doi:10.3772/j.issn.1002-0470.2014.09.010

基于最小交叉熵的相关向量机①

程丹松② 杨剑哲 李思倩 石大明③ 王 君 黄庆成

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 研究了传统相关向量机(RVM)的性能,分析了传统 RVM 的性能完全取决于先验假设的连接权值和参数的平滑性,因而其稀疏性实际上仍受核函数或核参数选择的控制,这在某些情况下可能会导致严重的欠拟合或过拟合现象的问题,在此基础上,提出了明确地给出基函数优化过程中的目标数量,并通过最小化训练阶段前向"假定"概率分布和测试阶段反向"真实"概率分布间的交叉熵来构建 RVM 的方法。实验结果表明,这种方法不但可以构建最小复杂度的基于最小交叉熵的 RVM 结构,而且构建的 RVM 能很好地对数据进行拟合,提高预测的准确性,增强其稀疏性。

关键词 相关向量机(RVM),贝叶斯推理,最小交叉熵,径向基函数(RBF)网络

0 引 言

支持向量机(support vector machine, SVM)是一种经典的核函数方法。核函数方法因具有稀疏性和数学处理能力,在解决一些机器学习和模式识别问题方面获得了广泛的应用。支持向量机(SVM)是由只依赖一些与训练样本相关联的核函数子集(支持向量)构成的模型函数[1]。虽然支持向量机也可以产生一个稀疏模型,但在相同情况下,支持向量的数目仍会随训练样本数目的增加而线性增长,这可能会造成过拟合,另外还会浪费计算时间。针对这两个缺点,Tipping 在文献[2]中提出了相关向量机(relevance vector machine , RVM)方法,它是一种基于贝叶斯框架的概率学习模型。传统的 RVM 是应用核函数方法,在使用最少相关基函数的情况下构建的基函数网络。同 SVM 相比, RVM 方法具有更好的稀疏性,并可以对超参数进行自动估计。RVM 极大地减少了核函数的计算量, 并且在核函数的选择上,不受梅西定理的限制,可以构建任意的核函数。所以使用贝叶斯公式的 RVM 模型不但可以提供更好的稀疏性,而且还可以实现参数和概率输出的自动估计。RVM 使用核函数作为候选基函数,每个基函数对应训练集 $X = \{x_i\}_{i=1}^{N}$ 中的一个实例。在完成训练后,大多数的权值被训练成 0,只有少数对应非 0 权值的"相关矢量"。为了实现这样的稀疏并避免出现过度拟合,RVM 在权值方面使用球形高斯先验,在方差方面使用 Gamma 先验分布。因此,RVM 算法可以被看作是一个在权重上使用 student-t 分布的贝叶斯回归学习过程[3,4]。

近年来,许多学者在提高 RVM 方法的处理效率和稀疏性方面提出了一些新的观点[5-8]。如 Chen[5]提出了一种分层的 RVM 回归方法,由粗到细对数据进行处理,与传统 RVM 相比,不但缩短了处理时间,还提高了模型的精度。此外,Mehrotra 在文献[6]中,使用组合核函数的 RVM 算法来进行分类。

传统的 RVM 在稀疏方面的性能都是由先验的平滑性决定的。然而一个明确的先验结构在权值变化上的不足意味着稀疏实际上取决于核函数和(或)核参数的选择。它可能会导致严重的过度拟合或欠拟合。为了解决这个问题,Schmolck 等人[9]提出了一种在贝叶斯回归过程中控制稀疏性的有效方案,即通过灵活的噪声相关平滑函数来取代经典的 RVM 的 Gamma 先验。在他们的方法中,一个基于 symmlet 小波的平滑先验模型使 RVM 回归适用

① 国家自然科学基金(61370162)和中国航天科技集团公司哈尔滨工业大学联合技术创新中心(CASC-HIT13-1004)资助项目。
② 男,1972 年生,博士,副教授;研究方向:机器学习,图像处理,计算机视觉,电磁场可视化;E-mail: cdsinhit@hit.edu.cn
③ 通讯作者, E-mail: dshi@hit.edu.cn
(收稿日期:2014-02-27)

