doi: 10.12068/j.issn.1005-3026.2019.02.006

基于改进 GA-BP 的移动通信用户流失预测算法

于瑞云, 薛 林, 安轩邈, 夏兴有
(东北大学 软件学院, 辽宁 沈阳 110169)

摘 要: BP 神经网络(BPNN)模型对移动通信用户流失的预测有较好的效果,但其全局搜索能力相对较弱,对初始网络权重非常敏感,因此本文通过对用户通信行为的分析,提出一种基于改进 GA-BP 的移动用户流失预测算法:用改进的遗传算法对 BPNN 的权值和阈值进行初始化,从而提高预测模型的准确率. 改进的遗传算法采用一种自适应的交叉概率和变异概率计算策略,提高了遗传算法寻找全局最优解的能力. 通过对比实验发现,本文构建的移动用户流失预测模型,在预测准确率上有着很好的表现.

关 键 词: 移动通信;行为分析;用户流失;BP 神经网络;遗传算法
中图分类号: TP 18 **文献标志码**: A **文章编号**: 1005-3026(2019)02-0180-06

Mobile Communications Customer Churn Prediction Algorithm Based on Improved GA-BP Network

YU Rui-yun, *XUE Lin*, *AN Xuan-miao*, *XIA Xing-you*
(School of Software, Northeastern University, Shenyang 110169, China. Corresponding author: XUE Lin, E-mail: 1601688@stu.neu.edu.cn)

Abstract: A customer churn prediction model based on BP neural network (BPNN) has achieved well enough results. However, it has relatively weak global search ability and is very sensitive to the initial network weights. A prediction algorithm based on improved genetic algorithm (IGA) and BPNN (IGA-BP) is proposed by analyzing users' communication behavior, where the weights and thresholds of BPNN are initialized with IGA, thus improving the accuracy of the prediction model. The improved algorithm adopts a self-adapting probability of crossover and mutation, which enhances the global optimum search ability of GA. The proposed IGA-BP model has obvious improvement on customer churn prediction, compared with existing algorithms.

Key words: mobile communication; behavior analysis; customer churn; BP neural network; genetic algorithm

面对日益激烈的市场竞争以及多变的用户需求,如何最大程度地降低用户流失率成为电信运营商首要关注的问题. 当今解决这一问题的主流方式是运用数据挖掘技术对移动用户的行为进行分析[1-5],建立用户流失预测模型,对用户流失情况进行预测,并对可能流失的用户采取相应的挽留措施.

用户流失预测问题本质上是对用户分类的问题:流失和未流失. 在用户流失预测领域,主要的分类方法有逻辑回归[6]、贝叶斯网络[7]、决策树[8]、支持向量机[9],以及神经网络等[10-12]. BP 神经网络(BP neural network, BPNN)是目前应用最为广泛的神经网络模型,并具有完备的理论体系.

BPNN 作为一种样本数据有标签情况下的预测和分类模型,可以被应用于解决用户流失的预测问题,但其存在全局搜索能力相对较弱,对初始网络权重非常敏感等问题. 为了提高 BPNN 的全局搜索能力,本文提出用一种改进的遗传算法(improved genetic algorithm, IGA)来初始化

收稿日期: 2017-11-15
基金项目: 国家自然科学基金资助项目(61672148,61502092); 教育部-中国移动科研基金资助项目(MCM20160201); 辽宁省百千万人才工程项目(201514).
作者简介: 于瑞云(1974-),男,辽宁丹东人,东北大学教授.

BPNN 的权值和阈值，使其更靠近全局最优，从而提高算法的准确率. IGA 提出了一种自适应的交叉概率和变异概率计算策略，提高了 GA 寻找全局最优解的能力. 基于 IGA 优化的 BPNN 预测模型的预测准确率有了很大提高.

1　算法设计

1.1　遗传算法的改进

标准 GA 在复杂优化问题及多峰值的函数优化求解过程中存在收敛速度慢、容易陷入局部最优的问题，本文针对移动用户话单数据的特点以及要解决的问题，基于标准 GA，提出一种拥有更强全搜索能力的改进 GA.

