集成特征选择的最优化支持向量机分类器模型研究

赵 宇[1] 陈 锐[1] 刘 蔚[1,2]

(中国科学院科技政策与管理科学研究所 北京 100190)[1] (中国科学院大学 北京 100049)[2]

摘 要 考虑将特征选择集成到支持向量机分类器中,提出集成特征选择的最优化支持向量机分类器——FS-SDP-SVM(Feature Selection in Semi-definite Program for Support Vector Machine)。该模型将每个特征分别在核空间中做特征映射,然后通过参数组合构成新的核矩阵,将特征选择过程与机器分类过程统一在一个优化目标下,同时达到特征选择与分类最优。在特征筛选方面,根据模型参数提出用于特征筛选的特征支持度和特征贡献度,通过控制二者的上下限可以在最优分类和最少特征之间灵活取舍。实证中分别将最优分类(FS-SDP-SVM1)和最少特征(FS-SDP-SVM2)两类集成化特征选择算法与 Relief-F、SFS、SBS 算法在 UCI 机器学习数据和人造数据中进行对比实验。结果表明,提出的 FS-SDP-SVM 算法在保持较好泛化能力的基础上,在多数实验数据集中实现了最大分类准确率或最少特征数量;在人工数据中,该方法可以准确地选出真正的特征,去除噪声特征。

关键词 特征选择,集成化方法,支持向量机分类器,特征核子空间,半正定规划

中图法分类号 TP393.0 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.036

Research on Optimal Support Vector Classifier Model Integrating Feature Selection

ZHAO Yu[1] CHEN Rui[1] LIU Wei[1,2]

(Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China)[1]

(University of Chinese Academy of Sciences, Beijing 100049, China)[2]

Abstract Considering taking the feature selection process into the support vector machine classifier, a new model called feature selection in semi-definite program for support vector machine(FS-SDP-SVM) was proposed in this paper for integrating the target of feature selection and machine classifier. The key to this model is to split the kernel space into several subspace by each feature. With the linear combination of these subspaces, the new kernel matrix was constructed and optimized with the support vector classifier by semi-definite programing. Two parameters for the feature choosing are announced, namely feature supporter and feature contributor, which can be flexibly adjusted for the need of maximizing accurate rate (FS-SDP-SVM1) or minimizing feature quantity (FS-SDP-SVM2). The empirical study analyzed the difference between two model types and other feature selection algorithms Relief-F, SFS and SBS on the UCI machine learning data and man-made data. Results show that FS-SDP-SVM can achieve maximum accurate rate or minimum feature quantity in majority of UCI data in consistent with the good ability of generalization. This method precisely gets rid of the noise data and preserves the real features in man-made data test.

Keywords Feature selection, Ensemble method, Support vector classifier, Sub-kernel space, Semi-definite programming

1 引言

基于数据的机器学习是现代智能技术中的一个重要方面,从观测数据中寻找规律,利用这些规律对未来数据或无法观测的数据进行分析和预测是机器学习领域发展的趋势之一。特征选择模型和以支持向量机(Support Vector Machine,SVM)模型为代表的分类器机器学习领域的核心研究内容。特征选择模型是根据某一特定的准则,从既有样本中选择合适的变量或提取新的特征来更好地实现数据分类的过程。特征选择算法已在网络安全信息挖掘[1]、金融信贷、生物医学、文本识别等领域取得了较好的应用效果[2]。SVM 分类器基于统计学习理论,借助最优化方法,选择与不同类别数据间距离最大的分类间隔来解决数据分类的问题[3],在实际应用中已表现出很多优越的性能,目前已成为研究的热点。随着研究的深入,标准的 SVM 模型也衍生出多种形式,包括控制最大误分类率的 C-SVM、控制支持向量个数的 v-SVM、能够显著提高计算速度的最小二乘 SVM(LS-SVM)[4]和简化 SVM(R-SVM)、多视角 SVM 以及 twin svm 等,可以最优化核参数的半正定规划 SVM(SDP-SVM)[5]等相关新形式都是在参数选择、核函数选取、降低计算时间等方面进行改进。特

到稿日期:2015-07-21 返修日期:2015-10-16 本文受 2013 质检公益性行业科研专项课题(201310118),2015 国家质检公益性行业科研专项课题(201510041),中科院重大任务专项课题(Y201161Z04)资助。

赵 宇(1982—),男,博士,助理研究员,主要研究方向为数据挖掘、最优化算法,E-mail:zhaoyu.05@163.com;陈 锐(1975—),男,博士,研究员,主要研究方向为城市运行管理、智慧城市建设(通信作者);刘 蔚(1986—),女,博士生,主要研究方向为网络优化。

征选择往往被用于机器学习前期的数据预处理过程，之后用处理过的数据进行分类研究，然而在分类问题中，特征选择算法所依据的准则或目标经常与分类器的目标不一致，容易产生选择的特征并不能满足分类模型的最优化条件的问题。

针对这一问题，许多学者在机器学习中研究嵌入(Embedded)和集成化(Ensemble)方法，这是近年来机器学习领域的热点问题，通过将单个分类器模型如神经网络、决策树、朴素贝叶斯等以某种方式(投票等)组合起来，对新的样本进行分类，只要能够保持子分类器具有较高的正确率和一定的差异性[6]，就能得到比单个分类器更优的分类效果，提高学习能力和泛化特性[7-9]。随着研究的深入，在特征选择过程中也经常采用多方法集成的技术，以根据目标(如特征数量最少、分类效果最好)获得不同的特征集；而特征选择嵌入分类器的研究也可以看作是一种集成化方法。Wrapper 方法就是早期经典的集成化特征选择方法[10]，该方法保证了特征选择和机器学习的统一性，但计算时间较长。针对 SVM 分类器，通过"0-范数(最优变量选择)[11]"和"1-范数(套索模型，LASSO)[12]"控制特征数量；之后的研究方向集中在多核函数[13]、多学习器[14]、多种数据的交叉验证、多种特征选择[15]方法上，如图 1 所示。

