具有同步化特征选择的迭代紧凑非平行支持向量聚类算法

方佳艳[1,2],刘 峤[1,2]
(1. 电子科技大学信息与软件工程学院,四川成都 611731;2. 中电科大数据研究院有限公司,贵州贵阳 550022)

摘 要: 本文提出了一种新的带有同步化特征选择的聚类算法,称为"具有同步化特征选择的迭代紧凑非平行支持向量聚类算法"(IT-NHSVC-SFS).在具有两个非平行超平面的学习模型中使用迭代(交替)优化算法完成聚类,同时引入两种类型的正则项,分别是欧几里得范数和无穷范数,欧几里得范数用于提升聚类模型的泛化能力,无穷范数实际上是对两个非平行超平面进行同步化地隐式特征抽取,从而降低来自于不相关特征的聚类噪音,保证了模型的聚类精度,并引入一组束缚变量(bounding variables)避免无穷范数的最大化操作,将非凸优化问题转化成二次凸优化问题.同时,由于新提出的模型体现着"最大间隔"的思想,因此具有良好的泛化能力.为了方便实现两个非平行超平面同步化的特征选择过程,文中将非平行超平面 SVM(Nonparallel Hyperplane SVM,NHSVM)作为 IT-NHSVC-SFS 算法的基础模型,因此和 TWSVM 以及它的变体模型不同的是:只需要求解一个二次规划问题(QP 问题)就可以同时得到两个最优超平面.同时,新算法在原有的 NHSVM 模型的约束条件集合中新添加了两组等式约束条件,从而无需进行原有模型中的两个大矩阵的求逆操作,降低了计算复杂度.此外,在 IT-NHSVC-SFS 模型中,用拉普拉斯损失函数(Laplacian loss measure)代替了 NHSVM 模型原有的铰链损失函数(hinge loss function),避免了算法早熟收敛(premature convergence).在一组标准数据集上的数值实验结果表明,相对于其他已有的聚类算法,IT-NHSVC-SFS 算法在聚类精度方面具有更好的表现.

关键词: 聚类;特征选择;非平行超平面支持向量机;无穷范数
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2020)01-0044-15
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.01.006

Iterative Tighter Nonparallel Hyperplane Support Vector Clustering with Simultaneous Feature Selection

FANG Jia-yan[1,2], LIU Qiao[1,2]
(1. *School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan* 611731, *China*; 2. *CETC Big Data Research Institute Co., Ltd., Guiyang, Guizhou* 550022, *China*)

Abstract: In this paper, a new clustering algorithm with simultaneous feature selection is proposed, which is called iterative tighter nonparallel support vector clustering with simultaneous feature selection (IT-NHSVC-SFS). In learning with two nonparallel hyperplanes model, we use the iterative (alternating) optimization algorithm to achieve clustering, and at the same time introduce two types of regularizes, the Euclidean norm and the infinite norm, respectively. Euclidean norm clustering model is used to improve the generalization ability and the infinite norm actually fulfills implicit feature extraction for the two nonparallel hyperplanes in order to reduce data noises from irrelevant features, and the clustering precision of the model is guaranteed. We also introduce a set of bounding variables to avoid maximization operation of the infinite norm, converting the non-convex optimization problem into a quadratic convex optimization problem. Meanwhile, because the new model embodies the idea of "maximum margin", it has good generalization ability. IT-NHSVC-SFS chooses nonparallel hyperplanes SVM (NHSVM) as the basis of the algorithm model. Unlike TWSVM and its variant models, only a quadratic

收稿日期:2018-04-09;修回日期:2019-04-18;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61772117);"十三五"装备预研领域基金(No. 6140312010203);军委科技委前沿探索项目(No. 1816321TS00105301);四川省科技服务业示范项目(No. 2018GFW0150);中电集团公司第五十四研究所开放课题(No. 185686)

programming problem(QP problem) needs to be solved to get the two optimal hyperplane simultaneously. This property is helpful to design a synchronous feature selection process for two nonparallel hyperplanes. The new algorithm adds two sets of equality constraints in the constraint set of the original NHSVM model, which can avoid the inverse operation of two large matrices and reduce the computational complexity. In addition, in the IT-NHSVC-SFS model, the Laplacian loss function replaces the original hinge loss function in NHSVM to avoid premature convergence. Numerical experiments on a set of benchmark data sets show that IT-NHSVC-SFS algorithm performs better in terms of clustering accuracy than other existing clustering algorithms.

Key words: clustering; feature selection; nonparallel hyperplane support vector machine; L-infinite norm

1 引言

聚类,作为一项重要的机器学习任务,由于数据样本规模的不断增长,人工添加数据标签的代价在不断增大,因此,聚类已经引起了越来越多的关注. 许多研究人员已经对在无监督和半监督的学习领域建立有效的聚类算法进行了广泛的研究[1~4]. 给定一组样本实例,一种特定的聚类算法是为了在大量数据实例中寻找隐藏在其中的结构信息,从而将具有相似结构或属性的数据实例分组到一起. 一些优秀的聚类方法包括 k-均值[5]、规格化切割[6]、混合模型[7]和光谱聚类[8]. 最近,也陆续出现了大量的聚类方法. 例如,噪声稀疏子空间聚类[9]考虑了子空间聚类问题的噪声存在. 它是稀疏子空间聚类(SSC)的改进版本,可以有效地识别底层子空间,甚至可以处理噪声数据. John R Hershey 提出了在深度聚类框架下进行分割和分离的识别性嵌入[10],在此基础上,一个深度网络被训练为在每一个时间频域内分配不同的嵌入向量,从而对来自输入混合的目标谱图的分割标签进行隐式的预测. 另外,有学者还提出了一种自适应的多任务聚类(SAMTC)[11]方法,以一种自动的方式识别和转移任务中的可重用实例,从而避免了负迁移. 作为一种优秀的光谱聚类的改进版本,自适应的矩阵分解矩阵[12]从矩阵分解的观点看待观测数据的聚类问题,同时得到关联矩阵(affinity matrix),使矩阵分解正则化.

受"支持向量机"(SVM)[13]所取得的巨大成功的驱动,在监督式学习中,越来越多在研究者建立了很多高效的 SVM 变体模型[14~18]. 此后,在非监督式学习中,Xu 等人将最大间隔的思想嵌入到聚类任务中,并提出了最大间隔聚类(MMC)[19]来实现在实例特定的标签值分配和特定于 SVM 的参数之间的同步优化,从而获得一组最优的标签值,这组最优的标签值在所有可能的标签值分配中,其相应的间隔(margin)是最大的. 受双子支持向量机(TWSVM)[20]分类模型的启发,Wang 等人将 TWSVM 的概念引入集群任务中,提出了用于聚类的双子支持向量模型(TWSVC)[21],通过解决一系列的二次规划问题来确定 k 个群中心平面. 此外,基于 TWSVC,有学者还提出了模糊最小二乘双支持向量聚类(F-LS-TWSVC)[22],在此基础上,优化了每个数据样本的模糊隶属度值,并利用它将每个数据样本分配给一个或多个群. 由于在上述的聚类策略中均采用了"最大间隔"思想,因此可以保证聚类模型具有良好的泛化性能.

在模式识别和机器学习中,另一个日益重要的研究领域是特征选择. 由于数据收集和存储技术的快速发展,越来越多的大型、高维度、复杂和异构的数据集正不断涌现,对现有的机器学习方法构成了严峻的挑战,因此,为了解决这样的问题,将特征选择技术纳入到学习模型中以适应高维复杂数据集的处理显得尤为重要. 除了维数灾难之外,现有的学习模型普遍存在这样一个情形,即并非所有的特征都具有完全相同的判别或决策能力(discriminative power). 在特定的类别概念下,一些非信息性的特征的确构成了数据噪声,会降低预测的准确性. 因此,特征选择是消除特征不相关性和冗余性的重要工具,这将提高最终判别规则(discriminative rule)的准确性,可解释性和可理解性.

特性子集选择策略本质上分为以下三种类型:过滤器(filters)、封装器(wrappers)和嵌入式方法(embedded methods)[23,24]. 过滤方法利用特征的统计特性,在训练任何学习模型之前,过滤出具有较差的识别能力的特征维度. 由于滤波方法是利用数据的固有性质,将特征抽取过程作为预处理步骤,因此它们的计算速度很快,并且可以很容易地扩展到高维数据空间.

封装的方法是在整个的特征变量的集合中搜索,对每个特征维度的重要性进行评估. 通常会根据整个学习模型后续的分类算法的特性,预先定义一种性能衡量准则,然后针对每个维度,对该性能衡量准则进行优化,从而逐一确定每一个特征维度在最终决策能力(discriminative power)方面的重要性. 由于封装方法中存在着特征选择过程和后续学习算法的交互[24,25],所以,尽管封装方法对计算能力要求较高[23],但是相比于过滤器方法,它们通常可以得到更加准确的特征抽取结果.

与滤波和包装方法不同,嵌入的方法将特征选择过程合并到最终模型识别规则(model discriminative rules)的推导过程中. 它们不仅可以实现与后续模型训

练过程的交互,而且比封装方法具有更少的密集计算[26~28].

上述的特征选择策略大多数都是用于分类问题,然而,在非监督式学习中,对于处在高维度的数据空间中的样本点,无关的和冗余的特征属性同样会成为数据噪音,影响最终的聚类精度.因此,很有必要将嵌入式的特征选择过程和聚类方法结合在一起.本文提出了一种新的聚类算法,称为"具有同步特征选择的迭代紧凑非平行支持向量聚类算法"(IT-NHSVC-SFS).该算法引入了组罚函数[29,30]来实现正则化和自动特征抽取.和一些比较流行的用于特征选择的正则项(l_1-范数 and l_0-范数)不同的是,组罚函数的目的是在两个非平行分割超平面上更加协调同步地对那些与某一给定的特征维度相关的权重施加惩罚.由于本文的新算法是建立在 NHSVM[31]模型上的,因此只需要求解唯一的二次优化问题,方便了在两个超平面上同步进行特征选择的过程.另外,在 NHSVM 模型上新添加了两组等式约束条件,所以训练过程无需进行大矩阵的求逆运算,减轻了模型的计算复杂度.此外,由于采用了交替优化方法对标签变量和模型参数进行优化,就避免了非凸优化问题的复杂求解过程.然而,在这种交替优化过程中,为了避免算法的早熟收敛(premature convergence),新模型中将 NHSVM 中原有的铰链损失函数(hinge loss function)替换成拉普拉斯损失函数(Laplacian loss function),防止算法过早地卡在局部最优解上.

2 相关领域概述

2.1 双子支持向量聚类算法

考虑一个聚类问题,双子支持向量聚类算法(Twin Support Vector for Clustering,TWSVC)旨在找到 k 个集群中心平面(cluster center plane).

$$\text{Center-plane}_i:\boldsymbol{w}_i^{\mathrm{T}}x+b_i=0,\quad i=1,\cdots,k \tag{1}$$

其中,$\boldsymbol{w}_i$ 是集群中心平面 i 的法向量,b_i 为偏移量,$x\in\mathbf{R}^n$.