1.1.1　染色体编码的设计

染色体编码是 GA 优化 BPNN 的关键. 具体的编码过程以一个有具体结构的 BPNN 为例，图 1 是一个 3－3－3 结构的 BP 网络，其中包含了输入层、输出层以及单隐层.

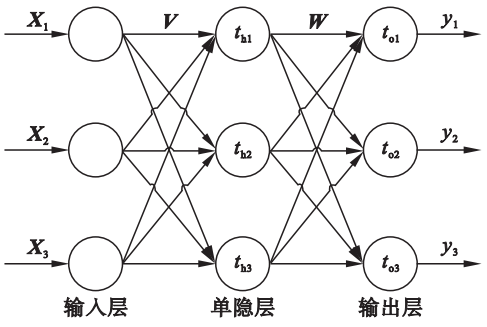

图 1　3－3－3 BPNN 结构图

Fig. 1　Structure of a 3－3－3 BPNN

染色体基因向量的分量是由该网络模型中所有权重和阈值构成的，起初基因向量的各个分量是随机分配的值，这些值的范围应该是在要解决问题的值域内. 染色体的基因向量是由该 BPNN 的 4 个参数矩阵编码构成的，这些矩阵具体定义如下：

$$\boldsymbol{V}=\begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}, \tag{1}$$

$$\boldsymbol{W}=\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}, \tag{2}$$

$$\boldsymbol{t}_{\mathrm{h}}=[\,t_{\mathrm{h1}} \quad t_{\mathrm{h2}} \quad t_{\mathrm{h3}}\,], \tag{3}$$

$$\boldsymbol{t}_{\mathrm{o}}=[\,t_{\mathrm{o1}} \quad t_{\mathrm{o2}} \quad t_{\mathrm{o3}}\,]. \tag{4}$$

式中：$\boldsymbol{V}$ 和 $\boldsymbol{W}$ 是两个权值矩阵；$\boldsymbol{t}_{\mathrm{h}}$ 和 $\boldsymbol{t}_{\mathrm{o}}$ 分别是隐含层和输出层的阈值矩阵.

这 4 个矩阵编码生成染色体基因向量的公式如下，其中 $\boldsymbol{X}_i$ 是第 i 个染色体的基因向量：

$$\boldsymbol{X}_i=[\,v_{11},\cdots,v_{33},w_{11},\cdots,w_{33},t_{\mathrm{h1}},t_{\mathrm{h2}},t_{\mathrm{h3}},t_{\mathrm{o1}},t_{\mathrm{o2}},t_{\mathrm{o3}}\,]. \tag{5}$$

本文采用浮点数编码方案，染色体向量的维度由 BPNN 中权值和阈值的数量决定，而权值和阈值的数量由 BPNN 的结构决定.

1.1.2　适应度函数的确定

BPNN 的均方误差 E 是评价 BPNN 预测模型的重要指标. E 值越小，表示 BPNN 的预测性能越好. 在 GA 中，个体的适应度值是评价个体表现的重要指标，假设第 i 个个体的适应度值为 F_i，其对应的 BPNN 的均方误差为 $E(\boldsymbol{X}_i)$，则个体的适应度函数为

$$F_i=E(\boldsymbol{X}_i). \tag{6}$$

从上述公式可以发现，个体的表现越好，其适应度值就会越小；使种群中最优个体适应度值接近 0 是 GA 不断进化的目标.

1.1.3　选择操作的设计

选择操作是从父代中选出优秀的个体直接进入到子代. 不同于交叉操作和变异操作，选择操作并不会破坏染色体中的基因. 在本文中，染色体的适应度值是 BPNN 的输出误差，个体表现越好，其适应度值的绝对值就应该越接近于 0，因此标准的轮盘赌法并不适用. 个体被选中概率的计算公式为

$$p_i=\frac{l}{F_i}\Big/\sum_{i=1}^{N}\frac{l}{F_i},\quad (i=1,2,\cdots,N). \tag{7}$$

式中：F_i 是第 i 个个体的适应度值；l 是调节因子，用以控制适应度反比值的量纲，保证计算的合理性. 从式(7)中可以发现，个体被选中的概率与其适应度值成负相关. 在实际计算过程中，样本数量比较大，不存在个体适应度为 0 的情况，因此不需要考虑适应度为 0 导致的计算错误.

