智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于邻域互信息与K-means特征聚类的特征选择

孙林, 梁娜, 徐久成

引用本文:
孙林, 梁娜, 徐久成. 基于邻域互信息与K-means特征聚类的特征选择[J]. 智能系统学报, 2024, 19(4): 983-996.
SUN Lin, LIANG Na, XU Jiucheng. Feature selection using neighborhood mutual information and feature clustering with K-means[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(4): 983-996.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202208012

DOI: 10.11992/tis.202208012
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240409.1332.002

基于邻域互信息与 K-means 特征聚类的特征选择

孙林[1], 梁娜[2], 徐久成[2]
（1. 天津科技大学 人工智能学院, 天津 300457; 2. 河南师范大学 计算机与信息工程学院, 河南 新乡 453007）

摘　要: 针对多数邻域系统通过人工调试很难搜索到最佳邻域半径, 以及传统的 K-means 聚类需要随机选取簇中心和指定簇的数目等问题, 提出了一种基于邻域互信息与 K-means 特征聚类的特征选择方法。首先, 将样本在各特征下与其他样本距离的平均值作为自适应邻域半径, 确定样本的邻域集, 并由此构建自适应邻域熵、邻域互信息、归一化邻域互信息等度量, 反映特征之间的相关性; 然后, 基于归一化邻域互信息构建自适应 K 近邻集合, 利用 Pearson 相关系数表示特征的权重定义加权 K 近邻密度, 实现自动选取 K-means 算法的簇中心, 进而完成 K-means 特征聚类; 最后, 给出加权平均冗余度, 选出每个特征簇中加权平均冗余度最大的特征构成最优特征子集。实验结果表明所提算法不仅可以有效提升特征选择的分类结果而且可以获得更好的聚类效果。

关键词: 特征选择; 邻域互信息; K-means; 特征聚类; 自适应 K 近邻; 特征权重; 加权 K 近邻密度
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)04−0983−14

中文引用格式: 孙林, 梁娜, 徐久成. 基于邻域互信息与 K-means 特征聚类的特征选择 [J]. 智能系统学报, 2024, 19(4): 983–996.
英文引用格式: SUN Lin, LIANG Na, XU Jiucheng. Feature selection using neighborhood mutual information and feature clustering with K-means[J]. CAAI transactions on intelligent systems, 2024, 19(4): 983–996.

Feature selection using neighborhood mutual information and feature clustering with K-means

SUN Lin[1], LIANG Na[2], XU Jiucheng[2]
(1. College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China; 2. College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: Aiming at the problems that it is difficult to search the optimal neighborhood radius through manual debugging in most neighborhood systems, and that traditional K-means clustering requires random selection of cluster centers and the number of specified clusters, this paper proposed a feature selection method using neighborhood mutual information and feature clustering with K-means. Firstly, the average distance of the sample from other samples under each feature is taken as the adaptive neighborhood radius, and the neighborhood set of the sample is determined. Then to reflect the correlation between features, some metrics are presented, such as adaptive neighborhood entropy, neighborhood mutual information, normalized neighborhood mutual information, etc. Secondly, an adaptive K neighbor set is constructed based on the normalized neighborhood mutual information, and the weighted K neighbor density is defined by using the feature weight with the Pearson correlation coefficient so that the K-means algorithm can automatically select the cluster center. The K-means feature clustering is completed well. Finally, the weighted average redundancy degree is given, and the feature with the largest weighted average redundancy in each feature cluster is selected to form the optimal subset of features. Experimental results show that the developed algorithm can not only effectively improve the classification results of feature selection, but also obtain better clustering effects.

Keywords: feature selection; neighborhood mutual information; K-means; feature clustering; adaptive K-nearest neighbor; feature weight; weighted k-nearest neighbor density

随着大数据时代的快速发展, 特征选择寻找最小特征子集提高模型分类精度和运行时效[1-3]。作为一种有效的特征选择模型, 邻域粗糙集理论是处理连续型数据的高效工具之一[4]。Hu 等[5] 提

收稿日期: 2022−08−12.　**网络出版日期:** 2024−04−10.
基金项目: 国家自然科学基金项目（62076089, 61772176, 61976082）; 河南省科技攻关计划项目 (212102210136).
通信作者: 孙林. E-mail: sunlin@tust.edu.cn.

出了基于邻域互信息的特征选择算法。但是其性能受邻域半径的影响较大。孙林等[6]提出了基于自适应邻域互信息与谱聚类的特征选择算法，引入标准差自适应地确定邻域半径。但是在一定范围内仍然需要调试邻域半径改进分类性能。Rahmanian 等[7]通过正则化回归模型提出了一种子空间聚类的无监督特征选择算法。但是，该算法在特征聚类时需要固定 K 近邻，且不适应于高维数据集。孙林等[8]提出了基于 K 近邻和优化分配策略的密度峰值聚类算法。但是 K 近邻数需要通过网格搜索策略在一定范围内调试，因而未凸显每个样本的局部几何结构。张新元等[9]提出了共享 K 近邻和多分配策略的密度峰值聚类算法。但是其邻域半径仍需要预先指定或者在一定范围内通过不断的人工调试才能获取较优值。为了解决上述问题，受文献 [6, 10] 自适应邻域的启发，在邻域决策系统中，本文将样本在特征子集下到其他样本距离的平均值作为邻域半径，自适应地确定样本在不同特征下的自适应邻域半径，设计归一化邻域互信息，有效度量特征之间的相关性。

截至目前，从聚类角度进行特征选择已经逐渐引起学者们的关注[11-13]。Chen 等[14]提出了基于冗余互补维度分析的聚类特征子集选择算法。但是，该算法在处理连续型数据时需要对数据进行离散化处理。Sun 等[15]提出了基于最近邻优化分配策略的自适应密度峰值聚类算法。但是仍需要指定簇的数目，导致聚类性能受人的主观影响较大。辛永杰等[16]提出来一种基于跨结构特征选择和图循环自适应学习的多视图聚类。但是该算法忽略了特征间的权重。Song 等[1]提出了基于相关性引导聚类和粒子群优化的高维数据特征选择算法。但是该算法在对特征进行聚类时需要预先指定阈值。总的来说，尽管上述算法在性能上有一定的提升，但是一些问题并未得到有效解决，如需预先指定簇的数目、随机选取簇中心以及处理连续型数据需要预先进行离散化等。

为解决邻域粗糙集需要调试邻域参数的问题，计算样本与其他样本间距离，定义自适应邻域半径，实现自动选择最优邻域半径；针对互信息通常选择多值特征的问题，定义归一化邻域互信息计算特征间的对称不确定性，度量特征间的相关性；为弥补 K-means 需要指定簇数目和随机选取簇中心的缺陷，将 K 近邻集与自适应邻域集结合，构建自适应 K 近邻集和加权 K 近邻密度，自动选取 K-means 的簇中心，实现 K-means 特征聚类，由此设计基于邻域互信息与 K-means 特征聚类的特征选择算法。

1　邻域粗糙集与 K-means 聚类

1.1　邻域粗糙集

假设邻域决策系统 $I = <U, C, D, H>$，其中，U 为 m 个样本构成的集合，C 为含连续型属性的条件属性集，D 为决策属性集，$A = C \cup D$，V 为属性的值域，H 为 $U \times A \to V$ 的映射函数。对于任意样本 x_i, $x_j \in U$, $B \subseteq C$, x_i 在 B 下的邻域集[5]表示为

$$N_B(x_i) = \{x_j \in U | d_B(x_i, x_j) \leqslant \delta\}$$

式中：预设邻域半径 $\delta > 0$，d 表示任意 2 个样本 x_i 和 x_j 在属性 $a_t \in B$ 下的欧氏距离，其表达式为

$$d_B(x_i, x_j) = \sqrt{\sum_{t=1}^{|B|} (H(x_i, a_t) - H(x_j, a_t))^2}$$

在 $I = <U, C, D, H>$ 中，样本子集 $X \subseteq U$, $B \subseteq C$，X 关于 B 的邻域下近似、上近似集[17]为

$$\underline{N_B^\delta}(X) = \{x_i \in U | \delta_B(x_i) \subseteq X\}$$

$$\overline{N_B^\delta}(X) = \{x_i \in U | \delta_B(x_i) \cap X \neq \varnothing\}$$

1.2　K-means 聚类

假设簇中心为 $F_1, F_2, \cdots, F_c$，随机选择 c 个样本点作为初始簇中心，计算其他每个样本到各个簇中心的距离[18]表示为

$$d_B(x_i, F_b) = \sqrt{(x_i - F_b)^2}$$

式中：F_b 为第 b 个簇中心，$b = 1, 2, \cdots, c$；x_i 为第 i 个样本，根据距离远近将它们划分到距离其最近的簇中。根据每个簇中所有样本点的均值，更新调整簇中心的计算公式[19]表示为

$$F_b^{\text{new}} = \frac{\sum_{x_i \in F_b} x_i}{|F_b|}$$

式中：$|F_b|$为划分到簇 F_b 中的样本数目，F_b^{new}为新的簇中心。如果F_b^{new}和 F_b 不相同，需要不断更新调整簇中心，重新划分样本点到距离其最近的簇中，迭代结束的条件[19]表示为

$$F_b^{\text{new}} = F_b^{\text{old}}$$

2　提出的特征选择算法

2.1　邻域互信息

为了解决传统的邻域互信息中邻域半径在一定范围内不断调试的问题，将样本在特征子集下到其他样本距离的平均值作为邻域半径，基于此可以自适应确定样本在不同特征下的自适应邻域半径，并确定样本的自适应邻域集。

