第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于遗传算法和BP神经网络的孤立性肺结节分类算法

胡 强 郝晓燕 雷 蕾

(太原理工大学计算机科学与技术学院 太原 030024)

摘 要 为了提高计算机辅助诊断系统中孤立性肺结节的良恶性诊断的准确性,提出了一种基于遗传算法和BP神经网的分类算法。该算法针对BP神经网络容易陷入局部最优的问题,综合考虑孤立性肺结节的医学诊断特性,采用遗传算法对基于BP神经网络的分类器进行优化,并通过对PET/CT图像进行处理,提取病灶的功能特征、结构特征以及临床信息作为神经网络分类器的输入样本,实现孤立性肺结节的良恶性分类。对医院以及网络公共数据库中的大量实验数据进行分类实验,结果表明优化后的算法在分类准确性上有较大的提高,说明该方法在肺结节临床分类方面是有效的。

关键词 孤立性肺结节,BP神经网络,遗传算法,分类

中图法分类号 TP391 **文献标识码** A

Solitary Pulmonary Nodules Classification Based on Genetic Algorithm and Back Propagation Neural Networks

HU Qiang HAO Xiao-yan LEI Lei

(College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract In order to improve the accuracy of benign and malignant diagnosis of the solitary pulmonary nodules in the computer aided diagnosis system, this paper proposed a novel classification algorithm based on genetic algorithm and back propagation neural networks. Considering the local optimum problem of the BP neural networks and the medical diagnosis features of solitary pulmonary nodules, the proposed algorithm uses genetic algorithm to optimize the classifier based on BP neural networks. Through the PET/CT image processing, the functional characteristics, structural characteristics and clinical information of the lesions are extracted as input samples of the neural network based classifier. Then, the benign and malignant diagnosis of the solitary pulmonary nodules is realized by the novel classifier. Classify experimental results on a large number of experiment data from a hospital and public databases on network show that the optimized algorithm is greatly improved on the classification accuracy, indicating that this method is effective in clinical classification of pulmonary nodules.

Keywords Solitary pulmonary nodules, Back propagation neural networks, Genetic algorithm, Classification

1 引言

孤立性肺结节是肺癌的一种早期表现,表现为肺实质内单发的、最大直径小于3厘米且不存在肺炎等其它病变的类圆形结节影[1]。CT,即计算机断层扫描成像,以高分辨率的结构成像显示组织形态密度,用于病灶的定位;PET,即正电子发射断层扫描成像,以分子水平的代谢显像显示组织代谢,用于辅助病灶定性。PET/CT成像将二者结合起来,在精确定位的同时辅助定性,是最有效的影像方式之一。为了在早期发现肺癌并及时对肺癌患者给出诊疗建议,采用计算机辅助孤立性肺结节的检测和诊断是十分必要的。计算机辅助诊断系统中采用的医学图像处理技术和相关算法一直是研究的热点问题,其中孤立性肺结节良恶性诊断中涉及的分类算法的准确性也亟待提高。

大量学者就孤立性肺结节的分类方法进行了研究,其中常用的分类器主要有贝叶斯分类器[2]、支持向量机[3]以及神经网络分类器等。与泛化性较差的贝叶斯分类器和内存需求大的支持向量机相比,人工神经网络以并行方式进行信息处理,以其自学习、自组织等优势得到了广泛的应用。其中,BP神经网络在理论研究和应用方面是最成熟的网络模型之一,在解决非线性分类问题中表现出了较大的优势[4-7],但容易陷入局部最小。Li M等人[8]研究了BP神经网络的优化算法及其在模式识别中的应用。国内外学者将遗传算法和BP神经网络相结合,能够实现两种算法的优势互补,在传感器温度补偿[9]、人工时预测[10]以及钢轨磨损量预测[11]等应用中表现出较好地收敛性和稳定性。

鉴于BP神经网络在解决分类问题中的优势,本文将其应用到PET/CT图像中的孤立性肺结节分类诊断中,针对其易陷入局部最小值的缺点,引入遗传算法对基于BP神经网络的分类器进行优化。通过提取肺结节的医学征象,利用基于遗传算法和BP神经网络的分类方法对肺结节进行分类,以期及早发现恶性病变,以免延误最佳治疗时机。

