文章编号:1001 - 9081(2005)06 - 1357 - 03

基于关联函数的动态聚类算法及应用

朱群雄[1],宣达婧[1],顾祥柏[1,2]
(1. 北京化工大学 信息科学与技术学院,北京 100029; 2. 中国石化工程建设公司,北京 100101)
(zhuqx@mail.buct.edu.cn)

摘 要:根据时序立体数据的特点,提出了基于关联函数一致性矩阵的动态聚类算法。给出了适用于时序立体数据关联函数的改进标准关联函数计算公式,并将该算法应用于乙烯裂解炉报警系统,结合流程的时序立体数据,得到了裂解炉报警系统的动态聚类分类结果,并验证了提出算法的有效性。本文算法对于时序数据的聚类具有较强的鲁棒性。

关键词:动态聚类; 关联函数; 一致性矩阵; 乙烯裂解炉
中图分类号: TP311.131 **文献标识码**:A

Dynamic clustering algorithm based on dependent function and its application

ZHU Qun-xiong[1], XUAN Da-jing[1], GU Xiang-bai[1,2]
(1. *School of Information Science & Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*;
2. *Sinopec Engineering Inc, Beijing* 100101, *China*)

Abstract: A dynamic clustering algorithm was proposed based on consistent matrix of dependent function for time series multi-dimensional data. Further, improved standard dependent function calculation formula was proposed for time series cubic data. The provided algorithm was adopted to cluster ethylene cracking furnace alarm information by time series data. The effective of the proposal algorithm has been verified by result of above case study. Proposed dynamic clustering algorithm has strong robustness in clustering of time series multi-dimensional data.

Key words: dynamic clustering algorithm; dependent function; consistent matrix; ethylene cracking furnace

0 引言

在实际应用的问题中,随着时间的发展与数据的积累,可以产生大量的按时间顺序排列的平面数据表序列,这样按照时间顺序排放的平面数据序列形成了一个数据匣,称之为时序立体数据表,并记为:$\{x_{ij}(t_k)\}$。例如对于 m 个属性 $C_1, C_2, \cdots, C_m$,取定 n 个参数 $x_1, x_2, \cdots, x_n$,且按照时间顺序 $t_1, t_2, \cdots, t_N$ 获得原始数据 $\{x_{ij}(t_k)\}$ 就构成了一个时序立体数据表。

$$\begin{array}{c|cccccccc} & t_1 & & & & \cdots & t_N & & \\ & x_1 & x_2 & \cdots & x_n & \cdots & x_1 & x_2 & \cdots & x_n \\ C_1 & x_{11}(t_1) & x_{12}(t_1) & \cdots & x_{1n}(t_1) & \cdots & x_{11}(t_N) & x_{12}(t_N) & \cdots & x_{1n}(t_N) \\ C_2 & x_{21}(t_1) & x_{22}(t_1) & \cdots & x_{2n}(t_1) & \cdots & x_{21}(t_N) & x_{22}(t_N) & \cdots & x_{2n}(t_N) \\ \vdots & & & & & & & & & \vdots \\ C_m & x_{m1}(t_1) & x_{m2}(t_1) & \cdots & x_{mn}(t_1) & \cdots & x_{m1}(t_N) & x_{m2}(t_N) & \cdots & x_{mn}(t_N) \end{array} \quad (1)$$

对于时序数据可以将动态的综合评价问题转化为静态的综合评价问题,即对数据进行时间维上的综合,为了体现实时性,可以采用一个遗忘因子 λ 对近期的数据加权,即:

$$x_{ij}(\lambda) = \frac{1}{N}\sum_{k=1}^{N} x_{ij}(t_k)\exp(\lambda t_k),$$
$$i = 1,2,\cdots,m;\ j = 1,2,\cdots,n;\ k = 1,2,\cdots,N \quad (2)$$

当 λ 取较大值时,表明看重近期的数据;当 $\lambda = 0$ 时,表明各个时期的数据被看重的程度相同。一般可取 $\lambda = (2N)^{-1}$[1]。

上面的方法是先将立体数据 $\{x_{ij}(t_k)\}$ 转化为平面数据 $\{x_{ij}\}$,应用比较简单方便,但不是直接利用立体数据表 $\{x_{ij}(t_k)\}$ 对属性 $C_1, C_2, \cdots, C_m$ 在不同时刻 t_k 的运动状态所进行的综合评价,在实际应用中受到一定的限制。为此有必要研究适合时序立体数据的动态聚类算法。

