Journal of Computer Applications
计算机应用, 2020, 40(7): 1896 - 1900
ISSN 1001-9081
CODEN JYIIDU
2020-07-10
http://www.joca.cn

文章编号:1001-9081(2020)07-1896-05
DOI:10.11772/j.issn.1001-9081.2019122075

基于正则互表示的无监督特征选择方法

汪志远,降爱莲*,奥斯曼·穆罕默德
(太原理工大学 信息与计算机学院,山西 晋中 030600)
(*通信作者电子邮箱ailianjiang@126.com)

摘　要:针对高维数据含有的冗余特征影响机器学习训练效率和泛化能力的问题,为提升模式识别准确率、降低计算复杂度,提出了一种基于正则互表示(RMR)性质的无监督特征选择方法。首先,利用特征之间的相关性,建立由Frobenius范数约束的无监督特征选择数学模型;然后,设计分治-岭回归优化算法对模型进行快速优化;最后,根据模型最优解综合评估每个特征的重要性,选出原始数据中具有代表性的特征子集。在聚类准确率指标上,RMR方法与Laplacian方法相比提升了7个百分点,与非负判别特征选择(NDFS)方法相比提升了7个百分点,与正则自表示(RSR)方法相比提升了6个百分点,与自表示特征选择(SR_FS)方法相比提升了3个百分点;在数据冗余率指标上,RMR方法与Laplacian方法相比降低了10个百分点,与NDFS方法相比降低了7个百分点,与RSR方法相比降低了3个百分点,与SR_FS方法相比降低了2个百分点。实验结果表明,RMR方法能够有效地选出重要特征,降低数据冗余率,提升样本聚类准确率。

关键词:特征选择;无监督学习;分治算法;岭回归;正则化
中图分类号:TP181　　**文献标志码:**A

Unsupervised feature selection method based on regularized mutual representation

WANG Zhiyuan, JIANG Ailian*, Osman MUHAMMAD
(*College of Information and Computer, Taiyuan University of Technology, Jinzhong Shanxi* 030600, *China*)

Abstract: The redundant features of high-dimensional data affect the training efficiency and generalization ability of machine learning. In order to improve the accuracy of pattern recognition and reduce the computational complexity, an unsupervised feature selection method based on Regularized Mutual Representation (RMR) property was proposed. Firstly, the correlations between features were utilized to establish a mathematical model for unsupervised feature selection constrained by Frobenius norm. Then, a divide-and-conquer ridge regression optimization algorithm was designed to quickly optimize the model. Finally, the importances of the features were jointly evaluated according to the optimal solution to the model, and a representative feature subset was selected from the original data. On the clustering accuracy, RMR method is improved by 7 percentage points compared with the Laplacian method, improved by 7 percentage points compared with the Nonnegative Discriminative Feature Selection (NDFS) method, improved by 6 percentage points compared with the Regularized Self-Representation (RSR) method, and improved by 3 percentage points compared with the Self-Representation Feature Selection (SR_FS) method. On the redundancy rate, RMR method is reduced by 10 percentage points compared with the Laplacian method, reduced by 7 percentage points compared with the NDFS method, reduced by 3 percentage points compared with the RSR method, and reduced by 2 percentage points compared with the SR_FS method. The experimental results show that RMR method can effectively select important features, reduce redundancy rate of data and improve clustering accuracy of samples.

Key words: feature selection; unsupervised learning; divide-and-conquer algorithm; ridge regression; regularization

0　引言

各种智能电子设备和信息系统的广泛使用产生和收集了大量的高维无标签数据。利用相应的机器学习算法对这些数据进行处理和分析,可以发掘数据的潜在价值。在模式识别与机器学习研究领域中,对于机器学习算法而言,样本数据集是训练模型的重要前提,但是收集到的原始数据通常含有大量的冗余特征,导致其并不适合直接被使用。冗余特征不必要地增加了数据的维数,降低了训练性能[1]。如果不剔除数据中的冗余特征,将会导致机器学习模型的训练时间变长,训练后得到的预测模型泛化能力差,甚至因遭遇维数灾难而使训练无法进行。文献[2]表明,去除冗余特征可以明显地提升训练效率。

特征选择是一种有效的数据降维方法[3],它可以选出数据中的重要特征,剔除不重要的冗余特征,从而得到更加低维

收稿日期:2019-12-09;**修回日期:**2020-02-24;**录用日期:**2020-03-03。　**基金项目:**山西省回国留学人员科研资助项目(2017-051)。
作者简介:汪志远(1992—),男,安徽宿州人,硕士研究生,主要研究方向:机器学习、特征选择;　降爱莲(1969—),女,山西太原人,副教授,博士,CCF会员,主要研究方向:人工智能、大数据、特征选择、计算机视觉;　奥斯曼·穆罕默德(1993—),男,埃塞俄比亚人,硕士研究生,主要研究方向:深度学习、图像处理。