本文受山西省自然科学基金(2012011011-2)资助。

胡 强(1990—),男,硕士生,主要研究方向为计算机语言学、自然语言处理,E-mail:1393983650@qq.com;郝晓燕(1970—),女,博士,副教授,主要研究方向为计算机语言学、自然语言处理;雷 蕾(1992—),女,硕士生,主要研究方向为图像处理。

2 理论基础

BP 神经网络是一种多层前馈神经网络，按照误差逆向传播算法进行训练，由数据流的向前计算（正向传播）和误差信号的反向传播两个过程构成。

设网络中输入层、隐含层以及输出层中的节点个数分别为 n、q 和 m，令 v_{ki} 为输入层与隐含层之间的权值，w_{jk} 为隐含层与输出层之间的权值，$f_1(\cdot)$、$f_2(\cdot)$ 分别为隐含层和输出层的传递函数，则正向传播时，样本从输入层经过全部隐含层传向输出层，得到隐含层节点和输出层节点的输出分别为 z_k、y_j。

$$z_k = f_1\left(\sum_{i=0}^{n} v_{ki} \times x_i\right), k=1,2,\cdots,q \tag{1}$$

$$y_j = f_2\left(\sum_{k=0}^{q} w_{jk} \times z_k\right), j=1,2,\cdots,m \tag{2}$$

用 x^p 表示第 p 个学习样本，其输出值为 $y_j^p (j=1,2,\cdots,m)$，采用均方误差函数定义误差值 E_p：

$$E_p = \frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2 \tag{3}$$

其中，t_j^p 为期望输出。则样本总体在整个网络学习过程中的总误差可以表示为：

$$E = \frac{1}{2}\sum_{p=1}^{p}\sum_{j=1}^{m}(t_j^p - y_j^p)^2 = \sum_{p=1}^{p} E_p \tag{4}$$

比较实际输出与期望输出，若二者关系满足定义的误差要求，则算法结束，表示由输入空间向输出空间的映射完成；否则进入反向传播过程。

在误差信号的反向传播过程中，采用累计求和误差 BP 神经网络算法调整每层的各个神经元的权值和阈值，以期减小全局误差 E。输出层和隐含层中神经元的权值调整公式分别为：

$$\Delta w_{jk} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta(t_j^p - y_j^p) f_2'(S_j) z_k \tag{5}$$

$$\Delta v_{ki} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta(t_j^p - y_j^p) f_2'(S_j) w_{jk} f_1'(S_k) x_i \tag{6}$$

其中，η 为学习率。

通过信号的前向传播和反向传播两个过程交替迭代进行，神经网络中的样本空间的误差呈梯度下降。经过这种交替过程，最终得到最优的网络结构，同时得到最小的误差函数值，表征 BP 神经网络的学习过程完成。BP 神经网络的这种训练机制要求神经元的特性函数必须是可导的，且对应的两次迭代采用正交的搜索方向，导致了一种“锯齿”型寻优，不利于网络得到全局最优解。同时，BP 神经网络这种反向传播机制使得其收敛速度较慢。

3 本文方法

3.1 肺结节特征提取

对孤立性肺结节进行特征提取和优化是实现准确分类的基础和前提。在 PET/CT 混合成像中，PET 图像中显示的是组织的代谢特征，表现为组织的标准摄取量（Standard Uptake Value，SUV）；CT 图像中显示的是组织的结构信息，可以提取到肺结节区域的几何特征、灰度特征、纹理特征以及空间位置等特征。将上述图像特征与病例临床信息进行组合，构成 23 维的特征向量，并将其作为分类器的样本输入值。将第 j $(j=1,2,\cdots,23)$ 个特征记为 F_j，具体定义如下：

