第 43 卷 第 4 期 计 算 机 工 程 2017 年 4 月
Vol. 43 No. 4 Computer Engineering April 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)04-0183-05 文献标志码：A 中图分类号：TP18

基于稀疏贝叶斯模型的特征选择

祝 璞[a,b,c]，黄章进[a,b,c]

（中国科学技术大学 a. 计算机科学与技术学院；
b. 安徽省计算与通信软件重点实验室；c. 先进技术研究院，合肥 230027）

摘 要： 通过采用稀疏贝叶斯推理方法，设计出可同时进行学习最优分类器与选取最优特征子集的特征选择概率分类向量机算法。该算法是对概率分类向量机特征选择的扩展，可提高其在高维数据集上的性能。通过选取零均值的高斯分布作为先验，在模型中起到正则项的作用，同时在核函数和特征中引入稀疏，得到泛化性更好的分类模型。在高维度和低维度数据集中的实验结果表明，该算法同时具有较好的分类和特征选择能力。

关键词： 机器学习；核函数；稀疏贝叶斯；特征选择；概率分类向量机；自动相关性检测

中文引用格式： 祝 璞，黄章进. 基于稀疏贝叶斯模型的特征选择[J]. 计算机工程，2017，43(4)：183-187，193.

英文引用格式： Zhu Pu，Huang Zhangjin. Feature Selection Based on Sparse Bayesian Model[J]. Computer Engineering，2017，43(4)：183-187，193.

Feature Selection Based on Sparse Bayesian Model

ZHU Pu[a,b,c]，HUANG Zhangjin[a,b,c]

（a. School of Computer Science and Technology；b. Anhui Province Key Laboratory of Computing and Communication Software；
c. Institute of Advanced Technology，University of Science and Technology of China，Hefei 230027，China）

【Abstract】 Through using sparse Bayesian inference thought，a Feature Selection Probabilistic Classification Vector Machine（FPCVM）is designed which can learn optimal classifier and automatically select the most relevant feature subset. FPCVM is an extension of Probabilistic Classification Vector Machine（PCVM），which improves the performance of PCVM on high dimension datasets. It uses zero-mean Gaussian distribution as priori to introduce sparseness both in kernel functions and feature space；these priors are preformed as regularization items in the likelihood function to acquire more generalized model. Experimental results on high dimension datasets and low dimension datasets show that the algorithm has better classification and feature selection.

【Key words】 machine learning；kernel function；sparse Bayesian；feature selection；Probabilistic Classification Vector Machine（PCVM）；automatic relevance determination

DOI： 10.3969/j.issn.1000-3428.2017.04.031

0 概述

在二类监督学习中，给出样本数据集特征 $x \in \boldsymbol{R}^M$，目标是训练出可以区分样本类别（不失一般性记为 $y=-1$，$y=+1$）的分类器。通常为了完成上述目标，需要给出一个训练集 $D=\{x^{(i)},y^{(i)}\}_{i=1}^{N}$，其中，$x^{(i)} \in \mathbf{R}^M$ 是样本的特征；$y^{(i)} \in \{-1,+1\}$ 是样本对应的类别。有了给定训练集后，训练任务分为以下2个：1）如何确定分类函数 $f(\cdot)$，即选择分类模型；2）如何选取特征子集，即特征选择。分类器的性能将会通过其泛化性能来评价：在面对未知数据集时，分类器能否可以做出正确的判断。

本文提出特征选择概率分类向量机（Feature Selection Probabilistic Classification Vector Machine，FPCVM）模型，介绍通过 EM 算法获取模型的参数，并通过 FPCVM 在不同数据集中的实验与现有算法的对比验证其性能。

1 研究背景

近年来，基于核方法的模型成为了机器学习的主流模型，如文献[1-2]，该方法可以看作是核函数 $\boldsymbol{\Phi}_\theta(x)$ 的线性组合：

基金项目： 安徽省自然科学基金（1408085MKL06）；高等学校学科创新引智计划项目（B07033）。
作者简介： 祝 璞（1991—），男，硕士研究生，主研方向为机器学习；黄章进（通信作者），副教授。
收稿日期： 2016-03-09 **修回日期：** 2016-05-23 **E-mail：** zhupu@mail.ustc.edu.cn

$$f(x;\boldsymbol{w}) = \sum_{i=1}^{N} w_i \phi_{i,\theta}(x) + b = \boldsymbol{\Phi}_\theta(x)\boldsymbol{w} + b \quad (1)$$

