基于高斯核的 SVM 的参数选择①

王行甫, 陈家伟

(中国科学技术大学 计算机学院, 合肥 230027)

摘 要: 基于高斯核的支持向量机应用很广泛, 高斯核参数 σ 的选择对分类器性能影响很大, 本文提出了从核函数性质和几何距离角度来选择参数σ, 并且利用高斯函数的麦克劳林展开解决了参数σ的优化选择问题. 实验结果表明, 该方法能较快地确定核函数参数σ, 且 SVM 分类效果较好, 解决了高斯核参数σ在实际应用中不易确定的问题.

关键词: 支持向量机; 高斯核; 参数选择; 几何距离; 麦克劳林展开

Parameter Selection of SVM with Gaussian kernel

WANG Xing-Fu, CHEN Jia-Wei

(School of Computer Science, University of Science & Technology of China, Hefei 230027, China)

Abstract: Support vector machine based on Gaussian kernel has been used in many areas. The parameter σ of the Gaussian kernel has great impact on the performance of the classifier. This paper proposes an approach to choose an optimal parameter σ based on the properties of the kernel function and the angle of geometric distance. What is more, we have solved the problem of the optimal option of the parameter σ by means of the McLaughlin expansion of the Gaussian kernel function. The experiment results indicate that this method can get parameter σ very quickly and can achieve high efficiency. Thus the difficulty of the estimation of the parameter σ can be solved by our method.

Key words: support vector machine; Gaussian kernel; parameter selection; geometric distance; McLaughlin expansion

1 引言

支持向量机(SVM)是 90 年代中期发展起来的基于统计学习理论的一种机器学习方法, 通过寻求结构化风险最小来提高学习机泛化能力, 实现经验风险和置信范围的最小化, 从而达到在统计样本量较少的情况下, 亦能获得良好统计规律的目的. 通俗来讲, 它是一种二类分类模型, 其基本模型定义为特征空间上的间隔最大的线性分类器, 即 SVM 的学习策略便是间隔最大化, 最终可转化为一个凸二次规划问题的求解.

SVM 是 Cortes 和 Vapnik 于 1995 年首先提出的, 它在解决小样本、非线性及高维模式识别中表现出许多特有的优势, 并能够推广应用到函数拟合等其他机器学习问题中.

SVM 学习中, 核函数的选择非常重要, 因为特征空间的结构由核函数决定, 它设计的好坏直接影响到分类效果. SVM 通过引入了核函数, 有效地解决了分类中的线性不可分问题. SVM 的理论研究主要以核函数方面的研究为主, 包括核函数的构造和核函数的选择. 由于核函数构造的复杂性, 目前对核函数的研究取得实质性进展的还是在核函数的选择上.

核函数的选择包括核函数类型的选择以及核函数参数的确定. 目前, 较常用的核函数主要有 3 类: (1)线性核函数; (2)多项式核函数; (3)高斯核函数. 其中, 高斯核函数具有较好地普适性, 在实际中应用最广泛, 并且效果很好.

选择高斯核函数来进行 SVM 学习后, 最重要的是核函数参数 σ (高斯径向基函数的宽度)的确定. σ 对分类器的性能影响很大, 若 σ 太小, 则所有的训练样

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(61272472,61232018,61202404);安徽省教学研究计划 2010
收稿时间:2013-11-16;收到修改稿时间:2013-12-12

本点都是支持向量，且它们全部能被正确的分类，但容易出现“过拟合”的现象，推广能力差；若 σ 太大，高斯核支持向量机对所有样本一视同仁，容易出现“欠拟合”的现象.

2 现有的方法及存在的问题

给定样本集 $X=\{x^{(1)},x^{(2)},...,x^{(i)},...,x^{(m)}\}$，$x^{(i)}$ 为样本的 n 维输入特征，样本标签 $y\in\{-1,1\}$. m 为样本的数目，m_+ 和 m_- 分别表示正样本和负样本的数目. 这是一个二分类问题，我们采用 SVM 来对其进行训练学习，选择高斯核作为 SVM 的核函数，其对应的映射函数为 $z=\phi(x)$. 其中，任意两个样本的输入特征分别为 $x^{(i)}$ 和 $x^{(j)}$，对应的样本标签分别为 $y^{(i)}$ 和 $y^{(j)}$.

训练一个 SVM 其实就是求解下面的二次规划(QP)问题:

$$\begin{aligned}&\min_{\gamma,\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}\xi_i\\ &s.t.\quad y^{(i)}\left(\omega^{(i)}z^{(i)}+b\right)\geq 1-\xi_i\\ &\quad\xi_i\geq 0,i=1,2,...m.\end{aligned}\tag{1}$$

其中，样本点 $x^{(i)}$ 通过映射函数 $z=\phi(x)$ 映射到高维空间中的 $z^{(i)}$，C 是对于错分样本的惩罚因子.

