本文引用格式:丰昌武,孙林. 融合 K-means 聚类和标记相关性的多标记 Relief 特征选择[J]. 聊城大学学报(自然科学版),2025,38(1):122-134.

Citation format:FENG Changwu,SUN Lin. Multilabel relief-based feature selection via fusing K-Means clustering and label correlation[J]. Journal of Liaocheng University(Natural Science Edition),2025,38(1):122-134.

文章编号 1672-6634(2025)01-0122-13 **DOI** 10.19728/j.issn1672-6634.2024080010

融合 K-means 聚类和标记相关性的多标记 Relief 特征选择

丰昌武,孙林

(天津科技大学 人工智能学院,天津 300457)

摘要 现有 Relief 算法在利用标记相关性方面存在不足,忽视了局部标记相关性所提供的宝贵信息。针对这一问题,提出了一种融合 K-means 聚类与标记相关性的多标记 Relief 特征选择方法。首先,为充分考虑样本标记相关性,采用 K-means 聚类算法对样本进行聚类,将其划分到不同的簇中,从而构建样本的局部标记空间。其次,定义了所有样本在特征上的欧式距离,以此衡量样本的全局标记相关性。同时,改进了传统的余弦相似度,使用 L1 范数的平方根进行优化,并在局部标记空间中应用改进的余弦相似度,以有效获取样本的局部标记相关性。最后,在 Relief 算法的基础上,融合了样本的全局标记相关性与局部标记相关性,以此作为衡量样本相似度的依据,进而判别最近邻同类样本与最近邻异类样本,最终获得特征权重。为评估所提算法的性能,在 10 个多标记数据集上进行了对比测试,实验结果证明,与其他多标记特征选择算法相比,本算法具有显著优势。

关键词 多标记学习;特征选择;K-means 聚类;标记相关性;Relief 算法

中图分类号 TP 18 **文献标志码** A **开放科学(资源服务)标识码(OSID)**

Multilabel Relief-Based Feature Selection via Fusing K-Means Clustering and Label Correlation

FENG Changwu, SUN Lin

(College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract Existing Relief algorithms are deficient in exploiting label correlation and often ignore the valuable information provided by local label correlation. To address this problem, this work proposed a multilabel Relief feature selection method that fuses K-means clustering with label correlation. First, to fully consider the relevance of sample labels, the K-means clustering algorithm was used to cluster the samples and divide them into different clusters, thereby constructing the local label space of the samples. Second, the Euclidean distance of all samples is defined to measure the global labeling correlation of the samples.

收稿日期:2024-08-20

基金项目:国家自然科学基金项目(61772176)资助

通信作者:孙林,男,汉族,博士,教授,河南省科技创新杰出青年(河南省杰青),研究方向:数据挖掘、机器学习、深度学习,E-mail:sunlin@tust.edu.cn。

At the same time, the traditional cosine similarity was improved by using the square root of the L1 norm for optimization, and this improved cosine similarity was applied in the local label space to efficiently obtain the local label correlation of the samples. Finally, on the basis of the Relief algorithm, the global label correlation and local label correlation of the samples were fused, as the basis for measuring the similarity of the samples. Then, the nearest-neighbor similar samples and nearest-neighbor dissimilar samples were discriminated to finally obtain the feature weights. To verify the effectiveness of the proposed algorithm, comparison tests were conducted on 10 publicly available multilabel datasets, and the experimental results proved that the proposed algorithm shown significant advantages over other multilabel feature selection algorithms.

Key words　multilabel learning; feature selection; K-means clustering; label correlation; Relief algorithm

0　引言

在当前复杂的现实场景中，多标记学习已逐渐成为文本挖掘、图像标注、生物医学等领域的研究热点[1, 2]。然而，大数据环境下的多标记学习一直面临高维诅咒的挑战，这些影响主要包括学习复杂性的提高和降低分类性能[3]。因此，特征选择工作变得愈发重要[4]，通过筛选相关特征并剔除冗余或不相关的特征，有效减少特征数量，提升分类精度[5]。针对高维多标记数据展开特征选择的研究是当前多标记学习中的重要环节。近年来，许多学者将聚类算法应用于特征选择中，以用来构建局部标记空间[6]。其中，K-means 聚类因其思想简单、聚类效果佳和易于实现而备受关注[7]。例如，孙林等[8]提出了基于邻域互信息与 K-means 特征聚类的特征选择方法。实验结果显示，K-means 聚类不仅能显著提升特征选择后的分类性能，还能获得更优的聚类效果。受此启发，本文将 K-means 聚类应用于多标记数据的特征选择中。

到目前为止，学者们已经提出了多种多标记学习算法。其中，Li 等[9]设计了基于 $L_{2,1}$ 正则化的自表达系数矩阵来描述一种多标记分类方法，旨在去除噪声和冗余特征。然而，该方法忽略了标记之间的相关性。众多研究已证明，利用标记之间的相关性能够显著提高多标记分类的性能[10]。结合不同情况下的标记相关性，已成为当前多标记特征选择研究的重要措施[11]。全局标记相关性存在于所有训练样本中[12]。基于此，Ji 等[13]提出了一种共享结构的多标记分类框架。但该算法对计算性能要求较高，可能影响算法时效性。由于局部标记相关性能够为多标记学习提供更多的有用信息，容斌元等[14]提出了一种融合局部标记相关性的标签分布学习模型，利用 K-means 聚类获得不同类别描述的局部信息。该方法结合了神经网络结构，因此该算法的时间复杂度较高。目前已有的多标记特征选择算法中，鲜有同时考虑标记的全局相关性与局部相关性的多标记分类算法。朱赛赛等[15]提出了一种基于全局和局部标记相关性的多标记分类算法。但其使用传统的余弦相似度度量局部标记相似性，可能在样本数值差异较大时效果不理想。为了有效利用样本的标记相关性，本文提出了一种将样本全局标记相关性与局部标记相关性相结合的多标记特征选择算法，减少冗余信息，有效提高分类器的预测性能。

Relief 算法是一种经典的特征权重评估方法[16]。其主要思想是通过计算特征间的距离来评估对目标变量的重要性[17]。然而，传统的 Relief 算法仅适用于二分类问题。这限制了其应用范围，为了扩展 Relief 算法的多标记学习的应用场景，孙林等[18]在 Relief 算法上提出了基于样本全局相似度和缺失标记的特征选择算法，该方法利用不同样本标记的重叠度来度量样本标记的相关性，从而在特征选择过程中考虑了全局标记相关性。但是，该方法仍然忽略了样本的局部特征所蕴含的丰富信息。Relief 算法使用余弦相似度来度量不同样本的相似性，有效衡量两个向量在方向上的相似程度[19]。然而，传统的余弦相似度仅关注向量方向，忽略了向量大小，这可能影响某些情况下相似性评估的准确性。为了解决这个问题，Fan 等[20]构造了一种基于特征冗余的控制函数来约束特征之间的联系，其中使用余弦相似度来计算两个特征之间的相似度，以保留原始数据集中的特征关系。然而，传统的余弦相似度仅使用 L2 范数衡量两个非零向量之间的相似度，且没有考虑数值差异较大的特征[21]。针对上述问题，Sohangir 等[22]提出了一种基于 L1 范数的 Hellinger 距离来改进余弦相似度，并证明了在高维数据处理中，L1 范数比 L2 范数具有更好的表现。受此启发，本文

在局部标记相关性的度量中引入了基于 L1 范数的余弦相似度。通过结合 L1 范数的优势,本文基于全局标记相关性和局部标记相关性,设计新的特征权值迭代公式,改进 Relief 算法,使其能够更加准确地评估样本间的相似性,特别是在处理数值差异较大的特征时。

为了解决传统 Relief 算法忽略局部标记相关性的问题,本文提出了一种基于 K-means 聚类和标记相关性的多标记 Relief 特征选择方法。该方法的核心在于同时考虑样本的全局标记相关性和局部标记相关性,以全面挖掘和利用标记信息,从而提高分类器的性能。首先,在全局样本上度量全局标记相关性,此外,通过 K-means 聚类方法对样本集进行了分群处理,进而形成了多个簇类,并计算每个簇的质心。随后,使用样本所在簇的质心与随机样本进行相似度度量,以此得到样本的局部标记相关性,这样不仅考虑了样本间的局部关系,而且还通过质心的引入,使得局部标记相关性的度量更加准确和全面。在局部标记相关性的度量中,本文采用了基于 L1 范数的余弦相似度。与传统的余弦相似度相比,L1 范数的平方根能够更好地处理数值差异较大的特征,从而提高局部标记相关性度量的准确性。最后,将上述方法运用于 Relief 算法中,利用样本的全局标记相关性与局部标记相关性作为衡量样本相似度的依据,进而判别最近邻同类样本与最近邻异类样本,最终获得特征权重。