1 时序数据的关联函数

对于时序数据,可以按照参数的方式来考虑建立关联函数[2]。

定义 1 设有 n 个参数,每个参数有 m 个指标,S 个采样样本,第 i 个参数的第 j 个指标在 t_k 时刻的观测值 $x_{ij}(t_k)$,则第 i 个参数的物元 R_i 表示为:

收稿日期:2004 - 12 - 09;**修订日期**:2005 - 03 - 14
基金项目:教育部科学技术研究重点项目(01024);中石化科学技术研究开发项目(E03007)
作者简介:朱群雄(1960 -),男,江苏无锡人,教授,博士,博士生导师,主要研究方向:过程系统工程、人工智能、数据挖掘、决策与控制等;宣达婧(1981 -),女,吉林榆树人,硕士研究生,主要研究方向:信息与决策; 顾祥柏(1966 -),男,江苏盐城人,博士研究生,高级工程师,主要研究方向:决策与控制、CIMS、数据挖掘、经济管理等.

$$R_i = \begin{pmatrix} v_i & C_1 & x_{i1}(t_k) \\ & C_2 & x_{i2}(t_k) \\ & \vdots & \vdots \\ & C_m & x_{im}(t_k) \end{pmatrix}, i = 1,2,\cdots,n \tag{3}$$

其中的 v_i 为第 i 个参数。

定义 2 设 i 参数 j 指标的关联函数为 $k_{ij}(x(t_k))$，则：

$$k_{ij}(x(t_k)) = \begin{cases} 0 & x \notin [x_{ij}(1),x_{ij}(4)]; \\ \dfrac{x_{ij}(t_k) - x_{ij}(1)}{x_{ij}(2) - x_{ij}(1)} & x \in [x_{ij}(1),x_{ij}(2)]; \\ 1 & x \in [x_{ij}(2),x_{ij}(3)]; \\ \dfrac{x_{ij}(4) - x_{ij}(t_k)}{x_{ij}(4) - x_{ij}(3)} & x \in [x_{ij}(3),x_{ij}(4)]; \end{cases}$$

$$i = 1,2,\cdots,n; j = 1,2,\cdots,m \tag{4}$$

称此关联函数为标准关联函数，称 $x_{ij}(1),x_{ij}(2),x_{ij}(3),x_{ij}(4)$ 为 $k_{ij}(x(t_k))$ 的节点。

定义 3 若标准关联函数 $k_{ij}(x(t_k))$ 的第一和第二个节点 $x_{ij}(1)$ 与 $x_{ij}(2)$ 为零，则：

$$k_{ij}(x(t_k)) = \begin{cases} 0 & x \notin [0,x_{ij}(4)]; \\ 1 & x \in [0,x_{ij}(3)]; \\ \dfrac{x_{ij}(4) - x_{ij}(t_k)}{x_{ij}(4) - x_{ij}(3)} & x \in [x_{ij}(3),x_{ij}(4)]; \end{cases}$$

$$i = 1,2,\cdots,n; j = 1,2,\cdots,m \tag{5}$$

称此关联函数 $k_{ij}(x(t_k))$ 为下侧关联函数。

定义 4 若标准关联函数 $k_{ij}(x(t_k))$ 的第三和第四个节点 $x_{ij}(3)$ 与 $x_{ij}(4)$ 重合，则：

$$k_{ij}(x(t_k)) = \begin{cases} 0 & x \notin [x_{ij}(1),x_{ij}(4)]; \\ \dfrac{x_{ij}(t_k) - x_{ij}(1)}{x_{ij}(2) - x_{ij}(1)} & x \in [x_{ij}(1),x_{ij}(2)]; \\ 1 & x \geqslant x_{ij}(2); \end{cases}$$

$$i = 1,2,\cdots,n; j = 1,2,\cdots,m \tag{6}$$

称此关联函数 $k_{ij}(x(t_k))$ 为上侧关联函数。

2 基于关联函数的动态聚类算法

建立关联函数信息矩阵 $K(t_k) = (k_{ij}(x(t_k)))_{n\times m}$：

$$K(t_k) = \begin{array}{c} \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{array} \begin{array}{c} \begin{array}{ccccc} C_1 & & C_2 & \cdots & C_m \end{array} \\ \begin{bmatrix} k_{11}(x(t_k)) & k_{12}(x(t_k)) & \cdots & k_{1m}(x(t_k)) \\ k_{21}(x(t_k)) & k_{22}(x(t_k)) & \cdots & k_{2m}(x(t_k)) \\ \vdots & & & \vdots \\ k_{n1}(x(t_k)) & k_{n2}(x(t_k)) & \cdots & k_{nm}(x(t_k)) \end{bmatrix} \end{array} \tag{7}$$