于多种信号,无需再通过交叉验证(cross validation, CV)来确定核参数。但是即使先验足够光滑,该方法也无法保证输出函数的绝对光滑,这就是导致过(欠)拟合的原因。最近,Kanamori 等又提出了一个基于准确度和计算效率差异的学习方法,称为无约束最小二乘拟合(unconstrained least- squares importance fitting)[10],它被证明具有最优的非参数收敛速度和数值稳定性。但是,无约束最小二乘拟合的实际性能也取决于核函数和正则化参数的选择。张宇航在文献[7]中利用 Fisher 线性鉴别分析(FLDA) 技术,在分类前对高光谱数据作可分性预处理来提高 RVM 方法的速度和效果,但该方法的实际性能也取决于核函数的选择。针对这个问题,我们在前期的工作中[8],使用 Kullback-Leibler (KL)分布来最小化两个用期望最大化(expectation maximum, EM)算法得到的概率分布偏差。然而,基于 KL 的学习只有当两个概率分布间达到最佳匹配时,才能产生最好的数据拟合。所以本文在基函数优化过程中,使用最小交叉熵来明确目标的数目,进而达到稀疏。与传统 RVM 方法相比,它在与数据依赖性相关的参数优化调整和自动模型选择方面具有较好的优势。这也意味着 RVM 基函数的最优数目和核参数一样,可在最和谐原则下被自动确定。

1 RVM 的稀疏学习

在关于分类和回归问题的有监督学习中,对于每个训练事例 x_i 都对应一个期望的目标 t_i($i = 1, \cdots, N$),其中 N 是训练事例的总数。我们使用 $D = \{(x_i, t_i)\}_{i=1}^{N}$ 来定义训练数据集。$X = \{x_i\}$ 是训练输入,$T = \{t_i\}$ 是训练输出。因此,我们简单定义 $D = \{X, T\}$。

在 RVM 中,训练事例和输入是多维矢量,目标和输出是标量,在回归时它是实数值,在分类时它是整数形式的分类标签。根据标准概率公式,目标可以被视为高斯噪声扰动的输出。因此,数据 (x_i, t_i) 的似然可以表示成下式所示的形式:

$$p(t_i \mid x_i, \boldsymbol{w}) = \mathbb{N}(t_i \mid y(x_i, \boldsymbol{w}), \sigma^2) \tag{1}$$

其中 σ^2 是加性噪声的方差。

1.1 参数学习的最大边缘似然

在给定训练集 D 后,通过计算后验概率分布 $p(\boldsymbol{w}, \alpha, \sigma^2 \mid D) = p(\boldsymbol{w}, \alpha, \sigma^2 \mid t)$,可以得到目标学习模型的权重 $\boldsymbol{w}$、权重方差 α 和噪声方差 σ^2,但后验概率不能直接计算,所以需要对后验概率

$$p(\boldsymbol{w}, \alpha, \sigma^2 \mid t) = p(\boldsymbol{w} \mid t, \alpha, \sigma^2) p(\alpha, \sigma^2 \mid t) \tag{2}$$

进行分解,其中:

$$p(\boldsymbol{w} \mid t, \alpha, \sigma^2) = \frac{p(t \mid \boldsymbol{w}, \sigma^2) p(\boldsymbol{w} \mid \alpha)}{p(t \mid \alpha, \sigma^2)} \tag{3}$$

这样相关矢量的学习就变成寻找超参数的后验概率。由于 $p(\alpha, \sigma^2 \mid t) \propto p(t \mid \alpha, \sigma^2) p(\alpha) p(\sigma^2)$,且 $p(\alpha)$ 和 $p(\sigma^2)$ 是常量,所以我们只需要在归一化超先验的情况下最大化 $p(t \mid \alpha, \sigma^2)$ 项,并称 $p(t \mid \alpha, \sigma^2)$ 为模型证据,它可以像下式那样在贝叶斯模型框架下进行计算:

$$p(t \mid \alpha, \sigma^2) = \int p(t \mid \boldsymbol{w}, \sigma^2) p(\boldsymbol{w} \mid \alpha) \mathrm{d}\boldsymbol{w} \tag{4}$$

在 RVM 中学习超级参数 α 和 σ^2 等价于在式(4)中最大化 $p(t \mid \alpha, \sigma^2)$,这就是最大边缘似然方法[11]。当进行超参数学习时,模型权重 $\boldsymbol{w}$ 可以通过式(2)所示的后验期望被估计。

