心，从而确保划分中心选择的确定性。具体而言，选择与数据集的中值向量差异最小的样本作为k-division的划分中心具有以下优势：首先，中值向量对异常值和离群点具有较低的敏感性，表现出较强的鲁棒性，能够提供更为稳健的集中趋势描述，从而使得所选中心周围的样本更加集中，这在一定程度上提高了粒球的质量。更重要的是，该方法有效避免了k-division方法中随机选择划分中心所带来的不确定性，以及由此导致的分类结果不一致性和不稳定性。因此，SAAGBG方法在要求结果稳定性的场景下具有更高的适应性和可靠性。



(a) 初始数据集 (b) 第1次迭代

Fig. 2 Stable division centers

图2 稳定的划分中心

2.2 基于粒球覆盖度的自适应条件优化

1.2.1节中已指出，原始基于k-division的自适应粒球生成方法需要对异质粒球进行重叠检测与去重叠处理。重叠检测基于预划分阶段所获得的全部粒球进行，而去重叠操作则要求将相互重叠的粒球中所包含的样本合并后重新划分。需注意的是，即便执行了重新划分，仍无法确保重叠现象被完全消除。因此，重叠检测和去重叠过程通常具有较高的计算开销。此外，自适应条件3将初始粒球的纯度作为粒球质量的最低下限，以控制粒球生成质量。初始粒球由原始数据集直接构造而得，其纯度即等于数据集中多数类样本所占的比例。如图1(b)所示，初始粒球由50个圆形样本和50个三角形样本构成，依据式(3)计算可得其纯度仅为0.5。由此可见，该条件对粒球质量的要求较为宽松，可能难以有效引导粒球划分过程的优化，从而无法充分保障粒球的质量。

SAAGBG方法摒弃了粒球重叠检测及将初始粒球纯度作为质量下限的自适应条件，转而引入粒球覆盖率的概念，依据覆盖率大小决定是否对粒球进行进一步划分。

定义 7. 粒球覆盖率。给定一个数据集$D=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^n$及粒球$gb=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbb{R}^f\}_{i=1}^m$，$gb$的覆盖率$CR(gb)$如式(11)所示。

$$CR(gb)=\frac{|gb|}{|D|}\,。 \tag{11}$$

覆盖率的大小反映了粒球中所包含样本的数量，覆盖率越大，表示粒球包含的样本数越多，反之亦然。Cheng等人[15]指出，当粒球中所包含的样本足够少时，可以保证该粒球中的大多数样本属于同一类别。这类粒球的覆盖率较小，但纯度较高，从而具有较好的

质量，并能够较准确地刻画数据集的分布特征。然而，粒球中的样本越少，覆盖率越低，生成的粒球数量将会增加，导致算法效率降低；相反，当粒球中的样本增多、覆盖率提高时，粒球中可能包含来自多个类别的样本，从而导致纯度下降。在文献 [15] 中，粒球内样本数的阈值设定为 $\sqrt{n}$，其中 n 为原始数据集样本总数。为确保粒球内样本分布更为集中，本文提出的 SAAGBG 方法进一步将粒球内样本数阈值调整为 $2\ln(n)$，并据此得到任意粒球的覆盖率阈值

$$CR_{\mathrm{T}} = \frac{2\ln(n)}{n}, \tag{12}$$

其中 n 为初始数据集中的样本总数。

基于 CR_{T}，SAAGBG 引入了一个新的自适应条件：对于任意粒球 gb，需满足 $CR(gb) \leqslant CR_{\mathrm{T}}$，否则粒球 gb 需被进一步划分。

在粒球生成过程中，尽管某些粒球可能会发生部分重叠，但通过条件 $CR(gb) \leqslant CR_{\mathrm{T}}$ 有效控制了粒球的规模，从而使得不同类别的数据簇所形成的粒球尽可能保持分离，因此从效率角度出发，SAAGBG 方法摒弃了重叠检测及去重叠过程。

如上所述，原始基于 k-division 的自适应粒球生成方法通过 3 个自适应条件来决定粒球是否需要进一步划分。而 SAAGBG 方法提出了基于粒球覆盖率的自适应条件，取代了原方法中的最低质量下限条件和异质粒球重叠检测，从而仅依赖以下 2 个自适应条件。

1）给定粒球 gb，其纯度为 $p(gb)$ 且子粒球的加权纯度和为 $W(gb)$，若 $p(gb) < W(gb)$，则 gb 需进一步划分。

2）给定粒球 gb，其覆盖率为 $CR(gb)$，若 $CR(gb) > CR_{\mathrm{T}}$，则 gb 需进一步划分。

根据条件 1）和条件 2），仅当粒球 gb 同时满足条件 $p(gb) = W(gb)$ 和 $CR(gb) \leqslant CR_{\mathrm{T}}$ 时，gb 才停止划分。基于粒球覆盖率的自适应条件并未引入额外的计算复杂度，且有效避免了粒球重叠检测与去重叠过程的计算开销，因此 SAAGBG 方法能够显著提升粒球生成的效率。