1.1.4　交叉操作的设计

在实际情况中发现，标准 GA 中基因按照固定概率进行交叉操作的策略并不可取. 因此，本文提出一种自适应的交叉概率，主要想法是根据个体适应度值及迭代次数，对个体的交叉概率进行自适应调整：对于表现较好(适应度值较低)的个体，适当降低交叉概率，避免破坏优良的基因；如果个体表现较差，则增大该个体的交叉概率，让其更多地进行交叉操作，对其进行优化，破坏其表现一般的基因结构. 另外，为了保证算法的收敛性、前期的种群多样性及后期的局部搜索能力，这种交叉概率也应随着算法的迭代不断减小. 在算法

迭代的前期,应赋予一个较大的交叉概率,保证种群的多样性,加快算法的搜索速度;在算法迭代的后期,种群的个体表现趋于稳定,大部分的优秀基因已经被确定,为了优秀基因的延续,应适当降低交叉概率,确保算法在极值点处不会出现震荡,保证算法的收敛性.

基于上述考虑,个体交叉概率的设定如下:

$$p_{ci}=p_{\mathrm{cmax}}-(p_{\mathrm{cmax}}-p_{\mathrm{cmin}})\cdot\frac{t\cdot F_{\min}}{T\cdot(F_i+F_{\min})},\quad(i=1,2,\cdots,N)\ .\tag{8}$$

式中:t 是算法当前的迭代次数;T 是算法的总迭代次数;p_{ci}是第 i 个个体在第 t 次进行交叉时的概率;F_i 是个体的适应度值;$F_{\min}$是种群当前表现最好的个体的适应度值;p_{cmax}是最大交叉概率,取值为 0.6;p_{cmin}是最小交叉概率,取值为 0.3. 从式(8)中可以发现,个体的交叉概率会随着适应度值的降低而降低,以保证优秀的基因有更大的概率延续;个体的交叉概率也会随着迭代次数的增加而降低,以保证算法的稳定性.

1.1.5 变异操作的设计

变异操作的目的是在算法迭代前期保证算法的全局搜索能力,在算法迭代后期保证算法的局部搜索能力和稳定性,因此,本文个体进行变异操作的概率和进行交叉操作的概率在设计上是相同的,都是根据个体的适应度值及算法的迭代次数来决定概率的大小,具体计算公式如下:

$$p_{\mathrm{m}i}=p_{\mathrm{mmax}}-(p_{\mathrm{mmax}}-p_{\mathrm{mmin}})\cdot\frac{t\cdot F_{\min}}{T\cdot(F_i+F_{\min})},\quad(i=1,2,\cdots,N)\ .\tag{9}$$

式中变量 $t,T,F_i,F_{\min}$ 都和式(8)相同;$p_{\mathrm{m}i}$是第 i 个个体在第 t 次迭代时基因发生变异的概率;p_{mmax}是最大变异概率,取值 0.005;p_{mmin}是最小变异概率,取值 0.001.

基于 BPNN 权值和阈值的特点和取值范围,本文采用在基因取值范围内随机取值的变异操作,当个体染色体的某个基因值要发生变异时,在[-1,1]之间随机选取一个实数值来替换发生变异的基因位. 这种基于随机取值的小概率变异操作有时会给算法的全局搜索能力带来很大的提高.

1.1.6 种群规模和迭代次数的确定

种群规模的设计和实际问题的复杂度有关. 为了保证算法的效率,一般情况下 30 个染色体就可以满足大部分需求,但由于本文用 GA 对 BPNN 的权值和阈值进行初始化,待定变量比较多,所以本文选取的种群规模为 100,以保证复杂问题中解的全局最优性.

迭代次数往往要设定得多一些,以保证算法收敛. 根据实验观察,算法迭代到 30 次之后趋于平稳,为了保证算法的完全收敛,本文设定 GA 迭代次数为 100 次.

