微生物学通报
Microbiology China
tongbao@im.ac.cn

JUN 20, 2012, 39(6): 865–871

生物实验室

聚类分析在黄霉素发酵过程中的应用

吴家鑫[1*] 张国栋[1] 刘晓洁[2] 齐鹏[3] 郑应华[1] 何继红[1]
宋敏[1] 葛辛玫[1] 王华丽[1] 曹芹[1]

(1. 中牧实业股份有限公司 北京 100091)
(2. 佰瑞星通讯技术(北京)有限公司 北京 100015)
(3. 中国牧工商(集团)总公司研究院 北京 102206)

摘 要: 【目的】将聚类分析的方法应用于黄霉素摇瓶发酵条件的优化过程中。【方法】通过系统聚类算法、K 均值聚类算法和模糊 C 均值聚类算法对不同批次黄霉素发酵的摇瓶数据的聚类分析进行比较, 发现模糊 C 均值聚类算法优于其他聚类算法, 确定了以模糊 C 均值聚类算法对黄霉素摇瓶发酵数据进行聚类分析。【结果】然后利用模糊 C 均值聚类算法选取优质组样本, 并利用优质样本优化了黄霉素摇瓶发酵的控制参数分布范围。【结论】这充分证明了聚类分析在发酵过程的优化过程中有良好的实用性。

关键词: 聚类分析, 黄霉素, 发酵

Application of cluster analysis on Flavomycin fermentation

WU Jia-Xin[1*] ZHANG Guo-Dong[1] LIU Xiao-Jie[2] QI Peng[3]
ZHENG Ying-Hua[1] HE Ji-Hong[1] SONG Min[1] GE Xin-Mei[1]
WANG Hua-Li[1] CAO Qin[1]

(1. *China Animal Husbandry Industry Co., Ltd., Beijing* 100091, *China*)
(2. *Polystar Communication Technology* (*Beijing*) *Co., Ltd., Beijing* 100015, *China*)
(3. *China Animal Husbandry Group Research Institute*, *Beijing* 102206, *China*)

Abstract: [Objective] The cluster analysis was applied to the optimization of the Flavomycin flask fermentation. **[Methods]** The data of Flavomycin flask fermentation from different batches was clustered by using system cluster, K-means cluster and fuzzy C-means cluster.

*通讯作者：Tel: 86-10-62882135; ✉: icebergwujiaxin@163.com
收稿日期：2011-10-10; 接受日期：2011-11-21

The result showed that fuzzy C-means cluster was better than the other two. **[Results]** Then the classes were separated and the excellent samples were distilled, moreover, the range of control parameter of Flavomycin flask fermentation was optimized. **[Conclusion]** The experiment results indicated that cluster analysis was effective in the optimization of fermentation process.

Keywords: Cluster analysis, Flavomycin, Fermentation

黄霉素(Flavomycin), 又称默诺霉素(Moenomycin)、黄磷脂素(Flavopholopo1)、斑伯霉素(Bambermycin), 于 1955 由 Lindner 等人从 *Streptomyces bambergiensis* 发酵产物中分离得到。黄霉素主要有 5 个组分, 即 A、A_{12}、C_1、C_3 和 C_4, 以 A 组分为主, 含量超过 50%, 5 个组分均有类似的抗菌活性[1−3]。黄霉素通过阻断细菌细胞壁构成物质肽聚糖层的生物合成而抑制细菌的生长, 主要抑制或杀死革兰氏阳性菌, 对革兰氏阴性菌的抑制作用较弱, 对真菌、病毒等基本无效[4−6]; 同时黄霉素也是一种很好的动物生长促进剂, 在很低剂量下即可显著促进动物生长, 提高饲料的转化率[6−8]。

聚类(Clustering)是数据挖掘、模式识别等研究方向的重要研究内容之一, 在识别数据的内在结构方面具有极其重要的作用[9]。聚类分析(Cluster analysis)是一种多变量分析程序, 其目的在于将数据分成几个相异性最大的群组, 同时使得群组内部的相似程度最高, 而群组间的差异尽可能的大。主要用于对没有分类信息的数据进行分析和挖掘, 是一种偏向于探索性的分析方法[10]。近年来聚类分析在生物学和化工领域的运用也越来越广泛[11−13]。

本文利用系统聚类算法、K 均值聚类算法(K-means)和模糊 C 均值聚类算法(Fuzzy C-means, FCM) 3 种不同的聚类分析算法对不同批次黄霉素发酵的摇瓶数据进行处理, 结果发现模糊 C 均值聚类算法具有较好的分类效果, 并利用发酵实验验证了模糊 C 均值聚类算法在发酵条件优化中的实用性。

