基于基因表达小样本数据的级联森林分类模型

范怡敏 齐 林 帖 云
(郑州大学信息工程学院 河南 郑州 450000)

摘 要 针对分类模型在处理基因表达小样本高维度数据集上存在的分类准确性不足、过拟合、计算复杂度大等问题,提出一种改进模型 Two Boosting Deep Forest(TBDForest)。在多描部分采用均等式特征利用方法对原始特征进行变换;在分类过程中考虑到模型所集成的每个森林的拟合质量,将上层最重要的部分判别特征输入到下一级联层,在层间改善类分布问题;对原级联层采用子层级联的结构,增加样本训练机会,减少训练开销,避免模型对参数的依赖。通过在五种疾病基因表达小样本数据集上的验证结果表明,改进的模型增强分类算法在小样本数据集的分类性能上达到了更好的分类效果。

关键词 基因表达数据 深度森林 小样本 分类模型

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.028

CASCADING FOREST CLASSIFICATION MODEL BASED ON GENE EXPRESSION SMALL SAMPLE DATA

Fan Yimin Qi Lin Tie Yun
(*School of Information Engineering*, *Zhengzhou University*, *Zhengzhou 450000*, *Henan*, *China*)

Abstract In order to solve the problems of classification accuracy, over fitting and calculation complexity in the processing of high-dimensional data set of small samples of gene expression, an improved model, two boosting deep forest (TBDForest), is proposed. In the multi description part of the model, the original features were transformed by using the equality of feature utilization; in the process of classification, considering the fitting quality of each forest integrated by the model, the most important part of the distinguishing features in the upper layer were input to the next cascade layer to improve the class distribution between layers; for the original level, the sub level cascade structure was adopted to increase the training opportunities of samples, reduce the training cost and avoid the dependence of the model on parameters. The validation results on 5 disease gene expression small sample datasets show that the improved model enhances the classification performance of the classification algorithm in the small sample datasets, and achieves a better classification effect.

Keywords Gene expression data Deep forest Small sample Classification model

0 引 言

目前,对生物医学和生物信息学数据的分析与分类越来越重要,如:疾病的诊断,癌症的分类[1],物种分类等。如今,在基因表达水平上的数据集常被用来鉴定和提取一些生物信息;生物医学通过基因分析来了解疾病发生与发展机制以及可能性等问题,从而进一步进行基因诊断、亚型分类等。然而,多数生物信息学数据的获取非常困难,通常只能得到小规模的样本数据集。基因表达数据集普遍存在样本个数少、维度高、噪声大等问题,复杂的分类处理模型很容易导致过拟合。当前常用于分类问题的典型机器学习算法包括:支持向量机(Support Vector Machine,SVM)[2]、随机森

收稿日期:2019-06-26。国家自然科学基金项目(61640214)。**范怡敏**,硕士生,主研领域:机器学习,生物信息处理。**齐林**,教授。**帖云**,教授。

林(Random Forest,RF)[3]、深度森林[4](Deep Forest),以及一些集成方法分类模型,例如 Adaboost[5]、XGBoost[4]等。文献[2]使用了 SVM 对基因表达数据进行分类。近年来深度神经网络(Deep Neural Network,DNN)、卷积神经网络(Convolutional Neural Network,CNN)[6]在很多应用领域中取得了较大的发展。然而,尽管深度学习网络模型为数据分析提供了有利的方法,但是由于神经网络算法自身的一些特点,在基因等小规模数据集中的应用受到了限制:一方面,深度神经网络模型比较复杂,通常需要大量的数据进行训练,而基因数据的样本量通常不足;另一方面,深度神经网络模型中参数过多,多数情况下模型的性能受参数影响较大[6]。因此其在小规模数据集的分类中通常难以获得预期的性能。为了改善深度神经网络的这些缺陷[6],2017 年出现了深度森林、多粒度级联森林(multi-Grained Cascade Forest,gcForest)模型,这是基于深度模型提出的一种通过集成学习方法提高分类性能的结构,能有效解决小样本数据分类等问题。文献[6]的实验结果验证了该模型相比于深度神经网络模型,能够避免训练所需数据量大、模型复杂性高、超参数过多等问题,可以取得更好的分类效果。但原始模型依然有很多需要改进的地方,例如:对于小样本高维度数据集,模型易有过拟合现象;分类精度较低;模型对所集成的森林多样性衡量不足,未对每个分类单元的分类质量进行区分;模型的泛化能力和分类稳定性有待提升等。

