竞争式模糊聚类学习在函数逼近中的应用*)

黄媛媛　傅　彦
(电子科技大学计算机科学与工程学院　成都 610054)

摘　要　本文通过对RCA算法中遗忘函数的修正,抑制了类间竞争迭代中的病态发散,从而实现了算法的稳健收敛。用该算法分析数据,采用TSK模糊模型对函数进行逼近,仿真表明该方法能有效地排除噪声及孤立点对系统逼近的干扰。

关键词　竞争聚类,模糊聚类,TSK模糊规则,函数逼近,RCA

Competitive Fuzzy Clustering with Application in Function Approach

HUANG Yuan-Yuan　ZHANG Jian　FU Yan
(School of Computer Science and Engineering of UESTC, Chengdu 610054)

Abstract　This paper modifies the loss function of RCA algorithm, gets a robust convergence by restraining the ill condition in the iteration of competitive agglomeration among the clusters. We use this algorithm for function approximation with TSK model. The simulation shows that the algorithm has good performance of restraining the interference of outliers.

Keywords　Competitive clustering, Fuzzy clustering, TSK fuzzy rule, Function approximation, RCA

1　引言

在模糊建模中,对于复杂非线性系统的识别,一般是将系统的输入输出数据空间划分为若干个子集,而对于每个子集构造较简单的局部模型,再通过所有局部模型的集结得到复杂非线性系统的全局模型。目前模糊聚类在对数据空间的划分中得到了广泛的应用。

多数的聚类算法是基于K均值和FCM[1],这些算法一般都要求事先知道分类的数目,并且在有噪声或有孤立点存在的情况下,聚类的效果往往很不理想。这些算法对于初始值及其他参数都较为敏感。近来提出了一种基于类间竞争的模糊学习思想RCA[2],该方法初始时生成大量的类来避免算法对初始化的敏感性,而最终分类的数目通过类间的竞争确定。本文从输入输出数据的分析出发,吸收了RCA的模糊学习思想,并对算法中的遗忘函数进行了合理的修正,抑制了类间竞争迭代中的病态发散,从而实现了算法的稳健收敛。通过典型的函数逼近和非线性系统建模的仿真,表明该算法能有效地排除孤立点对函数逼近的干扰。

2　竞争式模糊聚类学习

RCA竞争式模糊学习采用统计方法来消除孤立点的影响,利用分类间的竞争组合来确定适合的分类数目[2]。令$X=\{x_j|j=1,\cdots,N\}$为N个n维向量,目标准则函数定义如下:

$$J=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^2\rho_i(r_{ij}^2)-\alpha\sum_{i=1}^{C}[\sum_{j=1}^{N}w_{ij}u_{ij}]^2 \tag{1}$$

且满足条件

$$\sum_{i=1}^{C}u_{ij}=1,1\leqslant j\leqslant N \tag{2}$$

其中r_{ij}^2表示输入向量x_j到第i个类的距离,u_{ij}表示输入向量x_j对于第i个类的隶属度,$\rho_i()$是第i类的遗忘函数,$w_{ij}=w_i(r_{ij}^2)=\partial\rho_i(r_{ij}^2)/\partial r_{ij}^2$表示输入向量$x_j$对于$i$类的权值函数。遗忘函数减小了孤立点对整体模型的影响,而权值函数在计算类集合的势时排除掉孤立点。合理地选择距离r_{ij}及参数α,应用目标准则函数可以得到对输入数据的合理聚类。利用拉格朗日方法,由式(1)及(2)可得

$$J=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^2\rho_i(r_{ij}^2)-\alpha\sum_{i=1}^{C}[\sum_{j=1}^{N}w_{ij}u_{ij}]^2-\sum_{j=1}^{N}\lambda_j(\sum_{i=1}^{C}u_{ij}-1) \tag{3}$$

用上式对u_{ij}求导可得

$$\frac{\partial J}{\partial u_{st}}=2u_{st}\rho_s(r_{st}^2)-2\alpha w_{st}\sum_{j=1}^{N}w_{sj}u_{sj}-\lambda_t=0$$

$1\leqslant s\leqslant C,1\leqslant t\leqslant N$

则

$$u_{st}=\frac{2\alpha(w_{st}\sum_{j=1}^{N}w_{sj}u_{sj})+\lambda_t}{2\rho_s(r_{st}^2)} \tag{4}$$

