第16卷 第2期
2006年2月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 2
Feb. 2006

基于混合核函数的SVM及其应用

张 芬[1,2],陶 亮[1,2],孙 艳[1,2]

(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 电子科学与技术学院,安徽 合肥 230039)

摘 要:支持向量机可以很好地应用于函数拟合中,其中核函数的选择尤其重要。由于普通核函数各有其利弊,为了得到学习能力和泛化性能都很强的核函数,文中采用了混合核函数,并将由其构造的支持向量机运用于函数拟合中,且与普通核函数构造的支持向量机的实验结果进行了比较。结果表明其性能明显优于由普通核函数构造的支持向量机。

关键词:支持向量机;混合核函数;局部性核函数;全局性核函数

中图分类号:TP18 **文献标识码**:A **文章编号**:1005-3751(2006)02-0176-03

SVM and Its Application Based on Mixtures of Kernels

ZHANG Fen[1,2], TAO Liang[1,2], SUN Yan[1,2]

(1. Ministry of Edu. Key Lab. of Intelligent Computing and Signal Processing of Anhui Univ., Hefei 230039, China;
2. Institute of Electronic Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Support vector machine (SVM) can be used in function regression. It is important to choose an optimal kernel in order to enhance the characteristics of the SVM. Since every traditional kernel has its advantages and disadvantages for the SVM, in this paper, choose mixtures of kernels which have the desirable characteristics for SVM learning and generalization, and adopt it to function regression, then compare with the SVM using traditional kernels. The results show that the SVM performance by using mixtures of kernels is much better than that by using traditional kernels.

Key words: support vector machine; mixtures of kernels; local kernels; global kernels

0 引 言

支持向量机通过核函数定义的非线性特征映射,将待分类数据映射到一个高维的特征空间中,从而能够线性可分,然后在新特征空间中构造(广义)最优分类面,形成样本分类的决策规则。支持向量机的许多特性是由所选择的核函数来决定的,为了得到性能更为优良的支持向量机,一种改进的方法是把多个核函数组合起来,形成一种混合核函数[1],由这种混合核函数构造的支持向量机不仅学习能力强,而且具有很好的推广性。文中将这种支持向量机用于血浆脂蛋白样本与其血浆胆固醇的含量的测定中,并将结果与由其它核函数构造的支持向量机方法进行比较,意在提出一个更合适的核函数来解决函数拟合问题。

1 基于混合核函数的支持向量机

1.1 用于函数拟合的支持向量机

SVM的方法在函数拟合中表现出很好的效果[2,3],其思路与在模式识别中十分类似。这里考虑给定训练数据$\{(x_i,y_i),i=1,2,\cdots,n\}$,其中$x_i\in R^d$是第$i$个学习样本的输入值,且为一$d$维列向量$x_i=[x_i^1,x_i^2,\cdots,x_i^d]^T$,$y_i\in R$为对应的目标值。对于非线性不可分问题,通过非线性变换$\phi(\cdot)$将x映射到某个特征空间,因而转化成线性可分问题,线性估计函数可定义为:

$$y=f(x,w)=w^T\phi(x)+b \tag{1}$$

假设所有训练数据都可以以精度ε无误差地用线性函数拟合,即

$$|y-f(x)|_\varepsilon=\begin{cases}0 & |y-f(x)|\leqslant\varepsilon\\ |y-f(x)|-\varepsilon & |y-f(x)|>\varepsilon\end{cases} \tag{2}$$

则可以通过求下列代数式的最小值来获得最小风险:

$$\frac{1}{2}\|w\|^2+\frac{C}{n}\sum_{i=1}^{n}|y_i-f(x_i,w)|_\varepsilon \tag{3}$$

常数$C>0$,C表示对超出误差ε的样本的惩罚程度。采用优化方法可以得到其对偶问题。

$$\begin{cases}W(\alpha^{(*)})=-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)y_i\\ \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)\\ \text{s.t.}\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)=0;\alpha_i^{(*)}\in[0,C]\end{cases} \tag{4}$$

