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

孙林, 马天娇

引用本文

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]. 计算机科学, 2024, 51(7): 96-107.

SUN Lin, MA Tianjiao. Multilabel Feature Selection Based on Fisher Score with Center Shift and Neighborhood IntuitionisticFuzzy Entropy [J]. Computer Science, 2024, 51(7): 96-107.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400018

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

孙　林[1]　　马天娇[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　现有多标记 Fisher score 模型中边缘样本会影响算法分类效果。鉴于邻域直觉模糊熵处理不确定信息时具有更强的表达能力与分辨能力的优势,文中提出了一种基于中心偏移的 Fisher score 与邻域直觉模糊熵的多标记特征选择方法。首先,根据标记将多标记论域划分为多个样本集,计算样本集的特征均值作为标记下样本的原始中心点,以最远样本的距离乘以距离系数,去除边缘样本集,定义了新的有效样本集,计算中心偏移处理后的标记下每个特征的得分以及标记集的特征得分,进而建立了基于中心偏移的多标记 Fisher score 模型,预处理多标记数据。然后,引入多标记分类间隔作为自适应模糊邻域半径参数,定义了模糊邻域相似关系和模糊邻域粒,由此构造了多标记模糊邻域粗糙集的上、下近似集;在此基础上提出了多标记邻域粗糙直觉隶属度函数和非隶属度函数,定义了多标记邻域直觉模糊熵。最后,给出了特征的外部和内部重要度的计算公式,设计了基于邻域直觉模糊熵的多标记特征选择算法,筛选出最优特征子集。在多标记 K 近邻分类器下、9 个多标记数据集上的实验结果表明,所提算法选择的最优子集具有良好的分类性能。

关键词:多标记学习;特征选择;Fisher score;多标记模糊邻域粗糙集;邻域直觉模糊熵
中图分类号　TP181

Multilabel Feature Selection Based on Fisher Score with Center Shift and Neighborhood Intuitionistic Fuzzy Entropy

SUN Lin[1] and MA Tianjiao[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, Henan 453007, China

Abstract　The edge samples in the existing multilabel Fisher score models affect the classification effect of the algorithm. It has the available virtues of stronger expression and resolution when using neighborhood intuitive fuzzy entropy to deal with uncertain information. Therefore, this paper develops a multilabel feature selection based on the Fisher score with center shift and neighborhood intuitionistic fuzzy entropy. Firstly, the multilabel domain is divided into multiple sample sets according to the labels, the feature mean of the sample set is calculated as the original center point of the samples under the labels, and the distance of the furthest samples is multiplied by the distance coefficient, the edge sample set is removed, and then a new effective sample set is defined. The score of each feature under the labels is calculated after center migration processing and the feature score of the label set. Then, a multilabel Fisher score model is established based on center migration to preprocess multilabel data. Secondly, the multilabel classification interval is introduced as the adaptive fuzzy neighborhood radius parameter, the fuzzy neighborhood similarity relation and fuzzy neighborhood particle are defined, and the upper and lower approximate sets of the multilabel fuzzy neighborhood rough sets are constructed. On this basis, the rough intuitive membership function and non-membership function of multilabel neighborhood are proposed, and the multilabel neighborhood intuitionistic fuzzy entropy is defined. Finally, the formulas for calculating the external and internal significance of features are obtained, and a multilabel feature selection algorithm based on neighborhood intuitive fuzzy entropy is designed to screen the optimal feature subset. Under the multilabel K-nearest neighbor classifier, experimental results on nine multilabel datasets show that the optimal subset selected by the proposed algorithm has great classification effect.

Keywords　Multilabel learning, Feature selection, Fisher score, Multilabel fuzzy neighborhood rough sets, Neighborhood intuitionistic fuzzy entropy

到稿日期:2023-04-04　返修日期:2023-10-25
基金项目:国家自然科学基金(62076089,61772176)
This work was supported by the National Natural Science Foundation of China(62076089,61772176).
通信作者:孙林(slinok@126.com)

1　引言

目前，维度灾难是多标记学习面临的重要挑战之一[1]。特征选择是大规模数据降维的有效手段，可分为过滤、包裹、嵌入等[2]。过滤法是对特征集进行筛选，使用学习算法进行训练，其过程与学习算法无关，计算效率较高、通用性强、冗余度小，适用于处理大规模数据集[3]。包裹法不适合高维数据，通用性弱且计算复杂度高[4]。嵌入法过度依赖具体的学习算法，会出现过拟合现象，缺乏通用性[5]。为了有效处理高维多标记数据集，提升计算效率和避免出现过拟合的情况，本文使用过滤法设计多标记特征选择方法。

Fisher score 算法计算简单、快速高效，适用于大规模数据的过滤式特征选择[5-6]。截至目前，已有较多的研究和应用。例如，Sun 等[6]定义了差异系数改进 Fisher score 算法，并结合邻域粗糙集算法处理数据中的分布不平衡问题。但该算法没有考虑边缘样本的影响。Huang 等[7]将 Fisher score 算法用于多实例学习，来提取固定长度的表示向量。Xu 等[8]将 Fisher score 算法用于彩色图像隐写来评估每个特征的重要性。但上述算法均用于单标记学习。Wu 等[9]引入交叉系数改进 Fisher score 算法进行特征选择，解决了两种类别交叉重复的问题。Zhu 等[10]将牛顿法中的 Hessian 矩阵替换为 Fisher score 矩阵，以消除二阶导数的重复计算并提高收敛鲁棒性。Xu 等[11]引入归一化方法来改进 Fisher score 算法，以提高图像的分类性能。但这 3 种方法都没有考虑边缘样本的影响。同时，上述 6 种改进的 Fisher score 算法都只局限于传统单标记的研究。由此，Fisher score 算法在多标记学习领域得到了深入研究。例如，Sun 等[6]通过计算正、负标记的互信息考虑标记之间的相关性，设计了一种基于互信息的多标记 Fisher Score 算法，对多标记数据进预处理。但是，标记正、负的数量差异较大，造成最终计算结果更偏向数量占比较大的标记，进而影响最终的分类效果。Wang 等[12]发现极值样本会导致标记类中心出现偏差，因此针对文本分类提出了快速多标记 Fisher score 的特征选择算法。但是，样本拥有标记集合在对样本空间划分时会导致类别被多次重复计算。受到上述研究的启发，本文针对多标记数据集，使用结合中心偏移的 Fisher score 模型，对各标记样本计算得分，过滤掉边缘样本，从而对多标记数据进行预处理。

近年来，基于多标记邻域粗糙集模型的特征选择方法逐渐在多标记学习领域中得到了研究和扩展[13]。Liu 等[14]利用多标记中相似实例的分布信息定义新的自适应邻域关系来避免邻域参数的设置。Wu 等[15]通过考虑标记相关性，将相关标记划分为多个标记子集，进而将标记相关性引入邻域粗糙集模型中。然而，这种邻域粗糙集模型无法描述模糊背景下多标记样本的不确定性[16]。Xu 等[17]引入模糊邻域逼近精度，考虑了逼近中的不确定性，并提出了一种针对多标记数据的模糊邻域粗糙集模型。Sun 等[18]通过邻域相似关系定义了模糊邻域粒，提出了一种基于多标记模糊邻域粗糙集的特征选择算法。上述研究表明，模糊邻域粗糙集可以优化邻域相似性关系，从而降低数据分类的错误率，且已逐渐被扩展到多标记学习领域中。Yin 等[19]基于模糊邻域粗糙集提出了一种抗噪声的启发式多标记特征选择算法。但是，模糊邻域相似关系不能准确地反映现实生活中的数据普遍存在的优势关系，也不能根据不同的条件选择所需的数据[20]。基于此，Zhang 等[21]基于直觉模糊粗糙集研究了直觉模糊邻域优势关系，定义了直觉模糊有序信息系统中的邻域优势粗糙集模型。Xin 等[22]指出，直觉模糊集极大地改进了对样本特征的描述，对样本的描述更加准确和具体。信息熵[23]作为信息系统中一种重要的不确定性度量方式，学者们将模糊熵和直觉模糊熵应用于信息系统的不确定性度量，成果显著。Zheng 等[24]基于邻域空间下的二元关系建立了一对粗糙直觉隶属度函数，在邻域粗糙集基础上构造了模糊熵和直觉模糊熵来进行不确定性度量。Yao 等[25]提出了基于自适应邻域空间粗糙集模型的直觉模糊熵特征选择。由此可知，邻域直觉模糊熵能够更细腻地刻画客观世界模糊对象的本质，并在处理不确定信息时具有更强的表达能力与分辨能力。同时，截至目前，基于直觉模糊集的多标记特征选择鲜有研究。本文受上述传统直觉模糊粗糙集特征选择模型的启发，结合模糊集和邻域粗糙集，使用模糊相似关系定义了模糊邻域粒，并将其扩展到多标记特征选择中，建立了多标记直觉隶属度函数和非隶属度函数，提出了一种基于邻域直觉模糊熵的多标记特征选择方法，对直觉模糊粗糙集在多标记学习上的研究具有很好的参考价值。

为了解决上述问题，首先，针对边缘样本对 Fisher score 算法的影响，使用基于中心偏移的多标记 Fisher score 模型对多标记数据集进行预处理，过滤掉边缘样本，从而确保样本分布更集中。然后，引入分类间隔作为自适应模糊邻域半径参数，定义了模糊邻域相似关系以及邻域粒，并构造了多标记模糊邻域上、下近似集，进而构造了多标记模糊邻域粗糙集模型。在此基础上提出了多标记邻域粗糙直觉隶属函数和非隶属函数，由此定义了多标记邻域直觉模糊熵，扩展了直觉模糊熵在多标记模糊邻域粗糙集上的应用。最后，设计了一种基于中心偏移的 Fisher score 与邻域直觉模糊熵的多标记特征选择算法。

2　基础理论

2.1　Fisher score 模型

Fisher score[6]是过滤式特征选择模型，其关键思想是找到一个特征子集，使不同类间距离尽可能大，同一类别的距离尽可能小。Fisher score 具有计算简单、高效，适用于处理大规模数据的优点[12]。

给定训练数据集 $X\in R^{m\times n}$，对于类别 k，特征 i 的 Fisher score 计算公式[6]为：

$$FS(f_i)=\frac{S_B^{(k)}(f_i)}{S_w^{(k)}(f_i)} \tag{1}$$

其中，$S_B^{(k)}(f_i)=\sum_{k=1}^{c}\frac{n_k}{n}(\mu_i^{(k)}-\mu_i)^2$ 指特征 f_i 的类间散度，$S_w^{(k)}(f_i)=\frac{1}{n}\sum_{k=1}^{c}\sum_{x\in w_k}(x_i-\mu_i^{(k)})^2$ 指标记 l_k 下特征 f_i 的类内散度，n 为样本数，n_k 为标记 l_k 下的样本数量，$\mu_i(^k)$ 为标记 l_k 下样本在特征 f_i 下的值，μ_i 表示所有类别样本在特征 f_i 下的均值，x_i 表示在特征 f_i 下的值，$x\in w_k$ 表示标记 l_k 下的样本。