用(4)式和(2)式可以解出λ_t,并代入(4)式就得到了u_{ij}的迭代公式

$$u_{st}=\frac{1/\rho_s(r_{st}^2)}{\sum_{k=1}^{C}1/\rho_k(r_{kt}^2)}+\frac{\alpha w_{st}}{\rho_s(r_{st}^2)}(N_s-\overline{N}_t) \tag{5}$$

其中$N_s=\sum_{j=1}^{N}w_{sj}u_{sj}$表示$s$类集合的势,$\overline{N}_t=(\sum_{k=1}^{C}(N_k/\rho_k(r_{kt}^2)))/(\sum_{k=1}^{C}(1/\rho_k(r_{kt}^2)))$表示$s$类集合的加权平均势。(5)式中的遗忘函数为Tukey的双权函数[3]。

$$\rho_i(\hat{r}_{ij}^2)=\begin{cases}\frac{1}{3}[1-(1-\hat{r}_{ij}^2)^3] & |\hat{r}_{ij}^2|\leqslant 1\\ \frac{1}{3} & |\hat{r}_{ij}^2|>1\end{cases} \tag{6}$$

则权函数为

*)基金项目:国家自然科学基金资助项目(10476006)。**黄媛媛**　硕士研究生,主要研究方向为数据挖掘中的模式识别等。**傅　彦**　教授,硕士导师,主要研究方向为人工智能,数据挖掘等。

$$w_{ij}(\hat{r}_{ij}^2)=\begin{cases}(1-\hat{r}_{ij}^2)^2 & |\hat{r}_{ij}^2|\leqslant 1\\ 0 & |\hat{r}_{ij}^2|>1\end{cases} \tag{7}$$

这里 $\hat{r}_{ij}^2$ 是规则化的误差距离，可以由下式得到

$$\hat{r}_{ij}^2=\frac{r_{ij}-med_i}{\zeta\times MAD_i} \tag{8}$$

其中 med_i 是第 i 类误差的中值，而 MAD_i 是第 i 类误差绝对值的中值[4]。ζ 为调节系数，它的取值与迭代次数有关，随着迭代的增加而减小

$$\zeta(t)=\max\{\zeta_{\min},\zeta(t-1)-\Delta\zeta\} \tag{9}$$

其中 $\zeta(0)=8,\zeta_{\min}=4,\Delta\zeta=1$。由(5)式中可以看出 u_{ij} 的迭代公式实际上用了遗忘函数的倒数，而对于每一类在迭代运算中，总有一些点将位于类的中心，从而使这些点对应的误差 $\hat{r}_{ij}^2$ 趋近于 0，$\rho_i(\hat{r}_{ij}^2)$ 的倒数值将非常大，从而使(5)式中的第二项值出现较大的波动，导致 u_{ij} 的值小于 0 或远大于 1，u_{ij} 将很难正常收敛。为了克服上述的缺陷，这里对遗忘函数的倒数值作了归一化处理

$$\rho'_i(r)=\frac{1}{\rho_i(r)\sum_{k=1}^{C}1/\rho_k(r)} \tag{10}$$

相应 u_{ij} 的迭代公式修改为

$$u_{st}=\rho'_s(r_{st}^2)+\alpha w_{st}\rho'_s(r_{ij}^2)(N_s-\overline{N}_t) \tag{11}$$

其中 $\overline{N}_s=\sum_{k=1}^{C}(\rho'_k(r_{kt}^2)N_k)$。(11)式中参数 α 的选择应使在迭代开始时聚类的速度较慢，从而有利于小类的聚集，接着速度应增快，促使类间的合并。当合并到一定数目时聚类速度应逐渐放慢，使算法趋于收敛。合适的 α 由下式得到

$$\alpha(t)=\eta e^{\frac{-t}{\tau}}\frac{\sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij}^{(t-1)})^2\rho(r_{ij}^2)^{(t-1)}}{\sum_{i=1}^{C}\left[\sum_{j=1}^{N}w_{ij}^{(t-1)}u_{ij}^{(t-1)}\right]^2} \tag{12}$$