1 材料与方法

1.1 材料

1.1.1 菌种和仪器: 菌种: 斑伯链霉菌(*Streptomyces bambergiensis*), 菌株名称为 ZM4。

实验仪器: 摇床: HZQ-Y, 哈尔滨市东联电子技术开发有限公司; 旋转蒸发仪: R-124, 瑞士 Buchi 公司; 水浴锅: B-480, 瑞士 Buchi 公司; 离心机: TDL-40B, 上海安亭科学仪器厂; 高效液相色谱: LC-2010A HT, 日本 Shimadzu 公司。

1.1.2 培养基: 种子培养基(g/L): 玉米浆 3.0, 黄豆饼粉 30.0, 葡萄糖 40.0, 氯化钠 3.0, 碳酸钙 3.0, 磷酸氢二钾 1.0, 1×10^5 Pa 灭菌 30 min。

发酵培养基(g/L): 玉米浆 8.0, 黄豆饼粉 30.0, 玉米淀粉 30.0, 色拉油 35.0−60.0, 硫酸铵 3.0, 磷酸氢二铵 0.1, 硫酸镁 0.4, 碳酸钙 4.0, 1×10^5 Pa 灭菌 30 min。

1.2 方法

1.2.1 培养条件: 种子培养: 将 100 mL 种子培养基装入 500 mL 的三角瓶中, 用 8 层纱布包扎瓶口, 灭菌后将平板上的菌落接种于三角瓶内, 摇床转速 200 r/min, 37 ℃ 培养 48 h。

发酵培养: 将 100 mL 发酵培养基装入 500 mL 的三角瓶中, 用 8 层纱布包扎瓶口, 灭菌后将种子液以 10%的接种量接入发酵摇瓶进行培养, 摇床转速 200 r/min, 37 ℃ 培养 168 h。

1.2.2 分析方法: 黄霉素效价的高效液相色谱测定方法: 用十八烷基硅烷键合硅胶为填充物(YMC-Pack ODS-A, S-5 μm, 12 nm, 250 mm×4.6 mm 色谱柱或等效色谱柱); 以 0.2%甲酸铵溶液(用 10%

磷酸溶液调节 pH 至 4.9)-乙腈(55:45, *V*/*V*)为流动相, 流速为 0.5 mL/min, 检测波长为 258 nm。

生物量的测定方法: 将 10 mL 发酵液(体积为 *a*) 5 000 r/min 离心 10 min 后, 倒出上清液, 并准确测量上清液体积(体积为 *b*), 然后计算出菌体的生物量。

计算公式: 生物量=$\frac{a-b}{a}\times 100\%$

发酵液中氨基氮的测定方法: 利用甲醛滴定法进行测量。

发酵液中总糖、还原糖的测定方法: 利用菲林试剂法进行测量。

发酵液中残油的测定方法: (1) 称取一定量发酵液(质量为 *m*, 大约在 5 g 左右, 要求精确到小数点后第 3 位)于 50 mL 塑料离心管; (2) 依次加入 10 mL 25%氨水、5 mL 无水乙醇、25 mL 乙醚, 然后在往复床 200 r/min 上振荡 10 min, 再在离心机上 5 000 r/min 离心 10 min; (3) 称量 250 mL 蒸馏瓶的重量 *M*; (4) 将上清液小心吸入蒸馏瓶, 然后加 15 mL 正己烷、5 mL 无水乙醇于离心管。(5) 振荡 10 min, 离心 10 min 后小心吸出上清液与上次的上清液合并。(6) 在 80 ℃ 水浴锅中进行真空浓缩(真空度约为 1×10^5 Pa), 直到有机溶剂蒸发, 残油被全部析出。(7) 在温度为 100 ℃ 的电热套上烤 20 min, 放入干燥器中。(8) 冷却到室温后称量蒸馏瓶重量 *N*。

计算公式: 残油=$\frac{N-M}{m}\times 100\%$

1.2.3 数据处理软件: 使用 Matlab7.0 进行数据的聚类处理。

2 聚类

2.1 系统聚类

系统聚类的基本思想: *N* 个样本的聚类可以分级逐步进行, 最初是每个样本为一个集团, 然后按一定判据标准将某两个样本连成一个集团, 成为 $N-1$ 个集团, 这样连下去使集团数进一步减少, 一直到 *N* 个样本连成一个大集团为止。然后根据需要或者根据给出的距离临界值(阈值)确定分类数及最终要分的类。由于类与类之间的距离有多种定义方法, 不同的定义法就产生了不同的系统聚类法[14]。

