

# 数据挖掘方法在生物实验数据上的应用

辛月振,孙贝贝,夏盛瑜

(中国石油大学(华东)计算机与通信工程学院,山东 青岛 266580)

**摘要:**桑黄是一种具有很大药用价值的真菌,其产物黄酮具有抗癌作用。现阶段对桑黄黄酮的研究主要集中在多糖的药用机理、组成等方面。鉴于桑黄很少存在于野生环境,桑黄黄酮类化合物大多是从实验室培养提取,因此桑黄的实验室培养成为一个非常有前景的研究方向。为了解决生物实验试验周期长、实验数据难以利用的问题,利用桑黄生物实验所得到的数据,包括接种量、PH值、初始液量、温度、种龄、发酵时间和转速等参数,利用数据挖掘的方法,建立高产、低产的分类模型对数据进行分类。随后建立了基于高产数据集的BP神经网络预测模型。最后用遗传算法寻找最佳培养条件。结果表明,预测准确率达90%以上,且预测产量略高于实际产量。

**关键词:**生物信息学;培养条件优化;数据分类;BP神经网络;遗传算法

**中图分类号:**TP39

**文献标识码:**A

**文章编号:**1673-629X(2018)09-0143-04

doi:10.3969/j.issn.1673-629X.2018.09.029

## Application of Data Mining Method in Biological Experiment Data

XIN Yue-zhen, SUN Bei-bei, XIA Sheng-yu

(School of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580, China)

**Abstract:** Phellinus is a kind of fungus with great medicinal value, which is known as one of the elemental components in drugs avoiding cancers. The research on Phellinus focuses on polysaccharides, proteoglycans medicinal mechanism, composition and other aspects. Since Phellinus rarely exists in the wild environment, Phellinus flavonoids are mostly extracted from laboratory cultures. Cultivating Phellinus in the lab becomes a promising research branch. In order to solve the problem of long biological experiment period and difficult use of the experimental data, we use the data obtained by Phellinus experiment including inoculum size, PH value, initial liquid volume, temperature, seed age, fermentation time and speed and other parameters with data mining method to establish a high yield and low yield classification model. Then a BP neural network prediction model based on high yield dataset is established. Finally, the best culture condition is found by genetic algorithm. The result shows that forecasting accuracy rate is more than 90% and the yield we forecast is a slight increase than the real yield.

**Key words:** bioinformatics; optimization of culture conditions; data classification; BP neural network; genetic algorithm

## 0 引言

随着大规模生物实验技术的发展和数据累积,如何处理数据,从全局和系统水平研究和分析生物学系统,揭示其发展规律已成为一个新的研究热点。传统生物数据分析方法受限于其处理能力与时间复杂度,已逐渐不适用于当前的生物数据分析。将计算机技术与生物实验相结合,采用生物信息学的思想与方法成为目前生物数据处理的新途径<sup>[1]</sup>。

近年来,机器学习方法已应用于生物数据处理。在生物数据处理领域,人工神经网络与数据挖掘算法

已应用于产量的优化<sup>[2]</sup>,特别是在培养条件的优化方面。张梅等利用BP神经网络优化杜鹃花黄酮的提取工艺<sup>[3]</sup>。Khaouane L等利用神经网络和粒子群优化算法寻找最优截短侧耳素培养条件<sup>[4]</sup>。最近,随着生物数据的增加,数据分类思想也应用于生物数据处理方面<sup>[5-7]</sup>。分类的概念是在现有数据的基础上使用分类函数,或者构造一个分类模型(即通常称之为分类器)。函数或模型可以将数据库中的数据记录映射到给定的类别,它可以应用于数据预测。在文献[8]中,应用在这些实验中收集的数据,以统计方法建立数学

收稿日期:2017-10-25

修回日期:2018-03-07

网络出版时间:2018-05-16

基金项目:国家自然科学基金(61402187);山东省重点研发项目(2017GGX10147)

