文章编号: 1001-0920(2008)05-0560-03

基于核模糊聚类的多模型 LSSVM 回归建模

李 卫, 杨煜普, 王 娜
(上海交通大学 自动化系, 上海 200240)

摘 要: 针对大规模数据采用单模型回归存在精度差和计算量较大的问题,提出一种基于核模糊聚类的多模型最小二乘支持向量回归建模方法. 该方法首先使用基于条件正定核的模糊 C 均值聚类算法对数据集做出聚类划分;然后针对每个聚类做最小二乘支持向量回归估计;同时根据每个聚类内数据分布的特征,给出了一种简单的核参数选择方法. 利用数值仿真实验进行非线性函数估计,实验结果表明了所提出的方法具有良好的精度和泛化能力.

关键词: 核模糊聚类; 多模型; 最小二乘支持向量机
中图分类号: TP18 **文献标识码**: A

Multi-model LSSVM regression modeling based on kernel fuzzy clustering

LI Wei, YANG Yu-pu, WANG Na
(Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China. Correspondent: LI Wei, E-mail: wei_lea@sjtu.edu.cn)

Abstract: In dealing with massive data, a single model usually suffers from bad accuracy and big computation. Therefore, a multi-model LSSVM(least square support vector machine) based on kernel fuzzy clustering for nonlinear modeling is presented. A conditionally positive definite kernel-based fuzzy C-means clustering algorithm is used to make a partition for the data set. Then LSSVM is used to achieve regression for each cluster. According to the character of data distribution in clusters, a simple parameter selection criterion for kernel function is proposed. The numerical simulation results illustrate the effectiveness of the proposed approach.

Key words: Kernel fuzzy clustering; Multi-model; Least square support vector machine.

1 引 言

基于统计学习理论的支持向量机(SVM)是近几年最流行的机器学习工具之一,已在模式识别、信号处理和函数估计等领域得到了成功应用. SVM 的训练涉及到求解一个二次规划问题,在处理较大规模数据时不可避免地存在计算量和内存消耗过大的缺点. 最小二乘支持向量机(LSSVM)是 SVM 的扩展,通过选用不同的损失函数来避免求解二次规划问题,转而求解一组线性方程,从而提高运算速度. 但是,在处理实际问题时,由于数据的复杂性以及野点的存在,单一的回归模型往往不尽如人意,而采用多模型则显得更加合理.

文献[1]提出了基于满意模糊 C 聚类的多模型辨识方法,得到的多模型系统能在全局拟合和局部特性间作出权衡;文献[2]采用模糊 C 均值聚类和 RBF 神经网络相结合来进行多模型软测量建模,结果证明多模型方法具备更好的精度和泛化能力.

传统模糊 C 均值聚类方法依赖于数据分布形状,并且对孤立点敏感,这势必影响到最终的模型性能. 对此,本文提出一种基于条件正定核函数的模糊 C 均值聚类算法,使用核技巧将样本映射到高维特征空间聚类,实现对不规则形状数据聚类;再将聚类算法和最小二乘支持向量机相结合,得出了一种多模型回归建模方法.

2 核模糊 C 均值聚类

文献[3,4]给出过核聚类的算法,取得了良好的聚类效果. 但这些方法所求得的聚类中心存在于高维特征空间中,由于映射函数未知,最后只能采用

收稿日期: 2007-01-16; **修回日期**: 2007-05-23.
基金项目: 国家 973 重点基础研究发展基金项目(2004CB720703).
作者简介: 李卫(1979—),男,安徽怀宁人,博士生,从事复杂系统建模的研究; 杨煜普(1957—),男,西安人,教授,博士生导师,从事智能控制、智能信息处理等研究.

估算的方法得到聚类中心的位置. 本文给出的聚类算法能够克服这一缺点.

