智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于加权信息粒化的多标记数据特征选择算法

胡军, 王海峰

引用本文:
胡军,王海峰. 基于加权信息粒化的多标记数据特征选择算法[J]. 智能系统学报, 2023, 18(3): 619–628.
HU Jun,WANG Haifeng. Feature selection algorithm of multi-labeled data based on weighted information granulation[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(3): 619–628.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202111058

DOI: 10.11992/tis.202111058
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.tp.20230317.1408.004.html

基于加权信息粒化的多标记数据特征选择算法

胡军[1,2], 王海峰[1,2]
（1. 重庆邮电大学 计算机科学与技术学院, 重庆 400065; 2. 重庆邮电大学 计算智能重庆市重点实验室, 重庆 400065）

摘　要: 特征选择能去除不相关和冗余的特征, 是解决多标记数据维度灾难的有效工具。现有的多标记特征选择算法没有考虑标记空间存在的相关性, 认为每个样本的相关标记的重要性相同, 并且忽略了特征空间可能是标记重要性差异形成的内在因素, 使得选择的特征不能精确全面地刻画样本且计算过程复杂。为此, 本文利用标记间的相关性对标记空间进行划分以简化计算, 并定义标记重要性度量和特征权重, 在此基础上提出了一种基于加权信息粒化的多标记特征选择算法。通过在真实多标记数据集上的实验对比分析, 本文提出的算法在各项评价指标上均优于其他对比算法, 验证了算法的有效性和可行性。

关键词: 邻域粗糙集; 信息粒化; 多标记学习; 标记重要性; 标记关系; 特征权重; 特征选择; 谱聚类
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2023)03−0619−10

中文引用格式: 胡军, 王海峰. 基于加权信息粒化的多标记数据特征选择算法 [J]. 智能系统学报, 2023, 18(3): 619–628.
英文引用格式: HU Jun, WANG Haifeng. Feature selection algorithm of multi-labeled data based on weighted information granulation[J]. CAAI transactions on intelligent systems, 2023, 18(3): 619–628.

Feature selection algorithm of multi-labeled data based on weighted information granulation

HU Jun[1,2], WANG Haifeng[1,2]
(1. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Feature selection can remove irrelevant and redundant features. It is an efficient tool to solve the disaster of multi-labeled data dimensions. Existing multi-labeled feature selection algorithms did not take the correlation of label space into account, and considered that the relevant labels of each sample have the same importance, and ignored that the feature space may be the internal factor caused by the difference of label importance, so that the selected features can not accurately and comprehensively describe the samples and the calculation process is very complex. In this paper, the correlation between labels is used to divide the label space to simplify the calculation. Then, the label importance measure and feature weight are defined. And further, a feature selection algorithm of multi-label data based on weighted information granulation is proposed. The comparison and analysis on real multi-labeled data set of experiment show that the proposed algorithm is superior to other comparison algorithms in all evaluation indicators, which verifies effectiveness and feasibility of the algorithm.

Keywords: neighborhood rough set; information granulation; multi-label learning; label significance; label relationship; feature weight; feature selection; spectral clustering

在传统的分类学习中, 每个实例只属于一个类别标记。但在现实世界中, 一个样本可能涉及多个语义信息, 这种情况符合多标记学习中单个样本多义性的特征。因此, 多标记学习更贴近实际应用场景, 可以更准确地描述和解决现实世界中的问题。例如, 一幅图像往往同时包含“天空”“湖泊”“建筑”等多种重要的语义概念, 一种蛋白质可能同时具有多个功能, 一篇新闻报道可能同

收稿日期: 2021−11−30.　**网络出版日期**: 2023−03−20.
基金项目: 国家自然科学基金项目(61936001,62276038); 重庆市自然科学基金项目(cstc2019jcyj-cxttX0002, cstc2021ycjh-bgzxm0013); 重庆市教委重点合作项目(HZ2021008).
通信作者: 胡军. E-mail: hujun@cqupt.edu.cn.

时与“体育”“社会”“娱乐”“财经”等多个话题相关。这类复杂数据很难用单一的语义标记进行描述，合理的处理方式是为每个样本赋予一个标记集合，进而建模和学习。在多标记学习框架下，样本由特征集和相关的标记集构成，学习的目标是将由特征集描述的样本映射到多个类别标记，现已被广泛应用在机器学习、数据挖掘、模式识别等领域[1-4]。

在多标记学习中，数据特征的高维性容易导致维度灾难，特征选择技术是解决该问题的有效工具。其解决思路之一是通过问题转换的方法来处理多标记数据的特征选择问题，例如，利用二元关系（binary relevance，BR）[5]、标记幂集（label powerset，LP）[6]、剪枝问题转换（pruned problem transformation，PPT）等将多标记问题转换为单标记问题。其中，BR 将多标记数据转化为多个二分类数据，运用单标记特征选择算法对每个二分类数据进行特征排序。LP 将多标记数据中的每种标记组合看作一种类别，进而将多标记数据转化为多类别单标记数据。在剪枝问题转换方法的基础上，Read 等[7] 运用卡方检验方法对转换后的数据做特征选择，Doquire 等[8] 结合互信息提出了一种多标记数据特征选择算法，避免了 LP 中类别指数增加及类别不平衡的问题。然而，基于问题转化的特征选择算法往往导致转化后的数据不能真正反映多标记数据之间的真实分布，普遍存在转化过程繁琐耗时，转化后类别数量巨大，忽略标记相关性等问题。

为了避免转化数据带来的信息丢失，基于算法自适应的特征选择算法成为近年来的研究热点。Lee 等[9] 运用互信息度量特征与标记的相关性，使用多元互信息近似计算高维联合熵，大大降低了高维信息熵的计算复杂度。在此基础上，Lee 等[10] 从理论上分析了为何考虑低阶交叉信息的得分函数能获得有效的特征子集，并提出了新的得分函数以考虑任何程度的交叉信息。由于算法的相关性度量和冗余性度量是独立计算的，在处理大规模多标记数据时，冗余性不能得到很好地体现进而引入冗余特征，为此 Lee 等[11] 在冗余性度量计算中融合相关性度量，提出了新的特征度量准则。考虑到一些多标记特征选择算法[12-15] 中运用互信息来评估特征的相关性，忽略了已选特征对于相关性度量的影响和标记关系对特征相关性度量的影响，Zhang 等[16] 引入条件互信息综合考虑已选特征和标记关系的影响，使用基于两种条件互信息的双条件相关度量特征的相关性。Zhang 等[17] 提出了标记依赖、标记冗余、标记互补 3 种标记关系，通过对 3 种标记关系和标记关系随着不同特征变化的分析为多标记特征选择提供了新的视角。Qian 等[18] 认为在处理多标记数据的特征选择问题时还应该考虑特征之间的互补关系，并结合标签分布学习和特征互补关系提出了一种多标签特征选择算法。