2.2 K 均值聚类

K 均值聚类的基本思想: 首先从 *n* 个数据对象任选 *k* 个对象作为初始聚类中心; 剩下的对象, 则根据它们与这些聚类中心的距离(相似度), 分别将它们分配给与其最相似的类(聚类中心所代表的); 然后再计算每个所获新类的聚类中心(该聚类中所有对象的均值); 一直重复此过程直至标准测度函数收敛为止[15]。

2.3 模糊 C 均值聚类

模糊 C 均值聚类的基本思想: 首先随机初始化隶属度矩阵 $U=[u_{ij}]$, 记为 U^0; 然后根据隶属度矩阵, 利用公式(1)计算聚类中心向量 $C^{(k)}=[c_j]$; 用公式(2)更新隶属度矩阵 $U^{(k)}$, 记为 $U^{(k+1)}$, 当 $\|U^{(k+1)}-U^{(k)}\|<\varepsilon$ 时, 停止迭代, 否则重新计算聚类中心向量, 反复迭代直到收敛[16]。

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^m\cdot x_i}{\sum_{i=1}^{N}u_{ij}^m}, \tag{1}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i-c_j\|}{\|x_i-c_k\|}\right)^{\frac{2}{m-1}}} \tag{2}$$

3 结果与分析

本文将黄霉素摇瓶发酵的 132 组数据以黄霉素效价的高低划分为 3 类, 黄霉素效价大于或者等于 4 000 mg/L 为优质类, 黄霉素效价小于或者等于 3 000 mg/L 为劣质类, 黄霉素效价介于 3 000−4 000 mg/L 之间的为普通类, 利用不同聚类方法对发酵摇瓶的参数进行聚类分析。

3.1 利用系统聚类算法对数据进行分析

在利用系统聚类算法对数据进行聚类的过程中，为了能够比较多种系统聚类方法的聚类效果，本文以欧氏距离作为计算样本亲疏程度的距离，选用了最小距离法、最长距离法、未加权平均距离法、加权平均距离法、质心距离法和内平方距离法等 6 种不同的计算类间距离的方法进行系统聚类。

表 1 不同系统聚类方法比较
Table 1 The comparison of different system cluster

系统聚类方法 Algorithm		数据个数 Sample number	优质组数据个数 The number of excellent samples	优质组数据百分比 The percentage of excellent samples (%)	普通组数据个数 The number of normal samples	普通组数据百分比 The percentage of normal samples (%)	劣质组数据个数 The number of bad samples	劣质组数据百分比 The percentage of bad samples (%)
最小距离法 Single method	第一类 No.1	2	0	0.0	1	50.0	1	50.0
	第二类 No.2	128	46	35.9	58	45.3	24	18.8
	第三类 No.3	2	0	0.0	1	50.0	1	50.0
最长距离法 Complete method	第一类 No.1	17	0	0.0	15	88.2	2	11.8
	第二类 No.2	104	45	43.3	41	39.4	18	17.3
	第三类 No.3	11	1	9.1	4	36.4	6	54.5
未加权平均距离法 Average method	第一类 No.1	2	0	0.0	1	50.0	1	50.0
	第二类 No.2	71	20	28.2	46	64.8	5	7.0
	第三类 No.3	59	26	44.1	13	22.0	20	33.9
加权平均距离法 Weighted method	第一类 No.1	11	1	9.1	4	36.4	6	54.5
	第二类 No.2	70	41	58.6	15	21.4	14	20.0
	第三类 No.3	51	4	7.8	41	80.4	6	11.8
质心距离法 Centroid method	第一类 No.1	48	25	52.1	9	18.8	14	29.2
	第二类 No.2	11	1	9.1	4	36.4	6	54.5
	第三类 No.3	73	20	27.4	47	64.4	6	8.2
内平方距离法 Ward method	第一类 No.1	11	1	9.1	4	36.4	6	54.5
	第二类 No.2	48	25	52.1	9	18.8	14	29.2
	第三类 No.3	73	20	27.4	47	64.4	6	8.2

通过表 1 的比较可以看出, 在以上 6 种系统聚类算法的中, 加权平均距离法的分类效果最好: 第二类为优质组数据, 黄霉素效价大于或者等于 4 000 mg/L 的数据占到该类数据总数的 58.6%, 平均效价为 3 782 mg/L; 第三类为普通组数据, 黄霉素效价介于 3 000−4 000 mg/L 的数据占到该类数据总数的 80.4%, 平均效价为 3 452 mg/L; 第一类为劣质组数据, 黄霉素效价小于或者等于 3 000 mg/L 的数据占到该类数据总数的 54.5%, 平均效价为 3 026 mg/L。