现今集成学习被广泛应用,集成学习方法可以通过将多个学习模型组合,使组合后的模型具有更强的泛化能力[7]和更好的效果。综上所述,本文在原始深度森林模型的基础上提出一种改进模型 TBDForest,主要对多粒度扫描部分作了特征均等性利用的优化以及级联森林部分两个方面的改进。在五组基因表达 DNA 微阵列小样本数据集上进行特征选择后进行分类验证实验。实验结果显示改进后模型处理小规模数据集时的分类性能相比于常用的支持向量机、随机森林、gcForest、XGBoost、Adaboost 等方法有所提升,进一步实现分类模型在基因表达小样本数据中的应用。

1 传统方法与原理

1.1 特征选择

基因表达数据通常有数千甚至上万个特征基因,有高维度的特点,然而在这些特征中只有小部分基因与癌症亚型分类、疾病判别等相关,其余大部分是冗余或噪声特征,因此本文先对基因表达数据使用特征选择方法进行数据降维处理。Least absolute shrinkage and selection operator(Lasso)算法,是一种基于惩罚方法对样本数据进行变量选择,通过将原本的系数压缩,把原本非常小的系数压缩至零,从而将这部分系数所对应的变量视为不显著的变量并直接舍弃[8]。这种方法能够在保持原始基因的分类准确性的同时选择出重要的基因,降低时空消耗,更易于测试分类器性能。本文中实验数据样本小维度高,先将原始数据通过 Lasso 算法进行特征选择降维处理,然后使用选出的重要基因特征进行分类。

1.2 随机森林模型

Breiman 等[3]提出了随机森林算法 RF,其构建在单一决策树基础上,同时又将单一决策树方法进行延伸和改进,其基本思想是构造多棵决策树,组成一个森林,之后通过这些决策树共同决定输出的类别。整个 RF 算法中有两个随机性的过程:(1) 原始输入的数据随机从所有训练数据中有放回地选出一些建立一个决策树;(2) 建立每个决策树所用的特征是从整体的特征集随机性选取[9]的。这两方面的随机非常有利于 RF 模型避免过拟合。

RF 是综合考虑多个决策树而形成的一种基于集成学习思想的机器学习方法。每个森林通过多棵决策树对样本进行训练,由每棵决策树给出分类值,然后按照少数服从多数的原则表决完成最终的分类,不仅被用于分类还可以解决回归问题。RF 的投票决策过程如下:

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (1)$$

式中:$H(x)$表示组合分类模型;h_i 表示单棵决策树;Y 为输出变量;$I(\cdot)$为指示性函数。算法根据最大投票判断得票数最多的一类作为最后的分类结果[8]。

随机森林算法实现比较容易,不用设置过多参数,且应用广泛。随机森林对数据随机选取和特征随机选取这些随机性的设置,使得随机森林有非常好的抗噪声性能,也不容易过拟合。RF 能够处理高维度的数据,对不同数据集的适应能力强,既能处理离散型数据,也能处理连续型数据。RF 的训练效率高,能获得各个变量的重要性排序,训练时可以检测到特征相互之间的影响,从而使用并行化方法。随机森林的生成步骤如下:

(1) 从原始训练集中随机、有放回地采样 p 个训练样本,进行 p 次采样后生成 p 个训练集。

(2) 用 p 个训练集分别训练 p 个决策树模型。

(3) 将产生的 p 个决策树建立为随机森林。

(4) 对于分类问题,测试的样本由 p 个决策树以投票表决方式产生最终的分类结果。

1.3 深度森林模型

深度森林、多粒度级联森林是周志华教授提出的多个森林组成的深度树集成算法。该模型主要包括两个部分:多粒度扫描(Multi-Grained Scanning)部分和级联森林(Cascade Forest)部分[6]。模型主要有以下几个方面优势[10]:

(1) 模型级数自动调节,可扩展性强;

(2) 超参数少,且模型对其不敏感;

(3) 有很低的训练消耗,不仅可用在大规模数据集上,还能用在小样本数据集中;

