

Improved Wine Quality Recognition Based on Particle Swarm Optimization Neural Network

Weishu Li, Jianchao Chen, Guiwu Hu, Heng Zhao, Sitong Yang

School of Statistics and Mathematics Institute, Guangdong University of Finance and Economics, Guangzhou Guangdong
Email: 458662910@qq.com

Received: Jun 6th, 2018; accepted: Jul. 5th, 2018; published: Jul. 12th, 2018

Abstract

With the rise of China's economy, the wine industry has also got on the fast train of China's economic rise. The scale of wine industry has grown, but the quality evaluation of the wine has not kept pace with the development of the wine industry. Today's wine quality assessment methods lag far behind market demand. Aiming at this problem, particle swarm optimization is used to replace the process of BP network self-training, and to establish the PSO optimization BP network model, and then classify and evaluate the quality of wine. Compared with the literature, particle swarm optimization can effectively replace the process of BP neural network training.

Keywords

Wine Industry, Wine Quality Assessment, BP Neural Network, Particle Swarm Optimization

改进粒子群优化神经网络的葡萄酒质量识别

李伟书, 陈建超, 胡桂武, 赵恒, 杨思桐

广东财经大学, 统计与数学学院, 广东 广州
Email: 458662910@qq.com

收稿日期: 2018年6月6日; 录用日期: 2018年7月5日; 发布日期: 2018年7月12日

摘要

随着我国经济的崛起,葡萄酒业也搭上了我国经济崛起的快速列车。葡萄酒产业规模不断壮大,但葡萄酒质量评定却没跟上酒业发展的脚步。现今的葡萄酒质量评定方法远远落后于市场需求。针对这个问题,本文用PSO优化算法代替BP网络自身训练过程,建立PSO优化BP网络模型,进而对葡萄酒质量进行分类评定。经过实证与文献的对比,PSO优化算法的确能够有效的代替BP神经网络自身训练过程。

关键词

葡萄酒业,葡萄酒质量评, BP神经网络, PSO优化算法

Copyright © 2018 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,我国经济水平得到了飞跃的发展,人们的生活水平也随之提高。进而对养生的需求也越来越高。研究表明,适量的饮酒对人体机能起到很好的作用,能调节人体新陈代谢。而葡萄酒则是养生品之一,人们进而喜欢喝葡萄酒,市场对葡萄酒的需求量也随之越来越大。然而市场上葡萄酒的品质参差不齐,如何区分葡萄酒的质量,是现今许多消费者和生产厂家都急切盼望解决的问题,因此鉴别葡萄酒的品质就显得尤为重要。现今的鉴别方法常常是通过理化指标、感官指标来确定。而感官指标主要是通过聘请一批有资质的评酒员进行品评,用感官分析法来对葡萄酒进行等级分类。但这种方法易受到评酒人员嗜好、心理、经验等因素的影响,通常使得评定存在一定主观性和不确定性。因此对葡萄酒质量的评定更多的通过理化指标来确定。通过对葡萄酒的部分理化指标进行分析,建立相应的模型,进而区分出葡萄酒的品质。参考国内外的文献,有许多学者对葡萄酒等级的分类进行研究。

2017年,杨云等人通过构建AdaBoost-SVM集成学习方法将多个基分类器集成,建立以多分类器优化集成为核心的品质分类模型,进而对葡萄酒质量进行分类[1];2017年,屈玉涛等人基于SVM实现了对意大利葡萄酒的分类和预测[2];2016年,霍双红等人先用主成分分析,提取了主要的影响因素,再利用k均值和自组织神经网络算法分别对葡萄酒进行分类和比较[3];2012年,张浩彬在其硕士论文中,基于数据挖掘方法,建立了许多模型,对葡萄酒品质进行鉴定[4]。2009年,Cortez等人提出运用支持向量机对超过上千个葡萄酒样本进行质量评定[5]。

BP神经网络作为一种模式识别方法,具有强大的自学习、自适应、自组织等能力,从而使其能拟合任意的非线性映射。因此,我们将其应用在模式识别、信号处理、优化计算,图像处理,分类、自动目标识别、机器人等领域。而在分类领域,许多学者用BP网络方法进行分类,均取得不错效果,如张喜红运用改进后的BP神经网络,对西洋参进行等级分类[6];王利平将BP神经网络运用到了生物医学上,其用BP网络对红细胞进行识别分类[7]。但由于BP算法是基于梯度的最速下降法,以误差平方为目标函数,因此,不可避免的存在一个缺点——易陷入局部极小值。而粒子群算法作为一种新兴的群智能方法,该算法具有很好的优化能力。将其与BP神经网络结合,可以很好的解决BP网络易陷入局部极小值的情况[8]。