2.3 异常值处理与局部优化

在粒球生成过程中，部分样本可能被划分为孤立粒球，此类样本通常分布于不同类别的边界区域或作为离群点远离主体数据分布。当这些样本形成仅包含单个样本的“孤儿粒球”时，其标签特征往往与邻近粒球存在显著差异，从而导致对整体数据分布的错误表征。尤其在数据集存在标签噪声的情况下，稀疏的噪声数据容易被划分成只包含单个样本的粒球，从而影响分类性能。因此，SAAGBG 方法将包含单个样本的粒球视为异常粒球，并在迭代过程中剔除这类异常粒球。

此外，在粒球计算理论中，粒球包含的样本分布特征由其中心和半径 2 个参数描述。在 k-division 方法中，粒球的半径被定义为粒球中心到粒球内所有样本点的最大距离。然而，这一半径的计算易受到离群点的干扰，为了使粒球能够更加准确地描述数据集的集中趋势，SAAGBG 方法在进行异常粒球检测的同时，基于定义 1 重新计算粒球中心和半径，对每个非异常粒球进行局部优化，从而使粒球内的样本点更加紧凑，分布更加集中。

图 3 展示了 SAAGBG 方法在 fourclass 数据集上的迭代过程。图 3(a) 为未添加噪声的标准数据集，其中包含三角形和矩形 2 个类别样本。在首次迭代中，划分中心基于 2 个样本类别的中值向量计算得到，而非随机选取，进而将初始数据集划分为 2 个粒球，结果如图 3(b) 所示。经过最终迭代后，数据集被划分为多个尺度不一的粒球，如图 3(c) 所示，这些粒球将替代原始样本点用于后续过程学习。图 3(d) 为添加 10% 比例噪声后的数据集，其中噪声样本以圆点标记。图 3(e) 与图 3(f) 分别展示了在噪声数据集上未执行异常值处理和执行异常值处理后的粒球划分结果。可以观察到，在未执行异常值处理时，部分噪声样本形成孤立的“孤儿粒球”，混杂于正常粒球之间，对分类决策造成干扰。而在执行异常值处理后，大量噪声数据被有效剔除，所得粒球能够更准确地反映原始数据的分布结构，从而有助于提升分类精度。

2.4 算法步骤

根据 2.1~2.3 节，SAAGBG 方法的算法步骤如算法 1 所示。

算法 1. SAAGBG 方法。

输入：数据集 D；

输出：粒球集合 GBs，初始 GBs 为 $\varnothing$。

① 根据式 (7)~(10) 计算 D 中每个样本类别的划分中心 $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k$，其中 k 为样本类别数；

② 根据样本到划分中心的距离，将 D 划分成 k 个粒球 $gb_1^i, gb_2^i, \cdots, gb_k^i$，其中 i 表示迭代次数，初始化 i 为 1；

