基于 RBF 核的 SVM 分类应用研究

朱　芳，赵庆平，王江涛

（淮北师范大学 物理与电子信息学院，安徽 淮北 235000）

摘要：支持向量机（Support Vector Machine，SVM）在解决小样本、非线性及高维模式识别中具有优势，但核函数的选取没有定论，且其参数对 SVM 模型的性能起重要作用。针对这些问题，文章建立了基于 SVM 的分类模型，并通过 UCI 数据集验证了径向基核函数（Radial Basis Function，RBF）较其他核函数的有效性，其中核参数的选取采用改进的网格搜索法进行寻优。分类实验结果表明，选择 RBF 核函数的分类准确度较其他核函数提高了 2.5％到 35％。

关键词：支持向量机；径向基核函数；网格法；分类

中图分类号：O234；TP273^{+}.22　**文献标志码**：A　**文章编号**：1672－349X(2017)03－0013－05

DOI：10.16160/j.cnki.tsxyxb.2017.03.003

The Classification and Application of SVM Based on the RBF Kernel

ZHU Fang，ZHAO Qing-ping，WANG Jiang-tao

（School of Physics and Electrical Information，Huaibei Normal University，Huaibei 235000，China）

Abstract：The support vector machine（SVM）has some advantages in the small-sample，nonlinear and high dimensional pattern recognition，but the selection of kernel function is not conclusive，and its parameters have an important influence on the performance of the SVM model. To solve these problems，the authors of this paper established a classification model based on SVM，and verified the greater effectiveness of the radial basis function（Radial Basis Function，RBF）than that of other nuclear functions through the analysis of UCI data sets and kernel parameters were determined with the improved method of grid search. The experiment results show that the classification accuracy of RBF kernel function has been improved by 2.5％ to 35％ in comparison with other kernel functions.

Key Words：SVM；RBF；grid search method；classification

0　引言

支持向量机（Support Vector Machine，SVM）是 20 世纪 90 年代 Vapnik 在统计学习理论体系基础之上提出的思维新型机器学习方法，它利用风险最小化原则来拟合分类目标[1]。其优势是：①基于有限的目标数据获得样本分类后的最佳解，避免过学习；②不断寻找一个可行条件下的最优分类面或线，避免陷入局部最优；③数学形式简单，将不可分函数拟合到高维状态使其在高维变成可分状态，解决非线性问题[2]。在目前常用的分类算法中，基于决策树学习的分类算法，易于理解和实现，但是对于多

作者简介：朱芳（1986－），女，湖南衡阳人，讲师，硕士，主要从事信息与信号处理研究。

分类数据,容易导致过度拟合,错误增加;基于贝叶斯定理的分类算法有着稳定的数学理论,但是模型必须建立在属性相互独立的基础之上;K 近邻算法简单有效,但计算速度慢且量大;人工神经网络具有自学习、强非线性、鲁棒性、并行处理等能力,但学习速率慢、参数多且选择困难,易陷局部极小点。

核函数用于实现样本低维到高维空间的转换,是支持向量机对于非线性不可分问题的重要解决办法。尽管有研究表明不存在对所有领域的问题都具有最优泛化能力的核函数,但在某些问题上采用不同类型核函数,模型表现出不同的性能。目前对于核函数的选择或构造还没有统一的规则,只能凭借经验。一般情况下,只要满足 Mercer 条件的函数在理论上都可选为核函数。所以,只有选择合适的核函数,将数据投影到合适的高维空间,才可能得到性能更优的支持向量机模型[3-4]。相较与线性、多项式、Sigmoid 等常用核函数,RBF 函数具有较宽的收敛域,适应性更广,且只有一个超参数 σ 需要优化,计算难度和复杂度小,是较为理想的映射核函数[5]。除此之外,Brailovsky 等人还构造了全局、局部、混合及邻域等各种形式的混合核函数,但是在实际应用中还是主要以 RBF 核作为核函数[6],因为混合核又引入了权重参数,所以增加了模型中参数的选择困难。