(4) 可以进行并行处理。

1.3.1 多粒度扫描模块

受神经网络影响,gcForest 模型通过多粒度扫描流程处理数据特征关系,以增强级联森林部分的性能[6]。该模块使用不同尺寸的滑动窗口进行扫描,首先对原始的输入数据提取局部特征,产生一系列局部低维特征向量,然后经过森林的集合(随机森林和完全随机森林)训练出类向量[5]。例如,对于有 c 个类别的分类问题,一维特征向量长度为 n,长度是 m 的窗口每次滑动一个单位长度,产生 $n-m+1$ 个 m 维特征向量的数据子集,经过一个随机森林和一个完全随机森林后产生长度为 $2c(n-m+1)$ 的类向量;对于一个 $n\times n$ 的图像数据,$m\times m$ 大小的窗口一次滑一个单位尺寸,将产生 $(n-m+1)^2$ 个 $m\times m$ 的特征向量数据子集,经过一个随机森林和完全随机森林后将变成 $2c(n-m+1)^2$ 的类特征向量。将这些特征向量与初始样本特征组合起来,输入后面级联森林中[11]。深度森林模型的多粒度扫描模块如图 1 所示。

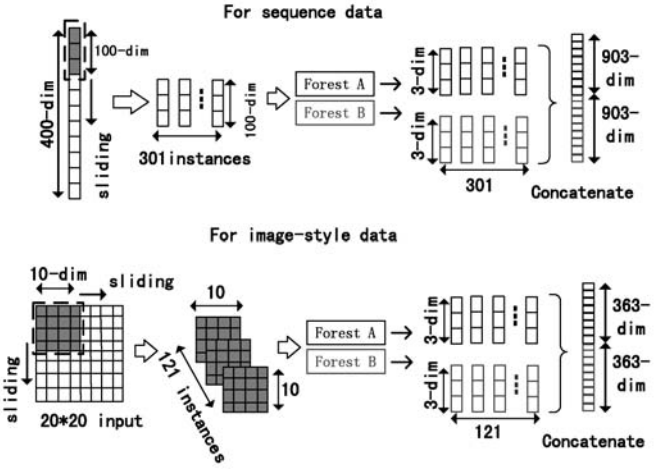


图 1　深度森林模型的多粒度扫描模块

1.3.2 级联森林模块

级联森林部分是一个通过加入新特征来对原始特征进行扩展的结构,由多个级联层组成,每个级联层包括两个随机森林和两个完全随机森林,如图 2 所示。原始特征经过每个级联层后的输出与原特征组合成扩展后的特征作为下一个级联层的输入[6]。该模型在一级结束后做一个分类性能的测试,然后继续生成下一级,当扩展一个新的级联层后,将整个模型的性能在验证集中进行测试,若没有显著的分类性能增益,训练过程终止[11],级联层数就确定了。级联结构增加了模型的深度而不引入额外的参数,通过评估每层的性能自适应地确定级联层的数量,因此超参数较少,而且超参数设定具有很好的鲁棒性。

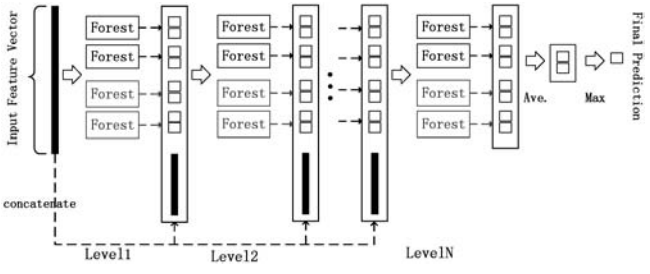


图 2　深度森林模型的级联模块

2 小样本数据集分类模型的改进

2.1 均等性多粒度扫描

原始深度森林模型的多粒度扫描部分对原始输入特征进行变换以加强特征表达能力,通过滑动窗口来扫描初始输入特征。用窗口滑动提取的实体分别训练一个随机森林与一个完全随机树森林,然后将这两种森林输出的类特征向量组合为初始输入特征的转换特征向量。

多粒度扫描部分处理空间上相关联的特征(如图像匹配数据、语音识别数据等)时具有明显的效果,但在空间上无相关的特征数据(如疾病判别、文本分类等)的应用中就可能丢失一些重要信息[10]。原因是多粒度扫描处理空间无关联特征时在某种程度上降低了两端(第一个与滑动的最后一组)特征重要性。多粒度扫描时,首个特征和滑动窗口滑到的最后一组特征都只被扫描了一次,即:这些特征都仅被利用了一次,如果这些特征的重要性比较大,该模块则无法有效对这些重要特征进行利用。

