改进谱聚类算法在多模型软测量中的应用

Application of Improved Spectral Clustering Algorithm in Multi-model Soft Sensing

王灿灿　李丽娟

(南京工业大学自动化与电气工程学院,江苏 南京　211816)

摘　要：针对工业过程的非线性、多工况的特点,提出了一种基于改进谱聚类算法的多模型软测量建模方法。采用改进的谱聚类算法对样本数据集进行聚类;根据最小二乘支持向量机(LS-SVM)算法对各类样本建立子模型,并采用粒子群算法(PSO)求解多模型的权值;将所建子模型按照"加权方式"进行组合,得到软测量模型。仿真试验表明,该方法具有较高的模型精度和良好的泛化性能。

关键词：多模型　谱聚类　软测量　粒子群算法(PSO)　最小二乘支持向量机(LS-SVM)

中图分类号：TP274　　**文献标志码**：A

Abstract: In accordance with the features of nonlinear and multiple working conditions of the industrial processes, the multi-model soft sensing modeling method based on improved spectral clustering algorithm is proposed. The sampling data are clustered by using the improved spectral clustering algorithm, and the sub-models of various types of samples are built in accordance with least square-support vector machine(LS-SVM) algorithm, the weights of multiple models are solved by using particle swarm optimization (PSO), then the sub-models are combined in accordance with the "weighted mode" to obtain soft sensing model. The simulation tests show that the method proposed possesses higher model accuracy and excellent generalization performance.

Keywords: Multiple models　Spectral clustering　Soft sensing　Particle swarm optimization(PSO) algorithm　Least square-support vector machine(LS-SVM)

0　引言

在化工过程和很多其他工业应用领域中,由于大多数系统存在机理复杂、高度非线性、强耦合、大时滞等特点,采用单一的软测量模型无法全面地描述复杂系统的全局特性,并且存在回归精度低和泛化能力差等问题。

为了解决上述问题,一种能够提高系统模型精度和泛化能力的多模型软测量建模方法应运而生[1-3]。仲蔚等[4]提出模糊C均值聚类和径向基核函数(radial basis function,RBF)网络相结合的策略来进行多模型建模。现场应用表明,该方法易于实现且具有更好的泛化结果和预报精度。周立芳等[5]提出基于K均值聚类算法的多模型预测控制,试验证明了多模型建模的模型精度和泛化特性。然而,传统的聚类算法如K均值算法、模糊C均值算法等都是建立在凸球形的样本空间上,当样本空间不为凸球形时,算法将会陷入局部最优。

针对传统聚类方法存在的问题,本文提出一种基于改进谱聚类的多模型建模算法。该算法具有识别非凸分布聚类的能力,不会陷入局部最优解,且能避免数据的维数过高所造成的奇异性问题,以便得到更加精确的聚类结果,提高模型精度。样本聚类后,采用最小二乘支持向量机(least square-support vector machine,LS-SVM)建立各子类模型,并采用粒子群(particle swarm optimization,PSO)算法对多模型权值进行寻优,系统软测量模型输出可视作各子模型的加权组合。本文所研究方法在丙烯精馏塔塔顶丙烯含量软测量中进行了应用研究,结果表明,该方法具有较高的精度和良好的泛化性能。

1　谱聚类算法

1.1　标准谱聚类算法

谱聚类是建立在图论中谱图理论的基础上[6],将聚类问题转化为一个无向图的最优划分问题的过程,其本质是通过Laplacian Eigenmap实现降维的过程。谱聚类的思想来源于谱图划分理论,将每个样本数据看作图中的顶点V,根据样本间的相似度将相应顶点之间的连接边E赋权重W,从而得到基于样本相似度的无向加权图$G=(V,E,W)$。此时,可以将聚类问题转化为在图G上的图划分问题。根据图论的划分理论来看,谱聚类的本质就是使得划分后的子图之间相似度最小[7],子图内部相似度最大。

国家自然科学基金资助项目(编号:61203072);

工业控制国家重点实验室开放课题基金资助项目(编号:ICT1234)。

修改稿收到日期:2013-11-21。

第一作者王灿灿(1988-),男,现为南京工业大学控制工程专业在读硕士研究生;主要从事工业过程先进控制、过程建模及优化等方向的研究工作。

根据准则函数和谱映射方法的不同,谱聚类算法有多种实现方法。实现过程一般分为三步:①定义数据样本点之间的相似性度量,建立数据点之间的相似矩阵;②通过计算相似矩阵的前 k 个特征值与特征向量,构建新的数据特征空间;③采用 K 均值或者其他传统聚类算法对特征空间中的特征向量进行聚类。

