基于稀疏系数矩阵重构的多标记特征选择

李永豪 胡 亮 高万夫

(吉林大学计算机科学与技术学院 长春 130012)

(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 处理复杂的多标记数据对于特征选择而言是一项挑战性任务.然而,现存的多标记特征选择方法存在三个问题未解决.首先,现有的多标记特征选择方法利用样例层流形正则化项保持样例的相似性结构或借助标签关联来指导特征选择,但两者对于特征选择的指导存在互补关系.其次,早期方法基于样例相似性所构造的近邻矩阵来探索标签关联,却忽略了成对标签本身的关联性.最后,早期方法整合多个未知变量,导致目标函数的求解变得困难.为解决上述问题,本文基于最小二乘回归模型构建经验损失函数,然后在目标函数中引入标签正则化项探索标签之间的关联,同时利用特征矩阵与重构稀疏系数矩阵的乘积表示预测标签并保留数据本身的局部几何结构.上述各项被整合在一个联合学习框架内.针对该学习框架,一套证明可收敛的优化方案被设计.在 13 个真实的多标记基准数据集上进行实验,实验结果验证了所提方法的有效性.

关键词 特征选择;多标记学习;流形学习;稀疏化学习;分类

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.001827

Multi-Label Feature Selection Based on Sparse Coefficient Matrix Reconstruction

LI Yong-Hao HU Liang GAO Wan-Fu

[1)] (*College of Computer Science and Technology, Jilin University, Changchun* 130012)

[2)] (*Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun* 130012)

Abstract Dealing with complicated multi-label data is a challenging task for feature selection in practical applications. However, there exist three unsolved issues in the existing multi-label feature selection methods. First, previous multi-label feature selection methods either employ instance-level manifold regularization terms to maintain the instance similarity or exploit the correlations among labels to guide feature selection process, however, both two are complementary to each other in feature selection process. Second, existing methods explore label correlations based on the affinity matrix of instance similarity, ignoring the pairwise label correlations. Third, previous methods involve several unknown variables, which makes the solution of the objective function difficult. To tackle the issues mentioned above, an empirical loss function model is constructed based on the least square regression model. And then, we introduce the label regularization term to exploit label correlations, meanwhile employing the product of feature matrix and weight coefficient matrix to represent predicted labels so that the local geometric structure of data set is stored. Finally, we integrate the terms mentioned above into one joint learning framework. An effective optimization method with provable convergence is designed to solve our proposed method. In summary, the novelties and main contributions of this paper can

收稿日期:2021-05-05;在线发布日期:2022-01-24.本课题得到国家重点研发专项(2017YFA0604500)、吉林省重点科技研发项目(20180201103GX)、吉林省科技厅联合基金项目(2020122209JC)资助.**李永豪**,博士研究生,主要研究方向为多标记特征选择.E-mail:yonghao17@mails.jlu.edu.cn.**胡 亮**,博士,教授,主要研究领域为机器学习、信息安全等.**高万夫**(通信作者),博士,副教授,主要研究方向为特征选择、多标记学习、信息论、因果关系.E-mail:gaowf@jlu.edu.cn.

be summarized as follows: the proposed method uses the instance-level manifold regularization term to maintain the instance similarity. At the same time, the proposed method introduces label-level manifold regularization term to exploit the label correlations. Moreover, the proposed method can store the geometric structure of labels in the weight coefficient matrix and employ the weight coefficient matrix to guide feature selection process, because the sparse coefficient matrix can maintain the geometric relationship between the data space and label space, as well as the relationship between labels, the proposed method can obtain superior classification ability on the test data set by using the sparse coefficient matrix that is learned by the training process. Furthermore, the proposed method introduces the $L_{2,1}$-norm that integrates the advantages of L_1-norm and L_2-norm to select important features in each iteration. Finally, the proposed method integrates all the above terms into one joint learning framework and develops a method to solve the constrained problem, i. e., regulating regression coefficient matrix based on instance-similarity and label-similarity for multi-label feature selection that is named as RMLFS, while an optimal scheme is designed. In addition, we can obtain a globally optimal solution by this learning framework because the objective function only incorporates one unknown variable unlike other existing methods that incorporate multiple unknown variables that lead to the local optimal solution in most cases, and the objective function is a convex function. This method conducts multiple evaluation criteria on thirteen benchmark data sets to show the superiority of the proposed multi-label feature selection method. In order to verify the classification superiority of the proposed method, numerous experiments are conducted on thirteen different multi-label data sets. Eight competitive methods including MIFS, MDMR, SCLS, LRFS, mRMR, RALM-FS, TRCFS and GMM are compared to the proposed method. The extensive experimental results show that the classification performance of the proposed RMLFS outperforms other compared methods in these experiments.

Keywords feature selection; multi-label learning; manifold learning; sparse learning; classification

1 引 言

高维数据的复杂性给机器学习算法带来巨大的挑战,它不仅增加了机器学习算法的计算成本,而且导致这些模型过拟合.特征选择技术可以约减冗余特征数目、改善模型分类性能[1-7],因此特征选择已经变成了处理高维数据至关重要的一项技术,并且该技术已经被广泛应用于多个领域,例如图像和视频标注[8-10]、文本挖掘、蛋白质功能预测等[11-13].

通常,特征选择方法包括过滤式模型[6]、包裹式模型[14-15]以及嵌入式模型[16-17].不同于过滤式和包裹式两类模型[18],嵌入式模型融合特征选择和学习算法在统一框架内,以至于嵌入式模型可以借助学习算法来进一步改善模型的性能.鉴于此,文章聚焦于嵌入式特征选择模型.然而对多标记数据进行特征选择处理仍然存在一些困难,原因在于多标记数据不仅涉及高维的特征空间,也涉及大量相互依存的标签.为有效处理此类多标记数据,学者提出算法适应性策略来直接处理多标记数据,取代直接将多标记数据转化成单标签数据再处理的方案[19-22].

早期的多标记特征选择方法利用样例层正则化项来保持样例的相似性结构或借助标签关联来指导特征选择.然而在这些多标记特征选择算法中依然存在三个问题未被解决.第一,尽管样例层正则化和标签之间的关联对于特征选择具有重要的指导作用,前期的方法仅关注两者中的一种,而两者对于特征选择的指导存在互补关系;第二,以往的多标记特征选择方法大多是基于样例相似性来构造近邻矩阵,借此来探索标签关联,却忽略了成对标签间固有的关联性;第三,早期多标记特征选择的目标函数包含多个未知变量,而多个变量的存在使得要优化的目标函数易于陷入局部最优解困境.

为了处理上述存在的问题,本文在利用最小二乘回归模型的基础上最小化经验损失函数,然后引入基于标签相似性的近邻矩阵来构造标签层正则化项,以便于挖掘标签相关性[23].与此同时,利用特征矩阵与重构稀疏系数矩阵的乘积预测标签来拟合现

存的标签，并通过这两个矩阵的乘积和流形正则化项将局部几何结构存储于重构稀疏系数矩阵中. 此外，我们也引入样例层流形正则化项保持样例相似性[24]. 最后，上述各项被整合到一个联合学习框架中，该框架被称为可调多标记特征选择(RMLFS).

综上，本文的主要贡献概括如下：

(1) 同时利用样例层正则化项和标签层正则化项，借助两者的互补关系来探索样例相似性和标签之间的关联.

(2) 将重构标签的几何结构存储于稀疏系数矩阵中，借此来指导特征选择，同时引入 $L_{2,1}$ 范数在每轮迭代中选择最重要的特征.

(3) 整合上述各项进入一个联合学习框架，并设计了一套被证明可收敛的优化方案来优化该学习框架.

(4) 目标函数被设计为只包含一个未知变量的凸函数，利用该函数可以得到全局最优解，并在 13 个基准数据集上利用不同度量准则对目标函数进行评估来验证所提方法的优越性.

本文第 2 节对相关工作和基础知识进行详细描述；第 3 节对目标函数和优化方案的设计过程展开详细介绍；第 4 节描述并分析对比算法在 13 个真实数据集上的实验结果，以验证提出方法在分类性能方面的优越性；最后第 5 节对全文进行总结并指明后续的研究方向.

2　相关工作

首先，表 1 对文章采用的主要符号进行描述，然后多标记学习、特征选择基础知识、代表性算法以及流形学习相关概念被总结.

表 1　符号描述

符号	描述
$\boldsymbol{A}$	任意矩阵
$\boldsymbol{a}$	任意向量
a	任意标量
$\boldsymbol{A}_{i\cdot}$	$\boldsymbol{A}$ 的第 i 行
$\boldsymbol{A}_{\cdot j}$	$\boldsymbol{A}$ 的第 j 列
$\boldsymbol{A}_{ij}$	$\boldsymbol{A}$ 的第 i 行第 j 列元素
n	矩阵 $\boldsymbol{A}$ 的行数
m	矩阵 $\boldsymbol{A}$ 的列数
$\boldsymbol{A}^{\mathrm{T}}$	矩阵 $\boldsymbol{A}$ 的转置
$\mathrm{Tr}(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的迹，其中 $\boldsymbol{A}$ 是方阵
$\Vert\boldsymbol{A}\Vert_F$	$\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{A}_{ij}^2}$
$\Vert\boldsymbol{A}\Vert_{2,1}$	$\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}\boldsymbol{A}_{ij}^2}$
$\boldsymbol{X}$	表示特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$
$\boldsymbol{Y}$	表示标签矩阵 $\boldsymbol{Y}\in\mathbb{R}^{n\times c}$

在多标记学习中，每一个样例包含多个不同的类标签，同时这些类标签之间也存在不同程度的关联，即标签之间有强相关，也有弱相关，甚至不相关. 因此当大量多标记问题需要被处理时，标签关联的有效探索被认为是至关重要的. 基于标签关联关系，多标记学习通常包括一阶关联、二阶关联以及高阶关联(即多于两个以上的标签之间的关联)这三种情况[25]. 在第一种情况下，标签与标签之间是相互独立的，即标签与标签之间不存在联系，利用这一思想的代表性算法是二元关联(Binary Relevance，BR)[19]. 在第二种情况中，代表性算法包括联合特征选择与分类(Joint Feature Selection and Classification，JFSC)和学习特定于标签的特征(Learning Label Specific Features，LLSF)[26-27]，即成对标签之间的关系被考虑，以至于多标记数据被转化成多个成对标签关联问题. 在高阶关联情况下，代表性方法包括分类器链(Classifier Chain，CC)和公共与类属特征选择(Common and Label-specific Feature Selection，CLFS)[28-29]，即多于两个标签之间的关联被考虑.

随着多标记数据的快速增长，多标记特征选择吸引了越来越多研究人员的注意. 而且多标记数据的复杂性必然使特征选择任务比处理单标签数据任务时更复杂. 通过广泛调研相关文献[18,25,30-33]，一种直接策略就是将多标记数据转化成多个独立的单标签数据，然后利用单标签特征选择算法来处理这些被转化的单标签数据. 经典代表性算法如最小冗余最大相关(minimal Redundancy Maximal Relevance，mRMR)[7]是一种基于信息理论的单标签特征选择方法，而且 mRMR 可以从特征全集中得到最优特征子集. 其中 mRMR 目标函数如下：

$$I(x_k)=I(x_k;y)-\frac{1}{|S|}\sum_{x_k\in S}I(x_j;x_k) \tag{1}$$

x_k 和 y 分别表示一个候选特征和一个类标签，S 表示已经选择的特征子集，$|S|$ 表示 S 包含的特征个数，x_j 表示一个已选特征.

