文章编号:1671-4598(2020)04-0222-05 DOI:10.16526/j.cnki.11—4762/tp.2020.04.045 中图分类号:E92 文献标识码:A

基于蒙特卡洛 k—means 聚类算法的舰船器材分类研究

吴雯雯，陈振林
(海军航空大学 岸防兵学院，山东 烟台 264001)

摘要：器材合理分类是建立预测模型的基础，某型舰船仪表器材数据较少、分类指标因素不足，使用传统方法易产生过拟合的问题；提出蒙特卡洛 K—means 算法，利用样本方差进行器材消耗聚类分析；该方法首先利用 MC 法计算初始聚类中心，参考 SBC 分类法制定聚类种类数 k，通过方差聚类建模来优化仪表器材的分类，最终得到仪表器材的聚类结果；实例计算表明，该方法能够有效改进 K—means 方法的分类结果，无需考虑其他备件指标因素影响，适用于数据量过小和存在白噪声的模型。

关键词：仪表器材；聚类分析；蒙特卡洛法；K—means

Research on Warship Spare Parts Cluster method Based on Monte Carlo K—means Cluster Algorithm

Wu Wenwen, Chen Zhenlin
(Naval Aviation University Coast Defence Academy, Yantai 264001, China)

Abstract: The reasonable classification of spare—parts is the basis of establishing the prediction model, the data of a certain warship spare—parts is less and the classification index factor is insufficient, so it is easy to use the traditional method to produce the problem of overfitting. A Monte Carlo K—means algorithm is proposed, and the sample variance is used for the spare—parts consumption volatility cluster analysis. Firstly, using Monte Carlo to calculate the initial clustering center, and refers to the SBC method to formulate the number of clustering categories k. The classification of instrument spare—parts is optimized by variance clustering modeling, and finally the cluster results of instrument spare—parts are obtained. The example shows that the method can effectively improve the classification results of K—means method without considering other index factors. It is suitable for the model with too small amount of data and white noise.

Keywords: instrumentation spare—parts; cluster algorithm; Monte—Carlo; K—means

0 引言

仪表器材是指用于检出、测量、观察、计算各种物理量、物质成分、物性参数等的器具。舰船仪表器材按照工作原理可以分为电磁式与机械式，按照测量类型可以分为力学、电磁、热工、化学、几何量、时频等六大类。遍布舰船各个工作部位，其主要作用是监测舰船运行状态，为舰船运行提供压力、电流、舵角、温度、风速、功率等信息。仪表器材的精确化保障对舰船运行至关重要。

舰船仪表种类繁多，数量庞大，消耗规律复杂，针对每一类器材进行分类预测并不现实，对仪表器材合理分类是提高效率的重要手段，是消耗预测的基础[1-4]。目前，针对器材的分类方法有定性方法与定量方法：定性方法有 ABC 分类法、VED 分类法等，这类方法操作简单，只需要考虑价值、关键性等一个或少数几个准则就能分类，但也存在过于粗放的问题；定量方法有基于器材消耗规律的 SBC 分类法、考虑多种分类因素的模糊综合评价法、层次分析法等，这些方法适用于样本容量大，影响因素复杂的情况。随着研究的不断深入，定性与定量相结合以及数据挖掘技术成为热点。

基于 VED 的 ABC 分类法将备件所属设备的重要程度等因素纳入了考虑范围[5-6]。文献［7-8］对备件品种的主要影响因素运用模糊综合评估方法进行综合评价，采用专家系统量化主要指标。基于 AHP 的 ABC 分类法，在两种方法结合的过程中，可以将定性因素和定量因素都转化成数值形式加以对比，在一定程度上能改进管理，但是备件关键性因素的确定受主观影响较大，不可避免地包含了主观性的不利影响[9-12]。