本文提出基于 RBF 核函数的 SVM 分类模型,以两个常用的 UCI 数据集来验证模型的有效性,并将其与几个常用的核函数进行对比。

1 基于 RBF 核的 SVM 模型的建立

1.1 支持向量机的分类原理

1.1.1 线性可分

一个二维空间里仅有两类样本的分类问题,如图 1 所示。圆点和方块表示两类样本;L_1,L_2,L_3 为分类线性函数。能够将两类分开的线性函数可以有很多个,而支持向量机要求有一条最优的分类线,即分类间隔最大。如图 2 所示,位于 H_1,H_2 上的训练样本点就称作支持向量,用于支撑最优分类超平面。

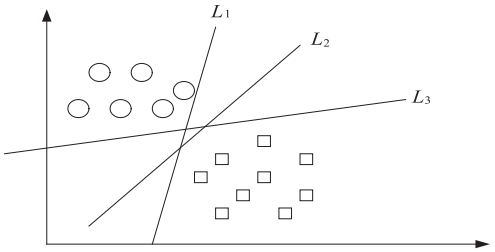

图 1 二分类示意图

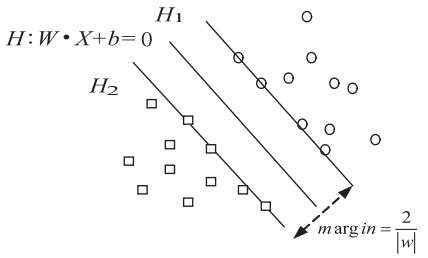

图 2 最优分类示意图

寻找最优分类面用函数表示为:

$$\min \frac{1}{2}\|w\|^2,$$

Subject to $y_i[(w \cdot x_i)+b]-1 \geqslant 0, i=1,2,\cdots,n$。 (1)

这是一个二次规划问题,将式(1)转换为无约束求解问题,引入拉格朗日乘子可得:

$$\min_{w,b,\alpha}\left\{\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i-b)-1]\right\}, \quad (2)$$

其中,$\alpha_i \neq 0$ 为样本 x_i 对应的拉格朗日系数。

对式(2)中的 w,b 分别求偏导,并等于 0,即将寻优问题转换为对偶问题,此时 $w=\sum_{i=1}^{n}\alpha_i y_i x_i$,带入到式(2)中,可得不等式二次函数寻优问题,存在唯一解。

$$\begin{cases} Q(\alpha)=\sum_{k=1}^{n}\alpha_k-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle \\ s.t \quad \sum_{i=1}^{n}\alpha_i y_i=0, 0\leqslant\alpha_i, i=1,\cdots,n \end{cases}。 \quad (3)$$

获得支持向量和相关参数后,最终可得上述问题的最优判别函数:

$$f(x)=\mathrm{sgn}[(w \cdot x)+b]=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i(x_i \cdot x)+b]。 \quad (4)$$

1.1.2 线性近似可分

线性可分的另一种情况是样本近似可分,但是存在少量的样本被错分或不可分的中间地带,此时引入松弛变量 $\zeta_i \geqslant 0, i=1,2,\cdots,n$,则

目标函数变为：

$$\begin{cases}\min\limits_{w,b,\zeta}\dfrac{1}{2}\|w\|^2+C\sum\limits_{i=1}^{n}\zeta_i \\ s.t\quad y_i[(x_i\cdot x_I)+b]\geqslant 1-\zeta_i\end{cases}, \tag{5}$$

其中 C 为惩罚因子。C 的大小决定了对离群样本的重视程度，C 越大，越重视。

1.1.3　非线性不可分

非线性不可分情况的分类线，如图 3 所示。没有线性的函数能将样本 ab 与其他线段分开。此时，只能引入一条非线性曲线：

$$g(x)=c_0+c_1x+c_2x^2。 \tag{6}$$

新建向量 y 和 a，使 $y=[1\quad x\quad x^2]^{\mathrm{T}}$，$a=[c_0\quad c_1\quad c_2]^{\mathrm{T}}$，则 $g(x)$ 转化为 $f(x)=\langle a,y\rangle$。利用映射函数可将原本二维的线性不可分问题，映射到一个四维空间而可分，关键是要找到映射方法。

