基于中心对齐多核学习的稀疏多元逻辑回归算法

雷大江*① 唐建烊① 李智星① 吴 渝②
①(重庆邮电大学计算机科学与技术学院 重庆 400065)
②(重庆邮电大学网络智能研究所 重庆 400065)

摘 要：稀疏多元逻辑回归(SMLR)作为一种广义的线性模型被广泛地应用于各种多分类任务场景中。SMLR通过将拉普拉斯先验引入多元逻辑回归(MLR)中使其解具有稀疏性，这使得该分类器可以在进行分类的过程中嵌入特征选择。为了使分类器能够解决非线性数据分类的问题，该文通过核技巧对SMLR进行核化扩充后得到了核稀疏多元逻辑回归(KSMLR)。KSMLR能够将非线性特征数据通过核函数映射到高维甚至无穷维的特征空间中，使其特征能够充分地表达并最终能进行有效的分类。此外，该文还利用了基于中心对齐的多核学习算法，通过不同的核函数对数据进行不同维度的映射，并用中心对齐相似度来灵活地选取多核学习权重系数，使得分类器具有更好的泛化能力。实验结果表明，该文提出的基于中心对齐多核学习的稀疏多元逻辑回归算法在分类的准确率指标上都优于目前常规的分类算法。

关键词：稀疏优化；核技巧；多核学习；稀疏多元逻辑回归
中图分类号：TN911.7; TP181 **文献标识码**：A **文章编号**：1009-5896(2020)11-2735-07
DOI: 10.11999/JEIT190426

Sparse Multinomial Logistic Regression Algorithm Based on Centered Alignment Multiple Kernels Learning

LEI Dajiang① TANG Jianyang① LI Zhixing① WU Yu②
①(*College of Computer*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)
②(*Institute of Web Intelligence*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: As a generalized linear model, Sparse Multinomial Logistic Regression (SMLR) is widely used in various multi-class task scenarios. SMLR introduces Laplace priori into Multinomial Logistic Regression (MLR) to make its solution sparse, which allows the classifier to embed feature selection in the process of classification. In order to solve the problem of non-linear data classification, Kernel Sparse Multinomial Logistic Regression (KSMLR) is obtained by kernel trick. KSMLR can map nonlinear feature data into high-dimensional and even infinite-dimensional feature spaces through kernel functions, so that its features can be fully expressed and eventually classified effectively. In addition, the multi-kernel learning algorithm based on centered alignment is used to map the data in different dimensions through different kernel functions. Then center-aligned similarity can be used to select flexibly multi-kernel learning weight coefficients, so that the classifier has better generalization ability. The experimental results show that the sparse multinomial logistic regression algorithm based on center-aligned multi-kernel learning is superior to the conventional classification algorithm in classification accuracy.

Key words: Sparse optimization; Kernel trick; Multiple kernels learning; Sparse Multinomial Logistic Regression(SMLR)

收稿日期：2019-06-11；改回日期：2020-03-28；网络出版：2020-08-27
*通信作者： 雷大江 leidj@cqupt.edu.cn
基金项目：重庆市留学归国人员创新创业项目支持人选(cx2018120)，国家社会科学基金(17XFX013)，重庆市基础研究与前沿探索项目(cstc2015jcyjA40018)
Foundation Items: The Chongqing Innovative Project of Overseas Study(cx2018120), The National Social Science Foundation of China(17XFX013), The Natural Science Foundation of Chongqing(cstc2015jcyjA40018)

1 引言

多元逻辑回归(Multinomial Logistic Regression, MLR) 算法是统计分析、机器学习和数据挖掘领域的一个经典多分类算法[1]。它相对其他分类算法来说模型较简单、容易理解，也适用于大规模的分类问题，被广泛应用于诸如医学检测[2]、地质测量[3]、文本分类等领域。引入了ℓ_1正则项的多元逻辑回归称作稀疏多元逻辑回归(Sparse Multinomial Logistic Regression, SMLR)，它通过将拉普拉斯先验引入多元逻辑回归中可以使其解具有稀疏性，这让它可以在进行分类的过程中嵌入特征选择，所以在高维数据[4]和稀疏数据集[5]的处理上都具有很大的优势。在多类别分类任务中，二分类算法采用one-vs-rest或者one-vs-one[6]的策略进行多类别分类时会受到在样本不平衡的影响[7]，而SMLR只需要训练1次即可[8]，它继承了稀疏逻辑回归稀疏解的同时也较好地处理了多分类问题。因此，SMLR被广泛应用在图像中的多类物体识别、高光谱图像分类[9]、生物信息学[10]等领域。

为了解决非线性的特征数据分类的问题，SMLR经过核技巧核化扩充[11]之后的分类器称为核稀疏多元逻辑回归(Kernel Sparse Multinomial Logistic Regression, KSMLR)[12]。核方法通过将原数据的样本映射到1个或多个高维甚至无穷维的特征空间，使得样本在这个特征空间内线性可分[13]。但是这种单核学习的核函数只有1个，其结构单一，为了将数据映射到不同的高维空间[14]中从而使得这些数据特征能够更好地表达[15]，本文将多核学习与SMLR算法结合起来，通过融合多个核函数形成多核稀疏多元逻辑回归算法(Multiple Kernels Sparse Multinomial Logistic Regression algorithm, MKSMLR)[16]。目前，多核学习的经典方法有SimpleMKL[17,18]和SPF-GMKL[19,20]等，而基于中心对齐的学习核算法(Algorithms for Learning Kernels Based on Centered Alignment)是当前比较流行的多核学习方法。该方法通过中心对齐思想学习能得到一组核函数的权重系数，由此可以更灵活地组合出新的核函数[21]。