3.2 利用 K 均值聚类算法对数据进行分析

本文采取划分式聚类算法中较为常见的 K 均值聚类算法来对发酵摇瓶的参数进行聚类分析。

通过表 2 可以看出, K 均值聚类算法的分类效果不理想, 很难根据已有分类对应产品效价的分类, 所以可以认为 K 均值聚类算法不适用于该批次黄霉素发酵数据的聚类分析。

3.3 利用模糊 C 均值聚类算法对数据进行分析

本文采用模糊聚类算法中较为常见的模糊 C 均值聚类算法来对发酵摇瓶的参数进行聚类分析。

通过表 3 可以看出, 模糊 C 均值聚类算法的分类效果很好, 且优于加权平均距离法: 第三类为优质组数据, 黄霉素效价大于或者等于 4 000 mg/L 的数据占到该类数据总数的 78.8%, 平均效价为 4 157 mg/L; 第二类为普通组数据, 黄霉素效价介于 3 000−4 000 mg/L 的数据占到该类数据总数的 80.4%, 平均效价为 3 452 mg/L; 第一类为劣

表 2 K 均值聚类算法聚类结果
Table 2 The results of K-means cluster

聚类方法 Algorithm		数据个数 Sample number	优质组数据个数 The number of excellent samples	优质组数据百分比 The percentage of excellent samples (%)	普通组数据个数 The number of normal samples	普通组数据百分比 The percentage of normal samples (%)	劣质组数据个数 The number of bad samples	劣质组数据百分比 The percentage of bad samples (%)
K 均值聚类算法 K-means	第一类 No.1	59	26	44.1	13	22.0	20	33.9
	第二类 No.2	56	20	35.7	32	57.1	4	7.1
	第三类 No.3	17	0	0.0	15	88.2	2	11.8

表 3 模糊 C 均值聚类算法聚类结果
Table 3 The results of fuzzy C-means cluster

聚类方法 Algorithm		数据个数 Sample number	优质组数据个数 The number of excellent samples	优质组数据百分比 The percentage of excellent samples (%)	普通组数据个数 The number of normal samples	普通组数据百分比 The percentage of normal samples (%)	劣质组数据个数 The number of bad samples	劣质组数据百分比 The percentage of bad samples (%)
模糊 C 均值聚类算法 Fuzzy C-means	第一类 No.1	29	1	3.4	8	27.6	20	69.0
	第二类 No.2	51	4	7.8	41	80.4	6	11.8
	第三类 No.3	52	41	78.8	11	21.2	0	0.0

表 4 实验发酵批次的控制参数分布范围
Table 4 The range of control parameter of Flavomycin flask fermentation

种子液生物量 Biomass(%)	pH	总糖 Total sugar (%)	还原糖 Reducing sugar (%)	氨基氮 Amino nitrogen (mg/L)	残油 Residual oil (%)
15−30	6.63−7.21	3.14−4.06	0.48−1.26	840−1036	3.42−5.96

表 5 优质组样本的控制参数分布范围
Table 5 The range of control parameter of the excellent samples

种子液生物量 Biomass (%)	pH	总糖 Total sugar (%)	还原糖 Reducing sugar (%)	氨基氮 Amino nitrogen (mg/L)	残油 Residual oil (%)
19−30	6.70−7.21	3.36−3.92	0.81−1.26	924−952	3.66−5.61

质组数据, 黄霉素效价小于或者等于 3 000 mg/L 的数据占到该类数据总数的 69.0%, 平均效价为 2 822 mg/L。

3.4 优化黄霉素摇瓶发酵条件

黄霉素的发酵过程中控制参数主要有种子液生物量、pH、总糖、还原糖、氨基氮和残油 6 个因素, 且这些控制参数的分布范围很宽泛(表 4), 本文剔除掉模糊 C 均值聚类分析得到的优质组中的异常数据后得到了优质组样本的控制参数范围, 其控制参数的分布范围明显缩小(表 5)。

4 结论

本文通过系统聚类算法(共 6 种算法)、K 均值聚类算法和模糊 C 均值聚类算法对不同批次黄霉素发酵的摇瓶数据的聚类分析进行比较, 确定了以模糊 C 均值聚类算法对本批次黄霉素摇瓶发酵 132 组数据进行聚类分析; 然后利用模糊 C 均值聚类算法分类得到的优质组样本缩小了黄霉素摇瓶发酵的控制参数分布范围, 这充分证明了聚类分析在发酵过程的优化过程中有良好的实用性。为了更好的应用聚类分析的方法, 今后应对连续发酵过程进行深入研究, 以期取得更好的实用意义。

参 考 文 献

[1] Scherkenbeck J, Hiltmann A, Hobert K, et al. Structures of some moenomycin antibiotics-inhibitors of peptidoglycan biosynthesis[J]. Tetrahedron, 1993, 49(15): 3091−3100.