文献［13-14］采取基于属性的备件品种确定方法，将关键性、可更换性、消耗性、维修性等因素引入备件决策，利用粗糙集理论对备件属性进行因素选取，体现了定量与

收稿日期:2020-02-11； 修回日期:2020-03-06。

作者简介:吴雯雯(1988-)，女，山东烟台人，硕士，助理会计师，主要从事航空特种勤务技术方向的研究。

陈振林(1969-)，男，山东青岛人，硕士，教授，主要从事雷达应用方向的研究。

定性相结合的特点，有较好的工程实用性。

由英国学者 Syntetos 等人提出的 SBC 分类法[15]应用广泛。该方法基于器材消耗规律进行分类，通过两个截断值 ADI 和 CV^2（x）将需求分为 4 类。其中 ADI（average demand interval）是需求发生时间间隔的平均值，反应的是 0 需求量发生的频率，ADI 值越大，说明需求中 0 需求发生的越频繁，间断性越明显；CV（coefficient of variation）为需求量变异程度系数，反映非零序列偏离均值的严重度，值越大，序列越不稳定。图 1 中的 A、B、C、D 分别代表不稳定型消耗、块状型消耗、平稳型消耗、间断型消耗。

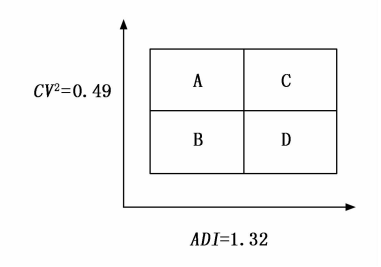

图 1　基于 SBC 分类法的备件消耗类型图

SBC 方法在处理大量数据时有着较为优越的解释效果，但在处理少量数据集的时候，往往容易产生较大的误差。虽然 SBC 方法对本文所研究的数据不太适用，但是它所包含的 4 种器材类型对有一定的通用指导意义。例如一部分价格昂贵、更换周期较长的仪表，就符合间断型消耗器材的特征，工作环境恶劣、大批量消耗的仪表其消耗特征也与平稳型消耗类型比较接近。

对于种类多、品种杂、消耗规律多样的器材，聚类分析作为一种定量方法，从数据分析角度，给出了更准确、细致的分类[16-17]。文献［18］运用主成分分析对分类准则进行降维，得到约简后准则再进行聚类处理。文献［19］从同一类器材中选择样本对网络进行训练，然后再用该网络对该类器材进行消耗预测，节省了训练时间。文献［20］基于器材消耗波动性进行聚类分析，采用层次划分聚类，使算法更稳定高效。

分析某型舰船仪表器材消耗数据，聚类分析方法适用性更好，主要有以下原因：1）某型舰船服役年限较短，数据量过少，器材属性、可靠性、影响因素等信息缺乏相关数据。如果采取 AHP、主成分分析法、灰色关联分析、支持向量机等方法，在数据量过少时，容易产生过拟合问题；2）仪表器材长期处于高温、高湿、高盐的工作环境，变化规律比较复杂，其损耗往往具有很大的偶然性，各种不同工况的影响或者操作的失误都有可能直接或间接地产生噪声影响。作为具有多量值特征的器材，采用聚类方法对其数学特征进行分析处理会更加准确、方便、科学。

1　蒙特卡洛 K－means 算法

聚类分析是一种重要的数据挖掘技术，是依据“物以类聚”的思想，对样本或者指标进行分类。其目的是把大量数据点的集合分成若干类自然分组，使得组内相似度最大化，组间相似度最小化，将目标集合分成由类似的个体组成的多个类的无监督分析过程，可有效地分析数据分布，广泛应用于模式识别、机器学习、航空航天等多个领域。聚类分析的分析思路为：在一批样本的多个观测指标中，找出一个统计量，该统计量可以度量样本间或者指标间的相似程度，构成一个对称的相似性矩阵，以此为基础，将各样本逐一归类。

k－means 聚类是最为常用的一种聚类方法，是基于原型的聚类。每一个簇都由某个中心点数据代表，这个中心点就是所谓的原型，该算法事先设置簇的个数，即 k 的值，k－means 聚类的目标是找出各簇的质心，然后与各质心相邻的数据点聚成各簇，以实现聚类。将所有点的均值作为簇的质心。k－means 聚类的优势在于对低维度数据聚类有着良好的解释效果，适用于数据的初步分析，是一种较为成熟的聚类方法。

k－means 聚类实现过程非常便捷，但它的一大弊端在于，该方法对初始聚类中心的选择十分敏感，不同的初始中心点会造成聚类结果的波动。随机初始化质心是该算法的基础，之后的工作都是围绕这一基础开展的，如果更换不同的初始化设置，那么就有可能得到更好的解。对于给定的数据，局部最优解往往不是全局最优解，因此，质心初始化对 k－means 聚类的结果有直接影响。为了有效地克服局部最优问题，可以采取多次初始化的方法。k－means 聚类在处理高维数据分类问题时，它更多表示为点的数据特性，而对多元线性的聚类列则存在缺陷，导致聚类中心散列，效果不佳。因此，本文通过引入 Monte－Carlo（MC）法对质心进行多次初始化，选出最好的那一次作为最终聚类中心。

