

基于 MATLAB 遗传算法的清酒质量等级优化

吴政治 李 劲

(昆明理工大学质量发展研究院 昆明 650093)

摘要 酒的质量等级评定是一件十分重要的工作。由于酒的质量等级是分类变量,因此不能利用传统的回归模型,但可采用 logistic 回归模型进行建模。文章在结合一次对葡萄牙清酒全面调查所获得的实际数据的基础上,利用有序 logistic 回归构建了清酒质量等级预测模型,并利用了带惩罚函数的遗传算法(GA)进行优化,找出了一组能获得最优质量等级的数据。

关键词 logistic 模型,优化设计,惩罚函数,遗传算法

中图分类号 TP301 **文献标识码** A

Study on Grade of Wine Quality Based on GA Toolbox of MATLAB

WU Zheng-zhi LI Jin

(Quality Development Institution, Kunming University of Science and Technology, Kunming 650093, China)

Abstract The wine quality grade evaluation is a very important work. Because the quality level of wine is a kind of classified variable, traditional regression model can't be able to use, but the logistic regression model can. This paper used order logistic regression to build a prediction model about wine quality grade based on the actual data of a complete investigation of the Portuguese wine. Then it applied genetic algorithm with penalty function to optimize, finding out a group of factors which get the best quality level.

Keywords Logistic model, Design of optimization, Penalty function, Genetic algorithm

1 引言

随着生活质量的提高,人们对酒的质量要求越来越高,如何使酒的质量得到提升,这个问题越来越引起人们关注。线性模型是如今解决一般问题的最简便的方法,但现实中许多实际问题都不能表达为容易解决的线性模型,如酒的质量等级评定。质量等级是一个分类,不具有连续性。对于这种等级分类的模型一般运用逻辑回归进行建模。有序逻辑回归是逻辑回归的一种,该方法能有效获得质量等级概率的累计模型,对于只需要最高质量等级的概率来说,比其他逻辑回归节省了计算步骤。解决非线性约束问题的方法有很多,如可行方向法、惩罚函数法,但这几种方法计算复杂且精度不高。遗传算法是近年来新兴的一种方法。20 世纪 70 年代,美国 Michiga 大学的 Holland 教授以达尔文的生物进化论和孟德尔的遗传变异理论为基础,提出了遗传算法。但是传统的遗传算法有一个大的缺点,即无法解决带约束条件的非线性问题,因此将遗传算法与惩罚函数结合是不错的选择。针对该领域的研究已有不少,如孙仁范^[1]等探讨了遗传算法在桁架结构优化设计中的应用,通过对经典遗传算法引入惩罚函数进行改进,建立了桁架结构的优化设计模型;王帅华等^[2]研究了惩罚函数结合遗传算法在油气储运工程经济中应用的可行性,并针对一个具体的实例进行了求解;柳敏飞^[3]针对单级直齿

圆柱齿轮减速器,以体积最小为目标,用遗传惩罚函数法建立了减速器的体积优化目标函数,用 MATLAB 语言编写了优化程序,并进行了实例优化设计,结果表明:这种优化方法合理有效;王庆荣等^[4]研究了城市公交车调度优化问题,根据公交车排班和调度运行要求,兼顾到乘客和公交公司的利益,建立了基于遗传算法和惩罚函数法的公交智能排班调度模型,并进行了计算机仿真;杨昆等^[5]针对某舰用电站机组负荷组合的特点,提出了一种基于浮点数和二进制数统一编码的改进遗传算法,并对算法中的编码解码方式、约束条件处理、遗传算子映射等作了改进,改进后的算法不仅较好地处理了机组优化组合中的各种约束条件,而且改善了算法的收敛性,优化结果表明机组油耗率降幅最大可达 2%。虽然遗传算法结合惩罚函数用于解决复杂优化问题的研究已有不少,但在酒类质量等级优化领域的应用还很少。本文通过 SPSS 与 MATLAB 工具,应用带惩罚函数的遗传算法对酒类质量等级进行建模与优化,期望有助于企业快速获得酒类质量等级最优时之参数。

2 清酒质量等级预测模型的建立

本节将应用有序 logistic 回归模型,研究清酒的各项因素对清酒质量等级的影响模式。本文的数据来自于葡萄牙米尼奥大学(Univ. Minho)的保罗·科尔特斯(Paulo Cortez)以及

