文章编号:1001-2486(2011)03-0164-05

LRE试车数据挖掘中基于最大散度差的模糊聚类分析方法*

王 珉,胡茑庆,秦国军
(国防科技大学 机电工程与自动化学院,湖南 长沙 410073)

摘 要:在对液体火箭发动机试车数据进行聚类分析时,为解决故障数据样本与正常样本类间差异不大的问题,引入最大散度差准则,提出基于最大散度差的聚类算法MSD-CA。该算法以散度度量样本间的相似性,使样本的类内散度最小化和类间散度最大化同时进行。在此基础上,应用模糊理论对最大散度差准则进行模糊化,提出基于最大散度差的模糊聚类算法MSD-FCA,用于对试车样本进行"软划分",以提高聚类的正确性。实验结果证明了MSD-FCA的有效性。

关键词:液体火箭发动机;试车数据;数据挖掘;最大散度差准则;软划分;模糊聚类
中图分类号:TP391;V434.21 **文献标识码**:A

Fuzzy Cluster Analysis Based on Maximum Scatter Difference in LRE Test Data Mining

WANG Min, *HU Niao-qing*, *Qin Guo-jun*
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: In the clustering analysis of test data of liquid rocket engine, in order to solve the problem that there is insignificant difference between fault sample and normal sample, the maximum scatter difference criterion was introduced and the maximum scatter difference based clustering algorithm (MSD-CA) was presented. In MSD-CA, the similarity of samples was measured by divergence, and the minimizing of the within-class divergence and maximizing of the between-class divergence were processed together. After that, fuzzy theory was introduced to maximum scatter difference criterion, and the maximum scatter difference based fuzzy clustering algorithm (MSD-FCA) was presented and used to do "soft partition" for test data to improve the precision of cluster. The method is verified with experimental results.

Key words: liquid rocket engine; test data; data mining; maximum scatter difference criterion; soft partition; fuzzy clustering

液体火箭发动机是运载器和航天器中应用最广泛的动力系统,恶劣的工作环境使其成为航天器系统故障的敏感多发部位,因此发展液体火箭发动机故障检测和诊断技术研究具有重要军事意义和经济效益,目前,制约该技术发展的主要"瓶颈"仍是诊断知识的获取[1-2]。数据挖掘技术为解决这一"瓶颈"问题提供了有效的途径,液体火箭发动机历次地面热试车过程中产生的数据蕴涵着巨大的技术财富,通过对这些数据的挖掘,获取其中隐藏的大量有价值的信息,可以为发动机的故障检测与诊断提供决策支持[3]。

聚类分析是数据挖掘的一种重要工具,它基于"物以类聚"的观点,用数据方法分析样本模式的分布形式,挖掘数据集的结构特征[4]。目前,C-均值聚类和模糊C-均值聚类是两种最基本、在试车数据挖掘中最常用的聚类算法[1]。这两种算法强调尽可能地使类内相似性最大,适合处理不同样本模式差异较大的数据集[5]。而在试车数据挖掘中,更常见的是样本模式差异不大的数据集,这主要是因为液体火箭发动机地面热试车代价高昂,安全性要求高,一旦检测到数据特征变化较大,立即启动了紧急关机措施,使得很难获取与正常样本差异较大的数据。如果能提出一种新的聚类算法,在使类内相似性最大的同时,考虑使类间的相似性最小,以在样本模式差异不大的情况下提高聚类分析的准确性,则无疑是很有实际意义的。

散布矩阵是描述样本模式距离的重要工具,文献[6]提出了FCS算法,基于散布矩阵进行聚类分析;文献[7]将模糊理论引入Fisher线性判决,

* 收稿日期:2010-09-12
基金项目:国家自然科学基金资助项目(50675219);湖南省杰出青年科学基金资助项目(08JJ1088)
作者简介:王珉(1980—),男,博士生。

