极限学习机在机场旅客吞吐量预测中的应用①

廖洪一, 王 欣

(中国民用航空飞行学院 计算机学院, 广汉 618300)

摘 要: 极限学习机(ELM)是一种新型单馈层神经网络算法, 在训练过程中只需要设置合适的隐藏层节点个数, 随机赋值输入权值和隐藏层偏差, 一次完成无需迭代. 结合遗传算法在预测模型参数寻优方面的优势, 找到极限学习机的最优参数取值, 建立成都双流国际机场旅客吞吐量预测模型, 通过对比支持向量机、BP 神经网络, 分析遗传-极限学习机算法在旅客吞吐量预测中的可行性和优势. 仿真结果表明遗传-极限学习机算法不仅可行, 并且与原始极限学习机算法相比, 在预测精度和训练速度上具有比较明显的优势.

关键词: 极限学习机; 遗传算法; 旅客吞吐量; 预测模型

Application of Extreme Learning Machine Algorithm in Predicting the Airport Passenger Throughput

LIAO Hong-Yi, WANG Xin

(School of Computer Science, Civil Aviation Flight Univercity of China, Guanghan 618300, China)

Abstract: Extreme learning machine (ELM) is a new type of single feed layer neural network algorithm. In the training process ELM only needs to set the hidden layer node number of suitable, random set the input weights and hidden layer deviation, finish in one time without iteration. Now use the genetic algorithm to optimize the extreme learning machine to find the optimal parameter values, so as to establish the Chengdu Shuangliu International Airport passenger throughput prediction model. Then through the comparison of support vector machine, BP neural network, analysis the feasibility and advantage of genetic-extreme learning machine algorithm. The simulation results show that the genetic-extreme learning machine algorithm is not only feasible, and compared with the original extreme learning machine algorithm, it has obvious advantages in prediction accuracy and training speed.

Key words: ELM; genetic algorithm; passenger throughput ; prediction model

近年来, 随着我国交通基础设施的大力建设, 国民收入水平的提高以及旅游业的快速发展, 交通运输市场需求日渐增大. 其中便捷的航空运输作为一种现代化的交通运输方式在经济社会中的重要性也愈发突显, 在满足国民日常出行需求的同时如何达到利益最大化, 正是现阶段需要深入研究的问题.

民航机场旅客吞吐量决定了航空公司对运输市场的规划调整和人力财力投入, 机场旅客吞吐量的增长速度不仅是民航发展的关键指标, 也是衡量我国经济发展水平的的一个重要标志, 关系到民航发展的长远战略决策. 综上所述, 对机场旅客吞吐量的预测, 对于民航业的系统规划、航空公司的投资、市场的合理运行都有着极其重要的现实意义. 目前, 机场旅客吞吐量的预测研究方法主要有灰色系统理论[1]、支持向量回归[2]、BP 神经网络[3]等, 这些方法各有优点, 在机场旅客吞吐量的研究方面都有着深远的影响.

文献[4]对机场旅客吞吐量预测方法进行了总结分析, 认为一元线性回归法的模型过于简单, 精度远远达不到实际应用要求; 计量经济法数据搜集工作量大, 影响因素复杂多变; 时间序列平滑法预测值滞后, 预

① 基金项目:国家自然科学基金民航联合基金(U1233105)

收稿时间:2015-03-13;收到修改稿时间:2015-05-12

测周期过长; 灰色预测方法计算繁琐, 增加了计算量; 传统的如人工神经网络算法等仍存在训练速度慢, 参数的选择容易出现局部最小值等问题, 导致预测的精度不高.

为解决传统神经网络存在的问题, Huang 等人提出了一种新的单馈层神经网络算法--极限学习机算法, 该算法随机设置内权值和偏置值, 一次完成不需要迭代, 大大节约了计算时间缩小了搜索空间. 现提出一种基于遗传算法的极限学习机机场旅客吞吐量预测方法(GA-ELM), 并通过 Matlab7.0 软件对预测模型进行仿真实现.

