面向不平衡数据的多层神经网络模型

张雪，石志国，刘璇
（北京科技大学计算机与通信工程学院，北京 100083）

摘　要：传统的不平衡数据分类问题往往会因为类间数据不平衡造成分类器的性能下降。利用 AUC（ROC 曲线下的面积）为评价指标，结合单类 F-score 特征选择和遗传算法建立多层神经网络模型，选出对于不平衡数据分类更有利的特征子集，从而建立更适用于不平衡数据分类的深度模型。基于 Tensor Flow 建立多层神经网络模型，通过对 4 组不同 UCI 数据集进行测试，并与传统的机器学习算法如朴素贝叶斯、*K* 最近邻、神经网络等进行对比验证。实验证明，所提模型在处理不平衡数据分类问题上的表现更优秀。

关键词：不平衡数据；单类 F-score 特征选择；遗传算法；多层神经网络
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.2096-3750.2018.00055

Multilayer neural network model for unbalanced data

ZHANG Xue, SHI Zhiguo, LIU Xuan
School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract: Classification of unbalanced data often has low performance of the classifier because of the unbalance of data between classes. Using AUC (the area under the ROC curve) as evaluation index, combined with one class F-score feature selection and genetic algorithm, a multilayer neural network model was established, and a more favorable feature set for unbalanced data classification was selected, so as to establish a deeper model suitable for classification of unbalanced data. Based on Tensor Flow, a multilayer neural network model was established. Using four different UCI datasets for testing, and comparing with the traditional machine learning algorithms such as Naive Bayesian, *K*NN, neural networks, etc, the performance of the proposed model built on the unbalanced data classification is more excellent.

Key words: unbalanced data, one class F-score feature selection, genetic algorithm, multilayer neural network

1　引言

随着信息技术的快速发展和存储设备容量的不断扩大，越来越多的行业开始意识到数据的重要性。新的数据形式不断出现，尤其是不平衡数据的广泛存在，给传统的机器学习算法带来了极大的挑战。尽管现有的机器学习算法如决策树[1-5]、支持向量机[4,6-8]、贝叶斯网络[2]、最近邻算法[3,9]等在处理很多数据分类问题方面表现优秀，但在处理不平衡问题时，往往会因为类间数据不平衡使学习到的分类器倾向于将样本分成多数类，造成对少数类的识别精度降低。在很多应用领域中，如金融欺诈检测、网络入侵检测[10]、骚扰电话的识别[11]、文本自动分类[12]以及医学癌症分类[13]等，将少数类误分为多数类的代价往往远大于将多数类误分为少数类的代价。如何处理不平衡数据、提升少数类的识别精度也成为目前机器学习领域一个新的研究热点。

目前，已有大量文献研究不平衡数据的分类问题，主要从数据层面和算法层面进行研究。数据层面通过采样[14]或欠采样技术[15]等重新平衡类间比例分布，使每类的样本比例大致相同。例如，Chawla 等[16]提出的合成少数过采样技术（SMOTE，syn-

收稿日期：2018-03-10；**修回日期：**2018-05-15
基金项目：国家重点研发计划基金资助项目（No.2016YFC0901303）
Foundation Item: The National Key R&D Program of China (No.2016YFC0901303)

thetic minoritye over-sampling technique）算法，利用已有少数类样本及其近邻，合成新样本数据，可以看作一种改进的过采样方法。算法层面则通过改进现有的分类算法提升分类器对少数类的识别能力。例如，Wu 等[8]通过对支持向量机（SVM, support vector machine）的边界进行调准，修改 SVM 的核函数来调整分类面，改进了传统 SVM 算法在处理不平衡数据时分类面偏向少数类的缺点，从而提升对少数类的识别精度。除此之外，一些集成学习方法如 Adaboost 等在处理不平衡分类问题上也取得了不错的成果[17]。

