文章编号:1001 - 9081(2006)07 - 1709 - 04

用 P-BP 预测网络模型预测通信网络指标

林 森,李志蜀
(四川大学 计算机学院,四川 成都 610065)
(linsenlsjs@ hotmail. com)

摘 要:结合某通信企业业务数据的特点,为其通信网络数据预测业务建立了一套通用的 P-BP 预测网络模型。它以时间序列分析为建模依据和指导,并改变 BP 神经网络的学习方法,提出 BP-L 网络用作模型中挖掘数据依赖性的工具,它的预测精度、运算速度、泛化能力明显高于 BP 网络。此外,P-BP 模型能依据历史数据自动计算最合适的预测阶数;根据业务数据特点设计的消除非平稳因素的方法,使其在平稳化的同时能很好地提高并行运算性能;用区间估计过滤异常数据,具有较强的抗干扰能力,能适应实际的工作环境。用业务数据测试该模型,得到了快速的、非常精确的预测效果和完备的预测值置信区间。

关键词:神经网络;时间序列分析;机器学习
中图分类号: TP183 **文献标识码**:A

Data prediction in communication network using P-BP predicting network

LIN Sen, LI Zhi-shu
(*College of Computer Science, Sichuan University, Chengdu Sichuan* 610065, *China*)

Abstract: A data predicting model in communication network field called P-BP was proposed. It was guided by time series analysis theory. BP-L neural network modified from BP network was proposed to improve its generalization ability, precision and speed in data predicting applications. BP-L was used as machine learning method in P-BP model. P-BP model could automatically leach abnormal data and get a most proper exponent. This model was tested for data predicting in CNET field and got a good result.

Key words: neural network; time series analysis; machine learning

在通信企业,通信网络数据的采集、维护、分析、预测是日常而十分重要的工作,其话务量、负荷等指标数据的预测对于企业经营和维护具有十分重要的指导意义。

在数据预测领域较早提出的有时间序列分析理论,它对时间序列的本质、特性等做了广泛而深入的研究,构建了一套全面的理论,并提出了以自回归滑动平均(Autoregressive Moving Average, ARMA)模型等为核心的具体方法。近年来机器学习理论不断发展,神经网络、支持向量机等也被应用于数据预测。机器学习方法具有对任意未知非线性函数关系进行学习的能力。比较而言,时间序列分析方法理论完善但手段较为简单,它指出了时间序列为什么可以预测、预测的原则、理论上涉及建模、参数估计、处理非平稳因素等各个方面,可以作为建立预测模型的依据和指导;机器学习方法学习能力强,却并非为数据预测而提出,直接用于数据预测具有盲目性,不一定能收到较好的效果,因此需要进行修改并在时间序列理论指导下作为挖掘工具使用。

我们为某通信企业的网络数据预测业务建立了一套 P-BP 预测网络模型,用于各种通信网络指标预测。它以时间序列理论指导建模、预测的全过程;用 BP-L 神经网络完成时间序列分析中挖掘数据依赖关系的步骤;能自动实时计算最合适的预测阶数;根据业务数据特点消除非平稳因素并提高并行运算能力,以统计方法过滤数据,具备抗干扰性,能应用于实际工作环境。和神经网络相比,支持向量机具有泛化能力较强、不会陷入局部最优解、能自行确定中间层节点数等优点;但考虑到支持向量机仅有的三层结构对我们的业务略显简单,而神经网络的应用已比较成熟,并有遗传算法等手段对其进行优化,因此,我们在 BP 神经网络的基础上,对它做了若干改进得到 BP-L 神经网络,使其更适于数据预测而不是模式识别;以它作为我们模型中使用的机器学习工具,使预测的速度、精确性和泛化能力都得到较大改善。

1 时间序列分析和遗传神经网络

1.1 时间序列分析理论和 ARMA 模型

时间序列分析理论认为,最基本的序列是平稳的,每一个数据组的概率结构不会逐点变化,并不依赖于时间起始点;观测值之间存在依赖性。一旦由历史数据计算出这种“依赖”关系,就可以通过系统的过去预测未来。这种依赖关系用 $ARMA(p,q)$ 模型来进行描述和求解:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} + a_t \quad (1)$$

该模型表示,未来的数据依赖于过去的连续 p 个时刻的观测值,以及连续 q 个时刻的干扰。p 为对过去观测值的依赖阶数,q 为对过去干扰的依赖阶数,a_t 为当前的干扰即残差。残差构成的序列称为残差序列。

ARMA 模型是非线性的。时间序列分析理论指出,在高阶的情况下,可以用线性的自回归(Auto-Regressive, AR)模

收稿日期:2006 - 01 - 13

作者简介:林森(1981 -),男,四川广汉人,硕士研究生,主要研究方向:计算机网络与信息系统、知识发现; 李志蜀(1946 -),男,重庆人,教授,博士生导师,主要研究方向:计算机网络与信息系统、自动控制.