的数据。数据维数降低,可以减少机器学习模型的训练时间,提高模型训练效率[4]。根据学习任务类型的不同,特征选择方法分为有监督特征选择方法和无监督特征选择方法两大类[5]。有监督特征选择方法依据特征与标签之间的相关性来选出重要的特征子集,因为重要特征与标签之间的相关性更强。然而,并非所有的数据集都带有类别标签信息,数据无标签使得有监督特征选择方法不再适用,也使得特征选择变得更加困难[6-7]。

较早提出的无监督特征选择方法,如最小化方差法[8]和Trace Ratio方法[9],这些方法单独计算每个特征的分数,按分数排名选出特征,此类方法评价标准单一,对数据的解释能力较差。之后发展出基于谱分析技术[10]的无监督特征选择方法,此类方法通过对邻接图拉普拉斯矩阵进行谱分解,利用关联矩阵的特征值度量各个特征的重要性,但是该方法在处理高维数据时面临计算量过大的问题。目前,正则回归也被应用到特征选择,Zheng等[11]提出的自步正则化无监督特征选择方法(Unsupervised Feature Selection by Self-Paced Learning Regularization, UFS_SP)和刘艳芳等[12]提出的邻域保持学习特征选择方法(Neighborhood Preserving Learning Feature Selection, NPLFS)将特征选择问题建模为损失函数最小化问题,对权重矩阵施加正则约束,表现出较好的鲁棒性,但是现有的正则特征选择方法优化困难,计算复杂度较高。

正则自表示(Regularized Self-Representation, RSR)方法[13]首次提出特征自表示性质,即高维数据的每个特征可由全部特征线性近似表示,该性质考虑了特征间的相关性,具有较强的数据解释能力,但该理论也存在不足之处,在目标函数优化时,特征权重容易向自身倾斜,导致无法合理地为特征分配权重。其他的基于自表示性质的特征选择方法均无法避免自表示性质的缺陷。为此,本文提出特征互表示性质,即高维数据中的每个特征可由除该特征之外的其他特征线性近似表示,而不是由全体特征线性近似表示,这克服了特征权重容易向自身倾斜的缺点。然后,基于特征互表示性质,提出一种新的正则化无监督特征选择方法,该方法能够有效提升数据聚类准确率,降低数据冗余率,并且计算复杂度较低。

1　特征互表示性质

假定$\boldsymbol{X}$是数据矩阵,$\boldsymbol{X}\in\mathbf{R}^{m\times n}$,$m$是样本数量,$n$是特征数量,$\boldsymbol{f}_i$代表$\boldsymbol{X}$的第$i$个特征,则$\boldsymbol{X}$可以表示为特征的集合:$\boldsymbol{X}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$。

高维数据中的每个特征可以由数据中的其他特征线性近似表示,特征之间存在相关性,因此可以很好地互相近似表示,把高维数据所满足的这一数学性质称为特征互表示性质。利用特征互表示性质,数据矩阵$\boldsymbol{X}$中的每一个特征$\boldsymbol{f}_i$可形式化表示为:

$$\boldsymbol{f}_i=\sum_{j=1\sim n\,\&\,j\neq i}\boldsymbol{f}_j w_j+\boldsymbol{e}_i \tag{1}$$

式中:w_j是第j个特征的权重系数;$\boldsymbol{e}_i$为残差向量,代表特征$\boldsymbol{f}_i$重构前后的残差。本文期望重构后的特征可以很好地近似表示原特征,所以残差应当尽可能小,使用向量的l_2-范数的平方度量残差大小,将损失函数定义为:

$$L(\boldsymbol{e}_i)=\left\|\boldsymbol{f}_i-\sum_{j=1\sim n,j\neq i}\boldsymbol{f}_j w_j\right\|_2^2$$

2　特征选择模型及其优化算法

2.1　正则互表示无监督特征选择模型

通过将$\boldsymbol{X}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$中的每一个特征进行重构,可得如下结果:

$$\boldsymbol{X}=\boldsymbol{X}\boldsymbol{W}^{\circ}+\boldsymbol{E} \tag{2}$$

式中:$\boldsymbol{W}^{\circ}$为权重矩阵,上标$^{\circ}$是指权重矩阵$\boldsymbol{W}$的对角线元素都为0,这是因为特征不参与自身的线性近似表示,所以权重系数为0;$\boldsymbol{E}$为残差矩阵,表示原始数据$\boldsymbol{X}$重构前后的残差。