当选择高斯核函数来进行 SVM 学习时，SVM 优化问题(1)中的惩罚因子 C 和高斯核函数的参数 σ 是两个可以人为调节的参数，参数取值不同，对应的分类器性质以及推广识别率也将有很大差别.

目前确定高斯核函数参数 σ 的主要方法是：对 σ 取不同的值，然后分别对样本集使用选取的 σ 进行 SVM 训练，选择分类错误率最小的一组 σ. 典型的方法有交叉验证法.

利用交叉验证法来确定 σ 时，首先需要给定一组 $\sigma_i,i=1,2,...,n$ 的值，然后分别对每一个 σ_i 分别进行 SVM 训练，计算各自的实际风险估计的性能指标，选择性能指标最好的 σ_i 作为最终的高斯核宽度 σ. 在计算性能指标时，采用 k-折交叉验证法.

利用交叉验证法来确定 σ 时，首先需要选择一组 $\sigma_i,i=1,2,...,n$ 的值，这组 σ_i 选择的好坏不仅直接影响最终的 σ，然而在实际过程中往往只能依靠经验来确定. 交叉验证实际上就是参数空间穷尽搜索法，也就是说用枚举参数空间的每一组可能的参数去训练和测试 SVM，找出效果最好的参数. SVM 的求解是比较耗时的，当样本达到一定规模时，交叉验证法将无法计算.

3 利用核函数性质和几何距离来选择 σ

设 $x,z\in X$，X 属于 $R(n)$ 空间，非线性函数 ϕ 实现输入空间 X 到特征空间 H 的映射，其中 H 属于 $R(m)$，$n\ll m$. 根据核函数有:

$$K(x,z)=\langle\phi(x),\phi(z)\rangle$$

其中，$\langle,\rangle$ 为内积，$K(x,z)$ 为核函数. 核函数即是通过映射函数 $\phi(x)$ 把样本点从 X 特征空间映射到 H 特征空间，再进行点积运算.

高斯核函数的定义如下:

$$K\left(x^{(i)},x^{(j)}\right)=\exp\left\{-\frac{\left\|x^{(i)}-x^{(j)}\right\|^2}{2\sigma^2}\right\}\tag{2}$$

映射函数 $\phi(x)$ 和核函数 K 之间具有以下关系:

$$\begin{aligned}&\left\|\phi\left(\mathrm{x}^{(i)}\right)-\phi\left(\mathrm{x}^{(j)}\right)\right\|^2\\ &=K\left(\mathrm{x}^{(i)},\mathrm{x}^{(i)}\right)+K\left(\mathrm{x}^{(j)},\mathrm{x}^{(j)}\right)-2K\left(\mathrm{x}^{(i)},\mathrm{x}^{(j)}\right)\end{aligned}\tag{3}$$

式(3)表示样本点 $\mathrm{x}^{(i)}$ 和 $\mathrm{x}^{(j)}$ 经 $\phi(x)$ 映射后，在高维特征空间 H 上的希尔伯特空间距离的平方.

由高斯核函数的定义(2)可知:

$$K\left(x^{(i)},x^{(i)}\right)=K\left(x^{(j)},x^{(j)}\right)=1$$

因此，式(3)在高斯核下可简化为:

$$\left\|\phi\left(\mathrm{x}^{(i)}\right)-\phi\left(\mathrm{x}^{(j)}\right)\right\|^2=2-2K\left(\mathrm{x}^{(i)},\mathrm{x}^{(j)}\right)\tag{4}$$

核函数的本质是把低维空间的不可分数据映射到高维空间中，使之在高维空间上可分. 因此，我们希望样本通过 $\phi(x)$ 从特征空间 X 映射到高维特征空间 H 后，具有更好的可分性.

我们可以从几何距离上来衡量数据的可分性，进而得到最优的 σ. 我们选取希尔伯特空间距离的平方(4)来作为衡量数据可分性的标准. 我们希望：同一类别的样本的距离尽量小，不同类别的样本的距离尽量大. 数学形式化表示如下:

$$\begin{cases}\min\left\|\phi\left(x^{(i)}\right)-\phi\left(x^{(j)}\right)\right\|^2,y^{(i)}y^{(j)}=1\\ \max\left\|\phi\left(x^{(i)}\right)-\phi\left(x^{(j)}\right)\right\|^2,y^{(i)}y^{(j)}=-1\end{cases}\tag{5}$$

结合(4)和(5)，我们定义评估函数:

$$
\begin{aligned}
&L(\sigma) \\
&= 2\left(m_+ m_- - C_{m_+}^2 C_{m_-}^2\right) + 2\sum_{i=1}^{m}\sum_{j=1}^{i-1} K\left(x^{(i)}, x^{(j)}\right) * \left(y^{(i)} y^{(j)}\right)
\end{aligned} \tag{6}
$$