多标记数据往往是高维的连续型数据，以上提到的方法处理连续型数据需要先对数据进行离散化，由此会造成部分信息丢失。邻域粗糙集[19] 通过距离公式和邻域半径来进行信息粒化，可以直接处理连续型数据，已被广泛应用于单标记数据的特征选择[20-23]。为将邻域粗糙集应用于多标记数据，段洁等[24] 在现有邻域粗糙集基础上重新定义了下近似和依赖度，设计了基于邻域粗糙集的多标记特征选择算法。Lin 等[25] 引入大间隔来粒化样本空间，从 3 种不同的认知观点构造多标记邻域信息系统及其不确定性度量，为多标记特征选择提供了新的邻域粗糙集模型。Long 等[26] 和黄锦涛等[27] 使用标记增强方法，将逻辑标记转化为标记分布作为辅助监督信息，由此构建面向标记分布数据的代价敏感特征选择算法。Jorge 等[28] 提出了一种分布式模型来计算评估特征的指标和多个标记间的互信息，为大规模多标记数据特征选择提供了方法。

上述研究均假定同一实例的相关标记重要程度相同，并且忽略了特征空间是标记重要程度可能存在差异的内在因素。然而，现实应用中多标记数据的标记间存在潜在的语义关系，不同标记的重要程度也有所不同，并且同一特征与不同标记的相关程度也存在差异，也就是说每个标记都存在相关性高的专属特征。因此，不能区分每个标记对样本的分辨能力，也忽略了每个特征对于不同标记所提供分类信息的差异，进而影响特征选择算法的性能。为此，本文使用谱聚类挖掘标记集中潜在的结构信息，量化标记的重要性和特征权重，进而提出了一种基于加权信息粒化的多标记特征选择算法（multi-label feature selection based on weighted information granulation，MFWIG）。最后，实验结果表明，本文通过融合标记的重要性和特征权重，能有效选择出重要特征，同时避免了高维联合熵的复杂运算，验证了本文算法的有效性。

1　多标记邻域熵

1.1　大间隔

一个单标记决策系统可以表示为$<U,F,D>$, 其中, $U=\{x_1,x_2,\cdots,x_n\}$表示样本集合, $F=\{f_1,f_2,\cdots,f_m\}$为描述样本的一组特征, D是类别标记。

定义 1[19]　设U是非空样本集合, 若存在U上的距离函数Δ, 使得$\forall x_i,x_j,x_k\in U$满足:

1) $\Delta(x_i,x_j)\geqslant 0$, 当且仅当$x_i=x_j$时, $\Delta(x_i,x_j)=0$;

2) $\Delta(x_i,x_j)=\Delta(x_j,x_i)$;

3) $\Delta(x_i,x_k)+\Delta(x_k,x_j)\geqslant\Delta(x_i,x_j)$。

则称$<U,\Delta>$是一个度量空间。

在m维特征空间中, 给定样本$x_i=(f_{1i},f_{2i},\cdots,f_{mi})$和$x_j=(f_{1j},f_{2j},\cdots,f_{mj})$,样本间的闵可夫斯基距离定义为

$$\Delta_P(x_i,x_j)=\left(\sum_{k=1}^{m}|f_{ki}-f_{kj}|^P\right)^{\frac{1}{P}}$$

当$P=1$时, Δ函数表示曼哈顿距离; 当$P=2$时, Δ函数表示欧氏距离; 当$P\to\infty$时, $\Delta_p(x_i,x_j)=\max_k\left|x_{ki}-x_{kj}\right|$。

定义 2[25]　设U是单标记下的非空样本集合, $\forall x\in U$, 则样本x的类间隔定义为

$$m(x)=\Delta(x,M(x))-\Delta(x,H(x)) \tag{1}$$

式中: $H(x)$表示样本空间U中距离样本x最近的同类样本, $M(x)$表示样本空间U中距离样本x最近的异类样本, $\Delta(x,M(x))$表示样本x到$M(x)$的距离, $\Delta(x,H(x))$表示样本x到$H(x)$的距离。分类间隔为样本x到$M(x)$和$H(x)$的距离之差, 如图 1 所示。

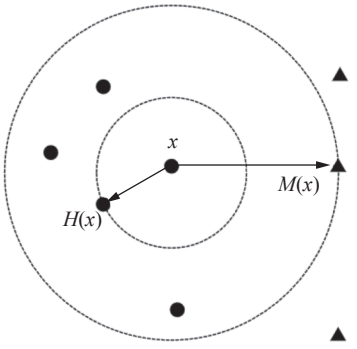


图 1　样本x的分类间隔$m(x)$
Fig. 1　The $m(x)$ of sample x

单标记系统中一个样本只有一个类别标记, 但在一些实际问题中一个样本可能存在多个类别标记, 即多标记系统。一个多标记决策系统可以表示为$<U,F,L>$, 其中,$U=\{x_1,x_2,\cdots,x_n\}$表示样本集合, $F=\{f_1,f_2,\cdots,f_m\}$为描述样本的一组特征, $L=\{l_1,l_2,\cdots,l_k\}$代表标记集合。$\forall x\in U$, $\forall l_i\in L$, 则样本x在标记l_i下的分类间隔为

$$m_{l_i}(x)=\Delta(x,M_{l_i}(x)-\Delta(x,H_{l_i}(x)) \tag{2}$$

式中: $H_{l_i}(x)$和$M_{l_i}(x)$分别表示在标记l_i下样本空间U中距离样本x最近的同类样本和异类样本。

1.2　多标记邻域熵

给定一个多标记决策系统$<U,F,L>$, 根据式(2)可知, $m_{l_i}(x)$表示样本x在单个标记l_i下的分类间隔, 则样本x在标记集L下的分类间隔根据 3 种不同的认知观点可分为 3 类[25], 以下给出中立观点下多标记邻域定义。

定义 3[25]　给定多标记决策系统$\mathrm{MDT}=<U,F,L>$, $\forall x\in U$, 样本x在中立观点下的多标记邻域定义为

$$\delta^{\mathrm{neu}}(x)=\{y|\Delta(x,y)\leqslant m^{\mathrm{neu}}(x)\} \tag{3}$$

式中: $m^{\mathrm{neu}}(x)=\dfrac{1}{|L|}\sum_{i=1}^{L}m_{l_i}(x)$表示在中立观点下的分类间隔。

定义 4[25]　设$U=\{x_1,x_2,\cdots,x_n\}$表示多标记下的非空样本集合, F为样本的特征集合, 设$f\subseteq F$为任意的特征子集, $\forall x_i\in U$, 样本x_i由f导出的邻域表示为$\delta_f^{\mathrm{neu}}(x_i)$,则$f$的邻域熵定义为

$$\mathrm{NH}^{\delta^{\mathrm{neu}}}(f)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta_f^{\mathrm{neu}}(x_i)\|}{n}\right)$$

