第 37 卷 第 14 期
Vol.37 No.14

计 算 机 工 程
Computer Engineering

2011 年 7 月
July 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)14—0169—03 文献标识码：A 中图分类号：TP301.6

基于 AP 的 LS-SVM 多模型建模算法

宋 坤，李丽娟，赵英凯

(南京工业大学自动化与电气工程学院，南京 210009)

摘 要：针对多工况对象的单模型建模中存在的回归精度差和泛化能力弱的问题，提出基于仿射传播聚类的 LS-SVM 多模型建模方法。该方法用仿射传播聚类算法对样本进行聚类，采用 LS-SVM 的方法对子类样本分别建立模型。测试样本根据相似性的测度进行归类，并用所属子类的模型进行预测输出。将该建模方法用在丙烯浓度的软测量建模实验中，结果表明该方法有较高的回归精度和较好的泛化能力。

关键词：多模型；仿射传播聚类；最小二乘支持向量机；建模

LS-SVM Multi-model Modeling Algorithm Based on AP

SONG Kun, LI Li-juan, ZHAO Ying-kai

(School of Automation and Electrical Engineering, Nanjing University of Technology, Nanjing 210009, China)

【Abstract】The single model of the object with multiple working positions usually suffer from bad accuracy. To solve the problem, a Least Squares-Support Vector Machine(LS-SVM) multi-model modeling method based on affinity propagation clustering is presented. In this method, affinity propagation clustering is used to cluster training samples. The sub-models are trained by LS-SVM. The predicted values of the test samples are estimated by the sub-models after it is classified by similarity measurement. The proposed method is applied for soft-sensing modeling to predict the propylene concentration. Experimental results indicate that the proposed method has a superior regression accuracy and good generalization ability.

【Key words】multi-model; affinity propagation clustering; Least Squares-Support Vector Machine(LS-SVM); modeling

DOI: 10.3969/j.issn.1000-3428.2011.14.056

1 概述

工业过程的监测、控制和优化往往要依赖于高性能的系统模型。随着工业过程的复杂化，出现了非线性、多工况等特点。使用单一模型存在回归精度低和推广能力差的问题。为了解决这个问题，有学者提出了多模型建模方法，根据先验知识将样本聚类，分别对子类样本建立模型，将各子模型的输出加权平均后作为最终的输出。

引入数据挖掘中的聚类思想，对解决多工况过程的建模有非常重要的意义。但来自现场的建模样本分类数目的先验知识一般是未知的，而传统的 k-means、模糊 C 均值等聚类方法对初始聚类中心的选取较敏感，很难根据对象自身情况准确地确定聚类中心和聚类。为了增加模型的鲁棒性，多模型建模采取加权输出的方法。该方法假设全局模型为子模型的线性组合，而实际上对于复杂对象，该假设不一定成立[1]。针对这个问题，有学者提出对各子模型用切换开关的方式组合作为最终模型的输出[2]，即样本分类后各子类独立建模并输出。由于分类后，各子模型训练样本会减少，小样本使传统的基于统计方法和神经网络的建模方法存在回归精度不高和泛化能力差的问题。

针对上述问题，本文提出了基于仿射传播聚类的最小二乘支持向量机多模型建模算法，并将该算法应用于丙烯精馏塔塔顶丙烯浓度的软测量建模中。

2 基于仿射传播聚类的 LS-SVM 多模型建模算法

基于仿射传播聚类(Affinity Propagation clustering, AP)[3]的最小二乘支持向量机(Least Squares-Support Vector Machine, LS-SVM)多模型建模算法如图 1 所示。首先采用仿射传播聚类的方法对样本数据聚类，得到若干子类，然后用最小二乘支持向量机对各子类分别训练建模，得到各子模型。测试样本先根据相似性度量的方法进行归类，再分别用对应的子模型预测并输出。