此外研究者还提出了多种多标记特征选择方法来直接处理多标记数据，这类方法统称算法适应性方法. 接下来，我们回顾了几种代表性的多标记特征选择方法. 这些方法采用不同的准则，例如基于互信息准则和基于稀疏学习准则的方法.

Lee 和 Kim[34]提出了一种大标签集的可伸缩准则法(Scalable Criterion for Large Label Set，SCLS). SCLS 采用可扩展的相关性评价标准来评价条件相关性，便于进行特征选择. 另外 Lin 等人[35]提出了一种基于最大依赖和最小冗余(Max-Dependency

and Min-Redundancy,MDMR)的方法,即该方法同时考虑特征依赖和特征冗余.

此外,Jian 等人[36]提出一种基于稀疏化的多标记特征选择方法,该方法被命名为多标记信息特征选择(MIFS).MIFS 方法利用标签空间的低秩潜在表示的子空间来削弱无关特征对模型造成的负面影响.而且 MIFS 通过流形正则化项来保持原始标签空间和低秩潜在表示的子空间的局部几何结构一致性.MIFS 具有如下形式的目标函数:

$$\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{B}}\{\|\boldsymbol{XW}-\boldsymbol{V}\|_F^2+\alpha\|\boldsymbol{Y}-\boldsymbol{VB}\|_F^2+\beta\mathrm{Tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{LV})+\gamma\|\boldsymbol{W}\|_{2,1}\} \tag{2}$$

其中 $\boldsymbol{W}\in\mathbb{R}^{d\times c}$ 表示特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$ 的特征选择矩阵,$\boldsymbol{Y}\in\mathbb{R}^{n\times c}$ 表示标签矩阵.$\boldsymbol{V}\in\mathbb{R}^{n\times k}$ 和 $\boldsymbol{B}\in\mathbb{R}^{k\times c}$ 分别表示潜在语义矩阵和其系数矩阵.α,β 和 γ 表示相关项的三个正则化参数,而 k 表示秩数.Cai 等人[37]也提出了一种基于稀疏学习的特征选择方法,被称为 RALM-FS.RALM-FS 方法认为 $L_{2,0}$ 范数比 $L_{2,1}$ 范数具有更稀疏的解集.因此为有效地利用 $L_{2,0}$ 范数,RALM-FS 方法引入了增广拉格朗日乘子,从而形成如下的函数:

$$\begin{aligned}&\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{B}}\|\boldsymbol{Y}-\boldsymbol{XW}-\boldsymbol{1}\boldsymbol{b}^{\mathrm{T}}\|_{2,1}\\&\text{s.t. }\|\boldsymbol{W}\|_{2,0}=a\end{aligned} \tag{3}$$

其中 $\boldsymbol{1}$ 表示数值全为 1 的列向量,$\boldsymbol{b}$ 表示偏差列向量,a 表示选择的特征数,$\|\boldsymbol{W}\|_{2,0}=\sum_{i=1}^{d}\left\|\sum_{j=1}^{c}w_{ij}^2\right\|_0$.另外,Liu 等人[38]提出了一种新的对噪声不敏感的基于迹率比准则的特征选择方法(Trace Ratio Criterion for Feature Selection,TRCFS).TRCFS 方法解决迹率比准则倾向于选择方差较小的特征的问题.同时特征选择技术被扩展到一些新的研究领域[39-41],例如多视图学习中,Zhang 等人[42]在 MIFS 的基础上通过自适应分配权重将多个视图的拉普拉斯矩阵融合,从而设计多视图特征选择方法.在类属特征学习中,Guo 等人[43]利用标签特定的鉴别映射特性进行多标记学习.在集成学习中,Guo 等人[44]设计的集成特征选择方法能明确地找出标签之间的相关性,而且可以通过多标记分类器和评价指标充分利用标签关联等.

另外,近些年流形学习被广泛地研究[45-46],流形正则化的目的是在保持高维数据的几何结构的同时,将高维数据投影到一个低维的数据空间,从而在保持高维数据的几何结构的同时有效实现维数约简.而且样例层流形正则化项的基本思想是如果两个样例相似性越高,那么由它们所获得的低维标签空间中相应的样例标签也应该是相似的.该正则化项被广泛应用在各种模型学习中[47-48].本文中标签层流形正则化项被引入工作方案内.

3 关键实现技术

3.1 设计方法

在这一节,我们提出了一种新的多标记特征选择方法.首先,我们学习从特征空间到标签空间的映射.如下优化问题被获得:

$$\min_{\boldsymbol{W}}\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2 \tag{4}$$

其中 $\boldsymbol{W}\in\mathbb{R}^{d\times c}$ 表示特征矩阵 $\boldsymbol{X}$ 的系数矩阵.该系数矩阵 $\boldsymbol{W}$ 可以度量矩阵 $\boldsymbol{X}$ 中每一个特征的贡献程度.本文利用最小二乘回归模型,通过不断迭代特征选择矩阵 $\boldsymbol{W}$,以此来最小化特征矩阵和标签矩阵之间的误差.其次根据样例层流形的基本思想[49]为利用样例层流形的几何结构来挖掘样例间的相似性.因此,如下形式的正则化项被构建:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{S}_{ij}^{\boldsymbol{X}}\|(\boldsymbol{XW})_{i\cdot}-(\boldsymbol{XW})_{j\cdot}\|_2^2\\&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{S}_{ij}^{\boldsymbol{X}}((\boldsymbol{XW})_{i\cdot}-(\boldsymbol{XW})_{j\cdot})((\boldsymbol{XW})_{i\cdot}-(\boldsymbol{XW})_{j\cdot})^{\mathrm{T}}\\&=\mathrm{Tr}((\boldsymbol{XW})^{\mathrm{T}}(\boldsymbol{A}^{\boldsymbol{X}}-\boldsymbol{S}^{\boldsymbol{X}})(\boldsymbol{XW}))\\&=\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})\end{aligned} \tag{5}$$

其中 $\boldsymbol{L}^{\boldsymbol{X}}=\boldsymbol{A}^{\boldsymbol{X}}-\boldsymbol{S}^{\boldsymbol{X}}$ 被称为关于特征矩阵 $\boldsymbol{X}$ 的图拉普拉斯矩阵.$\boldsymbol{S}^{\boldsymbol{X}}$ 表示近邻矩阵,$\boldsymbol{A}^{\boldsymbol{X}}$ 表示对角矩阵.而流形学习的成功依赖于高质量的近邻矩阵[23].其中近邻矩阵通常可以通过如下几种方式来构建[48]:

通过布尔值所获得的近邻矩阵:

$$\boldsymbol{S}_{ij}^{\boldsymbol{X}}=\begin{cases}1, & \text{若}\boldsymbol{X}_{i\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{j\cdot})\\ & \text{或}\boldsymbol{X}_{j\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{i\cdot})\\0, & \text{其他}\end{cases} \tag{6}$$

通过余弦距离计算两个向量之间的相似关系所获得的近邻矩阵:

$$\boldsymbol{S}_{ij}^{\boldsymbol{X}}=\begin{cases}\dfrac{\boldsymbol{X}_{i\cdot}^{\mathrm{T}}\boldsymbol{X}_{j\cdot}}{\|\boldsymbol{X}_{i\cdot}\|_2\|\boldsymbol{X}_{j\cdot}\|_2}, & \text{若}\boldsymbol{X}_{i\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{j\cdot})\\ & \text{或}\boldsymbol{X}_{j\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{i\cdot})\\0, & \text{其他}\end{cases} \tag{7}$$

通过热核函数①构建一个最近邻图[50],即热核函数通过确定点与点之间的权重大小,如果点 i 和点 j 相连,那么它们关系的权重通过如下公式获得:

$$\boldsymbol{S}_{ij}^{\boldsymbol{X}}=\begin{cases}e^{-\frac{\|\boldsymbol{x}_{i\cdot}-\boldsymbol{x}_{j\cdot}\|_2^2}{\sigma}}, & \text{若}\boldsymbol{X}_{i\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{j\cdot})\\ & \text{或}\boldsymbol{X}_{j\cdot}\in(\mathcal{N})_p(\boldsymbol{X}_{i\cdot})\\0, & \text{其他}\end{cases} \tag{8}$$

① 热核函数最早源于热力学的热传导方程,可以概括一个点周围的几何信息,从而起到探测模型表面几何变化的功能[49].

其中$(\mathcal{N})_p(\boldsymbol{X}_j.)$表示样例$\boldsymbol{X}_j.$的$p$个最近邻节点.$\sigma$表示热核函数的一个调节参数.

基于上述信息,式(4)和(5)被整合重构为如式(9):

$$\min_{\boldsymbol{W}}\{\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})\} \tag{9}$$

其中α是一个正则化参数,用来调控样例层流形正则化项对损失函数的贡献度.为了从标签层角度利用标签之间的相似性关系,我们采用如下形式的标签层正则化项:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}\boldsymbol{S}_{ij}^{\boldsymbol{Y}}\|(\boldsymbol{XW})_{\cdot i}-(\boldsymbol{XW})_{\cdot j}\|_2^2\\&=\frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}\boldsymbol{S}_{ij}^{\boldsymbol{Y}}((\boldsymbol{XW})_{\cdot i}-(\boldsymbol{XW})_{\cdot j})((\boldsymbol{XW})_{\cdot i}-(\boldsymbol{XW})_{\cdot j})^{\mathrm{T}}\\&=\mathrm{Tr}((\boldsymbol{XW})(\boldsymbol{A}^{\boldsymbol{Y}}-\boldsymbol{S}^{\boldsymbol{Y}})(\boldsymbol{XW})^{\mathrm{T}})\\&=\mathrm{Tr}(\boldsymbol{XWL}^{\boldsymbol{Y}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})\end{aligned} \tag{10}$$

其中$\boldsymbol{L}^{\boldsymbol{Y}}=\boldsymbol{A}^{\boldsymbol{Y}}-\boldsymbol{S}^{\boldsymbol{Y}}$被称为关于矩阵$\boldsymbol{Y}$的图拉普拉斯矩阵.$\boldsymbol{S}^{\boldsymbol{Y}}$表示近邻矩阵,$\boldsymbol{A}^{\boldsymbol{Y}}$表示对角矩阵.因此,如下函数被获得:

$$\begin{aligned}\min_{\boldsymbol{W}}\{&\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})+\\&\beta\mathrm{Tr}(\boldsymbol{XWL}^{\boldsymbol{Y}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})\}\end{aligned} \tag{11}$$

其中β是用来控制标签层正则化项贡献度的正则化参数.$\boldsymbol{W}$可以被用来度量每一个特征的贡献度,也就是说,$\|\boldsymbol{W}_{i}.\|_2$的值越大表示第$i$个特征越重要.并且目标函数中引入$L_{2,1}$范数正则化项来保证$\boldsymbol{W}$的行稀疏性,由于$L_{2,1}$范数应用于$\boldsymbol{W}$,使得目标函数具有强鲁棒性[11].最终的目标函数被构造如下:

$$\begin{aligned}\min_{\boldsymbol{W}}\{&\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})+\\&\beta\mathrm{Tr}(\boldsymbol{XWL}^{\boldsymbol{Y}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})+\gamma\|\boldsymbol{W}\|_{2,1}\}\end{aligned} \tag{12}$$

其中γ是用来调整目标函数稀疏度的正则化参数.接下来的小节设计了一种适用于目标函数(12)的优化方案.这一方案可以确保RMLFS的收敛性,在下一小节我们具体描述了该优化方案.提出算法借助最小二乘回归模型可以拟合特征矩阵和标签集的映射关系,同时联合目标函数各项通过迭代更新规则来学习系数矩阵$\boldsymbol{W}$.由于$\boldsymbol{W}$可以保持数据与标签以及标签之间的几何结构关系,因此利用训练所学习到的稀疏系数矩阵$\boldsymbol{W}$,算法可以在测试数据上得到优越的分类能力,而分类能力通过后续的实验加以验证.

