第48卷第2期 中国科学技术大学学报 Vol.48,No.2
2018年2月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Feb. 2018

文章编号:0253-2778(2018)02-0154-07

基于超限学习机的快速癌症检测方法

林宇鹏[1],谢智歌[2,3],徐 凯[3],陈 飞[4],刘利刚[1]

(1.中国科学技术大学数学科学学院,安徽合肥 230026;2.中国人民解放军71939部队,山东济南 250300;
3.国防科技大学计算机学院,湖南长沙 410073;4.中南大学湘雅二医院,湖南长沙 410073)

摘要:利用基于局部感受野的超限学习机(ELM-LRF)算法从给定的基因表达数据中提取有效的特征来进行癌症检测与分类.首先使用主成分分析(PCA)方法对原数据进行适当预处理,减少数据中存在的冗余,然后构建特定的特征映射,将得到的数据映射到相应特征空间中去,最后对得到的数据特征进行训练学习,得到最终训练好的特征提取模型.实验表明,ELM-LRF的学习效率更高,取得的癌症检测效果比以往方法更好.

关键词:超限学习机;特征学习;机器学习;分类;癌症检测

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.02.010

引用格式:林宇鹏,谢智歌,徐凯,等.基于超限学习机的快速癌症检测方法[J].中国科学技术大学学报,2018,48(2):154-160.

LIN Yupeng, XIE Zhige, XU Kai, et al. Fast cancer diagnosis based on extreme learning machine[J]. Journal of University of Science and Technology of China, 2018,48(2):154-160.

Fast cancer diagnosis based on extreme learning machine

LIN Yupeng[1], XIE Zhige[2,3], XU Kai[3], CHEN Fei[4], LIU Ligang[1]

(1. *School of Mathematical Sciences, University of Science and Technology of China, Hefei* 230026, *China*;
2. *PLA* 71939 *Unit, Jinan* 250300, *China*; 3. *School of Computer, National University of Defense Technology, Changsha* 410073, *China*;
4. *The Second Xiangya Hospital, Central South University, Changsha* 410073, *China*)

Abstract: The local receptive fields based extreme learning machine (ELM-LRF) method was utilized to learn the effective features from the acquired gene expression data to help enhance cancer diagnosis and classification. Firstly, the principal component analysis (PCA) method was implemented to process the dataset. Secondly, the features mapping to map our dataset were constructed to the specific feature space. Finally, the features to train the learning model were used to get the final ELM feature extraction model. The experiment shows that the proposed algorithm outperforms almost all the existing methods in accuracy and efficiency.

Key words: extreme learning machine (ELM); feature learning; machine learning; classification; cancer diagnosis

收稿日期:2016-12-12;**修回日期**:2017-06-05

基金项目:国家自然科学基金(61672482, 11526212, 61572507, 61540065),中国科学院"百人计划"资助.

作者简介:林宇鹏,男,1992年生,硕士.研究方向:计算机图形学,机器学习.E-mail: lypeng@mail.ustc.edu.cn

通讯作者:刘利刚,博士/教授.E-mail: lgliu@ustc.edu.cn

0　引言

进入 21 世纪，随着现代社会的发展，人们受到环境污染与不良生活习惯等因素的影响，癌症的发病率持续攀升.如何快速准确地对癌症进行诊断已经成为一个急需解决的问题.癌症的成因往往是由于调控细胞生长的基因发生突变或损伤导致基因的表达出现异常，致使细胞不正常生长而出现癌变.传统的癌症诊断方法一般是医生通过肉眼对患者的细胞进行观察来判断患者的患病程度，诊断的准确率在很大程度上依赖于医生的经验.近年来，基因芯片(微阵列)技术得到了快速的发展，该技术可以同时对大量的基因进行检测，并得到相应基因的基因表达谱.该基因表达谱记录了在微阵列实验中得到的相关基因的表达水平.通过比较正常细胞与患病细胞基因表达水平的异同点，可以研究其关联性，从而对相关疾病的病理有更深入的理解[1-2].如今，基因表达谱已经广泛应用于癌症的诊断与分类[2-5].