由此,可以总结本文的主要贡献包括 3 个方面:(1) 基于 K-means 聚类的局部标记空间。引入 K-means 聚类算法,将样本分为不同的簇,计算每个簇的质心;利用质心与随机样本的相似性度量,得到样本的局部标记空间,使得 K-means 聚类能够利用样本局部标记相关性提供的更全面的标记信息,从而提高分类器的性能。(2) 基于改进余弦相似度的标记相关性。本文使用余弦相似度对所有训练样本进行相似度度量,得到样本的全局标记相关性。同时,针对高维数据样本的处理,采用了基于 L1 范数的余弦相似度来度量样本的局部标记相关性,使得该模型在处理数值差异较大的特征时更加有效。(3) 多标记 Relief 特征选择模型。基于传统的 Relief 算法,本文结合了全局标记相关性和局部标记相关性,通过判别最近邻同类样本与最近邻异类样本,构建了新的特征权值迭代公式。最终,设计了基于全局标记相关性与局部标记相关性的多标记 Relief 特征选择算法,实现了对多标记数据的有效特征选择。

1 Relief 模型

在 Relief 模型[16]中,当同类样本在特征上的距离越小,异类样本在特征上的距离越大,则表明该特征对样本的区分能力越强,可以为分类器提供更多的学习信息,则就可以赋予该特征较大的权重;通过数次迭代,每次迭代随机选取样本,并确定该样本的最近邻同类和最近邻异类,代入特征权值迭代更新公式,迭代结束后,得到最终的特征权值。在 Relief 模型中,任意两个样本 x 和 $y\in U$ 在特征 f_i 上的距离[16]表示为

$$\mathrm{diff}(x,y)=\frac{x(f_i)-y(f_i)}{\max(f_i)-\min(f_i)}, \tag{1}$$

式中 $x(f_i)$表示样本 x 在特征 f_i 上的值,$y(f_i)$表示样本 y 在特征 f_i 上的值,$\max(f_i)$表示在特征 f_i 上的最大取值,$\min(f_i)$表示在特征 f_i 上的最小取值。

Relief 模型的特征权值迭代公式[16]表示为

$$w_{f_i}=w_{f_i}-(\mathrm{diff}(x,NH,f_i))^2+(\mathrm{diff}(x,NM,f_i))^2, \tag{2}$$

式中 NH 和 NM 分别代表样本 x 的最近邻同类样本和最近邻异类样本; $\mathrm{diff}(x, NH, f_i)$表示样本 x 与 NH 在特征 f_i 上的距离,$\mathrm{diff}(x, NM, f_i)$表示样本 x 与 NM 在特征 f_i 上的距离,w_{f_i} 表示特征 f_i 的特征权重。

2 本文提出的方法

2.1 基于 K-means 聚类的局部标记空间

在多标记学习范畴内,标记间的相关性构成了对模型性能提升具有显著价值的信息资源,这些信息对增强分类效果至关重要。由此,本文致力于在特征空间内计算样本的局部标记相关性。考虑到 K-means 聚类算法[23]作为一种有效的迭代数据划分方法,采用此算法在特征空间中将样本划分为 K 个簇。然后,运用改

进的余弦相似度来度量某一随机样本所在簇与其他簇之间的相似程度，从而实现对局部标记相关性的精确度量。

定义 1　给定一个多标记决策系统 $MDS = <U, C, D>$，其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合，$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合，$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合，则任意两个样本 x 和 $y \in U$ 在特征 f_i 上的距离表示为

$$\mathrm{diff}(x,y) = \frac{|x(f_i) - y(f_i)|}{\max(f_i) - \min(f_i)}, \tag{3}$$

式中 $x(f_i)$ 表示 x 在特征 f_i 上的值，$y(f_i)$ 表示 y 在特征 f_i 上的值，$\max(f_i)$ 与 $\min(f_i)$ 分别表示所有样本在特征 f_i 上的最大值与最小值。

定义 2　给定一个多标记决策系统 $MDS = <U, C, D>$，其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合，$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合，$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合。使用 K-means 聚类初始质心 $c \in U$ 为随机样本，对于任意样本 $x \in U$，基于欧式距离计算所有样本 x 到质心 c 的距离表示为

$$\mathrm{d}(x,c) = \sqrt{\sum_{i=1}^{n} (x(f_i) - c(f_i))^2}, \tag{4}$$

式中 $\sqrt{\sum_{i=1}^{n} (x(f_i) - c(f_i))^2}$ 表示样本 x 与质心 c 在特征 f 上的欧氏距离[16]，$x(f_i)$ 和 $c(f_i)$ 分别表示样本 x 与质心 c 在特征 f_i 上的取值。

定义 3　给定一个多标记决策系统 $MDS = <U, C, D>$，其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合，$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合，$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合。使用 $R = \{r_1, r_2, \cdots, r_k\}$ 表示簇集合，簇内样本平均特征值的计算公式表示为

$$r(f_i) = \frac{1}{n} \sum_{i=1}^{n} x(f_i), \tag{5}$$

式中 $r(f_i)$ 代表簇 r 在特征 f_i 上的值，$x(f_i)$ 表示簇 r 中的单个样本 x 在特征 f_i 上的值。具体来说，式(5)使用簇 r 中所有的样本在特征 f_i 的平均值代表簇 r 在特征 f_i 上的值。

由定义 2 和定义 3 可知，样本通过 K-means 聚类并得到簇内的平均特征值，使用簇内的平均特征值代表所在簇的特征值，由此得到了样本的局部标记空间。利用 K-means 聚类选择 k 个初始簇中心，利用式(4)计算所有样本到每个质心的距离，将每个样本分配给离质心最近的簇，利用式(5)计算每个簇中样本的平均值，以获得 k 个新质心位置，重复迭代，直到簇分配不变或达到设置的最大迭代次数。

2.2　基于余弦相似度的标记相关性

Sohangir 等[22]指出，通过计算两个向量夹角的余弦值，可以有效衡量个体之间的相似性。于是，使用余弦相似度作为近似度量的获取方法，最小化潜在特征间的相关性，进而表示样本在特征上的相似度。

定义 4　给定一个多标记决策系统 $MDS = <U, C, D>$，其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合，$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合，$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合。对于任意两个样本 x 和 $y \in U$，则 x 和 y 在特征 f_i 上余弦相似度[19]的计算公式为

$$\cos(x,y) = \frac{\sum_{i=1}^{n} (x(f_i) \times y(f_i))}{\sqrt{\sum_{i=1}^{n} (x(f_i))^2} \times \sqrt{\sum_{i=1}^{n} (y(f_i))^2}}。 \tag{6}$$

由式(6)可知，$-1 < \cos(x, y) < 1$。由此，当 $\cos(x, y)$ 越接近 -1，说明 x 和 y 相似度越低；当 $\cos(x, y)$ 越接近 1，说明 x 和 y 的相似度越高。由于 x 和 y 都在全局样本空间上，因此，可以使用 $\cos(x, y)$ 来度量样本的全局标记相关性。

Aggarwal 等[24]指出：从理论角度来看，L2 范数通常不是高维数据挖掘应用的理想度量方式，其中传统的余弦相似度常使用 L2 范数来作为距离度量，高维样本对 K 的取值非常敏感，意味着在对高维样本的处理中 L1 范数比 L2 范数更佳。由此，本文使用 L1 范数的平方根改进了传统的余弦相似度来度量局部标记相关性。

定义5 给定一个多标记决策系统 $MDS = <U, C, D>$,对于任意两个样本 x 和 $y \in U$,则基于L1范数平方根的余弦相似度计算公式表示为

$$\mathrm{lcos}(x,y) = \frac{\sum_{i=1}^{n} \sqrt{|x(f_i)y(f_i)|}}{\sqrt{(\sum_{i=1}^{n}|x(f_i)|)}\sqrt{(\sum_{i=1}^{n}|y(f_i)|)}}, \tag{7}$$

式中 $x(f_i)$ 和 $y(f_i)$ 分别表示两个样本 x 与 y 在特征 f_i 上的取值。由式(7)可知,$\mathrm{lcos}(x, y)$ 可以用来度量样本的局部标记相关性,其值越大则相似性越高。