1 极限学习机

1.1 极限学习机算法理论

Hornik K 的研究[5]表明在紧集(compact input sets)输入的情况下, 单馈层神经网络可以逼近任何连续函数; 在有限集(infinite input sets)输入的情况下, 如果该有限集含有 N 个不同实例, 那么一个具有非线性激励函数的单馈层神经网络, 只需要 N 个隐藏层节点, 就可以无误差的逼近该 N 个样本. 基于以上理论, Huang 等人提出了使用一次学习思想的极限学习机算法[6].

1.2 极限学习机算法模型

一个具有 $\hat{N}$ 个隐藏层节点的单馈层神经网络 SLFNs 的基本模型如下:

$$\sum_{i=1}^{n}\beta_i G\left(a_i x_j+b_i\right),\quad j=1,\cdots,n \tag{1}$$

其中训练样本为 $(\mathrm{x}_i,\mathrm{y}_i)$, $x_i=[\mathrm{x}_{i1},\cdots,\mathrm{x}_{in}]^T\in R^n$ 是网络的训练输入值; $y_i=[\mathrm{y}_{i1},\cdots,\mathrm{y}_{in}]^T\in R^n$ 是网络的训练期望输出集; $G(\mathrm{x})$ 是网络的激励函数; $a_i=[\mathrm{a}_{i1},\cdots,\mathrm{a}_{in}]^T$ 是第 i 个节点的输入权值; b_i 是第 i 个节点的偏置值; $\beta_i=[\beta_{i1},\cdots,\beta_{in}]^T$ 是第 i 个节点的输出权值; $G(\mathrm{x})$ 是激励函数, 可以选择“Sigmoid”、“Sine”或“RBF”等函数.

用 H 表示单隐层的输出矩阵[6]:

$$H(a_1,...,a_{\hat{N}},b_1,...b_{\hat{N}},x_1,...,x_N)=\begin{bmatrix} g(a_1\cdot x_1+b_1) & \cdots & g(a_{\hat{N}}\cdot x_1+b_{\hat{N}}) \\ M & \cdots & M \\ g(a_1\cdot x_N+b_1) & \cdots & g(a_{\hat{N}}\cdot x_N+b_{\hat{N}}) \end{bmatrix}_{N\times\hat{N}} \tag{2}$$

由此极限学习机的训练过程实质上就转换成了求解方程组 $H\beta=Y$, 即 $\left\|H(\hat{\mathrm{a}}_1,\cdots,\hat{\mathrm{a}}_{\hat{N}},\hat{\mathrm{b}}_1,\cdots,\hat{\mathrm{b}}_{\hat{N}})\hat{\beta}-Y\right\|=\min\left\|H(\mathrm{a}_1,\cdots,\mathrm{a}_{\hat{N}},\mathrm{b}_1,\cdots,\mathrm{b}_{\hat{N}})\beta-Y\right\|$ 的最小二乘方值 $\hat{\beta}$.

其中,

$$\beta=\begin{bmatrix}\beta_1^T\\ M\\ \beta_N^T\end{bmatrix}_{N\times M} \tag{3}$$

$$Y=\begin{bmatrix}y_1^T\\ M\\ y_N^T\end{bmatrix}_{N\times M} \tag{4}$$

在理想情况下, 当隐藏层的节点数等于训练样本数, 即 $\hat{N}=N$ 时, 该单隐层神经网络可以无误差的逼近训练样本. 但是在实际情况中, 隐藏层的节点数 $\hat{N}$ 远远小于 N, 因此满足 $H\beta=Y$ 的最小范数二乘方解是 $\beta=H^{+}Y$, 其中 H^{+} 是输出矩阵 H 的 Moore-Penrose 广义逆矩阵.

2 基于遗传算法优化的极限学习机参数选择

由式子(1)-(4)可以看出, 输出权值矩阵 β 是由输入权值矩阵和隐含层偏置值决定的, 理论上在极限学习机算法模型中, 输入权值矩阵和隐含层偏置值是随机设置的, 不需要进行调节以提高算法的学习速度. 但是随机选择的参数并不是最优的参数, 在实际的应用过程中, 可能存在输入权值矩阵和隐含层偏置值为零的无效节点, 需要设置大量的隐含层节点来满足理想的精度, 增加了训练的复杂度, 现使用遗传算法来优化极限学习机算法的参数.