针对这一特征利用不均等性的问题,本文做相应的改进:假设有一个 400 维的向量,利用大小为 100 的滑动窗口,滑动步长为 1,则 301 次滑动后获得 301 个 100 维类向量,在此基础上加入一组 100 维的类向量,

即有 302 组 100 维特征向量,其中第 302 组 100 维特征的前 99 个特征是第 301 组的后 99 个特征,最后一个特征为原始特征的第一个特征,这样第一个与最后一组的后 99 个特征就有与其他特征均等的利用机会,所有特征具有相同的重要度,因此不会丢失重要特征信息,从而可以将原始特征信息均等地传输到随机森林与完全随机森林部分。

2.2 对原始模型添加拟合质量

深度森林原始模型的一些缺点可能会限制其在生物学基因数据集上的效果:集合中的每个森林对最终预测都有相同的贡献,在学习过程中未考虑拟合质量。在小规模数据中模型的最终预测可能受到低质量森林投票结果的影响。因此,基于这种新型的深度结构算法,本文使用改进的级联结构做逐层的表征学习,增强特征表达能力。

原始深度森林模型中级联部分包含的随机森林和完全随机森林都是决策树的集合,均是由随机选择一个特征在决策树的各节点来分割,树不断生长,每个决策树输出一个类向量,最后随机森林组合所有决策树的投票结果后取平均值,得到森林整体的分类结果。本文的级联网络中各层使用两个随机森林和两个完全随机森林,两种森林均由 500 个决策树以及完全随机决策树构成。每个决策树决策过程[12]如图 3 所示,假定有三个类,n 个决策树,每个决策树将确定一个三维类向量,然后取 n 个三维类向量的平均值,最后得到最大值对应的类别作为决策树最终的分类结果。

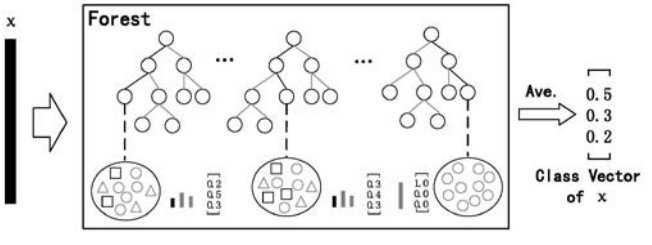


图 3 随机森林集成决策树生成类向量

随机森林中的子树是在整体特征中随机选取的部分特征,因此这些决策树彼此不同,并且各随机森林也有各自的多样性。那么,选取合适的判别特征来构建决策树的分裂点非常重要。然而在原始模型的级联层中,每个森林对最终预测结果的贡献是相同的,在学习过程中未考虑各自的拟合质量,忽略了一些重要特征,因此可能导致在小规模数据集中的模型整体性能对森林数量较为敏感。本文改进时考虑到随机森林在作特征选择时,能够隐含地提供分类过程中的特征相对重要性排序列表,从而新模型基于级联森林模块进行如下的改进:

(1) 选取各层中每个森林的前 k 个重要特征;

(2) 分别计算各森林这 k 个特征的标准差作为一个新特征;

(3) 将这些新特征与该层输出的类分布矢量以及原始特征结合起来;

(4) 将结合后的特征作为下一层的输入。

上述过程通过考虑模型中各森林对最终分类结果的不同影响,以新特征的形式加入到级联森林模块中,从而在层间传输高质量的判别特征,达到提升分类性能的目的。

2.3 深度森林级联层的改进

对于深度神经网络模型,网络的层数往往比层中神经元的个数对整体结构影响更大,基于这种思想,本文采用级联层展开的结构,在标准深度森林模型的级联层中使用子层级联的方法,对级联森林各层进行分子层的改进。将每个级联层改为两个子层级联的形式,原来各层所包含的两个随机森林、两个完全随机森林平均地放在两子层中,即每个子层包含两种森林各一个,如图 4 所示。这种分层监督学习的方法能够获得更精确的分类特征向量,该结构能够进行并行化计算,增加模型训练机会,有明显的效率和性能优势。