1.2 基于先验权重的参数正规化

在上一节已经介绍了参数学习过程中的先验概率分布权重 $p(\boldsymbol{w} \mid \alpha)$,同时明确地指出该权重是由一个均值为 0、对角协方差为 α 的球形高斯确定,如下式所示:

$$p(\boldsymbol{w} \mid \alpha) = \prod_{i=1}^{N} \mathbb{N}(\omega_i \mid 0, \alpha_i^{-1}) \tag{5}$$

其中 α_i 是超参数。为了避免出现过度拟合,RVM 对 α_i 使用 Gamma 超先验分布,同样在噪声方差上也施加 Gamma 超先验分布。这样 $p(\boldsymbol{w} \mid \alpha) p(\alpha)$ 被整合成 α_i 的形式,它的结果是 student-t 密度函数。所以 RVM 算法可以被看作是一个在权重上使用 student-t 分布的贝叶斯回归学习过程[3]。

每个独立的超参数或基函数的使用是相关向量机的关键特征,并决定最终的稀疏特性。在训练中,由于许多 α_i 趋向于无穷大,所以导致式(5)中相应 $p(\boldsymbol{w}_i \mid \alpha)$ 变为零,进而可以忽略这些权值为零的基函数,从此实现了模型的稀疏性。

2 基于最小交叉熵的 RVM

RVM 学习的目的是对数据进行和谐拟合,以获得相关向量的最佳设置,即获得最少的相关向量,从而达到稀疏。在本节最小交叉熵理论被用于实现这个目标。

2.1 候选相关向量

对于在 RVM 训练阶段给定的输入数据集 $X =$

$\{x_i\}_{i=1}^{N}$，其输出结果是在目标矢量 $\boldsymbol{t} = (t_1, t_2, \cdots, t_N)^{\mathrm{T}}$ 中找到 k 个相关矢量。在每个训练事例中，核函数被当作候选相关向量（基函数），并用标签 j 进行索引。由于在训练阶段，具有非零权重的基函数在时间上是各异的，所以候选的相关向量可以被看作是一个随机过程。

这样 RVM 学习问题就可以等价为一个模型选择问题，即在候选模型（每个训练实例的基函数）中选择一个候选子集（相关向量）。通常，模型选择方法分为两个阶段，即，模型的参数学习和特定准则（如贝叶斯信息准则[12]）下的最优评价模型。但这两个阶段的算法都是十分耗时的，因为在参数学习的过程必须对每个被列出的候选模型都进行处理。所以为了解决模型选择问题，三个贝叶斯相关方法被依次提出，它们分别是最小信息长度[13]、变分贝叶斯算法[14]和贝叶斯阴阳和谐理论[15]。这三种方法都具有自动选择模型的能力，并且在贝叶斯学习过程中都能利用一个内在能量来减少无关紧要的或不相关的模型维数。但通过文献[16]的研究对比可知，贝叶斯阴阳和谐理论在处理效果上大大优于其它参数最优化选择的方法。而变分贝叶斯方法和最小信息长度方法，在没有参数先验时，会被退化为最大似然算法。换句话说，这两个方法缺乏一个优化 Dirichlet- Normal- Wishart 先验的超参数。在本项研究中，交叉熵方法被用于参数学习和稀疏学习时的自动模型选择。

2.2 训练和测试阶段的最小交叉熵

本节首先考虑在 RVM 训练阶段的三个随机过程：$\boldsymbol{x}$, $\boldsymbol{t}$, $\boldsymbol{j}$。它们三者的联合分布可以通过以下两个公式进行计算：

$$p(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) = p(\boldsymbol{x})p(\boldsymbol{t} \mid \boldsymbol{x})p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x}) \quad (6)$$

$$p(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) = p(\boldsymbol{j})p(\boldsymbol{x} \mid \boldsymbol{j})p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{j}) \quad (7)$$

