doi: 10.13241/j.cnki.pmb.2019.05.037

·技术与方法·

肿瘤亚型识别研究中智能算法的应用 *

程慧杰[1] 陈 滨[1] 刘芷余[1] 何 颖[1] 卜宪庚[1] 高 越[2△]

（1 哈尔滨医科大学基础医学院 黑龙江 哈尔滨 150086；2 哈尔滨医科大学附属第四医院 黑龙江 哈尔滨 150001）

摘要 目的：为解决肿瘤亚型识别过程中易出现的维数灾难和过拟合问题，提出了一种改进的粒子群 BP 神经网络集成算法。**方法：**算法采用欧式距离和互信息来初步过滤冗余基因，之后用 Relief 算法进一步处理，得到候选特征基因集合。采用 BP 神经网络作为基分类器，将特征基因提取与分类器训练相结合，改进的粒子群对其权值和阈值进行全局搜索优化。**结果：**当隐含层神经元个数为 5 时，候选特征基因个数为 110 时，QPSO/BP 算法全局优化和搜索，此时的分类准确率最高。**结论：**该算法不但提高了肿瘤分型识别的准确率，而且降低了学习的复杂度。

关键词：特征基因；BP 神经网络；粒子群优化算法；肿瘤亚型识别；集成分类器

中图分类号：R73-3；Q-33 **文献标识码：**A **文章编号：**1673-6273(2019)05-960-05

Application of An Intelligent Algorithm in Tumor Subtype Recognition*

CHENG Hui-jie[1], CHEN Bin[1], LIU Zhi-yu[1], HE Ying[1], BU Xian-geng[1], GAO Yue[2△]

(1 Basic Medical College, Harbin Medical University, Harbin, Heilongjiang, 150086, China;

2 The Fourth Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, 150001, China)

ABSTRACT Objective: In order to solve the dimension disaster and over-fitting problems in the process of tumor subtype recognition, a particle swarm optimization (PSO) BP neural network ensemble algorithm was proposed. **Methods:** The Euclidean distance and mutual information was used to preliminarily filter redundant genes, and then Relief algorithm was adopted to further process the candidate feature genes set. The BP neural network was used as the base classifier, which combines feature genes extraction with classifier training. **Results:** When the number of hidden layer neurons is 5 and the number of candidate feature genes is 110, the QPSO/BP algorithm can optimize and search globally. **Conclusion:** The algorithm not only improves the accuracy of tumor classification and recognition, but also reduces the complexity of learning.

Key words: Feature gene; BP neural network; Particle swarm optimization (PSO); Tumor subtype recognition; Ensemble classifier

Chinese Library Classification(CLC): R73-3; Q-33 Document code: A

Article ID: 1673-6273(2019)05-960-05

前言

在肿瘤亚型分类过程中，由于每个样本都记录了组织细胞中成千上万个基因的表达水平，如此高维数、高冗余的特点导致容易出现“维数灾难”和“过拟合现象”，同时又大大降低分类精度、增加学习和训练的时间与空间复杂度[1-5]。因此，准确识别肿瘤亚型的关键在于提取有效可靠的特征基因(feature genes)。相关系数、Fisher 比率等传统的信息基因提取算法和决策树、支持向量机等单分类器在分类精度和学习复杂度等方面效果欠佳[6-12]。

为解决肿瘤亚型识别中易出现的维数灾难和过拟合问题，本文针对基因表达谱数据的特点，提出了一种用于肿瘤亚型识别的新算法 - 改进的粒子群 BP 神经网络集成算法 EC-PSO/BP (Ensemble Classifier of PSO/BP)。该算法采用欧式距离和互信息来初步过滤冗余基因，Relief 算法对过滤集进一步处理，得到候选特征基因集合。采用 BP 神经网络作为基分类器，将特征基因提取与分类器训练相结合，改进的粒子群训练 BP 网络逼近最优解，同时将特征基因的提取和基分类器的训练结合在一起。

