演化超网络在多类型癌症分子分型中的应用

王 进*① 丁 凌① 孙开伟① 李钟浩②
①(计算智能重庆市重点实验室(重庆邮电大学) 重庆 400065)
②(韩国仁荷大学信息与通信工程系 仁川 402-751)

摘 要：该文提出一种用于多类型癌症分子分型的演化超网络模式识别方法。首先采用“一对多”方法，将一个多类分型问题转化为多个二类分型问题；然后利用信噪比方法对 DNA 微阵列数据进行信息基因选择；经过超网络对训练集的演化学习，构造一系列二类分类器并进行集成，最终构建一个多类型癌症分型系统并对待测样本进行分类。对急性白血病、儿童小圆蓝细胞肿瘤和 GCM 数据集实验结果表明：演化超网络留一交叉验证(LOOCV)识别率分别为：98.61%，100%和 85.35%。演化超网络有利于挖掘癌症相关基因，具有良好的学习结果可读性。

关键词： 模式识别；机器学习；演化超网络；微阵列；癌症多类型分类
中图分类号：TP18; TP39 **文献标识码**： A **文章编号**：1009-5896(2013)10-2425-07
DOI: 10.3724/SP.J.1146.2012.01171

Applying Evolutionary Hypernetworks for Multiclass Molecular Classification of Cancer

Wang Jin① Ding Ling① Sun Kai-wei① Lee Chong ho②
①(*Chongqing Key Laboratory of Computational Intelligence* (*Chongqing University of Posts and Telecommunications*), *Chongqing* 400065, *China*)
②(*Department of Information and Communication Engineering*, *Inha University*, *Incheon* 402-751, *Republic of Korea*)

Abstract: This paper presents a pattern recognition method for multiclass cancer molecular classification using evolutionary hypernetworks. A multiclass classification issue is decomposed into a set of binary classification issues by One-Versus-All (OVA) approach. The signal-to-noise ratio method is employed for informative genes selection from the DNA microarray. A series of binary classifiers are evolved and used to build a final ensemble classifier for multiclass classification through an evolutionary learning procedure of the hypernetwork. The test sample is classified by using the ensemble classifier. Experimental results show that the Leave One Out Cross Validation (LOOCV) accuracy of the acute leukemia dataset, the small, round blue cell tumor dataset, and the GCM dataset is 98.61%, 100% and 85.35%, respectively. The evolutionary hypernetworks is fit to find cancer-related genes and has a good readability of the learned results.

Key words: Pattern recognition; Machine learning; Evolutionary hypernetworks; Microarray; Cancer molecular multiclass classification

1 引言

癌症治疗面临的重大挑战是如何针对病原上各自独立的癌症类型制定具体的治疗方案，以在达到最大疗效的同时降低药物的副作用。因此癌症的准确分型是癌症治疗的关键。一直以来，癌症诊断主要是基于肿瘤的表现形态。但是这种诊断方式具有很大的局限性，因为具有相似组织病理学表现的肿瘤可能表现出不同的临床发展过程[1]。DNA 微阵列技术的出现为从分子水平研究疾病的发病机理和临床诊断提供了强有力的手段[2]。基于 DNA 微阵列的海量基因表达谱数据，为寻找基因之间表达调控的复杂关系网络以及研究功能基因组和癌症检测提供了依据。

DNA 微阵列数据具有样本数量少、高维度、高噪音、高相关冗余、数据分布不均衡等特点。同时 DNA 微阵列数据包含着不同基因之间庞大而复杂的并行交互作用，这些基因间的交互作用对研究癌症的复杂发展机制有着重要意义。因此，在消除微阵列数据中冗余信息的同时，需要相应的计算方法分析数据，并从中发掘与癌症相关的基因以及基因间的相互作用。传统的模式识别方法已在癌症分子

2012-09-10 收到，2013-07-01 改回
国家自然科学基金(61203308, 61075019)，教育部留学回国人员科研启动基金(教外司留[2010]1174 号)和国家大学生创新创业训练计划(201210617003)资助课题
*通信作者：王进 wangjin_liips@yahoo.com.cn

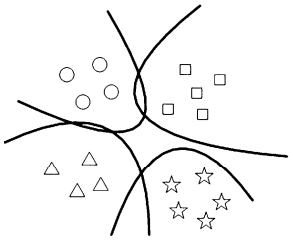
分型应用中表现出良好的分类性能[1–6]，但是这些方法大多没有考虑发掘基因间内在的相互作用，同时也普遍存在着数据处理能力低以及学习结果不易分析等局限。