3.2　优化方案

本节观察到目标函数(12)关于变量$\boldsymbol{W}$是凸的,因此本文可以得到函数(12)的全局最优解.另外,由于目标函数包含$L_{2,1}$范数项而难于直接求解,因此目标函数是非光滑的.为了解决这个问题,我们引入了一种松弛化的方法[11].对目标函数中的$L_{2,1}$范数项求关于变量$\boldsymbol{W}$的导数,可以获得如下公式:

$$\frac{\partial\|\boldsymbol{W}\|_{2,1}}{\partial\boldsymbol{W}}=2\boldsymbol{DW} \tag{13}$$

其中$\boldsymbol{D}\in\mathbb{R}^{d\times d}$表示一个对角矩阵,其第$(i,i)$个对角元素可通过如下公式进行计算:

$$\boldsymbol{D}_{ii}=\frac{1}{2\|\boldsymbol{W}_{i}.\|_2+\epsilon},\ \epsilon\to 0 \tag{14}$$

其中ϵ表示一个正的极小常数.它可以阻止非负性问题导致的负面干扰,以至于$\|\boldsymbol{W}\|_{2,1}$可以被松弛化为$2\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{DW})$.因此目标函数(12)可以被重构成如下形式:

$$\begin{aligned}\min_{\boldsymbol{W}}\{&\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})+\\&\beta\mathrm{Tr}(\boldsymbol{XWL}^{\boldsymbol{Y}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})+2\gamma\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{DW})\}\end{aligned} \tag{15}$$

进一步,非负约束条件$(\boldsymbol{W}\geqslant 0)$被整合到函数(15)中,从而一个有效的优化方案被获得.因此目标函数变成如下的拉格朗日函数形式:

$$\begin{aligned}\mathcal{L}(\boldsymbol{W})=&\|\boldsymbol{XW}-\boldsymbol{Y}\|_F^2+\alpha\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW})+\\&\beta\mathrm{Tr}(\boldsymbol{XWL}^{\boldsymbol{Y}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})+2\gamma\mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{DW})-\\&\mathrm{Tr}(\boldsymbol{\Psi W}^{\mathrm{T}})\end{aligned} \tag{16}$$

其中$\boldsymbol{\Psi}\geqslant 0$是一个拉格朗日乘子.为了更新$\boldsymbol{W}$,函数(17)关于$\boldsymbol{W}$的导数被获得:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}}=&2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XW}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+2\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW}+\\&2\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWL}^{\boldsymbol{Y}}+2\gamma\boldsymbol{DW}-\boldsymbol{\Psi}\end{aligned} \tag{17}$$

然后应用KKT条件$\boldsymbol{\Psi}_{ij}\boldsymbol{W}_{ij}=0$得到如下关于$\boldsymbol{W}$的公式:

$$\begin{aligned}&(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XW}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+2\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\boldsymbol{X}}\boldsymbol{XW}+\\&2\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWL}^{\boldsymbol{Y}}+2\gamma\boldsymbol{DW})\circ\boldsymbol{W}=0\end{aligned} \tag{18}$$

其中$\circ$表示哈达玛积.而且$\boldsymbol{L}^{\boldsymbol{X}}=\boldsymbol{A}^{\boldsymbol{X}}-\boldsymbol{S}^{\boldsymbol{X}}$和$\boldsymbol{L}^{\boldsymbol{Y}}=\boldsymbol{A}^{\boldsymbol{Y}}-\boldsymbol{S}^{\boldsymbol{Y}}$.因此我们获得如下形式的更新公式:

$$\boldsymbol{W}_{ij}^{t+1}\leftarrow\boldsymbol{W}_{ij}^{t}\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{\boldsymbol{X}}\boldsymbol{XW}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWS}^{\boldsymbol{Y}})_{ij}}{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XW}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}^{\boldsymbol{X}}\boldsymbol{XW}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XWA}^{\boldsymbol{Y}}+\gamma\boldsymbol{DW})_{ij}} \tag{19}$$

其中t表示当前的迭代次数.然后我们可以通过所提出的特征选择算法RMLFS来获得最优的l个特征组成的特征子集.RMLFS的伪代码在算法1中被描述.另外我们根据如下收敛性定理可证明该方案的收敛性.

收敛性定理1.　式(12)的目标函数值单调减小,直到算法1收敛.

证明.　相关证明详见附录部分.

《计算机学报》

算法 1. RMLFS.

输入：特征矩阵 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$，标签矩阵 $\boldsymbol{Y}\in\mathbb{R}^{n\times c}$；

正则化超参数 α,β,γ

输出：返回排序后的前 l 个特征

1. 随机初始化 $\boldsymbol{W}\in\mathbb{R}_{+}^{d\times c}$；
2. $t=0$；
3. 计算 $\boldsymbol{A}^{\boldsymbol{X}},\boldsymbol{S}^{\boldsymbol{X}},\boldsymbol{A}^{\boldsymbol{Y}}$ 和 $\boldsymbol{S}^{\boldsymbol{Y}}$；
4. 重复迭代：
5. 更新 $\boldsymbol{D}_{ii}=\dfrac{1}{2\|\boldsymbol{W}_{i\cdot}\|_2+\epsilon}$；
6. 更新：

$$\boldsymbol{W}_{ij}^{t+1}\leftarrow\boldsymbol{W}_{ij}^{t}\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{W}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{S}^{\boldsymbol{Y}})_{ij}}{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}^{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{W}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{A}^{\boldsymbol{Y}}+\gamma\boldsymbol{D}\boldsymbol{W})_{ij}}$$

7. $t=t+1$；
8. 直至满足收敛条件：$\dfrac{|\mathcal{L}^{t}-\mathcal{L}^{t-1}|}{\mathcal{L}^{t}}\leqslant 10^{-3}$；
9. Return 矩阵 $\boldsymbol{W}$；
10. 通过 $\|\boldsymbol{W}_{i\cdot}\|_2$，$i=1,2,\cdots,d$，选前 l 个特征.

4 实验结果及分析

本节使用 13 个真实的多标记基准数据集和 8 个对比算法验证 RMLFS 算法有效性. 所有实验均在 3.4GHz 的英特尔酷睿 i7-6700 且运行内存为 16GB 的计算机设备上进行.

4.1 数据集描述及实验设置

本文从 Mulan 数据库[51]中获取的 13 个多标准数据集已经被广泛使用在多标记学习方法中[3,23,46,52-53]. 我们使用这 13 个基准数据集来验证 RMLFS 的有效性. 例如，Flags 数据集属于图像域，该数据集包含 194 个样例和 7 个类标签，如红色、绿色和蓝色等；Science 数据集属于文本域，该数据集包含 5000 个样例，每个样例表示 743 个特征和 40 个标签. Genbase 数据集属于生物学领域，包含 662 个蛋白质，其中一个蛋白质具有 1185 个特征. 此外，还有 27 个类别标签，如受体和氧还原酶等. 关于这些数据集的详细信息见表 2.

表 2 实验数据集描述

编号	数据名称	样例数	特征数	标签数
1	Flags	194	19	7
2	Arts	5000	462	26
3	Education	5000	550	33
4	Entertain	5000	640	21
5	Health	5000	612	32
6	Recreation	5000	606	22
7	Reference	5000	793	33
8	Science	5000	743	40
9	Society	5000	636	27
10	Genbase	662	1185	27
11	Medical	978	1449	45
12	Enron	1702	1001	53
13	Social	5000	1047	39

为了全面验证 RMLFS 的分类性能，将其与以下的特征选择方法进行了比较：MIFS[36]、MDMR[35]、SCLS[34]、mRMR[7]、RALM-FS[37]、TRCFS[38]、LRFS[39]和 GMM[40]. 此外，参照 MIFS[36]，实验部分选择热核函数构造近邻矩阵，其中设置 p 和 σ 分别为 5.0 和 1.0. 进一步为公平起见，我们在相同的网格中调整各个带有正则化参数算法的参数，该网格为 $\{0.01,0.1,0.3,0.5,0.7,0.9,1.0\}$. 然后利用正则化参数的最优值来评价分类性能. 接下来，多标记问题通过 BR 模型转化为多个相互独立的单标签分类问题. 然后我们使用线性支持向量机（SVM）和 K 最近邻分类器（KNN）来学习这些相互独立的单标签分类问题，其中 SVM 的参数 C 通过网格 $\{10^{-4},10^{-3},\cdots,10^{3},10^{4}\}$ 进行调整. 另外采用考虑标签关系评测的学习器 ML-KNN[41]. 同时，在实验中采用了五折交叉验证. 而且采用以下评估标准，基于 $F1$ 度量的 Micro-average 和 Macro-average（Micro-$F1$ 和 Macro-$F1$）[54]和考虑标签关系的 Hamming Loss 评估指标. 它们的定义公式如下：

$$\begin{cases}\text{Micro-}F1=\dfrac{\sum\limits_{i=1}^{z}2TP^{i}}{\sum\limits_{i=1}^{z}(2TP^{i}+FP^{i}+FN^{i})}\\ \text{Macro-}F1=\dfrac{1}{z}\sum\limits_{i=1}^{z}\dfrac{2TP^{i}}{(2TP^{i}+FP^{i}+FN^{i})}\end{cases}\tag{20}$$

$$\text{Hamming Loss}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{z}\sum_{j=1}^{z}|Y'_{ij}\oplus Y_{ij}|\tag{21}$$

其中，z 表示标签的数目，对于第 i 个标签，TP 表示真阳性，FP 表示假阳性，FN 表示假阴性. $\oplus$ 表示异或操作. Y'_{ij} 和 Y_{ij} 分别表示第 (i,j) 个预测标签和原始标签.

4.2 实验结果及分析

提出的多标记特征选择方法 RMLFS 与 8 个代表性的特征选择算法在 13 个数据集上进行比较. 而且根据参考文献 MIFS[36]给定的经验值，即通过特征总量的前 20% 来计算每一个要对比方法的均值与标准差. 所选特征的数量步长设为 1%（Flags 数据集仅有 19 个特征，因此使用其所有特征）. 使用两个分类器的所有方法在 Micro-$F1$、Macro-$F1$ 和 Hamming Loss 准则下的实验结果被展示在表 3～表 7 中. 其中，相应数据集在 9 种方法下的最优结果用加粗黑体表示. 带有下划线的字体表示相应数据下 RMLFS 方法所得结果为次优值. 我们也将各

《计算机学报》

方法在所有数据集下的“平均值”给出. 此外，准则 Micro-$F1$、Macro-$F1$ 对应表中的值越大代表对应算法的分类性能越好，而 Hamming Loss 准则对应表中的值越小代表对应算法的分类性能越好（其中 mRMR 属于单标记特征选择方法，不适用于基于多标记学习器 ML-KNN 的 Hamming Loss 准则）.

