一种基于变分相关向量机的特征选择和分类结合方法

徐丹蕾 [1] 杜 兰 [1] 刘宏伟 [1] 洪 灵 [1] 李彦兵 [1]

摘 要 相关向量机 (Relevance vector machine, RVM) 是一种函数形式等价于支持向量机 (Support vector machine, SVM) 的全概率模型, 利用变分贝叶斯 (Variational Bayesian, VB) 方法求解的 RVM 可以给出所有参数的后验分布. 进一步, 通过对样本所在原始特征空间的稀疏化, 基于线性核的 RVM 可以在分类的同时实现对原始特征的线性选择. 本文在传统 VB-RVM 的基础上提出一种特征选择和分类结合方法. 该方法采用 Probit 模型将分类问题与回归问题有机地结合起来, 同时, 通过对特征维的幂变换扩展, 不仅在分类时增加了样本的信息量, 可以构造非线性分类面, 而且实现了非线性特征选择的功能. 通过对仿真数据和实测数据分别进行实验, 证明了该特征选择和分类结合方法的实用性和有效性.

关键词 特征选择, 稀疏化, 相关向量机, Probit 模型, 变分贝叶斯

DOI 10.3724/SP.J.1004.2011.00932

Joint Feature Selection and Classification Design Based on Variational Relevance Vector Machine

XU Dan-Lei[1] DU Lan[1] LIU Hong-Wei[1] HONG Ling[1] LI Yan-Bing[1]

Abstract The relevance vector machine (RVM) is a fully probabilistic model with an equivalent functional form as the support vector machine (SVM), which can give posterior distributions over all parameters through the variational Bayesian (VB) method. Moreover, the RVM with linear kernel can realize not only classification but also linear feature selection by imposing sparsity in feature space where data is originally represented. In this paper, a joint feature selection and classification design is proposed based on the traditional VB-RVM. In the proposed framework, the Probit model is utilized to connect the regression problem with the classification problem, and the feature dimension extension by power transformation can make full use of the samples form the nonlinear classification boundary, and can realize nonlinear feature selection as well. The experiments based on the synthetic data and measured data demonstrate the practicability and effectiveness of the proposed method.

Key words Feature selection, sparsity, relevance vector machine (RVM), Probit model, variational Bayesian (VB)

在过去的十几年里数据收集存储能力的发展导致许多学科的信息过载[1], 工作在不同领域的研究人员每天都要面对越来越多的观测数据和仿真数据, 与过去被广泛应用的传统且信息量少的数据相比, 这些数据给数据分析带来了新的挑战. 由于观测数据数目的增加, 特别是每个观测数据中变量数目的增加, 使得传统的统计方法不再适用. 在模式识别领域, 观测数据一般指代样本, 观测数据中的变量则代表特征, 样本的维度即特征的个数. 高维数据带来挑战的同时, 也带来了机遇, 而且必然能促进一些新算法的产生. 数据维度的增加, 不仅使计算量变大, 而且有些特征对于分类是无效的, 相当于噪声, 这样就会影响分类识别的精度, 因此, 在许多实际应用中减少原始数据的维度变得很重要. 一般维度减少的方法[2] 是在对原始特征进行变换或组合的基础上产生新的特征, 这种方法被称为特征提取算法, 比如 PCA (Principal component analysis)[3]、LDA (Linear discriminant analysis)[4]、FA (Factor analysis)[5] 等; 另一种方法是特征选择, 即从原始特征中选出对分类有利的特征子集, 常见的方法有 Fisher[6]、Relief[7]、搜索法 (穷举法)[6]、基于微粒群和粗糙集的特征选择算法 (Particle swarm optimization and rough set-based feature selection, PSORSFS)[8] 等, 其中, PSORSFS 是根据微粒群优化的思想提出的, 能解决最优特征选择问题, 而且有快速收敛的能力, 在问题空间中有较强的搜索能力, 能有效找到最小约简. 但是这些现有的减少原始特征维度的方法都是与分类器设计分离的, 这样就存在提取或选择的特征与分类器可能不是最佳匹配的问题, 因此, 将特征提取

收稿日期 2010-09-16 录用日期 2011-02-01
Manuscript received September 16, 2010; accepted February 1, 2011
国家自然科学基金 (60901067, 61001212), 新世纪优秀人才支持计划 (NCET-09-0630), 长江学者和创新团队发展计划 (IRT0954), 中央高校基本科研业务费专项资金资助
Supported by National Natural Science Foundation of China (60901067, 61001212), Program for New Century Excellent Talents in University (NCET-09-0630), Program for Changjiang Scholars and Innovative Research Team in University (IRT0954), and the Fundamental Research Funds for the Central Universities

1. 西安电子科技大学雷达信号处理国家重点实验室 西安 710071
1. National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071

或选择与分类器设计结合起来成为发展的趋势.

本文提出的就是一种特征选择和分类结合的方法. 2000 年, Tipping[9−10] 首次提出了基于贝叶斯框架的相关向量机 (Relevance vector machine, RVM) 的概念, 它的函数形式与支持向量机 (Support vector machine, SVM) 一样, 识别性能也不亚于 SVM. 相对于 SVM 来说, RVM 更稀疏, 不用进行交叉验证估计惩罚参数, 而且能够得到概率式的预测; 此外, 在核函数的选择上, RVM 不受梅西定理的限制, 可以构建任意的核函数. 同年, Bishop 等[11] 在 RVM 的基础上提出了 VB-RVM 的概念, 就是用变分 (Variational Bayesian, VB) 推导对 RVM 进行公式化描述, 与之前 RVM 对超参数只进行点估计相比, VB-RVM 可以给出所有参数和超参数的后验分布. 鉴于 RVM 具有良好的稀疏性, Carin 等[12] (2003 年, 采用 VB 方法) 以及 Li 等[13] (2006 年, 采用传统 RVM 中使用的 II 型最大似然方法) 用基于原始特征空间的线性核代替基于核空间的非线性核, 使之能对原始特征空间进行稀疏, 也即实现了特征选择. 本文在传统 VB-RVM 的基础上用 Probit 模型[14−16] 代替传统 VB-RVM 分类中的 Logistic 模型, 使分类问题与回归问题有机地结合起来, 避免了 Logistic 模型从连续输出到离散输出映射时的近似推导, 使得 RVM 回归模型的推理算法可以直接运用于分类模型, 而且利用 Probit 模型可以很容易地把二元分类推广到多元分类[14−16]; 同时在对原始特征空间进行线性稀疏的基础上, 对特征维进行幂变换扩展, 这样增加了样本的信息量, 在构造非线性分类界面的基础上, 使线性的特征选择变为非线性特征选择, 能在保证良好识别率的同时得到更稳健的特征选择结果.

本文的主要安排如下: 第 1 节介绍并比较 Tipping 提出的传统 RVM 分类器和本文采用的基于 Probit 模型的分类器; 第 2 节介绍本文提出的基于 RVM 的特征选择和分类结合模型, 以及如何用 VB 实现该模型, 并且给出了算法的复杂度分析; 第 3 节中, 分别用仿真数据和实测数据对该方法进行了实验. 其中, 仿真数据实验验证了该方法在进行特征选择方面的有效性, 实测数据实验对比了本文提出的方法与其他一些方法 (比如传统 RVM 分类器、SVM 分类器[17] 等) 的分类与 ROC 曲线下的面积 (Area under curve, AUC) 性能优劣; 第 4 节对文章进行了总结.

1　相关向量机 (RVM) 模型

1.1　传统的 RVM 分类器

在监督学习中, 给定一组输入样本向量 $X = \{\boldsymbol{x}_n\}_{n=1}^N$, 其中, N 为样本的个数, 对应的目标输出是 $\boldsymbol{t} = \{t_n\}_{n=1}^N$, 对回归问题, t_n 可以是任意值, 对分类问题, t_n 是类别标号 (二元分类时可以是 0 或 1).