1 候选特征基因集合

EC-QPSO/BP 算法根据欧氏距离、互信息指标来初步过滤冗余基因，然后采用 Relief 算法对过滤集合进一步处理，得到

* 基金项目：黑龙江省教育厅科学技术研究项目（12521258）

作者简介：程慧杰（1979-），女，博士研究生，副教授，主要研究方向：模式识别、生物信息，

电话：17703645500, E-mail：77050957@qq.com

△通讯作者：高越（1987-），女，硕士研究生，主治医师，主要研究方向：小儿疾病，电话：18645071396，E-mail：419535183@qq.com

（收稿日期：2018-12-08 接受日期：2018-12-31）

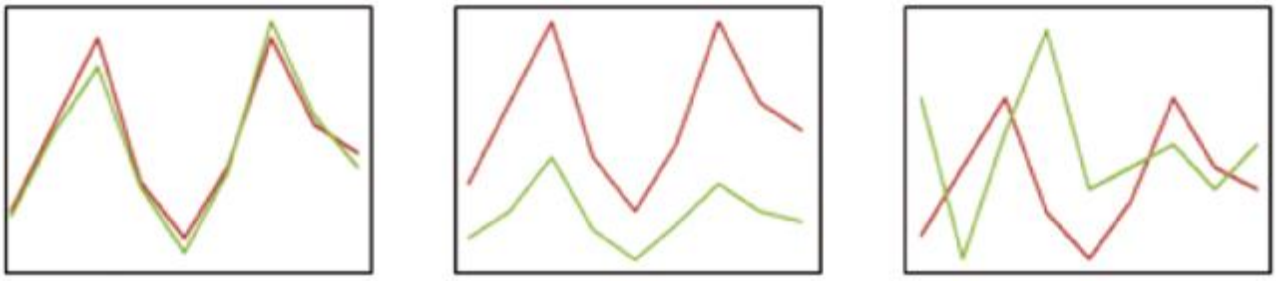
候选特征基因集合。

1.1 基因表达谱数据

肿瘤细胞在某一条件下的全基因组表达数据，可通过一次微阵列实验获得，其结果包含了成千上万个基因在细胞中的绝对或相对丰度[13-16]。多个条件下的全基因组表达数据就构成了一个 M× N 的矩阵 X(公式 1)，M>>N，元素表示第 i 个样本中第 j 个基因的表达水平。

$$X=\begin{bmatrix} x_{11} & x_{12}\cdots & x_{1N} \\ \vdots & \vdots & \vdots \\ x_{M1} & x_{M2}\cdots & x_{MN} \end{bmatrix} \quad (1)$$

两基因间的关系大致可分为三种：相似；变化趋势一致；调控关系输入一样，但是调控结果不同(如图 1)。



a.similar　　b.Consistence of trend　　c. Different regulatory outcomes

1.2 初步过滤

大量功能相关的基因在某些条件下会表现的异常相似，欧式距离可以反映出肿瘤基因之间的共表达关系。当两个基因的表达谱距离小于给定的阈值，则认为它们之间是共表达的。N 维空间中两点 x_i 和 x_j 间的欧式距离公式为

$$D(x_i,x_j)=\frac{\sqrt{\sum_{i=1}^{m}(x_i-x_j)^2}}{m} \quad (2)$$

两个基因间除共表达关系外，还存在着调控关系，互信息正是用来衡量这种调控程度的一个指标。x_i 和 x_j 的熵分别是 $H(x_i)$和 $H(x_j)$，x_i 和 x_j 的联合熵为 $H(x_i,x_j)$，x_i 和 x_j 的互信息定义为 $MI(x_i,x_j)$，公式如下

$$H(X)=-\sum_{k=1}^{m}P(x_k)\log_2 P(x_k)$$

$$MI(x_i,x_j)=H(x_i)+H(x_j)-H(x_i,x_j) \quad (3)$$

1.3 候选特征基因集合

肿瘤基因表达数据经过欧式距离和互信息的初步过滤后，大量的噪声信号、异常数据点和冗余数据被过滤掉[17-19]。Relief 算法(Recursive Feature Elimination)基于属性区分相近样本的能力，将其作为评估属性权重的标准，在一定程度上考虑了肿瘤基因表达谱数据之间的相关性。算法步骤如下：