由于通过基因芯片技术得到的基因数据量较大，很难进行人工分析，所以需要借助于自动化方法(例如机器学习)来进行分析与诊断.到目前为止，已经有许多机器学习的判别技术被应用于该领域.利用这些机器学习技术，可以对基因表达数据进行分析检测，从而将细胞分成正常细胞与癌变细胞两类，来达到对癌症进行诊断的目的.基因数据集一般具有高维度、样本数少的特点，这些特点给现有的机器学习技术带来了极大的困扰.因此如何有效地对基因表达数据进行分析，仍是一个巨大的挑战.

文献[6]将 3 种有监督机器学习技术：C4.5 决策树，Bagging 以及 Adaboost 决策树，应用于癌症检测中，并取得了一定的成效.由于在基因检测过程中，基因检测数据往往是高维且不平衡的，因此文献[7]使用了如下 4 种类别不平衡的分类器：对角线性判别分析(diagonal linear discriminant analysis)、随机森林(random forests)、支持向量机(support vector machine，SVM)以及阈值调整的支持向量机，但是这些方法仅仅着重于数据类别的不平衡性.文献[8]提出了一种特征选取的方法，该方法首先将基因数据集分成多个小的子集，然后从子集中选取出包含更多信息的基因并对其进行合并.该过程可以不断重复地进行直到最终仅留下一个子集，即可筛选出有效的特征.而文献[9]提出了一个用于基因微阵列分类的方法，作者利用各种不同的特征降维方法的组合来达到对癌症进行检测的目的.在文献[10]中，作者针对癌症分类问题提出了一个有效的系统.然而，该系统需要考虑结构与参数的设置问题，导致使用极为复杂.

本文使用基于局部感受野的超限学习机(ELM-LRF)算法来解决癌症检测问题.尽管癌症检测问题中的基因表达数据往往是病态的(样本数比较少但数据维数非常高)，但本文的方法依然能取得良好的结果.

1　相关理论与工作

1.1　基因表达数据

利用基因芯片技术获得的基因表达数据衡量了给定细胞中的目标基因的表达水平[10].对 n 个样本进行微阵列实验，各样本都进行一次基因检测，每次基因检测都能得到该样本中的 m 个基因的基因表达水平.这些从实验中获得的数据可以用一个基因表达矩阵来表示.n 个样本的实验数据可以排列成一个 $m \times n$ 的基因表达矩阵，其中矩阵的每一列都记录了一个特定的样本，由 m 个基因来表示，而矩阵的每一行记录了每个基因在 n 个样本里的活跃水平.图 1 展示了基因表达矩阵数值的热图.通过观察基因表达数据并进行分析，可以得到正常细胞与异常细胞之间的异同点，从而进行癌症的检测.然而，由于基因数目相对较大而样本数据相对较小，亦即基因数据是不平衡的，因此生成的基因矩阵往往是狭长的，给后续的分类工作带来极大的困难.

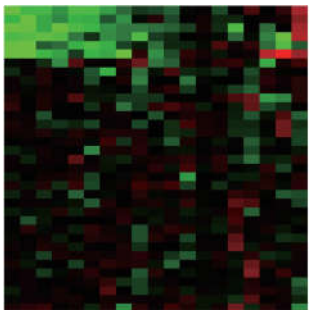

图 1　基因表达数据的热点图

Fig.1　Heat map of gene expression data

1.2　超限学习机(ELM)

超限学习机(extreme learning machine，ELM)是由黄广斌等[11-13]提出的一种单隐层前馈神经网络(SLFNs).在大多数的学习方法理论中，网络的隐层节点在训练阶段都将不断地进行迭代调整.而在 ELM 算法中，隐层节点参数是以任意的连续概率分