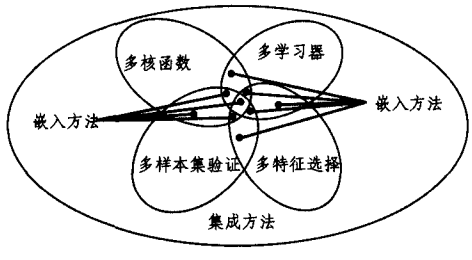

图 1　集成化方法示意图

近年来上述方法的研究也很多，但都有一定的局限性。如文献[16]提出基于有监督的递归分类树学习的特征基因选择算法，其本质上也是一种 Wrapper 方法；文献[17]基于广义粗集的集成特征选择方法，尝试通过特征选择融合多分类器，同时获取各特征空间中的多类模式可分性信息，并提出关于多决策表的相对优势决策约简，然而其特征选择过程基于 Filter 方法，与分类器目标也不一致；文献[18]提出基于选择性集成策略的嵌入式特征选择方法，根据选择性集成策略选取部分特征选择器集成，再改进序列前向搜索和封装器组合方法二次搜索最优特征子集，相当于进行两次特征选择，增加了计算复杂度。为了克服这些问题，本文提出一种新的特征选择和分类模型相结合的方法，利用 SDP-SVM 模型，将特征选择过程集成到核函数的构建和优化中，通过选择最优的特征子空间达到特征选择和最优分类的目标统一。

本文提出的算法基于对 SDP-SVM 模型的改造，在 SDP-SVM 的基础上融入特征选择过程。第 2 节先介绍 SDP-SVM 的基本模型并对其分析；第 3 节介绍在核矩阵构建过程中引入特征选择的过程，提出集成特征选择的最优化支持向量机分类器模型(FS-SDP-SVM)，并提出通过参数控制进行特征选择的规则；第 4 节基于 UCI 数据集开展应用实践；最后进行总结和展望。

2　基于半正定规划的 SVM 模型(SDP-SVM)

支持向量机(SVM)学习模型是从特征空间线性可分条件下寻求最优分界面发展而来的，考虑 1-范数软间隔(1-norm soft margin)的 SVM 最优化模型式(1)为：

$$\begin{aligned}&\min_{w,b,\xi}\langle w,w\rangle+C\sum_{i=1}^{n}\xi_i\\&\text{s. t. } y_i(\langle w,\phi(x_i)\rangle+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,n\end{aligned}\tag{1}$$

引入拉格朗日乘子 α_i 和核函数 $K(x_i,x_j)$，利用 KT 条件，求解式(2)中 α_i 的最大值。

$$\begin{aligned}\omega s1(K,\alpha)&=\langle w^*,w^*\rangle+C\sum_{i=1}^{m}\xi_i\\&=\max_{\alpha}2\alpha^{\mathrm{T}}e-\alpha^{\mathrm{T}}G(K)\alpha\\&\alpha^{T}y=0,0\leqslant\alpha\leqslant C\end{aligned}\tag{2}$$

对应的分类函数为：

$$f(x)=\mathrm{sgn}\{\alpha_i^* y_iK(x_i,x)+b^*\}$$

其中，$K(x_i,x_j)$是核函数，核函数的采用使得线性的 SVM 很容易推广到非线性领域。通过核函数的映射作用避免特征空间的内积运算[19]。常用的核函数形式包括内积核函数、多项式核函数、高斯核函数与 Sigmoid 核函数等。人为指定核函数形式没有基于样本自身特性，容易引起偏差，文献[4]提出的基于半正定规划的 SDP-SVM 模型通过多核矩阵的组合，提高 SVM 模型的预测能力，其 1-范数软间隔 SDP-SVM 模型如式(3)所示：

$$\begin{aligned}&\min_{K,t,\lambda,v,\delta}t\\&\text{s. t. } K\in\mathcal{K}\\&\mathrm{trace}(K)=c\\&\begin{pmatrix}G(K)&(e+v-\delta+\lambda y)\\(e+v-\delta+\lambda y)^{\mathrm{T}}&t-2C\delta^{\mathrm{T}}e\end{pmatrix}\geq 0\\&v\geqslant 0,\delta\geqslant 0\end{aligned}\tag{3}$$

式(3)将 SVM 的核函数参数选择问题与分类最优化问题结合在一起，其中 $\mathrm{trace}(K)=c$ 用于控制核矩阵以及变量 μ 的规模，加入对核函数的额外矩阵迹约束，c 为根据需要预先指定的常数。

式(3)基于若干核函数 k_i 的组合 K，对每一种核函数赋予一定的权重 μ_i 并纳入模型中以求整体最优解。这是多核函数的集成化方法。

3　集成特征选择的 SDP-SVM 模型

3.1　基于特征选择的核矩阵构建

核矩阵是将变量全体从原空间向特征空间映射，基于特征选择构建核矩阵的思路是将样本中每个单独的变量分别进行特征空间映射，并将多个特征子空间通过参数组合构成核矩阵，将该核矩阵代入 SDP-SVM 模型中求解，得到相应的最优条件下的参数值，根据参数值设计特征提取规则实现特征筛选。

定义某一样本集的特征空间 Ω，考虑针对单一变量构建的特征子空间 $\varphi_j(j=1,2,\cdots,n)$，$\Omega=\bigcup_{j=1}^{n}\varphi_j$。给定一种核函数形式 $k(\varphi)$，在 n 个特征子空间内分别计算核矩阵，令 μ_j'为特征子空间核矩阵的权重系数，通过核组合 $K'=\sum_{j=1}^{n}\mu_j'k(\varphi_j)$构建全体变量的特征映射。