其中 $\eta=8000$ 为初始值 $\tau=10$ 为时间常数，t 为迭代的次数。对于具体的某一次迭代，$\alpha(t)$ 实际上是个常数。

3 竞争学习在 TSK 模糊模型中的应用

本文中的模糊模型采用文[5]中的 TSK 模糊模型。该模型包含如下 IF-THEN 规则

$$R^i:\text{if } x_1 \text{ is } A_1^i(\vec{\theta}_1^i) \text{ and } x_2 \text{ is } A_2^i(\vec{\theta}_2^i),\cdots,x_n \text{ is } A_n^i(\vec{\theta}_n^i)$$
$$\text{then } h^i=f_i(x_1,x_2,\cdots,x_n;\vec{a}^i)=a_0^i+a_1^i x_1+\cdots+a_n^i x_n \tag{13}$$

其中 R^i 表示第 i 条规则，$A_j^i(\vec{\theta}_j^i)$ 是输入域上的模糊子集，若模糊模型有 C 条规则，则对于输入 $\vec{x}$，其估计输出为

$$y=\frac{\sum_{i=1}^{C}h^i u^i}{\sum_{i=1}^{C}u^i} \tag{14}$$

其中 u^i 是第 i 条规则的加权系数，可以通过上小节的竞争学习得到。定义 r_{ij} 为第 j 个真实输出值与第 i 条规则输出之间的误差[6]

$$r_{ij}=y_j-h^i=y_j-f_i(\vec{x}(j);\vec{a}) \tag{15}$$

其中 $i=1,2,\cdots,C,j=1,2,\cdots,N$。利用(3)式

$$\frac{\partial J}{\partial\vec{a}^i}=\sum_{j=1}^{N}(u_{ij})^2 w_{ij}\frac{\partial r_{ij}^2}{\partial\vec{a}^i}=0 \tag{16}$$

由(13)，(15)及(16)式可以得到

$$\sum_{j=1}^{N}(u_{ij})^2 w_{ij}\frac{\partial r_{ij}^2}{\partial\vec{a}^i}-\sum_{j=1}^{N}(u_{ij})^2 w_{ij}\frac{\partial r_{ij}^2}{\partial\vec{a}^i}f_i(\vec{x}_j;\vec{a}^i)=0 \tag{17}$$

而 $\frac{\partial r_{ij}}{\partial\vec{a}^i}=\vec{x}_j$。若定义 $X\in R^{N\times(n+1)}$，其中 $(1,\vec{x}_k)$ 是它的行向量，$Y\in R^{N\times 1}$ 是由 $y_k,k=1,\cdots,N$ 组成的列向量，$D_i\in R^{N\times N}$ 是对角阵，其第 k 个对角线上的元素为 $u_{ik}^2 w_{ik}$。则(17)式可以写成矩阵的形式

$$X^T D_i Y-(X^T D_i X)\vec{a}^i=0 \tag{18}$$

其中 $i=1,2,\cdots,C$，则多项式系数可以由下式求得

$$\vec{a}^i=[X^T D_i X]^{-1}X^T D_i Y \tag{19}$$

整个算法的流程可以归纳为如下几个步骤，$C(t)$ 是在第 t 次迭代时类的数目：

1. 设初始值 $t=0,w_{ij}=1,\forall i,j,C(0)=C_{\max}$，及迭代次数 $t_{\max}$。用 $C(0)$ 个类均匀覆盖输入空间，初始化 u_{ij}。

2. 用式(19)计算出多项式系数 $\vec{a}^i$，用式(15)计算 r_{ij}，再用式(8)，(6)，(7)计算 $\rho_i(\hat{r}_{ij}^2)$，$w_{ij}(\hat{r}_{ij}^2)$。用式(12)得到当前的 $\alpha(t)$ 值。

3. 用式(10)，(11)计算当前的 u_{ij}，计算每个分类的势 $N_i=\sum_{j=1}^{N}w_{ij}u_{ij}$，如果 $N_i<N_\epsilon$ 则删除该类，更新 $C(t+1)$。用式(9)更新 $\zeta(t+1)$。

