# 基于改进 SVM 的叶元数目预测

王德吉[1,2] 熊范纶[1] 王儒敬[1] 查世红[1,2]

[1](中国科学院合肥智能机械研究所 合肥 230031)
[2](中国科学技术大学 自动化系 合肥 230026)

**摘 要** 根据外界温度预测叶元数目在建立虚拟植物生长模型中有着重要意义.但是由于环境存在高噪声,不能通过简单的 SVM 或者最小二乘进行回归预测.本文从信息几何角度,构造具有数据依赖性的核函数,克服建模数据的高噪声、非线性,从而能准确预测叶元数目与温度函数关系.最后把模型应用于棉花生长模型的叶元预测,并和标准 SVM、最小二乘进行比较.实验证明新模型在准确度上有较大提高.

**关键词** 叶元,支持向量机,信息几何
**中图法分类号** TP301.6

## The Metamer Number Prediction Based on Improved SVM

WANG De-Ji[1,2], XIONG Fan-Lun[1], WANG Ru-Jing[1], ZHA Shi-Hong[1,2]
[1](*Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei* 230031)
[2](*Department of Automation, University of Science and Technology of China, Hefei* 230026)

**ABSTRACT**

The relation between the temperature and the metamer is very important for the virtual plant growth model. However, it is difficult to predict it just by SVM because there are too many noises in the raw data. In this paper, a new kernel function based on the information geometry is established to overcome the high noise and nonlinear data. The relation between number of metamer and temperature can thus be gotten precisely. The method is applied to the cotton growth model. Compared with the methods of least square and SVM, the improved SVM can predict the number of metamer more precisely.

**Key Words** Metamer, Support Vector Machine, Information Geometry

## 1 引 言

虚拟植物生长模型在研究植物生长、园林规划和生态平衡中,具有重要的理论意义和实际应用价值[1-3].但大部分模型都侧重于计算机图形学,研究真实感植物图形的模拟,不能真实反应植物生长过程.由 de Reffye 等人提出的"参考轴技术"模型,是第 1 个适合模拟真实植物生长过程的模型,但该模

收稿日期:2005-03-10;修回日期:2005-09-29
**作者简介** **王德吉**,男,1975 年生,博士研究生,主要研究方向为数据挖掘、支持向量机. E-mail: wangdejiboy@yahoo.com.cn. **熊范纶**,男,1940 年生,研究员,博士生导师,主要研究方向为知识工程. **王儒敬**,男,1964 年生,研究员,博士生导师,主要研究方向为决策支持系统. **查世红**,男,1969 年生,博士研究生,主要研究方向为 CAD/CAM、虚拟样机.

型存在不易理解和使用的缺点[4,5]. 赵星等人建立基于 Markov Chain 拓扑结构的模型,但由于该模型没有考虑拓扑结构中叶元数目与实际植物生长温度的关系,因此也不能真正地反映植物生长. 本文试图建立温度和叶元函数关系,通过该函数去控制 Markov Chain 拓扑结构的生成,从而建立虚拟植物模型与外界的关系. 关于温度和叶元函数关系,一些学者通过最小二乘回归建立,还有一些学者通过 SVM 进行回归. 但是由于实验数据具有高噪声、非线性,其回归的效果不是很理想. 为此,我们考虑通过改进 SVM 核函数进行抑制噪声的干扰.

## 2 模型建立

### 2.1 叶元数目的确定

SVM 是 Vapnik 等人在 1992 年的 COLT 会议上首次提出,最初用于模式识别问题[6]. 概括地说,SVM 是通过非线性映射将输入空间映射到一个高维特征空间(figure space),在这个空间中构造最优分类超平面的实现过程. 所谓最优分类面,就是不但能将所有样本正确分类,而且使训练样本中离分类面最近的点到分类面的间隔(margin)最大. 通过使间隔最大化来实现较好的泛化能力.

SVM 拓扑结构叶元数目模型可以看成一个非线性函数估计器,通过一个非线性映射 $\Phi$,将输入数据 $x$ 映射到高维特征空间,并在这个空间进行线性回归,即

$$f(x)=\langle\omega\cdot\Phi(x)\rangle+b, \tag{1}$$

其中 $\Phi:\mathbf{R}^n\to F$,$\omega\in F$,$b$ 为偏置. 这样在高维特征空间的线性回归便对应于低维输入空间的非线性回归. 利用结构风险最小化 SRM 原则,则有

$$\begin{aligned}R(\omega)&=R_{\text{emp}}(\omega)+\lambda\|\omega\|^2\\&=\frac{1}{l}\sum_{k=1}^{l}e(f(x_k)-y_k)+\lambda\|\omega\|^2,\end{aligned} \tag{2}$$