构造拉格朗日函数求解式(4),可得到支持向量机回

收稿日期:2005-05-27
基金项目:教育部优秀成果青年教师资助计划(教人司[2002]40号);安徽省自然科学基金项目(01042210)
作者简介:张 芬(1980—),女,安徽巢湖人,硕士研究生,研究方向为数字信号与图像处理。

归函数为:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(\boldsymbol{x},\boldsymbol{x}_i) + b \tag{5}$$

其中 $K(\boldsymbol{x},\boldsymbol{x}_i)$ 称为核函数,α_i,α_i^* 将只有小部分不为0,它们对应的样本就是支持向量。

1.2 核函数

所谓核函数就是存在一非线性变换 $\phi(\cdot)$,使 $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)$ 成立的一类函数。正是核函数的引入使SVM得以实用化,因为它避免了显示高维空间中向量内积而造成的大量运算。目前研究最多的核函数主要有3类:

1) 多项式核函数:

$$K(\boldsymbol{x},\boldsymbol{x}_i) = [(\boldsymbol{x}\cdot\boldsymbol{x}_i)+1]^q \tag{6}$$

2) 径向基核函数(RBF):

$$K(\boldsymbol{x},\boldsymbol{x}_i) = \exp(-\|\boldsymbol{x}-\boldsymbol{x}_i\|^2/\sigma^2) \tag{7}$$

3)Sigmoid 函数:

$$K(\boldsymbol{x},\boldsymbol{x}_i) = \tanh(v(\boldsymbol{x}\cdot\boldsymbol{x}_i)+c) \tag{8}$$

式(6～8)中 q,σ,c 等参数都是实常数。在实际运用中,通常要根据问题的具体情况选择合适的核函数以及相应的参数。

1.3 局部性核函数和全局性核函数

SVM的许多特性都是由所用核函数的类型决定的[1],其非线性水平是由核函数决定的。在SVM中,通常所选的核函数必须满足Mercer条件。核函数的类型有许多,解释它们各自的特性比较困难,然而,归结起来,核函数有两种主要类型,即:局部性核函数和全局性核函数。RBF函数即式(7)就是一个典型的局部性核函数。图1为当 σ 分别取0.1,0.2,0.3,0.4,0.5时RBF函数的曲线图(其中图例中的p即为 σ),0.2为测试输入,从中可以看出,局部性核函数仅仅在测试点附近小领域类对数据点有影响。而多项式核函数即式(6)是一个典型的全局性核函数。图2为当q分别取1,2,3,4,5时多项式核函数的曲线图,这里依然取0.2为测试输入,从图中可以看出,全局性核函数允许远离测试输入的数据点对核函数的值也有影响。

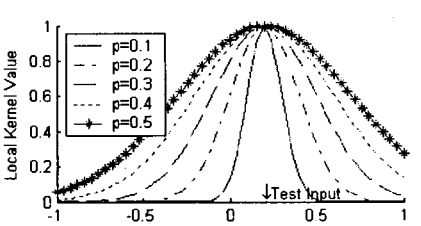

图1　RBF函数曲线图

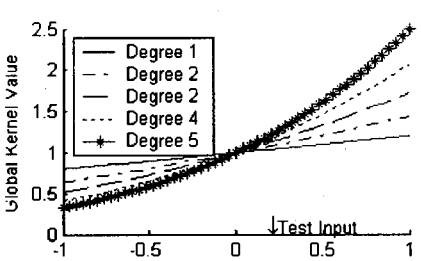

图2　多项式核函数曲线图

1.4 混合核函数

因为局部性核函数学习能力强、泛化性能较弱,而全局性核函数泛化性能强、学习能力较弱,因此考虑把这两类核函数混合起来[4]。函数 $K_{mix} = \lambda K_{poly} + (1-\lambda)K_{rbf}$ 就是混和核函数其中的一种,并且满足Mercer条件。这里还需要确定最优混合系数 $\lambda \in (0,1)$,通过实验得出 λ 一般在0.50～0.99之间,且当 λ 值较大时(例如 λ 取0.98),更能体现混合核函数的性能。为了保证混合核函数具有更好的学习能力和推广性,RBF核函数即 $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_j-\boldsymbol{x}_i\|^2/\sigma^2)$ 中 σ^2 取值宜在0.01～0.5之间;对于多项式核函数 $K(\boldsymbol{x},\boldsymbol{x}_i) = [(\boldsymbol{x}\cdot\boldsymbol{x}_i)+1]^q$,$q$ 值一般取1或2。