$FS(f_i)$越大表示特征鉴别类别的能力越强。

2.2 直觉模糊集

直觉模糊集作为模糊集的展开式，其不确定性既包括已知信息的模糊性，也包括未知信息的直观性，是 Atanassov 教授[26]在模糊集理论基础上的推广，其分别定义了一对关于样本的粗糙直觉隶属度函数，即隶属度和非隶属度。已知信息模糊性由隶属度与非隶属度的绝对偏差决定，未知信息的直觉性由犹豫度决定。

设在一个非空样本集 U 上，具有如下形式的集合 $B=\{\langle x,\mu_B(x),\nu_B(x)\rangle \mid x\in U\}$称为 U 上的一个直觉模糊集。样本 x 在集合 B 上的隶属度函数 $\mu_B(x)$和非隶属度函数 $\nu_B(x)$满足条件[27]：

$$\begin{cases}0\leqslant\mu_B(x)\leqslant 1\\0\leqslant\nu_B(x)\leqslant 1\\0\leqslant\pi_B(x)\leqslant 1\\\mu_B(x)+\nu_B(x)+\pi_B(x)=1\\x\in X\end{cases}\tag{2}$$

其中，$\pi_B(x)$表示 U 上元素 x 在 B 上的犹豫度。

假设直觉模糊信息系统 $IFIS=\langle U,B,V_{IF},IF\rangle$，其中 U 是非空有限集合，$C=\{c_1,c_2,\cdots,c_{|C|}\}$，$B$ 是 C 的非空特征子集，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，$V_{IF}$ 是所有直觉模糊值的集合，函数 $IF:U\times B\to V_{IF}$，使得 $IF(x,f)=\langle\mu_B(x),\nu_B(x)\rangle$，任意特征 $f\in B$。假设存在决策集 $D=\{l_1,l_2,\cdots,l_z\}$，当 $B=C\cup D$ 且 $C\cap D=\varnothing$ 时，则$\langle X,C\cup D,V_{IF},IF\rangle$被称为直觉模糊决策系统。为了便于表示，将直觉模糊决策系统记为 $IFDS=\langle U,C\cup D,V_{IF}\rangle$。

2.3 多标记邻域粗糙集

给定一个多标记邻域决策系统 $MNDS=\langle U,C\cup D,\cdots,\delta\rangle$，样本集 $U=\{x_1,x_2,\cdots,x_n\}$，特征集 $C=\{f_1,f_2,\cdots,f_{|C|}\}$，标记集 $D=\{l_1,l_2,\cdots,l_{|D|}\}$，$D^j$ 表示标记 l_j 的样本集，D_i 表示样本 x_i 所具有的标记集合，存在映射关系 $f:U\times\{C\cup D\}\to V$，$V=\bigcup_{a\in\{C\cup D\}}Va$，样本在特征 a 下的取值为 V_a，距离函数为 Δ，邻域参数 δ 的取值为$[0,1]$。下文中为了表述方便，将 $MNDS=\langle U,C\cup D,\cdots,\delta\rangle$简写为 $MNDS=\langle U,C\cup D\rangle$。对于任意特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，样本 x 在特征子集 B 下的邻域粒 $N_B^\delta(x)$为：

$$N_B^\delta(x)=\{y\in U\mid\Delta_B(x,y)\leqslant\delta\}\tag{3}$$

其中，$\Delta_B(x,y)$表示两个样本 x 与 y 在 B 下的欧氏距离。基于多标记邻域粗糙集关于 B 的邻域上、下近似集表示为：

$$\overline{N_BD}=\{x_i\in U\mid\forall l_j\in D_i,N_B^\delta(x_i)\cap D^j\neq\varnothing\}\tag{4}$$

$$\underline{N_BD}=\{x_i\in U\mid\forall l_j\in D_i,N_B^\delta(x_i)\subseteq D^j\}\tag{5}$$

其中，$i=1,2,\cdots,n$。

3 多标记特征选择方法

3.1 基于中心偏移的多标记 Fisher score 模型

为了说明边缘样本会给 Fisher score 计算每个特征得分带来具有负作用的信息，参照文献[12]，图 1 给出了两类样本在某一特征下的分布，聚集的样本用圆圈划分出 A 类和 B 类样本。假设每个类别样本下的特征平均值表示每个圆的中心点，可以看出，类别样本越集中，圆的半径越小，对应类内距离越小，类间的距离越大，则式(1)对应的 $FS(f_i)$越大，表示该特征鉴别类别的能力越强。

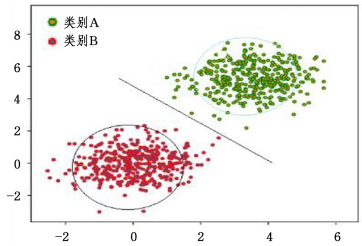

图 1 类中心偏差示意图

Fig. 1 Schematic diagram of class center deviation

由图 1 可以看出，圆的外部存在着部分样本，使用 Fisher score 算法计算类间距离与类内距离时，圆半径以及中心点会受这些极值样本的远近的影响，距离越远，影响越大。但是，理想的情况是类别样本尽可能地集中，因此边缘样本是具有负作用的信息，应该考虑去除这些边缘样本，保留有效样本，使得样本中心发生偏移，得到各个特征的得分。由此可见，部分处于边缘位置的样本会影响 Fisher score 算法的分类效果。

定义 1 在多标记决策系统 $MLDS=\langle U,C\cup D\rangle$中，$U$ 是非空有限集合，$C=\{f_1,f_2,\cdots,f_{|C|}\}$表示特征集，$D=\{l_1,l_2,\cdots,l_{|D|}\}$表示标记集，$D^k$表示具有标记 l_k的样本集，则标记 l_k将多标记论域划分表示为：

$$U^k=U/l_k=\{D_0^k,D_1^k\}\tag{6}$$

$$D_c^k=\{x_s\mid y_s^k=c,c\in\{0,1\}\}\tag{7}$$

其中，U^k表示由标记 $l_k\in D$ 划分得到的样本子集，D_c^k表示标记 l_k下取值为 c 的样本集，y_s^k 表示样本 x_s在标记 l_k下的标记值，c 表示标记取值。

定义 2 在多标记决策系统 $MLDS=\langle U,C\cup D\rangle$中，特征集 $C=\{f_1,f_2,\cdots,f_{|C|}\}$，标记集 $D=\{l_1,l_2,\cdots,l_{|D|}\}$，假设任意样本 $x_s\in U$，由标记 $l_k\in D$ 划分的样本子集 $U^k\subseteq U$，则样本集 D_c^k在特征 f_j下的均值表示为：

$$\overline{D(f_j)_c^k}=\frac{1}{|D_c^k|}\sum_{s=1}^{|D_c^k|}x(f_j)_s\tag{8}$$

其中，$|D_c^k|$表示样本集在标记 l_k下取值为 c 的样本数，$x(f_j)_s$表示样本 x_s在特征 f_j下的特征值。

根据文献[12]，将$\overline{D(f_j)_c^k}$看作样本集原始的中心点，则距离样本中心最远的样本点表示为：

$$x_{\max}=\arg\max\left(\sqrt{(x(f_j)_c^k-\overline{D(f_j)_c^k})^2}\right)\tag{8}$$

其中，$x_{\max}\in D_c^k$，$\sqrt{(x(f_j)_c^k-\overline{D(f_j)_c^k})^2}$表示在特征 f_j上标记 l_k取值为 c 的样本集 D_c^k中，样本到样本中心$\overline{D(f_j)_c^k}$的距离；$\arg\max(\ )$表示取值最大的点集。

根据文献[12]，样本集原始的中心点$\overline{D(f_j)_c^k}$和样本到$\overline{D(f_j)_c^k}$的距离结合距离系数可以重新定义有效样本。

定义 3 在多标记决策系统 $MLDS=\langle U,C\cup D\rangle$中，特征集 $C=\{f_1,f_2,\cdots,f_{|C|}\}$，标记集 $D=\{l_1,l_2,\cdots,l_{|D|}\}$，存在一个随机样本 x_s，在标记 l_k划分的论域空间中存在一个样本集 D_c^k，将$\overline{D(f_j)k_c}$看作样本集 D_c^k的中心点，则以$\overline{D(f_j)_c^k}$为中心点的有效样本集表示为：

$$VDf_c^k=\{x_q\mid\sqrt{(x(f_j)_q-\overline{D(f_j)_c^k})^2}\leqslant dis\_x_{\max}\cdot\lambda\}\quad(10)$$

其中，$x_q\in D_c^k$，$dis_x_{\max}$表示最远的样本点 $x_{\max}$到类中心的距离；$\lambda\in(0,1]$表示距离系数，且 λ 与样本数量呈正相关，当 $\lambda=1$ 时，即相当于不做中心偏移操作。参照文献[12]，设置 $\lambda=0.9$，此时选择的特征子集能够得到最佳的分类效果。

参照文献[14]的基于中心偏移的 Fisher score 模型，结合式(10)将原样本集 D_c^k 作为类中心，以距离最远的样本距离集合乘以距离系数为半径，将距离大于该半径的样本剔除，留下新的样本集。新样本集 $VDf_c{}^k$ 较原始样本分布更为紧密，与原始样本集 D_c^k 相比，可以看作中心发生了偏移，即可理解为进行了中心偏移的操作。

定义 4　在多标记决策系统 $MLDS=\langle U,C\cup D\rangle$ 中，特征集 $C=\{f_1,f_2,\cdots,f_{|C|}\}$，标记集 $D=\{l_1,l_2,\cdots,l_{|D|}\}$，假设任意样本 x_s，由标记 $l_k\in D$ 划分的样本子集 $U^k\subseteq U$，U^k 中任一类样本集 D_c^k 在特征 f_j 上经过中心偏移之后的样本集为 VDf_c^k，则特征 f_j 在样本子集 U^k 经过中心偏移处理之后的 Fisher score 得分表示为：

$$F(f_j^k)=\frac{(\overline{VD(f_j)_{c_i}^k}-\overline{VD(f_j)_{c_t}^k})^2}{\frac{|}{|Df_c^k|}}\Big/\sum_{c_i,c_t\in c,c_i\neq c_t}|VD(f_j)_c^k|\sum_{q=1}^{}(x(f_j)_q^k-\overline{VD(f_j)_c^k})^2\quad(11)$$

其中，$\overline{VD(f_j)_c^k}$ 表示标记 l_k 下新的有效样本集 $VDf_c{}^k$ 在特征 f_j 上的平均值，$|VD(f_j)_c^k|$ 表示新的有效样本集在 l_k 下取值为 c 的样本个数。