其中，$\boldsymbol{w} = (w_1, w_2, \cdots, w_N)^{\mathrm{T}}$ 作为模型的权值；$\boldsymbol{\Phi}_\theta(x) = (\phi_{1,\theta}(x), \phi_{2,\theta}(x), \cdots, \phi_{N,\theta}(x)) \in \mathbf{R}^{M\times M}$ 作为模型的核函数矩阵，其中，$\phi_{i,\theta}(x) = (k(x^{(1)}, x^{(i)}), (k(x^{(2)}, x^{(i)}), \cdots, k(x^{(N)}, x^{(i)}), \theta \in \mathbf{R}^M$ 是模型的核函数。对于高斯核，有：

$$k(x^{(i)}, x^{(j)}) = \exp\left(-\sum_{k=1}^{M} \theta_k (x_k^{(i)} - x_k^{(j)})^2\right) \quad (2)$$

从式(2)中可以得出，如果 $\theta_k = 0$ 意味着在模型中除去了 k 维特征，从而达到了特征选择的目的。

目前学术界主流的基于核函数的学习方法，包括支持向量机(Support Vector Machine，SVM)[3]、相关向量机(Relevant Vector Machine，RVM)[2]和概率分类向量机(Probabilistic Classification Vector Machines，PCVM)[1]等。这些方法都是通过用核函数的映射得到了最优的模型权值 w。在上述 3 种方法中，SVM 采用了最大化分类间隔的方法，训练得到分类器，但是没有能够给出概率的输出。文献[2]经过研究 SVM 这一缺陷后，通过稀疏贝叶斯推导，在模型中引入了零均值的高斯先验，得到了可同时获得稀疏模型和概率输出的分类方法，即 RVM。随后文献[4]研究发现，对于不同的类别，引入相同的零均值高斯先验，会造成模型的不稳定，并在随后的研究中指出对于不同的类别 y 应当使用不同的截断高斯分布作为先验，从而得出了更稳定的贝叶斯分类器，即 PCVM。Suykens 提出了最小二乘支持向量机(LS-svm)，在克服 SVM 不足(计算复杂)的同时丢失了 SVM 的稀疏，为此文献[5]提出了基于混合核函数稀疏 LS-SVM，贝叶斯模型应用十分广泛，文献[6]深入研究了基于关系选择的多关系朴素分类器。

上述 3 种算法都有优秀的分类性能，然而 3 种方法只优化了模型的权值参数 w，忽略了优化核函数的参数 θ。在高维数据中往往存在着大量冗余和与分类不相关的维度，由于这些维度的存在，使得上述 3 个算法很难在这些数据集中获取良好的性能，因此在算法中加入特征选择十分必要。当今学术界一些主流的特征选择算法，例如基于 T 检验特征选择(T-test)[4]、最大相关最小冗余(max-Dependency min-Redundancy，mRMR)[7]和主成分分析(Principal Component Analysis，PCA)[8]等都是与训练过程分离的，这样使最优特征子集大小的确定变得十分困难。基于这些不足，本文采用贝叶斯推理在核参数 θ 中引入截断高斯先验，并将其融入 PCVM 中得到可同时进行分类训练和相关特征检测的分类算法——特征选择概率分类向量机。

2 特征选择概率分类向量机

2.1 分类模型

在二项分类中，输入的数据是 $\{x_i, y_i\}_{i=1}^N$，其中，$y_i \in \{-1, +1\}$。模型需要使用一个连接函数，使输出的连续值平滑地映射到二值变量 $\{-1, +1\}$ 中，本文采用的连接函数如下：

$$\Psi(x) = \int_{-\infty}^{x} N(t|0,1)\,\mathrm{d}t$$

其中，$\Psi(x)$ 是高斯积分函数。采用此模型，可以在下一节的 EM 算法中快速地获得隐含变量的模型[9]。经过此连接函数和式(1)的组合后，FPCVM 的最终分类模型为：

$$\begin{aligned} l(x;\boldsymbol{w},b) &= \Psi\left(\sum_{i=1}^{N} w_i \phi_{i,\theta}(x) + b\right) \\ &= \Psi(\boldsymbol{\Phi}_\theta(x)\boldsymbol{w} + b) \end{aligned} \quad (3)$$

2.2 截断高斯先验

FPCVM 为了获得稀疏模型，为每个权重 w_i 分别引入一个截断高斯先验 α_i 作为超参数。获得最优特征子集，为每个 θ_i 引入一个截断高斯先验 τ_i 作为超参数，同时在 b 中引入零均值的高斯分布，如下：