型来逼近 ARMA 模型,通过拟合阶数渐增的 AR 模型,当残差平方和减小到无意义时停止,仍然可以得到要求的预测精度:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t \quad (2)$$

上述模型对平稳的时间序列适用。而实际的数据总含有周期性、趋势性等非平稳因素,在分析时需要先消除周期性、趋势性或修改模型使之含有周期项、趋势项,再与其他参数一起进行估算。

1.2 人工神经网络和遗传算法

人工神经网络作为一种广泛使用的机器学习方法,不需要设计任何数学模型,只靠过去的经验来学习、拟合任意非线性函数关系,广泛用于模式识别等领域。BP 模型是应用较广的一种神经网络,其结构如图 1 所示,是带权的前向网络,每个节点模拟一个神经元,按照神经元函数对输入信号进行处理,给出输出。输入层各节点提供输入数据,输出层各节点映射出函数值。隐层层数及层内节点数不定,决定了网络结构。

BP 神经网络的工作过程分为前向输出和误差反向传播两种基本步骤。前向输出是从输入层开始,每个节点对输入值以神经元函数计算,得到输出;除输入层节点外,各节点把上层相邻节点的输出值乘以边上的权再求和,作为自身的输入。输出层节点提供的输出值即为网络的输出。误差反向传播过程为,以输出值与训练样本对比,以误差为各边权值的多元复合函数,以求函数最小值的方式逐层反向调整各边的权值。

BP 神经网络的学习过程是,提供一组包括输入以及对应输出的训练样本,以样本反复进行前向输出与反向权值调整,直至输出数据达到要求的精度,得到能拟合该映射关系的 BP 网络。

神经网络隐层层数与节点数决定了网络结构,结构简单的网络无法精确拟合复杂的模型;过于复杂的网络又会造成过学习等问题,性能脆弱,没有推广性。不恰当的初始权值易使网络陷入局部最优解而不是全局最优解。对此可以采用遗传算法进行优化。遗传算法是模拟种群进化过程的全局性概率搜索算法,它以某种编码方案表示网络结构与初始权值,给出初始种群,模拟进化过程,对种群中的个体一代一代地执行选择、交叉、变异等操作,使得个体对问题的适应度不断提高。

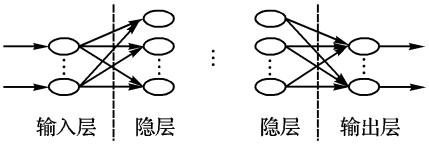

图 1 BP 神经网络的结构

2 P-BP 预测网络模型

如上所述,我们建立一套模型来为某通信企业实现通信网络数据预测业务,适用于各种通信网络指标。该模型以时间序列分析理论为总体指导,在每步预测时实时计算最合适的预测阶数,能进行数据过滤以识别、修正历史序列中的异常点,能根据业务数据特点在消除非平稳因素时很好地改善并行性能;用适于预测运算的 BP-L 神经网络替换(1),(2)式来挖掘数据间的依赖关系,并以遗传算法为其确定较为合适的网络结构。预测结果由精确预测值与预测置信区间构成,称为预报。这样最终得到的模型,我们将其称为 P-BP 预测网络模型:

$$(\hat{x}_{a+p}, \hat{x}_{a+p+1}, \cdots) = (\text{P-BP})(x_a, x_{a+1}, \cdots, x_{a+p-1}) \quad (3)$$

2.1 适于预测的 BP-L 神经网络

从理论上说,机器学习方法样本数越多越好,而在实际应用中样本数不可能是无限的,考虑运行效率等因素,不可能提供过多的样本,因此需要考虑其泛化能力,即由样本建立的模型是否具有好的推广性。而且,数据预测不等于模式识别。因此我们改进 BP 网络的学习方法使其更适于预测工作,称为 BP-L 神经网络。