Step1 随机取样，记为样本 α；

Step2 从样本 α 所在的分类样本组内，随机取 k_1 个最近邻样本 x；

Step3 从样本 α 所在的分类样本组外，随机取 k_2 个最近邻样本 y；

Step4 计算每个特征的权重

$$w_\alpha=w_\alpha+(p_{out}(\alpha,x,y)-p_{in}(\alpha,x,y)) \quad (4)$$

其中 $p_{in}(\alpha,x,y)=|x-y|$，$p_{out}(\alpha,x,y)=|x+y|$。

2 EC-QPSO/BP 集成分类器

EC-QPSO/BP 算法在保证泛化误差一定的条件下，通过增大基分类器之间的差异，来有效降低集成神经网络的泛化误差。基分类器的差异通过增大候选特征子集差异和粒子群算法优化 BP 神经网络，调整隐含层神经元的数目来实现[20-23]。

2.1 粒子群算法 PSO

t 个粒子组成的种群记为 $Z=\{z_1,z_2,\cdots,z_t\}$，其中每个粒子的位置 $z_i=\{z_{i1},z_{i2},\cdots,z_{id}\}$都表示问题的一个潜在解，解 z_i 的优劣由目标函数所确定的适应度值来衡量。粒子将在解空间中运动，并由速度 $V_i=\{v_{i1},v_{i2},\cdots,v_{id}\}$决定其飞行的方向和距离[24-26]。粒子速度更新和位置更新公式如下：

$$v_{id}^{t+1}=wv_{id}^{t}+c_1r_1(p_{id}-z_{id}^{t})+c_2r_2(p_{gd}-z_{id}^{t}) \quad (5)$$

$$Z_{id}^{t+1}=z_{id}^{t}+v_{id}^{t+1} \quad (6)$$

粒子速度的更新由三部分决定，每一部分的相对重要性由权重系数 w，c_1，c_2 决定。第一部分：wv_{id}^{t}粒子先前的速度；第二部分：$c_1r_1(p_{id}-z_{id}^{t})$“认知”部分，根据粒子自身的经验，判断与其自身最佳位置，即个体极值 $P_i=\{p_{i1},p_{i2},\cdots,p_{id}\}$的距离。若 $c_1=0$，则粒子只具有社会经验，此时收敛速度加快、但容易陷入局部最优。第三部分：$c_2r_2(p_{gd}-z_{id}^{t})$“社会”部分，根据粒子群体共享的信息，判断其与群体的最佳位置，即全局极值 $P_g=\{p_{g1},p_{g2},\cdots,p_{gd}\}$的距离。若 $c_2=0$，则粒子没有共享全体信息，得到正确解的几率大大下降。

2.2 BP 神经网络

BP 神经网络(back propagation)一种基于误差逆向传播的多层前馈神经网络，它的拓扑结构模型由输入层(input layer)、一个或多个隐含层(hide layer)以及一个输出层(output layer)组成(如图 1)。理论表明在隐含层神经元可以根据需要自由设置的情况下，三层 BP 神经网络能够以任意精度逼近任意连续函数[27-30]。

BP 神经网络具有较强的自主学习能力，并行的结构系统哦你个，大大提高数据处理的效率，同时能够协调好各个不同学习数据之间的差异，容错性强。但其隐含层的神经元数目难以确定，容易陷入局部极值。