他的同事在 2009 年对葡萄牙清酒进行的一次全面调查。他们收集了所有调查的清酒的 11 项指标,并根据清酒专家对所有的清酒进行的质量等级的评定,从而得到了一张具有 11+1 项因素的清酒数据。11 项数据分别是不挥发酸度、挥发性酸度、柠檬酸、残余糖份、氯化物、自由二氧化硫、总二氧化硫、密度、酸碱度、硫酸盐、酒精。接下来我们用 SPSS 软件进行分析。

在建模过程中,选择质量等级作为因变量,根据专家评定,将清酒质量划分为 6 个等级,等级越高越好,相应地,等级 6 为最好,等级 1 为最差。

针对上述数据,利用 SPSS 软件的有序 logistic 回归命令来建立回归模型。在初步建模中,我们发现除挥发性酸度、氯化物、总二氧化硫、硫酸盐、酒精这几项显著外,其他因素显著水平皆大于 0.05,故我们剔除这几项不显著因子重新建模得到的数据如表 1—表 3 所列。

表 1 模型拟合信息

模型	-2 对数似然值	卡方	df	显著性
仅截距	1618.136			
最终	1356.359	261.777	5	0.000

表 2 拟合度

	卡方	df	显著性
Pearson	2995.586	3040	0.714
偏差	1356.359	3040	1.000

表 3 参数估计值(部分)

阈值	等级 5	估计	标准误差	Wald	df	显著性
	x_1	-2.993	0.463	41.806	1	0.000
	x_2	-4.968	1.631	9.284	1	0.002
位置	x_3	-0.15	0.002	39.259	1	0.000
	x_4	2.284	0.454	25.338	1	0.000
	x_5	0.780	0.082	91.309	1	0.000

由表 1 与表 2 知,该模型的拟合优度较好。在参数估计值表中,各变量的显著水平都小于 0.05,说明变量对于因变量有显著的影响意义。由参数估计值得到 5 个概率累计模型,但本文研究的是使清酒质量等级最高的概率模型,即等级 6 的模型,故我们只用到最后一个等级 5 的模型(如表 3 所列)。等级 5 的模型如下,其中 p_1 为质量等级为 1 的概率, p_2 为质量等级为 2 的概率,以此类推, x_1 为挥发性酸度、 x_2 为氯化物、 x_3 为总二氧化硫、 x_4 为硫酸盐、 x_5 为酒精。

$$\begin{aligned} & \ln\left\{\frac{P(\text{质量等级} \leq 5 | x_1, x_2, x_3, x_4, x_5)}{1 - P(\text{质量等级} \leq 5 | x_1, x_2, x_3, x_4, x_5)}\right\} \\ &= 11.921 - (-2.993x_1 - 4.968x_2 - 0.15x_3 + 2.284x_4 + 0.78x_5) \end{aligned}$$

由于 $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$,所以 $p_6 = 1 - P(\text{质量等级} \leq 5 | x_1, x_2, x_3, x_4, x_5)$ 。即:

$$p_6 = \frac{1}{1 + e^{11.921 - (-2.993x_1 - 4.968x_2 - 0.15x_3 + 2.284x_4 + 0.78x_5)}}$$

3 遗传算法设计

3.1 优化目标与约束处理

在第 2 节我们建立了一个清酒被评为质量等级 6 的概率

模型。在实际问题中,我们需要考虑到各项因素的取值范围。我们用 SPSS 的描述统计得到各项因素在现实中的取值范围。我们要获得的目标是清酒被评为等级 6 的最大概率,所以优化目标为:

$$\text{MAX } p_6 = \frac{1}{1 + e^{11.921 - (-2.993x_1 - 4.968x_2 - 0.15x_3 + 2.284x_4 + 0.78x_5)}}$$

$$\text{s. t. } 0.18 \leq x_1 \leq 1.33$$

$$0.034 \leq x_2 \leq 0.611$$

$$8 \leq x_3 \leq 165$$

$$0.33 \leq x_4 \leq 2$$

$$8.4 \leq x_5 \leq 14.9$$

我们采用惩罚函数处理约束条件。即用下式对个体的适应度进行调整。

$$F'(X) = \begin{cases} F(X), & X \text{ 满足约束条件时} \\ F(X) - P(X), & X \text{ 不满足约束条件时} \end{cases}$$