针对非线性分类问题，除了本文使用到的核方法外，还可以采用核PCA[22]，与核方法类似，都是将数据特征从低维度空间转化到高维度空间从而达到线性可分的目的。一般这种操作都是在分类器进行学习前，先对数据进行的一种预处理方法。而本文则是将核方法嵌套到了算法的学习过程中，在每次优化损失函数的时候都会引入核技巧，这样能更好地对数据进行映射和充分地表达。除此之外，神经网络由于引入了非线性的激活函数，因此也能很好地解决非线性问题[23]，但是训练一个神经网络需要大量的样本数据，所以对于小批量的数据集，训练效果往往不尽如人意。综上，对于小批量的甚至是稀疏的数据集，多核学习是目前为止较为有效可行的解决方法。

而对于稀疏优化问题，本文采用快速迭代软阈值收缩算法(Fast Iterative Shrinkage-Thresholding Algorithm, FISTA)去优化求解[24]。考虑到多核稀疏多元逻辑回归算法的适用性，本文在不同类型和不同规模的数据集上对各种分类算法进行了分类精度的比较，给出了如何根据不同的场景来选择合适的优化算法的一些建议。本文所做的工作包括以下3点：(1)提出了基于多核学习的稀疏多元逻辑回归分类算法。(2)给出了在线性不可分情况下稀疏多类别分类任务的求解建议。(3)针对不同类型和不同规模数据的场景，比较了多核学习对分类效果的影响。

2 多核稀疏多元逻辑回归模型

本节首先介绍基于单核的稀疏多元逻辑回归算法，接着介绍基于中心对齐的多核学习算法，最后将多核学习和稀疏多元逻辑回归算法相结合，提出基于中心对齐多核学习的稀疏多元逻辑回归算法。

2.1 核稀疏多元逻辑回归算法

假设数据集$\boldsymbol{D}=\{\boldsymbol{X},\boldsymbol{Y}\}$包含$m$个样本，$n$个特征，其中$\boldsymbol{X}\in R^{m\times n}$, $\boldsymbol{Y}\in R^{k\times m}$, k为类别数。$\boldsymbol{Y}$为One-Hot编码后的标签矩阵，即每个标签对应一个k维向量，该向量中仅有1个元素取值为1，其他元素取值均为0。对于给定一个样本$\boldsymbol{X}^{(i)}\in R^{1\times n}$和标签$\boldsymbol{Y}^{(i)}\in R^{k}$, SMLR的损失函数[25]：

$$L(\boldsymbol{W})=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}1\left\{\boldsymbol{Y}_j^{(i)}=1\right\}\lg p\left(\boldsymbol{Y}_j^{(i)}=1|\boldsymbol{X}^{(i)};\boldsymbol{W}\right)+\lambda\|\boldsymbol{W}\|_1 \tag{1}$$

其中，$1\{\boldsymbol{Y}_j^{(i)}=1\}$为示性函数，且有$1\{\boldsymbol{Y}_j^{(i)}=1\}=1$, $1\{\boldsymbol{Y}_j^{(i)}\neq 1\}=0$。

一般而言，解的稀疏性通常通过引入ℓ_1范数来获得，但求解ℓ_1问题是NP难的组合优化问题。因此，通常的做法是对ℓ_0问题进行松弛优化，即求解凸优化ℓ_1问题。本文采用迭代软阈值收缩算法(Iterative Shrinkage Thresholding Algorithm, ISTA)，也称作近端梯度下降法(Proximal Gradient Method, PGM)来求解稀疏优化问题[26]。对于SMLR，优化目标式(1)可以写成

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}L(\boldsymbol{W})=l(\boldsymbol{W})+\lambda\|\boldsymbol{W}\|_1 \tag{2}$$

其中，$\boldsymbol{W}\in R^{n\times k}$。直接对式(2)进行求解并不容易，因此可以先将其转换为比较容易求解的形式。本文将$l(\boldsymbol{W})$在$\boldsymbol{W}^t$处做2阶泰勒展开，并取黑塞矩阵$\boldsymbol{H}_l(\boldsymbol{W}^t)=\boldsymbol{I}/\tau$，其中$\boldsymbol{I}\in R^{n\times n}$的单位矩阵，变量$\tau$为步长。则有

$$l(\boldsymbol{W})\approx l(\boldsymbol{W}^t)+(\boldsymbol{W}-\boldsymbol{W}^t)\nabla_l(\boldsymbol{W}^t)+\frac{1}{2\tau}(\boldsymbol{W}-\boldsymbol{W}^t)^{\mathrm{T}}(\boldsymbol{W}-\boldsymbol{W}^t) \tag{3}$$

那么SMLR的目标函数就可以重新写成

$$\hat{l}(\boldsymbol{W},\boldsymbol{W}^t)=l(\boldsymbol{W}^t)+(\boldsymbol{W}-\boldsymbol{W}^t)\boldsymbol{\nabla}_l(\boldsymbol{W}^t)+\frac{1}{2\tau}(\boldsymbol{W}-\boldsymbol{W}^t)^{\mathrm{T}}(\boldsymbol{W}-\boldsymbol{W}^t)+\lambda\|\boldsymbol{W}\|_1 \tag{4}$$