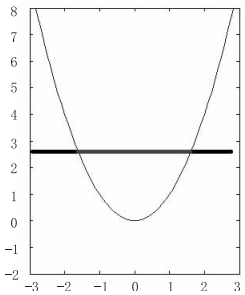

图 3　非线性不可分情况下的分类线

最终的判别函数为：

$$f(x)=\mathrm{sgn}[(w\cdot x)+b]=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b]。 \tag{7}$$

1.2　径向基核函数(RBF)的映射原理

SVM 分类应用中最重要的内容就是核函数的引入，并利用核函数免去高维变换，直接用低维度的参数代入核函数来等价高维度的向量的内积，并使不可分的模式通过非线性映射到高维空间转化为线性可分的问题[7]。其式如下：

$$K(x,z)=\langle\Phi(x),\Phi(z)\rangle, \tag{8}$$

其中〈,〉为内积，$K(x,z)$为核函数。

核函数有多种，主要有线性核函数(Linear kernel)，RBF 核函数(Radial Basis Function)，多项式核函数(Polynomial)，Sigmoid 核函数。其中，RBF 核函数是分类器中使用普遍的一种核函数，公式如下：

$$K(x_i,x_j)=\exp(\frac{-\|x_i-x_j\|}{2\sigma^2}), \tag{9}$$

其中 σ 为核参数。

1.3　建模过程

1.3.1　模型结构

利用 SVM 模型[7−9]解决分类问题，首先要从原始数据里把训练集和测试集提取出来，进行一定的数据预处理；然后使用训练集训练模型，在训练模型的过程中包括了核函数的选取和核参数的寻优，从而得到模型结构；最后使用测试集通过已训练好的模型进行分类，预测结果，将预测值与实际值进行对比得到分类准确率，以评估模型的可行性和有效性。SVM 模型建立流程如图 4 所示。

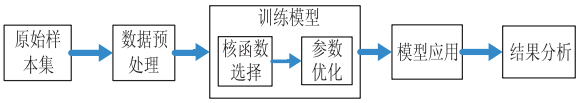

图 4　模型整体流程

1.3.2　参数的选择

RBF 核中的 σ 和 SVM 分类器中的 C 是可调节的参数，在空间区域中 (C,σ^2) 参数所取的值不同，大致分为 4 种情况：①σ^2 确切，C 太小，训练误差大，此时模型泛化能力低，属于欠训练。②σ^2 确切，$C\to\infty$，此时出现过训，模型接近硬边缘。③C 确切，$\sigma^2\to 0$，此时模型曲线趋于平坦，出现欠训练。④C 确切，$\sigma^2\to\infty$，不敏感损失带越敏感，越容易导致训练的分类器过学习现象。因此，每个数据集都存在一组最优的 (C,σ^2) 值，使得支持向量机具有较好的泛化性能[9]。

网格法[10]是将 C 和 σ^2 分别取 N 个值和 M 个值，用 $N\times M$ 个 (C,σ^2) 的组合，来测试各种分类器，然后判断其准确度，最后在网格中众多的 (C,σ^2) 样本下找到准确度最高的一组 (C,σ^2)。其缺点是网格划分得越细、范围越广，计算的工作量也就越大。其他寻找最优 (C,σ^2) 的方法，比如双线性法[9−10]，是先求解线性 SVM 中的最佳参数 C'，再获得满足 $\log\sigma=\log C-$

log C'的(C,σ^2)。与搜索法相比,双线性法运算简单但学习精度略低。而鉴于 Matlab 强大的运算能力,即使网格范围为$(2^{-10},2^{10})$,步长为 0.8,搜寻时间才只有 3.8 s。因此,本文主要采用二级网格搜索(Multiple Grid Searching, MGS)的方法来调整模型中的参数。参数寻优流程如图 5 所示。

