基于正交多项式核函数方法

赵金伟[1],冯博琴[1],闫桂荣[2]

(1.西安交通大学 电子与信息工程学院 计算机科学与技术系,陕西 西安 710049;
2.西安交通大学 机械结构强度与振动国家重点实验室,陕西 西安 710049)

摘 要:针对实际工程中小样本数据的稀疏性、分布特征不明显等问题,分析了现有的一些方法并指出了现有方法存在的问题,重点讨论了一类基于切比雪夫多项式的核方法。由于切比雪夫多项式的正交性,使得这些核函数在高维特征空间能得到更优的超平面。通过实验测试了这一类核函数的泛化性能以及学习效率。证明它们比其它的核函数需要更少的支持向量并能保证更好的学习性能。最后论文讨论了这类核函数方法存在的问题,并指出切比雪夫多项式核函数在解决小样本回归问题时具有很大的潜力,值得进一步研究。

关键词:切比雪夫多项式;回归问题;小样本;泛化能力

中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2012)05-0177-03

Review of Chebyshev Kernel Functions

ZHAO Jin-wei[1],FENG Bo-qin[1],YAN Gui-rong[2]

(1. Department of Computer Science,School of Elec. and Info. Engineering,Xi'an Jiaotong University,
Xi'an 710049,China;
2. State Key Lab of Strength and Vibration of Mechanical Structures,Xi'an Jiaotong University,
Xi'an 710049,China)

Abstract:In practical engineering,small-scale data sets are usually sparse and contaminated by noise. Analyze some new methods and their problem. Furthermore,discuss the Chebyshev kernel functions which were proposed recently. Because of the orthogonality of Chebyshev polynomials,the new kernels can find the best hyperplane in the feature space. To evaluate the performance of the new kernels,applied it to learn some benchmark data sets,and compared them with other conventional SVM kernels. The experiment results show that the Chebyshev kernels have excellent generalization performance and prediction accuracy,and do not cost much less support vectors compared with other kernels. Point out the problem of the new kernels and the research direction.

Key words:Chebyshev polynomials;regression problem;small scale data set;generalization performance

0 引 言

目前,随着科学技术的发展,利用支持向量机学习算法解决有限样本的分类、回归和概率密度估计问题的应用越来越广泛,如机器故障诊断、力学环境预测、地震预报等。因为它基于VC维理论和结构风险最小化原则,所以较好地解决了小样本、非线性、高维数和局部极小点等实际问题[1]。

然而,在工程中常常会遇到这样的小样本数据。这些数据具有以下特点:数量比较少、分布稀疏,不能完全有效地覆盖样本空间,实测数据噪声的影响较大等。例如一些重要的工程领域中的测量数据,如火箭、导弹的遥测数据,医学领域中的疫苗人体实验数据等。这些数据若不能很好地反映样本空间的分布特征,那么它们必定会影响支持向量机的学习效果。对于这种小样本问题,通常的做法是改进算法结构,如最小二乘支持向量机(LSSVM, Least Squares Support Vector Machine)[2]、近似支持向量机(PSVM, Proximal Support Vector Machine)[3]、模糊支持向量机(FSVM, Fuzzy Support Vector Machine)[4]、概率支持向量机(PSVM, Possibilistic Support Vector Macines)[5]等。另一种做法是寻找更好的核函数或核函数的组合,并优化它,来提高算法的泛化性能。于是文献[6~10]将各种不同的核组合在一起,利用自适应优化策略来控制参数,对于不同的训练数据集给出不同的核函数组合。由于核参数比较多,所以必须要有一种高效的寻参机制和核选择机制,这必定会导致学习算法效率

收稿日期:2011-10-17;**修回日期**:2012-01-18

基金项目:国家自然科学基金(61173040)

作者简介:赵金伟(1974-),男,辽宁开源人,博士研究生,从事机器学习、模式识别研究;冯博琴,教授,博士生导师,研究方向为编译理论;闫桂荣,教授,博士生导师,研究方向为控制科学与工程、机器学习。