从向量的角度分析,残差矩阵$\boldsymbol{E}$由残差向量的有序集合$\{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3,\cdots,\boldsymbol{e}_n\}$组成,每个特征重构前后的残差使用$\left\|\boldsymbol{e}_i\right\|_2^2$度量。将残差度量标准由向量推广到矩阵时,可以使用$\|\boldsymbol{E}\|_{\mathrm{F}}^2$度量原始数据$\boldsymbol{X}$重构前后的残差大小。本文期望重构后的数据$\boldsymbol{X}\boldsymbol{W}^{\circ}$能够很好地近似表示原始数据$\boldsymbol{X}$,因此残差越小越好,对应的目标函数为:

$$\hat{\boldsymbol{W}}=\arg\min_{\boldsymbol{W}^{\circ}}\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2 \tag{3}$$

目标函数式(3)的优化问题属于无偏估计回归问题,在无偏估计回归问题中,当数据矩阵$\boldsymbol{X}$不是列满秩的,或者某些列之间线性相关性比较强时,得到的最优解将会不稳定。为了解决这个问题,在目标函数中加入一个正则项$\lambda\left\|\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2$对权重矩阵的解空间进行约束,使计算出的最优解更稳定。基于上述分析,改进后的目标函数为:

$$\hat{\boldsymbol{W}}=\arg\min_{\boldsymbol{W}^{\circ}}\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2+\lambda\left\|\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2 \tag{4}$$

本文把包含权重矩阵正则项的目标函数式(4)称为正则互表示(Regularized Mutual Representation, RMR)无监督特征选择模型。

RMR模型的最优解$\hat{\boldsymbol{W}}$可以用于识别原始数据$\boldsymbol{X}$中具有代表性的特征,这用到了$\hat{\boldsymbol{W}}$具有的数学性质:权重矩阵$\hat{\boldsymbol{W}}$的第i行与数据矩阵$\boldsymbol{X}$的第i个特征是相互对应的,因为$\hat{\boldsymbol{W}}$中第i行的每个元素都是$\boldsymbol{X}$中第i个特征在近似表示其他特征时的权重系数,权重系数越大,该特征越重要。反之亦然,如果某个特征$\boldsymbol{f}_i$是重要的,或者说它在所有特征中更具代表性,那么与该特征对应的权重矩阵$\hat{\boldsymbol{W}}$的第i行也将更大。可以用$\hat{\boldsymbol{W}}$的行向量l_2-范数反映对应特征的重要程度,这是一种度量特征重要性的联立式评估指标,因为它考虑了特征间的相关性,这一指标与传统特征选择方法所采用的评估指标有明显的区别,因为传统方法采用的是孤立式的评估指标,未考虑特征间的相关性。因此,RMR特征选择模型具有更强的数据解释能力,能够有效地识别原始数据中具有代表性的特征子集。

2.2　分治-岭回归优化算法

针对提出的RMR特征选择模型,为之设计一种高效的优化算法,该算法结合了分治算法[14]和岭回归[15]优化算法,本文称之为分治-岭回归优化算法。本节首先证明以下定理。

定理1　给定如下两个优化问题的目标函数:

$$\hat{\boldsymbol{W}}=\arg\min_{\boldsymbol{W}^{\circ}}\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2+\lambda\left\|\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2$$

$$\hat{\boldsymbol{\omega}}_i=\arg\min_{\boldsymbol{\omega}_i}\left\|\boldsymbol{f}_i-\boldsymbol{X}\boldsymbol{\omega}_i\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i\right\|_2^2$$

其中:$\boldsymbol{f}_i$是$\boldsymbol{X}$的第i列;$\hat{\boldsymbol{\omega}}_i$是$\boldsymbol{W}^{\circ}$的第$i$列。则整体优化问题的最优解$\hat{\boldsymbol{W}}$可通过计算所有岭回归子优化问题的最优解$\hat{\boldsymbol{\omega}}_i$得到。

证明　将数据矩阵$\boldsymbol{X}$按列数分解为特征的有序集合$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$,将权重矩阵按列数分解为权重向量的有序集