将不同特征子集构成的子集空间核矩阵代入模型(3)，令 $G'(K)=y^{\mathrm{T}}K'y$，构建特征选择与 SDP-SVM 分类器的集成模型 FS-SDP-SVM(见式(4))。

$$\min_{K,t,\lambda,v,\delta} t$$

$$\text{s. t. } k\in K'$$

$$\text{trace}(K')=c \tag{4}$$

$$\begin{pmatrix} \sum_{i=1}^{n}\sum_{j=1}^{n}\mu_j' k(\varphi_j) y_i y_j & (e+v-\delta+\lambda y) \\ (e+v-\delta+\lambda y)^{\mathrm{T}} & t-2C\delta^{\mathrm{T}}e \end{pmatrix} \geq 0$$

$$v\geqslant 0,\delta\geqslant 0$$

用原始-对偶内点算法求解(4)获得核子空间的参数 μ_j' 以及此时的最优分界面。

3.2 模型分析

FS-SDP-SVM 模型将每个变量看作一个独立的特征子空间,其实质是对样本特征空间进行分解,并对每个特征子空间分别加权,通过对特征空间不同维度的尺度进行"缩放"来改变数据空间分布构型,以利于寻求最优分界面。

引理 1 FS-SDP-SVM 模型中参数 μ_j' 的大小决定了特征分类显著性的强弱。

证明:FS-SDP-SVM 模型属于半正定凸规划,当 t 达到最小值 $t_{\min}$ 时,未知变量 μ_j',λ,v,δ 尽量取其下限值,根据模型(4)的半正定约束条件,

$$(e+v-\delta+\lambda y)^{\mathrm{T}}\sum_{j=1}^{n}\frac{1}{\mu_j}\left(\sum_{i=1}^{n}k(\varphi_j)y_i y_j\right)^{-1}$$

$$(e+v-\delta+\lambda y)+\leqslant(t-2C\delta^{\mathrm{T}}e)$$

当取=号时,

$$(e+v-\delta+\lambda y)^{\mathrm{T}}(e+v-\delta+\lambda y)=(t_{\min}-2C\delta^{\mathrm{T}}e)\sum_{j=1}^{n}\mu_j\left(\sum_{i=1}^{n}k(\varphi_j)y_i y_j\right)^{-1}$$

显然,μ_j' 与 t 成反比,当 t 达到最小值 $t_{\min}$ 时,μ_j' 达到最大值 $\mu'_{j\_\max}$。$\mu'_{j\_\max}$ 所对应的特征就是需要选出的显著特征,具有较强的分类能力。

证毕。

下面根据 μ_j' 制定特征选择规则进行特征筛选。

3.3 特征选择规则

定义特征贡献度 FS_g 和特征支持度 FS_z。在已知的训练样本中,令 $\omega_j=\frac{|\mu_j'|}{\sum_{j=1}^{n}|\mu_j'|}$,根据 3.2 节的分析结果,将 ω_j 由大到小排列,尽可能选取那些 ω_j 较大的特征。定义 $FS\_g\geqslant\sum\omega_j$,表示特征贡献度是特征参数贡献率的累积,$FS\_g$ 越大(接近于 1),则选到的特征越多,但除了强显著性特征外,还可能会包含不必要的特征;而 FS_g 越小,则仅会选到少量的重要性特征,可能会漏掉必要的特征。

给定一个 FS_g 值,k 次训练可以得到 k 组特征子集 X_i' $(i=1,2,\cdots,k)$,每组特征子集包含的特征有差异,那些对分类有显著性影响的特征子集会在多次训练中被重复选中。根据投票表决的思想,k 次训练后统计所有 X' 包含的每个特征的数目 Num_j 并计算其在 k 次训练中的占比,可以找出显著特征。令 $Num\_rate_j=\frac{Num_j}{k}$,定义特征支持度 FS_z 为支持某一特征入选最终特征子集的上界,$0.5\leqslant FS\_z\leqslant 1$,即 k 次训练中某一特征出现次数至少要过半才可能被选中。当 $Num\_rate_j\geqslant FS\_z$ 时,该特征被选入最终的特征子集 X_{final} 中,进入新样本分类测试过程。

3.4 SVM-SDP 算法步骤

综上所述,基于特征选择的核矩阵分类模型算法的步骤如下所示。

Step1:样本选择。将其分为训练样本集 $\{X_{train},y_{train}\}$ 和检验样本 $\{X_{test},y_{test}\}$。

Step2:核参数预估计。选取一种核函数 $k_\theta(\cdot)$,在训练集上采用 C-SVM 或是 v-SVM 模型训练其参数 θ。

Step3:将 n 个变量每个单独构成一个特征子集,对每个子集分别计算其核矩阵 $k(\varphi_j)$,$j=1,2,\cdots,n$;构造 Gram 核矩阵:

$$G'(K)=y^{T}K'y=\sum_{j=1}^{n}\mu_j' y^{T}k(\varphi_j)y$$

Step4:将训练样本集分成 k 组,分别计算每组的 Gram 矩阵 $G'(K)$ 并代入模型(4)求解得到 k 组 μ_j',λ,v,δ,t 的值。

Step5:给定特征子集的贡献度 FS_g 和支持度 FS_z,计算 $\{FS\_g,FS\_z\}$ 下每个特征子集的贡献度 $\omega_j=\frac{|\mu_j'|}{\sum_{j=1}^{n}|\mu_j'|}$,将 ω_j 由大到小排列后选择刚好满足 $\sum\omega_j\geqslant FS\_g$ 的特征集合作为新的样本集并统计每个特征在 k 次交叉验证计算中出现的次数 Num_j,从中选取 $Num_j\geqslant FS\_z$ 的特征集合作为最终变量子集 X_{final}。