③ $GBs \leftarrow GBs \cup \{gb_1^1, gb_2^1, \cdots, gb_k^1\}$；

/*迭代划分*/

④ repeat

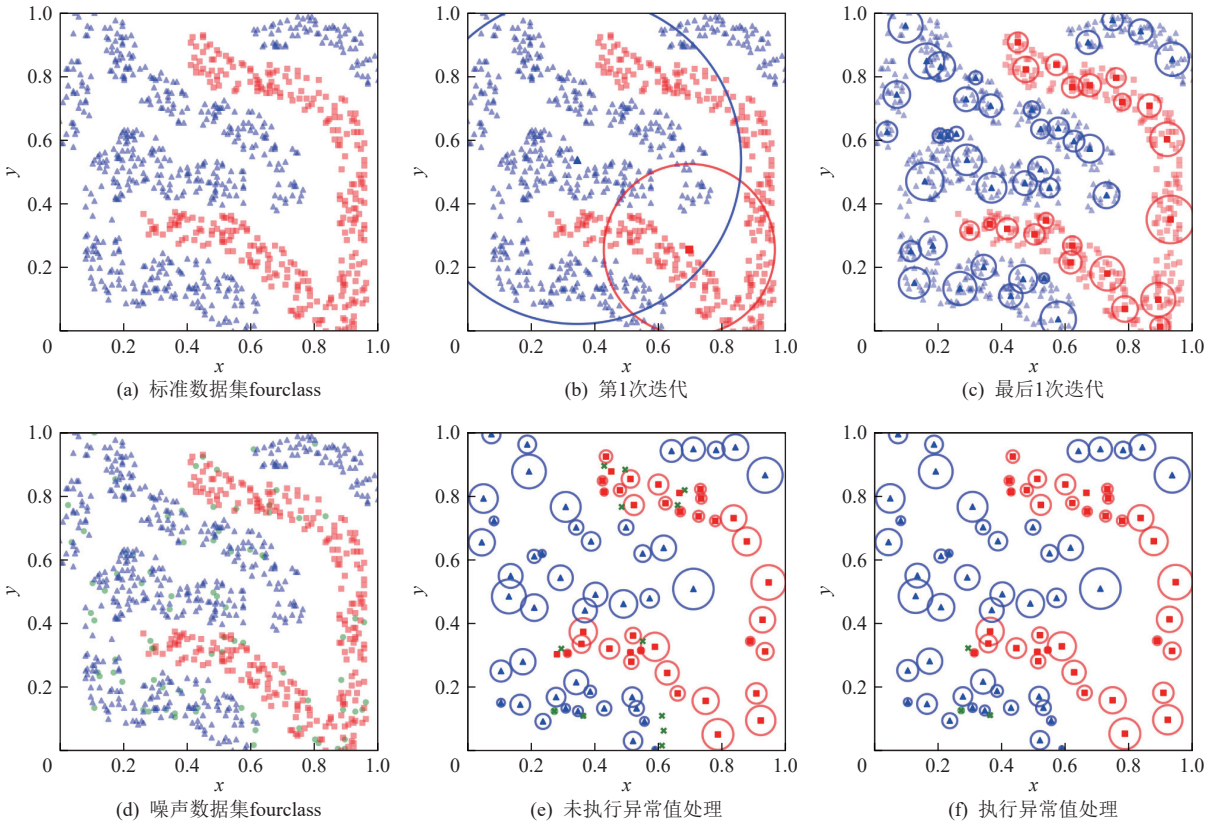


Fig. 3 Example of SAAGBG's iterative process on dataset fourclass

图 3 SAAGBG 在数据集 fourclass 上的迭代过程示例

⑤ for each $gb_j^i \in GBs$ do

⑥ 根据式(3)和式(11)分别计算gb_j^i的纯度$p(gb_j^i)$和覆盖率$CR(gb_j^i)$;

⑦ 根据式(7)~(10)计算与gb_j^i划分中心异质的$k-1$个样本类别的划分中心,将gb_j^i预划分成k个子粒球$gb_1^{i+1}, gb_2^{i+1}, \cdots, gb_k^{i+1}$,其中$k$表示$gb_j^i$中样本的类别数;

⑧ 根据式(5)计算gb_j^i的子粒球的加权纯度和$W(gb_j^i)$;

⑨ if $p(gb_j^i) < W(gb_j^i)$ or $CR(gb_j^i) > CR_T$ then

⑩ $GBs \leftarrow GBs - \{gb_j^i\}$;

⑪ $GBs \leftarrow GBs \cup \{gb_1^{i+1}, gb_2^{i+1}, \cdots, gb_k^{i+1}\}$;

⑫ end if

⑬ end for

⑭ until $|GBs|$不再增加;

/*异常值检测*/

⑮ $GBs_{\text{outlier}} \leftarrow \varnothing$;

⑯ for each $gb \in GBs$ do

⑰ if $|gb| = 1$ then

⑱ $GBs_{\text{outlier}} \leftarrow GBs_{\text{outlier}} \cup \{gb\}$;

⑲ end if

⑳ end for

/*异常值剔除*/

㉑ $GBs \leftarrow GBs - GBs_{\text{outlier}}$;

/*局部优化*/

㉒ for each $gb \in GBs$ do

㉓ 根据式(1)重新计算 gb 的中心;

㉔ 根据式(2)重新计算 gb 的半径;

㉕ end for

㉖ return GBs。

算法 1 中,基于粒球覆盖率的自适应条件未引入额外的计算开销,因此 SAAGBG 方法的时间复杂度与 k-division 的时间复杂度相同。具体而言,SAAGBG 的时间复杂度为$T(n) = O((kt - t + 1)n)$,其中 k 表示初始数据集样本类别数,t表示粒球生成过程的迭代次数,n表示初始数据集的样本数量。

3 实验与分析

3.1 实验设置

为了评估 SAAGBG 方法的性能,本文在 15 个常用的基准分类数据集上进行了对比实验。数据集的

基本信息如表 1 所示。基于 SAAGBG 的 GB*k*NN 分类方法(SAA-GB*k*NN)分别与传统的 *k*NN 算法以及 2 个基于自适应粒球生成的 GB*k*NN 分类方法进行了比较，分别是原始的基于 *k*-division 的自适应 GB*k*NN 方法(ADP-GB*k*NN)[30]和基于最短异质距离的自适应 GB*k*NN 方法(IGB*k*NN)[29]。

Table 1 Basic Information of Fifteen Datasets

表 1 15 个数据集的基本信息

数据集编号	数据集	样本数	特征数	类别数
D1	housevotes	232	16	2
D2	spectfheart	267	44	2
D3	breastCancer	683	10	2
D4	creditApproval	690	15	2
D5	diabetes	768	8	2
D6	fourclass	862	2	2
D7	raisin	900	7	2
D8	phoneme	5 404	5	2
D9	optdigits	5 620	64	10
D10	thyroid	7 000	21	3
D11	svmguid1	7 089	4	2
D12	electrical	10 000	13	2
D13	pen	10 992	16	10
D14	HTRU2	17 898	8	2
D15	codrna	49 466	8	2

如 1.2.2 节所述，自适应 GB*k*NN 方法无需设置任何参数，而传统的 *k*NN 算法需要预先指定最近邻参数 *k*。为了全面评估 *k*NN 算法的性能，本节将其参数 *k* 分别设置为 1, 3, 5, 7, 9, 11, 13, 15，并以各参数下分类准确率的平均值作为最终结果进行对比。此外，为了减小 ADP-GB*k*NN 以及 IGB*k*NN 中随机性带来的评估误差，所有实验均采用 5 折交叉验证进行评估。具体流程为：将各数据集随机划分为 5 个互斥子集，轮流以其中 4 个作为训练集、1 个作为测试集，最终以 5 轮实验的平均准确率作为模型的性能指标。