于是，最小化问题变为

$$\boldsymbol{W}^{t+1}=\underset{\boldsymbol{W}}{\operatorname{argmin}}\hat{l}(\boldsymbol{W},\boldsymbol{W}^t) \tag{5}$$

通过简单的代数变换，最小化式(5)可以被重写为

$$p_\tau(\boldsymbol{W}^t)=\underset{\boldsymbol{W}}{\operatorname{argmin}}\frac{1}{2\tau}\left\|\boldsymbol{W}-(\boldsymbol{W}^t-\tau\boldsymbol{\nabla}_l(\boldsymbol{W}^t))\right\|_2^2+\lambda\|\boldsymbol{W}\|_1 \tag{6}$$

其中，式(6)忽略了与最小化$\boldsymbol{W}$无关的常量。对于最小化问题式(6)，可以快速地进行求解。

令

$$\boldsymbol{W}'^t=\boldsymbol{W}^t-\tau\nabla_l(\boldsymbol{W}^t) \tag{7}$$

则式(6)的解可以表示为$S_{\lambda\tau}\left(\boldsymbol{W}'^t\right)$。其中，软阈值操作$S_\lambda:R^n\to R^n$被定义为

$$[S_\lambda(a)]_i=\begin{cases}a_i-\lambda, & a_i>\lambda\\ 0, & |a_i|\le\lambda\\ a_i+\lambda & ,a_i<-\lambda\end{cases} \tag{8}$$

其中，下标i表示逐元素执行软阈值操作。

一般而言，步长通常可以取固定值$\tau=1/L$，其中L为利普希茨常数(Lipschitz Constant)。但使用固定步长的一个缺点是利普希茨常数L不总是已知的，在大规模问题下其值也不易计算。因此，Beck等人[24]给出了一个简单的线性搜索策略，步长τ的取值可以通过线性搜索(line search)的方式确定。

为了解决非线性问题，在稀疏多元逻辑回归中，我们引入表达定理：$w^*=\sum_{i=1}^n\alpha_i\phi(x_i)$, 此处的$\phi(x_i)$为$\boldsymbol{X}^{(i)}$在特征空间的投影，于是将特征投影到高维空间有

$$z^{(i)}=\sum_{j=1}^n\alpha^{(j)}\left\langle\phi\left(x^{(i)}\right),\phi\left(x^{(j)}\right)\right\rangle=\sum_{j=1}^n\alpha^{(j)}k(x^{(i)},x^{(j)})=\boldsymbol{k}^{(i)}\boldsymbol{\alpha} \tag{9}$$

其中，$\boldsymbol{k}^{(i)}$是核矩阵$\boldsymbol{k}$的第i行向量，$z^{(i)}$关于$\boldsymbol{X}^{(i)}$是非线性的，关于$\phi(x_i)$是线性的。$\boldsymbol{\alpha}$和$z^{(i)}$之间也是线性关系。核函数相当于使用$\phi(\boldsymbol{X})$预处理所有的输入，然后在高维或无穷维的特征空间中学习线性模型[27]。

于是对于给定一个样本$\boldsymbol{X}^{(i)}\in R^{1\times n}$和标签$\boldsymbol{Y}^{(i)}\in R^k$，样本$\boldsymbol{X}^{(i)}$属于类别$j$的概率可以表示为

$$p\left(\boldsymbol{Y}_j^{(i)}=1|\boldsymbol{X}^{(i)};\boldsymbol{\alpha}\right)=\exp\left(\boldsymbol{k}^{(i)}\boldsymbol{\alpha}_j^{\mathrm{T}}\right)\Big/\sum_{l=1}^k\exp\left(\boldsymbol{k}^{(i)}\boldsymbol{\alpha}_l^{\mathrm{T}}\right) \tag{10}$$

也一定有

$$\sum_{j=1}^k p(\boldsymbol{Y}_j^{(i)}=1|\boldsymbol{X}^{(i)};\boldsymbol{\alpha})=1 \tag{11}$$

同理，将SMLR的损失函数中的所有内积计算都用核函数去替代，于是可以得到KSMLR的损失函数

$$L(\boldsymbol{\alpha})=-\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^k1\left\{\boldsymbol{Y}_j^{(i)}=1\right\}\cdot\lg\left(\exp\left(\boldsymbol{k}^{(i)}\boldsymbol{\alpha}_j^{\mathrm{T}}\right)\Big/\sum_{l=1}^k\exp\left(\boldsymbol{k}^{(i)}\boldsymbol{\alpha}_l^{\mathrm{T}}\right)\right)+\lambda\|\boldsymbol{W}\|_1 \tag{12}$$

通过核技巧，可以在保证有效收敛的条件下用凸优化技术来学习关于$\boldsymbol{X}^{(i)}$的非线性模型。因$\phi(\boldsymbol{X})$是固定不变的，需要求解的变量只有$\boldsymbol{\alpha}$，且有$\boldsymbol{\alpha}\in R^{n\times k}$。针对KSMLR问题，采用回溯ISTA算法进行优化求解，其迭代步骤如算法1所示。

算法1：KSMLR问题的回溯ISTA算法

输入：

初始化步长：$\tau=1/L$, $L>0$,

初始化参数：$\boldsymbol{\alpha}\in R^{n\times k}$，初始化核函数参数：$\sigma=2$,

最大迭代次数：Iter = 500,

回溯参数：$\beta\in(0,1)$

输出：

算法最终的参数：$\boldsymbol{\alpha}^{t+1}$

迭代步骤：

步骤1　由样本$\boldsymbol{X}^{(i)}$计算得到核矩阵$\boldsymbol{k}$;