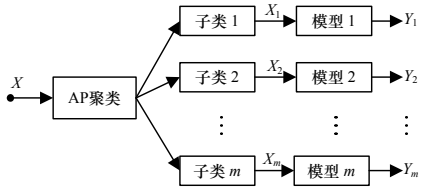

图 1 多模型建模方法

2.1 仿射传播聚类算法

仿射传播聚类是一种新的快速、有效的聚类方法。它不需要事先确定聚类个数，是各个样本点通过迭代竞争聚类中心，并达到最优的聚类结果。仿射传播聚类能更好地根据样本自身的情况聚类，使聚类结果与对象的特征更加吻合。该方法能有效地解决多模型建模中存在的聚类准确度不高的问题。

仿射传播聚类最初将所有样本点都看作潜在的聚类中心，通过循环迭代，每个样本点竞争聚类中心。循环迭代是基于相似度计算进行寻找聚类中心的过程。设样本个数为 N 的样本空间，任意 2 个样本点 x_i 和 x_j 之间的相似度 $S(i,j)$用负

基金项目：江苏省自然科学基金资助项目(BK2009356)；江苏省高校自然科学基金资助项目(09KJB510003)；南京工业大学青年教师学术基金资助项目(39710005)

作者简介：宋 坤(1983－)，男，硕士，主研方向：软测量建模；李丽娟，副教授；赵英凯，教授、博士生导师

收稿日期：2010-12-30 **E-mail：**ljli@njut.edu.cn

的欧氏距离(或欧氏距离的平方)度量，其值储存于 $N\times N$ 的相似度矩阵 $\boldsymbol{S}$ 中。矩阵对角线上的元素为偏向参数 p，它表示各样本点被选作聚类中心的倾向性。p 可取 $\boldsymbol{S}$ 元素的中值[4]，调节 p 可以改变聚类的结果，当 p 小于一定的阈值时会引起分类数目的改变。

定义 $R(i,k)$为 x_k 适合作为 x_i 的聚类中心的程度(responsibilty，吸引度)；$A(i,k)$为 x_i 选择 x_k 作为其聚类中心的适合程度(availability，归属度)。为了找到合适的聚类中心 x_k，AP 算法不断地从数据样本中搜集证据 $R(i,k)$和 $A(i,k)$。$R(i,j)$和 $A(i,j)$的迭代公式为：

$$R(i,k)=S(i,j)-\max_{j\neq k}\{A(i,j)+S(i,j)\} \tag{1}$$

$$A(i,k)=\min\left\{0,R(k,k)+\sum_{j\notin\{i,k\}}\max\{0,R(j,k)\}\right\} \tag{2}$$

AP 聚类过程就是主要根据式(1)和式(2)不断循环迭代从而更新证据的过程。迭代更新的快慢可以通过调节阻尼系数 λ 实现。对于数据点 x_i，若数据点 x_k 使得 $R(i,k)+A(i,k)$为 $R(i,j)+A(i,j), j=1,2,\cdots,N$ 中的最大值，那么数据点 x_k 就是 x_i 的聚类中心。通过迭代竞争的方式，仿射传播聚类可以得到最优的聚类中心和各个样本点的类属情况。

为了评价聚类的效果，引入 Silhouette 指标[4]，它反映了聚类结构的类内紧密性和类间可分性。设一个具有 n 个样本、k 个聚类 $\boldsymbol{C}_i(i=1,2,\cdots,k)$的数据集，某一个样本 t 的 Silhouette 指标为：

$$S_{il}(t)=\frac{\min\{d(t,\boldsymbol{C}_i)\}-a(t)}{\max\{a(t),\min\{d(t,\boldsymbol{C}_i)\}\}} \tag{3}$$

其中，$d(t,\boldsymbol{C}_i)$为 $\boldsymbol{C}_i$ 的样本 t 到另一个类 $\boldsymbol{C}_j$ 的所有样本的平均不相似度或距离；$a(t)$为聚类 $\boldsymbol{C}_i$ 中的样本 t 与 $\boldsymbol{C}_i$ 内所有其他样本的平均不相似度或距离。一个数据集中所有样本点 Silhouette 指标的平均值为 $S_{il\text{-}av}=mean[sum(S_{il}(t))]$，$S_{il\text{-}av}$ 可以反映整个数据集聚类的质量。$S_{il\text{-}av}$ 的值越大表示聚类质量越高，$S_{il\text{-}av}>0.5$ 说明各个类能明显地分开，$S_{il\text{-}av}<0.2$ 说明缺乏实质性的聚类结构。本文将 $S_{il\text{-}av}$ 作为评价聚类质量的指标。