3.2 实验结果与分析

本节主要从分类精度、分类效率以及分类稳定性 3 个角度评估各分类模型的性能。其中分类精度使用准确率(*Accuracy*)作为评价指标，其计算公式如式(13)所示。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}, \tag{13}$$

其中 TP 表示正确预测为正类的样本数，TN 表示正确预测为负类的样本数，FP 表示错误预测为正类的样本数，FN 表示错误预测为负类的样本数。*Accuracy* 的值越大，表示分类方法的准确率越高，分类结果越好。

此外，本文以 5 折交叉验证的平均执行时间 t 作为分类效率的评价指标。为了便于结果的比较，本文采用 t 的以 10 为底的对数值 $\lg(t)$ 作为分类效率的对比指标，记作 LGT：

$$LGT=\lg(t), \tag{14}$$

其中 LGT 的值越小，表示分类方法的效率越高。

3.2.1 标准数据集上的性能对比

表 2 展示了 SAA-GB*k*NN, ADP-GB*k*NN, IGB*k*NN 以及传统的 *k*NN 方法在标准数据集上的 *Accuracy* 对比。从表 2 中可以看出，尽管与 IGB*k*NN 方法相比，SAA-GB*k*NN 的总体平均 *Accuracy* 略低，但相较于 ADP-GB*k*NN, SAA-GB*k*NN 在大多数数据集上的 *Accuracy* 表现更优。总体而言，SAA-GB*k*NN 在所有数据集上的平均 *Accuracy* 也显著高于 ADP-GB*k*NN，显示出更优的分类效果。

Table 2 *Accuracy* Comparison of Different Methods on the Standard Datasets

表 2 不同方法在标准数据集上的 *Accuracy* 对比

数据集编号	SAA-GB*k*NN（本文）	ADP-GB*k*NN	IGB*k*NN	*k*NN
D1	0.869	0.881	0.856	**0.899**
D2	**0.805**	0.742	0.790	0.729
D3	0.948	0.964	**0.966**	0.964
D4	**0.846**	0.828	0.845	0.845
D5	0.697	0.685	**0.715**	0.694
D6	0.998	0.971	0.999	**1.000**
D7	**0.848**	0.829	0.833	0.836
D8	0.851	**0.888**	0.880	0.872
D9	0.978	0.957	0.977	**0.986**
D10	0.926	0.926	**0.932**	0.929
D11	0.944	0.956	**0.960**	0.950
D12	0.831	0.856	0.892	**0.912**
D13	**0.985**	0.984	0.983	0.983
D14	**0.976**	0.97	0.972	0.972
D15	0.866	0.865	0.883	**0.904**
平均值	0.891	0.887	**0.899**	0.898

注：黑体数值表示最优值。

图 4 展示了 SAA-GB*k*NN, ADP-GB*k*NN, IGB*k*NN 方法在标准数据集上的 LGT 对比，其中图 4 的纵轴表示 LGT，该值越低表示效率越高。从图 4 中可以看出，SAA-GB*k*NN 在所有数据集上的 LGT 均低于其他

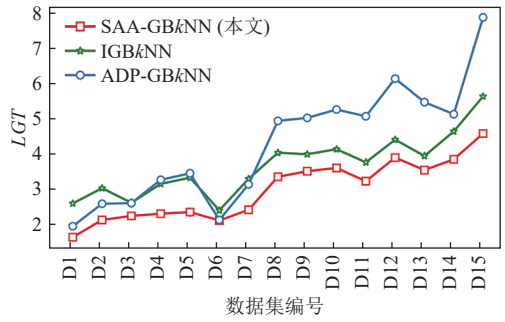
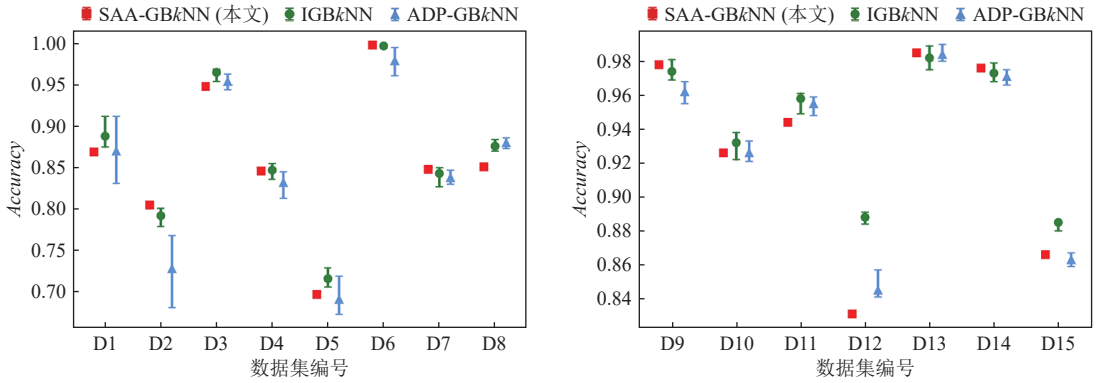


