文章编号：1007-1423(2016)19-0006-03　　DOI：10.3969/j.issn.1007-1423.2016.19.002

# 一种基于高斯分布的 SVM 回归方法

郭金玲

（山西大学商务学院信息学院，太原　030031）

**摘要：**
核函数参数选择是支撑向量机(SVM)研究的主要问题之一。提出检验样本是否呈高斯分布的方法，确定最优核参数选择的依据，采用两组数据集分别进行回归实验，验证所提出方法的有效性。

**关键词：**
支撑向量机；回归；高斯分布

**基金项目：**
山西省科技厅自然科学基金资助项目(No.2014011018-1)、山西大学商务学院院基金(No.2015009)

## 0　引言

支撑向量机是上世纪 90 年代 V. Vapnik 提出的一种机器学习方法，该方法可用于解决大数据领域中的单分类、多分类以及预测问题等[1-3]。许多学者将该技术应用于空气监测、金融评测、医学分析、地质勘查等实际问题的解决过程中。胡世前等利用 SVM 构建了预测精度较高、有效检测大气质量的预警系统，实验表明该预警系统的高效性[4]。蔡丹莉等利用 SVM 技术，结合蛋白质特征，对蛋白质相互药理作用及性能影响进行了高效预测[5]。王奉伟等在分析了大坝变形有关数据的特定规律基础上，利用 SVM 方法实现了对大坝变形的高精度、多尺度预测[6]。

SVM 方法通过引入核函数，将样本映射到高维空间实现预测及分类，其预测最优化过程可描述为：

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

$$s.t. \quad y_i[w.x-b] \geqslant 1 \quad i=1,2,3,\cdots,l \tag{1}$$

经过转化处理，采用最小二乘法求得 $a$ 和 $b$ 的值，回归函数如下：

$$f(x)=\sum_{i=1}^{l} \sigma_i k(x_i,x)+b \tag{2}$$

由于高斯核具备计算量小、泛化性能高等优点，目前被广泛应用于 SVM 分类及预测模型中[7-8]。在实际问题的解决过程中，核函数的参数选取是最为关键的，而核参数的选取是一直以来的研究热点。本文探讨了在样本基本符合高斯分布时，如何高效正确选取核参数的过程，实验结果证明该方法的有效性。

## 1　实验样本集

文中选取了两组样本集进行实验，样本集 D1 是人工构造的高斯分布数据集，具体分布见图 1；样本集 D2 呈不规则分布，具体见图 2。

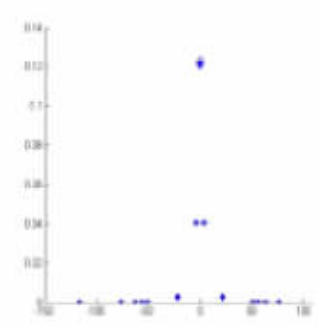

图 1 高斯分布数据集 D1

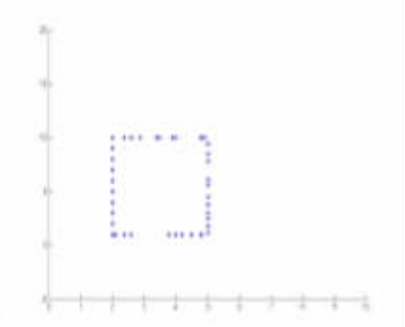

图 2 不规则分布数据集 D2

## 2 核参数选取方法

对于实验样本集$\{(x_1,y_1),\cdots,(x_l,y_l)\}$,采用以下算法检验其是否呈高斯分布,具体过程如下[9-10]:

步骤 1:取 $m=n$,将实轴分为 $n+1$ 个区间;

步骤 2:采用极大似然法计算出 $\alpha,\sigma$ 的估计值;

步骤 3:计算出统计量 $v'$

$$v'=\sum_{i=1}^{n=1}\frac{(v_i-lp_i)^2}{lp_i}$$

步骤 4:若 $v'$ 近似服从 $x^2$ 分布,则断定该样本集呈高斯分布,同时在以上判断过程中,可计算出形状分布参数。

结论:如果实验样本集基本呈高斯分布,采用高斯核进行回归实验时,其最优核参数可以选取样本集的形状分布参数。

## 3 数值实验

采用文中的方法对样本集 D1、D2 分别检测,通过以上四个步骤的计算,可得到结论:D1 呈高斯分布,且形状参数为 0.7;D2 不呈高斯分布。分别采用高斯核 SVM 和多项式核 SVM 对 D1 和 D2 进行回归实验,采用不同核函数参数进行多次回归实验,具体实验结果图见图 3、图 4、图 5、图 6、图 7 及图 8。