定义 5　在多标记决策系统 $MLDS=\langle U,C\cup D\rangle$ 中，特征集 $C=\{f_1,f_2,\cdots,f_{|C|}\}$，标记集 $D=\{l_1,l_2,\cdots,l_{|D|}\}$，假设任意样本 x_q，在标记 l_k 上的 Fisher score 得分为 $F(f_j^k)$，则样本 x_q 在标记集 D 上经过中心偏移的特征 f_j 的得分表示为：

$$F(f_j)=\sum_{k=1}^{|D|}F(f_j^k)\quad(12)$$

其中，$|D|$ 表示标记个数。这里的 $F(f_i)$ 越大，则表明该特征鉴别标记的能力越强。通过对最终得分的逆序排序，参照文献[6]选择前 60% 为最优。

3.2　多标记邻域直觉模糊熵

直觉模糊集相较于模糊集增加了非隶属度函数，可以表示“支持、反对、中立”3 种状态，能够更加细腻地刻画客观世界模糊对象的本质[27]，因而在处理不确定信息时具有更强的表达能力与分辨能力。同时，其在信息融合、模式识别、图像处理、最优化理论、多目标进化优化等领域展现出独特优势[28-29]。在此基础上，将其扩展到多标记特征选择中。在多标记邻域决策系统中，基于模糊相似关系构建自适应邻域空间，使用隶属函数和非隶属函数提出邻域直觉模糊熵，以此来对多标记数据进行不确定性度量，从而准确和细腻地对多标记数据进行特征选择。

定义 6　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，其中 $L=\{l_1,l_2,\cdots,l_{|L|}\}$，对于任意样本 $x\in U$，则自适应模糊邻域半径 δ^F 表示为：

$$\delta^F=\frac{1}{|U||L|}\sum_{i=1}^{|U|}\sum_{k=1}^{|L|}\left(\frac{\Delta_{l_k}(x,NS_{l_k}(x))}{|NS_{l_k}(x)|}-\frac{\Delta_{l_k}(x,NT_{l_k}(x))}{|NT_{l_k}(x)|}\right)\quad(13)$$

其中，$NS_{l_k}(x)$ 和 $NT_{l_k}(x)$ 分别表示样本 x 在标记 l_k 下的异类和同类样本，$\Delta_{l_k}(x,NS_{l_k}(x))$ 和 $\Delta_{l_k}(x,NT_{l_k}(x))$ 分别表示样本 x 在标记 l_k 下关于 $NS_{l_k}(x)$ 和 $NT_{l_k}(x)$ 的距离。

定义 7　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，其中 $L=\{l_1,l_2,\cdots,l_{|L|}\}$。对于任意样本 $x,y\in U$，特征 $f\in B$，则两个样本 x 和 y 关于特征 f 的模糊邻域相似关系表示为：

$$R_f(x,y)=\begin{cases}0, & |F(x,f)-F(y,f)|>\delta^F\\ 1-|F(x,f)-F(y,f)|, & |F(x,f)-F(y,f)|\leqslant\delta^F\end{cases}\quad(14)$$

定义 8　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，其中 $L=\{l_1,l_2,\cdots,l_{|L|}\}$。对于任意样本 $x,y\in U$，则样本 x 关于特征子集 B 的模糊邻域粒表示为：

$$FN_B^\delta(x)=[x]_B(y)=\begin{cases}0, & R_B(x,y)<1-\delta^F\\ R_B(x,y), & R_B(x,y)\geqslant1-\delta^F\end{cases}\quad(15)$$

定义 9　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，$L=\{l_1,l_2,\cdots,l_{|L|}\}$，$D^j$ 表示标记 l_j 的样本集，D_i 表示样本 x_i 所具有的标记集合。根据自适应邻域阈值，特征子集 B 下的多标记模糊邻域的上、下近似集分别表示为：

$$\overline{FN_BD}=\{x_i\in U\mid\forall l_j\in D_i,FN_B^\delta(x_i)\cap D^j\neq\varnothing\}\quad(16)$$

$$\underline{FN_BD}=\{x_i\in U\mid\forall l_j\in D_i,FN_B^\delta(x_i)\subseteq D^j\}\quad(17)$$

由此，可以将论域划分为正域：$POS_B(D)=\underline{FN_BD}$，负域：$NEG_B(D)=U-\overline{FN_BD}$，边界域：$BND_B(D)=\overline{FN_BD}-\underline{FN_BD}$，进而可以定义多标记模糊邻域粗糙集模型。

定义 10　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，$B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，$L=\{l_1,l_2,\cdots,l_{|L|}\}$，$D^j$ 表示标记 l_j 的样本集，D_i 表示样本 x_i 所具有的标记集合，对于样本子集 $X\subseteq U$，任意样本 $x\in U$ 在 B 下关于 D 的多标记邻域粗糙直觉隶属度函数和非隶属度函数$\langle\mu_{IF_X^B}(x),\nu_{IF_X^B}(x)\rangle$分别定义为：

$$\mu_{IF_X^B}(x)=\begin{cases}1, & x\in POS_B(D)\\ \min\limits_{1\leqslant i\leqslant|B|}\dfrac{|\delta_B(x)\cap D|}{|\delta_B(x)|}, & x\in BND_B(D)\\ 0, & x\in NEG_B(D)\end{cases}\quad(18)$$

$$\nu_{IF_X^B}(x)=\begin{cases}0, & x\in POS_B(D)\\ 1-\max\limits_{1\leqslant i\leqslant|B|}\dfrac{|\delta_B(x)\cap D|}{|\delta_B(x)|}, & x\in BND_B(D)\\ 1, & x\in NEG_B(D)\end{cases}\quad(19)$$

其中，$\mu_{IF_X^B}(x)$ 表示邻域粗糙直觉模糊集中样本 x 隶属于 D 的程度，$\nu_{IF_X^B}(x)$ 表示样本 x 不隶属于 D 的程度。根据这两个函数可以得到样本 x 关于 D 的犹豫度，表示为：

$$\pi_{IF_X^B}(x)=1-\mu_{IF_X^B}(x)-\nu_{IF_X^B}(x)\quad(20)$$

通过多标记邻域粗糙直觉隶属度函数、非隶属度函数和犹豫度描述样本与集合之间的关系，不仅符合人类的认知观，而且能够更好地反映出集合的不确定性。

性质 1　在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$

中，特征子集 $B\subseteq C, B=\{f_1,f_2,\cdots,f_{|B|}\}$，样本子集 $X\subseteq U$，IF_X^B 是由邻域决策系统产生的直觉模糊集，可以得到如下性质：

(1)当 $x\in POS_B(X)$，有 $\mu_{IF_X^B}(x)=1,\nu_{IF_X^B}(x)=1$；

(2)当 $x\in NEG_B(X)$，有 $\mu_{IF_X^B}(x)=0,\nu_{IF_X^B}(x)=1$；

(3)当 $x\in BND_B(X)$，有 $0\leqslant\mu_{IF_X^B}(x)\leqslant 1, 0\leqslant\nu_{IF_X^B}(x)\leqslant 1$。

定义 11 多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C, B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D$，$L=\{l_1,l_2,\cdots,l_{|L|}\}$，$D^j$ 表示标记 l_j 的样本集，D_i 表示样本 x_i 所具有的标记集合。对于任意样本子集 $X\subseteq U$，则样本 $x\in U$ 在特征子集 B 下关于 X 的多标记邻域直觉模糊熵 $IE(IF_X^B)$ 表示为：

$$IE(IF_X^B)=\frac{1}{|L|}\frac{1}{|D^j|}\sum_{k=1}^{|L|}\sum_{x_i\in D^j}\Phi(x_i) \tag{21}$$

其中，$|L|$ 为标记数量。$\Phi(x)$ 与 $G(t)$ 分别为：

$$\Phi(x)=\begin{cases}\dfrac{1}{\pi_{IF_X^B}(x)}\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t, & \pi_{IF_X^B}(x)\neq 0\\ G(\mu_{IF_X^B}(x)), & \pi_{IF_X^B}(x)=0\end{cases} \tag{22}$$

$$G(t)=\begin{cases}-t\times lbt-(1-t)\times lb(1-t), & t\in(0,1)\\ 0, & t=0\vee 1,\end{cases} \tag{23}$$

$$\pi_{IF_X^B}(x)=1-\mu_{IF_X^B}(x)-\nu_{IF_X^B}(x) \tag{24}$$

性质 2 在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，存在 B 诱导的自适应邻域空间，任意样本子集 $X\subseteq U$ 关于 B 的邻域直觉模糊熵将满足如下性质：

(1)$0\leqslant IE(IF_X^B)\leqslant 1$；

(2)$IE(IF_X^B)=0$ 当且仅当任意样本 $x\in U$ 时，有 $\mu_{IF_X^B}(x)=1,\nu_{IF_X^B}(x)=0$ 或者 $\mu_{IF_X^B}(x)=0,\nu_{IF_X^B}(x)=1$；

(3)$IE(IF_X^B)=1$ 当且仅当任意样本 $x\in U$ 时，有 $\mu_{IF_X^B}(x)=0.5$。

证明：首先，为证明性质(1)，分析式(23)即函数 $G(t)=\begin{cases}-xlbx-(1-x)lb(1-x), & x\in(0,1)\\ 0, & x=0\vee 1\end{cases}$ 的单调性，对 $G(t)$ 求导，可得 $G'(t)=\ln\left(\dfrac{1}{t}-1\right)$。当 $1<t<0.5$ 时，函数 $G(t)$ 单调递增；当 $0.5<t<1$ 时，函数 $G(t)$ 单调递减，所以 $0\leqslant G(t)\leqslant 1$，$0\leqslant\mu_{IF_X^B}(x)\leqslant 1, 0\leqslant\nu_{IF_X^B}(x)\leqslant 1$。当 $\pi_{IF_X^B}(x)\neq 0$ 时，$0\leqslant\dfrac{1}{\pi_{IF_X^B}(x)}\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t\leqslant\frac{1}{\pi_{IF_X^B}(x)}\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}\mathrm{d}t$，即 $0\leqslant\dfrac{1}{\pi_{IF_X^B}(x)}\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t\leqslant 1$。性质(1)得证。接下来，运用反证法，假设存在任意样本 $x\in U, 0<\mu_{IF_X^B}(x)\leqslant 0.5$ 且 $0<\nu_{IF_X^B}(x)\leqslant 0.5$，根据定义 9 有 $\Phi(x)>0$，则 $IE(IF_X^B)>0$。性质(2)成立。假设存在任意样本 $x\in U, 0\leqslant_{IF_X^B}(x)<0.5$ 且 $0\leqslant\nu_{IF_X^B}(x)<0.5$，根据定义 8 有 $\Phi(x)<1$，则 $IE(IF_X^B)<1$。性质(3)成立。

性质 3 在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，任意特征子集 $B_1\subseteq B_2\subseteq C$，标记子集 $L\subseteq D, L=\{l_1,l_2,\cdots,l_{|L|}\}$，对于任意样本子集 $X\subseteq U$ 关于特征子集 B_1 和 B_2 的多标记邻域直觉模糊熵 $IE(IF_X^B)$，有 $IE(IF_X^{B_2})\leqslant IE(IF_X^{B_1})$。