在理想的情况下，式(6)和式(7)将返回相同的结果。但实际上 j 对应的相关向量是最优设置，否则，这两个等式不可能得到相同的结果。等式(6)等号的右半部分表明它是一种从已知的训练数据和目标来获取解决方案的前向途径；而等式(7)等号的右半部分表明它是测试当前解性能的反向途径。假如我们用 $p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})$ 和 $p_{\mathrm{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})$ 来分别定义式(6)和式(7)的联合概率分布，两者的最佳匹配可通过典型的均方误差（MSE）和最大似然（ML）来衡量，并通过最小化 KL 分布来实现

$$KL(p_{\mathrm{train}} \mid p_{\mathrm{test}}) = \int_{x,t,j} p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) \log \frac{p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})}{p_{\mathrm{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{j} \quad (8)$$

在本文我们使用 $p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})$ 来匹配 $p_{\mathrm{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})$，从而使结果更接近于训练数据。但是由于缺乏对最小复杂性的控制，基于 KL 学习的泛化能力很弱[17]。同时根据文献[18]可知，KL 距离可以被视为一个特殊类型的交叉熵。在式(8)中，p_{train} 的分布是固定的先验分布；p_{test} 的分布是在 KL 散度最小化时与 p_{train} 的最佳匹配。然而，在构建 RVM 时，我们寻找的不仅是 p_{train} 和 p_{test} 间的最佳匹配，还有其最紧凑的结构。为了实现这一目标，将进行另一方面的最小化，把 p_{test} 分布看作是固定的参考分布，对 p_{train} 进行优化，从而使其在最少相关向量的情况下尽可能接近于 p_{test}。在本文我们引入贝叶斯阴阳和谐学习理论[15,17]来处理负交叉熵，并在下面的讨论中交替使用最小化的交叉熵和最大化的负交叉熵。并通过最大化贝叶斯阴阳和谐函数来构建最紧凑的 RVM 结构：

$$H(p_{\mathrm{train}} \parallel p_{\mathrm{test}}) = \int_{x,t,j} p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) \log p_{\mathrm{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{j} \quad (9)$$

比较式(8)和式(9)，我们可以得到如下式所示的训练-测试-熵表达式：

$$H(p_{\mathrm{train}} \parallel p_{\mathrm{test}}) = -KL(p_{\mathrm{train}} \parallel p_{\mathrm{test}}) + H(p_{\mathrm{train}} \parallel p_{\mathrm{train}}) \quad (10)$$

最大化的训练-测试-熵由两部分组成：训练和测试间的最小 KL 分布和训练空间的最大熵。前者使数据具有较好的拟合性，后者在一些限制条件下使复杂性变得最小。

当训练事例 $\boldsymbol{x}$ 和目标 $\boldsymbol{t}$ 是已知的情况下，前向训练阶段的“假定”概率分布为 $p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) = p(\boldsymbol{x})p(\boldsymbol{t} \mid \boldsymbol{x})p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x})$，当训练数据以 (x_i, t_i) 形式出现时，我们认为 $p_{\mathrm{train}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) = p(\boldsymbol{x}, \boldsymbol{t})p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x})$ 模型结构可以表示成如下的形式：

$$p(\boldsymbol{x}, \boldsymbol{t}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - x_n)\delta(\boldsymbol{t} - t_n) \quad (11)$$

$$p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x}) \text{ 是无拘束的} \quad (12)$$

反向测试阶段的“真实”概率分布 $p_{\mathrm{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j}) = p(\boldsymbol{j})p(\boldsymbol{x} \mid \boldsymbol{j})p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{j})$ 对应反向测试路径。当对由候选关联向量 $\boldsymbol{j}$ 组成的 RVM 结构进行测试时，它的输出目标 $\boldsymbol{t}^*$ 由给定一个输入 $\boldsymbol{X}$ 构建。本文定义

$p(\boldsymbol{j}) = p_j$,其中 $p_j \geqslant 0$ 且 $\sum p_j = 1$,并进行下面的合理假设:

$$p(\boldsymbol{x} \mid \boldsymbol{j}) = \mathbb{N}(\boldsymbol{x} \mid x_j, \sigma^2 \boldsymbol{I}) \tag{13}$$

$$p(\boldsymbol{t}^* \mid \boldsymbol{x}, \boldsymbol{j}) = \mathbb{N}(\boldsymbol{t}^* \mid t_j, \sigma_j^2) \tag{14}$$