MC 法亦被称作随机抽样技术，广泛应用于对物理过程或生化过程的模拟，也可以求解一些最优化问题。在利用计算机在统计抽样理论的基础上，通过有关随机变量的统计抽样检验或随机模拟，估计和描述函数的统计量、求解问题近似解的一种数值计算方法。MC 法不但能够解决随机性问题，也能解决确定性问题。其基本原理是：为解决某一实际问题，首先建立与所求解问题相应的一个随机模型，形成随机变量，使随机变量的某个数字特征（如期望值等）正好是问题的解；然后按照模型进行大量的随机实验，以获得随机变量的大量抽样值，用统计方法作出所求数字特征的估计值，就得到问题的解。MC 法计算程序简单，其收敛是统计意义上的收敛，收敛速度和问题维数无关。MC 法误差仅与方差和样本容量有关，而与样本中元素所在的空间无关[21]。MC 法具有程序结构简单、不受问题条件限制、模拟过程灵活、适于求解多维问题等优点，所以有着广泛的应用。

MC 法进行质心初始化的思想是：利用 MC 法计算机模

拟构建一个 n 维向量 $Y=(q_1,q_2,q_3,\cdots,q_n)$，与方差矩阵 $u_i=\frac{1}{|C_i|}\sum_{X\in C_i}X$（$u_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$，是簇 C_i 的均值向量）结合建立初始质心预测模型，利用 SSE 对模型参数进行动态调整，同时再与新采集的模型进行匹配，直到模型趋于平稳，此时确定初始质心。理想情况下，k－means 聚类初始质心应该落在每一类簇的中心附近，可以有效降低迭代次数并避免局部最优。

最终结果可表示为：

$$Z_t=SSE(\overrightarrow{Y}) \tag{1}$$

其中：SSE 代表方差函数，Z_t 表示在时间为 t 年下的模型所得误差，随后引入变量 $j\in(1,2)$，如果 $t+1$ 下的模型误差比 t 模型下的误差更小，则替代模型为：

$$f(\overrightarrow{Y_{t1}},\overrightarrow{Y_{t2}}) \tag{2}$$

2 算法流程

1）对数据进行特征选择。舰船器材具有品种繁多、影响因素多、波动性大的特点，器材的消耗因为影响因素的变动会存在一定程度的波动。舰船器材因其应用目的的特殊性，其影响因素复杂多变，使得波动性表现得更加明显，主要体现在消耗的规模波动和结构波动两方面。规模波动是指需求总量的波动，包含收缩和扩张两种情况；结构波动则比较复杂，主要体现在器材品种的不断改变。SBC 方法中的用到了两个波动性指标：需求发生间隔的平均值、需求量变异程度系数。但是分析本文数据可知，目前对该型舰船仪表器材的消耗数据是以年为单位进行统计，若采取 SBC 方法分类指标，数据过少，将会出现很大误差。因此，本文采用计算样本总体方差描述器材波动性。表达式为：

$$\sigma^2=\frac{\sum(X-\mu)^2}{N} \tag{3}$$

能够反应出曲线的变化规律和数据离散的分布特性，因此适用于 k－means 聚类。该方法的优点在于，解决了 k－means 处理多维数据噪声过大以及消耗器材数据时间轴数据过少无法采用合适模型的问题，同时为后期的模型更新做出了铺垫。

2）确定 k 值，即聚类种类。直观地看 k－means 就是把数据空间划分为 k 个区域或者划出 k 条边界，其中各区域以其原型为质心。通常情况下，增大 k 值就能减小 SSE，但这种方法容易出现过拟合，失去聚类分析的意义。k 值要事先指定，并且在很大程度上影响聚类结果。在先验知识不足的情况下，该参数的选取比较困难，需要进行多次试验才能找到最佳类别数。在实际应用中，往往需要与别的算法组合使用来确定合适的类别数，这些算法可能比 K－means 算法要复杂得多，抵消了 K－means 算法简便易行的优势。因此，本文根据对某型舰船机电仪表器材属性、工作原理及消耗情况的大致了解，参考 SBC 分类法的种类数，令聚类种类 $k=4$，使得聚类结果更加贴合器材管理实际。