证明：如果 $\pi_{IF_X^{B_2}}(x)=\pi_{IF_X^{B_2}}(x)=0$，根据定义 10，显然 $IE(IF_X^{B_2})=IE(IF_X^{B_1})$ 成立。如果 $\pi_{IF_X^{B_1}}(x)=\pi_{IF_X^{B_2}}(x)=\varepsilon>0$，由性质 1 可知，当 $0<t<0.5$，函数 $G(t)$ 单调递增；当 $0.5<t<1$，函数 $G(t)$ 单调递减。如果 $\pi_{IF_X^B}(x)$ 保持不变，那么当 $\mu_{IF_X^B}(x)=\nu_{IF_X^B}(x)$ 时，$\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t$ 达到最大。当 $\mu_{IF_X^B}(x)\leqslant\nu_{IF_X^B}(x)$ 时，$\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t$ 是关于 $\mu_{IF_X^B}(x)$ 的单调递增函数；当 $\mu_{IF_X^B}(x)\geqslant\nu_{IF_X^B}(x)$，$\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t$ 是关于 $\mu_{IF_X^B}(x)$ 的单调递减函数。则 $|\mu_{IF_X^{B_1}}(x)-\nu_{IF_X^{B_1}}(x)|\geqslant|\mu_{IF_X^{B_2}}(x)-\nu_{IF_X^{B_2}}(x)|$ 分如下 4 种情形进行讨论：

(1)当 $\mu_{IF_X^{B_1}}(x)\leqslant\nu_{IF_X^{B_1}}(x)$ 且 $\mu_{IF_X^{B_2}}(x)\leqslant\nu_{IF_X^{B_2}}(x)$，可得 $\nu_{IF_X^{B_1}}(x)-\mu_{IF_X^{B_1}}(x)\geqslant\nu_{IF_X^{B_2}}(x)-\mu_{IF_X^{B_2}}(x)$，又 $\nu_{IF_X^{B_1}}(x)+\mu_{IF_X^{B_1}}(x)=\nu_{IF_X^{B_2}}(x)+\mu_{IF_X^{B_2}}(x)$，则有 $\mu_{IF_X^{B_1}}(x)\leqslant\mu_{IF_X^{B_2}}(x)$。

(2)当 $\mu_{IF_X^{B_1}}(x)\geqslant\nu_{IF_X^{B_1}}(x)$ 且 $\mu_{IF_X^{B_2}}(x)\geqslant\nu_{IF_X^{B_2}}(x)$，可得 $\mu_{IF_X^{B_1}}(x)-\nu_{IF_X^{B_1}}(x)\geqslant\mu_{IF_X^{B_2}}(x)-\nu_{IF_X^{B_2}}(x)$，又 $\nu_{IF_X^{B_1}}(x)+\mu_{IF_X^{B_1}}(x)=\nu_{IF_X^{B_2}}(x)+\mu_{IF_X^{B_2}}(x)$，则有 $\mu_{IF_X^{B_1}}(x)\geqslant\mu_{IF_X^{B_2}}(x)$。

(3)当 $\mu_{IF_X^{B_1}}(x)\leqslant\nu_{IF_X^{B_1}}(x)$ 且 $\mu_{IF_X^{B_2}}(x)\geqslant\nu_{IF_X^{B_2}}(x)$，可得 $1-\mu_{IF_X^{B_2}}(x)\leqslant 1-\nu_{IF_X^{B_2}}(x)$ 和 $\nu_{IF_X^{B_1}}(x)-\mu_{IF_X^{B_1}}(x)\geqslant\mu_{IF_X^{B_2}}(x)-\nu_{IF_X^{B_2}}(x)$，又 $\nu_{IF_X^{B_1}}(x)+\mu_{IF_X^{B_1}}(x)=\nu_{IF_X^{B_2}}(x)+\mu_{IF_X^{B_2}}(x)=1-\lambda$，则有 $\mu_{IF_X^{B_1}}(x)\leqslant\nu_{IF_X^{B_2}}(x)\leqslant\mu_{IF_X^{B_2}}(x)$ 且 $\mu_{IF_X^{B_2}}(x)<1-\nu_{IF_X^{B_2}}(x)$。

(4)当 $\mu_{IF_X^{B_1}}(x)\geqslant\nu_{IF_X^{B_1}}(x)$ 且 $\mu_{IF_X^{B_2}}(x)\leqslant\nu_{IF_X^{B_2}}(x)$，可得 $1-\mu_{IF_X^{B_1}}(x)\leqslant 1-\nu_{IF_X^{B_1}}(x)$ 和 $\mu_{IF_X^{B_1}}(x)-\nu_{IF_X^{B_1}}(x)\geqslant\nu_{IF_X^{B_2}}(x)-\mu_{IF_X^{B_2}}(x)$，又 $\nu_{IF_X^{B_1}}(x)+\mu_{IF_X^{B_1}}(x)=\nu_{IF_X^{B_2}}(x)+\mu_{IF_X^{B_2}}(x)=1-\lambda$，则有 $\mu_{IF_X^{B_1}}(x)\leqslant\mu_{IF_X^{B_2}}(x)\leqslant\nu_{IF_X^{B_2}}(x)$ 且 $\nu_{IF_X^{B_2}}(x)<1-\mu_{IF_X^{B_2}}(x)$。

综上所述，可以推出 $\displaystyle\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t\leqslant\int_{\mu_{IF_X^B}(x)}^{1-\nu_{IF_X^B}(x)}G(t)\mathrm{d}t$，因此，$IE(IF_X^{B_2})\leqslant IE(IF_X^{B_1})$ 成立，其中 $B_1=B_2$ 时等号成立。

从上述性质可知，随着特征个数增加，多标记邻域直觉模糊熵单调递增。

3.3 多标记特征选择算法描述

定义 12 在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C, B=\{f_1,f_2,\cdots,f_{|B|}\}$，标记子集 $L\subseteq D, L=\{l_1,l_2,\cdots,l_{|L|}\}$，$D^j$ 表示标记 l_j 的样本集，D_i 表示样本 x_i 所具有的标记集合，则特征子集 B 为 C 的相对约简当且仅当如下条件成立：

(1)$IE(IF_{U/D}^B)=IE(IF_{U/D}^C)$；

(2)对任意特征 $a\in B, IE(IF_{U/D}^{B-\{a\}})>IE(IF_{U/D}^B)$。

由定义 12 可知，相对约简既限定了约简特征子集 B 的邻域直觉模糊熵不能低于特征集 C 的邻域直觉模糊熵，又限定了约简集的极小性。

定义 13 在多标记邻域决策系统 $MNDS=\langle U,C\cup D,\delta^F\rangle$ 中，特征子集 $B\subseteq C$，标记子集 $L\subseteq D, L=\{l_1,l_2,\cdots,l_{|L|}\}$，任意特征 $a_i\in C-B$，则特征 a_i 在 B 中相对决策 D 的外部重要度表示为：

$$SIG_{\text{out}}(a_i,B,D)=|IE(IF_{U/D}^B)-IE(IF_{U/D}^{B\cup\{ai\}})| \tag{25}$$

定义 14 在多标记邻域决策系统 $MNDS=\langle U,C\cup D,$

$\delta^F\rangle$中，特征子集 $B\subseteq C$，标记子集 $L\subseteq D$，$L=\{l_1,l_2,\cdots,l_{|L|}\}$，任意特征 $a_i\in B$，特征 a_i 在 B 中相对决策 D 的内部重要度表示为：

$$SIG_{in}(a_i,B,D)=|IE(IF^B_{U/D})-IE(IF^{B-\{a_i\}}_{U/D})| \tag{26}$$

式(26)反映了从当前特征子集 B 中删去特征 a_i 后，多标记邻域直觉模糊熵的变化程度。在此基础上，借助正向贪心搜索算法迭代地选择具有最大重要度的特征，当加入特征后，确定性规则生成不再受到影响时，则算法终止。

接下来，为实现多标记特征选择算法，首先构建基于中心偏移的多标记 Fisher score 算法(Multilabel Fisher Score Algorithm with Center Shift，MFSCS)，如算法 1 所示。在算法 1 的基础上，基于中心偏移的 Fisher score 与邻域直觉模糊熵设计了多标记特征选择算法(Multilabel Feature Selection Algorithm based on Fisher Score with Center Shift and Neighborhood Intuitionistic Fuzzy Entropy，MFSNE)，如算法 2 所示。算法 1 和算法 2 的伪代码分别描述如下。

算法 1　MFSCS 算法

输入：多标记决策系统 MLDS=〈U，C∪D〉

输出：候选特征子集 S

1. 初始化 F(f_j)=0
2. For 每个标记 k=1:z
3.　For 每个特征 j=1:m
4.　　For 每个样本 i=1:n
5.　　根据式(8)和式(9)计算标记 l_k 下的特征 Fisher score 得分
6.　　F(f_j)=F(f_j) +F(f_j^k)
7.　　End
8.　End
9. End
10. 对 F(f_j)进行逆序，输出前 60%特征子集 S

在算法 1 中，主要的计算消耗在于对每个标记计算得分，假设多标记数据集有 n 个样本、m 个特征和 l 个标记，则计算全体特征得分的代价为 O(m)，步骤 3 计算全体标记下特征得分的计算复杂度为 O(mz)，则步骤 4－步骤 9 得到的计算复杂度为 O(nmz)，步骤 10 对最终特征得分进行排序的计算复杂度为 O($mlbm$)，参照文献[6]，选择前 60%的特征作为最佳选择。则算法 1 总的计算复杂度为 O($mz+nmz+mlbm$)。在大多情况下 $n\geqslant m$，算法 1 的计算复杂度接近 O(nmz)。

算法 2　MFSNE 算法

输入：多标记邻域决策系统 MLDS=〈U，S∪D〉

输出：最优特征子集 red

1. 初始化 red=Ø 且 $E(F^{\emptyset}_{U/D})=1$
2. For 每个样本 $x_i\in U$
3.　根据式(13)和式(15)得到 x_i 的模糊邻域半径和模糊邻域粒
4. End For
5. For 每个候选特征 $a_k\in S-$ red
6.　For 每个标记 $l_j\in D$ 和 $D^j\in U/D$
7.　If x_i 是否位于 D^j 边界域中
8.　根据式(18)和式(19)计算隶属度函数 $\mu_{IF^B_X}(x)$ 与非隶属度函数 $\nu_{IF^B_X}(x)$ 以及犹豫度 $\pi_{IF^B_X}(x)$
9.　If $\pi_{IF^B_X}(x)=0$
10.　$IE(IF^{red}_D)\leftarrow IE(IF^{red}_D)+\Phi(\mu_{IF^B_X}(x))$
11.　Else If
$IE(IF^{red}_{D^j})\leftarrow IE(IF^{red}_{D^j})+\Phi(\mu_{IF^B_X}(x))$
12.　Else If
$$IE(IF^{red}_{D^j})\leftarrow IE(IF^{red}_{D^j})+\frac{1}{\pi_{IF^B_X}(x)}\int_{\mu_{IF^B_X}(x)}^{1-\nu_{IF^B_X}(x)}G(t)dt$$
13.　End If
14.　End If
15.　End For
16. $IE(IF^{red}_D)\leftarrow IE(IF^{red}_D)+IE(IF^{red}_{D^j})$
17. End For
18. 根据式(26)计算特征 a_k 的 $SIG_{out}(a_k,red,D)$
19. 选择 a_t 满足 $SIG_{out}(a_t,red,D)=\max\{SIG_{out}(a_k,red,D)\}$
20. If $SIG_{out}(a_t,red,D)>\lambda$ // * 根据文献[25]设置终止阈值 $\lambda=0.001$ * //
21. Then red←red∪a_t 转向步骤 7
22. Else 返回最优特征子集 red