合$\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\}$,则有:

$$\hat{\boldsymbol{W}}=\arg\min_{\boldsymbol{W}^{\circ}}\left\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2+\lambda\left\|\boldsymbol{W}^{\circ}\right\|_{\mathrm{F}}^2=$$

$$\arg\min_{\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]}\left\|\left[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\right]-\boldsymbol{X}\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]\right\|_{\mathrm{F}}^2+$$

$$\lambda\left\|\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]\right\|_{\mathrm{F}}^2=$$

$$\arg\min_{\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]}\left\|\left[\boldsymbol{f}_1-\boldsymbol{X}\boldsymbol{\omega}_1,\boldsymbol{f}_2-\boldsymbol{X}\boldsymbol{\omega}_2,\cdots,\boldsymbol{f}_n-\boldsymbol{X}\boldsymbol{\omega}_n\right]\right\|_{\mathrm{F}}^2+$$

$$\lambda\left\|\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]\right\|_{\mathrm{F}}^2$$

因为矩阵的Frobenius范数的平方等于矩阵中所有列向量l_2-范数平方的和,因此上式中$\hat{\boldsymbol{W}}$可继续变形为:

$$\hat{\boldsymbol{W}}=\arg\min_{\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]}\left(\left\|\boldsymbol{f}_1-\boldsymbol{X}\boldsymbol{\omega}_1\right\|_2^2+\left\|\boldsymbol{f}_2-\boldsymbol{X}\boldsymbol{\omega}_2\right\|_2^2+\cdots+\left\|\boldsymbol{f}_n-\boldsymbol{X}\boldsymbol{\omega}_n\right\|_2^2\right)+\lambda\left(\left\|\boldsymbol{\omega}_1\right\|_2^2+\left\|\boldsymbol{\omega}_2\right\|_2^2+\cdots+\left\|\boldsymbol{\omega}_n\right\|_2^2\right)=$$

$$\arg\min_{\left[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_n\right]}\left(\left\|\boldsymbol{f}_1-\boldsymbol{X}\boldsymbol{\omega}_1\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_1\right\|_2^2\right)+\left(\left\|\boldsymbol{f}_2-\boldsymbol{X}\boldsymbol{\omega}_2\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_2\right\|_2^2\right)+\cdots+\left(\left\|\boldsymbol{f}_n-\boldsymbol{X}\boldsymbol{\omega}_n\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_n\right\|_2^2\right)=$$

$$\sum_{i=1}^{n}\arg\min_{\boldsymbol{\omega}_i}\left(\left\|\boldsymbol{f}_i-\boldsymbol{X}\boldsymbol{\omega}_i\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i\right\|_2^2\right)$$

至此可知,只要使得分解后的每个岭回归子优化问题都取得最优解,即可保证整体优化问题取得最优解。

定理1证毕

分治-岭回归优化算法首先利用分治算法的思想,将整体优化问题分解为若干个子优化问题;然后针对每个子优化问题利用岭回归优化算法计算最优解;最后将各个子优化问题的最优解进行整合,得到整体优化问题的最优解。

分治-岭回归优化算法的具体计算步骤如下。

首先,将RMR模型的整体优化问题按照数据矩阵$\boldsymbol{X}$的每个特征,分解为如下的n个岭回归子优化问题:

$$\hat{\boldsymbol{\omega}}_i=\arg\min_{\boldsymbol{\omega}_i}\left\|\boldsymbol{f}_i-\boldsymbol{X}\boldsymbol{\omega}_i\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i\right\|_2^2\tag{5}$$

式中:$i=1,2,\cdots,n$。第i个子优化问题的最优解$\hat{\boldsymbol{\omega}}_i$用于重构特征$\boldsymbol{f}_i$,使得重构前后的误差最小;$\hat{\boldsymbol{\omega}}_i$的第$i$个元素为0,因为特征$\boldsymbol{f}_i$不参与自身的线性近似表示。

在实际计算各个子问题最优解$\hat{\boldsymbol{\omega}}_i$的过程中,已知原始数据$\boldsymbol{X}$中的每个特征都不参与自身的线性近似表示,因此可以将子优化问题的目标函数进行变换,得到如下表达式:

$$\hat{\boldsymbol{\omega}}_i'=\arg\min_{\boldsymbol{\omega}_i'}\left\|\boldsymbol{f}_i-\boldsymbol{X}_i\boldsymbol{\omega}_i'\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i'\right\|_2^2\tag{6}$$