超网络(hypernetworks)是受到生物分子网络的启发而建立的一种基于超图(hypergraph)的认知学习模型。组成超网络的超边包含多个特征，超边表达特征变量之间的高阶关联性。通过演化学习，超网络可以有效获取与模式识别相关的关键特征，从而表达复杂数据的内在结构和相互之间的关系，也可以表示对象的不同属性以及属性和对象之间的关联程度[7]，因此非常适用于解决基于 DNA 微阵列数据的癌症分子分型问题，同时有效挖掘癌症相关基因和基因间的相互作用。在癌症分子分型应用中，Park 等人[8,9]已成功利用超网络模型实现了对癌症样本与正常样本分型和对癌症基因间相互作用的有效挖掘。然而，针对临床诊断中常见的多类癌症分类问题[10,11]，在超网络领域还缺乏相应的探索。

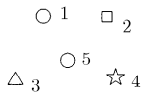
本文提出了一种基于演化超网络的多类癌症分子分型方法。针对急性白血病 (acute leukemia) 数据集、儿童小圆蓝细胞肿瘤(Small, Round Blue Cell Tumor, SRBCT)数据集以及 GCM 数据集，首先根据一对多(One-Versus-All, OVA)方法把单个多类分型问题转化为一系列二类分型问题，然后采用信噪比(Signal-to-Noise Ratio, SNR)方法进行特征基因选择，再通过超网络对训练集的学习建立起癌症分子分型系统，实现对待测样本的分类。通过与其它癌症分子分型方法的对比实验，验证了基于演化超网络的多类型癌症分子分型方法的可行性和有效性。

2 多类型分类

对于多类型癌症分型问题，由于算法需要构建更多的分离界限，从本质上来说要比二类分型困难。同时，许多基于排序的基因选择方法，如果直接应用到多类分型，其分型精度都有所降低[10]。由此，一般是把多类分型问题转化为二类分型问题再进行解决。其中，最常见的有一对多和成对(All-Pairs, AP)两种方法[11]。将多类型癌症分子分型问题转化成二类分型问题后，可以把很多较好的二类分类算法直接应用到多分类问题中。采用 AP 方法，随着癌症类数的增加，分类器数量也会急剧增加，会导致决策速度慢，而且由于训练样本少，容易造成不可分区域。由此，本文采用相对简单的一对多的变换方法，构造一系列的二类分类器，然后把这些二类分类器进行集成，从而达到多类型分型的目的。每一个分类器都把其中的一类同余下的混合类分开。对 k(k 大于 2)种癌症类型，可以构造 k 个独立的二类分类器。图 1 给出了采用 OVA 的方法解决多类型分类的具体过程。图 1(a)是训练集样本，表示 4 组二类分类问题被学习。第 1 组二类分类是把圆形类型样本从所有样本中区分出来；第 2 组把正方形类型样本从所有样本中区分出来；第 3 组是区分三角形类型的样本；第 4 组是区分五角星类型的样本。图 1(b)表示等待分类的 5 个新样本。图 1(c)展示的是 OVA 分类方法的编码表。表格第 1 行表示 4 种分类器，第 1 列表示类型标签，此表列出了不同类型在不同分类器中的理想输出。图 1(d)显示了 5 个待分类样本在训练好的 4 种分类器中的实际输出与最终判定类型。



(a)训练集样本



(b)待测试新样本

类型	○	□	△	☆
○	+1	−1	−1	−1
□	−1	+1	−1	−1
△	−1	−1	+1	−1
☆	−1	−1	−1	+1

(c) OVA编码表

	○	□	△	☆	类型
1	0.7	−0.2	−3	−4	○
2	−0.5	0.5	−1	−3	□
3	−6	−0.3	0.7	−1	△
4	−2	−1	−0.2	0.4	☆
5	0.8	−3	−4	−0.7	○

(d)测试样本输出表

图 1 一对多分类方法

3 信息基因选择

假定训练集中有 $M\times N$ 个基因表达值，其中 M 是训练集样本数，N 为每个样本包含的基因数。则第 i 个基因表示为 $\boldsymbol{g}_i$：

$$\boldsymbol{g}_i=\left(e_1,e_2,\cdots,e_M\right) \tag{1}$$

在式(1)中，$i=1,2,\cdots,N$，$e_m(m=1,2,\cdots,M)$ 表示第 i 个基因在第 m 个样本中的表达水平。

3.1 信噪比信息基因选择方法

DNA 微阵列数据有着样本数量少、高维度、高噪音、高相关等特征，进行信息基因选择可以有效减少数据噪音和降低计算复杂度。近年来，研究人员已经提出了大量的特征基因选择方法，主要可以分为 3 类：Filter 方法，Wrapper 方法和 Embedded 方法[12]。其中 Filter 方法根据基因本身的特性来选择基因，不依赖分类器，且计算方法比较简单[13]。本文采用基于 Filter 的信噪比方法[14]选择与分类相