布来随机产生的，并不会迭代调整，因此，隐层节点参数的产生是独立于应用存在的[11,19]．ELM 方法可以有效地避免局部极小解的出现，同时拥有全局逼近性以及良好的泛化性．

1.2.1 基本超限学习机

图 2 展示了基本超限学习机的网络结构．给定 n 个输入数据样本 x_i，超限学习机的输出函数为

$$f(x)=\sum_{i=1}^{p}\beta_i h_i(x)=h(x)\beta \tag{1}$$

式中，$\beta=[\beta_1,\cdots,\beta_p]^{\mathrm{T}}\in R^{p\times t}$ 表示隐藏层与输出层之间的输出权重的向量．$h(x)=[h_1(x),\cdots,h_p(x)]\in R^{n\times p}$ 表示隐藏层的输出向量．h 是特征映射函数，其作用是将 R^d 的输入数据映射到 R^p 的特征空间中去．对于隐藏层中不同的隐层节点，可以使用不同种类的激活函数．对于隐藏层节点，还可以选取分段非线性连续函数作为激活函数，例如：Sigmoid 函数，高斯函数或者傅里叶函数．

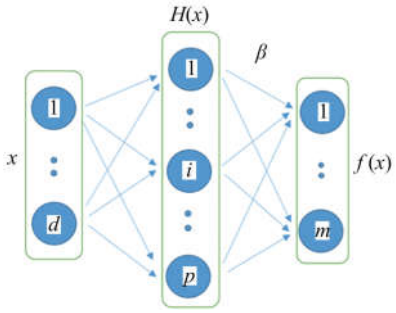

图 2 基本超限学习机

Fig.2 Basic ELM

在经过特征映射的步骤之后，可以按如下方式定义能量函数来最小化输出权重的范数与训练误差的加权和：

$$\omega\|H\beta-T\|_2^2+\|\beta\|_2^2. \tag{2}$$

式中，参数 T 表示存储训练数据的目标矩阵，参数 ω 的作用是用来调整这两项间的权重，可以由用户根据特定应用进行自行调整．通过在能量函数中增加正则项（权重的范数），可以在减小模型训练误差的同时令训练得到的模型复杂度尽量小，来减少过拟合现象的发生．以上问题拥有如下所示的闭合形式解．

$$\beta=\begin{cases}H^{\mathrm{T}}\left(\dfrac{I}{C}+HH^{\mathrm{T}}\right)^{-1}T,\text{if } N\leqslant p;\\ \left(\dfrac{I}{C}+HH^{\mathrm{T}}\right)^{-1}H^{\mathrm{T}}T,\text{if } N>p\end{cases} \tag{3}$$

1.2.2 ELM-LRF

自超限学习机方法（ELM）提出以来，全连接形式的超限学习机方法已经得到了广泛的研究探讨，然而，很少有研究涉及局部连接形式的超限学习机方法．最近，黄广斌等[19]提出了一种新型的网络结构：基于局部感受野的超限学习机（ELM-LRF）．ELM-LRF 拥有以下两个重要的特征：①ELM-LRF 在网络节点间使用了局部连接的方式，并在超限学习机的网络结构里引入了局部感受野的概念．②ELM-LRF 网络隐层的每个节点都可以看作是多个节点的组合或者一个子网络．

2 算法流程

本部分将详细介绍如何对给定的基因表达数据进行特征学习来对模型进行训练，训练得到的模型可以对给定的输入样本（基因表达数据）进行分类，预测出样本的患病程度．图 3 展示了本文方法的具体流程．在数据预处理阶段，首先将使用降维技术对基因表达数据进行预处理，从而到降维的目的，方便后续操作．在特征映射阶段，将会构建特征映射来将数据映射到一个特定的特征空间．在 ELM 学习阶段，将利用闭合形式的公式解来计算出模型的输出权重，最终结束训练阶段，得到训练好的超限学习机特征提取模型．