但上述的 2 类方法都有各自的长处和不足。数据层面的抽样方法，无论是过采样还是欠采样都在一定程度上改变了原始数据的分布，样本增加导致分类器过拟合，样本减少导致有效信息丢失。特别是对于医学领域来说，出于数据真实性的考虑，一些重新构造新样本的算法如 SMOTE 算法等并不合适。而基于传统分类器的改进方法，其基本原理是使分类器更注重少数类，因而当少数类样本不能反映其真实分布时，这类算法容易出现过学习的现象[18]。

针对上述方法存在的难点，本文并没有选择采用上述 2 类方法或其结合方式，而是创新性地选择从偏向少数类分类的特征子集出发，基于 UCI 医学数据进行研究，提出了一种基于单类 F-score 特征选择和遗传算法改进的多层神经网络模型。本文使用单类 F-score 特征选择替代传统 F-score 对于特征的评价方法，以此选择更新遗传算法的种群个体，在使用多层神经网络进行适应度函数计算时，对于每一个个体网络输入都会以 7:3 的概率重新选出训练集和测试集，同时遗传算法每次通过交叉变异生成新种群时都具有一定随机性，从而确保本文模型能探索的空间尽可能大，且模型能在更大的空间里选出最优的分类子集。从医学上不平衡数据的处理来说，本文使用单类特征选择和遗传算法改进传统多层神经网络的方法，在一定程度上为多层神经网络解决不平衡数据分类问题提供了一种新的研究思路。实验在不同 UCI 医学数据集进行测试，并与传统的机器学习算法如朴素贝叶斯、*K* 最近邻（*K*NN，*K*-nearest neighbor）、神经网络等进行对比验证，结果表明本文模型在处理不平衡数据分类问题上的表现更优秀。

2 相关工作

2.1 单类 F-score 特征选择

理论上，随着特征的增加分类器的识别性能会更好。但在实际建模过程中，训练数据一般具有有限的、过多的特征，尤其是一些与类别不相关的特征和冗余特征，会大大降低分类器的学习速度，同时也会导致分类器对训练数据的“过适应”问题[19-21]的出现。因此需要进行特征选择，考虑如何从原始特征集合中选出一个最优特征子集。

传统的单类 F-score 特征选择方法采用向后选择的启发式方法，通过衡量原始特征集合中每个特征的 F-score 值，选出对分类识别最有效的特征子集。研究证明，F-score 方法在处理二分类问题上具有简单有效的特点[22]。其算法具体描述如下。

给定训练样本集 X_k, k=1,2,…,N，n_+和 n_-分别代表正类和负类的样本大小，则样本的第 i 个特征的 F-score 值 $F(i)$被定义为

$$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2} \tag{1}$$

其中，$\bar{x}_i$ 代表第 i 个特征在整个数据集上的平均值，$\bar{x}_i^{(+)}$ 和 $\bar{x}_i^{(-)}$ 分别表示第 i 个特征在正类和负类上的平均值，$x_{k,i}^{(+)}$ 为第 k 个正类样本点的第 i 个特征的特征值，$x_{k,i}^{(-)}$ 为第 k 个负类样本点的第 i 个特征的特征值。在式(1)中，分子可以看作 2 类样本的近似类间距之和，而分母表示总类内样本的协方差，因此，F-score 值越大，表示类间距离远、类内紧密，则此特征在 2 类样本中的分辨能力越强[22]。如图 1 所示。