定义 1　在 $I = <U, C, D, H>$ 中，特征子集 $B \subseteq C$，则其他所有样本到 x_i 之间的距离表示为

$$P_B(x_i)=\{d_Bx_{(i,1)},d_Bx_{(i,2)},\cdots,d_Bx_{(i,j)},\cdots,d_Bx_{(i,m)}\}$$

式中: $d_Bx_{(i,j)}$ 表示在特征子集 B 上样本 x_i 和 x_j 之间的欧氏距离, $j=1,2,\cdots,m$。

定义 2　在 $I=<U,C,D,H>$ 中, 特征子集 $B\subseteq C$, 对于任意样本 $x_i\in U$, 样本 x_i 在 B 上的自适应邻域半径 $\delta_B(x_i)$ 和邻域集 $N_B(x_i)$ 表示为

$$\delta_B(x_i)=\frac{\sum_{j=1}^{m}d_B(x_i,x_j)}{(m-1)} \tag{1}$$

$$N_B(x_i)=\{x_p\,|\,d_B(x_i,x_p)\leqslant\delta_B(x_i)\} \tag{2}$$

Hu 等[5] 构建邻域互信息度量特征与决策之间的相关性。受此启发, 基于自适应邻域集研究邻域熵及其互信息等度量方法。

定义 3　在 $I=<U,C,D,H>$ 中, A 和 B 是 C 的 2 个特征子集, 对于任意样本 $x_i\in U$, B 上的自适应邻域熵、A 和 B 的邻域联合熵、B 相对于 A 的邻域条件熵、B 和 A 的自适应邻域互信息分别表示为

$$E_\delta(B)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|N_B(x_i)|}{|U|}\right) \tag{3}$$

$$E_\delta(A,B)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|N_{A\cup B}(x_i)|}{|U|}\right)$$

$$E_\delta(B|A)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|N_{A\cup B}(x_i)|}{|N_A(x_i)|}\right)$$

$$E_\delta(B;A)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|N_A(x_i)||N_B(x_i)|}{|U||N_{A\cup B}(x_i)|}\right) \tag{4}$$

性质 1　互信息与熵满足的关系为

$$E_\delta(B;A)=E_\delta(B)+E_\delta(A)-E_\delta(B,A)$$

$$E_\delta(B;A)=E_\delta(B)-E_\delta(B|A)=E_\delta(A)-E_\delta(A|B)$$

文献 [20] 强调: 对称不确定性克服互信息对具有更多值要素的偏向。为了弥补这种偏差, 基于自适应邻域互信息定义归一化邻域互信息。

定义 4　在 $I=<U,C,D,H>$ 中, A 和 B 是 C 的 2 个特征子集, 则特征 A 和 B 的归一化邻域互信息表示为

$$M(A,B)=\frac{2\times E_\delta(B;A)}{E_\delta(A)+E_\delta(B)} \tag{5}$$

式中: $E_\delta(A)$ 和 $E_\delta(B)$ 分别是 A 和 B 的自适应邻域熵, $E_\delta(B;A)$ 是 A 和 B 的自适应邻域互信息。$M(A,B)$ 的值越大, 表示 A 和 B 越相似。

定义 5　对于任意特征 $f_s\in C$, 则特征 f_s 与其他所有特征的归一化邻域互信息的有序序列表示为

$$\boldsymbol{L}(f_s)=(M(f_s,f_{(s,1)}),M(f_s,f_{(s,2)}),\cdots,M(f_s,f_{(s,n-1)}))$$

式中: $M(f_s,f_{(s,1)})\geqslant M(f_s,f_{(s,2)})\geqslant\cdots\geqslant M(f_s,f_{(s,n-1)})$, $f_{(s,r)}$ 是特征 f_s 降序序列的第 r 个特征, $s=1,2,\cdots,n$, $r=1,2,\cdots,n-1$。为了表述方便, 用 $M_s(f_{(s,r)})$ 表示 $M(f_s,f_{(s,r)})$。

2.2　K-means 特征聚类

文献 [21] 基于归一化互信息度量特征之间的对称不确定性, 通过 K 近邻关系和 K 近邻密度实现对特征的聚类。但是, 其需要预先指定 k 的数目, 并且也忽略了特征之间的影响。因此, 为了解决这些缺点, 本文提出自适应 K 近邻集和加权 K 近邻密度, 实现 K-means 特征聚类。

定义 6　在 $I=<U,C,D,H>$ 中, 任意特征 $f_s\in C$, 特征 f_s 的 K 近邻集和自适应邻域集分别为

$$\gamma(f_s)=f_{(s,k)}|k\in M_s(f_{(s,k)}),1\leqslant k\leqslant\sqrt{n}$$

$$\varepsilon(f_s)=f_{(s,r)}|M_s(f_{(s,r)})>Q_{\text{avg}}(f_s)$$

式中: $1\leqslant r\leqslant n-1$, $Q_{\text{avg}}(f_s)$ 为特征 f_s 与其他特征的对称不确定性平均值。

文献 [9] 表明 K 近邻在稀疏密度区域表现不佳, 自适应邻域关系在高密度区域表现不佳。由此将 K 近邻集与自适应邻域集结合, 提出自适应 K 近邻集。

定义 7　在 $I=<U,C,D,H>$ 中, 任意特征 $f_s\in C$, 则特征 f_s 的自适应 K 近邻集合表示为

$$\zeta(f_s)=f_{(s,l)}|f_{(s,l)}\in(\gamma(f_s)\cap\varepsilon(f_s)) \tag{6}$$

式中 $1\leqslant l\leqslant n-1$。

由定义 6 至定义 7 可知, 如果满足 $M_s(f_{(s,l)})>M(f_s)$ 且 $f_{(s,l)}\in\gamma(f_s)$, 则特征 $f_{(s,l)}$ 属于特征 f_s 的近邻; 否则, 特征 $f_{(s,l)}$ 不属于特征 f_s 的近邻。

借鉴文献 [22] 加权 K 近邻思想, 基于 Pearson 相关系数定义近邻特征之间的权重, 并由此设计加权 K 近邻密度与加权平均冗余度。

定义 8　在 $I=<U,C,D,H>$ 中, 任意特征 $f_s\in C$, 特征 $f_{(s,l)}$ 是特征 f_s 的第 l 个近邻, 则特征 $f_{(s,l)}$ 相对于特征 f_s 的权重表示为

$$\omega(f_s,f_{(s,l)})=\left|\frac{R(f_sf_{(s,l)})-R(f_s)\times R(f_{(s,l)})}{\sqrt{R(f_s^2)-R^2(f_S)}\times\sqrt{R(f_{(s,l)}^2)-R^2(f_{(s,l)})}}\right|$$

式中: $w_{(fs,f(s,l))}$ 度量 2 个变量 f_s 和 $f_{(s,l)}$ 之间的相关程度, $R()$ 为数学期望值。

文献 [21] 中定义的 K 近邻密度、平均冗余度假设特征之间的影响是相同的。为了克服该局限性, 受文献 [8] 中加权 K 近邻思想的启发, 考虑特征之间的影响存在不一定相同的情形, 将基于 Pearson 相关系数表示的特征权重与特征的密度、平均冗余度分别进行结合, 定义加权 K 近邻密度。

定义 9　在 $I=<U,C,D,H>$ 中, 任意特征 $f_s\in C$, 则特征 f_s 的加权 K 近邻密度表示为

$$Z(f_s)=\frac{\sum_{l=1}^{|\zeta(f_s)|}\omega(f_s,f_{(s,l)})\times M(f_s,f_{(s,l)})}{|\zeta(f_s)|} \tag{7}$$

由定义 9 可知，加权 K 近邻密度 $Z(f_s)$ 反映了特征 f_s 与其 K 个近邻的冗余程度，因此，近邻密度越大，越有可能成为簇中心。

定义 10　在 $I=<U, C, D, H>$中，任意特征 $f_s\in C(s=1, 2, \cdots, n)$，则所有特征自适应加权 K 近邻密度的有序序列表示为

$$L(Z)=(f_1, f_2, \cdots, f_n) \tag{8}$$

式中 $Z(f_1)>Z(f_2)>\cdots>Z(f_n)$。

依据定义 10，$\boldsymbol{L}(Z)$ 中的第一个特征 f_1 添加到特征簇中心集合 F_c 中，对于任意特征 $f_s\in C-F_c$ 和 $f_t\in F_c$，如果特征 $f_s\in\zeta(f_t)$ 且 $f_t\in\zeta(f_s)$，那么 f_s 将被放到特征簇中心集合 F_c 中，依次重复上述步骤直到结束，特征簇中心集 F_c 实现了初始化。因此，根据特征的自适应 K 近邻集合、加权 K 近邻密度来确定簇中心，使得提出的 K-means 特征聚类既不需要随机选取也不需要预先确定簇中心的数目。由此，对于任意特征簇 $F_e(e=1, 2, \cdots, g)$，若任意特征 f_s、f_t、f_p 和 f_q 属于特征簇 F_e，且 $M(f_s, f_t)\geqslant M(f_p, f_q)$，那么 $M(f_s, f_t)=\max M(F_e)$，否则 $M(f_s, f_t)=\min M(F_e)$。假设 F_c 是初始特征簇中心集合，对于任意特征 $f_s\in C-F_c$ 和 $f_t\in F_c$，并且 f_t 属于特征簇 F_t，如果 f_s 和 f_t 具有最大对称不确定性且 $M(f_s, f_t)>\max M(F_t)$，则将 f_s 放到特征簇 F_t 中，否则 f_s 将被放到 F_c 中成为一个新的簇中心。由此可知，对于任意特征 $f_s\in F_s$ 和 $f_t\in F_t$，如果 f_s 和 f_t 是簇中心，则 $M(f_s, f_t)<\min M(F_t)$ 且 $M(f_s, f_t)<\min M(F_s)$，这表示簇中心之间的对称不确定性小于特征簇的最小对称不确定性。通过上述过程，将相似度较大的特征分配到一个簇中，将相似度较小的特征分配到其他的不同簇中，进而实现 K-means 特征聚类。