1.2 遗传算法优化 BPNN 的基本过程

首先根据问题的复杂度确定 GA 的种群规模、迭代次数及部分参数,然后确定种群个体的染色体编码规则,以后的步骤如下:

1) 将个体的基因转化成 BPNN 的权值和阈值,通过 BPNN 的前馈传播计算种群个体的适应度值.

2) 根据适应度值对个体进行选择操作,根据适应度值和迭代次数对个体进行交叉和变异操作,并将保留下来的个体作为下一代.

3) 根据终止条件判断种群是否完成进化,如未完成则回到步骤 1).

IGA-BP 算法的伪代码如下:

```
1: for each chromosome do
2:    Randomly initialize chromosome vector X_i
3: end for
4: while maximum iterations or the fitness of any individual is attained do
5: Decode the chromosome vector X_i into matrices (1),(2),(3), and(4), and train the BP neural network using data samples
6: Update the fitness value of X_i by Eq. (6)
7: for  i = 1 to M  do        //M 是初始种群的个数
8:    Select operation by Eq. (7)
9: end for
10: for  i = 1 to M/2  do
11:    Crossover operation by Eq. (8)
12: end for
13: for  i = 1 to M  do
14:    Mutation operation by Eq. (9)
15: end for
16: end while
```

2 仿真实验与分析

2.1 数据样本

以辽宁移动公司的真实通话记录作为数据样本. 每条数据有上百个属性,但由于大部分的属性和用户流失的相关性很低,所以没必要将所有属性都作为样本属性. 采用关联性分析算法提取样

本属性，最终选取入网时间、基本费总和、通话时长、通话次数、长途费总和、呼叫类型比例，以及掉话比率这 6 个与用户流失相关性最高的属性作为 BPNN 的输入属性，如表 1 所示.

表 1　仿真实验的用户属性

Table 1　Customer features of simulation

字段名称	类型及长度	说明(按月)
入网时间	NUMBER(8)	用户电话号注册时间
基本费总和	NUMBER(8)	用户基本费用总和
通话时长	NUMBER(8)	用户接打通话总时长
通话次数	NUMBER(8)	用户接打通话总次数
呼叫类型比例	DOUBLE(16)	打电话次数/接电话次数
掉话比率	DOUBLE(16)	用户掉话次数/总通话次数

2.2　BPNN 结构的设计

1）隐层神经元数量的确定.

一个拥有 S 型传递函数的单隐层和线性输出层的 BPNN 可以近似模拟任何函数. 考虑到网络结构的复杂度和整个网络的训练时间，在设计 BPNN 的结构时选择单隐层.

网络的性能受隐层神经元数量的影响. 在单隐层的前提下，可以通过增加隐层神经元的数量来适当提高网络的预测准确度. 采用 cut-and-try 方法来确定隐层神经元的数量. 起初设置较少的数量，训练网络并记录网络的预测准确度，然后再逐渐增加隐层神经元的数量. 用同样的样本数据进行训练，能使网络输出误差最小的隐层神经元数量就是最终要确定的.

隐层神经元的数量范围可按下式计算：

$$l = \sqrt{n+m} + a\ . \tag{10}$$

式中：l，n 和 m 分别是隐层、输入层和输出层的神经元数量；a 是调节因子，通常取 1 到 10.

隐层神经元数量与网络性能的关系通过对比实验得到(见表 2)，在确定隐层神经元数量时，BPNN 的训练函数和传递函数都是统一的. 从表 2 中能发现，当隐层神经元数量为 8 时，IGA－BP 算法的网络误差最小，即网络预测准确度最好.

表 2　隐层神经元数量对误差的影响

Table 2　Effect of the number of hidden-layer neurons on error

隐层神经元数量	网络的输出误差	隐层神经元数量	网络的输出误差
4	0.112 113	9	0.097 431
5	0.103 223	10	0.097 251
6	0.098 832	11	0.097 470
7	0.099 231	12	0.098 190
8	0.095 886	13	0.098 274

2）训练函数的选择.

在仿真实验中，BPNN，GA－BP 和 IGA－BP 都采用相同的网络结构：单隐层，隐层神经元的数量为 8，隐层的传递函数为 S 型函数；输出层的传递函数是线性激活函数. 三种算法的各层传递函数是一致的.