对于式(7)中的关联函数集 $\{k_{ij}(x(t_k))\}$ $i = 1,2,\cdots,n; j = 1,2,\cdots,m; k = 1,2,\cdots,N$，在进行标准化处理[3]以后，对参数 $k_{li}(x(t_k))$ 与 $k_{lj}(x(t_k))$，$i,j = 1,2,\cdots,n$; $l = 1,2,\cdots,n$ 作重要性比较[4]，以 $f^l_{ij}(t_k)$ 表示重要性标度，其取值原则为：

$$f^l_{ij}(t_k) = \begin{cases} 0 & k_{lj}(x(t_k)) \text{ 比 } k_{li}(x(t_k)) \text{ 重要} \\ 0.5 & k_{li}(x(t_k)) \text{ 与 } k_{lj}(x(t_k)) \text{ 同样重要} \\ 1 & k_{li}(x(t_k)) \text{ 比 } k_{lj}(x(t_k)) \text{ 重要} \end{cases} \tag{8}$$

令 $E^l_i(t_k) = \sum_{j=1}^{n} f^l_{ij}(t_k)$，则对应于各属性的模糊一致性矩阵 $A^l_i(t_k)$ 中的元素 $a^l_{ij}(t_k)$ 为：

$$a^l_{ij}(t_k) = \frac{E^l_i(t_k) - E^l_j(t_k)}{2^\alpha n} + 1$$

$i,j = 1,2,\cdots,n; k = 1,2,\cdots,N$;

其中 α 为敏感因子，可以取为 0,1,2,3。则：

$$A^l(t_k) = \begin{bmatrix} a^l_{11}(t_k) & a^l_{12}(t_k) & \cdots & a^l_{1n}(t_k) \\ a^l_{21}(t_k) & a^l_{22}(t_k) & \cdots & a^l_{2n}(t_k) \\ \vdots & & & \vdots \\ a^l_{n1}(t_k) & a^l_{n2}(t_k) & \cdots & a^l_{nn}(t_k) \end{bmatrix} \tag{9}$$

根据一致性判断矩阵的可综合性的特点，可以得到时序立体信息数据的综合一致性矩阵为：

$$A^l = \sum_{k=1}^{N} A^l(t_k) =$$

$$\begin{bmatrix} \sum_{k=1}^{N} a^l_{11}(t_k) & \sum_{k=1}^{N} a^l_{12}(t_k) & \cdots & \sum_{k=1}^{N} a^l_{1n}(t_k) \\ \sum_{k=1}^{N} a^l_{21}(t_k) & \sum_{k=1}^{N} a^l_{22}(t_k) & \cdots & \sum_{k=1}^{N} a^l_{2n}(t_k) \\ \vdots & & & \vdots \\ \sum_{k=1}^{N} a^l_{n1}(t_k) & \sum_{k=1}^{N} a^l_{n2}(t_k) & \cdots & \sum_{k=1}^{N} a^l_{nn}(t_k) \end{bmatrix} =$$

$$\begin{bmatrix} A^l_{11} & A^l_{12} & \cdots & A^l_{1n} \\ A^l_{21} & A^l_{22} & \cdots & A^l_{2n} \\ \vdots & & & \vdots \\ A^l_{n1} & A^l_{n2} & \cdots & A^l_{nn} \end{bmatrix}, k = 1,2,\cdots,N \tag{10}$$

根据一致性矩阵的可综合性，可以得到综合一致化矩阵：

$$A' = \sum_{l=1}^{m} w_l A^l \tag{11}$$

将(11)式中各列归一化，并将其中的零采用 $+\varepsilon$(ε 为一任意小的正数)代替得到综合信息矩阵为：

$$A'' = \begin{bmatrix} \dfrac{A'_{11}}{\sum_{i=1}^{n} A'_{i1}} & \dfrac{A'_{12}}{\sum_{i=1}^{n} A'_{i2}} & \cdots & \dfrac{A'_{1n}}{\sum_{i=1}^{n} A'_{in}} \\ \dfrac{A'_{21}}{\sum_{i=1}^{n} A'_{i1}} & \dfrac{A'_{22}}{\sum_{i=1}^{n} A'_{i2}} & \cdots & \dfrac{A'_{2n}}{\sum_{i=1}^{n} A'_{in}} \\ \vdots & & & \vdots \\ \dfrac{A'_{n1}}{\sum_{i=1}^{n} A'_{i1}} & \dfrac{A'_{n2}}{\sum_{i=1}^{n} A'_{i2}} & \cdots & \dfrac{A'_{nn}}{\sum_{i=1}^{n} A'_{in}} \end{bmatrix} \tag{12}$$

