基于改进的人工神经网络对存储系统性能进行预测的方法

郭 佳

(北京交通大学计算机与信息技术学院 北京 100044) (国家保密科技测评中心 北京 100044)

摘 要 测量和评估网络存储系统的性能是用户和企业普遍关心的重点问题之一,因 BP 神经网络具有强大的非线性映射能力,文中提出了一种利用改进的 BP 神经网络实现对网络 IO 性能进行预测的方法。改进的主要内容包括:1)利用马尔科夫链进行预测,更新输出层输出;2)当算法选择概率达到一定值后,利用人工蜂群算法对权值进行优化。最后模拟预测模型的实现过程,将预测结果与传统的 BP 神经网络进行对比。实验结果证明:该算法能够在基本不增加算法运行时间的情况下提高存储性能预测的求解精度和收敛速度。

关键词 存储系统,BP 神经网络,马尔科夫链,人工蜂群算法

中图法分类号 TP389 **文献标识码** A

Method of Predicting Performance of Storage System Based on Improved Artificial Neural Network

GUO Jia

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

(National Secrecy Science and Technology Evaluation Center, Beijing 100044, China)

Abstract Measuring and evaluating the performance of network storage system is one of the key problems to users and corporations. For the strong nonlinear mapping function of the BP-ANN, a new improved algorithm for network I/O performance prediction was proposed by improved BP-ANN, and the new algorithm includes two aspects. Firstly, Markov Chain is used to forecast and update the output of output layer. Secondly, the artificial bee colony algorithm is used to optimize the weights when the probability of algorithm selection reaches a certain value. The implementation process of evaluation model was simulated, and the results were compared with BP-ANN. The experimental results show that the presented approach can significantly improve the solution accuracy and convergence speed of evaluating the performance of network storage system almost without increasing the running time.

Keywords Storage systems, BP-ANN, Markov chain, ABC

1 引言

随着存储技术的飞速发展,海量数据处理需求的日益迫切,用户对数据存储系统在性能方面提出了更高的要求。对于存储系统性能方面的评价和研究,国内外都做了大量工作[1],利用神经网络对网络存储过程进行建模,可以实现对网络性能的预测。神经网络的学习就是对权值的调整过程,因此,权值的存储和修改是人工神经网络学习算法实现的关键技术之一[2]。

Karaboga 于 2005 年提出人工蜂群算法(Artificial Bee Colony algorithm, ABC)用于优化问题[3],之后于 2007 年首次提出将人工蜂群算法应用于人工神经网络最优权值阈值的改进过程[4]。文献[5]将人工神经网络的结构及激励函数也作为人工蜂群算法寻找最优解的一部分,并提出将网络连接率与均方差一起作为适应度函数。人工蜂群算法是较为优越的全局优化算法,但具有更新值选取随机、最差蜜源进化几率小等缺点。部分研究将人工蜂群算法与已有的算法相结合,文献[6]提出基于混沌搜索与人工蜂群算法的混合算法。部分研究着重改进人工蜂群算法的自身参数,如文献[7]受 PSO 的启发,在雇佣蜂和跟随蜂搜索邻域解时,用全局最优解 gbest 代替随机生成的邻域解。文献[8]在文献[7]的基础上,采用正交实验设计的方式来生成新的食物源,保证侦察蜂同时保存放弃的解和全局最优解(gbest)在不同维度上的有益信息。文献[9]在文献[7]的基础上加入选择因子,避免算法过快地收敛至局部最优解。文献[10]改进了侦察蜂的随机侦察机制,用反向学习策略(OBL)产生侦察蜂的新蜜源,降低了蜜源搜索的随机性,文献[11]和文献[12]对跟随蜂选择过程中的轮盘赌算法进行了改进。

马尔科夫链(Markov Chain)建立在系统"状态"和"状态转移"的概念之上[13-14],是一种描述动态随机现象的数学模型。其根据概率分布从一种状态转变为另外一种状态[15],已应用于众多实际领域[16-17]。文献[18]结合 BP 神经网络和马尔科夫链,建立了预测存储系统 IO 负载的模型,利用 BP 神经网络实现预测,利用马尔科夫链对相对误差进行修正。文献[19]采用黄金分割原理对误差修正过程中的状态空间进行分割,但均将马尔科夫链算法用于网络训练后的预测,未运用于训练过程中权值的更改。