因此, 本文通过葡萄酒的一些理化指标建立了 BP 神经网络模型和 PSO 优化 BP 神经网络模型。通过模型的对比, 能客观的反映葡萄酒质量的评定效果。

2. 数据来源

实证所用数据来自 UCI 数据库(<http://archive.ics.uci.edu/ml/datasets/Wine+Quality>), 此处采用的数据是红葡萄酒的数据, 该数据由 Paulo Cortez 教授提供。数据包含 1599 个样本, 其中有 11 种主要理化指标成分以及一个关于葡萄酒等级的变量指标, 该等级指标有 6 个级别, 分别是级别 3 到级别 8。由表 1 可知, 该原始数据的红酒, 在类别为 3 的样本数量有 10 个, 类别为 8 的样本数量有 18, 均少于其他类别的样本数。为了研究方便, 将这六个级别重新划分, 分为低等, 中等, 高等三类, 并分别用 1/6、1/2、5/6 来表示这三类, 其对应的数据量分别有 63、1319、217 [5] [9]。该数据样本的变量说明如表 2。

3. 模型建立与结果展示

3.1. 模型建立

随着数据种类和数量的增加, 一般的人工处理数据, 就变得不现实。人工智能算法的提出, 可以为广大人民解决许多问题, 能大大的缩短处理时间和减少人力。而 BP 神经网络是人工智能算法中被运用最多的神经网络模型之一。BP 神经网络是 1986 年由 Rumelhart 和 McClland 为首的科学家小组提出, 是一种按误差逆传播算法训练的多层前馈网络。BP 网络能学习和存储大量的输入到输出模式映射关系, 而无需提前揭示这种映射关系的数学方程。它的学习规则是使用最速下降法, 通过反向传播来不断调整网络的权重和阈值, 使网络的误差平方和最小。数学理论已经证明, 三层的 BP 神经网络可以拟合任意的非线性函数。

Table 1. Number of categories

表 1. 各类别数量

类别	3	4	5	6	7	8
数量	10	53	681	638	199	18

Table 2. Variable descriptor

表 2. 变量说明

非挥发性酸含量(x1)	Fixed acidity, g/L, 连续变量
挥发性酸含量(x2)	Volatile Acidity, g/L, 连续变量
柠檬酸含量(x3)	Citric Acid, g/L, 连续变量
残余糖分含量(x4)	Residual Sugar, g/L, 连续变量
氯化钠含量(x5)	Chlorides, g/L, 连续变量
游离二氧化硫含量(x6)	Free Sulfur Dioxide, mg/L, 连续变量
总二氧化硫含量(x7)	Total Sulfur Dioxide, mg/L, 连续变量
密度(x8)	Density, g/mL, 连续变量
酸碱度(x9)	pH, 连续变量
硫酸钾含量(10)	Sulphate, g/L, 连续变量
酒精浓度(x11)	Alcohol, %, 连续变量
品质(y)	Quality, 离散变量

BP 神经网络虽然具有以上的优点，但是具有一个明显的缺点，容易陷入局部极小值。传统的 BP 神经网络为一种局部搜索的优化方法，它要解决的是一个复杂非线性化问题，网络的权值是通过沿局部改善的方向逐渐进行调整，这样会使算法陷入局部极值，权值收敛到局部极小点，从而导致网络训练失败。加上 BP 神经网络算法对初始网络权重非常敏感，以不同的权值初始化网络，其往往会收敛于不同的局部极小，这也是许多学者每次训练得到不同结果的根本原因。因此，针对 BP 神经网络的该缺点，提出用粒子群优化算法优化 BP 神经网络的权值与阈值。

粒子群优化(PSO)算法是近年来发展起来的一种新的进化算法。PSO 算法属于进化算法的一种，和遗传算法相似，它是从随机解出发，通过迭代寻找最优解，它是通过适应度来评价解的品质。它通过追随当前搜索到的最优解来寻找全局最优值。将 BP 神经网络与 PSO 算法结合，通过把 BP 神经网络的权值与阈值看作 PSO 算法里的粒子，经过迭代更新，寻找出全局最优的适应值，最后将最优适应值返回到 BP 神经网络，作为 BP 网络的初始权值。