关的 n 个基因。公式为

$$P(\boldsymbol{g})=\frac{\mu_1(\boldsymbol{g})-\mu_2(\boldsymbol{g})}{\sigma_1(\boldsymbol{g})+\sigma_2(\boldsymbol{g})} \tag{2}$$

其中 $\mu_1(\boldsymbol{g})$，$\mu_2(\boldsymbol{g})$ 分别表示基因在类型 1 和类型 2 中的平均值；$\sigma_1(\boldsymbol{g})$，$\sigma_2(\boldsymbol{g})$ 分别表示基因在类型 1 和类型 2 中的标准差。式(2)中 $P(\boldsymbol{g})$ 绝对值越大，则基因对分类的相关性越好。根据求得的 $P(\boldsymbol{g})$ 值，我们从正值和负值中分别选择绝对值较大的 $n/2$ 个基因。针对每一个二类分类器，分别用信噪比进行信息基因的选择，k 种类型的癌症可以选出 k 个信息基因子集。

3.2 信息基因归一化处理

从大量的 DNA 微阵列数据中选取了信息基因之后，必须进行归一化处理。采用式(3)进行处理：

$$\mathrm{nor}_n=\frac{e_m-g_{\mathrm{avg}_i}}{g_{\mathrm{SD}_i}} \tag{3}$$

式(3)中，e_m 即是所选中的基因 $\boldsymbol{g}_i$ 在样本 m 中的表达水平，g_{avg_i} 表示选中的基因 $\boldsymbol{g}_i$ 在训练集所有样本中的平均表达水平，g_{SD_i} 表示选中的基因 $\boldsymbol{g}_i$ 在训练集所有样本中的标准差。求出归一值之后，进行二值化处理，即如果 $\mathrm{nor}_n\geq 0$，把该值定义为 1；否则定义为 0。

4　演化超网络多类型癌症分型系统

4.1 超网络模型

超网络是一种基于超图的认知学习模型[15,16]，是由大量超边(hyperedge)组成的任意超图结构。超图 $G=(V,E)$ 是一个无向图。该无向图中，顶点集 $V=\{v_1,v_2,\cdots,v_n\}$，超边集 $E=\{e_1,e_2,\cdots,e_m\}$，任意一条超边 e_i 包含非零个顶点。图 2 表示一个包含 7 个顶点 4 条超边的超图，图中 $V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7\}$，$E=\{e_1=\{v_1,v_2,v_3\},e_2=\{v_2,v_3\},e_3=\{v_3,v_5,v_6\},e_4=\{v_4,v_7\}\}$。

超网络可被定义为一个三元组 $H=\{V,E,W\}$，其中 V, E, W 分别表示超网络的顶点集合、超边集合以及超边权值的集合。超边所连接的顶点个数称为超边的阶数(order)。超边的权值就是边的重复数目。超网络经过演化学习，可以作为一个分类器。超边连接的顶点被看做决策属性，超边被看做决策规则，超边可以复制，超边对分类的重要性取决于超边权值的大小。

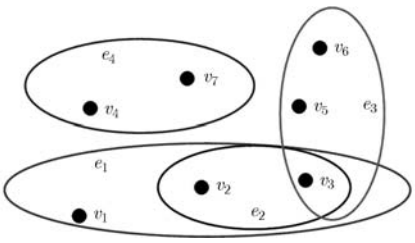


图 2 一个简单的超图模型

给定一个 k 维的训练集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，类型 $Y=\{0,1\}$，样本 $\boldsymbol{x_i}=(x_{i1},x_{i2},\cdots,x_{ik},y_i)$，其中 $x_{ij}\in\{0,1\}$ 表示样本所包含的特征。$y_i\in Y$，是类别标签，表示样本所属类型。从样本 $\boldsymbol{x}_i$ 中随机选取任意个特征 x_{ij}，和 $\boldsymbol{x}_i$ 的标签 y_i 一起组成一条超边；多条随机生成超边组成的超边集合即为初始超网络。经过对该超边集合进行演化学习，超网络就可以作为一个分类器，其分类过程如下：