其中,$l$ 为样本数目,$e(\cdot)$ 是损失函数,$\lambda$ 是调整的常数. 最小化 $R(\omega)$,可得

$$\omega=\sum_{k=1}^{l}(\alpha_k-\alpha_k^*)\Phi(x_k), \tag{3}$$

其中 $\alpha_k$ 和 $\alpha_k^*$ 是最小化 $R(\omega)$ 的解. 考虑式(1)和式(3),$f(x)$ 可表示为

$$f(x)=\sum_{k=1}^{l}(\alpha_k-\alpha_k^*)\langle\Phi(x_k)\cdot\Phi(x)\rangle+b. \tag{4}$$

令 $K(x,x')$ 满足

$$K(x,x')=\langle\Phi(x)\cdot\Phi(x')\rangle \tag{5}$$

称为核函数,它是满足 Mercer 条件的对称正实数函数. 把式(5)代入式(4)得

$$f(x)=\sum_{k=1}^{l}(\alpha_k-\alpha_k^*)K(x_k,x)+b, \tag{6}$$

这即是基于SVM的拓扑结构叶元数目模型. 我们这里采用 Spline 核函数[10,11]:

$$K(x,x')=1+\langle x\cdot x'\rangle+\frac{1}{2}\langle x\cdot x'\rangle\min\langle x\cdot x'\rangle-\frac{1}{6}\min\langle x\cdot x'\rangle^3. \tag{7}$$

### 2.2 一种新的核函数设计

在上述的 SVM 中,核函数的选择以及 Lagrange 算子的优化都未考虑数据的影响,对于本实验具有高噪声,非线性的数据预测很不准确. 因此,我们从信息几何的角度进行新的核函数的设计.

从几何的观点看,非线性映射 $\Phi(x)$ 是一个子流形,它定义从输入空间 $S$ 到特征空间 $F$ 的一个嵌入. 一般 $F$ 为再生核 Hilbert 空间(RKHS),RKHS 是 Hilbert 空间的子空间,因此可以在 $S$ 空间中引入一个黎曼度量 $G_{ij}(x)$(具体见文献[7] ~ 文献[12]),这个黎曼度量可以用核函数 $k(x,x')$ 近似地表示为

$$G_{i,j}=\frac{\partial}{\partial_{k_i}}\frac{\partial}{\partial_{k_j}}k(x,x')\mid x'=x. \tag{8}$$

本文将核函数取为 Spline 核函数,即

$$K(x,x')=1+\langle x\cdot x'\rangle+\frac{1}{2}\langle x\cdot x'\rangle\min\langle x\cdot x'\rangle-\frac{1}{6}\min\langle x\cdot x'\rangle^3, \tag{9}$$

其中 $\sigma$ 为归一化参数.

此时,黎曼度量为 $\delta_{i,j}=\frac{\delta_{i,j}}{\sigma^2}$,

$$\delta_{i,j}=\begin{cases}1, & i=j\\0, & i\neq j\end{cases} \tag{10}$$

下面在核函数中引入一个保角映射 $D(x)$,于是得新的核函数为

$$\tilde{k}(x,x')=D(x)D(x')k(x,x'). \tag{11}$$

由式(11),可得到变换后的黎曼度量:

$$\tilde{G}_{i,j}(x)=\frac{\partial D(x)}{\partial_{k_i}}\frac{\partial D(x)}{\partial_{k_j}}+[D(x)]^2G_{i,j}(x). \tag{12}$$

基于上述讨论,如果适当选取一个保角映射,就可以在保持原来空间拓扑结构不变的情况下,对非线性数据中的重要样本点附近的区域实现有效放大,从而提高预测效果. 关键问题是如何选择有序样本的重要样本点和重要样本点的数目. 最优分割算法(OPA)是有效方法. 采用最优分割算法,将保角映射取为

$$D(x)=\frac{1}{n}\sum_{i=1}^{n}\exp[-\frac{\|x-\tau_i\|^2}{T_i}],\qquad(13)$$

其中:$n$、$\tau_i$、$T_i$ 分别为OPA算法的分类点数目、第 $i$ 类的中心与对应的宽度.

新的核函数 $\tilde{k}(x,x')$ 满足 Mercer 定理中级数展开的条件,因此可以作为SVM中的核函数.

最后,将依赖于数据的SVM方法的执行步骤简述如下:

1、用OPA方法确定所要研究的非线性数据序列的分类数以及各类的中心与宽度.

2、对核进行修正,首先用式(9)计算基本核函数,然后用式(11)与式(13)对其修正.

3、用改进的核函数训练SVM.