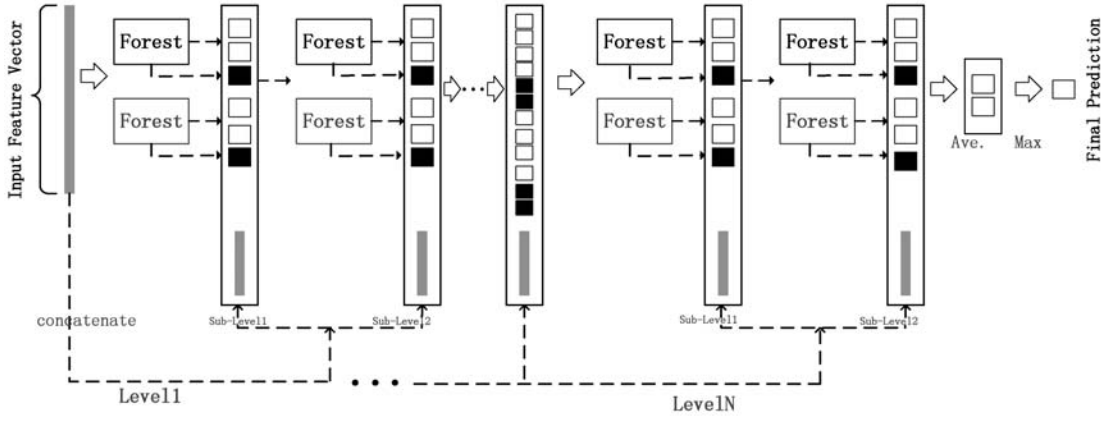


图 4 改进模型级联部分结构

2.4　TBDForest 模型

综合以上三个方面,本文提出的 TBDForest 模型整体结构如图 5 所示。假设输入的原始数据是 400 维,在多粒度扫描模块中分别使用大小为 100、200、300 的滑块进行滑动,得到 302、202、102 个 100 维、200 维、300 维的数据特征。将这些特征输入到一个随机森林与一个完全随机森林的级联中。假设有两种类别,那么,每个 100 维的特征向量被变换为两维的类向量,也就得到一个 604 + 604、404 + 404、204 + 204 的概率分布,将这些概率向量组合起来就得到 2 424 维的数据特征向量。完成了多粒度扫描过程后将得到的 2 424 维数据输入改进的级联结构中。假设选取每个森林的前三个重要特征(k 的值为 3)来提取偏差特征。第一个子层中的每个森林输出各自的类分布以及标准偏差特征向量,然后与该子层的输入特征组合在一起,第一子层就输出 2 426 维特征向量,作为训练数据输入下一子层,第二个子层重复第一个子层的过程,最后输出两个子层的类分布和偏差特征作为级联部分第一层的输出。以后的各层依次重复上述过程,直到模型的性能验证结果表明可以终止级联层。