通过观察这些表，我们发现所提出的方法 RMLFS 在“平均值”方面比其他 8 种方法获得了更优异的分类结果. 在表 3～表 6 中，除数据集 Medical 和 Genbase 之外，在 Micro-$F1$、Macro-$F1$ 的评价标准下，与 8 种对比方法相比，我们的方法在 11 个数据集上获得了最优或次优的结果. 在表 7 中，除数据集

表 3　9 种特征选择方法在 SVM 分类器上的 Micro-$F1$ 结果

Data sets	MIFS	MDMR	SCLS	LRFS	mRMR	RALM_FS	TRCFS	GMM	RMLFS
Flags	**0.7225**±0.0499	0.6491±0.0425	0.6746±0.031	0.6659±0.0405	0.6651±0.0322	0.6315±0.0481	0.6955±0.0203	0.6603±0.0321	0.7136±0.0424
Arts	0.1391±0.0783	0.0946±0.0524	0.106±0.046	0.1041±0.0448	**0.2582**±0.0576	0.1018±0.0607	0.229±0.0917	0.0482±0.0259	0.2545±0.06
Education	0.0733±0.0587	0.127±0.0811	0.0733±0.0593	0.1495±0.0806	0.2357±0.0576	0.1934±0.0558	0.2181±0.0956	0.0569±0.0415	**0.276**±0.0548
Entertain	0.2278±0.1121	0.1898±0.0865	0.1132±0.0717	0.1843±0.0821	0.3445±0.0682	0.2143±0.1004	0.331±0.0996	0.1163±0.0769	**0.35**±0.0869
Health	0.4681±0.0795	0.4346±0.0761	0.3451±0.0441	0.4754±0.037	0.5381±0.0682	0.5157±0.0439	0.5282±0.058	0.3618±0.0552	**0.5594**±0.0378
Recreation	0.2524±0.0698	0.0331±0.0382	0.0644±0.02	0.0392±0.0374	0.2745±0.0445	0.2279±0.0711	0.2678±0.0488	0.007±0.0108	**0.2808**±0.047
Reference	0.3593±0.1046	0.3559±0.069	0.2661±0.0829	0.3768±0.0664	0.4114±0.0911	0.4042±0.1322	0.4391±0.0999	0.305±0.0649	**0.4396**±0.0957
Science	0.1286±0.0569	0.1269±0.0743	0.0365±0.035	0.1286±0.0776	0.2071±0.0448	0.0969±0.0539	0.2249±0.0662	0.0368±0.032	**0.2338**±0.0537
Society	0.3001±0.0424	0.3191±0.0172	0.216±0.028	0.321±0.016	0.2708±0.0511	0.2231±0.0594	0.3106±0.0645	0.3041±0.0403	**0.3221**±0.0452
Genbase	0.9717±0.101	0.9785±0.0675	0.541±0.0137	0.9785±0.0676	**0.9937**±0.0006	0.9582±0.1365	0.9748±0.0875	0.0000±0.0000	0.9711±0.1054
Medical	0.7149±0.1058	0.7556±0.0546	0.3697±0.0094	0.7567±0.0548	0.7572±0.0081	0.3632±0.1474	**0.7613**±0.1081	0.0000±0.0000	0.7372±0.1349
Enron	0.3723±0.0274	0.4739±0.0492	0.4884±0.0314	0.4457±0.0555	**0.5346**±0.0241	0.3891±0.0588	0.4051±0.0505	0.385±0.0357	0.5274±0.049
Social	0.2756±0.1361	0.465±0.1058	0.3631±0.12	0.4648±0.1194	0.5313±0.067	0.149±0.1124	**0.5462**±0.0595	0.3843±0.1113	0.5458±0.0998
Average	0.3851	0.3849	0.2813	0.3916	0.4632	0.3437	0.4563	0.2051	**0.4778**

表 4　9 种特征选择方法在 SVM 分类器上的 Macro-$F1$ 结果

Data sets	MIFS	MDMR	SCLS	LRFS	mRMR	RALM_FS	TRCFS	GMM	RMLFS
Flags	**0.5697**±0.0998	0.5035±0.0535	0.5388±0.042	0.5106±0.0442	0.5052±0.0495	0.4872±0.0436	0.5221±0.0397	0.5233±0.0329	0.5638±0.0842
Arts	0.055±0.034	0.0393±0.0219	0.0381±0.0161	0.0433±0.0187	**0.1142**±0.0252	0.0385±0.0256	0.1025±0.0359	0.0194±0.0107	0.1131±0.0275
Education	0.0195±0.017	0.0353±0.0249	0.0195±0.0149	0.0485±0.0279	**0.0803**±0.0204	0.0525±0.0144	0.0765±0.0267	0.0132±0.0106	0.0767±0.0168
Entertain	0.0971±0.0474	0.0772±0.032	0.0436±0.03	0.0788±0.0307	0.1465±0.026	0.0815±0.0384	0.145±0.0405	0.043±0.0294	**0.1573**±0.037
Health	0.1181±0.0461	0.1226±0.0519	0.0671±0.0346	0.1454±0.0378	0.1836±0.0339	0.1592±0.0415	**0.1875**±0.0454	0.0441±0.0164	0.1862±0.0434
Recreation	0.1336±0.0334	0.0193±0.0227	0.0301±0.0108	0.0231±0.0225	0.1449±0.0283	0.1218±0.0421	0.1459±0.0299	0.0043±0.0065	**0.1506**±0.0289
Reference	0.0628±0.0236	0.0408±0.0205	0.0214±0.0068	0.051±0.0207	0.0929±0.0194	0.0628±0.0278	0.094±0.0234	0.0186±0.0044	**0.0969**±0.0198
Science	0.0341±0.0164	0.0378±0.0285	0.0074±0.0071	0.0388±0.0287	0.0758±0.0151	0.0357±0.0199	**0.0939**±0.0265	0.0087±0.0081	0.0866±0.023
Society	0.0554±0.0197	0.049±0.0159	0.021±0.0035	0.0493±0.0155	0.0641±0.0114	0.0318±0.0116	0.0788±0.0195	0.0346±0.008	**0.0909**±0.0183
Genbase	0.6894±0.1091	0.7642±0.1291	0.2413±0.0221	0.7636±0.1304	**0.8148**±0.0003	0.7375±0.1527	0.7651±0.142	0.0000±0.0000	0.7837±0.1169
Medical	0.2216±0.0483	0.3157±0.0736	0.0793±0.0127	0.3172±0.0735	0.3054±0.0117	0.1288±0.0629	**0.3358**±0.0681	0.0000±0.0000	0.3252±0.0812
Enron	0.0743±0.0169	0.1076±0.0324	0.1189±0.031	0.0988±0.0325	0.1463±0.0303	0.0741±0.0268	0.0822±0.0289	0.0798±0.0237	**0.1381**±0.0366
Social	0.0308±0.0164	0.0754±0.0314	0.0362±0.0143	0.0832±0.0312	0.1042±0.0131	0.0144±0.012	**0.1263**±0.0246	0.0349±0.011	0.1223±0.0256
Average	0.1663	0.1683	0.0971	0.1732	0.2137	0.1558	0.212	0.0634	**0.2224**

表 5　9 种特征选择方法在 3NN 分类器上的 Micro-$F1$ 结果

Data sets	MIFS	MDMR	SCLS	LRFS	mRMR	RALM_FS	TRCFS	GMM	RMLFS
Flags	0.6615±0.0302	0.6107±0.0165	0.6122±0.0115	0.6157±0.018	0.6078±0.018	0.6057±0.0122	0.6604±0.0343	0.6129±0.0136	**0.6829**±0.0351
Arts	0.2016±0.0521	0.1893±0.0343	0.1641±0.0243	0.1963±0.0273	0.2332±0.0232	0.1824±0.0435	0.2612±0.0643	0.1838±0.0336	**0.2829**±0.0323
Education	0.1827±0.055	0.2308±0.0404	0.1764±0.0371	0.2431±0.0455	0.263±0.0256	0.2605±0.0505	0.2913±0.0685	0.2223±0.0361	**0.3061**±0.0431
Entertain	0.2764±0.0652	0.2832±0.0304	0.2341±0.0334	0.2807±0.0339	0.3353±0.0268	0.2735±0.0651	0.3418±0.0816	0.2802±0.0323	**0.3842**±0.0649
Health	0.435±0.0706	0.407±0.0514	0.348±0.0389	0.4365±0.0204	0.4823±0.0412	0.4739±0.0587	0.4751±0.0728	0.3901±0.0446	**0.4889**±0.0652
Recreation	0.2824±0.0535	0.1401±0.0257	0.1483±0.0215	0.1406±0.0264	0.257±0.0252	0.2717±0.0635	0.3048±0.0465	0.1343±0.0238	**0.314**±0.0482
Reference	0.3816±0.0546	0.3784±0.0443	0.3121±0.0304	0.3859±0.0456	0.4067±0.0424	0.3857±0.089	0.4478±0.0723	0.371±0.041	**0.4497**±0.0447
Science	0.1711±0.0365	0.1737±0.0345	0.1154±0.0162	0.1765±0.0346	0.2119±0.0203	0.1597±0.0476	0.2556±0.0377	0.161±0.0301	**0.2598**±0.0342
Society	0.3055±0.043	0.3121±0.0324	0.2449±0.0207	0.3153±0.0282	0.2977±0.0235	0.2546±0.0499	0.3189±0.0506	0.3117±0.0294	**0.334**±0.045
Genbase	0.9665±0.1032	0.9767±0.0655	0.5175±0.0118	0.9766±0.0657	**0.9915**±0.0009	0.9559±0.134	0.9724±0.0879	0.0721±0.0000	0.9662±0.1012
Medical	0.6104±0.0953	0.6411±0.0342	0.3528±0.0132	0.6427±0.0349	0.5811±0.0456	0.2939±0.1077	0.6561±0.1288	0.0000±0.0000	**0.7008**±0.1275
Enron	0.4102±0.0243	0.4435±0.0437	0.4365±0.0265	0.4194±0.048	0.505±0.0093	0.365±0.0729	0.3923±0.0347	0.4145±0.0502	**0.4932**±0.0386
Social	0.3378±0.0811	0.4509±0.0578	0.3763±0.0569	0.4549±0.0543	0.4978±0.0243	0.3154±0.054	**0.5383**±0.0307	0.4295±0.0578	0.5272±0.0757
Average	0.4017	0.4029	0.3107	0.4065	0.4362	0.3691	0.4551	0.2756	**0.4761**