作者简介:辛月振(1992-),男,硕士研究生,研究方向为生物信息学、不平衡数据处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20180515.1655.054.html>

模型来预测桑黄产黄酮产量,并取得了较好的效果。但在这个过程中,发现统计方法在处理生物实验数据具有模型建立依赖先验知识,数据受误差样本扰动大,信息易丢失等缺点。因此,文中采用分类算法对整个样本集进行高产和低产的数据分类,取得了良好的分类精度。在高产数据集的基础上,采用 BP 神经网络和遗传算法对产量进行优化。最终得出了最优产量与实验条件。

1 数据采集与分类

1.1 数据采集

首先从生物单因素试验中采集数据。文中所采集的实验数据来源于桑黄实验室发酵实验<sup>[9]</sup>,包括接种量、PH 值、初始液量、温度、种龄、发酵时间和转速等参数。共获取了 90 组实验数据。

1.2 数据分类

将数据集划分为高产量数据集和低产量数据集两部分。由之前的生物数据处理经验,来自生物实验的数据具有不同实验梯度数据相似度高、实验梯度有限等特点。传统的预测方法在整个数据集中很难取得好的结果。所以文中使用分类的方法,针对高产的数据,增加分类数据集中的样本差。选择分类时必须考虑到两个关键因素。

第一,保持两个数据集之间的平衡。较大的不平衡可能导致分类器中更多的偏差<sup>[10]</sup>。类别数据不平衡是分类任务中一个典型存在的问题。简而言之,即数据集中,每个类别下的样本数目相差很大。例如,在一个二分类问题中,共有 100 个样本(100 行数据,每一行数据为一个样本的表征),其中 80 个样本属于 class<sub>1</sub>,其余的 20 个样本属于 class<sub>2</sub>,class<sub>1</sub>:class<sub>2</sub>=80:20=4:1,这便属于类别不平衡。如果使用这种模型,分类器就不能找到高产因子,也不能为 BP 神经网络建立训练数据集。

第二,高产数据集和低产数据集必须覆盖所有单因素实验的实验条件。文中考虑两种分类策略:第一个,取黄酮类化合物产量的中位数作为分类边界(在实验数据中是 1 100 μg/ml),这样获得了数目相同的高产和低产数据集。通过大量实验,证明在此分类边界下分类效果是可以接受的。但是这种方法将会使某些单因素实验因素完全划分为某低产类或高产类当中;另一个策略是在每一组单变量实验中选择一个边界。保持每个单因素实验数据在两个不同的类中,并且尽量使两个类别中的元素数量尽可能接近。结合上述条件,选择黄酮产量为 1 273 μg/ml 作为边界条件。在这个边界条件下,得到 20 组高产量数据和 30 组低产量数据。

分类结果如表 1 所示。

表 1 分类准确率(逻辑回归)

类别	0	1	准确率/%
0	21	10	67.7
1	4	16	80
总计			72.5

2 模型建立

BP(back propagation)神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络,是目前应用最广泛的神经网络之一<sup>[11]</sup>。

基本 BP 算法包括信号的前向传播和误差的反向传播两个过程。即计算误差输出时按从输入到输出的方向进行,而调整权值和阈值则从输出到输入的方向进行<sup>[12]</sup>。

2.1 正向传递子过程

现在设节点*i*和节点*j*之间的权值为 $w_{ij}$ ,节点*j*的阈值为 $b_j$ ,每个节点的输出值为 $x_j$ ,而每个节点的输出值是根据上层所有节点的输出值、当前节点与上一层所有节点的权值和当前节点的阈值还有激活函数来实现的。具体计算方法如下:

$$S_j = \sum_{i=0}^{m-1} w_{ij}x_i + b_j \tag{1}$$

$$x_j = f(S_j) \tag{2}$$

其中, $f$ 为激活函数,一般选取 S 型函数或者线性函数。

2.2 反向传递子过程

反向传递是将输出误差通过隐含层向输入层逐层反传,并将误差分摊给各层所有单元,以从各层获得的误差信号作为调整各单元权值的依据。通过调整输入节点与隐层节点的连接强度和隐层节点与输出节点的连接强度以及阈值,误差沿梯度方向下降,经过反复学习训练,确定与最小误差相对应的网络参数(权值和阈值),训练即告停止。