虽然 K 均值算法实现简单,可以重复运行而不受初始化的影响,但是 K 均值算法也有一些致命的缺点,如不能处理非球型簇以及不同尺寸和不同密度的簇,对于含有离群点的聚类也不太适应。同时,K 均值算法是一种贪心算法,经常会陷入局部最优解。然而,谱聚类算法只需要数据之间的相似度矩阵,不必像 K 均值算法那样要求数据必须是 N 维欧氏空间向量。因此,本文采用改进谱聚类算法来克服 K 均值算法的缺点,从而得到准确的聚类结果,提高模型精度。

1.2　改进的谱聚类算法

在谱聚类算法中,特征值和特征值向量的选择对聚类结果影响很大,而特征值和特征向量的选择以聚类分组数为依据[8]。因此,本文采用基于特征差值与正交特征向量的改进谱聚类算法,对标准谱聚类算法进行改进,实现聚类数目的自动确定。谱聚类算法的基本思想是:先利用样本数据构建相似矩阵,然后对由相似矩阵生成的规范化相似矩阵进行谱分解,从而得到相应的特征值和特征向量;随后对特征值按降序排列,并用本征间隙来表述相邻特征值之间的差,通过第一个极大本征间隙出现的位置来自动确定类个数;最后结合获得的类个数和特征向量之间的夹角实现数据分类。本文选择规范割集划分准则来实现谱聚类算法。

设输入数据集为 $S=\{S_1,S_2,\cdots,S_n\}$,且 S 为 R^n 空间中待聚类的数据集;输出为聚类分组数 k 和聚类结果。聚类算法具体步骤如下。

① 构造数据集的亲和(相似)矩阵 $\boldsymbol{W}_{ij}$, $\boldsymbol{W}_{ij}=\exp(-\|S_i-S_j\|^2/2\delta^2)$,$i\neq j$,且 $W_{ii}=0$,其中 δ 为阈值参数。

② 构造拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}$,$\boldsymbol{D}$ 为对角矩阵,其对角元素为 $D_{ii}=\sum W_{ik}$。

③ 求解拉普拉斯矩阵 $\boldsymbol{L}$ 的特征值$(\lambda_1,\lambda_2,\cdots,\lambda_n)$及其对应的特征向量$(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\cdots,\boldsymbol{\eta}_n)$。

④ 计算特征值差值 $g_1,g_2,\cdots,g_{n-1}$,其中 $g_i=\lambda_i-\lambda_{i+1}$。

⑤ 计算本征间隙序列$\{g_1,g_2,\cdots,g_{n-1}\mid g_i=\lambda_i-\lambda_{i+1}\}$,本征间隙序列中第一个极大值对应的下标就是类个数 k,即 $k=\arg\min_i\{g_i-g_{j\mid j<i}>0\&g_i-g_{i+1}>0\}$。

⑥ 构造矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k]$,其中 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k$ 为拉普拉斯矩阵 $\boldsymbol{L}$ 的前 k 个特征值对应的特征向量。

⑦ 对矩阵 $\boldsymbol{X}$ 中的每一行进行单位化处理,得到矩阵 $\boldsymbol{Y}$。

⑧ 将 $\boldsymbol{Y}$ 中的每一行看作 R^k 空间中的一个点,并对其使用 K 均值算法,得到 k 类样本子集,如果矩阵 $\boldsymbol{Y}$ 中第 i 行属于第 j 类,则 X_i 也属于第 j 类。

根据矩阵的摄动理论,当第 k 和$(k+1)$个特征值之间的差值越大时,所选的 k 个特征向量构成的子空间就越稳定。此时,以矩阵 $\boldsymbol{Y}$ 中的每一行作为 k 维空间中的一个点,形成 k 个聚类。它们将彼此正交地分布于 k 维空间中的单位球上,且在单位球上形成的这 k 个聚类对应着原来空间中所有点形成的 k 个聚类。由此根据拉普拉斯矩阵的特征值之间的差值来确定类个数,其中差值取[0.06,0.1]。