下降。

1 切比雪夫正交多项式

切比雪夫多项式是一种高等超越函数。它是以递归方式定义的一系列正交多项式序列。常见的是第一类切比雪夫多项式和第二类切比雪夫多项式。它们来自于方程(1)和方程(2)。

$$(1-x^2)y'' - xy' + n^2y = 0 \tag{1}$$

$$(1-x^2)y'' - 3xy' + n(n+2)y = 0 \tag{2}$$

令 $x=\cos(\theta)$,则方程(1)的解为式(3),称为第一类切比雪夫多项式;方程(2)的解为式(4),称为第二类切比雪夫多项式。

$$T_n(\cos(\theta)) = \cos(n\theta)\ ,\ n=0,1,2,3,\cdots \tag{3}$$

$$U_n(\cos(\theta)) = \frac{\sin((n+1)\theta)}{\sin\theta}\ ,\ n=0,1,2,3,\cdots \tag{4}$$

其中,第一类切比雪夫多项式的母函数为 $\frac{1-xt}{1-2xt+t^2}$,展开可得(5)式。

$$\frac{1-xt}{1-2xt+t^2} = \sum_{n=0}^{\infty} T_n(x)t^n, |x| \leqslant 1, |t| \leqslant 1 \tag{5}$$

第二类切比雪夫多项式的母函数为 $\frac{1}{1-2xt+t^2}$,展开可得(6)式。

$$\frac{1}{1-2xt+t^2} = \sum_{n=0}^{\infty} U_n(x)t^n, |x| \leqslant 1, |t| \leqslant 1 \tag{6}$$

两类切比雪夫多项式分别关于权函数 $\frac{1}{\sqrt{1-x^2}}$ 和 $\sqrt{1-x^2}$ 正交。数学形式如下。

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}\mathrm{d}x = \begin{cases} 0(m \neq n) \\ \frac{\pi}{2}(m=n\neq 0) \\ \pi(m=n=0) \end{cases} \tag{7}$$

$$\int_{-1}^{1} \sqrt{1-x^2}U_m(x)U_n(x)\mathrm{d}x = \begin{cases} \frac{\pi}{2}(m=n\neq 0) \\ 0(m\neq n,\text{或 } m=n=0) \end{cases} \tag{8}$$

2 切比雪夫正交多项式核函数

切比雪夫多项式是一组非常重要的正交函数,理论上已被证明它能很好地逼近任何一条曲线。正是这一特点,引起了国内外一些学者的关注。

2.1 正交切比雪夫多项式核

业宁在第一类切比雪夫多项式的基础上构造了正交切比雪夫多项式核函数[11]。假设多项式的最大阶数为 n 。

核函数的一维形式为:

$$K(x,y) = \frac{\sum_{i=0}^{n} T_i(x)T_j(y)}{\sqrt{1-xy}}\text{。}$$

向量形式为:

$$K(x,y) = \prod_{j=1}^{m} \frac{\sum_{i=0}^{n} T_i(x_j)T_i(y_j)}{\sqrt{1-x_jy_j}}$$,其中 m 表示向量的维数。

若 $n=3$,则该核函数一维变量形式为:

$$K(x,y) = \frac{1+xy+(2x^2-1)(2y^2-1)}{\sqrt{1-xy}} \tag{9}$$

该核函数向量形式为:

$$K(x,y) = \prod_{j=1}^{m} \frac{1+x_jy_j+(2x_j^2-1)(2y_j^2-1)}{\sqrt{1-x_jy_j}} \tag{10}$$

2.2 泛化切比雪夫多项式核函数

Sedat 基于第一类切比雪夫多项式,将自变量 x 为向量 $\boldsymbol{x}$ 定义了一种泛化的切比雪夫多项式[12,13]。这时,其递推公式变为:

$$T_0(\boldsymbol{x}) = 1$$

$$T_1(\boldsymbol{x}) = \boldsymbol{x}$$

$$T_n(\boldsymbol{x}) = 2\boldsymbol{x}T_{n-1}^T(\boldsymbol{x}) - T_{n-2}(\boldsymbol{x}), n=2,3,\cdots$$

