基于聚类分析的煮茧工艺参数

段春稳[1,3],任强胜[1,3],卜献鸿[1,2,3],李 帆[2,3],黎 钢[1,3],王建平[1,3]
(1.四川省丝绸科学研究院有限公司,四川 成都 610031;
2.四川省丝绸工程技术研究中心,四川 成都 610031;
3.现代茧丝绸制造技术资源四川省科技资源共享服务平台,四川 成都 610031)

摘 要:介绍一种 K-means 聚类数据分析方法,用于对制丝生产过程中形成的数据进行分析,为煮茧工艺设计提供指向性参考方案。

关键词:聚类分析;制丝数据;煮茧工艺
中图分类号:TS 142.2 **文献标志码**:B **文章编号**:1673-0356(2022)05-0036-04

制丝是将蚕茧加工成生丝的过程,包括混剥选茧、煮茧、缫丝、复摇整理、生丝检测等工序。煮茧是制丝生产的重要环节,煮茧质量直接影响到生丝质量、原料茧消耗、产量完成水平[1]。煮茧需要根据蚕茧原料特性,结合专业知识和煮茧经验来确定煮茧工艺,但由于蚕茧原料特性包含指标较多,专业技术人员水平及经验局限,煮茧工艺设计往往不能充分发挥蚕茧原料特性,并且在“试煮”过程中造成了大量原料浪费。

随着信息技术的发展,制丝行业生产过程中形成了大量繁杂的数据,包括茧质调查数据、煮茧生产数据、缫丝生产数据等,这些数据都是用于指导生产的重要技术指标,但由于数据的多样性、动态性、复杂性,以及目前制丝行业的技术局限,这些数据对煮茧工艺并未形成实质性的关联和指导。

介绍一种基于 K-means 聚类分析方法[2],对制丝行业形成的生产数据进行分析,为煮茧工艺参数设计提供指向性设计方案。

1 煮茧工艺

煮茧的目的是通过水、蒸汽等介质对蚕茧的作用,使干胶变成明胶[3-4],降低茧丝的胶着力,使茧丝能够依次离解,为缫丝创造条件。煮茧工艺的设计过程是根据蚕茧指标初步确定煮茧工艺,包括渗透、吐水、蒸煮、调整、保护过程的温度及时间。煮茧工艺设计方案目前参考解舒率、茧层率、蚕茧干燥程度、净度(环颣、颣结)等指标进行设置。不同原料的煮茧方法按解舒好、茧层厚的煮茧方法;解舒好、茧层薄的煮茧方法;解舒不良、茧层厚的煮茧方法;解舒不良、茧层薄的煮茧方法;干燥程度不同的煮茧方法;洁净差的煮茧方法[5]进行。煮茧结果按偏生、偏熟、适度、白斑、瘪茧、浮茧等状态来区分。目前煮茧工艺以蚕茧指标大概范围,来指导大概的参数区间,以经验判断为主。这种方法获得的煮茧工艺参数是一种随机、模糊的估算结果,其中反复调整和人为因素等不确定性导致了生产的不稳定和大量浪费且效率低下。

随着信息技术在缫丝行业的应用,在制丝大生产过程中,采集、存储了大量的茧质数据、工艺数据,这些数据若能有效用于煮茧工艺设计,能提高生丝的产量和质量,降低茧耗。

2 制丝生产数据的特点

2.1 多样性

制丝过程中形成了大量的茧质调查数据、煮茧生产数据、缫丝生产数据,这些数据既有关联,又有交叉,还有差异,其类型、形态、来源具有多样性。例如茧源特性一项,涉及的指标就非常多,包括蚕茧原料的基础数据和测试数据两部分,其中原料基础数据包含品种、季别、产地、饲养方式、茧型大小、茧层厚薄等指标;测试数据包含茧丝长、解舒率、茧层率、解舒丝长、单丝纤度、清洁、洁净、万米吊糙等指标。

2.2 时效性与动态性

制丝生产数据时效性、动态性较强,但由于技术局限,制丝生产数据往往存在滞后性,对煮茧工艺不能起到一对一的指导作用,往往通过滞后的数据指导下一

收稿日期:2021-12-03
基金项目:四川省应用基础研究项目(2020YJ0269)
第一作者:段春稳(1983—),女,高级工程师,主要从事丝绸工程技术研究。