定义 5[25]　设$f_1\subseteq F$, $f_2\subseteq F$为任意两组特征, $\forall x_i\in U$, 样本x_i由$f_1\cup f_2$导出的邻域表示为$\delta_{f_1\cup f_2}^{\mathrm{neu}}(x_i)$, 则$f_1$、$f_2$的联合熵定义为

$$\mathrm{NH}^{\delta^{\mathrm{neu}}}(f_1,f_2)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta_{f_1\cup f_2}^{\mathrm{neu}}(x_i)\|}{n}\right)$$

定义 6[25]　设$f_1\subseteq F$, $f_2\subseteq F$为任意两组特征, 在已知特征f_1的条件下, f_1, f_2的条件熵定义为

$$\mathrm{NH}^{\delta^{\mathrm{neu}}}(f_2|f_1)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta_{f_1\cup f_2}^{\mathrm{neu}}(x_i)\|}{\|\delta_{f_1}^{\mathrm{neu}}(x_i)\|}\right)$$

定理 1[28]　设$f_1\in F,f_2\in F$为样本x_i的两组特征, NMI为特征f_1和特征f_2的互信息, 那么有:

$$\mathrm{NMI}(f_1,f_2)=\mathrm{NMI}(f_2,f_1)=\mathrm{NH}(f_1)-\mathrm{NH}(f_1|f_2)=\mathrm{NH}(f_2)-\mathrm{NH}(f_2|f_1)=\mathrm{NH}(f_1)+\mathrm{NH}(f_2)-\mathrm{NH}(f_1,f_2)$$

由定理 1 可得, $\forall f_1\subseteq F$, $f_2\subseteq F$, f_1, f_2的邻域互信息为

$$\mathrm{NMI}^{\delta^{\mathrm{neu}}}(f_1,f_2)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta_{f_1}^{\mathrm{neu}}(x_i)\|\cdot\|\delta_{f_2}^{\mathrm{neu}}(x_i)\|}{n\|\delta_{f_1\cup f_2}^{\mathrm{neu}}(x_i)\|}\right)$$

2　多标记特征选择算法

目前多数多标记特征选择方法都假定每个样

本的相关标记的重要性相同，并且忽略了特征空间对标记重要性的影响，不能区分每个标记对样本的分辨能力，也忽略了每个特征对于不同标记所提供分类信息的差异，进而影响特征选择算法的性能。此外，多标记特征选择算法大多都考虑在标记全集下进行特征选择，这往往包含大量高维信息熵的计算，高维信息熵的计算复杂，并且从分类性能的方面看，考虑标记全集得出的特征子集对于某个标记或某些标记可能并不重要，即忽略了某些标记的专属特征。为此，本节利用标记空间中潜在的标记关系将标记集划分以简化计算和选择专属特征，使选择的特征提供更全面的分类信息；并在此基础上构造融合标记重要性和特征权重的多标记加权邻域关系，提出了一种基于加权信息粒化的多标记数据特征选择算法 MFWIG。图 2 给出了算法的基本框架。首先，运用谱聚类挖掘标记集划分为标记簇。其次，利用标记间互信息度量簇中各标记的重要性，并根据特征与标记的相关性赋予特征相应的权重。最后结合标记重要度和特征权重设计新的特征选择度量标准，并在该标准的基础上构造特征选择算法。

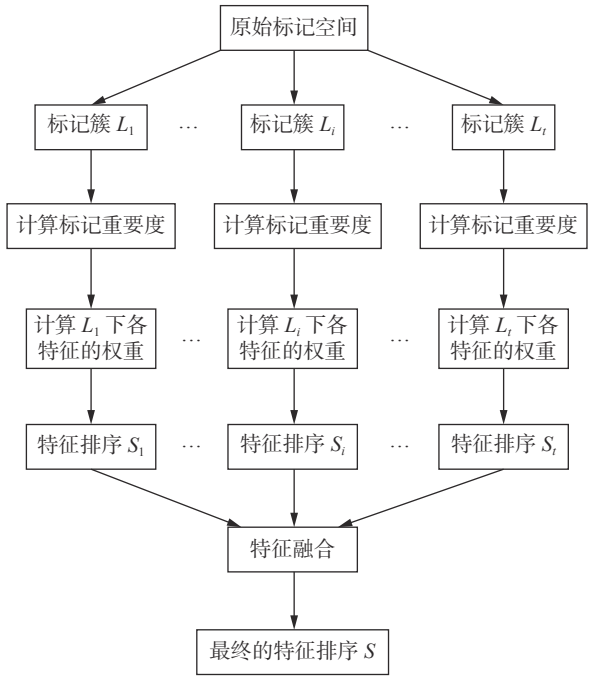


图 2　MFWIG 算法框架

Fig. 2　Framework of MFWIG

2.1　标记簇的划分

为了得到合适的标记划分，本文使用谱聚类来挖掘标记空间中潜在的语义信息将标记空间划分为若干标记簇。相较于传统的聚类方法谱聚类更能适应数据的分布，聚类过程更高效，更适合处理标记集这样的稀疏数据。

设图 $G=(V,E)$，其中节点集合 V 由标记集 $L=\{l_1,l_2,\cdots,l_k\}$ 构成，边集合 E 中的每条边 e 连接的两个标记，表示为两个标记间的权重。谱聚类通过对图进行切割，让切割后的不同子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到划分标记簇的目的。

2.2　多标记加权邻域

给定一个多标记决策系统 $\mathrm{MDT}=<U,F,L>$，通过谱聚类获得若干簇，令得到的标记划分为 $L^P=\{L_1,L_2,\cdots,L_t\}$。

定义 7　设 $\forall L_c\in L^P$，$\forall l_i\in L_c$，标记 l_i 的重要度定义为

$$L_{\mathrm{sig}}(l_i)=\frac{\sum_{l_j\in L_c,i\neq j}I(l_i,l_j)}{\sum_{l_k\in L_c}\sum_{l_j\in L_c,k\neq j}I(l_k,l_j)}\tag{4}$$

式中：$I(l_i,l_j)$ 为标记 l_i 和标记 l_j 之间互信息[29]。

给定标记 l_i，设单标记决策系统为 $<U,F,l_i>$，$\forall f\in F,\forall x\in U,\phi(x,a)$ 为样本 x 的特征值，$\boldsymbol{A}$ 为特征矩阵，决策向量 $\boldsymbol{Y}=[\phi(x_1,l_i)\ \phi(x_2,l_i)\ \cdots\ \phi(x_n,l_i)]^{\mathrm{T}}$，特征的分配系数 $\boldsymbol{v}=[v(f_1)\ v(f_2)\ \cdots\ v(f_m)]^{\mathrm{T}}$。特征分配系数越大特征与标记 l_i 的相关性越高，则赋予其更高的权重。为了得到最优的特征分配系数，求解以下最优问题[22]：

$$v^*=\mathrm{argmin}\|\boldsymbol{A}\boldsymbol{v}-\boldsymbol{Y}\|^2$$

其中 $\boldsymbol{A}=\begin{bmatrix}\phi(x_1,f_1) & \cdots & \phi(x_1,f_m)\\ \vdots & \ddots & \vdots\\ \phi(x_n,f_1) & \cdots & \phi(x_n,f_m)\end{bmatrix}$。