[2] 李蕴玉. 黄霉素的应用效果[J]. 中国饲料, 1999(16): 17−18.

[3] 卜仕金, 徐小艳. 专用抗生素饲料添加剂-黄霉素[J]. 兽药与饲料添加剂, 2003, 8(3): 14−16.

[4] Kurz M, Guba W, Vértesy L. Three-dimensional structure of moenomycin A-a potent inhibitor of penicillin-binding protein 1b[J]. European Journal of Biochemistry, 1998, 252(3): 500−507.

[5] Kempin U, Hennig L, Knoll D, et al. Moenomycin a: new chemistry that allows to attach the antibiotic to reporter groups, solid supports, and proteins[J]. Tetrahedron, 1997, 53(52): 17669−17690.

[6] 陈新凤, 张春永. 黄霉素及其在动物生产中的应用[J]. 兽药与饲料添加剂, 2005, 10(2): 19−21.

[7] 李爱科, 陈清明, 肖希龙, 等. 黄霉素对生长育肥猪的饲养效果研究[J]. 中国饲料, 1995(22): 14−16.

[8] 沈建忠, 肖希龙, 朱蓓蕾, 等. 黄霉素促进肉牛生长的研究[J]. 中国饲料, 1994(10): 13−15.

[9] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件

学报, 2008, 19(1): 48−61.

[10] 吴明隆. SPSS统计应用实务[M]. 北京: 科学出版社, 2003: 234−234.

[11] 孔德翀, 杨雪莲, 严明, 等. 酿酒酵母糖酵解途径中酶量变化对乙醇浓度影响的模拟分析[J]. 生物工程学报, 2007, 23(2): 332−336.

[12] 罗致强, 宫衡, 付水林. 聚类分析在红霉素摇瓶培养基无机盐分析中的应用[J]. 中国抗生素杂志, 2006, 31(3): 172−175.

[13] 李修亮, 苏宏业, 褚健. 基于在线聚类的多模型软测量建模方法[J]. 化工学报, 2007, 58(11): 2834−2839.

[14] 李永刚, 蒋爱平. 结合主元分析和系统聚类的丙烯腈反应器优化[J]. 吉林大学学报: 信息科学版, 2004, 22(4): 310−313.

[15] 周爱武, 于亚飞. K-means聚类算法的研究[J]. 计算机技术与发展, 2011, 21(2): 62−65.

[16] 刘晓洁. 基于PCA的贝叶斯网络构造算法研究与应用[D]. 北京: 北京化工大学硕士学位论文, 2009: 18−19.

稿件书写规范

高校教改纵横栏目简介及撰稿要求

“高校教改纵横”栏目, 是中国微生物学会主办的科技期刊中唯一的教学类栏目, 也是中国自然科学核心期刊中为数不多的教学栏目。该栏目专为微生物学及其相关学科领域高校教师开辟, 一方面为高校微生物学科的教师提供一个发表论文的平台, 同时微生物关联学科的一部分确实优秀的论文也可以在此发表, 是微生物学及相关领域教学研究、交流、提高的园地。

本栏目的文章有别于其他实验类研究报告, 特色非常鲜明。要求作者来自教学第一线, 撰写的稿件内容必须要有新意、要实用, 不是泛泛地叙述教学设计与过程, 而是确实有感而发, 是教学工作中的创新体会, 或者在教学中碰到的值得商榷的、可以与同行讨论的有价值的论题。在内容选材上应该有鲜明的特点和针对性, 做到主题明确、重点突出、层次分明、语言流畅。教师的教学思路应与时俱进, 注意将国内外新的科技成果和教学理念贯穿到教学之中, 只有这样才能真正起到教与学的互动, 促进高校生物学教学的发展, 更多更好地培养出国家需要的高科技创新人才。这也是本栏目的目的所在。

同时, 为了给全国生物学领域的教学工作者提供一个更广阔更高层次的交流平台, 本栏目还开辟了“名课讲堂”版块, 邀约相关生命科学领域, 如微生物学、分子生物学、生物医学、传染病学、环境科学等的教学名师、知名科学家就教学和学生培养发表观点, 推荐在教学改革、教学研究、引进先进教学手段或模式以及学生能力培养等方面有突出成绩的优秀论文, 为高校教师以及硕士、博士研究生导师提供一个可资交流和学习的平台, 促进高校教学和人才培养水平的提高。

欢迎投稿! 欢迎对本栏目多提宝贵意见!