2.3　特征选择算法描述

文献 [21] 指出：特征聚类之后，冗余特征被分到同一个特征簇中，具有最大平均冗余度的特征包含了同一簇中其他特征的大多数信息量。由此将同一簇中每个特征相对于其他特征的权重引入到平均冗余度中，定义加权平均冗余度，选出每个特征簇中加权平均冗余度最大特征构成所选特征子集。

定义 11　在 $I=<U, C, D, H>$中，任意特征 f_s 和 f_t 属于特征簇 $F_e(e=1, 2, \cdots, g)$，则特征 f_s 的加权平均冗余度表示为

$$T(f_s, F_e)=\frac{\sum_{t=1}^{|F_e|}\omega(f_s, f_t)\times M(f_s, f_t)}{|F_e|}$$

式中：$|F_e|$是特征簇 F_e 中包含的特征数，$w(f_s, f_t)$ 为 f_t 相对于 f_s 的权重。

由定义 11 可知，在一个特征簇中，具有最大加权平均冗余度 $T(f_s, F_e)$ 的特征 f_s 比同一个特征簇中的其他特征更重要。由此，选择特征簇 F_e 中的特征 f_s 作为代表特征，而 F_e 中的其余特征作为冗余特征而被剔除。从每个特征簇选出具有最大 $T(f_s, F_e)$ 的特征，构成最终所选特征集。

为了提高其计算效率，对于高维数据集使用文献 [17]Fisher 得分算法对高维数据集进行初步降维处理，利用文献 [23] 中的数据离散度代替类内散度与类间散度，避免不同量纲对统计结果的影响。

定义 12　在 $I=<U, C, D, H>$中，任意特征 $f_s\in C$，$U/D=\{D^1, D^2, \cdots, D^d\}$，假设样本 $x_i\in D^a$，则决策类 D^a 中的样本在特征 f_s 上的 Fisher 得分和离散度分别表示为

$$W(f_s)=\frac{\sum_{a=1}^{d}|D_a|\times\dfrac{(R(f_s^a)-R(f_s))^2}{R(f_s)}}{\sum_{a=1}^{d}|D_a|\times(Y_s^a)^2} \tag{9}$$

$$Y_s^a=\frac{\sqrt{\sum_{x_i\in D^a}(H(x_i, f_s)-R(f_s^a))^2}}{|D^a||R(f_s^a)|}$$

式中：$|D^a|$为决策类 D^a 中样本的数目，$1\leqslant a\leqslant c$，$R(f_s^a)$为决策类 D^a 中样本在特征 f_s 下的平均值，$H(x_i, f_s)$ 为样本 x_i 在特征 f_s 上的值，$R(f_s)$ 为全部样本在特征 f_s 下的平均值，Y_s^a为决策类 D^a 中样本在特征 f_s 下的离散度。

接下来，基于邻域互信息与 K-means 特征聚类设计特征选择算法 (feature selection algorithm using neighborhood mutual information and feature clustering with K-means, FSNFK)，其伪代码描述如算法 1 所示。

算法 1　FSNFK 算法

输入　$I=<U, C, D, H>$

输出　最优特征子集 S

1) 初始化特征簇中心 $F_c=\varnothing$，特征子集 $J_{\text{sub}}=S=\varnothing$

2) 依据式 (1) 和式 (2) 计算每个样本 x_i 的自适应邻域集

3) For 任意特征 f_s、$f_t\in C$ 且 $s\neq t$ do

4) 根据式 (3) 计算 f_s 和 f_t 的自适应邻域熵

5) 根据式 (4) 计算 f_s 和 f_t 的自适应邻域互信息

6) 根据式 (5) 计算 f_s 和 f_t 的归一化邻域互信息 $M(f_s, f_t)$

7) End for

8) For 任意特征 f_s do

9) 降序排序 f_s 与其他特征的归一化邻域互信息得到 $L(f_s)$

10) 根据式 (6) 计算 f_s 的自适应 K 近邻集合

11) 根据式 (7) 计算 f_s 的加权 K 近邻密度 $Z(f_s)$

12) 降序排序所有特征的加权 K 近邻密度得到 $\boldsymbol{L}(Z)$

13) End for

14) 降序排序所有特征的加权 K 近邻密度对应的特征集为 $J_{\text{sub}} = \{f_1, f_2, \cdots, f_n\}$, 并且 $Z(f_s) > Z(f_t)$, 其中, $s < t \leqslant n$, 特征簇中心 $F_c = F_c \cup \{f_1\}$

15) For 任意特征 $f_s \in J_{\text{sub}}$ do

16) For 任意特征 $f_t \in J_{\text{sub}}$ do

17) If $f_t \in \zeta(f_s)$ 且 $f_s \in \zeta(f_t)$

18) 特征簇中心 $F_c = F_c \cup \{f_s\}$

19) $\max M(F_c) = \max(\max M(F_c), M(f_s, f_t))$

20) End if

21) End for

22) End for

23) For $d = 1$: length(F_c) do

24) 特征子集 $F_d = f_d$, 其中, 特征 f_d 是 F_c 中的第 d 个特征

25) End for

26) For 任意特征 $f_s \in J_{\text{sub}}$ 且 $f_s \in F_c$ do

27) $d = \text{argmax}x_{f_d \in F_c}(M(f_s, f_d))$

28) If $M(f_s, f_d) > \max M(F_c)$ then

29) 特征子集 $F_d = F_d \cup \{f_s\}$

30) Else 特征簇中心 $F_c = F_c \cup \{f_s\}$, 跳转到步骤 23)

31) End if

32) End for

33) For 任意特征簇 F_d do

34) 特征 $f_s = \text{argmax}x_{f_s \in F_d}(T(f_s, F_d))$

35) 特征子集 $S = f_s \cup S$

36) End for

37) 返回最优特征子集 S

2.4　计算复杂度分析与比较

假若一个数据集含有 m 个样本和 n 个特征, 步骤 1)~2) 的计算复杂度为 $O(mn^2)$, 步骤 3)~7) 的复杂度为 $O(n^2/2)$。步骤 8)~13) 的复杂度为 $O(n|\zeta(f_s)|)$, 这里 $|\zeta(f_s)| << n$。步骤 14)~32) 中最多有 n 个特征不属于聚类中心, 则将特征分配到所在簇中的计算复杂度最差为 $O(n + m + n^2)$, 因而其计算复杂度是 $O(n^2)$。如果特征簇的数量是 w, 步骤 33)~36) 计算加权平均冗余度的复杂度是 $O(w^2)$; 如果所有特征都被划分为一个特征簇, 步骤 32)~步骤 36) 计算加权平均冗余度的复杂度是 $O(n^2)$。因此, FSNFK 算法的计算复杂度为 $O(mn^2)$。由分析可知, FSNFK 算法的计算复杂度明显低于 MSU-FS[7]、FMSU-FS[7]、FSFC[21]、MICIMR[24]、IG-RFE[25]、GSMI[26] 和 UFSMI[27], 与 DRGS[28] 和 MRI[29] 相同, 略高于 K-means[30]、K-medoids[31]、DBSCAN[32]、OPTICS[33]、IWFS[34] 和 CRFS[35]。

3　实验和结果分析

3.1　实验准备

实验配置为 Windows 7 操作系统、Intel(R) i7 CPU 3.20 GHz 处理器, 8 GB 内存。依据文献 [7, 24, 26], 本文选用 19 个实验数据集 (https://archive.ics.uci.edu/ml 和 http://portals.broadinstitute.org/cgi-bin/cancer/datasets. cgi), 如表 1 所示。对于特征数大于 500 的高维数据集, 采用式 (9) 进行初步降维提升算法的计算效率, 由此设置 Ovarian-PBSII 数据集为 10 维, Isolet、CLL-SUB-111、lung、lung-cancer-203、ALLAML、COLON、warpPIE10P 和 warpAR10P 这 8 个数据集为 50 维, TOX171 数据集为 200 维。

表 1　19 个数据集的描述

Table 1　Description of nineteen datasets

编号	数据集	样本数	属性数	类别数
1	Wpbc	198	31	2
2	Isolet	1560	617	26
3	CLL-SUB-111	1440	1024	20
4	lung	203	3312	5
5	movement-libras	360	90	15
6	SPAMBASE	4601	57	2

