

文章编号: 1001-0920(2009)02-0309-04

一种过程支持向量机及其在动态模式分类中的应用

许少华^{1,2}, 何新贵², 周继¹, 王兵¹

(1. 大庆石油学院 计算机与信息技术学院, 黑龙江 大庆 163318; 2. 北京大学 信息科学技术学院, 北京 100871)

摘要: 针对一般 SVM 在机制上难以直接对动态模式进行分类的问题, 提出了一种基于函数正交基展开的过程支持向量机. 该模型的输入为时变函数, 输出为模式类别. 在输入函数空间中选择一组适当的正交函数基, 将输入函数在该组函数基下进行有限项展开, 把展开式系数作为核函数的输入. 由于时变函数在基函数映射下与展开式系数一一对应, 从而可利用 SVM 的变换机制实现动态模式分类. 给出了基于 SMO 的求解算法, 实验结果验证了模型和算法的有效性.

关键词: 过程支持向量机; 时变信号; 模式分类; 正交函数基展开; 求解算法

中图分类号: TP183 **文献标识码:** A

Process support vector machine and its applications in dynamic pattern classification

XU Shaohua^{1,2}, HE Xin-gui², ZHOU Ji¹, WANG Bing¹

(1. College of Computer and Information Technology, Daqing Petroleum Institute, Daqing 163318, China; 2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China. Correspondent: XU Shaohua, E-mail: xush62@163.com)

Abstract: Aiming at the problem that support vector machine (SVM) is difficult to solve dynamic pattern classification directly in mechanism, a process support vector machine (PSVM) model based on orthogonal function basis expansion is presented in this paper. The input of PSVM can be functions with time-varying, and its output can be pattern classifications. A group of proper orthogonal function basis is chosen in input function space. The input functions with finite terms of the function basis are expanded, and the expansion coefficients are considered as the inputs of kernel function. As time-varying functions under basis-function mapping are in one-to-one correspondence with expansion coefficients, the transformation mechanism of SVM is used to implement classification of dynamic patterns. The solving algorithm based on SMO is given, the results of simulation experiments show the effectiveness of the model and algorithm.

Key words: Process support vector machine; Time-varying signal; Pattern classification; Orthogonal function basis expansion; Solving algorithm

1 引言

支持向量机(SVM)是 Vapnik 等人在 1995 年提出的一种基于统计学习理论的机器学习方法, 具有数学理论完备、全局优化性能好、泛化能力强、算法复杂度与特征空间维数无关等优点^[1,2], 在解决小样本、非线性及高维空间模式识别问题中表现出许多特有的优势, 在许多领域得到了成功的应用.

在科学研究与实际工程应用中, 存在大量的时变信号动态模式分类问题, 例如, 旋转机械运动故障

诊断、飞机发动机飞行状况实时监测、心电图过程信号分析等. 而在现有的 SVM 模型中, 系统的输入一般是与时间无关的常量, 即是几何点式的瞬时输入, 在机制上无法对时变过程信号直接进行分类处理. 本文将传统 SVM 模型和信息处理机制向时域空间进行扩展, 提出了一种基于正交函数基展开的过程支持向量机模型(PSVM). PSVM 的输入可为时变函数或函数向量, 输出为模式类别. 在时变信号模式分类时, 首先在时变输入函数空间中选择一组适当

收稿日期: 2007-11-28; 修回日期: 2008-05-27.

基金项目: 国家自然科学基金项目(60473051); 黑龙江省自然科学基金项目(ZA2006-11).

作者简介: 许少华(1962—), 男, 河北邢台人, 教授, 博士后, 从事模式识别、神经网络等研究; 何新贵(1938—), 男, 浙江浦江人, 中国工程院院士, 教授, 博士生导师, 从事模糊逻辑、神经网络等研究.

的正交函数基(如三角函数基、Walsh 函数基等),将系统输入函数按照给定的拟合精度在该组函数基下进行有限项展开,然后以输入函数展开式系数作为核函数的输入.由于系统输入函数在确定的函数基下与基展开式系数之间存在着——对应关系,从而可借助于一般 SVM 的分类机制完成对时变信号的模式分类.