通过求解一系列原始优化问题(2)可以得到这 k 个集群中心平面.

$$\begin{aligned}&\min_{\boldsymbol{w}_i,b_i,\boldsymbol{\xi},\boldsymbol{x}_i}\frac{1}{2}\|X_i\boldsymbol{w}_i+b_i\boldsymbol{e}\|^2+c\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\xi}_i\\&\text{s.t.}\quad|\hat{\boldsymbol{X}}_i+b_i\boldsymbol{e}|\geqslant\boldsymbol{e}-\boldsymbol{\xi}_i,\boldsymbol{\xi}_i\geqslant0\end{aligned} \tag{2}$$

其中,$\boldsymbol{w}_i$ 代表的是第 i 个子簇中的数据矩阵,矩阵的第 m 行表示该子簇的第 m 个样本点.$\boldsymbol{\xi}_i$ 是松弛向量,每个维度代表松弛变量,用来记录其他所有子簇的样本点的距离违反量.在上述原始优化问题(2)的目标函数中,第一项是第 i 个子簇中的所有样本点到该子簇的中心平面 $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}+b_i=0$ 的距离之和.最小化该项就相当于使得第 i 个子簇的中心平面离该子簇的所有样本点尽可能地接近.目标函数的第二项代表其他所有子簇的样本点的距离违反量的总和,最小化该项的目的在于确保第 i 个子簇的样本点集合与剩余样本点的距离尽可能地远.TWSVC 算法通过对样本标签值变量和 k-PC[32]中的子簇中心平面参数进行交替优化,不断地更新样本标签值.初始化所有子簇的样本标签值之后,接下来以固定的标签值求解优化问题(2).值得注意的是,问题(2)是一个非凸优化问题,需要使用凹凸过程(Concave-Convex Procedure,CCCP)[33]求解.在该过程中,问题(2)被分解成有限个二次凸优化的子问题(3),变量初始值为 $\boldsymbol{w}_i^0$ 和 b_i.

$$\begin{aligned}&\min_{\boldsymbol{w}_i^{j+1},b_i^{j+1},\boldsymbol{\xi}_i^{j+1}}\frac{1}{2}\|\boldsymbol{X}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e}\|^2+c\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\xi}_i^{j+1}\\&\text{s.t.}\quad\mathrm{T}(|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e}|)\geqslant\boldsymbol{e}-\boldsymbol{\xi}_i^{j+1},\boldsymbol{\xi}_i^{j+1}\geqslant0\end{aligned} \tag{3}$$

其中,$j=0,1,2,\cdots$ 表示的是每个子问题的索引值,T(.)代表一阶泰勒展开式.通过计算 $|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}|$ 关于 $\boldsymbol{w}_i^j$ 和 b_i^j 的次梯度(subgradient)[34],可得:$\nabla(|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}|)=\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))[\hat{\boldsymbol{X}}_i,\boldsymbol{e}]$.由于 $|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}|=\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e})$,则:

$$\begin{aligned}&\mathrm{T}(|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e}|)\\&=|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}|+\nabla(|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e}|)([\boldsymbol{w}_i^{j+1};b_i^{j+1}]-[\boldsymbol{w}_i^j;b_i^j])\\&=\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))[\hat{\boldsymbol{X}}_i,\boldsymbol{e}](\boldsymbol{w}_i^{j+1}+b_i^{j+1})\\&\quad+(|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}|-\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))[\hat{\boldsymbol{X}}_i,\boldsymbol{e}][\boldsymbol{w}_i^j;b_i^j])\\&=\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e})\end{aligned} \tag{4}$$

因此,问题(3)就可以等价地写成:

$$\begin{aligned}&\min_{\boldsymbol{w}_i^{j+1},b_i^{j+1},\boldsymbol{\xi}_i^{j+1}}\frac{1}{2}\|\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e}\|^2+c\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\xi}_i^{j+1}\\&\text{s.t.}\quad\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^{j+1}+b_i^{j+1}\boldsymbol{e})\\&\qquad\geqslant\boldsymbol{e}-\boldsymbol{\xi}_i^{j+1},\boldsymbol{\xi}_i^{j+1}\geqslant\boldsymbol{0}\end{aligned} \tag{5}$$

受 SVM[35,36]和 TWSVM[20,37]的启发,问题(5)的模型参数可以通过求解以下对偶问题得出:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\alpha}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}\\&\text{s.t.}\quad0\leqslant\boldsymbol{\alpha}\leqslant c\boldsymbol{e}\end{aligned} \tag{6}$$

其中,$\boldsymbol{G}=\mathrm{diag}(\mathrm{sign}(\hat{\boldsymbol{X}}_i\boldsymbol{w}_i^j+b_i^j\boldsymbol{e}))[\hat{\boldsymbol{X}}_i\boldsymbol{e}]$,$\boldsymbol{H}=[\boldsymbol{X}_i\boldsymbol{e}]$,$\boldsymbol{\alpha}\in\mathbf{R}^{m-m_i}$ 是拉格朗日乘子向量.由于问题(6)是一个凸的二次规划问题,所以可以采用高效的迭代方法——SOR(Successive Overrelaxation)技术[38],求解一组线性等式即可.因此,通过求解对偶问题(6),我们可以借助 KKT 条件从对偶变量最优解得出原始变量最优解.原始变量最优解如下:

$$[\boldsymbol{w}_i^{j+1};b_i^{j+1}]^{\mathrm{T}}=(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\alpha} \tag{7}$$

因此,求解问题(2)的主要流程如下:

(1)选择变量初始值$[\boldsymbol{w}_i^0, b_i^0]$.

(2)对于$j=0,1,\cdots$,通过式(7)找到$[\boldsymbol{w}_i^{j+1}, b_i^{j+1}]$. 若$\|[\boldsymbol{w}_i^{j+1}, b_i^{j+1}]-[\boldsymbol{w}_i^j, b_i^j]\|$比预先设定的阈值小,则终止算法,并执行$\boldsymbol{w}_i=\boldsymbol{w}_i^{j+1}, b_i=b_i^{j+1}$.

研究证明,CCCP 可以找到问题(2)的局部最优解[33]. 当求出最优解$[\boldsymbol{w}_i, b_i]$$(i=1,\cdots,k)$时,就可以根据以下的判别式(8)对数据样本的标签值作进一步更新.

$$y=\arg\min_i\{|\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}+b_i|, i=1,\cdots,k\} \tag{8}$$

其中,$|\cdot|$是绝对值符号. 然后,不断地对样本的类别标签值和子簇中心平面参数进行交替优化,直到满足某一特定的终止条件.

2.2　非平行超平面支持向量机

非平行超平面支持向量机(Nonparallel Hyperplane SVM ,NHSVM)模型继承了 TWSVM 的一些特性,但是 NHSVM 模型是通过求解一个 QP 问题同时得到两个非平行分割超平面的,需要求解的 QP 问题如下:

$$\begin{aligned}\min_{\substack{\boldsymbol{w}_k,b_k,\boldsymbol{\xi}_k\\k=1,2}}\ &\frac{1}{2}(\|\boldsymbol{A}\boldsymbol{w}_1+\boldsymbol{e}_1b_1\|^2+\|\boldsymbol{B}\boldsymbol{w}_2+\boldsymbol{e}_2b_2\|^2)\\&+\frac{c_1}{2}(\|\boldsymbol{w}_1\|^2+b_1^2+\|\boldsymbol{w}_2\|^2+b_2^2)\\&+c_2(\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{\xi}_1+\boldsymbol{e}_2^{\mathrm{T}}\boldsymbol{\xi}_2)\\\text{s.t.}\ \ &\boldsymbol{A}\boldsymbol{w}_1+\boldsymbol{e}_1b_1-\boldsymbol{A}\boldsymbol{w}_2-\boldsymbol{e}_1b_2\geqslant\boldsymbol{e}_1-\boldsymbol{\xi}_1\\&\boldsymbol{B}\boldsymbol{w}_1+\boldsymbol{e}_2b_2-\boldsymbol{B}\boldsymbol{w}_1-\boldsymbol{e}_2b_1\geqslant\boldsymbol{e}_2-\boldsymbol{\xi}_2\\&\boldsymbol{\xi}_1\geqslant0,\ \boldsymbol{\xi}_2\geqslant0\end{aligned} \tag{9}$$

其中,$c_k>0(k=1,2)$是正常数,用来对经验风险和结构风险[39]进行折中. $\boldsymbol{A}\in\mathbf{R}^{m_1\times n}$和$\boldsymbol{B}\in\mathbf{R}^{m_2\times n}$分别是正类别和负类别样本的数据矩阵. $\boldsymbol{e}_1$ 和 $\boldsymbol{e}_2$ 是维度合适的单位向量. b_1 和 b_2 分别是两个超平面相对于原点的偏移量. $\boldsymbol{\xi}_1$ 和 $\boldsymbol{\xi}_2$ 是由松弛变量组成的向量,分别记录负类别样本点和正类别样本点的距离违反值. NHSVM 最终的决策规则和 TWSVM 模型是一致的.

2.3　用于 SVM 的特征选择技术

正如引言部分所述,特征选择策略可以被分为三种类型:过滤器(filters)、封装器(wrappers)和嵌入式方法(embedded methods)[23,24].

费舍尔得分(Fisher Score)[40]是用于二元分类中的特征选择策略的一个代表. 它通过计算特征和预测值之间的自定义相关性来评估每一个特征的重要性,关于某个特征的自定义相关性可以被定义为关于该特征的两个类别的均值之差和方差之和的比值. 这种相关性如下:

$$F(j)=\left|\frac{\mu_j^+-\mu_j^-}{(\sigma_j^+)^2+(\sigma_j^-)^2}\right| \tag{10}$$

其中,μ_j^+ 和 μ_j^- 分别代表正类别和负类别样本中第 j 个特征的均值;σ_j^+ 和 σ_j^- 分别是正类别和负类别样本中第 j 个特征的方差. 因此,我们可以将式(10)作为特征维度重要性的衡量标准,根据重要性,对所有的特征维度进行排序,然后抽取前 r 个最重要的特征维度(费舍尔得分最高),并在这 r 个特征子集上进行 SVM 模型的训练. 由于这种过滤器特征选择方式分离了特征抽取过程和学习模型的参数训练过程,因此它可以和任何一种二元分类算法如 TWSVM,NHSVM 结合在一起使用,同时允许使用核技巧将模型推广到非线性情形.

基于 SVM 的递归特征消除(The Recursive Feature Elimination SVM,SVM-RFE)算法是封装式方法(wrapper methods)的一种典型代表[41]. 该算法并不是定义一种和训练模型完全独立的相关性对特征进行排序,而是逐一移除一些特征属性,使得这些属性的移除可以导致最大的类分隔. 算法每次只移除一个特征来最小化 $W^2(\boldsymbol{\alpha})$的变化量,直到仅剩下最后 r 个特征才终止特征消除过程. $W^2(\boldsymbol{\alpha})$的定义如下:

$$W^2(\boldsymbol{\alpha})=\sum_{i,s=1}^{m}\alpha_i\alpha_s y_i y_s\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_s \tag{11}$$

其中,那些出现在标准 SVM 构造的对偶变量全部集中在向量 $\boldsymbol{\alpha}$ 中. 然后,SVM-RFE 算法根据$|W^2(\boldsymbol{\alpha})-W^2_{-p}(\boldsymbol{\alpha})|$衡量标准对每个维度的特征进行重要性排序,$W^2_{-p}(\boldsymbol{\alpha})$是从 $W^2(\boldsymbol{\alpha})$的每个样本点中移除了特征 p[41]. 由于该特征消除策略是建立在标准 SVM 模型的对偶形式基础之上的,因此可以通过引入核函数将 SVM-RFE 模型推广到非线性分类情形中.