3.1.1. BP 神经网络模型

BP 神经网络，也叫误差反向传播算法的学习过程，由两个过程组成——信息的正向传播过程和误差的反向传播过程。外界将数据输入到输入层，然后网络通过隐层的处理，再映射到输出层，然后经过与期望值的对比判断，此为正向传播过程。当实际输出与期望值有差距时，将进入误差传播过程。误差通过输出层，按误差梯度下降的方式修正各层权值，向隐层、输入层逐层反传。此两个过程周而复始的进行，直至误差精度达到预先设置的值或者学习次数达到预先设置的次数为止。

BP 神经网络的核心是 BP 算法，一种对于多基本子系统构成的大系统进行微商计算的严格而有效的方法，采用最小均方差学习方式。它的基本思想是梯度下降法，利用梯度搜索技术，以期使网络的实际输出值和期望输出值的误差均方差为最小。

BP 网络中，隐层的节点数一直是研究的难点。实际上，隐层节点数的多少对网络的性能是有影响的，若隐层节点数少了，则网络的拟合度很难达到所要的精度，但节点数多了，训练网络的时间会增加，且有可能出现过拟合的情况。经查阅相关文献资料知，有一个经验公式可以确定隐含层节点数目，其公式如下：

$$h = \sqrt{m+n} + a \quad (1)$$

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。

3.1.2. 改进后的 PSO 优化算法

粒子群优化算法是一种进化计算技术，在 1995 年由 Eberhart 博士和 Kennedy 博士提出，源于对鸟群捕食的行为研究。该算法最初是根据鸟集群活动的规律性启发，进而利用群体智能建立的一个简化模型。其基本思想是：通过群体中个体间的协作和信息共享来寻找最优解。由于该算法简单容易实现且没有许多参数的调节。因此被广泛应用于函数优化、神经网络训练、模糊系统控制以及其他算法的应用领域。

PSO 算法的更新规则是先随机产生一群粒子(随机解)，然后通过迭代找到最优解。在每一次的迭代过程中，粒子通过跟踪两个“极值”(pbest、gbest)来更新自己的速度和位置。在找到最优值(最优适应值)或到达设置的步长前，粒子将通过以下公式来更新自己的速度和位置：

速度更新：

$$v_i(t+1) = v_i(t) + P + G \quad (2)$$

位置更新

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

其中

$$P = c_1 * r_1 * (pbest_i - x_i(t)) \quad (4)$$

$$G = c_2 * r_2 * (gbest - x_i(t)) \quad (5)$$

公式(2)、(3)中, $i=1,2,\dots,n$, n 为此群中粒子的总数, v_i 为粒子的速度, v_i 被限制在 v_{\max} 内, 当速度超出 v_{\max} 时, v_i 将被赋值为 v_{\max} 。 x_i 为粒子当前位置, 与速度同理, 被限制在 x_{\max} 内。 r_1, r_2 为[0,1]内的随机数, c_1, c_2 为常数, 称其为学习因子, t 为当前迭代步数。公式(2)、(3)为 PSO 的标准形式。其中, 公式(2)的第一部分称为记忆项, 表示上一次粒子速度大小和方向的影响; 第二部称为自身认知项, 是从当前点指向粒子自身最好点的一个矢量, 表示粒子的动作来源于自身经验的部分; 第三部分则称为群体认知项, 是一个从当前点指向种群最好点的矢量, 反映了粒子间的协同合作和知识共享。

PSO 算法经过多年的研究, 许多学者对其进行继续优化。其中优化方法之一则是优化公式(2), 对公式(2)添加一个惯性因子 w , 优化后的公式(2)为以下公式:

$$v_i(t+1) = w \times v_i(t) + P + G \quad (6)$$

惯性因子 w 的值为非负。若其值较大, 全局寻优能力强, 局部寻优能力弱; 若其值较小, 全局寻优能力弱, 局部寻优能力强[10]。动态 w 能获得比固定值更好的寻优结果。因此将通过以下公式来动态改变 w 的值:

$$w = w_{\max} - w_e \quad (7)$$

$$w_e = \frac{(w_{\max} - w_{\min}) * t}{t_{\max}} \quad (8)$$

其中, w_{\max} 为取最大惯性权值, 通常取为 0.85, w_{\min} 为取最小惯性权值, 通常取为 0.35。 t 为当前迭代步数, t_{\max} 为最大迭代步数。该模型用误差函数来做 PSO 算法的适应度函数, 该误差函数为一个 account 函数, 该函数即为若预测值 \hat{Y}_i 与期望值 Y_i 之间的差值小于 1/6, 则产生一个值 1, 否则为 0, 最后求 1 的个数, 该数值即为误差函数。产生适应度函数的过程如下:

Step 1: 预测值与期望值的作差;