Step6:对照 X_{final} 剔除 X_{test} 中的多余变量并进行分类测试。

通过上述 6 步实现特征子集的选择,这种基于半正定规划 SVM 的特征选择算法称作 SDP-SVM 特征选择算法。

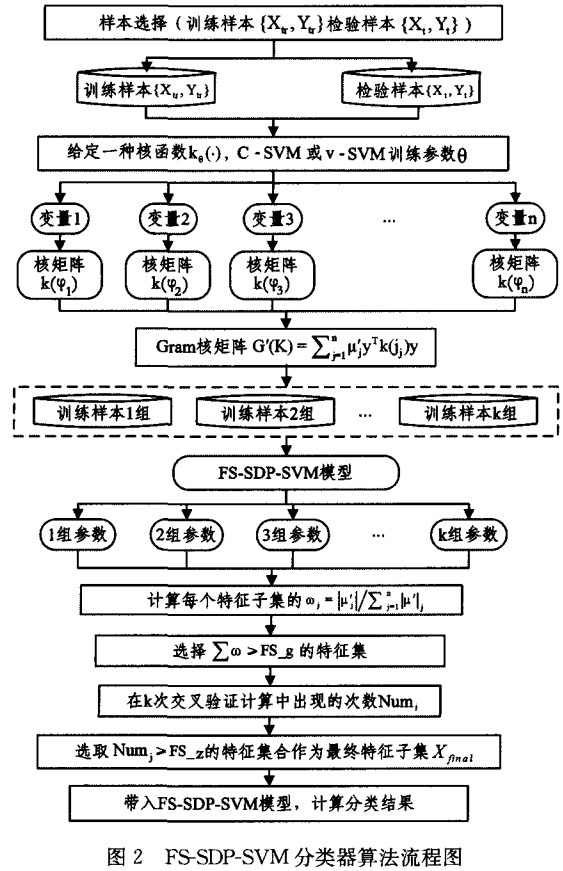

图 2 FS-SDP-SVM 分类器算法流程图

4 模型实验

本文基于 UCI 机器学习数据库和人工构建的多元函数方程进行特征选择的实证分析。采用 5-*fold* 交叉验证训练参数 μ_j',λ,v,δ,t,以降低训练样本选择不当所带来的方差,选取平均误差最小的子集和模型参数作为最终的结果。

实验中，FS_g 越大则特征越多，FS_z 越大则特征越少。通过实验可以选择合适的一组$\{FS_g, FS_z\}$，这组数值可以是取得最优分类准确率的一组（特征数量不一定最少），也可以是特征数目最少的一组（分类准确率在可接受范围内），本文分别对这两种情况展开实验。

4.1 数据介绍

选取的 UCI 数据包括 Breast cancer、Musk、Spam、Vote 数据以及德国和澳大利亚的信用统计数据（下文用 German 和 Australia 代替）。实证考察模型(4)的特征选择情况和分类情况。Breast cancer 数据集是由威斯康星大学附属医院的 William H. Wolberg 博士提供的临床 10 个不同观察因素下病人肺癌肿瘤是良性或恶性的情况；Musk 数据集是专家归纳出的麝香分子可能的特性[20]。数据从不同角度刻画气体分子的特征，根据这些特征找出哪些是真正的麝香分子。Spam 数据是用于根据某些字段判断一封邮件是否为垃圾邮件的数据，4601 个样本中垃圾邮件占 1813 个，58 个变量主要围绕邮件中字或词频统计来刻画邮件属性。Vote 数据是 1984 年美国国会针对众议院 435 名议员在 16 项工作情况中表现的投票记录数据，其中民主党 267 人，共和党 168 人。这 16 项工作主要是残疾人和孤儿保障、移民政策、教育开支、犯罪等，均为布尔型数据（0 和 1）。目的是根据每名众议员在 16 项工作中的表现情况明确其为共和党人还是民主党人。German 包含 1000 个样本，其中有 700 个可信客户和 300 个违约客户。Australia 包含 690 个样本，包括 383 个信用客户和 307 个违约客户。

多元函数方程是人为给定一种自变量与因变量之间的关系表达式：$y=x_1+2x_2+\sin(x_3)+x_4^2+e^{x_5}$，初始的基础自变量为 x_1 至 x_5，这 5 个变量之间相互独立，按正态分布分别独立随机生成 1000 组$[-1,1]$之间的随机数，计算因变量 y 值，取这 1000 个 y 的中位数$\bar{y}$，并加入 10^{-6}量级的扰动项，当$y\geqslant\bar{y}$时，令$y=+1$，当$y<\bar{y}$时，令 $y=-1$，据此得到所有 1000 组基础自变量 x_1 至 x_5 和类别标号变量 y。为了验证算法的有效性，需要增加人工变量进行“干扰”，以此监测算法的学习能力和泛化特性。为此，在 5 个自变量的基础上，新增 7 组人工变量。新增的原则就是加入与原变量相关的变量，加入随机噪声变量。第 6 个变量定义为 $x_6=x_1+1$，显然它与 x_1 的高度相关；第 7 个变量定义为 $x_7=x_2x_3$，与 x_2 和 x_3 相关；最后 5 组变量 $NV_1, NV_2, \cdots, NV_5$ 是$[-1,1]$之间纯随机白噪声变量，按照均匀分布生成。

相关变量和样本数目信息统计如表 1 所列。

表 1 数据集相关信息统计

数据集	变量个数	交叉验证分组数目	每组训练样本个数	样本总量	每组检验样本数量	检验样本组数
Breast cancer	9	5	50	699	100	200
Musk	166	5	50	6598	200	200
Spam	58	5	50	4601	200	200
Vote	16	5	50	435	100	100
German	24	5	50	1000	100	200
Australia	14	5	50	690	100	100
Artificial data	12	5	50	1000	200	200