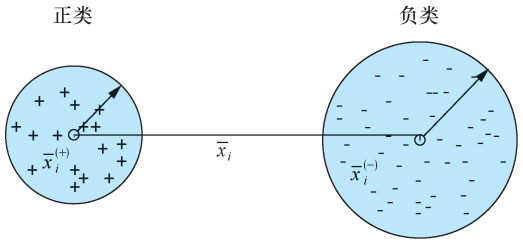

图 1 二分类中的传统 F-score 特征选择

但是当正负类样本比例相差过大时，F-score 算法会倾向于选择能识别多数类的特征而忽略少数类。因此，在处理非平衡数据问题上，应充分考虑

少数类的信息在所选特征子集中的表达问题。本文选择改进后的单类 F-score 特征选择法。该算法的具体描述如下[23]。

给定训练样本集 X_k, k=1,2,⋯,N, n_+和 n_-分别代表正类和负类的样本大小，则样本的第 i 个特征的 F-score 值 $F(i)$被定义为

$$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2} \tag{2}$$

其中，$\bar{x}_i$ 代表第 i 个特征在整个数据集上的平均值，$\bar{x}_i^{(+)}$ 和 $\bar{x}_i^{(-)}$ 分别表示第 i 个特征在少数类和多数类上的平均值，$x_{k,i}^{(+)}$ 为第 k 个少数类类样本点的第 i 个特征的特征值。在式(2)中，分子可以看作 2 类样本的近似类间距之和，而分母表示少数类内样本的协方差。当第 i 个特征在整个数据集上的平均值 $\bar{x}_i$ 不变时，如果 $\bar{x}_i^{(+)}$ 和 $\bar{x}_i^{(-)}$ 距离越接近，说明实际上第 i 个特征在 2 类样本之间的取值范围相差不大，第 i 个特征对于 2 类问题分类来说不具有明显的区分性，甚至有可能是无关特征。故而分子越大，表示该特征在多数类和少数类中的类间距离越大，在 2 类之间就具有更好的区分性。而分母越小，意味着每一个求和项就越小，少数类样本集整体样本的第 i 个特征取值和第 i 个特征在少数类上的平均值 $\bar{x}_i^{(+)}$ 越接近，即少数类样本在该特征维度上围绕 $\bar{x}_i^{(+)}$ 密集分布。理想情况下，分子趋于 0 时，代表少数类样本在该特征维度上趋于少数类的中心样本，实际模型训练和测试过程中，测试样本和训练样本在第 i 个特征上越相似，越容易向同一类靠近。因此，特征的单类 F-score 值越大，代表该特征在 2 类问题上分类能力更强，同时识别出少数类的能力更好。

2.2　遗传算法与多层神经网络

遗传算法（GA, genetic algorithm）是一种模拟生物群体进化的优化算法，1975 年由美国的 Holland 教授[24]首先提出。GA 是一种启发式的寻优算法，通过模拟生物进化的自然选择和变异过程，一代一代循环迭代进化，使种群朝着最优解的方向进化。遗传算法具有一定的自适应性和智能性，在特征选择中，遗传算法常用来解决寻找分类模型最优特征子集问题。

其次，在现实生活中，许多系统的输入和输出之间存在复杂的非线性关系，这类系统往往很难用传统的数理方法建立模型。而一个合理的多层神经网络能够实现对系统输入输出样本的自动学习，理论上可以逼近任意形式的目标函数。由于多层神经网络存在梯度内在不稳定的问题，因而直到 2006 年，Hinton 等[25-26]才在《Science》及《Neural Computation》上发表文章，强调多隐层深度神经网络相比浅层网络具有更优异的特征学习能力，并可以通过逐层无监督的预训练有效解决深度神经网络训练困难的问题。多层神经网络才迎来了新一轮的研究热潮。至今，多层神经网络的应用领域不断扩大，潜力日趋明显，很多由传统机器学习方法无法解决的问题在采用多层神经网络后取得了良好的效果。