所以式(9)中的和谐函数是一个由自由分布 $\hat{p} = p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x})$,模型参数 $\boldsymbol{\Theta} = \{\sigma_x^2, (p_j, \sigma_j^2)_{j=1}^N\}$ 和非零权重基函数的数目 k 构成的函数:

$$\begin{aligned} H(\hat{p}, \boldsymbol{\Theta}, k) &= H(p_{\text{train}} \parallel p_{\text{test}}) \\ &= \int_{x,j,t} p(\boldsymbol{x}, \boldsymbol{t}) p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x}) \log \\ &\quad [p(\boldsymbol{j}) p(\boldsymbol{x} \mid \boldsymbol{j}) p(\boldsymbol{t}^* \mid \boldsymbol{x}, \boldsymbol{j})] \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{t} \mathrm{d}\boldsymbol{j} \end{aligned} \tag{15}$$

在文献[19]中,对应 $\hat{p}$ 的 H 最大值为

$$\hat{p}^* = p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x}) = \frac{p_{\text{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})}{\sum_j p_{\text{test}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{j})} \tag{16}$$

根据等式(13)、(14)指定的模型格式,我们可知

$$\hat{p}^* = = \frac{p_j \mathbb{N}(\boldsymbol{x} \mid x_j, \sigma_x^2 \boldsymbol{I}) \mathbb{N}(\boldsymbol{t} \mid t_j, \sigma_j^2)}{\sum_{h=1}^N p_h \mathbb{N}(\boldsymbol{x} \mid x_h, \sigma_x^2 \boldsymbol{I}) \mathbb{N}(\boldsymbol{t} \mid t_h, \sigma_h^2)} \tag{17}$$

把式(11)、(17)和(13)、(14)代入式(15),我们可以重写它的表达式,注意:此处 $p(\boldsymbol{j} \mid \boldsymbol{t}, \boldsymbol{x})$ 被它的最大值 $\hat{p}^*$ 所替代:

$$H(\boldsymbol{\Theta}, k) = H(\hat{p}^*, \boldsymbol{\Theta}, k) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N Q_j(x_i) \cdot J_j(x_i) \tag{18}$$

其中,

$$Q_j(x_i) = \frac{p_j \mathbb{N}(x_i \mid x_j, \sigma_x^2 \boldsymbol{I}) \mathbb{N}(t_i \mid t_j, \sigma_j^2 \boldsymbol{I})}{\sum_{h=1}^N p_h \mathbb{N}(x_i \mid x_h, \sigma_x^2 \boldsymbol{I}) \mathbb{N}(t_i \mid t_h, \sigma_h^2 \boldsymbol{I})},$$

$J_j(x_i) = \log p_j \mathbb{N}(x_i \mid x_j, \sigma_x^2 \boldsymbol{I}) \mathbb{N}(t_i \mid t_j, \sigma_j^2 \boldsymbol{I})$。

2.3 RVM 模型预测

本节可以很容易地以概率形式来构建 RVM 的预测函数:$p(\boldsymbol{t}^* \mid \boldsymbol{x}^*) = \sum_{j=1}^N p(j \mid \boldsymbol{x}^*) p(\boldsymbol{t}^* \mid \boldsymbol{x}^*, j)$,其中 $\boldsymbol{x}^*$ 和 $\boldsymbol{t}^*$ 分别表示测试数据和 RVM 的预测结果,它的输出如下式所示:

$$\begin{aligned} E(\boldsymbol{t}^* \mid \boldsymbol{x}^*) &= \sum_{j=1}^N \frac{p_j \mathbb{N}(\boldsymbol{x}^* \mid x_j, \sigma_x^2 \boldsymbol{I})}{\sum_{h=1}^N p_h \mathbb{N}(\boldsymbol{x}^* \mid x_h, \sigma_x^2 \boldsymbol{I})} \\ &\quad E(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{j}) \\ &= \sum_{j=1}^N p_j t_j \frac{\mathbb{N}(\boldsymbol{x}^* \mid x_j, \sigma_x^2 \boldsymbol{I})}{\sum_{h=1}^N p_h \mathbb{N}(\boldsymbol{x}^* \mid x_h, \sigma_x^2 \boldsymbol{I})} \end{aligned} \tag{19}$$