在算法 2 中，假设算法 1 得到的候选特征个数为 s，算法 2 得到的最终特征子集为 r。步骤 4－步骤 6 计算模糊邻域半径与模糊邻域粒，其复杂度为 O($s(nlbn+z)$)；步骤 9－步骤 13 计算多标记邻域直觉模糊熵，其复杂度为 O(nmz)，则步骤 4－步骤 13 总的计算复杂度为 O($s(nlbn+z)nml$)；步骤 7－步骤 21 使用多标记邻域直觉模糊熵进行特征选择过程，假设最终约简的特征子集个数为 r，则其计算复杂度为 $O(szn+z+nlbn)+(s-1)(2nz+2(z+nlbn))+\cdots+(s-r+1)(rnz+r(z+nlbn)))=O((nz+z+nlbn)\sum_{i=1}^{r}(r-I+1))$。因此算法 2 的计算复杂度为 $O((nz+z+nlbn)\sum_{i=1}^{r}(r-i+1))$。由于先进行算法 1 得出候选子集，因此算法 1 和算法 2 总的计算复杂度为 $O(mz+nz+mlbm)+O((nz+z+nlbn)\sum_{i=1}^{r}(r-i+1))=O(((3sr^2-2r^3+(3s+2)r)/6)\times(nz+z+nlbn)+nmz)$，当 $r=s$ 时为最坏的计算复杂度，为 $O((s^3+3s^2+2s)/6)\times(nz+z+nlbn)+nmz))$。通常情况下，$z\ll n$，因此算法 2 最终的计算复杂度不超过 O($s^3nz$)。

4　实验结果与分析

4.1　实验准备

为了测试 MFSNE 算法的有效性，从 Mulan 数据库[1)]中选择了 9 个多标记数据集，如表 1 所列。采用多标记 K 最近邻(Multilabel K-Nearest Neighbor，MLKNN)分类器和 5 个评价指标对算法性能进行分析，包括平均分类精度(Average Precision，AP)、汉明损失(Hamming Loss，HL)、排序损失(Ranking Loss，RL)、1-错误率(One Error，OE)、覆盖率(Coverage，CV)。另外，MLKNN 分类器的平滑参数 $s=1$，近邻数 $k=10$，将选择的特征个数记为 FN(Number of Selected Features)。实验中粗体表示最佳值。实验环境为 Windows 10，RAM 16GB，Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz，编程软件为 MATLAB R2016a。

1) http://mulan.sourceforge.net/datasets.html

表 1 9 个多标记数据集描述

Table1 Description of nine multilabel datasets

序号	数据集	样本数	特征数	标记数	所属领域	训练集	测试集
1	Birds	645	260	19	Audio	322	323
2	Cal500	502	68	174	Music	335	167
3	Computer	5000	681	159	Text	2000	3000
4	Emotion	593	72	6	Music	391	202
5	Flags	194	19	10	Images	129	65
6	Image	600	294	5	Image	400	200
7	Enron	1702	640	21	Text	1123	579
8	Scene	2407	294	6	Image	1211	1196
9	Yeast	2417	103	14	Biology	1500	917

4.2 MLKNN 分类器下的实验结果对比

为了检验 MFSNE 算法的分类效果，本文选择 8 种算法进行比较，包括：基于多标准决策的多标记特征选择算法(MFS-MCDM)[30]、基于共享的潜在特征和标记结构特征选择算法(SSFS)[31]、基于可扩展准则的大型标记集多标记特征选择算法(SCLS)[32]、基于最大依赖和最小冗余的多标记特征选择算法(MDMR)[33]、基于多元互信息的多标记特征选择算法(PMU)[34]、基于信息论特征排序的快速多标记特征选择算法(FIMF)[35]、基于改进 ReliefF 的多标记特征选择算法(MFSR)[36]和基于 Fisher score 与模糊邻域熵的多标记特征选择算法(MLFSF)[37]。另外，为了阐明 MFSCS 算法的有效性，我们增加了 3 种对比算法：多标记 Fisher score 算法(MLFS)[6]、基于中心偏移的多标记 Fisher score 算法(MFSCS)和基于邻域直觉模糊熵的多标记特征选择算法(MLNIE)。需要说明的是，MLNIE 算法不采用本文算法 1 的结果而是直接利用算法 2 做特征选择。在 9 个数据集上通过 5 个指标(AP，HL，RL，CV 和 OE)评估算法的分类性能。为了保持实验结果的一致性，MFS-MCDM，SSFS，SCLS，MDMR，PMU 和 FIMF 这 6 种算法的实验数据均出自文献[38]。为了保证实验的公平性，实验参数均按照文献[38]设置，9 个数据集的训练集和测试集与其一致，FN 设置为[5，10，15，20，25，30，35，40，45，50]，对于特征数小于 50 的 Flags 数据集的 FN 设置为[2，4，6，8，10，12，14，16，18]。表 2 列出了 12 种算法在 9 个多标记数据集上的 5 个指标的实验对比结果。“↑”表示指标值越大越好，“↓”表示指标值越小越好。

表 2 12 种算法在 9 个多标记数据集上的 5 种评价指标的实验结果

Table 2 Experimental results of twelve algorithms on nine multilabel datasets in terms of five metrics

指标	算法	Birds	Cal500	Computer	Emotion	Enron	Flags	Image	Scene	Yeast
AP(↑)	MFS-MCDM	0.5302	0.4966	0.6134	0.7815	0.6103	0.8425	0.7288	0.8058	0.7551
	SSFS	0.5143	0.4945	0.6059	0.7584	0.6585	0.8372	0.7208	0.7727	0.7312
	SCLS	0.4435	0.4942	0.6317	0.7496	**0.6589**	0.8024	0.7437	0.8163	0.7563
	MDMR	0.4158	0.4959	0.6304	0.7551	0.6566	0.8462	0.7058	0.7633	0.7579
	PMU	0.4435	0.4930	0.6093	0.7143	0.6483	0.8411	0.7002	0.8034	0.7562
	FIMF	0.4074	0.4959	0.6203	0.7510	0.6548	0.8410	0.6791	0.6906	0.7552
	MFRS	0.6976	0.4881	0.6218	0.8130	0.5576	0.8113	0.6588	0.7403	0.7543
	MLFSF	0.6903	0.4942	0.6169	0.7750	0.5760	0.8385	0.6735	0.8078	0.7457
	MLFS	0.7019	0.4950	0.6061	0.7999	0.5917	0.8239	0.7257	0.7481	0.7471
	MFSCS	0.7014	0.4956	0.6333	0.8020	0.5809	0.8321	0.7240	0.7433	0.7562
	MLNIE	0.7106	0.4951	0.6181	0.7943	0.6214	0.8321	0.7010	0.8108	0.7421
	MFSNE	**0.7327**	**0.4972**	**0.6387**	**0.8175**	0.6404	**0.8502**	**0.7698**	**0.8231**	**0.7614**
HL(↓)	MFS-MCDM	**0.0471**	0.9655	0.0421	0.2302	0.0544	0.5802	0.2050	0.1066	0.2014
	SSFS	0.0495	0.9648	0.0427	0.2418	0.0498	0.6413	0.2164	0.1290	0.2137
	SCLS	0.0499	0.9651	0.0398	0.2500	0.0495	0.6330	0.2110	0.1073	0.2006
	MDMR	0.0505	0.9657	0.0398	0.2409	0.0505	0.5934	0.2240	0.1348	0.1999
	PMU	0.0504	0.9667	0.0416	0.2673	0.0505	0.5934	0.2270	0.1137	0.2006
	FIMF	0.0520	0.9657	0.0409	0.2252	0.0501	0.5934	0.2340	0.1587	0.2021
	MFRS	0.0502	0.1392	0.0400	0.2063	0.0578	0.2812	0.2440	0.1405	0.1998
	MLFSF	0.0562	0.1379	0.0407	0.2062	0.0562	0.2803	0.2279	0.1161	0.2064
	MLFS	0.0526	0.1378	0.0425	0.1889	0.0544	0.2835	0.1980	0.1388	0.2036
	MFSCS	0.0556	0.1379	0.0390	0.2104	0.0557	0.2817	0.2090	0.1469	0.1976
	MLNIE	0.0529	0.1371	0.0408	0.1964	0.0525	0.2813	0.2239	0.1102	0.2091
	MFSNE	0.0491	**0.1365**	**0.0389**	**0.1561**	0.0511	**0.2462**	**0.1841**	**0.1045**	**0.1947**
RL(↓)	MFS-MCDM	0.1944	0.1823	0.0982	0.1864	0.1034	0.1823	0.2258	0.1201	0.1742
	SSFS	0.2138	0.1831	0.1004	0.2010	0.0913	0.2078	0.2477	0.1463	0.1925
	SCLS	0.2655	0.1833	0.0909	0.2056	0.0921	0.2482	0.2167	0.1129	0.1745
	MDMR	0.2591	**0.1818**	0.0903	0.1994	0.0944	**0.1813**	0.2550	0.1444	0.1710
	PMU	0.2585	**0.1818**	0.0980	0.2570	0.0949	0.1977	0.2483	0.1290	0.1723
	FIMF	0.2586	**0.1818**	0.0955	0.2012	0.0935	0.1823	0.2662	0.1994	0.1747
	MFRS	0.1278	0.1879	0.0929	0.1579	0.1025	0.2000	0.3117	0.1529	0.1777
	MLFSF	0.1288	0.1893	0.0997	0.1921	0.1025	0.2023	0.2732	0.1138	0.1832
	MLFS	0.1153	0.1874	0.1001	0.1650	0.1035	0.2067	0.2354	0.1498	0.1832
	MFSCS	0.1211	0.1884	0.0862	0.1676	0.0973	0.1944	0.2479	0.1515	0.1738
	MLNIE	0.1169	0.1882	0.0973	0.1630	0.1003	0.1979	0.2732	0.1169	0.1847
	MFSNE	**0.1127**	0.1819	**0.0842**	**0.1560**	**0.0913**	0.1907	**0.2102**	**0.1082**	**0.1691**