(1)对训练集 $\boldsymbol{X}$ 进行演化学习；

(2)记录所有与待分类样本 $\boldsymbol{z}$ 匹配的超边；

(3)分别计算样本 $\boldsymbol{z}$ 属于不同类型的条件概率 $P(Y/\boldsymbol{z})$；

(4)根据最大条件概率判定样本 $\boldsymbol{z}$ 的类型。

求样本属于某一类型的条件概率利用式(4)~式(6)计算：

$$P\left(Y/\boldsymbol{z}\right)=\frac{P(\boldsymbol{z},Y)}{P(\boldsymbol{z})} \tag{4}$$

$$P(\boldsymbol{z},Y)\approx\frac{1}{|L|}\sum_{i=1}^{|L|}f_i^{(k)}\left(z_{i1},z_{i2},\cdots,z_{ik},Y\right) \tag{5}$$

$$P(\boldsymbol{z})\approx\frac{|M|}{|L|} \tag{6}$$

其中，其中 $|L|$ 表示超边总数，$|M|$ 表示测试样本 $\boldsymbol{z}$ 中与超网络匹配的超边数。$f_i^{(k)}(z_{i1},z_{i2},\cdots,z_{ik},Y)$ 是样本 $\boldsymbol{z}$ 生成的阶数为 k 的第 i 条超边。如果该超边与输入模式 X 正确匹配，则 $f_i^{(k)}(z_{i1},z_{i2},\cdots,z_{ik},Y)=1$，否则 $f_i^{(k)}(z_{i1},z_{i2},\cdots,z_{ik},Y)=0$。所谓正确匹配，可以这样理解：假设空间维度为 5 的样本 $\boldsymbol{z}_i=(z_{i1}=1,z_{i2}=0,z_{i3}=0,z_{i4}=1,z_{i5}=1,y_i=1)$，有一阶数为 3 的超边 $E_j=\{e_{j1}=1,e_{j3}=0,e_{j4}=1,e_{jy}=1\}$。因 $e_{ji}=z_{ii}$，即超边中除了最后的标签变量外所有变量与样本 z_i 都相等，则超边 E_j 与样本 z_i 匹配；若最后的标签变量也相等，即 $y_i=e_{jy}$，则超边与样本正确匹配。

4.2 超网络的演化学习

超网络的演化学习可以基于权值调整[8,9]，也可以基于超边替代[15]。本文采用超边替代的方式，即演化过程中超边不断地进行匹配、选择。给定一个训练集，可以创建一个超边集合，构成一个初始化的超网络。对这个集合中的超边不断进行匹配、选择和替代，即对初始化超网络进行调整，最终形成一种能够再现训练数据的概率分布。超网络的演化学习具体步骤如表 1 所示。

表 1 超网络的演化学习步骤

输入：训练样本集 $\boldsymbol{X}$；n 为样本个数；t 为每个样本生成的超边数；k 为超边阶数；

MaxSubstCnt=500

输出：超网络分类器 $H(X)$

(1)初始化：构造一个初始超边集合 $\mathrm{LIB}=\{L_1,L_2,\cdots,L_n\}$

(2)计算适应值：

For i=1 to n do

For j=1 to t do

计算每条超边的适应值：$\mathrm{fit}_w(l_{ij})$ 和 $\mathrm{fit}_c(l_{ij})$；

End For

End For

(3)计算超边替代数目 substCnt:

If substCnt > maxSubstCnt

substCnt=maxSubstCnt

maxSubstCnt=0.8 × substCnt

End if

(4)替代超边：按适应值排序，把排在末尾的 substCnt 条超边用新超边替代；

(5)if substCnt = 0

return

上述算法步骤(1)中：$L_i=\{l_{i1},l_{i2},\cdots,l_{it}\}$，是针对样本 $\boldsymbol{x}_i$ 随机生成的一个超边子集，l_{ij} 表示一条阶数为 k 的超边。全部样本的超边子集构成初始超边集合 LIB。每条超边定义两个初始适应值：表示超边能正确分类样本的适应值 $\mathrm{fit}_c=0$ 和不能正确分类样本的适应值 $\mathrm{fit}_w=0$。若训练集中共有 n 个样本，则此超边集合共包含 $n\times t$ 条超边。maxSubstCnt 为每次迭代过程中允许替代超边数的上限值。