传统模糊 C 均值聚类是将平方范数作为聚类相似性衡量标准的. 若存在样本集 $X = \{x_i \mid i = 1,2,\cdots,n\}$,模糊 C 均值聚类的价值函数为

$$J = \sum_{j=1}^{k} Ji = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{m} \| x_i - c_j \|^2, \quad (1)$$

约束于

$$\sum_{j=1}^{k} u_{ij} = 1,\ 0 \leq u_{ij} \leq 1. \quad (2)$$

其中: k 为聚类个数, c_j 为聚类中心, u_{ij} 为样本 x_i 对应于第 j 个聚类的隶属度值, $m \in [1,\infty)$ 是一个加权指数. 通过拉格朗日乘子法,构造目标函数如下:

$$\bar{J}(U, c_1, \cdots, c_k, \lambda_1, \cdots, \lambda_n) = J(U, c_1, \cdots, c_k) + \sum_{i=1}^{n}\lambda_i\Big(\sum_{j=1}^{k} u_{ij} - 1\Big) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m} \| x_i - c_j \|^2 + \sum_{i=1}^{n}\lambda_i\Big(\sum_{j=1}^{k} u_{ij} - 1\Big).$$

最小化 $\bar{J}$ 即可导出聚类中心 c 和隶属度矩阵 U 的更新公式.

引入非线性映射 $\Phi: x \to \Phi(x)$,特征空间中的样本距离则定义为

$$\| \Phi(x_i) - \Phi(c_j) \|^2 = K(x_i, x_i) + K(c_j, c_j) - 2K(x_i, c_j), \quad (3)$$

其中 K 为核函数. 在 SVM 中应用的核函数一般需满足 Mercer 条件[5]. 但在一般的核学习过程中,并非都需满足 Mercer 条件,而且 Mercer 条件也显得过于苛刻了. 这里,讨论一种条件正定核[6],它可以用于核学习,且能简化距离计算.

定义 1[6]　一个对称函数 $K: \chi \times \chi \to R$ 对所有的 $m \in N$, $x_i \in \chi$ 和所有的 $c_i \in R$, $\sum_{i=1}^{m} c_i = 0$,产生一个正定的 Gram 矩阵,即

$$\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geq 0, \quad (4)$$

则 K 为一个条件正定核. 下面就是一个常见的条件正定核:

$$K(x, y) = -(\| x - y \|^2 + b^2)^{1/2}, b \in R. \quad (5)$$

不妨设 $b = 1$,将以上核代入式(3),改写聚类价值函数(1),得到

$$J_\Phi = \sum_{i=1}^{k} Ji = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m} \| \Phi(x_j) - \Phi(c_i) \|^2 = -2\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m}[1 + K(x_i, c_j)], \quad (6)$$

则新的目标函数为

$$\bar{J}_\Phi(U, \Phi(c_1), \cdots, \Phi(c_k), \lambda_1, \cdots, \lambda_n) = -2\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m}[1 + K(x_i, c_j)] + \sum_{j=1}^{n}\lambda_j\Big(\sum_{i=1}^{k} u_{ij} - 1\Big). \quad (7)$$

分别对 $\bar{J}_\Phi$ 关于 u, c 和 λ 求偏导,得到新的聚类中心 c 和隶属度矩阵 U 的更新公式如下:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} K^{-1}(x_i, c_j)\, x_i}{\sum_{i=1}^{n} u_{ij}^{m} K^{-1}(x_i, c_j)}, \quad (8)$$

$$u_{ij} = \frac{(1 + K(x_i, c_j))^{-1/m-1}}{\sum_{j=1}^{k}(1 + K(x_i, c_j))^{-1/m-1}}. \quad (9)$$

根据以上结果,一种核模糊 C 均值聚类算法可归纳如下:

Step1: 设定参数 m, k 初始值,用 $[0,1]$ 间的随机数初始化隶属矩阵 U,使其满足式(2)中的约束条件.

Step2: 计算核矩阵 $K(x_i, c_j)$.

Step3: 用式(8)计算 k 个聚类中心 c_j ($j = 1,2,\cdots,k$).

Step4: 根据式(6)计算价值函数. 如果价值函数值或变化值小于某个既定的阈值,则算法停止,样本按所属隶属度最高值分类;否则,转至 Step5.

