

# 基于数据依赖核支持向量机回归的风速预测模型

王定成,倪郁佳,陈北京,曹智丽

(南京信息工程大学计算机与软件学院,江苏 南京 210044)

**[摘要]** 针对风速随机性大、影响因素多、预测准确度不高的情况,基于支持向量机与信息几何的统计学关联性,从信息几何学角度分析核函数的几何结构,构造数据依赖核函数,并与支持向量机回归相结合,形成数据依赖核支持向量机回归(Data Dependent Kernel-SVR,DDK-SVR)方法.将该方法用于风速预测中,建立 DDK-SVR 风速预测模型,并将预测结果与传统支持向量机、神经网络方法进行对比.结果表明,DDK-SVR 方法具有更高的预测精度.

**[关键词]** 风速预测,数据依赖核,支持向量机回归

**[中图分类号]** TP181 **[文献标志码]** A **[文章编号]** 1001-4616(2014)03-0015-06

## A Wind Speed Forecasting Model Based on Support Vector Regression with Data Dependent Kernel

Wang Dingcheng, Ni Yujia, Chen Beijing, Cao Zhili

(College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

**Abstract:** Wind is random and has many factors. Besides, the prediction accuracy of wind is not high. Therefore, based on the statistics relationship between Support Vector Machine(SVM) and information geometry, the geometry of kernel function is analyzed. A data dependent kernel is constructed and combined with Support Vector Regression(SVR). Then, the support vector regression machine with data dependent kernel is proposed. We build a wind speed forecasting model and forecast the wind speed. Compared with SVM and neural networks, DDK-SVR method has higher prediction accuracy.

**Key words:** wind speed forecasting, data dependent kernel, support vector regression machine

近年来,由于全球能源危机和令人日益担忧的空气、水和土壤污染,风能作为一种清洁的可再生能源,越来越受到人们的关注.相较于其他的可再生能源,风能投入少收益大,并且可利用的风能在全世界范围内分布都很广泛.风能的转化直接受风速的影响,风的间接性和波动性会对电网的运行产生各种不良影响<sup>[1]</sup>.因此,对风速的准确预测提出了更高的要求.

风速的预测方法有多种,主要有持续法、时间序列法、卡尔曼滤波法、Grey 预测法和人工神经网络<sup>[2-5]</sup>.本文将数据依赖核替代支持向量机回归(SVR)算法中的传统核函数,构造基于数据依赖核支持向量机回归的风速预测模型,利用风速模型<sup>[6]</sup>生成模拟风速数据,使用该数据集对构造的风速预测模型进行试验,将试验结果与人工神经网络、支持向量机方法预测的结果进行比较评价.

## 1 支持向量机回归

支持向量机(SVM)是 Vapnik 等人提出的一种机器学习方法.这种方法是基于统计学习理论发展起来的,统计学习理论是专门研究有限样本情况下机器学习规律的理论<sup>[7]</sup>.支持向量机最初被提出是用于解决模式识别问题,因其能很好地解决非线性、高维数等实际问题,慢慢也将它用于回归问题.其基本思想是通过内积函数定义的非线性变换将输入空间变换到一个高维空间,在这个高维空间中寻找输入变量和输

收稿日期:2013-12-11.

基金项目:国家自然科学基金(61103141)、江苏省自然科学基金(BK2012858)、江苏省高校自然科学基金资助项目(13KJB520015).

通讯联系人:王定成,博士,研究员,研究方向:智能计算. E-mail:dcwang2005@126.com

出变量之间的一种非线性关系.用于非线性回归的SVR算法如下:

对于给定的训练样本集  $D: \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in R^n, y_i \in \mathbf{R}, i=1, 2, \dots, l\}$ , 如果该样本是非线性的, 使用非线性映射函数将数据映射到一个高维特征空间, 然后再在高维空间中进行线性回归, 构造线性回归函数

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b, \quad (1)$$

上式中  $\varphi(\mathbf{x})$  为非线性映射函数;  $\mathbf{w}$  为权向量;  $b$  为常数.