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{\alpha}) &= \prod_{i=1}^{N} p(w_i|\alpha_i) = \prod_{i=1}^{N} N_t(w_i|0,\alpha_i^{-1}) \\ p(b|\beta) &= N(b|0,\beta^{-1}) \\ p(\boldsymbol{\theta}|\boldsymbol{\tau}) &= \prod_{i=1}^{M} p(\theta_i|\tau_i) = \prod_{i=1}^{M} N_t(\tau_i|0,\tau_i^{-1}) \end{aligned} \quad (4)$$

其中，$N_t(w_i|0,\alpha_i^{-1})$ 和 $N_t(\tau_i|0,\tau_i^{-1})$ 是截断高斯分布。当 $y_i = +1$ 时，先验是一个非负的左截断高斯分布；当 $y_i = -1$ 时，先验是一个非正的右截断高斯分布，而对于核参数 θ 始终是用非负的左截断高斯分布，如图 1 所示。

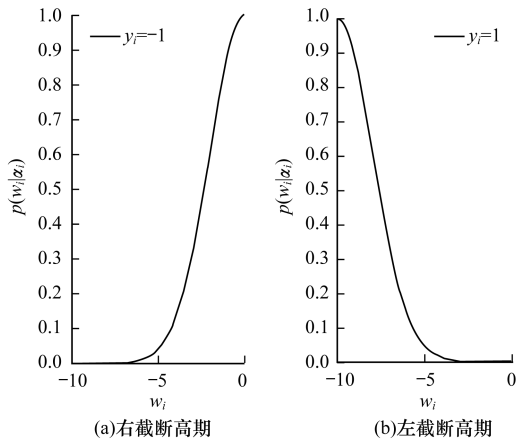

图 1 权值的截断高斯先验

当 $y = -1$ 时，使用的是图 1(a)所示的右截断高斯；当 $y = +1$ 时，使用的是图 1(b)所示的左截断高斯。对于核参数 θ 使用图 1(b)所示的左截断高斯。

这样可以得到如下的先验分布：

$$p(w_i|\alpha_i)=\begin{cases}2N(w_i|0,\alpha_i^{-1}),y_iw_i\geqslant 0\\0,y_iw_i<0\end{cases}$$

$$p(\theta_i|\tau_i)=\begin{cases}2N(\theta_i|0,\tau_i^{-1}),\theta_i\geqslant 0\\0,\theta_i<0\end{cases}\tag{5}$$

本节给出了模型的先验,根据贝叶斯推理求解模型参数应采用最大化后验(Maximum a Posterior, MAP),即$(\boldsymbol{w},b,\boldsymbol{\theta})=\arg\max_{(\boldsymbol{w},b,\theta)}p(\boldsymbol{w},b,\boldsymbol{\theta}|y,x,\boldsymbol{\alpha},\beta,\boldsymbol{\tau})$。由于模型中存在大量未知的变量,本文将在下一节阐述如何使用期望最大化(Expectation-Maximization,EM)算法求解包含隐变量的后验。

2.3　最大化后验

这一节将给出 EM 算法求解最大化后验的详细推导过程。文献[10]已经证明,在数据集不完整或者模型的先验或似然函数中存在隐变量时,EM 算法是可以有效的求解模型的最大化后验。EM 算法通过在计算期望(E-step)和最大化(M-step)两部分中进行迭代计算从而完成最大化后验。在实际应用中,E 和 M 部分应当根据实际问题进行相应的调整,接下来将会介绍 FPCVM 如何通过 EM 算法求解模型的参数。

在贝叶斯求解过程中,通常假设计算模型 $\boldsymbol{\Phi}_\theta(x)\boldsymbol{w}+b$ 被一个噪音变量 ϵ 干扰,其中,$\epsilon\sim N(0,1)$。将其加入之前引入的概率连接函数式(3)可得 $h_\theta(x)=\boldsymbol{\Phi}_\theta(x)\boldsymbol{w}+b+\epsilon\geqslant 0, l=1, h_\theta(x)=\boldsymbol{\Phi}_\theta(x)\boldsymbol{w}+b+\epsilon\leqslant 0, l=-1$。这样,得到如下概率模型:

$$\begin{aligned}p(l=1|x,\boldsymbol{w},b,\boldsymbol{\theta})&=p(\boldsymbol{\Phi}_\theta(x)\boldsymbol{w}+b+\epsilon\geqslant 0)\\&=\Psi(\boldsymbol{\Phi}_\theta(x)\boldsymbol{w}+b)\end{aligned}\tag{6}$$