2 实验结果与分析

在文献[5]中曾利用264个病人血样值样本,构造了标准算法的支持向量机用于回归估计3种血浆脂蛋白(HDL,LDL和VLDL)的胆固醇含量,取得了不错的效果。文中考虑用混合核函数的SVM算法对它进行学习,并分别与各个核函数学习的结果进行了比较。

随机取数据样本的三分之一进行训练,测试时使用全部数据样本。实验中同样利用回归估计出的血样值样本对应的每种血浆脂蛋白的胆固醇含量与相应的实际含量之间的相关系数 R 来反映回归估计性能的优劣,即

$$R = \frac{\mathrm{Cov}[f(\boldsymbol{x}),\boldsymbol{y}]}{\|f(\boldsymbol{x})\|\cdot\|\boldsymbol{y}\|} \tag{9}$$

其中 $|R|\leqslant 1$,$f(\boldsymbol{x})$ 为回归估计出的血样值样本对应的每种血浆脂蛋白的胆固醇含量(264 * 1列向量),$\boldsymbol{y}$ 为相应的实际含量(264 * 1列向量)。如果存在一种理想的回归估计过程,也就是说回归估计值和相应的实际值恰好完全相同,此时 R 的值为1,也就是说,R 越接近于1,回归估计的精度就越高。

在前面的讨论中知道,对于混合核函数 $K_{mix} = \lambda K_{poly} + (1-\lambda)K_{rbf}$,$\lambda$ 一般在0.50～0.99之间,因此实验中取 $\lambda = 0.98$,取 $\varepsilon = 0.1$,$C = 1000$,同时为了保证混合核函数有更好的学习能力和推广性,对于核函数RBF,实验中 $\sigma = 0.01$ 时不同训练样本训练后进行拟合的 R 值最好,而多项式核函数 $K(\boldsymbol{x},\boldsymbol{x}_i) = [(\boldsymbol{x},\boldsymbol{x}_i)+1]^q$ 中 $q = 1$。由于 $\sigma = 0.01$ 时单个核函数的结果很差,为了更好地体现混合核函数的效果,实验中对采用不同训练样本训练后进行回归的 R 值的进行了比较,并与单个核函数的最好结果进行了比较,如表1所示。从表1中可以看出采用混合核函数进行函数拟合的效果更好,尤其是VLDL的精确度得到了大幅度的提高。

表2是混合核函数在 $\lambda = 0.98$,$q = 1$,σ 取不同值时进行拟合的结果,由表2可以看出 σ 在0.01～0.5之间时效果较好。

表 1 用不同训练样本训练后进行拟合的 R 值的比较

	训练样本取法	多项式核函数	rbf 核函数	混合核函数
HDL	隔值取法 1	0.903	0.737	0.92
	隔值取法 2	0.886	0.719	0.908
	前三分之一	0.883	0.661	0.893
LDL	隔值取法 1	0.886	0.703	0.898
	隔值取法 2	0.823	0.703	0.879
	前三分之一	0.816	0.613	0.859
VLDL	隔值取法 1	0.51	0.594	0.685
	隔值取法 2	0.532	0.547	0.59
	前三分之一	0.437	0.589	0.644

表 2 不同 σ 参数值下进行回归的 R 值的比较

σ	HDL	LDL	VLDL
0.01	0.92	0.898	0.685
0.1	0.92	0.898	0.683
0.2	0.923	0.898	0.675
0.5	0.929	0.886	0.601
1	0.895	0.856	0.435
2	0.874	0.859	0.26
5	0.57	0.582	0.0861