提出了基于模糊 Fisher 准则的半模糊聚类算法 FFC-SFCA。上述算法通过参数调节平衡类内相似性和类间相似性两个目标,使对类内相似性最大化和类间相似性最小化同时进行,取得了比 FCM 等更好的效果,但在应用中存在散布矩阵的奇异性问题。

最大散度差准则是一种基于散布矩阵的线性判别准则[8-9],与 Fisher 准则一样都是寻找最优的鉴别矢量,使得各类之间尽可能地分开,该准则使用类间散度减去 η 倍类内散度作为判别标准,能在一定程度上克服 Fisher 准则类内散布矩阵奇异性问题[9]。本文首先提出基于最大散度差的聚类算法 MSD-CA(Maximum scatter difference based clustering algorithm),然后引入模糊理论对最大散度差准则进行模糊化,提出基于最大散度差的模糊聚类算法 MSD-FCA(Maximum scatter difference based fuzzy clustering algorithm),以散度度量样本模式间的相似性,通过最小化类内散度和最小化类间散度,解决试车数据集中样本模式差异小的问题,进一步提高试车数据聚类的准确性。

1　基本概念

1.1　散布矩阵

设数据集 X 包含 N 个 d 维样本,即 $X=\{x_1, x_2, \cdots x_N\}$,将其划分为 C 类,第 i 类的样本个数为 N_i,聚类中心(均值向量)$m_i = \frac{1}{N_i}\sum_{k=1}^{N_i} x_k$,总体样本均值 $m = \frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{N_i} x_k$,则类内散布矩阵 S_i,总体类内散布矩阵 S_W,总体类间散布矩阵 S_B 分别定义为[4]

$$S_i = \sum_{k=1}^{N_i}(x_k - m_i)(x_k - m_i)^{\mathrm{T}} \tag{1}$$

$$S_W = \sum_{i=1}^{C}\sum_{k=1}^{N_i}(x_k - m_i)(x_k - m_i)^{\mathrm{T}} \tag{2}$$

$$S_B = \sum_{i=1}^{C} N_i(m_i - m)(m_i - m)^{\mathrm{T}} \tag{3}$$

1.2　最大散度差

最大散度差(MSD)的基本目的是寻找一个最优投影方向,实现分类的类内散度最小、类间散度最大,属于有监督的分类算法[8-9]。

令 ω 为投影空间法向量,样本 x_k 的投影为 $y_k = \omega^{\mathrm{T}} x_k$,投影空间的聚类中心 $\tilde{m}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} y_k = \omega^{\mathrm{T}} m_i$,总体样本均值 $\tilde{m} = \frac{1}{N}\sum_{i=i}^{C}\sum_{k=1}^{N_i} y_k = \omega^{\mathrm{T}} m$,投影空间的类内散布矩阵 $\tilde{S}_i$、总体类内散布矩阵 $\tilde{S}_W$ 和总体类间散布矩阵 $\tilde{S}_B$ 分别定义如下:

$$\tilde{S}_i = \sum_{k=1}^{N_i}(y_k - \bar{m}_i)(y_k - \tilde{m}_i)^{\mathrm{T}} = \omega^{\mathrm{T}} S_i \omega \tag{4}$$

$$\tilde{S}_W = \sum_{i=1}^{C}\sum_{k=1}^{N_i}(y_k - \tilde{m}_i)(y_k - \tilde{m}_i)^{\mathrm{T}} = \omega^{\mathrm{T}} S_W \omega \tag{5}$$

$$\tilde{S}_B = \sum_{i=1}^{C} N_i(\tilde{m}_i - \tilde{m})(\tilde{m}_i - \tilde{m})^{\mathrm{T}} = \omega^{\mathrm{T}} S_B \omega \tag{6}$$

