研究简报

改进 k-means 聚类算法多模型建模的一种新的评价函数

周立芳，周芦文，赵豫红
（浙江大学信息学院控制系工业控制研究所，浙江 杭州 310027）

关键词：k-means 聚类；性能评价函数；pH 中和过程；偏最小二乘算法
中图分类号：TP 27 **文献标识码**：A **文章编号**：0438－1157（2007）08－2051－05

Multi-modeling of pH neutralization processes using improved k-means clustering based on new validity function

ZHOU Lifang, ZHOU Luwen, ZHAO Yuhong
(*Institute of Industrial Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou* 310027, *Zhejiang, China*)

Abstract: The modeling and control of pH neutralization processes is a difficult problem in the field of process control. A multi-modeling method using an improved k-means clustering based on a new validity function is proposed in this paper. There are some common problems, including the number of clusters assumed as a priori knowledge and initial cluster centers selected randomly for classical k-means clustering. The proposed algorithm is used to compute initial cluster centers and a new validity function is added to determine the appropriate number of clusters, then partial least squares (PLS) is used to construct the regression equation for each local cluster. Simulation results showed that multiple models using the proposed algorithm gave good performance, and the feasibility and validity of the proposed algorithm was verified.

Key words: k-means clustering algorithm; validity function; pH neutralization processes; partial least squares

引 言

过程系统的控制、仿真与优化往往都是依赖于高性能的模型。特别是对于基于模型的控制方案中，模型不仅要高精度地拟合过程的稳态特性，还必须具有大范围描述过程动态行为的能力。对于化工非线性对象 pH 中和过程，由于 pH 中和滴定曲线的严重非线性、pH 反应的滞后性以及外部干扰的复杂性，使得其成为过程控制中典型的控制难题。因此，获得良好的 pH 模型将有助于提高 pH 中和过程的控制品质。Buchholt 等[1]采用线性差分方程作为系统模型来描述 pH 过程；Karr 等[2]对系统建立模糊模型；Norquay 等[3]采用 Wiener 模型对 pH 过程进行模型结构辨识；Nie 等[4]则采

2006－09－08 收到初稿，2007－04－02 收到修改稿。
联系人及第一作者：周立芳（1973—），女，副教授。
基金项目：国家自然科学基金项目（60503065）。

Received date: 2006－09－08.
Corresponding author: ZHOU Lifang, associate professor. **E－mail**: lfzhou@iipc.zju.edu.cn
Foundation item: supported by the National Natural Science Foundation of China (60503065).

用模糊神经网络对 pH 过程进行建模和辨识。这些模型从不同角度反映了 pH 中和过程的动态特性。

聚类算法[5]是一种被广泛关注与研究的数据挖掘技术，已经被广泛地应用于许多领域[6-12]，如商务上分析客户群特征，生物学上对基因进行分类等，在过程建模领域，聚类算法也早有所应用。它能够作为一个独立的工具来分析数据分布的情况，观测每一簇的特点，集中针对特定的簇进行分析，因此它能够为多模型建模提供一种划分子空间区域的准则。

本文采用基于优化性能指标的聚类方法对 pH 过程进行模型辨识。首先针对 k-means 聚类算法当中的聚类个数难以事先给定以及算法对初始点的强依赖性等缺点，对传统的 k-means 聚类算法，利用一种新的评价函数对聚类结果进行判断，从而实现结构参数 k 的优化，同时对 k-means 聚类的初始点进行优化选取；然后针对每一个子区域采用偏最小二乘回归算法（PLS）辨识模型，获得 pH 中和过程的多模型，最后进行仿真研究，验证了改进算法的可行性以及有效性。

1 基于一种新评价函数聚类算法的多模型建模方法

k-means 聚类算法由 Mac Queen[13] 提出，具有算法结构简单、收敛速度快的优点，十分适用于大规模的数据分析。但是 k-means 聚类算法具有两大突出缺点：一是必须事先已知或者给定簇的个数 k；二是 k-means 聚类算法是一种迭代算法，其聚类效果的好坏与初始点的选取有着相当密切的关系。由此，本文提出一种基于结构参数 k 优化的 k-means 聚类算法的多模型建模方法。

1.1 算法基本思想

该算法的基本思想是先设定 k-means 聚类的初始分类个数 $k=2$，然后根据 CCIA 算法[14] 初始化聚类中心，获取数据样本的初始分类，然后再利用基于密度的多尺度空间数据融合算法（DBMSDC）[15] 对初始分类进行相似簇的融合过程，获取对应 k 值下的样本簇分类结果，按所定义的评价函数进行聚类结果的性能指标值计算，如果性能指标值增加，则认为目前的聚类效果比之前的好，增加新的簇是有利的，则进入下一个循环当中，直到性能指标值随着 k 的增加出现负增长，结束聚类过程，进入模型结构和参数的辨识阶段，利用偏最小二乘回归 PLS 对子区域建立子模型，从而构成系统的多模型集。