步骤2　初始化计数器 $t\leftarrow0$;

步骤3　初始化参数$\boldsymbol{\alpha}^{\boldsymbol{t}}\leftarrow\boldsymbol{\alpha}$;

步骤4　$\boldsymbol{\alpha}^{t+1}=p_\tau(\boldsymbol{\alpha}^t)$;

步骤5　$\tau=\beta\tau$;

步骤6　当满足$l(\boldsymbol{\alpha}^{t+1})\le\hat{l}(\boldsymbol{\alpha}^{t+1},\boldsymbol{\alpha}^t)$或迭代到指定次数时算法终止，执行步骤7。否则，令$t\leftarrow t+1$，并返回到步骤4;

步骤7　返回更新完成的算法参数$\boldsymbol{\alpha}^{t+1}$。

2.2 基于中心对齐的多核学习算法

2.2.1 中心对齐定义

如果将样本$\boldsymbol{x}$映射到高维的特征空间上，表示为$\boldsymbol{\phi}(\boldsymbol{x})$，那么中心核函数$K_c$可以定义为

$$\begin{aligned}K_c(\boldsymbol{x},\boldsymbol{x}')&=(\boldsymbol{\phi}(\boldsymbol{x})-E_x[\boldsymbol{\phi}])^{\mathrm{T}}(\boldsymbol{\phi}(\boldsymbol{x}')-E_{x'}[\boldsymbol{\phi}])\\&=\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')-E_x[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')]-E_{x'}[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')]\\&\quad+E_{x,x'}[\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')]\end{aligned}\tag{13}$$

从式(13)定义可以看出，核矩阵$\boldsymbol{K}$不依赖于特征映射规则。

根据中心核函数的定义，可以给出类似的中心核矩阵的定义，假定有样本$\boldsymbol{D}=(x_1,x_2,\cdots,x_m)$，且在高维空间上的特征向量为$\boldsymbol{\phi}(x_i),i\in[1,m]$，通过从中减去经验期望来对齐核矩阵，于是可以写作$\boldsymbol{\phi}(x_i)-\bar{\boldsymbol{\phi}}$，这里$\bar{\boldsymbol{\phi}}=\dfrac{1}{m}\sum_{i=1}^{m}\cdots(x_i)$。于是用$\boldsymbol{K}_c$来替代原来的核矩阵$\boldsymbol{K}$，使得样本$\boldsymbol{D}$是居中的。定义所有的$i,j\in[1,m]$，于是有

$$[\boldsymbol{K}_c]_{ij}=\boldsymbol{K}_{ij}-\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{K}_{ij}-\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{K}_{ij}+\frac{1}{m^2}\sum_{i,j=1}^{m}\boldsymbol{K}_{ij}\tag{14}$$

接着，令$\boldsymbol{\phi}=[\boldsymbol{\phi}(x_1),\boldsymbol{\phi}(x_2),\cdots,\boldsymbol{\phi}(x_m)]$，且$\bar{\boldsymbol{\phi}}=[\bar{\phi},\cdots,\bar{\phi}]$，于是可以得到$\boldsymbol{K}_c=(\boldsymbol{\phi}-\bar{\boldsymbol{\phi}})(\boldsymbol{\phi}-\bar{\boldsymbol{\phi}})^{\mathbf{T}}$，这里$\boldsymbol{K}_c$是一个半正定的矩阵。

另外，跟核函数一样，让$\langle\cdot,\cdot\rangle_F$表示内积操作，$\|\cdot\|_{\mathrm{F}}$表示范数，于是有

$$\forall\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{m\times m},\langle\boldsymbol{A},\boldsymbol{B}\rangle_{\mathrm{F}}=\mathrm{Tr}\left[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}\right]$$

并且有$\|\boldsymbol{A}\|_{\mathrm{F}}=\sqrt{\langle\boldsymbol{A},\boldsymbol{A}\rangle_{\mathrm{F}}}$。

定义一个所有元素都为1的向量，$\boldsymbol{1}\in\mathbb{R}^{m\times1}$，和一个单位矩阵$\boldsymbol{I}$，于是对于任意一个核矩阵$\boldsymbol{K}\in\mathbb{R}^{m\times m}$，这个中心核矩阵都可以写成

$$\boldsymbol{K}_c=\left(\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{m}\right)\boldsymbol{K}\left(\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{m}\right)\tag{15}$$

并且对于任意的两个核矩阵$\boldsymbol{K}$和$\boldsymbol{K}'$，都有

$$\langle\boldsymbol{K}_c,\boldsymbol{K}_c'\rangle_{\mathrm{F}}=\langle\boldsymbol{K},\boldsymbol{K}_c'\rangle_{\mathrm{F}}=\langle\boldsymbol{K}_c,\boldsymbol{K}'\rangle_{\mathrm{F}}\tag{16}$$

2.2.2 内核的线性组合

本文通过中心对齐方法来学习内核的最佳线性组合，假定学习得到了p个不同的核矩阵$\boldsymbol{K}_1,\boldsymbol{K}_2,\cdots,\boldsymbol{K}_p$，于是最终希望得到的多核线性组合为

$$\boldsymbol{K}_{c\mu}=\sum_{q=1}^{p}\mu_q\boldsymbol{K}_{cp}\tag{17}$$

这里$\boldsymbol{K}_{cp}$是中心核矩阵，其中有

$$\boldsymbol{K}_{cp}=\left(\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{n}\right)\boldsymbol{K}_q\left(\boldsymbol{I}-\frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{n}\right)\tag{18}$$