一方面,我们的目的是进行数据预测,这里神经网络用于拟合数据间的依赖性。与模式识别不同,数据点间可能会有、但不是一定会有高精度的函数关系。这是非常重要的,如果一定要达到指定的精度,可能带来完全无法接受的时间开销,甚至会陷入死循环;而且以样本拟合出总体并不存在的高精度函数关系,预测结果必定不准。因此,模型每次预测都应动态寻找最合适的精度要求,不能高也不能低,避免上述问题出现的同时应很好地反映数据间的依赖关系。在 BP-L 神经网络中我们采用动态调节误差阈值的方法来自动寻找这一最合适的精确度:

1) 在训练初期设置一个较高的精度要求 σ。训练时对训练轮数计数。当训练次数达到一个较大的值而仍然不能达到要求的精度时,扩大 σ。

2) 模型中为提高速度,在修正权值时给予相对较大的初始步长 $\eta = \eta_0$,在权值修正发生错误时降低步长 η 并恢复刚才的权值。在这里,若 η 已低于指定的标准,则放弃精度要求,放大 σ,并令 $\eta = \eta_0$,恢复原始步长。

另一方面,对一批训练样本,BP 网络一般以误差平方和作为总体误差。而实际上样本是有限的,且邻近的数据总是存在依赖关系,一个数据点可能持续对后面的数据点产生影响;数据特征具有局部性。这样,对邻近时间段的样本,其训练精度应该提出更高的要求,反之,相隔较久的样本则放宽精度。在 BP-L 神经网络中,一轮前向输出后,对各样本的输出误差改为使用加权平均方法统计。近期的外部干扰其影响仍然可能延续,权重更大;越远的数据对总体误差的贡献越小,久远的临时性干扰对预测的影响就被削弱,还可以补偿因近期样本精确拟合带来的开销以保证速度。

最后,如前所述,过于简单和过于复杂的网络结构都会影响学习或泛化能力。遗传算法可以进行优化,但需要耗费大量时间,不能实时使用。对此我们进行折中,先采用遗传算法等手段进行试验,确定一个对于实际业务,适应度与时间性能都能达到要求的网络结构,在模型中使用。在后面实测时网络结构定为 7 层,节点神经元使用 Sigmoid 函数,由于该函数值域为(0,1),因此计算前以数据最大值的 1.2 倍为标准对数据进行归一。

2.2 实时计算模型阶数

久远的历史数据对预测点没有影响,做一步预测时,用于预测一个数据点的历史数据有多少个最为合适,称为模型的阶数。若只凭反复试验或经验选择阶数,不精确、不科学、不通用,需要有一种每步预测时实时确定阶数的方法。

我们的 P-BP 预测模型把序列处理后得到一组平稳序列,作为基本的预测对象。残差是否独立是模型的阶数是否合适的一个重要判据。平稳的序列由存在依赖关系的数据组以及随机的残差构成,一个合适的模型,其残差应相互独立。对此可以使用 Bartlett 提出的估计自相关的标准(误)差的公式:

$$\sigma(\hat{\rho}_k) = N^{-1/2}\left[1 + 2(\hat{\rho}_1^2 + \hat{\rho}_2^2 + \cdots + \hat{\rho}_p^2)\right]^{1/2} \approx 1/\sqrt{N} \quad (4)$$

以 $|\hat{\rho}_k| < 1.96/\sqrt{N}$ 作为 5% 水平上残差自相关是否足够小的快速检验标准[4]。此外,Box 和 Jenkins 提出一种多用途的检验,即对于统计量:$Q = N\sum_{k=1}^{K}\hat{\rho}_k^2$,应在相应的概率水平上小于 $\chi^2(K - n - m)$。

在初始指定一个较低的阶数时,随着阶数的增加,残差平方和可能会得到改善。在残差不断改善的情况下,可使用具有 s 和 $N-r$ 个自由度的 F-检验来判断阶数的增加是否必要,这样可减少计算阶数时尝试的次数,提高时间性能:

$$F = \frac{A_1 - A_0}{s} \Big/ \frac{A_0}{N-r} \sim F(s, N-r) \tag{5}$$

用它来检查随着阶数的变化,残差值是不是显著地减小了。其中 A_1 和 A_0 分别是 n 阶和 n' 阶模型的残差平方和,$n' > n$,s 为模型参数之差。用它来检验假设:n' 阶模型的 r 个参数中有 s 个为0。从2阶开始建立阶数递增的模型,每步对序列建模,分离出残差序列;若 n 阶时残差在5% 的水平上近似独立,并且在阶数增加到 n' 后,使用 F- 检验证明在5% 的显著性水平上,阶数的增加并未显著改善残差,此时就认为阶数 n 是合适的,以它作为本次预测使用的阶数。