定义6 给定一个多标记决策系统 $MDS = <U, C, D>$,其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合,$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合,$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合,使用 $R = \{r_1, r_2, \cdots, r_k\}$ 表示簇集合。基于定义4和定义5,可以构建度量随机样本 x 与其余样本 y 在局部标记空间之间新的相似度,则其度量公式表示为

$$\mathrm{lcs}(x,y) = \frac{\sum_{i=1}^{n} \sqrt{|x(f_i)r(f_i)|}}{\sqrt{(\sum_{i=1}^{n}|x(f_i)|)}\sqrt{(\sum_{i=1}^{n}|r(f_i)|)}}, \tag{8}$$

式中 x 为随机样本,r 为其余样本 y 所在的簇,$r(f_i)$ 表示样本 y 所属簇在特征 f_i 上的值。由式(8)可知,$\mathrm{lcs}(x, y)$ 可以有效衡量两个样本的局部标记相关性,且其值越大相似性越高。

2.3 多标记 Relief 特征选择模型

定义7 给定一个多标记决策系统 $MDS = <U, C, D>$,对于任意两个样本 x 和 $y \in U$,则样本 x 和 y 的全局与局部标记相似度计算公式表示为

$$\mathrm{sim}(x,y) = \cos(x,y) + \mathrm{lcs}(x,y), \tag{9}$$

式中 $\cos(x, y)$ 可以度量样本的全局标记相关性,$\mathrm{lcs}(x, y)$ 可以度量样本的局部标记相关性。由式(9)可知,$\mathrm{sim}(x, y)$ 结合了样本的全局标记相关性与局部标记相关性,其值越大则表明两个样本的相似度越高。

定义8 给定一个多标记决策系统 $MDS = <U, C, D>$,对于任意两个样本 x 和 $y \in U$,则 x 和 y 的最近邻同类和异类样本的判别关系表示为

$$\mathrm{d}_r(x,y) = \begin{cases} NHC, & \mathrm{sim}(x,y) \geqslant \mathrm{avg}_x(\mathrm{sim}), \\ NMC, & \mathrm{sim}(x,y) < \mathrm{avg}_x(\mathrm{sim}), \end{cases} \tag{10}$$

式中 NHC 表示样本 x 与 y 为最近邻同类样本,NMC 表示样本 x 与 y 为最近邻异类样本,$\mathrm{avg}_x(\mathrm{sim})$ 表示计算样本 x 与剩余样本间的相似度均值。式(10)中 $\mathrm{d}_r(x, y)$ 能够判别任意两个样本 x 和 y 是最近邻同类样本或最近邻异类样本。

定义9 给定一个多标记决策系统 $MDS = <U, C, D>$,其中 $U = \{x_1, x_2, \cdots, x_m\}$ 表示样本集合,$C = \{f_1, f_2, \cdots, f_n\}$ 表示特征集合,$D = \{l_1, l_2, \cdots, l_s\}$ 表示标记集合。对于任意样本 $x_i \in U$ 和特征 $f_i \in C$,则 f_i 的特征权重迭代公式表示为

$$w_{f_i} = w_{f_i} - \sum_{j=1}^{k1}\mathrm{diff}(x_i, NHC_j, f_i) + \sum_{j=1}^{k1}\mathrm{diff}(x_i, NMC_j, f_i), \tag{11}$$

式中 NHC_j 和 NMC_j 分别代表样本 x_i 的最近邻同类样本和最近邻异类样本;$k1$ 表示样本 x_i 近邻样本数,$\mathrm{diff}(x_i, NHC_j, f_i)$ 表示样本 x_i 与 NHC_j 在特征 f_i 上的距离,$\mathrm{diff}(x_i, NMC_j, f_i)$ 表示样本 x_i 与 NMC_j 在特征 f_i 上的距离,w_{f_i} 表示特征 f_i 的特征权重,w_{f_i} 值越大则说明在特征 f_i 上就越能够有效区分异类样本,并且同类样本在特征 f_i 上的差异性越小。通过式(11)的迭代,可以得到最终的特征权重矩阵。

2.4 多标记特征选择算法描述

针对当前多标记特征选择算法在结合全局标记相关性与局部标记相关性方面的不足,尤其是未能充分利用样本的局部标记相关性这一问题,本文在传统 Relief 算法的基础上进行了创新,设计了一种新的多标记 Relief 特征选择算法,即结合全局标记相关性和局部标记相关性的 RGLC(Multilabel Relief-Based Feature Selection via Fusing K-Means Clustering and Label Correlation)算法。以下是该算法的详细步骤概述,具体

实现细节请参见算法 1 的完整描述。

算法 1. RGLC 算法

输入：给定多标记决策系统 $MDS = <U, C, D>$，设置迭代次数为 e、近邻样本数为 $k1$、聚类数量为 k

输出：排序后特征序列 S

1：通过式(6)计算所有样本之间的距离

2：利用 K-means 聚类将 MDS 中的所有样本划分为 k 个簇

3：随机选取样本 x

4：通过式(9)计算随机样本 x 与其余样本之间在局部标记空间中的相似度

5：赋值特征权重 $w_{f_i} = 0$，其中 $i = 0, 1, 2, \cdots, n$

6：**For** $i = 1 : e$

7：　根据式(10)计算 sim

8：　if $\mathrm{sim}(x, y) \geqslant \mathrm{avg}_x(\mathrm{sim})$ then x 与 $y \in NHC$

　　else　x 与 $y \in NMC$

9：　逆序排序 sim，将相似度最高的 $k1$ 个样本作为样本 x 的同类样本集，将相似度最低的 $k1$ 个样本作为样本 x 的异类样本集

10：　根据式(11)计算每个特征的权重 w_{f_i}

11：**End for**

12：将特征权重降序排序

13：输出排序后特征序列 S

在 RGLC 算法中，假定样本数量、特征数量和聚类数量分别为 m、n 和 k，步骤 1 计算所有样本间距离的计算复杂度为 $O(m^2n)$，步骤 2 中 K-means 聚类的计算复杂度为 $O(kmn)$，步骤 3、步骤 5 和步骤 13 复杂度为常数，步骤 4 的时间复杂度约为 $O(m^2n)$，步骤 6 到步骤 11 的时间复杂度约为 $O(mn)$，步骤 12 的计算复杂度为 $O(n\log n)$，总的最坏的计算复杂度为 $O(m^2n + kmn + m^2n + mn + n\log n) \approx O(m^2n)$。在本文使用的对比算法中，MFS-MCDM 算法[25]的计算复杂度近似为 $O(m^2n)$、PMFS 算法[26]的计算复杂度为 $O(m^2)$、PUM 算法[27]的计算复杂度为 $O(m^2)$、GLFS 算法[28]的计算复杂度约为 $O(m^3)$、WFSNR 算法[29]的计算复杂度为 $O(m^3)$、NRFSFN 算法[30]的计算复杂度为 $O(m^2n)$、ENFSFN 算法[30]的计算复杂度为 $O(m^2n)$、SRMFS[18]算法的计算复杂度为 $O(m^2)$、LSFRS 算法[31]和 SFSLH 算法[32]的计算复杂度均为 $O(m^3)$。在所有对比算法中最高计算复杂度为 $O(m^3)$，最低计算复杂度为 $O(m^2)$，本文算法的计算复杂度介于中间。

3　实验分析

3.1　实验准备

为了评估 RGLC 算法的分类效果，从 MULAN 数据库(https://mulan.sourceforge.net/datasets.html)中选择 10 个多标记数据集，数据集描述如表 1 所示。实验环境为 Windows 11 Intel(R) i5-500@3.00GHz 16GB RAM，以及 MATLAB R2022b。参照文献[6]和文献[18]的实验设置，RGLC 算法中迭代次数 $e = 200$ 且最近邻样本数量 $k1 = 5$；选用多标记 K 最近邻(Multilabel K-Nearest Neighbors，MLKNN)作为多标记分类器，其平滑系数为 1 且样本最近邻数量为 10。参照文献[33]，选择平均分类精度(Average Precision，AP)、排序损失(Ranking Loss，RL)、汉明损失(Hamming Loss，HL)和覆盖率(Coverage，CV)为评价指标；其中 AP 值越大(↑)则表明算法性能越好，其它 3 个指标值越小(↓)则表明算法性能越优。