通过以最大化居中对齐核矩阵来优化这个$\boldsymbol{\mu}$，于是优化目标为

$$\max_{\boldsymbol{\mu}}\frac{\langle\boldsymbol{K}_{c\mu},\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\rangle_{\mathrm{F}}}{n\langle\boldsymbol{K}_{c\mu},\boldsymbol{K}_{c\mu}\rangle_{\mathrm{F}}}\tag{19}$$

用Align方法[28]求解得到$\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_p)$,这里$\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$是基于目标向量$\boldsymbol{y}\boldsymbol{y}$的理想内核，$\langle\cdot,\cdot\rangle$是矩阵的内积操作。这里设置的每一个$\mu_q$都是独立的，$\mu_q=\left\langle\mathbf{K}_{c\mu},\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\right\rangle_{\mathrm{F}}^{d}$，而$d$是用户定义的超参数，并且有$\sum\mu_q=1$且$\mu_q>0,\forall q$。

2.3 多核稀疏多元逻辑回归算法

本文通过使用样本$\boldsymbol{x}$的全部或者部分特征并以不同的核函数或同种类型核函数的不同参数来构建p个不同的核矩阵$\boldsymbol{K}_1,\boldsymbol{K}_2,\cdots,\boldsymbol{K}_{\boldsymbol{p}}$，然后利用中心对齐思想和Align方法去求得这$p$个核函数基于中心核函数的最优组合系数$\boldsymbol{\mu}$,其中$\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_p)$。于是最后可以得到一个基于多核学习的核矩阵$\boldsymbol{K}_{c\mu}$，其形式为：$\boldsymbol{K}_{c\mu}=\sum_{q=1}^{p}\mu_q\boldsymbol{K}_{cp}$，这里的$\boldsymbol{K}_{cp}$是中心核矩阵，并有$\sum\mu_q=1$且$\mu_q>0,\forall q$。

于是，对于样本$\boldsymbol{X}^{(i)}\in\mathbb{R}^{1\times n}$和标签$\boldsymbol{Y}^{(i)}\in\mathbb{R}^{k}$，样本$\boldsymbol{X}^{(i)}$属于类别$j$的概率可以表示为

$$\begin{aligned}&p\left(\boldsymbol{Y}_j^{(i)}=1|\boldsymbol{X}^{(i)};\boldsymbol{\alpha}\right)\\&=\exp\left(\boldsymbol{K}_{c\mu}^{(i)}\boldsymbol{\alpha}_j^{\mathrm{T}}\right)\Big/\sum_{l=1}^{k}\exp\left(\boldsymbol{K}_{c\mu}^{(i)}\boldsymbol{\alpha}_l^{\mathrm{T}}\right)\end{aligned}\tag{20}$$

自然地可以得到基于多核学习的MKSMLR损失函数

$$\begin{aligned}L(\boldsymbol{\alpha})=&-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\boldsymbol{1}\left\{\boldsymbol{Y}_j^{(i)}=1\right\}\\&\cdot\lg\left(\exp\left(\boldsymbol{K}_{c\mu}^{(i)}\boldsymbol{\alpha}_j^{\mathrm{T}}\right)\Big/\sum_{l=1}^{k}\exp\left(\boldsymbol{K}_{c\mu}^{(i)}\boldsymbol{\alpha}_l^{\mathrm{T}}\right)\right)\\&+\lambda\|\boldsymbol{W}\|_1\end{aligned}\tag{21}$$

由于引入了多个核矩阵，线性搜索迭代速度会非常的缓慢，为了解决MKSMLR求解速度缓慢的问题，本文采用快速迭代软阈值收缩算法对其进行优化求解。FISTA通常也被称为快速近端梯度法(Fast Proximal Gradient Method, FPGM)，它使用了Nesterov加速策略对原始ISTA算法进行加速，能够将ISTA算法在最坏情况的收敛率由$O(1/T)$优化为$O(1/T^2)$，其中T为迭代次数。ISTA算法使用I/τ来近似估计黑塞矩阵，而FISTA算法则利用了梯度$\boldsymbol{\nabla}_l(\boldsymbol{W}^t)$的最小利普希茨常数来近似估计黑塞矩阵。FISTA与ISTA的另一个区别在于，FISTA在最小化式(5)时并不是只使用了$\boldsymbol{W}^t$，而是使用了前两次参数$\{\boldsymbol{W}^{t-1},\boldsymbol{W}^t\}$的线性组合。基于同样的原因，FISTA也给出了带有回溯的版

本。于是，我们便可以应用基于广泛的快速1阶训练方法FISTA来快速优化求解具有多个核矩阵的MKSMLR，其迭代步骤如算法2所示。

3　实验结果与分析

本节将通过使用不同领域和不同规模的数据集来比较SMLR算法和KSMLR算法以及MKSMLR算法在分类上的效果差异。

3.1　实验设置

本文所有的实验都采用同一台服务器，中央处理器为12核、主频为2.0GHz的Intel(R) Xeon(R) E5-2620，并且具有64GB的随机存取存储器，编程实验环境为Python 2.7。为了有效地比较不同优化算法的性能，实验选取了多个领域的不同大小的数据集，包括非线性特征数据集Banana；多类物体识别数据集COIL20；人脸识别数据集ORL和GT-32；小规模的手写体数字识别数据集MNIST-S；肺部基因数据集Lung；高光谱图像分类数据集Indian-pines以及图像分割数据集Segment。