对于 RVM 回归问题[9−11], 模型定义为

$$t_n = y(\boldsymbol{x}_n;\boldsymbol{w}) + \varepsilon_n = \Big[\sum_{n=1}^{N} w_n K(\boldsymbol{x},\boldsymbol{x}_n) + w_0\Big] + \varepsilon_n = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n) + \varepsilon_n \tag{1}$$

其中, $K(\boldsymbol{x},\boldsymbol{x}_n)$ 是选用的核函数, $\boldsymbol{\phi}(\boldsymbol{x}_n) = [1, K(\boldsymbol{x}_n,\boldsymbol{x}_1), K(\boldsymbol{x}_n,\boldsymbol{x}_2),\cdots,K(\boldsymbol{x}_n,\boldsymbol{x}_N)]$, $\{w_n\}_{n=0}^N$ 代表不同的权重, ε_n 是噪声, 假设是服从均值为零, 方差为 τ^{-1} 的高斯分布.

文献 [9−11] 为参数 $\boldsymbol{w}$ 加上一个零均值的高斯先验分布:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{n=0}^{N} N(w_n|0,\alpha_n^{-1}) \tag{2}$$

注意这里每一个权值 w_n 都独立地对应一个参数 α_n.

为了使参数学习更灵活, 对 $\boldsymbol{\alpha}$ 和噪声方差 τ^{-1} 分别定义超先验, 合适的先验分布是伽马分布:

$$p(\boldsymbol{\alpha}) = \prod_{n=0}^{N} \mathrm{Gamma}(\alpha_n|a_n,b_n) \tag{3}$$

$$p(\tau) = \mathrm{Gamma}(\tau|c,d) \tag{4}$$

其中

$$\mathrm{Gamma}(\alpha|a,b) = \Gamma(a)^{-1} b^a \alpha^{a-1}\mathrm{e}^{-b\alpha} \tag{5}$$

这里 $\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t$ 为 gamma 函数, 通常定义超参数为很小的值, 比如令 $a_n = b_n = c = d = 10^{-6}$, 这样的超参数先验对后验学习不提供信息, 使后验完全取决于数据.

模型建立后利用 VB 方法[11, 18] 可以求出 $\boldsymbol{w}$、$\boldsymbol{\alpha}$ 和 τ 的后验分布. 在迭代求解的过程中, 大部分的 α_n 会趋于无穷大, 对应的 w_n 为零, 实现了稀疏化.

RVM 分类与回归在本质上服从一样的模型框架, 只不过改变了目标值的条件分布.

对于二元分类的情况, 文献 [9−11] 对连续隐变量 $y(\boldsymbol{x}_n;\boldsymbol{w})$ 采用 Logistic 映射函数 $\sigma(y) = 1/(1+\mathrm{e}^{-y})$, 并且假设 $p(\boldsymbol{t}|X)$ 为伯努利分布, 则似然函数如下

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{n=1}^{N} \sigma\{y(\boldsymbol{x}_n;\boldsymbol{w})\}^{t_n}[1-\sigma\{y(\boldsymbol{x}_n;\boldsymbol{w})\}]^{1-t_n} \tag{6}$$

需要注意的是这里没有考虑噪声变量 ε_n. 如果直接采用 VB 方法对上述模型进行求解很困难, 因此, 文献 [11] 参考 Jaakkola 和 Jordan 的方法, 利用下面

的不等式引进一个下界:

$$\sigma\{y(\boldsymbol{x}_n;\boldsymbol{w})\}^{t_n}[1-\sigma\{y(\boldsymbol{x}_n;\boldsymbol{w})\}]^{1-t_n}=\sigma(z_n)\geq \sigma(\xi_n)\exp\left(\frac{z_n-\xi_n}{2}-\lambda(\xi_n)(z_n^2-\xi_n^2)\right) \quad (7)$$

其中, $z_n=(2t_n-1)y(\boldsymbol{x}_n;\boldsymbol{w})$, $\lambda(\xi_n)=(1/4\xi_n)\times\tanh(\xi_n/2)$, ξ_n 是一个变分的参数, 当 $\xi_n=z_n$ 时, 式 (7) 中的等式成立. 最后再利用 VB 方法对这个下界进行求解.

1.2 Probit 模型

前面介绍的二元分类的 Logistic 模型为

$$p(t_n=1|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n))=\frac{1}{1+\exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n))} \quad (8)$$

Logistic 函数是从连续的变量到二值输出 t_n 的映射, 尽管 Logistic 映射函数很容易理解, 但它不是一个标准的概率函数, 在推理的过程中会引起许多的困难. 此外, 传统的 RVM 分类模型对似然函数引进了一个下界, 是一个近似推导, 因此, 不能准确地估计出模型的真实值.

针对上面的问题, 我们提出用 Probit 模型[14−16] 实现从连续量到离散量的映射, 如下:

$$p(t_n,y_n|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n),\tau)=p(t_n|y_n)N(y_n;\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n),\tau^{-1}) \quad (9)$$

其中, $\{y_n\}_{n=1}^N$ 是隐藏在 t_n 后面的连续随机变量, 基于模型的需要, 这里目标值 $\{t_n\}_{n=1}^N$ 假定为 1 或 −1, 概率关系如下:

$$p(t_n|y_n)=I(t_n=\mathrm{sign}(y_n))=\begin{cases}1, & t_n=\mathrm{sign}(y_n)\\ 0, & \text{否则}\end{cases} \quad (10)$$

其中, $I(\cdot)$ 是一个指示函数. 通过对 y_n 进行积分, 发现:

$$\begin{aligned}p(t_n=1|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n),\tau)&=\int p(t_n=1,y_n|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n),\tau)\mathrm{d}y_n=\\ &\int_0^\infty N(y_n;\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n),\tau^{-1})\mathrm{d}y_n=\\ &\mathrm{normcdf}\left(\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n)}{\tau^{-\frac{1}{2}}}\right)\end{aligned} \quad (11)$$

其中, normcdf 为正态累积分布函数. 图 1 给出了式 (8) 和式 (11) 分别表示的 Logistic 模型和 Probit 模型的比较, 可以看出 Probit 模型很好地近似了 Logistic 模型.

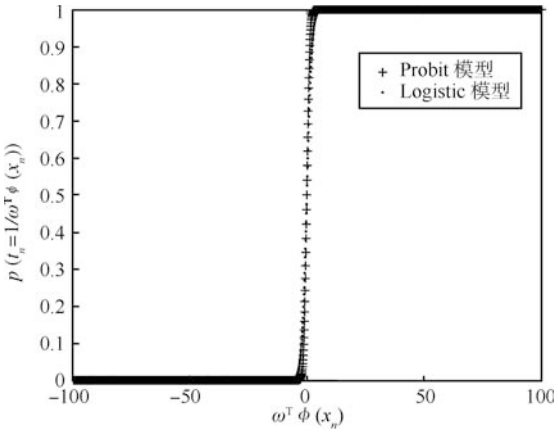

图 1 Probit 模型与 Logistic 模型的比较 ($\tau=1$)

Fig. 1 Comparison of Probit model and logistic model ($\tau=1$)

Probit 模型的优点就是通过引入隐变量把二值输出的分类问题转变为回归问题, 使分类和回归的模型完全等价; 而使用 Logistic 模型必须忽略噪声变量. 因此, 基于 Probit 模型可以灵活地把回归模型的推理算法直接运用到分类模型. 此外, 利用 Probit 模型也可以很容易地把二元分类情况推广到多元分类情况[14−16], 类似于二元的情况, 相对于多元 Logistic 模型[19], 多元 Probit 模型也可以避免复杂的近似运算, 具有更简单、实用的特点, 近些年被广泛采用[14−16].

2 基于 RVM 的特征选择和分类结合模型

2.1 线性特征选择

RVM 的结果是高度稀疏的, 意味着许多样本对应的权值趋于零, 需要强调是式 (1) 描述的模型是在核空间进行稀疏, 而不是在原始特征空间, 为了利用 RVM 进行特征选择, 文献 [12−13] 采用线性核 RVM, 线性核 RVM 的模型输出定义为

$$y(\boldsymbol{x}_n;\boldsymbol{w})=\boldsymbol{x}_n\boldsymbol{w}=\sum_{p=1}^{P}x_{pn}w_p \quad (12)$$

其中, $\boldsymbol{x}_n=[x_{1n},x_{2n},\cdots,x_{Pn}]$, $\boldsymbol{w}=[w_1,w_2,\cdots,w_P]^{\mathrm{T}}$ 为其对应的权值, P 是样本的特征维数. 利用第 1 节介绍的 RVM 模型可以很容易地求出 $\boldsymbol{w}$ 的稀疏解, 不为零的 w_p 对应的特征即为要选择的特征, 这样就实现了线性特征选择和分类的结合.