MSD 以投影后数据样本的类间散度与类内散度的广义差为目标函数,如式(7)所示,在式(8)的约束下,求使目标函数最大化的向量 ω^*,即最优投影方向。

$$J(\omega) = \tilde{S}_B - \eta\tilde{S}_W = \omega^{\mathrm{T}} S_B \omega - \eta\omega^{\mathrm{T}} S_W \omega \tag{7}$$

$$\| \omega \| = 1 \tag{8}$$

参数 η 为一正实数,用来平衡最大化类间散度和最小化类内散度两个不同的目标。选定 η 后,在矢量 ω^* 上的投影可以保证达到类内散度最小,类间散度最大。

2　基于 MSD 的聚类分析

与 Fisher 准则一样,MSD 也是一种在有监督的情况下,进行投影方向上投影点类内和类间散布矩阵优化运算的分类方法[8]。基于 Fisher 准则,文献[10]提出了 KIF(K-means Iterative Fisher)方法,巧妙地将 Fisher 准则应用于无监督聚类。本文借鉴文献[10]的思想,将 MSD 应用于无监督的聚类分析,提出最大散度差聚类算法(Maximum Scatter Difference based Clustering Algorithm, MSD-CA)。

定义 Lagrange 函数:

$$L = \omega^{\mathrm{T}} S_B \omega - \eta\omega^{\mathrm{T}} S_W \omega - \lambda(\omega^{\mathrm{T}}\omega - 1) \tag{9}$$

式中,λ 为 Lagrange 乘子。

将 L 分别对 ω 和 m_i 求偏导数,并令偏导数为零,即

$$\frac{\partial L}{\partial \omega} = 0 \tag{10}$$

$$\frac{\partial L}{\partial m_i} = 0 \tag{11}$$

对于式(10)可以求得

$$(S_B - \eta S_W)\cdot\omega = \lambda\cdot\omega \tag{12}$$

解式(12)为求矩阵 $S_B - \eta S_W$ 的本征值问题,λ 为

该矩阵的特征值,ω 为对应的特征向量。

对于式(11)可以求得

$$m_i = \frac{N_i \cdot m - \eta \sum_{k=1}^{N_i} x_k}{N_i(1-\eta)} \tag{13}$$

基于 MSD 的聚类算法 MSD-CA:

步骤 1 输入数据集 X,指定参数 η、阈值 ε 和迭代次数 M,任选初始聚类中心 $\{m_i(j)\}(i=1,2,\cdots,C)$,迭代序号 $j=0$;

步骤2 计算所有样本与聚类中心的距离 $\| x_k - \{m_i(j)\} \|(i=1,2,\cdots,C)$,并按最小距离原则进行分类,计算 S_B 和 S_W,计算矩阵 $S_B - \eta S_W$ 的最大特征值 λ,并取 ω 为矩阵 $S_B - \eta S_W$ 属于 λ 的模为 1 的特征向量,使用式(9)计算 $J_j(\omega)$;

步骤 3 使用式(13)生成新的聚类中心 $\{m_i(j+1)\}$,按最小距离原则重新划分数据集 X;

步骤4 重复步骤 2,计算 $J_{j+1}(\omega)$,若 $|J_{j+1}(\omega) - J_j(\omega)| < \varepsilon$,算法结束;

步骤5 若 $j=M$,算法结束,否则 $j=j+1$,返回步骤 3。

3 基于 MSD 的模糊聚类分析

3.1 散布矩阵模糊化

MSD-CA 和 C – 均值聚类、KIF 一样,属于对数据对象的"硬划分",它把每个待聚类的对象严格地划分到某个类中,体现了非此即彼的性质,因此这种聚类的类别界限是分明的,然而事物之间的界限往往不是那么分明,客观世界中存在着大量模糊划分的现象,模糊理论为这种"软划分"提供了有力的数学工具[11]。

模糊类内散布矩阵 S_{fW} 和模糊类间散布矩阵 S_{fB} 定义如下:

$$S_{fW} = \sum_{i=1}^{C} \sum_{k=1}^{N} \mu_{ik}^{\beta} (x_k - m_i)(x_k - m_i)^{\mathrm{T}} \tag{14}$$

$$S_{fB} = \sum_{i=1}^{C} \sum_{k=1}^{N} \mu_{ik}^{\beta} (m_i - m)(m_i - m)^{\mathrm{T}} \tag{15}$$