2 神经网络及其参数优化算法的相关概念

2.1 BP 神经网络

误差反向传播算法(Error Back Propagtion, BP)的提出,

郭 佳(1983—),女,博士生,工程师,主要研究方向为信息安全,E-mail:m13581902161@163.com(通信作者)。

解决了多层神经网络的学习问题,其包含输入层、输出层和一个或多个隐层,输入层和输出层根据实际需求确定神经元个数。标准的 BP 网络属于有监督学习,训练时先使用随机值作为权值,根据输出值与目标输出计算误差,沿着误差性能函数梯度的反方向修改权值,直至误差不再下降。其网络结构如图 1 所示。

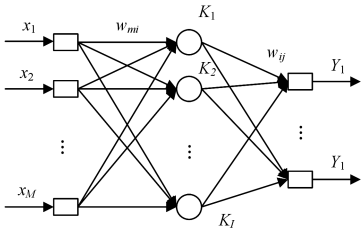

图 1　三层 BP 网络结构

2.2　马尔科夫链

如果对任一 $n>1$,任意 $i_1,i_2,\cdots,i_{n-1},j\in S$,恒有 $P\{X_n=j|X=i_1,X_2=i_2,\cdots,X_{n-1}=i_{n-1}\}=P\{X_n=j|X_{n-1}=i_{n-1}\}$,则称离散型随机过程$\{X_t,t\in T\}$为马尔科夫链。

如果在时刻 t_n 系统的状态为 $X_n=i$ 的条件下,在下一时刻 t_{n+1} 系统状态为 $X_{n+1}=j$ 的概率 $P_{ij}(n)$ 与 n 无关,则称马尔科夫链是齐次马尔科夫链,由状态 E_i 经过 n 步转移到状态 E_j 的转移概率为:

$$P_{ij}^{(n)}=\frac{l_{ij}^{(n)}}{L_i} \tag{1}$$

其中,L_i 表示状态 E_i 出现的总次数,$l_{ij}^{(n)}$ 表示状态 E_i 经过 n 步转移到状态 E_j 的次数。

马尔科夫链是否适用于预测值的改进需要进行检验,使用 χ^2 统计量来检验马尔科夫的适用性。假设状态序列包含 n 个可能的状态,计算“边际概率”:

$$P\cdot j=\frac{\sum_{i=1}^{m}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m}f_{ij}} \tag{2}$$

其中,f_{ij} 表示状态序列 $x_1,x_2,\cdots,x_n$ 中从某一状态 i 经过一次跳转到达状态 j 的频数。

当 n 充分大时,统计量

$$\chi^2=2\sum_{i=1}^{m}\sum_{j=1}^{m}f_{ij}\left|\log\frac{P_{ij}}{P_{.j}}\right| \tag{3}$$

服从自由度为$(m-1)^2$ 的 χ^2 分布,其中 P_{ij} 为状态转移概率矩阵。

给定显著性水平 α,查表得分位点 $\chi_\alpha^2((m-1)^2)$的值,计算后得统计量 χ^2。如果计算结果 $\chi^2>\chi_\alpha^2((m-1)^2)$,则证明符合马氏性,否则状态空间不适合使用马尔科夫链进行优化。

2.3　人工蜂群算法

人工蜂群算法(ABC)将蜜蜂分为雇佣蜂、跟随蜂和侦察蜂[20]。雇佣蜂负责勘探蜜源,并在蜂巢的舞蹈区通过摇摆舞的形式将蜜源信息分享给蜜蜂,这些信息包括蜜源的位置和蜜蜂量的多少;之后,跟随蜂根据收到的信息,以一定的概率选择某一蜜源继续开采,若该蜜源的蜜蜂量越多,则被选中的概率就越高;如果一个蜜源的适应值连续 $limit$ 次未更新,那么说明它已被开采殆尽,需要放弃该位置,在此,采蜜的雇佣蜂就转成侦察蜂,再重新随机搜索一个新的蜜源。