式中, $F(X)$ 为原适应度函数, $F'(X)$ 为考虑了惩罚函数之后的新适应度函数, $P(X)$ 为惩罚函数。

对于惩罚函数 $P(X)$,适用于 MATLAB 遗传工具箱的有以下两种模式。

$$\text{第一种模式: } P(x) = \sum_{j=1}^k (R_j \times G(x_j))$$

其中, R_j 为第 j 项因素的惩罚系数, k 等于全部因子个数。 $G(x)$ 为因子 x 违反约束的程度,通常为该因子超出界限的程度。这种方法的本质是定义一组不同的系数来单独权衡每个约束,优点是能够比较精确地淘汰不可行解以获得较好的目标函数,缺点是有大量的系数需要设定。

$$\text{第二种模式: } P(x) = C \times \sum_{j=1}^k G(x_j)$$

其中, C 为惩罚系数, $G(x)$ 与第一种模式相同。这种方法的优点显而易见,就是只需要设置一个系数,缺点是过早收敛,不能精确地获得较好的目标函数。

3.2 算法设计

遗传算法工具箱能自动寻找最小值,但本文需要求最大概率,于是将公式稍作更改,将整个函数乘以 -1,从而转成求最小值问题。即:

$$F'(X) = \begin{cases} -F(X), & X \text{ 满足约束条件时} \\ -F(X) + P(X), & X \text{ 不满足约束条件时} \end{cases}$$

根据 3.1 节的约束处理,适应度函数如下,其中 $R_1 = 10, R_2 = 10, R_3 = 0, R_4 = 10, R_5 = 0, C = 10$ 。

第一种模式:

$$p_6 = \frac{-1}{1 + e^{11.921 - (-2.993x_1 - 4.968x_2 - 0.15x_3 + 2.284x_4 + 0.78x_5)}} + 10 \times \max(0, 0.18 - x_1, x_1 - 1.33) + 10 \times \max(0, 0.034 - x_2, x_2 - 0.611) + \max(0, 8 - x_3, x_3 - 165) + 10 \times \max(0, 0.33 - x_4, x_4 - 2) + \max(0, 8.4 - x_5, x_5 - 14.9)$$

第二种模式:

$$p_6 = \frac{-1}{1 + e^{11.921 - (-2.993x_1 - 4.968x_2 - 0.15x_3 + 2.284x_4 + 0.78x_5)}} + 10 \times [\max(0, 0.18 - x_1, x_1 - 1.33) + \max(0, 0.034 - x_2, x_2 - 0.611) + \max(0, 8 - x_3, x_3 - 165) + \max(0, 0.33$$

$$-x_4 x_4 - 2) + \max(0.8, 4 - x_5 x_5 - 14.9)]$$

3.3 遗传算法参数设置

只对几个关键参数进行设置,其余参数使用遗传算法工具箱默认值。

初始种群范围。初始种群范围是指初始种群中的所有个体均处于的范围。如果已知函数的最小值点在某个范围附近,那么就可以直接设置一个初始种群范围,这样可以加快遗传算法的搜索速度。根据多次实验,种群范围设置在 $[0,1]$ 最佳。

种群尺度。种群尺度就是随机生成 N 个初始结构数据,遗传算法以这 N 个初始点开始迭代。种群尺度越大,那么收敛到最优解的可能性越大,但计算时间会大大增加。根据反复实验,将种群尺度设置为 20 时,收敛的效果最好且花费的计算时间较合理。

终止条件。终止条件是指当遗传算法运算达到某项终止条件时,停止运输并给出最佳值。工具箱中包含的终止条件有:代数、时限、适应度限、停滞代数、停滞时限。这里我们将设置代数为 200,停滞代数为 80,停滞时间为 50 秒,其余为默认值。

4 实例计算

(1)首先,在 MATLAB 的 M 文件中编写目标函数,如下所示。

第一种模式:

```
function p=wine_fun(x)
p=-1/(1+exp(11.9221+2.993*x(1)+4.968*x(2)
+0.15*x(3)-2.284*x(4)-0.78*x(5)))+10
*max([0 0.18-x(1) x(1)-1.33])+10*max([0
0.034-x(2) x(2)-0.611])+max([0 8-x(3) x
(3)-165])+10*max([0 0.33-x(4) x(4)-2])+
max([0 8.4-x(5) x(5)-14.9])
```

第二种模式:

```
function p=wine_fun(x)
p=-1/(1+exp(11.9221+2.993*x(1)+4.968*x(2)
+0.15*x(3)-2.284*x(4)-0.78*x(5)))+10
*(max([0 0.18-x(1) x(1)-1.33])+max([0
0.034-x(2) x(2)-0.611])+max([0 8-x(3)
x(3)-165])+max([0 0.33-x(4) x(4)-2])+
max([0 8.4-x(5) x(5)-14.9]))
```