式中,μ_{ik} 为样本 x_k 属于 i 类的隶属度,定义隶属度矩阵 $U = \{\mu_{ik}\}_{C \times N}$,$\beta \in [1, \infty]$ 为权指数,又称为平滑因子,控制样本在模糊类之间的分享程度。

定义模糊 MSD 目标函数:

$$J^F(\omega) = \omega^{\mathrm{T}} S_{fB} \omega - \eta \omega^{\mathrm{T}} S_{fW} \omega \tag{16}$$

问题转化为最大化目标函数 $J^F(\omega)$,求最优投影方向 ω^*,约束条件为

$$\begin{aligned} &\| \omega \| = 1 \\ &\sum_{i=1}^{C} \mu_{ik} = 1, k = 1,2,\cdots,N \end{aligned} \tag{17}$$

3.2 基于最大散度差的模糊聚类算法 MSD-FCA

定义 Lagrange 函数:

$$L = \omega^{\mathrm{T}} S_{fB} \omega - \eta \omega^{\mathrm{T}} S_{fW} \omega - \lambda(\omega^{\mathrm{T}} \omega - 1) - \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{C} \mu_{ik} - 1 \right) \tag{18}$$

式中,λ 和 $\lambda_k (k=1,2,\cdots,N)$ 为 Lagrange 乘子。

将 L 分别对 ω、m_i 和 μ_{ik} 求偏导数,并令偏导数为零,即式(10)、(11)和(19)。

$$\frac{\partial L}{\partial \mu_{ik}} = 0 \tag{19}$$

对于式(10)可以求得

$$(S_{fB} - \eta S_{fW}) \cdot \omega = \lambda \cdot \omega \tag{20}$$

即求矩阵 $S_{fB} - \eta S_{fW}$ 的特征值 λ 及对应的特征向量 ω。

对于式(11)可以求得

$$m_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{\beta} \cdot (\eta x_k - m)}{\sum_{k=1}^{N} \mu_{ik}^{\beta} \cdot (\eta - 1)} \tag{21}$$

对于式(19)可以求得

$$\mu_{ik} = \left[\frac{\lambda_k}{\beta \omega^{\mathrm{T}} [(m_i - m)(m_i - m)^{\mathrm{T}} - \eta (x_k - m_i)(x_k - m_i)^{\mathrm{T}} \omega]} \right]^{\frac{1}{\beta - 1}} \tag{22}$$

由 $\sum_{i=1}^{C} \mu_{ik} = 1, k = 1,2,\cdots,N$,可得

$$\begin{aligned} &\sum_{i=1}^{C} \mu_{ik} \\ &= \sum_{i=1}^{C} \left[\frac{\lambda_k}{\beta \omega^{\mathrm{T}} [(m_i - m)(m_i - m)^{\mathrm{T}} - \eta (x_k - m_i)(x_k - m_i)^{\mathrm{T}} \omega]} \right]^{\frac{1}{\beta - 1}} \\ &= 1 \end{aligned} \tag{23}$$

式(22)和(23)相除,得

$$\mu_{ik} = \frac{\left[\frac{1}{\beta \omega^{\mathrm{T}} [(m_i - m)(m_i - m)^{\mathrm{T}} - \eta (x_k - m_i)(x_k - m_i)^{\mathrm{T}} \omega]} \right]^{\frac{1}{\beta - 1}}}{\sum_{q=1}^{C} \left[\frac{1}{\beta \omega^{\mathrm{T}} [(m_q - m)(m_q - m)^{\mathrm{T}} - \eta (x_k - m_q)(x_k - m_q)^{\mathrm{T}} \omega]} \right]^{\frac{1}{\beta - 1}}} \tag{24}$$