（续表）

指标	算法	Birds	Cal500	Computer	Emotion	Enron	Flags	Image	Scene	Yeast
OE(↓)	MFS-MCDM	0.5465	**0.0838**	0.4650	0.3119	0.3472	0.1406	0.4200	0.3169	0.2356
	SSFS	0.5872	**0.0838**	0.4730	0.3455	0.2435	0.1500	0.4300	0.3661	0.2508
	SCLS	0.6454	0.0898	0.4580	0.3614	0.2470	0.2031	0.4000	0.2977	0.2298
	MDMR	0.6744	0.0898	0.4543	0.3564	**0.2435**	0.1563	0.4450	0.3905	0.2366
	PMU	0.6395	0.0898	0.4700	0.3614	0.2694	0.1719	0.4700	0.3904	0.2366
	FIMF	0.7035	0.0898	0.4627	0.3515	0.2453	0.1250	0.5000	0.4983	0.2366
	MFRS	0.3746	0.1557	0.4537	0.2426	0.3869	0.1846	0.5300	0.4323	0.2290
	MLFSF	0.3622	0.0958	0.4587	0.2970	0.3748	0.1385	0.5075	0.3227	0.2388
	MLFS	0.3467	0.0898	0.4730	0.2723	0.3282	0.2154	0.4200	0.4189	0.2399
	MFSCS	0.3467	0.0898	0.4523	0.2525	0.3627	0.1692	0.4229	0.4281	0.2377
	MLNIE	0.3529	0.0958	0.4570	0.2723	0.2902	0.1846	0.4627	0.3069	0.2563
	MFSNE	**0.2977**	0.0898	**0.4500**	**0.2327**	0.2850	**0.0635**	**0.3333**	**0.2901**	**0.2257**
CV(↓)	MFS-MCDM	**2.2755**	**127.990**	4.6820	1.9851	13.9260	3.6769	1.1750	0.7032	6.4569
	SSFS	2.5542	129.300	4.7290	2.0842	**12.8950**	3.8400	1.2520	0.8329	6.6772
	SCLS	3.2012	129.370	4.3697	2.1139	13.0415	4.0460	1.1650	0.6681	6.4482
	MDMR	3.0495	128.990	4.3480	2.0891	13.1606	3.7308	1.3200	0.8253	6.3642
	PMU	3.0526	128.990	4.6520	2.3614	13.4128	3.7231	1.2900	0.7492	6.3708
	FIMF	3.0526	128.900	4.5580	2.0545	13.2038	3.7000	1.3550	1.0953	6.3740
	MFRS	3.4644	130.371	4.4272	1.7973	13.9827	3.7231	1.4750	0.8637	6.3522
	MLFSF	3.4768	130.120	4.6510	1.9307	14.1813	3.7692	1.3532	0.6622	6.5529
	MLFS	3.3684	130.186	4.6620	1.8317	14.3903	3.7846	1.1900	0.8428	6.4613
	MFSCS	3.4644	129.988	4.1140	1.9257	13.6235	3.7385	1.2289	0.8411	6.2814
	MLNIE	3.3189	130.551	4.5837	1.8218	13.8843	3.7385	1.4030	0.6823	6.4798
	MFSNE	2.9674	128.766	**4.0180**	**1.7970**	13.0147	**3.4923**	**1.1443**	**0.6229**	**6.2650**

分析表 2 的实验结果，从 AP 指标来看，与其他 11 种对比算法相比，MFSNE 算法在 Birds，Cal500，Computer，Emotion，Flags，Image，Scene 和 Yeast 这 8 个数据集上表现最优；在 Enron 数据集上，与最优的 MDMR 算法相差 1.85%；同样，MLFS 和 MFSCS 这两个使用 Fisher score 算法的 AP 值都较低，推测可能是 Fisher score 算法不适用于这种离散型数值的数据集。因此，在 AP 指标上，MFSNE 在大部分数据集上表现较好。从 HL 指标来看，与其他 11 种算法相比，MFSNE 算法在 Cal500，Computer，Emotion，Flags，Image，Scene 和 Yeast 这 7 个数据集上表现最优。其中，在 Cal500 数据集上，MFSNE，MFRS，MLFSF，MLFS，MFSCS 和 MLNIE 这 6 种算法的平均值为 0.1377，比其他 6 种算法的平均值低 0.8278；在 Emotion 和 Flags 数据集上，MFSNE 比次优的 MLFSF 算法分别低了 3.28%和 3.42%；在 Birds 数据集上，仅比最优的 MFS-MCDM 算法高出 0.2%；在 Enron 数据集上，比最优的 MFSCS 算法略低 0.16%，原因是 MFSNE 漏选了部分重要特征，使得该指标下降。因此，在 HL 指标上，MFSNE 算法整体表现较好。从 RL 指标来看，与其他 11 种算法相比，MFSNE 算法在 Birds，Computer，Emotion，Enron，Image，Scene 和 Yeast 这 7 个数据集上表现最优。在 Cal500 数据集上，与同为最优的 MDMR 算法、PMU 算法和 FIMF 算法仅相差 0.01%；在 Flags 数据集上，由于漏选和已选特征相关性高的特征，导致 MFSNE 算法未达到最优，与最优的 MDMR 算法相差 0.94%。因此，在 RL 指标上，MFSNE 算法整体上效果良好。从 OE 指标来看，与其他 11 种算法相比，MFSNE 算法在 Birds，Computer，Emotion，Flags，Image，Scene 和 Yeast 这 7 个数据集上表现最优。其中，在 Birds，Flags 和 Image 这 3 个数据集上表现较好，分别比次优的 MLFS 算法、FIMF 算法和 SCLS 算法低了 4.91%，6.15%和 6.67%；在 Cal500 数据集上，与同时取得最优的 MFS-MCDM 算法和 SSFS 算法相差 0.60%；在 Enron 数据集上，与最优的 SCLS 算法相差 4.15%。从 OE 指标的含义来看，这可能是因为选择的部分特征导致部分预测序列靠前的标记预测错误。因此，在 OE 指标上，MFSNE 算法整体表现较好。从 CV 指标来看，与其他 11 种算法相比，MFSNE 算法在 Computer，Emotion，Flags，Image，Scene 和 Yeast 这 6 个数据集上表现最优。在 Birds 数据集上，MFSNE 算法比最优的 MFS-MCDM 算法高 0.6919，高于 SSFS 算法，低于其他 9 种算法；在 Cal500 数据集上，MFSNE 算法比最优的 MFS-MCDM 算法高 0.776，排名第二；在 Enron 数据集上，与最优 SSFS 算法相差 0.1197，低于其他 10 种对比算法，这可能是部分标记特定特征被漏选，导致 CV 值下降。因此，MFSNE 算法在 CV 指标上的大部分数据集上性能均有提升。

从表 2 的 5 个指标整体上来看，MFSNE 算法在大部分数据集上表现较好且有明显提升。具体来讲，在 Computer，Emotion，Image，Scene 和 Yeast 这 5 个数据集上的 5 个指标均取得最优结果；在 Birds 数据集上，在 AP，RL 和 OE 这 3 个指标上，优势较为明显，在 HL 指标和 CV 指标上，虽未取得最优值，但排名分别为第二和第三，与最优值的差距较小；在 Cal500 数据集上，在 AP 和 HL 这 2 个指标上均取得最优值，在 RL，OE 和 CV 指标上的排名分别为第四、第四和第二，算法之间的差距较小；在 Enron 数据集上整体表现较差，仅在 RL 指标上取得最优，在 AP，HL，OE 上排名皆为第六，在 CV 指标上排名第二，与最优算法相差较小，从数据集本身来看，Enron 属于离散型数值数据，因此，本文模型在此数据集上表现欠佳的原因可能是使用 Fisher score 去除边缘样本时去掉了具有重要信息的样本；在 Flags 数据集上，在 AP，HL，OE 和 CV 这 4 个指标上均取得最优，在 RL 指标上排名第四，与最优值差距不足 1%，这可能是选择了部分无关特征导致指标值降低。综合上述实验结果分析，MFSNE 算法在 AP，

HL,RL 和 OE 这 4 个指标上整体性能较好,而在 CV 指标上只有少部分未取得最优值。从指标本身以及 MFSNE 算法来看,在这部分数据集中,MFSNE 算法忽略了与已选特征有强相关性但重要度低的特征,导致分类效果不理想。总体来看,虽然在个别数据集上部分指标未得到显著提升,但在大部分数据集上,MFSNE 算法能够带来良好的分类效果。另外,尤其要分析比较 MLFS,MFSCS 和 MLNIE 这 3 种算法。在 AP 指标上,MLFS 算法在 Image 数据集上高于另外两种算法,MFSCS 算法在 Cal500,Computer,Emotion,Enron,Flags 和 Yeast 这 6 个数据集上高于另外两种算法,MLNIE 算法只在 Birds,Flags 和 Scene 这 3 个数据集上高于另外两种算法。整体来看,MFSCS 算法和 MLNIE 算法在大部分数据集上较 MLFS 算法均有提升。在 HL 指标上,MLFS 算法在 Birds,Emotion 和 Image 这 3 个数据集上低于 MLNIE 算法和 MFSCS 算法,MFSCS 算法在 Computer 和 Yeast 这两个数据集上低于另外两种算法,MLNIE 算法在 Cal500,Enron,Flags 和 Scene 这 4 个数据集上低于另外两种算法。整体来看,这 3 种算法在 Emotion,Image,Scene,Image 和 Yeast 这 5 个数据集上的差值在 1.15%～3.66%之间,在其余 5 个数据集上的差值均在 0.08%～0.3%之间,整体差距较小。在 RL 指标上,MLFS 算法在 Birds,Cal500,Image 和 Yeast 数据集上均比另外两种算法低,MFSCS 算法在 Computer,Enron 和 Flags 这 3 个数据集上低于另外两种算法,MLNIE 算法在 Emotion 和 Scene 这两个数据集上均低于另外两种算法。整体看,3 种算法在 9 个数据集上的差值在 0.1%～3.78%之间,MLFS 算法和 MFSCS 算法能够在大部分数据集上取得最优值,而 MLNIE 算法提升效果不明显。在 OE 指标上,MLFS 算法在 Birds,Cal500 和 Image 数据集上低于另外两种算法,MFSCS 算法在 Birds,Cal500,Computer,Emotion,Flags 和 Yeast 数据集上低于另外两种算法,MLNIE 算法在 Enron 和 Scene 这两个数据集上高于另外两种算法,其中,在 Enron 数据集上,MLNIE 与 MFSCS 算法相差 7.25%,提升较为明显。整体上,MFSCS 和 MLNIE 这两种算法相较于 MLFS 算法在 8 个数据集上均有提升;在 CV 指标上,MLFS 算法只在 Image 数据集上比另外两种算法低,MFSCS 算法在 Cal500,Computer,Enron,Flags 和 Yeast 数据集上低于另外两种算法,其中在 Cal500,Computer 和 Enron 数据集上,MFSCS 与 MLFS 算法的差值在 0.5480～0.7668 之间,提升效果明显,MLNIE 算法在 Birds,Emotion,Flags 和 Scene 这 4 个数据集上均低于另外两种算法。MFSCS 和 MLNIE 这两种算法在大部分数据集上均有提升。总体来看,这 3 种算法均不如 MFSNE 算法,但 MFSCS 算法和 MLNIE 算法在 5 个指标的大部分数据集上均有提升,其中 MFSCS 算法在大多数的数据集上高于其他算法。由此可见,MFSCS 算法对多标记数据进行预处理之后,使用 MLNIE 算法对评估候选特征子集获取最优特征子集,这样可以有效提升分类效果。