由于噪声 ϵ 是不可观测的,可得到隐变量 $h_\theta(x)$。由核函数矩阵 $\boldsymbol{\Phi}_\theta=(\boldsymbol{\Phi}_\theta(x^{(1)})^{\mathrm{T}},(\boldsymbol{\Phi}_\theta(x^{(2)})^{\mathrm{T}},\cdots,\boldsymbol{\Phi}_\theta(x^{(N)}))^{\mathrm{T}}$,其中,$\boldsymbol{\Phi}_\theta(x^{(i)})=(\phi_\theta(x^{(1)},x(i)),\phi_\theta(x^{(2)},x(i)),\cdots,\phi_\theta(x^{(N)},x(i)))$和隐变量向量 $\boldsymbol{H}_\theta(x)=(h_\theta(x^{(1)}),\cdots,h_\theta(x^{(N)}))^{\mathrm{T}}$,这样可以得到如下模型:

$$\begin{aligned}&p(\boldsymbol{H}_\theta|\boldsymbol{w},b,\boldsymbol{\theta})\\&=(2\pi)^{N/2}\exp\left\{-\frac{1}{2}\|\boldsymbol{H}_\theta-(\boldsymbol{\Phi}_\theta\boldsymbol{w}+b\boldsymbol{I})\|^2\right\}\end{aligned}$$

其中,$\boldsymbol{I}=(1,1,\cdots,1)^{\mathrm{T}}$ 是一个 N 维的全 1 向量。

将 α,β 和 τ 当做隐含变量,可以得到模型后验的对数公式:

$$\begin{aligned}&\ln p(\boldsymbol{w},b,\boldsymbol{\theta}|y,\boldsymbol{H}_\theta,\alpha,\beta,\tau)\\&\propto\ln p(\boldsymbol{H}_\theta|w,b,\boldsymbol{\theta})+\ln p(\boldsymbol{w}|\alpha)\\&+\ln p(b|\beta)+\ln p(\boldsymbol{\theta}|\boldsymbol{\tau})\\&\propto w^{\mathrm{T}}\boldsymbol{\Phi}_\theta^{\mathrm{T}}(2\boldsymbol{H}_\theta-\boldsymbol{\Phi}_\theta\boldsymbol{w})+2b\boldsymbol{I}^{\mathrm{T}}\boldsymbol{H}_\theta\\&-2b\boldsymbol{I}^{\mathrm{T}}\boldsymbol{\Phi}_\theta\boldsymbol{w}-b^2N-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w}-\beta b^2-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\theta}\end{aligned}\tag{7}$$

其中,$\boldsymbol{A}=\mathrm{diag}(\alpha_1,\alpha_2,\cdots,\alpha_N)$;$\boldsymbol{\Delta}=\mathrm{diag}(\tau_1,\tau_2,\cdots,\tau_M)$。

期望步:推导出模型的对数似然函数后,标记为 Q,计算隐变量期望的公式。

$$\begin{aligned}&Q(\boldsymbol{w},b,\boldsymbol{\theta}|\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}})\\&=E_{\boldsymbol{H}_\theta,\alpha,\beta,\tau}[\ln p(w,b,\theta|y,\boldsymbol{H}_\theta,\alpha,\beta,\tau)|y,\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}}]\\&=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}_\theta^{\mathrm{T}}(2\boldsymbol{H}_\theta-\boldsymbol{\Phi}_\theta\boldsymbol{w})+2b\boldsymbol{I}^{\mathrm{T}}\boldsymbol{H}_\theta\\&-2b\boldsymbol{I}^{\mathrm{T}}\boldsymbol{\Phi}_\theta\boldsymbol{w}-b^2N-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w}-\beta b^2-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{\theta}\end{aligned}\tag{8}$$

其中,$\boldsymbol{H}_\theta=E[H_\theta|y_i,\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}}]$;$\boldsymbol{A}=\mathrm{diag}(E[\alpha_i|y_i,\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}}])$;$\boldsymbol{\beta}=E[\beta|y_i,\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}}]$;$\boldsymbol{\Delta}=\mathrm{diag}(E[\tau_i|y_i,\boldsymbol{w}^{\mathrm{old}},b^{\mathrm{old}},\boldsymbol{\theta}^{\mathrm{old}}])$。

假设超参数 α,β,τ 都服从伽玛分布,可以通过共轭先验得到超参数的分布,具体推导过程可以参考文献[1]中的附录 B。

最大化步:这一部分将通过更新 $\boldsymbol{w},b$ 和 $\boldsymbol{\theta}$,求解公式 Q 的最大值。

式(8)对参数 $\boldsymbol{w},b$ 和 $\boldsymbol{\theta}$ 的偏微分如下:

$$\frac{\partial Q}{\partial w}=-2\boldsymbol{\Phi}_\theta^{\mathrm{T}}\boldsymbol{\Phi}_\theta w+\boldsymbol{\Phi}_\theta^{\mathrm{T}}\boldsymbol{H}_\theta-2\boldsymbol{A}w\tag{9}$$