Fig. 4 Efficiency comparison of different methods on the standard datasets

图 4 不同方法在标准数据集上的效率对比

方法, 且随着数据集规模的增大, SAA-GB*k*NN 与其他方法在*LGT*上的差距愈加明显。尤其在较大规模数据集上, SAA-GB*k*NN 与 ADP-GB*k*NN 的*LGT*的差值已超过 1, 这表明相较于原始的基于 *k*-division 的自适应粒球生成方法, 本文所提的 SAAGBG 方法在效率上提高了数十倍甚至更多。这一性能提升归因于 SAAGBG 方法通过引入粒球覆盖度的自适应条件, 消除了粒球重叠检测过程, 从而避免了异质粒球重叠检测及去重叠所带来的性能消耗。

为了评估 SAAGBG 方法的稳定性, 本文针对 SAA-GB*k*NN, ADP-GB*k*NN, IGB*k*NN 这 3 种粒球分类方法, 在每个数据集上再进行了 9 次 5 折交叉验证实验(共 10 次)。以 10 次实验所得 *Accuracy* 的最大值、平均值及最小值作为稳定性评价指标, 结果对比如图 5 所示。



(a) 第1组数据的稳定性对比 (b) 第2组数据的稳定性对比

Fig. 5 Stability comparison of different methods on the standard datasets

图 5 不同方法在标准数据集上的稳定性对比

从图 5 可以看出, SAA-GB*k*NN 每次实验的 *Accuracy* 的最大值、平均值以及最小值均完全一致; 相比之下, ADP-GB*k*NN 和 IGB*k*NN 在各数据集上的结果均存在一定波动。这一结果表明, SAA-GB*k*NN 具有更优的稳定性。该稳定性差异源于粒球划分中心选择策略的不同。具体而言, 原始基于 *k*-division 的自适应粒球生成方法和基于最短异质距离的自适应粒球生成方法均采用随机策略选取划分中心, 导致每次实验生成的粒球存在差异, 进而导致 ADP-GB*k*NN 和 IGB*k*NN 在 10 次实验中的分类结果不一致。而 SAA-GB*k*NN 方法采用基于数据集中值向量的确定性中心选择策略。在每次交叉验证中, 由于数据集的分布保持不变, 数据集的中值向量也不会发生变化, 因此每次实验中生成的粒球保持一致, 从而保证了基于粒球的分类结果一致性, 显著提升了算法的稳定性。

3.2.2 噪声数据集上的性能对比

为了评估粒球分类方法在噪声数据环境下的鲁棒性, 本文对 SAA-GB*k*NN, ADP-GB*k*NN, IGB*k*NN 这 3 种方法进行了对比实验。实验在噪声比例分别为 10%, 20%, 30%, 40% 的数据集上进行, 其中噪声数据的生成方式如下: 对于表 1 中的每个数据集, 随机选取一定比例(10%~40%)的样本并修改其标签。各方法在不同噪声水平下的 *Accuracy* 比较结果展示于表 3 和表 4。

从表 3 和表 4 的结果可以看出, 在相同噪声比例下, SAA-GB*k*NN 在大多数数据集上的 *Accuracy* 明显优于 ADP-GB*k*NN。此外, 综合所有数据的平均 *Accuracy*, SAA-GB*k*NN 在不同噪声率下的表现始终优于 ADP-GB*k*NN。这一差异可归因于 SAA-GB*k*NN 方法在粒球生成过程中采取的异常值处理策略: 该方法将仅包含 1 个样本的粒球视为异常值并进行剔除。在对标准数据集注入噪声后, 大量噪声样本被划分至“孤儿粒球”中, 进而被有效识别和剔除。这一机制显著降低了噪声样本对分类过程的干扰, 从而提升了模型在噪声环境下的鲁棒性和 *Accuracy*。

Table 3 *Accuracy* Comparison of Different Methods on Datasets with 10%, 20% Noise Ratios

表 3 不同方法在 10%，20% 噪声率数据集上的 *Accuracy* 对比

数据集编号	10% 噪声率			20% 噪声率		
	SAA-GB*k*NN	IGB*k*NN	ADP-GB*k*NN	SAA-GB*k*NN	IGB*k*NN	ADP-GB*k*NN
D1	**0.781**	0.775	0.706	0.594	**0.644**	0.550
D2	**0.723**	0.712	0.708	0.670	**0.678**	0.596
D3	0.839	**0.865**	0.810	0.720	**0.733**	0.693
D4	**0.761**	0.744	0.732	**0.625**	0.601	0.577
D5	0.661	**0.667**	0.646	0.534	**0.557**	0.517
D6	0.889	**0.898**	0.862	**0.768**	0.737	0.740
D7	**0.741**	0.738	0.718	**0.612**	0.600	0.570
D8	0.771	**0.785**	0.769	**0.622**	0.621	0.590
D9	0.877	**0.882**	0.811	0.693	**0.703**	0.590
D10	**0.833**	**0.833**	0.767	**0.650**	0.635	0.550
D11	0.845	**0.853**	0.817	**0.664**	0.663	0.626
D12	0.761	**0.795**	0.749	0.629	**0.645**	0.593
D13	0.885	**0.891**	0.847	0.694	**0.703**	0.582
D14	**0.886**	0.872	0.815	0.669	**0.679**	0.610
D15	0.785	**0.793**	0.752	**0.632**	0.628	0.591
平均值	0.803	**0.807**	0.767	0.652	**0.655**	0.598