与标准谱聚类算法相比,改进后的谱聚类算法可以自动确定聚类个数[9],对样本数据建立规范化相似矩阵并进行谱分解;利用本征间隙自动确定样本数据的类个数;根据确定的类个数和谱分解后的特征向量间的夹角,实现样本数据的分类。

2　基于 LS-SVM 的子模型建模

支持向量机是一种小样本学习理论,它的基本思想是采用结构风险最小化原则构造最优决策函数[10],解决样本空间中高度非线性回归问题。最小二乘支持向量机(LS-SVM)是一种改进的支持向量机,它将求解二次规划问题转化为求解线性方程组[11],解决了一般标准支持向量机求解凸二次规划问题所带来的复杂度的问题,提高了学习速度。LS-SVM 算法的优化问题描述如下。

$$\begin{cases}\min Q(\boldsymbol{\omega},\boldsymbol{e})=\dfrac{1}{2}\|\boldsymbol{\omega}\|^2+\dfrac{\gamma}{2}\sum_{i=1}^{l}e_i^2\\ y_i=\boldsymbol{\omega}\varphi(x_i)+b+e_i,\ i=1,2,\cdots,l\end{cases}\tag{1}$$

式中:$\varphi(x_i)$为核空间映射函数;$\boldsymbol{\omega}$ 为权矢量;e_i 为误差变量;b 为偏差量;γ 为正则化参数。

为了求解上述优化问题,需要把有约束优化问题变为无约束优化[14]。为此建立相应的 Lagrange 函数:

$$L=Q(\boldsymbol{\omega},\boldsymbol{e})-\sum_{i-1}^{l}\alpha_i[\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i)+b+e_i-y_i]\tag{2}$$

根据 KKT(Karush-Kuhn-Tucher)最优条件,得到如下线性方程组:

$$\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\\ \boldsymbol{e} & \Omega+\dfrac{I}{\gamma}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix}\tag{3}$$

式中:$\boldsymbol{y}=y(y_1,y_2,\cdots,y_l)$;$\boldsymbol{e}=[1,1,\cdots,1]^{\mathrm{T}}$;$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)^{\mathrm{T}}$ 为拉格朗日乘子;$\Omega_{ij}=\varphi^{\mathrm{T}}(x_i)\varphi(x_j)=K(x_i,x_j)$ 为核函数(核函数采用径向基核函数)。

因此,LS-SVM 算法的优化问题转化为求解线性方程组(3),最终可得到 LS-SVM 的模型表达式为:

$$y(x)=\sum_{i=1}^{l}\alpha_i K(x,x_i)+b \tag{4}$$

3　软测量多模型建模

本文软测量多模型建立步骤如下。首先,采用改进谱聚类算法对训练样本集 X_1 聚类,得到 n 个类别,对各类建立 LS-SVM 子模型;然后,根据欧氏距离将测试样本集 X_2 中的测试样本点划分到相应的子模型中,得到相应的子模型输出 $y_i(i=1,\cdots,n)$;最后,将各子模型按照"加权方式"进行组合,得到系统模型输出 Y,完成多模型的建立。基于改进谱聚类算法的软测量多模型结构如图 1 所示。

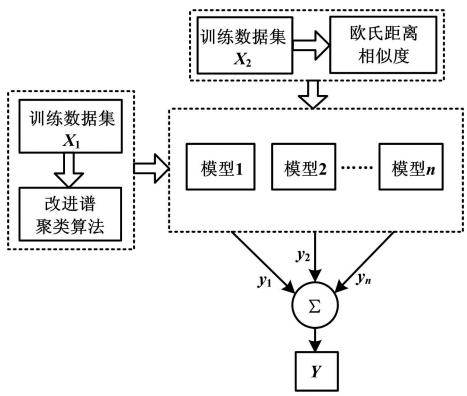

图 1　多模型软测量系统结构图

Fig. 1　System structure of multiple models soft-sensing

在辨识出系统各个子模型的基础上,按照"加权方式"进行组合,得到系统的软测量模型输出。描述如下:

$$f(t)=\boldsymbol{W}^{\mathrm{T}}Z(t) \tag{5}$$

式中:$f(t)$ 为模型输出结果;权值向量 $\boldsymbol{W}=(w_1,w_2,\cdots,w_i)$,$w_i$ 为每个子模型的加权值,且满足 $w_1+\cdots+w_n=1$,$w_i\geqslant 0,i=1,\cdots,n$。