2.1 遗传算法简介

遗传算法是由美国 J. Holland 教授提出的一类由达尔文生物进化论自然淘汰法则而借鉴来的随即搜索算法[7], 自 1975 年遗传算法被提出以来在解决各类组合优化、机器学习、信息处理等方面做出了巨大贡献. 相较于传统的搜索算法, 遗传算法的初始解是随机产生的, 并随之开始搜索, 再通过一定的选择、交叉、变异操作等逐渐迭代来产生新解, 不需要确定的目标函数. 遗传算法的特点在于不依赖于问题的具体领域, 可以进行全局搜索, 避免达到局部最优; 在迭代过程中并不是对参数本身进行优化, 而是优化编码参数的染色体, 从而可以自适应地调整搜索方向, 不再受到约束条件的限制; 遗传算法的兼容性强, 可以嵌入已有的模型之中或和其他优化方法组合, 有效地增强了种群的多样性和进化速度.

遗传算法的迭代过程可由如下步骤描述:

① 根据编码随机产生初始解;

② 适应性函数评价: 计算初始化种群中每个个

体的适应度值;

③ 对父代染色体进行选择、交叉、变异操作得到子代染色体;

④ 适应性函数评价: 计算子代染色体中每个个体的适应度值;

⑤ 检测是否达到精度要求, 如果满足则终止循环, 如果不满足则回到步骤 3 继续执行.

2.2 极限学习机参数优化

为了消除无效节点, 提高模型的泛化能力, 现结合遗传算法和极限学习机算法, 利用遗传算法优化输入权值矩阵$[a_1,a_2,\cdots,a_{\tilde{N}}]$和隐含层偏置值$[b_1,b_2,\cdots,b_{\tilde{N}}]$, 获得最优模型, 优化参数步骤如下:

① 随机产生一组极限学习机算法的输入权值矩阵和隐含层偏置值, 采用二进制编码方式对该组参数进行编码, 随机确定初始种群, 设置种群进化代数 k=0, 最大迭代数 T=100.

② 样本的均方根误差作为适应度函数, 将初始种群个体$\{x_k^i\}_{i=1}^N$带入极限学习机算法中计算输出权值矩阵, 通过训练样本得到均方根误差, 有$R(a,b,x,g)=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-y_i'\right)^2}{n}}$, 误差越大, 适应度越小. 判断是否满足精度要求, 如果满足则是最优参数; 如果不满足则继续知道达到最大迭代数.

③ 选择误差较小的参数, 直接遗传给下一代, 优化问题可以数学描述为:

$$\begin{cases}\min R(a,b,x,g)\\ a\in R^n,b\in R^n,x\in R^n\end{cases}$$

对当前种群再进行交叉、变异操作得到子代染色体, 其中:

$$\text{交叉算法为}\begin{cases}X_i^{t+1}=c_i\cdot X_i^t+(1-c_i)\cdot X_{i+1}^t\\ X_{i+1}^{t+1}=(1-c_i)\cdot X_i^t+c_i\cdot X_{i+1}^t\end{cases}\tag{5}$$

$$\text{变异算法为: } X_i^{t+1}=X_i^t+c_i\tag{6}$$

④重复步骤②, 这样初始种群不断进化, 最终达到训练精度, 得到参数的最优选择.

3 实验研究

本文选取成都双流国际机场 1970 年-2014 年之间旅客吞吐量为样本数据, 其中 1970 年-2009 年之间的数据作为模型的训练样本, 2010 年-2014 年之间的数据作为模型的测试样本. 同时使用支持向量机(SVM)和 BP 神经网络作为对比实验.

3.1 样本数据预处理

由于不可靠的样本数据可能会导致错误的输出, 所以对数据进行预处理对于预测模型来说具有重要的影响, 有效地数据预处理可以避免噪声和不可靠因素的副作用. 数据的归一化处理可以使每一个特性在相同的范围内开始训练过程, 从而加快训练速度. 现对训练样本数据集中的每个数据进行式(7)的计算:

$$S_i'=\frac{S_i-mean(\mathrm{S}_i)}{std(\mathrm{S}_i)}\tag{7}$$

相应的, 预测完成后的反归一化处理为:

$$S_i=mean(\mathrm{S}_i)+S_i'\cdot std(\mathrm{S}_i)\tag{8}$$

式中, S_i为原始数据; $mean(\mathrm{S}_i)$为样本均值; $std(\mathrm{S}_i)$为样本标准差.