注：黑体数值表示最优值。

Table 4 *Accuracy* Comparison of Different Methods on Noise Datasets with 30%, 40% Noise Ratios

表 4 不同方法在 30%，40% 噪声率数据集上的 *Accuracy* 对比

数据集编号	30% 噪声率			40% 噪声率		
	SAA-GB*k*NN	IGB*k*NN	ADP-GB*k*NN	SAA-GB*k*NN	IGB*k*NN	ADP-GB*k*NN
D1	**0.575**	0.544	0.431	**0.494**	0.488	0.488
D2	0.531	**0.573**	0.528	**0.517**	0.491	0.483
D3	0.596	**0.628**	0.601	**0.536**	0.535	0.502
D4	0.519	**0.542**	0.510	**0.496**	0.483	**0.496**
D5	0.512	**0.514**	0.513	0.484	0.483	**0.494**
D6	**0.622**	0.617	0.585	0.524	0.541	**0.546**
D7	**0.527**	0.507	0.521	0.514	**0.526**	0.511
D8	0.528	**0.534**	0.530	0.498	**0.505**	0.503
D9	**0.462**	0.446	0.352	**0.249**	0.236	0.170
D10	**0.475**	0.436	0.390	**0.361**	0.352	0.337
D11	**0.524**	0.518	0.512	0.511	**0.513**	0.507
D12	0.521	**0.536**	0.518	0.500	**0.502**	0.501
D13	**0.464**	0.462	0.330	**0.264**	0.239	0.166
D14	0.524	**0.531**	0.525	0.497	0.495	**0.498**
D15	**0.521**	0.520	0.508	0.500	0.499	**0.502**
平均值	**0.527**	**0.527**	0.490	**0.463**	0.459	0.447

注：黑体数值表示最优值。

SAA-GB*k*NN, ADP-GB*k*NN, IGB*k*NN 在不同噪声率数据集下的效率对比如图 6 所示。其中图 6(a)~(d)分别展示了 3 种粒球分类方法在噪声率为 10%, 20%, 30%, 40% 时的对比结果。从图中可以看出，在相同噪声率的数据集上，SAA-GB*k*NN 与 IGB*k*NN 的 *LGT* 的差值基本在 0.5~1.5 波动，表明与 IGB*k*NN 相比，SAA-GB*k*NN 在效率上提升了数倍至数十倍。而相较于 ADP-GB*k*NN，尤其在数据集规模较大的情况下，如 D13 和 D15 数据集，SAA-GB*k*NN 与 ADP-GB*k*NN 的 *LGT* 差值超过 2，意味着 SAA-GB*k*NN 的效率提升可达数百倍。此外，随着数据集规模的增大，SAA-GB*k*NN 在效率上的优势愈加明显。这是由于随着数据集规模的扩展，生成的粒球数量增多，同时噪声数据使得类别边界变得模糊，导致相互重叠的异质粒球增多。ADP-GB*k*NN 方法需要进行重叠检测和去重叠处理，这使得计算复杂度显著提高，从而降低了其效率。与此不同，SAA-GB*k*NN 摒弃了粒球重叠检测与去重叠过程，且基于粒球覆盖率的自适应条件未引入额外计算开销，因此在效率上显著优于现有的自适应粒球分类器。

4 总　　结

本文提出了一种稳定的加速自适应粒球生成方法(SAAGBG)。具体而言，在划分中心的选择策略上，SAAGBG 摒弃了传统方法中从各类别随机选取样本的方式，转而采用基于中值向量的中心确定机制，即选择与类别中值向量差异最小的样本作为划分中心。中值向量对异常值和离群点具有较低的敏感性，表现出较强的鲁棒性，能够提供更为稳健的集中趋势描述。这一策略使得所选中心周围的样本更加集中，从而在一定程度上提高了粒球的质量。更重要的是，该方法有效避免了现有自适应粒球生成方法随机选择划分中心所带来的不确定性，这种不确定性往往导致分类结果的不一致和不稳定。此外，SAAGBG 引入了粒球覆盖率作为关键指标，并据此提出了基于粒球覆盖率的自适应条件。与原始基于 *k*-division 的自适应粒球生成方法相比，SAAGBG 省去了异质粒球重叠检测和去重叠过程，将自适应条件由 3 项简化为 2 项，显著提高了粒球生成的效率。在标准数据集及不同比例噪声数据集上的实验表明，