**表1　棉花的叶元数、积温观测表**

Table 1　Records of the number of metamer and accumulated temperature

| 叶元数目 | 积温(℃) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 植株1 | 植株2 | 植株3 | 植株4 | 植株5 | 植株6 | 植株7 | 植株8 | 植株9 | 植株10 |
| 0 | 142 | 122 | 123 | 202 | 129 | 145 | 199 | 175 | 161 | 156 |
| 1 | 189 | 182 | 214 | 219 | 179 | 236 | 228 | 192 | 167 | 182 |
| 2 | 243 | 261 | 259 | 256 | 233 | 291 | 230 | 294 | 241 | 234 |
| 3 | 325 | 297 | 261 | 281 | 298 | 328 | 301 | 334 | 252 | 308 |
| 4 | 333 | 337 | 331 | 342 | 299 | 334 | 330 | 336 | 368 | 372 |
| 5 | 341 | 421 | 379 | 402 | 382 | 378 | 374 | 384 | 415 | 438 |
| 6 | 451 | 457 | 402 | 437 | 453 | 435 | 430 | 388 | 458 | 465 |
| 7 | 457 | 533 | 464 | 448 | 483 | 514 | 475 | 479 | 501 | 486 |
| 8 | 575 | 575 | 527 | 529 | 540 | 516 | 564 | 556 | 562 | 572 |
| 9 | 577 | 579 | 546 | 536 | 550 | 549 | 584 | 569 | 607 | 584 |
| 10 | 660 | 599 | 648 | 639 | 594 | 672 | 672 | 585 | 648 | 656 |
| 11 | 685 | 709 | 662 | 649 | 715 | 710 | 699 | 696 | 652 | 694 |
| 12 | 747 | 745 | 678 | 689 | 724 | 766 | 711 | 766 | 747 | 719 |
| 13 | 759 | 802 | 786 | 786 | 728 | 769 | 761 | 774 | 760 | 814 |
| 14 | 816 | 808 | 805 | 808 | 830 | 818 | 839 | 792 | 772 | 859 |
| 15 | 818 | 828 | 855 | 893 | 873 | 858 | 887 | 880 | 883 | 895 |
| 16 | 877 | 871 | 888 | 942 | 869 | 913 | 889 | 928 | 924 | 960 |
| 17 | 961 | 981 | 931 | 954 | 987 | 944 | 974 | 931 | 998 | 1001 |
| 18 | 971 | 999 | 977 | 1047 | 997 | 987 | 1010 | 978 | 1096 | 1063 |
| 19 | 1073 | 1067 | 1042 | 1120 | 1042 | 998 | 1060 | 1006 | 1067 | 1069 |
| 20 | 1078 | 1094 | 1107 | 1135 | 1098 | 1061 | 1078 | 1116 | 1075 | 1129 |
| 21 | 1151 | 1100 | 1120 | 1172 | 1157 | 1165 | 1133 | 1134 | 1101 | 1181 |
| 22 | 1162 | 1194 | 1231 | 1212 | 1173 | 1186 | 1141 | 1200 | 1141 | 1190 |
| 23 | 1242 | 1264 | 1235 | 1213 | 1266 | 1270 | 1225 | 1255 | 1250 | 1249 |
| 24 | 1294 | 1327 | 1323 | 1258 | 1268 | 1299 | 1256 | 1312 | 1316 | 1282 |
| 25 | 1295 | 1343 | 1345 | 1324 | 1358 | 1380 | 1358 | 1344 | 1394 | 1290 |

# 3　实验与结果分析

## 3.1　实验场所和设备

田间实验于2004年在郑州市农科所实验田进行.土壤类型为红褐土,土质为壤土.所用棉花品种为豫棉10号.种植面积为100m²,种植行距和株距均为0.6m.施用磷酸二铵300 kg/h为基肥,在播前整地时翻入土壤.在实验期间根据土壤含水量适时灌水,保证植株不受干旱胁迫并经常进行除草.温度是影响植物发育最重要的因素,积温被广泛应用于植物发育速度的预测,本模型采用由从出苗期开始的每日平均气温计算的积温预测棉花生长速度.共采集10株棉花的数据,前9株用于训练模型,最后1株用于验证预测结果.积温是作为输入量,叶元作为输出量.

**表2　最小二乘、标准SVM和改进SVM预报结果与实测结果对比**

Table 2　Comparison among records and the predictive results by LS, SVM and improved SVM