表 1 成都双流国际机场旅客吞吐量

年份	旅客吞吐量(人次)	归一化
1970	44680	-0.694354881
1971	51960	-0.693654408
1972	65264	-0.692374313
1973	68224	-0.692089505
1974	82608	-0.690705494
1975	116968	-0.687399416
1976	14654	-0.697243946
1977	169162	-0.682377372
1978	225310	-0.676974879
1979	283868	-0.671340498
1980	292906	-0.670470873
1981	326932	-0.667196932
1982	322116	-0.667660322
1983	238924	-0.675664956
1984	450488	-0.655308522
1985	627578	-0.638269134
1986	866080	-0.615320758
1987	1094198	-0.593371518
1988	1017616	-0.600740146
1989	1115746	-0.591298195
1990	1414098	-0.562591124
1991	1669747	-0.537992884
1992	2232857	-0.483811115
1993	2840049	-0.425387829
1994	3485757	-0.36325858
1995	4155243	-0.298841441
1996	4295283	-0.28536696
1997	4376255	-0.277575931
1998	4384842	-0.2767497
1999	4985883	-0.218918256
2000	5524709	-0.167073065

2001	6244726	-0.097793893
2002	7548680	0.027670996
2003	8196742	0.090026744
2004	11685643	0.425724612
2005	13899929	0.638780554
2006	16280225	0.86780978
2007	18574284	1.088541384
2008	17246806	0.960813044
2009	22637762	1.479524362
2010	25805815	1.784350621
2011	29073719	2.098784423
2012	31595130	2.341391564
2013	33444618	2.51934708
2014	37675232	2.926411683

3.2 实验结果及分析

将经过预处理的数据作为样本集进行遗传算法训练, 计算出最优参数后再用极限学习机预测模型进行仿真实验. 将遗传-极限学习机算法和支持向量机、BP 神经网络两种对比算法的预测结果进行汇总, 结合成都双流国际机场的实际旅客吞吐量, 得到最终的预测结果如表 2 所示.

表 2　旅客吞吐量三种模型预测值

年份	GA-ELM	SVM	BP
2010	26117852	26413587	29987523
2011	30274318	30762421	33841128
2012	33186214	33971365	36928446
2013	34881198	34184692	38745378
2014	39242158	39717953	44581684

从图 1、2、3 可以看出, GA-ELM 模型的预测值与实际值基本保持一致, 即使随着时间的推移, 误差增大, 预测值也能和实际值保持同样的增减趋势. 与此相比 SVM 模型在 2013 年的预测值的增长趋势不明显. BP 模型误差明显大于前两种算法, 且在 2014 年的预测值增长明显过大, 与实际情况不符.