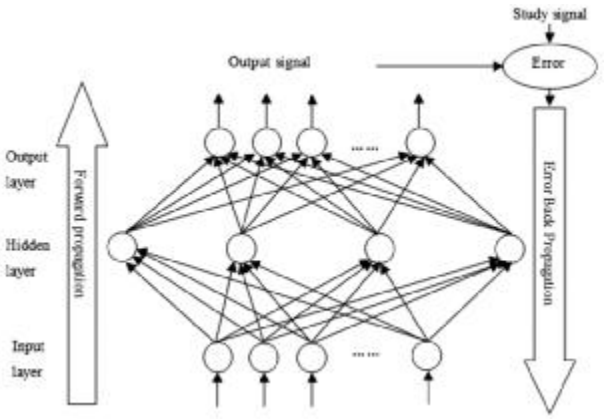


图 2 BP 神经网络结构模型

Fig.2 BP Neural Network Structural Model

2.3 基分类器 PSO/BP

传统 PSO 算法没有考虑到群体中粒子个体之间的相互影响，本文将 PSO 算法做如下改进。将惯性权重系数 w 设定为服从某种分布的随机数，改进后的 w 将避免陷入局部最优解，同时又可在全局解空间范围内进行搜索，加快了学习的速度。

$$w=w_{min}+(w_{max}-w_{min})rnd() \quad (7)$$

w_{max} 和 w_{min} 分别为惯性权重平均值的最大值和最小值，rand()为均匀分布函数，使得在区间内取得最优值与惯性权重平均值的最大值和最小值的概率相等。基分类器算法流程如图 3 所示。

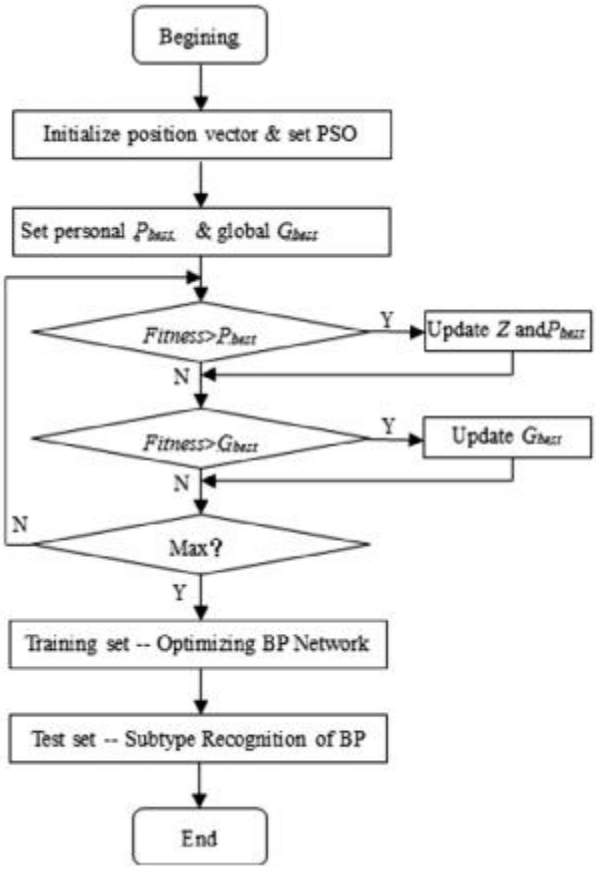


图 3 PSO/BP 基分类器算法流程

Fig.3 The Algorith chart of PSO/BP Classifier

2.4 EC-QPSO/BP 集成模型

该算法采用欧式距离和互信息来初步过滤冗余基因，Relief 算法对过滤集进一步处理，得到候选特征基因集合。采用 BP 神经网络作为基分类器，将特征基因提取与分类器训练相结合，改进的粒子群训练 BP 网络逼近最优解，对其权值和阈值进行全局搜索优化，以选出自适应特征子集、有效提取亚型识别中的特征基因，同时将特征基因的提取和基分类器的训练结合在一起。