2.3　非平稳因素的消除与改善并行运算能力

在 ARMA 模型中,时间序列必须是平稳的,因为它假定每个数据组在围绕一个标准起伏,依赖关系不随时间起点变化,即均值与协方差与时间起点无关。而数据周期性、长期趋势等非平稳因素会破坏这一条件。对通信网络数据业务,预测指标的时间单位为小时或日。分析业务的特点,对有限样本进行中、短期预测,序列本身基本没有固定的趋势;长期的总体性趋势在序列中的微弱反应则可放在网络中学习。因此主要是考虑去除周期因素。

时间序列分析方法中把周期因素加入回归方程一起求解。但对我们的业务数据,其序列均遵循弱周期性,它们大致具有以日或小时为周期的特点,但各周期相比较,曲线的趋势走向并没有多大的相似性,这一特点从图 2 即可看出,用强周期方式处理效果就不好。

因此我们采用以下方法来消除周期因素,用它既可去除周期性,还可提高模型的并行运算能力。即把周期为 n 的序列分为子序列 $\{x_1, x_{n+1}, \cdots\}, \{x_2, x_{n+2}, \cdots\}, \cdots, \{x_n, x_{2n}, \cdots\}$,对各子序列分别预测。这样的子序列,既保留了数据依赖性,又消除了周期性。而且,由于每个子序列是一个独立的处理对象,独立提供历史数据,独立完成预测,这样的设计就提高了模型的并行运算能力,在多 CPU 的环境下能很好地发挥机器性能,提高求解速度。

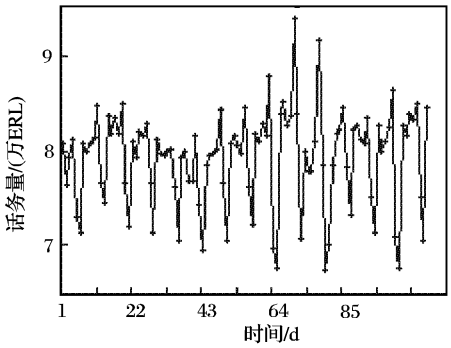

图2　通信网络数据示例——话务量指标

2.4　异常采样值过滤

业务数据不可避免会出现异常采样值,异常值的存在一定程度上会影响预测的准确性。在模型中我们使用区间估计识别异常点。以日数据序列 $(x_1, x_2, \cdots)$ 为例,其周期为7,将其分为 $(x_1, x_8, \cdots), (x_2, x_9, \cdots), \cdots, (x_7, x_{14}, \cdots)$ 7个子序列。对子序列,在置信水平 $1-\alpha$ 的条件下,大样本时用正态分布,小样本时样本用自由度为 $n-1$ 的 t 分布求得边际误差,分别为 $z_{a/2}\frac{s}{\sqrt{n}}$ 和 $t_{a/2}\frac{s}{\sqrt{n}}$。因此,总体均值的置信区间为:

$$\begin{cases} \hat{x} \pm z_{a/2}\dfrac{s}{\sqrt{n}}, & n \geqslant 30 \\ \hat{x} \pm t_{a/2}\dfrac{s}{\sqrt{n}}, & z < 30 \end{cases}$$

$z_{a/2}$, $t_{a/2}$ 分别表示标准正态分布与 t 分布右侧面积为 $a/2$ 时的 z 和 t 值。为具有较强的容错性,我们以各子序列上限的最大值作水平阈线 Max,下限的最小值作水平阈线 Min。阈线之间以及阈线邻近的点视为正常点,其余视为异常点。对于图 2 所示的日话务量数据,其 95% 置信区间为(69 504.53 ERL,87 294.404 ERL)。如图 3 所示,6 个主要的异常点都被阈线正确隔离。

