用于微阵列数据分类的子空间融合演化超网络

王 进,刘 彬,张 军,陈乔松,邓 欣
(重庆邮电大学计算智能重庆市重点实验室,重庆 400065)

摘 要: 针对传统模式识别方法在学习具有小样本特性的 DNA 微阵列数据时存在的过拟合问题,本文提出了一种子空间融合演化超网络模型. 该模型通过子空间划分、超边全覆盖和子空间融合三种方法降低模型对初始化的依赖,减少了对数据空间的拟合误差,提高了演化超网络的泛化能力. 对四个 DNA 微阵列数据集的实验结果表明,子空间融合演化超网络的识别率和在小样本训练集下的泛化能力均优于参与对比的其他传统模式识别方法.

关键词: 模式识别; 微阵列数据分类; 演化超网络; 子空间; 过拟合
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2016)10-2308-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.10.004

Classification of Microarray Data Using Evolutionary Hypernetworks with Subspace Fusion

WANG Jin, LIU Bin, ZHANG Jun, CHEN Qiao-song, DENG Xin
(*Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: In order to solve the over-fitting problem of the traditional pattern recognition approaches under the DNA microarray data with small train samples, a subspace fusion-based evolutionary hypernetwork model is proposed in this paper. With the methods of subspace division, hyperedge coverage, and subspace fusion, the proposed scheme reduces the dependence on the initialization, decreases the fitting error of the data space, and enhances the generalization ability of the evolutionary hypernetwork. The experimental results on four DNA microarray datasets show that the proposed model achieves higher classification accuracy and stronger generalization ability than other compared traditional pattern recognition method.

Key words: pattern recognition; microarray data classification; evolutionary hypernetwork; subspace; over-fitting

1 引言

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,特别是在临床诊断白血病[1]、结肠癌[2]等恶性肿瘤上具有较高的应用价值. 与传统基于形态学信息的癌症诊断方法相比,基于 DNA 微阵列技术获得的基因表达谱的癌症诊断方法具有更高的准确率和可信度[1].

传统的模式识别方法在学习具有小样本特性的 DNA 微阵列数据时存在过拟合问题[3],这导致模型分类的泛化能力下降. 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我们研究癌症的复杂发展机制有着重要意义. 传统模式识别方法[4~7]虽然取得了较好的分类效果,却难以深度挖掘基因之间的相互作用.

超网络(Hypernetwork, HN)是受生物分子网络启发而建立的一种基于超图(Hypergraph)的认知学习模型[8,9]. 通过演化学习,超网络可以有效获取与分类相关的关键特征,拟合输入模式空间中数据的分布概率,从而表达复杂数据的内在结构和相互之间的关系. 因能有效挖掘与癌症分类相关的基因以及基因间的相互作用,演化超网络模型已成功用于 DNA 微阵列数据分类[10,11],然而该模型的分类效果与泛化能力受超边库初始化质量的影响较大.

针对上述问题,本文将子空间概念引入到演化超网络模型中,提出了一种子空间融合演化超网络(Evolutionary Hypernetworks with Subspace Fusion, SF-HN). 通过子空间超边覆盖,弱化模型对超边初始化过程的依

收稿日期:2015-03-11;修回日期:2015-06-30;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61203308, No. 61403054);重庆教委科学技术研究项目(自然科学类)(No. KJ1400436);重庆市基础与前沿研究计划项目(No. cstc2014jcyjA40001)

赖,提升其在小样本训练集下的泛化能力. 为验证子空间融合演化超网络的性能,本文根据部分替代整体思想提出了一种分类器泛化能力评价方法. 通过对四个 DNA 微阵列数据集进行试验,证明了该模型具有更优的准确性和泛化能力.