SVM-RFE 的思想同样也运用在了 TWSVM 模型中[29]. 对于每一个特征属性,TWSVM-RFE 定义了一种对特征重要性的排序准则 $W(j)=w_{1j}^*+w_{2j}^*$,其中,w_{1j}^* 和 w_{2j}^* 分别是归一化的两个分割超平面的第 j 个特征的权重值. w_{ij}^* 被定义为 $w_{lj}^*=|w_{lj}|/\|\boldsymbol{w}_l\|_2$,其中 $l=1,2$. 然而,TWSVM-RFE 中存在的主要问题是:我们仅仅根据每个特征在两个超平面上对应的权重值的平均结果的绝对值来判断该特征的重要性. 在这一前提下,某一个特征在其中一个超平面的权重值可能正向很大(代表正相关性很大),而在另一个超平面的权重值是负向很大(代表负相关性很大),由于平均的效果,该特征最终的重要性可能很低,所以就可能会被移除. 因此,这种方式并没有真正实现两个非平行分割超平面的协同特征抽取. 此外,不论是 SVM-RFE 还是 TWSVM-RFE,模型的计算复杂度都比较高,因此都不适用于处理维度过高的样本数据.

对于嵌入式的特征选择策略,一种常用的方法是用套索惩罚(l_1-范数)来代替欧几里得范数同时进行正则化和特征选择过程,从而对预测精度和稀疏性[42]进行折中. 同样地,l_1-范数和 l_0-范数均可以运用在 TWSVM 模型

中. 计算结果说明,特征消除实际上是由于选择了适当的范数而产生的间接效果[42]. 这两种范数均可以用在TWSVM模型的两个子问题中来代替欧几里得范数,从而得到一对稀疏化的非平行的最优超平面. L_1-TWSVM模型的两个子问题的公式构造如下:

$$\begin{aligned}\min_{\boldsymbol{w}_1,b_1,\boldsymbol{\xi}_2} & \|\boldsymbol{w}_1\|_1 + c_1\|A\boldsymbol{w}_1 + \boldsymbol{e}_1 b_1\|_1 + c_3\boldsymbol{e}_2^{\mathrm{T}}\boldsymbol{\xi}_2 \\ \text{s.t.} \quad & -(B\boldsymbol{w}_1 + \boldsymbol{e}_2 b_1) + \boldsymbol{e}_2 - \boldsymbol{\xi}_2 \\ & \boldsymbol{\xi}_2 \geqslant \boldsymbol{0} \end{aligned} \tag{12}$$

$$\begin{aligned}\min_{\boldsymbol{w}_2,b_2,\boldsymbol{\xi}_1} & \|\boldsymbol{w}_2\|_1 + c_2\|B\boldsymbol{w}_2 + \boldsymbol{e}_2 b_2\|_1 + c_4\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{\xi}_1 \\ \text{s.t.} \quad & -(A\boldsymbol{w}_2 + \boldsymbol{e}_1 b_2) \geqslant \boldsymbol{e}_1 - \boldsymbol{\xi}_1 \\ & \boldsymbol{\xi}_1 \geqslant \boldsymbol{0} \end{aligned} \tag{13}$$

然而,这种方法存在一个很大的弊端. 由于求解两个子优化问题的过程是相互独立的,所以对于两个分类器的特征抽取过程实际上并不协同,这就意味着从两个分割超平面抽取得到的两个特征属性集合并不相同,且这两个集合的并集会非常庞大,不利于进行样本预测. 尽管这种方法某种程度上可以取得较好的预测效果,但是如果能够获取一组规模较小的特征集合,则有利于提高模型最终的判别式(discriminative rule)的可解释性,同时减小特征变量的存储消耗,降低预测新样本时获取特征变量的开支. 因此,需要对两个分割超平面进行协同性的特征消除过程,实现特征选择同步化. 我们可以采用组罚函数作为模型的正则项从两个超平面中尽可能地抽取出共同的特征.

组罚函数[43]的原理如下:假设所有的 n 个特征变量被划分成多个互不相交的哑变量集合 I_j,$|I_j| = p_j$($j = 1,\cdots,J$),$\sum_{j=1}^{J} p_j = n$ 以欧几里得范数为例,组罚函数的构造如下:

$$\Gamma(\boldsymbol{w}) = \sum_{j=1}^{J} \sqrt{p_j}\,\|\boldsymbol{w}^{(j)}\|_2 \tag{14}$$

其中,$\|\boldsymbol{w}^{(j)}\|_2 = \sqrt{\sum_{l\in I_j} w_l^2}$ 除了欧几里得范数,无穷范数同样可以用来惩罚成组的变量,例如 F_∞ – SVM 算法[44]. L-无穷范数的形式如下:

$$\Gamma(\boldsymbol{w}) = \sum_{j=1}^{J} \|\boldsymbol{w}^{(j)}\|_\infty \tag{15}$$

其中,$\|\boldsymbol{w}^{(j)}\|_\infty = \max_{l\in I_j}\{|w_l|\}$. $F_{-\infty}$ – norm SVM 算法通过引入一组束缚变量(bounding variables),将优化问题转化成线性规划问题,同时添加一组约束条件 $|w_l| \leqslant t_j, l \in I_j (j = 1,\cdots,J)$.

3 具有同步化特征选择的迭代紧凑非平行支持向量聚类算法

在该部分我们将提出一种新的聚类算法,称为“具有同步化特征选择的迭代紧凑非平行支持向量聚类算法(IT-NHSVC-SFS)”. 该算法建立在NHSVM模型的基础之上,并使用迭代优化策略使得NHSVM模型可以处理聚类任务. 之所以选择NHSVM而不是TWSVM作为算法的基础模型是因为NHSVM只需要求解一个QP问题就可以同时得到两个非平行超平面,方便将L-无穷范数的惩罚项同时施加在两个超平面上,有利于实现特征选择同步化. 在TWSVM模型中,由于求解孪生非平行超平面是两个相互独立的过程,所以无法真正实现两个分割超平面的特征抽取过程的同步化. 此外,新算法IT-NHSVC-SFS在NHSVM模型的基础上还作出了一系列新的改进:①为了在聚类训练的同时能够实现自动特征消除过程,在NHSVM的目标函数中添加了L-无穷范数,并引入一组束缚变量避免无穷范数的最大化操作,将非凸优化问题转化成二次凸优化问题;②为了避免原始NHSVM模型中的大规模矩阵的求逆运算,减小计算复杂度,IT-NHSVC-SFS在原有的约束条件集合中另外增加了两组等式约束条件;③由于聚类标签变量和分割超平面参数的交替优化过程中存在算法的早熟收敛(premature convergence),因此使用拉普拉斯损失函数(Laplacian loss function)来代替原先的铰链损失函数(hinge loss function),从而有效防止算法过早地卡在不太好的局部最优解上.

首先我们简要说明一下本文后续部分所涉及的数学符号. 考虑一个二元的聚类问题,假设有 m 个正类别样本点和 n 个负类别样本点. 由于这是聚类问题,所以训练时 m 和 n 并不是固定值,在每一轮的交替优化过程中会有所变化. $\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 是两个超平面对应的权值向量,分别包含了偏移量 b_1 和 b_2. 上标(1)和(2)标识了样本点属于哪个类别,(1)表示正类别,(2)表示负类别. 下标 i 和 j 分别是正类别和负类别的样本索引. 比如,$\boldsymbol{x}_i^{(1)} \in \mathbf{R}^{\tilde{d}}$ 是 $\tilde{d}$ 维空间中的第 i 个正类别样本点,$y_i^{(1)}$ 则是其相应的类别标签. IT-NHSVC-SFS算法的原始问题如下:

$$\begin{aligned}\min_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2\\ \boldsymbol{\xi}^{(1)},\boldsymbol{\xi}^{(2)}}} & \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)} + b_1)^2 + \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)} + b_2)^2 \\ & + \frac{1}{2}(\|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2) \\ & + c_1\Big(\sum_{i=1}^{m}\boldsymbol{\xi}_i^{(1)} + \sum_{j=1}^{m}\boldsymbol{\xi}_j^{(2)}\Big) + c_2\sum_{d=1}^{\tilde{d}}\|\boldsymbol{w}^{(d)}\|_\infty \\ \text{s.t.} \quad & y_i^{(1)}(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)} + b_1) - y_i^{(1)}(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)} + b_2) \geqslant 1 \\ & -\boldsymbol{\xi}_i^{(1)}, \boldsymbol{\xi}_i^{(1)} \geqslant 0, i = 1,\cdots,m \\ & y_j^{(2)}(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)} + b_1) - y_j^{(2)}(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)} + b_2) \geqslant 1 \\ & -\boldsymbol{\xi}_j^{(2)}, \boldsymbol{\xi}_j^{(2)} \geqslant 0, j = 1,\cdots,n \end{aligned} \tag{16}$$

其中,$c_1, c_2 > 0$ 是常数,用来权衡模型拟合度,最大间隔,松弛变量惩罚和特征选择过程. $\boldsymbol{\xi}_i^{(1)}$ 和 $\boldsymbol{\xi}_i^{(2)} > 0$ 是松弛变量,分别记录正类别和负类别样本点的距离违反

量. 优化问题(16)中的目标函数的最后一项是所有特征的权值对应的 L-无穷范数之和,同时实现两个分割超平面的正则化和隐式特征选择. w_{kd}表示权值向量 $\boldsymbol{w}_k$ 的第 d 个维度. 记 $\boldsymbol{w}^{(d)}=(w_{1d},w_{2d})\in\mathbf{R}^2$, $\|\boldsymbol{w}^{(d)}\|_\infty=\max_{k=1,2}\{|w_{kd}|\}$ $(d=1,\cdots,\tilde{d})$ 其他项和 NHSVM 保持一致.