批蚕茧进行工艺设计，使得煮茧工艺设计较依赖经验判断。

2.3　复杂性

制丝过程是一个复杂的系统。数据之间的关系呈现不确定性，数据间可能无法通过数学形式表示，数据关系较为复杂。

3　数据挖掘技术

数据挖掘(Data Mining)，又称作数据库知识发现(Knowledge Discovery from Database，KDD)，是从数据中获取价值的一个过程，可以形式化地表示为“数据＋工具＋方法＋目标＋行动＝价值”[6]。数据挖掘分为有指导的数据挖掘和无指导的数据挖掘。有指导的数据挖掘是利用可用的数据建立一个模型，这个模型是一个特定属性的描述；无指导的数据挖掘是在所有的属性中寻找某种关系。其中，分类、估值和预测属于有指导的数据挖掘；关联规则和聚类属于无指导的数据挖掘。

作为无指导的数据挖掘方法的一种，聚类分析是从无标记数据集中获取信息和知识的重要手段，是数据挖掘、统计学、模式识别等领域的重要研究内容[7]。聚类分析算法可以作为一种强有力的能够发现制丝数据之间内在关系的、隐含的信息和知识的工具。结合制丝生产数据特点，探索将聚类分析方法用于制丝数据分析，为煮茧工艺参数设计提供指向性设计方案，是实现煮茧数字化、智能化的有效途径之一。

3.1　K-means 聚类分析算法的实现探讨

K-means(K 均值算法)一种经典的划分方法，是目前应用较广泛的聚类分析方法，K-means 算法的步骤如下[8]：

(1)从数据集中随机选出 k 个数据对象作为初始的聚类中心；

(2)将其他的数据对象按照某种聚类度量划分最近的聚类中心，从而将数据集划分为 k 个簇；

(3)计算每个簇中的数据对象的均值作为新的聚类中心；

(4)重新划分数据对象，不断重复这个过程，直到每个簇中数据对象不再变化为止。

其算法公式为：

$$\sum_{i=0}^{n} \min_{\mu j \in C} (\| x_i - \mu_j \|^2) \tag{1}$$

3.2　制丝数据聚类分析

3.2.1　架构设计

运用 Python 计算机编程语言实现聚类算法，具体通过 Scikit-learn 机器学习框架进行制丝数据分析，在实施方案中可以使用 K-means＋＋ 初始化方案，来解决 K-means 高度依赖于质心初始化和运算效率低的问题，并借助轮廓系数法提高聚类效率。

K-means 聚类分析方法在制丝工艺数据分析中的模式探索：煮茧工艺设计目标最终体现在生丝品质和产量、茧耗上，选择以洁净成绩为聚类核心，对洁净成绩优秀的缫丝工艺进行特征提取，通过构建解舒率、茧层率、蒸煮温度等特征提取算法，测算出能够体现最优煮茧工艺的数值结果。

选择洁净成绩在 94.5 以上的制丝生产样本数据 50 条进行特征提取。对解舒率、茧层率和蒸煮温度 3 个特征值进行数据分析。通过 pandas 库将数据导入 Python，对数据进行清洗和处理，在 Scikit-learn 框架下进行 K-means 聚类。

导入的样本数据的散点图如图 1 所示。

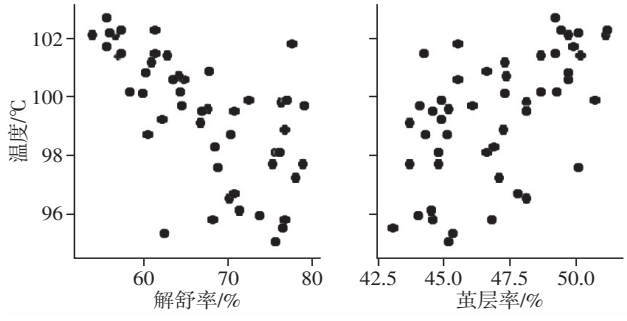


图 1　样本数据散点图

从数据集的散点图不能直观地看出他们之间的规律，将通过聚类分析来探究数据之间的关系。

3.2.2　K-means 算法实现

随机选择 k 值，以 $k=4$ 为例，对数据进行针对解舒率、茧层率、蒸煮温度值的三维聚类代码及聚类结果，如图 2 所示。

```
In [10]: from sklearn.cluster import KMeans
         julei=KMeans(n_clusters=4)#制丝数据聚类分析
         julei.fit(data)

Out[10]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
             n_clusters=4, n_init=10, n_jobs=1, precompute_distances='auto',
             random_state=None, tol=0.0001, verbose=0)

In [11]: label=julei.labels_
         center=julei.cluster_centers_
         label#获得聚类标签

Out[11]: array([1, 0, 3, 1, 2, 1, 1, 0, 0, 1, 3, 3, 0, 2, 1, 0, 3, 0, 3, 2, 3, 3,
                0, 3, 2, 2, 2, 3, 0, 3, 0, 2, 3, 2, 0, 2, 3, 2, 0, 0, 2, 0, 3, 2,
                1, 1, 2, 2, 3, 1])
```