其中，m_+ 和 m_- 都可看作是与 σ 无关的常数，因此式(6)可简化为：

$$
L(\sigma) = \sum_{i=1}^{m}\sum_{j=1}^{i-1} K\left(x^{(i)}, x^{(j)}\right) * \left(y^{(i)} y^{(j)}\right) \tag{7}
$$

为了满足式(5)的对数据可分性的衡量标准，我们只需将式(7)的评估函数 $L(\sigma)$ 最大化. 我们选取式(7)的评估函数 $L(\sigma)$ 来作为选择高斯核函数参数 σ 的性能指标，从而高斯核函数的参数 σ 的确定问题就转化为最优化的求解问题.

4　求解最优化问题

为了表述方便，我们定义：

$$
\lambda_{ij} = \left\| x^{(i)} - x^{(j)} \right\|^2, c = -\frac{1}{2\sigma^2} \tag{8}
$$

将式(2)和式(8)代入式(7)，得到：

$$
L(\sigma) = L(c) = \sum_{i=1}^{m}\sum_{j=1}^{i-1} e^{c\lambda_{ij}} y^{(i)} y^{(j)} \tag{9}
$$

我们用麦克劳林公式对 e^x 进行展开：

$$
e^x \approx 1 + x + \frac{1}{2}x^2 \tag{10}
$$

结合(9)和(10)，得到：

$$
\begin{aligned}
L(\mathrm{c}) &= \sum_{i=1}^{m}\sum_{j=1}^{i-1}\left(1 + \lambda_{ij}\mathrm{c} + \frac{1}{2}\lambda_{ij}^2 \mathrm{c}^2\right) \mathrm{y}^{(i)} y^{(j)} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{i-1} \mathrm{y}^{(i)} y^{(j)} + \left(\sum_{i=1}^{m}\sum_{j=1}^{i-1}\lambda_{ij}\mathrm{y}^{(i)} y^{(j)}\right)\mathrm{c} + \left(\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{i-1}\lambda_{ij}^2 \mathrm{y}^{(i)} y^{(j)}\right)\mathrm{c}^2
\end{aligned} \tag{11}
$$

令 $A = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{i-1}\lambda_{ij}^2 \mathrm{y}^{(i)} y^{(j)}$，在实际分类问题中，一般有 $A < 0$.

式(11)是简单的二次函数，对 $L(c)$ 求极值，在实际应用中通常就是最大值($A < 0$). 对 $L(c)$ 求极值，对应的 c 为：

$$
c = -\frac{\sum_{i=1}^{m}\sum_{j=1}^{i-1}\lambda_{ij} y^{(i)} y^{(j)}}{\sum_{i=1}^{m}\sum_{j=1}^{i-1}\lambda_{ij}^2 y^{(i)} y^{(j)}}
$$

对于不平衡分类问题(某些类的样本数量远远少于其他类)，可能有 $A > 0$. 我们可以先采用重采样方法(如 smote)使得不平衡的样本分布变得比较平衡，再应用本文的方法来确定高斯核的参数 σ .

在极少的情况下，得到的 σ 可能为复数，此时我们取 σ 的模来作为高斯核的宽度.

下面简单分析下该方法的时间复杂度和空间复杂度.

时间复杂度：我们只需要计算 λ_{ij} 和 λ_{ij}^2，计算的时间复杂度为 $O\left(m^2/2\right)$，m 为训练样本个数.

空间复杂度：该算法不需要存储矩阵，只需要几个临时变量，空间复杂度是 $O(1)$.

由上述分析可知，用该方法确定高斯核参数 σ ，不需要涉及 SVM 的求解. 而 SVM 的求解是比较耗时的，所以当样本达到一定规模时，交叉验证法将无法计算. 该方法相比于交叉验证,在训练时间和训练空间上都小很多. 下面将在 matlab 上进行仿真实验.

5　实验及分析

实验在 UCI 的 EEG Eye State Data Set 数据集上完成. 所有的样本数据都是通过 Emotiv EEG 脑电波测量仪不间断地测量得到，测量的持续时间为 117 秒. 在脑电波测量过程中，通过摄像头检测眼部的状态，通过分析视频帧把眼部的状态作为样本标签添加到样本数据中. “1”表示眼睛闭合, “0”表示眼睛睁开.

我们需要建立一个模型，用于根据某一时刻的脑电波值来预测眼睛的状态. 样本的输入特征空间为 14 维，表示在某一时刻所测得的 14 处的脑电波值. 样本标签为 0 或 1，分别表示眼睛闭合或睁开. 这是一个典型的二分类问题，我们采用 SVM 进行训练学习，并选取高斯核作为 SVM 的核函数.

为了简化训练过程，我们选取其中的 2724 条样本数据进行训练. 为了消除量纲影响，我们对输入数据进行了归一化处理.