式(6)中的$\boldsymbol{X}_i$是由$\boldsymbol{X}$去掉特征$\boldsymbol{f}_i$得到,式(6)中的$\boldsymbol{\omega}_i'$由式(5)中的$\boldsymbol{\omega}_i$删除第i个元素得到。目标函数式(6)的优化问题是岭回归优化问题,该优化问题的损失函数以及推导过程如下:

$$L(\boldsymbol{e}_i)=\left\|\boldsymbol{f}_i-\boldsymbol{X}_i\boldsymbol{\omega}_i'\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i'\right\|_2^2=$$

$$(\boldsymbol{f}_i-\boldsymbol{X}_i\boldsymbol{\omega}_i')^{\mathrm{T}}(\boldsymbol{f}_i-\boldsymbol{X}_i\boldsymbol{\omega}_i')+\lambda\boldsymbol{\omega}_i'^{\mathrm{T}}\boldsymbol{\omega}_i'=$$

$$\boldsymbol{\omega}_i'^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\omega}_i'-\boldsymbol{f}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\omega}_i'-\boldsymbol{\omega}_i'^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i+\boldsymbol{f}_i^{\mathrm{T}}\boldsymbol{f}_i+\lambda\boldsymbol{\omega}_i'^{\mathrm{T}}\boldsymbol{\omega}_i'$$

由于岭回归优化问题的损失函数是一个光滑的凸函数,因此可以在导数为0处求得损失函数的最优解,即能够使残差$L(\boldsymbol{e}_i)$最小的解析解。最优解的推导过程如下:

令$L(\boldsymbol{e}_i)$对$\boldsymbol{\omega}_i'$求偏导,得:

$$\frac{\partial L(\boldsymbol{e}_i)}{\partial\boldsymbol{\omega}_i'}=2\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\omega}_i'-2\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i+2\lambda\boldsymbol{\omega}_i'$$

令$\frac{\partial L(\boldsymbol{e}_i)}{\partial\boldsymbol{\omega}_i'}=0$,得:

$$\hat{\boldsymbol{\omega}}_i'=(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i\tag{7}$$

式中:$\boldsymbol{I}$是单位矩阵;λ是正则项参数,且$\lambda>0$。

计算出所有子优化问题的最优解$\hat{\boldsymbol{\omega}}_i'(i=1,2,\cdots,n)$后,在每个$\hat{\boldsymbol{\omega}}_i'$的第$i$个位置补0得到$\hat{\boldsymbol{\omega}}_i$,这一步与式(5)到式(6)的变换过程正好相反。然后,将转换后的各个子问题最优解进行整合,得到RMR模型的整体最优解:

$$\hat{\boldsymbol{W}}=\{\hat{\boldsymbol{\omega}}_1,\hat{\boldsymbol{\omega}}_2,\cdots,\hat{\boldsymbol{\omega}}_n\}$$

总结上述分治-岭回归优化算法的具体计算步骤,得到RMR特征选择方法的计算框架1。

输入 数据集$\boldsymbol{X}$,正则项参数λ,特征选择数量k;

输出 特征子集$\boldsymbol{S}$。

1)将$\boldsymbol{X}$划分为特征集合$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$

2)将整体优化问题分解为n个子优化问题,$i=1,2,\cdots,n$:

① 定义损失:$L(\boldsymbol{e}_i)=\left\|\boldsymbol{f}_i-\boldsymbol{X}_i\boldsymbol{\omega}_i'\right\|_2^2+\lambda\left\|\boldsymbol{\omega}_i'\right\|_2^2$

② 计算$L(\boldsymbol{e}_i)$最优解:$\hat{\boldsymbol{\omega}}_i'=(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i$

3)对每个$\hat{\boldsymbol{\omega}}_i'$进行变换得到$\hat{\boldsymbol{\omega}}_i$,整合所有$\hat{\boldsymbol{\omega}}_i$,得到RMR模型最优解:$\hat{\boldsymbol{W}}=\{\hat{\boldsymbol{\omega}}_1,\hat{\boldsymbol{\omega}}_2,\cdots,\hat{\boldsymbol{\omega}}_n\}$。

4)依次计算$\hat{\boldsymbol{W}}$各个行向量的l_2-范数,得到有序集合$\{s_1,s_2,\cdots,s_n\}$,依据s_i值的大小,从$\boldsymbol{X}$中选出对应的前k个特征,组成特征子集$\boldsymbol{S}$。

3 实验验证与结果分析

3.1 实验数据集

实验中使用的数据集来源于UCI标准数据集,共5个数据集,关于这些数据集的详细描述信息请参考表1。

表1 实验数据集

Tab. 1 Experimental Datasets

数据集	样本数	特征数	类别数	描述信息
Iris20	150	20	3	合成数据
CNAE9	1 080	856	9	文本分类
COIL20	1 440	1 024	20	物体图像
Isolet5	1 560	617	26	字母读音
Semeion	1 593	256	10	手写数字

表1中的Iris20数据集为合成数据集,它是由UCI标准数据集Iris扩展而成,前4列是Iris数据集初始数据,后16列数据是由前4列进行随机线性组合得到,线性组合系数和为1,并在后16列数据中加入高斯白噪声。