4.2 初始核矩阵参数选取

在进行特征选择前，先通过经典的 SVM 模型对选定的 RBF 函数参数给予初始值。常用的分类模型有 C-SVC 和 v-SVC，分别侧重于控制误分类样本点个数和控制支持向量个数。本文的核心在于特征子集的组合，因此直接选用高斯核函数作为 SDP-SVM 模型的核矩阵基础形式，用 C-SVC 模型遴选初始核矩阵参数 σ 和罚参数 C。核参数 C 和 σ 是通过 5-*fold* 交叉验证得到的平均值；在校验分类结果准确性上采用特异度（Specificity）、敏感度（Sensitivity）准则以及总分类正确率进行判定。

$$\text{特异度}=\frac{\text{实际 } y=-1 \text{ 被模型正确识别} \hat{y}=-1 \text{ 的数量}}{y=-1 \text{ 的样本总量}}$$

$$\text{敏感度}=\frac{\text{实际 } y=1 \text{ 被模型正确识别} \hat{y}=1 \text{ 的数量}}{y=1 \text{ 的样本总量}}$$

$$\text{总分类正确率}=\frac{\text{样本正确分类的数量}}{\text{总样本量}}$$

4.3 特征选择方法比较和参数设定

采用特征选择方法处理数据是为了提高学习能力或最大限度降低数据规模。基于 SDP-SVM 的特征选择方法按照最大分类准确率和最少特征数量命名为 SDP-SVM1 算法和 SDP-SVM2 算法。为了便于对比，分别采用连续前向选择（Sequential Forward Selection，SFS）、Relief-F 和连续后向选择（Sequential Backward Selection，SBS）3 种特征选择方法进行比较。SFS 和 SBS 属于 wrapper 特征选择中的两种子集搜索算法；Relief-F 属于 filter 特征选择算法的一种。

SFS 是从空子集开始逐次添加变量进入子集，评估每一步变量子集的分类能力，直到添加任何子集都无法改善分类效果为止。

Relief-F 方法考察每个变量对不同类别数据的区分能力，据此确定变量的特征相关度，将特征相关度强的变量选入子集。

SBS 则是 SFS 的反向操作，从所有变量开始逐次删减变量，考察剩余变量构成的子集的分类能力，直到删减任何变量都会降低分类效果为止。表 3 记录了 4 种特征选择方法对 Brease cancer、Musk、Spam、Vote 数据集及人工数据集的筛选结果，限于篇幅，只列出了每组数据选中变量的个数。

贡献度 FS_g 选择为以 0.05 的步进从 0.5 至 1 递增，共 11 个贡献度值；支持度 FS_z 选择为以 0.1 的步进从 0.5 至 1 递增，共 6 个贡献度值，考虑到是 5-*fold* 交叉验证法，支持度的步进不需要过于细致。

4.4 实验结果

在 4 种特征选择的子集下进行基于 RBF 核函数的分类评估工作，计算平均分类正确率。通过 FS_g 和 FS_z 在 0.5～1 区间内的变化，找出最优的分类与特征选择结果，如图 3 所示。

Breast cancer 数据的分类准确率和特征数量均随着贡献度 FS_g 的增加而增加，但支持度 FS_z 的增加却对分类正确率基本无影响，而且特征数目随着 FS_z 的增加下降幅度也不甚明显。在 0.55～0.9 的区间内，FS_g 作用下的分类准确率和 FS_z 作用下的分类准确率基本相同。考虑分类正确率最大的原则，SDP-SVM1 模型会选择所有特征；按照合理分类准确率，尽量少特征的原则，SDP-SVM2 模型选择出准确率为 95.895%的 4 个关键特征。

Musk 数据在分类准确率方面呈现出较多的稳态性，FS_g 作用下的分类准确率在 0.6 之后就几乎没有再增加，对应的 FS_z 在 0.5～0.8 区间内的分类准确率始终处于高位，没有变化。可见 Musk 数据在 SDP-SVM 特征选择算法下可以实现最优效果，即特征数量少，分类准确率高。特征数量随 FS_g 的增加而增长得很快，但对分类准确率呈负贡献，因此，Musk 必然有变量只起了干扰作用。对应最大分类准确率 87.49%的特征数目为 25 个；对应合理的准确率 85.41%的特征数目是 14 个。显然二者很接近，特征差异较为明显。

Spam 数据的一个显著特点是在 0.8 以上，分类准确率会随着 FS_z 的增加而显著增长，说明垃圾邮件筛选需要的仅仅是几个核心特征，对应位置的 FS_g 则会造成分类准确率的波动。在最大分类准确率 75.205%的位置，对应的特征数目仅仅有 4 个；在合理的准确率 74.025%处，仅靠 3 个特征就可以实现，这也证实了我们的想法，即在 SDP-SVM 特征选择模型下，Spam 数据仅仅需要几个核心特征就能实现垃圾邮件的分类处理。

Vote 数据在 FS_g 和 FS_z 位于 0.5～0.9 区间内时二者对分类准确率的作用非常相似。在最大分类准确率 95.87%处，需要的特征数量也较多(15 个)；在合理的分类准确率 95.6%时，对应的特征仅需要 4 个，分类准确率差异并不大，但特征数目差异明显，有较大的特征剔除空间。