将这 2 个程序分别命名为 wine1.m 与 wine2.m,以供优化时调用。

(2)设置工具箱相关参数。

以第一种模式为例。在 Fitness function 栏里调用已经编辑好的 M 文件作为适应度函数,格式为“@文件名”,本例为“@wine1”。

Number of variables 项填写适应度函数所包含的变量个数,此处填 5。在 plots 选项卡中,勾选 Best fitness。在右边 option 中的 population 选项里,将 population size 设置成 20,

initial range 改为 $[0 1]$ 。最后在 stopping criteria 选项卡中将 Generation 设置为 200, Stall generation 设置为 80, Stall time limit 设置为 50。其余工具箱选项皆为默认;第二种模式唯一改变的地方是在 Fitness function 栏中将文件名改为 wine2,其他选项与上同。

(3)运行及结果分析

点击 Run solver 选项卡中的 star 开始运算,由于遗传算法的随机性,每种模式运行 3 次取最好值。得到的两种模式 Best fitness 图如图 1、图 2 所示。

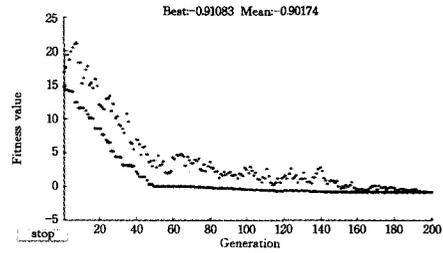


图 1 第一种模式的 Best fitness 图

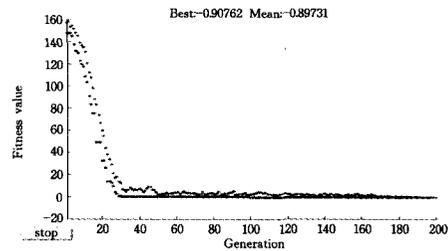


图 2 第二种模式的 Best fitness 图

从图 1、图 2 可知,两种模式的收敛性不相同。模式一得到的初始值比较小,且初始搜索得到的适应值之间的距离比较接近,收敛速度较一般,在 150 代附近开始稳定并趋向于最佳值点。模式二搜索的初始点比较大,且初始得到的适应值之间的距离比较大,但收敛速度非常快,在 30 代左右就已经非常接近最佳值。造成这种情况的原因是两种模式的惩罚系数不同;模式二使用的是一个统一的惩罚系数,因此个别因素受到的惩罚力度大于第一种模式(因子不同惩罚力度也不同),这样就使得收敛性大大加快。但模式二有它自身的缺陷,即由于收敛速度过快,有很大几率会陷入局部最优解,因此要在停滞条件中尽量调高停滞代数以避免过早停滞运算。模式一得到的最佳值为 0.91083,模式二得到的最佳值是 0.90762。模式一的效果略好,但两者相差不大。

取模式一得到的最佳值为准,当最佳值为 0.91083 时,获得的各因素的取值如表 4 所列。

表 4 模式一优化结果各项因素的值

因素	x_1	x_2	x_3	x_4	x_5
取值	0.183	0.037	8.009	1.998	14.89

由表 4 可知,当清酒 5 项因素取以上值时,该清酒将会有很大的概率(约 91%)被评为质量等级 6(即最优)。

结束语 本文以获得质量等级最优为目标,以清酒的 5 项参数为约束条件,利用有序 logistic 回归构建了清酒产品质量最优的数学模型。针对模型的非线性且有约束条件的特

点,使用了带有惩罚函数的遗传算法来寻找最优解。通过对遗传算法的遗传算子、群体规模、初始种群范围以及惩罚因子的选择与调整,获得了寻优结果。但本文也有一点不足之处,即由于数据版权的限定,本文没能得到各项指标的具体单位,以至于无法推广到实践当中,但对于优化酒类质量等级具有较好的指导作用。

参 考 文 献

[1] 孙仁范,牟在根,颜谋,等. 遗传算法在桁架结构优化设计中的应用[J]. 建筑结构学报,2004,25(3):75-81
 [2] 王帅华,秦晓霞,孙海,等. 惩罚函数结合遗传算法在工程经济中的应用[J]. 油气储运,2009,28(10):68-73
 [3] 柳敏飞. 基于 MATLAB 遗传算法的齿轮减速器的优化设计[J]. 组合机床与自动化加工技术,2009,28(10):36-39
 [4] 王庆荣,朱昌盛,梁剑波,等. 基于遗传算法的公交智能排班系统应用研究[J]. 计算机仿真,2011,28(3):345-349