3.2 评价指标

3.2.1 聚类准确率

使用K-Means算法[16]对数据进行聚类,并计算数据样本聚类之后的准确率,聚类准确率(ACCuracy of clustering,ACC)的值越大说明聚类结果越好。ACC计算公式如下:

$$ACC=\frac{1}{m}\sum_{i=1}^{m}f(l_i,map(r_i))\tag{8}$$

式中:m是样本个数;l_i是第i个样本的实际标签;$map(r_i)$是第i个样本的聚簇标签,该函数使用Kuhn-Munkres算法计算样本的聚簇标签;函数$f(a,b)$用于判断标签a和标签b的值是否相等,若相等则函数值为1,不相等则函数值为0。

3.2.2　归一化互信息

归一化互信息(Normalized Mutual Information, NMI)是评价聚类结果好坏的常用指标之一,用于度量聚簇标签向量与实际标签向量的一致程度,NMI值越大说明聚类结果越好。*NMI*计算公式如下:

$$NMI(\boldsymbol{p},\boldsymbol{q})=\frac{I(\boldsymbol{p},\boldsymbol{q})}{\sqrt{H(\boldsymbol{p})H(\boldsymbol{q})}} \tag{9}$$

式中:$\boldsymbol{p}$是聚簇标签向量;$\boldsymbol{q}$是实际标签向量;$I(\boldsymbol{p},\boldsymbol{q})$是向量$\boldsymbol{p}$与$\boldsymbol{q}$之间的互信息;$H(\boldsymbol{p})$和$H(\boldsymbol{q})$分别是$\boldsymbol{p}$和$\boldsymbol{q}$的信息熵。

3.2.3　冗余率

冗余率(Redundancy rate, Rr)用于度量数据的特征之间是否具有较强的相关性,相关性越强说明数据冗余程度越高,因此冗余率越小越好。*Rr*计算公式如下:

$$Rr(S)=\frac{2}{n(n-1)}\sum_{f_i,f_j\in S,i>j}\left|\beta_{i,j}\right| \tag{10}$$

式中:S是特征子集;n是特征子集的特征数量;$\beta_{i,j}$是特征f_i和特征f_j的皮尔逊相关系数。

3.3　实验设置

为了验证所提出的RMR无监督特征选择方法的有效性和高效性,实验中使用另外四种无监督特征选择方法用于对比,对比方法分别为经典的Laplacian特征选择方法[17]、使用了谱分析技术的非负判别特征选择(Nonnegative Discriminative Feature Selection, NDFS)方法[18]、引入了正则约束的RSR方法和自表示特征选择(Self-Representation Feature Selection, SR_FS)方法[19]。

每种特征选择方法从每个数据集中选出9个不同维数的特征子集,选出特征数量依次是特征总数量的1/10,2/10,…,9/10,然后取这9个不同维数的特征子集的聚类准确率平均值、归一化互信息平均值以及冗余率平均值,作为评价该特征选择方法效果优劣的三项指标。对于目标函数中有正则项的特征选择方法,将正则项参数的值域设置为{0.01,0.1,1,10,100},对每个正则项参数均进行实验并选取最佳结果。

另外,针对聚类准确率这一评价指标,由于*K*-Means聚类算法得出的聚类准确率会受初始质心选取的影响,为减少随机误差的影响,提升结果准确度,对每个特征子集进行100次聚类,然后取100次聚类准确率的平均值作为该特征子集的聚类准确率。

3.4　结果分析

首先,将本文提出的RMR方法应用在Iris20合成数据集上,因为Iris20数据集的后16列数据是由前4列数据组合得到,且加入了一定的噪声,因此后16列数据是应当剔除的冗余特征,而前4列数据为特征选择方法应当识别出的重要特征。

图1是RMR方法计算出的Iris20特征权重直方图,横轴代表Iris20数据集的20个特征,纵轴代表计算出的各个特征的权重。权重越大,特征越重要,越具有代表性。从图1可以直观地看出,前4个特征的权重明显大于后16个特征的权重,这说明RMR方法可以有效地识别出数据集中具有代表性的特征。从表2的实验结果数据可以得知,RMR特征选择方法与其他四种特征选择方法相比,其在5个标准数据集上选出的特征子集的平均聚类准确率最大,说明RMR方法对聚类准确率的提升能力更强。从表3的实验结果数据可以看出,RMR方法选出的特征子集的归一化互信息NMI值最大,说明RMR方法选出的特征子集的聚类效果更好。无论是聚类准确率还是归一化互信息,都是度量聚类结果好坏的常用指标,值越大越好。因此从表2和表3可以得出相同的结论,使用RMR方法对原始数据集进行特征选择,可以选出数据中的具有代表性的特征,有效地改善数据在聚类时的表现,并且同其他四种对比方法相比效果更好。