$$\frac{\partial Q}{\partial b}=2\boldsymbol{I}^{\mathrm{T}}\boldsymbol{H}_\theta-2bN-2\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Phi}_\theta-b\boldsymbol{\beta}\tag{10}$$

$$\frac{\partial Q}{\partial\theta_k}=-2\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{(\boldsymbol{\Phi}_\theta\boldsymbol{w}-\boldsymbol{H}_\theta)\boldsymbol{w}^{\mathrm{T}}\odot\frac{\partial\boldsymbol{\Phi}_\theta}{\partial\theta_k}\right\}_{(i,j)}-2\theta_k\tau_k\tag{11}$$

其中,$\odot$表示哈达姆矩阵乘子,即 2 个矩阵中相同位置的元素进行乘法操作。

在通常情况下,很难通过解析的方法来同时获得 Q 在 $\boldsymbol{w},b$ 和 θ_k下的最大值,但是可以通过迭代的方法,首先令$\frac{\partial Q}{\partial w}=0$ 和$\frac{\partial Q}{\partial b}=0$,得到如下公式:

$$\boldsymbol{w}^{\mathrm{new}}=(\boldsymbol{\Phi}_\theta^{\mathrm{T}}\boldsymbol{\Phi}_\theta+\boldsymbol{A})^{-1}(\boldsymbol{\Phi}_\theta^{\mathrm{T}}\boldsymbol{H}_\theta-b\boldsymbol{\Phi}_\theta^{\mathrm{T}}\boldsymbol{I})\tag{12}$$

$$b^{\mathrm{new}}=\frac{\boldsymbol{I}^{\mathrm{T}}\boldsymbol{H}_\theta-\boldsymbol{I}^{\mathrm{T}}\boldsymbol{\Phi}_\theta\boldsymbol{w}}{\beta+N}\tag{13}$$

将 $\boldsymbol{w}^{\mathrm{new}}$和 b^{new}加入 Q 中,然后使用现有的凸优化方法(如梯度下降、共轭下降或牛顿法)获得 Q 关于 $\boldsymbol{\theta}$ 的最大值。本文采用的是共轭下降法来优化 $\boldsymbol{\theta}$。

FPCVM 计算过程伪代码如下:

算法 1　FPCVM

输入　训练集 $D=(x_n,y_n)_{n=1}^N$,其中,$x_n\in\mathbf{R}^M$,$y_n\in\{-1,1\}$;*maxIts* 是最大迭代次数;*threshold* 是收敛阈值;θ 是模型核参数;*initParameter* 是模型初始化值

输出　输出模型的权值 $\boldsymbol{w}$、间隔 b 和核参数 $\boldsymbol{\theta}$

```
1.  [w,b,θ] = initialize(initParameter);
2.  Vector = determine_usefull_vector(w);
3.  for i = 1 to maxItes do
4.      Φ = Kernel(X,Y,θ);
5.      [w^new b^new] = model_parameter_update{Φ,w,b,
Y,vector};
6.      θ^new = kernel_parameter_update(X,Y,θ,w^new,
b^new,vector);
7.      vector^new = determine_usefull_vector(w^new);
8.      if max(abs(w^new - w)) < threshold then
```

9. break;
10. end
11. end

上述伪代码包括如下 5 个过程：

1)初始化模型参数 $\boldsymbol{w}$,b 和 $\boldsymbol{\theta}$ 并创建一个 N 维指示向量 ***vector***,如果 $w_i \neq 0$,则 $vertor_i = 1$(第 1 行～第 2 行)；

2)计算核函数矩阵(第 4 行)；

3)根据式(12)、式(13)更新模型参数($\boldsymbol{w}$,b)(第 5 行)；

4)根据式(11)更新核函数参数 $\boldsymbol{\theta}$(第 6 行)；

5)更新相关向量(第 7 行)；

6)比较新权重 $\boldsymbol{w}^{new}$ 和旧权重 $\boldsymbol{w}^{old}$,来确定函数是否收敛。如果两者绝对值小于阈值,则退出循环返回 $\boldsymbol{w}$,b,$\boldsymbol{\theta}$;否则程序继续(第 7 行～第 10 行)。

3 实验与结果分析

在实验部分,FPCVM 的分类性能和特征选择性能将会分别被检验。首先 FPCVM 的分类性能将会在低维度数据集中与 RVM 和 PCVM 进行对比。随后 FPCVM 将在高维数据集中测试其特征选择的效率,性能将会与主流的特征选择算法进行比较。实验中错误率作为检验标准。在实验部分所需的阈值都是通过在训练集上的交叉验证获得。