在以下分类实验中，引入了其他多分类算法与本文的算法进行对比，其中包括：SVM、用于稀疏的线性可分数据集上的稀疏逻辑回归(Sparse Logistic Regression, SLR)算法、采用L2正则的权重衰减多元逻辑回归(Weight-Decay Multinomial Logistic Regression, WDMLR)算法、采用ISTA优化算法求解的稀疏多元逻辑回归(Sparse Multinomial Logistic regression-ISTA, SML-ISTA)算法、采用FISTA优化算法求解的稀疏多元逻辑回归(Sparse Multinomial Logistic regression-FISTA, SML-FISTA)算法和基于单核学习的核稀疏多元逻辑回归(Kernel Sparse Multinomial Logistic Regression, KSMLR)算法。

核函数主要使用径向基函数(Radial Basis Function, RBF), $K(\boldsymbol{x},\bar{\boldsymbol{x}})=\exp\left(-\dfrac{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^2}{2\sigma^2}\right)$，收敛条件参数$\tau=10^{-5}$。对于优化参数$\boldsymbol{\alpha}$，将其初始值取为$\boldsymbol{\alpha}^{(0)}$。另外，超参数$t$和$\alpha$表示线性搜索参数，$\lambda$表示正则惩罚项，$\sigma$为高斯核函数中的参数。其他算法如SVM, SLR和WDMLR的参数均为默认超参数。

3.2　实验结果

实验结果中的分类准确率如表1所示，算法运行时间如表2所示。

3.3　实验分析

从表1的实验结果中我们可以得出以下结论，

算法2：MKSMLR问题的回溯FISTA算法

输入：

初始化步长：$\tau=1/L,\ L>0$,

初始化参数：$\boldsymbol{\alpha}\in R^{n\times k}$,

初始化核函数参数：$\sigma=2$,

最大迭代次数：$\mathrm{Iter}=500$,

回溯参数：$\beta\in(0,1)$

输出：

算法最终的参数：$\boldsymbol{\alpha}^{t+1}$

迭代步骤：

步骤1　由样本$\boldsymbol{X}^{(i)}$计算得到p个不同的核矩阵；

步骤2　用Align方法计算得到多核学习参数$\boldsymbol{\mu}$并生成新的核矩阵$\boldsymbol{K}_{c\mu}$；

步骤3　初始化计数器 $t\leftarrow 0$；

步骤4　初始化参数$\boldsymbol{\alpha}^{\boldsymbol{t}}\leftarrow\boldsymbol{\alpha}$, $\mu^t\leftarrow 1$,$v^t\leftarrow\boldsymbol{\alpha}^{\boldsymbol{t}}$；

步骤5　$\boldsymbol{\alpha}^{t+1}=p_\tau\left(v^t\right)$;

步骤6　$\mu^{t+1}=\dfrac{1+\sqrt{1+4(\mu^t)^2}}{2}$;

步骤7　$v^{t+1}=\boldsymbol{\alpha}^{t+1}+\dfrac{\mu^t-1}{\mu^{t+1}}(\boldsymbol{\alpha}^{t+1}-\boldsymbol{\alpha}^t)$;

步骤8　$\tau=\beta\tau$;

步骤9　当满足$l\left(\alpha^{t+1}\right)\le\hat{l}\left(\alpha^{t+1},\ \alpha^t\right)$或迭代到指定次数时算法终止，执行步骤10。否则，令$t\leftarrow t+1$，并返回到步骤5；

步骤10　返回更新完成的算法参数$\boldsymbol{\alpha}^{t+1}$。

表 1　分类准确率

数据集	SVM	SLR	WDMLR	SML-ISTA	SML-FISTA	KSMLR	MKSMLR
Banana	0.9069	–	–	–	–	0.9069	**0.9107**
COIL20	0.8032	0.9676	0.9832	0.9895	0.9958	0.9977	**1**
ORL	0.9507	0.9420	**0.9545**	0.9242	**0.9545**	0.9000	0.9167
GT-32	–	–	0.7823	0.7580	0.7621	**0.8044**	**0.8044**
MNIST-S	0.9113	0.9001	0.9109	0.9036	0.9048	0.9360	**0.9400**
Lung	0.7705	0.9344	0.9104	0.9104	0.9254	0.9180	**0.9344**
Indian-pines	0.7980	0.8182	0.7599	0.8120	0.8120	0.8218	**0.8237**
Segment	0.5989	0.9235	0.8268	0.8925	0.9253	0.9538	**0.9567**

注：表中的“–”符号表示未能正确分类或分类效果接近于随机选择。

表 2 算法运行时间(s)

数据集	SML-ISTA	SML-FISTA	KSMLR	MKSMLR
Banana	–	–	0.78	**1.19**
COIL20	1.71	**0.39**	7.61	13.46
ORL	142.05	7.5	10.43	**2.73**
GT-32	88.19	**2.03**	37.94	10.77
MNIST-S	**0.12**	0.14	0.14	22.98
Lung	42.71	**1.4**	2.12	3.08
Indian-pines	427.62	**18.58**	68.31	909.1
Segment	21.33	20.71	**13.68**	33.35