由于无穷范数的存在,问题(16)是一个非凸优化问题,因此需要引入一组束缚变量(bounding variables)$z\in\mathbf{R}^{\tilde{d}}$ 来避免无穷范数中的最大化操作,将非凸优化问题转化成二次凸优化问题. 另外,还需要引入两组原始变量 $\eta_i(i=1,\cdots,m)$ 和 $\varphi_j(j=1,\cdots,n)$ 来代替目标函数中的前两项,避免大矩阵的求逆操作,因此需要在原有的约束条件集合中增加两组等式约束条件. 由于铰链损失函数(hinge-loss = $\max(0,1-y_sf_i)$, $f_i=f(x_i)$)的数学特性[45],当使用交替优化的方式进行二元聚类时,算法有很大的可能性会发生早熟收敛,这就有必要采用一种具有较为友好的数学形式的损失函数. 本文采用的是拉普拉斯函数($L=|f_i-y_i|$)来表达松弛变量惩罚值,不仅对那些离超平面较近的样本点施加惩罚,对那些离超平面较远的样本点同样也要施加一定的惩罚值,这种方式在实际中可以取得比铰链损失函数更好的聚类效果. 另外,在不改变拉普拉斯函数值的情况下,将绝对值符号中的每一项分别乘以 y_i. 经过上述步骤,问题(16)可以重写为如下形式:

$$
\begin{aligned}
&\min_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2\\ \xi^{(1)},\xi^{(2)},\eta,\varphi}}\ \frac{1}{2}\sum_{i=1}^{m}\eta^2+\frac{1}{2}\sum_{j=1}^{n}\varphi_j^2\\
&\quad+\frac{1}{2}(\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2)\\
&\quad+c_1\Big(\sum_{i=1}^{m}\xi_i^{(1)}+\sum_{j=1}^{m}\xi_j^{(2)}\Big)+c_2\\
&\text{s.t.}\quad \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1=\eta_i\\
&\qquad \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2=\varphi_j\\
&|1-(y_i^{(1)}(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)-y_i^{(1)}(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2))|\leqslant\xi_i^{(1)},\\
&\xi_i^{(1)}\geqslant0,i=1,\cdots,m\\
&|1-(y_j^{(2)}(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-y_j^{(2)}(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(2)}+b_2))|\leqslant\xi_j^{(2)},\\
&\xi_i^{(1)}\geqslant0,j=1,\cdots,n\\
&|\boldsymbol{w}_k|\leqslant z,k=1,2
\end{aligned}
\tag{17}
$$

为了简化原始问题(17)的求解过程,消去了绝对值符号并且分别用一对松弛变量代替式(17)中的松弛变量 $\xi_i^{(1)}$ 和 $\xi_j^{(2)}$. 因此,问题(17)就可以进一步演化为如下形式:

$$
\begin{aligned}
&\min_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2\\ \hat{\xi}^{(1)},\check{\xi}^{(1)},\hat{\xi}^{(2)}\check{\xi}^{(2)}\\ \eta,\varphi,z}}\ \frac{1}{2}\sum_{i=1}^{m}\eta_i^2+\frac{1}{2}\sum_{j=1}^{n}\varphi_j^2+\frac{1}{2}(\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2)\\
&\quad+\frac{c_1}{2}\Big(\sum_{i=1}^{m}(\hat{\xi}_i^{(1)}+\check{\xi}_i^{(1)})+\sum_{j=1}^{n}(\hat{\xi}_j^{(2)}+\check{\xi}_j^{(2)})\Big)+c_2\boldsymbol{e}^{\mathrm{T}}z\\
&\text{s.t.}\quad \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1=\eta_i\\
&\qquad \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2=\varphi_j\\
&1-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)+(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)\leqslant\hat{\xi}_i^{(1)},\hat{\xi}_i^{(1)}\geqslant0\\
&(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)-1\leqslant\check{\xi}_i^{(1)},\check{\xi}_i^{(1)}\geqslant0,i=1,\cdots,m\\
&1-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)+(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)\leqslant\hat{\xi}_j^{(2)},\hat{\xi}_j^{(2)}\geqslant0\\
&(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-1\leqslant\check{\xi}_{ij}^{(1)},\check{\xi}_j^{(1)}\geqslant0,j=1,\cdots,n\\
&|\boldsymbol{w}_k|\leqslant z,k=1,2
\end{aligned}
\tag{18}
$$

在原始问题的求解中使用对偶理论有如下优点:①由于推导过程是建立在对偶形式之上的,因此可以直接运用核技巧[46]将模型推广到非线性情形中;②如果样本数据是处在高维度的数据空间中,并且样本总数远小于特征维数,则对偶理论就可以将原始优化问题从高维空间转化到一个低维度的空间中,低维空间的维数由样本总量决定;③将原始问题转化成为对偶形式可以引申出很多的几何解释[47],提高实际训练过程的可理解性和可解释性. 因此,在本文提出的新的聚类算法中,利用了强大的对偶理论完成数学推导.

首先,关于原始问题(18)的拉格朗日函数如下:

$$
\begin{aligned}
&L(\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2,\hat{\xi}^{(1)},\check{\xi}^{(1)},\hat{\xi}^{(2)}\check{\xi}^{(2)},\eta,\varphi,z)\\
&\quad=\frac{1}{2}\sum_{i=1}^{m}\eta_i^2+\frac{1}{2}\sum_{j=1}^{n}\varphi_j^2+\frac{1}{2}(\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2)\\
&\quad+c_1\Big(\sum_{i=1}^{m}(\hat{\xi}_i^{(1)}+\check{\xi}_i^{(1)})+\sum_{j=1}^{n}(\hat{\xi}_j^{(2)}+\check{\xi}_j^{(2)})\Big)\\
&\quad+c_2\boldsymbol{e}^{\mathrm{T}}z+\sum_{i=1}^{m}\alpha_i(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1-\eta_i)\\
&\quad+\sum_{j=1}^{n}\beta_i(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2-\varphi_j)\\
&\quad+\sum_{i=1}^{m}\hat{s}_i[1-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)+(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)-\hat{\xi}_i^{(1)}]\\
&\quad+\sum_{i=1}^{m}\check{s}_i[(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)-1-\check{\xi}_i^{(1)}]\\
&\quad+\sum_{i=1}^{m}[\hat{p}_i(-\hat{\xi}_i^{(1)})+\hat{p}_i(-\check{\xi}_i^{(1)})]\\
&\quad+\sum_{j=1}^{n}\hat{t}_i[1-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)+(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-\hat{\xi}_j^{(2)}]\\
&\quad+\sum_{j=1}^{n}\check{t}_i[(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-1-\check{\xi}_j^{(2)}]\\
&\quad+\sum_{j=1}^{n}[\hat{q}_j(-\hat{\xi}_j^{(2)})+\check{q}_j(-\check{\xi}_j^{(2)})]
\end{aligned}
$$

$$+\hat{\boldsymbol{v}}_1^{\mathrm{T}}(\boldsymbol{w}_1-\boldsymbol{z})+\breve{\boldsymbol{v}}_1^{\mathrm{T}}(-\boldsymbol{w}_1-\boldsymbol{z})$$
$$+\hat{\boldsymbol{v}}_2^{\mathrm{T}}(\boldsymbol{w}_2-\boldsymbol{z})+\breve{\boldsymbol{v}}_2^{\mathrm{T}}(-\boldsymbol{w}_2-\boldsymbol{z}) \tag{19}$$

其中,$\boldsymbol{\alpha}_i,\boldsymbol{\beta}_j,\hat{\boldsymbol{s}}_i,\breve{\boldsymbol{s}}_i,\hat{\boldsymbol{p}}_i,\breve{\boldsymbol{p}}_i,\hat{\boldsymbol{t}}_j,\breve{\boldsymbol{t}}_j,\hat{\boldsymbol{q}}_j,\breve{\boldsymbol{q}}_j,\hat{\boldsymbol{v}}_1,\breve{\boldsymbol{v}}_1,\hat{\boldsymbol{v}}_2,\breve{\boldsymbol{v}}_2$ 是每个约束条件分别对应的拉格朗日乘子. 除了 $\boldsymbol{\alpha}_i(i=1,\cdots,m)$ 和 $\boldsymbol{\beta}_j(j=1,\cdots,n)$ 之外,所有的拉格朗日乘子都是非负变量. 原始问题(18)可以等价地写成式(20).

$$\min_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2,\boldsymbol{z},\boldsymbol{\eta},\\ \boldsymbol{\varphi},\hat{\boldsymbol{\xi}}^{(1)},\breve{\boldsymbol{\xi}}^{(1)},\hat{\boldsymbol{\xi}}^{(2)},\breve{\boldsymbol{\xi}}^{(2)}}}\ \max_{\substack{\boldsymbol{\alpha},\boldsymbol{\beta},\hat{\boldsymbol{s}},\breve{\boldsymbol{s}},\hat{\boldsymbol{p}},\breve{\boldsymbol{p}},\\ \hat{\boldsymbol{t}},\breve{\boldsymbol{t}},\hat{\boldsymbol{q}},\breve{\boldsymbol{q}},\hat{\boldsymbol{v}}_1,\breve{\boldsymbol{v}}_1,\hat{\boldsymbol{v}}_2,\breve{\boldsymbol{v}}_2}}\ \{L(\boldsymbol{w}_1,\boldsymbol{w}_2,b_1,b_2,\boldsymbol{z},\boldsymbol{\eta},\boldsymbol{\varphi},\hat{\boldsymbol{\xi}}^{(1)},\breve{\boldsymbol{\xi}}^{(1)},\hat{\boldsymbol{\xi}}^{(2)}\breve{\boldsymbol{\xi}}^{(2)},\boldsymbol{\alpha},\boldsymbol{\beta},\hat{\boldsymbol{s}},\breve{\boldsymbol{s}},\hat{\boldsymbol{p}},\breve{\boldsymbol{p}},\hat{\boldsymbol{t}},\breve{\boldsymbol{t}},\hat{\boldsymbol{q}},\breve{\boldsymbol{q}},\hat{\boldsymbol{v}}_1,\breve{\boldsymbol{v}}_1,\hat{\boldsymbol{v}}_2,\breve{\boldsymbol{v}}_2):\hat{\boldsymbol{s}},\breve{\boldsymbol{s}},\hat{\boldsymbol{p}},\breve{\boldsymbol{p}},\hat{\boldsymbol{t}},\breve{\boldsymbol{t}},\hat{\boldsymbol{q}},\breve{\boldsymbol{q}},\hat{\boldsymbol{v}}_1,\breve{\boldsymbol{v}}_1,\hat{\boldsymbol{v}}_2,\breve{\boldsymbol{v}}_2\geqslant\boldsymbol{0}\} \tag{20}$$

因此,原始问题(18)的沃尔夫-对偶(Wolfe-dual)问题可以由 Karush-Kuhn-Tucker(KKT)定理得到. KKT 条件如下:

$$\frac{\partial L}{\partial \boldsymbol{w}_1}=\boldsymbol{w}_1+\sum_{i=1}^{m}\alpha_i\boldsymbol{x}_i^{(1)}-\sum_{i=1}^{m}\hat{s}_i\boldsymbol{x}_i^{(1)}+\sum_{i=1}^{m}\breve{s}_i\boldsymbol{x}_i^{(1)}+\sum_{j=1}^{n}\hat{t}_j\boldsymbol{x}_j^{(2)}-\sum_{j=1}^{n}\breve{t}_j\boldsymbol{x}_j^{(2)}+(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)=0 \tag{21}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_2}=\boldsymbol{w}_2+\sum_{j=1}^{n}\beta_i\boldsymbol{x}_j^{(2)}+\sum_{i=1}^{m}\hat{s}_i\boldsymbol{x}_i^{(1)}-\sum_{i=1}^{m}\breve{s}_i\boldsymbol{x}_i^{(1)}-\sum_{j=1}^{n}\hat{t}_j\boldsymbol{x}_j^{(2)}+\sum_{j=1}^{n}\breve{t}_j\boldsymbol{x}_j^{(2)}+(\hat{\boldsymbol{v}}_2-\breve{v}_2)=0 \tag{22}$$

$$\frac{\partial L}{\partial b_1}=\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\hat{s}_i-\sum_{i=1}^{m}\breve{s}_i-\sum_{j=1}^{n}\hat{t}_j-\sum_{j=1}^{n}\breve{t}_j=0 \tag{23}$$

$$\frac{\partial L}{\partial b_2}=\sum_{j=1}^{n}\beta_j+\sum_{i=1}^{m}\hat{s}_i-\sum_{i=1}^{m}\breve{s}_i-\sum_{j=1}^{n}\hat{t}_j+\sum_{j=1}^{n}\breve{t}_j=0 \tag{24}$$

$$\frac{\partial L}{\partial \boldsymbol{\eta}_i}=\eta_i-\alpha_i=0,\alpha_i=\eta_i \tag{25}$$

$$\frac{\partial L}{\partial \boldsymbol{\varphi}_j}=\varphi_j-\beta_j=0,\beta_j=\varphi_j \tag{26}$$

$$\frac{\partial L}{\partial \xi_i^{(1)}}=c_1-\hat{s}_i-\hat{p}_i=0,\quad \hat{p}_i=c_1-\hat{s}_i,0\leqslant\hat{s}_i\leqslant c_1 \tag{27}$$

$$\frac{\partial L}{\partial \xi_i^{(1)}}=c_1-\breve{s}_i-\breve{p}_i=0,\quad \breve{p}_i=c_1-\breve{s}_i,0\leqslant\breve{s}_i\leqslant c_1 \tag{28}$$

$$\frac{\partial L}{\partial \xi_i^{(2)}}=c_1-\hat{t}_j-\hat{q}_j=0,\quad \hat{q}_j=c_1-\hat{t}_j,0\leqslant\hat{t}_j\leqslant c_1 \tag{29}$$

$$\frac{\partial L}{\partial \xi_i^{(1)}}=c_1-\breve{t}_j-\breve{q}_j=0,\quad \breve{q}_j=c_1-\breve{t}_j,0\leqslant\breve{t}_j\leqslant c_1 \tag{30}$$

$$\frac{\partial L}{\partial \boldsymbol{z}}=c_2\boldsymbol{e}-\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1-\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2=0 \tag{31}$$