3.2　实验结果比较

本节采用 AP、RL、HL 和 CV 这 4 种指标进行实验比较，为了保持实验结果客观一致，选取表 1 中的 10

个多标记数据集,选择的特征数量与文献[16]中 SRMFS 算法保持一致。首先,与目前 10 种先进的多标记特征选择算法进行比较,具体包括 MFS-MCDM 算法[25]、PMFS 算法[26]、PUM 算法[27]、GLFS 算法[28]、WFSNR 算法[29]、NRFSFN 算法[30]、ENFSFN 算法[30]、SRMFS[18]算法、LSFRS 算法[31]和 SFSLH 算法[32]。使用的对比算法涉及的有关参数设置如下:MFS-MCDM 算法的正则化参数为 0.1;PMFS 算法的正形参为 -10^3;GLFS 算法的参数$\{\alpha,\beta,\gamma,\lambda\}$设置为$\{1,0.1,100,-4\}$; NRFSFN 算法和 ENFSFN 算法的模糊邻域半径和阈值分别为$\{1,0.9\}$和$\{0.3,0.5\}$;SRMFS 算法的数迭代次数为 200 最近邻样本数量等于 5。具体的实验结果见表 2 至表 5 所示。另外,为了避免实验的随机性和偶然性引起的误差,RGLC 算法在 10 个多标记数据集均单独重复 20 次实验,分别取 AP、RL、HL 和 CV 这 4 种指标结果的平均值。需要说明的是:在下面所有实验结果中,表中粗体代表在所比较的算法中呈现的最优值,average 表示该算法在所有多标记数据集上实验结果的平均值。

表 1 10 个多标记数据集的信息

Tab. 1 Information of ten multilabel datasets

NO.	Datasets	Samples	Features	Labels	Lcard	Lden	Domain
1	Arts	5 000	462	26	1.636	0.063	Text
2	Computer	5 000	681	33	1.508	0.046	Text
3	Enron	1 702	1 001	53	3.378 4	0.064	Text
4	Health	5 000	612	32	1.633	0.052	Text
5	Business	5 000	438	30	1.588	0.053	Text
6	Reference	5 000	793	33	1.169	0.035	Text
7	Science	5 000	743	40	1.451	0.036	Text
8	Yeast	2 417	103	14	4.237	0.303	Biology
9	Medical	978	1 499	45	1.245	0.028	Text
10	Image	2 000	294	5	1.236	0.247	Image

表 2 11 种算法在 AP(↑)指标下 10 个多标记数据集上的实验结果

Tab. 2 Experimental results of AP(↑) index with the eleven algorithms on the ten multilabel datasets

Datasets	MFS-MCDM	PMFS	PUM	GLFS	WFSNR	NRFSFN	ENFSFN	SRMFS	LSFRS	SFSLH	RGLC
Arts	0.480 6	0.472 2	0.480 0	0.450 8	0.484 9	0.470 3	0.484 9	0.489 0	0.489 6	0.476 7	**0.508 2**
Computers	0.619 5	0.610 0	0.619 4	0.607 0	0.621 3	0.600 3	0.621 3	0.629 7	0.606 7	0.629 3	**0.633 5**
Enron	0.610 7	**0.657 8**	0.649 5	0.538 8	0.594 0	0.539 1	0.594 0	0.616 4	0.584 6	0.620 3	0.630 2
Health	0.621 3	0.617 2	0.633 3	0.627 5	0.649 1	0.658 2	0.649 1	0.672 9	0.667 2	0.650 2	**0.689 6**
Business	0.862 9	0.865 2	0.860 6	0.870 0	0.868 8	0.865 4	0.868 8	0.873 3	0.866 9	0.867 2	**0.875 9**
Reference	0.577 3	0.599 5	0.600 1	0.578 1	0.586 0	0.589 9	0.586 0	0.601 0	0.592 1	0.587 5	**0.609 3**
Science	0.392 6	0.426 4	0.421 9	0.416 4	0.434 4	0.409 3	0.434 4	0.435 2	0.440 5	0.421 6	**0.446 3**
Yeast	0.757 7	0.697 6	0.758 3	0.755 7	0.756 6	0.707 7	0.756 6	0.750 3	0.742 4	0.750 1	**0.759 1**
Medical	0.458 2	0.72 16	0.652 5	0.596 3	0.397 0	0.385 3	0.633 7	0.620 1	0.657 8	0.687 2	**0.750 9**
Image	0.423 7	0.383 3	0.420 7	0.409 8	0.364 6	0.381 2	0.410 0	0.408 8	0.423 4	0.425 6	**0.440 7**
average	0.580 5	0.605 1	0.609 6	0.585 0	0.575 7	0.560 7	0.603 9	0.609 7	0.607 1	0.611 6	**0.634 4**

从表 2 至表 4 来看,在 AP、RL 和 HL 这 3 个指标上,本文的 RGLC 算法除了在 Enron 数据集上未取得最优值,在其余的 9 个数据集上均取得最优值。详细来讲,在 Enron 数据集上,PMFS 算法在 AP、RL 和 HL 指标上均取得了最优值,RGLC 算法排名第 3,与 PMFS 算法分别相差 0.027 6、0.003 8 和 0.002 7。其原因可能是 RGLC 算法所选的特征会导致分类器预测了无关的标记。从表 5 的 CV 指标上来看,RGLC 算法除在 Science 数据集上排名第 3 外,在其余的 7 个数据集上均取得最优值,且与最优的 SRMFS 算法仅相差 0.000 9,这可能是由于 Science 数据集中存在更多的负标签。在这 10 个数据集上,从 AP、RL、HL 和 CV 这 4 个指标的平均值以及最优值的数量上来说,本文的 RGLC 算法的排名均最优。综上所述,与其它 10 种算

法相比，RGLC 算法呈现了更优的分类性能，由此证明了结合了全局标记相关性与局部标记相关性的多标记 Relief 特征选择算法是有效的。

表 3　11 种算法在 RL(↓)指标下 10 个多标记数据集上的实验结果

Tab. 3　Experimental results of RL(↓) index with the eleven algorithms on the ten multilabel datasets

Datasets	MFS-MCDM	PMFS	PUM	GLFS	WFSNR	NRFSFN	ENFSFN	SRMFS	LSFRS	SFSLH	RGLC
Arts	0.164 0	0.160 2	0.158 9	0.169 7	0.157 5	0.160 8	0.157 5	0.158 2	0.168 8	0.162 9	**0.154 4**
Computers	0.097 7	0.098 0	0.099 2	0.100 5	0.094 4	0.103 7	0.094 4	0.091 5	0.102 1	0.099 8	**0.091 4**
Enron	0.102 3	**0.092 1**	0.095 2	0.109 2	0.102 2	0.114 2	0.102 2	0.112 1	0.101 5	0.113 2	0.095 9
Health	0.078 8	0.073 7	0.076 3	0.077 6	0.070 7	0.071 1	0.070 7	0.067 7	0.069 8	0.068 1	**0.063 8**
Business	0.046 5	0.046 0	0.048 2	0.044 6	0.043 8	0.043 8	0.043 8	0.042 5	0.046 8	0.041 9	**0.041 2**
Reference	0.099 7	0.095 1	0.094 5	0.100 0	0.097 7	0.098 5	0.097 7	0.094 1	0.097 6	0.099 2	**0.091 6**
Science	0.155 3	0.146 7	0.150 7	0.146 9	0.148 2	0.282 4	0.266 5	0.158 6	0.165 4	0.142 9	**0.133 5**
Yeast	0.172 0	0.220 6	0.171 2	0.172 4	0.172 0	0.214 6	0.172 2	0.185 6	0.182 1	0.196 7	**0.170 4**
Medical	0.126 5	0.067 4	0.082 5	0.082 5	0.141 3	0.138 1	0.080 8	0.095 6	0.077 1	0.080 9	**0.052 3**
Image	0.366 6	0.366 9	0.290 2	0.442 1	0.372 6	0.360 2	0.398 0	0.338 7	0.354 9	0.321 4	**0.224 3**
average	0.140 9	0.136 7	0.126 7	0.144 6	0.140 0	0.158 7	0.148 4	0.134 5	0.136 6	0.132 7	**0.111 9**

表 4　11 种算法在 HL(↓)指标下 10 个多标记数据集上的实验结果

Tab. 4　Experimental results of HL (↓) index with the eleven algorithms on the tent multilabel datasets