根据模糊隶属度的要求,$\mu_{ik} \in [0,1]$,可对式(24)做出如下限定[7],若

$$\omega^{\mathrm{T}} (x_k - m_i)(x_k - m_i)^{\mathrm{T}} \omega \leqslant \frac{1}{\eta} \omega^{\mathrm{T}} (m_i - m)(m_i - m)^{\mathrm{T}} \omega \tag{25}$$

则 $\mu_{ik} = 1$,且 $\forall\, q \neq i$,$\mu_{qk} = 0$,此时对样本 x_k 采用"硬划分"。式(25)表示,当某一样本 x_k、第 i 类聚类中心 m_i 和样本总体均值 m 分别沿鉴别矢量 ω

方向投影后,如果样本投影点到聚类中心投影点的距离小于或等于聚类中心投影点到样本总体均值投影点距离的 $1/\eta$ 倍,则样本 x_k 严格隶属于第 i 类。

可用图 1 对式(25)进行说明,图中“☆”和“×”表示两类的聚类中心,硬划分区内的点满足式(25),将这些点按图中“分界线”方向向最优投影面投影,根据图示,硬划分区域为垂直最优投影面且对称分布于聚类中心的两带状区域。该区域的大小与参数 η 密切相关,因此在聚类过程中选择合适的参数 η 非常重要。

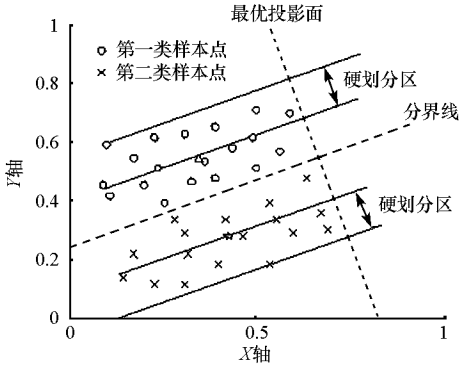

图 1　硬划分区示意图
Fig.1　Sketch map of the crisp section

3.3　MSD-FCA 的聚类步骤

由上述分析总结 MSD-FCA 的聚类步骤如下:

步骤 1　输入数据集 X,指定参数 η、阈值 ε 和迭代次数 M,随机产生隶属度矩阵 $U(j)$,迭代序号 $j=0$;

步骤 2　根据式(21)计算初始聚类中心 $\{m_i(j)\}$;

步骤 3　计算 S_{fB} 和 S_{fW},计算矩阵 $S_{fB}-\eta S_{fW}$ 的最大特征值 λ,并取 ω 为矩阵 $S_{fB}-\eta S_{fW}$ 属于 λ 的模为 1 的特征向量,使用式(16)计算 $J_j(\omega)$;

步骤 4　根据式(24)生成新的隶属度矩阵 $U(j+1)$;

步骤 5　重复步骤 2、3,计算 $J_{j+1}(\omega)$,若 $|J_{j+1}(\omega)-J_j(\omega)|<\varepsilon$,算法结束;

步骤 6　若 $j=M$,算法结束,否则 $j=j+1$,返回步骤 4。

4　算法应用研究及结果分析

4.1　数据样本描述

液体火箭发动机地面试验又称为试车,通常分为 120s 的短试车和 500s 的长试车,在试车过程中,发动机涡轮泵要经过 3s 的启动阶段和 117s/497s 的稳态运行阶段[12]。本文针对稳态运行阶段,提取特征空间的数据样本进行聚类分析。TF619 次热试车过程中,燃料涡轮泵的涡轮叶片分别在 120.8s 和 127s 附近发生了两次脱落,相应地,涡轮泵振动信号在 120.8s 和 127s 附近出现了两次冲击。但是试车并没有终止,所幸没有造成严重事故。使用两个时域特征(均方根值 x_{RMS} 和裕度因子 x_{CF})描述 TF619 次试车样本轴向振动信号的统计特性,振动信号的采样频率 $f_s=50\mathrm{kHz}$,设置特征的计算步长 $M=5000$ 点,相邻特征之间的时间间隔为 $\Delta t=M/f_s=0.1\mathrm{s}$,如图 2 所示。