定义 8[22]　设 $\forall L_c\in L^P$，$\forall l_i\in L_c$，,则单标记决策系统可以表示为 $<U,F,l_i>$，$\forall f\in F$，特征 f 关于类别标记 l_i 的权重定义为

$$w(f)=|F||v(f)|\Big/\sum_{f_i\in F}|v(f_i)|$$

定理 2[22]　设 $\forall L_t\in L^P$，$\forall l_i\in L_t$，单标记决策系统可以表示为 $<U,F,l_i>$，特征权重向量为 $\boldsymbol{\omega}=[\omega(f_1)\ \omega(f_2)\ \cdots\ \omega(f_m)]^{\mathrm{T}}$，$\forall f\in F$ 有：

1）$\omega(f)\geqslant 0$；

2）$\sum_{f_i}\omega(f_i)=|F|$。

定义 9　设 $\forall L_c\in L^P$，$L_c=\{l_s,l_{s+1},\cdots,l_h\}$，则多标记决策系统可以表示为 $<U,F,L_c>$，$\forall f\in F$，特征 f 关于标记簇 L_c 的权重定义为

$$w_m(f)=\frac{|F|\sum_{l_i\in L_c}L_{\mathrm{sig}}(l_i)\left|v_{l_i}(f)\right|}{\sum_{f_j\in F}\sum_{l_i\in L_c}L_{\mathrm{sig}}(l_i)\left|v_{l_i}(f_j)\right|}\tag{5}$$

式中：$L_{\mathrm{sig}}(l_i)$ 表示标记 l_i 的重要度，$v_{li}(f)$ 表示特征 f 在标记 l_i 下的特征分配系数。在此基础上，$\forall x_1$，$x_2\in U$，样本间的加权距离为

$$\Delta^{\mathrm{wei}}(x_1,x_2)=\sqrt{\sum\nolimits_{f\in F}(w_m(f)(\phi(x_1,f)-\phi(x_2,f)))^2}$$

对于标记l_i, 样本x的加权分类间隔$m'_{l_i}(x)$为

$$m'_{l_i}(x)=\Delta^{\mathrm{wei}}(x,\mathrm{NM}_{l_i}(x))-\Delta^{\mathrm{wei}}(x,\mathrm{NH}_{l_i}(x)) \tag{6}$$

定义 10 设$\forall L_c\in L^P$, $L_c=\{l_s,l_{s+1},\cdots,l_h\}$, 则多标记决策系统可以表示为$<U,F,L_c>$, $\forall x\in U$, 样本x在标记簇L_c下的加权分类间隔定义为

$$m^{\mathrm{wei}}_{L_c}(x)=\sum\nolimits_{l_i\in L_c}L_{\mathrm{sig}}(l_i)*m'_{l_i}(x) \tag{7}$$

$\forall L_c\in L^P$, $L_c=\{l_s,l_{s+1},\cdots,l_h\}$,则多标记决策系统可以表示为$<U,F,L_c>$, $\forall x\in U$, 根据式(6)和式(7)样本x在标记簇L_c下的加权邻域为

$$\delta^{\mathrm{wei}}(x)=\{y|\Delta^{\mathrm{wei}}(x,y)\leqslant m^{\mathrm{wei}}_{L_c}(x),y\in U\} \tag{8}$$

若$m^{\mathrm{wei}}_{L_c}(x)\leqslant 0$, 则令$m^{\mathrm{wei}}_{L_c}(x)=0$。

2.3 多标记加权邻域熵

定义 11 设$U=\{x_1,x_2,\cdots,x_n\}$表示多标记下的非空样本集合, F为样本的特征集合, 设$f\subseteq F$为任意的特征子集, $\forall x\in U$, 样本x_i在融合标记重要度和特征权重下由f导出的邻域表示为$\delta^{\mathrm{wei}}_f(x_i)$,那么$f$的邻域熵定义为

$$\mathrm{NH}^{\mathrm{wei}}(f)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_f(x_i)\|}{n}\right)$$

定义 12 设$f_1\in F,f_2\in F$为样本x_i的两组特征, $\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)$表示样本$x_i$在$f_1\cup f_2$的特征空间下的邻域, 则融合标记重要度和特征权重下的联合熵定义为

$$\mathrm{NH}^{\mathrm{wei}}(f_1,f_2)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)\|}{n}\right)$$

当f_1为输入特征f, f_2为类别标记l时, $\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)$改写为$\delta'^{\mathrm{wei}}_{f_1\cup f_2}(x_i)=\delta^{\mathrm{wei}}_f(x_i)\cap l_{x_i},l_{x_i}=\delta_l=\{y|\Delta(x,y)=0,y\in U\}$, 由此, 特征$f$和类别标记$l$的联合熵可定义为

$$\mathrm{NH}^{\mathrm{wei}'}(f,l)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_{c_1}(x_i)\cap l_{x_i}\|}{n}\right)$$

定义 13 设$f_1\in F,f_2\in F$为样本x_i的两组特征, 其中f_1为已知特征, 则f_2在已知特征f_1下的条件熵定义为

$$\mathrm{NH}^{\mathrm{wei}}(f_2|f_1)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)\|}{\|\delta^{\mathrm{wei}}_{f_1}(x_i)\|}\right)$$

由定理 1 可得, $\forall f_1\in F,f_2\in F$, f_1、f_2的加权邻域互信息定义为

$$\mathrm{NMI}^{\mathrm{wei}}(f_1,f_2)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_{f_1}(x_i)\|\cdot\|\delta^{\mathrm{wei}}_{f_2}(x_i)\|}{n\|\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)\|}\right)$$

当f_1为输入特征f, f_2为类别标记l时, $\delta^{\mathrm{wei}}_{f_1\cup f_2}(x_i)$改写为$\delta'^{\mathrm{wei}}_{f_1\cup f_2}(x_i)=\delta^{\mathrm{wei}}_f(x_i)\cap l_x$, $l_{x_i}=\delta_l=\{y|\Delta(x,y)=0,y\in U\}$, 此时, 特征$f$和类别标记$l$的互信息可定义为

$$\mathrm{NMI}^{\mathrm{wei}'}(f,l)=-\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{\|\delta^{\mathrm{wei}}_f(x_i)\|\cdot\|l_{x_i}\|}{n\|\delta^{\mathrm{wei}}_f(x_i)|\cap l_{x_i}\|}\right)$$