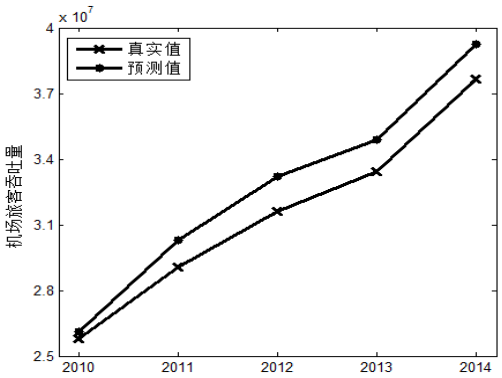

图 1　旅客实际吞吐量与 GA-ELM 预测结果对比

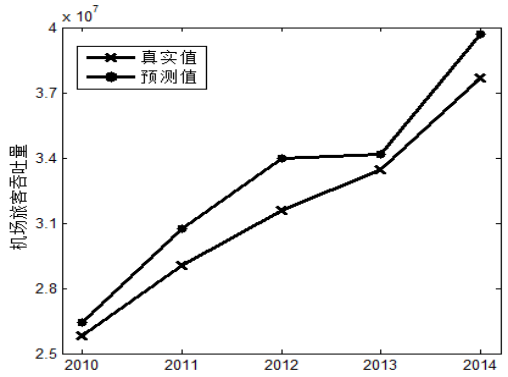

图 2　旅客实际吞吐量与 SVM 预测结果对比

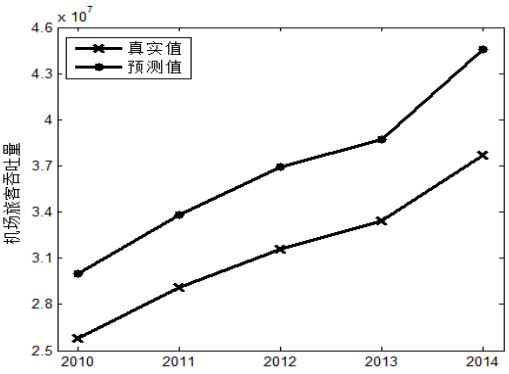

图 3　旅客实际吞吐量与 BP 预测结果对比

表 3　三种模型误差(%)比较

年份	GA-ELM	SVM	BP
2010	1.21	2.36	16.21
2011	4.13	5.81	16.39
2012	5.04	7.52	16.88
2013	4.30	2.21	15.85
2014	4.16	5.42	18.33

为更好的比较三种算法的预测结果现引入下面三种评价指标对不同预测模型的结果做出评价:

①均方根相对误差(RMSE):

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/y_i^2}{n}}\times 100\% \tag{9}$$

②平均绝对百分误差(MAPE):

$$MAPE=\left(\frac{100}{n}\right)\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i} \tag{10}$$

③相关系数(R):

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\times\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}} \tag{11}$$

式中: y_i 为样本实际值; $\hat{y}_i$ 为训练模型预测值; T_0

表示起始时间.

表 4 三种模型评价指标(%)对比

预测模型	RMSE	MAPE	R
GA-ELM	3.99	3.77	0.9981
SVM	5.10	4.66	0.9899
BP	16.76	16.73	0.9989

表 4 的实验数据表明根据三种评价指标来看, 与支持向量机算法、BP 神经网络算法相比较, 遗传-极限学习机算法的平均绝对百分误差较小, 预测能力和预测精度明显优于其他两种, 速度更快并且更为稳定, 因此预测的整体性能优于两种对比算法. 同时从实验结果也可以看出改进的极限学习机在单一时间变量预测方面可以有更加广泛的应用.

4 结语

对机场旅客吞吐量的精确预测不仅能够给航空公司带来巨大的经济效益, 也对民航的科学合理发展规划提供了重要的参考数据. 针对极限学习机对没有在训练集中出现的样本反应能力较差的问题, 本文建立了一种基于遗传算法优化的极限学习机算法模型, 以成都双流国际机场的旅客吞吐量预测为例, 对模型进行了应用研究, 取得了比较满意的效果. 预测结果表明与支持向量机、BP 神经网络相比, 该方法表现出了更高的预测精度. 本文只考虑了旅客吞吐量时间单一变量, 而机场旅客吞吐量一般要受到当年城市 GDP、人口总数、社会消费水平等复杂因素的影响, 增加了预测的难度, 因此预测模型还有待进一步的改进.

参考文献

1 屈拓.组合模型在机场旅客吞吐量预测中的应用.计算机仿真,2012,29(4):108–111.

2 关静.基于灰色支持向量机的民航旅客吞吐量预测.大连交通大学学报,2013,34(3):41–43.

3 基于 BP 神经网络的港口吞吐量预测模型.系统科学学报, 2012, 20(4): 88–91.

4 张慧,王喆.机场吞吐量预测方法探讨.中国民用航空,2008, 10(94): 67–68.

5 Hornik K. Approximation capabilities of multilayer feed-forward networks. Neural Networks, 1991, 4(2): 251–257.

6 Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and application. NeuroComputing, 2006, 70(1-3): 489– 501.

7 马永杰,云文霞.遗传算法研究进展.计算机应用研究, 2012,29(4):1201–1210.