2.2 对特征维进行幂变换扩展

文献 [12−13] 提出的特征选择方式为线性特征选择, 它比较简单, 无法保证在得到高识别率的同时进行有效的特征选择, 因此, 我们考虑用幂变换的方

式对特征维进行扩展, 这样可以增加样本的信息量, 使线性特征选择变为非线性特征选择, 具体扩展方式如下:

第 n 个样本 $\boldsymbol{x}_n = [x_{1n}, x_{2n}, \cdots, x_{Pn}]$ 扩展为 $\boldsymbol{x}_n = [\tilde{\boldsymbol{x}}_{1n}, \cdots, \tilde{\boldsymbol{x}}_{mn}, \cdots, \tilde{\boldsymbol{x}}_{Mn}]$, 其中, $\tilde{\boldsymbol{x}}_{mn} = [{x_{1n}}^m, {x_{2n}}^m, \cdots, {x_{Pn}}^m]$, $m = 1, 2, \cdots, M$, M 为扩展的次数.

由前面描述的模型知, 特征扩展 M 次后, 相应的 $\boldsymbol{w}$ 和 $\boldsymbol{\alpha}$ 也要扩展 M 次. 需要注意的是: 对于 M 次的扩展, $\{w_{pm}\}_{m=1}^{M}$ 可以变化, 即同一特征的不同幂次对分类面的构造所起的作用可以不同, 但是我们希望同一特征的不同幂次对应相同的 α_p, 这样, 对于给定的每一个特征, 它在各种幂次下被选择的先验由相同的 α_p 决定, 但对应的构造分类面的具体系数 $\{w_{pm}\}_{m=1}^{M}$ 在不同幂次间有差别, 从而既实现了特征选择, 又可以构造非线性分类面.

$\boldsymbol{w}$ 和 $\boldsymbol{\alpha}$ 的扩展用数学模型可以表示如下:

$$\boldsymbol{w} = [\tilde{\boldsymbol{w}}_1, \cdots, \tilde{\boldsymbol{w}}_m, \cdots, \tilde{\boldsymbol{w}}_M]^{\mathrm{T}} \tag{13}$$

其中, $\tilde{\boldsymbol{w}}_m = [w_{1m}, w_{2m}, \cdots, w_{Pm}]$.

$$\begin{aligned}\boldsymbol{\alpha} = [&\alpha_1, \cdots, \alpha_P, \cdots, \alpha_1, \cdots, \\ &\alpha_P, \cdots, \alpha_1, \cdots, \alpha_P]^{\mathrm{T}}_{(M\times P)\times 1}\end{aligned} \tag{14}$$

在实际应用中, 根据 $\{\alpha_p\}_{p=1}^{P}$ 的大小 (把 $\{\alpha_p\}_{p=1}^{P}$ 按照从小到大顺序排列为 $\{B_i\}_{i=1}^{P}$, 如果第一次出现 B_i/B_{i+1} 小于一个很小的值 e, 比如 $e = 10^{-2}$, 则在 B_i 和 B_{i+1} 之间选择一个值作为门限), α_p 小的特征将对应非零的 $\{w_{pm}\}_{m=1}^{M}$, 而 α_p 很大的特征对应的 $\{w_{pm}\}_{m=1}^{M}$ 将自动趋于零, 实现特征的稀疏化选择.

这里需要说明的是 $M = 1$ 即等价于文献 [12] 采用的线性特征选择和分类结合方法, 只是回归的连续输出到分类的离散输出间的映射方式不同, 但这对分类器性能影响不大. 因此, 在后面的实验中会用 $M = 1$ 代表文献 [12] 的方法和我们提出的幂次变换后的方法进行性能比较. 另外, 我们要强调, $M \neq 1$ 时, 本文提出的方法与直接在文献 [12] 的基础上对特征维进行幂变换扩展的方法是不同的, 本文的方法限定了同一特征的不同幂次对应相同的 α_p, 而直接幂次扩展时, 同一特征的不同幂次对应不同的 α_{pm}, 这会导致某个特征在某个幂次对应较小的 α_{pm}, 而在另一幂次却对应较大的 α_{pm} (如后面仿真实验图 4 所示), 显然, 这种直接幂次扩展法不具有特征选择的功能, 而只是构造了非线性分类面, 对于给定的一组训练样本集其性能可能较好, 但很难保证对测试样本稳健地推广性能. 在实验部分, 我们将进一步比较本文提出的方法和这种直接特征扩展法的分类和 AUC 性能.

2.3　基于变分贝叶斯 (VB) 的模型实现

经过前面的介绍, 对于本文提出的模型我们已经有了一个完整的概率说明. 下面用一个有向图表示这个概率模型, 如图 2. 和文献 [11] 中的图 2 比较, 我们可以更清楚地看出本文提出的模型和传统 RVM 模型的区别: 通过采用 Probit 模型引入隐变量 $\{y_n\}_{n=1}^{N}$, 把二值输出的分类问题转变为回归问题; 之后对特征维进行了幂变换扩展, 相应的 $\boldsymbol{w}$ 和 $\boldsymbol{\alpha}$ 也得到了扩展, 但是对于 M 次的扩展, $\{w_{pm}\}_{m=1}^{M}$ 可以变化, 同一特征的不同幂次对应的 α_p 要保持不变.

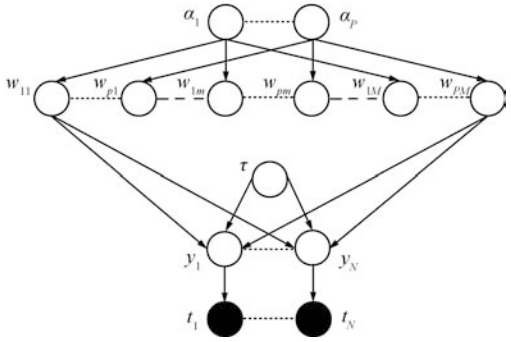

图 2　有向无环图表示基于变分相关向量机的特征选择和分类结合模型

Fig. 2　Directed acyclic graph representing the joint feature selection and classification model based on the VB-RVM

2.3.1　变分贝叶斯 (VB) 方法

VB-RVM 是利用变分贝叶斯的方法推导 RVM 模型参数和超参数的后验分布. 下面基于 RVM 模型介绍 VB 方法[9, 18] 的基本思想.

模型中的观测变量为 $\{\boldsymbol{x}_n\}_{n=1}^{N}$ 和 $\{t_n\}_{n=1}^{N}$, 隐变量为 $\{y_n\}_{n=1}^{N}$, 参数为 $\{w_{pm}, \alpha_p, \tau\}_{p=1}^{P}{}_{m=1}^{M}$, 因此, 对数边缘似然函数 $\ln p(t)$ 可写为

$$\begin{aligned}&\ln p(\boldsymbol{t}) = \ln \iiiint p(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau = \\ &\ln \iiiint q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)\frac{p(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)}{q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau \geq \\ &\iiiint q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)\ln\frac{p(\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)}{q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau\end{aligned} \tag{15}$$

其中, $q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau)$ 为隐变量和参数间的联合概率分布函数. VB 算法假定 $\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau$ 之间是相互独立的, 因此, 4 个变量的联合概率分布可以近似写为

$$q(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\alpha}, \tau) \approx q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau) \tag{16}$$

基于式 (16) 的假设, 式 (15) 可以改写如下:

$$\ln p(\boldsymbol{t}) \geq \iiiint q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau) \ln\frac{p(\boldsymbol{t},\boldsymbol{y},\boldsymbol{w},\boldsymbol{\alpha},\tau)}{q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau)}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau = F(q(\boldsymbol{y}),q(\boldsymbol{w}),q(\boldsymbol{\alpha}),q(\tau)) \tag{17}$$

可以看出对数边缘似然函数具有一个下界, 可以通过最大化下界 $F(q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau))$ 来逼近真实值.