1）代谢特征

PET 图像提供的功能影像反映了孤立性肺结节分子水平的代谢信息，所提取的特征记为 F_1，即 SUVmax。

2）形态特征

肺结节的形态特征主要提取其几何特征和医学征象特征。其中几何特征包括肺结节面积 $Ar(F_2)$、最大直径 $Dx(F_3)$、圆形度 $Cl(F_4)$，由式(7)—式(9)确定：

$$Ar = \sum_{(x,y)\in S} I(x,y) \tag{7}$$

$$Dx = \max_{(x_1,y_1)\in E,(x_2,y_2)\in E} \sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{8}$$

$$Cl = \frac{4\pi Ar}{(Count(I(x_i,y_i))+\sqrt{2}Count(I(x_j,y_j)))^2} \tag{9}$$

其中，S 表示肺结节区域，$I(x,y)$ 表示肺结节内的像素点，E 为肺结节边界点的集合，(x_1,y_1) 和 (x_2,y_2) 为 E 上的任意两点。$I(x_i,y_i)$ 和 $I(x_j,y_j)$ 分别表示两边界点间处于水平或垂直关系和处于对角关系的像素点集。

医学征象特征包括空洞个数 (F_5)、分叶等级 (F_6)、毛刺等级 (F_7) 以及钙化度 $Cpr(F_8)$，其中前 3 个特征由病理标注信息读入，钙化度由式(10)计算得到：

$$Cpr = \frac{Count(I(x,y)>T)}{Ar} \tag{10}$$

其中，$Count$ 表示计数，T 为设定的钙化灰度阈值。

3）灰度特征

肺结节的灰度特征主要包括灰度均值 $Gm(F_9)$、灰度标准差 $Gu(F_{10})$ 及区域对比度 $Const(F_{11})$，由式(11)—式(14)确定：

$$Gm = \frac{1}{Ar}\sum G(x,y) \tag{11}$$

$$Gu = \sqrt{\frac{1}{Ar}\sum_{(x,y)\in R}(G(x,y)-Gm)^2} \tag{12}$$

$$Const = \mathrm{abs}\left(\frac{Gm-bM}{Gm+bM}\right) \tag{13}$$

其中，bM 是一个中间变量：

$$bM = \frac{\dfrac{\sum_{(x,y)\in(M\times N)} G(x,y)}{M\times N}\times M\times N - Gm\times Ar}{M\times N - Ar} \tag{14}$$

其中，$G(x,y)$ 表示 CT 图像中像素的灰度值，$M\times N$ 为结节外接矩形域。

4）纹理特征

灰度共生矩阵主要用来描述图像上不同区域灰度级的差异与空间分布情况，本文采用基于外接矩形的灰度共生矩阵所提取的 ASM 能量 (F_{12})、对比度 (F_{13})、熵 (F_{14})、局部平稳性 (F_{15})、惯量 (F_{16}) 来分别描述图像灰度分布均匀度或平滑度、图像清晰度、图像具有的信息量、图像局部像素的平均水平以及图像纹理的粗细程度，分别由式(15)—式(19)确定：

$$f_1 = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P_\delta^2(i,j) \tag{15}$$

$$f_2 = \sum_{n=0}^{L-1} n^2 \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P_\delta(i,j) \tag{16}$$

$$f_3 = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P_\delta(i,j)\log P_\delta(i,j) \tag{17}$$

$$f_4 = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\frac{P_\delta(i,j)}{1+(i-j)^2} \tag{18}$$

$$f_5 = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-j)^2 P_\delta(i,j) \tag{19}$$

其中，i 和 j 分别表示像素的灰度，δ 表示两个像素的位置关系，L 表示图像的灰度级，$P_\delta(i,j)$ 是从图像中灰度级为 i 的点离开固定位置 δ 达到灰度级为 j 的点的概率。

5）空间位置

空间位置特征主要提取肺结节距肺壁最小距离 $Dst(F_{17})$ 和质心距 $Pcl(F_{18})$，由式(20)、式(21)确定：

$$Dst=\min_{(x,y)\in E}\sqrt{(x_0-x)^2+(y_0-y)^2} \tag{20}$$

$$Pcl=\sqrt{(x_0-x_A)^2+(y_0-y_A)^2} \tag{21}$$

其中，(x_0,y_0) 表示疑似结节的质心，(x,y) 表示疑似结节所在肺实质（左肺或右肺）的边界点集 E 中的元素，(x_A,y_A) 表示左右肺实质的质心

6）临床信息

影响肺结节良恶性诊断的临床特征记为 $(F_{19}-F_{23})$：年龄、性别、吸烟量、戒烟史、恶性肿瘤病史。

3.2 基于 GA-BP 神经网络的分类算法

为了避免在使用 BP 神经网络进行分类的过程中陷入局部最优，并提高收敛速度，采用遗传算法（GA）对 BP 神经网络的训练过程进行优化和改进，使得搜索过程始终遍及整个解空间，从而得到全局最优解。GA 与 BP 神经网络的结合方式主要有：将 GA 作为 BP 神经网络的一种学习算法对网络进行训练，在不改变网络结构的同时实现网络权值的进化；利用 GA 对 BP 神经网络连接方式进行编码，实现网络的结构进化。在这种方式下，一般选择其它的学习算法进行权值训练，来评估网络结构的适应度；采用 GA 实现 BP 神经网络的学习规则进化。