为了验证本文提出的方法确定的 σ 相比于交叉验证法在分类错误率相差不大的前提下具有较小计算量，我们将对两者进行实验比较. 另外，我们采用交叉验证法(5-fold)来选取惩罚因子 C . C 分别在 0.01~1 之间均匀地取 10 个值, 10~1000 之间均匀地取 100 个值.

图 1 是利用本文方法确定 σ ，分类正确率 CR 与惩罚因子 C 的对应曲线图. 从曲线图可以看出，本文方法确定的 σ 的参数敏感性相对 C 较好. 由于我们确定

的σ依赖于训练样本，采用 5-fold 交叉验证来选择惩罚因子C时，每一折对应的σ分别为 0.6092，0.5984, 0.5967, 0.6280, 0.6130. 我们得到的最终模型的分类正确率达到了 90.31%，训练时间为 1136.88s.

图 2 是利用交叉验证法(5-fold)来选择σ，分类正确率 CR 与σ的对应曲线图. σ分别在 0.01~1 之间均匀地取 10 个值, 10~1000 之间均匀地取 100 个值. 从曲线图可以看出，最优的σ大概在 0.5 到 1 之间，而利用本文方法确定的σ恰好在这个区间内. 我们得到的最终模型的分类正确率为 90.68%，训练时间为 75283.73s.

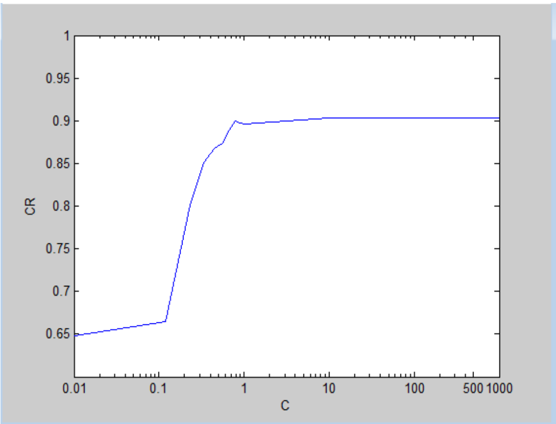

图 1 利用本文的方法确定σ，分类正确率 CR 与惩罚因子C的对应曲线图

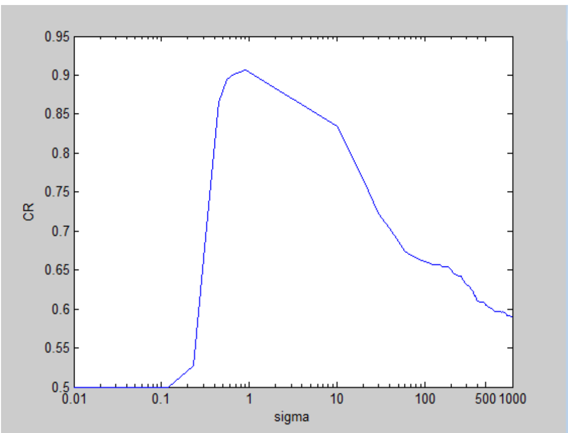

图 2 利用交叉验证法确定σ，分类正确率 CR 与σ的对应曲线图

由以上的实验结果可以得到：在分类准确率上，本文提出的方法和交叉验证法相差不大；在训练时间上，本文的方法要比交叉验证法小很多. 在保证分类准确率的基础上，本文提出的方法可以有效地降低高斯核SVM在参数选择上的时耗，并且本文提出的方法只需要常数级的空间开销.

6 结语

本文提出了从核函数性质和几何距离角度来选择高斯核函数的参数σ，并且利用高斯函数的麦克劳林展开解决了参数σ的优化选择问题. 实验结果表明，该方法能较好且较快地确定高斯核函数的参数σ.

本文虽然解决了σ的优化选择问题，但另一个参数惩罚因子C的选择仍然主要依靠经验，在将来的研究工作中需要研究惩罚因子C的选择问题.

参考文献

1 Vapnik V. Statistical Learning Theory. New York, USA. Wiley. 1998.

2 Lanckriet G, Cristianini N, Bartlett PL, et al. Learning the kernel matrix with semi-definite programming. Journal of Machine Learning Research, 2004, 5: 27–72.

3 奉国和.SVM 分类核函数及参数选择比较.计算机工程与应用,2011,47(3):123–128.

4 杨紫微,王儒敬,檀敬东,应磊,苏雅茹.基于几何判据的SVM 参数.计算机工程,2010,36(17):206–209.

5 刘向东,骆斌,陈兆乾.支持向量机最优模型选择的研究.计算机研究与发展,2005,42(2):576–581.

6 朱树先,张仁杰.支持向量机核函数选择的研究.科学技术与工程,2008,8(16):4513–4516.