3　基于改进的神经网络算法预测模型

BP 网络在训练过程中通过输出误差调整权值,为更好获取 IO 响应时间的规律,提高预测精度,本文从两方面提出 BP 算法的改进模型:1)基于马尔科夫链适用于多种复杂因素影响时间序列预测的特性,将马尔科夫链滚动预测方法应用于训练过程中输出值的调整,训练过程的前 N 次采用 BP 算法,第 $N+1$ 次采用马尔科夫算法更新 BP 网络输出,将新的输出与训练样本比对并计算误差,根据误差更新权值;2)利用人工蜂群算法改进连接权值,在避免取得局部最优解的同时提高算法的收敛速度,在 BP 神经网络对各层的连接权值进行训练调整时,启用人工蜂群算法对网络的权值进行优化。网络性能预测算法的流程如图 2 所示。

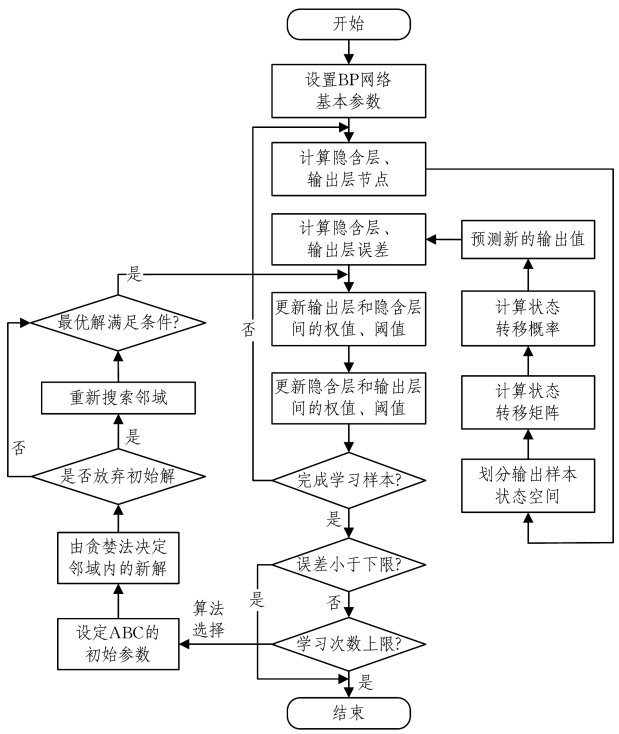

图 2　网络性能预测算法流程图

预测步骤如下:

步骤 1　确定 BP 神经网络初始结构及参数

(1)确定 BP 神经网络结构为 3 层,输入节点为 5 个,输出节点为 1 个、隐含层 11 个,激励函数为 Sigmoid 函数,输出层使用线性函数,初始阈值为(0,1)之间的随机数,最大训练次数为 50 次,训练误差性能目标值为 0.01。

(2)确定初始权值为 ω_{ij},ν_{ij},初始值为较小的非零随机值,范围为$(-3/\sqrt{F},3/\sqrt{F})$,其中 F 为权值输入端连接的神经元个数。

(3)选定训练样本数据,输入参数为 IO 请求数、IO 请求到达数和读写百分比,输出参数为 IO 响应时间。

步骤 2　利用训练数据对 BP 神经网络进行训练

(4)利用 BP 神经网络模型对数据进行预测,对响应时间实测值和预测值进行拟合,计算误差 $f(t)$、相对误差 $f'(t)$和误差均值$\bar{f}$,用式(4)对 $f'(t)$进行归一化处理:

$$F^{1\prime}(t)=\frac{f'(t)-0.5(f'_{\max}(t)+f'_{\min}(t))}{0.5(f'_{\max}(t)-f'_{\min}(t))} \tag{4}$$

其中，$f'_{\max}(t)$和 $f'_{\min}(t)$分别为相对误差的最大值和最小值，$F^{1\prime}(t)$为归一化处理后的相对误差。

(5)根据标准的 BP 网络沿着误差性能函数梯度的反向修改权值(最速下降法)，得到权值修正量 $\Delta\omega_{ij}$ 和 $\Delta\upsilon_{ij}$。