2　演化超网络

超网络是一种由大量超边组成的概率图模型,通过超边表达模式空间中数据的分布概率[8]. 超边所连接的顶点数称为超边的阶数(Order),所有超边阶数都为 k 的超网络称为 k 阶超网络[12]. 超网络演化学习通过调整超边库,提高模型与数据在模式空间概率分布的拟合度. 超边替代法[11]和梯度下降法[12]是常用的演化学习方法. 在分类模式下,超网络通过输入样本 X 与输出类别 Y 的联合概率 $P(X,Y)$ 以及 X 的分布概率 $P(X)$,得到最终的决策输出:

$$Y^* = \arg\max_Y(P(X,Y)/P(X)) = \arg\max_Y(P(Y|X)) \tag{1}$$

细粒度演化超网络(Fine-Grain Evolutionary Hypernetwork,FG-HN)[11]将最优类别信息离散化(Optimal Class-Dependent Discretization,OCDD)算法与超网络结合,采用多位二进制来表述特征属性,降低了数据离散化过程中的信息损失. 然而 FG-HN 仍无法解决在学习具有小样本特性的 DNA 微阵列数据时存在的过拟合问题.

3　子空间融合演化超网络

传统演化超网络只对输入模式中的训练样本集进行学习,处理小样本数据时,其泛化性将受到影响. 为了提高模型的泛化能力,本文在 FG-HN[11]的基础上提出了一种子空间融合演化超网络.

令 $S = A_1 \times A_2 \times \cdots \times A_D$ 表示 D 维数据空间,$A_j(j = 1,\cdots,D)$ 表示 S 中的一个属性域,k 维空间 $P_i = A_{i1} \times A_{i2} \times \cdots A_{ik}(ik \leqslant D)$ 为 S 的一个子空间. $S = P_1 \cup P_2 \cup \cdots \cup P_i \cup \cdots$ 为空间 S 的一个子空间划分. 在分类过程中,$X = A_1 \times A_2 \times \cdots \times A_D$ 表示特征属性空间,Y 表示类别标签空间. 对于空间 X 的数据进行离散化处理,特征 A_j 的离散区间数为 m_j,则子空间 P_i 包含的总数据点为 $m = m_{i1} \times m_{i2} \times \cdots \times m_{ik}$,$m$ 也称为 P_i 的秩,空间中的数据点也称为单元格. 超边所包含的特征空间可表示为 $Ei = A_{i1} \times A_{i2} \times \cdots \times A_{ik}$. 将超边看作输入模式空间的子空间,超边库表示特征属性空间 X 的一个划分 $X = E_1 \cup E_2 \cup \cdots \cup E_{|L|}$,其中 $|L|$ 表示超边总数.

偏斜度 $SOD(T,P)$[13]是衡量子空间划分效果的评价指标,其定义如下:

$$SOD(T,P) = \sum_{i=1}^{m} |pi - \mu| \Big/ 2N \tag{2}$$

其中,N 为训练集 T 的样本数,P 为子空间,p_i 为训练集 T 投影在子空间 P 的第 i 个单元格上的样本数,m 是子空间的秩,$\mu = N/m$ 表示平均分布在单元格上的数据点数. $SOD(T,P)$ 的取值范围为[0,1],其值越小,数据点的分布越均匀;反之,则分布越集中.

在 SF-HN 中,首先进行子空间划分,选择样本分布均匀的子空间集合;其次,生成超边并把超边决策范围覆盖到整个子空间;接着融合子超边簇,生成初始化模型;最终通过梯度下降方法对模型进行演化学习,提高模型对输入数据的拟合精度. 子空间融合演化超网络流程如图 1 所示,其中超网络中的每种连线代表一条超边(例如实线表示一条包含顶点 A_4,A_1 和 A_6 的 3 阶超边).