计算机学报

表 6 9 种特征选择方法在 3NN 分类器上的 Macro-*F*1 结果

Data sets	MIFS	MDMR	SCLS	LRFS	mRMR	RALM_FS	TRCFS	GMM	RMLFS
Flags	0.5265±0.0776	0.4432±0.0175	0.4538±0.0137	0.4531±0.022	0.4513±0.0227	0.4386±0.0109	0.5102±0.0487	0.4452±0.0156	**0.5522**±0.0658
Arts	0.0951±0.033	0.0952±0.0245	0.0737±0.0139	0.1006±0.025	0.1122±0.015	0.0714±0.0253	**0.1434**±0.0402	0.0887±0.0229	0.1382±0.0309
Education	0.0426±0.0176	0.0823±0.0187	0.0592±0.0208	0.087±0.0202	0.0956±0.0124	0.0838±0.0234	**0.1047**±0.0227	0.0796±0.0185	0.101±0.0194
Entertain	0.1379±0.0418	0.148±0.0182	0.112±0.0202	0.1448±0.0198	0.1748±0.017	0.1305±0.0402	0.181±0.0462	0.1455±0.0175	**0.1977**±0.039
Health	0.157±0.0407	0.1437±0.0328	0.0985±0.0305	0.1611±0.0225	0.1953±0.0201	0.1859±0.0317	0.2008±0.0352	0.1313±0.0277	**0.2054**±0.0368
Recreation	0.1703±0.0381	0.0945±0.0219	0.0854±0.0161	0.0932±0.0213	0.1644±0.018	0.1597±0.0447	0.1856±0.0354	0.0909±0.0205	**0.1868**±0.0331
Reference	0.088±0.0238	0.0809±0.0174	0.0497±0.0155	0.0852±0.017	0.0955±0.0117	0.0766±0.0257	0.1047±0.0197	0.0763±0.0172	**0.1155**±0.0202
Science	0.0618±0.0147	0.0674±0.0238	0.0372±0.0098	0.0673±0.0231	0.1071±0.0103	0.0567±0.0211	0.1123±0.0217	0.0611±0.021	**0.1131**±0.0263
Society	0.0888±0.0211	0.0839±0.0139	0.055±0.0112	0.0847±0.0129	0.0777±0.0077	0.0534±0.0156	0.1031±0.0199	0.0836±0.0152	**0.1124**±0.0188
Genbase	0.6661±0.1075	0.7092±0.1036	0.2238±0.0178	0.709±0.1043	**0.7403**±0.002	0.6887±0.1322	0.7023±0.1234	0.0055±0.0000	0.7052±0.0927
Medical	0.161±0.0209	0.1856±0.0257	0.0626±0.006	0.1872±0.0265	0.1526±0.0393	0.0694±0.0287	0.2472±0.0427	**0.2582**±0.0575	**0.2582**±0.0575
Enron	0.0873±0.014	0.1152±0.0203	0.1106±0.0126	0.1093±0.0219	0.1454±0.0078	0.0809±0.0256	0.0968±0.0219	0.1038±0.0199	**0.1313**±0.0189
Social	0.0506±0.0167	0.1066±0.0319	0.0494±0.0132	0.1114±0.0269	0.1161±0.0057	0.0382±0.0123	**0.1528**±0.0239	0.0957±0.0293	0.1459±0.0301
Average	0.1795	0.1812	0.1131	0.1841	0.2022	0.1641	0.2188	0.1281	**0.2279**

表 7 8 种特征选择方法在 ML-KNN 分类器上的 Hamming Loss 结果

Data sets	MIFS	MDMR	SCLS	LRFS	RALM_FS	TRCFS	GMM	RMLFS
Flags	0.292±0.0205	0.3146±0.0111	0.3227±0.0061	0.3169±0.0087	0.3239±0.0087	0.2942±0.0193	0.3181±0.0082	**0.2877**±0.0204
Arts	0.0632±0.0012	0.0647±0.0007	0.0634±0.0006	0.0647±0.0009	0.0633±0.0008	0.06±0.0015	0.0648±0.0007	**0.0592**±0.0007
Education	0.0449±0.0007	0.0449±0.0011	0.0456±0.0008	0.0447±0.0011	0.0433±0.0005	0.0423±0.0013	0.045±0.0008	**0.0412**±0.0007
Entertain	0.0653±0.0018	0.0655±0.0016	0.0677±0.001	0.0657±0.0015	0.0638±0.0021	0.0606±0.002	0.0661±0.0018	**0.0585**±0.0022
Health	0.0448±0.0026	0.0466±0.0023	0.0497±0.0011	0.0454±0.0013	0.0423±0.0022	0.0417±0.003	0.0476±0.0021	**0.0395**±0.0019
Recreation	0.0601±0.0017	0.0671±0.0007	0.0654±0.0007	0.0667±0.0007	0.0605±0.0013	**0.0586**±0.0012	0.067±0.001	0.0591±0.0016
Reference	0.0313±0.0015	0.0325±0.0014	0.0339±0.0007	0.0326±0.0013	0.0325±0.0013	0.029±0.0014	0.0327±0.0017	**0.0286**±0.0015
Science	0.036±0.0006	0.0363±0.0005	0.0368±0.0005	0.0362±0.0005	0.0361±0.0004	0.0343±0.0004	0.0363±0.0004	**0.0341**±0.0004
Society	0.0582±0.0012	0.0585±0.0011	0.0603±0.0004	0.0584±0.001	0.0593±0.0006	0.0573±0.0012	0.0586±0.0011	**0.0562**±0.0011
Genbase	**0.0026**±0.0055	0.0027±0.004	0.0309±0.0004	0.0027±0.004	0.0038±0.0068	0.003±0.0047	0.0456±0.0000	0.003±0.0056
Medical	0.0165±0.002	0.0175±0.001	0.0233±0.0002	0.0175±0.001	0.027±0.0007	0.0149±0.002	0.0276±0.0000	**0.0139**±0.003
Enron	0.0574±0.0012	0.0531±0.0027	0.0532±0.0012	0.055±0.003	0.057±0.0023	0.056±0.0023	0.0577±0.0007	**0.0506**±0.0022
Social	0.0311±0.0016	0.0275±0.0018	0.0292±0.0018	0.0275±0.002	0.0324±0.001	0.0238±0.001	0.0284±0.0019	**0.0233**±0.0018
Average	0.0618	0.064	0.0679	0.0642	0.065	0.0597	0.0689	**0.0581**

Genbase 之外，RMLFS 取的最优或次优的结果. mRMR 方法在生物数据集 Genbase 上获得最优结果，这是因为 mRMR 方法是关注生物和医学的数据集的方法[7]. 总体而言，RMLFS 在实验中取得了良好的分类性能. 接下来，本文选取六个具有代表性的数据集（Arts、Education、Entertain、Health、Recreation 和 Society）来展示分类效果. 实验结果如图 1～图 5 所示. 在这些图中，我们使用 X 轴表示特征的数量，而 Y 轴表示分类性能. 分类性能首先随着已选特征的增加而提高，然后逐渐趋于平坦甚至稳定. 并且可以直观地观察到 RMLFS 在大多数情况下优于其他六种特征选择方法. 在图 5 中，RMLFS 等方法的标签预测误差率首先随着已选特征的增加而变小，然后逐渐趋于平坦甚至稳定.

4.3 参数敏感性分析

类似于其他方法[5,36]，我们对提出的 RMLFS 目标函数中的正则化参数进行参数敏感性分析. 在本小节，仅使用数据集 Arts 来分析正则化参数 (α,β,γ) 这些参数在上文提到的网格中进行调整，并且调整其中一个参数时，其他参数被固定不变. 其中被固定的参数的值被设置为 0.5. 整个参数敏感性分析实验过程中，在 SVM 和 K-NN 分类器下，我们得到了所有参数的 Micro-*F*1 和 Macro-*F*1 结果. 为了方便起见，我们只使用 SVM 分类器来分析参数的敏感性. 实验结果被展示在图 6 中. 可以看出，分类性能对这些参数 (α,β,γ) 的变化不是很敏感. 而且还观察到，该方法的分类性能先提高，然后随着 SVM 分类器已选特征数的增加而逐渐趋于稳定，这与第 4.2 节实验结果相符合.