(6)利用式(5)、式(6)对初始权值进行修正：

$$\omega_{ij}=\omega_{ij}(0)+\Delta\omega_{ij} \tag{5}$$

$$\upsilon_{ij}=\upsilon_{ij}(0)+\Delta\upsilon_{ij} \tag{6}$$

步骤 3 利用人工蜂群算法改进网络连接权值

(7)初始化人工蜂群算法参数，将修正后的权值作为人工蜂群算法的初始解(X_0，X_1，…，X_N)，其中 N 为蜜源的个数；每个解的维度为输入层节点与隐层节点之积再加上隐层节点与输出层节点之积，即蜜源的维度 D；雇佣蜂的数量 N_c 与跟随蜂的数量 N_g 一样，蜜源的个数 N 即为解的个数，蜂群的总数量 $N=N_c+N_g$。

(8)将神经网络误差函数的倒数作为人工蜂群算法的目标函数，如式(7)和式(8)所示，按函数值大小进行排列，$f(X_i)$越大说明误差越小，排列越靠前。

$$J=\frac{1}{2}\sum_{p=1}^{P}f_p^2(t) \tag{7}$$

$$f(X_i)=\frac{1}{J} \tag{8}$$

其中，P 为样本的次数。

(9)雇佣蜂对第 n 次蜜源搜索找到可行性解 $X_i(0)$，在其周围区域寻找更高收益率的蜜源：

$$V_i^j=X_i^j+rand(-1,1)(X_i^j-X_k^j) \tag{9}$$

其中，$i\in(1,2,3,\cdots,NC)$，$k\in(1,2,3,\cdots,NC)$，随机选择 $k\neq i$，$j\in(1,2,3,\cdots,D)$。

(10)所有雇佣蜂完成搜索后，跟随蜂根据接收到的信息随机选择一个蜜源进行下一步的开采，本文采用轮盘赌机制，一个蜜源被选中的概率为：

$$P=\frac{f(X_i)}{\sum_{i=1}^{SN}f(X_i)}i \tag{10}$$

(11)当雇佣蜂对应的蜜源适应值连续 $limit$ 次未更新，说明该蜜源已经开采完，相应的雇佣蜂采用混沌搜索产生新的蜜源：

$$X_i^j=X_{\min}^j+rand(-1,1)(X_{\max}^j-X_{\min}^j) \tag{11}$$

其中，$X_{\max}^j$和 $X_{\min}^j$分别为解的上下界。

步骤 4 利用马尔科夫链改进网络输出值

(12)以前 N 次循环输出的 I/O 响应时间为基准时间进行预测，利用黄金分割率原理，按照式(12)计算分割单位长度：

$$\lambda_i=0.618^q\overline{f(x)} \tag{12}$$

其中，$f(x)$为平均值，q 为划分的状态空间个数。

(13)根据分割单位长度，在最大值和最小值范围内划分数据的状态区间。

(14)计算马尔科夫链的状态转移矩阵，选择离预测时步最近的前 3 个时步，合计概率最大的状态区间为下一个时步可能出现的状态区间。

(15)确定该状态空间的中点为最可能的响应时间马尔科夫链预测值，利用式(13)进行：

$$f'(t)=(1+\bar{\Delta})f(t)=\left[1+\frac{\Delta_U+\Delta_D}{2}\right]f(t) \tag{13}$$

其中，Δ_U 与 Δ_D 分别为某一变动区间的上下限，$\bar{\Delta}$为平均相对误差，$f(t)$为第 $N+1$ 次循环神经网络的输出值，$f'(t)$为利用马尔科夫链进行预测的预测值。

(16)将 $f'(t)$返回至步骤 3 作为神经网络输出值，计算其与实测值的误差。

4 模拟预测实验

4.1 样本数据的选择

本实验采用的存储设备配置为 RAID5，测试主机 CPU 为 1.8 GHz，AMD Athlon 处理器，2.00 GB 内存。通过回放合成 I/O 负载值评估模型，回放 6 次然后取算术平均值。将 20 组数据分为两个部分，前 16 组用于训练网络，后 4 组用于检验模型的有效性。根据式(12)计算状态空间范围。