为了更直观地展示各种算法的性能,本文绘制了相关图形,但由于篇幅限制,只选择了 6 个代表性数据集(Birds,Cal500,Emotion,Image,Scene 和 Yeast)上的 AP 指标情况,其余 4 个指标的变化曲线图可以单独提供。图 2 展示了 12 种算法在 6 个数据集上的 AP 指标变化曲线对比结果图。横轴表示 FN,纵轴表示各算法在 AP 指标上的结果值。图 2 中,在 Cal500 数据集和 Image 数据集上,由于 MFSNE 算法使用搜索策略最终得到的特征数为随机值,所以 $FN<50$。

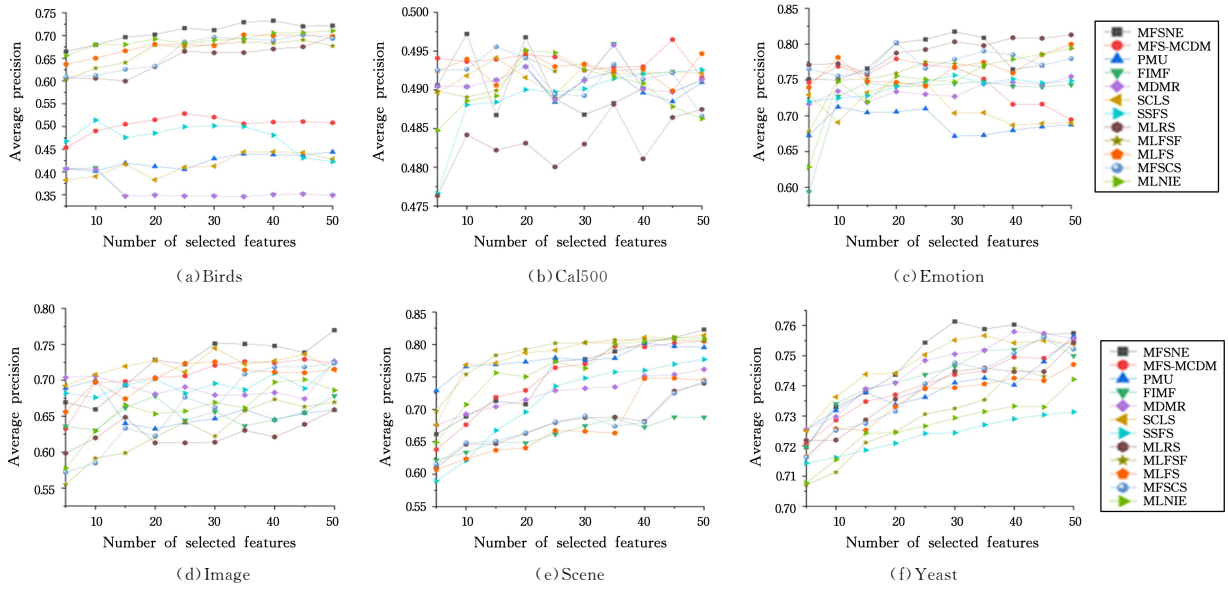

图 2 12 种算法在 6 个多标记数据集上的 AP(↑)指标比较

Fig. 2 Comparison of twelve algorithms on six multilabel datasets in terms of AP(↑)

从图 2 可知,在 Birds,Cal500,Emotion,Image,Scene 和 Yeast 这 6 个数据集上,MFSNE 算法均能取得最优值。其中,在 Birds 数据集上,MFSNE 算法整体上优于其他 11 种算法;在 Cal500 数据集上,当 $5<FN<10$ 时,AP 处于一般水平。但随着 FN 增加,在 $FN=10$ 和 $FN=20$ 时,MFSNE 的 AP 值高于其他 11 种算法,其中在 $FN=10$ 时取得最优值。其余情况下,MFRS 算法的 AP 值基本低于其他算法。从图中还可以看到,MFSNE 算法在 $FN<25$ 时相比其他算法波动较大,特别是当 $FN=15$ 时,原因可能是该序列选择的特征和已选特征构成冗余,导致最终标记预测错误。但 MFSNE

算法在 FN 较小时，其 AP 已达到全局最优。在 Emotion 数据集上，在 $FN\leqslant 20$ 时，MFSNE 算法基本高于其他大部分算法。随着 FN 的增加，在 $20<FN\leqslant 35$ 时，MFSNE 算法的 AP 值高于其他 11 种算法并在 $FN=30$ 时取得最优值，在 $FN>35$ 后低于 MLRS，MFS-MCDM，MLNIE 和 MFSCS 这 4 种算法；在 Image 数据集上，在 $FN<30$ 时，MFSNE 算法 AP 值基本处于一般水平，这可能是该部分特征序列存在冗余信息，导致指标值未增加，但随着 FN 的增加，在 $FN\geqslant 30$ 时，MFSNE 算法高于其他 11 种算法并在 $FN=50$ 时取得最优值；在 Scene 数据集上，在 $FN\leqslant 45$ 时，随着 FN 的增加，MFSNE 算法的 AP 值由最初居于中间阶段逐渐增大，在 $FN=45$ 处与其他算法持平，在 $FN>45$ 后均高于其他 11 种算法并取得最优值，这可能是因为在特征子集排序靠前的特征虽然重要度高，但是对标记相关性有重要作用的特征未被选择，从而出现最初 AP 值不高的结果。因此，结合 FN 和 AP 这两种指标的评价结果来看，在大多数情况下，相对于其他 11 种算法，MFSNE 算法在 AP 上具有良好的分类效果。

4.3　统计分析

为了分析所有算法在每个评价指标上的统计学的性能，采用 Friedman 测试和 Nemenyi 测试。其中 Friedman 统计量[39]表示为：

$$\chi_F^2=\frac{12T}{s(s+1)}\left(\sum_{i=1}^{s}R_i^2-\frac{s(s+1)^2}{4}\right) \tag{27}$$

$$F_F=\frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2} \tag{28}$$

其中，T 和 s 分别为数据集和算法的数量；$R_i(i=1,2,\cdots,s)$ 表示第 i 个算法在所有数据集上的平均排序。这里的临界值域(Critical Difference，CD) 的计算公式为：

$$CD_\alpha=q_\alpha\sqrt{\frac{s(s+1)}{6T}} \tag{29}$$

其中，q_α 表示测试的临界列表值，α 为显著性级别。参照文献[40-41]，采用 CD 图可视化显示所有对比算法之间的差异性。如果两种算法的平均排名差在一个误差之内，则使用连线将它们连接起来，否则在统计学上认为它们之间具有显著差异，其中不同颜色的连线是为了区分不同的两种算法之间存在显著差异。

根据表 2 的实验结果，12 种算法在 5 种指标的平均排名及其 Friedman 检验的 χ_F^2 和 F_F 如表 3 所列，相应的 CD 图如图 3 所示，横轴显示了所有对比算法的平均等级，其中左侧为最好。从图 3 可以看出，MFSNE 算法在 5 种评价指标上均明显优于其他算法。通过计算，当显著性水平 $\alpha=0.05$，$F(11,88)=2.045$，拒绝零假设。对于 Nemenyi 检验，当 $\alpha=0.05$ 时，$q_\alpha=3.2680$，$CD=5.5545$，其中 $s=9$ 且 $T=12$。MFSNE 算法在 5 种评价指标上均明显优于其他 11 种对比算法。

表 3　12 种算法的 5 种评价指标的统计结果

Table3　Statistical results of five metrics for twelve algorithms

指标	*AP*	*HL*	*RL*	*OE*	*CV*
χ_F^2	27.4786	23.8376	24.7991	16.7051	28.4530
F_F	3.0736	2.5372	2.6737	1.6239	3.2266

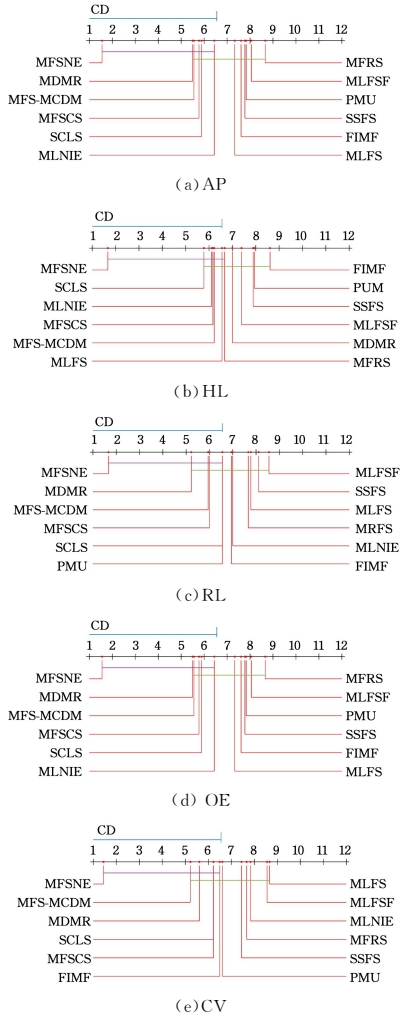

图 3　12 种算法在 5 种指标上的 Nemenyi 检验结果

Fig. 3　Nemenyi test results of twelve algorithms in terms of five metrics

结束语　本文提出了一种基于中心偏移的 Fisher score 与邻域直觉模糊熵的多标记特征选择方法。首先，为了去除边缘样本的影响，本文根据标记将多标记论域划分成了多个样本集，将计算得到的样本集的特征均值作为标记下样本的原始中心点，以最远样本的距离乘以距离系数，去除边缘样本，集从而定义了新的有效样本集，计算中心偏移处理后的标记下每个特征的得分以及标记集的特征得分，由此得到了基于中心偏移的多标记 Fisher score 模型，对多标记数据进行了预处理；然后，将分类间隔作为自适应模糊邻域半径参数，定义了模糊邻域相似关系和模糊邻域粒，并构造了多标记模糊邻域粗糙集的上、下近似集；在此基础上提出了多标记邻域粗糙直觉隶属度函数和非隶属度函数，由此定义了多标记邻域直觉模糊熵；最后给出了内部和外部重要度的特征评估方式，由此设计了一种基于中心偏移的 Fisher score 和邻域直觉模糊熵的多标记特征选择算法。在 MLKNN 分类器下的 9 个多标记数据集上的实验结果表明，相较于 11 种先进的多标记特征选择算法，MFSNE 算法选择的最优子集具有良好的分类性能。但是，本文仍存在一些问题：在第一阶段，Fisher score 算法考虑了边缘样本，对于离散型和数值型的混合数据，未充分考虑两种不同类型数据之间的差异，导致损失了

某些重要特征;在第二阶段,基于模糊相似关系的多标记邻域直觉模糊熵计算模糊邻域粒的复杂度较高。因此,使用 Fisher score 算法在去除边缘样本时有效区分离散型和数值型数据的边缘样本和类别中心,以及计算模糊邻域粒时选择高效的搜索策略来有效降低计算成本,都是下一步考虑的研究工作。

参 考 文 献

[1] SUN L, HUANG M M, XU J C. Weak label feature selection method based on neighborhood rough sets and relief[J]. Computer Science, 2022, 49(4): 152-160.