本文给出了由时变信号输入层、正交函数基表示层、核函数变换层和模式输出层构成的 4 层 PSVM 模型以及相应的求解算法,证明了 PSVM 在函数空间的动态模式分类在一组正交函数基映射下与传统 SVM 在高维数量空间模式分类问题的等价性,并以三角函数类模式识别和油田测井水淹层判别问题为例,验证了模型和算法的有效性.

2 时变函数空间与高维数量空间模式分类问题的等价性分析

定义 1 数量空间 R^n 中内积定义如下:

$$\begin{aligned} X, Y &= \sum_{i=1}^n x_i y_i, \\ X &= (x_1, x_2, \dots, x_n) \in R^n, \\ Y &= (y_1, y_2, \dots, y_n) \in R^n. \end{aligned} \quad (1)$$

定义 2 n 维连续函数空间 $(C[a, b])^n$ 中内积定义如下:

$$\begin{aligned} X(t), Y(t) &= \int_a^b X(t) (Y(t))^T dt = \\ &= \int_a^b \left(\sum_{i=1}^n x_i(t) y_i(t) \right) dt, \quad X(t), Y(t) \in (C[a, b])^n, \end{aligned} \quad (2)$$

式中 $(Y(t))^T$ 表示函数向量 $Y(t)$ 的转置.

定理 1 设 U 为连续函数空间 $(C[a, b])^n$ 的一个子空间, $X(t) = (x_1(t), x_2(t), \dots, x_n(t))$ 为 U 中的任意函数向量, $b_1(t), b_2(t), \dots, b_L(t)$ 为 $C[0, T]$ 空间中一组满足 U 中函数拟合精度要求的有限标准正交基函数, $X(t)$ 在该组基函数下的展开式系数记为 $A = (a_{11}, a_{12}, \dots, a_{1L}, \dots, a_{n1}, a_{n2}, \dots, a_{nL})$, 所有 A 组成的集合记为 $\tilde{H}, \tilde{H} \subset R^{n \times L}$. 定义 U 通过函数基展开变换到 \tilde{H} 的映射为 \cdot . 分别用式(1)和式(2)定义 \tilde{H} 和 H 中的内积,则泛函空间 $\{U, \cdot, \cdot\}$ 与 $\{\tilde{H}, \cdot, \cdot\}$ 在映射 \cdot 下同构.

证明 设函数向量 $X(t) = (x_1(t), x_2(t), \dots, x_n(t))$ 和 $Y(t) = (y_1(t), y_2(t), \dots, y_n(t))$ 通过映射到高维数量空间中的点分别为 $A = (a_{11}, a_{12}, \dots, a_{1L}, \dots, a_{n1}, a_{n2}, \dots, a_{nL})$ 和 $B = (b_{11}, b_{12}, \dots, b_{1L}, \dots, b_{n1}, b_{n2}, \dots, b_{nL})$, 则有:

$$1) \quad X(t) + Y(t) = \sum_{l=1}^L (a_{1l} + b_{1l}) b_l(t), \dots, \sum_{l=1}^L (a_{nl} + b_{nl}) b_l(t), \dots$$

$$\begin{aligned} &= \sum_{l=1}^L (a_{1l} + b_{1l}) b_l(t), \\ (X(t) + Y(t)) &= ((a_{11} + b_{11}), (a_{12} + b_{12}), \dots, (a_{1L} + b_{1L}), \dots, \\ &= (a_{n1} + b_{n1}), (a_{n2} + b_{n2}), \dots, (a_{nL} + b_{nL})) = \\ A + B &= (X(t)) + (Y(t)); \end{aligned}$$

$$\begin{aligned} 2) \quad (kX(t)) &= \sum_{l=1}^L (k a_{1l} b_l(t)), \dots, \sum_{l=1}^L (k a_{nl} b_l(t)) = \\ &= (ka_{11}, ka_{12}, \dots, ka_{1L}, \dots, \\ &= (ka_{n1}, ka_{n2}, \dots, ka_{nL})) = \\ k(a_{11}, a_{12}, \dots, a_{1L}, \dots, a_{n1}, a_{n2}, \dots, a_{nL}) &= \\ kA &= k(X); \end{aligned}$$