Datasets	MFS-MCDM	PMFS	PUM	GLFS	WFSNR	NRFSFN	ENFSFN	SRMFS	LSFRS	SFSLH	RGLC
Arts	0.061 5	0.062 0	0.061 2	0.063 1	0.061 8	0.062 3	0.061 8	0.061 2	0.071 2	0.069 2	**0.060 9**
Computers	0.041 6	0.042 9	0.040 9	0.042 0	0.041 1	0.041 9	0.041 1	0.042 2	0.048 9	0.040 6	**0.039 4**
Enron	0.055 1	**0.049 8**	0.050 0	0.059 9	0.053 5	0.057 6	0.053 5	0.055 0	0.052 9	0.057 6	0.052 5
Health	0.049 9	0.048 1	0.049 2	0.049 8	0.047 1	0.045 8	0.047 1	0.044 3	0.047 0	0.059 8	**0.043 6**
Business	0.028 8	0.028 2	0.028 7	0.028 4	0.028 4	0.028 0	0.028 4	0.028 6	0.028 4	0.027 9	**0.027 6**
Reference	0.035 3	0.032 0	0.032 4	0.034 2	0.034 6	0.031 7	0.034 6	0.031 9	0.031 8	0.032 9	**0.030 9**
Science	0.035 7	0.035 1	0.035 0	0.035 5	0.035 3	0.035 2	0.035 3	0.035 5	0.035 2	0.048 9	**0.034 9**
Yeast	0.197 9	0.248 8	0.197 5	0.196 5	0.201 0	0.233 1	0.2010	0.202 5	0.227 3	0.216 5	**0.193 2**
Medical	0.227 5	0.236 7	0.208 9	0.239 7	0.197 5	0.033 2	0.024 4	0.021 6	0.022 5	0.022 1	**0.016 0**
Image	0.240 4	0.226 5	0.238 5	0.248 9	0.238 7	0.230 4	0.243 2	0.228 5	0.214 9	0.234 9	**0.190 5**
average	0.097 4	0.101 0	0.094 2	0.099 8	0.093 9	0.079 9	0.077 0	0.075 1	0.078 0	0.081 0	**0.068 9**

表 5　11 种算法在 CV(↓) 指标下 10 个多标记数据集上的实验结果

Table 5　Experimental results of CV (↓) index with the eleven algorithms on the tent multilabel datasets

Datasets	MFS-MCDM	PMFS	PUM	GLFS	WFSNR	NRFSFN	ENFSFN	SRMFS	LSFRS	SFSLH	RGLC
Arts	5.786 7	5.667 3	5.652 0	5.923 7	5.625 0	5.634 3	5.625 0	5.691 3	5.786 7	5.698 7	**5.530 8**
Computers	4.634 0	4.692 0	4.708 0	4.754 7	4.521 0	4.915 3	4.521 0	4.370 6	4.548 1	4.584 9	**4.365 2**
Enron	13.887 7	13.1485	13.475 0	14.336 8	14.012 1	14.839 4	14.012 1	13.588 6	14.022 9	13.647 8	**13.024 3**
Health	3.9237	3.765 3	3.867 0	3.929 7	3.693 0	3.678 0	3.693 0	3.568 4	3.598 0	3.558 2	**3.408 6**
Business	2.509 0	2.495 3	2.586 7	2.424 0	2.494 7	2.598 7	2.404 7	2.341 7	2.397 6	2.790 5	**2.338 1**
Reference	3.798 3	3.643 3	3.625 3	3.811 3	3.721 3	3.753 0	3.721 3	3.586 6	3.640 1	3.694 1	**3.495 8**
Science	7.647 0	7.293 0	7.465 0	7.283 0	7.349 3	7.725 7	7.349 3	**7.288 1**	7.339 3	7.368 2	7.289 0
Yeast	6.399 1	7.426 4	6.377 3	6.402 4	6.395 9	7.031 6	6.395 9	6.362 4	6.315 9	6.357 4	**6.306 0**
Medical	6.595 3	3.879 1	4.586 0	4.584 5	7.243 4	7.023 3	4.440 3	4.340 4	5.562 1	3.391 3	**3.340 8**
Image	2.548 8	2.890 0	2.698 8	2.658 8	3.055 0	2.847 5	2.678 8	3.312 2	2.621 5	2.450 0	**2.250 6**
average	5.773 0	5.490 0	5.504 1	5.610 9	5.811 1	6.004 7	5.484 1	5.445 0	5.583 2	5.354 1	**5.134 9**

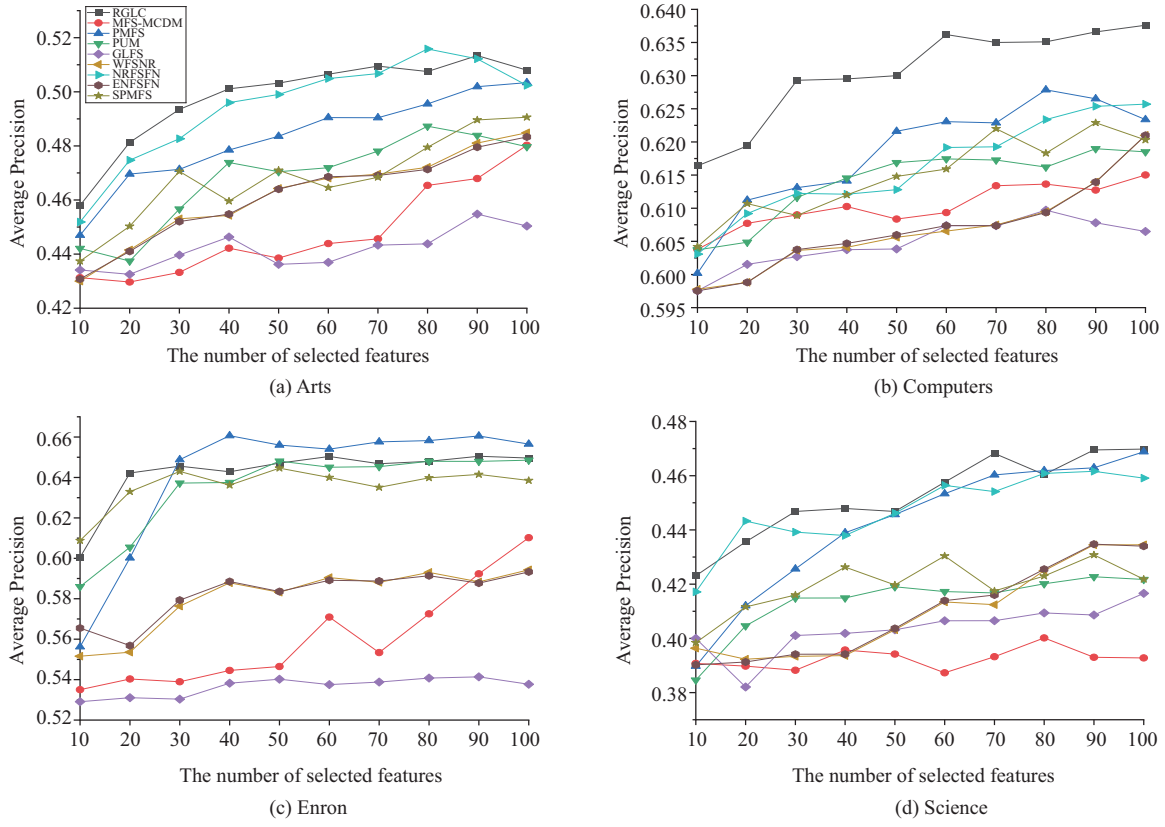

图 1　9 种算法在 4 个多标记数据集上的 AP(↑)指标比较

Fig. 1　Comparison of nine algorithms on four multilabel datasets in terms of AP (↑)