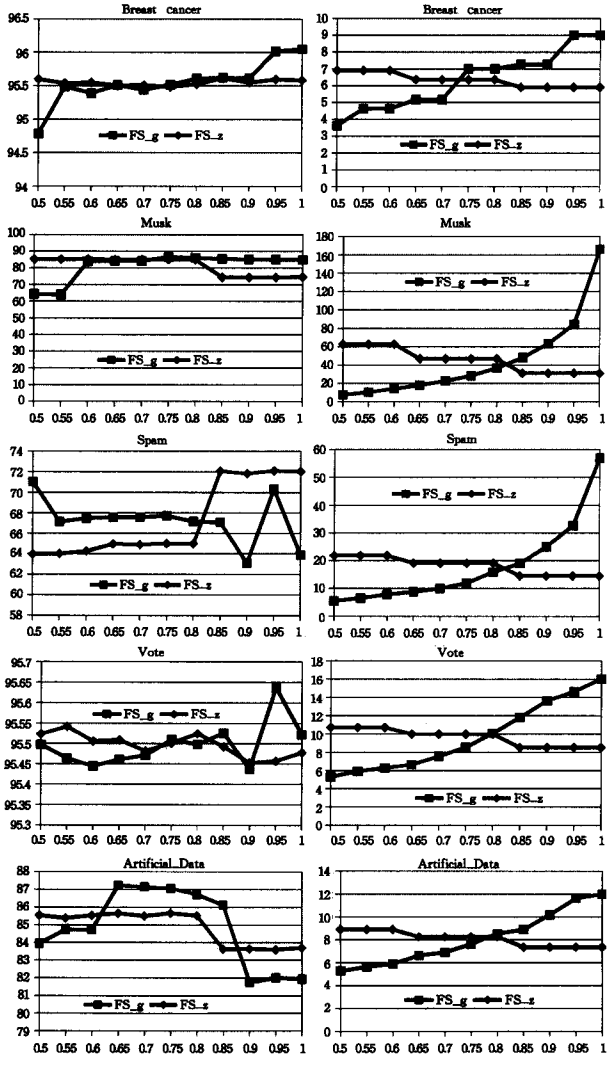

图 3　5 类数据平均分类正确率和特征数量变化曲线图

人工数据集的目的是检验 SDP-SVM 能否选出真正有意义的变量。在 FS_g 和 FS_z 的限制下，在 0.5～0.8 区间内 FS_z 主导的分类准确率处于高位无变化，FS_g 主导的分类准确率从 0.65 的位置就开始下降，因此最优分类准确率 87.64%对应的特征为 1,2,3,4,5,6,8；合理准确率 87.62%对应的特征为 1,2,3,4,5,6。在相差无几的情况下，合理分类准确率能够找出所有的真实特征和一个高度相关特征 (X_6)。

样本的总分类正确率、特征数量、敏感度和特异度如表 2—表 5 所列。

表 2　4 种特征选择算法在不同数据集上的总分类正确率(%)

特征选择方式	Relief-F	SFS	SBS	SDP-SVM	
				1 类	2 类
Breast cancer	94.35	95.52	95.18	96.13	95.64
Musk	85.21	85.03	85.05	86.42	85.56
Spam	63.5	65.25	64.22	75.205	74.025
Vote	95.52	95.44	95.63	95.865	95.645
Artificial data	74.71	87.74	87.59	87.57	87.29

表 3　4 类特征选择方法选中的特征数量

特征选择方式	Relief-F	SFS	SBS	SDP-SVM	
				1 类	2 类
Breast cancer	3	4	5	9	4
Musk	91	54	89	15	14
Spam	27	39	6	4	3
Vote	3	1	5	15	4
Artificial data	4	6	5	7	6

表 4　4 种特征选择算法在不同数据集上的特异度

特征选择方式	Relief-F	SFS	SBS	SDP-SVM	
				1 类	2 类
Breast cancer	97.75	98.18	97.16	97.66	96.86
Musk	63.26	62.86	63.95	63.36	64.15
Spam	79.59	60.54	82.57	62.69	19.24
Vote	96.68	96.85	96.64	96.92	96.79
Artificial data	86.96	81.32	90	91.17	89.11

表 5　4 种特征选择算法在不同数据集上的敏感度

特征选择方式	Relief-F	SFS	SBS	SDP-SVM	
				1 类	2 类
Breast cancer	93.65	90.95	91.7	94.25	93.75
Musk	99.06	98.06	98.61	99.94	98.52
Spam	57.13	99.65	48.25	81.18	98.66
Vote	95.05	94.45	94.69	94.71	94.74
Artificial data	81.19	77.18	89.11	80.46	79.94

下面分别用分类正确率最大，敏感度、特异度差异最小与合理分类准确率下最少特征 3 个评判准则分析 5 种特征选择模型的优劣性。

1)分类正确率最大准则

从表 2—表 5 中可以看出，提出的 SDP-SVM1 算法在 5 组数据中取得了较好的结果。分类结果中 Breast cancer 数据位于第 2；Musk 数据位于第 1；Spam 数据位于第 1；Vote 数据位于第 1；人工数据 Artificial data 位于第 2。特别值得注意的是，在 Musk 数据和 Spam 数据上，所提出的 SDP-SVM 算法能够以 2%以上的优势领先其它特征选择算法(考虑到其它数据的整体差异性并不明显，这种差异已经较为显著)。其它 3 种方法在不同数据上表现各有优劣，在特征选择数量方面，这种考虑最大分类准确率的算法在 Breast cancer 数据集上无法实现特征选择，Breast cancer 来源于真实数据也不一

定需要移除某些特征，毕竟每个特征对分类都有贡献，因此不进行特征选择的分类准确率依然很高。在人工变量数据上SDP-SVM1选择了7个变量，分别是X_1-X_6，X_8，X_{10}，并没有选出真实的变量集。Relief-F表现最差，差异明显，其他方法则都能够通过剔除若干干扰提高分类准确率。

2)合理的分类准确率、尽量少的特征原则

在这种情形下，合理的分类准确率指的是其它3种方法的平均分类准确率，在达到这一准确率的前提下找出尽量少的特征。从表5可以看出，除了Vote和人工数据集外，SDP-SVM2方法在其它3项数据集的特征选择中都取得了最少特征数量，优势明显；特别是Musk数据、Spam数据，选出的特征数量大约仅占总特征数量的2/9，1/7和1/4。SFS虽然在人工数据集上取得了最少特征集，但分类正确率较低，原因在于选中了噪声NV_1。从实用性来讲其并不比SBS好，SBS选择了X_2-X_6，为最好结果。SDP-SVM2在人工数据集选择的变量是X_1-X_6，X_6作为X_1的线性变换，具有完全替代性，可见SDP-SVM2提出的特征选择算法在人造变量数据集上也具有较大的优势。

