基于模糊聚类分析的控制图变点估计研究

沈维蕾　赵　韩　周　蓉
合肥工业大学，合肥，230009

摘要：针对传统控制图不能解决生产过程发生变化的确切时间点即变点估计的问题，基于模糊聚类和统计学理论知识提出了变点估计的混合模糊统计聚类方法。在模拟仿真试验下，验证了该方法无论在固定抽样策略还是在可变抽样策略下，与传统的修哈特控制图法和 FCM 方法相比，对于控制图的变点估计都有着良好的效果。最后将该方法运用于发动机缸体制造的过程控制之中，检验了其有效性和可行性。

关键词：统计过程控制；变点估计；模糊集理论；模糊聚类；可变抽样控制图
中图分类号：TP273　　DOI：10.3969/j.issn.1004－132X.2012.23.013

Research on Change－point Estimation of Control Charts Based on Fuzzy Clustering Approach

Shen Weilei　Zhao Han　Zhou Rong
Hefei University of technology，Hefei，230009

Abstract：Traditional control charts did not indicate the real time of the process changes which were change－point. A fuzzy－statistical clustering approach was given to estimate the time of changes based on the fuzzy cluster and statistical theory. The proposed approach has a good result of change －point compared with traditional Shewhart control chart method and FCM method to estimate the change－point of different control charting methods with or without variable sampling strategies in a good deal of simulation. Finally，the method was used in the engine cylinder manufacturing process. The effectiveness and feasibility of this study were tested.

Key words：statistical process control；change－point estimation；fuzzy set theory；fuzzy clustering；variable sampling control chart

0　引言

统计过程控制（statistical process control，SPC）是一种通过监控工序过程来保证工序能够在充分发挥其过程能力的基础上，制造出合格产品的方法。SPC 的主要工具是控制图[1]。当控制图发出一个失控状态的信号时，表示生产过程发生了变化[2]。但是在确定过程变化的实际时间点时，控制图并不是很有效的工具。为了解决这个问题，有些学者提出了变点模型，但是并不能很好地处理实际应用中很普遍的可变抽样策略[3]。因此，本文基于模糊聚类理论以及统计方法，提出一种新的模糊统计聚类方法来处理实际应用中的变点问题。

收稿日期：2012—06—20

1　模糊聚类与变点统计分析

1.1　聚类分析

聚类就是按照一定的要求和规律对事物进行区分和分类的过程[4]。聚类分析则是指用数学的方法研究和处理给定对象的分类，它是多元统计分析方法中的一种。传统的聚类分析是一种硬划分，它把每个待辨识的对象严格划分到某类中，具有非此即彼的性质，因此这种类别划分的界限是分明的。而实际上大多数对象并没有严格的属性，它们在性态和类属方面存在着中介性，具有亦此亦彼的性质，因此适合进行软划分。模糊集理论的提出为这种软划分提供了有力的分析工具，并称之为模糊聚类分析。

[12]　顾临怡，胡志刚，刘莹冰. 用 SIMULINK 实现脉宽/脉频调制中的占空比控制[J]. 液压与气动，2003(9)：1-3.

（编辑　何成根）

作者简介：刘　忠，男，1968 年生。常熟理工学院机械工程学院教授、博士，湖南师范大学机电技术装备研究所教授。主要研究方向为现代工程装备机电液系统集成与节能控制。出版专著 3 部，发表论文 50 篇。刘　灯，男，1988 年生。湖南师范大学机电技术装备研究所硕士研究生。金　耀，男，1972 年生。湖南师范大学机电技术装备研究所副教授、博士。朱　浩，男，1972 年生。湖南大学机械与运载工程学院副教授、博士。

1.2 变点统计分析

1.2.1 变点的定义

变点指的是某一时刻，在此前后的观测值或数据遵循两个不同的分布模型[5]。对于均值控制图来说，当工序过程中的均值存在一个突变时，变点模型可按下式来构造：

$$X_i \sim \begin{cases} N(\mu_0,\sigma_0^2) & i=1,2,\cdots,t-1 \\ N(\mu_1,\sigma_0^2) & i=t,t+1,t+2,\cdots,n \end{cases} \tag{1}$$