3) 由函数基的标准正交性和定义 2, 可得

$$\begin{aligned} X(t), Y(t) &= \int_0^T \left(\sum_{i=1}^n x_i(t) y_i(t) \right) dt = \\ &= \int_0^T \left(\sum_{i=1}^n \left(\sum_{l=1}^L a_{il} b_l(t) \right) \cdot \left(\sum_{l=1}^L b_{il} b_l(t) \right) \right) dt = \\ &= \int_0^T \left(\sum_{i=1}^n \sum_{l=1}^L \sum_{s=1}^L a_{il} b_{is} b_l(t) b_s(t) \right) dt = \\ &= \sum_{i=1}^n \sum_{l=1}^L a_{il} b_{il} = A, B = \\ &= (X(t)), (Y(t)), \end{aligned}$$

定理得证.

上述定理表明,泛函空间 $\{U, \cdot, \cdot\}$ 与 $\{\tilde{H}, \cdot, \cdot\}$ 在映射 \cdot 下同构的,因此当 PSVM 与 SVM 的核函数中的聚合运算分别为式(2)和式(1)定义的内积函数时,函数空间 U 与数量空间 \tilde{H} 在映射 \cdot 下具有相同的模式可分性.故时变信号模式分类问题在映射 \cdot 下可转换为数量空间的模式分类,从而可借助于传统 SVM 的分类机制进行动态模式识别.将时变函数基函数展开运算与 SVM 信息变换过程整合在一起,构成基于函数正交基展开的 PSVM 模型.

3 基于正交函数基展开的 PSVM 模型

设 $U = (C[0, T])^n$ 为时变信号输入空间, $b_1(t), b_2(t), \dots, b_L(t)$ 为 $C[0, T]$ 中一组满足 U 中函数拟合精度要求的标准正交函数基, $X(t) \in U$ 通过该组基函数展开映射到高维空间的点记为 $A = (a_{11}, a_{12}, \dots, a_{1L}, \dots, a_{n1}, a_{n2}, \dots, a_{nL})$. 基于正交函数基展开的 PSVM 拓扑结构如图 1 所示.

图 1 中, $X(t) = (x_1(t), x_2(t), \dots, x_n(t)) \in (C[0, T])^n$ 为 PSVM 的时变输入函数, $K(\cdot, \cdot)$ 为核函数, X_j 为 \tilde{H} 中 SVM 的支持向量, $j (j = 1, 2, \dots, m)$ 为核变换层节点与输出层节点的连接权,

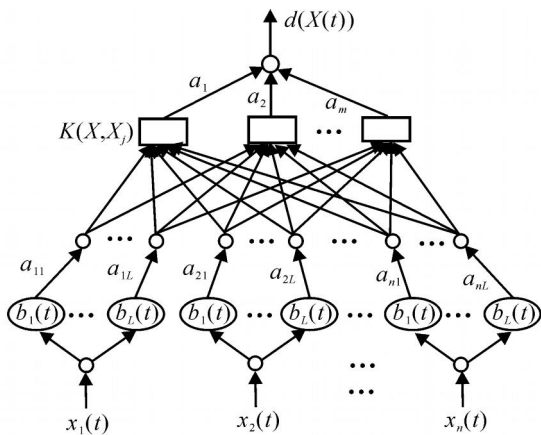


图 1 基于函数正交基展开的过程支持向量机

狄 犴 犸) 为 PSVM 的输出。

由定理 1, 泛函空间 $\{U, \cdot, \cdot\}$ 与 $\{\tilde{H}, \cdot, \cdot\}$ 在基函数展开映射 \tilde{H} 下具有相同的模式可分性。定义 \tilde{H} 中 SVM 的支持向量 X_j 与基函数 $B(t) = (b_1(t), b_2(t), \dots, b_l(t))$ 的组合 $X_j(t) = X_j * B(t)$ 为 PSVM 的过程支持向量, 则 PSVM 的决策规则可表示为

$$y = \text{sgn} \left(\sum_{j \in C} d(X(t))_j K(X(t), X_j(t)) \right) \quad (3)$$

其中: C 表示所有过程支持向量构成的集合, $K(X(t), X_j(t))$ 表示以 $X(t), X_j(t)$ 的函数基展开式系数构成的数值向量的核函数值。