3.1 低维数据分类测试

第一阶段的数据集采用的是 Ripley 的合成数据集和拉舍尔 Rätsch 的香蕉数据集。其中 Ripley 的合成数据集是由 2 个部分重合的二维高斯分布混合得到,有着 8% 的固有错误率[11],而香蕉数据集是由 Rätsch 采用更复杂的方法合成的[12],详细信息见表 1。这一阶段实验的对比算法为 RVM[2] 和 PCVM[1]。

表 1 度数据集

数据集	训练集个数	测试集个数	正类样本占比例/%	维度
合成数据集(synthetic)	250	1 000	44.83	2
香蕉数据集(banana)	400	4 900	50.00	2
拼接数据集(splice)	1 000	2 175	44.93	60
肿瘤数据集(colon)	61	1	35.48	2 000

在香蕉数据集实验中 RVM 和 PCVM 中的阈值参数通过交叉认证获得[4],实验结果如图 2 所示(上半部分为 Ripley 的合成数据集,下半部分为 Rätsch 的香蕉数据集)[4]。在实验中由于样本维度过低,FPCVM 没有进行特征选择,但从结果中可以看到 FPCVM 有着与另 2 个分类器相当的分类性能,而 FPCVM 不需要交叉验证可以自动优化参数,在实际应用中可以省去交叉验证所带来的额外开销。

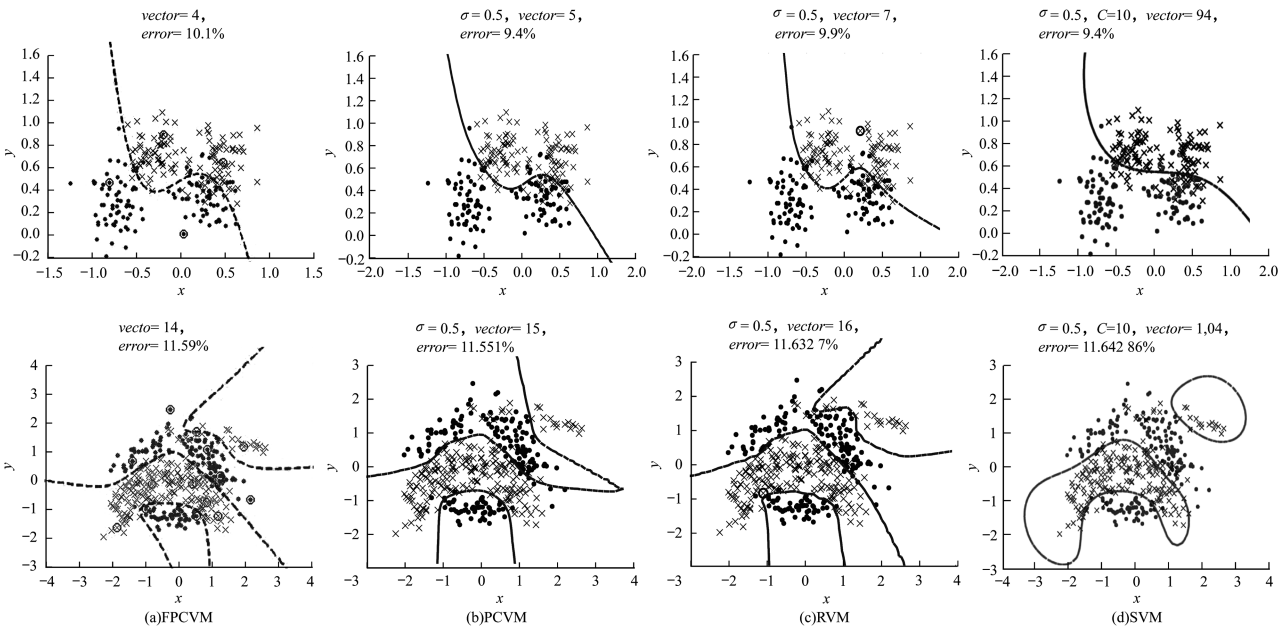

图 2 在低维度数据集下的测试结果

3.2 高维数据集测试

这一阶段的实验主要是验证 FPCVM 在高维度数据集下的分类性能。使用的 2 个数据集 DNA 拼接(splice)数据集和肿瘤数据集(colon_cancer)。DNA 拼接(splice)数据集是由 DNA 片段拼接成的 60 维数据集,由加州大学欧文分校(University of California at Irvine,UCI)[13] 提供。肿瘤数据集[14](colon_cancer)包含 2 000 维特征,包括了 22 个正