图 3 算法流程

Fig.3 Algorithm flow

2.1 数据预处理

对于得到的 n 个样本的基因数据，每个样本包含有 m 个基因的基因表达水平，可以用一个 $m\times n$ 的基因表达矩阵来进行表示．由于各类不可避免的因素的影响，基因芯片技术得到的基因表达数据往往维数 m 较高，样本数量 n 偏少，并带有噪声，因而会对后续的分类工作造成影响．为了方便于后续步骤的进行，本文方法将对数据进行一系列的预处理．数据预处理阶段，将使用一些特征提取的方法来对数据进行降维处理．通过使用特定的数据降维算法，可以减少数据的冗余，使后面的训练过程更加顺利地进行．本文方法中使用成分分析方法（PCA）来进行数据的降维处理．PCA 降维技术可以有效地降低数据维数，同时又能尽可能减少有用信息的缺失．

2.2 特征映射

在超限学习机网络中，节点间的连接数目往往会比节点数目多得多．为了增强网络提取特征的能

力，可以将网络节点的数目设置较多，而节点间的连接将不使用全局连接，而是使用局部连接的方式，从而大大减少网络之中节点间的连接数目，得到一个稀疏网络.对于输入层与隐藏层的节点之间，可以设置全连接或稀疏连接，还可以设置不同密度的连接来满足需求.由于基因数目的巨大以及基因之间关系的稀疏性，本文的网络结构连接使用的是稀疏连接的方式而不选择全连接.在特征映射阶段，实验利用超限学习机技术将输入数据映射至特定的特征空间.超限学习机理论表明，隐层节点并不需要不断地调整，相反，隐层节点的参数可以以任意的概率分布进行随机产生.

卷积(convolution)操作：卷积操作已经被广泛地应用于卷积神经网络(convolutional neural network)中，并取得了很好的效果[15-18].通过使用卷积操作，可以有效地对数据进行特征提取，并具有很好的局部敏感性.在本文方法中，将随机生成输入层与输出层间的节点权重，并借助卷机操作进行特征映射.网络结构中的 k 个输入权重首先将随机生成，然后利用奇异值分解(SVD)的方法进行正交化，最终将得到 k 个特征映射作为卷积核，其中第 k 个卷积核为

$$a_k \in R^{d\times d}$$

给定数据样本 x_i 以及生成的特征映射 a_k，即可以进行卷积操作：

$$\widehat{c_{i,k}} = x_i * a_k \tag{4}$$

池化(pooling)操作：在特征映射过程中得到的特征数目往往比较大，通过采用池化操作可以有效地进行特征压缩来提取主要的特征，减少特征的维数，简化网络的计算复杂度[19].为了便于理解，以下将利用 $\hat{c}$ 对上述步骤中得到的 $\widehat{c_{i,k}}$ 进行表示，对 $\hat{c}$ 进行如下形式的池化操作：

$$h_{x,y} = \sqrt{\sum_{i=x-ej=y-e}^{x+ey+e} c_{i,j}^2} \tag{5}$$

式中，$c_{i,j}$ 表示 $\hat{c}$ 中位于 (i,j) 处的元素，而 $h_{x,y}$ 表示经过池化操作处理后的数据在 (x,y) 位置处的元素，参数 e 代表的是池化的尺寸大小，文中选取的池化操作执行的将是平方和操作.

以上的操作可以在网络各层中重复进行，拥有更多层数的网络将拥有更强的特征提取能力.如图 4 所示，本文中使用含有三层隐藏层的网络结构来进行特征提取.这三层网络是复合网络层，每层均包含了卷积层与池化层.网络中每一层的输出都是网络下一层的输入，从而将数据映射进入一个更高层次的特征空间.通过已经构建训练好的超限学习机的深度网络，即可以对输入的基因表达数据进行有效的特征提取.