本文设计的分类方法主要是针对 BP 神经网络的学习过程进行优化的。首先将网络的连接权重和阈值进行编码，作为遗传算法中的染色体，然后构造对应的适应函数，进行算法迭代，直到达到终止条件，网络优化完成。与普通 BP 学习算法相比，通过 GA 优化后的 BP 神经网络算法能够较好地收敛到全局最优解，在利用神经网络泛化的映射能力的同时，可以提高神经网络的收敛性和学习能力。遗传算法在进行算法优化的过程中涉及到了两个关键的问题：染色体的编码设计和评价函数的构造。

染色体编码设计：设网络中输入层、隐含层以及输出层中节点个数分别为 i、j 和 k，则 BP 算法学习后生成的网络结构可用以下 4 个矩阵表示：输入层到隐含层和 M 隐含层到输出层的权值矩阵 M、N，以及隐含层和输出层的阈值矩阵 t、t'。

$$M=\begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1j} \\ M_{21} & M_{22} & \cdots & M_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{i1} & M_{i2} & \cdots & M_{ij} \end{pmatrix}$$

$$N=\begin{pmatrix} N_{11} & N_{12} & \cdots & N_{1j} \\ N_{21} & N_{22} & \cdots & N_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ N_{i1} & N_{i2} & \cdots & N_{ij} \end{pmatrix}$$

$$t=\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_j \end{pmatrix},t'=\begin{pmatrix} t_1' \\ t_2' \\ \vdots \\ t_k' \end{pmatrix}$$

采用遗传算法对神经网络中的权值和阈值进行优化，就可以转换为对以上矩阵的优化，需要将它们进行编码，构成遗传算法操作中的染色体。在这里，将 M、t、N、t' 中的元素按顺序串联起来，构成一个二进制串，作为染色体的编码，映射关系如图 1 所示。

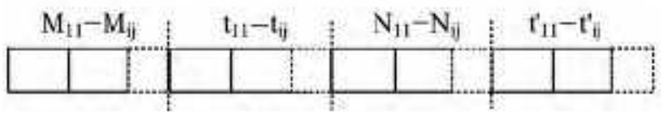

图 1 染色体编码设计

将网络的权系数通过编码编入遗传算法的染色体后，遗传算法的寻优过程还需要设计适应度函数，即评价函数，这里采用基于 BP 网络误差值的适应度函数 $f(x)$，定义为均方误差的倒数：

$$f(x)=\frac{1}{\frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{L}(d_{pk}-o_{pk})^2} \tag{22}$$

将基于遗传算法的 BP 神经网络算法描述如下：

Step1 初始化。输入样本数据，分别生成 BP 网络的输入、输出矩阵，定义学习速率；随机生成遗传算法中的染色体数据，定义遗传操作的交叉、变异概率。

Step2 按照染色体编码方式，将其映射为对应的权系数矩阵 M、t、N。

Step3 通过生成的矩阵计算输出矩阵，计算样本的均方误差。

Step4 根据式(22)计算个体的适应度函数值。

Step5 采用轮盘法复制生成新种群，并对个体执行交叉和变异操作。

Step6 循环执行上述步骤直到达到最大迭代次数；循环结束后转到 Step7。

Step7 利用上述优化过程中得到的全局解进一步优化 BP 神经网络中的权系数，并通过 BP 算法自身反演调节，直到网络在分类过程中的误差精度足够小为止。

4 实验结果及分析

为了验证设计的分类算法的有效性，本文在 MALAB 2012b 环境下进行分类对比实验。实验中使用来自山西某医院 PET/CT 中心 2011 年 10 月—2015 年 10 月之间的共 150 例肺部 PET/CT 序列图像数据（其中恶性 90 例，良性 60 例）。图像采集设备为 Discovery ST16 PET-CT，CT 采集参数为 150mA、140kV，层厚 3.75mm，图像大小为 512×512。纳入研究数据集的标准为：肺内单发肺结节，结节最大直径≤30mm，边界清楚不透光影，不伴有肺不张、局部淋巴结肿大和肺炎，有明确的病例诊断。