其中，X_i 为该过程的第 i 个输出值。开始时，X_i 服从于正态分布 $N(\mu_0,\sigma_0^2)$，当该过程到达 t 点时，X_i 却服从于另一个不同均值的正态分布 $N(\mu_1,\sigma_0^2)$，其中 $\mu_0 \neq \mu_1$。我们把这种在过程中引起分布变化的点 t 称为变点。

1.2.2 变点估计中的聚类方法

将聚类方法应用于变点估计之中，其本质就是将控制图中的观测值进行分类，以找出变点。只有将所有独立而又连续的观测值全部分为在控状态聚类和失控状态聚类这两种类别，才能准确地找出变点。这种分类在研究中又称为模式分类[6]。

使用聚类方法来估计变点，其实质也是聚类方法应用于模式识别的一个方面。在以前的研究中，经常采用变点模型来进行变点的估计，而本文采用聚类分析来找出变点，这两种方法有很多相似点[7]。

2 混合模糊统计聚类方法

2.1 混合模糊统计聚类方法概述

当控制图接受一个失控的信号时(即观测值超出控制界限)，可以得出不同的聚类所确定的目标函数[7]。在基于目标函数的基础上，可以确定出最佳的两种聚类，即在控状态类和失控状态类。

假设控制图在某 t 时刻发出了一个失控信号，现研究的主要目的是找出均值从 μ_0 变到 $\mu_1(\mu_0 \neq \mu_1)$ 的这一点即变点 t。考虑到控制图对于聚类的约束条件，将变点 t 之前的所有观测值都归为在控状态类，而将变点 t 之后的所有观测值都归为失控状态类。所有点中使得目标函数最小的那一点，就认为它是变点 t。

2.2 聚类的隶属函数

隶属函数是用于表征模糊集合的数学工具，它是用来刻画处于中介过渡事物对差异双方所具有的倾向性。本文讨论的问题主要是估计正态分布下连续生产过程中的均值变点问题，因此一个样本对于其所属聚类的隶属概率等于该样本所服从分布的概率。因此我们假设真正的变点发生在 t 时刻，于是每个样本的隶属度可按下式计算：

$$\left.\begin{aligned} P(Y_i \in IFC) &= P(Y_i \in f(Y;\theta_0)) \\ P(Y_i \in OFC) &= P(Y_i \in f(Y;\theta_1)) \end{aligned}\right\} \tag{2}$$

式中，IFC 为在控状态类；OFC 为失控状态类；Y_i 为第 i 个样本相适应的随机变量；$f(Y;\theta_0)$ 为在控状态参数为 θ_0 的过程分布函数；$f(Y;\theta_1)$ 为失控状态参数为 θ_1 的过程分布函数。

按照式(2)，可以推导出正态分布下的单因素计数值控制图的具体隶属函数。先假设生产过程总体处于在控状态下，服从正态分布 $N(\mu_0,\sigma_0^2)$，即 $X \sim N(\mu_0,\sigma_0^2)$。该分布指的是生产过程中零件的个体分布，由于控制图上所研究的点为某一抽样样本的平均值 $\bar{x}_i$(第 i 个样本的平均值)，因此可得

$$P(\bar{x}_1 \in IFC) = P(\bar{x}_i \in N(\mu_0,\frac{{\sigma_0}^2}{n_i})) =$$

$$\begin{cases} 2P(Y \leqslant \bar{x}_i \mid Y \sim N(\mu_0,\dfrac{\sigma_0^2}{n_i})) & \bar{x}_i \ll \mu_0 \\ 2P(Y \leqslant -\bar{x}_i \mid Y \sim (\mu_0,\dfrac{\sigma_0^2}{n_i})) & \bar{x}_i > \mu_0 \end{cases} \tag{3}$$