表 3 不同的 C 值下平均训练时间的比较

C 值	∞	10^7	10^5	10^3	10^2
平均训练时间	11.1	3.6	2.9	2.5	2.4

不仅如此，实验中还发现，C 值的选取影响训练时间的长短，C 值越小平均训练时间越短，但当 C 值过小（$C \leqslant 10^2$）时，实验结果的精度会下降。表 3 是 C 取不同值时平均训练时间的值，因此本实验中采用 $C = 1000$ 是较合理的，并且有效提高了 SVM 方法的训练速度。以上实验均利用 Matlab 6.1 编程，运行于 Pentium Ⅳ/2G，256M 内存 PC。

3 结束语

文中简要介绍了由混合核函数构造的支持向量机，并将其运用于函数拟合中。通过对 3 种不同类别血浆脂蛋白样本与其血浆胆固醇的含量的测定，验证了选择这种混合核函数的实验具有很好的效果，实验中 VLDL 的精确度有明显提高，而且本实验中训练时间只有 2.5 秒左右，很好地解决了训练速度慢的问题。当然，还可以考虑用其它的核函数来进行混合，形成不同的混合核函数，譬如将两个或多个 RBF 核函数混合，或者将两个或多个多项式核函数混合形成混合核函数，或许会得到更好的效果，以便找到选择最优核函数的某些规律，这是以后值得研究的一个课题。

参考文献：

[1] Smits G F, Jordaan E M. Improved SVM Regression using Mixtures of Kernels[A]. Proceedings of the 2002 International Joint Conference on Neural Networks[C]. Hawaii: IEEE, 2002.2785 - 2790.

[2] 边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，1999.

[3] Zhang Li, Zhou Weida, Jiao Licheng. Wavelet Support Vector Machine[J]. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 2004, 34(1): 34 - 39.

[4] Zhang Sheng, Liu Jian, Tian Jin - wen. An SVM - based Small Target Segmentationand Clustering Approach[A]. Proceedings of the Third International Conference on Machine Learning and Cybernetics[C]. Shanghai: IEEE, 2004. 3318 - 3323.

[5] 丁 蕾，陶 亮. 支持向量机在胆固醇测定中的应用[J]. 安徽大学学报(自然科学版)，2005(2): 60 - 63.

(上接第 89 页)

Select association_type. association_type_id, 'role_type_' & association_type. role_type_1 as role_1, associaton. member_1 as member_1, 'role_type_' &associaton_type. role_type_2 as role_2, association. member_2 as member_2 from association_type, association where association_type. name = association. instanceOf

4 结束语

主题地图是一个新兴的 ISO 标准，它提供了一种用于组织信息的系统。文中首先提出了主题地图的概念，描述了主题地图的 3 要素：主题、联系和事件的含义，并介绍了其描述语言 XTM。最后，将主题地图的概念方法应用到扬州宝军电子有限公司的信息系统扩建项目中，对其现有数据库，设计了一种扩展关系数据库，实现多个关系数据库之间数据交换的方法。

随着人们对主题地图的不断认识了解，相信这样一种数据描述与组织的技术，必将在更加广泛的领域中得以应用。

参考文献：

[1] 秦铁辉，郭延吉，孙 琳. 信息时代的"全球定位系统"——主题地图[J]. 江西图书馆学刊，2005, 35(1): 1 - 3.

[2] 张佩云，吴 江，贾 晖. 主题地图标准及其应用研究[J]. 安徽大学学报(自然科学版)，2004, 28(3): 19 - 22.

[3] Pepper S. Chief Strategy Officer The TAO of Topic Maps [EB/OL]. http://www. ontopia. net/topicmaps/materials/tao. html, 2002.

[4] Garshol L M. What Are Topic Maps? [EB/OL]. http://www. xml. com/pub/a/2002/09/11/topicmaps. html, 2002.

[5] TopicMap. Org XML Topic Maps (XTM) 1.0. [EB/OL] http://www. topicmaps. org/xtm/index. html, 2000.