Step 2: 判断该差值的绝对值是否小于 1/6, 若是, 则取该差值为 1, 否则取为 0;

Step 3: 对该差值进行求和, 命名为 account, 并与此时网络训练样本数作商, 命名为 accuracy, 接着用 1 对 accuracy 作差, 结果为 mse。以 mse 作为 PSO 算法的适应度函数。

通过将 PSO 优化算法应用到 BP 神经网络中, 优化 BP 网络中权值与阈值。用 PSO 算法具有全局寻优能力的特点, 来代替 BP 网络自行训练的过程。PSO 算法替代 BP 网络训练的具体过程步骤如下:

步骤 1: 初始化粒子群。随机产生各粒子的位置与速度;

步骤 2: 构建好 BP 网络架构, 将 BP 网络的权值与阈值看作粒子的解。根据公式(7), 计算各个粒子的适应值;

步骤 3: 将各个粒子的适应值与遍历过的对应最好位置 pbest 的值进行比较, 若此时的适应值较好, 则将该值视作当前粒子最好的位置 pbest;

步骤 4: 将各个粒子的最好位置 pbest 与群体遍历过的最好位置 gbest 作比较, 若较好, 则将对应的 pbest 当作全局最好值 gbest;

步骤 5: 通过公式(2)与公式(3), 对粒子的位置和速度进行更新;

步骤 6: 判断是否达到目标精度或最高迭代次数;

步骤 7: 若步骤 6 判断为是或者程序达到最高迭代步数, 则输出全局最优解; 否则, 则返回步骤四。具体流程如图 1。

3.2. 结果展示

实证所用的计算软件是版本为 2016b 的 MATLAB。BP 模型的网络结构为三层网络。所用实证数据有 11 个指标变量, 1 个质量指标变量。由于变量 x_2 的取值区间为 [0.12, 1.58], 而 x_7 的取值区间为 [6, 289], 由此可知各变量间的量纲不同, 进行网络训练前, 先对各变量进行归一化。所用的归一化公式如下:

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (9)$$

因此, 根据数据特征, 网络的架构为输入层节点数为 11 个, 输出节点数为 1 个。隐层节点数按公式 (1), 确定了隐层节点数在 4~14 间的整数, 最后经过试验验证, 隐层节点数为 8 时, 效果最好。因此, 神经网络整体架构为: 输入是 11 个, 输出是 1 个, 隐层节点个数为 8 个。网络其余参数见表 3。

两个模型的实证总样本为 1599 个, 按 1:4 的原则将数据分为测试集与训练集, 即测试集的样本数为 399 个, 训练集样本数为 1200 个。

由于 BP 网络易陷入局部极小值, 则用 PSO 算法来替代 BP 网络自身训练过程。PSO 优化算法是将 BP 神经网络的权值与阈值当作粒子, 然后通过适应度函数来更新迭代粒子的位置与速度, 最终寻找到全局最优的粒子。粒子群算法的各个参数见下表 4。

为了使结果有可比性, 运用 MATLAB 软件, 设置了每一次运行的训练集数据与测试集数据在两个模型中均一致, 且 BP 网络的构架也一致。

为了降低随机带来的偏差, 对两个模型进行了三十次的模拟, 然后取其均值作为最终的结果。两个模型的运算结果见下表 5。

实证结果表明, 在各自训练三十次后, PSO 优化 BP 网络模型得到的平均准确率均要比单纯的使用 BP 网络模型要高, 两个极值均要优于单纯的 BP 网络模型, 说明 PSOBP 模型的性能要优于单纯的使用 BP 网络模型。但是由方差来看, PSO 优化 BP 网络模型比单纯的 BP 网络要高, 说明 PSOBP 网络的稳定性要低。导致 PSOBP 网络的稳定性比单纯的 BP 网络差的原因, 可能是个别粒子在寻优过程中, 误入了陷阱点, 此陷阱点并非是极小值点, 进而影响 BP 算法在测试过程达不到最优。PSO 算法优化 BP 网络除了在稳定性能上不能优于 BP 算法, 但从其他方面与 BP 算法比较, 在葡萄酒优化方面, 均取得足够的优势, 这也进一步说明该方法是可以代替 BP 网络自身的训练过程。所用的数据与文献[9]相同, 且同样将葡萄酒等级分为三类。因此与该文献的结果做对比最合适。结果对比如下表 6。