其中，$\bar{x}_i$ 为控制图上的第 i 个点，也就是从工序过程中抽取的第 i 个样本中的 n_i 个数据的平均值。

同样可以推导出失控状态下的隶属概率，它与样本点服从于正态分布 $N(\mu_1,\dfrac{\sigma_0^2}{n_i})$(其中 $\mu_1 \neq \mu_0$) 的概率相等。因此可得

$$P(\bar{x}_1 \in OFC) = P(\bar{x}_i \in N(\mu_1,\frac{\sigma_0^2}{n_i})) =$$

$$\begin{cases} 2P(Y \leqslant \bar{x}_i \mid Y \sim N(\mu_1,\dfrac{\sigma_0^2}{n_i})) & \bar{x}_i \ll \mu_1 \\ 2P(Y \leqslant -\bar{x}_i \mid Y \sim (\mu_1,\dfrac{\sigma_0^2}{n_i})) & \bar{x}_i > \mu_1 \end{cases} \tag{4}$$

这里的 μ_1 为该过程处于失控状态下的均值。

2.3 聚类目标函数

为了在众多可能的分类中找出合理的分类结果，要确立合理的聚类准则，即聚类的目标函数。聚类的目标函数很多，硬聚类常使用的目标函数有最小平方误差和。而模糊聚类常用的目标函数众多，如熵函数、似然函数、似然对数函数等[8]。本文采用的目标函数为

$$F = -\sum_{i=1}^{t-1}\ln(P(\bar{x}_i \in IFC)) - \sum_{i=t}^{n}\ln(P(\bar{x}_i \in OFC)) \tag{5}$$

当 t 使得该函数取得最小值时，则 t 最有可能为变点。该函数主要评估在控状态聚类和失控状态聚类不同组合之间的效果。

3 模拟对比试验

3.1 模糊聚类方法在固定抽样控制图下的试验

首先生成服从 $\mu=100$、$\sigma=5$(本节模拟运算中

变量全用量纲一单位）的正态分布的随机数列，在$t=100$时，将生成服从$\mu=100+\delta$（其中$\delta=1,1.5,2,3$）、$\sigma=5$的正态分布的随机数列。然后分别采用三种方法（模糊统计聚类法、休哈特控制图法、FCM）[9]来找出该随机数列的变点，每一种方法循环1000次并求出其平均值和标准差。抽样方法采用的固定抽样样本(n_1,n_2)分别为(2,2)、(4,4)、(8,8)。表1给出了该组试验的结果数据。

表1　三种方法在随机模拟1000次之后的数据对比

估计方法	样本大小	估计量	变点值			
			1	1.5	2	3
模糊统计聚类法	(2,2)	μ	103.233	101.212	99.702	99.421
		σ	23.241	7.922	5.123	4.981
	(4,4)	μ	98.268	101.664	101.462	99.668
		σ	11.345	8.882	6.136	3.646
	(8,8)	μ	99.831	99.761	99.662	99.992
		σ	12.371	3.432	1.411	0.546
休哈特控制图法	(2,2)	μ	265.231	142.343	115.456	109.334
		σ	155.211	48.981	35.234	13.889
	(4,4)	μ	136.231	116.221	109.234	100.551
		σ	89.541	10.421	2.172	0.042
	(8,8)	μ	111.451	100.422	100.089	100.011
		σ	20.121	1.092	0.221	0.011
改进过的FCM	(2,2)	μ	118.679	86.672	79.342	74.609
		σ	52.871	25.543	24.012	26.112
	(4,4)	μ	79.891	75.676	74.112	72.198
		σ	25.991	22.675	28.743	29.987
	(8,8)	μ	75.341	73.654	71.654	64.653
		σ	27.791	27.848	28.682	29.693

通过表1可以看出，本文方法显著地提高了控制图中的变点估计的精度与稳定性，说明本文方法在固定抽样样本的变点检测上显然要好于其他两种方法。

3.2　模糊聚类方法在可变抽样控制图中的试验

首先生成服从$\mu=100$、$\sigma=5$的正态分布的随机数列，在$t=100$时，将自动生成$\mu=100+\delta$（其中$\delta=1,1.5,2,3$）、$\sigma=5$的正态分布的随机数列。试验中将上警告线设置为UCL，下警告线设置为LCL，抽样策略采用$(n_1,n_2)=(2,8)$。即在前一样本点处于上警告线和下警告线之间时，取小样本$n_1=2$进行平均运算求得样本点；如果前一样本点在上警告线和下警告线之外但又未超出上下控制线时，取大样本$n_2=8$进行平均运算取得样本点。为了减小随机误差对试验的影响，将试验随机模拟1000次，求出最后的平均值和标准差。