由式(21)可得:

$$\boldsymbol{w}_1=\sum_{i=1}^{m}(\hat{s}_i-\breve{s}_i-\alpha_i)\boldsymbol{x}_i^{(1)}-\sum_{j=1}^{n}(\hat{t}_j-\breve{t}_j)\boldsymbol{x}_j^{(2)}-(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1) \tag{32}$$

同样地,由式(22)可得:

$$\boldsymbol{w}_2=\sum_{j=1}^{n}(\hat{t}_j-\breve{t}_j-\beta_i)\boldsymbol{x}_j^{(2)}-\sum_{i=1}^{m}(\hat{s}_i-\breve{s}_i)\boldsymbol{x}_i^{(1)}-(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2) \tag{33}$$

由式(23)可知,有如下关系存在:

$$\sum_{i=1}^{m}(\alpha_i-\hat{s}_i+\breve{s}_i)+\sum_{j=1}^{n}(\hat{t}_j-\breve{t}_j)=0 \tag{34}$$

同样地,由式(24)可知,有如下关系成立:

$$\sum_{j=1}^{n}(\beta_j-\hat{t}_j+\breve{t}_j)+\sum_{i=1}^{m}(\hat{s}_i-\breve{s}_i)=0 \tag{35}$$

根据式(31)可得:

$$c_2\boldsymbol{e}=\hat{\boldsymbol{v}}_1+\breve{\boldsymbol{v}}_1+\hat{\boldsymbol{v}}_2+\breve{\boldsymbol{v}}_2 \tag{36}$$

将以上 KKT 条件代入式(19)中,可以得到简化的拉格朗日函数式(37).

$$\begin{aligned}L&=-\frac{1}{2}\sum_{i=1}^{m}\alpha_i^2-\frac{1}{2}\sum_{j=1}^{n}\beta_j^2-\frac{1}{2}\|\boldsymbol{w}_1\|^2-\frac{1}{2}\|\boldsymbol{w}_2\|^2+\sum_{i=1}^{m}(\hat{s}_i-\breve{s}_i)+\sum_{j=1}^{n}(\hat{t}_j-\breve{t}_j)\\&=-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}-\frac{1}{2}\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{w}_1-\frac{1}{2}\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_2+\boldsymbol{e}^{\mathrm{T}}(\hat{\boldsymbol{s}}-\breve{\boldsymbol{s}})+\boldsymbol{e}^{\mathrm{T}}(\hat{\boldsymbol{t}}-\breve{\boldsymbol{t}})\end{aligned} \tag{37}$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_m,)^{\mathrm{T}},\boldsymbol{\beta}=(\alpha_1,\cdots,\alpha_m,)^{\mathrm{T}},\hat{\boldsymbol{s}}=(\hat{s}_1,\cdots,\hat{s}_m)^{\mathrm{T}},\breve{\boldsymbol{s}}=(\breve{s}_1,\cdots,\breve{s}_m)^{\mathrm{T}},\hat{\boldsymbol{t}}=(\hat{t}_1,\cdots,\hat{t}_n)^{\mathrm{T}},\breve{\boldsymbol{t}}=(\breve{t}_1,\cdots,\breve{t}_n)^{\mathrm{T}}$.

$\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 可以用矩阵和向量的形式表示出来,如式(38)和(39)所示.

$$\begin{aligned}\boldsymbol{w}_1&=\boldsymbol{A}^{\mathrm{T}}(\hat{\boldsymbol{s}}-\breve{\boldsymbol{s}}-\boldsymbol{\alpha})-\boldsymbol{B}^{\mathrm{T}}(\hat{\boldsymbol{t}}-\breve{\boldsymbol{t}})-(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)\\&=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{s}-\boldsymbol{s})-\boldsymbol{B}^{\mathrm{T}}\boldsymbol{t}-(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)\end{aligned} \tag{38}$$

$$\begin{aligned}\boldsymbol{w}_2&=\boldsymbol{B}^{\mathrm{T}}(\hat{\boldsymbol{t}}-\breve{\boldsymbol{t}}-\boldsymbol{\beta})-\boldsymbol{A}^{\mathrm{T}}(\hat{\boldsymbol{s}}-\breve{\boldsymbol{s}})-(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)\\&=\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{\beta})-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{s}-(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)\end{aligned} \tag{39}$$

其中,$\boldsymbol{s}=\hat{\boldsymbol{s}}-\breve{\boldsymbol{s}},\boldsymbol{t}=\hat{\boldsymbol{t}}-\breve{\boldsymbol{t}}$. 将拉格朗日函数式(37)写成矩阵的形式如下:

$$\boldsymbol{L}=-\left\{\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}+\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}+(\boldsymbol{s}-\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{A}\,\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{s}-\boldsymbol{\alpha})\right\}$$

$$
\begin{aligned}
&-2(\boldsymbol{s}-\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{t}-2(\boldsymbol{s}-\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{A}(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)\\
&+\boldsymbol{t}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{t}+2\boldsymbol{t}^{\mathrm{T}}\boldsymbol{B}(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)+(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)^{\mathrm{T}}(\hat{\boldsymbol{v}}_1-\breve{\boldsymbol{v}}_1)\\
&+(\boldsymbol{t}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{t}-\boldsymbol{\beta})-2(\boldsymbol{t}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{s}\\
&-2(\boldsymbol{t}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{B}(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)+\boldsymbol{s}^{\mathrm{T}}\boldsymbol{A}(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)\\
&+(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)^{\mathrm{T}}(\hat{\boldsymbol{v}}_2-\breve{\boldsymbol{v}}_2)
\end{aligned}
\tag{40}
$$

最终,问题式(18)的对偶问题可以写成如下简洁的形式:

$$
\begin{aligned}
\max\ &-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\theta}+\boldsymbol{e}^{\mathrm{T}}\boldsymbol{s}+\boldsymbol{e}^{\mathrm{T}}\boldsymbol{t}\\
\text{s.t.}\ \ &-c_1\boldsymbol{e}\leqslant\boldsymbol{s}\leqslant c_1\boldsymbol{e}\\
&-c_1\boldsymbol{e}\leqslant\boldsymbol{t}\leqslant c_1\boldsymbol{e}\\
&c_2\boldsymbol{e}=\hat{\boldsymbol{v}}_1+\breve{\boldsymbol{v}}_1+\hat{\boldsymbol{v}}_2+\breve{\boldsymbol{v}}_2
\end{aligned}
\tag{41}
$$

其中,$\boldsymbol{\theta}=(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{s},\boldsymbol{t},\hat{\boldsymbol{v}}_1,\breve{\boldsymbol{v}}_1,\hat{\boldsymbol{v}}_2,\breve{\boldsymbol{v}}_2)^{\mathrm{T}}\in\mathbf{R}^{2m+2n+4d}$

$$
\boldsymbol{Q}=\begin{pmatrix}
\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{I} & 0 & \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} & \boldsymbol{A} & -\boldsymbol{A} & \boldsymbol{\theta} & 0\\
0 & \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{I} & \boldsymbol{B}\boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} & 0 & 0 & \boldsymbol{B} & -\boldsymbol{B}\\
\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} & \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} & -\boldsymbol{A} & \boldsymbol{A} & \boldsymbol{A} & -\boldsymbol{B}\\
\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} & -\boldsymbol{B}\boldsymbol{A}^{\mathrm{T}} & 2\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} & \boldsymbol{B} & -\boldsymbol{B} & -\boldsymbol{B} & \boldsymbol{B}\\
\boldsymbol{A}^{\mathrm{T}} & 0 & -\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{I} & -\boldsymbol{I} & 0 & 0\\
-\boldsymbol{A}^{\mathrm{T}} & 0 & \boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{B}^{\mathrm{T}} & -\boldsymbol{I} & \boldsymbol{I} & 0 & 0\\
0 & \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{B}^{\mathrm{T}} & 0 & 0 & \boldsymbol{I} & -\boldsymbol{I}\\
0 & -\boldsymbol{B}^{\mathrm{T}} & -\boldsymbol{A}^{\mathrm{T}} & \boldsymbol{B}^{\mathrm{T}} & 0 & 0 & -\boldsymbol{I} & \boldsymbol{I}
\end{pmatrix}_{(2m+2n+6d)\times(2m+2n+6d)}
$$

矩阵 $\boldsymbol{Q}$ 的大小为$(2m+2n+6d)\times(2m+2n+6d)$,和用于二元分类的特征选择算法中的二次型矩阵的大小处在相同的数量级上[48]. 因此,当固定标签值变量时,通过求解对偶问题(41)就可以获得拉格朗日乘子的最优解,并借助式(38)和(39)从中进一步得到两个非平行超平面的权值向量最优解. 最优偏移量可以从以下的互补松弛定理中得到:

$$
\hat{s}_i[1-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)+(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)-\boldsymbol{\xi}_i^{(1)}]=0 \tag{42}
$$

$$
\breve{s}_i[(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_1)-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_i^{(1)}+b_2)-1-\boldsymbol{\xi}_i^{(1)}]=0 \tag{43}
$$

$$
\hat{t}_j[1-(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)+(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-\boldsymbol{\xi}_j^{(2)}]=0 \tag{44}
$$

$$
\hat{t}_j[(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_2)-(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j^{(2)}+b_1)-1-\boldsymbol{\xi}_j^{(2)}]=0 \tag{45}
$$

我们将这一寻找最优分割超平面的过程称为紧凑非平行超平面 SVM(tigher NHSVM),该过程结合了拉普拉斯函数和 L-无穷范数. 最终的计算结果说明,从两个非平行超平面的权值向量中抽取得到的两个特征子集基本一致,因此实现了两个超平面上的特征选择同步化.