1.2 一种新的评价函数

由于聚类算法是一种无监督方法，所以通常都会采用性能评价函数对其聚类结果进行有效性判断。如果所选取的评价函数无法正确地反映出聚类的质量，将会大大减弱聚类算法的有效性和正确性。因此，根据样本数据本身的结构特征以及本文算法的特殊性，将自定义一种新的性能评价函数引入到改进的 k-means 聚类算法当中，实现结构参数 k 值的优化。

自定义的性能评价函数分为两部分：

第一，聚类目的是希望将数据对象分组成为多个类或簇，在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。因此一个好的分类结果应该使得类间相异度大，类内相似度大。根据聚类的结果，类间相异度定义为

$$R_{\_\text{out}}(c_i,c_j)=\frac{2\sum_{i=1}^{q}\sum_{j=i+1}^{q}d(c_i,c_j)^2}{q(q-1)}$$
$$i,j\in[1,q] \tag{1}$$

类内相异度定义为

$$R(x_{ij},c_j)=\frac{1}{N_j}\sum_{i=1}^{N_j}d(x_{ij},c_j)^2 \tag{2}$$

$$R_{\_\text{in}}=\frac{1}{q}\sum_{j=1}^{q}R(x_{ij},c_j) \tag{3}$$

其中，c_i、c_j 表示簇的质心点；x_{ij} 表示隶属于簇 j 的样本点 i；q 表示目前簇的个数；N_j 表示簇 j 所包含样本点的个数；$d(x,y)$ 用于计算 x 和 y 之间的欧氏距离。

第二，希望所得到的簇基本都能够满足密度阈值的要求且样本点均匀，定义一个关于簇的样本点分布差异度

$$R_{\_N}(N_i,N_j)=\frac{2\sum_{i=1}^{q}\sum_{j=i+1}^{q}(N_i-N_j)^2}{q(q-1)}$$
$$i,j\in[1,q] \tag{4}$$

另外，为了尽量使得不满足密度阈值要求的簇能够按照最近邻的原则归入到相邻的簇当中去，所以再增加一项反映出目前不满足密度阈值的簇的个数 q'，认为其值越大则聚类质量越差。因此，定义聚类效果评价函数为

$$\text{fun}(\cdot)=\widetilde{\omega}_1 f_1+\widetilde{\omega}_2 f_2+\widetilde{\omega}_3 f_3=$$

$$\tilde{\omega}_1 \frac{R_{\_\text{out}}}{1+R_{\_\text{in}}} + \frac{\tilde{\omega}_2}{R_{\_N}} + \frac{\tilde{\omega}_3}{q'+1} \tag{5}$$

2　pH 中和过程仿真算例

本文主要针对多输出 pH 过程进行建模分析。考虑一个包含三股进料的 pH 系统：流入股包括强酸（HNO_3）、强碱（NaOH）以及缓冲流入股（$NaHCO_3$），流出股包括罐的液位 h 以及流出物的 pH 值（见图 1）。