基于此,构造了泛化切比雪夫多项式核函数:

$$K(\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{j=0}^{n} T_j(\boldsymbol{x})T_j^T(\boldsymbol{z})}{\sqrt{m-\langle \boldsymbol{x},\boldsymbol{z}\rangle}} \tag{11}$$

其中 m 表示向量的维数。

在等式(11)中为了保证分母为非负,将分母中的 1 变为维数 m,并不会影响学习效果。

当时,核函数为:

$$K(\boldsymbol{x},\boldsymbol{z}) = \frac{1+\langle \boldsymbol{x},\boldsymbol{z}\rangle + (2\langle \boldsymbol{x},\boldsymbol{x}\rangle - 1)(2\langle \boldsymbol{z},\boldsymbol{z}\rangle - 1)}{\sqrt{m-\langle \boldsymbol{x},\boldsymbol{z}\rangle}} + \frac{\langle \boldsymbol{x},\boldsymbol{z}\rangle(4\langle \boldsymbol{x},\boldsymbol{x}\rangle - 3)(4\langle \boldsymbol{z},\boldsymbol{z}\rangle - 3)}{\sqrt{m-\langle \boldsymbol{x},\boldsymbol{z}\rangle}} \tag{12}$$

2.3 修正的切比雪夫多项式核函数

由于指数函数比平方根形式的函数衰减更快,Sedat 以类似于高斯核的指数形式的权函数 $\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{z}\|^2)$ 来替换泛化的切比雪夫多项式核函数的平方根形式的权函数 $\sqrt{m-\langle \boldsymbol{x},\boldsymbol{z}\rangle}$,得到了一种非线性形式的核函数,以便更好更快地捕获非线性超平面[14]。其核函数形式为:

$$K(\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{j=0}^{n} T_j(\boldsymbol{x})T_j^T(\boldsymbol{z})}{\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{z}\|^2)} \tag{13}$$

其中 γ 为衰减因子。

3 实验分析

本节以 UCI 的 Servo 数据集和 STRD 中的 Longley、Misra1c 和 MGH10 数据集为例,比较正交切比雪

夫多项式核(OCK)[11]、泛化切比雪夫多项式核(GCK)[12,13]、改进的切比雪夫多项式核(MCK)[14]、径向基核(RBF)、sigmoid 核(SK)、普通多项式核(GPK)、线性核(LK)和小波核(WLK)的泛化性能。其中 Servo 数据集有 4 个属性,总样本数为 167 个,取前 30 行作为实验数据。Longley 数据集有 6 个属性,总样本数据为 17 个。Misra1c 数据集有 1 个属性,总样本数为 14 个。MGH10 数据集有 1 个属性,总样本数为 17 个。实验前,对所有数据进行归一化处理。

每组实验做 10 次 10 折交叉验证,求出平均预测均方差值(MSE 值)和相关性系数(SCC 值),进行比较。其中相关性系数的计算公式如式(14)所示。

其中 l 表示样本个数,$f(x_i)$ 表示预测结果,y_i 表示真实结果。相关系数 SCC 表示了预测结果与真实结果的相关程度,也是泛化能力的表征量。相关系数 SCC 的值越大,表示预测结果与真实结果越相关,说明回归模型的泛化能力越强。为比较各种核函数的性能,以传统支持向量机 ε - SVM 算法为例,其源程序使用 Libsvm[15]。ε - SVM 算法利用 PSO 算法和十折交叉验证进行参数优化。结果如表 1 和表 2 所示。其中 NSV 表示支持向量个数。

从表 1 和表 2 中可以发现,这三种切比雪夫多项式核的 MSE 值较小,而 SCC 值较其它核函数的大。还可以看到,这三种核函数的支持向量较少。说明这三种核函数在解决小样本回归问题时,具很高的泛化性能。使得支持向量机的算法效率得到了提高。预测结果与真实结果更相关。