本研究只考虑高斯核函数 $K(\boldsymbol{x}^* \mid x_j) = \exp(-\frac{\| \boldsymbol{x}^* - x_j \|^2}{2\sigma_x^2})$,让 $E(\boldsymbol{t}^* \mid \boldsymbol{x}^*) = \sum_{j=1}^N \omega_j K(\boldsymbol{x}^*, x_j)$,我们通过更新参数 σ_x^2 和 p_j 来计算权重 ω_j,式为

$$\omega_j = \frac{p_j t_j}{\sum_{h=1}^N p_j \exp(-\frac{\| \boldsymbol{x}^* - x_h \|^2}{2\sigma_x^2})} \tag{20}$$

2.4 参数和稀疏的自动优化

考虑到上面的约束 $\sum p_j = 1$,引入拉格朗日乘数 λ 到 $H(\boldsymbol{\Theta}, k)$ 中,得到表达式

$$L(\boldsymbol{\Theta}, k, \lambda) = H(\boldsymbol{\Theta}, k) + \lambda(1 - \sum_{j=1}^N p_j) \tag{21}$$

这样参数和稀疏的优化就变成一个迭代过程。我们利用固定点迭代的方式来进行参数更新,即:在每次迭代时,分别对 $p_j, \sigma_x^2, \sigma_j^2$ 和 λ 计算 $L(\boldsymbol{\Theta}, k, \lambda)$ 的方差,然后对参数进行更新。这里需要注意,相关向量机的数目 k,并不是真正的参数,而是每次迭代 $p_j \neq 0$ 时基函数的个数。这样,一方面,每个 p_j 将尽可能接近式(15)给出的和谐函数 $H(\boldsymbol{\Theta}, k)$;另一方面,在式(21)的 $\sum p_j = 1$ 约束下,一些 p_j 将趋向于 0。因此,使用贝叶斯阴阳学习的最大化和谐函数将产生最小数目的 k 值。即 RVM 的稀疏结构可根据其模型自动选择构成。

3 实 验

本研究使用新的交叉熵方法对合成数据集进行测试。本实验将用来评估基于交叉熵方法在超参数学习方面的能力,同时将本文提出的交叉熵和贝叶斯阴阳和谐学习方法与基于 LASSO 的贝叶斯稀疏核学习(BSKL- LASSO)方法[10]、文献[20]方法和 Tipping 的 RVM[7]方法进行比较。在这里,我们将文献[9]中实现的 RVM 算法称作改进的 RVM(mRVM)。

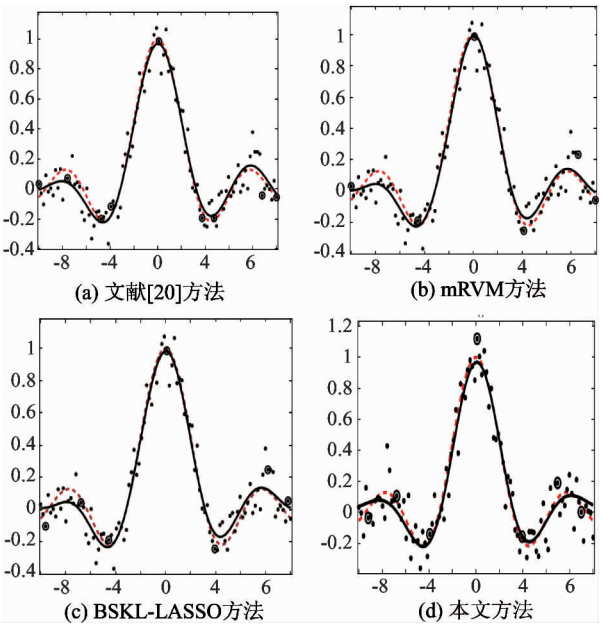
例子 1:在本例中,本文使用高斯径向基函数(RBF)网络作为模型的标量函数:

$$\operatorname{sinc}(x) = \sin(x)/x, \quad -10 \leqslant x \leqslant 10$$

在实验中使用的 RBF 核函数满足如下形式:$K(\boldsymbol{x}, \boldsymbol{x}'; \{\sigma_x\}) = \exp\{-\frac{(\boldsymbol{x} - \boldsymbol{x}')^2}{2\sigma_x^2}\}$,其中 σ_x^2 是 RBF 核的宽度,完整的 RBF 神经网络模型(1)是由所有位于输入训练数据中心的 RBF 回归定义的。训练数据集 $\{(x_i, t_i)\}_{i=1}^{100}$ 是由[-10,10]区间均匀