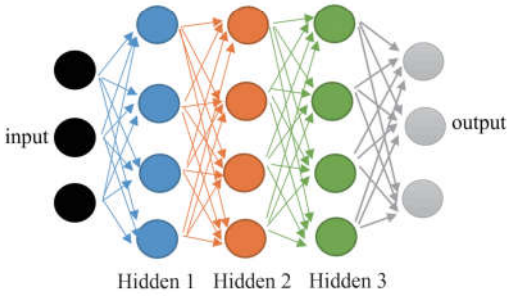

图 4　三层隐藏层的网络结构

Fig.4　Network structure of three hidden layers

2.3　ELM 特征学习

所有的训练数据都有训练标签，代表该样本的患病程度.样本的患病程度由轻到重可以分为 t 类.第 i 个样本若其标签为 j 类，则该标签可用一个 t 维单位列向量来表示，该列向量的第 j 维分量为 1，其他分量为 0.将所有这些标签排列在一起即可得到训练数据的目标矩阵 $T \in R^{N\times t}$.

在经过特征映射阶段，输入的基因表达数据将被映射入一个新的特征空间，最终可以得到一个矩阵数据 $H \in R^{N\times kp}$.根据 ELM 学习理论，可以利用以下形式的闭合形式的公式解来求得网络的输出权值：

$$\beta = \begin{cases} H^{\mathrm{T}}\left(\dfrac{I}{C} + HH^{\mathrm{T}}\right)^{-1} T, \text{if } N \leqslant kp; \\ \left(\dfrac{I}{C} + H^{\mathrm{T}}H\right)^{-1} H^{\mathrm{T}}T, \text{if } N > kp \end{cases} \tag{6}$$

在得到已有的网络模型后，对于一个新的基因表达数据输入，其输出为一个 t 维的列向量.由于患病程度类别为渐变的，该 t 维列向量的最大分量即为该样本的预测类别.

3　实验结果

本文中的实验平台如下：中央处理器(CPU)为 Intel(R) Core(TM) i7-4790K 4.00GHz，安装内存(RAM)容量为 8.00GB，编程环境为 Matlab2016a.

实验将在 12 个基因数据集上进行测试来评估本文方法的性能，测试过程采用交叉验证的方法.在对每个基因数据集进行测试时，该数据集里的样本

先被随机打乱并进行重排，然后这些样本将被平均分成10份.每次测试的过程都将使用其中的一份样本作为测试数据，另外的9份样本数据将合在一起作为训练数据.该过程将重复进行10次，每次都重新选择新的一份数据作为测试集进行测试.直到每一份数据都被当成测试集训练过一次为止.

在经过以上的10次训练测试后，将得到10次实验结果.这10次实验结果可以统计其精确度的平均值与标准差，来作为衡量算法性能好坏的依据.对于不同的方法，都采取该策略得到最终结果参数，之后比较各方法优劣.

表1中列举了实验中的5个基因数据集.可以看出，基因数据集往往基因维数达到上万，而数据样本数却至多只有上百.本文方法在这些数据集上所花费的时间主要集中在网络的训练过程，测试时间基本只占总时间的十分之一，方法需要花费的总时间与数据规模有关.由于本文方法中的网络参数是随机产生的，在训练过程不需要进行迭代调整，因此方法的运行时间量级为秒级.

表2中展示了不同方法在试验的12个数据集上的表现.可以看出，稀疏自动编码器(sparse autoencoder)方法尽管在某些数据集上可以得到良好的预测精确度，但在Leukemia[23]和、Seminoma[26]等数据集上都未能取得很好的效果.而栈式自动编码器(stacked autoencoder)方法尽管可以通过逐层训练得到最佳参数，但在AML[25]和Prostate Cancer[29]等数据集上却表现糟糕.支持向量机(SVM)方法虽在这些实验数据集上均取得了良好的预测精确度，但效果上有所波动，与其他方法相比并非一直是最佳的选择.而本文的方法在11个数据集上都取得了超过0.8的预测精确度，在多数数据集上都取得了0.9以上的预测精确度而且在12个数据集上均取得了最高的预测精确度.