步骤(2)中适应值的计算方法为：提取与每个样本匹配的超边，用这些超边分别对样本进行分类，被正确分类的样本加入到集合 X^c 中，被错误分类的样本加入到集合 X^w 中。对于 X^c 中样本 $\boldsymbol{x}_i$，如果超边 l_{ij} 可以正确分类样本 $\boldsymbol{x}_i$ 则将其适应值设置为 $\mathrm{fit}_c(l_{ij})=\mathrm{fit}_c(l_{ij})+1$，否则将其适应值设置为 $\mathrm{fit}_c(l_{ij})=\mathrm{fit}_c(l_{ij})-1$；对于 X^w 中样本 $\boldsymbol{x}_i$，如果该边可以正确分类 $\boldsymbol{x}_i$ 则将适应值设为 $\mathrm{fit}_w(l_{ij})=\mathrm{fit}_w(l_{ij})+1$，否则将适应值设置为 $\mathrm{fit}_w(l_{ij})=\mathrm{fit}_w(l_{ij})-1$。

步骤(3)中对超边适应值排序时，先按照 fit_w 降序排序；对于有相同 fit_w 的超边，按照 fit_c 降序排序。超边替代数量 substCnt 式(7)，式(8)求得：

$$\mathrm{substCnt}=w\times r\times|L| \tag{7}$$

$$r=\frac{|X^w|}{|X|} \tag{8}$$

$|X|$ 表示训练集的样本数量。w 用于控制被替代的超边数目，$|L|$ 表示超边的总数。适应值排序最后的 substCnt 个超边，用与之关联样本重新产生的新超边替代。

超网络的演化学习过程中，超网络的最大可能超边数目是 $2^k\times C(n,k)$，n 是样本包含特征向量个数，k 是超边的阶数。演化过程中，随机产生超边集合，这样不同相互作用的基因之间就可以任意组合，根据适应值，去掉分型相关性不好的超边，保留分型相关性好的超边，有利于发现对分型贡献大的基因组合。在选择替代超边过程中把那些适应值低的超边用新的超边代替，增加了算法的问题解空间，对提高演化超网络的性能有一定帮助。

4.3 演化超网络构造多分类系统

$k(k>2)$种类型的癌症分型，采用一对多的方法，建立 k 个独立的二类分类器，并把它们集成为一个多类型分子分型系统。其中，每一个二类分类器都是经过演化学习的超网络。整体结构如图 3 所示：

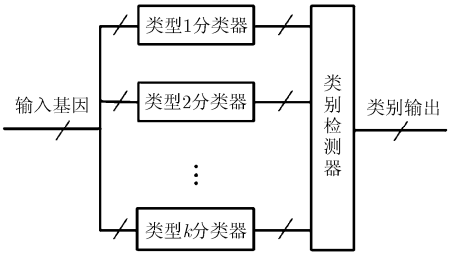


图 3 基于演化超网络集成的多类型分子分型系统

对样本进行分类，针对不同的类型分类器输入样本与之关联的基因。经过类型分类器的处理，每个分类器输出样本所属类型：单一类型 c_d 或者混合类型 c_m。把 k 个分类器的输出结果分别输入到类别检测器进行检测，判断样本的最终类型并输出。本文中，类别检测器中判定样本类型方法是：若类型 1 分类器输出的结果是单一类型 c_d，而其它类型分类器均输出混合类型，则可以判断该样本属于类型 1 分类器识别的单一类型；若经过 k 个类型分类器处理后，有大于 1 个的类型分类器输出单一类型，我们计算每一个单一类型下该样本正确匹配超边的比例，把比例较大的类型判定为该样本的类型。

5 实验结果与分析

为验证基于演化超网络的多类型癌症分子分型系统性能，我们采用急性白血病数据集(leukemia)[14]，儿童小圆蓝细胞肿瘤数据集(SRBCT)[17]和GCM[18]数据集进行测试。首先采用一

对多的方法，将多类型问题转化为一系列的二类分类问题。针对二类分类问题，首先根据信噪比方法进行信息基因选择，选出与分类相关度高的信息基因子集并进行归一化处理；然后利用训练集样本对超网络进行演化，生成多个二类分类器；通过分类器集成，构建一个多类型的癌症分子分型系统。为验证分型系统的性能，本文采用留一交叉验证法(Leave One Out Cross Validation, LOOCV) [2]，每个实验结果取30次实验的平均结果，对比了不同参数设定对系统性能的影响。

表 2 给出了实验所用到的急性白血病数据集，儿童小圆蓝细胞肿瘤数据集和GCM数据集类别数，样本数和基因个数的描述。

表 2 实验所用数据集的基本描述

	类别数	训练集数目	测试集数目	基因个数
Leukemia	3	38	34	7129
SRBCT	4	63	20	2308
GCM	14	144	54	16063

图 4 给出了 3 个数据集在样本选择基因个数为 32，初始超边数为 100, w=1.5 时，不同阶数下的平均识别率。由图中可以看出，在超网络模型中，超边的阶数对系统的分类性能有较大影响，尤其是对急性白血病数据集和 GCM 数据集而言。不同的数据集获得最好的识别率都对应不同的阶数。