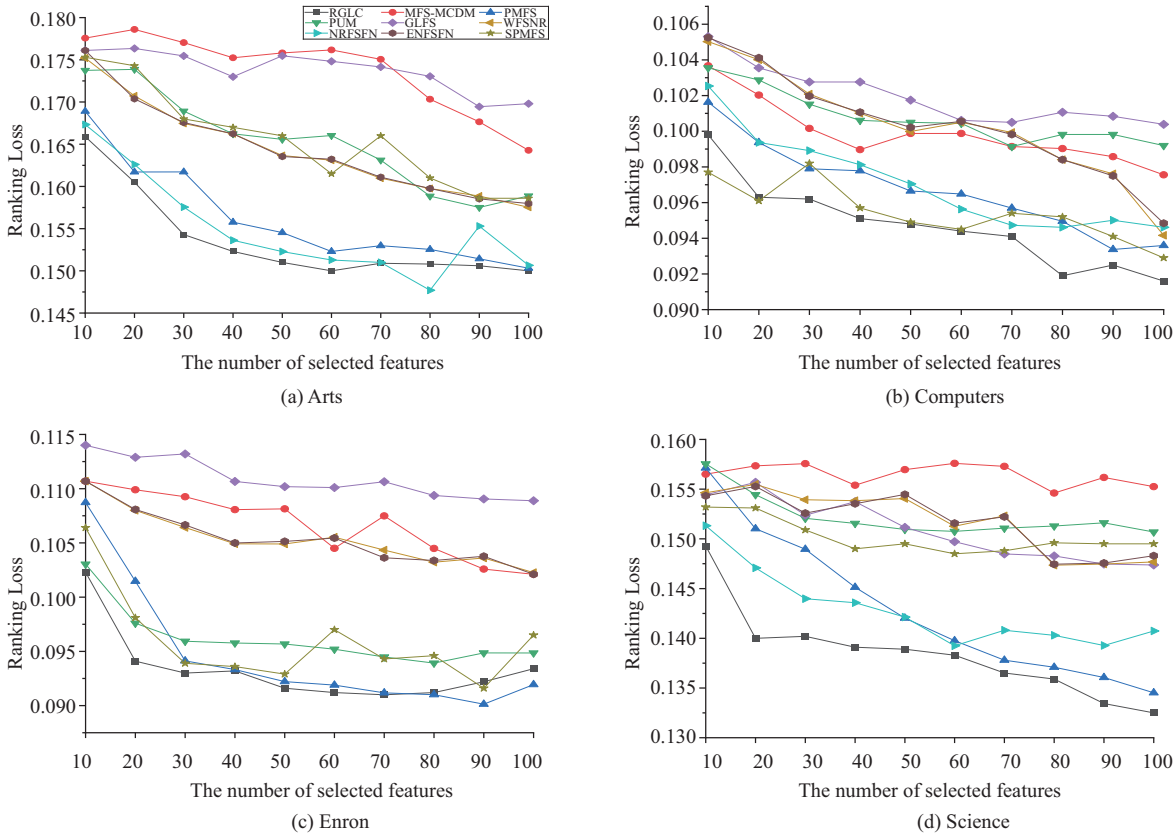

图 2　9 种算法在 4 个多标记数据集上的 RL(↓)指标比较

Fig. 2　Comparison of nine algorithms on four multilabel datasets in terms of RL(↓)

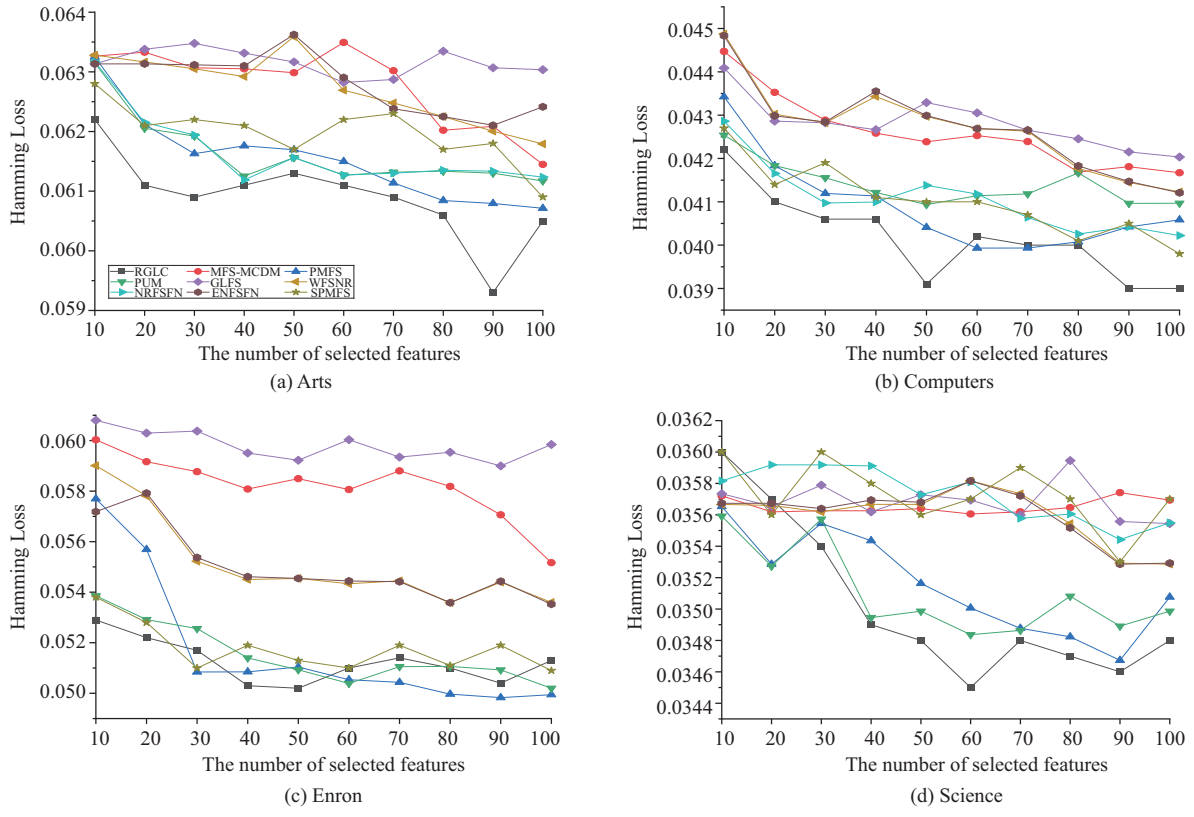

图 3　9 种算法在 4 个多标记数据集上的 HL(↓)指标比较

Fig. 3　Comparison of nine algorithms on four multilabel datasets in terms of HL(↓)

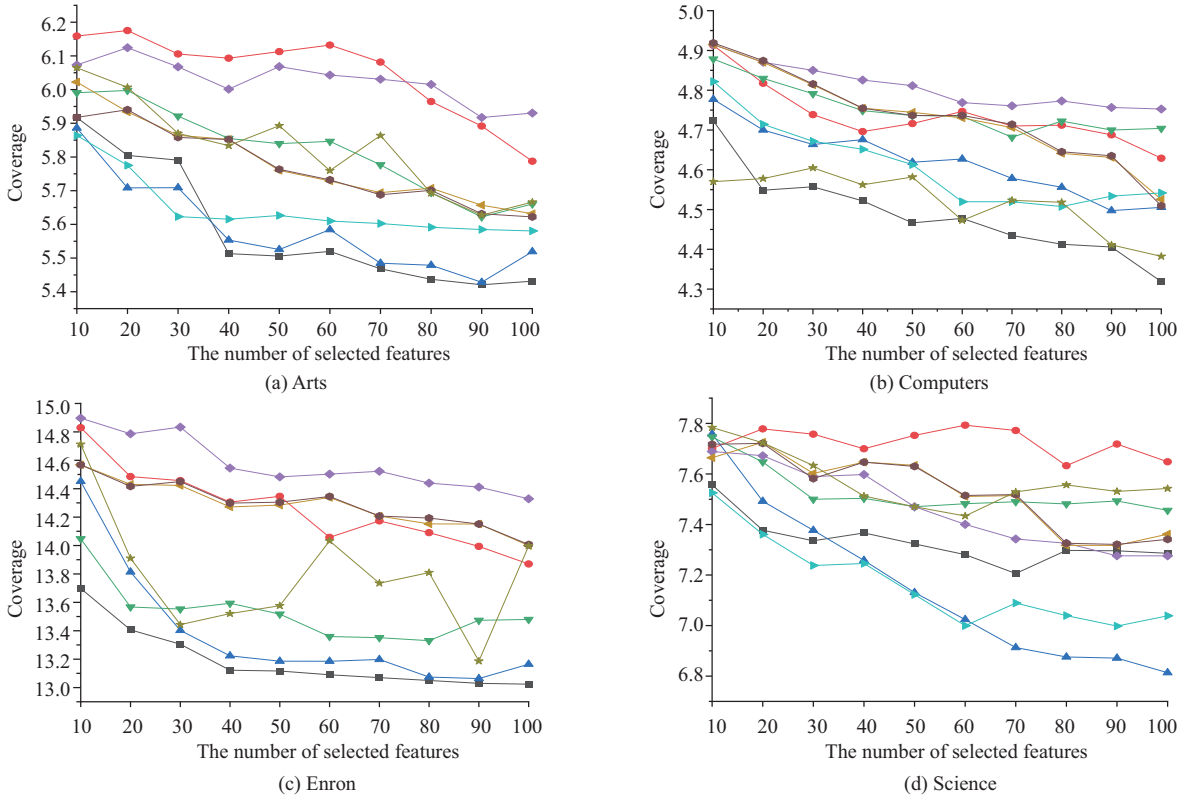

图 4　9 种算法在 4 个多标记数据集上的 CV(↓)指标比较

Fig. 4　Comparison of nine algorithms on four multilabel datasets in terms of CV(↓)