假设输出层的所有结果为 $d_j$ ,误差函数如下:

$$E(w,b) = 1/2 \sum_{j=0}^{n-1} (d_j - y_j)^2 \tag{3}$$

其中, $E(w,b)$ 为当前位置的梯度。

由经验公式可以确定隐含层节点数目,如下:

$$h = \sqrt{m + n} + a \tag{4}$$

其中, $h$ 为隐含层节点数目; $m$ 为输入层节点数目; $n$ 为输出层节点数目; $a$ 为 1-10 之间的调节常数。经过反复试验确定中间层节点数为 9。

每个隐层传递函数设置为“tansig”(双极性 S 函数)、“logsig”(单极性 S 函数)。训练方法设定为“trainlm”。trainlm 是指 L-M 优化算法<sup>[13]</sup>。

Sigmoid 函数如下:

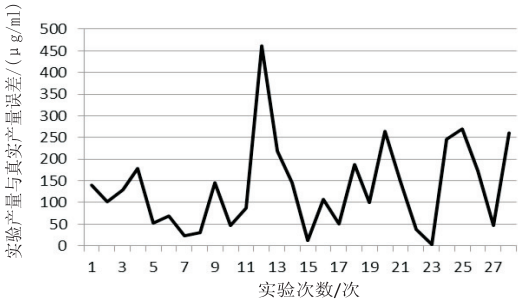
$$S = \frac{1}{1 + e^x}$$

(5)

每次选择 15 组数据进行建模,选择 5 组数据进行验证。训练次数设定为 1 000,训练收敛误差设定为 0.000 01。重复 7 次实验的结果如表 2 所示。平均误差为 133.53,误差百分比为 8.7%。误差值如图 1 所示,误差百分比如图 2 所示。可以判断模型取得了很好的效果。

表 2 BP 预测结果

实际产量	预测产量	误差	错误率/%
1 447.519	1 587.9	140.381	9.70
1 374.983	1 273.6	101.382 6	7.37
1 502.487	1 632	129.513	8.62
1 453.231	1 274.9	178.330 6	12.27
1 506.056	1 453.09	52.966 09	3.52
1 489.364	1 420.734	68.63	4.61
2 127.726	2 103.793	23.932 99	1.12
1 453.231	1 423.269	29.961 77	2.06
1 467.791	1 321.5	146.290 5	9.97
1 273.595	1 320.8	47.205 01	3.71
1 447.519	1 360.8	86.719 17	5.99
1 841.729	1 380.6	461.129 4	25.04
1 374.983	1 592.9	217.917 4	15.85
1 619.554	1 473.6	145.954	9.01
1 597.995	1 586.4	11.595	0.73
1 502.487	1 394.3	108.187	7.20
1 506.056	1 454.8	51.255 6	3.40
1 465.664	1 278.7	186.964	12.76
1 477.273	1 376.9	100.373 5	6.79
1 631.99	1 368.2	263.790 4	16.16
1 447.519	1 300.5	147.019 2	10.16
1 597.995	1 560.9	37.095	2.32
1 320.795	1 317	3.794 994	0.29
1 453.231	1 699.8	246.569 4	16.97
1 841.729	1 571.4	270.329 4	14.68
1 489.364	1 315.7	173.664	11.66
1 320.795	1 274	46.794 99	3.54
1 546.336	1 285.3	261.036	16.88