初始超边数为 100, w=1.5，固定阶数的设定下，选择的信息基因数不同也会对识别率造成影响，如表 3 所示。随着信息基因个数增加，急性白血病和GCM 数据集的识别率都有所降低。过多的信息基因将导致冗余和噪声基因的增加，降低超网络识别率。而对 SRBCT 数据集，其识别率保持在 100%，这可能是由于该数据经过 SNR 选出的基因子集是非常规律的，所以基因个数的改变对其影响不大。

表 3 不同信息基因个数下的平均 LOOCV 识别率(%)

数据集	信息基因个数			
	32	64	100	200
Leukemia	98.61	97.22	97.22	97.22
SRBCT	100	100	100	100
GCM	81.31	80.30	72.72	74.75

图 5 给出了 3 个数据集在初始超边数为 100，阶数固定，信息基因个数为 32 的设定下，w 取不同值时的平均识别率。由图中可以看出，w 值的改变只对 GCM 数据集的识别率有所影响。w 取值限制了超网络演化过程中每次迭代可替代的超边总数。超网络的演化是以不断挖掘对分类相关的超边，并且保留对分类效果最好的超边为目的。所以，对于某些数据集，w 值仍然有一定影响：w 值的改变，可能导致演化过程中对样本分类高相关的超边的替代。

传统机器学习方法的一个固有缺点在于其学习结果通常难于理解。例如神经网络学习的结果是一些权值。超网络通过演化学习不断对特征空间进行搜索，可以得到与分类高度关联的基因组合，而最终演化得到的超边集合直接反映了样本中的不同基因组合对分类的关键程度。表 4 以急性白血病数据集为例，列出了多类型分型系统中某一二类分类器超边权值排前 10 的基因组合。排列名次越靠前，说明这组基因组合与样本分类关联度越大。

表 5 列出了不同阶数设定下，演化超网络针对上述 3 种癌症数据集的演化学习时间。学习时间在以下实验平台获得：C 语言(VC6.0)， Pentium(R) Dual-Core CPU, 2.70 GHz CPU 时钟频率，2 GB 内存。由表 5 可见，随着超网络阶数增大，系统学习时间也相应增加。

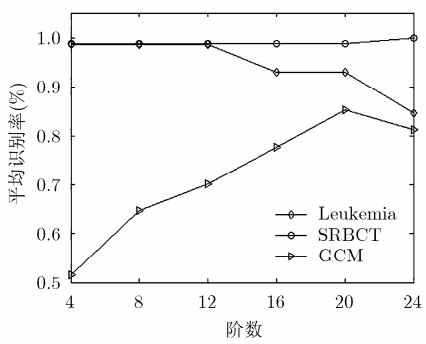


图 4 不同阶数下的平均 LOOCV 识别率

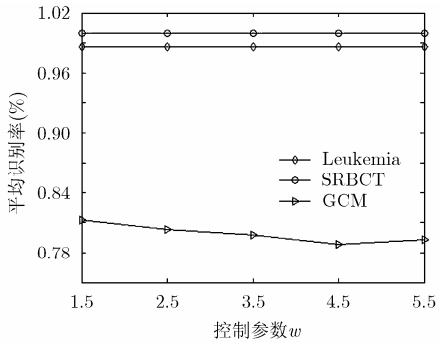


图 5 w 不同取值下的平均 LOOCV 识别率

表 4 急性白血病数据集下超边权值排名前 10 基因组合实例

序号	基因组合							
1	J03077_s	M63483	U50327_s	M33197_5	U59878	X99584	J03801_f	U64675
2	M33197_5	J03077_s	U59878	M27891	M63483	U50327_s	J03801_f	X99584
3	U50327_s	U64675	M19045_f	J03077_s	J03801_f	U59878	X99584	D11327_s
4	D87743	U48251	J03473	M28170	L06797_s	X15949	M89957	X82240_rna1
5	U50327_s	U64675	U09578	D21261	D11327_s	M19045	U59878t	M33197_5
6	X82240_rnal	L06797_s	Z49194	U05259_rna1	M89957	X58529	M28170	X15949
7	X89985	U50327_s	X99584	U79253	M27891	D11327_s	M19045_f	M33197_5
8	U64675	X89985	D11327_s	M27891	J03801_f	U79253	U59878	M19045_f
9	U64675	D11327_s	M33197_5	J03801_f	X89985	X99584	U79253	D21261
10	D21261	X89985	M19045_f	D11327_s	J03801_f	U59878	M27891	J03077_s