2.3.2 基于 RVM 的非线性特征选择和分类结合模型更新

由上一节知, 可以通过对对数边缘似然的下界使用 EM (Expectation-maximization) 算法求解, 得出隐变量和所有参数的后验分布. 下界的表达形式[11, 18] 如下:

$$\begin{aligned} F(q(\boldsymbol{y}),q(\boldsymbol{w}),q(\boldsymbol{\alpha}),q(\tau)) = & \iiiint q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau)\ln\frac{p(\boldsymbol{t},\boldsymbol{y},\boldsymbol{w},\boldsymbol{\alpha},\tau)}{q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau)} \\ & \mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau = \\ & \iiiint q(\boldsymbol{y})q(\boldsymbol{w})q(\boldsymbol{\alpha})q(\tau)\{\ln p(\boldsymbol{t}|\boldsymbol{y}) + \\ & \ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1}) + \ln p(\boldsymbol{w}|0,\boldsymbol{\alpha}^{-1}) + \\ & \ln p(\boldsymbol{\alpha}|\boldsymbol{a},\boldsymbol{b}) + \ln p(\tau|c,d) - \ln q(\boldsymbol{y}) - \\ & \ln q(\boldsymbol{w}) - \ln q(\boldsymbol{\alpha}) - \ln q(\tau)\}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\tau \end{aligned} \tag{18}$$

因为设定的变量分布 $q(\boldsymbol{y})$、$q(\boldsymbol{w})$、$q(\boldsymbol{\alpha})$、$q(\tau)$ 均为共轭先验分布, 因此, 它们与其后验分布具有相同的分布形式, 下面为它们的概率分布形式:

$$q(\boldsymbol{y}) \propto \prod_{n=1}^{N} p(t_n|y_n)N(y_n;\tilde{m}_n,\tilde{\sigma}_n) = \prod_{n=1}^{N} N^{t_n}(y_n;\tilde{m}_n,\tilde{\sigma}_n) \tag{19}$$

$$q(\boldsymbol{w}) = N(\boldsymbol{w};\tilde{\boldsymbol{\mu}}_w,\tilde{\Sigma}_w) \tag{20}$$

$$q(\tau) = \mathrm{Gamma}(\tau;\tilde{c},\tilde{d}) \tag{21}$$

$$q(\boldsymbol{\alpha}) = \Big\{\prod_{p=1}^{P}\mathrm{Gamma}(\alpha_p;\tilde{a}_p,\tilde{b}_p)\Big\}^{M} \tag{22}$$

其中, $N^{t_n}(\cdot)$ 表示截断正态分布[20], 截断的方向 ($N^{+}(\cdot)$ 或 $N^{-}(\cdot)$) 由 t_n 决定.

由 EM 算法的理论知识可知, 求变量的后验分布其实就是求完全似然函数的对数关于其他变量的期望 (用 "$\langle\cdot\rangle$" 表示), 再提取与该变量有关的项. 下边分别对 $\boldsymbol{y}$, $\boldsymbol{w}$, τ, $\boldsymbol{\alpha}$ 的后验分布进行求解.

1) $\boldsymbol{y}$

提取式 (18) 中与变量 $\boldsymbol{y}$ 有关的项:

$$\tilde{F}(q(\boldsymbol{y})) = \int q(\boldsymbol{y})\Big\{\ln p(\boldsymbol{t}|\boldsymbol{y}) + \iint q(\boldsymbol{w})q(\tau)\ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1})\mathrm{d}\boldsymbol{w}\mathrm{d}\tau - \ln q(\boldsymbol{y})\Big\}\mathrm{d}\boldsymbol{y} \tag{23}$$

令式 (23) 关于 $q(\boldsymbol{y})$ 的导数为 0, 因此

$$\ln p(\boldsymbol{t}|\boldsymbol{y}) + \iint q(\boldsymbol{w})q(\tau)\ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1})\mathrm{d}\boldsymbol{w}\mathrm{d}\tau - \ln q(\boldsymbol{y}) - 1 = 0 \tag{24}$$

求解式 (24), 得出:

$$\tilde{m}_n = \boldsymbol{x}_n\langle\boldsymbol{w}\rangle, \qquad \tilde{\sigma}_n = \langle\tau\rangle^{-1} \tag{25}$$

2) $\boldsymbol{w}$

提取式 (18) 中与变量 $\boldsymbol{w}$ 有关的项:

$$\tilde{F}(q(\boldsymbol{w})) = \int q(\boldsymbol{w})\Big\{\iint q(\boldsymbol{y})q(\tau)\ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1})\mathrm{d}\boldsymbol{y}\mathrm{d}\tau + \int q(\boldsymbol{\alpha})\ln p(\boldsymbol{w}|0,\boldsymbol{\alpha}^{-1})\mathrm{d}\boldsymbol{\alpha} - \ln q(\boldsymbol{w})\Big\}\mathrm{d}\boldsymbol{w} \tag{26}$$

令式 (26) 关于 $q(\boldsymbol{w})$ 的导数为 0, 因此

$$\tilde{\Sigma}_w = (\mathrm{diag}(\langle\boldsymbol{\alpha}\rangle) + \langle\tau\rangle\sum_{n=1}^{N}\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_n)^{-1} \tag{27}$$

$$\tilde{\boldsymbol{\mu}}_w = \langle\tau\rangle\tilde{\Sigma}_w\sum_{n=1}^{N}\boldsymbol{x}_n^{\mathrm{T}}\langle y_n\rangle \tag{28}$$

3) τ

提取式 (18) 中与变量 τ 有关的项:

$$\tilde{F}(q(\tau)) = \int q(\tau)\Big\{\iint q(\boldsymbol{y})q(\boldsymbol{w})\ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1})\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{w} + \ln p(\tau|c,d) - \ln q(\tau)\Big\}\mathrm{d}\tau \tag{29}$$

令式 (29) 关于 $q(\tau)$ 的导数为 0, 因此

$$\tilde{c} = c + \frac{1}{2}, \qquad \tilde{d} = d + \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(\langle {y_n}^2\rangle - 2\langle y_n\rangle\langle\boldsymbol{w}^{\mathrm{T}}\rangle{\boldsymbol{x}_n}^{\mathrm{T}} + \boldsymbol{x}_n\langle\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle{\boldsymbol{x}_n}^{\mathrm{T}}) \tag{30}$$

4) $\boldsymbol{\alpha}$

提取式 (18) 中与变量 $\boldsymbol{\alpha}$ 有关的项:

$$\tilde{\mathrm{F}}(q(\boldsymbol{\alpha}))=\int q(\boldsymbol{\alpha})\Big\{\int q(\boldsymbol{w})\ln p(\boldsymbol{w}|0,\boldsymbol{\alpha}^{-1})\mathrm{d}\boldsymbol{w}+\ln p(\boldsymbol{\alpha}|\boldsymbol{a},\boldsymbol{b})-\ln q(\boldsymbol{\alpha})\Big\}\mathrm{d}\boldsymbol{\alpha} \tag{31}$$

令式 (31) 关于 $q(\boldsymbol{\alpha})$ 的导数为 0, 因此

$$\tilde{a}_p=a_p+\frac{1}{2},\qquad \tilde{b}_p=b_p+\frac{1}{M}\sum_{m=1}^{M}\frac{1}{2}\langle {w_{pm}}^2\rangle \tag{32}$$

上边求出了每个变量后验分布函数中的参数迭代公式, 但是它们都是由其他变量的期望表示的, 因此, 需要求出一些参数的期望[11, 21−22].

对于式 (19), t_n ($t_n\in\{-1,+1\}$) 决定截断正态分布的截断方向, 因此:

$$\langle y_n\rangle=\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w+t_n\frac{1}{\langle\tau\rangle^{\frac{1}{2}}}\frac{\mathrm{normpdf}\left(\frac{(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)}{\langle\tau\rangle^{-\frac{1}{2}}}\right)}{\mathrm{normpdf}\left(\frac{t_n(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)}{\langle\tau\rangle^{-\frac{1}{2}}}\right)} \tag{33}$$

$$\langle {y_n}^2\rangle=\frac{1}{\langle\tau\rangle}+(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)^2+t_n(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)\langle\tau\rangle^{-\frac{1}{2}}\frac{\mathrm{normpdf}\left(\frac{(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)}{\frac{1}{\langle\tau\rangle}^{\frac{1}{2}}}\right)}{\mathrm{normpdf}\left(\frac{t_n(\boldsymbol{x}_n\tilde{\boldsymbol{\mu}}_w)}{\langle\tau\rangle^{-\frac{1}{2}}}\right)} \tag{34}$$

其中, normpdf 为正态概率密度函数, normcdf 为正态累积分布函数.