至此,我们可以描述 IT-NHSVC-SFS 算法的整体框架,如下.

(*a*)初始化:先用简单的聚类算法例如 k-means(KM),对所有的标签值进行初始化[6].

(*b*)固定标签值,通过训练 tighter NHSVM 模型得到两个稀疏的非平行超平面的权值向量,实现特征抽取协同性.

(*c*)利用当前的最优稀疏超平面,根据判别式 $y=\arg\min\limits_{k=1,2}\dfrac{|\boldsymbol{w}_k^{\mathrm{T}}+b_k|}{\|\boldsymbol{w}_k\|}$对样本数据进行重新标记.

(*d*)通过轮流进行聚类过程(*c*)和非平行超平面的参数训练过程(*b*),就可以实现样本标签和模型参数之间的交替优化.

(*e*)当满足预先设定的停止条件时,即可终止算法. 然后分别从得到的两个最优分割超平面的权值向量中抽取出前 r 个绝对值最大的权值对应的特征维度. 最后返回经过特征抽取的低维度的最优分割超平面.

值得一提的是,原始问题的对偶形式(41)有两个重要的特性:第一,由于第2.3 部分介绍的 RFE 算法是建立在对偶形式之上的,所以 RFE 算法同样可以适用于对偶问题(41)来进行特征选择;第二,由于 IT-NHSVC-SFS 算法的公式构造是建立在对偶形式上的,因此可以很方便地利用核函数将算法从线性情形推广到非线性情形.

4　数值实验

本文提出的聚类算法 IT-NHSVC-SFS 具有以下三种特性:首先,由于它是建立在 NHSVM 模型的基础之上的,因此继承了“最大间隔”的性质,一定程度上保证了模型能够具有良好的泛化能力;其次,它实现了两个分割超平面的特征选择同步化,使得算法收敛时,从两个超平面抽取得到的特征子集是相似的;最后,由于该聚类算法通过特征选择过程剔除了那些可能会带来数据噪音的特征,因此理论上这种聚类算法相比于其他算法具有更高的聚类精度. 为了实际验证上述的特性(第一个特性在[20]中已经被证实过),在第 4.1 部分,我们先在一些二元分类的数据集上进行一系列数值实验,目的是为了先验证:当标签值固定时,IT-NHSVC-SFS 算法为两个非平行分割超平面选择的两组特征子集是基本一致的,能够保证特征消除过程在两个超平面上的协同性. 在第 4.2 部分,本文在几个无标签值的

UCI 数据集[49]上设计了一组比较实验来验证 IT-NHSVC-SFS 算法具有良好的聚类表现,尤其是在聚类精度方面,相比于其他算法具有明显的优越性.

4.1 特征选择与同步化

本文算法提出的其中一种假设是它可以实现特征选择同步化,即能够确保两个分类器最终保留的相关特征的子集是相似的. 相比之下,像 twin SVM 这类方法,两个最优分割超平面是分别通过求解两个相互独立的 QP 问题得到的,所以最终针对两个分割超平面进行特征选择得到的特征子集是很不一样的. 为了验证本文算法在特征选择上的协同性(同步化),我们将在已知样本标签值的情况下,运行 IT-NHSVC-SFS 算法,选取的数据集是一个 UCI 数据集,一个常用于验证特征选择算法的 Sonar 数据集,六个 microarray 数据集,分别是 Gravier's breast cancer[50], Alon's colon cancer[51], Alizadeh's lymphoma[52], West's breast cancer[53], Pomeroy's, central nervous system embryonal tumor[54] 和 Shipp's lymphoma[55]. 表 1 提供了这些数据集的特性. 我们选择了 TWSVM-RFE [56](使用 twin SVM),以及建立在 NHSVM 基础上的 NHSVM-RFE 和 L1-TWSVM(见式(12)和(13))三种模型与本文算法 IT-NHSVC-SFS 作比较. 在每个数据集上使用 10-折交叉验证技术进行参数选择和验证. 在进行特征选择之前,参数 $c_i(i=1,\cdots,4)$(分别对应于 Twin SVM, NHSVM, L1-TWSVM 和 IT-NHSVC-SFS)的值在 $c_i \in \{2^{-7},2^{-7},\cdots,2^{6},2^{7}\}$范围内选择. 对于那些基于核方法的算法,我们可以使用径向基函数(RBF),核宽度参数设为 $\gamma = 1/2\sigma^2 = 1/r$,其中 r 是选出的特征总数. 对于这四种算法,令 $c_1 = c_2$,对于 Twin SVM 和 L1-TWSVM,设 $c_3 = c_4$,这样就可以缩小网格搜索法的搜索空间.

前面提到的所有特征选择策略都是先在所有的特征变量集合上进行训练的,然后根据重要性进行特征维度的排序. 因此特征选择是在模型参数训练之后进行的,但是对于 IT-NHSVC-SFS 算法来说,特征选择过程在两个非平行超平面上确实协同的. 需要选出的特征子集的集合大小 n 从{20,50,100,250,500,1000}范围内选择,而对于 Sonar 数据集,$n \in \{5,10,20,30,40,50\}$. 对于四种算法,我们首先对模型参数进行训练,对于 L1-TWSVM 和 IT-NHSVC-SFS,根据两个超平面的每个特征维度的权值大小的绝对值(体现每个特征的重要性),对所有特征进行降序排列;而对于 TWSVM-RFE 和 NHSVM-RFE,则根据模型自定义的重要性准则对所有特征进行降序排列. 然后针对每一个分割超平面,抽取出前 n 个最重要的特征. 另外引入了两组二元值的示性变量,用来说明每个超平面的特征选择的情况(对于某一超平面,若特征 j 被选中,则记为 1;否则记为 0). 最后,计算两个示性向量(每个示性向量对应一个超平面)之间的皮尔森相关系数(Pearson's correlation)[57],该相关系数可以用来判别两个分割超平面的特征抽取过程的同步化程度. 同步化程度和该相关系数成正比关系. 也就是说,相关系数越大,同步化程度越高;相关系数越小,甚至为负数,则说明经过特征选择过程之后,两个超平面留下来的两组相关特征很不相同. 对于所有的 n,表 2 和表 3 分别从最大皮尔森相关系数和平均皮尔森相关系数的角度总结了所有数据集上的实验结果.

表 1 7 个数据集各自的特征总数,样本总数以及每个类别的样本数(minority;majority)

Dataset	#features	#examples	#class(min. ,maj.)
SONAR	60	208	(97;111)
ALON	2,000	62	(22;40)
GRAVIER	2,905	168	(57;111)
ALIZADEH	4,026	96	(35;61)
POMEROY	7,128	60	(21;39)
WEST	7,129	49	(24;25)
SHIPP	7,129	77	(19;58)

表 2 在 7 个数据集上,各算法在所有可能的 n 的取值上的最大皮尔森相关系数

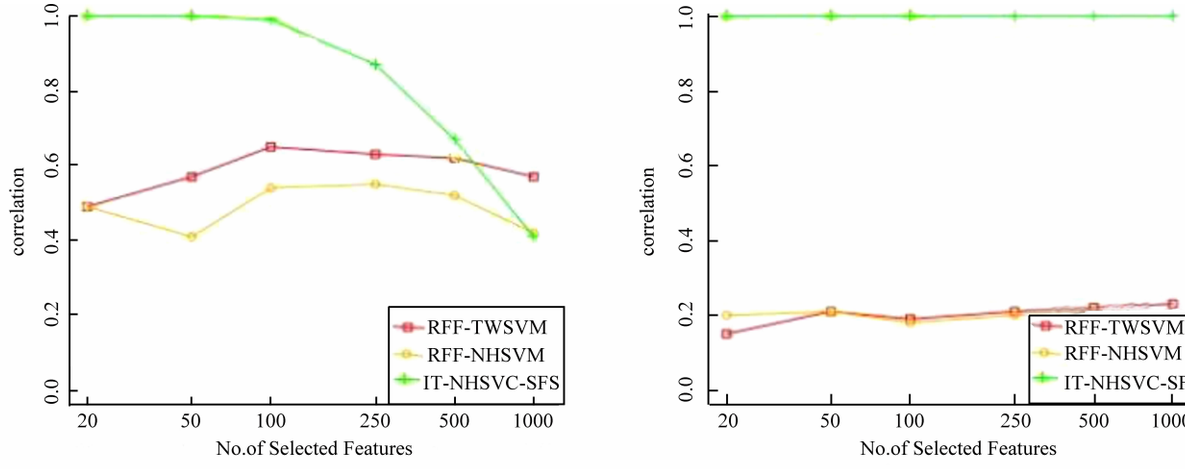
Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
TWSVM-RFE	0. 83	0. 69	0. 27	0. 35	0. 66	0. 80	0. 42
NHSVM-RFE	0. 47	0. 53	0. 69	0. 33	0. 79	0. 84	0. 48
L1-TWSVM	1. 00	0. 92	0. 95	0. 91	0. 97	0. 98	0. 92
IT-NHSVC-SFS	1. 00	1. 00	0. 95	1. 00	1. 00	1. 00	1. 00

表 3 在 7 个数据集上,各算法在所有可能的 n 的取值上的平均皮尔森相关系数

Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
TWSVM-RFE	0. 54	0. 57	0. 26	0. 23	0. 60	0. 64	0. 25
NHSVM-RFE	0. 46	0. 48	0. 52	0. 29	0. 83	0. 74	0. 41
L1-TWSVM	0. 39	0. 78	0. 73	0. 64	0. 72	0. 84	0. 57
IT-NHSVC-SFS	1. 00	0. 84	0. 86	1. 00	0. 98	1. 00	0. 89

从表 2 和表 3 中,可以看出,相比于其他算法,IT-NHSVC-SFS 在最大和平均两个衡量角度上均取得了最大的皮尔森相关系数值,且接近于 1. TWSVM-RFE 和 NHSVM-RFE 的相关系数要低很多,但是 NHSVM-RFE 模型在识别两个超平面的共同相关特征方面比 TWSVM-RFE 要好. 这样的实验结果并没有超出我们的预期,因为 IT-NHSVC-SFS 算法是选择 NHSVM 作为基础模型的,仅仅需要求解一个 QP 问题就可以同时得到两个最优分割超平面,因此避免了 TWSVM 中两个独立的求解过程,而正是由于这种独立性,导致了经过特征选择之后,两个超平面剩下的特征子集的不一致性. 另外,实验中 IT-NHSVC-SFS 识别共同相关特征的能力比 NHSVM-RFE 强,说明了 L-无穷范数在协同性特征选择中发挥的巨大作用,和 RFE 思想相比,很大程度上降低了计算复杂度,适合高维空间数据的降维处理.

同时,我们还研究了皮尔森相关系数随着特征子集大小的函数变化情况,实验结果绘制于图 1. 图中标出了 n 的所有可能取值对应的函数值. 由于空间的限制,我们仅仅在特征选择领域最为熟知的 Alon 和 Alizadeh 数据集[58~60]上进行实验. 从图 1 中可以明显得出,当 n 小于等于 100 时,IT-NHSVC-SFS 的相关系数接近于 1. 如图 1(a)所示,对于 Alon 数据集,当减小特征选择子集的大小时,IT-NHSVC-SFS 的相关系数就会增大;而在图 1(b)中,对于 Alizadeh 数据集而言,即使特征选择子集的大小在不断增加,IT-NHSVC-SFS 的同步化程度依然很高,且相关系数接近于 1. 因此,我们可以得出结论,选择有限数量的特征数 n 有利于提高两个分割超平面上特征选择的同步化程度,这样就可以使得两个超平面的特征子集的并集尽可能地小,从而降低存储消耗,在预测新样本的标签值时减少获取特征维度数据的时间开支.