实验采用对比的方式，分别采用文献[2,3,8]中的方法、基本 BP 神经网络与本文中提出的基于优化的 BP 神经网络分类方法对上述 150 例数据构成的数据集进行分类。分类实验对每个算法采用 5 次五折交叉验证，即将数据集分为 5 份（每份中恶性 18 例、良性 12 例），轮流将其中的 1 份作为训练集，剩下的 4 份作为测试集，求其均值以评估分类器的有效性。

不同方法的分类结果如表 1 所列。与文献中常用的方法及基本的 BP 神经网络算法相比，本文方法在对数据集中的

（下转第 54 页）

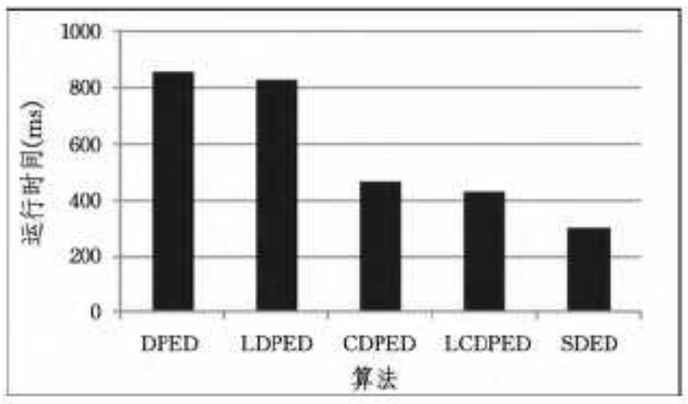

图 1　5 种算法的平均查询时间

下面对 LCDPED 和 SDED 算法在不同长度字符串中的性能进行比较，分别将 LCDPED 和 SDED 算法运用于 Author、Author1、Author2 中，得到的结果如图 2 所示。从图中可知，与 LCDPED 算法相比，本文算法不论在那个数据上都有显著的优势。

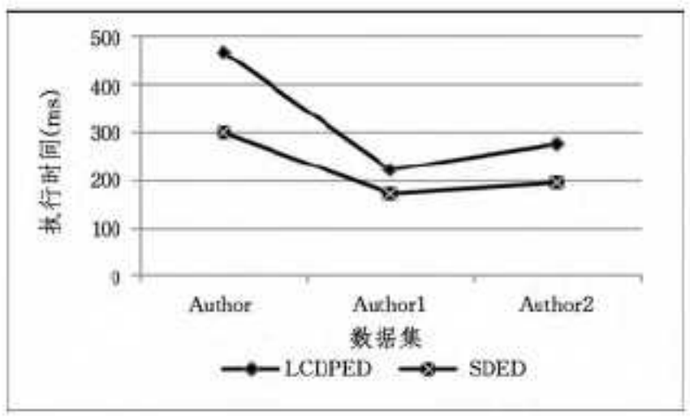

图 2　LCDPED 和 SDED 在不同长度字符串中的执行时间

结束语　本文提出了基于基本操作序列的编辑距离顺序验证方法。在阈值为 2 的近似串搜索实验中发现，相较于动态规划法及其改进算法，本文方法不论在长字符还是短字符中都具有良好的效果。

阈值较大时，本文方法需要验证的次数较多，时间效率会下降，所以本文算法适用于阈值较小的情况。

参 考 文 献

[1] Deng D, Li G, Feng J, et al. Top-k string similarity search with edit-distance constraints[C]// ICDE. 2013: 925-936

[2] Bayardo R, Ma Y, Srikant R. Scaling up all-pairs similarity search[C]// WWW Conference. 2007

[3] Chaudhuri S, Ganjam K, Ganti V, et al. Robust and Efficient Fuzzy Match for Online Data Cleaning[C]// SIGMOD. 2003: 313-324