另外

$$\langle\boldsymbol{w}\rangle=\tilde{\boldsymbol{\mu}}_w,\quad \langle\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\rangle=\tilde{\Sigma}_w+\tilde{\boldsymbol{\mu}}_w\tilde{\boldsymbol{\mu}}_w^{\mathrm{T}} \tag{35}$$

$$\langle\alpha_p\rangle=\frac{\tilde{a_p}}{\tilde{b_p}},\qquad \langle\ln\alpha_p\rangle=\Psi(\tilde{a_p})-\ln\tilde{b_p} \tag{36}$$

$$\langle\tau\rangle=\frac{\tilde{c}}{\tilde{d}},\qquad \langle\ln\tau\rangle=\Psi(\tilde{c})-\ln\tilde{d} \tag{37}$$

其中, Ψ 函数定义如下:

$$\Psi(a)=\frac{\mathrm{d}\ln\Gamma(a)}{\mathrm{d}a} \tag{38}$$

模型训练好之后, 对于一个测试样本 $\boldsymbol{x}_*$, 它的预测概率可以计算如下:

$$p(t_*=1|\boldsymbol{x}_*)=\int p(t_*=1|y_*)p(y_*|\boldsymbol{x}_*,\langle\boldsymbol{\omega}\rangle,\langle\tau\rangle)\mathrm{d}y_*=\int_0^{\infty}N(y_*;\boldsymbol{x}_*\langle\boldsymbol{\omega}\rangle,\langle\tau\rangle^{-1})\mathrm{d}y_*=\mathrm{normcdf}\left(\frac{\boldsymbol{x}_*\tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}}{\langle\tau\rangle^{-\frac{1}{2}}}\right) \tag{39}$$

根据预测概率就可以对测试样本进行分类识别, 当 $p(t_*=1|\boldsymbol{x}_*)\geq 0.5$ 时, 测试样本被判为一类, 当 $p(t_*=1|\boldsymbol{x}_*)<0.5$ 时, 则被判为另外一类.

另外, 通过式 (18) 可以计算对数似然函数的下界:

$$\begin{aligned}F(q(\boldsymbol{y}),q(\boldsymbol{w}),q(\boldsymbol{\alpha}),q(\tau))=&\\ &\langle\ln p(\boldsymbol{t}|\boldsymbol{y})\rangle+\langle\ln p(\boldsymbol{y}|X\boldsymbol{w},\tau^{-1})\rangle+\\ &\langle\ln p(\boldsymbol{w}|0,\boldsymbol{\alpha}^{-1})\rangle+\langle\ln p(\boldsymbol{\alpha}|\boldsymbol{a},\boldsymbol{b})\rangle-\langle\ln q(\boldsymbol{\alpha})\rangle+\\ &\langle\ln p(\tau|c,d)\rangle-\langle\ln q(\tau)\rangle-\langle\ln q(\boldsymbol{y})\rangle-\langle\ln q(\boldsymbol{w})\rangle\end{aligned} \tag{40}$$

对应于每一次的迭代, $F(q(\boldsymbol{y}),q(\boldsymbol{w}),q(\boldsymbol{\alpha}),q(\tau))$ 会一直增加, 直至收敛. 通过观察 $\langle\boldsymbol{w}\rangle=\tilde{\boldsymbol{\mu}}_w$ 或 $\langle\alpha_p\rangle=\tilde{a}_p/\tilde{b}_p$ 的值, 可以确定哪些特征会被选择, 即实现了非线性特征选择. 但是在对特征维进行扩展后, 对应于每一次的扩展, $\tilde{\boldsymbol{\mu}}_w$ 的值不一样, 那么用 $\tilde{\boldsymbol{\mu}}_w$ 来描述特征选择性的随机性比较大, 因此, 我们可以用 $\boldsymbol{\alpha}$ 来描述, 当 α_p 比较小时 (在实际应用中我们可以设定一定的门限), 对应的特征就会被选择.

2.4　算法复杂度分析

根据前面对模型更新过程的介绍, 发现本文提出的方法对各参数更新的计算复杂度主要取决于计算权值 $\boldsymbol{w}$ 的方差矩阵, 其求逆操作的复杂度为 $\mathrm{O}((M\times P)^3)$, 其中, P 为特征的维度, M 为特征幂变换扩展的次数; 传统的 RVM 也需要类似的求逆运算, 其复杂度为 $\mathrm{O}((N+1)^3)$[10], N 为训练样本的个数. 因此, 两种方法的复杂度差别主要取决于具体数据对应的 $N+1$ 与 $M\times P$ 的大小.

3　实验结果与分析

3.1　仿真数据

这一节是对仿真数据进行实验, 我们的目的是检测本文提出的非线性特征选择方法能否选出代表一个给定输出的所有特征. 仿真数据的产生方

法[13, 23] 如下:

$$\boldsymbol{y} = \boldsymbol{x}_1 + 2\boldsymbol{x}_2 + 3\boldsymbol{x}_3 + 2\sin(\boldsymbol{x}_4) + \mathrm{e}^{\boldsymbol{x}_5} + 0\boldsymbol{x}_6 + \cdots + 0\boldsymbol{x}_{12} \tag{41}$$

$$\boldsymbol{t} = \mathrm{sign}(\boldsymbol{y}) \tag{42}$$

这里的特征 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_5$ 是分别服从标准正态分布且独立同分布的随机变量, 第 6 个变量与 $\boldsymbol{x}_1$ 有关, 定义为 $\boldsymbol{x}_6 = \boldsymbol{x}_1 + 1$, 第 7 个变量为 $\boldsymbol{x}_7 = \boldsymbol{x}_2\boldsymbol{x}_3$, 与 $\boldsymbol{x}_2$ 和 $\boldsymbol{x}_3$ 都有关系, 另外的 5 个特征 $\boldsymbol{x}_8 \sim \boldsymbol{x}_{12}$ 也服从标准正态分布, 但是与 $\boldsymbol{y}$ 没有任何关系. 如式 (41), 我们随机产生 200 个样本.

下面用本文介绍的方法对仿真数据进行实验, 如图 3, 可以看出当 $M = 1$ (特征不扩展, 即文献 [12] 采用的方法) 时, 比较小的 5 个 $\log\alpha_p$ 对应的特征为 $\boldsymbol{x}_3, \boldsymbol{x}_2, \boldsymbol{x}_1, \boldsymbol{x}_6, \boldsymbol{x}_5$, 表明特征 $\boldsymbol{x}_1, \boldsymbol{x}_6$ 都被选择出来了, 而想要的特征 $\boldsymbol{x}_4$ 却没被选择出来, 因此, 没有达到理想的特征选择效果, 然而在对特征维进行幂变换扩展后, 比较小的 5 个 $\log\alpha_p$ 对应的特征均为 $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, $\boldsymbol{x}_5$, $\boldsymbol{x}_1$ 或 $\boldsymbol{x}_6$, 表明所有想要的特征都被选出来了. 证明了基于 VB-RVM 的非线性特征选择方法的有效性.

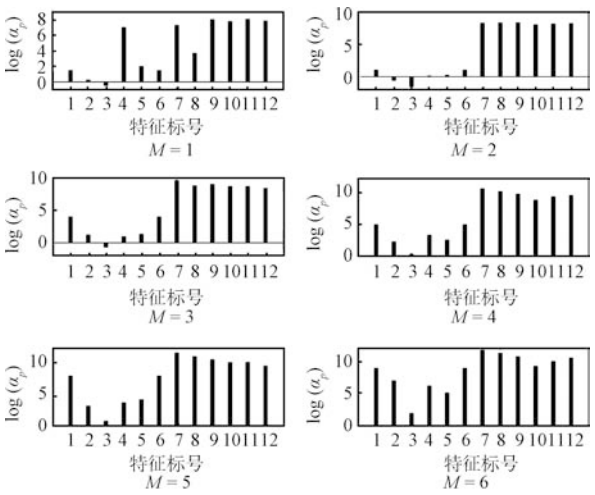

图 3 分别给出当 M 从 1 到 6 时, 通过 VB-RVM 得到的 $\log\alpha_p$ 以及它们对应的特征 ($M = 1$ 等价于文献 [12] 采用的方法)

Fig. 3 $\log\alpha_p$ and corresponding selected features through VB-RVM respectively for $M = 1, 2, \cdots, 6$ ($M = 1$ equals the method which is adopted by [12])

接下来用直接对特征维进行幂变换扩展 (不同幂次对应不同的 α_{pm}) 的方法对仿真数据进行实验, 设定 $M = 3$, 图 4 给出了迭代更新之后 $\log\alpha_{pm}$ 的值, 从图 4 中可以看出, 不同的幂次对应的比较小的 $\log\alpha_{pm}$ 的特征是不一样, 比如特征 1 在 $m = 1$ 时比较小, 但在 $m = 2$ 和 $m = 3$ 时却很大, 因此, 直接进行幂变换扩展的方法不能确定哪些特征被选出了, 即不具备特征选择的功能.