4. 如果 $t>t_{\max}$ 则停止，否则跳到第二步去执行。其中 N_ϵ 的选择由下式得到

$$N_\epsilon=1+800\alpha(t)^2+0.725(\max(N_i)-\min(N_i))/C(t)+\min(N_i)/1.2 \tag{20}$$

4 仿真实例

用上述方法对带有噪声和孤立点的数据进行处理，模型 1 为

$$y=x^{2/3}\quad -2\leqslant x\leqslant 2 \tag{21}$$

数据点数为 300，噪声为正态分布 $N(0,0.04)$，孤立点为 20 个随机点，$t_{\max}=40,C_{\max}=40$。

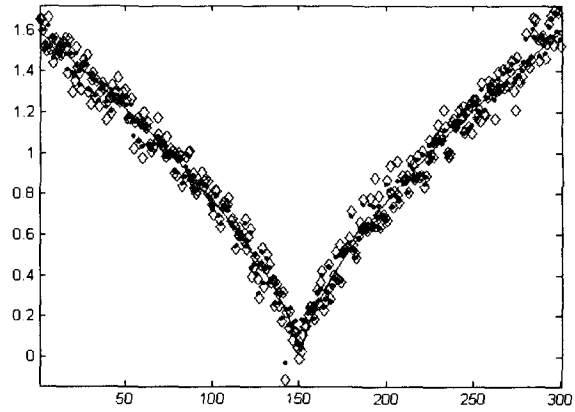

图 1　模型 1 的仿真结果

仿真结果如图 1 所示，其中黑实线为模型轨迹，方框点为带有噪声和孤立点的输入数据，黑点为处理后的轨迹。

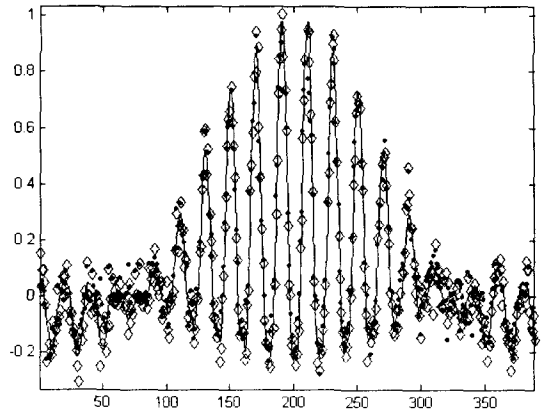

图 2　模型 2 的仿真结果

(下转第 212 页)

插队算法约好的平均路径长度;对于 CPU 的运行时间，基于禁忌表的定位算法每次所用时间的变化很小，而模拟退火算法和两阶段模拟退火算法，CPU 所用时间的变化却较大，特别是当采用自适应冷却进度表时，变化就更大。显然，在不降低解的质量的前提下，基于禁忌表的定位算法明显优于模拟退火算法和两阶段模拟退火算法，也优于嵌套插队算法。

3.2 用基于禁忌表的定位算法对 FL417 的求解结果

FL417(417 个城市)问题是一个已知最优解长度为 11861 的 TSP 问题，用最优解长度为 11861 标准来判断所得解的质量，更客观地说明问题。文[13]中，用嵌套划分算法(NP 算法)和其他启发式算法相结合(2-opt，3-opt)计算了 FL417 问题，结果如表 5 所示。文[8]中，用嵌套插队算法对 F1417 计算 10 次所得的结果如表 6 所示。用基于禁忌表的定位算法对 FL417 的求解结果如表 7 所示。

表 5 文[13]用嵌套划分算法和启发式算法结合(2opt，3opt)计算 F1417 的结果

方法	路径质量=(最优解-已获解)/ 最优解	CPU 时间(s)
2-opt，起始路径随机选择	14.6%	112
NP 算法结合 2-opt	14.0%	35
	4.9%	4199
3-opt，起始路径随机选择	0.78%	206298
NP 算法结合 3-opt	1.36%	25368
	0.51%	133129

表 6 文[8]用嵌套插队算法计算 10 次所得的结果

实例	平均		最好		最差	
FL417	路径质量	cpu 时间(s)	路径质量	cpu 时间(s)	路径质量	cpu 时间(s)
	0.31%	240	0.084%	238	0.548%	242