3）利用 MC 法确定初始聚类中心，通过迭代，利用计算机快速运算，不断进行重复性操作，重复执行建立初始质心预测模型，在每次执行这组命令时，都从变量的原值推断出它的新值，直到各数据点不再变更自己所属的簇，或者这个变更不再显著，这样最后确定的质心就是数据内部各簇的代表或者原型。

4）选取 SSE 来作为误差检验指标。SSE 是拟合数据和原始数据对应点的误差的平方和，计算公式为：

$$SSE=\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2 \tag{4}$$

SSE 越接近于 0，则模型选择和拟合更好，数据预测也越成功。

该统计参数是预测数据和原始数据对应点误差的平方和的均值，计算公式为：

$$MSE=\frac{SSE}{n}=\frac{1}{n}\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2 \tag{5}$$

对于样本集 $D=\{x_i,x_i,\cdots,x_m\}$。K－means 聚类方法将聚类划分为 $C=\{C_1,C_2,\cdots,C_k\}$，最小平方误差为：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-u_i\|_2^2 \tag{6}$$

公式（6）刻画了簇内样本围绕簇均值向量的紧密程度，E 值越小，簇内样本的相似度越高。

3 仿真试验与比较

现以某型舰船 2015～2019 年 49 种仪表器材年消耗数据为例进行分析。使用 Matlab 软件进行仿真试验，通过 STDEVP 函数计算样本总体方差，得到结果如表 1 所示。

表 1 仪表器材年消耗数据

编号	2015	2016	2017	2018	2019	σ^2
A1	2	6	5	5	4	1.35
A2	1	4	6	4	7	2.05
A3	3	2	1	2	1	0.74
A10	11	8	6	7	9	1.72
…	…	…	…	…	…	…
A42	2	9	11	10	8	3.16
A43	4	1	6	5	7	2.05
A44	1	0	1	2	1	0.63
…	…	…	…	…	…	…
A48	8	7	6	8	9	1.01
A49	2	3	5	5	4	1.16

从表 1 可以看出，数据方差 σ^2 总体偏小，在一定范围内波动，对此建立需求方差变量 $T=[\sigma_1{}^2\sigma_2{}^2\cdots\sigma_{49}{}^2]$。对该变量分别进行经典 k－means 聚类及蒙特卡洛 k－means 聚类分析，对比两种方法的结果，验证本文方法的有效性和精确性。

3.1　经典 k－means 聚类分析

经典 k－means 聚类过程如下：

首先，初始化质心。随机初始化 k 个质心。

第二步，划分数据点，质心确定后，找出距离最近质心的数据点，形成簇，此处采取欧氏距离进行度量，有 n 个特征的数据点 $X(x_1,x_2,\cdots,x_n)$ 与点 $C(c_1,c_2,\cdots,c_n)$ 之间的欧式距离计算公式为：

$$d=\sqrt{(x_1-c_1)^2+(x_2-c_2)^2+\cdots+(x_n-c_n)^2} \quad (7)$$

各点找到相距最近的质心之后，就归属于该簇，数据空间就被划分成 k 个子区域。

第三步，找出该簇最有代表性的点，作为新的质心，即求解所有点到质心距离误差平方和最小化问题。

第四步，反复计算并更新质心。新的质心确定之后，更新各数据点至最近的质心，确定新簇并再一次更新质心。重复这个过程。直至各数据点所从属的簇不再变化或者变化不再显著，那么最后确定的质心就是各簇的代表，可以描述整个模型。

使用 Matlab 软件进行 k－means 聚类，所得结果见表 2，聚类图如图 2 所示。从表 2 中可以看出，k－means 方法针对每年数据都产生不同的聚类中心。从图 2 可以看出，k－means 聚类无法合理处理多维数据，聚类效果不明显。

表 2　最终聚类中心

聚类类型	1	2	3	4
2015	3	2	8	3
2016	15	2	8	3
2017	16	2	7	5
2018	14	3	7	6
2019	12	2	8	6