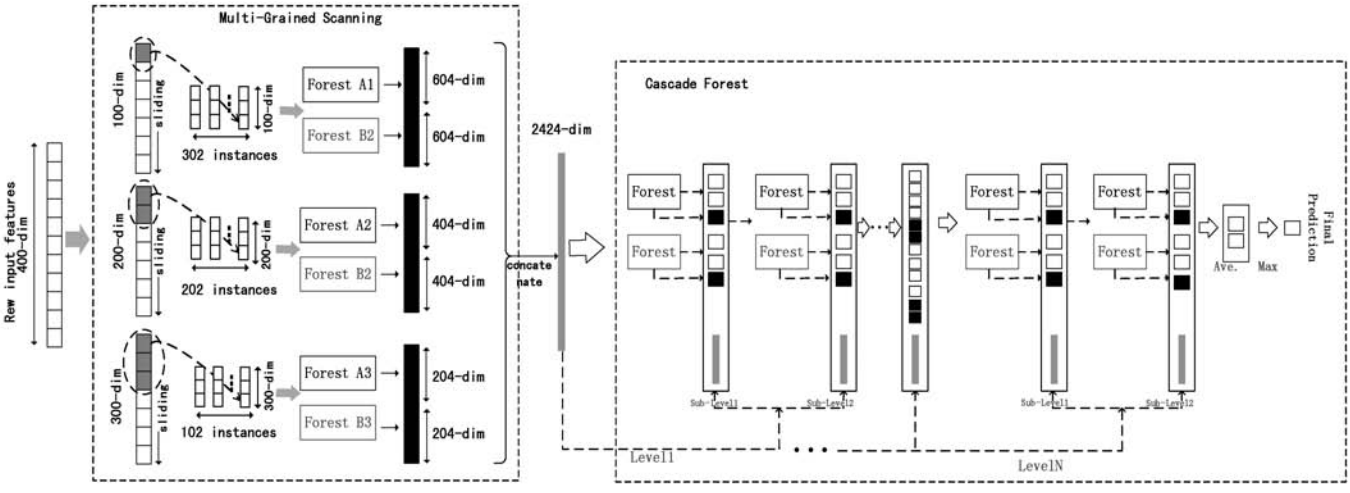


图 5　TBDForest 模型的整体结构

3　实　验

3.1　实验环境

为检验 TBDForest 算法对于基因表达小样本数据集的适用性,本文结合 5 个数据集使用 Python 语言来实现特征选择与分类实验。实验使用十折交叉验证方法,模型训练前将数据随机划分成 10 份,每次取 9 份来用于分类模型的训练,留下 1 份数据用来检测模型性能,每次做十轮实验,十轮结果的平均值就是最终分类的结果[10]。本文对各数据集先特征选择,之后进行分类性能测试。

3.2　实验数据

本实验下载了 EBI、NCBI 数据库中的五组高维基因表达 DNA 微阵列数据来验证改进模型的性能。这五种疾病数据的样本量均较小。数据的基本信息如表 1 所示,Gravier 和 West 是乳腺癌数据、Pomeroy 是中枢神经系统疾病数据、Alon 是结肠癌数据、Gordon 是肺癌数据,这些数据用于检测癌症亚型或患病与否的分类[13]。

表 1　数据集的基本信息表

数据集	样本规模	属性个数	类别数
Gravier	168	2 905	2
Pomeroy	60	7 128	2
West	49	7 129	2
Alon	62	2 000	2
Gordon	181	12 533	2

3.3　评价标准

实验结果综合考虑准确率(Accuracy)、精确度(Precision)、召回率(Recall)、F-1 Score 这四个分类方法中最常用的分类性能评价指标[14]。这些指标建立在混淆矩阵的基础上,如表 2 所示。

表 2　混淆矩阵

分类	实际为正类	实际为负类
预测为正类	TP	FP
预测为负类	FN	TN

准确率(Accuracy)[14] 即分类准确的样本数量与样本总量的比值。定义如下:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

精确度(Precision)即查准率;召回率(Recall)即查全率[14]。定义如下:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F-1 Score 即一种综合考虑查准率和查全率的分类评价指标,其中查全率与查准率权重相同[15]。定义如下:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

3.4 结果与分析

基因表达数据样本小,特征维度高,直接通过模型分类会导致效果不理想并且缺乏稳定性,如表 3 所示。表中五个数据集中 gcForest 和 TBDForest 模型分类准确率相对较好,但是整体分类准确率不高。因此对五个数据集分别采用 Lasso 算法进行特征选择,然后将选择后的特征数据使用 SVM、RF、XGBoost、Adaboost、gcForest、TBDForest 进行分类测试以及结果比较。数据经过 Lasso 算法挑选的重要特征数量基本信息如表 4 所示。

表 3 特征选择前五个数据集在六种分类方法上的 Accuracy 对比值 %

分类方法	数据集				
	Gravier	Pomeroy	West	Alon	Gordon
SVM	78.627	83.666	83.333	79.157	99.636
RF	69.793	65.045	81.139	73.680	90.837
gcForest	81.424	82.005	87.561	75.430	95.100
XGBoost	70.486	74.018	84.121	73.419	91.355
Ababoost	72.381	71.778	83.206	72.990	89.672
TBDForest	89.973	89.796	90.667	82.911	97.035