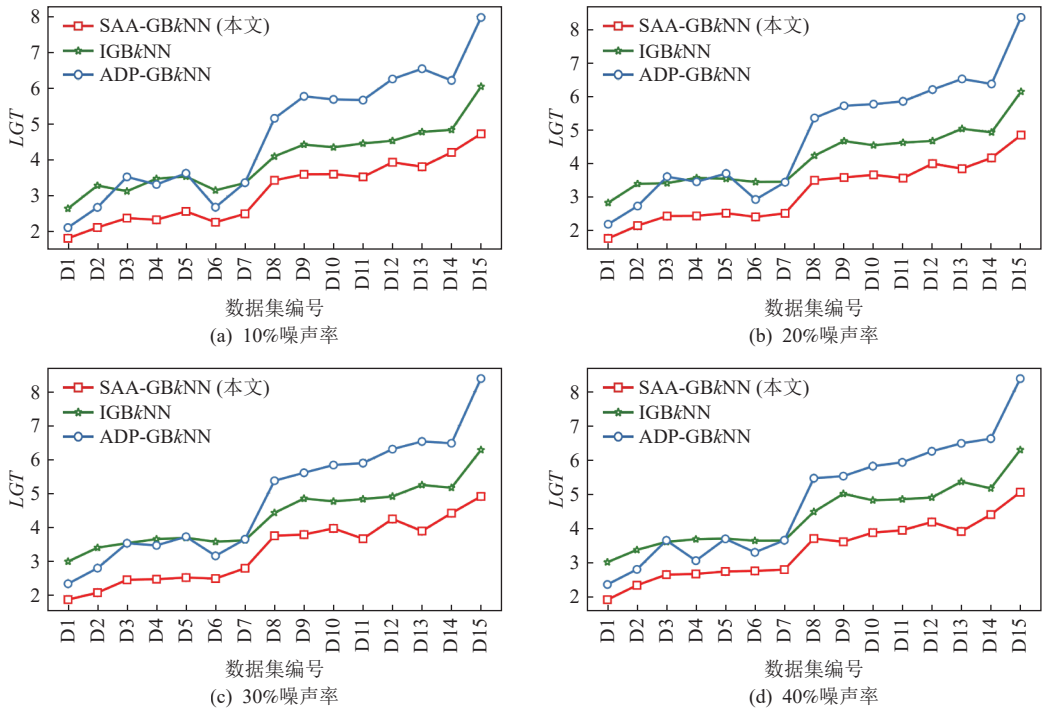


Fig. 6 Efficiency comparison of different methods on datasets with different noise ratios

图 6 不同方法在不同噪声率数据集上的效率对比

SAAGBG 在保持 *Accuracy* 的同时，相较于基于最短异质距离的自适应粒球生成方法[29]和原始基于 *k*-division 的自适应粒球生成方法[30]，展现出了更优的稳定性。同时，SAAGBG 在粒球生成效率上较现有自适应方法提升了数倍至数十倍，且该优势随数据集规模增大而更为显著。本研究目前主要聚焦于粒球计算理论的探讨，未来工作将致力于粒球应用研究的拓展，包括粒球强化学习与粒球神经网络等方向。

作者贡献声明：张清华提出了算法思路和实验方案；廖伟负责完成实验并撰写论文；刘凡、高满提出指导意见并修改论文。

参 考 文 献

[1] Shu Xiaoling, Ye Yiwan. Knowledge discovery: Methods from data mining and machine learning[J]. Social Science Research, 2023, 110: 102817

[2] Dogan A, Birant D. Machine learning and data mining in manufacturing[J]. Expert Systems with Applications, 2021, 166: 114060

[3] Zadeh L A. Fuzzy sets and information granularity[J]. Advances in Fuzzy Set Theory and Applications, 1979, 11: 3−18

[4] Wang Guoyin, Fu Shun, Yang Jie, et al. A review of research multi-granularity cognition based intelligent computing[J]. Chinese Journal of Computers, 2022, 45(6): 1161−1175 (in Chinese)
(王国胤, 傅顺, 杨洁, 等. 基于多粒度认知的智能计算研究[J]. 计算机学报, 2022, 45(6): 1161−1175)

[5] Zhu Xiubin, Wang Dan, Pedrycz W, et al. A design of granular classifier based on granular data descriptors[J]. IEEE Transactions on Cybernetics, 2023, 53(3): 1790−1801

[6] Xia Shuyin, Liu Yunsheng, Ding Xin, et al. Granular ball computing classifiers for efficient, scalable and robust learning[J]. Information Sciences, 2019, 483: 136−152

[7] Chen Yan, Wang Pingxin, Yang Xibei, et al. Granular ball guided selector for attribute reduction[J]. Knowledge-Based Systems, 2021, 229: 107326

[8] Xie Jiang, Kong Weiyu, Xia Shuyin, et al. An efficient spectral clustering algorithm based on granular-ball[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(9): 9743−9753

[9] Peng Xiaoli, Wang Ping, Xia Shuyin, et al. VPGB: A granular-ball based model for attribute reduction and classification with label noise[J]. Information Sciences, 2022, 611: 504−521

[10] Zhang Kai, Zhong Huajia, Chen Tianning, et al. Dissipation behaviors of granular balls in a shaken closed container[J]. Mechanical Systems and Signal Processing, 2022, 172: 108986

[11] Wu Chengying, Zhang Qinghua, Zhao Fan, et al. Hyper-interval granulation approach based on the density peaks clustering for classification[J]. Chinese Journal of Computers, 2023, 46(8): 1620−1635 (in Chinese)
(吴成英, 张清华, 赵凡, 等. 基于密度峰值聚类的超区间粒化方法及其分类模型[J]. 计算机学报, 2023, 46(8): 1620−1635)

[12] Su Xinyu, Wang Xiwen, Peng Dezhang, et al. Granular-ball computing guided anomaly detection for hybrid attribute data[J]. International Journal of Machine Learning and Cybernetics, 2025,