续表 1

编号	数据集	样本数	属性数	类别数
7	lung-cancer-203	203	12 600	5
8	ALLAML	72	7 129	2
9	COLON	62	2 000	2
10	TOX171	171	5 748	4
11	Ovarian-PBSII	253	15 155	2
12	Wine	178	13	3
13	Sonar	208	60	2
14	Breast Cancer	569	31	2
15	Breast Tissue	106	9	6
16	Parkinsons	197	22	2
17	warpPIE10P	210	2 420	2
18	warpAR10P	130	2 400	10
19	Lung cancer	32	56	2

3.2　特征选择结果分析

为了验证 FSNFK 在不同数据集上选择特征进行分类的有效性，第一部分选择 FSNFK 与文献 [24] 中的 6 种方法进行对比，包括 MICIMR[24]、IG-RFE[25]、DRGS[28]、MRI[29]、IWFS[34] 和 CRFS[35]。为了与文献 [24] 实验保持一致，采用 10 折交叉验证，从表 1 中选择 6 个数据集，选用 K 最近邻 (K-nearest neighbor, KNN)、决策树 (C4.5) 和随机森林 (random forest, RF) 作为分类器，使用文献 [24] 中的 F_1_Macro 作为特征选择分类结果的评价指标，其中 F_1_Macro 越大则分类效果越好。表 2 给出了 3 种分类器下 8 种算法在 6 个数据集上的 F_1_Macro，其中 Fullset 表示直接对原始数据集处理的算法。

表 2　3 种分类器下 8 种算法在 6 个数据集上的 F_1_Macro

Table 2　F_1_Macro of eight algorithms on six datasets under three classifiers　%

分类器	数据集	FullSet	MICIMR	IG-RFE	IWFS	CRFS	DRGS	MRI	FSNFK
KNN	Wpbc	10.51	14.23	9.15	11.20	10.31	14.80	10.31	**61.02**
	ISOLET	26.76	66.08	46.82	37.05	44.58	55.94	56.96	**72.72**
	CLL-SUB-111	68.30	76.66	66.02	60.70	64.11	75.67	64.43	**77.75**
	lung	62.48	75.42	66.63	52.42	68.43	76.23	76.34	**82.89**
	movement-libras	68.03	82.57	70.52	67.65	74.07	74.65	74.07	**84.13**
	SPAMBASE	81.37	**89.77**	86.78	81.32	87.19	83.02	88.21	88.56
	平均	52.91	67.46	57.65	51.72	58.12	63.39	61.72	**77.85**
C4.5	Wpbc	13.43	14.66	12.41	11.81	10.04	13.93	12.11	**60.60**
	ISOLET	26.14	66.40	42.26	38.22	38.97	50.97	50.72	**66.42**
	CLL-SUB-111	62.71	**68.95**	62.45	63.30	64.62	67.60	57.10	61.50
	lung	58.28	81.57	74.94	54.83	66.35	68.19	70.65	**82.60**
	movement-libras	56.75	66.53	63.33	57.31	59.52	62.33	62.42	**68.03**
	SPAMBASE	87.33	90.28	89.36	86.11	88.99	89.45	89.68	**91.18**
	平均	50.77	64.73	57.46	51.93	54.75	58.75	57.11	**69.55**
RF	Wpbc	16.63	17.62	16.82	17.51	16.14	17.10	14.50	**64.25**
	ISOLET	29.01	71.28	47.68	47.98	46.58	57.07	60.25	**82.80**
	CLL-SUB-111	68.30	**76.66**	66.02	60.70	64.11	75.67	64.43	72.85
	lung	60.64	80.13	70.81	51.35	64.47	73.11	74.96	**87.02**
	movement-libras	65.55	76.01	67.83	62.94	69.37	71.49	66.38	**84.27**
	SPAMBASE	89.86	92.56	91.85	89.68	91.84	90.89	92.02	**93.70**
	平均	55.00	69.04	60.17	55.03	58.75	64.22	62.09	**80.82**

注：加黑代表最优结果，下同。

根据表 2 可知, 在 KNN 分类器下, 在 Wpbc、ISOLET、CLL-SUB-111、lung 和 movement-libras 这 5 个数据集上, FSNFK 的 F_1_Macro 优于其他对比算法, 尤其是在 Wpbc、ISOLET 和 lung 这 3 个数据集上优势非常明显, 其 F_1_Macro 值分别比其他算法优 46.22%~51.87%、6.64%~45.96% 与 6.25%~30.47%; 在 SPAMBASE 数据集上, FSNFK 取得了次优的 F_1_Macro, 比最优的 MICIMR 低 1.21%, 但是比其他算法高 0.35%~7.24%。究其原因可能是因为 FSNFK 在 SPAMBASE 数据集上丢失了个别重要特征而造成的。在 C4.5 分类器下, 在 5 个数据集 Wpbc、ISOLET、lung、movement-libras 和 SPAMBASE 上, FSNFK 的 F_1_Macro 优于其他对比算法, 特别是在 Wpbc、lung 和 movement-libras 这 3 个数据集上优势非常明显; 在 CLL-SUB-111 数据集上, FSNFK 的 F_1_Macro 尽管比最优的 MICIMR 低 7.45%, 但是比 MRI 高 4.4%。在 RF 分类器下, 在 Wpbc、ISOLET、lung、movement-libras 和 SPAMBASE 这 5 个数据集上, FSNFK 的 F_1_Macro 优于其他对比算法, 特别是在 Wpbc、ISOLET 和 lung 数据集上优势非常明显; 在 CLL-SUB-111 数据集上, FSNFK 尽管比表现最好的 MICIMR 低 3.81%, 但是比 FullSet、IG-RFE、IWFS、CRFS 和 MRI 这 5 种对比算法高 4.55%~12.15%。在 C4.5 和 RF 这 2 种分类器下 CLL-SUB-111 数据集, MICIMR 均比 FSNFK 优秀, 究其原因是 FSNFK 在 CLL-SUB-111 数据集上未充分考虑类与特征的相关性与冗余性而导致的。综合分析可知, FSNFK 的性能表现良好。

实验的第 2 部分选择 FSNFK 与文献 [26] 中的 5 种方法进行对比, 包括: 基于目标函数驱动的 2 种特征选择算法(K-means[30] 和 K-medoids[31])、基于密度的 2 种特征选择算法(DBSCAN[32] 和 OPTICS[33])和文献 [26] 的多元归一化互信息的无监督基因选择算法 (unsupervised gene selection algorithm based on multivariate normalized mutual information of genes, GSMI)[26]。为了与文献 [26] 实验保持一致, 从表 1 中选择 5 个数据集, 选用文献 [26] 的分类精度指标, 使用支持向量机 (support vector machine, SVM)、KNN 和 RF 分类器评估所有对比算法, 采用 5 折交叉验证。表 3 给出了 3 种分类器下 6 种算法在 5 个数据集上的分类精度。

表 3　3 种分类器下 6 种算法在 5 个数据集上的分类精度

Table 3　Classification accuracy of six algorithms on five datasets under three classifiers　%

分类器	数据集	K-means	K-medoids	DBSCAN	OPTICS	GSMI	FSNFK
SVM	lung-cancer-203	90.12±03.92	87.67±03.43	90.62±04.85	78.33±02.39	78.33±00.86	**93.07±02.88**
	ALLAML	65.24±01.17	80.38±11.51	81.90±11.62	76.38±09.64	83.1±10.84	**90.18±06.45**
	COLON	78.97±06.71	67.56±12.15	78.97±08.54	74.49±08.39	77.38±15.14	**85.26±08.13**
	TOX171	19.88±05.70	46.20±07.71	55.58±10.78	56.81±09.94	46.73±10.53	**97.06±02.63**
	Ovarian	97.63±01.48	98.81±00.96	**99.60±00.80**	77.04±10.29	97.64±02.28	97.22±03.12
	平均	70.37±03.80	76.12±07.15	81.33±07.32	72.61±08.07	76.64±07.93	**92.56±04.64**
KNN	lung-cancer-203	**92.57±04.21**	89.64±05.72	90.59±08.09	88.17±01.05	89.17±04.55	91.12±03.25
	ALLAML	68.09±05.09	77.61±07.31	81.81±09.73	80.38±11.51	77.90±10.96	**83.14±05.53**
	COLON	80.25±16.51	79.23±07.67	77.43±08.06	72.56±08.42	79.10±06.01	**83.33±10.54**
	TOX171	53.78±05.20	45.00±03.68	50.27±10.45	47.36±04.59	44.45±06.25	**94.12±04.92**
	Ovarian	95.24±01.62	94.88±02.92	**97.23±02.04**	79.84±04.90	94.88±04.21	96.04±02.81
	平均	77.99±06.53	77.27±05.46	79.47±07.67	73.66±06.09	77.10±06.40	**89.55±05.41**
RF	lung-cancer-203	86.21±01.19	82.78±02.99	82.28±03.55	78.83±00.96	82.28±01.69	**90.60±04.71**
	ALLAML	69.33±09.41	86.00±12.01	**95.81±03.42**	87.33±09.45	93.14±04.22	91.43±07.00
	COLON	82.05±12.39	73.97±06.77	85.38±06.32	74.36±07.39	77.56±09.17	**86.67±08.50**
	TOX171	49.78±11.69	54.45±06.50	53.88±13.34	45.04±05.53	53.83±05.99	**92.94±04.40**
	Ovarian	97.61±01.94	94.87±03.64	**97.62±02.32**	78.65±05.43	96.05±04.11	97.22±03.47
	平均	77.00±07.32	78.41±06.38	82.99±05.79	72.84±05.75	80.57±05.04	**91.77±05.62**