2.2 基于 LS-SVM 的子模型建模

支持向量机(Support Vector Machine, SVM)[5]是一种小样本学习理论，它基于结构风险最小化原则，具有泛化能力强的优点。最初 SVM 用来解决分类问题，后来被引入到回归建模中。最小二乘支持向量机(Least Squares-Support Vector Machine, LS-SVM)[6]是一种改进的支持向量机。与标准支持向量机相比，LS-SVM 避免了求解非线性规划问题，降低了计算复杂度，求解速度加快。LS-SVM 建模能有效解决子模型训练样本少和工业过程复杂化带来的问题。用于回归问题的 LS-SVM 的描述如下：

设训练数据集$\{(x_i,y_i)|i=1,2,\cdots,N\}$，$N$为训练样本数，$x_i\in\boldsymbol{R}^N$是第 i 个样本的输入，$y_i\in\boldsymbol{R}$ 是对应第 i 个样本的输出。输入空间 $\boldsymbol{R}^N$ 通过非线性函数 $\varphi(x_i)$被映射到一个高维的特征空间 $\boldsymbol{Z}$。在 $\boldsymbol{Z}$ 中采用形如式(4)的表达式估计未知的非线性函数，其中，$\boldsymbol{w}$ 和 b 是待定参数。

$$y(x)=\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b,\boldsymbol{w}\in\boldsymbol{Z},b\in\boldsymbol{R} \tag{4}$$

LS-SVM 的优化问题被定义为：

$$\min_{\boldsymbol{w},e}J(\boldsymbol{w},e)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{\gamma}{2}\sum_{i=1}^{N}e_i^2,\gamma>0 \tag{5}$$

满足等式约束：

$$y_i=\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b+e_i,i=1,2,\mathrm{L}\ ,N \tag{6}$$

其中，目标函数的第 1 项对应于模型的泛化能力；第 2 项代表了模型的精确性；γ 是模型泛化能力和精度之间的一个折中参数，可以人为调整；e_i 是第 i 个数据的实际输出和预测输出间的误差。

解上述优化问题的 Lagrange 函数，得到最优解为：

$$\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0 & \boldsymbol{1}^{\mathrm{T}}\\ \boldsymbol{1} & \boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{I}\end{bmatrix}^{-1}\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \tag{7}$$

其中，向量 $\boldsymbol{y}=[y_1\ y_2\ \cdots\ y_N]^{\mathrm{T}}$；$\boldsymbol{1}=[1\ 1\ \cdots\ 1]^{\mathrm{T}}$；$\boldsymbol{\alpha}=[\alpha_1\ \alpha_2\ \cdots\ \alpha_N]^{\mathrm{T}}$；$\boldsymbol{\Omega}$ 是一个 $N\times N$ 对称矩阵；$\Omega_{ij}=\boldsymbol{\varphi}(x_i)^{\mathrm{T}}\boldsymbol{\varphi}(x_j)=K(x_i,x_j), i,j=1,2,\cdots,N$，$K(\cdot,\cdot)$为核函数，最终得到的 LS-SVM 模型表达式为：

$$y(x)=\sum_{i=1}^{N}\alpha_iK(x,x_i)+b \tag{8}$$

各 LS-SVM 子模型建立后，测试样本就可以根据相似性的度量结果，选择不同的子模型测试输出，并将其作为最终的输出。

2.3 多模型建模步骤

基于仿射传播聚类的 LS-SVM 的多模型建模的步骤如下：

(1)使用 AP 聚类算法将训练样本聚类。首先将样本数据进行归一化处理，并初始化 AP 聚类算法的参数 λ 和 p。λ 的值可以根据对象大小在 0.5~0.9 之间设置，偏向参数 p 取 $p_m=median(\boldsymbol{S})$。