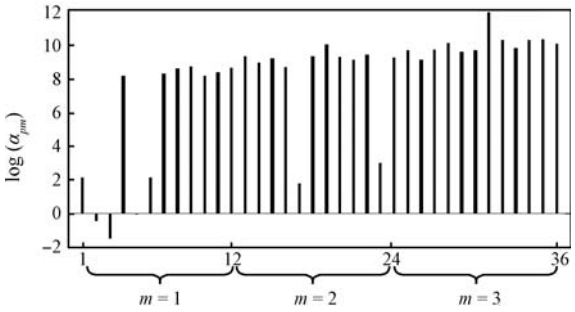

图 4 直接进行幂变换扩展 (不同幂次对应不同的 α_{pm}) 方法得到的 $\log\alpha_{pm}$

Fig. 4 $\log\alpha_{pm}$ through direct power transformation extension (a different power corresponds to a different α_{pm})

3.2 实测数据

本节我们用雷达实测数据测试该特征选择和分类结合方法的性能. 数据包括两类目标: 民航和汽车, 样本数分别为 553 和 546, 特征维数为 21, 提取的特征主要包括雷达目标时域回波的方差、多普勒谱的熵、幂变换后的熵、目标峰值的主瓣宽度以及主副瓣比等. 为了证明该方法的有效性和稳健性, 把民航和汽车样本分别均分为 5 组, 取其中的一组固定作为测试样本, 剩余 4 组中的任意一组作为训练样本, 另外 3 组分别作为训练阶段的 3 次交叉验证样本, 用来选择稳健的参数, 如果同时选出多种参数, 在测试阶段用这几种参数得到识别结果的平均值作为最终的识别结果. 这样, 根据训练样本的选取不同, 总共可以进行 4 次实验. 设定 α_p 的门限为 100 (根据第 2.2 节中介绍的准则选取门限为 100. 当门限大于 100 时, 识别率不会有太大改变, 因为多选取出来的特征对应的 w_{pm} 很小, 对分类的贡献很小; 当门限远小于 100 时, 识别率会明显下降, 因为必要的特征没有被选择), 当 α_p 小于 100 时, 认为对应的特征被选出来.

为了证明本文提出方法的有效性, 下面用它和其他一些常见的特征选择和分类方法进行比较, 这些方法包括基于所有特征的传统 RVM 分类器[9−10]、基于所有特征的 SVM 分类器[17]、Fisher 特征选择[6] 结合传统 RVM 分类器[9−10]、Relief 特征选择[7] 结合传统 RVM 分类器[9−10]、特征搜索法[6] (穷举法) 结合 SVM 分类器[17]、直接对特征维进行幂变换扩展 (不同幂次对应不同的 α_{pm}) 的方法、文献 [12] 提出的方法, 以及 PSORSFS[8] 特征选择结合 SVM 分类器[17].

3.2.1 特征选择和分类性能比较

为了与本文提出的方法进行准确的比较, 这里 Relief、Fisher、搜索法 (穷举法) 选择的特征个数与本文提出方法选择出的特征个数设为一致. 表 1 为这几种方法的比较结果, 实验结果分别为选择特征的个数、选择特征的标号、错分率. 表 1 中从 (a) 到 (d) 分别为 4 次实验的结果.

表 1 中粗体字分别表示 4 次实验中不同方法性能最好的结果, 各种方法选出的特征标号的排列顺序是由它们对分类的贡献决定的 (在实验 (a)、(c)、(d) 中特征搜索法结合 SVM 没有给出选择的特征标号, 是因为通过交叉验证选出了多个错分率相同的特征组合, 这相当于没有选出稳健的特征组合), 通过表 1 可以看出: 1) 采用 Relief 或 Fisher 特征选择与传统 RVM 分类器结合得到的错分率在所有的方法中是比较高的 (除了实验 (b) 中 Fisher 特征选择结合 RVM 分类器与本文提出方法的错分率一样); 2) 当特征全部选用时, RVM 的错分率要略高于 SVM 的错分率, 另外, 除了实验 (a) 中 SVM 的错分率略低于本文提出的方法, 其他实验中基于全部特征的 SVM 和 RVM 的错分率都要高于本文提出的方法, 这也证明了特征选择对提高识别性能的重要性; 3) 在相同特征个数的情况下, 采用特征搜索法结合 SVM 的错分率要高于本文提出方法的错分率, 而且当搜索法 (穷举法) 选择的特征个数为 3、4 时分别需要 1 330、5 985 次搜索实验, 其运算量远远大于本文提出方法的运算量; 4) 文献 [12] 采用的方法在 4 次实验中的错分率都要高于本文提出方法的错分率, 而且它选择的特征个数要比本文提出方法选出的特征个数多, 且选出的特征在 4 次实验中也不稳定, 这说明了本文提出的通过特征维幂变换扩展构造分类面的方法对提高分类性能和稀疏化特征选择均是有效的; 5) 直接对特征维进行幂变换扩展 (不同的幂次对应不同的 α_{pm}) 方法的错分率一般都要高于本文提出方法的错分率 (实验 (c) 中两种方法的错分率一样), 这是由于直接扩展法只能构造非线性分类面但没有特征选择的功能, 这直接影响了该方法对测试样本的推广性能; 6) PSORSFS 特征选择方法在识别率方面通常优于 Relief 或 Fisher 特征选择方法 (除了实验 (b) 中 Fisher 特征选择方法的错分率更低), 但和本文提出的方法相比, PSORSFS 特征选择方法除了在实验 (a) 中的错分率比较低外, 在其他三个实验中的错分率都比本文提出方法的错分率高, 而且在 4 次实验中它选出的特征都不一样, 说明该方法在特征选择方面还不太稳定.

表 1 只给出了不同特征选择和分类方法在 4 次实验中的单次实验结果, 为了更充分地比较各种方法的性能优劣, 表 2 给出了 4 次实验错分率平均值和标准差, 同样地, 粗体字表示不同方法中性能最好的结果.

通过表 2 可以看到: 1) 本文提出方法在所有方法中的平均错分率最低, 表明该方法的分类性能较好; 2) 本文提出方法的错分率的标准差在所有方法中是最小的, 说明鲁棒性较好; 另外, 观察表 1 中本文方法选择的特征, 可以看到选择的特征在 4 次实验中也很稳定, 这表明该方法对于特征选择具有很好的稳健性.

3.2.2 AUC 性能比较

除了识别率, ROC 曲线下的面积 (AUC) 也是判断分类器性能优劣的一个重要指标[24−25]. AUC 评价标准可以衡量数据类别在任何分布或任何代价下分类算法的总体性能, 对类别分布比例和错误代价具有不敏感性, 因此, 在类别分布未知时和错误代价敏感时, AUC 成为分类器性能的有效评价方法[25]. AUC 的值在 0 ∼ 1 之间, 越趋近于 1 说明性能越好. 下面以本文提出方法的第 2 个实验 (实验 (b)) 为例详细说明如何实现 AUC 性能评估.