万方数据 图 1 误差值

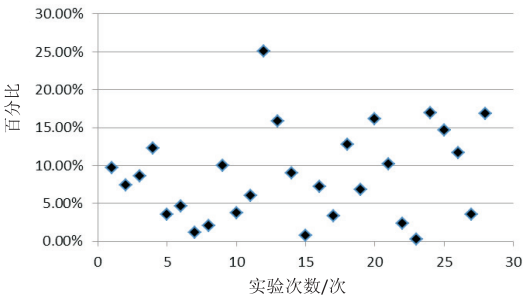


图 2 误差百分比

3 实验仿真与寻优

文中采用遗传算法( genetic algorithm,GA)来优化产量。GA 是模拟达尔文生物进化论中自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法<sup>[14]</sup>。GA 是从代表问题可能潜在的解集的一个种群( population)开始,而一个种群则由经过基因( gene)编码的一定数目个体( individual)组成。每个个体实际上是染色体( chromosome)带有特征的实体。染色体作为遗传物质的主要载体,即多个基因的集合,其内部表现(即基因型)是某种基因组合,决定了个体的形状的外部表现<sup>[15]</sup>。因此,在一开始需要实现从表现型到基因型的映射即编码工作。由于仿照基因编码的工作很复杂,往往进行简化,如二进制编码。遗传算法过程如图 3 所示。

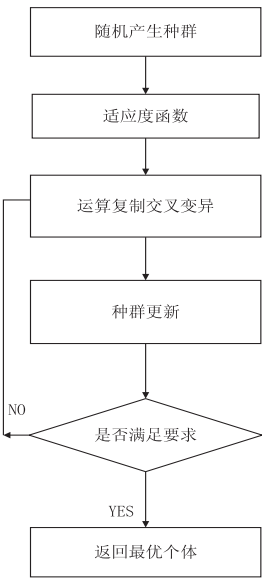


图 3 遗传算法流程

设置 GA 算法的参数如下:种群大小设置为 300,染色体大小设置为 6,交叉速率设置为 1,变异率设置为 0.01。提取 BP 神经网络的隐藏阈值作为 GA 算法的适应度函数。在大约 30 到 500 次迭代之后,GA 过程返回最佳个体。训练过程如图 3 所示。重复测试 7 次,结果如表 3 所示。可以看到,得到的收益比实际收益略有增加。

表3 7次实验预测结果

PH	温度 /℃	装液 量/ml	转速	接种 量	种龄	发酵 时间	桑黄产量 /(μg/ml)	迭代 次数
6	29	100	150	12%	7	8	2164	39
6	28	901	150	12%	8	11	2204	31
6	30	90	150	12%	7	12	2121	208
6	30	90	141	9%	8	8	2045	430
6	28	90	150	12%	8	11	2204	52
6	29	100	150	12%	9	11	2207	44
6	29	100	150	12%	8	8	2171	56

4 结束语

利用桑黄实验数据作为载体,提出了一种利用计算机技术处理生物实验数据的方法。实验结果表明,模型预测的最优条件与生物实验结果一致,证明该方法对培养条件优化具有良好的可预测性。机器学习与数据挖掘的算法在处理大量数的生物数据具有独特优势,是生物信息学潜在的发展方向<sup>[16-17]</sup>。

参考文献:

[1] 王勇献,王正华.生物信息学导论[M].北京:清华大学出版社,2011.

[2] LAVECCHIA A. Machine-learning approaches in drug discovery: methods and applications[J]. Drug Discovery Today,2015,20(3):318-331.

[3] 张梅,潘大仁,周以飞,等. BP神经网络结合正交试验法优选锦锈杜鹃黄酮的提取工艺[J]. 信阳师范学院学报:自然科学版,2011,24(2):261-264.

[4] KHAOUANE L,SI-MOUSSA C,HANINI S,et al. Optimization of culture conditions for the production of Pleuromutilin from Pleurotus Mutilus, using a hybrid method based on central composite design, neural network, and particle swarm optimization[J]. Biotechnology & Bioprocess Engineering, 2012,17(5):1048-1054.

[5] TSAI M F,YU S S. Data mining for bioinformatics: design

with oversampling and performance evaluation[J]. Journal of Medical & Biological Engineering,2015,35(6):775-782.