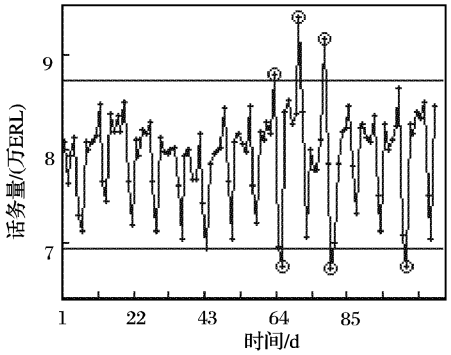

图3　异常点的识别

对于异常点,可用历史数据对其进行预测,以预测值修正。从后面的实测可以看到,本文的预测方法有较强的抗干扰能力,预测精度很高,这样做是可行的。

2.5　预测和预报

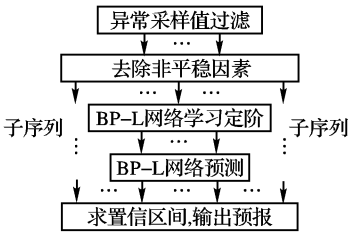

图4　P-BP 预测网络观测步骤

P-BP 预测网络的预测步骤可以用图 4 说明。历史数据经过滤、消除非平稳因素之后,分为若干平稳的子序列,这些子序列是预测的基本单元。这时对子序列使用 BP-L 网络学习数据间的依赖关系。如前所述,这是一个反复的过程。对第 i 个子序列 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$,当指定在阶数 p 下做学习时,训练样本为:

$$(x_{i1}, x_{i2}, \cdots, x_{ip}) \rightarrow x_{i(p+1)}$$
$$(x_{i2}, x_{i3}, \cdots, x_{i(p+1)}) \rightarrow x_{i(p+2)}$$
$$\cdots$$
$$(x_{i(m-p)}, x_{i(m-p+1)}, \cdots, x_{i(m-1)}) \rightarrow x_{im}$$

创建输入层节点数为 p 的 BP-L 神经网络,以上述 $m-p$ 组样本训练,拟合出数据间的依赖关系;让 p 从2开始递增,按照 2.2 节论述的方法找到最合适的阶数 $p = t$,即完成了第3步的学习、定阶过程。于是进入第4步,创建一个输入层节点数为 t 的 BP-L 神经网络完成子序列 X_i 的预测任务:

$$(x_{i(m-t+1)}, x_{i(m-t+2)}, \cdots, x_{im}) \rightarrow \hat{x}_{i(m+1)}$$
$$(x_{i(m-t+2)}, x_{i(m-t+3)}, \cdots, \hat{x}_{i(m+1)}) \rightarrow \hat{x}_{i(m+2)}$$
$$\cdots$$

每次对各子序列单独预测下一个数据点,若原始序列周期为 n,则 P-BP 模型一步并行预测可得到 n 个预测点,n 个点

并入一个序列即恢复了原来的非平稳性。把预测点加入原始序列作为历史数据使用,即可进一步预测后面的数据点。反复执行该过程即可完成一定时间长度的中、短期预测。

一个数据点的预报由两个要素构成:预测值和预测值置信区间。预测值总会存在一定的误差,预测值置信区间则给出预测值可能的取值范围。预测误差受随机因素的影响,可以用正态分布近似描述其分布规律。由此可得出预测值的95%置信区间:$(\hat{y}-1.96\times s/\sqrt{n},\ \hat{y}+1.96\times s/\sqrt{n})$,其中$\hat{y}$为预测值,$s$为残差序列的样本方差。

3 实际测试与分析

实测数据使用图2所示的,某通信企业下辖某市某区某网元的日话务量指标,数据采集自2004年11月19日起;使用前11个周期共77个真实点作为历史数据,连续预测4个周期共28点,与真实点进行对比分析。真实值与预测值数量比为11:4。预测分4次完成,每次并行预测7个点。预测中,第77个点以后的真实数据是不可见的,比如,预测第13周期的数据时,使用第12周期各点的预测值而不是真实值来作为历史数据。同时,不进行异常点处理,保留真实数据以验证抗干扰能力。

以(预测值-真实值)/预测值计算预测误差,如表1所示;并把预测结果汇总到Excel表中,与原始数据用折线图进行对比,如图5所示。

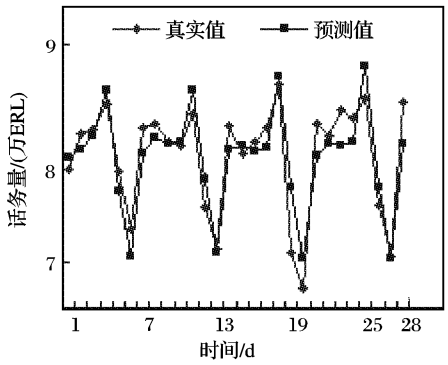

图5 预测值/实际值对比

由表1统计,预测误差绝对值低于1%有7个点,占25%;低于2%有13个点,占46.4%;低于4%有26个点,占93%;误差取绝对值,最小为0.191%,平均仅为2.3%,预测值是很准确的。真实数据均落入了预测值的95%置信区间内,如图6所示。