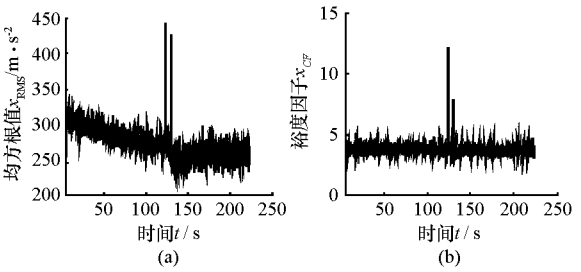

图 2　TF619 次试车轴向振动时域特征
Fig.2　Time-domain features of axial vibration during test TF619

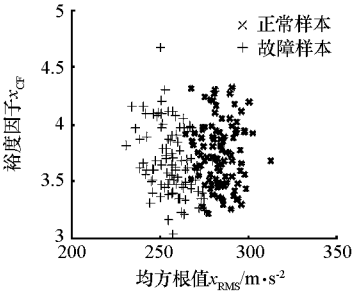

图 3　特征空间状态描述
Fig.3　State description in feature space

叶片脱落后,涡轮泵处于故障工作状态,由于试车没有终止,因此 TF619 记录了非常宝贵的故障数据。图 2 中,冲击特征前的记录为正常状态下的振动特征,冲击特征后的记录为故障状态下的振动特征,由图 2 可知,两类特征的差异不大,这也正是紧急关机系统出现漏警的原因。从两类特征样本中随机抽取 100 个记录,用于描述试车中涡轮泵轴向振动信号在特征空间的分布,如图 3 所示。从图 3 中同样可知,两类特征样本的分布十分紧凑,样本模式差异小,使用 FCM、MSD-CA、MSD-FCA 三种方法对该数据样本进行聚类分析,其中 $C=2$。

4.2　聚类结果分析

常用欧式距离描述样本的相似性和算法迭代误差,但是,特征空间中不同特征轴的值量级不同会导致欧氏距离产生错误的结论,因此,在聚类分析前,使用式(26)对样本进行归一化。

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{26}$$

式中，x_i 为特征轴上第 i 个特征值，$x_{\max}$ 和 $x_{\min}$ 分别为该特征轴上的最大值和最小值。算法参数设置为：$\beta = 2, \eta = 2, \varepsilon = 0.05$。

聚类结果如图 4 所示，图中“◇”和“☆”分别表示两类的聚类中心，标“○”的点表示分类错误的样本。将聚类结果与原数据集样本标记进行对比，统计各算法错分样本数，进而计算其准确率，结果如表 1 所示。由表 1 可知，本文采用的基于最大散度差的聚类算法，同时考虑了类内散度最小化和类间散度最大化，其聚类效果优于基于误差平方和的 FCM 方法。在聚类中，引入模糊理论对样本进行“软划分”，使得 MSD-FCA 的错分率要小于“硬划分”MSD-CA 的错分率。对图 3 所示特征样本，MSD-FCA 取得了比较好的聚类效果，较好地将差异小的故障样本和正常样本进行了分类。