表 5 不同阶数下演化超网络学习时间对比(s)

数据集	阶数					
	4	8	12	16	20	24
Leukemia	87.89	96.50	108.36	130.16	134.28	157.74
SRBCT	89.01	26.80	21.56	135.84	207.78	372.20
GCM	2537.56	2299.76	3450.45	3427.78	7521.66	6015.84

为验证演化超网络多类型癌症分型方法的性能，表 6，表 7，表 8 给出了演化超网络与其它模式识别方法的识别率(包含 LOOCV 识别率和独立测试集识别率)对比结果。表中本文方法识别率后的 3 个数字分别表示超网络阶数、信息基因数、w 值。从表 6，表 7，表 8 可见，本文方法拥有与其它方法可比的识别率。表 7 中 SRBCT 数据集本文方法的识别率达到了 100%，主要是该数据集本身有效信息含量较多，数据处理过程中信息损失较少的结果。

6 结束语

本文探索了演化超网络在多类型癌症分子分型中的应用。首先把多类型癌症分型问题转化成二类癌症分型问题，采用演化超网络构建分型系统，最终实现基于 DNA 微阵列数据的多类型癌症诊断。在超网络演化学习过程中，随机选取超边构成超边集合，并采用基于替代的演化学习，不断地把初始化超边集中未包括的超边加入到搜索空间，扩大了搜索范围，适用于发现基因之间的相互作用。相较于传统模式识别方法学习结果不易理解和分析的局限，超网络学习结果有良好的可读性。同其它传统分类方法的对比实验表明，演化超网络有较高的分类准确度。

为进一步提高演化超网络性能，有效处理高维 DNA 微阵列数据，有待探索更优的超网络计算模型和硬件实现方法。此外，数据处理过程中的特征选

表 6 不同分类方法对急性白血病数据集的平均识别率对比(%)

方法	独立测试集	LOOCV
本文方法	95.74(6,32,2)	98.61(12,32,1.5)
TPCR[19]	–	98.61
TSG[20]	97.06	–
QDA[21]	–	100
ENRHC[22]	97.06	–
MSVM[23]	97.06	–

表 7 不同分类方法对 SRBCT 数据集的平均识别率对比(%)

方法	独立测试集	LOOCV
本文方法	100(4,32,2)	100(24,32,1.5)
FNN[1]	95	–
TPCR[19]	–	100
TSG[20]	100	–
ENRHC[22]	95	–
MSVM[23]	100	–

表 8 不同分类方法对 GCM 数据集的平均识别率对比(%)

方法	独立测试集	LOOCV
本文方法	70.37(9,20,1.5)	85.35(20,32,1.5)
SVM[24]	78	–
OAA [18]	69.57	–
GA/SVM[10]	–	85.19
ENRHC[22]	63.04	–
TSG [20]	67.39	–

择和归一化等还有待进一步研究，以有效地和超网络模型相结合挖掘癌症相关基因，提高分型精度。

参 考 文 献

[1] Wang L, Chu F, and Xie W. Accurate cancer classification using expressions of very few genes[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2007, 4(1): 40–53.

[2] 任江洪, 韩露. 基于高斯过程的 DNA 微阵列分类算法[J]. 计算机工程与应用, 2011, 47(33): 26–29.
Ren J H and Han L. Classification algorithm for DNA microarray base on Gaussian process[J]. *Computer Engineering and Applications*, 2011, 47(33): 26–29.

[3] Resson H W, Varghese R S, Zhang Z, *et al.*. Classification algorithms for phenotype prediction in genomics and proteomics[J]. *Frontiers in Bioscience*, 2008, 13: 691–708.

[4] 明利特, 蒋芸, 王勇, 等. 基于临域粗糙集和概率神经网络集成的基因表达谱分类方法[J]. 计算机应用研究, 2011, 28(12): 4440–4444.
Ming L T, Jang Y, Wang Y, *et al.*. Gene expression profiles classification method based on meighborhood rough set and probabilistic neural networks ensemble[J]. *Application Research of Computers*, 2011, 28(12): 4440–4444.

[5] 陈磊, 刘毅慧. 基于 CART 算法的肺癌微阵列数据的分类[J]. 生物信息学, 2011, 9(3): 229–234.
Chen L and Liu Y H. Classification based on CART algorithm for microarray data of lung cancer[J]. *China Journal of Bioinformatics*, 2011, 9(3): 229–234.