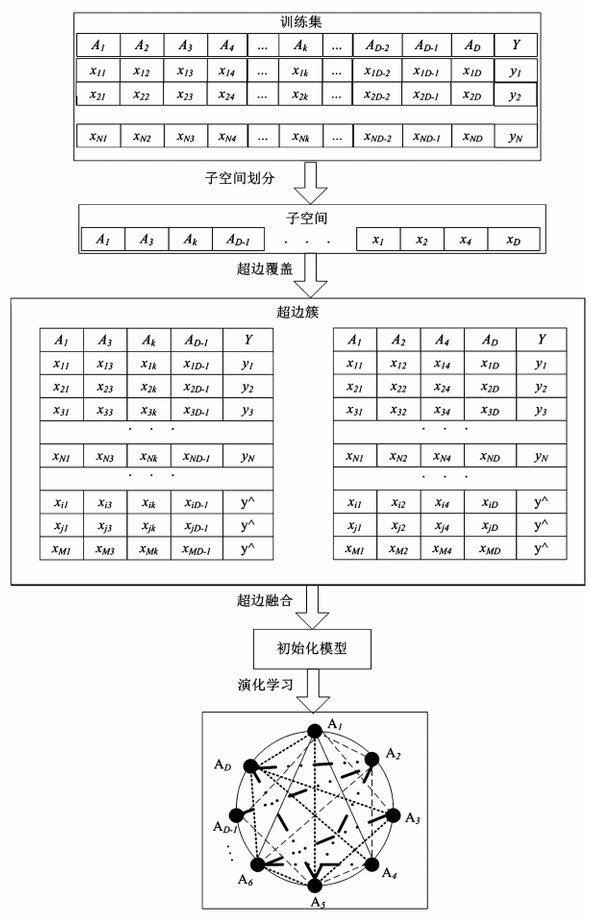


图1 子空间融合演化超网络

3.1　子空间划分算法

超网络是对输入模式空间数据分布概率的拟合,当数据散列分布时,其覆盖的数据空间广,拟合时的误差小. 因此在子空间划分算法中,采用训练集在子空间上的偏斜度对子空间的优劣进行评价,并选取样本分

布信息较多的子空间集合.

子空间划分算法的流程为:首先生成 $sn*d$ 条 k 阶超边,sn 表示超边总数(既划分子空间数),d 表示冗余倍数;然后提取每条超边所表述的子空间,将训练集在该子空间上进行投影操作,计算子空间的偏斜度;最后根据偏斜度对子空间排序,选取偏斜度最低的 sn 个子空间作为模式空间的划分. 此外,为了平衡冗余倍数与子空间性能的关系,在算法中加入阈值 $tsod$,当选取的子空间偏斜度值大于该阈值时,就剔除该子空间并重新生成. $tsod$ 的取值为 $SOD(T,E_t)$,其中 E_t 为只包含 k 个相同特征 A_{ind} 的子空间,$ind=\underset{i\in\{1,2,\cdots,D\}}{\operatorname{argmin}}(SOD(T,ASi))$ 是具有最小 SOD 值的特征下标,AS_i 为只包括一个特征 A_i 的子空间.

算法 1 子空间划分算法

输入:训练集 T,超边阶数 k;子空间数 sn,冗余倍数 d,阈值 $tsod$.
输出:划分的子空间集合 E.
步骤 1 $num\leftarrow sn$,$E\leftarrow\Phi$,计算 $tsod$.
步骤 2 初始化 $num*d$ 条阶数为 k 的超边.
步骤 3 将 T 向每条超边对应的子空间 EE_i 投影,并计算 $SOD(T,EE_i)$.
步骤 4 将所有子空间按 SOD 值升序排序.
步骤 5 若选择前 num 个子空间的 SOD 值均小于 $tsod$,则将前 num 个子空间加入 E;否则将满足 $SOD(T,EE_i)<tsod$ 的子空间加入 E,并记录仍需生成的子空间个数 $tsum\leftarrow sn-|E|$.
步骤 6 若 $tsum>0$,$num\leftarrow tsum$,转入步骤 2;否则转入步骤 7.
步骤 7 返回 E.

3.2 子空间超边生成算法

子空间超边生成算法通过产生由训练样本映射得到的与训练集完全拟合的映射超边和通过映射超边信息确定类别的预测超边,加入样本关联信息,扩展超边的决策范围,对子空间进行超边全覆盖.

子空间超边生成算法的流程为:将子空间 E_i 中对应的单元格转化为超边加入到子超边簇 LS_i 中,此时超边不包含类别信息;将训练集 T 在子超边簇 LS_i 上投影,并确定至少有一个样本映射到对应单元格的超边类别;最后对剩余未知类别信息的超边进行类别预测. 由于输入模式空间数据为连续分布,因此对模式空间中的数据点,其类别可由其相邻数据点的类别确定. 故对每条未知类别超边,统计其相邻超边的类别,并将包含超边最多的类别赋给待预测类别超边;若不同类别包含的超边数相等,则此超边处在类别分界线上,不对其类别赋值. 当无新确定类别的超边时,算法终止.