常人的数据和 40 肿瘤患者的数据。数据集的详细信息参考表 1。肿瘤数据集的实验重复 62 次,采用 leave-one-out 交叉验证的思想,每次留出一个样本当作测试集,剩余 61 个当作训练集,计算平均错误率;拼接数据集的实验重复 100 次,每次选出 1 000 个样本当作训练集剩余的 2 175 个样本当作测试集,计算平均错误率。最终分类器的性能将于主流的特征选择算法进行对比,对比算法包括基于最小平方误差的特征选择(LS)、T-test [4],mRMR [7]和贝叶斯的稀疏多值罗辑回归(Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation, SBMLR)[15]。LS, T-test 和 mRMR 算法选出的特征子集将会使用 LIBSVM 进行分类测试,并统计错误率。测试结果如图 3 所示。

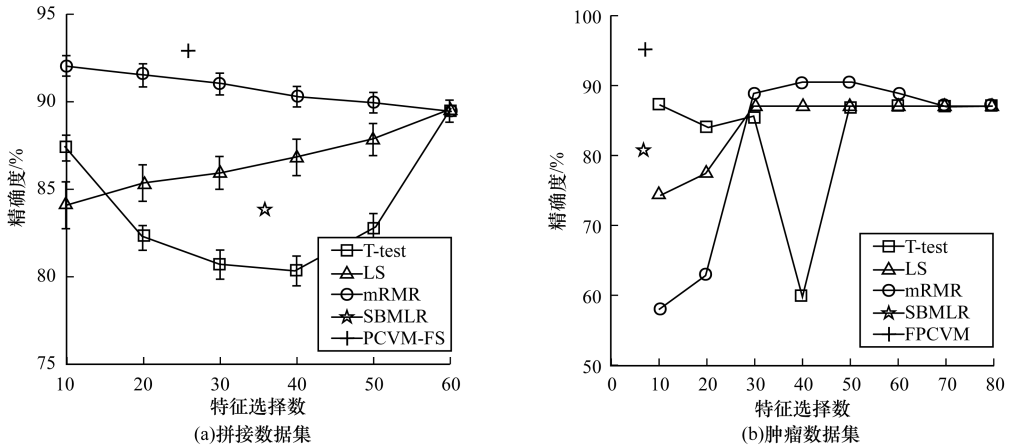

图 3　FPCVM 在高维度数据集下的对比测试结果

在高维度数据集的测试中,FPCVM 表现出良好的特征选择能力,在 60 维的拼接数据集中和 2 000 维的肿瘤数据集中分别选取了 26 维的特征子集和 7 维的特征子集进行分类,其分类性能也明显超过其余几个算法。相比于其他特征选择算法,FPCVM 不仅能够得到更稀疏的模型,而且可以同时进行分类器的学习和特征子集的选取。在 4 个数据集中的实验中,在高维度包含大量不相关特征的数据集中,FPCVM 的性能明显优于其余算法,见表 2,而在低维度不需要特征选择的数据集中,FPCVM 的分类性能也与主流的分类算法性能相当并且可以省去检查验证带来的额外时间(错误率作为性能指标,括号内为选取的特征子集的维度)。这样证明了 FPCVM 是一个优秀的分类器。

表 2　高维度实验结果

数据集	FPCVM	LS	T-test	mRMR	SBMLR
拼接数据集(splice)	0.071 2(26)	0.140 4(30)	0.193 0(30)	0.089 7(30)	0.162 6(35.86)
肿瘤数据集(colon)	0.048 4(7)	0.258 1(1)	0.129 0(10)	0.419 4(10)	0.194 5(6.70)

4　结束语

本文提出了一个可以同步进行分类器学习和特征子集选取的分类算法 FPCVM。该算法采用稀疏贝叶斯推理模型,通过在模型的参数和核参数中引入截断高斯分布作为先验,可同时在样本空间和特征空间中引入稀疏,得到一个稳定的分类器。通过分别在高维度和低维度数据集中的实验,证明 FPCVM 同时拥有较好的分类和特征选择能力。

参考文献

[1] Chen Huanhuan, Tino P, Yao Xin. Probabilistic Classification Vector Machines[J]. IEEE Transactions on Neural Networks,2009,20(6):901-914.

[2] Tipping M E. Sparse Bayesian Learning and the Relevance Vector Machine[J]. Journal of Machine Learning Research,2001(1):211-244.

[3] Vapnik V. Statistical Learning Theory[M]. London, UK:London University Press,1998.

[4] Zhou Nina, Wang Lipo. A Modified T-test Feature Selection Method and Its Application on the Hapmap Genotype Data[J]. Genomics, Proteomics & Bioinformatics,2007,5(3):242-249.