表 7 基于禁忌表的定位算法计算 10 次所得的结果

实例	平均		最好		最差	
FL417	路径质量	cpu 时间(s)	路径质量	cpu 时间(s)	路径质量	cpu 时间(s)
	0.312%	121.2	0.105%	112.3	0.539%	134.5

比较表 5、表 6、表 7 结果显示：基于禁忌表的定位算法所得解的质量明显好于文[13] 结果，也好于文[8]嵌套插队算法结果。说明基于禁忌表的定位算法是解决 TSP 问题的一种非常有效的算法。

结论 本文提出基于禁忌表的定位算法是快速、高效的近似算法，在合理的计算时间内求解较大规模 TSP 问题，实例验证在求解质量和求解速度方面表现较好。基于禁忌表的定位算法是一种很有竞争力的求解 TSP 问题的算法。

参 考 文 献

1 王凌. 智能优化算法及其应用[M]. 北京：清华大学出版社，2001

2 Glover F, Laguna M. Tabu Search. Boston: Kluwer Academic publishers, 1997

3 贺一，刘光远. 变异操作对禁忌搜索性能的影响研究[J]. 计算机科学，2002，29(5)

4 贺一，刘光远，邱玉辉. Tabu search 中集中性和多样性的自适应搜索策略. 计算机研究与发展[J]. 2004，41(1)：162～166

5 周培德. 货郎担问题的几何解法[J]. 软件学报，1995(6)

6 于志伟，陶波，王元美. 一种竞争算法及其在组合优化问题上的应用[J]. 软件学报，1998(10)

7 鄢烈祥. 用列队竞争法解旅行商问题[J]. 运筹与管理，1999(3)

8 翟东海，靳蕃. 用嵌套插队算法解决 TSP 问题[J]. 运筹与管理，2003，12(04)

9 高国华，沈林成，常文森. 求解 TSP 的空间锐化模拟退火算法[J]. 自动化学报，1999，25(3)

10 万颖瑜，周智，陈国良，顾钧. SizeScale：求解旅行商问题(TSP)的新算法. 计算机研究与发展[J]，2002，39(10)

11 邹鹏，周智，陈国良，顾钧. 求解 TSP 问题的多级归约算法[J]. 软件学报，2003，14(1)

12 James M, James C P. A fast mothod for generalized startingt temperature determjnation in homogeneous two-stage simulated annealing system [J]. Computer and Operation research, 1999, 26: 481～503

13 Shi Leyuan, et al. New parallel randomized algorithms for the traveing sales man problem [J]. Operation and research, 1999, 26: 371～394

(上接第 192 页)

模型 2 为

$$y=\frac{\sin(x_1)\sin(x_2)}{x_1x_2} \tag{22}$$

其中 $-5\leqslant x_1, x_2\leqslant 5$，数据点数为 400，噪声为正态分布 $N(0, 0.04)$，孤立点为 20 个高斯分布的随机点，$t_{max}=40$，$C_{max}=40$。仿真结果如图 2 所示，其中黑实线为模型轨迹，方框点为带有噪声和孤立点的输入数据，黑点为处理后的轨迹。

结论 本文采用 RCA 的模糊学习思想，对算法中遗忘函数进行了合理的修正，实现了算法的稳健收敛。由仿真结果可以看出，该算法在输入数据有噪声和孤立点的干扰下依然能自动地向理想的模型逼近。

参 考 文 献

1 Bezdek J C. Pattern Recognition With Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981

2 Frigui H, Krishnapuram R. A robust competitive clustering algorithmwith applocations in computer vision. IEEE Trans. Pattern Anal. Mach. Intell, 1999, 21: 450～465

3 Hampel F R, Ronchetti E M, Rousseeuw P J, Stahel W A. Robust Statistics: The Approach Based on Influence Functions. NewYork: John Wiley & Sons, 1986

4 Hawkins D M. Identification of Outliers. London, U. K.: Chapman & Hall, 1980

5 Sugeno M, Yasukawa T. A fuzzy-logic-based approach to qualitativemodeling. IEEE Trans. Fuzzy Syst, 1993, 1: 7～31

6 Chuang C-C, Su S-F, Chen S-S. Robust TSK Fuzzy Modeling for Function Approximation With Outliers. IEEE Trans. Fuzzy Syst, 2001, 9: 810～821