分布的输入 x_j 生成的。目标 t_j 的噪声由均值为 0、方差为 0.1 的高斯给出。在完整的 RBF 模型中 N 的取值为 100。

本文使用改进的 RVM 方法、BSKL- LASSO 方法和本文方法对数据集的 σ_x^2 值进行学习，其中 BSKL-LASSO 方法的 RBF 核函数宽度设定为 3。在实验结果上，mRVM 方法获得的 $\sigma_x^2 = 2.6193$，BSKL- LASSO 方法获得的 $\sigma_x^2 = 3.2258$，基于最小交叉熵方法获得的 $\sigma_x^2 = 3.6523$。模型所产生的回归向量和未知数据的预测如图 1 所示。图 1(d)显示了本文方法产生的模型，其显著矢量(被选择的回归)用○表示。



点是噪声的训练数据，曲线是 sinc(x)函数，实曲线是由不同算法生成的模型；标记○显示了每种算法的关键回归函数

图 1　单一尺度 sinc 函数模型的结果

表 1 显示了文献[20]方法、mRVM 模型、BSKL-LASSO 模型和本文的交叉熵模型在训练集和测试集中的归一化均方根误差(NRMSE)。从表 1 中可以看出在未知核宽度的情况下，基于最小交叉熵模型的结果比其它三种方法好。在实际应用中，本文的模型在式(21)归一化约束下，进行 19 次迭代循环后，就已经去掉了大量的不必要核基函数。

表 1　不同算法的性能比较

方法(回归的数目)	训练集的 NRMSE	测试集的 NRMSE
文献[20](10)	1.3747e-3	1.3737e-3
mRVM(6)	1.5889e-3	1.5975e-3
BSKL-LASSO(7)	1.5369e-3	1.5209e-3
本文方法(7)	1.3128e-3	1.3196e-3

例子 2：文献[21]给出了下面的合成函数：

$$t(\boldsymbol{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \sum_{k=6}^{10} 0 \cdot x_k$$

输入 $x = (x_1, x_2, \cdots, x_{10})$ 被定义为 10 维的超立方体，其中输入变量 x_6 到 x_{10} 没有对目标值 $\boldsymbol{t}$ 作出任何贡献。我们选用从 10 维的超立方体得到包含 300 个输入 $\{x_i\}$ 的训练数据集；对应的目标 $\{t_i\}$ 是加入标准方差为最大函数值 20% 高斯噪声的函数值。对于测试数据集，本例使用 1000 个无噪声的事例。

本例分别运行 mRVM、BSKL-LASSO 和本文的交叉熵模型。表 2 显示了模型中核基函数的平均回归值、训练和测试的平均归一化均方根误差。结果表明，我们的模型不仅提高了平均测试的归一化均方根误差，而且与其它模型相比使用更少的回归系数。

表 2　回归数目的比较

方法	训练的 NRMSE	测试的 NRMSE	回归的平均数目
mRVM	0.0261 ±0.0039	0.0286 ±0.0067	59.8 ±10.7
BSKL-LASSO	0.0105 ±0.0018	0.0179 ±0.0019	10.7 ±3.7
本文方法	0.0122 ±0.0011	0.0139 ±0.0011	8.4 ±2.6

4　结　论

在本研究中，相关向量机的稀疏学习实现了在前向训练途径“假设”概率分布和反向检测途径“真实”概率分布的最小化差别。我们的思想可以概括如下，首先，将训练集候选的相关向量视为随机变量，使用贝叶斯阴阳和谐的学习方法将这种 RVM 稀疏学习问题转换成自动模型选择问题；第二，使用贝叶斯优化过程来进行参数迭代更新；第三，本文提出的最佳解决方案是在最小复杂性结构和良好数据拟合之间的一个折衷结构。实验显示我们提出

的 RVM 模型不仅提高了预测的准确性,同时也加强了其稀疏性。

参考文献

[1] Scholkopf B, Smola A. Learning with Kernels. Cambridge, Massachusetts: The MIT Press, 2002