4 结束语

文中对最近提出的三种切比雪夫正交多项式核函数进行了研究与讨论,发现这三种核函数在小样本回归问题中具有较高的泛化性能。

然而,这类核函数还存在以下值得研究的方向:

1)这些核函数只是针对第一类切比雪夫多项式,然而发现用第二类切比雪夫多项式在进行回归分析时,效果也很好,所以若能考虑第二类切比雪夫多项式核函数也是一个很好的选择。

2)通过实验,发现第二类切比雪夫多项式与第一类切比雪夫多项式在进行回归分析时,特点不同。如何在二者之间进行选择,以及如何评价这两种多项式在针对具体数据集时的泛化性能,也是一个值得研究的问题。

3)权函数的选择和核函数的构造形式,不但需要其能保证正交,还能保证满足 Mercer 条件也是很有挑战性的研究工作。

总之,切比雪夫多项式核函数在解决小样本回归问题时,具有很大的潜力。还有很多值得研究与发展的东西。

$$\mathrm{SCC} = \frac{\left(l \times \sum_{i=1}^{l} (f(x_i) y_i) - \sum_{i=1}^{l} f(x_i) \sum_{i=1}^{l} y_i\right)^2}{\left(\left(l \times \sum_{i=1}^{l} (f(x_i) f(x_i)) - \sum_{i=1}^{l} f(x_i) \sum_{i=1}^{l} f(x_i)\right)\left(l \times \sum_{i=1}^{l} y_i y_i - \sum_{i=1}^{l} y_i \sum_{i=1}^{l} y_i\right)\right)} \tag{14}$$

表 1　不同的核函数和算法在各种标准数据集上的性能比较(1)

核函数	RFB			LK			GPK			SK		
评价	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV
Servo	2.99	92.639	10	16.517	61.14	22	3.16	92.6	24	22.521	35.824	23
Longley	15.86	0. 24	11	9.1	96	8	10.1	85.3	13	8.155	88	8
Misra1c	7.603	66.	7	1.582	99.978	6	5.12	70.3	10	1.19	100	6
MGH10	4.61E-02	99.9	12	5.38	83.027	8	0. 2919	99.1916	14	7.656	76.411	8

表 2　不同的核函数和算法在各种标准数据集上的性能比较(2)

核函数	WLK			OCK			GCK			MCK		
评价	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV	MSE(%)	SCC(%)	NSV
Servo	3.356	91.842	12	3.157	92.495	14	2.8457	93.2758	12	2.7376	93.	11
Longley	0. 3328	99	10	0.6849	91.1	11	1.0038	84.93	10	0.5052	98.78	12
Misra1c	0	100	7	5.6113	84.82	6	0. 181	100	7	0	100	8
MGH10	0. 626	98	5	0	100	11	0	100	13	0.4899	98.58	5

(下转第 184 页)

较大的影响。当水平分辨率提高时,会模拟出低分辨率下所没有呈现的风场的垂直速度区域,并且整体上对风场和气压场的描述更加细致精确。

文中的目的在于针对分辨率的不同分析数值模式对风场的模拟效果,而若要分析分辨率的不同与实际大气状况之间的差别,以便得出适合该区域或者某种天气过程的最优分辨率,则需要与实际观测数据进行对比分析,这是下一步的工作方向。

参考文献:

[1] 苑海燕,杜继稳,侯建忠,等."神舟六号"飞船着陆时段主着陆场区风场的数值模拟[J].气象科学,2008,28(1):56-61.

[2] 孙 宁.采用WRF模式模拟风场分析中尺度暴雨[J].科技信息,2007(17):41-43.

[3] Liu Yubao, Warner T. Simultaneous nested modeling from the synoptic scale to the LES scale for wind energy applications [J]. Journal of Wind Engineering and Industrial Aerodynamics,2011,99(4):308-319.