图 2　三维聚类代码及聚类结果

由聚类结果可以看出：上述所列 50 个样本分别按要求划分为 4 簇，所属的聚类标签如图 2 所示，即第一个样本属于第 1 簇，第二个样本属于第 0 簇，第三个样本属于第 3 簇，以此类推。

3.2.3 基于轮廓系数的聚类簇数确定

上述过程得到的聚类结果，我们不能判定其聚类效果，需要多次运行 K-means 算法来确定聚类的簇数 k，聚类效率低。轮廓系数法结合内聚度和分离度 2 种因素，可以对相同原始数据上的不同聚类结果进行评价[9]。因此，我们利用轮廓系数来确定聚类簇数：若 $s(i)$ 的类内内聚度为 $a(i)$，类间分离度为 $b(i)$，则 $s(i)$ 的轮廓系数为：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{2}$$

轮廓系数 $s(i)$ 在 -1 和 1 之间变化，$s(i)$ 的值越接近 1 聚类效果越好。根据公式(2)，对聚类结果进行轮廓系数计算，其代码及运算结果如图 3 所示。

```
In [12]: from sklearn.metrics import silhouette_samples
         xs=silhouette_samples(data,label)#获取制丝数据聚类轮廓系数
         xs
Out[12]: array([0.63058634, 0.47939118, 0.48322759, 0.70513502, 0.23617652,
                0.63295142, 0.63154445, 0.46644834, 0.14188542, 0.68491031,
                0.46528999, 0.61133038, 0.30927693, 0.0963683 , 0.29533563,
                0.40537541, 0.55748961, 0.37379907, 0.617526  , 0.47455292,
                0.60501978, 0.59792075, 0.41573922, 0.50238586, 0.16367357,
                0.32877874, 0.38489503, 0.63139328, 0.42591683, 0.5344568 ,
                0.53467496, 0.4970364 , 0.54274824, 0.34845726, 0.18232048,
                0.3289451 , 0.2818661 , 0.32477498, 0.36382076, 0.22233188,
                0.44474771, 0.21308336, 0.60135652, 0.0539109 , 0.60485943,
                0.13375112, 0.44368351, 0.15125218, 0.58595931, 0.70795992])

In [13]: means=np.mean(xs)
         means
Out[13]: 0.4291264163664125
```

图 3 聚类结果轮廓系数代码及运算结果

$k=4$ 时的轮廓系数为 0.429 126 416 366 412 5。之后，分别计算 $k=2\sim(n-1)$ 的聚类结果的轮廓系数，其代码和结果如图 4 所示。

```
In [14]: def juleipingjia(n):
             julei=KMeans(n_clusters=n)
             julei.fit(data)
             label=julei.labels_
             xs=silhouette_samples(data,label,metric='euclidean')
             means=np.mean(xs)
             return means

In [17]: y=[]
         for n in range(2,49):
             means=juleipingjia(n)
             y.append(means)
         y
Out[17]: [0.5411131546764456,
          0.446184008083087,
          0.4291264163664125,
          0.4195808561053043,
          0.35061654939538545,
          0.34102378561641894,
          0.3283446368604865,
          0.328118921655737,
          0.3271233098336915,
          0.3098502092939366,
          0.3087029797217844,
```

图 4 聚类代码及结果($k=2$)

从结果可以看出当 $k=2$ 时轮廓系数最大，为 0.541 113 154 676 445 6，聚类效果最好。因此，对数据集进行 $k=2$ 的聚类，其聚类代码及结果如图 5 所示。

```
In [19]: from sklearn.cluster import KMeans
         julei=KMeans(n_clusters=2)#制丝数据聚类分析
         julei.fit(data)
Out[19]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
             n_clusters=2, n_init=10, n_jobs=1, precompute_distances='auto',
             random_state=None, tol=0.0001, verbose=0)

In [20]: label=julei.labels_
         center=julei.cluster_centers_
         label#获得聚类标签
Out[20]: array([0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1,
                1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0,
                0, 0, 0, 0, 1, 0])
```

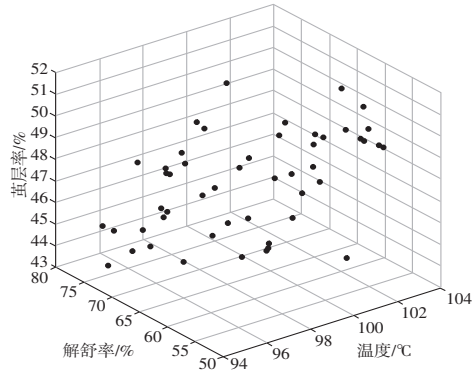
图 5 聚类代码及结果($k=2$)