由表 3 可以看出，在 SVM 分类器下，在 lung-cancer-203、ALLAML、COLON 和 TOX171 这 4 个数据集上，FSNFK 的分类精度优于其他算法，特别是在 ALLAML、COLON 和 TOX171 数据集上优势明显；在 Ovarian 数据集上，FSNFK 的分类精度比最优的 DBSCAN 低 2.38%，但是比 OPTICS 高 20.18%。同时，FSNFK 的平均分类精度优势明显。在 KNN 分类器下，在 ALLAML、COLON 和 TOX171 这 3 个数据集上，FSNFK 优于其他 5 种对比算法，特别是在 TOX171 数据集上优势明显；在 lung-cancer-203 和 Ovarian 数据集上，FSNFK 取得了次优的分类精度，分别比最优的 K-means 与 DBSCAN 低 1.45% 和 0.41%，但是比其他算法分别高 0.53%~2.95% 和 0.8%~16.2%。究其原因是 FSNFK 选取的特征仍存在部分冗余特征。同时，FSNFK 的平均分类精度优势凸出。在 RF 分类器下，在 lung-cancer-203、COLON 和 TOX171 这 3 个数据集上，FSNFK 的分类精度优于其他对比算法，特别是在 lung-cancer-203 和 TOX171 这 2 个数据集上，优势显著；在 ALLAML 和 Ovarian 这 2 个数据集上，FSNFK 的分类精度分别比最优的 DBSCAN 低 4.38 和 0.40，但是比其他 3 种算法分别高 4.1%~22.1% 和 1.17%~18.57%。同时，FSNFK 的平均分类精度表现最佳。另外，在这 3 种分类器下的 Ovarian 数据集以及在 RF 分类器下 ALLAML 数据集，FSNFK 均略次于最优的 DBSCAN，究其原因是这 2 个数据集存在个别异常值样本，而 DBSCAN 对异常值不敏感造成的。综合来看，FSNFK 能够表现出良好的分类性能。

3.3　聚类分析结果

为了充分展示在不同数据集上实施 FSNFK 后对所选特征进行 K-means 聚类的效果，本节实验第 1 部分选择 FSNFK 与文献 [26] 中的 5 种方法进行对比，包括：K-means[30]、K-medoids[31]、DBSCAN[32]、OPTICS[33] 和 GSMI[26]。为了确保与文献 [26] 实验结果的一致性，从表 1 中选择 5 个数据集，在 3.2 节第 2 部分特征选择的基础上，首先对上述 6 种算法选择的特征子集实施 K-means 聚类分析，指定簇数为数据集中类的数目，然后选用文献 [26] 中标准互信息 (normalized mutual information, NMI)、调整兰德系数 (adjusted rand index, ARI) 和 F−分数 (F-Score) 这 3 种聚类指标评价所有对比算法的聚类效果，采用 10 折交叉实现实验验证。表 4 给出了 3 种聚类指标下 6 种算法在 5 个数据集上的聚类结果。

表 4　3 种聚类分析指标下 6 种算法在 5 个数据集上的聚类结果

Table 4　Clustering results of six algorithms on five datasets under three clustering analysis metrics

聚类指标	数据集	K-means	K-medoids	DBSCAN	OPTICS	GSMI	FSNFK
NMI	lung-cancer-203	0.486	0.392	0.417	0.300	0.225	**0.492**
	ALLAML	0.022	0.136	0.179	0.033	0.206	**0.220**
	COLON	−0.001	−0.007	−0.008	0.007	0.007	**0.115**
	TOX171	0.026	0.193	0.163	0.258	0.244	**0.325**
	Ovarian	0.063	0.002	0.004	0.004	−0.003	**0.119**
	平均	0.119	0.143	0.151	0.120	0.136	**0.254**
ARI	lung-cancer-203	0.336	0.213	0.236	0.162	0.076	**0.371**
	ALLAML	0.037	0.210	−0.019	0.087	0.161	**0.233**
	COLON	**0.018**	−0.016	0.005	0.011	−0.006	−0.021
	TOX171	0.019	0.129	0.135	0.158	**0.290**	0.232
	Ovarian	0.062	0.006	0.021	−0.001	−0.004	**0.071**
	平均	0.094	0.108	0.076	0.083	0.103	**0.177**
F-Score	lung-cancer-203	0.138	0.315	0.030	0.222	0.212	**0.702**
	ALLAML	0.611	0.264	0.431	0.667	0.708	**0.745**
	COLON	0.403	0.516	0.419	0.419	0.452	**0.608**
	TOX171	0.292	0.333	0.211	0.328	0.135	**0.525**
	Ovarian	0.372	0.451	0.589	0.530	0.506	**0.642**
	平均	0.363	0.376	0.336	0.433	0.403	**0.644**

由表 4 分析可知，在 NMI 指标下，FSNFK 优于其他 5 种对比算法，特别是在 COLON、TOX171 和 Ovarian 数据集上，FSNFK 优势明显，其 NMI 值比其他对比算法分别高 0.108~0.123、0.067~

0.299 与 0.056~0.122。同时, FSNFK 的平均 NMI 优势明显。在 ARI 指标下, FSNFK 在 lung-cancer-203、ALLAML 和 Ovarian 数据集上表现最优, 特别是在 lung-cancer-203 和 ALLAML 数据集上, FSNFK 的优势极为明显; 在 COLON 与 TOX171 数据集上未能表现最优, 究其原因为噪声点导致了实例类别划分与聚类划分的重叠程度较小而引起。同时, FSNFK 的平均 ARI 明显最优。在 F-Score 指标下, FSNFK 在这 5 个数据集上表现最优, 特别是在 lung-cancer-203、COLON 和 TOX171 数据集上, 其 F-Score 值比其他对比算法分别高 0.387~0.672、0.092~0.205 与 0.192~0.390。同时, FSNFK 的平均 F-Score 最优。总的来说, FSNFK 在这 5 个数据集上的聚类效果优于其他 5 种算法, 这表明该算法在特征选择之后有效提升了数据集的聚类分析性能。

本节实验的第 2 部分选择 FSNFK 与文献 [7] 中 4 种方法进行对比, 包括: UFSMI[27]、FSFC[21]、FMSU-FS[7] 和 MSU-FS[7]。为了与文献 [7] 实验保持一致, 首先从表 1 中选择了 9 个数据集, 对上述 5 种算法选择的特征子集实施 K-means 聚类分析, 指定簇数为数据集中类的数目, 然后使用 NMI、ARI 和 F-Score 这 3 种聚类分析指标评价所有对比算法, 采用 5 折交叉验证方法实现。表 5 给出了 3 种聚类指标下 5 种算法在 9 个数据集上的聚类结果。

表 5　3 种聚类分析指标下 5 种算法在 9 个数据集上的聚类结果

Table 5　Clustering results of five algorithms on nine datasets under three clustering analysis metrics

聚类指标	数据集	UFSMI	FSFC	MSU-FS	FMSU-FS	FSNFK
NMI	Wine	0.1395	0.1953	0.0825	0.1570	**0.3288**
	Sonar	0.0036	0.0085	0.0116	0.2058	**0.3331**
	Breast Cancer	**0.5023**	0.2850	0.0456	0.3742	0.2400
	Breast Tissue	0.4075	0.3253	0.2301	0.3295	**0.5505**
	Parkinsons	0.1960	0.0693	0.0911	−0.0058	**0.2925**
	warpPIE10P	0.1240	0.2579	0.2462	0.0931	**0.4535**
	warpAR10P	0.1272	0.1442	0.1894	0.2836	**0.5103**
	TOX171	0.1925	0.2468	0.2348	0.0589	**0.3570**
	Lung-cancer	−0.0188	−0.0304	0.0857	**0.3708**	0.1470
	平均	0.1860	0.1669	0.1352	0.2341	**0.3570**
ARI	Wine	0.1099	0.1969	0.0522	0.0911	**0.3060**
	Sonar	−0.0030	0.0126	0.0158	0.0483	**0.3181**
	Breast Cancer	**0.5561**	0.2819	0.0836	0.3963	0.3000
	Breast Tissue	0.2480	0.2104	0.0939	0.2289	**0.2987**
	Parkinsons	0.1462	0.1667	0.0540	0.0655	**0.3854**
	warpPIE10P	0.0598	0.1017	0.1104	0.0295	**0.2162**
	warpAR10P	0.0651	0.0740	0.1071	0.1553	**0.2485**
	TOX171	0.1466	0.1109	0.1831	0.0546	**0.2661**
	Lung-cancer	−0.0098	−0.0294	0.0747	**0.3096**	0.1727
	平均	0.1652	0.1251	0.0861	0.1532	**0.2791**
F-Score	Wine	0.1858	0.2234	0.3228	0.3803	**0.6607**
	Sonar	0.4568	0.4481	0.4808	0.5208	**0.6772**
	Breast Cancer	0.1052	0.6477	0.5471	0.6888	**0.7550**
	Breast Tissue	0.0671	0.0629	0.0817	0.0147	**0.5618**
	Parkinsons	0.6387	0.1888	0.2611	0.2266	**0.8276**
	warpPIE10P	0.1141	0.0774	0.1466	0.1130	**0.4624**
	warpAR10P	0.1076	0.1141	0.1263	0.0993	**0.4855**
	TOX171	0.1734	0.1873	0.1296	0.2921	**0.5602**
	Lung-cancer	0.1406	0.2974	0.2455	0.4018	**0.7343**
	平均	0.2210	0.2497	0.2602	0.3042	**0.6361**