图 5 给出了所有测试样本 ("·" 表示样本的真实类别是民航, "+" 表示样本的真实类别是汽车) 的预测概率值 ($p(t_* = 1|\boldsymbol{x}_*)$), 按照前面的讨论我们将根据预测概率值的大小对它们进行判决, 大于等于 0.5 的点被判为民航目标, 小于 0.5 的点被判为汽车目标. 根据图 5 的结果, 假设民航样本为检测目标、汽车样本为虚警目标, 在计算 ROC 曲线时, 首先, 把所有测试样本的预测概率值按照从大到小的顺序排列; 其次, 在 0 ∼ 1 之间设定若干门限, 对每个给定的门限, 小于门限的认为是虚警样本, 大于门限的认为是检测样本, 可以得到相应的虚警概率和检测概率, 它们的定义如下:

$$\text{虚警概率} = \frac{\text{虚警样本被错分为检测目标的数目}}{\text{虚警样本总数}} \tag{43}$$

$$\text{检测概率} = \frac{\text{检测样本被正确分为检测目标的数目}}{\text{检测样本总数}} \tag{44}$$

最后, 所有虚警和检测概率对应起来就可以绘制出图 6 中用 "·" 标记的 ROC 曲线. 类似地, 我们也可以用 1 减去图 5 的结果, 这样预测概率大于等于 0.5 的就被判为汽车目标, 并假设汽车样本为检测目标、民航样本为虚警目标, 用相同的方法得到图 6 中 "+" 对应的 ROC 曲线. 显然, 由于对称关系, 图 6 中两条 ROC 曲线下的面积值, 即 AUC, 应该是完全相同的, 在实验中只需要计算一种情况.

表 1 (a) ∼ (d) 分别给出 4 次本文提出方法与其他方法比较的实验结果

Table 1 (a) to (d) respectively show the four experimental results for comparison between the method we proposed and other methods

	特征选择方法结合分类器	选择特征的个数 (选择特征的标号)	错分率 (%)	AUC
	本文提出方法	3 (2, 9, 6)	1.92	0.9975
	直接扩展特征 (不同幂次对应不同的 α_{pm})	\	2.73	**0.9979**
	$M = 1$ (文献 [12] 中方法)	4 (6, 2, 3, 9)	2.73	0.9908
	基于所有特征的传统 RVM	21 (1 ∼ 21)	4.55	0.9884
(a)	基于所有特征的 SVM	21 (1 ∼ 21)	1.82	0.9920
	Fisher 特征选择结合传统 RVM	3 (15, 3, 2)	5.45	0.9778
	Relief 特征选择结合传统 RVM	3 (9, 2, 15)	6.21	0.9760
	特征搜索法结合 SVM	3 (\)	2.27	0.9955
	PSORSFS 特征选择结合 SVM	3 (2, 12, 21)	**1.36**	0.9974
	本文提出方法	4 (2, 9, 6, 8)	**1.82**	**0.9979**
	直接扩展特征 (不同幂次对应不同的 α_{pm})	\	2.27	0.9888
	$M = 1$ (文献 [12] 中方法)	5 (2, 6, 9, 10, 21)	2.27	0.9937
	基于所有特征的传统 RVM	21 (1 ∼ 21)	2.50	0.9912
(b)	基于所有特征的 SVM	21 (1 ∼ 21)	3.18	0.9924
	Fisher 特征选择结合传统 RVM	4 (15, 2, 16, 6)	**1.82**	0.9973
	Relief 特征选择结合传统 RVM	4 (9, 2, 15, 11)	4.55	0.9805
	特征搜索法结合 SVM	4 (2, 9, 10, 20)	2.27	0.9964
	PSORSFS 特征选择结合 SVM	3 (9, 15, 20)	2.73	0.9924
	本文提出方法	3 (2, 9, 6)	**1.82**	0.9974
	直接扩展特征 (不同幂次对应不同的 α_{pm})	\	**1.82**	**0.9987**
	$M = 1$ (文献 [12] 中方法)	4 (2, 6, 9, 10)	3.18	0.9907
	基于所有特征的传统 RVM	21 (1 ∼ 21)	2.96	0.9907
(c)	基于所有特征的 SVM	21 (1 ∼ 21)	4.32	0.9893
	Fisher 特征选择结合传统 RVM	3 (15, 2, 3)	5.45	0.9801
	Relief 特征选择结合传统 RVM	3 (9, 2, 15)	5.45	0.9757
	特征搜索法结合 SVM	3 (\)	2.44	0.9958
	PSORSFS 特征选择结合 SVM	2 (2, 20)	3.18	0.9979
	本文提出方法	4 (2, 9, 8, 6)	**1.82**	**0.9975**
	直接扩展特征 (不同幂次对应不同的 α_{pm})	\	2.73	0.9935
	$M = 1$ (文献 [12] 中方法)	5 (6, 2, 3, 18, 9)	4.09	0.9931
	基于所有特征的传统 RVM	21 (1 ∼ 21)	3.79	0.9884
(d)	基于所有特征的 SVM	21 (1 ∼ 21)	2.05	0.9900
	Fisher 特征选择结合传统 RVM	4 (15, 3, 2, 16)	5.91	0.9811
	Relief 特征选择结合传统 RVM	4 (9, 2, 15, 11)	5.18	0.9769
	特征搜索法结合 SVM	4 (\)	2.73	0.9929
	PSORSFS 特征选择结合 SVM	2 (15, 20)	4.09	0.9871

表 2　4 次实验中错分率和 AUC 的平均值和标准差

Table 2　The means and standard deviations of error rate and AUC in the four experiments

特征选择方法结合分类器	平均错分率 (%)	错分率标准差	AUC 平均值	AUC 标准差
本文提出方法	**1.845**	**0.0433**	**0.9976**	**1.9203e−004**
直接扩展特征 (不同幂次对应不同的 α_{pm})	2.388	0.3776	0.9947	0.0040
$M=1$ (文献 [12] 中方法)	3.068	0.6723	0.9921	0.0013
基于所有特征的传统 RVM	3.450	0.7855	0.9897	0.0013
基于所有特征的 SVM	2.843	0.9963	0.9909	0.0013
Fisher 特征选择结合传统 RVM	4.658	1.6490	0.9841	0.0077
Relief 特征选择结合传统 RVM	5.348	0.5955	0.9773	0.0019
特征搜索法结合 SVM	2.428	0.1879	0.9952	0.0013
PSORSFS 特征选择结合 SVM	2.840	0.9850	0.9937	0.0044

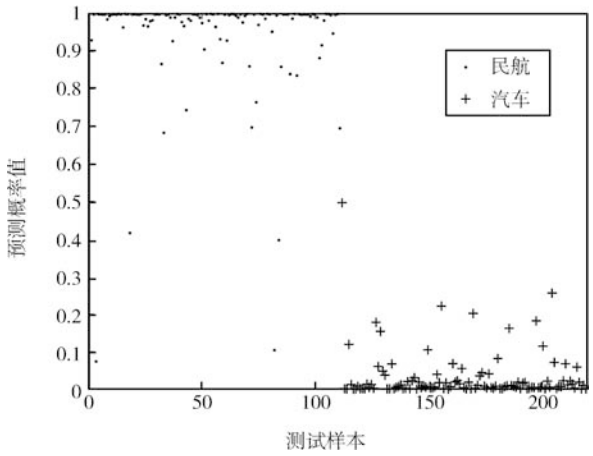

图 5　第 2 次实验中所有测试样本的预测概率值

Fig. 5　The predictive probability of all the test samples in the second experiment

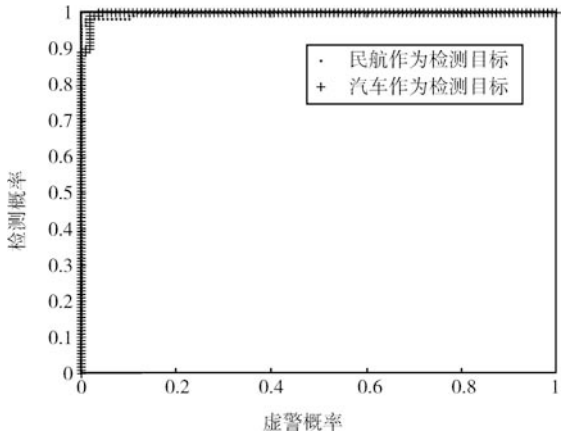

图 6　第 2 次实验的 ROC 曲线

Fig. 6　The ROC curve of the second experiment

为了比较前面提到的各种方法的 AUC 性能, 在表 1 的最后一列中加入了各种方法的 AUC 值, 在表 2 的后两列也加入了 4 次实验 AUC 的平均值和标准差, 从表 1 中可以看出, 在实验 (a) 和 (c) 中直接特征扩展方法的 AUC 略大于本文提出的方法的 AUC, 但是表 2 中本文提出方法的 AUC 平均值最大并且标准差最小, 表明该方法的 AUC 性能很好并且稳定.