3　基于单类特征选择和遗传算法改进的多层神经网络

3.1　整体框架

本文整体框架如图 2 所示。

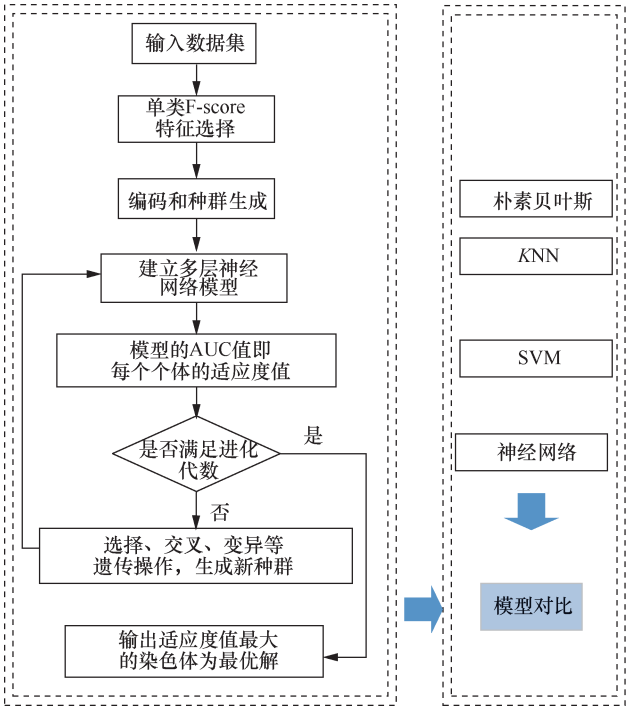

图 2　本文整体框架

本文的主要思想是：首先利用数据集的各个特征进行单类 F-score 特征选择，并依据从高到低的概率生成初始种群，即单类 F-score 值越大的特征，其在初始种群中被选中的概率越大；然后结合遗传算法的选择、交叉、变异等遗传操作，并利用多层神经网络的曲线下面积（AUC, area under curve）值

为评价指标，选出最优特征子集。

本文研究内容主要包括 2 部分：第一部分主要是根据图 2 左侧框图具体实现本文模型，这一部分将在 3.2 节和 3.3 节进行详细介绍；第二部分就是通过模型间的对比验证进行本文模型的评估，这一部分将在第 4 节的实验结果部分得到具体体现。

3.2　构造多层神经网络

本文构造的网络由一个输入层、2 个隐含层、一个输出层组成，2 个隐含层由 RELU 连接的 2 层神经网络实现，采取 Dropout 的方式强迫神经网络学习更多知识，需要丢掉 RELU 的部分结果，保留 75%的网络信息，同时在一定程度上缓解模型的过拟合问题。该网络具体构造如下。

```
def multilayer_perceptron(x, weights, biases):
# 2 个隐含层的多层神经网络
layer_1=tf.add(tf.matmul(x,weights['h1']), biases['b1'])
layer_1=tf.nn.relu(layer_1)
layer_2=tf.add(tf.matmul(layer_1,weights['h2']), biases['b2'])
layer_2=tf.nn.relu(layer_2)
drop_out=tf.nn.dropout(layer_2, 0.75)
out_layer=tf.matmul(drop_out,weights['out'])+biases['out']
return out_layer
```

3.3　详细设计

结合改进后的 F-score 和遗传算法进行特征选择，并结合多层神经网络进行模型评估，本文提出了基于单类 F-score 特征选择、遗传算法和多层神经网络的建模方法。该方法整体实现方案描述如下。

1) 初始种群生成方法

不同于传统遗传算法，本文首先计算每个特征的单类 F-score 值，并根据数值的大小进行降序排列，则初始种群中的个体按照特征单类 F-score 值降序排列的概率生成。这种特征选择方法类似于选择算子中具有排名的转盘式选择算子[27]。根据已有选择算子计算式进行修改[27-28]，若设排在第 i 位的特征被选中的概率为 p_i，则有

$$p_i=\frac{1}{n}\left(a-\frac{bi}{n+1}\right),\quad i=1,2,\cdots,n \tag{3}$$

其中，i 为特征的排名顺序，n 是特征总数量，a、b 是常数，通常 a=1.1，b=0.2（b=2(a−1)）。本文采用二进制编码法勾勒个体的所有特征，如“010101”表示个体的第 2、4、6 位特征被选中。

2) 适应度函数

现实中，少数类往往蕴含更重要的信息。而在样本不平衡的数据集中，即使少数类被完全错分，分类器的准确率可能还是很高。因此，以准确率为评价指标的不平衡数据分类模型并没有太大的实际意义。本文选择多层神经网络的 AUC 值作为个体适应度的评价标准。基本思想是：将每个个体放入适应度函数中进行评估，特征编码为 1 的特征被选中，而特征编码为 0 的个体该特征列全部置为 0，这样就保证了每次多层神经网络的输入维度是一样的。选择后的数据按照 7:3 分为训练集和测试集，使用多层神经网络训练模型，计算测试数据的 AUC 值，即个体的适应度值。算法流程如图 3 所示。