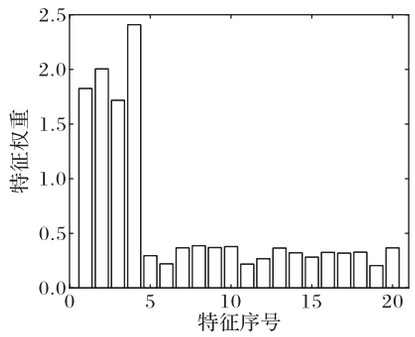

图1　Iris20特征权重直方图

Fig. 1　Histogram of feature weights of Iris20

表2　聚类准确率对比　　单位:%

Tab. 2　Comparison of clustering accuracy　　unit: %

数据集	Laplacian	NDFS	RSR	SR_FS	RMR
Iris20	86.3	84.9	85.4	87.0	89.3
COIL20	57.2	57.9	58.3	62.5	65.7
Isolet5	47.9	48.5	50.1	54.6	58.2
CNAE9	45.7	47.2	48.6	51.3	54.1
Semeion	56.1	57.3	56.8	60.2	63.7
平均值	58.6	59.2	59.8	63.1	66.2

表3　归一化互信息对比　　单位:%

Tab. 3　Comparison of normalized mutual information　　unit: %

数据集	Laplacian	NDFS	RSR	SR_FS	RMR
Iris20	68.4	65.8	66.7	71.5	73.9
COIL20	74.3	74.6	75.3	78.4	80.6
Isolet5	68.0	67.5	68.4	70.6	72.8
CNAE9	42.7	46.4	45.8	49.3	52.3
Semeion	48.6	51.1	50.2	54.4	57.5
平均值	60.4	61.1	61.3	64.8	67.4

从表4中的实验结果可知,RMR方法与另外四种特征选择方法相比,其所选出的特征子集的冗余率最低,说明RMR方法从原始数据中选出的特征相关性较弱,数据冗余程度更低。

表4　冗余率对比　　单位:%

Tab. 4　Comparison of redundancy rate　　unit: %

数据集	Laplacian	NDFS	RSR	SR_FS	RMR
Iris20	57.0	51.3	40.5	36.4	32.6
COIL20	26.5	25.7	22.8	23.1	20.5
Isolet5	28.1	21.6	16.3	15.4	12.2
CNAE9	2.2	1.8	1.9	1.7	1.3
Semeion	11.4	10.3	9.1	8.6	7.8
平均值	25.0	22.1	18.1	17.0	14.9

RMR方法选出的特征子集之所以具有较好的聚类表现和较低的冗余程度,是因为RMR模型是基于正则互表示性质,利用了特征间的相关性,同时克服了特征权重容易向自身倾斜的缺点。RSR方法和SR_FS方法虽然利用了特征间的相关性,但是忽视了特征权重容易向自身倾斜的问题。NDFS方

法结合谱聚类技术进行特征选择，该方法的效果依赖于样本相似性矩阵，容易受其影响。Laplacian方法没有考虑特征间的相关性，因此无法有效识别冗余特征。

在计算复杂度方面，RMR数学模型中的误差项和正则项都是使用矩阵Frobenius范数进行约束，使用分治算法将矩阵优化问题转换为若干个向量优化问题，如式(5)所示。此时的向量优化问题是岭回归优化问题，目标函数中的误差项和正则项均由向量l_2-范数约束，l_2-正则是凸形正则，可以在导数为0处直接取得使残差最小的解析解，如式(7)所示，式中：$\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i$的计算复杂度是$O(mn^2)$；$(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{I})^{-1}$求逆运算的复杂度是$O(n^3)$；$\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i$的复杂度是$O(mn)$；矩阵$(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i+\lambda\boldsymbol{I})^{-1}$与向量$\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{f}_i$相乘的复杂度是$O(n^2)$。考虑到样本数量大于特征数量(即$m>n$)，所以单个岭回归优化的总体渐进复杂度为$O(mn^2)$。因为整体优化问题被分解为$n$个岭回归子优化问题，所以RMR方法的整体优化问题的渐进计算复杂度为$O(mn^3)$。RMR方法计算特征选择模型最优解的过程仅相当于一次矩阵迭代运算，而非多次迭代，而NDFS方法、RSR方法以及SR_FS方法都是采用矩阵多次迭代的方式逼近最优解，每一次迭代都会涉及大量的矩阵运算，计算复杂度与迭代次数成正比。因此，RMR方法具有更低的计算复杂度，在计算性能上更具优势。