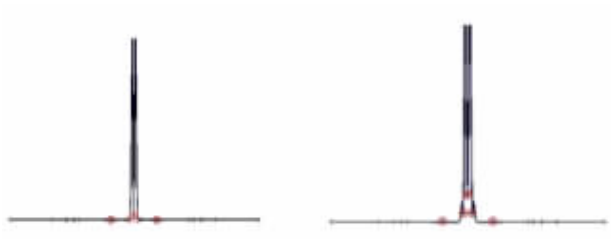

图 3 D1 回归图(高斯核,σ=0.7)　图 4 D1 回归图(高斯核,σ=1)

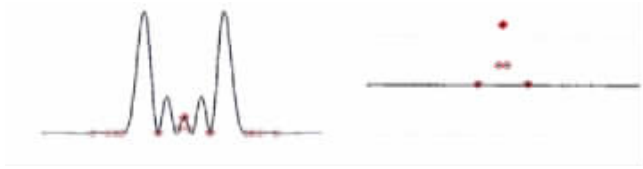

图 5 D1 回归图
(高斯核,σ=10)



图 6 D1 回归图
(多项式核,d=2)

图 7 D2 回归图
(高斯核,σ=1)

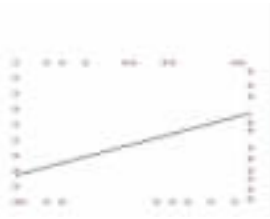

图 8 D2 回归图
(多项式核,d=3)

通过比较以上数值实验,可以看到:数据集呈高斯分布时,采用高斯核 SVM,且核参数和其形状参数一致时,回归效果最好,拟合度最高且支持向量个数较少。

## 4 结语

本文探讨了数据集呈高斯分布时,如何高效选取核函数及参数的过程。首先给出了判断数据呈高斯分布的方法,采用人工构造的数据集进行了数值实验,实验结果表明文中提出的方法的正确性及有效性。


**参考文献:**

[1]W.J. Wang , Z.B. Xu , W.Z. Lu , X.Y. Zhang. Determination of the Spread Parameter in the Gaussian Kernel for Classification and Regression[J]. Neurocomputing, 2003, 55:643-663.

[2]K.B. Duan, S. Keethi , A.N. Poo Evaluation of Simple Performance Measure for Tuning SVM Hyperparameters[J]. Neurocomputing, 2003, 51: 41-59.

[3]V. Cherkassky, Y.Q. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression[J]. Neural Networks, 2004, 17:113-126.

[4]胡世前, 姜倩雯, 凌冰, 尹伟东. 基于改进支持向量机的空气质量监测预警模型[J]. 江苏大学学报(自然科学版),2016,34(4):38-42.

[5]蔡丹莉, 郭 红. 基于混合核函数 SVM 的蛋白质相互作用预测方法[J]. 福州大学学报(自然科学版),2014,42(6):834-840.
[6]王奉伟, 周世健, 周清, 池其才. 局部均值分解结合支持向量回归的大坝变形预测[J]. 测绘科学, 2016,34(3):42-47.
[7]B. Krawczyk, M. Wozniak, F. Herrera. On the Usefulness of One-Class Classifier Ensembles for Decomposition of Multi-Class Problems[J]. Pattern Recognition, 2015, 48(12): 3969-3982.
[8]Wang Xiao-ming,Chung F L,Wang Shi-tong. Theoretical Analysis for Solution of Support Vector Data Description[J]. Neural Networks,2011,24(4): 360-369.
[9]A.T.Walden. NonGaussian, Reflectivity, Entropy, and Reconvolution[J]. Geophysics, 2011,50(12):2862-2888.
[10]赵倩,李宏伟等. 一种产生广义高斯分布随机数的算法[J]. 应用数学,2010,5: 64-69.

**作者简介:**
郭金玲(1982-),女,山西长子人,硕士研究生,讲师,研究方向为机器学习与数据挖掘
收稿日期:2016-06-25　　修稿日期:2016-07-01

# A Kind of SVM Regression Method Based on Gaussian Distribution

GUO Jin-ling

(School of Information, Business College of Shanxi University, Taiyuan Shanxi 030031)

**Abstract:**
The kernel parameter selection is one of the key problems for support vector machine (SVM). Presented a new way to select the kernel function and its parameter, it is based on the characteristics of data distribution. Presents an approach to determine Gauss distribution, and then on the basis of determining Gauss distribution, discusses how to select the kernel function and its parameter. The simulation experiments demonstrate the feasibility and the effectiveness of the presented approach.
**Keywords:**
Support Vector Machine; Regression; Gauss Distribution