2.4 算法描述

给定标记簇L_i, 特征f与标记簇L_i之间的相关性通常使用互信息$\mathrm{NMI}(f,L_i)$来度量, 然而, 当标记数量巨大时$\mathrm{NMI}(f,L_i)$计算复杂度指数增加且很难精确计算。由于同标记簇中的标记彼此高度相关, 使用$\max\limits_{l\in L_i}\{\mathrm{NMI}^{\mathrm{wei}}(f,l)\}$能够有效度量与标记簇$L_i$最相关的候选特征, 并且避免高维信息熵的复杂计算。选择的候选特征要具有对标记簇L_i的分辨信息, 并且已选择的特征没有提供这些分辨信息, 即使选择的候选特征与已选特征间的冗余性尽可能小。条件互信息$\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)$可以用来度量候选特征$f$和已选特征$f_j$冗余性, 若存在已选特征$f_j$提供了$l_{\max}$相关的分辨信息, $\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)$的值会很小, 表示特征$f_j$和特征$f$对于$l_{\max}$是存在冗余的。因此, 使用$\max\limits_{l\in L_i}\{\mathrm{NMI}^{\mathrm{wei}}(f,l)\}$来确保特征与标记的相关性, $\min\limits_{f_j\in S}\{\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)\}$来度量候选特征$f$和已选特征集$S$冗余性, 候选特征评估准则可定义为

$$J=\max_{l\in L_i}\{\mathrm{NMI}^{\mathrm{wei}}(f,l)\}+\min_{f_j\in S}\{\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)$$

多标记数据往往存在类别标记不平衡的问题, 此时与标记$l_{\max}$相关的特征不能很好地描述整个标记簇, 为此, 结合标记的重要度将式(4)中的评估标准改写为

$$J'=\left\{\max_{l\in L_i}\{\mathrm{NMI}^{\mathrm{wei}}(f,l)\}+\min_{f_j\in S}\{\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)\}\right\}*L_{\mathrm{sig}}(l_{\max}) \tag{9}$$

由文献[11,30]得, 式(8)中的条件互信息可由下式近似计算:

$$\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max}|f_j)=\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max})-\frac{\mathrm{NMI}^{\mathrm{wei}}(f,l_{\max})}{\mathrm{NH}^{\mathrm{wei}}(f_j)}\mathrm{NH}^{\mathrm{wei}}(f,f_j) \tag{10}$$

根据式(9)提出基于加权邻域粗糙集的多标记特征选择算法:

算法 1 基于加权信息粒化的多标记特征选择算法(MFWIG)

输入 多标记决策系统$\mathrm{MDT}=<U,F,L>$, 选择特征数N

输出 特征子集S

1)谱聚类将标记集L划分为若干标记簇$L^P=\{L_1,L_2,\cdots,L_t\}$;

2)for each $L_i\in L^P$:

3)根据式(4)计算每个标记$l\in L_i$的重度$L_{\mathrm{sig}}(l)$;

4)根据式(5)计算每个特征$f\in F$的权重$w_m(f)$;

5）$\forall x \in U$，根据式（7）计算样本x的加权分类间隔$m_{L_i}^{\text{wei}}(x)$；

6）$\forall x \in U$，根据式（8）计算样本x的加权邻域$\delta^{\text{wei}}(x)$；

7）初始化$S_i = \varnothing$；

8）while $|S_i| \leqslant N \times \dfrac{|L_i|}{|L|}$:

根据式（9）计算特征得分，选择特征$f_{\max} = \max\limits_{f} J'(f)$；$S_i = S_i \cup \{f_{\max}\}$；$F = F - \{f_{\max}\}$；

end while

end for

9）将每个专属特征子集按基数大小排列得，$S^P = \{S_{p1}, S_{p2}, \cdots, S_{pt}\}$；

10）每次从每个特征子集S_{pi}中依次选择一个新的特征加到S中，直到遍历完所有子集；

11）返回特征子集S。

其中，8）中分别对每个标记簇选择专属特征，主簇中包含的标记更多，其相关的特征在最终选择的特征子集也应该占更大比重。

算法复杂度分析：标记聚类的复杂度是$O(k^3)$，对于每个标记簇，标记重要度的复杂度$O(|L_i|^2)$，特征权重的复杂度是$O(m|L_i|)$，计算加权分类间隔的复杂度是$O(n^2|L_i|)$，特征排序的计算复杂度是$O(m^2)$。由此可知算法的时间复杂度为$O\left(\sum\limits_{i=1}^{t} n^2|L_i| + m^2 + k^3\right)$。

3　实验及结果分析

3.1　实验数据

为了验证本文所提算法的有效性，选取 6 个 Mulan 平台中的真实多标记数据集上进行对比实验。这些数据集涵盖了不同领域的多标记分类问题，包括国家标志、音乐标签、情感分类、鸟类识别等。各数据集的相关信息如表 1 所示。

表 1　实验数据集

Table 1　The datasets used in experiment

数据集	样本数	特征数	类别数
Flags	194	19	7
CAL500	502	68	174
Emotions	593	72	6
Birds	645	260	19
Scene	2317	294	6
Yeast	2417	103	14

3.2　评价指标

本文采用平均精度（average precision, AP）P_{A}、排序损失（ranking loss, RL）L_{R}、汉明损失（hamming loss, HL）L_{H} 和覆盖率（coverage, CV）C_{V} 作为算法性能的评价指标。令测试集为$Z = \{(x_i, Y_i)\}_{i=1}^{m} \subset R^d \times \{+1, -1\}^{|L|}$，根据预测函数$f_l(x)$可定义排序函数$\text{rank}_f(x, l) \in \{1, 2, \cdots, |L|\}$。

AP：表示所有标记的预测集合中，位置排在相关标记前面的标记且仍是相关标记的平均概率，定义为

$$P_{\text{A}} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i|} \times \sum_{l \in R_i}\frac{\{k|\text{rank}_f(x_i,k) \leqslant \text{rank}_f(x_i,l), k \in R_i\}}{\text{rank}_f(x_i,l)}$$

式中：$R_i = \{l|Y_{il} = +1\}$表示与样本x_i相关标记构成的集合。

RL：表示所有样本的不相关标记的排序在其相关标记前面的平均概率，定义为

$$L_{\text{R}} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i|\left|\overline{R_i}\right|} \times |\{(r,k)\left|\text{rank}_f(x_i,l) \geqslant \text{rank}_f(x_i,k), (l,k) \in R_i \times \overline{R_i}\right\}|$$

式中：$\overline{R_i} = \{l|Y_{il} = -1\}$表示与样本$x_i$不相关标记构成的集合。

HL：表示所有样本的预测标记与真实标记的平均差异值，定义为

$$L_{\text{H}} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{L}\sum_{l=1}^{L}|f_l(x_i) \oplus Y_i|$$