根据表 5 可知，在 NMI 指标下，Wine、Sonar、Breast Tissue、Parkinsons、warpPIE10P、warpAR10P 和 TOX171 数据集上，FSNFK 优于其他 4 种对比算法，特别是在 Breast Tissue、warpPIE10P 和 warpAR10P 数据集上，其 NMI 值分别比其他对比算法高 0.143~0.320 4、0.195 6~0.360 4 与 0.226 7~0.383 1；在 Breast Cancer 数据集上，FSNFK 的 NMI 值尽管比最优的 UFSMI 低 0.262 3，但是比 MSU-FS 高 0.194 4；在 Lung-cancer 数据集上，FSNFK 的 NMI 值比最优的 FMSU-FS 低0.223 8，但是比其他算法高 0.061 3~0.177 4。同时，FSNFK 的平均 NMI 优势明显。在 ARI 指标下，FSNFK 在 Wine、Sonar、Breast Tissue、Parkinsons、warpPIE10P、warpAR10P 和 TOX171 数据集上的表现最佳，特别是在 Wine、Sonar 和 Parkinsons 这 3 个数据集上，FSNFK 的优势明显；在 Breast Cancer 数据集上，FSNFK 的 ARI 值尽管比最优的 UFSMI 低 0.256 1，但是比 FSFC 和 MSU-FS分别高 0.018 1 和 0.216 4；在 Lung-cancer 数据集上，FSNFK 的 ARI 值为次优，比最优的 FMSU-FS低 0.136 9，但比其他对比算法高 0.098~0.202 1。同时，FSNFK 的平均 ARI 优势凸出。在 F-Score 指标下，FSNFK 在 9 个数据集上的表现均最优，尤其是在 Breast Tissue、warpPIE10P、warpAR10P 和 Lung-cancer 数据集上，优势特别明显。另外，在 Breast Cancer 数据集上，UFSMI 的 NMI 和 ARI 均比 FSNFK 优秀，究其原因是 UFSMI 采用互信息标准测量特征之间的统计相关性并去除不相关的特征。在 Lung-cancer 数据集上，FMSU-FS 的 NMI 和 ARI 均比 FSNFK 优秀，究其原因是 FSNFK 在进行特征选择时丢失了个别关键特征造成的。总的来说，FSNFK 在 9 个数据集上的 3 种聚类分析指标表现均优于其他对比算法，这说明通过 FSNFK 算法的特征选择之后能够有效提升聚类分析的性能。

最后，由于篇幅限制，为了直观展示在代表性数据集上实施 FSNFK 算法前后的聚类效果，从表 1 中选择 CLL-SUB-111、TOX171、Ovarian-PBSII 和 Wine 共 4 个代表性数据集，图 1 和图 2 分别给出了在 4 个数据集上未使用 FSNFK 和使用 FSNFK 的 K-means 聚类结果，其中，指定簇的数目为数据集中的类别数。通过比较图 1 和图 2 的聚类效果，当实施 FSNFK 后，在这 4 个数据集上的 K-means 聚类效果更优，不同簇之间的区分更明显。

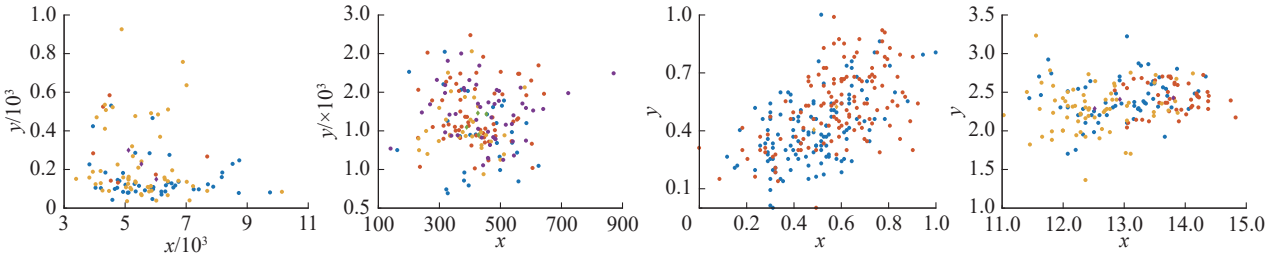

图 1　4 个数据集上的 K-means 聚类结果

Fig. 1　K-means clustering results of four datasets

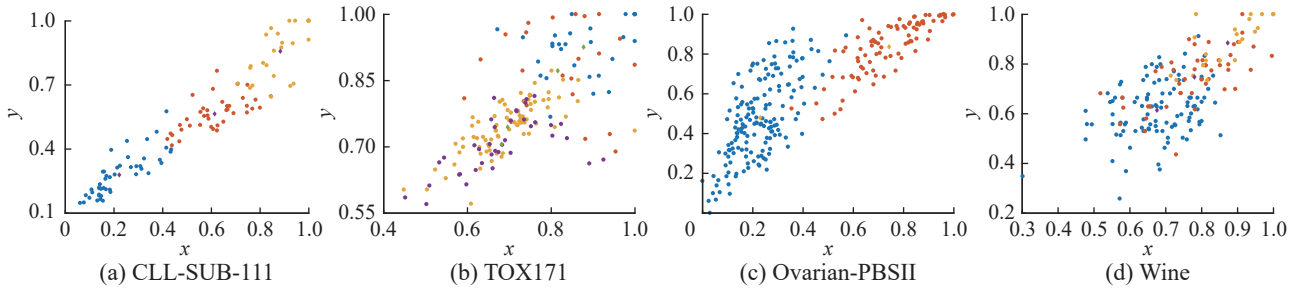

图 2　4 个数据集上实施 FSNFK 算法之后的 K-means 聚类结果

Fig. 2　K-means clustering results of four datasets after performing FSNFK algorithm

3.4　运行时间对比

为了说明 FSNFK 算法的运行效率，从表 1 中选择 6 个代表性数据集，与文献 [26] 中的 3 种算法 (K-means[30]、DBSCAN[32] 和 OPTICS[33]) 和 FSFC[21] 算法对比运行时间，如表 6 所示。根据表 6 分析可知，FSNFK 的运行时间最低，明显优于其他 4 种对比算法，说明 FSNFK 具有很好的时效性。

表 6　5 种算法在 8 个数据集上的运行时间

Table 6　Running time of five algorithms on eight datasets　　s

数据集	FSFC	OPTICS	DBSCAN	K-means	FSNFK
warpPIE10P	12.815	14.020	10.400	10.706	**9.089**
lung-cancer-203	151.281	179.581	142.383	153.626	**55.964**
ALLAML	47.807	66.210	48.005	46.552	**7.568**
Ovarian-PBSII	245.426	328.027	339.360	248.625	**3.990**
Isolet	94.777	336.256	63.123	111.533	**30.957**
CLL-SUB-111	102.180	101.186	103.766	101.776	**16.038**

3.5　统计分析

采用文献 [17] 中 Friedman 检验与 Nemenyi 测试展示对比算法的统计性能。根据文献 [36] 中统计方法, CD 图展示算法之间的差异性。表 7 给出了表 2 中 8 种算法在 KNN、C4.5 和 RF 分类器下 F_1_Macro 的统计结果。表 8 给出了表 3 中 6 种算法在 SVM、KNN 和 RF 分类器下分类精度的统计结果。表 9 给出了表 4 中 6 种算法在 NMI、ARI 和 F-Score 指标下聚类的统计结果。表 10 给出了表 5 中 5 种算法在 NMI、ARI 和 F-Score 指标下聚类的统计结果。

表 7　8 种算法在 KNN、C4.5 和 RF 分类器下 F_1_Macro 的统计结果

Table 7　Statistical results of eight algorithms in terms of F_1_Macro under the KNN, C4.5 and RF classifiers

分类器	FullSet	MICIMR	IG-RFE	IWFS	CRFS	DRGS	MRI	FSNFK	χ_F^2	F_F
KNN	6.33	2.33	5.83	7.17	5.33	3.50	4.33	1.17	29.78	12.18
C4.5	6.50	1.83	4.50	6.83	5.83	3.67	4.83	2.00	25.39	7.64
RF	6.50	1.83	5.00	6.67	6.00	3.83	4.83	1.33	28.89	11.02

表 8　6 种算法在 SVM、KNN 和 RF 分类器下分类精度的统计结果

Table 8　Statistical results of six algorithms in terms of classification accuracy under the SVM, KNN and RF classifiers