3)敏感度与特异度差异最小准则

从敏感度、特异度结果来看，Musk的敏感度和特异度差距较大，其中SBS特征选择方法的敏感度高于特异度，其它均是特异度高于敏感度，这主要是由于SBS从总特征中逐渐去掉某些特征造成的(其它都是增添特征或直接优化)；Breast cancer、Musk、数据集敏感度总体高于特异度，而Artificial data特异度较高；从数据特征上看，Musk数据二类样本数量有差异，造成敏感度和特异度的计算差异较大，说明特征选择后分类模型的泛化能力不强，根据敏感度和特异度的差异准则能够较好地选出泛化能力强的特征选择算法，表6给出这一规则下最优和次优的特征选择方法的对比。

表6　6种数据集依据敏感度、特异度差异最小的最优特征选择方式

数据集	最优特征选择方式	次优特征选择方式
Breast cancer	SDP-SVM2	SDP-SVM1
Musk	SDP-SVM2	SBS
Spam	SDP-SVM1	Relif-F
Vote	Relif-F	SDP-SVM2
Artificial data	SBS	SFS

从表6可以看出，在最优特征选择方式上，SDP-SVM2方法占据5项数据的2项；次优选择上，SDP-SVM2方法在Vote数据上取得，说明所提出的SDP-SVM特征选择方法具备较强的泛化能力。唯一的例外是在Spam数据上SDP-SVM2的敏感度和特异度差距很大，泛化能力较差，在其它数据上，SDP-SVM2由于选择的特征数量更少，其在泛化能力方面优于SDP-SVM1也是意料之中的事情。

综上所述，所提出的SDP-SVM方法在选择合适的贡献度和支持度后能够在保持较高分类准确率的情况下大幅减少特征数目，具备较强的泛化能力。不论是实际数据还是人工数据都有着非常好的效果，在二类样本差异较大的数据上还有待改进，不过SDP-SVM对特征数量的灵活控制是其一大特色，这对于降低数据集规模，保持一定的准确率和泛化特性，加快计算速度有着显著的正向贡献。就本文而言，所提方法与其它几种特征选择算法相比具有较大的优势。

结束语　随着大数据时代的到来，特征选择在剖析数据中的隐含信息及服务于决策支持方面发挥着越来越重要的作用。特征选择算法很大程度上依赖于数据本身的特性，每种算法对数据的要求各不相同，都有其适用的数据类型，目前来看并没有一套较为普适的特征选择算法出现，而且现在对特征选择多样化的需求也越加广泛，因此该领域还有很大的拓展与提升空间。本文通过特征子空间组合构建核函数，将特征选择集成在半正定支持向量机分类器中，可以看作是在这一领域的初步尝试，就实证研究情况来看取得了较好的效果。鉴于集成化的学习研究的日渐丰富，该项研究仍然具有很大的拓展空间。本文仅仅涉及了单变量特征子集组合的情况，未来还可以考虑多种特征空间子集情形，如考虑每个特征子空间由几个变量构成；还可以考虑几种特征选择算法选出的子集组合，得到类似于多核学习的集成化特征选择模式，这将是未来研究的重点方向之一。

参 考 文 献

[1] Zhao Y P, Li C. Feature Selection and Patent Analysis Research in Web Security Information Mining[J]. Chinese Journal of Management Science, 2004, 12(z1): 514-518(in Chinese)
赵燕平，李超. 网络安全信息挖掘中的特征选择与专利分析研究[J]. 中国管理科学，2004，12(z1)：514-518

[2] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2002, 3(6): 1157-1182

[3] Zhang X G. Introdction to Statistical Learning Theory and Support Vector Machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42(in Chinese)
张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26 (1)：32-42

[4] Wei L W, Chen Z Y, Li J P. Evolution strategies based adaptive L-p LS-SVM[J]. Information Science, 2011, 181(14): 3000-3016

[5] Lanckriet G, Cristianini N, Bartlett P, et al. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine Learning Research, 2002, 5(1): 323-330

[6] Dietterich T G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization[J]. Machine Learning, 2000, 40(2): 139-157

[7] Mason L, Bartlett P, Baxter J. Improved generalization through explicit optimization of margins[J]. Machine Learning, 2000, 38(3): 243-255

[8] Kong E B, Dietterich T G. Error-Correcting Output Coding Corrects Bias and Variance[C]// Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, 1995: 313-321

[9] Breiman L. Bias, variance and arcing classifiers[J]. Additives for Polymers, 2002(6): 10

[10] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97(1/2): 273-324

[11] Weston J, Elisseeff A, Schulkopf B, et al. Use of the zero norm with linear models and kernel methods[J]. Journal of Machine Learning Research, 2003, 3: 1439-1461

(下转第215页)

同样，初步估计泄漏源 T' 坐标(−0.2，−4.9)和流量 $Q'=1.28$，并求出 $Pinv_{T'}=0.0317$。为避免视觉处理的物体过少，设定 $K_M=0.01$，此时阈值 $K_M \cdot Pinv_{T'}=0.0003$。确定视觉处理范围后，将各物体投影到投影圆上，视觉处理在范围内的环境物体。通过最后的累计概率(顺时针方向分别为0.0025，0.0222，0.0095，0.0020)驱使移动机器人向最大概率物体靠近并可确认泄漏点 O 即为气体泄漏源。