如图2所表明的,历史数据点总体呈现较大的无规律波动,并且历史数据最后3个周期出现明显的数据异常,预测时间段紧随其后。为验证模型对异常点的处理能力,模型没有进行异常剔除,保留真实数据进行预测。结果表明,预测值没有多大误差,这说明该预测方法具有较强的抗干扰能力。

表1 预测值误差分析

周期	预测值/ERL	预测误差	周期	预测值/ERL	预测误差
1	79593.00897	0.013996	1	85779.62178	0.025818
2	80331.76377	-0.019100	2	77554.31168	0.032153
3	81615.32560	-0.007110	3	70958.97879	-0.003950
4	85736.47395	0.013004	4	80453.03483	-0.027070
5	76457.75544	-0.023880	5	80774.55420	0.010344
6	70472.62512	-0.038210	6	80218.31171	-0.010270
7	80055.96888	-0.027970	7	80507.41675	-0.023330
1	81384.41753	-0.017160	1	87162.89755	0.008408
2	80982.35157	-0.002360	2	76854.32296	0.077545
3	81101.27846	0.004037	3	70357.61098	0.039677
4	79807.94395	-0.036390	4	80989.38923	-0.009030
5	80674.06953	-0.040100	5	81093.59298	-0.026760
6	87977.27932	0.032968	6	76919.28681	0.022580
7	70369.11100	-0.001910	7	80962.68265	-0.045980

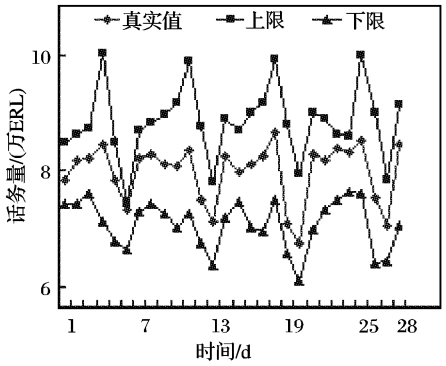

图6 真实值与预测值的95%置信区间

参考文献:

[1] BURGES JC. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998,2(2): 121-167.

[2] PLATT JC. Sequential minimal optimization: a fast algorithm for training support vector machines[A]. Advances in Kernel Methods-Support Vector Learning[C]. MIT Press, 1999.185-208.

[3] VAPNIK V. Statistical Learning Theory[M]. John Wiley and Sons Inc, 1998.

[4] (美)PANDIT SM,吴宪民. 时间序列及系统分析与应用[M]. 李昌琪,荣国俊,译. 北京:机械工业出版社, 1988.

[5] 张波. 人工神经网络在流量预测中的运用[J]. 水道港口, 2005, 26(2): 80-82.

[6] 柳进,于继来,唐降龙. 基于数据挖掘的电网高峰负荷预测系统[J]. 计算机工程, 2005, 31(1): 10-11.

[7] 邹柏贤 刘强. 基于ARMA模型的网络流量预测[J]. 计算机研究与发展, 2002, 39(12): 1645-1651.

(上接第1708页)

3 结语

本文在对传统遗传算法分析的基础上,利用JADE进行了分布式遗传算法的研究,提出了一种基于多Agent协同的并行遗传算法结构。该结构由若干计算单元组成,每个计算单元实际上就是一个运行简单遗传算法的独立的计算Agent。通过试验可以明显感到,分布式运行效率较高,通过主从式运行机制,简化了各部分之间的通讯环节,提高了运行效率,加快了实现速度。

参考文献:

[1] 江瑞, 罗予频, 胡东成, 等. 一种基于多Agent协同的准并行遗传算法[J]. 电子学报, 2002, 30(10), 1490-1491.

[2] BELLIFEMINE F, CAIRE G, TRUCCO T, *et al*. JADE ADMINISTRATOR'S GUIDE [EB/OL]. http://jade.cselt.it/, 2005-03.

[3] BELLIFEMINE F, CAIRE G, TRUCCO T, *et al*. JADE PROGRAMMER'S GUIDE[EB/OL]. http://jade.cselt.it/, 2005-03.

[4] CAIRE G. JADE PROGRAMMING FOR BEGINNERS[EB/OL]. http://jade.cselt.it/, 2003-11.

[5] 王小平,曹立明. 遗传算法——理论、应用与软件实现[M]. 西安:西安交通大学出版社, 2002.18-50.