在SVR中引入  $\varepsilon$ -不敏感二次损失函数, 将回归问题转化为一个优化问题, 优化函数为

$$\begin{aligned} \min \quad & \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i, \\ \text{s. t.} \quad & \|y_i - f(\mathbf{x}_i)\| \leq \varepsilon + \xi_i, \quad \xi_i \geq 0, \end{aligned} \quad (2)$$

式中:  $\xi_i$  为  $x_i$  的松弛变量;  $C$  为惩罚参数;  $\varepsilon$  为  $\varepsilon$ -不敏感损失函数的参数.

不难看出, 上面的优化函数为一个二次规划问题, 根据以往的算法, 会将这个二次规划问题引入Lagrange乘子, 并转化到它的对偶空间中解决. 在这里, 不使用上述的传统方法, 而是将损失函数定义在超球上, 直接在原始空间解决该问题. 损失函数定义为

$$L_\varepsilon(z) = \begin{cases} 0, & z \leq \varepsilon, \\ (z - \varepsilon)^2, & z > \varepsilon. \end{cases} \quad (3)$$

将定义的损失函数代入目标函数, 就使原本有约束的目标函数转化为无约束的优化函数, 转化后的优化函数为

$$\min \|\mathbf{w}\|^2 + C \sum_{i=1}^l L_\varepsilon(\|y_i - f(\mathbf{x}_i)\|), \quad (4)$$

对  $\mathbf{w}$  取偏导

$$2\|\mathbf{w}\| + C \sum_{i=1}^l \frac{\partial}{\partial \mathbf{w}} L_\varepsilon(\|y_i - \mathbf{w}\varphi(\mathbf{x}_i) - b\| \cdot \varphi(\mathbf{x}_i)), \quad (5)$$

式(5)取0时, 得到的  $\mathbf{w}^*$  为  $\mathbf{w}$  的最优解.

## 2 数据依赖核

传统的SVR中, 核函数或者是核参数的选择大多是通过经验来选择, 或是依靠遗传算法<sup>[8]</sup>、克隆选择算法(clonal selection algorithms, CSA)<sup>[9]</sup>、粒子群优化算法(PSO)<sup>[10]</sup>等优化算法进行核参数优化. 以上提到的优化算法在一定程度上提高了核学习算法的性能, 但都未考虑到数据的影响, 未改变数据在核映射空间内分布结构, 因而不能从根本上改变核学习算法的性能. 核函数的选择与构造很大程度上影响支持向量机的性能, 传统的学习算法对于不同的训练样本数据构造的核函数都是固定的, 因此能否实现根据所给数据选择相应的核函数或构造相应的核函数是个值得研究的问题. 本文介绍一种基于信息几何的方法<sup>[11]</sup>, 根据其与支持向量机的统计学相关性, 从信息几何学角度分析支持向量机核函数的几何结构, 对核函数进行数据依赖性改进.

从几何的角度看, 非线性映射是从输入空间  $S$  到特征空间  $F$  的一个曲面流形的嵌入. 通常情况下, 空间  $F$  是一个再生核 Hilbert 空间, 因此, 可在空间  $S$  中引入一个非线性映射的黎曼度量  $g_{ij}$ . 其公式为

$$g_{ij}(\mathbf{x}) = \left( \frac{\partial}{\partial x_i} \varphi(\mathbf{x}) \right) \cdot \left( \frac{\partial}{\partial x_j} \varphi(\mathbf{x}) \right), \quad (6)$$

由于  $K(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x}) \cdot \varphi(\mathbf{x}')$ , 因此

$$g_{ij}(\mathbf{x}) = \frac{\partial}{\partial x_i \partial x_j} K(\mathbf{x}, \mathbf{x}') \Big|_{\mathbf{x}'=\mathbf{x}}. \quad (7)$$

在实验中经常使用一些典型的核函数, 大致可以分为两类. 一种是内积函数, 可以表示为

$$K(\mathbf{x}, \mathbf{x}') = f\left(\frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|^2\right), \quad (8)$$

如高斯核函数  $K(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2}$ , 另一种是非径向函数, 通常表示为

$$K(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} \cdot \mathbf{x}'), \quad (9)$$

如  $d$  阶多项式核函数  $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^d$ .