式中：$\oplus$表示异或运算，Y_{il}表示与样本x_i相关标记的真实值。

CV：表示所有样本实际包含的标记所需的最大排序距离的平均值，定义为

$$C_{\text{V}} = \frac{1}{m}\sum_{i=1}^{m}\left[\max_{l \in R_i}\text{rank}_f(x_i,l) - 1\right]$$

4 种指标中，AP 取值越大表示算法性能越好，其他指标则取值越小表示算法性能越好。

3.3　结果分析

实验中选择了 5 种不同的对比方法，包括 PPT-CHI、PPT-MI、PMU、SCLS 和 CFSM[26]。采用的多标记分类器为 ML-KNN（multi-label learning based on k-nearest neighbors classifier）[31]，平滑参数s设置为 1，近邻个数k设置为 10，对分类指标采用 10 折交叉验证法。

实验首先对比各种算法诱导出的特定特征子集的分类效果，由于以上 5 种对比算法和本文算法均是对特征进行排序，为确保实验对比的公平性，将所有算法取特征排序的前N个特征作为特征子集，然后分析各算法的分类性能随特征数目的变化情况。其中，为了选择合适比例的特征数量[32]，本实验按以下规则设定选取比例：$m \leqslant 100$，取前40%，$100 < m \leqslant 500$，取前30%。各个数据集最

终所选取的特征子集的特征个数分别为: Flag 8 个, Emotions 28 个, CAL 50 027 个, Birds 78 个, Yeast 30 个, Scene 88 个。表 2~5 给出了各算法分别在 4 种评价指标上的实验结果。

表 2　6 种算法在 AP 上的性能比较

Table 2　Comparison of six algorithms in AP

数据集	PPT-CHI	PPT-MI	PMU	SCLS	CFSM	MFWIG
Flags	0.7864	0.7932	0.8035	0.7990	0.7849	**0.8176**
CAL500	0.4418	0.4418	0.4407	0.4409	0.4407	**0.4422**
Emotions	0.7352	0.7473	0.7067	0.7337	0.7353	**0.7475**
Birds	0.6908	0.6946	0.7061	0.7038	0.6839	**0.7090**
Scene	0.7445	0.7515	0.7541	0.7502	0.6347	**0.7583**
Yeast	0.6888	0.6956	0.7024	0.6966	0.6931	**0.7030**
平均值	0.6813	0.6873	0.6856	0.6874	0.6621	**0.6963**

表 3　6 种算法在 HL 上的性能比较

Table 3　Comparison of six algorithms in HL

数据集	PPT-CHI	PPT-MI	PMU	SCLS	CFSM	MFWIG
Flags	0.2970	0.2977	0.2756	0.2640	0.2768	**0.2593**
CAL500	0.1471	0.1471	0.1480	0.1474	0.1477	**0.1467**
Emotions	0.2005	0.1983	0.2259	0.2057	0.2062	**0.1966**
Birds	0.0489	0.0476	0.0492	0.0481	0.0510	**0.0472**
Scene	0.1132	0.1140	0.1064	**0.0983**	0.1431	0.1116
Yeast	0.2075	0.2090	0.2104	0.2068	**0.2050**	0.2080
平均值	0.1690	0.1690	0.1693	0.1617	0.1716	**0.1616**

表 4　6 种算法在 RL 上的性能比较

Table 4　Comparison of six algorithms in RL

数据集	PPT-CHI	PPT-MI	PMU	SCLS	CFSM	MFWIG
Flags	0.2613	0.2559	0.2403	0.2366	0.2550	**0.2177**
CAL500	0.2237	0.2237	0.2236	0.2236	0.2238	**0.2233**
Emotions	0.2214	0.2138	0.2525	0.2165	0.2190	**0.2080**
Birds	0.1251	0.1217	0.1208	0.1267	0.1253	**0.1187**
Scene	0.1425	0.1405	0.1431	0.1407	0.2217	**0.1361**
Yeast	0.2368	0.2299	0.2232	0.2253	0.2320	**0.2231**
平均值	0.2018	0.1976	0.2006	0.1949	0.2128	**0.1878**

表 5　6 种算法在 CV 上的性能比较

Table 5　Comparison of six algorithms in CV

数据集	PPT-CHI	PPT-MI	PMU	SCLS	CFSM	MFWIG
Flags	5.1789	5.1594	4.9807	4.8782	4.9992	**4.8342**
CAL500	136.3305	136.3305	**136.2089**	136.3225	136.2944	136.3185
Emotions	3.0792	3.0535	3.2107	3.0523	3.0676	**3.0173**
Birds	3.7731	3.7272	3.7233	3.9166	3.8068	**3.6650**
Scene	1.8010	1.7906	1.8015	1.7869	2.1937	**1.7702**
Yeast	8.5388	8.4697	8.3690	8.3745	8.5432	**8.3542**
平均值	26.4503	26.4218	26.3824	26.3885	26.4842	**26.3266**

于给定的评价指标斜体加粗表示个算法中性能最优的结果。根据表 2~5 的结果可得出：

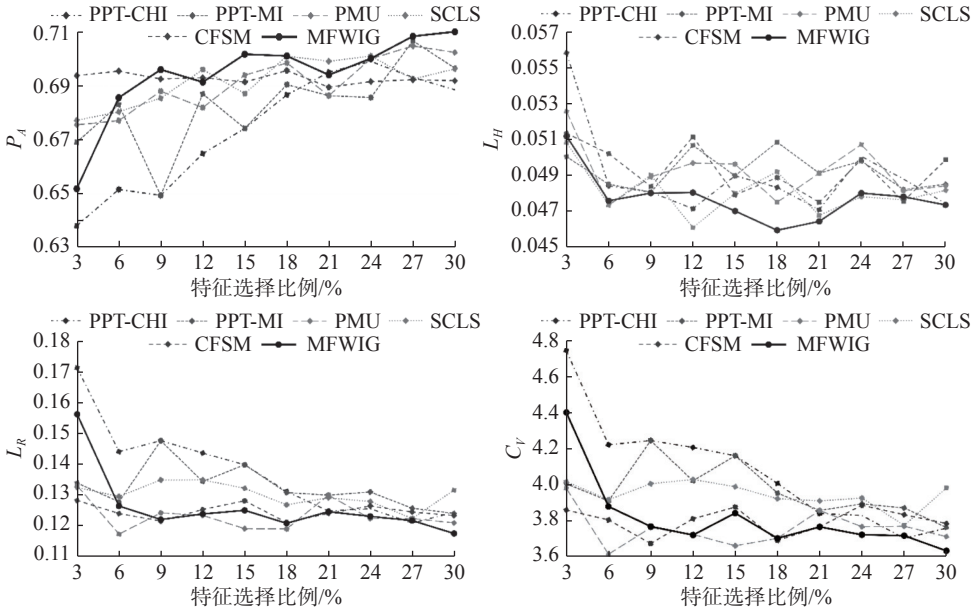
1）对于 AP 指标和 RL 指标，MFWIG 在 6 个数据集上都明显优于其他 5 种算法，即分类性能均取得最优；对于 HL 指标和 CV 指标，MFWIG 分别在 4 个数据集和 5 个数据集上均取得最优值。