表 1 测试数据

序号	I/O 请求数	是否随机读写	读写百分比	响应时间/ms
训练数据				
1	20	1	0.10	21.4462
2	30	1	0.15	21.9037
3	40	0	0.20	21.6896
4	50	0	0.25	21.1457
5	60	1	0.35	21.4657
6	70	1	0.55	21.9085
7	80	0	0.45	21.1318
8	90	0	0.55	21.5850
9	100	0	0.25	21.2790
10	110	1	0.15	21.7472
11	120	1	0.25	21.1235
12	130	0	0.45	21.0734
13	140	0	0.65	21.6754
14	150	1	0.75	21.2605
15	160	1	0.45	21.2109
16	170	0	0.35	21.8223
检验数据				
17	180	1	0.65	21.2567
18	190	1	0.75	21.4586
19	200	0	0.45	21.5782
20	210	0	0.35	21.1273

4.2 样本数据“马氏性”检验

按照预测步骤中的第(12)、(13)步，将训练数据的响应时间划分为不同状态区间：[3,4,3,1,3,4,1,3,2,4,1,1,3,2,1,4]，可得一步、二步和三步状态转移概率矩阵及 χ^2 概率分布表，分别如表 2 和表 3 所列。

表 2 状态转移概率矩阵及边际概率

	一步状态转移概率				二步状态转移概率				三步状态转移概率			
	1	2	3	4	1	2	3	4	1	2	3	4
1	0.200	0.000	0.600	0.200	0.000	0.500	0.250	0.250	0.500	0.250	0.000	0.250
2	0.500	0.000	0.000	0.500	0.500	0.000	0.000	0.500	1.000	0.000	0.000	0.000
3	0.200	0.400	0.000	0.400	0.400	0.000	0.400	0.200	0.400	0.000	0.200	0.400
4	0.667	0.000	0.333	0.000	0.667	0.000	0.333	0.000	0.000	0.333	0.667	0.000
边际概率	0.131	0.033	0.078	0.092	0.131	0.042	0.082	0.079	0.158	0.049	0.072	0.054

表 3　χ^2 概率分布表

	一步 χ^2 概率分布表				二步 χ^2 概率分布表				三步 χ^2 概率分布表			
1	1.527	0.000	7.692	2.174	0.000	11.905	3.049	3.165	3.165	5.102	0.000	4.630
2	3.817	0.000	0.000	5.435	3.817	0.000	0.000	6.329	6.329	0.000	0.000	0.000
3	1.527	12.121	0.000	4.348	3.053	0.000	4.878	2.532	2.532	0.000	2.778	7.407
4	5.092	0.000	4.269	0.000	5.092	0.000	4.061	0.000	0.000	6.796	9.264	0.000
合计	11.962	12.121	11.962	11.957	11.962	11.905	11.988	12.025	12.025	11.898	12.042	12.037
总合	48.011				47.880				48.002			

若给定显著性水平 α 为 0.05,已知 $m=4$,查表可得分位点 $\chi^2_{0.05}(3^2)$ 的值为 7.81,此时 $\chi^2>\chi^2_{0.05}(3^2)$ 证明经神经网络输出的 I/O 响应时间很好地符合马氏性,可通过马尔科夫算法改进神经网络的输出值,从而进一步对权值进行调整。

4.3　训练误差分析

使用马尔科夫算法和人工蜂群算法改进 BP 网络后,误差结果如图 1 所示。因前 8 次循环作为马尔科夫预测的基础数据,未进行训练,故误差为 0。从后 42 次循环可以看出,采用马尔科夫修正 BP 网络输出后,误差小于原 BP 网络。

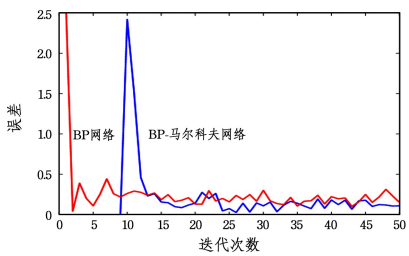

图 3　串行训练误差图

4.4　网络预测

BP 神经网络预测与改进网络的预测值如表 4 所列。

表 4　I/O 预测值对比

序号	BP 网络预测 I/O 时间	改进模型预测 I/O 时间	实测 I/O 时间
1	21.3153	21.3024	21.2567
2	21.3959	21.4244	21.4586
3	21.5142	21.5238	21.5782
4	21.2985	21.3045	21.1273