Step5: 用式(9)计算新的 U 矩阵,返回 Step2.

3　多模型 LSSVM 回归建模

最小二乘支持向量回归机思想可简单表述如下[7]:若 $\{(x_1, y_1), \cdots, (x_N, y_N)\} \subset R^n \times R$ 是一个训练样本集,线性回归函数为 $f(x) = \omega^{\mathrm{T}} x + b$, $\omega \in F$, b 为偏置. 利用结构风险最小化原则,优化问题为

$$\min\Big\{\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi_i\Big\},$$
$$\text{s.t. } y_i - (wx_i + b) = \xi_i,$$
$$i = 1,2,\cdots,N. \quad (10)$$

用拉格朗日乘子法求解,优化问题最终转化为求解方程组

$$\begin{bmatrix} 0 & 1 \\ 1 & x^{\mathrm{T}}x + c^{-1} \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}.$$

得到 $\omega = \sum_{i=1}^{N}\alpha_i x_i$, $\xi_i = \alpha_i / C$. 通过选择适当的核函数 K,所求得的非线性回归函数则形如

$$f(x) = \sum_{i=1}^{N}\alpha_i K(x_i, x) + b. \quad (11)$$

核函数可选取常用的高斯核

$$K(x, y) = \exp[-(x - y)^2/\sigma^2]. \quad (12)$$

与单模型方法不同,多模型可针对不同聚类来对子模型选取不同的核参数. 由于高斯核是典型的局部性核函数,其函数值受离测试点距离较近的数

据影响较大,可以根据聚类内数据和聚类中心的平均距离来确定核宽度参数,以保证该聚类内的数据对相应的核函数值影响较大,而非聚类内数据对其影响较小. 计算公式可定义为

$$\sigma_i = \sqrt{\frac{1}{l}\sum_{j=1}^{l} \| x_{ij} - c_i \|^2},$$
$$i = 1,2,\cdots,k,\ j = 1,2,\cdots,l. \qquad (13)$$

其中:c_i 为第 i 个聚类中心,x_{ij} 为属于第 i 聚类的第 j 样本数据,l 为第 i 个聚类所包含的样本总数.

对得到的 k 个子回归模型,采用模糊隶属度来综合最后的输出结果

$$F = \sum_{i=1}^{k}\sum_{j=1}^{N}\mu_{ij} f_i(x_j). \qquad (14)$$

对于训练样本,μ_{ij} 已知;而对新的测试样本,也无需重新做聚类来求 μ_{ij},可以用训练样本中离它最近的样本隶属度来代替.

多模型回归建模算法(如图 1 所示)可描述如下:

Step1: 应用核模糊聚类算法将样本数据划分为 k 个聚类($k \geq 2$);

Step2: 选高斯核函数,按式(13)计算核宽度参数;

Step3: 对每个聚类做最小二乘支持向量回归,得到 k 个子模型 $f_i, i = 1,2,\cdots,k$;

Step4: 按式(14)综合多模型输出.

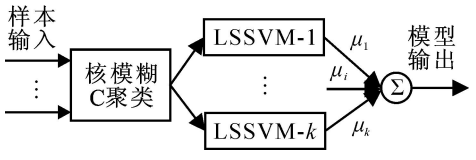

图 1 LSSVM 多模型回归建模

4 数值仿真

例 1 考虑如下非线性函数:

$$f_1(x_1,x_2) = (1 + x_1^{-2} + x_2^{-1.5})^2. \qquad (15)$$

在 $[1,5]\times[1,5]$ 上随机选取 100 个点,前 50 个为训练样本,后 50 个为测试样本. 设定聚类个数 $k = 3$,模糊矩阵加权指数 $m = 2$,LSSVM 参数 $C = 30$. 采用根均方差作为模型性能评价指标. 图 2 显示了模型对训练样本集实际函数值的逼近效果.