表1 方法运行效率

Tab.1 Operational efficiency of method

数据集	数据规模	训练耗时/s	测试耗时/s
AML[20]	54613 * 183	3.5304	0.3721
Adenocarcinoma[21]	34749 * 28	0.286	0.0255
Leukemia[23]	54675 * 125	2.407	0.2277
Prostate Cancer[29]	12600 * 34	0.1274	0.0146
Leukemia[30]	54613 * 230	4.4307	0.4197

表2 实验结果

Tab.2 Experimental results

数据集	sparse autoencoder	stacked autoencoder with fine tunning	PCA+Softmax/SVM (with Gaussian kernel)	本文方法
AML[20]	0.7463±0.062	0.9515±0.047	0.9404±0.03	0.9675±0.0265
Adenocarcinoma[21]	0.9167±0.18	0.8750±0.16	0.9333±0.14	0.9333±0.1333
Breast Cancer[22]	0.8667±0.219	0.8333±0.272	0.850±0.241	0.9000±0.2000
Leukemia[23]	0.5609±0.024	0.9365±0.049	0.9295±0.09	0.9596±0.0404
Leukemia[24]	0.4676±0.23	0.3371±0.038	0.4633±0.180	0.5667±0.1856
AML[25]	0.8167±0.298	0.550±0.137	0.7333±0.196	0.8667±0.2211
Seminoma[26]	0.350±0.337	0.8000±0.258	0.7667±0.251	0.8000±0.2449
Ovarian Cancer[27]	0.7545±0.135	0.9900±0.032	1±0	1±0
Colon Cancer[31]	0.6667±0.0	0.8333±0.176	0.8333±0.236	0.8667±0.1633
Medulloblastom[28]	0.6667±0.0	0.7667±0.225	0.7667±0.274	0.8000±0.3145
Prostate Cancer[29]	0.975+0.079	0.7333±0.102	0.94167±0.124	0.9750±0.0750
Leukemia[30]	0.6918+0.108	0.9126±0.055	0.9039±0.081	0.9609±0.0410

4 结论

本文利用基于局部感受野的超限学习机方法来解决癌症分析与检测问题.该问题与其他问题的不同点在于样本维数较高而数量较少,给问题的解决带来了极大的障碍.与传统的机器学习方法不同的是,本文方法中的网络结构参数是随机产生的,从而得到的网络模型是独立于特定应用的.实验结果表明,本文方法在精确度方面超过了以往的其他方法,拥有非常优秀的特征学习能力,同时由于该方法的网络参数不需要进行调整,因此学习效率更高,速度更快,在一些有性能需求的实际应用中将有重大的意义.在实验的过程中发现,本文方法尽管在大多数数据集上都达到了很高的精确度,但在某些数据集上,精确度仍有着一定的提升空间.因此探究本文方法在这些数据集上的表现形式,并进一步研究并优化本文方法的检测准确度,将是我们未来的研究方向.

致谢　作者感谢黄广斌教授对本文的技术支持和讨论,同时感谢 Rasool Fakoor 为本文提供数据集.

参考文献(References)

[1] WANG Z, PALADE V. Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis[J]. BMC Genomics, 2011, 12(Suppl 2): S5.

[2] CHEN H, ZHAO H, SHEN J, et al. Supervised machine learning model for high dimensional gene data in colon cancer detection[C]// 2015 IEEE International Congress on Big Data. IEEE, 2015: 134-141.

[3] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531-537.

[4] HONG J H, CHO S B. Gene boosting for cancer classification based on gene expression profiles[J]. Pattern Recognition, 2009, 42(9): 1761-1767.

[5] ALON U, BARKAI N, NOTTERMAN D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences, 1999, 96(12): 6745-6750.

[6] TAN A C, GILBERT D. Ensemble machine learning on gene expression data for cancer classification[J]. Applied Bioinformatics, 2003, 2(3 Suppl): S75-83.