令 $A'' = A' \times (A')^{\mathrm{T}}$，则得到可用于聚类的相似矩阵。

通过上述标定，得到的相似矩阵反映了样本间关于指标的相似关系，但它只具有自反性和对称性，不具有传递性，因此采用平方法得到 A'' 的传递闭包 $t(A'')$：

$$\begin{aligned} &A'' \circ A'' = A''^2 \\ &A''^2 \circ A''^2 = A''^4 \\ &\vdots \\ &A''^{2k} \circ A''^{2k} = A''^{2k+1} \end{aligned} \tag{13}$$

其中 $k \leqslant \lg n$。在 A'' 是相似矩阵的前提下，一定存在一 A''，使得 $A''^{2k} = A''^{2k+1}$，则 $t(A'') = A''^{2k+1}$ 就是论域上的一个等价矩阵，对于等价矩阵中的值选择不同的 λ 值，得到不同的水平截集，得到聚类结果。

3 关联函数动态聚类算法的应用

下面以某乙烯厂的裂解炉报警系统为例来说明关联函数

动态聚类算法在优化操作中的应用。

按照乙烯裂解炉的工艺流程和操作规程,可以得到裂解炉系统报警相关的参数集如下:

参数报警位号集 = {FC10632(原料油总流量),FC1061(A 组进料流量),FC1062(B 组进料流量),FC1063(C 组进料流量),FC1064(D 组进料流量),FC1065(A 组蒸汽流量),FC1066(B 组蒸汽流量),FC1067(C 组蒸汽流量),FC1068(D 组蒸汽流量),TC10629A(急冷气出口温度),TC10629B(急冷气出口温度),TI10624A(A 组炉管 COT),TI10625A(B 组炉管 COT),TI10626A(C 组炉管 COT),TI10627A(D 组炉管 COT),TC10632(4 组炉管平均 COT),TI10628A(TLE 出口温度),TI10628B(TLE 出口温度),TI10628C(TLE 出口温度),TI10628D(TLE 出口温度),PC1061(底部燃料压力),PC10635(侧壁燃料压力),FC10634(锅炉给水总流量),LC1063(液位),TI10671(超高压蒸汽温度),TC1065(超高压蒸汽温度),FI10635(超高压蒸汽流量)}。

对上述主要参数的实际生产数据采样,按照裂解炉的操作周期,假设 5 天更新一次报警的分类,即样本更新周期 N = 5 天,选择随机采样数据的长度为 1 440,按照文献[5] 根据 Chernoff bounds 提出的定理 1,取 $\delta = 0.01$,可以确定上述采样数据集可以代表报警动态实际状态。在此基础上,考虑到流程中各参数对于流程报警系统的影响程度不同,因此对于高高报与低低报属性的权重各取为 0.3,对于高报与低报属性的权重各取为 0.2。

为了应用动态聚类算法,有必要建立各流程参数对应报警系统的特征。流程的报警系统等级一般可以分为高高报(AHH)、高报(AH)、低报(AL)、低低报(ALL)共四个状态。对于高高报,其关联函数的四个节点可以按照如下方法选取:假设流程参数的最大值为 $a_4 = a_3$、高位联锁(IH) = a_2、高高报(AHH) = a_1;对于高报,其关联函数的四个节点可以按照如下方法选取:最大值为 $a'_4 = a'_3$、高高报(AHH) = a'_2、高报(AH) = a'_1,按照式(6) 的上侧关联函数的修正公式计算出各参数每个采样相应高报类的关联函数。对于低低报,其关联函数的四个节点可以按照如下方法选取:假设流程参数的最低值为 $b_1 = b_2$,则低低报(ALL) = b_4、低位联锁(IL) = b_3;对于低报,其关联函数的四个节点可以按照如下方法选取:最低值为 $b_1 = b_2$,则低报(AL) = b_4、低低报(ALL) = b_3,按照式(5) 的下侧关联函数的修正公式计算出各参数每个采样相应的低报类的关联函数。

按照以上算法,取 $\varepsilon = 0.0001$,对得到的参数传递闭包,选取不同的 λ 值进行分类,得到如图 1(a) 的参数聚类谱系图。水平截集 λ 越大,报警的分类越少,对于低级别的报警分辨率越低,会有效管理重要报警,但可能忽视低级别报警中的重要报警,水平截集 λ 越小,报警的分类越多,对于低级别的报警分辨率越高,会降低重要报警管理的有效性。