算法 2 子空间超边生成算法

输入:训练集 T,子空间 E_i.
输出:子超边簇 LS_i.
步骤 1 $LS_i\leftarrow\Phi$.
步骤 2 子空间 E_i 中每个单元格 f_j 转化为超边 l_j 并加入到子超边簇 LS_i,其中超边的类别标签为空.
步骤 3 将训练集 T 在子超边簇 LS_i 上投影.
步骤 4 遍历每条超边 l_j 对应的单元格 f_j,若至少有一个样本映射到 f_j,则将该超边类别赋为映射到相应单元格中数量最多的样本类别.
步骤 5 统计未知类别超边的数量 ln,$lt\leftarrow ln$.
步骤 6 统计每条未知类别超边的相邻超边类别,若不同类别超边数量不等,则将包含超边数量最多的类别赋给该超边.
步骤 7 统计未知类别的超边数量 ln,若 $lt\neq ln$,转入步骤 5;否则转入步骤 8.
步骤 8 返回 LS_i.

3.3 子空间融合算法

覆盖子空间的子超边簇既包含由训练集映射而成的超边,也包含由映射超边对未知类别超边进行预测扩展而成的超边. 子超边簇中由训练集映射而成的超边是对训练集样本分布的零误差拟合,而经预测扩展而成的超边则存在拟合误差,并且不同子空间中预测超边的拟合误差不同. 子空间融合算法通过融合不同子空间上的子超边簇,降低预测超边的拟合误差,进而提高模型对训练集的拟合精度. 由于偏斜度低的子空间生成的超边簇具有更优的拟合效果,在融合时采用加权集成的方式,将 $1\text{-}SOD(T,E_i)$ 作为子空间对应超边簇的权重.

算法 3 子空间融合算法

输入:训练集 T,超边簇集合 LS.
输出:超边库 L.
步骤 1 $L\leftarrow\Phi$.
步骤 2 遍历每个子超边簇,根据 T 在每个子超边簇 LS_i 对应子空间 E_i 的投影,计算 $SOD(T,E_i)$,并将 LS_i 中每个超边的权重设为 $1\text{-}SOD(T,E_i)$.
步骤 3 将赋予权重的超边簇加入到 L 中.
步骤 4 返回 L.

3.4 子空间融合超网络的演化学习

SF-HN 通过将子超边簇融合为一个超边库,拟合模式空间的数据分布. 因模型对子空间进行了超边全覆盖操作,无需替代操作,故采用梯度下降演化学习方法[12],通过训练集来控制超边权重的调整方向,调整模型结构,降低融合后模型对未知样本的预测误差. 权重变化值 Δw_j 计算公式为:

$$\Delta w_j=\eta\sum_{\boldsymbol{x}_i\in T}(P^*(y^*|\boldsymbol{x}_i)-P(y^*|\boldsymbol{x}_i))\times I_{l_j=(\boldsymbol{x}_i,y)}\quad(3)$$

其中,$P(y^*|\boldsymbol{x}_i)$ 和 $P^*(y^*|\boldsymbol{x}_i)$ 分别表示样本 $\boldsymbol{x}_i$ 属于类别 y^* 的实际概率和目标概率,y^* 是超网络对样本 $\boldsymbol{x}_i$ 的分类结果,y 是样本 $\boldsymbol{x}_i$ 的真实类别,$\boldsymbol{\eta}$ 是学习速率. I 为

匹配函数,当超边 l_j 与样本 $\boldsymbol{x}_i$ 匹配时值为 1;否则值为 0. 子空间融合超网络的分类方法与传统超网络的流程[11]相似,唯一的区别在于估计概率时统计超边权重之和而不是数量之和.

算法 4　子空间融合超网络演化学习算法

输入:训练集 T,子空间数 sn;超边阶数 k;梯度下降演化代数 $iternum$.
输出:超边库 L.