粒子群算法是一种适应性较强的全局优化算法,能够快速找到合适的权向量[12]。本文采用粒子群算法对多模型权值进行寻优,算法步骤如下。

① 初始化粒子种群及参数,设定多模型中加权系数个数 n、粒子数目 c;对于第 i 个粒子,将每个加权系数作为粒子 i 的位置编码;计算各粒子的适应度,设置粒子 i 的初始速度为 0。反复进行,生成 m 个粒子。

② 由初始化粒子群得到粒子的个体最优位置 $P_{\mathrm{id}}(i)$ 和全局最优位置 P_{gd}。

③ 更新初始化粒子的速度和位置,惯性因子 ξ 按式(6)计算:

$$\xi=\xi_{\max}-\frac{\xi_{\max}-\xi_{\min}}{D_{\max}}D \tag{6}$$

式中:D 为当前迭代次数;$D_{\max}$ 为最大迭代次数;$\xi_{\max}=1,\xi_{\min}=0$。

④ 对于每一个粒子 i,比较它们的适应度函数和经历过的最好位置的适应度值 $P_{\mathrm{id}}(i)$,若更好,则更新 $P_{\mathrm{id}}(i)$。

⑤ 对于每个粒子 i,比较它们的适应度值和群体所经历的最好位置 P_{gd} 的适应度值,若更好,则更新 P_{gd}。

⑥ 检查终止条件(是否到达设定迭代次数)。若条件满足,迭代终止,输出全局最优加权解,否则返回步骤③。

4　仿真实例

丙烯是重要的石油化工基础原料,用于生产聚丙烯、苯酚、丙酮等。丙烯精馏塔中丙烯浓度是重要的质量指标,人工采样离线分析的方法存在长时间滞后问题,不利于生产过程的在线检测与控制。因此,本文将基于改进谱聚类的多模型软测量建模方法应用于丙烯生产过程中质量指标的预测。

根据丙烯生产工艺,选择塔顶温度、进料量温度、回流温度、进料压力、塔釜压力、回流量、塔釜液位以及回流罐液位作为输入变量,丙烯含量作为输出变量。将现场采集的样本数据进行异常样本数据的剔除,对输入变量的样本数据进行归一化处理,得到 150 组样本数据,其中 100 组用于训练模型,50 组作为测试。采用改进谱聚类算法进行聚类的步骤如下。

首先,用改进谱聚类算法对训练样本聚类,样本数据被自动聚为 3 类,分别建立子模型;然后根据欧氏距离将测试样本归类,用相应的子模型预测输出;最后,根据粒子群算法求解多模型权值,将建立的子模型按照"加权方式"组合,得到丙烯质量指标的软测量模型。

为了验证本文方法的有效性,分别采用基于 K 均值-LS-SVM 多模型建模(分类参数取为 2)和基于 LS-SVM 单模型建模,然后与本文方法进行比较。同时,采用均方根误差(root-mean-square error,RMSE)和最大绝对误差(maximum absolute error,MAXE)来评价模型预测性能。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i}^{n}[f(x_i)-y(x_i)]^2} \tag{7}$$

$$MAXE=\max\left(\sum_{i}^{n}\frac{|f(x_i)-y(x_i)|}{y(x_i)}\right) \tag{8}$$

式中:$f(x_i)$ 和 $y(x_i)$ 分别为模型的输出值和真实值;n 为样本个数。

根据试验结果,3 种方法的模型测试误差比较结果如表 1 所示。

表 1　模型预测误差

Tab. 1　Predictive errors of models

方法	*RMSE*	*MAXE*
单一 LS-SVM	0.049 73	0.129 05
K 均值-LS-SVM	0.040 51	0.123 17
本文方法	0.033 62	0.102 53