注：表中的“–”符号表示未能正确分类或分类效果接近于随机选择。

对于非线性数据集Banana，常规的分类算法如SLR和WDMLR都无法将其正确分类，只有通过核技巧扩充后的算法诸如带高斯核的SVM, KSMLR和MKSMLR算法才能进行正常分类。另外，对于大多数的线性数据集，使用核技巧以后都能提高其分类器的精度，而对于数据集ORL和肺部基因数据集Lung，单核学习就表现得不如其他常规算法，但是在剩余的数据集上都能一定程度地提高分类效果。原因是数据集ORL和Lung都是特征数远大于样本数的数据集，因此对于这类数据集并不适用于核方法。采用多核学习的MKSMLR算法由于在核函数的选择上更加灵活，所以几乎在所有数据集上的表现都优于其他常规算法以及基于单核学习的KSMLR算法。对于某些具有稀疏特性的数据集诸如人脸识别数据集ORL和GT-32以及小规模的手写体数字识别数据集MNIST-S，采用稀疏优化算法ISTA求解相比如其他优化算法能得到更好的分类效果。

在核方法中，通常会引入一个稠密的核矩阵，其存储和计算代价都非常的高，存储稠密矩阵需要$O\left(m^2\right)$的空间，而计算这样的矩阵则需要$O\left(m^2n\right)$的代价，这里m和n分别代表了样本的个数和维度。稀疏优化算法ISTA在最坏的情况下的收敛率为$O(1/T)$，而FISTA优化算法采用了Nesterov加速策略并利用了梯度$\nabla_l\left(\boldsymbol{W}^t\right)$的最小利普希茨常数来近似估计黑塞矩阵，因此FISTA在最坏情况下的收敛率则为$O(1/T^2)$，这里T为迭代次数。对于多核学习，在单核学习的基础上，由于引入了多个核矩阵，于是为了在保证能得到稀疏解的同时减少算法的时间开销，所以对多核学习算法KSMLR采用的是FISTA优化算法。从实验结果表1和表2都能看出，FISTA优化求解算法运行效率都要高过于ISTA优化求解算法。而对于多核学习，由于需要求解多个核矩阵，所以相比于单核学习需要花费更多的运行时间，但是由于多核学习能学习到更多有用的特征，可能导致收敛得更快，因此迭代次数要少很多，最终在某些如ORL, GT-32等数据集上，总的运行时间相比于单核学习反而更少了。

4 结束语

本文提出了一种基于中心对齐多核学习的稀疏多元逻辑回归算法，通过核技巧解决了稀疏多元逻辑回归不能用于非线性数据分类的问题。然后在核函数的选择上，利用了中心对齐多核学习算法去灵活地选取核函数的权重系数，用不同核函数的线性组合去生成新的核函数，从而将数据映射到不同维度的高维空间中。在优化算法中，考虑到多核学习需要计算多个核矩阵，存在一定的时间开销，于是采用FISTA优化算法进行快速求解。最后在稀疏的人脸识别数据集以及其他领域的公共数据集上，本文提出的MKSMLR算法在分类准确率指标上都优于目前常规的分类算法。

参 考 文 献

[1] ZHOU Changjun, WANG Lan, ZHANG Qiang, *et al*. Face recognition based on PCA and logistic regression analysis[J]. *Optik*, 2014, 125(20): 5916–5919. doi: 10.1016/j.ijleo.2014.07.080.

[2] WARNER P. Ordinal logistic regression[J]. *Journal of Family Planning and Reproductive Health Care*, 2008, 34(3): 169–170. doi: 10.1783/147118908784734945.

[3] LIU Wu, FOWLER J E, and ZHAO Chunhui. Spatial logistic regression for support-vector classification of hyperspectral imagery[J]. *IEEE Geoscience and Remote Sensing Letters*, 2017, 14(3): 439–443. doi: 10.1109/LGRS.2017.2648515.

[4] ABRAMOVICH F and GRINSHTEIN V. High-dimensional classification by sparse logistic regression[J]. *IEEE Transactions on Information Theory*, 2019, 65(5): 3068–3079. doi: 10.1109/TIT.2018.2884963.

[5] CARVALHO C M, CHANG J, LUCAS J E, *et al*. High-dimensional sparse factor modeling: Applications in gene expression genomics[J]. *Journal of the American Statistical Association*, 2008, 103(484): 1438–1456. doi: 10.1198/016214508000000869.

[6] GALAR M, FERNÁNDEZ A, BARRENECHEA E, *et al*. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes[J]. *Pattern Recognition*, 2011, 44(8): 1761–1776. doi: 10.1016/j.patcog.2011.01.017.

[7] 曾志强, 吴群, 廖备水, 等. 一种基于核SMOTE的非平衡数据集分类方法[J]. 电子学报, 2009, 37(11): 2489–2495. doi: 10.3321/j.issn:0372-2112.2009.11.024.

ZENG Zhiqiang, WU Qun, LIAO Beishui, *et al*. A classfication method for imbalance data set based on kernel SMOTE[J]. *Acta Electronica Sinica*, 2009, 37(11): 2489–2495. doi: 10.3321/j.issn:0372-2112.2009.11.024.

[8] CAO Faxian, YANG Zhijing, REN Jinchang, *et al*. Extreme sparse multinomial logistic regression: A fast and robust framework for hyperspectral image classification[J]. *Remote Sensing*, 2017, 9(12): 1255. doi: 10.3390/rs9121255.