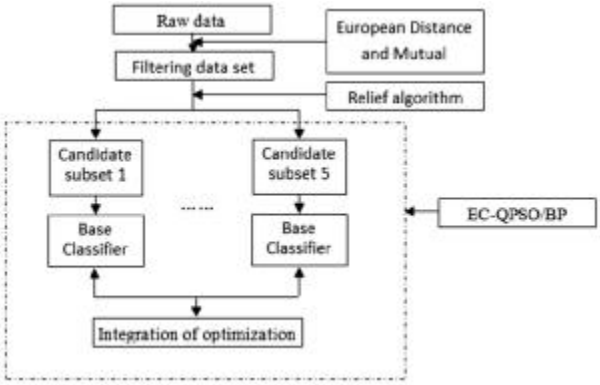


图 4 EC-QPSO/BP 集成模型

Fig.4 EC-QPSO/BP Ensemble Modul

3 结果

3.1 实验数据

为了评估 EC-QPSO/BP 算法的性能，采用 G. Gordon 等人公布的急性白血病数据集(Leukemia)进行实验研究。该数据集由 72 个样本组成，其中急性淋巴性白血病(Acute Lymphoblastic Leukemia, ALL) 样本 47 个，急性骨髓性白血病(Acute Myeloid Leukemia, AML)样本 25 个，此数据集中 72 个样本包含的基因数目为 7129 个。结肠癌数据集(Colon)每个样本都包含了 2000 个基因的表达水平数据，其中正常(Normal)样本 22 个，结肠癌(Tumor)样本 40 个。

3.2 参数设置

EC-QPSO/BP 算法需要对候选特征子集、QPSO 算法和 BP 神经网络三部分分别设置参数。候选特征子集的获得，以欧式距离、互信息和 Relief 算法为平台，这部分的参数设置主要是阈值，在实验中根据实际情况，向最优解不断调整。BP 神经网络的活化函数采用双曲线(Hyperbolic)函数，动量因子的范围在[0,1]。改进的粒子群中 50%的粒子编码进行随机初始化，50%的粒子编码根据特征基因设定。种群中粒子数目 t=30，加速常数 c1=0.15、c2=0.45，最大迭代次数 T=20，每个粒子的适应度从提取的信息基因数目和准确率两个角度来评价。

3.3 实验结果与讨论

EC-QPSO/BP 算法采用五个同型的基分类器进行全局优化搜索，为保证每个基分类器存在差异，从候选特征子集的基因个数和隐含层神经元个数两个方面着手。不同的候选特征子集基因个数和隐含层不同的神经元个数，经过基分类器模型全局优化搜索后，得到的特征基因个数和分类准确率如表 1 所示。

表 1 基分类器参数指标及性能

Table 1 Parameter Index and Performance of Base Classifier

No.	Candidate genes num	Hidden layer neurons	Feature gene num	Accuracy rate
1	125	4	86	84.5%
2	120	5	88	93.4%
3	110	5	68	96.5%
4	100	5	70	91.9%
5	95	4	74	92.1%

从表 1 可知,当设置不同的参数时,候选特征子集中的基因个数产生差异,当隐含层神经元个数为 5 时,候选特征基因个数为 110 时,QPSO/BP 算法全局优化和搜索后获得的特征基因个数为 68 个,此时的分类准确率最高,达到了 96.5%。而当隐含层神经元个数为 4 时, 候选特征基因个数为 125 时,该基分类器的分类准确率最低,仅有 84.5%;其它三个基分类器获得的特征基因个数分别为 88、70 和 74。

以分类准确率最高的基分类器 3 为例,其进化代数与适应度关系曲线入图 4 所示。当量子量子群优化 BP 神经网络,进化到第 60 代左右时,适应度曲线趋于平稳,无明显波动。在此过程中,EC-QPSO/BP 算法算法既没有陷入到局部最优解当中,有没有出现过拟合现象。