为了使聚类效果更好，根据 $S_{il\text{-}av}$ 指标调整 p，p 的下降步幅设置为[4]：

$$p_{\mathrm{step}}=\frac{0.01p_m}{0.1\sqrt{k+50}} \tag{9}$$

其中，k 为聚类中心数。选取 $S_{il\text{-}av}$ 最大时的聚类结果作为各子模型的训练样本。

(2)子模型采用 LS-SVM 训练建模并确定其模型参数。本文选用 RBF 核函数 $K(x,x_i)=\exp\{-\|x-x_i\|^2/\sigma^2\}$。采用交叉验证法优化子模型的参数 γ 和 σ，其中，模型预测能力指标为：

$$P_{\mathrm{PRESS}}=\sum_i(y_i-y_{-i})^2 \tag{10}$$

其中，y_i 为样本 i 的实际值；y_{-i} 为用除掉第 i 个样本的训练数据建立的模型对 y_i 的预测值。

(3)模型测试。对于新的测试样本 x_t，计算它与步骤(1)中的聚类中心的相似度，将 x_t 归为与其相似度最大的类中，并用相应的子模型预测输出。本文采用欧氏距离测量相似度。

3 实验和分析

将本文算法用于某石化企业丙烯精馏塔装置[7]的建模实验，该装置如图 2 所示。其作用是将进料分离成 2 个部分：(1)塔顶出聚合级丙烯产品；(2)塔釜采出 C_3 液化气产品。塔顶丙烯浓度是重要的质量指标，离线分析的方法往往不能及时地提供化验数据，因此，采用本文提出的方法进行多模型建模并估计。

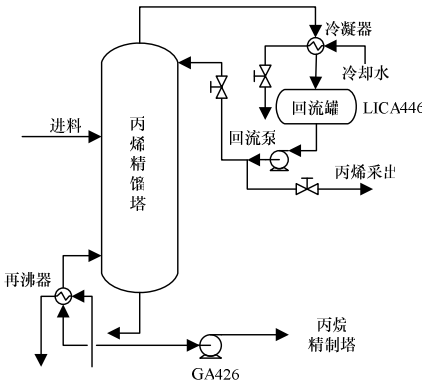

图 2 丙烯精馏塔

根据工程经验初选回流量、塔釜液位、回流罐液位、进料压力、塔釜压力、塔顶温度、进料温度、回流温度等 8 个输入变量，如表 1 所示，塔顶丙烯浓度为输出变量。样本为 2008 年 9 月-12 月现场数据剔除空值后的 151 组数据。

表 1　输入变量描述

输入变量	描述
FIC481	回流量/(t·h^{-1})
LIC445	塔釜液位/(V%)
LICA446	回流罐液位/(V%)
PDI460	进料压力/kpa
PRCA461	塔釜压力/Mpa
TUI482	塔顶温度/(℃)
TUI483	进料温度/(℃)
TUI486	回流温度/(℃)

采用本文第 2.3 节介绍的多模型建模和测试步骤进行软测量建模仿真实验。为方便比较，根据对象的工况先验知识，取原样本序列的 41 组~130 组作为训练样本，剩下的 61 组作为测试样本。

首先，用仿射传播聚类的方法对训练样本聚类，通过调节偏向参数 p，得到训练样本在分为 2 类时的 Silhouette 指标最高，$S_{il\text{-}av}$=0.6105。用这 2 类样本分别建立 2 个子模型，并用交叉验证法确定子模型的参数 γ 和 σ。

然后，根据各测试样本与训练样本各聚类中心的距离，将测试样本归类，并用相应的子模型预测输出。

最后，与其他建模方法进行比较，包括 k 均值 LS-SVM 多模型建模(k 均值分类参数取 2)、AP-RBF 神经网络(训练误差取 *goal*=0.09)多模型建模、基于 LS-SVM 单模型建模。基于 AP 的 LS-SVM 多模型的丙烯浓度预测曲线如图 3 所示，其中前 90 组的内推结果也画出。