分类器	K-means	K-medoids	DBSCAN	OPTICS	GSMI	FSNFK	χ_F^2	F_F
SVM	4.20	4.20	2.40	4.60	3.80	1.80	9.11	2.30
KNN	2.8	4.2	2.8	5	4.8	1.4	14.03	5.11
RF	3.6	4.2	2.2	5.4	3.8	1.8	12.54	4.03

表 9　6 种算法在 NMI、ARI 和 F-Score 指标下聚类的统计结果

Table 9　Clustering statistical results of six algorithms in terms of the NMI, ARI and F-Score metrics

指标	K-means	K-medoids	DBSCAN	OPTICS	GSMI	FSNFK	χ_F^2	F_F
NMI	4.00	4.40	4.00	3.60	4.00	1.00	11.17	3.23
ARI	3.20	4.00	3.80	3.80	4.00	2.20	3.51	0.65
F-Score	5.00	3.40	4.40	3.40	3.80	1.00	13.46	4.66

表 10　5 种算法在 NMI、ARI 和 F-Score 指标下聚类的统计结果

Table 10　Clustering statistical results of five algorithms in terms of the NMI, ARI and F-Score metrics

指标	UFSMI	FSFC	MSU-FS	FMSU-FS	FSNFK	χ_F^2	F_F
NMI	3.44	3.33	3.66	3.11	1.44	11.47	3.74
ARI	3.33	3.56	3.67	3.11	1.33	13.16	4.61
F-Score	3.89	3.89	3.11	3.11	1	20.18	10.20

当显著性水平为 0.1 时, 临界值为 1.90, 表 2 对应的 Nemenyi 测试结果如图 3 所示。由图 3(a) 可知, FSNFK 优于其他 7 种算法, 尽管 FSNFK 与 DRGS、IWFS 之间不存在显著性差异。在图 3(b)

中，尽管 FSNFK 排名略次于 MICIMR，但是明显优于 CRFS、FullSet 和 IWFS，另外 FSNFK 与 CRFS 之间不存在显著性差异。在图 3(c) 中，FSNFK 明显优于其他 7 种算法，尽管 FSNFK 与 IG-RFE 之间不存在显著性差异。当显著性水平为 0.1 时，临界值为 2.16，表 3 对应的 Nemenyi 测试结果如图 4 所示。由图 4 可以看出，FSNFK 在 SVM、KNN 和 RF 分类器下表现最优。在图 4(b) 和图 4(c) 中，FSNFK 与 K-medoids 之间不存在显著性差异。当显著性水平为 0.1 时，临界值为 2.16，表 4 对应的 Nemenyi 测试结果如图 5 所示。由图 5 看出，FSNFK 在 NMI、ARI 和 F-Score 指标下表现最优。在图 5(a) 中，FSNFK 与 K-means 之间不存在显著性差异；在图 5(c) 中，FSNFK 与 GSMI 不存在显著性差异。当显著性水平为 0.1 时，临界值为 1.87，表 5 对应的 Nemenyi 测试结果如图 6 所示。由图 6 可以看出，FSNFK 在 NMI、ARI 和 F-Score 指标上表现最佳。在图 6(a) 中，FSNFK 与 K-means 不存在显著性差异。在图 6(c) 中，FSNFK 与 GSMI 不存在显著性差异。总的来说，FSNFK 表现均优于其他对比算法。

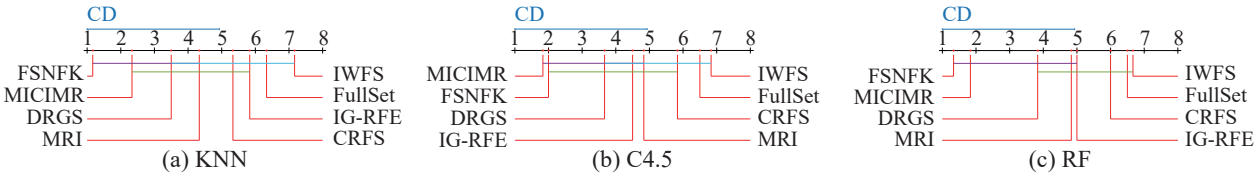

图 3　8 种算法在 KNN、C4.5 和 RF 分类器下 F_1_Macro 的 Nemenyi 测试结果

Fig. 3　Nemenyi test results of eight algorithms in terms of F_1_Macro under the KNN, C4.5 and RF classifiers

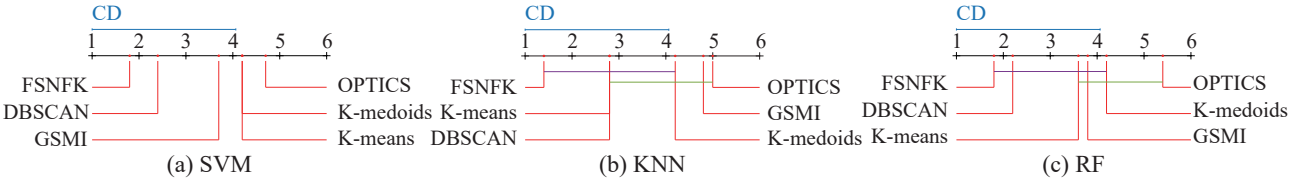

图 4　6 种算法在 SVM、KNN 和 RF 分类器下分类精度的 Nemenyi 测试结果

Fig. 4　Nemenyi test results of six algorithms in terms of classificaiton accuracy under the SVM, KNN and RF classifiers

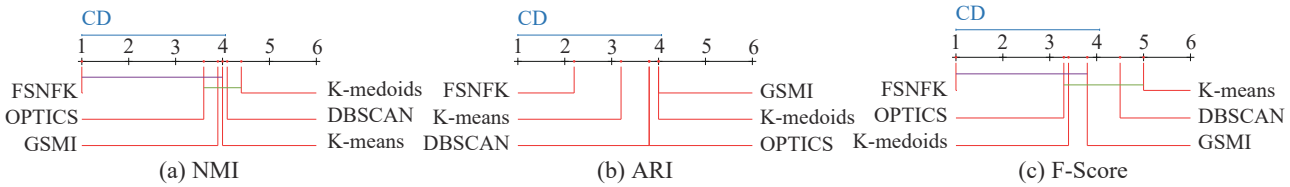

图 5　6 种算法在 NMI、ARI 和 F-Score 指标下的 Nemenyi 测试结果

Fig. 5　Nemenyi test results of six algorithms in terms of the NMI, ARI and F-Score metrics

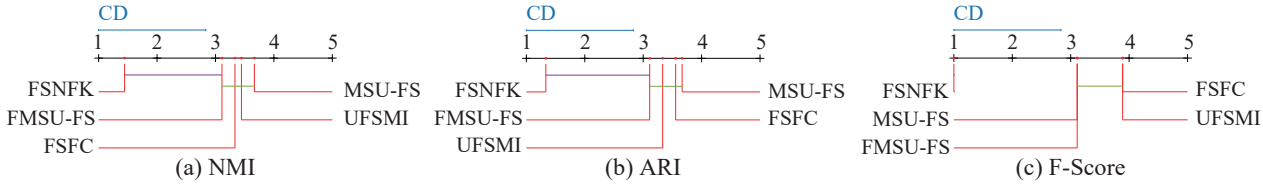

图 6　5 种算法在 NMI、ARI 和 F-Score 指标下的 Nemenyi 测试结果

Fig. 6　Nemenyi test results of five algorithms in terms of the NMI, ARI and F-Score metrics

4　结束语

本文提出了基于邻域互信息与 K-means 特征聚类的特征选择方法。首先，定义了每个样本的自适应邻域半径，由此找到每个样本的邻域集，并在此基础上构建了归一化邻域互信息度量特征之间的相关性。然后，定义了自适应 K 近邻集合，基于 Pearson 相关系数提出了加权 K 近邻密度，进而实现 K-means 特征聚类。最后，给出了加权平均冗余度，选出每个特征簇中加权平均冗余度

最大的特征构成特征子集。但是，所提算法未充分考虑特征之间的相关性、冗余性与互补性等问题，导致未能在所有数据集上取得最优的分类效果，因而，在未来的研究工作中将注重解决上述问题。

参考文献:

[1] SONG Xianfang, ZHANG Yong, GONG Dunwei, et al. A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high- dimensional data[J]. IEEE transactions on cybernetics, 2022, 52(9): 9573–9586.

[2] 梁云辉, 甘舰文, 陈艳, 等. 基于对偶流形重排序的无监督特征选择算法 [J]. 计算机科学, 2023, 50(7): 72–81.
LIANG Yunhui, GANJianwen, CHEN Yan, et al. Unsupervised feature selection algorithm based on dual manifold re-ranking[J]. Computer science, 2023, 50(7): 72–81.

[3] 侯天宝, 王爱银. 基于 Stacking 特征增强多粒度联级 Logistic 的个人信用评估 [J]. 河南师范大学学报(自然科学版), 2023, 51(3): 111–122.
HOU Tianbao, WANG Aiyin. Personal credit evaluation based on stacking feature enhancing multi-grained cascade logistic[J]. Journal of Henan Normal University (natural science edition), 2023, 51(3): 111–122.

[4] 杨洁, 匡俊成, 王国胤, 等. 代价敏感的多粒度邻域粗糙模糊集的近似表示 [J]. 计算机科学, 2023, 50(5): 137–145.
YANG Jie, KUANG Juncheng, WANG Guoyin, et al. Cost-sentitive multigranulation approximation of neighborhood rough fuzzy sets[J]. Computer science, 2023, 50(5): 137–145.