[2] LIU Y, CHENG L, SUN L. Feature selection method based on k-s test and neighborhood rough set[J]. Journal of Henan Normal University(Natural Science Edition), 2019, 47(2): 21-28.

[3] SUN L, XU F, LI S, et al. Multilabel feature selection algorithm using ReliefF and mRMR[J]. Journal of Henan Normal University(Natural Science Edition), 2023, 51(6): 21-29.

[4] CAO D T, SHU W H, QIAN J. Feature selection algorithm based on rough set and density peak clustering[J]. Computer Science, 2023, 50(10): 37-47.

[5] WANG Z K, SHEN D S, WANG C X. Fisher Score Fast Multi-Label Feature Selection Algorithm Based on Text Classification [J]. Computer Engineering, 2022, 48(2): 113-124.

[6] SUN L, WANG T X, DING W P, et al. Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification[J]. Information Sciences, 2021, 578: 887-912.

[7] HUANG S L, LIU Z, JIN W, et al. A fisher score-based multi-instance learning method assisted by mixture of factor analysis [J]. Neurocomputing, 2022, 507: 358-368.

[8] XU J C, YANG J, MA Y Y, et al. Feature selection method for color image steganalysis based on fuzzy neighborhood conditional entropy[J]. Applied Intelligence, 2022, 52(8): 9388-9405.

[9] WU D, GUO S Z. An improved Fisher Score feature selection method and its application [J]. Journal of Liaoning Technical University(Natural Science), 2019, 38(5): 472-479.

[10] ZHU J X, ZHU Z, AU S. Accelerating computations in two-slabele bayesian system identification with fisher information matrixand eigenvalue sensitivity[J]. Mechanical Systems and Signal Processing, 2023, 186: 109843.

[11] XU S X, MUSELET D, TREMEAU A. Sparse coding and normalization for deep fisher score representation[J]. Computer Vision and Image Understanding, 2022, 220: 103436.

[12] WANG Z K, SHEN D S, WANG C X. Fisher score fast multilabel feature selection algorithm based on text classification[J]. Computer Engineering, 2022, 48(2): 113-124.

[13] SUN L, WANG L Y, DING W P, et al. Feature selection using fuzzy neighborhood entropy-based uncertainty measuresfor fuzzy neighborhood multigranulation rough sets[J]. IEEE Transactions on Fuzzy Systems, 2021, 29(1): 19-33.

[14] LIU J H, LIN Y J, DU J X, et al. ASFS: A novel streaming feature selection for multilabel data based on neighborhood rough set[J]. Applied Intelligence, 2023, 53(2): 1707-1724.

[15] WU Y L, LIU J H, YU X H, et al. Neighborhood rough set based multilabel feature selection with label correlation[J]. Concurrency and Computation: Practice and Experience, 2022, 34(22): e7162.

[16] CAO J F, TIAN X D, JIA Y M, et al. Segmentation method of ancient murals based on improved PSPNet[J]. Journal of Henan Normal University (Natural Science Edition), 2022, 50(4): 65-75.

[17] XU J C, SHEN K L, SUN L. Multilabel feature selection based on fuzzy neighborhood rough sets[J]. Complex & Intelligent Systems, 2022, 8(3): 2105-2129.

[18] SUN L, CHEN Y S, DING W P, et al. AMFSA: Adaptive fuzzy neighborhood-based multilabel feature selection with ant colony optimization[J]. Applied Soft Computing, 2023, 138: 110211.

[19] YIN T Y, CHEN H M, YUAN Z, et al. Noise-resistant multilabel fuzzy neighborhood rough sets for feature subset selection [J]. Information Sciences, 2023, 621: 200-226.

[20] XUE Z A, PANG W L, YAO S Q, et al. Theproposed theory based intuitionistic fuzzy three-way decisions model[J]. Journal of Henan Normal University(Natural Science Edition), 2020, 48(5): 31-36.

[21] ZHANG X Y, HOU J L, LI J L. Multigranulation rough set methods and applications based on neighborhood dominance relation in intuitionistic fuzzy datasets[J]. International Journal of Fuzzy Systems, 2022, 24(8): 3602-3625.

[22] XIN X W, SHI C L, SUN J B, et al. A novel attribute reduction method based on intuitionistic fuzzy three-way cognitive clustering[J]. Applied Intelligence, 2023, 53(2): 1744-1758.

[23] MAO P D, XU D L. Efficient single-image super-resolution: deeply-supervised symmetric distillation network[J]. Journal of Henan Normal University (Natural Science Edition), 2023, 51(6): 57-65.

[24] ZHENG T T, ZHU L Y. Uncertainty measures of neighborhood system-based rough sets[J]. Knowledge-Based Systems, 2015, 86: 57-65.

[25] YAO S, XU F, ZHAO P, et al. Intuitionistic Fuzzy Entropy Feature Selection Algorithm Based on AdaptiveNeighborhood Space Rough Set Model[J]. Journal of Computer Research and Development, 2018, 55(4): 802-814.

[26] ATANASSOV K T. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20(1): 87-96.

[27] JAIN P, SOM T. Multigranular rough set model based on robust intuitionistic fuzzy covering with application tofeature selection [J]. International Journal of Approximate Reasoning, 2023, 156: 16-37.

[28] LI B X, WAN R Z, ZHU Y J, et al. Multi-strategy comprehensive article swarm optimization algorithm based on population partition [J]. Journal of Henan Normal University(Natural Science Edition), 2022, 50(3): 85-94.

[29] WAN F, WANG M S, HAN Y P, et al. Research and application of reservoir flood risk early warning and ecological dispatching [J]. Journal of Henan Normal University(Natural Science Edition), 2022, 50(3): 20-28.

[30] AMIN H, MOHAMMAD B D, HOSSEIN N. MFS-MCDM: Multilabel feature selection using multi-criteria decision making [J]. Knowledge-Based Systems, 2020, 206: 106365.

[31] GAO W F, LI Y H, HU L. Multilabel feature selection with constrained latent structure shared term[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34: 1253-1262.

[32] LEE J, KIM D. SCLS: Multilabel feature selection based on scalable criterion for large label set[J]. Pattern Recognition, 2017,

66：342-352.

[33] LIN Y J，HUQ H，LIU J H，et al. Multilabel feature selection based on max-dependency and min-redundancy[J]. Neurocomputing，2015，168：92-103.

[34] LEE J，KIM D. Feature selection for multilabel classification using multivariate mutual information[J]. Pattern Recognition Letters，2013，34(3)：349-357.

[35] LEE J，KIM D. Fast multilabel feature selection based on information-theoretic feature ranking[J]. Pattern Recognition，2015，48(9)：2761-2771.

[36] SUN L，CHEN Y S，XU J C. Multilabel feature selection algorithm based on improved ReliefF[J]. Journal of Shandong University(Natural Science)，2022，57(4)：1-11.

[37] SUN L，MA T J，XUE Z A. Multilabel feature selection algorithm based on fisher score and fuzzy neighborhood entropy[J]. Journal of Computer Applications，2023，43(12)：3779-3789.

[38] ZHANG Y，MA Y. Non-negative multilabel feature selection with dynamic graph constraints[J]. Knowledge-Based Systems，2022，238：107924.

[39] SUN L，SI S S，DING W P，et al. Multiobjective sparrow search feature selection with sparrow ranking and preference information and its applications for high-dimensional data[J]. Applied Soft Computing，2023，147：110837.

[40] SUN L，YIN T Y，DING W P，et al. Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy[J]. IEEE Transactions on Fuzzy Systems，2022，30(5)：1197-1211.

[41] HOU T B，WANG A Y. 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.

SUN Lin，born in 1979，Ph.D，professor，doctoral supervisor，is a member of CCF (No. 74144M). His main research interests include machine learning，data mining and bioinformatics.

（责任编辑：何杨）

CCF“中国计算机历史记忆”认定申报征集通知

中国计算机事业自 1956 年创建以来，已经走过了 60 多年的发展历程。国产的计算机装备、基础软件和应用系统为国家重大项目和经济发展做出了不可替代的贡献。这些散落于全国各地的计算机历史物件见证了我国计算机事业的蓬勃发展。为了推动这些珍贵历史物件的保护，CCF 于 2017 年启动了“中国计算机历史记忆”认定计划（简称“CCF 历史记忆认定”）。这项计划通过认证的方式，识别和挖掘国内的珍贵计算机物件，鼓励社会各界为中国计算机事业留下宝贵的历史记忆。

现向全国各单位及个人征集见证中国计算机发展历史的珍贵物件线索。CCF 将组织认定工作，向具有重要历史价值的物件颁发证书，并通过网站向公众分享被认定物件的信息。具体安排如下：

一、征集范围

CCF 历史记忆认定物件（下称“物件”）应在中国研制或生产，包括但不限于计算机相关的原型系统、部件、装置、书籍、软件、手稿等。物件应在推动中国计算机发展方面具有重要历史意义，其研制或生产时间应距离申请认定日期至少 20 年。

二、申报材料

CCF 历史记忆认定实行申报制，由物件所有者（单位或个人）提交申报材料，CCF 按年度进行认定工作。申报材料应包括以下内容：

1. 物件概述，1 000 字左右，描述物件起源、相关重大事件与人物、历史意义等；

2. 物件保存情况，可附若干当前照片；

3. 如有其他老照片、影音资料、采访等材料，可作为附件提供；

4. 保存单位或个人简介。

三、认定与发布

CCF 设立历史记忆认定机构，该机构由历史记忆认定委员会（以下简称“认定委员会”）与认定委员会秘书处组成。认定过程分为 3 个阶段：

1. 申报阶段。申报单位或个人可随时提交申报材料。

2. 初评阶段。认定委员会设立、监督并指导认定工作组开展初评。

3. 最终认定阶段。认定委员会每年召开一次认定工作会议，审议认定工作组初评结果，并评出最终认定结果。

认定结果由 CCF 理事长批准后生效。CCF 将制作带有“CCF 中国计算机历史记忆”LOGO 的牌匾与认定证书，授予申报单位或个人。CCF 历史记忆认定颁发仪式将于 CCF 颁奖典礼上举行。

四、申报方式与时间节点

申报截止日期为 2024 年 11 月 15 日。

请在此之前填写下附申报表并发送至：history@ccf.org.cn，联系人：吴树民

据 CCF 微信公众号