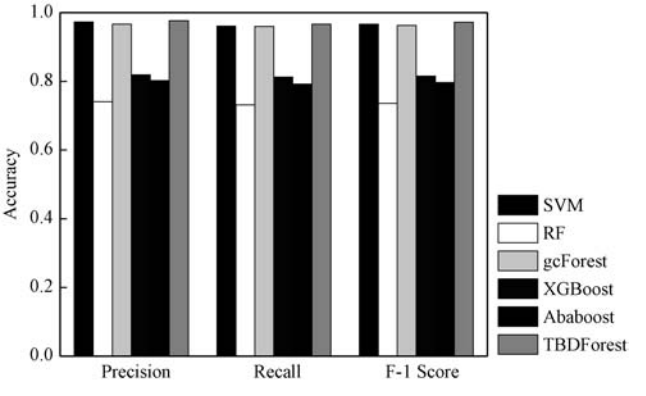
表 4 特征选择后的数据集的基本信息表

数据集	样本规模	属性个数
Gravier	168	128
Pomeroy	60	56
West	49	51
Alon	62	58
Gordon	181	97

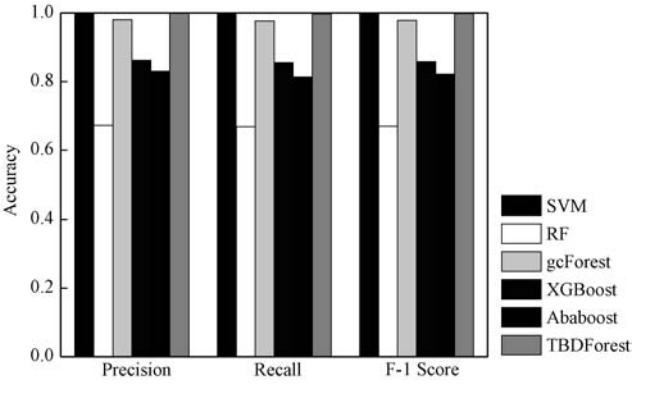
表 5 是五个数据集在几种分类方法上的 Accuracy 值,可以看出 DBForest 模型与传统方法 SVM、RF、gcForest、XGBoost、Adboost 相比具有更高的准确性。图 6 为六种算法在五个数据集上的 Precision、Recall、F-1 Score 指标,从图 6(a)图能够直观地看出,在 Gravier 数据集上 SVM、gcForest、TBDForest 模型的三个指标相当,本文的模型略高,而 RF 与 XGBoost、Adboost 两种集成分类模型效果并不是很好。图 6(b)Pomeroy 数据集上的指标结果显示 SVM 与 TBDForest 几乎可以达到 100% 的分类性能。图 6(c)West 数据集中 RF、gcForest 则有比较高的指标值,改进的模型依然处于领先的效果。图 6(d)的 Alon 数据集上 TBDForest 效果较为明显,优于其他几种方法。图 6(e)中对于 Gordon 数据集,六种方法相差并不大,都能取得良好的分类结果。综上所述,通过 Accuracy、Precision、Recall、F-1 Score 指标对比,TBDForest 模型在 Accuracy、Precision、Recall、F-1 Score 方面均优于其他五种方法,可见本文对模型的改进有效地增强分类性能,有良好的适用性。原因是一方面其充分考虑了特征重要性,另一方面子层增加了训练机会。

表 5 特征选择后五个数据集在六种分类方法上的 Accuracy 对比值 %

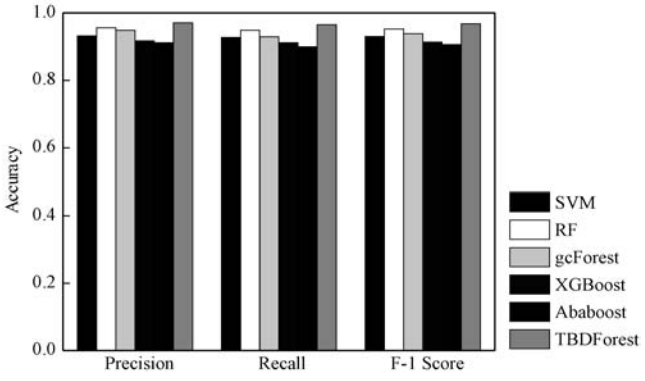
分类方法	数据集				
	Gravier	Pomeroy	West	Alon	Gordon
SVM	98.627	100.000	94.306	98.157	100.000
RF	74.396	67.845	95.941	83.930	99.137
gcForest	97.175	98.943	95.033	98.635	99.811
XGBoost	84.768	86.693	93.025	91.585	99.097
Ababoost	82.907	83.792	92.471	91.473	99.204
TBDForest	98.860	100.000	97.333	99.709	100.000