当核函数为径向函数,黎曼度量为

$$\begin{aligned} g_{ij}(\mathbf{x}) &= \frac{\partial}{\partial x_i \partial x_j} K(\mathbf{x}, \mathbf{x}') \Big|_{\mathbf{x}'=\mathbf{x}} = -\delta_{ij} f' \left( \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \right) \Big|_{\mathbf{x}'=\mathbf{x}} - \\ & f'' \left( \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \right) (\mathbf{x}_i - \mathbf{x}'_i) (\mathbf{x}_j - \mathbf{x}'_j) \Big|_{\mathbf{x}'=\mathbf{x}} = -f'(0) \delta_{ij}, \\ \delta_{ij} &= \begin{cases} 1, & i=j, \\ 0, & i \neq j, \end{cases} \end{aligned} \quad (10)$$

相应地,当核函数为内积函数时,黎曼度量为

$$g_{ij}(\mathbf{x}) = f'(0) \delta_{ij} + x_i x_j f''(0), \quad (11)$$

令  $g(\mathbf{x}) = \det(g_{ij}(\mathbf{x}))$ , 则空间  $F$  可以表示为

$$dV = \sqrt{g(\mathbf{x})} dx_1 \cdots dx_n, \quad (12)$$

式中系数  $\sqrt{g(\mathbf{x})}$  表示了空间  $F$  在非线形映射函数下局部区域的缩放性,称之为缩放系数.

将核函数进行共形变换,构造如下函数

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = c(\mathbf{x}) c(\mathbf{x}') K(\mathbf{x}, \mathbf{x}'), \quad (13)$$

其中  $c(\mathbf{x})$  将定义为  $\mathbf{x}$  的正定函数,则函数  $\tilde{K}(\mathbf{x}, \mathbf{x}')$  为满足 Mercer 条件的数据依赖的核函数.

**证明** 由于  $c(\mathbf{x})$  为  $\mathbf{x}$  的正定函数,因此  $\exists a > 0$ , 使得

$$c(\mathbf{x}) \geq a > 0. \quad (14)$$

定义  $\mathbf{X}$  为集合  $R^n$  中的紧集,  $T_k$  为希尔伯特空间上的积分算子,满足

$$\begin{aligned} T_k f(\mathbf{x}) &= \int_{\mathbf{X} \times \mathbf{X}} \tilde{K}(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' = \int_{\mathbf{X} \times \mathbf{X}} c(\mathbf{x}) c(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq \\ & a^2 \int_{\mathbf{X} \times \mathbf{X}} K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0, \quad \forall f \in L_2(\mathbf{x}), \end{aligned} \quad (15)$$

这里  $L_2(\mathbf{x})$  指的是由满足条件  $\int f(\mathbf{x})^2 d\mathbf{x} < \infty$  的所有函数  $f(\mathbf{x})$  组成的空间.

因而函数  $\tilde{K}(\mathbf{x}, \mathbf{x}')$  可表示为  $\mathbf{X} \times \mathbf{X}$  的一致收敛序列

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \tilde{\varphi}_i(\mathbf{x}) \tilde{\varphi}_i(\mathbf{x}'), \quad (16)$$

其中  $\lambda_i > 0$  是  $T_k$  的特征值,  $\tilde{\varphi}_i \in L_2(\mathbf{x})$  是对应  $\lambda_i$  的特征函数.

根据式(16)可知,函数  $\tilde{K}(\mathbf{x}, \mathbf{x}')$  为核函数得证.