有多种训练算法可以用来训练 BPNN，最具代表性的有标准梯度下降法(LGD)、有动量的梯度下降法(GDM)、Fletcher－Reeves 共轭梯度法(CGF)，LM 梯度下降法(LM)等. 这些不同的训练方法针对 IGA－BP 算法的性能表现见图 2.

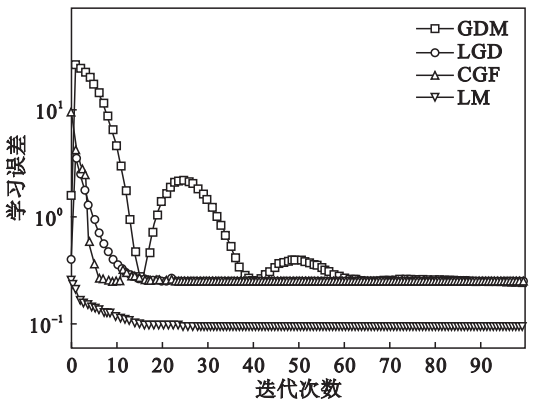

图 2　不同训练函数的性能

Fig. 2　Performance of different training functions

从图 2 能发现，GDM 算法有明显的震荡，并在网络迭代 60 次之后趋于稳定，当网络稳定时，GDM 算法输出误差最大. LGD 算法的曲线更平滑，但当网络收敛时，它的表现和 GDM 算法一样差. CGF 算法的曲线是最陡的，并且有着最快的收敛速率，当迭代到 10 次左右时网络就收敛了，但 CGF 算法的误差仍然很高，这说明 CGF 算法很容易陷入局部最优且无法逃逸. LM 算法无论是在准确率还是在收敛速度上都表现出了最好的性能，因此选择 LM 作为网络的训练函数.

2.3　算法参数的选取

由于 BPNN 的权值和阈值的取值范围在[－1，1]之间，所以 GA 算法中每个染色体基因向量的分量随机初始值都在[－1，1]之间. 因为已经确定 BPNN 输入神经元的数量为 6，隐层神经元的数量为 8，因此可得每个染色体基因向量的维度为 $6\times8+8\times1+8\times1+1=65$.

2.4　算法的输出误差分析

将 IGA－BP 的性能与标准 BP 以及 GA－BP 进行对比实验. GA－BP 算法中，个体进行交叉及变异的概率都是固定的，具体参数为：种群规模 30，进化次数 100，交叉概率 0.4，变异概率 0.002.

三种算法的输出误差如图 3 所示.

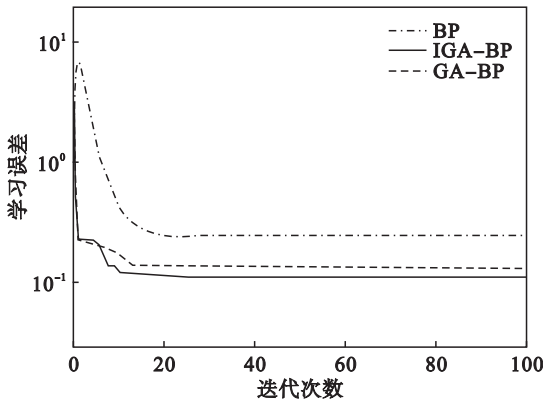

图 3 三种算法的学习误差

Fig. 3 Learning error of three algorithms

在三种算法中,BPNN 的结构和传递函数都是一样的,并且训练函数都为 LM 梯度下降函数. 从图 3 中可以发现,标准 BP 输出误差最大,这是由于其全局搜索能力差,网络性能依赖权值和阈值的随机初始值;IGA – BP 算法的输出误差最小. IGA – BP 算法与 GA – BP 算法在前三次迭代时曲线几乎是重合的,GA – BP 在之后的训练过程中平稳地进行搜索,而 IGA – BP 算法曲线有一个明显的抖动,这是因为 IGA – BP 摆脱局部最优的能力要强于 GA – BP,所以 IGA – BP 算法在用户流失预测问题上表现出了很好的性能.