| 实际叶元数目 | 最小二乘 | | 标准SVM | | 改进核函数的SVM | |
|---|---|---|---|---|---|---|
| | 预测数目 | 绝对误差 | 预测数目 | 绝对误差 | 预测数目 | 绝对误差 |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 2 | 1 | 1 | 2 | 0 | 2 | 0 |
| 3 | 2 | 1 | 1 | 2 | 3 | 0 |
| 4 | 2 | 2 | 2 | 2 | 4 | 0 |
| 5 | 4 | 1 | 5 | 0 | 5 | 0 |
| 6 | 5 | 1 | 4 | 2 | 7 | 1 |
| 7 | 7 | 0 | 6 | 1 | 8 | 1 |
| 8 | 7 | 1 | 7 | 1 | 8 | 0 |
| 9 | 8 | 1 | 7 | 2 | 9 | 0 |
| 10 | 9 | 1 | 9 | 1 | 10 | 0 |
| 11 | 11 | 0 | 10 | 1 | 12 | 1 |
| 12 | 11 | 1 | 10 | 2 | 12 | 0 |
| 13 | 11 | 2 | 11 | 2 | 12 | 1 |
| 14 | 14 | 0 | 14 | 0 | 14 | 0 |
| 15 | 14 | 1 | 15 | 0 | 15 | 0 |
| 16 | 16 | 0 | 16 | 0 | 16 | 0 |
| 17 | 16 | 1 | 16 | 1 | 17 | 0 |
| 18 | 16 | 2 | 16 | 2 | 18 | 0 |
| 19 | 18 | 1 | 16 | 3 | 19 | 0 |
| 20 | 21 | 1 | 21 | 1 | 21 | 1 |
| 21 | 21 | 0 | 21 | 0 | 21 | 0 |
| 22 | 23 | 1 | 27 | 5 | 22 | 0 |
| 23 | 24 | 1 | 27 | 4 | 23 | 0 |
| 24 | 25 | 1 | 27 | 3 | 24 | 0 |
| 25 | 27 | 2 | 27 | 2 | 24 | 1 |

### 3.2 模型分析

把第10株数据作为测试集，对训练好的模型进行测试，预报结果与实测结果对比如表2所示.

从表2可以看出改进的SVM模型效果较好，正确率为73%，实测值与预报值的误差小于等于1的命中率为96%，最大绝对误差不超过2. 而标准SVM则效果一般，正确率为31%，实测值与预报值的误差小于等于1的命中率为54%，最大绝对误差则超过5；最小二乘预测的效果最差，正确率仅为20%.

## 4 结束语

通过信息几何改进的核函数，在SVM的预测中能够有效防止高噪音干扰. 从而能够准确预测叶元的数目和温度的关系，进而成功建立反映外界环境的拓扑结构和虚拟植物生长模型. 在棉花生长模型实验的研究中证明了这一点.

## 参 考 文 献


[1] Hu B G, Zhao X, *et al*. Plant Growth Modeling and Visualization —Review and Perspective. Acta Automatica Sinica, 2001, 27(6): 816—835 (in Chinese)
(胡包钢，赵 星，等. 植物生长建造模型与可视化——回顾与展望. 自动化学报，2001，27(6)：816—835)

[2] Guo Y, Li B G. Research Summary about Virtual Plant. Chinese Science Bulletin, 2001, 46(4): 273—280 (in Chinese)
(郭 炎，李保国. 虚拟植物的研究进展. 科学通报，2001，46(4)：273—280)

[3] Ding W L. Research of the Agricultural Expert System Based on Artificial Plant Growth Model. Journal of Zhejiang University of Technology, 2005, 33(5): 525—533 (in Chinese)
(丁维龙. 基于虚拟植物生长模型的农业专家系统研究. 浙江工业大学学报，2005，33(5)：525—533)

[4] McKinion J M, Baker D N, Whisler F D, Lambert J R. Application of Gossym/Comax System to Cotton Crop Management. Agricultural Systems, 1989, 31: 55—65

[5] de Reffye P, Edelin C, Francon J, *et al*. Plant Models Faithful to Botanical Structure and Development. ACM Computer Graphics, 1988, 22(4): 151—158

[6] Ma Y, Huang D X, Jin Y H. Soft-Sensor Modeling Method Based on Support Vector Machine. Information and Control, 2004, 33(4): 417—421 (in Chinese)
(马 勇，黄德先，金以慧. 基于支持向量机的软测量建模方法. 信息与控制，2004，33(4)：417—421)

[7] Sun Y F, Liang Y C. An Improved Method for Kernel Function with Data-Dependent Type of Support Vector Machine. Journal of Jilin University (Science Edition), 2003, 41(3): 329—333 (in Chinese)
(孙延风，梁艳春. 支持向量机的数据依赖型核函数改进算法. 吉林大学学报(理学版)，2003，41(3)：329—333)

[8] Vapnik V N. Statistical Learning Theory. New York, USA: Wiley, 1998

[9] Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers. In: Proc of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, USA, 1992

[10] Cortes C, Vapnik V N. Support Vector Networks. Machine Learning, 1995, 20(3): 273—297

[11] Bishop C M. Training with Noise is Equivalent to Tikohonov Regularization. Neural Computation, 1995, 7: 108—116

[12] Vapnik V N, Chervonenkis A Y. Necessary and Sufficient Conditions for the Uniform Convergence of Means to Their Expectations. Theory of Probability and Its Applications, 1981, 26 (3): 532—553