此时的黎曼度量变为

$$\tilde{g}_{ij}(\mathbf{x}) = c_i(\mathbf{x}) c_j(\mathbf{x}) + c(\mathbf{x})^2 g_{ij}(\mathbf{x}). \quad (17)$$

核函数的使用是为了将非线性问题在高维中转化为线性问题而提出的,那么对于回归,就要争取使得样本数据转化后大部分点成为线性的,支持向量更接近超平面,减少离群点.因此,为了提高支持向量机的回归性能,要使空间内大部分区域压缩.这里令

$$c(\mathbf{x}) = \sum_{\mathbf{x}_i \in SV} h_i \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2 / 2\tau^2), \quad (18)$$

此时,可得缩放系数为

$$\sqrt{g(\mathbf{x})} = \frac{h_i}{\sigma^n} e^{m^2/2\tau^2} \sqrt{1 + \frac{\sigma^2}{\tau^4} r^2}, \quad (19)$$

其中,  $r = \|\mathbf{x} - \mathbf{x}_i\|$ .

从式(19)可以看出,当在训练样本  $\mathbf{x}_i$  附近,  $h_i < \sigma$  且  $\tau < \sigma/\sqrt{n}$  时这个区域被压缩;当远离训练样本  $\mathbf{x}_i$  时,该区域被压缩.这样就能通过控制  $h$  达到使空间压缩的目的,减弱核参数的影响,提高支持向量机回归的性能.

### 3 基于数据依赖核 SVR 的风速预测模型

#### 3.1 数据依赖核 SVR 算法

将数据依赖核应用于支持向量机回归,给出如下基于数据依赖核支持向量机回归(DDK-SVR)的预测

过程:

步骤一,给出训练样本  $D: \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbf{R}^n, y_i \in \mathbf{R}, i=1, 2, \dots, l\}$ , 设置参数值  $C, \varepsilon$ .

步骤二,初始化. 令  $\mathbf{w}=0, b=0$ , 计算损失函数, 当  $L_\varepsilon(z) \neq 0$  时, 该样本点为支持向量  $\mathbf{M}_{SV}$ .

步骤三, 根据步骤二中得到的支持向量, 按照式 (13) 用数据依赖的方法修改初始核函数, 形成新的数据依赖核函数.

步骤四, 用新的核函数再次训练数据, 从而得到更优的训练结果. 若未达到目标结果, 返回步骤三, 否则迭代结束.

3.2 数据依赖核 SVR 模型的预测结果与分析

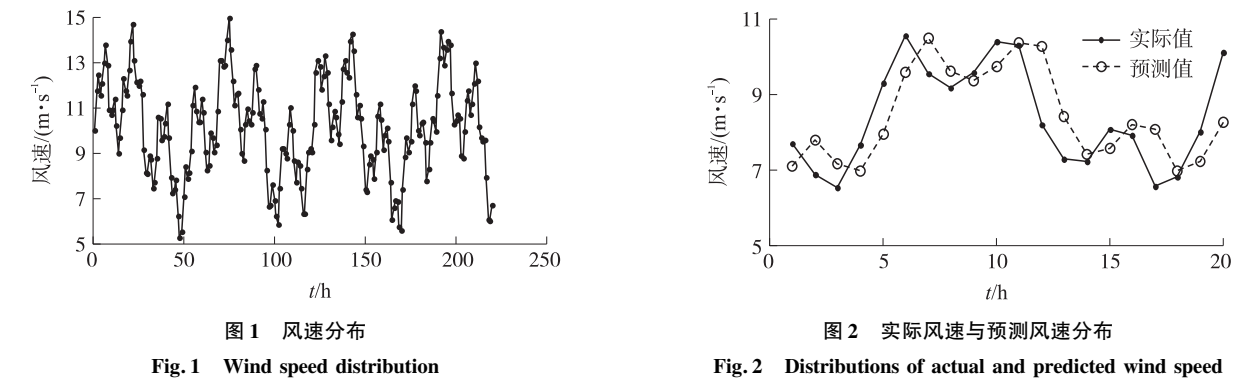
本文利用文献[6]中给出的风速模型:

$$V_w = 10 + 2\sin(wt) + 2\sin(3.5wt) + \sin(12.5wt) + 0.5\sin(35wt),$$
$$w = 2\pi/T_w, \quad T_w = 60s.$$