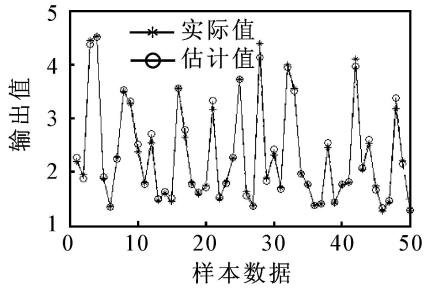

图 2 函数 1 训练样本逼近效果

为更好说明本方法的有效性,分别使用 SVM,LSSVM 以及自适应神经模糊推理系统(ANFIS) 对本例建模,并对结果作出比较. 参数选择采用经验参数设置:SVM 的不敏感损失参数 $e = 0.01$,$C = 30$,高斯核函数宽度参数 $\sigma = 2.1$;LSSVM 参数 $C = 30$,高斯核函数宽度参数 $\sigma = 2.1$;ANFIS 训练代数设为 100,每维数据划分为 2 个模糊子区间,采用钟形隶属度函数. 模型输出比较结果如表 1 所示. 可见,无论在训练集还是测试集上,本文提出模型与其他模型相比均有较高的精度.

表 1 例 1 仿真结果比较

模 型	训练集根均方差	测试集根均方差
SVM	0.254 1	0.235 2
LSSVM	0.112 2	0.146 9
ANFIS	0.032 4	0.046 1
本文方法	0.025 2	0.031 1

例 2 考虑非线性函数

$$f_2(x_1,x_2) = \sin(\pi x_1)\sin(\pi x_2). \qquad (16)$$

在 $[-1,1]\times[0,1]$ 上选取 300 个样本对,前 100 个为训练样本,后 200 个为测试样本. 利用本文方法实现函数估计,设定聚类个数 $k = 5$,其他参数设置同例 1,最后得到的对训练样本逼近效果如图 3 所示. 再分别使用 SVM,LSSVM 和 ANFIS 与之比较. SVM 参数 $e = 0.01$,$C = 15$,核宽度参数 $\sigma = 0.55$;LSSVM 参数 $C = 15$,$\sigma = 0.55$;ANFIS 参数设置同上例. 表 2 中列出了几种方法的比较结果. 可见,本文提出的多模型方法能够得到更好的训练精度,同时保持良好的泛化能力.

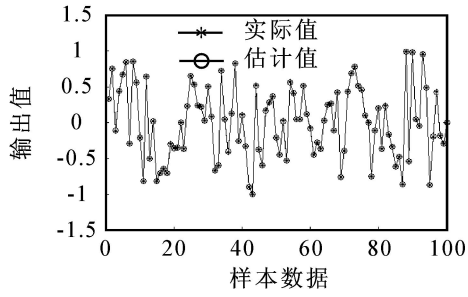

图 3 函数 2 训练样本逼近效果

表 2 例 2 仿真结果比较

模型	训练集根均方差	测试集根均方差
SVM	0.007 1	0.010 1
LSSVM	0.011 7	0.024 7
ANFIS	0.003 9	0.015 7
本文方法	0.003 4	0.007 3

(下转第 566 页)

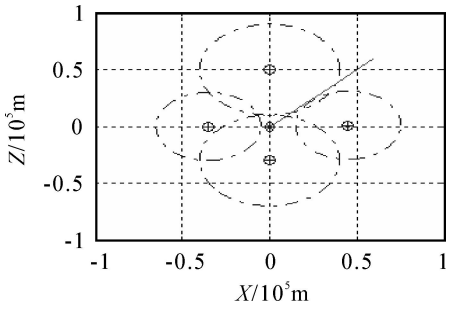

图 2 优化策略

2) 在不考虑目标时效性的前提下,按给定的目标优化航迹进行攻击,发射时间任意,发射精确制导武器数量为 2,作战效能为 0.931,满足期望.

通过算例检验了文中策略优化模型与优化算法的有效性. 对比 1) 和 2) 可知,精确制导武器打击时间敏感目标的关键是时效性的把握,增加精确制导武器发射数量可提高对时间敏感目标的捕捉能力和毁伤能力. 对比两种条件下的计算结果可知,提高精确制导武器的相对打击能力可提高对时间敏感目标的打击效果.