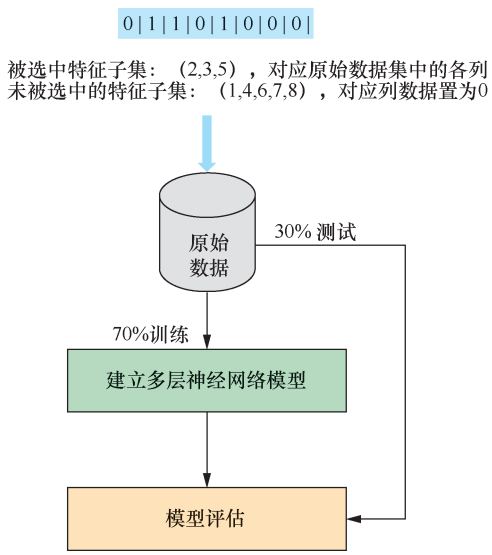

图 3　适应度值计算流程

3) 遗传操作

遗传算法是受达尔文生物进化论启发提出的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。在初代种群生成后，需要按照优胜劣汰的原则，逐代演化，借助于自然遗传学的遗传算子进行组合交叉、变异等遗传操作，产生代表新的解集的种群。该算法基本步骤如下。

```
pop=geneEncoding(pop_size,chrom_length,min_length)  #初始种群产生
for i in range(pop_epochs)
```

```
#终止迭代次数
    obj_value = calobjValue(pop)
    #计算适应度函数值
    fit_value = calfitValue(obj_value)
    #优胜劣汰
    best_individual,best_fit=best(pop,fit_value)
    #选出本次最优
    results.append([best_fit,best_individual])
    #记录每次迭代最优解及最优值
    selection(pop, fit_value) #新种群复制
    crossover(pop, pc) #交叉
    mutation(pop,pm) #变异
print(results)//打印记录结果
```

其中，选择的目的是选择出一些优良的个体遗传到下一代群体中，体现了达尔文的适者生存原则。交叉是最主要的遗传操作，将群体中的各个个体随机搭配成对，对每一个个体，以某个概率（称为交叉概率）把 2 个父代个体的部分结构加以替换重组。通过交叉，遗传算法的搜索能力得以飞跃提高。变异是对群体中的每一个个体，以某一概率（称为变异概率）改变某一个或某一些基因座上的基因值为其他的等位基因，对应于本文个体二进制编码上 0 或 1 的改变。

4) 上述循环结束后，从各模型的评估结果中选出分类结果最优的模型，即对应 AUC 值最大的模型。该模型的特征子集即被选出的最优特征子集。

4　实验和性能分析

4.1　实验配置

本文模型计算框架使用 TensorFlow。TensorFlow 是 Google 旗下开源的深度学习框架。与 Caffe、Theano、Torch、MXNet 等框架相比，TensorFlow 在 Github 上 Fork 数和 Star 数都是最多的，而且在图形分类、音频处理、推荐系统和自然语言处理等场景下都有丰富的应用。支持 Python 和 C++ 2 种编程语言，在 CPU 和 GPU 上均可流畅运行，高度灵活，并且具有很强的移植性。此外，实验室还配置了先进的 CPU+GPU 混合计算集群，为本课题提供了良好的计算平台。

4.2　实验数据集

本文只研究二分类问题，实验采用的数据集均为二分类问题数据集。为检验本方法的有效性，选取 UCI 机器学习数据集中的 4 个数据分布不平衡的数据集作为实验数据。测试数据集描述如表 1 所示。

表 1　　实验数据集描述

数据集	样本大小	少数类个数	特征个数	少数类标识	少数类比例
PIDD(UCI)	768	268	8	1	34.9%
hep	155	32	19	DIE	20.65%
breast-w	699	241	10	Malignant	34.5%
sick-euthyroid	3 163	293	25	sick-euthyroid	9.26%