(20)

用上述模型每小时生成一个数据, 从 0 时到 239 时共生产 240 个样本点. 其中前 220 个点作为训练数据样本, 对后 20 个点进行预测. 本文选用 RBF 核函数作为初始核函数, 根据经验选取惩罚参数  $C$  为 1 000,  $\varepsilon$  为 0.1, 高斯核  $\sigma^2$  为 1. 前 220 个风速数据分布如图 1 所示

利用本文建立的数据依赖核支持向量机回归模型对前 220 个风速进行训练, 并对后 20 个风速值进行预测, 预测风速值的分布如图 2 所示(实际值用实心点表示, 预测值用空心点表示).



根据模型对风速的预测结果如表 1 所示.

表 1 数据依赖核 SVR 模型预测结果

Table 1 Prediction with data dependent kernel SVR model				
t/h	实际风速/(m/s)	预测风速/(m/s)	绝对误差(AE)/(m/s)	相对误差(RE)/%
1	7.687	7.073	0.613	7.985
2	6.882	7.766	-0.884	-12.854
3	6.538	7.156	-0.617	-9.447
4	7.664	6.953	0.710	9.276
5	9.272	7.947	1.325	14.292
6	10.538	9.565	0.972	9.231
7	9.549	10.482	-0.932	-9.760
8	9.154	9.610	-0.456	-4.982
9	9.574	9.361	0.213	2.227
10	10.380	9.732	0.649	6.253
11	10.279	10.343	-0.063	-0.617
12	8.200	10.254	-2.053	-25.038
13	7.271	8.420	-1.148	-15.797
14	7.218	7.400	-0.181	-2.511
15	8.067	7.554	0.513	6.367
16	7.893	8.202	-0.308	-3.909
17	6.560	8.049	-1.487	-22.678
18	6.820	6.973	-0.152	-2.236
19	7.999	7.202	0.797	9.964
20	10.113	8.242	1.870	18.500

由表1可知,模型预测值对于真实值来说大多数数据的预测较为准确,只有少数几个点的误差较大,预测的绝对误差最小能达到0.063 m/s,相对误差百分比基本在10%以下或接近10%。

将本文中的模型预测结果与传统的支持向量机、神经网络的预测结果进行比较,这些方法预测结果的风速分布如图3。

定义平均绝对百分误差:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{y_i}, \quad (21)$$

上式中,  $y_i$  为真实值;  $\hat{y}_i$  为预测值;  $N$  为样本数据的个数。

3种方法的预测结果误差分析情况如表2所示。

表2 3种模型预测的误差

模型	平均绝对误差 (MAE)/(m/s)	均方根误差 (RMSE)/(m/s)	平均百分比误差 (MAPE)/%
DDK-SVR	0.798	0.930	9.696
SVM	0.890	1.156	10.603
NN	0.827	0.999	9.975

从表2中可以看到,数据依赖核支持向量机回归无论在平均绝对误差、均方根误差,还是平均百分比误差都小于传统支持向量机和神经网络。实验证明基于数据依赖核支持向量机回归的风速预测模型在风速预测中的可行性,具有一定的实用价值。