步骤 1　根据子空间划分算法,生产包含 sn 个子空间的子空间集合 E.
步骤 2　对每个子空间 E_i,利用子空间超边生成算法,生成子超边簇 LS_i. 最终得到包含 sn 个子超边簇的超边簇集合 LS.
步骤 3　对超边簇 LS,利用子空间融合算法,得到初始超边库 L.
步骤 4　t←0.
步骤 5　用当前子空间融合超网络模型对训练集分类.
步骤 6　对每个错分样本 $\boldsymbol{x}_i$,更新与 $\boldsymbol{x}_i$ 匹配的超边 l_j 的权重 $w_j = w_j + \Delta w_j$,其中通过式(3)计算 Δw_j.
步骤 7　t + +,若 $t < iternum$,转入步骤 5;否则转入步骤 8.
步骤 8　返回 L.

4　实验结果与分析

为验证子空间融合演化超网络的分类准确性和泛化性,本文采用结肠癌[2]、急性白血病[1]、肺癌[14]、前列腺癌[15]四个 *DNA* 微阵列数据集进行实验验证. 数据集的具体信息如表 1 所示.

4.1　分类性能测试

本文采用 OCDD 算法[16]对输入数据进行离散化处理,采用信噪比特征基因选择方法[1]对数据进行降维处理. 为了验证 SF-HN 的分类效果,将其与其他文献方法(GSVM-RFE[5], NN[4], Bagging[7])、传统分类方法(C4.5 决策树、朴素贝叶斯(Na¿ve Bayes,NB)、支持向量机(Support Vector Machine,SVM))以及 HN 和 FG-HN 进行对比. 本文的所有实验结果为 20 次实验的平均值. SF-HN 的参数通过训练集 5 折交叉验证来确定,其参数设置如表 2 所示. HN 和 FG-HN 采用文献[11]中的实验参数设定,分类算法 C4.5、NB、SVM 采用 Weka 机器学习开源项目提供的算法(http://www.cs.waikato.ac.nz/ml/weka/),其输入数据的特征维度与 FG-HN 相同. 此外,本文通过 t-检验来测试 SF-HN 在统计学上是否显著优于 C4.5、NB、SVM、HN 和 FG-HN 方法.

通过对完整的训练集进行学习,然后对独立测试集进行测试,所得结果如表 3 所示. 在表 3 中,"·"表示 SF-HN 在 $p < 0.01$ 的水平下显著优于对比方法. 相对于其它对比分类算法,SF-HN 具有较好的分类性能和显著性优势. 这主要是由于 SF-HN 在空间中进行超边覆盖,增加了模型的信息熵,从而更有效地拟合输入模式空间中的数据分布.

表 1　数据集信息表

数据集	训练集样本数	测试集样本数	特征维度
结肠癌	38	24	2000
急性白血病	38	34	7129
肺癌	32	149	12533
前列腺癌	102	34	12600

表 2　SF-HN 的参数设置

数据集	特征数	离散区间数	超边阶数
结肠癌	32	3	3
急性白血病	20	4	3
肺癌	45	4	5
前列腺癌	32	4	2

表 3　不同方法对 4 个 DNA 微阵列数据集的测试集分类结果

	GSVM-RFE[5]	NN[4]	Bagging[7]	C4.5	NB	SVM	HN[11]	FG-HN[11]	SF-HN
结肠癌	91.68%	87.90%	–	91.67% ·	91.67% ·	95.83%	89.58% ·	95.00% ·	95.83%
急性白血病	–	95.90%	–	88.23% ·	91.18% ·	94.11% ·	93.20% ·	95.60% ·	96.47%
肺癌	–	–	93.29%	93.95% ·	97.31% ·	99.32% ·	98.19% ·	99.40% ·	99.60%
前列腺癌	98.29%	–	73.53%	91.17% ·	94.12% ·	97.06% ·	96.47% ·	100.00%	99.70%

4.2　泛化性能测试

泛化能力表示分类器通过对训练集的学习,对未知样本的预测能力[17]. 但在据作者所知的文献中,还缺乏公认的对分类器泛化性能评价的定量指标. 机器学习领域通常认为泛化性能好的算法在小样本训练集下仍可获得较高的分类精度. 本文采用部分替代整体思想进行泛化性能测试的实验设计,通过拆分原始训练集获得小样本训练集,进而验证不同分类方法在小样本训练集下的泛化性能. 泛化性测试的主要流程为:首先是将训练集按原正负类别的比例平均分为 n 份;之后利用拆分后的每一份数据分别训练分类器并对独立测试集进行测试. 对得到的 n 个独立测试集测试结果求取平均,作为分类器泛化能力评价指标. 在本文中,训练集平均划分为 n 份以 n-bt 表示.