[7] LIN W J, CHEN J J. Class-imbalanced classifiers for high-dimensional data[J]. Briefings in Bioinformatics, 2013, 14(1): 13-26.

[8] SHARMA A, IMOTO S, MIYANO S. A top-r feature selection algorithm for microarray gene expression data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2012, 9(3): 754-764.

[9] NANNI L, BRAHNAM S, LUMINI A. Combining multiple approaches for gene microarray classification[J]. Bioinformatics, 2012, 28(8): 1151-1157.

[10] FAKOOR R, LADHAK F, NAZI A, ET AL. Using deep learning to enhance cancer diagnosis and classification[C]// Proceedings of the International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP, 2013, volume 28.

[11] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: A new learning scheme of feedforward neural networks[C]// 2004 IEEE International Joint Conference on Neural Networks. IEEE, 2004, 2: 985-990.

[12] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: Theory and applications[J]. Neurocomputing, 2006, 70(1): 489-501.

[13] HUANG G B, CHEN L, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Transactions on Neural Networks, 2006, 17(4): 879-892.

[14] HUANG G B, BAI Z, KASUN LL C, et al. Local receptive fields based extreme learning machine[J]. IEEE Computational Intelligence Magazine, 2015, 10(2): 18-29.

[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2012:1097-1105.

[16] SZEGEDY C, LIU W, JIAY, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016:770-778.

[18] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[19] SCHERER D, MÜLLER A, BEHNKE S. Evaluation of pooling operations in convolutional architectures for object recognition[C]// International Conference on Artificial Neural Networks. Berlin/ Heidelberg: Springer-Verlag, 2010:92-101.

[20] MILLS K I, KOHLMANN A, WILLIAMS M, et al. Microarray classification of myelodysplastic syndrome (MDS) identifies subgroups with distinct clinical outcomes and identifies patients with high risk of AML transformation[J]. Blood, 2009(114): 1063-1072.

[21] FUJIWARA T, HIRAMATSU M, ISAGAWA T, et al. ASCL1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis[J]. Lung Cancer, 2012, 75(1): 119-125.

[22] WOODWARD W A, KRISHNAMURTHY S, YAMAUCHI H, et al. Genomic and expression analysis of microdissected inflammatory breast cancer [J]. Breast Cancer Research & Treatment, 2013, 138 (3): 761-772.

[23] KLEIN H U, RUCKERT C, KOHLMANN A, et al. Quantitative comparison of microarray experiments with published leukemia related gene expression signatures[J]. BMC Bioinformatics, 2009, 10(1): 422.

[24] Wl C M Y, PUI C H, DOWNING J R, et al. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells[J]. Nature Genetics, 2003, 34(1): 85-90.

[25] YAGI T, MORIMOTO A, EGUCHI M, et al. Identification of a gene expression signature associated with pediatric AML prognosis[J]. Blood, 2003, 102 (5): 1849.

[26] GASHAW I, GRÜMMER R, KLEIN-HITPASS L, et al. Gene signatures of testicular seminoma with emphasis on expression of ets variant gene 4[J]. Cellular & Molecular Life Sciences Cmls, 2005, 62 (19-20): 2359-2368.

[27] PETRICOIN E F, ARDEKANI A M, HITT B A, et al. Use of proteomic patterns in serum to identify ovarian cancer[J]. Lancet, 2002, 359(9306): 572-577.

[28] POMEROY S L, TAMAYO P, GAASENBEEK M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression[J]. Nature, 2002, 415(6870): 436.

[29] SINGH D, FEBBO P G, ROSS K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002, 1(2): 203-209.

[30] VERHAAK R G, WOUTERS B J, ERPELINCK C A, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling [J]. Haematologica, 2009, 94(1): 131-134.

[31] ALON U, BARKAI N, NOTTERMAN D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(12): 6745-6750.