[4] Jiang Y, Li G, Feng J, et al. String similarity joins: An experimental evaluation[J]. Proceedings of the VLDB Endowment, 2014, 7(8): 625-636

[5] Li C, Lu J, Lu Y. Efficient merging and filtering algorithms for approximate string searches[C]// ICDE. 2008: 257-266

[6] 纳瓦罗. 柔性字符串匹配[M]. 北京：电子工业出版社，2007

[7] 姜华，韩安琪，王美佳，等. 基于改进编辑距离的字符串相似度求解算法[J]. 计算机工程，2014，40(1)：222-227

[8] 黄亮，赵泽茂，梁兴开. 基于编辑距离的 Web 数据挖掘[J]. 计算机应用，2012，32(6)：1662-1665

[9] 薛晔伟，沈钧毅，张云. 一种编辑距离算法及其在网页搜索中的应用[J]. 西安交通大学学报，2008，42(12)：1450-1454

(上接第 39 页)

孤立性肺结节进行分类实验时，样本正例率(TPR)、反例率(TNR)和总体准确率都有了较大的提高，表现出了较好的分类性能，说明该方法在孤立性肺结节分类中是有效的。

表 1　不同分类方法的分类结果

方法	TPR(%)	TNR(%)	诊断正确率(%)
Bayesian hybrid classifier[2]	90.21	92.53	91.37
SVM[3]	87.39	91.29	89.34
BP neural networks[7]	84.80	90.34	87.57
LM-based BP networks[8]	88.02	91.82	89.92
Our method	93.33	95.25	94.29

结束语　在肺部影像的计算机辅助诊断系统中，孤立性肺结节的良恶性诊断是关键。为了提高诊断的准确性，本文提出了一种基于改进 BP 神经网络的肺结节分类算法，该算法基于孤立性肺结节的医学诊断属性，将遗传算法与 BP 神经网络结合起来，构造分类器，在准确特征提取的基础上实现了孤立性肺结节的良恶性分类。实验中分别采用文献中的方法、基本的 BP 神经网络以及本文方法对医院及网络公共数据集中的肺结节数据进行分类实验，根据分类结果验证方法的有效性。结果表明，优化后的算法表现出较好的分类性能，在分类准确性上较其它方法有了较大的提高，是肺结节临床分类方面的有效性算法。

参 考 文 献

[1] 强彦，卢军佐，赵涓涓，等. 基于 PET/CT 的孤立性肺结节的自动分割方法[J]. 清华大学学报(自然科学版)，2013，(2)：200-204

[2] Calle-Alonso F, Pérez C J, Arias-Nicolás J P, et al. Computer-aided diagnosis system: A Bayesian hybrid classification method [J]. Computer methods and Programs in Biomedicine, 2013, 112 (1): 104-113

[3] Keshani M, Azimifar Z, Tajeripour F, et al. Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system[J]. Computers in Biology and Medicine, 2013, 43(4): 287-300

[4] Tian G, Qi W. Target Recognition Algorithm Based on BP Networks and Invariant Moments[J]. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2013, 11(2): 969-974

[5] 李松，刘力军，刘颖鹏. 改进 PSO 优化 BP 神经网络的混沌时间序列预测[J]. 计算机工程与应用，2013，49(6)：245-248

[6] Kai S, Zhikun L, Hang S, et al. A research of maize disease image recognition of corn based on BP networks[C]// 2011 Third International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2011, 1: 246-249

[7] 王先军，白国振，杨勇明. 复杂背景下 BP 神经网络的手势识别方法[J]. 计算机应用与软件，2013，30(3)：247-249

[8] D Shuo, C Xiao-heng, W Qing-hui, et al. Application of BP Networks Improved by LM Algorithm in Pattern Classification[J]. Electronic Test, 2014, 2: 16

[9] 彭基伟，吕文华，行鸿彦，等. 基于改进 GA-BP 神经网络的湿度传感器的温度补偿[J]. 仪器仪表学报，2013(1)：153-160

[10] 戴健伟，吉华，杨岗，等. 基于 GA-BP 算法的化工设备设计人工时预测[J]. 计算机集成制造系统，2013(7)：1665-1675

[11] 王平，王彩芸，王文健，等. GA-BP 网络在钢轨磨损量预测中的应用[J]. 润滑与密封，2011(2)：99-102，71