泛化性能测试中对训练集进行拆分后,训练集中样本数量减少,离散区间数过大将导致数据中同类别样本间的关联概率降低;而阶数过大的超边很难与样本进行匹配. 因此对四个数据集,特征选择数设为 32,特征最大离散区间数为设为 3,HN、FG-HN 和 SF-HN 的阶数分别设定为 5、4、3. 对每个数据集,采取 2-bt、3-bt、4-bt、5-bt 泛化性能实验.

泛化性测试结果如表 4 ~ 7 所示. 相对于其他方法,

在 3-bt、4-bt、5-bt 设定下 SF-HN 具有更高的泛化性能. 这是因为在分类器的学习过程中,SF-HN 通过对超边类别的预测,对子空间进行超边覆盖,在本质上类似于虚拟样本生成,通过增加样本的数量,实现了对数据分布的更优拟合. 而在 2-bt 时,由于某些数据集中正负类别的界限较宽,SVM 能够发挥更优的性能. 当 n – bt 中的 n 增大时,所有方法对独立测试集的识别率随之降低. 这是因为随着训练集样本数的减少,关于模式空间描述的信息量相应减少,从而导致分类器对模式空间的描述可信度降低. 然而相对其它方法,SF-HN 下降趋势最缓慢. 这是因为 SF-HN 通过子超边簇对子空间进行全覆盖,增加了超边对未知样本的匹配概率,避免了超边对其生成样本的过度依赖,不会出现对训练集的过拟合,在小样本数据中具有较高的优势.

表 4 结肠癌数据集不同分类器泛化性能测试结果

	C4.5	NB	SVM	HN	FG-HN	SF-HN
2-bt	87.50%	92.75%	94.17%	88.89%	89.72%	93.08%
3-bt	83.33%	86.11%	91.05%	87.48%	89.12%	91.80%
4-bt	81.94%	82.50%	89.67%	87.08%	88.88%	91.67%
5-bt	73.33%	81.25%	87.59%	85.83%	87.50%	88.65%

表 5 急性白血病数据集不同分类器泛化性能测试结果

	C4.5	NB	SVM	HN	FG-HN	SF-HN
2-bt	84.56%	79.41%	93.12%	92.08%	93.85%	95.28%
3-bt	84.31%	72.55%	93.53%	92.50%	91.47%	95.00%
4-bt	83.83%	70.59%	92.16%	91.47%	84.55%	93.70%
5-bt	79.41%	69.12%	90.44%	89.31%	84.35%	92.51%

表 6 肺癌数据集不同分类器泛化性能测试结果

	C4.5	NB	SVM	HN	FG-HN	SF-HN
2-bt	81.72%	99.32%	99.32%	98.59%	99.24%	99.32%
3-bt	81.21%	98.83%	99.13%	97.98%	98.82%	99.22%
4-bt	74.72%	97.85%	99.10%	97.24%	98.18%	99.19%
5-bt	71.53%	95.97%	98.89%	96.19%	97.94%	98.92%

表 7 前列腺癌数据集不同分类器泛化性能测试结果

	C4.5	NB	SVM	HN	FG-HN	SF-HN
2-bt	81.62%	90.44%	96.77%	85.22%	86.86%	95.18%
3-bt	80.88%	90.15%	92.29%	85.06%	86.54%	93.51%
4-bt	77.45%	89.2%	85.29%	84.71%	85.29%	89.25%
5-bt	62.35%	87.65%	83.53%	84.51%	85.00%	88.18%

5 结论

本文提出了一种子空间融合演化超网络模型. 通过将子空间概念引入到演化超网络中,把超边包含的特征看作是输入模式空间的子空间,在子空间进行超边覆盖,减弱了模型对超网络初始化效果的依赖;同时通过超边子空间覆盖和子空间融合,加入样本间的关联信息,提高了模型对未知样本的分类效果和泛化性能. 本文根据部分替代整体思想设计了分类器泛化性能测试实验,并提出了评价分类器泛化性能的方法. 通过四个 DNA 微阵列数据集下的对比实验表明,本文方法的识别率和在小样本训练集下的泛化能力均优于其他传统模式识别方法.