16(5): 2869−2884

[13] Xia Shuyin, Lian Xiaoyu, Wang Guoyin, et al. GBSVM: An efficient and robust support vector machine framework via granular-ball computing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5): 9253−9267

[14] Xia Shuyin, Peng Daowan, Meng Deyu, et al. Ball *k*-means: Fast adaptive clustering with no bounds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 87−99

[15] Cheng Dongdong, Li Ya, Xia Shuyin, et al. A fast granular-ball-based density peaks clustering algorithm for large-scale data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(12): 17202−17215

[16] Cong Hui, Sun Qiguo, Yang Xibei, et al. Enhancing graph convolutional networks with progressive granular ball sampling fusion: A novel approach to efficient and accurate GCN training[J]. Information Sciences, 2024, 676: 120831

[17] Xia Shuyin, Zheng Shaoyuan, Wang Guoyin, et al. Granular ball sampling for noisy label classification or imbalanced classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 2144−2155

[18] Xia Shuyin, Wang Cheng, Wang Guoyin, et al. GBRS: A unified granular-ball learning model of Pawlak rough set and neighborhood rough set[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1719−1733

[19] Qian Wenbin, Xu Fankang, Qian Jin, et al. Multi-label feature selection based on rough granular-ball and label distribution[J]. Information Sciences, 2023, 650: 119698

[20] Xia Shuyin, Zhang Hao, Li Wenhua, et al. GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1231−1242

[21] Xia Shuyin, Lian Xiaoyu, Wang Guoyin, et al. Granular-ball fuzzy set and its implement in SVM[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 6293−6304

[22] Yin Rui, Lu Wei, Yang Jianhua. A hypersphere information granule-based fuzzy classifier embedded with fuzzy cognitive maps for classification of imbalanced data[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(1): 175−190

[23] Yang Xin, Li Yanhua, Xia Shuyin, et al. Granular-ball three-way decision[C]//Proc of the 8th Int Joint Conf on Rough Sets. Berlin: Springer, 2023: 283−295

[24] Yang Jie, Kuang Juncheng, Wang Guoyin, et al. Adaptive three-way KNN classifier using density-based granular balls[J]. Information Sciences, 2024, 678: 120858

[25] Xia Deyou, Wang Guoyin, Zhang Qinghua, et al. Three-way approximations fusion with granular-ball computiong to guide multigranularity fuzzy entropy for feature selection[J]. IEEE Transactions on Fuzzy Systems, 2024, 32(10): 5963−5977

[26] Zhang Qinghua, Wu Chengying, Xia Shuyin, et al. Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(9): 9319−9332

[27] Qian Wenbin, Ruan Wenyong, Li Yihui, et al. Granular ball-based label enhancement for dimensionality reduction in multi-label data[J]. Applied Intelligence, 2023, 53: 24008−24033

[28] Gao Peng, Liu Qun, Jiang Haihuan. Multi-granular differential evolution algorithm for multi-agent task allocation[C]//Proc of the 9th Int Conf on Big Data and Information Analytics. Piscataway, NJ: IEEE, 2023: 897−904

[29] Liao Wei, Zhang Qinghua, Xie Qin, et al. A new adaptive and effective granular ball generation method for classification[J]. International Journal of Machine Learning and Cybernetics, 2025, 16(5): 3501−3520

[30] Xia Shuyin, Dai Xiaochuan, Wang Guoyin, et al. An efficient and adaptive granular-ball generation method in classification problem[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(4): 5319−5331

[31] Chen Yuming, Li Wei. Granular vectors and *k* nearest neighbor granular classifiers[J]. Journal of Computer Research and Development, 2019, 56(12): 2600−2611 (in Chinese)
(陈玉明, 李伟. 粒向量与 *k* 近邻粒分类器[J]. 计算机研究与发展, 2019, 56(12): 2600−2611)

[32] Pan Zhibin, Wang Yikun, Pan Yiwei. A new locally adaptive *k*-nearest neighbor algorithm based on discrimination class[J]. Knowledge-Based Systems, 2020, 204: 106185



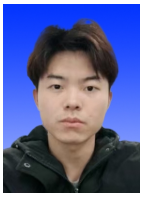
Zhang Qinghua, born in 1974. PhD, professor, PhD supervisor. His main research interests include rough set, fuzzy set, granular computing, and uncertain information processing.

张清华, 1974 年生。博士, 教授, 博士生导师。主要研究方向为粗糙集、模糊集、粒计算、不确定性信息处理。



Liao Wei, born in 1999. Master. His main research interests include machine learning, rough set, and uncertain information processing.

廖　伟, 1999 年生。硕士。主要研究方向为机器学习、粗糙集、不确定性信息处理。



Liu Fan, born in 2001. Master candidate. His main research interests include granular computing, three-way conflict analysis, and uncertain information processing.

刘　凡, 2001 年生。硕士研究生。主要研究方向为粒计算、三支冲突分析、不确定性信息处理。



Gao Man, born in 1994. PhD. His main research interests include granular computing, three-way decision, and uncertainty information processing.

高　满, 1994 年生。博士。主要研究方向为粒计算、三支决策、不确定性信息处理。