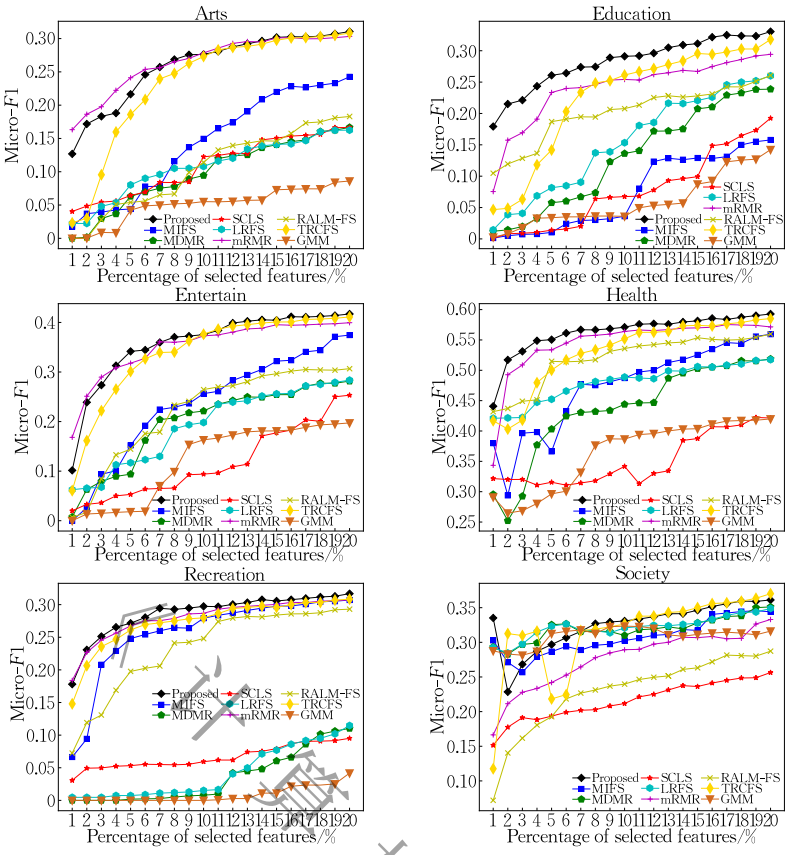

图 1　六个数据集在 Micro-$F1$(SVM)上的实验结果

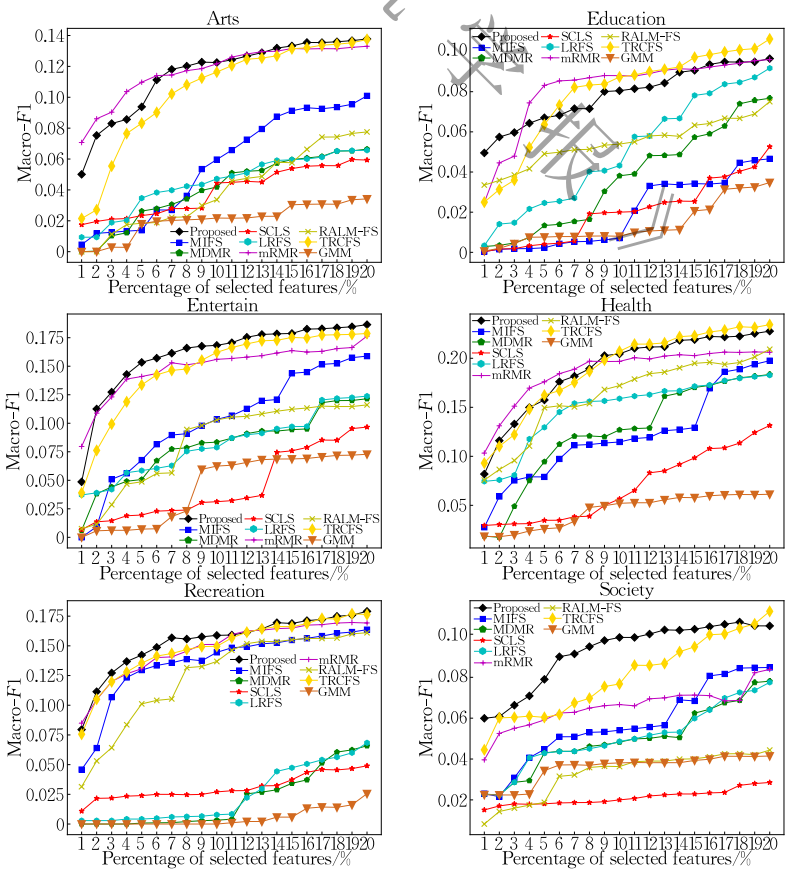

图 2　六个数据集在 Macro-$F1$(SVM)上的实验结果

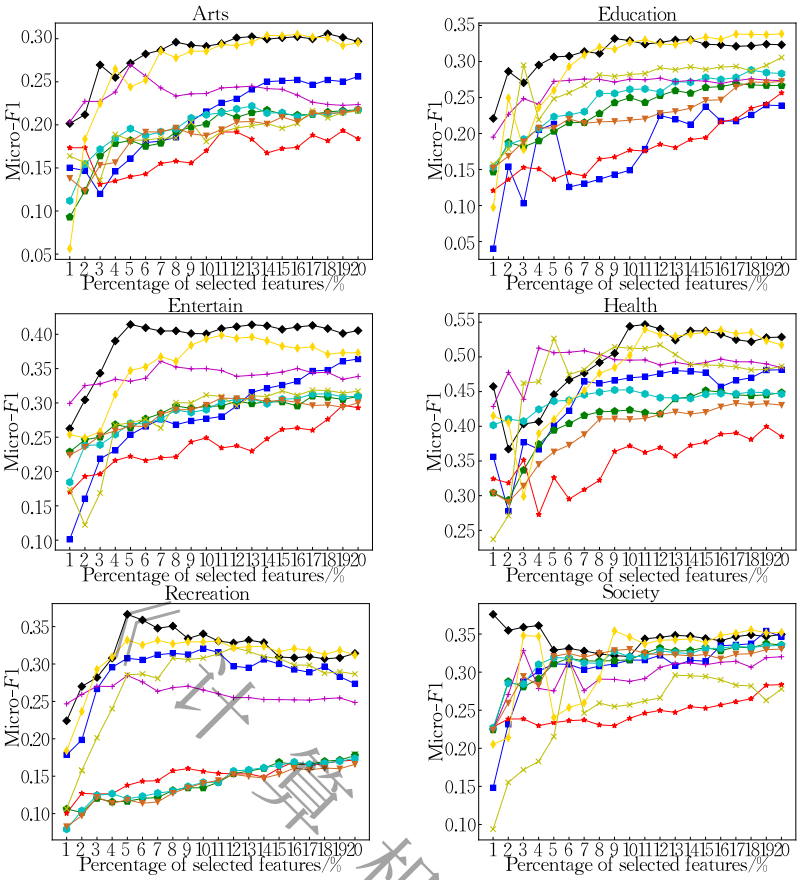

图 3 六个数据集在 Micro-$F1$(3NN)上的实验结果

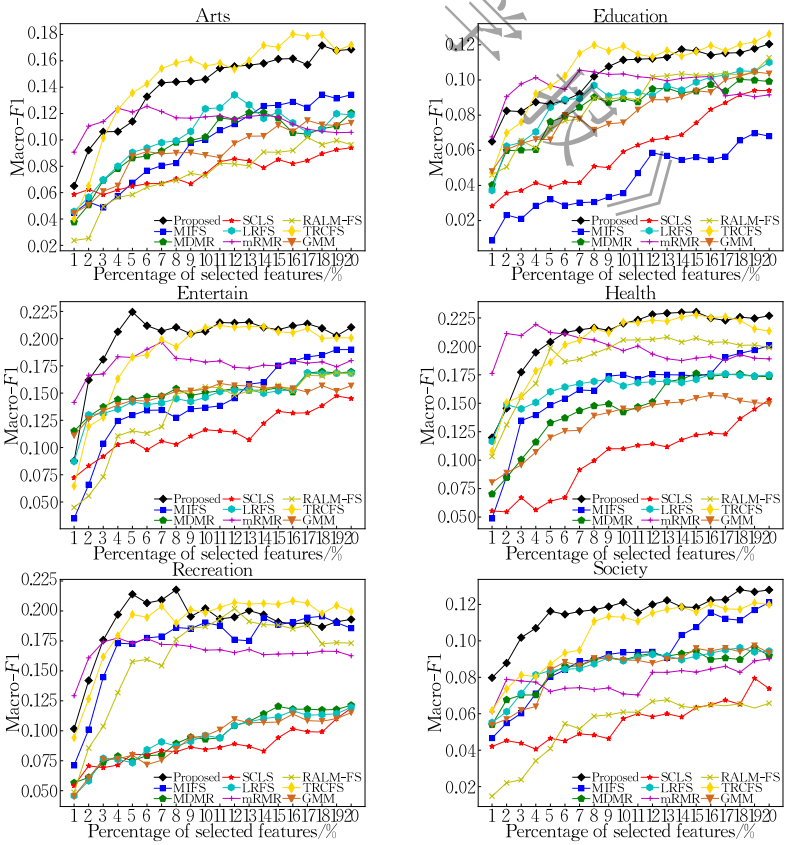

图 4 六个数据集在 Macro-$F1$(3NN)上的实验结果

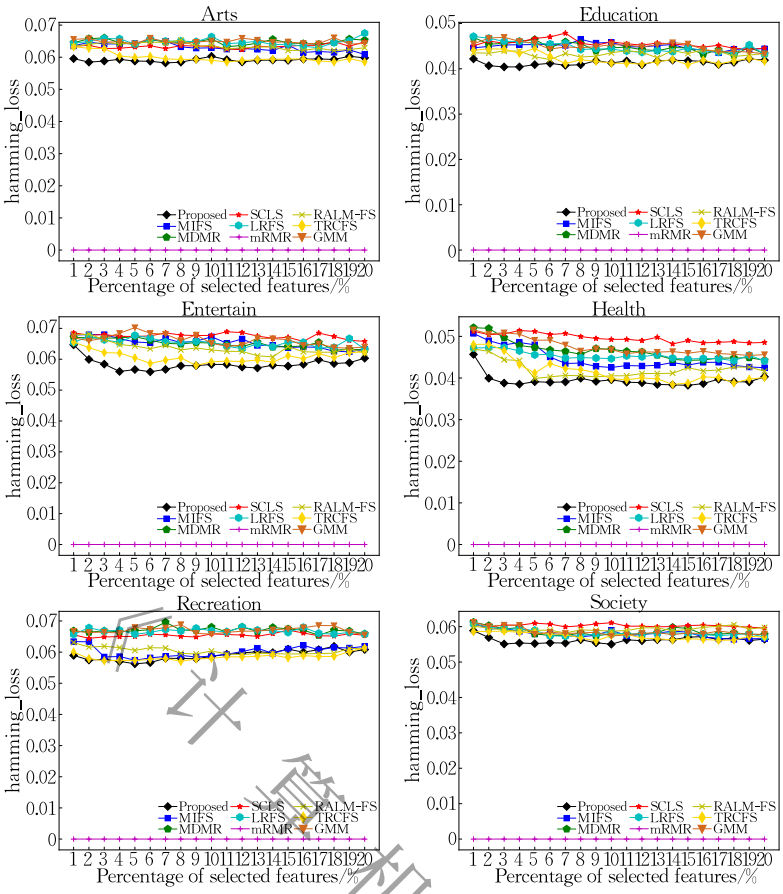

图 5　六个数据集在 Hamming Loss(ML-KNN)上的实验结果

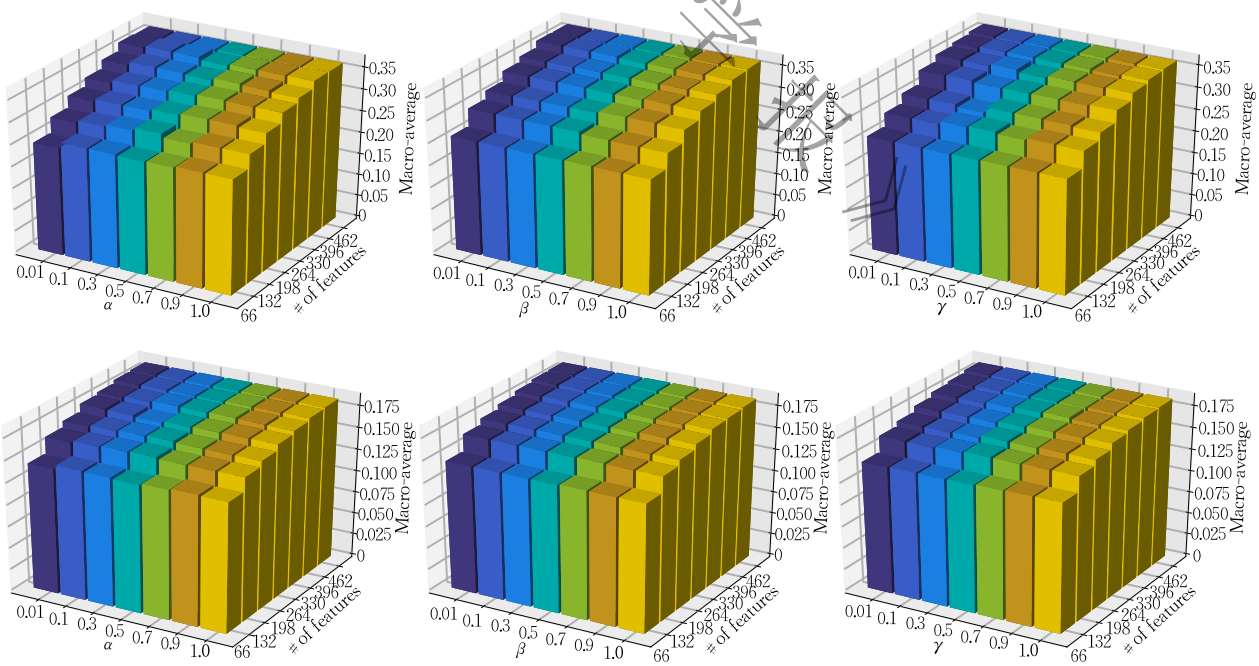

图 6　在 Arts 数据集上提出算法的 Micro-$F1$ 和 Macro-$F1$(SVM)