(a) ALON dataset　　(b) ALIZADEH dataset

图1　皮尔森相关系数关于n的函数变化趋势

4.2　聚类精度和训练速度

在该部分,我们进行了一系列对比实验,来验证本文提出的新算法具有良好的聚类表现. 选择一组无标签值的 UCI 数据集(ionosphere, letter, digits and satellite)和标准数据集(image and ringnorm). 对于 digits 数据集,我们采用和文献[61]一样的处理方式,只关注几对较难区分的数字(3 vs 8,2 vs 7,1 vs 7,8 vs 9). 对于 satellite 和 letter 数据集,由于存在多个类别,我们仅选择前两个类别对应的数据点来进行二元聚类实验,satellite 数据集中选择 C1 和 C2 类别,letter 数据集中选择 A 和 B 类别. 另外,考虑到 ringnorm 数据集的规模较大,所以随机抽取出 500 个样本点. 由于空间限制,并且为了简洁起见,该部分省略了利用网格搜索法进行参数(控制模型复杂度的参数)选择的过程. 后续的所有实验中,对于每一个数据集和每一个聚类算法,均是建立在最优参数组合的基础上进行的. 所有实验均采用了高斯核函数 $\exp(-\|x\|^2/\sigma^2)$. 为了选择核宽度参数值 σ,我们先对每个数据集的训练样本的最大距离作粗略估计(D),然后将 σ 值设为 D 的 2~5 倍. 其中,D 定义为 $D=\left(\sum_{k=1}^{d}[\max\{X^k\}-\min\{X^k\}]^2\right)^{1/2}$,$X^k$ 是数据集样本的第 k 个特征(属性)值.

4.2.1　聚类精度

我们选择以下的聚类算法进行对比实验:①k-Means clustering(KM)[5];②Normalized Cut(NC)[6](使用高斯关联函数);③ Maximum Margin Clustering(MMC)[19];④ Generalized Maximum Margin Clustering(GMMC)[61];⑤TWSVC[21];⑥IT-NHSVC-SFS. 在 TWSVC 中,使用的是铰链损失函数,而在 IT-NHSVC-SFS 中使用的是拉普拉斯函数作为损失度量. 对于所有具有

迭代交替优化性质的算法如 TWSVC 和 IT-NHSVC-SFS,初始的类别标签采用的是 k-means 聚类过程. 由于这些是具有局部优化性质的交替优化算法,它们对局部最优解较为敏感,所以在 k-means 算法中应该尽可能使初始的代表点相隔较远. 另外,对于这些局部优化的聚类算法,我们会在每个数据集上进行10 次重复实验,每次重复实验的 k-means 算法中的初始化步骤的随机种子不同,最后求取 10 次重复实验的平均值获得较为稳定的实验结果. 对于 TWSVC,设 $k=2$,仅仅建立两个子簇中心平面.

对于 NC/MMC/GMMC,数据集 inonosphere 和 digits 的实验结果来源于文献[61]. 由于 MMC 和 GMMC 在大数据集上的训练效率较低,所以没有列出相关的实验结果. 在聚类精度方面的实验结果见表 4. 由于 MMC、GMMC、TWSVC 和 IT-NHSVC-SFS 均具有“最大间隔”的性质,所以它们最终的聚类结果所对应的样本“间隔”是所有其他可能的聚类结果中最大的. 因此,可以推测出,MMC、GMMC、TWSVC 和 IT-NHSVC-SFS 在大多数的数据集上取得的聚类精度比那些没有“最大间隔”特性的聚类算法高. 从表 4 中,也可以看出,在大多数数据集上,较好的聚类结果基本是 MMC、GMMC、TWSVC 和 IT-NHSVC-SFS 算法取得的. 然而,由于巨大的计算开销,MMC 和 GMMC 在处理规模稍大一些的数据集(letter, satellite, image 和 ringnorm)时,无法在一个合理的时间范围内产生聚类结果,因此表格中的对应位置为空. 此外,由于本文提出的 IT-NHSVC-SFS 算法中包含了特征选择过程,因此来自于不相关特征的负面影响在求解两个非平行分割超平面的过程中就会被自动减弱,使得该算法的聚类精度比其他不含有特征选择过程的聚类算法精度要高. 正如表 4 中的实验结果所示,黑体部分代表的是每个数据集上最好的聚类结果,可以发现,IT-NHSVC-SFS 在所有数据集上的聚类精度都是最高的,这和前文的预期设想保持一致.

4.2.2 聚类速度

由于 MMC 和 GMMC 模型需要求解半正定规划问题,它们通常的计算复杂度会非常高. 而在本文的算法中,采用了迭代交替优化的策略,从而避免去求解关于整型标签值变量和连续型模型参数变量的非凸优化问题,使得新算法具有较好的计算效率. 表 5 记录了各个聚类算法的训练时间. 但是为了方便算法在速率上的比较,有必要去运行 MMC 算法. 考虑到 MMC 的计算复杂度问题,我们从数据集中分别随机抽取 50 个样本点组成四个小的数据子集. 实验的其他设定和文献[19]保持一致. 从表 5 中可得,TWSVC 算法的训练速度比 MMC 快上百倍. 由于 IT-NHSVC-SFS 算法在聚类训练过程的同时又另外实现了特征选择,所以比 TWSVC 的训练速度慢,但是仍然比 MMC 算法速度快大约两个数量级,结果如表 5 所示. 尽管 IT-NHSVC-SFS 算法的训练速度比 NC 和 TWSVC 都要慢一些,但是在聚类精度上却具有很明显的优势.

至此,基于上述的数值实验,我们可以得出结论:本文提出的 IT-NHSVC-SFS 算法的确可以实现两个非平行分割超平面的特征选择同步化,相比于那些缺乏特征选择机制的聚类算法,有利于提升模型的聚类精度. 尽管 IT-NHSVC-SFS 算法在训练速度比 TWSVC 要慢一些,但是时间效率仍然高于 MMC 算法.

表 4 各个数据集上的聚类错误率(%)

Data	size	KM	NC	MMC	GMMC	TWSVC	IT-NHSVC-SFS
digits 3-8	357	5.28 ±0	35	10	5.6	3.54 ±0	2.13 ±0
digits 1-7	361	0.57 ±0	45	31.25	2.2	0.59 ±0	0.25 ±0
digits 2-7	356	3.12 ±0	34	1.25	0.5	2.78 ±0	0.0 ±0
digits 8-9	354	9.45 ±0	48	3.75	16.0	7.63 ±0	1.37 ±0
ionosphere	351	34 ±18.6	25	21.25	23.5	22.78 ±2.15	18.65 ±12.26
letter	1555	18.23 ±0	23.2	–	–	7.26 ±0	4.19 ±0
satellite	2236	4.16 ±0	4.21	–	–	3.34 ±0	0.16 ±0
image	1010	42.55 ±0	41.2	–	–	40.3 ±0	22.4 ±0
ringnorm	1000	22 ±5.79	22.3	–	–	24.7 ±9.28	8.5 ±4.79

表 5　各个算法的运行时间(秒)和聚类精度(%)
(括号内的数字分别代表 TWSVC 和 IT-NHSVC-SFS
算法相对于 MMC 算法在速度上的提升)

	Data	NC	MMC	TWSVC	IT-NHSVC-SFS
Time	digits 3-9	0.04	1257	8.7 (144)	32.5(39)
	digits 8-9	0.12	968	7.4 (131)	27.6(35)
	digits 2-7	0.05	1054	7.5 (141)	25.2(42)
	ionosphere	0.19	774	7.2 (108)	24.9(31)
Error(%)	digits 3-9	24	8	5	2
	digits 8-9	9	4	4	0
	digits 2-7	7	15	9	3
	ionosphere	29	27	25	9

5　结论

本文提出了一种新的聚类算法,简称为 IT-NHSVC-SFS. 该算法在模型参数的求解过程中嵌入了特征选择机制. 选用了 NHSVM 作为算法的基础模型,不仅继承了"最大间隔"的特性来保证模型的泛化能力,而且将 TWSVM 中两个非平行超平面独立的求解过程合并成一个优化问题,有利于为两个非平行分割超平面设计一种同步化的特征选择策略. 而且,在 NHSVM 模型原有的目标函数中,增加了 L-无穷范数正则项来实现两个超平面的协同性特征选择,并引入一组束缚变量将带有无穷范数的非凸优化问题转化成二次凸优化问题. 这种同步化的特征选择方式有很多的优点:相比于其他非同步化的特征选择过程(如 L1-TWSVM),同步化的特征选择策略使得两个超平面对应选择出来的特征子集的并集是最小的,因此降低了存储消耗以及预测新样本时获取特征的开支. 此外,特征抽取过程可以减弱由不相关特征带来的数据噪音的负面影响,某种程度上保证了模型良好的聚类精度. 通过剔除那些不相关的特征维度,可以让我们对实际的数据产生过程有更加清晰的了解,显著增强模型的可理解性和可解释性. 该算法还采用了交替优化策略对样本实例的标签值变量和模型参数变量进行优化,从而避免了关于整型和连续型变量的非凸优化问题的求解. 另外,为了避免这种迭代优化算法的早熟收敛,该算法用拉普拉斯函数来代替铰链损失函数,尽可能地防止优化过程收敛在一个较差的局部最优解上.

IT-NHSVC-SFS 算法虽然是针对二元聚类问题的,但是仍然很有必要将其中的特征选择策略和最大间隔思想推广到多类别聚类问题中. 因此如何设计协同性的特征选择过程过滤掉多个超平面中不相关的和冗余的特征将会是我们下一步的研究重点. 此外,还有必要进一步研究适用于该聚类算法的更高效的实现方法,尤其是在高维空间领域. 运用线性规划公式(linear programming formulations)[62]或者一些高效的优化方法如 Frank-Wolfe 算法[1]可能可以改进本文的聚类算法,提高计算效率. 这些都将是我们后一步的研究工作.

参考文献

[1] 杭文龙,蒋亦樟,刘解放,等. 迁移近邻传播聚类算法[J]. 软件学报,2016,27(11):2796 – 2813.
HANG Wen-long, JIANG Yi-zhang, LIU Jie-fang, et al. Transfer affinity propagation clustering algorithm[J]. Journal of Software,2016,27(11):2796 – 2813. (in Chinese)

[2] 王岩,彭涛,韩佳育,等. 一种基于密度的分布式聚类方法[J]. 软件学报,2017,28(11):2836 – 2850.
WANG Yan, PENG Tao, HAN Jian-yu, et al. Density-based distributed clustering method [J]. Journal of Software, 2017,28(11):2836 – 2850. (in Chinese)

[3] 王卫卫,李小平,冯象初,等. 稀疏子空间聚类综述[J]. 自动化学报,2015,41(8):1373 – 1384.
WANG Wei-wei, LI Xiao-ping, FENG Xiang-chu, et al. A survey on sparse subspace clustering[J]. Acta Automatica Sinica,2015,41(8):1373 – 1384. (in Chinese)

[4] 李向丽,曹晓锋,邱保志. 基于矩阵模型的高维聚类边界模式发现[J]. 自动化学报,2017(11):1962 – 1972.
LI Xiang-li, CAO Xiao-feng, QIU Bao-zhi. Clustering boundary pattern discovery for high dimensional space based on matrix model[J]. Acta Automatica Sinica,2017, 43(11):1962 – 1972. (in Chinese)

[5] HARTIGAN J A, WONG M A. A K-means clustering algorithm[J]. Applied Statistics,1979,28(1):100 – 108.