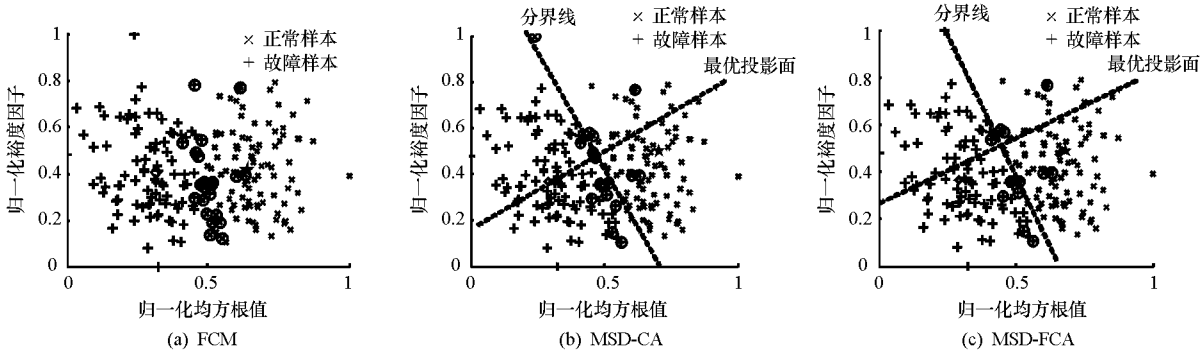

图 4 三种聚类算法的聚类结果

Fig.4 Cluster result of three cluster algorithm

表 1 三种算法的聚类准确率比较

Fig.1 Comparison of the accurate rate of three cluster algorithm

算法	样本数	错分样本数	分类错误率
FCM	200	23	11.5%
MSD-CA	200	15	7.5%
MSD-FCA	200	12	6%

5 结束语

提出了基于最大散度差准则的聚类分析算法 MSD-CA，以散度度量液体火箭发动机试车数据集中各样本模式的相似性，在聚类分析时，使类内散度最小化和类内散度最大化同时进行。引入模糊理论对最大散度差准则进行模糊化，提出了基于最大散度差的模糊聚类分析算法 MSD-FCA，以“软划分”的方式优化聚类效果，对实际试车数据的聚类分析结果表明，MSD-FCA 具有比常用聚类算法更好的效果，适于解决试车数据样本差异小的问题。由于 MSD-FCA 需要求解矩阵特征值和特征向量，且样本隶属度等需通过迭代公式进行计算，因而在高维大数据量情况下，需要进一步研究提高算法效率的途径。

参 考 文 献：

[1] William J M, Peter R, et al. Knowledge Mining Application in ISHM Test-bed[R]. IEEEAC 1296, Version 3, 2005.

[2] 胡小平，韩泉东，李京浩. 故障诊断中的数据挖掘[M]. 长沙：国防科技大学出版社，2009.

[3] 王珉，胡荳庆，杨思峰，等. 基于故障仿真的故障知识库应用研究[J]. 宇航学报，2010，31(4)：1253 – 1258.

[4] 温熙森. 模式识别与状态监控[M]. 北京：科学出版社，2007.

[5] Yu J, Li C X. Novel Cluster Validity Index for FCM Algorithm[J]. Journal of Computer Science and Technology, 2006, 21(1): 137 – 140.

[6] Wu K, Yu J, Yang M S. A Novel Fuzzy Clustering Algorithm Based on a Fuzzy Scatter Matrix with Optimality Tests[J]. Pattern Recognition Letters, 2005, 26(10): 639 – 652.

[7] 曹苏群，王士同，陈晓峰，等. 基于模糊 Fisher 准则的半模糊聚类算法[J]. 电子与信息学报，2008，30(9)：2162 – 2165.

[8] 宋枫溪，刘树海，杨静宇，等. 最大散度差分类器及其在文本分类中的应用[J]. 计算机工程，2005，31(5)：8 – 10.

[9] 宋枫溪，张大鹏，杨静宇，等. 基于最大散度差鉴别准则的自适应分类算法[J]. 自动化学报，2006，32(4)：541 – 549.

[10] Clausi D A. K-means Iterative Fisher (KIF) Unsupervised Clustering Algorithm Applied to Image Texture Segmentation[J]. Pattern Recognition, 2002, 35(9): 1959 – 1972.

[11] 杨小兵. 聚类分析中若干关键技术研究[D]. 杭州：浙江大学，2005.

[12] Hu L, Qin G J, Hu N Q. Online Support Vector Novelty Detection-algorithm for Turbopump of Liquid Rocket Engine Based on Vibration Signals[C]//Proceedings of ISMA2008 International Conference on Noise and Vibration Engineering. Leuven, 2008, 15 – 17: 3283 – 3292.