[6] BAZZAN A L C. Agents and data mining in bioinformatics: joining data gathering and automatic annotation with classification and distributed clustering[M]//Agents and data mining interaction. [s. l.]:Springer-Verlag,2009:3-20.

[7] SAEB A,DAVID S K,RUBEAN K A. Comparative analysis of data mining tools and classification techniques using WEKA in medical bioinformatics[J]. Computer Engineering & Intelligent Systems,2013,4(13):11-17.

[8] LI Zhongwei,XIN Yuezheng,WANG Xun,et al. Optimization to the culture conditions for phellinus production with regression analysis and gene-set based genetic algorithm[J]. Biomed Research International,2016,8(1):1-7.

[9] 刘伟. 药用菌桑黄代代谢黄酮的调控研究[D]. 青岛:中国石油大学(华东),2012.

[10] COHEN G,HILARIO M,SAX H,et al. Learning from imbalanced data in surveillance of nosocomial infection[J]. Artificial Intelligence in Medicine,2006,37(1):7-18.

[11] 梁循. 数据挖掘:建模、算法、应用和系统[J]. 计算机技术与发展,2006,16(1):1-4.

[12] 王崇骏,于汶滢,陈兆乾,等. 一种基于遗传算法的BP神经网络算法及其应用[J]. 南京大学学报:自然科学版,2003,39(5):459-466.

[13] 徐远芳,周旻,郑华. 基于MATLAB的BP神经网络实现研究[J]. 微型电脑应用,2006,22(8):41-44.

[14] 赵志鹏,董红斌. 一种新的基于遗传操作的改进型遗传算法[J]. 计算机应用与软件,2008,25(1):235-237.

[15] MARDLE S,PASCOE S. An overview of genetic algorithms for the solution of optimisation problems[J]. Cheminform, 2007,26(16):1785-1790.

[16] WANG Xun,SONG Tao,GONG Faming,et al. On the computational power of spiking neural P systems with self-organization[J]. Scientific Reports,2016,6(1):24-27.

[17] 张方舟,高晓松. 基于条件函数依赖的挖掘算法研究[J]. 计算机技术与发展,2015,19(5):56-59.

(上接第131页)

[6] 寇文龙,陈莉君. 通用高性能密码服务系统模型[J]. 微电子学与计算机,2016,33(10):87-90.

[7] 林璟铨. 一种基于虚拟化环境中提供密码服务的方法和系统:中国,104461678[P]. 2015-03-25.

[8] 李国,蔡成杭,马晓艳,等. 一种基于宿主机的密码机及密码运算实现方法:中国,105871540[P]. 2016-08-17.

[9] 张晏,岑荣伟,沈宇超,等. 云计算环境下密码服务资源池的应用[J]. 信息安全研究,2016,2(6):558-561.

[10] 涂俊. 云计算—安全资源池化[J]. 信息通信,2017(4):119-120.

[11] KREUTZ D,RAMOS F M V,VERISSIMO P E,et al. Software-defined networking: a comprehensive survey[J]. Proceeding of the IEEE,2015,103(1):10-13.

[12] ROTHSCHILD,ERRAR N,UHLIG S,et al. OFLOPS: an open

framework for OpenFlow switch evaluation[C]//Proceedings of the 13th international conference on passive and active measurement. [s. l.]:[s. n.],2012:85-95.

[13] LARA A,KOLASANI A,RAMAMURTHY B. Network innovation using OpenFlow: a survey[J]. IEEE Communications Surveys & Tutorials,2014,16(1):493-512.

[14] 齐保社. 面向数据中心的 VXLAN 系统设计与实现[D]. 南京:南京大学,2017.

[15] PAUL S,JAIN R,SAMAKA M,et al. Application delivery in multi-cloud environments using software defined networking[J]. Computer Networks,2014,68(11):166-186.

[16] VARADHARAJAN V,TUPAKULA U. Trust enhanced security for cloud environment[C]//IEEE international conference on trust, security and privacy in computing and communications. [s. l.]:IEEE,2012:145-152.