[6] SHI J, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2000,22(8):888 – 905.

[7] REDNER R A, WALKER H F. Mixture densities, maximum likelihood and the EM algorithm[J]. Siam Review, 1984,26(2):195 – 239.

[8] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[A]. International Conference on Neural Information Processing Systems: Natural and Synthetic[C]. US: MIT Press,2001. 849 – 856.

[9] WANG Y X, XU H. Noisy sparse subspace clustering[A]. International Conference on Machine Learning[C]. US: JMLR,2013. I – 89.

[10] HERSHEY J R, CHEN Z, ROUX J L, et al. Deep clustering: Discriminative embeddings for segmentation and separation[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. US: IEEE, 2016. 31 – 35.

[11] ZHANG X, ZHANG X, LIU H. Self-adapted multi-task clustering[A]. International Joint Conference on Artificial

Intelligence[C]. US:AAAI Press,2016. 2357 – 2363.
[12] ZHANG L, ZHANG Q, DU B, et al. Adaptive manifold regularized matrix factorization for data clustering[A]. Twenty-Sixth International Joint Conference on Artificial Intelligence[C]. Berlin:Springer,2017. 3399 – 3405.
[13] VAPNIK V N. Statistical learning theory[J]. Encyclopedia of the Sciences of Learning, 2008, 41(4): 3185 – 3185.
[14] 田中大,张超,李树江,等. 基于相空间重构与最小二乘支持向量机的时延预测[J]. 电子学报,2017,45(5):1044 – 1051.
TIAN Zhong-da, ZHANG Chao, LI Shu-jiang, et al. Time-delay prediction based on phase space reconstruction and least squares support vector machine[J]. Acta Electronica Sinica, 2017, 45(5): 1044 – 1051. (in Chinese)
[15] 陈素根,吴小俊. 改进的投影孪生支持向量机[J]. 电子学报,2017,45(2):408 – 416.
CHEN Su-gen, WU Xiao-jun. Improved projection twin support vector machine[J]. Acta Electronica Sinica, 2017, 45(2): 408 – 416. (in Chinese)
[16] 高雷阜,赵世杰,于冬梅,等. 耦合负类样本裁剪与非对称错分惩罚的非均衡 SVM 算法[J]. 电子学报,2017,45(12):2978 – 2986.
GAO Lei-fu, ZHAO Shi-jie, YU Dong-mei, et al. Unbalanced support vector machine coupling negative-samples cutting with asymmetric misclassification cost[J]. Acta Electronica Sinica, 2017, 45(12): 2978 – 2986. (in Chinese)
[17] 白海钏,鲍长春,刘鑫. 基于局部最小二乘支持向量机的音频频带扩展方法[J]. 电子学报,2016,44(9):2203 – 2210.
BAI Hai-chuan, BAO Chang-chun, LIU Xin. Audio bandwidth extension method based on local least square support vector machine[J]. Acta Electronica Sinica, 2016, 44(9): 2203 – 2210. (in Chinese)
[18] 储茂祥,王安娜,巩荣芬. 一种改进的最小二乘孪生支持向量机分类算法[J]. 电子学报,2014,42(5):998 – 1003.
CHU Mao-xiang, WANG An-na, GONG Rong-fen. Improvement on least squares twin support vector machine for pattern classification[J]. Acta Electronica Sinica, 2014, 42(5): 998 – 1003. (in Chinese)
[19] XU L, NEUFELD J, LARSON B, et al. Maximum margin clustering[J]. Advances in Neural Information Processing Systems, 2004, 17: 1537 – 1544.
[20] JAYADEVA, KHEMCHANDANI R, CHANDRA S. Twin support vector machines for pattern classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(5): 905 – 910.
[21] WANG Z, SHAO Y H, BAI L, et al. Twin support vector machine for clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2583 – 2588.
[22] KHEMCHANDANI R, PAL A, CHANDRA S. Fuzzy least squares twin support vector clustering[J]. Neural Computing & Applications, 2016, 29(2): 1 – 11.
[23] CHANDRASHEKAR G, SAHIN F. A Survey on Feature Selection Methods[M]. US:Pergamon Press, Inc. 2014.
[24] GUYON I. An Introduction to Variable and Feature Selection[M]. US:JMLR, 2003.
[25] MALDONADO S, WEBER R. A wrapper method for feature selection using support vector machines[J]. Information Sciences, 2009, 179(13): 2208 – 2217.
[26] HSU HH, HSIEH C W, LU M D. Hybrid feature selection by combining filters and wrappers[J]. Expert Systems with Applications, 2011, 38(7): 8144 – 8150.
[27] SEBBAN M, NOCK R. A hybrid filter/wrapper approach of feature selection using information theory[J]. Pattern Recognition, 2002, 35(4): 835 – 846.
[28] YANG C H, CHUANG L Y, YANG C H. IG-GA: A hybrid filter/wrapper method for feature selection of microarray data[J]. Journal of Medical & Biological Engineering, 2010, 30(1): 23 – 28.
[29] YANG Y, ZOU H. A fast unified algorithm for solving group-Lasso penalize learning problems[J]. Statistics & Computing, 2015, 25(6): 1129 – 1141.
[30] Moreno-VegaJ M. High-dimensional feature selection via feature grouping[J]. Information Sciences an International Journal, 2016, 326(C): 102 – 118.
[31] SHAO Y H, CHEN W J, DENG N Y. Nonparallel hyperplane support vector machine for binary classification problems[J]. Information Sciences, 2014, 263(3): 22 – 35.
[32] BRADLEY P S, MANGASARIAN O L. k-Plane clustering[J]. Journal of Global Optimization, 2000, 16(1): 23 – 32.
[33] YUILLE A L, RANGARAJAN A. The concave-convex procedure[J]. Neural Computation, 2003, 15(4): 915 – 936.
[34] CHEUNG P M, KWOK J T. A regularization framework for multiple-instance learning[A]. International Conference on Machine Learning[C]. US: DBLP, 2006. 193 – 200.
[35] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273 – 297.
[36] DENG N, TIAN Y, ZHANG C. Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions[M]. US: Chapman & Hall/CRC, 2012.

[37] SHAO Y H, ZHANG C H, WANG X B, et al. Improvements on twin support vector machines[J]. IEEE Transactions on Neural Networks, 2011, 22(6): 962 – 968.

[38] MANGASARIAN O L, MUSICANT D R. Successive overrelaxation for support vector machines[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1032 – 1037.

[39] BAI L, WANG Z, SHAO Y H, et al. A novel feature selection method for twin support vector machine[J]. Knowledge-Based Systems, 2014, 59(2): 1 – 8.

[40] DUDA R O, HART P E, STORK D G. Pattern Classification[M]. US: Wiley, 2001.

[41] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3): 389 – 422.

[42] BRADLEY P S, MANGASARIAN O L. Feature selection via concave minimization and support vector machines[A]. Fifteenth International Conference on Machine Learning[C]. US: Morgan Kaufmann Publishers Inc, 1998. 82 – 90.

[43] YUAN M, LIN Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society, 2006, 68(1): 49 – 67.

[44] CHEVILLARD S, LAUTER C. A certified infinite norm for the implementation of elementary functions[A]. International Conference on Quality Software[C]. US: IEEE, 2007. 153 – 160.

[45] ZHANG K, TSANG I W, KWOK J T. Maximum margin clustering made practical[J]. IEEE Transactions on Neural Networks, 2009, 20(4): 583 – 596.

[46] SCHÖLKOPF, BERNHARD, SMOLA A J. Learning with kernels[J]. IEEE Transactions on Signal Processing, 2002, 52(8): 2165 – 2176.

[47] BENNETT K P, BREDENSTEINER E J. Duality and geometry in SVM classifiers[A]. Seventeenth International Conference on Machine Learning[C]. US: Morgan Kaufmann Publishers Inc, 2000. 57 – 64.

[48] MALDONADO S, LÓPEZ J. Synchronized feature selection for support vector machines with twin hyperplanes[J]. Knowledge-Based Systems, 2017, 132: 119 – 128.

[49] BACHE K, LICHMAN M. UCI Machine Learning Repository[OL]. http://archive.ics.uci.edu/ml/index.php, 2013.

[50] GRAVIER E, PIERRON G, VINCENTSALOMON A, et al. A prognostic DNA signature for T1T2 node-negative breast cancer patients[J]. Genes Chromosomes & Cancer, 2010, 49(12): 1125 – 1134.

[51] ALON U, BARKAI N, NOTTERMAN D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(12): 6745.

[52] DAVIES A J, ROSENWALD A, WRIGHT G, et al. Transformation of follicular lymphoma to diffuse large B-cell lymphoma proceeds by distinct oncogenic mechanisms[J]. British Journal of Hematology, 2007, 136(2): 286.

[53] WEST M, BLANCHETTE C, DRESSMAN H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(20): 11462 – 11467.

[54] POMEROY S L, TAMAYO P, GAASENBEEK M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression[J]. Nature, 2002, 415(6870): 436.

[55] SHIPP M A, ROSS K N, TAMAYO P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nature Medicine, 2002, 8(1): 68 – 74.

[56] YANG Z M, HE J Y, SHAO Y H. Feature selection based on linear twin support vector machines[J]. Procedia Computer Science, 2013, 17: 1039 – 1046.

[57] PEARSON K. Note on regression and inheritance in the case of two parents[J]. Proceedings of the Royal Society of London, 2006, 58: 240 – 242.

[58] NEUMANN J, SCHNÖRR C, STEIDL G. Combined SVM-based feature selection and classification[J]. Machine Learning, 2005, 61(1-3): 129 – 150.

[59] RAKOTOMAMONJY A. Variable selection using SVM based criteria[J]. Journal of Machine Learning Research, 2003, 3(7-8): 1357 – 1370.

[60] SCHÖLKOPF B, PLATT J, HOFMANN T. Generalized maximum margin clustering and unsupervised kernel learning[A]. International Conference on Neural Information Processing Systems[C]. US: MIT Press, 2006. 1417 – 1424.

[61] DJURIC N, LAN L, VUCETIC S, et al. Budgeted SVM: A toolbox for scalable SVM approximations[J]. Journal of Machine Learning Research, 2013, 14(1): 3813 – 3817.

[62] ÑANCULEF R, FRANDI E, SARTORI C, et al. A novel frank-wolfe algorithm analysis and applications to large-scale SVM training[J]. Information Sciences, 2014, 285(C): 66 – 99.

作者简介

方佳艳 女,1997 年出生,安徽池州人. 电子科技大学信息与软件工程学院本科生,主要研究方向为机器学习、人工智能、计算复杂性理论.
E-mail:fyk80@163.com

刘 峤(通信作者) 男,1974 年生,教授,博导,研究方向为机器学习和知识图谱.