3.2.3　复杂度比较

根据第 2.4 节的讨论, 本文提出方法单次迭代的复杂度取决于幂次参数和特征维度, 经过交叉验证第一次实验选出了 3 个错分率相同的幂次参数 M, 分别为 2、3、5, 这里以 5 作为第一次实验的幂次参数, 第 2、3、4 次实验选出的幂次参数均为 3, 而总特征个数 P 为 21, 因此, 本文提出方法 4 次实验的复杂度分别为 $O(105^3)$、$O(36^3)$、$O(36^3)$、$O(36^3)$; 对于传统 RVM 算法单次迭代的复杂度主要取决于训练样本的个数, 4 次实验训练样本个数 N 分别为 221、220、219、219, 因此, 它们复杂度为分别为 $O(222^3)$、$O(221^3)$、$O(220^3)$、$O(220^3)$. 显然, 对 4 次实验, 传统 RVM 算法的复杂度都远大于本文提出方法的复杂度.

通过上面的分析可以证明本文提出的特征选择和分类结合方法是具备实用性和有效性的.

4　结论

本文提出的基于 VB-RVM 的非线性特征选择和分类结合方法是一种把特征选择和分类识别有效地结合在一起的方法, 该方法采用 RVM 的基本模型框架, 用 Probit 模型代替原 RVM 分类中的 Logistic 模型, 使分类与回归有机地结合起来, 避免了 Logistic 模型从连续输出到离散输出映射时的近似

推导, 使得 RVM 回归模型的推理算法可以直接运用于分类模型, 而且利用 Probit 模型可以很容易地把二元分类推广到多元分类; 求解的过程中用 VB 的方法对模型进行公式化描述, 这样能够得到所有参数和超参数的后验分布, 更重要的是对原始特征维进行了幂变换扩展, 而且限定扩展后不同幂次要对应相同的 α_p, 这样增加了样本的信息量, 在构造非线性分类面的同时实现了非线性特征选择. 通过对仿真数据和实测数据分别进行实验, 对比了本文提出的方法与其他一些常见的特征选择和分类方法, 证明了本文提出方法的有效性和实用性. 本文只是对特征维进行简单的幂变换扩展, 在下一步的工作中将会对扩展的方式进行更深入的研究, 使特征选择的方式更多样、更有效.

致谢

本文关于文献 [8] 的对比实验使用的是该文第一作者上海大学王向阳博士提供的程序代码, 并且所有参数设置均与文献 [8] 一致. 在此, 感谢王博士对本文工作的大力支持.

References

1 Fodor I K. A Survey of Dimension Reduction Techniques, Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, USA, 2002

2 Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(2): 153−158

3 Jackson J E. *A User's Guide to Principal Component*. New York: John Wiley and Sons, 1991

4 Yu H, Yang J. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 2001, **34**(10): 2067−2070

5 Mardia K V, Kent J T, Bibby J M. *Multivariate Analysis*. London: Academic Press, 1980

6 Duda R O, Hart P E, Stork D G. *Pattern Classification (Second Edition)*. New York: John Wiley and Sons, 1997. 94−99

7 Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence. California, USA: AAAI, 1992. 129−134

8 Wang X Y, Yang J, Teng X L, Xia W J, Jensen R. Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 2007, **28**(4): 459−471

9 Tipping M E. The relevance vector machine. *Advances in Neural Information Processing Systems 12*. Cambridge: The MIT Press, 2000. 652−658

10 Tipping M E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 2001, **1**: 211−244

11 Bishop C M, Tipping M E. Variational relevance vector machines. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 2000. 46−53

12 Carin L, Dobeck G J. Relevance vector machine feature selection and classification for underwater targets. In: Proceedings of the OCEANS. San Diego, USA: IEEE, 2003. 1110−1110

13 Li D F, Hu W C. Feature selection with RVM and its application to prediction modeling. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag, 2006. 1140−1144

14 Girolami M, Rogers S. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 2006, **18**(8): 1790−1817

15 Zhou X, Wang X, Dougherty E R. Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. *IEE Proceedings Systems Biology*, 2006, **153**(2): 70−78

16 Damoulas T, Girolami M. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, 2008, **24**(10): 1264−1270

17 Burges C J C. A tutorial on support vector machine for pattern recognition. *Data Mining and Knowledge Discovery*, 1998, **2**(2): 121−167

18 Hou Qing-Yu. Study of Radar Automatic Target Recognition Methods Based on High Resolution Profile [Ph. D. dissertation], Xidian University, China, 2009
(侯庆禹. 基于高分辨距离像的雷达自动目标识别方法研究 [博士学位论文], 西安电子科技大学, 中国, 2009)

19 Krishnapuram B, Carin L, Figueiredo M A T, Hartemink A J. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(6): 957−968

20 Zhang Hai-Juan, Zhang Xiao-Ran, Wen Yan-Qing, Guo Ming-Ming. Using Fisher information matrix to deal with parameter estimation for truncated samples from normal distribution. *Journal of Chongqing Technology Business University (Natural Science Edition)*, 2007, **24**(3): 228−229
(张海娟, 张晓冉, 温艳清, 郭明明. 用 Fisher 信息阵处理截断正态分布的参数估计. 重庆工商大学学报 (自然科学版), 2007, **24**(3): 228−229)

21 Beal M J. Variational Algorithms for Approximate Bayesian Inference [Ph. D. dissertation], London University, UK, 2003

22 Nielsen F B. Variational Approach to Factor Analysis and Related Models [Master dissertation], Technical University of Denmark, Denmark, 2004

23 Bi J B, Bennett K P, Embrechts M, Breneman C, Song M H. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 2003, **3**: 1229−1243

24 Xue Y, Liao X J, Carin L, Krishnapuram B. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 2007, **8**: 35−63

25 Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997, **30**(7): 1145−1159

徐丹蕾 西安电子科技大学雷达信号处理国家重点实验室博士研究生. 2009 年获得西安电子科技大学电子工程学院学士学位. 主要研究方向为雷达目标识别.
E-mail: xdlei5258@163.com
(**XU Dan-Lei** Ph. D. candidate at the National Key Laboratory of Radar Signal Processing, Xidian University. She received her bachelor degree from Xidian University in 2009. Her main research interest is radar target recognition.)

杜 兰 西安电子科技大学电子工程学院教授. 2007 年获得西安电子科技大学信息与通信工程博士学位, 主要研究方向为统计信号处理、雷达信号处理、机器学习及其在雷达目标检测与识别方面的应用. 本文通信作者.
E-mail: dulan@mail.xidian.edu.cn
(**DU Lan** Professor at the College of Electrical Engineering, Xidian University. She received her Ph. D. degree in information and communication engineering from Xidian University in 2007. Her research interest covers statistical signal processing, radar signal processing, and machine learning with application to radar target recognition. Corresponding author of this paper.)

刘宏伟 西安电子科技大学电子工程学院教授. 研究方向为雷达信号处理、MIMO 雷达、雷达目标识别、自适应信号处理、认知雷达.
E-mail: hwliu@.xidian.edu.cn
(**LIU Hong-Wei** Professor at the College of Electrical Engineering, Xidian University. His research interest covers radar signal processing, MIMO radar, radar target recognition, adaptive signal processing, and cognitive radar.)

洪 灵 西安电子科技大学雷达信号处理国家重点实验室博士研究生. 2008 年获得西安电子科技大学电子工程学院学士学位. 主要研究方向为雷达目标识别.
E-mail: linghong@mail.xidian.edu.cn
(**HONG Ling** Ph. D. candidate at the National Key Laboratory of Radar Signal Processing, Xidian University. She received her bachelor degree from Xidian University in 2008. Her main research interest is radar target recognition.)

李彦兵 西安电子科技大学雷达信号处理国家重点实验室博士研究生. 2009 年获得西安电子科技大学电子工程学院硕士学位. 主要研究方向为雷达目标识别和雷达信号处理理论.
E-mail: xidianlyb@163.com
(**LI Yan-Bing** Ph. D. candidate at the National Key Laboratory of Radar Signal Processing, Xidian University. He received his master degree from Xidian University in 2009. His research interest covers radar target recognition and radar signal processing.)