[2] Tipping M. The relevance vector machine. In: Advances in Neural Information Processing Systems, MIT Press, 2000

[3] Gao J, Kwan P, Shi D. Sparse kernel learning with LASSO and its Bayesian inference. *Neural Networks*, 2010, 23(2): 257-264

[4] Mohsenzadeh Y, Sheikhzadeh H, Reza A M. The relevance sample-feature machine: A sparse Bayesian learning approach to joint feature-sample selection. *IEEE Trans Cybern*, 2013,43(6): 2241-2254

[5] Chen X J, Hu T, Wang D D, et al. Research of small parts gesture estimation based on multilevel RVM regression. In: Proceedings of the IEEE International Conference on Electronic Measurement & Instruments, Harbin, China, 2013. 877-881

[6] Mehrotra H, Vatsa M, Singh R, et al. Biometric match score fusion using RVM: a case study in multi-unit iris recognition. In: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2012. 65-70

[7] 张宇航,张晔. SVM 和 RVM 对高光谱图像分类的应用潜能分析. 哈尔滨工业大学学报, 2012,44(3):34-39

[8] Shi D, Nguyen M N, Zhou S, et al. Fuzzy CMAC with incremental Bayesian Ying-Yang learning and dynamic rule construction. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 2010, 40(2):548-552

[9] Schmolck A, Everson R. Smooth relevance vector machine: a smoothness prior extension of the RVM. *Machine Learning*, 2007, 68 :107-135

[10] Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 2009,10 :1391-1445

[11] Berger J. Statistical Decision Theory and Bayesian Analysis, 2nd . New York: Springer, 1985

[12] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*, 1978, 6(2):461-464

[13] Figueiredo M A F, Jain A K. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002,24(3) :381-396

[14] Jaakkola T, Jordan M. Bayesian parameter estimation through variational methods. *Statistics and Computing*, 2000,(10):25-37

[15] Xu L. Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint de- composition of data covariance, and applications of network biology. *Journal of Frontiers of Electrical and Electronic Engineering in China* , 2011, 6(1):86-119

[16] Shi L, Tu S, Xu L. Learning Gaussian mixture with automatic model selection: A comparative study on three Bayesian related approaches. *Frontiers of Electrical and Electronic Engineering*, 2011, 6(2):215-244

[17] Xu L. On essential topics of BYY harmony learning: Current status, challenging issues, and gene analysis applications. *Frontiers of Electrical and Electronic Engineering*, 2012, 7(2): 147-196

[18] Rubinstein R, Kroese D. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning. New York: Springer-Verlag, 2004

[19] Xu L. Bayesian Ying Yang Learning (I): A Unified Perspective for Statistical Modeling. In: Zhong N, Liu J eds, Intelligent Technologies for Information Analysis, Springer, 2004. 615-659

[20] Mohsenzadeh Y, Sheikhzadeh H. Gaussian Kernel Width Optimization for Sparse Bayesian Learning. *IEEE Trans On Neural Networks And Learning Systems*, 2014, 99 (10): 1-11

[21] Friedman, J H. Multivariable adaptive regression splines. *The Annals of Statistics*,1991, 19 (1):1-57

The relevance vector machine based on cross entropy minimization

Cheng Dansong, Yang Jianzhe, Li Siqian, Shi Daming, Wang Jun, Huang Qingcheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

The performance of the classical relevance vector machine (RVM) was studied, and it was analyzed that the performance of the original RVM purely depends on the smoothness of the presumed prior of the connection weights and parameters, consequently its sparsity is actually still controlled by the choice of kernel functions or kernel parameters, leading to severe underfitting or overfitting in some cases, and based on these, the RVM based on cross entropy minimization was constructed by explicitly involving the number of basis functions into the objective of the optimization procedure, and by the minimization of the cross entropy between the "hypothetical" probability distribution in the forward training pathway and the "true" probability distribution in the backward testing pathway. The experimental results show that the proposed methodology can achieve the construction of the structure with the least complexity, and the constructed RVM has the good data fitting, the good detection precision and the good sparsity.

Key words: relevance vector machine (RVM), Bayesian inference, cross entropy minimization, radial basis function (RBF) networks