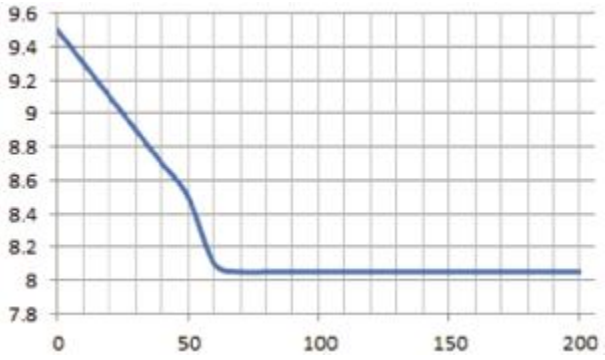


图 5 基分类器进化代数与适应度关系曲线图

Fig.5 Relation Curve of Evolutionary Algebra and Fitness of Base Classifier

4 结论

为解决肿瘤亚型识别过程中易出现的维数灾难和过拟合问题,提出了一种改进的粒子群 BP 神经网络集成算法。算法采用 BP 神经网络作为基分类器,将特征基因提取与分类器训练相结合, 改进的粒子群对其权值和阈值进行全局搜索优化。该算法提高了肿瘤分型识别的准确率, 降低了学习的复杂度。但关于如何提高各个基分类器分类结果的整合,以及进一步解决优化过程中出现的过拟合问题,有待于深入的研究[31]。

参 考 文 献(References)

[1] 李颖新, 李建更, 阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报, 2006, 29(2): 324-330

[2] 韦鹏宇,潘福成,李帅. 改进人工蜂群优化 BP 神经网络的分类研究[J]. 计算机工程与应用, 2018, 54(10): 158-163

[3] 陶晓玲,亢慈楠,刘丽燕. 基于选择性集成的并行分类器融合方法[J]. 计算机工程与科学, 2018, 40(5): 787-792

[4] Zeping Yang, Daqi Gao. Classification for imbalanced and overlapping classes using outlier detection and sampling tech niques [J]. Applied Mathematics & Information Sciences, 2013, 7: 375-381

[5] 张家敏, 许德章. 改进粒子群优化 BP 神经网络的六维力传感器解耦研究[J]. 仪表技术与传感器, 2016, 7: 8-11

[6] 吴志攀,赵跃龙,罗中良等. 基于 PSO-BP 神经网络的识别技术[J]. 中山大学学报, 2017, 56(1): 47-52

[7] 郑蒙蒙,李新利,巨汉基等. 基于 BP 神经网络的电能表软件故障分类研究[J]. 华北电力技术, 2016, 8: 8-12

[8] 王斯盾,琚生根,周刚等. 基于集成分类器的用户属性预测研究[J]. 四川大学学报, 2017, 54(6): 1195-1202

[9] 马超. 基于 FCBJ 特征选择和集成优化学习的基因表达数据分类算法[J]. 计算机应用研究, 2018, 36(10): 156-1164

[10] Kamali T, Boostani R. Paraei H. A multi-classifier approach to MUAP classification for diagnosis of neuromuscular disorders [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2014, 22(1): 191-200

[11] Tian J, Xing Y, Yao S, et al. Comparison of Landsat-TM Image Forest Type Classification Based on Cellular Automata and BP Neural Network Algorithm[J]. Scientia Silvae Sinicae, 2017, 2(1): 17-20

[12] Sun H, Wang R, Geng J. Thermal System Modeling Based on Entropy and BP Neural Network [J]. Journal of System Simulation, 2017: 32 (12): 220-226

[13] Li Z, Wang F, Sun T, et al. A constrained optimization method based on BP neural network[J]. Neural Computing & Applications, 2016, 7 (11): 1-9

[14] Liu B, Wang L, Jin Y H. An effective PSO-based memetic algorithm for flow shop scheduling.[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2007, 37(1): 18-27

[15] Lin S, Xu J, Liu S, et al. A robust image watermarking scheme using Arnold transform and BP neural network [J]. Neural Computing & Applications, 2017: 1-16

[16] Liu B, Wang L, Jin Y, et al. An Effective PSO-Based Memetic Algorithm for TSP [J]. Lecture Notes in Control & Information Sciences, 2006, 345: 1151-1156

[17] Garg H. A hybrid PSO-GA algorithm for constrained optimization problems [J]. Applied Mathematics & Computation, 2016, 274(11): 292-305

[18] Yu R, An X, Bo J, et al. Particle classification optimization-based BP network for telecommunication customer churn prediction[J]. Neural Computing & Applications, 2018, 29(3): 707-720

[19] Cao J, Chen L, Wang M, et al. A Parallel Adaboost-Backpropagation Neural Network for Massive Image Dataset Classification[J]. Sci Rep, 2016, 6(1): 3820-3821