为了验证基于关联函数动态聚类算法对于异常数据的鲁棒性,以参数 FC1066 为例,在采样的数据中人为加入异常数据,为了方便验证与程序实现,选用对本属于正常非报警但靠近报警点的数据,将靠近高报一定范围的数据提高 15%,而将靠近低报一定范围的数据降低 15%,由于在原采样数据中符合这样条件的数据并不集中,所以可以方便地人为加入异常数据。对 FC1066 加入异常数据举例如表 1。

对进行异常化处理后的采样数据,采用第 2 节的算法,取 $\varepsilon = 0.0001$,得到的传递闭包与原采样数据的传递闭包的最大偏差为 0.93%,因此在选用与未加入异常数据同样的 1 值时,得到与图 1(a) 相同的分类谱系图。说明本文提出的基于关联函数的动态聚类算法对异常数据具有较好的稳定性。

表 1　FC1066 异常数据举例

时间点	原数据	异常数据
586	3903	4488
658	3963	4557
785	3966	4560
863	3949	4541
890	2596	2210
893	2611	2220
923	2628	2235
961	2613	2225
1024	2501	2130
1141	2415	2140
1246	3952	4540

为了验证算法的动态跟踪性,选取报警状态动态变化了的第二组 1 440 个采样数据进行聚类,对得到的传递闭包采用与图 1(a) 相同的 λ 值进行分类,得到如图 1(b) 的聚类结果,通过对流程的分析,判断出算法的结果能够动态一致地跟踪流程实际的报警状态。

提出的方法可以在线跟踪乙烯裂解炉系统的运行状况。从不同数据段的聚类谱系图 1(a) 和(b) 可以看出,本算法对实时数据有很好的动态跟踪效果,即具有很好的一致跟踪性。对实时数据中含有异常点(如表 1) 的情况,关联函数变化不明显,说明本算法具有很好的鲁棒性。

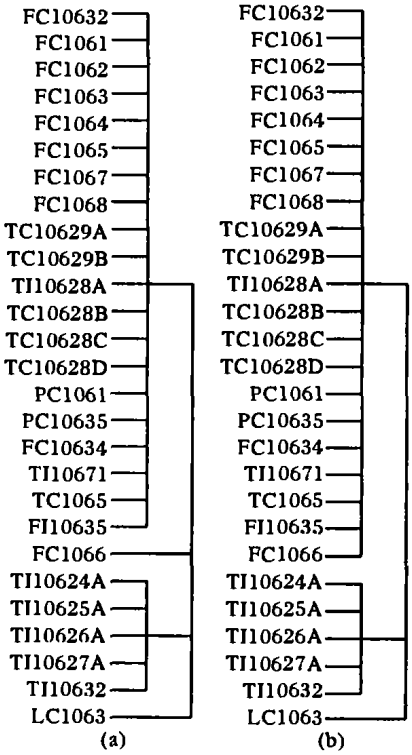


图 1　报警参数聚类谱系图

4　结语

时序数据随着时间会激增,对其进行聚类处理需要具有动态传递性和可综合性的算法,从而实现信息的一致性,并且聚类算法的计算量不能随着数据的增加而急剧增加,同时应对异常数据具有较好的鲁棒性。本文以文献[2]提出的关联函数为基础,提出了适合动态时序数据的关联函数计算公式,采用文献[3]的一致性信息矩阵建立的方法,提出了关联函数动态聚类算法。提出的算法具有较好的一致性与动态跟踪性,其计算量不随数据的增加而大幅度增加,可以广泛用于时序数据的聚类分类问题,同时,算法对实时数据中出现异常数据,具有很强的鲁棒性。

参考文献:

[1] 郭亚军. 综合评价理论与方法[M]. 北京:科学出版社,2002.

[2] 党耀国, 梁保松, 叶耀军. 聚类分析中关联函数的构造[J]. 广大工业大学学报,1998, 15(2):121 - 124.

[3] 李荣钧. 模糊多准则决策理论与应用[M]. 北京:科学出版社,2002.

[4] 孙才志,林学钰. 基于层次分析的模糊一致性判断矩阵及其应用[J]. 模糊系统与数学,2002,16(3):59 - 63.

[5] GUHA S, RASTOGI R, SHIM K. CURE: An efficient clustering algorithm for large database[J]. Information System, 2001, 26(1): 35 - 58.