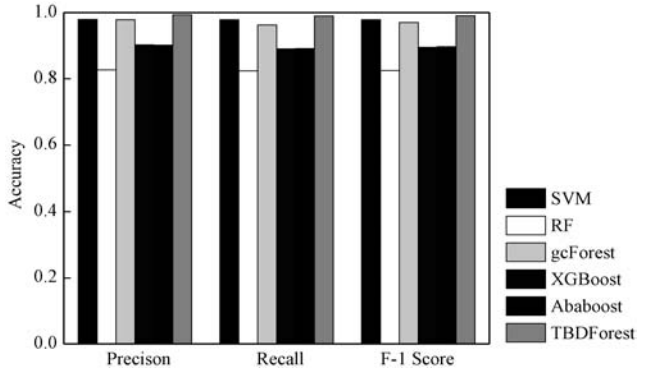
(a) Gravier



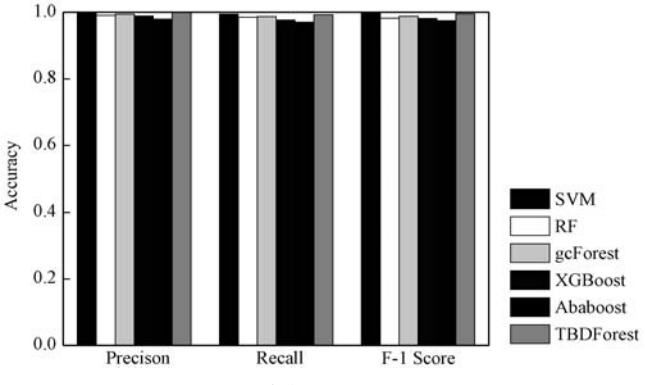
(b) Pomeroy



(c) West



(d) Alon



(e) Gordon

图 6　六种分类模型的对比图

4　结　语

针对目前应用比较广泛的一些分类模型算法对基因表达小样本数据分类时的不足,本文进行分类模型的优化,提出基于深度森林的改进模型 TBDForest,该模型从三个方面进行改进:(1) 采用特征均等性利用的多粒度扫描模块进行特征变换;(2) 考虑每个森林的拟合质量,通过加入标准差特征向量来改善类分布问题;(3) 在每个级联层采用子层结构,增加样本训练机会,进一步提升分类准确性。经过实验验证,改进的模型在基因表达数据小样本数据集上取得相对较高的 Accuracy、Precision、Recall、F-1 Score 值,对于小样本高维度数据有比较好的分类性能。未来将着重考虑将模型应用到各种类型的数据中;考虑所集成森林的多样性;对特征选择方法进行优化,减小模型算法的计算消耗,更好地避免"维度灾难",提高稳定性。

参　考　文　献

[1] Kido S, Hirano Y, Hashimoto N. Detection and classification of lung abnormalities by use of Convolutional Neural Network(CNN) and Regions with CNN features(R-CNN) [C]//2018 International Workshop on Advanced Image Technology(IWAIT). IEEE,2018.

[2] Furey T S, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906 - 914.

[3] Breiman L. Random forests [J]. Machine Learning,2001,45(1): 5 - 32.

[4] Han M, Li S, Wan X, et al. Scene recognition with convolutional residual features via deep forest[C]//2018 IEEE 3rd International Conference on Image, Vision and Computing(ICIVC). IEEE,2018.

[5] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences,1997,55(1):119 - 139.

[6] Zhou Z H, Feng J. Deep forest: towards an alternative to deep neural networks[EB]. arXiv:1702.08835, 2017.

[7] 怀听听. 随机森林分类算法的改进及其应用研究[D]. 杭州:中国计量大学,2016.

[8] 张航飞. 基于 Lasso-PCA 及改进自适应遗传神经网络的电力负荷预测研究[D]. 秦皇岛:燕山大学,2018.

[9] 赵清华,张艺豪,马建芬,等. 改进 SMOTE 的非平衡数据集分类算法研究[J]. 计算机工程与应用,2018,54(18):168 - 173.

[10] 薛参观,燕雪峰. 基于改进深度森林算法的软件缺陷预测[J]. 计算机科学,2018,45(8):160 - 165.

[11] 宫振华,王嘉宁,苏翀. 一种加权的深度森林算法[J]. 计算机应用与软件,2019,36(2):274 - 278.

[12] Zhu Q, Pan M, Liu L, et al. An ensemble feature selection method based on deep forest for microbiome-wide association studies[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). IEEE,2018.

[13] Güçkiran K, Cantürk İ, Özyilmaz L. DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO[J]. Süleyman Demirel üniversitesi Fen Bilimleri Enstitüsü Dergisi,2019,23(1):126 - 132.

[14] 姚登举. 面向医学数据的随机森林特征选择及分类方法研究[D]. 哈尔滨:哈尔滨工程大学,2016.

[15] 王超学,张涛,马春森. 改进 SVM-KNN 的不平衡数据分类[J]. 计算机工程与应用,2016,52(4):51 - 55.