接下来，为进一步比较 RGLC 算法选择的特征数量对分类性能的影响，从表 1 中选取了 Arts、Comput-

ers、Enron 和 Science 这 4 个代表性数据集，在 AP、RL、HL 和 CV 这 4 种指标上进行实验，呈现其变化趋势。实验结果如图 1 至图 4 所示，其中横轴表示选择的特征个数(Selected feature number，N_F)，纵轴表示各算法在 4 种指标上的结果值。由图 1 的 AP 指标来看，RGLC 算法在绝大部分情况下均能取得最优值；在 Enron 数据集上，当 $N_F \geqslant 30$，RGLC 算法取得次优值。由图 2 的 RL 指标来看，RGLC 算法在大部分情况下均能取得最优值；在 Enron 数据集上，当 $N_F=80$ 时，RGLC 算法取得次优值。从图 3 的 HL 指标来看，与其它算法相比，RGLC 算法在多数情况下取得最优值；在 Computers 数据集上当 $N_F=60$ 与 70 时取得次优值，除此之外均能取得最优值；在 Enron 数据集上，当 $N_F \geqslant 90$ 时随着所选特征数量的增加，HL 反而呈现了上升的趋势。根据图 4 的 CV 指标分析，在 Arts、Computers 和 Enron 这 3 个数据集上表现较好。总体来看，RGLC 算法在绝大多数情况下均可以取得最优值，即使未取得最优值的大多情况下也能取得次优值，且在这 4 个数据集上，RGLC 算法相比于其余对比算法表现出更稳定的分类效果。

3.3 参数分析

为了探究 K-means 聚类中 K 值对 RGLC 算法性能的影响，本节选取一个代表性数据集 Arts，针对不同的 K 值进行实验分析。通过表 6 可知，当 $K=5$ 至 1 000 时，RGLC 算法的分类效果大致呈现倒 U 型，且当 $K=200$ 时，RGLC 算法在 AP、RL、HL 和 CV 指标上效果最好。

表 6 Arts 数据集上不同 K 值的实验结果

Tab. 6 Experimental Results for Different K Values on the Arts Dataset

	$K=5$	$K=10$	$K=20$	$K=50$	$K=100$	$K=200$	$K=300$	$K=500$	$K=1\,000$
AP	0.502 2	0.501 1	0.488 4	0.501 0	0.510 8	**0.512 9**	0.510 6	0.490 4	0.502 2
RL	0.153 5	0.154 4	0.156 3	0.153 2	0.153 2	**0.149 7**	0.154 1	0.155 6	0.153 5
HL	0.061 5	0.061 2	0.061 2	0.061 6	0.060 7	**0.059 7**	0.061 7	0.061 2	0.0615
CV	5.579 8	5.599 8	5.589 9	5.544 3	5.496 6	**5.405 7**	5.559 3	5.560 9	5.579 8

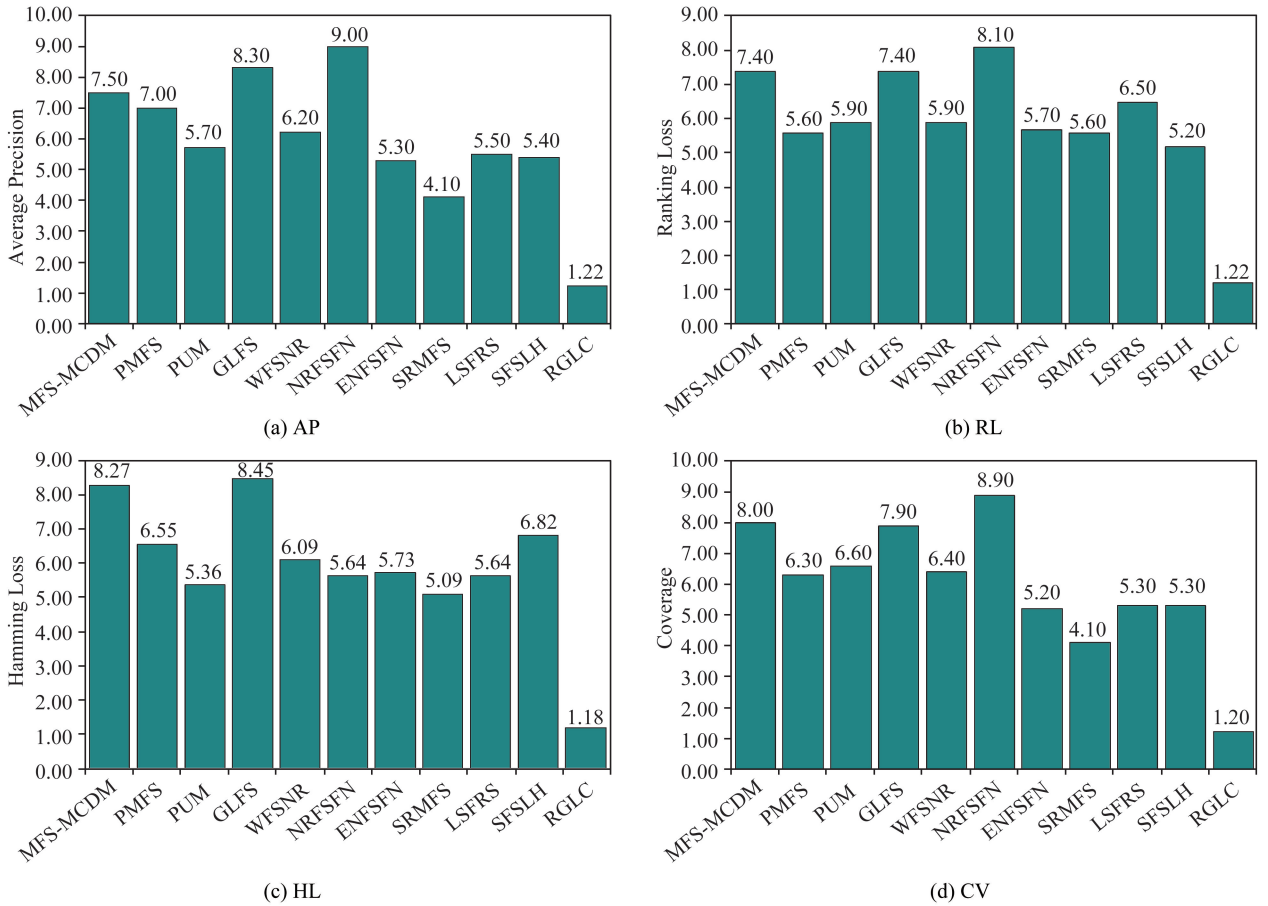

图 5 MLKNN 分类器下 11 种算法的平均排名

Fig. 5 Average ranking of eleven algorithms under the MLKNN classifier

3.4　统计结果分析

本节使用表 2 至表 5 结果进行统计结果分析。11 个算法在 10 个数据集上的平均排名如图 5 所示，本文的 RGLC 算法位于图最右侧，即在 AP、RL、HL 和 CV 这 4 种指标上排名最好，且显著好于其余对比算法。在 AP 指标上，ENFSFN 算法、LSFRS 算法和 SFSLH 算法取得相近的平均排名。在 RL 指标上，PUM 算法与 WFSNR 算法排名相同。在 CV 指标上，SRMFS 取得次优排名。综上所述，本文的 RGLC 算法与其余 10 种算法之间存在显著差异性，且分类性能优于其它算法。参照文献[34]的统计分析过程，通过 CD 图可视化 Nemenyi 检验，具体结果如图 6 所示。