[20] Lin Y C, Chen D D, Chen M S, et al. A precise BP neural network-based online model predictive control strategy for die forging hydraulic press machine[J]. Neural Computing & Applications, 2016: 1-12

[21] Zhang S, Lv J, Yuan X, et al. BP Neural Network with Genetic Algorithm Optimization for Prediction of Geo-Stress State from Wellbore Pressures [J]. International Journal of Computational Intelligence & Applications, 2016, 15(03): 80-85

[22] Zhao Z, Xu Q, Jia M. Improved shuffled frog leaping algorithm-based BP neural network and its application in bearing early fault diagnosis [J]. Neural Computing & Applications, 2016, 27(2): 375-385

[23] Wang R, Zha B. A Research on the Optimal Design of BP Neural Network based on Improved GEP [J]. International Journal of Pattern Recognition & Artificial Intelligence, 2018: S0218001419590079

[24] Xie F, Fan H, Li Y, et al. Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model [J]. IEEE Transactions on Medical Imaging, 2017, 36(3): 849-858

[25] Aburomman A A, Reaz M B I. A novel SVM-kNN-PSO ensemble method for intrusion detection system [J]. Applied Soft Computing, 2016, 38(C): 360-372

[26] Ou X, Yan P, Wei H, et al. Adaptive GMM and BP Neural Network Hybrid Method for Moving Objects Detection in Complex Scenes[J]. International Journal of Pattern Recognition & Artificial Intelligence, 2018, 32(5): 1102-1108

[27] Singh B K, Verma K, Thoke A S. Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images[J]. Expert Systems with Applications, 2016, 66: 114-123

[28] Novizon, Abdul-Malek Z. Neutral Networks for Fault Classification: Comparison between Feed-Forward Back-Propagation, RBF and LVQ Neural Network[J]. Applied Mechanics & Materials, 2016, 818: 96-100

[29] Fan Y, Li F W B. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining[J]. Computers & Education, 2018, 123: 97-108

[30] Khatami A, Mirghasemi S, Khosravi A, et al. A new PSO-based approach to fire flame detection using K-Medoids clustering [J]. Expert Systems with Applications, 2017, 68(3): 69-80

[31] Li Z, Wang F, Bing X, et al. Prediction of stock prices based on LM-BP neural network and the estimation of overfitting point by RDCI [J]. Neural Computing & Applications, 2017(9): 1-20

(上接第 995 页)

[27] Flogel U, Ding Z, Hardung H, et al. In vivo monitoring of inflammation after cardiac and cerebral ischemia by fluorine magnetic resonance imaging[J]. Circulation, 2008, 118(2): 140-148

[28] Shin SH, Kadayakkara DK, et al. In Vivo 19F MR Imaging Cell Tracking of Inflammatory Macrophages and Site-specific Development of Colitis-associated Dysplasia[J]. Radiology, 2016: 152387

[29] Hertlein T, Sturm V, Lorenz U, et al. Bioluminescence and ^{19}F magnetic resonance imaging visualize the efficacy of lysostaphin alone and in combination with oxacillin against Staphylococcus aureus in murine thigh and catheter-associated infection models[J]. Antimicrob Agents Chemother, 2014, 58(3): 1630-1638

[30] Zhong J, Narsinh K, Morel PA, et al. In Vivo Quantification of Inflammation in Experimental Autoimmune Encephalomyelitis Rats Using Fluorine-19 Magnetic Resonance Imaging Reveals Immune Cell Recruitment outside the Nervous System[J]. PLoS One, 2015, 10 (10): e0140238

[31] Bonner F, Merx MW, Klingel K, et al. Monocyte imaging after myocardial infarction with 19F MRI at 3 T: a pilot study in explanted porcine hearts [J]. Eur Heart J Cardiovasc Imaging, 2015, 16 (6): 612-620

[32] Temme S, Grapentin C, Quast C, et al. Noninvasive Imaging of Early Venous Thrombosis by 19F Magnetic Resonance Imaging with Targeted Perfluorocarbon Nanoemulsions[J]. Circulation, 2015, 131(16): 1405-1414