结束语　本文分析了影响 BP 网络算法性能的两个主要因素,并分别通过马尔科夫链及人工蜂群算法对其进行了改进,通过 Matlab 实现了算法的模拟,进一步比较了不同算法网络训练过程中的误差值和最终的预测值。通过实验得出:改进的算法有效地避免了局部最优值,提高了收敛速度和预测精度。但本文所选取的 ABC 算法为基本算法,未对其进行优化,故在神经网络训练过程中会引入 ABC 算法固有的缺点,如容易陷入局部最优、算法更新时随机性较大等问题,应进一步改进。

参 考 文 献

[1]　崔宝江,刘军,王刚,等. 网络存储系统 I/O 响应时间边界性能研究[J]. 通信学报,2006,27(1):69-74.

[2]　陈琼,郑启伦,凌卫新. 采用计数器存储权值的人工神经网络的实现[J]. 计算机工程与应用,2001,20:22-25.

[3]　KARABOGA D. An idea based on honey bee swarm for numerical optimization[R]. Erciyes University, Kayseri, Turkey, TechnicalReport-TR06, 2005.

[4]　KARABOGA D, AKAY B, OZTURK C. Artificial Bee Colony (ABC) optimization algotithm for training feed-forward neural networks[C]// LNCS: Modeling Decisions for Artificial Intelligence. Springer-Verlag, 2007: 18-329.

[5]　BEATRIZ A G, HUMBERTO S, ROBERTO A. VÁZQUEZ. Artificial neural network synthesis by means of artificial bee colony (ABC) algorithm[C]// 2011 IEEE Congress of Evolutionary Computation (CEC). 2011: 331-338.

[6]　暴励. 人工蜂群算法的混合策略研究[D]. 太原:太原科技大学,2010.

[7]　ZHU G, KWONG S. Gbest-Guided artificial bee colony algorithm for numerical function optimization[J]. Applied mathematics and Computation, 2010, 217(7): 3166-3173.

[8]　周新宇,吴志健,王文明. 基于正交实验设计的人工蜂群算法[J]. 软件学报,2015,26(9):2167-2190.

[9]　冷昕,张树群,雷兆宜. 改进的人工蜂群算法在神经网络中的应用[J]. 计算机工程与应用,2016,52(11):7-10.

[10]　王允霞. 蜂群算法的研究及其在人工神经网络中的应用[D]. 广州:华南理工大学,2013:25-27.

[11]　向万里,马寿峰. 基于轮盘赌反向选择机制的蜂群优化算法[J]. 计算机应用研究,2013(1):86-89.

[12]　魏波,喻飞,徐星,等. 基于改进轮盘赌策略的交互式演化算法[J]. 计算机与数字工程,2014(10):1762-1767.

[13]　ROMANOVSKII V. Discrete Markov's chains[M]. Moscow: Gostexizdat, 1949.

[14]　WHITTAKER J A, THOMASON M G. A Markov Chain Model for Statistical software testing[J]. IEEE Transactions on Software Engineering, 1994, 30(10): 812-824.

[15]　MAREK I, SZYLD D B. Algebraic schwarz methods for the numerical solution of Markov chains[J]. Linar Algebraic and its Applications, 2004, 386: 67-81.

[16]　POGGI P, NOTTON G, MUSELLI M. Stochastic study of hourly total solar radiation in Corsica using a Markov model[J]. International Journal of Climatology, 2000, 20(14): 1843-1860.

[17]　LI Y Z, LUAN R, NIU J C. Forecast of power generation for grid-connected photovoltaic system based on grey model and Markov chain[C]// 3rd IEEE Conference on Industrial Electronics and Applications. Singapore: IEEE, 2008: 1729-1733.

[18]　刁莹. 用数学建模方法评价存储系统性能[D]. 哈尔滨:哈尔滨工程大学,2013:82-88.

[19]　林已杰. 一种基于马尔科夫和神经网络的软件衰退预测方法研究[D]. 重庆:西南大学,2010:31-32.

[20]　KARABOGA D, GORKEMLI B, OZTURK C, Karaboga N. A comprehensive survey: Artificial bee colony (ABC) algorithm and applications[J]. Artificial Intelligence Review, 2014, 42(1): 21-57.