采用不同的核函数可构造出模式空间不同的分类曲面, 实际中采用较多的核函数有^[1,3,4]:

多项式核函数

$$K(x, y) = [(x \cdot y) + 1]^d, \quad (4)$$

径向基核函数

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad (5)$$

两层感知器核函数

$$K(x, y) = \tanh[\rho_0(x \cdot y) + \rho_1]. \quad (6)$$

实际应用中, 可根据动态系统的统计特性选择适当的核函数。一般当核函数的统计特性与实际系统的特征相近时, 可获得最佳的分类结果。

4 PSVM 的求解算法

PSVM 的学习过程分为两个阶段: 第 1 阶段为预处理阶段, 在时变函数空间中选择一组适当的正交函数基和展开方法, 将训练样本集中的样本函数根据需要的拟合精度要求按函数基进行有限项展开, 从而把每个时变输入函数在正交函数基映射下变换成一组函数基展开式系数向量; 第 2 阶段是进行图 1 所示的 PSVM 第 2 隐层到输出节点的核函数性质参数和连接权的学习。当时变输入函数确定后, 其基于函数基的展开系数也随之确定, 此时可采用传统 SVM 的学习机制和算法或其他快速优化算法

进行 PSVM 的求解。

5 仿真实验

例 1 两类三角函数模式分类

考虑如下两类三角函数的模式分类问题。第 1 类函数: $\sin(k t/20 + 0.33)$, $k = 1, 2, \dots, 30, t \in [0, 2]$, 对应输出类别为 1; 第 2 类函数: $\cos(k t/20 + 0.25)$, $k = 1, 2, \dots, 30, t \in [0, 2]$, 对应输出类别为 -1, 共得到 2 组 60 个函数样本。从每组中随机选取 20 个函数作为训练样本, 其余 10 个作为测试样本。采用标准正交三角函数系对 60 个样本函数在拟合精度 0.05 下进行基函数展开, 当基函数项数为 32 时满足拟合精度要求。PSVM 的核函数选择径向基函数。以 40 个时变输入函数的基展开系数作为非时变训练样本, 构成 32 维数量空间中的一个 SVM 训练样本集, 采用 SMO 算法^[5,6] 求解出 10 个支持向量和相应的连接权 $w_j (j = 1, 2, \dots, 10)$, PSVM 的结构确定为: 1-32-10-1。对测试集函数样本进行类型判别, 20 个测试函数样本类别全部判对。

例 2 在油田测井水淹层识别中的应用

油层测井水淹状况识别是油田开发过程中一项十分重要而又复杂的工作^[7]。笔者以大庆萨北油田高台子储油层为研究对象, 采用本文建立的 PSVM 进行水淹层判别。在该区块的北 2-357- 检 82 井、北 1- 丁 1- 检 254 井和北 2-6- 检 251 井等 3 口有水淹状况分析资料的取心井中, 选取 50 个小层组成 PSVM 的训练样本集, 27 个小层样本组成测试集, 每个学习样本由小层测井曲线和水淹类型组成。其中, 水淹油层 42 个样本, 未水淹油层 35 个样本, 分水淹和未水淹两种状况进行判别。

根据测井变量与水淹状况的相关分析结果并结合专家经验, 选取声波时差 AC, 自然电位 SP 等 4 条测井曲线和小层有效厚度 h 作为进行油层水淹判别的特征变量。利用 PSVM 进行小层水淹状况识别时, 系统要有统一的过程输入区间, 但实际中各小层厚度一般不同, 故要进行归一化处理。将待处理并所有小层的厚度最大值取整加 1, 构成统一的过程输入区间, 厚度较小的层厚度不足部分的值取一个基线值 0.2。这样处理, 小层厚度变量已隐含在其他输入变量中, 故在进行油水层识别时可去掉厚度参数。

实际测井中, 测井仪器每英寸 (2.54 cm) 记录 1 次测量值, 测井信号是随深度变化的离散采样数据。采用 Walsh 标准正交函数系^[8] 在精度 0.05 下对采样数据进行拟合。77 个小层样本在基函数项数为 32 时, 达到拟合精度要求。以 50 个训练样本的 Walsh 基函数拟合系数作为非时变训练样本, 构成 32 维数量空间中的一个 SVM 训练样本集, 采用基于样本取

样的 SMO 算法求解出 12 个支持向量以及 PSVM 的连接权系数,确定出结构为 4-32-12-1 的 PSVM. 对测试集样本进行水淹状况判别,判对 25 个,正确识别率为 92.59%. 而采用油田使用的基于 BP 网络水淹层判别软件对相同的训练集和测试集进行训练和判别,测试样本判对 21 个,正确识别率为 77.78%. 处理结果表明 PSVM 的正确识别率有显著提高.