精馏塔塔顶丙烯质量指标的 3 种方法模型预测结果如图 2 ~ 图 4 所示。

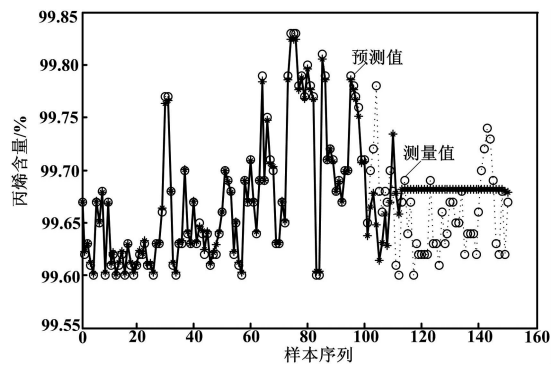

图 2　单一 LS-SVM 模型预测

Fig. 2　LS-SVM single model prediction

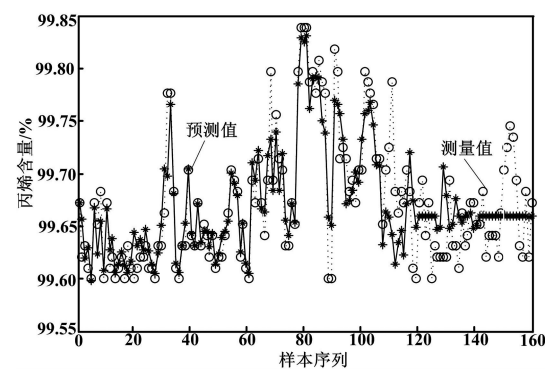

图 3　*K* 均值 LS-SVM 多模型预测

Fig. 3　*K*-means LS-SVM multiple models prediction

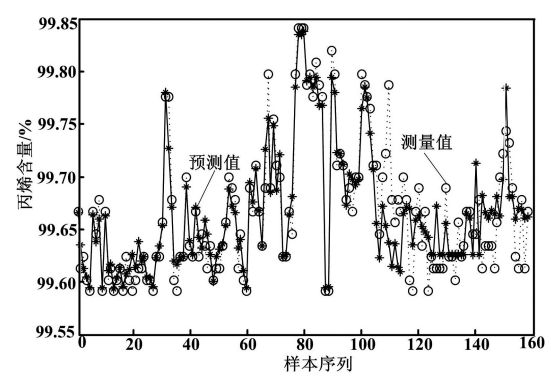

图 4　改进谱聚类 LS-SVM 多模型预测

Fig. 4　Improved spectral clustering LS-SVM multiple models prediction

由表 1 可以看出,采用本文提出的方法得到的模型均方根误差(RMSE)和最大绝大值误差(MAXE)均体现了该方法的优势。

比较图 2、图 3 和图 4 的预测结果可知,本文方法建立的模型较其他建模方法具有更好的跟踪效果,说明了本文方法的有效性。

5　结束语

本文提出的基于 SC-LS-SVM 多模型建模方法,通过改进谱聚类算法对样本数据进行聚类,对各子类样本建立 LS-SVM 子模型,采用粒子群算法对多模型权值进行寻优,并将子模型按照“加权方式”进行组合,得到系统软测量模型。将该方法应用于丙烯精馏塔塔顶丙烯含量的软测量建模中,通过仿真试验对比,本文方法可以较好地跟踪丙烯质量指标的变化,具有更好的模型预测精度。

参考文献

[1] 王孝红,刘文光,于宏亮. 工业过程软测量研究[J]. 济南大学学报:自然科学版,2009,23(1):80-86.

[2] Petr K,Bogdan G,Sibylle S. Data-driven soft sensors in the process industry[J]. Computers and Chemical Engineering,2009,33(4):795-814.

[3] Fortuna L, Graziani S, Rizzo A. Soft sensors for monitoring and control of industrial processes[M]. London:Springer,2007.

[4] 仲蔚,俞金寿. 基于模糊 c 均值聚类的多模型软测量建模[J]. 华东理工大学学报:自然科学版,2000,26(1):83-87.

[5] 周立芳,张赫男. 基于聚类多模型建模的多模态预测控制[J]. 化工学报,2008,59(10):2546-2552.

[6] Luxburg U V. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.

[7] 戴月明,高倩. 自适应半监督模糊谱聚类算法[J]. 计算机工程与应用,2010,46(33):212-214.

[8] Zhao F,Jiao L C,Liu H Q,et al. Spectral clustering with eigenvector selection based on entropy ranking[J]. Neurocomputing,2010,73(10-12):1704-1717.

[9] Ng A Y,Jordan M I,Weiss Y. On spectral clustering:analysis and an algorithm[C]//Cambridge:MIT Press,2002:121-526.

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

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

[12] Kennedy J, Eberhart R C. Particle swarm optimization[C]// Proceedings of 1995 IEEE International Conference on Neural Networks,New York,1995:1942-1948.