[5] HU Qinghua, ZHANG Lei, ZHANG David, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert systems with applications, 2011, 38: 10737–10750.

[6] 孙林, 梁娜, 徐久成. 基于自适应邻域互信息与谱聚类的特征选择 [J]. 山东大学学报(理学版), 2022, 57(12): 13–24.
SUN Lin, LIANG Na, XU Jiucheng. Feature selection using adaptive neighborhood mutual information and spectral clustering[J]. Journal of Shandong University (nature science edition), 2022, 57(12): 13–24.

[7] RAHMANIAN M, MANSOORI E. Unsupervised fuzzy multivariate symmetric uncertainty feature selection based on constructing virtual cluster representative[J]. Fuzzy sets and systems, 2022, 438: 148–163.

[8] 孙林, 秦小营, 徐久成, 等. 基于 K 近邻和优化分配策略的密度峰值聚类算法 [J]. 软件学报, 2022, 33(4): 1390–1411.
SUN Lin, QIN Xiaoying, XU Jiucheng, et al. Density peak clustering algorithm based on k-nearest neighbors and optimized allocation[J]. Journal of software, 2022, 33(4): 1390–1411.

[9] 张新元, 贠卫国. 共享 K 近邻和多分配策略的密度峰值聚类算法 [J]. 小型微型计算机系统, 2023, 44(1): 75–82.
ZHANG Xinyuan, YUN Weiguo. Sharing K-nearest neighbors and multiple assignment policies density peaks[J]. Journal of chinese computer systems, 2023, 44(1): 75–82.

[10] 孙林, 李梦梦, 徐久成. 二进制哈里斯鹰优化及其特征选择算法 [J]. 计算机科学, 2023, 50(5): 277–291.
SUN Lin, LI Mengmeng, XU Jiucheng. Binary Harris hawk optimization and its feature selection algorithm[J]. Computer science, 2023, 50(5): 277–291.

[11] 曹栋涛, 舒文豪, 钱进. 基于粗糙集与密度峰值聚类的特征选择算法 [J]. 计算机科学, 2023, 50(10): 37–47.
CAO Dongtao, SHU Wenhao, QIAN Jin. Feature selection algorithm based on rough setand density peak clustering[J]. Computer science, 2023, 50(10): 37–47.

[12] 徐天杰, 王平心, 杨习贝. 基于人工蜂群的三支 k-means 聚类算法 [J]. 计算机科学, 2023, 50(6): 116–121.
XU Tianjie, WANG Pingxin, YANG Xibei. Three-way k-means clustering based on artificial bee colony[J]. Computer science, 2023, 50(6): 116–121.

[13] 李冰晓, 万睿之, 朱永杰, 等. 基于种群分区的多策略综合粒子群优化算法 [J]. 河南师范大学学报(自然科学版), 2022, 50(3): 85–94.
LI Bingxiao, WAN Ruizhi, ZHU Yongjie, et al. Multi-strategy comprehensive particle swarm optimization algorithm based on population partition[J]. Journal of Henan Normal University (natural science edition), 2022, 50(3): 85–94.

[14] CHEN Zhijun, CHEN Qiushi, ZHANG Yishi, et al. Clustering-based feature subset selection with analysis on the redundancy-complementarity dimension[J]. Computer communications, 2021, 168: 65–74.

[15] SUN Lin, QIN Xiaoying, DING Weiping, et al. Nearest neighbors-based adaptive density peaks clustering with optimized allocation strategy[J]. Neurocomputing, 2022, 473: 159–181.

[16] 辛永杰, 蔡江辉, 贺艳婷, 等. 基于跨结构特征选择和图循环自适应学习的多视图聚类 [J/OL]. 计算机科学, 2024. [2024−05−14]. https://link.cnki.net/urlid/50.1075.tp.20240513.1427.017.
XIN Yongjie, CAI Jianghui, HE Yanting et al. Multi-view clustering based on cross-structural feature selection and graph cycle adaptive learning[J/OL]. Computer science, 2024. [2024−05−14]. https://link.cnki.net/urlid/50.1075.tp.20240513.1427.017.

[17] 孙林, 马天娇. 基于中心偏移的 Fisher score 与直觉邻域模糊熵的多标记特征选择 [J/OL]. 计算机科学, (2023-11-14)[2023-12-06]. https://link.cnki.net/urlid/50.1075.TP.20231113.1009.012.
SUN Lin, MA Tianjiao. Multilabel feature selection using fisher score with center shift and neighborhood intuitionistic fuzzy entropy[J/OL]. Computer science, (2023-11-14)[2023-12-06].https://link.cnki. net/urlid/50.1075.TP.20231113.1009.012.

[18] 赵燕伟, 朱芬, 桂方志, 等. 基于可拓距的改进 K-

means 聚类算法 [J]. 智能系统学报, 2020, 15(2): 344–351.
ZHAO Yanwei, ZHU Fen, GUI Fangzhi, et al. Improved K-means algorithm based on extension distance[J]. CAAI transactions on intelligent systems, 2020, 15(2): 344–351.

[19] 王雷, 杜亮, 周芃. 基于稀疏连接的层次化多核 K-Means 算法 [J]. 计算机科学, 2023, 50(2): 138–145.
WAND Lei, DU Liang, ZHOU Peng. Hierarchical multiple kernel K-means algorithm based on sparse connectivity[J]. Computer science, 2023, 50(2): 138–145.

[20] GUSTAVO S C, MIGUEL G T, SANTIAGO G G, et al. A multivariate approach to the symmetrical uncertainty measure: Application to feature selection problem[J]. Information sciences, 2019, 494: 1–20.

[21] ZHU Xiaoyan, WANG Yu, LI Yingbin, et al. A new unsupervised feature selection algorithm using similarity-based feature clustering[J]. Computational intelligence, 2019, 35(1): 2–22.

[22] ALHELALI B, CHEN Qi, XUE Bing, et al. A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data[J]. Soft computing, 2021, 25(8): 5993–6012.

[23] SUN Lin, ZHANG Jiuxiao, DING Weiping, et al. Mixed measure-based feature selection using the Fisher score and neighborhood rough sets[J]. Applied intelligence, 2022, 52: 17264–17288.

[24] ZHANG Li, CHEN Xiaobo. Feature selection methods based on symmetric uncertainty coefficients and independent classification information[J]. IEEE access, 2021, 9: 13845–13856.

[25] LIN Xiaohui, LI Chao, REN Wenjie, et al. A new feature selection method based on symmetrical uncertainty and interaction gain[J]. Computational biology and chemistry, 2019, 83: 107149.

[26] RAHMANIAN M, MANSOORI E G. An unsupervised gene selection method based on multivariate normalized mutual information of genes[J]. Chemometrics and intelligent laboratory systems, 2022, 222: 104512.

[27] FAIVISHEVSKY L, GOLDBERGER J. Unsupervised feature selection based on non-parametric mutual information[C]// 2012 IEEE international workshop on machine learning for signal processing. Santander: IEEE, 2012: 1−6.

[28] SUN Xin, LIU Yanheng, WEI Da, et al. Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis[J]. Journal of biomedical informatics, 2013, 46(2): 252–258.

[29] WANG Jun, WEI Jinmao, YANG Zhenglu, et al. Feature selection by maximizing independent classification information[J]. IEEE transactions on knowledge and data engineering, 2017, 29(4): 828–841.

[30] MACQUEEN J. B. Some methods for classification and analysis of multivariate observations[C]//Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Los Angeles: University of California Press, 1967: 281−297.

[31] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. Hoboken: Wiley Online Library, 1991.

[32] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2th International Conference on Knowledge Discovery and Data Mining. Muenchen: AAAI Press, 1996: 226–231.

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

[34] ZENG Zilin, ZHANG Hongjun, ZHANG Rui, et al. A novel feature selection method considering feature interaction[J]. Pattern recognition, 2015, 48(8): 2656–2666.

[35] 刘杰, 张平, 高万夫. 基于条件相关的特征选择方法 [J]. 吉林大学学报(工学版), 2018, 48(3): 874–881.
LIU Jie, ZHANG Ping, GAO Wanfu. Feature selection method based on conditional relevance[J]. Journal of Jilin University (engineering and technology edition), 2018, 48(3): 874–881.

[36] 孙林, 徐枫, 李硕, 等. 基于 ReliefF 和最大相关最小冗余的多标记特征选择 [J]. 河南师范大学学报(自然科学版), 2023, 51(6): 21–29.
SUN Lin, XU Feng, LI Shuo, et al. Multilabel feature selection algorithm using ReliefF and mRMR[J]. Journal of Henan Normal University (natural science edition), 2023, 51(6): 21–29.

作者简介：

孙林，教授，博士生导师，博士，计算机学会会员，主要研究方向为粒计算、大数据挖掘和机器学习。发表学术论文 60 余篇。E-mail：sunlin@tust.edu.cn。

梁娜，硕士研究生，主要研究方向为数据挖掘。E-mail：ms_liangna@126.com。

徐久成，教授，博士生导师，博士，计算机学会高级会员，主要研究方向为粒计算、大数据挖掘和智能信息处理。E-mail：xjc@htu.edu.cn。