从收敛速度上看,IGA – BP 算法比标准 BP 和 GA – BP 算法的收敛速度都快. IGA – BP 算法大约迭代 10 次就收敛了,而 GA – BP 大约迭代 17 次才收敛,标准 BP 算法大约迭代 20 次才收敛. 这是因为 IGA – BP 的交叉和变异概率与迭代次数有关,因此算法前期有着很强的全局搜索动力,然后又会很快进入稳定状态.

2.5 算法的预测准确率分析

算法预测准确率定义为样本数据正确分类的百分比,即 TP(true positive)值. 三种算法对用户流失的预测结果见表 3、表 4. 可见,IGA – BP 算法针对移动用户数据集展现了很强的流失预测能力.

表 3 三种算法的预测分类矩阵

Table 3 Prediction classification matrix of three algorithms

算法	用户状态	预测流失	预测未流失
BP	实际流失	22 880	27 120
	实际未流失	7 285	42 715
GA – BP	实际流失	25 413	24 587
	实际未流失	6 541	43 459
IGA – BP	实际流失	26 536	23 464
	实际未流失	6 147	43 853

表 4 三种算法的预测值指标

Table 4 Prediction indexes of three algorithms %

预测模型	TN	TP	FP	FN
BP	85.43	45.76	14.57	54.24
GA – BP	86.92	50.83	13.08	49.17
IGA – BP	87.71	53.07	12.29	46.93

注:TN, true negative;TP, true positive;FP, false positive;FN, false negative.

TP 是用户流失预测中最重要的指标,基于 BP,GA – BP 以及 IGA – BP 三种算法,用户数据 6 个属性对 TP 值的影响如图 4 所示.