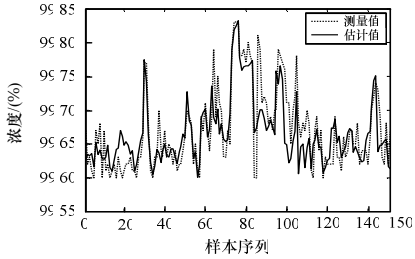

图 3　AP+LS-SVM 多模型对丙烯浓度的预测结果

采用泛化均方根误差评价模型的预测性能[8]:

$$r_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i}^{n}(y_i-f(x_i))^2} \tag{11}$$

其中，y_i 和 $f(x_i)$分别为测试样本的实际值和模型预测值；n 为测试样本数。根据实验结果，各模型预测的泛化均方根误差在表 2 中列出。

表 2　各模型预测误差

模型	误差
k 均值-LS-SVM 多模型	0.040 692 1
AP-LS-SVM 多模型	0.037 764 9
AP-RBF_NN 多模型	0.040 207 7
LS-SVM 单模型	0.053 376 5

通过比较和分析表 2 的泛化均方根误差，可以看出对于多工况的工业对象，多模型的回归精度更高。对于分类后子模型的小样本建模，基于 LS-SVM 的多模型比基于神经网络的多模型有更好的泛化能力。k 均值聚类分类参数取最优分类数时，聚类的效果仍然比 AP 聚类差些，从而使模型的预测精度变差。

4　结束语

本文针对复杂工业过程中单模型建模精度低和泛化性能差的问题，引入了基于仿射传播聚类的 LS-SVM 多模型建模方法。将该方法应用于丙烯精馏塔塔顶丙烯浓度的软测量建模中，实验结果表明该方法有较好的预测性能，具有一定理论指导意义和工程应用价值。针对实验过程中交叉验证法确定模型参数速度慢的问题，下一阶段的工作将是寻找合适的智能优化算法优化模型参数。

参考文献

[1] 李修亮, 苏宏业, 褚　健. 基于在线聚类和关联向量机的多模型软测量建模[J]. 化工自动化及仪表, 2008, 35(3): 34-37.

[2] 李雅芹, 杨慧中. 基于仿射传播聚类和高斯过程的多模型建模方法[J]. 计算机与应用化学, 2010, 27(1): 51-54.

[3] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814): 972-976.

[4] 王开军, 张军英, 李　丹, 等. 自适应仿射传播聚类[J]. 自动化学报, 2007, 33(12): 1242-1246.

[5] Vapnik V N. Statistical Learning Theory[M]. [S. l.]: John Wiley, 1998.

[6] Li Lijuan, Su Hongye, Chu Jian. Modeling of Isomerization of C8 Aromatics by Online Least Squares Support Vector Machine[J]. Chinese Journal of Chemical Engineering, 2009, 17(3): 437-444.

[7] 胡云苹, 赵英凯. 基于 3MAD-PCA 的软测量数据过失误差侦破[J]. 计算机工程与设计, 2010, 31(1): 184-186.

[8] 穆朝絮, 梁瑞鑫, 李训铭. 非线性系统的 LSSVM 联合逆控制器[J]. 计算机工程, 2009, 35(12): 181-183.

编辑　顾逸斐

(上接第 168 页)

参考文献

[1] Fukunaga K. Introduction to Statistical Pattern Recognition[M]. 2nd ed. [S. l.]: Academic Press, 1990.

[2] 范　燕. 对称 LDA 及其在人脸识别中的应用[J]. 计算机工程, 2010, 36(1): 201-202.

[3] Cai Deng, He Xiaofei, Han Jiawei. Semi-supervised Discriminant Analysis[C]//Proc. of IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: [s. n.], 2007: 1-7.

[4] Wright J, Ganesh A, Yang A. Robust Face Recognition via Sparse Representation[J]. IEEE Trans. on Pattern Analysis Machine Intelligence, 2009, 31(2): 210-227.

[5] Donoho D, Tsaig Y. Fast Solution of l1-norm Minimization Problems When the Solution May Be Sparse[D]. Stanford, USA: Stanford University, 2006.

[6] 黄　勇. 基于多种主元分析与信息融合的人脸表情识别[D]. 江门: 五邑大学, 2007.

编辑　顾逸斐