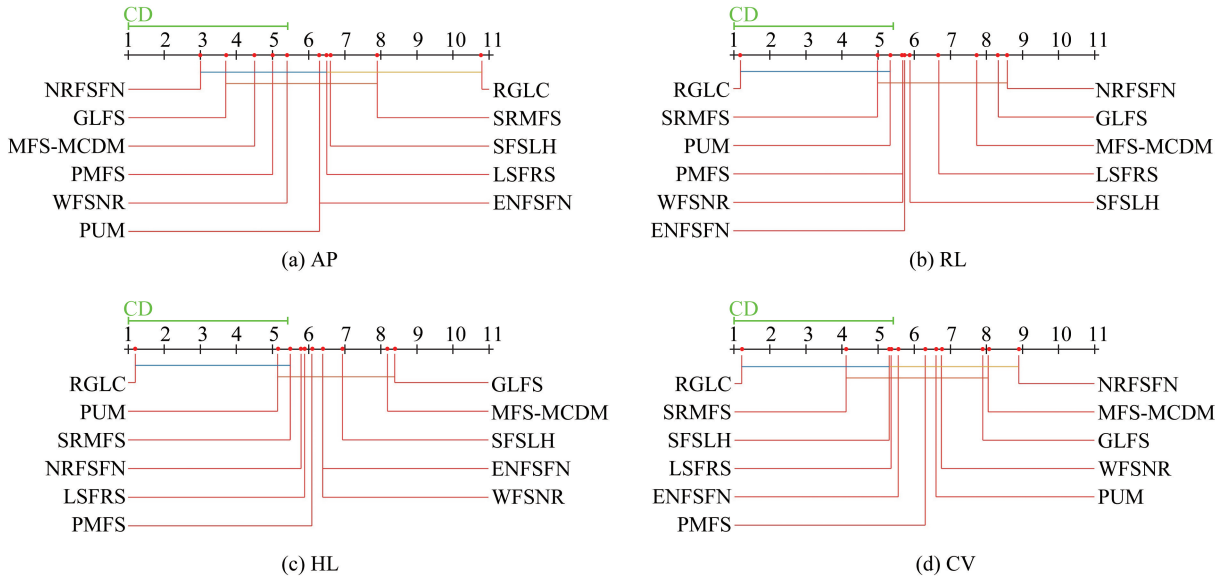

图 6　MLKNN 分类器下 11 种算法的 Nemenyi 检验结果

Fig. 6　Nemenyi test results of eleven algorithms under the MLKNN classifier

4　结束语

本文为了解决传统 Relief 算法在特征选择过程中忽视标记相关性以及现有余弦相似度在高维数据局部标记相关性度量的不足的问题，提出了一种新的多标记 Relief 特征选择方法，该方法将全局标记相关性与局部标记相关性相结合。首先引入了 K-means 聚类算法，通过聚类将样本划分为不同的簇，并利用欧式距离计算样本到各质心的距离。在此基础上，计算簇内特征的平均值，以此代表簇中样本在局部空间特征上的值，从而构建了样本的局部标记空间。接下来，定义了特征上的欧式距离来度量样本的全局标记相关性。在局部标记空间中，采用 L1 范数的平方根对余弦相似度进行改进，以更有效地捕捉样本的局部标记相关性。在 Relief 算法的基础上，结合了全局标记相关性与局部标记相关性来衡量样本的相似度，从而能对最近邻同类样本与最近邻异类样本进行更准确地判别。最后构建了特征权重的迭代公式，使 Relief 算法不仅能在全局样本标记空间上学习，还能从样本的局部标记空间上获取更全面的标记相关性信息。为了验证所提算法的有效性，在 10 个公开多标记数据集上进行了对比测试。实验结果表明，与其他 10 种多标记分类算法相比，本文所提算法在分类性能上表现出色。然而，在局部标记相关性的度量中，仍需手动设置聚类数量，而不同的聚类数量可能会对算法的分类效果产生影响。因此，如何设计一种自适应选择最佳聚类数量的方法，将是未来研究的重要方向。

参　考　文　献

[1] FAN Y L, LIU J H, TANG J N, et al. Learning correlation information for multilabel feature selection[J]. Pattern Recognition, 2024, 145: 109899.

[2] QIAN W B, XIONG Y S, DING W P, et al. Label correlations-based multilabel feature selection with label enhancement[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107310.

[3] MA J J, XU F, RONG X F. Discriminative multilabel feature selection with adaptive graph diffusion[J]. Pattern Recognition, 2024, 148: 110154.
[4] LI Y H, HU L, GAO W F. Multilabel feature selection with high-sparse personalized and low-redundancy shared common features[J]. Information Processing & Management, 2024, 61(3): 03633.
[5] 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, https://link.cnki.net/urlid/50.1075.tp.20240513.1427.017.
[6] 杨希洪, 郑群, 章佳欣, 等. 基于特征插值的深度图对比聚类算法[J/OL]. 计算机科学, 2024. https://link.cnki.net/urlid/50.1075.tp.20240422.1507.006.
[7] 徐天杰, 王平心, 杨习贝, 等. 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.
[8] 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(04): 983-996.
[9] LI Y H, HU L, GAO W F. Label correlations variation for robust multilabel feature selection[J]. Information Sciences, 2022, 609: 1075-1097.
[10] SUN L J, FENG S H, LIU J, et al. Global-local label correlation for partial multilabel learning[J]. IEEE Transactions on Multimedia. 2022, 24: 581-593.
[11] GAO W F, HU L, ZHANG P. Class-specific mutual information variation for feature selection[J]. Pattern Recognition, 2018, 79: 328-339.
[12] Zhu Y, Kwok J T, Zhou Z H. Multi-label learning with global and local label correlation[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(6): 1081-1094.
[13] JI S W, TANG L, YU S P, et al. Extracting shared Subspace for multilabel classification [C]. //Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas Nevada USA, August 24-27, 2008, New York: Association for Computing Machinery, 2008: 381-389.
[14] 容斌元, 徐媛媛, 吕亚兰, 等. 融合标签局部相关性的标签分布学习[J]. 山东大学学报(理学版), 2022, 57(7): 53-64.
[15] 朱赛赛, 贾修一, 李泽超. 一种基于全局和局部标记相关性的多标记分类算法[J]. 电子学报, 2020, 48(12): 2345-2351.
[16] 丁思凡, 王锋, 魏巍. 一种基于标签相关度的 Relief 特征选择算法[J]. 计算机科学, 2021, 48(4): 91-96.
[17] ZHANG J D, LIU K Y, YANG X B, et al. Multilabel learning with Relief-based label-specific feature selection[J]. Applied Intelligence, 2023, 53: 18517-18530.
[18] 孙林, 丰昌武, 陈雨生, 等. 基于样本全局相似度和 Relief 的缺失标记特征选择[J]. 昆明理工大学学报(自然科学版), 2024, 49(2): 39-48.
[19] XIA P, ZHANG L, LI F. Learning similarity with cosine similarity ensemble[J]. Information sciences, 2015, 307: 39-52.
[20] FAN Y L, CHEN B H, HUANG W Q, et al. Multilabel feature selection based on label correlations and feature redundancy[J]. Knowledge-Based Systems, 2022, 241: 108256.
[21] 周昌, 李向利, 李俏霖, 等. 余弦相似度的稀疏非负矩阵分解算法[J]. 计算机科学, 2020, 47(10): 108-113.
[22] SOHANGIR S, WANG D. Improved sqrt-cosine similarity measurement[J]. Journal of Big Data, 2017, 4: 1-13.
[23] ABIODUN M. IKOTUN, ABSALOM E. EZUGWU, LAITH A, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178-210.
[24] AGGARWAL C, ALEXANDER H, DANIEL K. On the surprising behavior of distance metrics in high dimensional space [C]//. Database Theory-ICDT 2001, Berlin, Heidelberg, January 4-6, 2001, Berlin, Heidelberg: Springer, 2001: 420-434.
[25] HASHEMI A, DOWLATSHAHI M B, NEZAMABADI-POUR H. MFS-MCDM: Multi-label feature selection using multi-criteria decision-making[J]. Knowledge-Based Systems, 2020, 206: 106365.
[26] HASHEMI A, DOWLATSHAHI M B, NEZAMABADI-POUR H. An efficient Pareto-based feature selection algorithm for multi-label classification[J]. Information Sciences, 2021, 581: 428-447.
[27] XU J, SHEN K, SUN L. Multi-label feature selection based on fuzzy neighborhood rough sets[J]. Complex & Intelligent Systems, 2022, 8(3): 2105-2129.
[28] ZHANG J, WU H, JIANG M, et al. Group-preserving label-specific feature selection for multi-label learning[J]. Expert Systems with Applications, 2023, 213: 118861.
[29] 孙林, 黄苗苗, 徐久成. 基于邻域粗糙集和 Relief 的弱标记特征选择方法 [J]. 计算机科学, 2022, 49 (4): 152-160.
[30] YIN T, CHEN H, YUAN Z, et al. Noise-resistant multilabel fuzzy neighborhood rough sets for feature subset selection[J]. Information Sciences, 2023, 621: 200-226.
[31] Sun L, Ma Y, Ding W, et al. LSFSR: Local label correlation-based sparse multilabel feature selection with feature redundancy[J]. Information Sciences, 2024, 667: 120501.
[32] Sun L, Ma Y, Ding W, et al. Sparse feature selection via local feature and high-order label correlation[J]. Applied Intelligence, 2024, 54(1): 565-591.
[33] 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.
[34] YIN T Y, CHEN H M, WAN J H, et al. Exploiting feature multi-correlations for multilabel feature selection in robust multi-neighborhood fuzzy β covering space[J]. Information Fusion, 2024, 104: 102150.

(责任编辑:刘庆松)