4.4　收敛性分析与时间复杂度

本节中，为了验证 RMLFS 的收敛性能，使用了四个不同的基准数据集进行了实验，包括 Arts、Education、Health 和 Reference. 图 7 显示在这些数据集上目标函数的迭代次数. 可观察到，所提出的 RMLFS 在 4 个数据集上可以在 40 次迭代内快速收敛，其他数据集也有相似的模式. 接下来，本节给出了 RMLFS 算法的计算复杂度，并与本文中提到

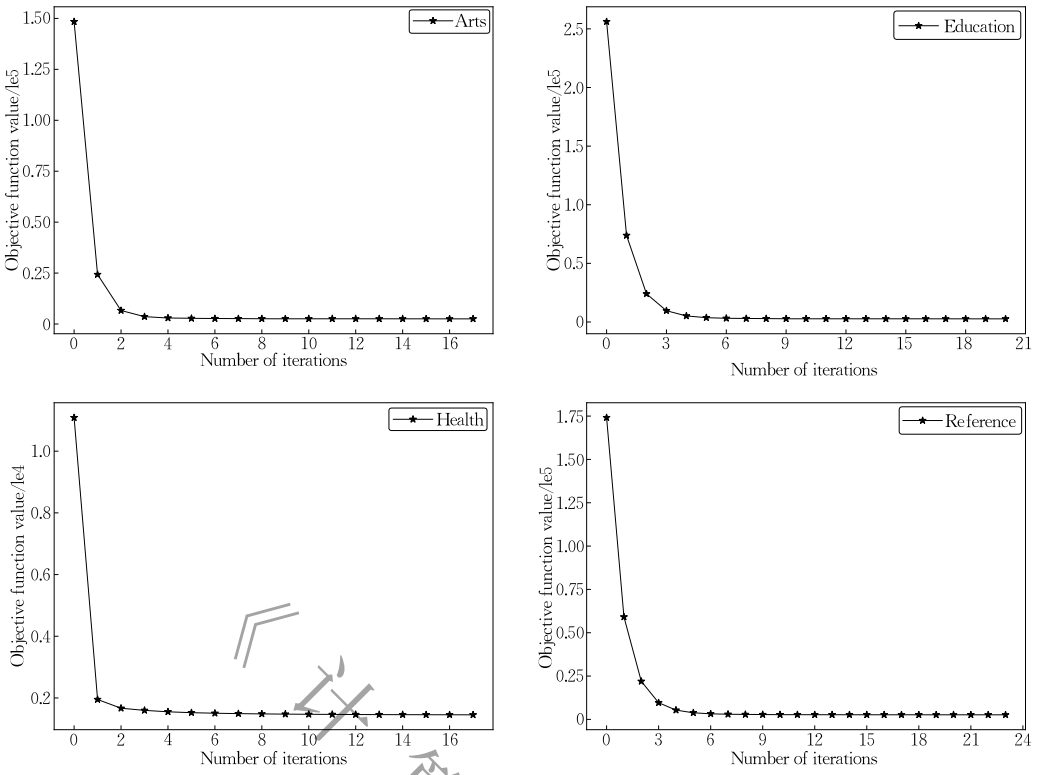

图 7 提出方法的收敛曲线图

的其他对比方法进行了比较. 假设 k 表示已经选择的特征数，d 表示特征总数，每个数据集有 n 个样例. mRMR 的计算复杂度是 $O(knd)$，SCLS 的计算复杂度是 $O(dl+kd)$，其中 l 表示标签数. 而 MIFS 每次迭代的计算复杂度为 $O(cnd+n^2)$，其中 c 表示标签矩阵的聚类数. 由于涉及到矩阵的逆操作，RALM-FS 的计算复杂度为 $O(d^3)$. 由于使用了启发式搜索过程，MDMR 的计算复杂度是 $O(k(d-k))$. TRCFS 的计算复杂度为 $O(n^3+d^2n+dn^2)$. LRFS 的计算复杂度为 $O(dc^2+kd)$. 在每次迭代过程中，RMLFS 的计算复杂度为 $O(dn^2+d^2n)$. 虽然RMLFS 的计算复杂度并不是最低的，但 RMLFS 的分类性能优于其他的比较方法.

5 总结与展望

本文提出了一种新的多标记特征选择方法 RMLFS. 现存的多标记特征选择方法要么采用样例层流形正则化方法来保持样例的相似性，要么利用标签间的相关性来指导特征选择过程. 然而这两种方法对于特征选择的指导存在互补关系，因此采用流形正则化方法同时考虑样例相似性和标签相似性，从而可得到从特征空间到标签空间对应关系的精确映射. 此外，该方法还引入了标签流形正则化项，该正则化项的近邻矩阵是基于标签之间的相似性来利用标签相关性的. 而且所提出的目标函数只有一个未知变量，即稀疏系数矩阵，可以有效约减变量过多导致的局部最优问题. 实验结果表明，本文提出的 RMLFS 算法在分类性能上优于其他同类算法. 由于非凸优化问题广泛存在于各个领域，是当下急需攻克的一项重要课题，因此是非常值得研究的方向. 在今后的工作中，非凸优化问题下的多标记特征选择问题将被重点研究.

参 考 文 献

[1] Zhang Rui, Nie Fei-Ping, Li Xue-Long. Self-weighted supervised discriminative feature selection. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(8): 3913-3918

[2] Gao Wan-Fu, Hu Liang, Zhang Ping. Class-specific mutual information variation for feature selection. Pattern Recognition, 2018, 79: 328-339

[3] Hu J, Li Y, Xu G, et al. Dynamic subspace dual-graph regularized multi-label feature selection. Neurocomputing, 2022, 467: 184-196

[4] Wang J, Zhang H, Wang J, et al. Feature selection using a neural network with group lasso regularization and controlled redundancy. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(3): 1110-1123

[5] Tang C, Bian M, Liu X, et al. Unsupervised feature selection

via latent representation learning and manifold regularization. Neural Networks, 2019, 117: 163-178

[6] Wang J, Wei J M, Yang Z, et al. Feature selection by maximizing independent classification information. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(4): 828-841

[7] Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Computer Society, 2005, 27(8): 1226-1238

[8] Hou C, Jiao Y, Nie F, et al. 2D feature selection by sparse matrix regression. IEEE Transactions on Image Processing, 2017, 26(9): 4255-4268

[9] Zhu R, Dornaika F, Ruichek Y. Learning a discriminant graph-based embedding with feature selection for image categorization. Neural Networks, 2019, 111: 35-46

[10] Shi C, An G, Zhao R, et al. Multiview Hessian semisupervised sparse feature selection for multimedia analysis. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 27(9): 1947-1961

[11] Nie F, Huang H, Cai X, et al. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Red Hook, USA, 2010: 1813-1821

[12] Kong Y, Yu T. A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. Bioinformatics, 2018, 34(21): 3727-3737

[13] Uysal A K. An improved global feature selection scheme for text classification. Expert systems with Applications, 2016, 43: 82-92

[14] Wang A, An N, Chen G, et al. Accelerating wrapper-based feature selection with K-nearest-neighbor. Knowledge-Based Systems, 2015, 83: 81-91

[15] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing, 2018, 62: 441-453

[16] Hu L, Li Y, Gao W, et al. Multi-label feature selection with shared common mode. Pattern Recognition, 2020, 104: 107344

[17] Shang R, Wang W, Stolkin R, et al. Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection. IEEE Transactions on Cybernetics, 2017, 48(2): 793-806

[18] Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A, et al. A review of microarray datasets and applied feature selection methods. Information Sciences, 2014, 282: 111-135

[19] Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification. Pattern Recognition, 2004, 37(9): 1757-1771

[20] Liu J, Lin Y, Wu S, et al. Online multi-label group feature selection. Knowledge-Based Systems, 2018, 143: 42-57

[21] Zhang J, Luo Z, Li C, et al. Manifold regularized discriminative feature selection for multi-label learning. Pattern Recognition, 2019, 95: 136-150

[22] Liu M, Zhang D. Pairwise constraint-guided sparse learning for feature selection. IEEE Transactions on Cybernetics, 2015, 46(1): 298-310

[23] Zhu Y, Kwok J T, Zhou Z H. Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(6): 1081-1094

[24] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006, 7(1): 2399-2434

[25] Zhang M-L, Zhou Z-H. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(8): 1819-1837

[26] Huang J, Li G, Huang Q, et al. Joint feature selection and classification for multilabel learning. IEEE Transactions on Cybernetics, 2017, 48(3): 876-889

[27] Huang J, Li G, Huang Q, et al. Learning label specific features for multi-label classification//Proceedings of the IEEE International Conference on Data Mining. Atlantic, USA, 2015: 181-190

[28] Ling Y, Wang Y, Wang X, et al. Exploring common and label-specific features for multi-label learning with local label correlations. IEEE Access, 2020, 8: 50969-50982

[29] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification. Machine Learning, 2011, 85(3): 333-359

[30] Gui J, Sun Z, Ji S, et al. Feature selection based on structured sparsity: A comprehensive study. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(7): 1490-1507

[31] Li J, Cheng K, Wang S, et al. Feature selection: A data perspective. ACM Computing Surveys (CSUR), 2017, 50(6): 1-45

[32] Sheikhpour R, Sarram M A, Gharaghani S, et al. A survey on semi-supervised feature selection methods. Pattern Recognition, 2017, 64: 141-158

[33] Kashef S, Nezamabadi-pour H, Nikpour B. Multilabel feature selection: A comprehensive review and guiding experiments. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(2): e1240

[34] Lee J, Kim D W. SCLS: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition, 2017, 66: 342-352

[35] Lin Y, Hu Q, Liu J, et al. Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing, 2015, 168: 92-103

[36] Jian L, Li J, Shu K, et al. Multi-label informed feature selection//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 1627-1633

[37] Cai X, Nie F, Huang H. Exact top-k feature selection via $l_{2,0}$-norm constraint//Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 1240-1246

[38] Liu Y, Nie F, Wu J, et al. Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. Neurocomputing, 2013, 105: 12-18

[39] Zhang P, Liu G, Gao W. Distinguishing two types of labels for multi-label feature selection. Pattern Recognition, 2019, 95: 72-82

[40] Gonzalez-Lopez J, Ventura S, Cano A. Distributed multi-label feature selection using individual mutual information measures. Knowledge-Based Systems, 2020, 188: 105052

[41] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7): 2038-2048

[42] Zhang Y, Wu J, Cai Z, et al. Multi-view multi-label learning with sparse feature selection for image annotation. IEEE Transactions on Multimedia, 2020, 22(11): 2844-2857

[43] Guo Y, Chung F, Li G, et al. Leveraging label-specific discriminant mapping features for multi-label learning. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(2): 1-23

[44] Guo Y, Chung F L, Li G, et al. Multi-label bioinformatics data classification with ensemble embedded feature selection. IEEE Access, 2019, 7: 103863-103875

[45] Wu W, Kwong S, Hou J, et al. Simultaneous dimensionality reduction and classification via dual embedding regularized nonnegative matrix factorization. IEEE Transactions on Image Processing, 2019, 28(8): 3836-3847

[46] Hou P, Geng X, Zhang M L. Multi-label manifold learning// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, 2016: 1680-1686

[47] Nie F, Wang Z, Wang R, et al. Submanifold-preserving discriminant analysis with an auto-optimized graph. IEEE Transactions on Cybernetics, 2019, 50(8): 3682-3695

[48] Cai D, He X, Han J, et al. Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(8): 1548-1560

[49] Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395-416

[50] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003, 15(6): 1373-1396

[51] Tsoumakas G, Spyromitros-Xioufis E, Vilcek J, et al. Mulan: A java library for multi-label learning. The Journal of Machine Learning Research, 2011, 12: 2411-2414

[52] Liu Hai-Yang, Wang Zhi-Hai, Zhang Zhi-Dong. ReliefF based pruning model for multi-label classification. Chinese Journal of Computers, 2019, 42(3): 483-496(in Chinese) (刘海洋, 王志海, 张志东. 基于 ReliefF 剪枝的多标记分类算法. 计算机学报, 2019, 42(3): 483-496)

[53] Huang K H, Lin H T. Cost-sensitive label embedding for multi-label classification. Machine Learning, 2017, 106(9): 1725-1746

[54] Yu K, Yu S, Tresp V. Multi-label informed latent semantic indexing//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, 2005: 258-265

[55] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22

附录 A.