[5] 杨昆,欧阳光耀,陈海龙. 改进型遗传算法在机组负荷优化组合中的应用[J]. 控制理论与应用,2011,28(5):722-726
 [6] 赵瑜,张建伟,张翌娜. GA 及惩罚函数思想在渡槽优化中的应用[J]. 灌溉排水学报,2005(8):73-76
 [7] 邹志红,云逸,王惠文,等. 基于 logistic 回归的水质预测研究[J]. 数学的实践与认识,2008(1):82-87
 [8] 李杨,张维竞. 基于 MATLAB 遗传算法工具箱的水鸟水翼控制研究[J]. 信息技术,2013(5):24-31
 [9] 熊巍,赵海娟,程红莉. 累计 logistic 回归在企业竞争力评价中的应用[J]. 统计与信息论坛,2004(1):85-88
 [10] 雷英杰,张善文,李续武,等. MATLAB 遗传算法工具箱及应用[M]. 西安:西安电子科技大学出版社,2005
 [11] Kurtaran H, Erzurumlu T. Efficient warpage optimization of thin shell plastic parts using response surface methodology and genetic algorithm[J]. The International Journal of Advanced Manufacturing Technology,2006,27(5/6):468-472

(上接第 56 页)

智能交通管理与服务系统”项目通过国家验收。该项目实现了车辆通行管理、出入收费管理、单位车辆管理、车流统计与分析。另外,通过整合已建或在建交通诱导系统、出租车电召系统、闪播网(公交 WiFi)、停车信息及管理系统、电子不停车收费(ETC)系统等,重庆市智能交通系统可向公众提供机动车出行服务、行业应用服务、大众出行路况信息等服务。

在政府智能公交建设思路影响下,重庆新城公交公司与中国联通合作,综合云计算、物联网以及移动互联网技术,建成了智能公交系统,实现了自动语音报站、班车路线管理、实时视频监控、车辆调度管理、超速报警、指定线路行驶、文字信息显示等功能。该系统具体组网方案如图 4 所示。

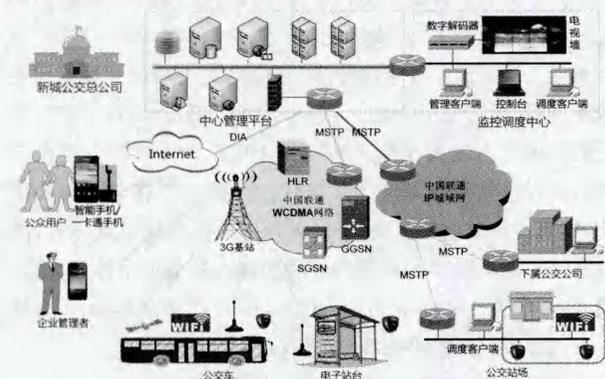


图 4 新城公交公司智能公交系统整体解决方案

智能公交系统整体解决方案主要包括中心管理平台、监控调度中心、前端采集设备、传输网络,其中传输网络由中国联通无线 VPDN 及相应专线组成,完成 3G 智能公交业务数据的采集及监控中心与中心平台设备的互连。

该系统利用重庆联通优势的 WCDMA 网络,同时融合了 GPS 定位系统,对公交行业的音视频数据及定位数据采集点进行联网传输,实现了跨区域的统一管理、资源共享^[7],从而为城市管理者及公交行业提供了一种全新直观的扩大视觉、听觉范围的全方位综合信息服务。

结束语 随着技术的成熟,以及智能交通方面的经验的不断增多,越来越多的城市将会采用全球领先的做法,综合运用物联网、云计算以及移动互联网等技术,实现交通管理的自动化和智能化,有效地缓解交通拥堵,预防交通事故,降低环境污染。

智能交通是智慧城市建设的重要组成部分,是未来城市建设的主方向,涉及相关的市民,以及公安交通管理、交通管理部门车辆管理、城市规划建设、通信等多个部门,若需要整合的服务交付,智能交通还需要政策、资金的支持。

参 考 文 献

[1] 李晖,贾辉然,等. 智能交通系统的研究现状及发展趋势[J]. 产业与科技论坛,2008,7(8):166-251
 [2] 岳梅樱. 智慧城市实践分享系列谈[M]. 北京:电子工业出版社,2012
 [3] 张永民. 智慧城市总体方案[J]. 中国信息世界,2011(3):12-21
 [4] 王家耀. 智慧让城市更美好[J]. 自然杂志,34(3):139-142
 [5] 赵新勇. 公安智能交通管理系统的发展与思考[J]. 中国交通信息产业,2008(2):130-132
 [6] 张开广,孟红玲,等. 洛阳智慧交通系统的设计与实现[J]. 河南科学,2012(2):217-220
 [7] 李勇. 智慧城市建设对城市信息安全的强化与冲击分析[J]. 图书情报工作,2012(3):20-24