[9] LIU Tianzhu, GU Yanfeng, JIA Xiuping, *et al*. Class-specific sparse multiple kernel learning for spectral–spatial hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(12): 7351–7365. doi: 10.1109/TGRS.2016.2600522.

[10] FANG Leyuan, WANG Cheng, LI Shutao, *et al*. Hyperspectral image classification via multiple-feature-based adaptive sparse representation[J]. *IEEE Transactions on Instrumentation and Measurement*, 2017, 66(7): 1646–1657. doi: 10.1109/TIM.2017.2664480.

[11] OUYED O and ALLILI M S. Feature weighting for multinomial kernel logistic regression and application to action recognition[J]. *Neurocomputing*, 2018, 275: 1752–1768. doi: 10.1016/j.neucom.2017.10.024.

[12] 徐金环, 沈煜, 刘鹏飞, 等. 联合核稀疏多元逻辑回归和TV-L1错误剔除的高光谱图像分类算法[J]. 电子学报, 2018, 46(1): 175–184. doi: 10.3969/j.issn.0372-2112.2018.01.024.
XU Jinhuan, SHEN Yu, LIU Pengfei, *et al*. Hyperspectral image classification combining kernel sparse multinomial logistic regression and TV-L1 error rejection[J]. *Acta Electronica Sinica*, 2018, 46(1): 175–184. doi: 10.3969/j.issn.0372-2112.2018.01.024.

[13] SCHÖLKOPF B and SMOLA A J. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond[M]. Cambridge: MIT Press, 2002.

[14] 汪洪桥, 孙富春, 蔡艳宁, 等. 多核学习方法[J]. 自动化学报, 2010, 36(8): 1037–1050. doi: 10.3724/SP.J.1004.2010.01037.
WANG Hongqiao, SUN Fuchun, CAI Yanning, *et al*. On multiple kernel learning methods[J]. *Acta Automatica Sinica*, 2010, 36(8): 1037–1050. doi: 10.3724/SP.J.1004.2010.01037.

[15] GÖNEN M and ALPAYDIN E. Multiple kernel learning algorithms[J]. *Journal of Machine Learning Research*, 2011, 12: 2211–2268.

[16] GU Yanfeng, LIU Tianzhu, JIA Xiuping, *et al*. Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(6): 3235–3247. doi: 10.1109/TGRS.2015.2514161.

[17] RAKOTOMAMONJY A, BACH F R, CANU S, *et al*. SimpleMKL[J]. *Journal of Machine Learning Research*, 2008, 9: 2491–2521.

[18] LOOSLI G and ABOUBACAR H. Using SVDD in SimpleMKL for 3D-Shapes filtering[C]. CAp - Conférence D'apprentissage, Saint-Etienne, 2017. doi: 10.13140/2.1.3091.3605.

[19] JAIN A, VISHWANATHAN S V N, and VARMA M. SPF-GMKL: Generalized multiple kernel learning with a million kernels[C]. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012: 750–758. doi: 10.1145/2339530.2339648.

[20] BAHMANI S, BOUFOUNOS P T, and RAJ B. Learning model-based sparsity via projected gradient descent[J]. *IEEE Transactions on Information Theory*, 2016, 62(4): 2092–2099. doi: 10.1109/TIT.2016.2515078.

[21] CORTES C, MOHRI M, and ROSTAMIZADEH A. Algorithms for learning kernels based on centered alignment[J]. *Journal of Machine Learning Research*, 2012, 13(28): 795–828.

[22] CHENG Chunyuan, HSU C C, and CHENG Muchen. Adaptive kernel principal component analysis (KPCA) for monitoring small disturbances of nonlinear processes[J]. *Industrial & Engineering Chemistry Research*, 2010, 49(5): 2254–2262. doi: 10.1021/ie900521b.

[23] YANG Hongjun and LIU Jinkun. An adaptive RBF neural network control method for a class of nonlinear systems[J]. *IEEE/CAA Journal of Automatica Sinica*, 2018, 5(2): 457–462. doi: 10.1109/JAS.2017.7510820.

[24] BECK A and TEBOULLE M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. *SIAM Journal on Imaging Sciences*, 2009, 2(1): 183–202. doi: 10.1137/080716542.

[25] KRISHNAPURAM B, CARIN L, FIGUEIREDO M A T, *et al*. Sparse multinomial logistic regression: Fast algorithms and generalization bounds[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(6): 957–968. doi: 10.1109/tpami.2005.127.

[26] CHEN Xi, LIN Qihang, KIM S, *et al*. Smoothing proximal gradient method for general structured sparse regression[J]. *The Annals of Applied Statistics*, 2012, 6(2): 719–752. doi: 10.1214/11-aoas514.

[27] LECUN Y, BENGIO Y and HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436–444. doi: 10.1038/nature14539.

[28] PÉREZ-ORTIZ M, GUTIÉRREZ P A, SÁNCHEZ-MONEDERO J, *et al*. A study on multi-scale kernel optimisation via centered kernel-target alignment[J]. *Neural Processing Letters*, 2016, 44(2): 491–517. doi: 10.1007/s11063-015-9471-0.

雷大江：男，1979年生，副教授，研究方向为机器学习.
唐建烊：男，1993年生，硕士生，研究方向为核机器学习.
李智星：男，1985年生，副教授，研究方向为自然语言处理.
吴　渝：女，1970年生，教授，研究方向为网络智能.

责任编辑：陈　倩