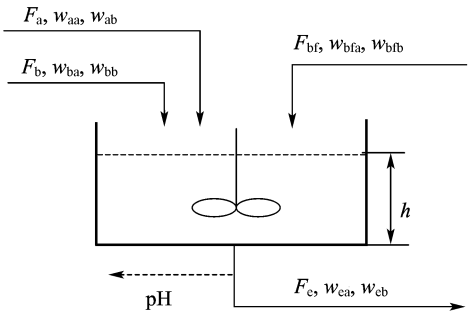


图 1　两输出 pH 中和过程示意图

Fig. 1　pH neutralization processes of two outputs

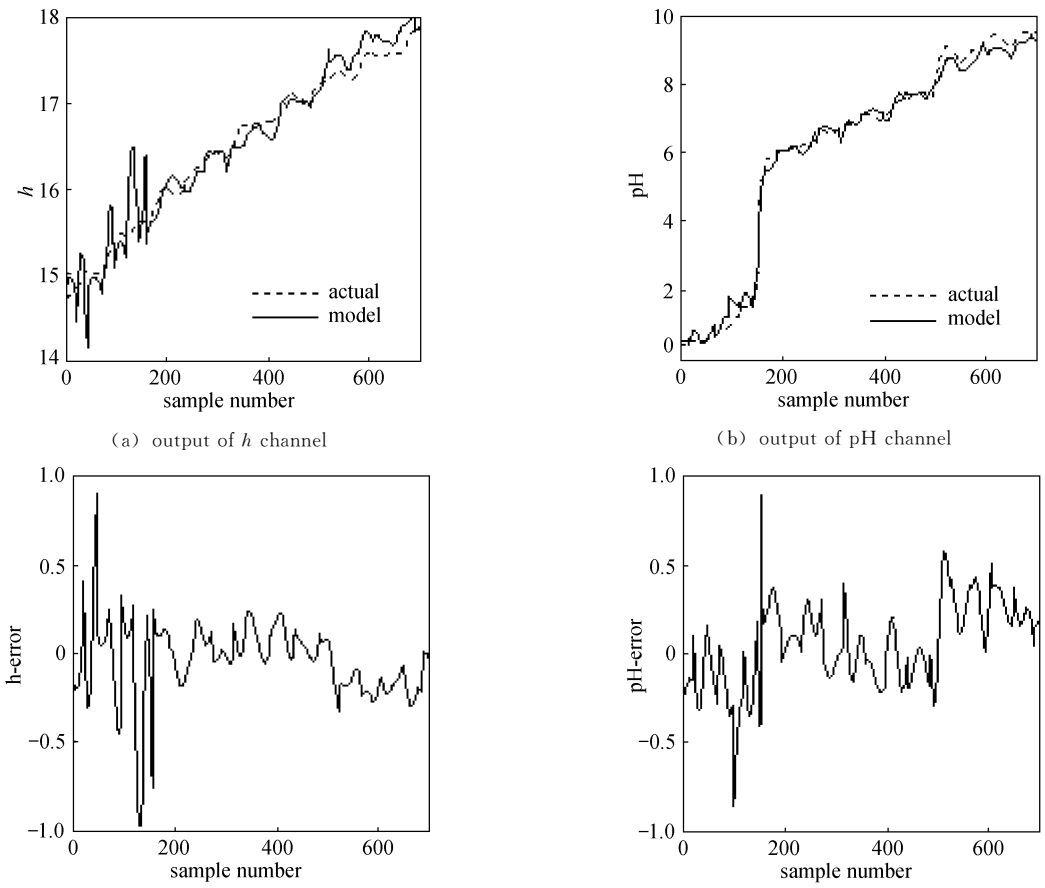
假设 CSTR 是完全混合且处处等温，McAvoy 等[16]给出了图 1 所示 pH 动态数学模型。该模型由两部分组成，动力学方程用于描述化学成分浓度的动态变化，代数方程用于描述这些浓度间的化学平衡条件的非线性静态特性。该系统的输入输出函数关系可以表述如下

$$\begin{aligned}\hat{y}_h(m) = {} & P_{10} + P_{11}F_a(m-1) + \\ & P_{12}F_b(m-1) + P_{13}F_{bf}(m-1) + \\ & P_{14}y_h(m-1) + P_{15}y_{pH}(m-1)\\ \hat{y}_{pH}(m) = {} & P_{20} + P_{21}F_a(m-1) + \\ & P_{22}F_b(m-1) + P_{23}F_{bf}(m-1) + \\ & P_{24}y_h(m-1) + P_{25}y_{pH}(m-1)\end{aligned} \tag{6}$$

设定反应物的流量按式（7）进行变化

$$\begin{aligned}F_a(m) &= 18 + 4\sin(2\pi m t_s/15),\\ F_b(m) &= 20 + 4\cos(2\pi m t_s/25),\\ F_{bf}(m) &= 0.55 + 0.055\sin(2\pi m t_s/10)\end{aligned} \tag{7}$$

图 2 给出了采用本文建模方法两通道的输出跟踪以及辨识误差曲线。从图 2 可以看出，两通道的辨识模型的输出跟踪效果都比较好，只是在子模型



（a）output of h channel　（b）output of pH channel

（c）error for h channel　（d）error for pH channel

图 2　基于本文方法的两输出 pH 过程辨识模型效果

Fig. 2　pH neutralization processes model based on method proposed in this paper

切换的时刻，系统的暂态性能表现得略微差一些，这方面值得进一步研究。

对于建模效果而言，往往通过模型的均方根误差指标

$$\mathrm{RMSE}=\left[\frac{1}{n_1}\sum_{j=1}^{n_1}(y_i-\hat{y}_i)^2\right]^{0.5} \tag{8}$$

来衡量模型拟合效果，采用本文方法得到 h 与 pH 通道模型的 RMSE 与文献［4］得到的结果进行比较，如表 1 所示。

表 1 两输出 pH 过程辨识结果比较

Table 1 Comparison of identification results for pH neutralization processes

Item	Rules		RMSE	
	USOCPN[4]	Method in this paper	USOCPN[5]	Method in this paper
h	44	4	0.114	0.215
pH	44	4	0.217	0.240