通过聚类算法，将 50 个样本数据分成了聚类标签为 0 和 1 的 2 类。

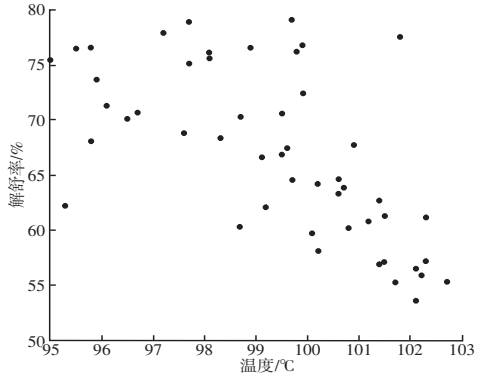
3.2.4 聚类结果有效性评价

在得出聚类结果后，需要对每一类对象进行描述分析，分析这一类对象最典型的共性是什么，从而理解为什么这些对象会被分到一类中，解读这些数据的相似性。这一步需要凭借技术和经验进行人为解读。

为了便于对聚类结果进行解读，通过可视化软件对上述聚类结果进行可视化呈现，如图 6 所示。



(a)三维聚类结果散点图



(b)上视角解舒与温度关系散点图

图 6 样本聚类结果散点图

从图 6 可以看出聚类结果将数据样本聚为边界清晰的 2 类。通过对聚类结果的解读，其聚类所得的 2 个簇，回到数据表本身进行分析，可以得出解舒率、茧层率、煮茧温度之间的对应关系大致区间范围，见表 1。

表 1　解舒率、茧层率及煮茧层对应关系

解舒率/%	茧层率/%	温度/℃
66.6～79.1	43.1～48.1	95.5～101.8
53.6～64.2	45.5～51.2	99.7～102.7

这一结果是通过数据分析得来的，与实际生产经验所得以及教科书指向意见一致。证明了采用聚类分析对制丝数据进行分析，用于指导煮茧工艺设计的可行性。由于 50 个样本数据代表性存在一定的局限，下一步将采用更多的样本数据做进一步分析。

4　结束语

K-means 聚类分析方法在对制丝过程中形成的大量数据进行聚类分析，能够挖掘出数据价值，提取出来的特征数据及其聚类结果对煮茧工艺设计具有一定的指向性。

参考文献：

[1] 陈祥平，卜献鸿，刘季平，等. 煮茧技术及设备的发展与展望[J]. 丝绸，2018，55(8)：21-28.

[2] 张冬梅. 基于轮廓系数的层次聚类算法研究[D]. 秦皇岛：燕山大学，2009.

[3] 王小英. 新编制丝工艺学[M]. 北京：中国纺织出版社，2001.

[4] 苏州丝绸工学院，浙江丝绸工学院. 制丝化学[M]. 北京：中国纺织出版社，1983.

[5] 苏州丝绸工学院，浙江丝绸工学院. 制丝学[M]. 北京：中国纺织出版社，1979.

[6] 邢海龙. 大数据联盟数据挖掘服务模式研究[D]. 哈尔滨：哈尔滨理工大学，2020.

[7] 张远翔. 聚类分析中的最佳聚类数确定方法研究[D]. 合肥：安徽大学，2020.

[8] 程东东，黄金龙，朱庆生. 基于自然邻居的聚类分析和离群检测算法研究[M]. 上海：上海交通大学出版社，2019.

[9] 舒浩浩，陈盛双，李石君. 基于最优 K 均值聚类的时空动态背景模型[J]. 小型微型计算机系统，2019，40(2)：413-419.

Cocoon Cooking Process Parameters Based on Clustering

DUAN Chunwen[1,3], REN Qiangsheng[1,3], BU Xianhong[1,2,3], LI Fan[2,3], LI Gang[1,3], WANG Jianping[1,3]

(1. Sichuan Academy of Silk Sciences Co., Ltd., Chengdu 610031, China;
2. Sichuan Provincial Silk Engineering Research Center, Chengdu 610031, China;
3. Modern Cocoon & Silk Manufacturing Technology Resources Sharing and Service Platform of Sichuan Province, Chengdu 610031, China)

Abstract: The K-means clustering analysis method was introduced, which was used to analyze the data formed in the process of silk production and provide a directional reference scheme for the design of cocoon cooking process.

Key words: clustering analysis; silk making data; cocoon cooking process