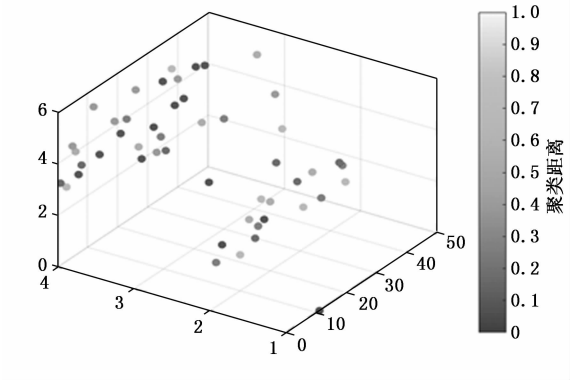

图 2　K－means 聚类图

3.2　蒙特卡洛 k－means 算法

对器材消耗进行蒙特卡洛 k－means 聚类分析，得到聚类结果见表 3，聚类图如图 3。从图 3 中可以看出，聚类效果显著，第 2、3 类消耗器材在总体样本中占比较高。

表 3　最终聚类中心

聚类类型	1	2	3	4
σ^2	2.56	0.85	1.73	4.69

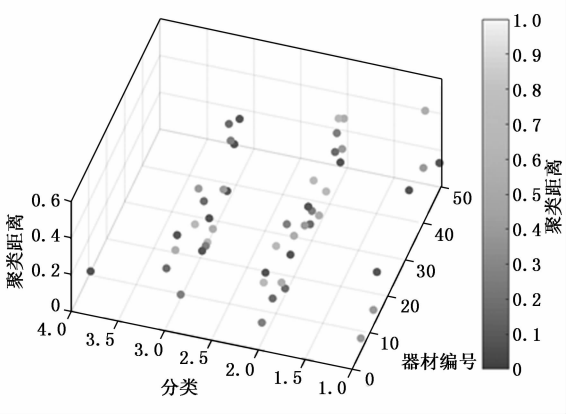

图 3　蒙特卡洛 K－means 聚类图

对比以上两种聚类结果及散点图可以看出，未进行蒙特卡洛 k－means 聚类结果散列，聚类图分类不明显，受时间序列的影响较大，不能够直观地分析出结果，而处理过后的数据聚类效果明显，该方法很好地将低数据量的消耗器材映射到了三维空间，同时解决了 k－means 算法无法处理高维数据的问题。

以（2.56，0.85，1.73，4.69）作为聚类中心得到聚类结果见表 4。

表 4　聚类成员

案例号	聚类	距离	案例号	聚类	距离
A1	3	0.32	A26	2	0.162
A2	2	0.109	A27	3	0.018
A3	2	0.308	A28	2	0.225
A4	2	0.342	A29	2	0.109
A5	1	0.069	A30	2	0.162
A6	3	0.32	A31	2	0.308
A7	3	0.382	A32	1	0.34
A8	2	0.122	A33	3	0.114
…	…	…	…	…	…
…	…	…	…	…	…
A15	2	0.037	A43	1	0.303
A16	2	0.368	A44	1	0.595
A17	3	0.242	A45	2	0.308
A18	2	0.162	A46	2	0.368
A19	3	0.397	A47	2	0.109
A20	2	0.225	A48	2	0.162
A21	3	0.018	A49	2	0.308
A22	3	0.269			
A23	3	0.005			
A24	3	0.005			
A25	3	0.116			

根据 4 种器材年消耗相对值，得出器材分类消耗折线图，如图 4 所示。从图中可以看出，2015～2019 年 4 类器材消耗均呈上升趋势，这与舰船遂行任务增多以及仪表到寿更换的客观事实是吻合的。从需求间隔和需求量上看，第 1 类与第 3 类器材波动性最强，第 4 类次之，第 2 类最为平稳。