1) PIDD（pima indians diabetes database），来自 UCI 医学数据；2) hep 是一个肝炎的小集合，在整个数据集中只有 155 个实例的数据，每个实例由 19 个属性描述；3) breast-w，来自 UCI 的乳腺癌数据集；4）sick-euthyroid 数据集的目标是预测甲状腺疾病。收集的数据有 25 个属性，7 个是连续的，18 个是布尔值。数据集包含 3 163 个实例，其中少数类比例为 9.26%。

4.3　不平衡数据评价指标

分类模型中常用准确率作为模型的评价标准，但是对于不平衡数据来说，准确率并不适合。而受试者特性（ROC, receiver operating characteristic）曲线由于不受样本分布影响，因此很多文献中常用 ROC 曲线下面积（AUC）作为不平衡数据分类模型的评价指标。曲线下的面积越大，AUC 值越大，表示模型的分类效果越好。以二分类问题为例，其混淆矩阵如表 2 所示。

表 2　　二分类的混淆矩阵

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

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

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

ROC 曲线以 *TPR* 为纵坐标、*FPR* 为横坐标。其中，*TPR* 表示真实的正例中，被预测正确的比例；*FPR* 表示真实的反例中，被预测正确的比例。最理想的分类器就是对样本分类完全正确，即 *FP*=0，*FN*=0。所以理想分类器 *TPR*=1，*FPR*=0。

4.4 实验结果

1) 本文通过计算特征集合的单类 F-score 值和传统 F-score 值，并通过 matplotlib 模块将其绘制成折线图。其中，纵轴表示 F-score 值，实折线是单类 F-score 值，虚折线是传统的 F-score 值；横轴表示特征集从左到右的特征序号，编号从 0 开始。2 种算法的计算结果分别如图 4~图 7 所示。