由表2可以看出，本文方法对于可变抽样控制图的变点估计同样有着很高的精确性。特别是在均值变点呈小幅度变化时，这种方法的优越性更加凸显。但是另一方面，在突变幅度变得很大时，无论是本文方法还是其他方法，其应用效果都趋于一致。这主要是由休哈特控制图的性质和它们在大幅度突变时都有着良好表现的相似性而决定的[10]。

表2　三种方法在随机模拟1000次以后的变点均值数据对比

抽样策略	试验方法	变点均值			
		1	1.5	2	3
$(n_1,n_2)=(2,8)$	模糊统计聚类方法	100.975	99.664	99.602	99.618
	改进过的FCM	136.749	88.973	78.963	74.622
	可变抽样控制图方法	163.654	104.637	101.623	100.121

4　实例分析

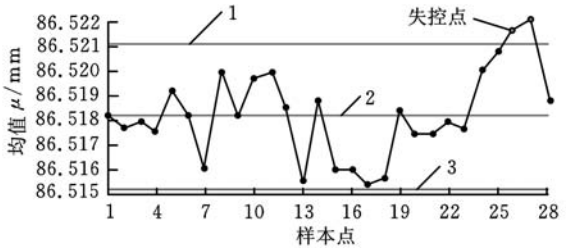
某发动机缸体的制造技术标准为$\phi86.5_0^{+0.03}$mm。通过对生产线上的发动机缸体直径进行持续记录，并研究这些记录数据的波动来判断该生产线是否处于正常的生产状态。如果该工序过程发生异常，则通过模糊统计聚类方法来找出工序过程发生异常的变点。通过与均值控制图和比较来评价模糊统计聚类方法。

下面给出该公司2011-05-05至2011-06-01发动机缸体生产线上的全部采样数据（每天依次采集4个数据，单位为mm）：

86.519,86.518,86.517,86.519,86.518,86.517,86.517,
86.519,86.519,86.518,86.518,86.517,86.518,86.518,
86.516,86.518,86.518,86.520,86.519,86.520,86.518,
86.517,86.518,86.520,86.514,86.519,86.512,86.520,
86.521,86.520,86.519,86.520,86.520,86.517,86.518,
86.518,86.520,86.519,86.519,86.518,86.520,86.516,
86.517,86.519,86.518,86.517,86.518,86.520,86.517,
86.517,86.517,86.518,86.519,86.520,86.522,86.518,
86.519,86.521,86.521,86.520,86.519,86.522,86.519,
86.526,86.520,86.521,86.519,86.525,86.523,86.519,
86.518,86.520,86.519,86.520,86.521,86.521,86.519,
86.519,86.517,86.519,86.520,86.518,86.512,86.515,
86.516,86.520,86.521,86.520,86.516,86.518,86.520,
86.513,86.512,86.520,86.514,86.519,86.519,86.513,
86.514,86.517,86.517,86.514,86.516,86.516,86.517,
86.515,86.519,86.518,86.518,86.519,86.517,86.515

根据以上数据绘制出均值控制图，如图1所示。

由图1似乎可以得出以下结论：5月30日（即样本点为26）以前生产过程处于正常状态，5月30日控制图发出失控信号，因此生产工序过程可能在5月30日或5月31日发生了变化。但是这个结论显然是不正确的，因为控制图发出失控



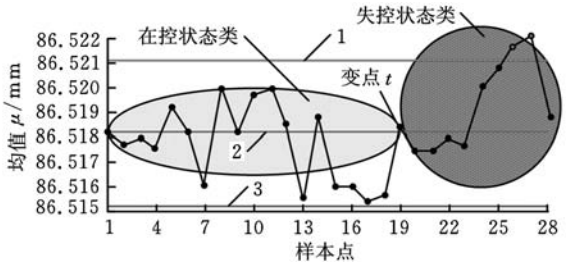
1. UCL=86.521 201mm　2. $\bar{x}$=86.518 268mm