可以看出，本文所提出方法拟合的模型数大大小于文献［4］中的模型数，从而可以减少控制器切换次数，且模型的均方根误差指标相差不大。因此，利用本文所提出的方法对非线性系统进行建模，进而设计先进控制器将是下一步研究的方向。

3 结 论

从系统输入输出数据出发，利用改进的 k-means聚类算法对样本空间进行子区域的划分，提出一种新的性能评价函数对聚类结果进行判断，能够快速、准确地对确定系统的子空间划分数目，并采用多因变量的偏最小二乘 PLS 算法对每个子空间进行模型辨识，构建出系统的多模型集。对两输出 pH 过程的仿真结果表明了本文算法是可行的。

符 号 说 明

F——流量，m^3

h——液位高度，m

k——聚类个数

t_s——离散化步长

w——浓度，$mg\cdot L^{-1}$

y——输出实测值

$\hat{y}$——输出估计值

$\tilde{\omega}$——权重系数

下角标

a——a 股进料

aa——a 股进料中 a 组分

ab——a 股进料中 b 组分

b——b 股进料

ba——b 股进料中 a 组分

bb——b 股进料中 b 组分

bf——bf 股进料

bfa——bf 股进料中 a 组分

bfb——bf 股进料中 b 组分

e——流出物

ea——流出物 e 中 a 组分

eb——流出物 e 中 b 组分

References

［1］ Buchholt F，Kummel M. Self-tuning control of a pH-neutralization process. *Automatica*，1979，**15**：665-671

［2］ Karr C L，Gentry E J. Fuzzy control of pH using genetic algorithms. *IEEE Transaction on Fuzzy Systems*，1993，**1**(1)：46-53

［3］ Norquay S L，Palazoglu A，Romagnoli J A. Nonlinear model predictive control of pH neutralization using Wiener models//Proc 13th IFAC World Cong. San Francisco，1996：31-36

［4］ Nie J H，Loh A P，Hang C C. Modeling pH neutralization processes using fuzzy-neural approaches. *Fuzzy Sets and Systems*，1996，**78**：5-22

［5］ Han J W，Kamber M. Data Mining Concepts and Techniques（数据挖掘概念与技术）. Fan Ming（范明），Meng Xiaofeng（孟小峰），trans. Beijing：Mechanic Industry Press，2001：223-261

［6］ Hao Meirui（郝梅瑞）. A study on regional characteristics and structural types of urban family consumption in China. *Consumer Economics*（消费经济），2004，**20**（6）：3-7

［7］ Li Junli（李俊莉），Wang Hui（王慧），Zheng Guo（郑国）. Assessment and clustering analysis of the influences of the development zones on China's urban development. *Human Geography*（人文地理），2006，**21**（4）：39-43

［8］ Liang Hong（梁宏）. The cluster analysis of the regional disparity of the population changes in China. *Population Journal*（人口学刊），2002，**5**：33-37

［9］ Ming Jian（明健），Zheng Huachuan（郑华川），Cui Lei（崔雷），Zhuo Renjie（卓仁杰）. Analyzing the research status of gastric precancerous lesions with the cluster analysis method. *Journal of Medical Intelligence*（医学情报工作），2002，**23**（5）：262-263

［10］ Zhang Han（张晗），Cui Lei（崔雷）. Study of bioinformatics through co-word analysis. *Journal of Medical Intelligence*（医学情报工作），2004，**25**（5）：327-330

［11］ Zhou Xiang（周祥），He Xiaorong（何小荣），Chen

Bingzhen (陈丙珍). Self-clustering algorithm for partitioning ANN samples. *Journal of Chemical Industry and Engineering (China)* (化工学报), 2002, **53** (9): 942-945

[12] Zhou Luwen (周芦文), Zhou Lifang (周立芳). Multiple modeling method based on advanced k-means clustering. *Journal of University of Science and Technology of China* (中国科学技术大学学报), 2005, **35** (suppl.): 62-67

[13] Mac Queen J. Some methods for classification and analysis of multivariate observations//Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, 1967: 281-297

[14] Khan S S, Ahmad A. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letter*, 2004, **25**: 1293-1302

[15] Mitra P, Murthy C A, Sankar K P. Density-based multi-scale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24** (6): 734-747

[16] Mc Avoy T J, Hsu E, Lowenthal S. Dynamics of pH in controlled stirred tank reactor. *Ind. Engng. Chem. Process Des. Develop*, 1972, **11**: 68-70

《化工进展》2007 年第 7 期目次