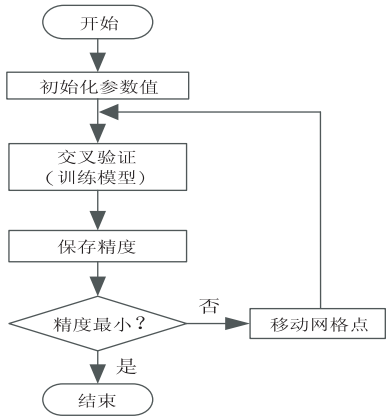

图 5 参数寻优流程图

2 基于 RBF 核的 SVM 模型验证实验

2.1 葡萄酒的二分类鉴别实验

实验仿真基于 MatlabR2010a 编程实现。

选择酒精度和苹果酸作为鉴别张裕葡萄酒和其他葡萄酒的特征指标。通过建立的 SVM 模型对不同葡萄酒中这两种指标依据含量的不同进行分类,并且找出分类准确率最高的参数(C,σ^2)。为了验证模型的准确性,输入新的指标数据集,利用已确定核参数的 SVM 模型识别相应参数的酒是张裕葡萄酒还是其他葡萄酒。

2.1.1 无参数优化的分类实验

根据经验选择固定参数组合$(C,g)=(1,0.5)$(本实验数据来源于 UCI 数据库,并进行整理,其中参数 $g=1/(2\sigma^2)$)。总样本集大小为 178 * 2,将样本集随机分为 90 个样本集作为训练集,88 个样本集作为测试集。其训练和测试所得分类准确率分别为 88.89% 和 88.64%。

2.1.2 改进网格法优化参数分类实验

本文运用二级网格法进行参数(C,σ^2)的寻优。其步骤如下:

①选定 $N\times M$ 为 15×15 的参数组合,C 取值分别为$[2^{-10},2^{-9},\cdots,2^{9},2^{10}]$,$g$ 取值分别为$[2^{-10},2^{-9},\cdots,2^{9},2^{10}]$,共 400 个$(C,\sigma^2)$组合,步长为 1.5,作为一级网格粗选范围。②使用 $K=5$ 的交叉验证法筛选出平均分类准确率大于 70%的(C,g)组合,此时 C 取值范围为$[-5,5]$,g 取值范围为$[-3,10]$,步长为 0.8,作为二级网格搜寻范围。

葡萄酒数据集的参数选择结果如图 6 所示。图中显示了葡萄酒数据集中训练集寻优过程中平均分类准确率大于 70%时所对应的不同参数组,且具有相同分类准确率的(C,g)组采用等高线表示。

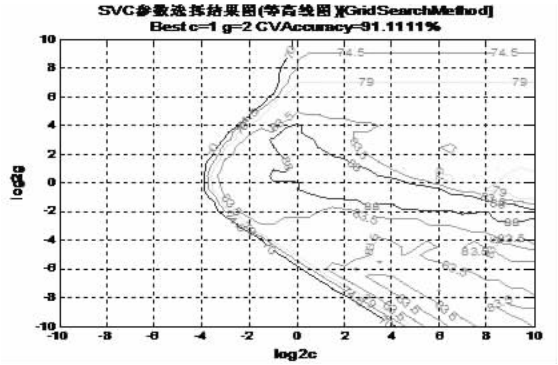

图 6 葡萄酒数据集参数选择结果图

由图 6 可知,经过网格法寻优之后,参数(C,g)从经验选择的$(1,0.5)$变为$(1,2)$,样本训练集的平均准确度从 88.89% 提高到 91.11%。因此利用网格法可以选出分类准确度更优的分类核参数。