3. LCL=86.515 370mm

图 1　发动机缸体生产线的均值控制图

信号并不能代表工序过程在此时发生了异常，控制图本身并不能给出变点的确切时间。因此，下面将使用模糊统计聚类方法来对这些数据进行再一次分析，并估计其变点即工序过程发生变化的确切时间点。

将数据作为一个数列，以控制图中显示的失控点(即样本点为 26)作为遍历算法的起点。通过计算每个样本点的隶属函数，可以求出所有样本点的目标函数[11]。通过对目标函数值进行比较判断，得出的结论是，当样本点为 19 时，即在 5 月 23 日工序过程发生了突变。这也就说明生产线发生异常变化的实际时间是 5 月 23 日，与控制图中的失控点 5 月 30 日相差很大。采集数据并进行分析的目的就是改善生产线，使之处于正常状态下，因此找出生产过程发生变化的实际点是非常重要的，图 2 是通过仿真后得出的聚类分析图。



1. UCL=86.521 201mm　2. $\bar{x}$=86.518 268mm

3. LCL=86.515 370mm

图 2　控制图上样本点的聚类分析

从以上的对比中可以发现，模糊统计聚类方法在实际应用中具有很大的意义。通过模糊统计聚类方法可以精确地找出均值或者标准差发生突变的确切时间，在仅有发生质量问题的趋势和隐患时，及时查找原因并进行整改，使生产过程回到正常水平。

5　结论

(1)将模糊聚类分析应用到控制图中，可以准确地知道均值或者标准差发生突变的确切时间，在仅有发生质量问题的趋势和隐患时，及时查找原因并进行整改，使生产过程回到正常水平。

(2)模糊聚类分析方法无论在固定抽样策略还是可变抽样策略下，对于控制图的变点估计都有着良好的效果。

(3)将模糊聚类分析方法运用于发动机缸体制造的过程控制之中，得到了满意的效果，证明该方法是有效性的、可行性的。

参考文献：

[1] 伍爱. 质量管理学[M]. 3 版. 广州：暨南大学出版社，2010.

[2] 袁太平. 基于成组技术的项目过程质量控制研究[D]. 天津：河北工业大学，2006.

[3] 张公绪，孙静. 新编质量管理学[M]. 2 版. 北京：高等教育出版社，2003.

[4] 高新波. 模糊聚类分析及应用[M]. 西安：西安电子科技大学出版社，2004.

[5] 陈希儒. 变点统计分析简介[J]. 数理统计与管理，1991(2)：53-54.

[6] 郑立伟. 基于成组技术的质量控制方法与工具研究[D]. 天津：天津大学，2004.

[7] Zarandi M H F, Alaeddini A. A General Fuzzy-statistical Clustering Approach for Estimating the Time of Change in Variable Sampling Control Charts[J]. Information Sciences, 2010, 180(16): 3033-3014.

[8] Alaeddini A, Ghazanfari M, Nayeri M A. A Hybrid Fuzzy-statistical Clustering Approach for Estimating the Time of Changes in Fixed and Variable Sampling Control Charts[J]. Information Sciences, 2009, 179(11): 1769-1784.

[9] 何清. 模糊聚类分析理论与应用研究进展[J]. 模糊系统与数学，1998，12(2)：89-90.

[10] 袁芳，田铮，苏晓丽，等. 独立序列均值与方差变点的累积和估计及应用[J]. 控制理论及应用，2010，27(3)：396-398.

[11] 张建明. 质量管理中的模糊聚类分析方法[J]. 科学管理，2001(2)：8-9.

(编辑　郭　伟)

作者简介：沈维蕾，女，1969 年生。合肥工业大学机械与汽车工程学院副教授。主要研究方向为质量管理与可靠性。发表论文 20 余篇。赵　韩，男，1957 年生。合肥工业大学机械与汽车工程学院教授、博士研究生导师。周　蓉，女，1977 年生。合肥工业大学机械与汽车工程学院讲师。