[6] 于化龙, 顾国昌, 赵靖, 等. 基于 DNA 微阵列数据的特征子空间集成分类[J]. 吉林大学学报(工学版), 2011, 41(4): 1071–1076.
Yu H L, Gu G C, Zhao J, *et al.*. Feature subspace ensemble classification based on DNA microarray data[J]. *Journal of Jilin University*(*Engineering and Technology Edition*), 2011, 41(4): 1071–1076.

[7] Ha J W, Lee B J, and Zhang B T. Text-to-image retrieval based on incremental association via multimodal hypernetworks[C]. Proceedings of the 2012 IEEE Conf. Systems, Man, and Cybernetics, Seoul, Korea, 2012: 3239–3244.

[8] Park C H, Kim S J, Kim S, *et al.*. Finding cancer-related gene combinations using a molecular evolutionary algorithm[C]. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, Boston, MA, USA, 2007: 158–163.

[9] Park C H, Kim S J, Kim S, *et al.*. Use of evolutionary hypernetworks for mining prostate cancer data[C]. Proceedings of the International Symposium on Advanced Intelligent Systems, Sokcho, Korea, 2007: 702–706.

[10] Peng S, Xu Q H, Ling X B, *et al.*. Molecular classifcation of cancer types from microarray data using the combination of genetic algorithms and support vector machines[J]. *Federation of European Biochemical Societies Letters*, 2003, 555(2): 358–362.

[11] Joseph S J, Robbins K R, Zhang W S, *et al.*. Comparison of two output-coding strategies for multi-class tumor classification using gene expression data and latent variable model as binary classifier[J]. *Cancer Informatics*, 2010, 9(3): 39–48.

[12] 段旭. 基于边缘分布模型的基因选择方法[J]. 计算机工程与设计, 2011, 32(11): 3836–3839.
Duan X. Marginal distribution model for gene selection[J]. *Computer Engineering and Design*, 2011, 32(11): 3836–3839.

[13] 耿耀君, 张军英. 一种基于监督降维和形状分析的基因选择方法[J]. 西安电子科技大学学报(自然科学版), 2011, 38(3): 121–127.
Geng Y J and Zhang J Y. Gene selection method based on supervised dimension reduction and procrustes analysis[J]. *Journal of Xidlan University*, 2011, 38(3): 121–127.

[14] 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.

[15] 王进, 金理雄, 孙开伟. 基于演化超网络的中文文本分类方法[J]. 江苏大学学报(自然科学版), 2013, 34(2): 196–201.
Wang J, Jin L X, and Sun K W. Chinese text categorization based on evolutionary hypernetwork[J]. *Journal of Jiangsu University* (*Natural Science Edition*), 2013, 34(2): 196–201.

[16] 王众托. 关于超网络的一点思考[J]. 上海理工大学学报, 2011, 33(3): 229–237.
Wang Z T. Reflection on supernetwork[J]. *Journal University of Shanghai for Science and Technology*, 2011, 33(3): 229–237.

[17] Khan J, Wei J S, Ringner M, *et al.*. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. *Nature Medicine*, 2001, 7(6): 673–679.

[18] Tapia E, Ornella L, Bulacio P, *et al.*. Multiclass classification of microarray data samples with a reduced number of genes[J]. *BMC Bioinformatics*, 2011, DOI:10.1186/1471–2105–12–59.

[19] Tan Y X, Shi L M, Tong W D, *et al.*. Multi-class cancer classification by total principal component regression(TPCR) using microarray gene expression data[J]. *Nucleic Acids Research*, 2005, 33(1): 56–65.

[20] Wang H Y, Zhang H Y, Dai Z J, *et al.*. TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection[J]. *BMC Medical Genomics*, 2013, 6(Suppl.1): S3.

[21] Nguyen D V and Rocke D M. Multi-class cancer classification via partial least squares with gene expression profiles[J]. *Bioinformatics*, 2002, 18(9): 1216–1226.

[22] Wei J M, *et al.*. Ensemble rough hypercuboid approach for classifying cancers[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(3): 381–391.

[23] Lee Y Y and Lee C K. Classification of multiple cancer types by multicategory support vector machines using gene expression data[J]. *Bioinformatics*, 2003, 19(9): 1132–1139.

[24] Rifkin R, Mukherjee S, Tamayo P, *et al.*. An analytical method for multiclass molecular cancer classification[J]. *SIAM Review*, 2003, 45(4): 706–723.

王　进：　男，1979 年生，教授，研究方向为演化计算、机器学习、智能信息处理.

丁　凌：　女，1986 年生，硕士生，研究方向为特征选择、集成学习.

孙开伟：　男，1987 年生，硕士生，研究方向为演化超网络、不平衡数据挖掘.