结束语 融合算法可有效地应用于气体泄漏源搜寻。以标准正态分布密度函数作为权值的加权平均法可提高气体测量结果的可靠性，即使个别气体传感器出现故障时，仍然可较好地估计环境气体浓度；最小二乘法能够最优估计未知参数，可应用于初步估计泄漏源的位置和流量信息；概率赋值方式可容纳多种信息途径共同判断泄漏源，更加全面的信息数据可提高搜寻效率。

参 考 文 献

[1] Shao Yun-ming, Zhu Ying, Huang De-xian, et al. Advances in study on parameter estimation of atmospheric contaminant dispersion[J]. CIESC Journal, 2011, 62(10): 2677-1681(in Chinese)
邵昀明，朱鹰，黄德先，等. 有毒气体扩散源参数估计方法综述[J]. 化工学报，2011，62(10)：2677-1681

[2] Liu Quan-yi, Su Bo-ni, Wang Sheng. Study on Fast Gas Source Identification Based on Wireless Sensor Network[J]. China Safety Science Journal, 2013, 23(1): 142-147(in Chinese)
刘全义，苏伯尼，王晟. 基于无线传感器网络的气体泄漏源快速定位方法研究[J]. 中国安全科学学报，2013，23(1)：142-147

[3] Meng Qing-hao, Li Fei. Review of Active Olfaction[J]. ROBOT, 2006, 28(1): 89-95(in Chinese)
孟庆浩，李飞. 主动嗅觉研究现状[J]. 机器人，2006，28(1)：89-95

[4] Jiang Ping, Meng Qing-hao, Zeng Ming, et al. A Novel Visual Search Method for Gas Leakage Source Based on Mobile Robot [J]. ROBOT, 2009, 31(5): 397-403(in Chinese)
蒋萍，孟庆浩，曾明，等. 一种新的移动机器人气体泄漏源视觉搜寻方法[J]. 机器人，2009，31(5)：397-403

[5] Ishida H, Ushiku T, Toyama S. Mobile robot path planning using vision and olfaction to search for a gas source[C]// IEEE Sensors. Irvine, 2005: 1112-1115

[6] Zhang Jian-hua, Zhang Xiao-jun, Sun Ling-yu, et al. Basing on the Olfaction and Vision Information Fusion for Robot's Odor Source Localization[C]// IEEE International Conference on Robotics and Biomimetics. Tianjin, China, 2010: 845-849

[7] Lu Qiang, He Yang, Wang Jian. Localization of Unknown Odor Source Based on Shannon's Entropy Using Multiple Mobile Robots[C]// Conference of the IEEE Industrial Electronics Society. Dallas, 2014: 2798-2803

[8] Ding Xin-wei, Wang Shu-lan, Xu Guo-qing. A Review of Studies on the Discharging Dispersion of Flammable and Toxic Gases [J]. Chemical Engineering, 2000, 28(1): 33-35(in Chinese)
丁信伟，王淑兰，徐国庆. 可燃及毒性气体泄漏扩散研究综述[J]. 化学工程，2000，28(1)：33-35

[9] Yu Chang, Tian Guan-san. Study of Indoor Flammable Gas Leakage Processes[J]. Journal of Shandong University of Architecture and Engineering, 2006, 21(3): 243-246(in Chinese)
于畅，田贯三. 可燃气体室内泄漏扩散的研究[J]. 山东建筑工程学院学报，2006，21(3)：243-246

[10] 童志权. 大气污染控制工程[M]. 北京：机械工业出版社，2006：395-401

[11] Zhou Ming-hui, Hu Shi-qiang, Chen Si-cong. Cylinder Unwarping and Real-Time Target Tracking Based on Omni-directional Camera[J]. Computer Engineering, 2013, 39(11): 1-4(in Chinese)
周明晖，胡士强，陈思聪. 基于全景摄像头的柱面展开及实时目标跟踪[J]. 计算机工程，2013，39(11)：1-4

(上接第 182 页)

[12] Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society1996, 58(1): 267-288

[13] Wang H Q, Sun F C, Cai Y N, et al. On Multiple Kernel Learning Methods[J]. Acta Automatica Sinica, 2010, 36(8): 1037-1050(in Chinese)
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

[14] Kittler J, Hatef M, Duin R P W, et al. On combining classifiers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239

[15] Tsymbal A, Pechenizkiy M, Cunningham P. Diversity in search strategies for ensemble feature selection[J]. Information Fusion, 2005, 6(1): 83-98

[16] Li X, Zhang T W, Guo Z, et al. An Novel Ensemble Method of Feature Gene Selection Based on Recursive Partition-tree[J]. Chinese Journal of Computers, 2004, 27(5): 675-682(in Chinese)
李霞，张田文，郭政，等. 一种基于递归分类树的集成特征基因选择方法[J]. 计算机学报，2004，27(5)：675-682

[17] Sun L, Han C Z, Shen J J, et al. Generalized Rough Set Method for Ensemble Feature Selection and Multiple Classifier Fusion [J]. Acta Automatica Sinica, 2008, 34(3): 298-304(in Chinese)
孙亮，韩崇昭，沈建京，等. 集成特征选择的广义粗集方法与多分类器融合[J]. 自动化学报，2008，34(3)：298-304

[18] Pan W B, Cheng G, Guo, X J, et al. On Embedded Feature Selection Using Selective Ensemble for Network Traffic[J]. Chinese Journal of Computers, 2014, 37(10): 2128-2138(in Chinese)
潘吴斌，程光，郭晓军，等. 基于选择性集成策略的嵌入式网络流特征选择[J]. 计算机学报，2014，37(10)：2128-2138

[19] Scholkopf B, Smola A J. Learning with Kernels[M]. MIT Press, 2002

[20] Wolberg W H, Mangasarian O L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology [J]. Proceedings of the National Academy of Sciences, 1990, 87(23): 9193-9196