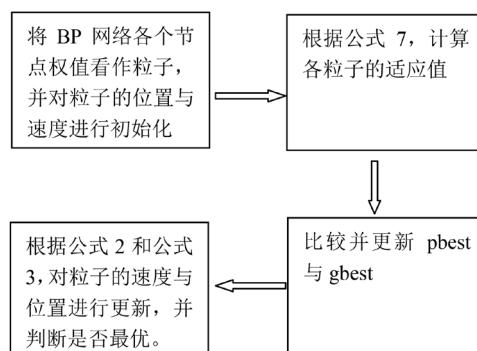


Figure 1. PSO flow diagram
图 1. PSO 算法流程图

Table 3. BP' parameter
表 3. BP 网络结构参数一览表

输入层到隐层的传递函数	logsig
隐层到输出层的传递函数	logsig
网络训练函数	traingd
net.trainParam.epochs	400
net.trainParam.goal	0.001
net.trainParam.lr	0.3

Table 4. PSO' parameter
表 4. 粒子群参数一览表

BP 网络输入层到隐层的传递函数	logsig
BP 网络隐层到输出层的传递函数	logsig
BP 网络训练函数	traingd
PSO 的学习因子 c1、c2	2
PSO 的最大惯性系数	0.85
PSO 的最小惯性系数	0.35
PSO 最大迭代步数	200
PSO 最大速度	1
PSO 最大位置	3
PSO 粒子数	30
PSO 粒子长度	73
PSO 目标精度	0.01

Table 5. Thirty operations
表 5. 三十次结果展示表

模型	均值	最大值	最小值	方差
BP 网络	0.8312	0.9624	0.1203	0.0245
PSOBP 网络	0.8874	0.9774	0.1504	0.0425

Table 6. Comparison of results
表 6. 结果对比表

	模型	总体结果
文献[9]所用模型结果	多元回归	0.7931
	主成分回归	0.8245
	判别分析	0.8119
	支持向量机	0.8276
	分类回归树	0.8213
	随机森林	0.8433
模型结果	PSO 优化 BP 网络	0.8874

由表 6 可知, 文献[9]所建立的众多模型中, 其分类准确率最高的模型是随机森林模型, 达到 84.33%。但该模型的分类准确率略差于的 PSO 优化 BP 网络模型。因此, 所建立的模型, 在葡萄酒分类的应用中, 是有参考意义的。

4. 结论

随着时代的发展,大数据成为了当今热门话题。而大数据的应用也越来越被社会所接受。现今,无论是硬件上对葡萄酒物理化学属性检测工具的出现,还是软件上对葡萄酒品质分类的数据挖掘算法的普及,均为用大数据方法鉴定葡萄酒质量提供了前提条件。目前,如何对智能算法进行优化的问题,是许多研究智能算法的学者在大数据领域关注的较多的方向之一。正是在大数据的背景下,我们运用 PSO 优化算法改进 BP 神经网络的自身易陷入局部极值的不足。结果表明 PSO 优化算法改进 BP 网络是可行的,且是可信的。该模型的不足之处是理论支撑不是很成熟,这也是未来需要继续往下深究的努力方向。

参考文献

- [1] 杨云, 卢美静. 基于集成支持向量机的葡萄酒品质分类方法[J]. 计算机工程与设计, 2017, 38(9): 2541-2545.
- [2] 屈玉涛, 邓万字. 基于 matlab 的 svm 分类预测实现[J]. 信息通信, 2017(3): 33-34.
- [3] 霍双红, 胡红萍, 白艳萍, 王建中. 基于 PCA-K-means 和 PCA-SOM 神经网络的葡萄酒分类[J]. 数学的实践与认识, 2016, 46(17): 168-173.
- [4] 张浩彬. 数据挖掘方法在葡萄酒品质鉴定中的应用[D]: [硕士学位论文]. 广州: 暨南大学, 2014.
- [5] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support Systems*, 47, 547-553. <https://doi.org/10.1016/j.dss.2009.05.016>
- [6] 张喜红. 基于 BP 神经网络的西洋参等级分类方法研究[J]. 云南民族大学学报(自然科学版), 2017, 26(4): 322-326.
- [7] 王利平. 基于 BP 神经网络的红细胞识别分类方法研究与系统实现[D]: [硕士学位论文]. 湘潭: 湘潭大学, 2017.
- [8] 陈雯柏. 神经网络原理与实践[M]. 西安: 西安电子科技大学出版社, 2016: 44-51.
- [9] 林翠香. 基于数据挖掘的葡萄酒质量识别[D]: [硕士学位论文]. 长沙: 中南大学, 2010.
- [10] 徐雅香. 粒子群算法及在神经网络分类器中的应用[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2008.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2163-145X, 即可查询
2. 打开知网首页 <http://cnki.net/>
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: hjdm@hanspub.org