2.1.3 不同核函数的分类实验

为了区分不同核函数对葡萄酒分类结果的影响,采用相同的样本训练集和测试集进行分类实验。由于不同的核函数需要优化的参数不同,为了减少参数带来的影响,不同核函数中的同一种参数均采用相同的值,经过网格法寻优之后 SVM 分类模型中的惩罚参数 $C=1$,RBF 核函数中的参数 $g=\frac{1}{2\sigma^2}=2$。选择几种常用核函数(线性、多项式、RBF,Sigmoid 核函数)构建 SVM 分类模型,再将测试集对应不同核函数的 SVM 分类模型进行分类预测实验。实验

仿真结果的分类准确率如表 1 所示。

表 1　葡萄酒数据集对不同核函数的分类准确率

核函数	线性核函数	多项式核函数	RBF 核函数	Sigmoid 核函数
分类准确率(%)	85.3	86.3	92.1	66.7

从表 1 中可以看出，SVM 分类器利用 RBF 核函数的准确度是 92.1%，比线性核函数的准确率高出 6.8%，比多项式核函数的准确率高出 5.8%，比 Sigmoid 核函数的准确率高出 25.4%。因此，对于二分类样本一定且小的情况下选择 RBF 核函数进行高维拟合，一定程度上提高了预测的准确度。

2.2　ZOO 数据集的多分类实验

ZOO 数据集描述了 17 种属性特征的 7 类动物，总样本大小为 101 * 17，属于多分类数据集。实验将数据集随机分为 61 个样本作为训练集，40 个样本作为测试集（本实验数据来源于 UCI 数据库网站，并进行预处理）。采用上节的网格搜索法寻找分类准确率最高，且 C 值最小的参数组合，寻优结果如图 7 所示。

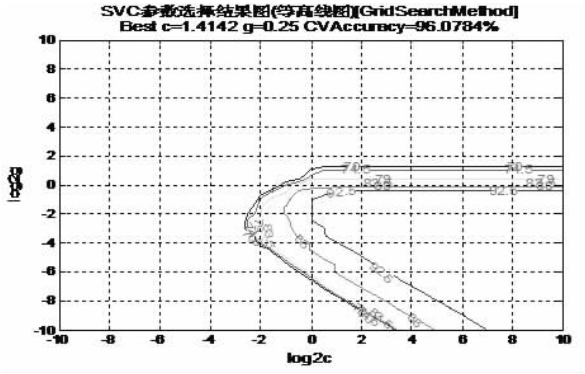

图 7　ZOO 数据集参数选择结果图

图 7 显示了 ZOO 数据集中训练集的平均分类准确率大于 70%时对应的不同参数组，图中等高线表示具有相同分类准确率时的(C,g)组。从图 7 可以看出，参数组合为(1.414 2，0.25)下的训练集平均分类准确率可达 96.078 4%，因此，选择此参数值构建分类模型。再选择不同核函数得到不同的分类模型，最后将各模型使用测试集进行分类预测。将预测结果与真实值进行对比，其分类准确率如表 2 所示。

表 2　ZOO 数据集不同核函数下的分类准确率

核函数	线性核函数	多项式核函数	RBF 核函数	Sigmoid 核函数
分类准确率(%)	80	85	87.5	52.5

从表 2 可以看出，SVM 分类器选择 RBF 核函数的准确度是 87.5%，比线性核函数的准确率高出 7.5%，比多项式核函数的准确率高出 2.5%，比 Sigmoid 核函数的准确率高出 35%。因此，对于多分类样本一定且小的情况下选择 RBF 核函数进行高维拟合，可明显提高分类的准确度。

3　实验结果分析

本文以真实的数据来源建立了基于 SVM 的分类模型，验证了 RBF 核函数较其他核函数具有更好的分类准确度，并得到以下结论：

(1)不管是二分类还是多分类，在相同样本集和参数不变的情况下，运用 RBF 核函数来进行高维拟合，较其他常用核函数的分类准确度提高了 2.5%～35%。此结果证明 RBF 核函数在 SVM 分类使用中具有明显优势。