参考文献

[1] 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.

[2] Alon U, Barkai N, Notterman D, 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.

[3] Reunanen J. Overfitting in making comparisons between variable selection methods[J]. Journal of Machine Learning Research, 2003, 3: 1371 – 1382.

[4] Cho S B, Won H. Cancer classification using ensemble of neural networks with multiple significant gene subsets[J]. Applied Intelligence, 2007, 26(3): 243 – 250.

[5] Mundra P A, Rajapakse J C. SVM-RFE with MRMR filter for gene selection [J]. IEEE Transactions on Nanobioscience, 2010, 9(1): 31 – 37.

[6] Prasartvit T, Banharnsakun A, Kaewkamnerdpong B, et al. Reducing bioinformatics data dimension with ABC-kNN [J]. Neurocomputing, 2013, 116: 367 – 381.

[7] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification[J]. Applied Bioinformatics, 2003, 2(3 suppl): 75 – 83.

[8] Zhang B T. Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory[J]. IEEE Computational Intelligence Magazine, 2008, 3(3): 49 – 63.

[9] Kim S J, Ha J W, Zhang B T. Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes[J]. Journal of Biomedical Informatics, 2014, 49(6): 101 – 111.

[10] Park C H, Kim S J, Kim S, et al. Use of evolutionary hypernetworks for mining prostate cancer data[A]. Proceedings of the 8th International Symposium on Advanced Intelligent Systems[C]. Springer, 2007. 702 – 706.

[11] 王进,张军,胡白帆. 结合最优类别信息离散的细粒度超网络微阵列数据分类[J]. 上海交通大学学报,2013,47(12):1856 – 1862.
Wang Jin, Zhang Jun, Hu Bai-fan. Optimal class-dependent discretization-based fine-grain hypernetworks for classification of microarray data[J]. Journal of Shanghai Jiaotong University, 2013, 47(12): 1856 – 1862. (in Chinese)

[12] Wang J, Huang P L, Sun K W, et al. Ensemble of cost-

sensitive hypernetwork for class-imbalance learning[A]. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics [C]. IEEE, 2013. 1883 – 1888.

[13] 孙焕良,鲍玉斌,于戈. 一种基于划分的孤立点检测算法[J]. 软件学报,2006,17(5):1009 – 1016.
Sun Huan-liang, Bao Yu-bin, Yu Ge. An algorithm based on partition for outlier detection[J]. Journal of Software, 2006,17(5):1009 – 1016. (in Chinese)

[14] Gordon G J, Jensen R V, Hsiao L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma [J]. Cancer research, 2002, 62 (17): 4963 – 4967.

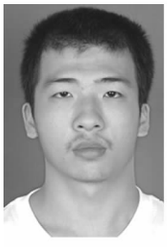
[15] Singh D, Febbo P G, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002,1(2):203 – 209.

[16] Liu L, Wong K C, Wang Y. A global optimal algorithm for class-dependent discretization of continuous data[J]. Intelligent Data Analysis, 2004,8(2):151 – 170.

[17] 张海,徐宗本. 学习理论综述(I):稳定性与泛化性[J]. 工程数学学报,2008,25(1):1 – 9.
Zhang Hai, Xu Zong-ben. A survey on learning theory (I): stability and generalization[J]. Chinese Journal of Engineering Mathematics, 2008, 25 (1): 1 – 9. (in Chinese)

作者简介

王　进　男,1979 年 1 月出生于重庆,教授. 主要研究方向为数据挖掘、机器学习.
E-mail:wangjin@ cqupt. edu. cn



刘　彬(通信作者)　男,1989 年 11 月出生于河北保定,硕士研究生. 主要研究方向为数据挖掘.
E-mail:nanfeizhilu@163. com