通过上面的实验,结果表明数据依赖核支持向量机的回归模型在风速预测上的精度高于另两种算法,下面将该模型应用于真实气象数据中。

使用某风电场2011年的日平均风速数据用于训练模型,2012年的日平均风速数据用于预测,检验该模型。图4为2011年日平均风速的分布图。

将训练好的模型用于预测2012年的日平均风速,与真实值进行比较,得到图5。

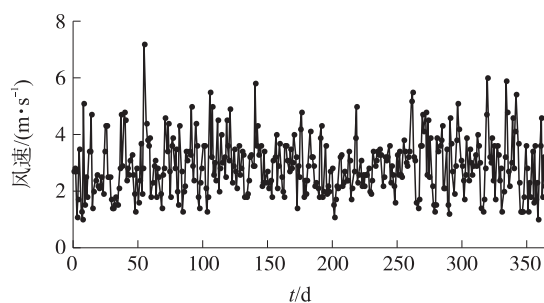


图4 2011年日平均风速分布图

Fig. 4 Distribution of day average wind speed in 2011

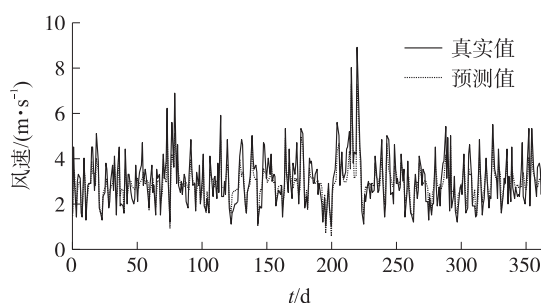


图5 实际风速与预测风速分布图

Fig. 5 Distributions of actual and predicted wind speed

从图上可以看出预测值与真实值的差距不大,通过计算,可以得到该模型的平均百分比误差为17.6%,在误差可接受范围内。

## 4 结论

本文从信息几何学上证明了数据依赖核对核函数的优化,并将其引入支持向量机,建立基于数据依赖核支持向量机回归的风速预测模型。对该模型进行试验分析,实验结果表明,数据依赖核下建立的回归模型降低了核参数对预测效果的影响,在风速预测中比传统的支持向量机和神经网络具有更高的预测精度和更好的泛化能力;该模型对于风速这种波动性大、影响因素多的数据的预测具有很大的优势,能更好地为风电事业服务。

## [参考文献]

- [1] 范伟,赵书强,胡炳杰.应用 STATCOM 提高风电场的电压稳定性[J].电网与清洁能源,2009,25(4):40-44.
- [2] Firat U, Engin S N, Saraclar M, et al. Wind speed forecasting based on second order blind identification and autoregressive model[C]//International Conference on Machine Learning and Applications (ICMLA). Washington, DC, USA, 2010: 686-691.
- [3] Babazadeh H, Gao Wenzhong, Lin Cheng, et al. An hour ahead wind speed prediction by Kalman filter[C]//Power Electronics and Machines in Wind Applications (PEMWA). Denver, USA, 2012: 1-6.
- [4] Huang Chiyo, Liu Yuwei, T Weichang, et al. Short term wind speed predictions by using the grey prediction model based forecast method[C]//Green Technologies Conference (IEEE-Green). Baton Rouge, Louisiana, 2011: 1-5.
- [5] Ghanbarzadeh A, Noghrehabadi A R, Behrang M A, et al. Wind speed prediction based on simple meteorological data using artificial neural network[C]//IEEE International Conference on Industrial Informatics. Cardiff, Wales, 2009: 664-667.
- [6] Adam Mirecki, Xavier Roboam, Frederic Richardean. Architecture complexity and energy efficiency of small wind turbines [J]. IEEE Transactions on Industrial Electronics, 2007, 54(1): 660-670.
- [7] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.
- [8] Liu H, Yaonan Wang, Xiaofen Lu. A method to choose kernel function and its parameters for support vector machines[C]//Machine Learning and Cybernetics. Guangzhou, China, 2005: 4 277-4 280.
- [9] Ming-Yuan Cho, Tsair-Fwu Lee, Shih-Wei Kau, et al. Fault diagnosis of power transformers using SVM/ANN with clonal selection algorithm for features and kernel parameters selection[C]//Innovative Computing, Information and Control. Beijing, China, 2006: 26-30.
- [10] Bai Jing, Guo Yueling. Speech recognition method based on linear descending inertia weight PSO algorithm optimizing SVM kernel parameters[J]. Natural Computation, 2009: 565-568.
- [11] William M Boothby. An Introduction to Differentiable Manifolds and Riemannian Geometry, Revised Second Edition[M]. Singapore: Elsevier, 2007.

[责任编辑:顾晓天]