[4] 王澄海,胡 菊,靳双龙,等.中尺度WRF模式在西北西部地区低层风场模拟中的应用和检验[J].干旱气象,2011,29(2):161-167.

[5] Rife D L. Verification of Temporal Variations in Mesoscale Numerical Wind Forecasts[J]. Monthly Weather Review,2005,133:3368-3381.

[6] Rife D L, Davis C A, Knievel J C. Temporal Changes in Wind as Objects for Evaluating Mesoscale Numerical Weather Prediction[J]. Weather and Forecasting, 2009,24:1374-1389.

[7] Rife D L, Davis C A, Liu Yubao. Predictability of Low-level Winds by Mesoscale Meteorological Models [J]. Monthly Weather Review,2004,132:2553-2569.

[8] Valari M, Menut L. Does an Increase in Air Quality Models' Resolution Bring Surface Ozone Concentrations Closer to Reality [J]. Journal of Atmospheric and Oceanic Technology, 2008,25(11):1955-1968.

[9] Salvador R, Calbo J, Millan M M. Horizontal Grid Size Selection and Its Influence on Mesoscale Model Simulations[J]. Journal of Applied Meteorology,1999,38(9):1311-1329.

[10] 张 宇,郭振海,张文煜,等.中尺度模式不同分辨率下大气多尺度特征模拟能力分析[J].大气科学,2010,34(3):653-660.

[11] 吕光辉,于恩涛,向伟玲,等.WRF模式分辨率对新疆异常降雨天气要素模拟的影响[J].气候与环境研究,2009,14(1):85-96.

[12] 刘 峰,刘式达,文丹青.广州白云机场"721"低空风切变天气过程综合分析[J].北京大学学报(自然科学版),2007,43(1):23-29.

(上接第179页)

参考文献:

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag,1995.

[2] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters,1999(9):293-300.

[3] Fung G, Mangasarian O L. Proximal support vector machine classifiers[C]//Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2001:77-86.

[4] Huang H P, Liu Y H. Fuzzy support vector machines for pattern recognition and data mining[J]. International Journal of Fuzzy Systems,2002(4):826-835.

[5] Lee K Y, Dae-Won K. Possibilistic support vector machines [J]. Pattern Recognition,2005,38(8):1325-1327.

[6] Rostamizadeh A. Theoretical Foundations and Algorithms for Learning with Multiple Kernels[D]. New York:New York University,2010.

[7] Kloft M, Brefeld U, Sonnenburg S, et al. Non-sparse Regularization for Multiple Kernel Learning[R]. USA:Cornell University,2010.

[8] Cortes C, Mohri M, Rostamizadeh A. Two-stage Learning Kernel Algorithms [C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel:[s. n.], 2010.

[9] Cristianini N, Kandola J, Elisseeff A, et al. On Kernel Target Alignment[J]. Advances in Neural Information Processing Systems,2002(14):367-373.

[10] Kandola J, Shawe-Taylor J, Cristianini N. Optimizing Kernel Alignment over Combinations of Kernels[R]. London:Department of Computer Science, Royal Holloway, University of London, UK,2002.

[11] Ye N, Sun R, Liu Y, et al. Support vector machine with orthogonal Chebyshev kernel[C]//Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06). Hong Kong:[s. n.],2006.

[12] Ozer S, Chen C H. Generalized Chebyshev Kernels for Support Vector Classification [C]//Proceedings of the 19th international conference on pattern recognition (ICPR'08). Tampa, Florida, USA:[s. n.],2008.

[13] Ozer S. On the Classification Performance of Support Vector Machines Using Chebyshev Kernel Functions[D]. Dartmouth, MA:University of Massachusetts,2007.

[14] Ozer S, Chen C H, Cirpan H A. A set of new Chebyshev kernel functions for support vector machine pattern classification[J]. Pattern Recognition,2011,44(7):1435-1447.

[15] Chang Chih-Chung, Lin Chih-Jen. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011(3):1-27.