[5] 李　炜,章　寅,赵小强. 混合核函数稀疏 LS-SVM 软测量建模与应用[J]. 控制工程,2012,19(1):81-85.

[6] 毕佳佳,张　晶. 基于关系选择的多关系朴素贝叶斯分类[J]. 计算机工程,2016,42(5):218-223.

[7] Peng Hanchuan, Long Fulmi, Ding Chris. Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226-1238.

[8] Jolliffe I. Principal Component Analysis[M]. [S. l.]: Wiley Online Library,2005.

(下转第 193 页)

同时,MDSD 方法与 Amazon 方法相比加入了领域相关性更高的词向量特征,实验结果表明,MDSD 在诸多评价指标上均优于 Amazon。

4　结束语

本文提出一种新的评估在线产品评论有用度的方法。新方法吸收了神经网络语言模型的思想,结合词向量模型,改善有用度评估的效果。本文使用真实数据集以及多种评价标准来评价基于词向量的有用度评估模型,证明了该模型的有效性。此外,本文探究了领域相关词向量模型与领域无关模型在产品评论有用度评估任务上的优劣性,通过多组实验对比,证明了领域相关的词向量可以更好地优化评估效果。下一步将侧重于训练更粗粒度的神经网络语言模型,生成句子向量或者文档向量。

参考文献

[1] Jindal N, Liu Bing. Opinion Spam and Analysis[C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. Los Angeles, USA: IEEE Press, 2008: 219-230.

[2] Liu Jingjing, Cao Yunbo, Lin Chin-yew, et al. Low-quality Product Review Detection in Opinion Summarization[C]//Proceedings of EMNLP-CoNLL'07. Prague, Czech Republic: [s. n.], 2007: 334-342.

[3] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. The Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[4] 邱云飞,王建坤,邵良杉,等. 基于用户行为的产品垃圾评论者检测研究[J]. 计算机工程,2012,38(11):254-257,261.

[5] Kim S M, Pantel P, Chklovski T, et al. Automatically Assessing Review Helpfulness[C]//Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: [s. n.], 2006: 423-430.

[6] 胡令传,陶晓鹏. 客户评论中用户体验信息自动提取研究[J]. 计算机工程,2015,41(1):49-53.

[7] O'Mahony M P, Smyth B. Using Readability Tests to Predict Helpful Product Reviews[C]//Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information. Paris, France: [s. n.], 2010: 164-167.

[8] Liu Yang, Huang Xiangji, An Aijun, et al. Modeling and Predicting the Helpfulness of Online Reviews[C]//Proceedings of the 8th IEEE International Conference on ICDM. Pisa, Italy: [s. n.], 2008: 443-452.

[9] Hong Yu, Lu Jun, Yao Jianmin, et al. What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, USA: ACM Press, 2012: 495-504.

[10] 孟佳娜,段晓东,杨　亮. 基于特征变换的跨领域产品评论倾向性分析[J]. 计算机工程,2013,39(10):167-171.

[11] Yang Yinfei, Yan Yaowei, Qiu Minghui, et al. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews[C]//Proceedings of ACL'15. Beijing, China: [s. n.], 2015: 38-44.

[12] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//Proceedings of ANIPS'. Lake Tahoe, USA: [s. n.], 2013: 3111-3119.

[13] Mikolov T, Chen Kai, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[C]//Proceedings of ICLR'13. Scottsdale, USA: [s. n.], 2013: 254-262.

[14] Smola A J, Schölkopf B. A Tutorial on Support Vector Regression[J]. Statistics and Computing, 2004, 14(3): 199-222.

[15] Rob H, Koehler J, Anne B. Another Look at Measures of Forecast Accuracy[J]. International Journal of Forecasting, 2006, 22(4): 679-688.

编辑　索书志

(上接第 187 页)

[9] Nelder J A, Baker R J. Generalized Linear Models[M]. [S.l.]: Wiley Online Library, 1972.

[10] Dempster A P, Laird N M, Rubin D B, et al. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[11] Ripley B D. Pattern Recognition and Neural Net-works[M]. [S.l.]: Cambridge University Press, 2007.

[12] Rätsch G, Onoda T, Müller K R. Soft Margins for Adaboost[J]. Machine Learning, 2001, 42(3): 287-320.

[13] Blake C, Merz C J. UCI Repository of Machine Learning Databases[2011-11-23]. http://www.ics.uci.edu/mlearn/Machine-Learning.html.

[14] 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[C]//Proceedings of National Academy of Sciences, 1999, 96(12): 6745-6750.

[15] Cawley G C, Talbot N L C, Girolami M. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation[C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: [s. n.], 2007: 209-216.

编辑　索书志