2）从 6 个数据集、4 个评价指标的 24 项对比结果可知，PPT-CHI、PPT-MI 和 CFSM 没有取得最优值的情况，PMU、SCLS 和 CFSM 取得最优值的情况均仅占 4.166 7%，而对比 5 种算法 MFWIG 明显优于其他 5 种对比算法，取得最优值的情况占 87.5%。

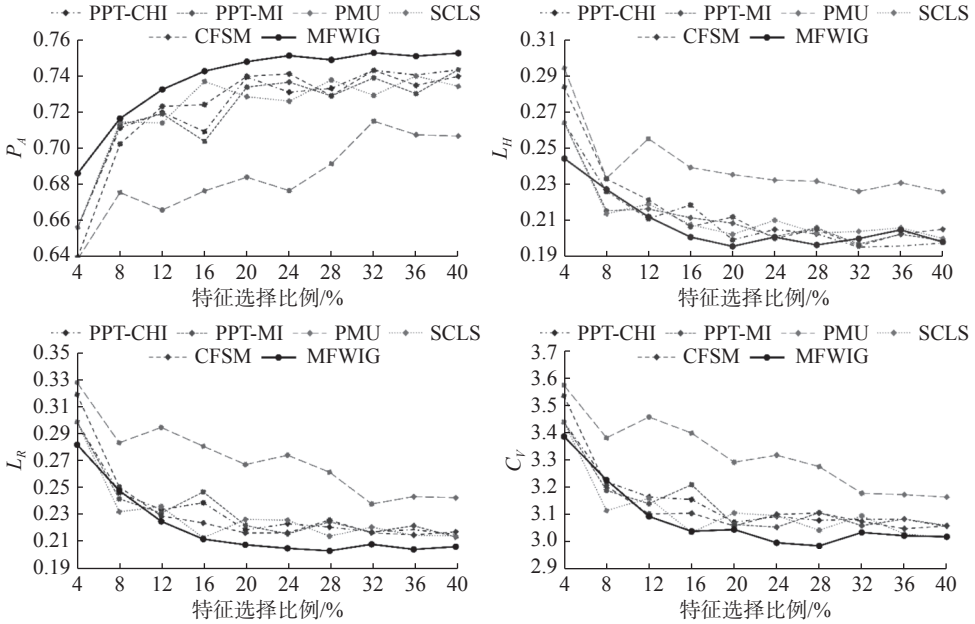
3）从各评价指标平均性能的对比发现，相对 AP 指标、RL 指标、CV 指标和 HL 指标，MFWIG 均取得最优值。

综上，在多数情况下 MFWIG 算法在各评价指标上优于对比算法，其原因在于 MFWIG 考虑了标记间潜在的语义关系和标记的不同重要度，在构造邻域关系时融合了特征权重，使得算法选择的特征能更好地描述标记集合，有效提升分类能力和减少标记预测时的损失，并且 MFWIG 采用条件互信息更好地度量特征间的冗余性。

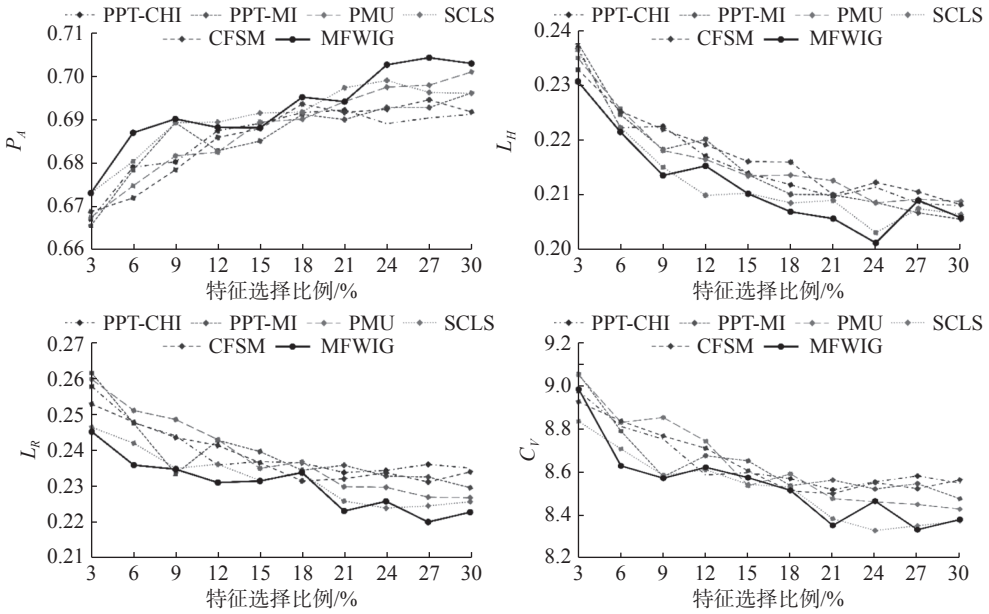
以上实验是根据指定特征数得到最终特征子集来进行静态的对比分析，图 3 为不同数据集上 6 种算法的 4 个评价指标随特征数目的变化趋势。



(a) Birds



(b) Emotions



(c) Yeast

图 3　各算法 4 个评价指标随特征数在 3 个数据集上的变化趋势

Fig. 3　Variation of 4 evaluation indicators of different algorithms with the number of selected features on 3 datasets

从实验结果可知, 选择的特征比例不同算法的性能也不同, 并且比例在 30%~40% 时各算法性能曲线都趋于平稳, 验证了之前实验设定的合理性。此外, MFWIG 在各数据集上的 4 个评价指标均取得较优的性能表现, 尤其在 Emotions 数据集上, 4 个评价指标的性能曲线都明显优于其他算法。最后, MFWIG 在选择特征比例过小时性能相对较差, 这是由于不同标记簇的专属特征间存在交集, 使得最终融合得到的子集数相较于其他算法更小。

4　结束语

针对现实应用场景中每个实例的相关标记的重要程度不同, 并且对于不同标记同一个特征的分类信息也存在差异的问题, 本文提出了一种基于加权信息粒化的多标记特征选择算法。该方法首先利用谱聚类挖掘标记间的相关性来划分标记集, 然后使用标记间的互信息度量标记的重要性, 并在构造邻域关系前融合特征权重, 最后利用标记簇内的高度相似性和条件互信息提出了新的特征选择度量标准, 避免了高维信息熵的计算, 并且构造了一种新的多标记特征选择算法。通过 6 个真实数据集上基于 4 种不同的评价指标的实验对比, 本文提出的算法相较其他对比方法能取得较优的性能表现, 验证了其有效性和可行性。此外, 本文未深入分析算法所得特征子集的内部结构与原特征集间的关系, 下一步将从多标记数据质量的角度来完善此研究。

参考文献:

[1] QIAN Wenbin, HUANG Jintao, WANG Yinglong, et al. Mutual information-based label distribution feature selection for multi-label learning[J]. Knowledge-based systems, 2020, 195(5): 105684.