(2)运用网格搜索法进行参数寻优，葡萄酒数据集二分类的最优参数 C 和 g 的值分别是 1 和 2，ZOO 多分类数据集的寻优结果为(1.414 2，0.25)，均能提高测试对象的分类准确率。

4　结论

对于不同的实验对象，不同核函数的适应性会有差别，并且相同实验对象选择不同的核函数，SVM 模型也会表现出不同的性能，但是总的来说，RBF 核函数相较其他核函数具有更高的适应性，灵活性也更强，还可以与其他函数混合，存在大量的可能性。除此之外，确定了某种核函数，其相应的超参数（如径向基函数中的 σ、多项式中的 d）不同，得到的模型和预测结果也不同[11]。因此，要提高 SVM 模型的推广性能，必须从这两方面进行考虑。

参考文献：

[1]　Vapnik V N. The nature of statistical learning theory[M]. New York：Springer-Verlag，1995.　（下转第 39 页）

比较，灼烧 600 ℃的催化剂对麦草畏的降解效果最好，所以，600 ℃可作为制备催化剂的最佳灼烧温度。

(2)三价铁离子和锰离子作为活性组分负载在 13X 载体上，对麦草畏的降解效果最为显著，并且铁和锰廉价易得，所以，13X 可作为最优的催化剂载体，三价铁离子和锰离子作为最佳的催化剂活性组分。

参考文献：

[1] 陆洪宇，马文成，张梁，等. 臭氧催化氧化工艺深度处理印染废水[J]. 环境工程学报，2013，7(8)，2873－2876.

[2] Hendric Nollet，Murie Roels，Pierre Lutgen，et al. Removal of PCBs from wastewater using fly ash[J]. Chemosphere，2003，53(6)：655－665.

[3] Schwarzenbach R P，Escher B I，Fenner K，et al. The challenge of micropollutants in aquatic systems[J]. Science，2006，313(5790)：1072－1077.

[4] 尚会建，张少红，赵丹，等. 分子筛催化剂的研究进展[J]. 化工进展，2011(S1)：407－410.

[5] 陈珊珊. 催化臭氧氧化工艺中粉煤灰基催化剂的制备及应用研究[D]. 苏州：苏州科技学院，2014.

(责任编校：李秀荣)

(上接第 17 页)

[2] Yendrapalli K，Mukkamala S，Sung A H，etal. Biased support vector machines and kernel methods for intrusion detection[J]. World Congress on Engineering，2007，1(2)：321－325.

[3] 王睿. 关于支持向量机参数选择方法分析[J]. 重庆师范大学学报：自然科学版，2007(2)：36－38.

[4] 张国宣，孔锐，郭立，等. 支持向量机中核函数的分类研究及组合使用[C]. 中国电子学会电路与系统学会第十八届年会论文集，2004.

[5] Smola A J，Schelkopf B. A tutorial on support vector regression[J]. Statistics and Computing，2004，14(3)：199－222.

[6] Goldsmith A J，Chua S G. Variable-power mqam for fading channels[J]. IEEE Trans. on Communications，1997，45(10)：1218－1230.

[7] 曹国超，郑刚. 用于心电波形度量及适于 K 近邻方法的核函数的选择[J]. 天津理工大学学报，2009(5)：42－45.

[8] 符欲梅，朱芳，咎昕武. 基于支持向量机的桥梁健康监测系统缺失数据填补[J]. 传感技术学报，2012，25(12)：1706－1710.

[9] 王鹏，朱小燕. 基于 RBF 核的 SVM 的模型选择及其应用[J]. 计算机工程与应用，2003，39(24)：72－73.

[10] 李琳，张晓龙. 基于 RBF 核的 SVM 学习算法的优化计算[J]. 计算机工程与应用，2006，42(29)：190－192.

[11] Deris A M，Zain A M，Sallehuddin R. Overview of support vector machine in modeling machining performances[J]. Procedia Engineering，2011，24(8)：308－312.

(责任编校：夏玉玲)