6 结 论

本文将传统 SVM 的信息处理机制向时间域进行扩展,建立了一种基于正交函数基展开的 PSVM 模型. PSVM 的输入为某一过程区间上的时变信号,可自适应提取输入信号的过程特征,并将多个变量的过程特征加以组合形成类别输出,在信息处理机制上对解决动态模式分类问题有很好的适应性. PSVM 拓宽了传统 SVM 的应用领域,对于时变对象的模式分类等问题的解决具有重要意义.

参考文献(References)

- [1] Vapnik V N. Statistical learning theory [M]. New York: Wiley, 1998.
- [2] Felipe Cucker, Steve Smale. On the mathematical foundations of learning [J]. Bulletin of the American Mathematical Society, 2001, 39(1): 1-49.
- [3] 王国胜. 核函数的性质及其构造方法. 计算机科学, 2006, 33(6): 172-174.
(Wang G S. Properties and construction methods of kernel in support vector machine[J]. Computer Science, 2006, 33(6): 172-174.)
- [4] Scholkopf, Sung K, Burges C, et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers [J]. IEEE Trans on Signal Processing, 1997, 45(11): 2758-2765.
- [5] Collobert R, Bengio S. SVM Torch: A support vector machine for large scale regression and classification problems[J]. J of Machine Learning Research, 2001, 1(2): 143-160.
- [6] 范玉刚, 李平. 基于样本取样的 SMO 算法[J]. 信息与控制, 2004, 33(6): 665-669.
(Fan Y C, Li P. A sampling based SMO algorithm[J]. Information and Control, 2004, 33(6): 665-669.)
- [7] 侯连华, 王京红, 刘泽容. 水淹层测井评价方法[J]. 石油学报, 1999, 20(3): 49-55.
(Hou L H, Wang J H, Liu Z R. Evaluation of water flooded interval well logging[J]. Acta Petrolei Sinica, 1999, 20(3): 49-55.)
- [8] 王能超. 同步并行算法设计[M]. 北京: 科学出版社, 1996: 100-105.
(Wang N C. Synchronous parallel algorithm design [M]. Beijing: Science Press, 1996: 100-105.)
- [4] Matsopoulos G, Marshall S. Feature migrateonin morphological scale space [C]. 1993 IEEE Int Conf on Acoustics, Speech, and Signal Processing. New York: Institute of Electrical and Electronics Engineers, 1993: 599-602.
- [5] 陈白帆, 蔡自兴. 基于尺度空间理论的 Harris 角点检测 [J]. 中南大学学报, 2005, 36(5): 751-754.
(Chen B F, Cai Z X. The Harris corner detection based on scale space[J]. J of Zhongnan University, 2005, 36(5): 751-754.)
- [6] 张剑清, 潘励, 王树根. 摄影测量学[M]. 武汉: 武汉大学出版社, 2003: 83-85.
(Zhang J Q, Pan L, Wang S G. Photogrammetry[M]. Wu Han: Wuhan University Press, 2003: 83-85.
- [7] Charris M S. A combined corner and wine detectors[C]. Proc of the 4th Alvey Vision Conf. Manchester: The Plessey Company, 1988: 147-151.
- [8] 陈静. 图像配准特征点提取算法研究[D]. 南京: 南京理工大学, 2006.
(Chen J. The algorithm studies of image registration feature points [D]. Nanjing: Nanjing University of Science and Technology, 2006.)
- [9] Lindebern T. Feature detection with automatic scale selection[J]. Int J of Computer Vision, 1998, 30(2): 79-116.
- [10] Mikolajczyk K, Schmid C. Scale and affine invariant interest point detectors[J]. Int J of Computer Vision, 2004, 60(1): 63-86.

(上接第 308 页)