[2] 高琪, 李德玉, 王素格. 基于模糊不一致对的多标记属性约简 [J]. 智能系统学报, 2020, 15(2): 374–385.
GAO Qi, LI Deyu, WANG Suge. Multi-label attribute reduction based on fuzzy inconsistency pairs[J]. CAAI transactions on intelligent systems, 2020, 15(2): 374–385.

[3] SUN Lin, WANG Tianxiang, DING Weiping, et al. Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification[J]. Information sciences, 2021, 578: 887–912.

[4] LIU Jinghua, LI Yuwen, WENG Wei, et al. Feature selection for multi-label learning with streaming label[J]. Neurocomputing, 2020, 387: 268–278.

[5] BOUTELL M R, LUO Jiebo, SHEN Xipeng, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757–1771.

[6] READ J, PFAHRINGER B, HOLMES G. Multi-label classification using ensembles of pruned sets[C]//2008 Eighth IEEE International Conference on Data Mining. Pisa: IEEE, 2009: 995−1000.

[7] READ J. A pruned problem transformation method for multi-label classification[C]// New Zealand Computer Science Research Student Conference 2008. Christchurch: YUMPU, 2008: 143150: 41.

[8] DOQUIRE G, VERLEYSEN M. Feature selection for multi-label classification problems[C]//International Work-Conference on Artificial Neural Networks. Berlin:

Springer, 2011: 9−16.
[9] LEE J, KIM D W. Feature selection for multi-label classification using multivariate mutual information[J]. Pattern recognition letters, 2013, 34(3): 349–357.
[10] LEE J, KIM D W. Mutual Information-based multi-label feature selection using interaction information[J]. Expert systems with applications, 2015, 42(4): 2013–2025.
[11] LEE J, KIM D W. SCLS: Multi-label feature selection based on scalable criterion for large label set[J]. Pattern recognition, 2017, 66: 342–352.
[12] WANG Yingyao, DAI Jianhua. Label distribution feature selection based on mutual information in fuzzy rough set theory[C]//2019 International Joint Conference on Neural Networks. Budapest: IEEE, 2019: 1−2.
[13] DAI Jianhua, CHEN Jiaolong, LIU Ye, et al. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation[J]. Knowledge-based systems, 2020, 207: 106342.
[14] QIAN Wenbin, HUANG Jintao, WANG Yinglong, et al. Label distribution feature selection for multi-label classification with rough set[J]. International journal of approximate reasoning, 2021, 128: 32–55.
[15] WANG Chenxi, LIN Yaojin, LIU Jinghua. Feature selection for multi-label learning with missing labels[J]. Applied intelligence, 2019, 49(8): 3027–3042.
[16] ZHANG Ping, GAO Wanfu. Feature relevance term variation for multi-label feature selection[J]. Applied intelligence, 2021, 51(7): 5095–5110.
[17] ZHANG Ping, LIU Guixia, GAO Wanfu, et al. Multi-label feature selection considering label supplementation[J]. Pattern recognition, 2021, 120: 108137.
[18] QIAN Wenbin, LONG Xuandong, WANG Yinglong, et al. Multi-label feature selection based on label distribution and feature complementarity[J]. Applied soft computing, 2020, 90: 106167.
[19] HU Qinghua, YU Daren, LIU Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information sciences, 2008, 178(18): 3577–3594.
[20] LIU Keyu, LI Tianrui, YANG Xibei, et al. Granular cabin: an efficient solution to neighborhood learning in big data[J]. Information sciences, 2022, 583: 189–201.
[21] CHEN Yan, LIU Keyu, SONG Jingjing, et al. Attribute group for attribute reduction[J]. Information sciences, 2020, 535: 64–80.
[22] HU Meng, TSANG E C C, GUO Yanting, et al. A novel approach to attribute reduction based on weighted neighborhood rough sets[J]. Knowledge-based systems, 2021, 220: 106908.
[23] JIANG Zehua, DOU Huili, SONG Jingjing, et al. Data-guided multi-granularity selector for attribute reduction[J]. Applied intelligence, 2021, 51(2): 876–888.
[24] 段洁, 胡清华, 张灵均, 等. 基于邻域粗糙集的多标记分类特征选择算法 [J]. 计算机研究与发展, 2015, 52(1): 56–65.
DUAN Jie, HU Qinghua, ZHANG Lingjun, et al. Feature selection for multi-label classification based on neighborhood rough sets[J]. Journal of computer research and development, 2015, 52(1): 56–65.
[25] LIN Yaojin, HU Qinghua, LIU Jinghua, et al. Multi-label feature selection based on neighborhood mutual information[J]. Applied soft computing, 2016, 38: 244–256.
[26] LONG Xuandong, QIAN Wenbin, WANG Yinglong, et al. Cost-sensitive feature selection on multi-label data via neighborhood granularity and label enhancement[J]. Applied intelligence, 2021, 51(4): 2210–2232.
[27] 黄锦涛, 钱文彬, 王映龙. 基于标记增强的多标记代价敏感特征选择算法 [J]. 小型微型计算机系统, 2020, 41(4): 685–691.
HUANG Jintao, QIAN Wenbin, WANG Yinglong. Multi-label cost-sensitive feature selection algorithm based on label enhancement[J]. Journal of Chinese computer systems, 2020, 41(4): 685–691.
[28] GONZALEZ-LOPEZ J, VENTURA S, CANO A. Distributed multi-label feature selection using individual mutual information measures[J]. Knowledge-based systems, 2020, 188: 105052.
[29] SUN Zhenqiang, ZHANG Jia, DAI Liang, et al. Mutual information based multi-label feature selection via constrained convex optimization[J]. Neurocomputing, 2019, 329: 447–456.
[30] KWAK N, CHOI C H. Input feature selection for classification problems[J]. IEEE transactions on neural networks, 2002, 13(1): 143–159.
[31] ZHANG Minling, ZHOU Zhihua. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038–2048.
[32] KASHEF S, NEZAMABADI-POUR H, NIKPOUR B. Multilabel feature selection: a comprehensive review and guiding experiments[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2018, 8(2): e1240.

作者简介：

胡军，教授，博士，主要研究方向为多粒度认知计算、人工智能安全和图分析与挖掘，近年来主持参与国家重点研发计划、国家自然科学基金、重庆市自然科学基金等科研项目 10 多项，授权国家发明专利 5 项，发表科学研究论文 60 多篇，出版专著 3 部。

王海峰，硕士研究生，主要研究方向为粒计算、粗糙集。