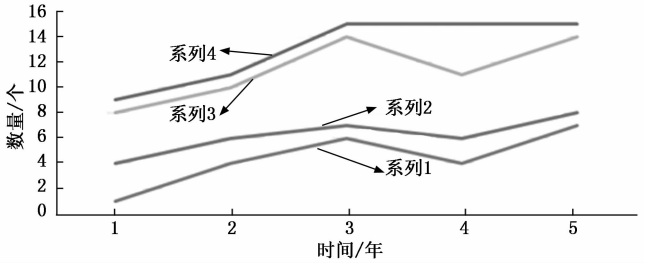

图 4 器材分类消耗折线图

4 结束语

准确的分类是消耗预测的基础，利用消耗波动性对器材进行分类符合实际工作需要，具有很强的借鉴意义。本文着力研究舰船仪表器材分类问题，针对某型舰船仪表器材数据量稀疏，采取需求量变异程度系数等其他波动性指标易造成过拟合的情况，考虑利用样本方差来体现器材消耗波动性，无需计算器材内在属性，不需要对数据进行时间序列 AR 建模，简化了仪表器材消耗分类模型，能够有效解决数据量过少时模型建立困难的问题，避免了复杂模型放大误差。本文基于蒙特卡洛法改进了初始聚类中心的选择，有效避免了传统算法随机选择初始聚类中心导致的结果不稳定性。与多尺度最小二乘 SVM 模型、AHP 理论相比，采用本文的方法，对数据不足的模型有着较好的适用性。后续研究将结合其他分类方法，对聚类结果进一步的量化分析。

参考文献：

[1] 侯甲凯. 航空公司航材周转件需求预测研究 [D]. 广汉：中国民用航空飞行学院，2015.

[2] 郭 峰. 强海滨. 航材统计预测与决策 [M]. 青岛：国防工业出版社，2017.

[3] Gu J，Zhang G，Li K W. Efficient aircraft spare parts inventory management under demand uncertainty [J]. Journal of Air Transport Management，2015，42：101－109.

[4] 王 芳. 基于不同航材分类的航材需求预测方法研究综述 [J]. 科技创新与应用，2015 (26)：58－59.

[5] 金夏芳，基于 VED 的备件 ABC 分类研究 [J]. 物流技术，2008，21 (12)：86－88.

[6] Kumar S，Chakravarty A. ABC － VED analysis of expendable medical stores at a tertiary care hospital [J]. Medical Journal Armed Forces India，2015，71 (1)：336－341.

[7] Chu C W，Liang G，Liao C. Controlling Inventory by Combining ABC Analysis and Fuzzy Classification [J]. Computers & Industrial Engineering，2008，55 (4)：841－851.

[8] Zhiwen Y，Hantao C，Jane Y，et al. Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2015，12 (4)：887－901.

[9] 徐向宇，李乃梁，王 晶，等. AHP－DEA 的备件 ABC 分类法 [J]. 机械设计与制造，2016 (8)：269－272.

[10] Cui N F，Luo X. ABC classfication based on AHP in servicing spare part [J]. Industrial Engineering & Management，2004，9 (6)：33－36.

[11] Nariswari N P A，et al. Testing an AHP model for aircraft spare parts [J]. Production Planning and Control，2019，30 (4)：329－344.

[12] Nariswari N P A，Bamford D R，DEHEB. Testing an AHP model for aircraft spare parts [J]. Production Planning and Control，2019，30 (4)：329－344.

[13] 张 帅，唐金国，俞金松，等. 基于属性的舰载机航材备件品种确定方法 [J]. 火力与指挥控制，2015.

[14] Braglia M，Grassi A，Montanari R. Multi－attribute classification method for spare parts inventory management [J]. Journal of Quality in Maintenance Engineering，2004，10 (1)：55－65.

[15] Syntetos A A，Keyes M，Babai M Z. Demand categorisation in a European spare parts logistics network [J]. International Journal of Operations &Production Management，2009，29 (3)：292－316.

[16] 王纵虎. 聚类分析优化关键技术研究 [D]. 西安：西安电子科技大学，2012.

[17] Nguyen T T，Krishnakumari P，Calvert S C，et al. Feature extraction and clustering analysis of highway congestion [J]. Transportation Research Part C：Emerging Technologies，2019，100：238－258.

[18] 张作刚，崔国伟，秦瑞清. 灰色聚类分析在航材分类中的运用 [J]. 四川兵工学报，2013，34 (9)：56－59.

[19] 虞文胜. 聚类分析在航材分类上的应用 [J]. 价值工程，2011，30 (30)：309－310.

[20] 薛永亮，陈振林. 基于消耗波动性聚类的航材分类研究 [J]. 系统工程与电子技术，2019，12：2802－2806.

[21] 付元钢. 气动热辐射的直接蒙特卡洛法模拟 [D]. 哈尔滨：哈尔滨工业大学，2007.