6 结 论

通过对时间敏感目标特性和精确制导武器对目标打击的过程模型的分析与建模,建立了基于效能的打击决策模型,并设计了优化算法. 通过算例检验了策略优化模型与算法的有效性. 本文方法可适用于不同类型的精确制导武器,具有一定的推广价值.

参考文献(References)

[1] Hewitt, Mark A. Time sensitive targeting-overcoming the intelligence gap in interagency operations[R]. US: Naval War College, 2003.

[2] 沈林成, 高国华, 常文森, 等. 开放式飞行任务规划方法[J]. 宇航学报, 1998, 19(4): 13-18.
(Shen Lin-cheng, Gao Guo-hua, Chang Wen-sen, et al. Anopen system approach to mission route planning[J]. J of Astronautics, 1998, 19(4): 13-18.)

[3] 蒋忠中, 汪定伟. 有时间窗车辆路径问题的捕食搜索算法[J]. 控制与决策, 2007, 22(1): 59-68.
(Jiang Zhong-zhong, Wang Ding-wei. Predatory search algorithm for vehicle routing problem with time windows [J]. Control and Decision, 2007, 22(1):59-68.)

[4] 樊晓平, 罗熊, 易晟, 等. 复杂环境下基于蚁群算法的机器人路径规划[J]. 控制与决策, 2004, 19(2): 166-170.
(Fan Xiao-ping, Luo Xiong, Yi Sheng, et al. Path planning for robots based on ant colony optimization algorithm under complex environment[J]. Control and Decision, 2004, 19(2): 166-170.)

[5] 张克, 刘永才, 关世义. 关于导弹武器作战效能能评估问题的探讨[J]. 宇航学报, 2002, 23(2):58-66.
(Zhang Ke, Liu Yong-cai, Guan Shi-yi. An investigation into the problem of evaluating combat effectiveness for missile weapon systems [J]. J of Astronautics, 2002, 23(2): 58-66.)

(上接第 562 页)

5 结 论

本文给出了一种基于核模糊 C 均值聚类的多模型 LSSVM 回归建模方法. 由于核技巧的使用,核模糊聚类避免了传统聚类方法对数据分布的依赖性. 应用条件正定核使得聚类中心仍在原始数据空间中,便于利用. 结合最小二乘支持向量机对每个聚类做非线性回归,并根据聚类内的样本数据分布来调整核参数,使得最后的回归模型在精度上优于传统方法.

参考文献(References)

[1] 薛振框, 李少远. MIMO 非线性系统的多模型建模方法[J]. 电子学报, 2005, 33(1): 52-56.
(Xue Zhen-kuang, Li Shao-yuan. A multi-model Modeling approach to MIMO nonlinear systems [J]. Acta Electronic Sinica, 2005,33(1): 52-56.)

[2] 仲蔚, 俞金寿. 基于模糊 C 均值聚类的多模型软测量[J]. 华东理工大学学报, 2000, 26(1): 83-87.
(Zhong Wei, Yu Jin-shou. Study on soft sensing modeling via FCM based multiple models[J]. J of East China University of Science and Technology, 2000, 26 (1): 83-87.)

[3] 李侃, 刘玉树. 模糊核聚类的自适应算法[J]. 控制与决策, 2004, 19(5): 595-597.
(Li Kan, Liu Yu-shu. Fuzzy kernel clustering self-adaptive algorithm[J]. Control and Decision, 2004, 19 (5): 595-597.)

[4] 孔锐, 张国宣. 基于核的 K-均值聚类[J]. 计算机工程, 2004, 30(11): 12-14.
(Kong Rui, Zhang Guo-xuan. Kernel-based K-means clustering[J]. Computer Engineering, 2004, 30 (11): 12-14.)

[5] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer Verlag, 1995.

[6] Scholkopf B. The kernel trick for distances [R]. Cambridge: Microsoft Research, 2000.

[7] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural Processing Letter, 1999, 9(3): 293-300.