4 结语

本文研究利用高维无标签数据特征之间的相关性进行特征选择，通过在特征选择时对特征权重施加正则约束，提升了特征选择结果的鲁棒性，解决了特征权重分配不合理导致无法有效识别冗余特征的问题。实验结果表明，RMR方法能够选出重要特征，减少数据冗余，提升聚类精度。算法理论复杂度分析表明，所提方法因使用分治-岭回归优化而具有较低的计算复杂度。RMR方法也存在不足之处，其未考虑数据集的样本数量少于特征数量的情况，未来研究可考虑高维小样本数据情况下如何改进。

参考文献(References)

[1] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers and Electrical Engineering, 2014, 40(1): 16-28.

[2] WANG S, PEDRYCZ W, ZHU Q, et al. Unsupervised feature selection via maximum projection and minimum redundancy [J]. Knowledge-Based Systems, 2015, 75: 19-29.

[3] 黄铉. 特征降维技术的研究与进展[J]. 计算机科学, 2018, 45(6A): 16-21, 53. (HUANG X. Research and development of feature dimensionality reduction [J]. Computer Science, 2018, 45(6A): 16-21, 53.)

[4] ZHU P, ZHU W, HU Q, et al. Subspace clustering guided unsupervised feature selection [J]. Pattern Recognition, 2017, 66: 364-374.

[5] LI J, CHENG K, WANG S, et al. Feature selection: a data perspective[J]. ACM Computing Surveys, 2018, 50(6): No. 94.

[6] MORADI P, ROSTAMI M. A graph theoretic approach for unsupervised feature selection[J]. Engineering Applications of Artificial Intelligence, 2015, 44: 33-45.

[7] NIE F, ZHU W, LI X. Unsupervised feature selection with structured graph optimization[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2016: 1302-1308.

[8] HE X, JI M, ZHANG C, et al. A variance minimization criterion to feature selection using Laplacian regularization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 2013-2025.

[9] NIE F, XIANG S, JIA Y, et al. Trace ratio criterion for feature selection[C]// Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2008: 671-676.

[10] 潘锋，王建东，牛奔. 基于谱分析的无监督特征选择算法[J]. 计算机应用, 2011, 31(8): 2108-2110, 2114. (PAN F, WANG J D, NIU B. Unsupervised feature selection algorithm based on spectral analysis[J]. Journal of Computer Applications, 2011, 31(8): 2108-2110, 2114.)

[11] ZHENG W, ZHU X, WEN G, et al. Unsupervised feature selection by self-paced learning regularization[J]. Pattern Recognition Letters, 2020, 132:4-11.

[12] 刘艳芳，叶东毅. 基于邻域保持学习的无监督特征选择算法[J]. 模式识别与人工智能, 2018, 31(12): 1096-1102. (LIU Y F, YE D Y. Unsupervised feature selection algorithm based on neighborhood preserving learning[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(12): 1096-1102.)

[13] ZHU P, ZUO W, ZHANG L, et al. Unsupervised feature selection by regularized self-representation [J]. Pattern Recognition, 2015, 48(2): 438-446.

[14] KHAN I M, ANDERSON K S. Performance investigation and constraint stabilization approach for the orthogonal complement-based divide-and-conquer algorithm[J]. Mechanism and Machine Theory, 2013, 67: 111-121.

[15] SHEN X, ALAM M, FIKSE F, et al. A novel generalized ridge regression method for quantitative genetics[J]. Genetics, 2013, 193(4): 1255-1268.

[16] COHEN M B, ELDER S, MUSCO C, et al. Dimensionality reduction for k-means clustering and low rank approximation[C]// Proceedings of the 2015 47th Annual ACM Symposium on Theory of Computing. New York: ACM, 2015: 163-172.

[17] HE X, CAI D, NIYOGI P. Laplacian score for feature selection [C]// Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2005: 507-514.

[18] LI Z, YANG Y, LIU J, et al. Unsupervised feature selection using nonnegative spectral analysis [C]// Proceedings of the 26th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2012: 1026-1032.

[19] HE W, ZHU X, CHENG D, et al. Unsupervised feature selection for visual classification via feature-representation property [J]. Neurocomputing, 2017, 236: 5-13.

This work is partially supported by the Research Project of Shanxi Scholarship Council of China (2017-051).

WANG Zhiyuan, born in 1992, M. S. candidate. His research interests include machine learning, feature selection.

JIANG Ailian, born in 1969, Ph. D., associate professor. Her research interests include artificial intelligence, big data, feature selection, computer vision.

Osman MUHAMMAD, born in 1993, M. S. candidate. His research interests include deep learning, image processing.