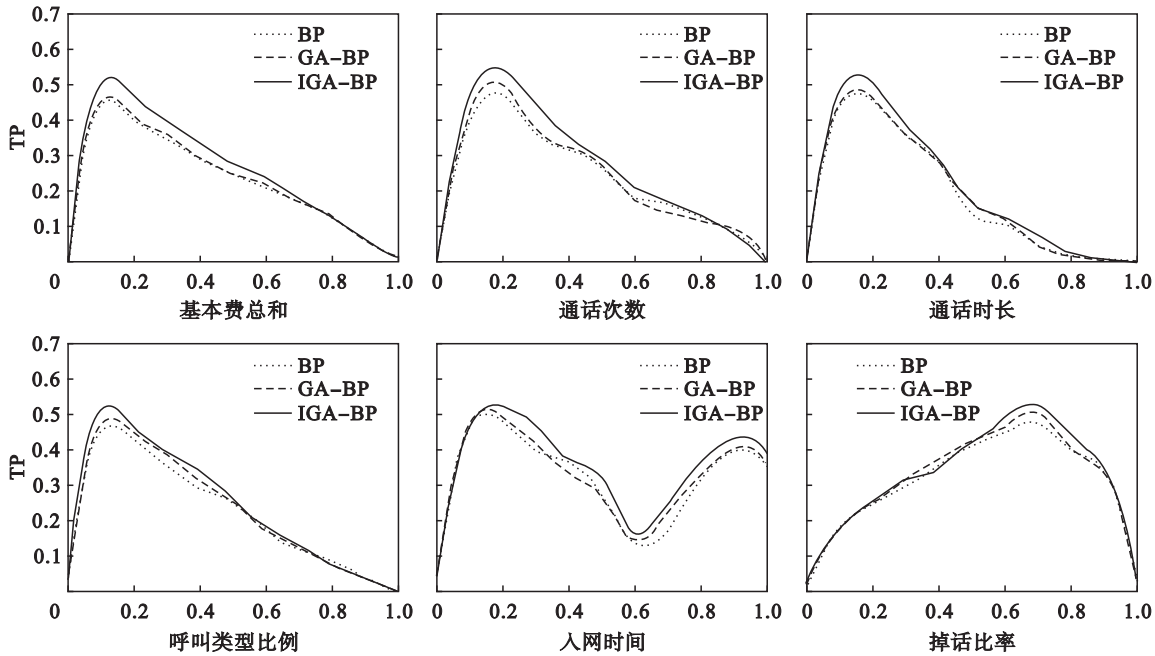

图 4 基于不同算法的不同属性对 TP 值的影响

Fig. 4 Effect of different attributes on TP based on different algorithms

为了消除指标之间的量纲影响，对数据进行了 min – max 标准化处理以方便指标之间的对比. 由图 4 可以发现，当用户的基本费总和、通话次数、通话时长以及呼叫类型比例这 4 个属性值很低时，三种算法都有较高的 TP 值. 但处理分布在两端的数据时，三种算法的表现都相对较差. GA – BP 曲线和 BP 曲线在很多分布区域都是重合的，这是因为标准 GA 的遗传操作类似于一种随机策略，在进行交叉和变异操作时，没有考虑个体的适应度值，很可能破坏了好的基因结构，因此没有起到很好的优化作用. 相比于 BP 和 GA – BP，IGA – BP 的 TP 值始终是最高的，这意味着 IGA – BP 模型的预测准确率也是最高的.

由图 4 还可以发现，随着入网时间的变化，3 条曲线有着相同的震荡趋势；入网时间较早和较晚的用户更有可能流失. IGA – BP 算法仍然有着最高的 TP 值. 此外，当用户遭遇频繁掉话时，用户流失的可能性更大. 相比其他两种算法，IGA – BP 有着更高的 TP 值.

综上，相比 BP 和 GA – BP，IGA – BP 算法在 TP 指标上有显著提高.

3　结　　语

本文提出用改进遗传算法来初始化 BPNN 的权值和阈值，弥补其全局搜索能力的不足. 由于 BPNN 待确定的权值和阈值变量较多，并且精度要求高，本文遗传算法的染色体采用浮点数编码，并将网络的输出误差作为个体的适应度函数. 改进遗传算法中提出一种自适应概率的交叉操作和变异操作，根据个体的适应度值以及种群当前迭代次数来计算自适应概率. 为了保证算法的稳定性，种群发生交叉和变异的概率也会随着进化的次数逐渐变小. 通过对预测准确率和网络输出误差的分析，证明了基于改进遗传算法优化 BPNN 的预测模型具有良好的性能.

参考文献：

[1] Luo C, Zeng J, Yuan M, et al. Telco user activity level prediction with massive mobile broadband data[J]. *ACM Transactions on Intelligent Systems & Technology*, 2016, 7(4): 1 – 30.

[2] Coussement K, Lessmann S, Verstraeten G. A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry [J]. *Decision Support Systems*, 2017, 95: 27 – 36.

[3] Keramati A, Jafari-Marandi R, Aliannejadi M, et al. Improved churn prediction in telecommunication industry using data mining techniques[J]. *Applied Soft Computing*, 2014, 24: 994 – 1012.

[4] Kim K, Jun C H, Lee J. Improved churn prediction in telecommunication industry by analyzing a large network [J]. *Expert Systems with Applications*, 2014, 41(15): 6575 – 6584.

[5] Vafeiadis T, Diamantaras K I, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction[J]. *Simulation Modelling Practice and Theory*, 2015, 55: 1 – 9.

[6] Lu N, Lin H, Lu J, et al. A customer churn prediction model in telecom industry using boosting[J]. *IEEE Transactions on Industrial Informatics*, 2014, 10(2): 1659 – 1665.

[7] Kisioglu P, Topcu Y I. Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey[J]. *Expert Systems with Applications*, 2011, 38(6): 7151 – 7157.

[8] García S, Fernández A, Herrera F. Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems[J]. *Applied Soft Computing*, 2009, 9(4): 1304 – 1314.

[9] Farquad M A H, Ravi V, Raju S B. Churn prediction using comprehensible support vector machine: an analytical CRM application[J]. *Applied Soft Computing*, 2014, 19: 31 – 40.

[10] Pendharkar P C. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services[J]. *Expert Systems with Applications*, 2009, 36(3): 6714 – 6720.

[11] Subramanian K, Suresh S. A meta-cognitive sequential learning algorithm for neuro-fuzzy inference system[J]. *Applied Soft Computing*, 2012, 12(11): 3603 – 3614.

[12] Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment[J]. *Expert Systems with Applications*, 2014, 41(4): 2052 – 2064.