根据 EM 算法概念[55],我们采用如下定义.

定义 1. 如果如下两个条件被满足,即$\mathcal{G}(x,x')\geq\mathcal{F}(x)$和$\mathcal{G}(x,x)=\mathcal{F}(x)$,那么$\mathcal{G}(x,x')$是$\mathcal{F}(x)$的一个辅助函数.

收敛性定理 1. 式(12)的目标函数值单调减小,直到算法 1 收敛.

证明. 要证明收敛性定理 1,我们首先需要证明如下两个引理成立.

引理 1. 如果$\mathcal{G}(x,x')$是$\mathcal{F}(x)$的一个辅助函数,那么$\mathcal{F}(x)$在如下条件下单调递减:

$$x^{t+1}=\arg\min_{x}\mathcal{G}(x,x^t) \tag{A1}$$

引理 1 的证明. 根据定义 1 条件和结论可得:$\mathcal{F}(x^{t+1})\leq\mathcal{G}(x^{t+1},x^t)\leq\mathcal{G}(x^t,x^t)=\mathcal{F}(x^t)$,引理 1 被证明. 接下来通过一个恰当设定的辅助函数,我们证明本文目标函数的收敛性. 这里我们仅仅考虑$\boldsymbol{W}_{ij}$,因为目标函数的更新规则本质是按照元素更新的. 这里$\mathcal{F}_{ij}$被用来表示关于$\boldsymbol{W}_{ij}$的$\Theta(\boldsymbol{W})$的部分. 然后我们得到如下公式:

$$\begin{aligned}\mathcal{F}'_{ij}=&(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+2\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}_x\boldsymbol{X}\boldsymbol{W}+\\&2\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{L}_y+2\gamma\boldsymbol{D}\boldsymbol{W})_{ij}\end{aligned} \tag{A2}$$

$$\begin{aligned}\mathcal{F}''_{ij}=&2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}+2\alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{X})_{ii}+2\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}(\boldsymbol{L}_{\boldsymbol{Y}})_{jj}+\\&2\gamma(\boldsymbol{D}_{ii}-\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{D}^3\mathrm{Diag}(\boldsymbol{e}_i\boldsymbol{\varepsilon}_j^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}})\boldsymbol{W}\boldsymbol{\varepsilon}_j)\end{aligned} \tag{A3}$$

其中$\boldsymbol{e}_i\in\mathbb{R}^d$和$\boldsymbol{\varepsilon}_j\in\mathbb{R}^c$表示标准列向量. 函数$\mathrm{Diag}(\boldsymbol{A})$可以设置任意矩阵$\boldsymbol{A}$的非对角元素为 0. 因此$\mathcal{F}_{ij}(\boldsymbol{W}_{ij})$的泰勒展开式为

$$\begin{aligned}\mathcal{F}_{ij}(\boldsymbol{W}_{ij})=&\mathcal{F}_{ij}(\boldsymbol{W}_{ij}^t)+\mathcal{F}'_{ij}(\boldsymbol{W}_{ij}^t)(\boldsymbol{W}_{ij}-\boldsymbol{W}_{ij}^t)+\\&\frac{1}{2}\mathcal{F}''_{ij}(\boldsymbol{W}_{ij}^t)(\boldsymbol{W}_{ij}-\boldsymbol{W}_{ij}^t)^2\end{aligned} \tag{A4}$$

引理 2. 如下函数是关于$\mathcal{F}_{ij}(\boldsymbol{W}_{ij})$的一个辅助函数.

$$\begin{aligned}\mathcal{G}(\boldsymbol{W}_{ij},\boldsymbol{W}_{ij}^t)=&\mathcal{F}_{ij}(\boldsymbol{W}_{ij}^t)+\mathcal{F}'_{ij}(\boldsymbol{W}_{ij}^t)(\boldsymbol{W}_{ij}-\boldsymbol{W}_{ij}^t)+\\&\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_x\boldsymbol{X}\boldsymbol{W}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{A}_y+\gamma\boldsymbol{D}\boldsymbol{W})_{ij}}{\boldsymbol{W}_{ij}^t}(\boldsymbol{W}_{ij}-\boldsymbol{W}_{ij}^t)^2\end{aligned} \tag{A5}$$

引理 2 的证明. 显然,当$\boldsymbol{W}_{ij}=\boldsymbol{W}_{ij}^t$时,$\mathcal{G}(\boldsymbol{W}_{ij},\boldsymbol{W}_{ij})=$

$\mathcal{F}(\boldsymbol{W}_{ij})$. 接下来只需要证明$(\boldsymbol{W}_{ij},\boldsymbol{W}_{ij}^t)\geqslant\mathcal{F}(\boldsymbol{W}_{ij})$，即只需证明如下不等式：

$$\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}+\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{W}+\beta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{A}_{\boldsymbol{Y}}+\gamma\boldsymbol{D}\boldsymbol{W})_{ij}}{\boldsymbol{W}_{ij}^t}$$
$$\geqslant(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}+\alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{X})_{ii}+\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}(\boldsymbol{L}_{\boldsymbol{Y}})_{jj}+$$
$$\gamma(\boldsymbol{D}_{ii}-\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{D}^3\mathrm{Diag}(\boldsymbol{e}_i\boldsymbol{\varepsilon}_j^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}})\boldsymbol{W}\boldsymbol{\varepsilon}_j) \quad (A6)$$

接下来我们证明该不等式成立：

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W})_{ij}=\sum_{l=1}^{d}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{il}\boldsymbol{W}_{lj}^t\geqslant(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}\boldsymbol{W}_{ij}^t \quad (A7.1)$$

$$\alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{W})_{ij}=\alpha\sum_{l=1}^{d}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{X}}\boldsymbol{X})_{il}\boldsymbol{W}_{lj}^t$$
$$\geqslant\alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_{\boldsymbol{X}}\boldsymbol{X})_{ii}\boldsymbol{W}_{ij}^t\geqslant\alpha(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{A}_{\boldsymbol{X}}-\boldsymbol{S}_{\boldsymbol{X}})\boldsymbol{X})_{ii}\boldsymbol{W}_{ij}^t$$
$$=\alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{X})_{ii}\boldsymbol{W}_{ij}^t \quad (A7.2)$$

$$\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{A}_{\boldsymbol{Y}})_{ij}=\beta\sum_{l=1}^{c}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W})_{il}(\boldsymbol{A}_{\boldsymbol{Y}})_{lj}$$
$$\geqslant\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W})_{ii}(\boldsymbol{A}_{\boldsymbol{Y}})_{jj}\geqslant\beta\sum_{l=1}^{d}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{il}\boldsymbol{W}_{lj}^t(\boldsymbol{A}_{\boldsymbol{Y}})_{jj}$$
$$\geqslant\boldsymbol{W}_{ij}^t(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}(\boldsymbol{A}_{\boldsymbol{Y}})_{jj}\geqslant\boldsymbol{W}_{ij}^t(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}(\boldsymbol{A}_{\boldsymbol{Y}}-\boldsymbol{S}_{\boldsymbol{Y}})_{jj}$$
$$=w_{ij}^t(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})_{ii}(\boldsymbol{L}_{\boldsymbol{Y}})_{jj} \quad (A7.3)$$

$$\gamma(\boldsymbol{D}\boldsymbol{W})_{ij}=\gamma\sum_{l=1}^{d}\boldsymbol{D}_{il}\boldsymbol{W}_{lj}^t\geqslant\gamma\boldsymbol{D}_{ii}\boldsymbol{W}_{ij}^t$$
$$\geqslant\gamma(\boldsymbol{D}_{ii}-\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{D}^3\mathrm{Diag}(\boldsymbol{e}_i\boldsymbol{\varepsilon}_j^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}})\boldsymbol{W}\boldsymbol{\varepsilon}_j)\boldsymbol{W}_{ij}^t \quad (A7.4)$$

因此可以证明设定的辅助函数成立，即引理 2 得以证明. 所以可证明目标函数(12)在算法 1 的更新规则下单调递减直至收敛. 因此收敛性定理 1 得以证明.　　证毕.

LI Yong-Hao, Ph. D. candidate. His research interest is multi-label feature selection.

HU Liang, Ph. D., professor. His research interests include machine learning, information security, etc.

GAO Wan-Fu, Ph. D., lecturer. His research interests include feature selection, multi-label learning, information theory and causality.

《计算机学报》

Background

The high-dimensional data is a challenge for machine learning because it not only increases computational cost but also over-fits training models. Feature selection technique can reduce redundant features and improve the classification performance. As a result, it has become a crucial technique for dealing with high-dimensional data, and it has been widely used in many domains, such as image annotation, video annotation and text mining, etc.

However, it is not easy to conduct feature selection on multi-label data because multi-label data not only involves the high-dimensional feature space but also a large number of labels. To address multi-label data, some researchers use the problem transformation strategy. On the other hand, some researchers propose the algorithm adaption strategy that is directly applied to multi-label data.

Existing multi-label feature selection methods either employ instance-level manifold regularizer to maintain the instance similarity or exploit the correlations among labels to guide feature selection process. However, there exist three main issues in most multi-label feature selection methods. First, although both instance-level manifold regularization and correlations among labels have important guiding significance for feature selection, previous multi-label feature selection methods focus on either of the two. Second, most previous multi-label feature selection methods exploit correlations based on the affinity matrix of instance similarity, ignoring the pairwise label correlations. Third, the objective functions of previous multi-label feature selection methods incorporate several unknown variables, which makes the solution of the objective function difficult, additionally, more than one variable usually leads to falling into local optimum. To tackle the problems mentioned above, we first minimize the empirical loss function in the least square regression model. And then, the label manifold regularizer whose affinity matrix is based on label similarity to exploit label correlations is introduced. Meanwhile, the product by feature matrix and weight coefficient matrix is adopted to represent predicted labels. As a result, the local geometric structure is stored in the product by the feature matrix and weight coefficient matrix, furthermore, stored in the weight coefficient matrix. Additionally, we adopt an instance-level manifold regularizer to maintain the instance similarity. Finally, the above terms are integrated into one multi-label objective function which is named Regulating Multi-Label Feature Selection (RMLFS).

This work is supported by the National Key R&D Plan of China under Grant No. 2017YFA0604500, the Key Science and Technology R&D Projects of Jilin Province (No. 20180201103GX), the Science Foundation of Jilin Province of China under Grant No. 2020122209JC.