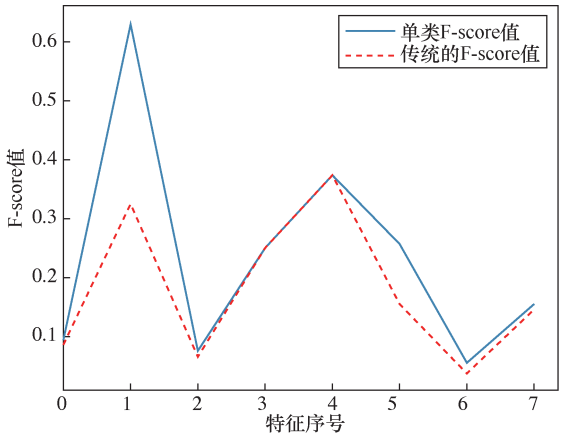

图 4　PIDD 的 2 种 F-score 值比较

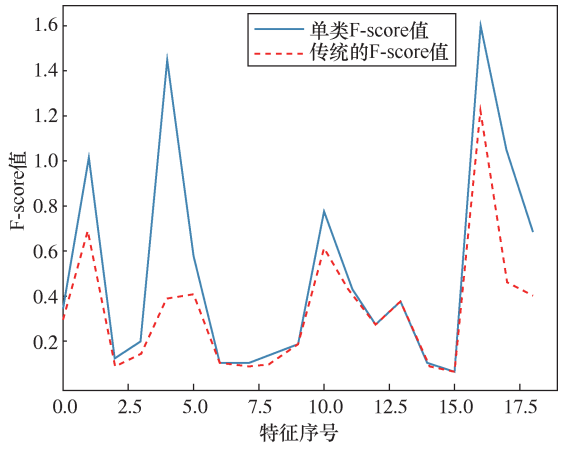

图 5　hep 的 2 种 F-score 值比较

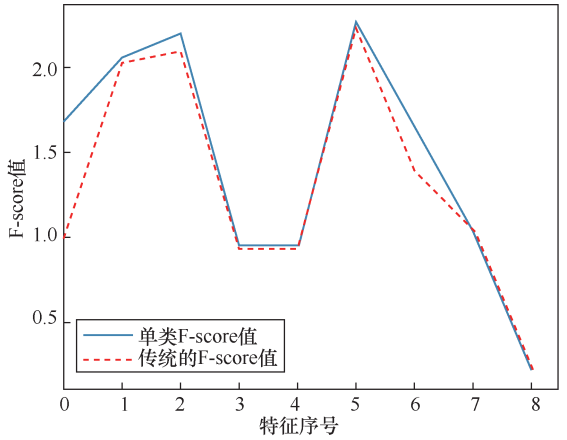

图 6　breast-w 的 2 种 F-score 值比较

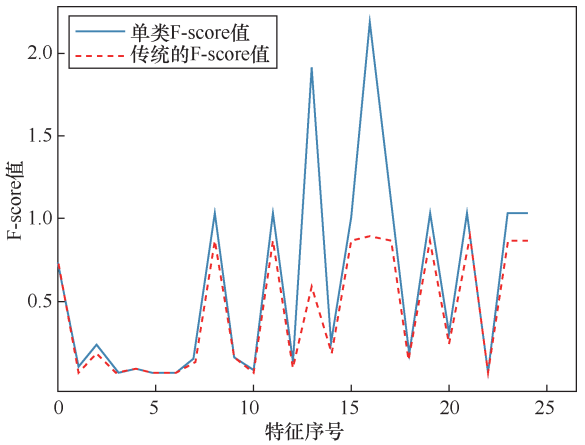

图 7　sick-euthyroid 的 2 种 F-score 值比较

从图 4~图 7 可以看出，对应于同一特征集合，特征的 2 种 F-score 值并不相同。某些特征的传统 F-score 值很小，但单类 F-score 值却很大，表示这类特征可能具有较强的分辨少数类的能力。

2) 本文通过计算测试集数据的 AUC 值进行模型性能评估。将本文模型分别与朴素贝叶斯、*K*NN、SVM、神经网络等进行对比验证。实验所得结果如表 3 所示。

下面是本文模型所选择的特征子集描述。其

表 3　模型的对比验证

数据集	朴素贝叶斯	*K*NN	SVM	神经网络	GA+神经网络	本文模型
PIDD	80.06%	78.81%	76.85%	77.08%	87.96%	87.73%
hep	66.96%	76.07%	70.18%	83.22%	87.86%	89.11%
breast-w	94.99%	94.32%	94.32%	96.07%	96.42%	96.43%
sick-euthyroid	64.13%	73.58%	50.00%	55.08%	66.86%	88.01%

中，数组中的元素位置对应于原始特征集从左到右的特征编号（从 0 开始），0 表示该位特征未被选中，1 表示该位特征被选中，则 4 个数据集所选特征子集如表 4 所示。

表 4　　模型所选特征子集

数据集	所选特征子集
PIDD	[0, 1, 0, 1, 1, 1, 0, 0]
hep	[0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0]
breast_w	[1, 0, 1, 0, 1, 1, 1, 0, 1]
sick-euthyroid	[1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1]

从表 3 和表 4 可以看出，本文模型相比于传统机器学习方法实验所得 AUC 值较大，且本文所选特征子集更利于不平衡数据建立模型，这说明本文模型更为优秀。

3) 本文建模过程中各个数据集的神经网络节点个数如表 5 所示。

表 5　　模型参数设置

数据集	输入层	隐含层 1	隐含层 2	RELU	输出层
PIDD	8	10	12	75%	2
hep	19	22	24	75%	2
breast-w	9	10	12	75%	2
sick-euthyroid	25	28	30	75%	2

此外，从训练时间上来说，本文模型和“遗传算法+神经网络”训练出来的模型相差不大，4 个模型的训练时间都在 3~4 h，远大于朴素贝叶斯、*K*NN、SVM 和神经网络的训练时间。本文模型时间效率整体都偏低，这也是笔者下一步工作亟待改进的要点。

5　讨论

本文模型相比于传统针对不平衡数据的处理算法来说，使用了单类特征选择和遗传算法改进了多层神经网络，使其能够从偏向不平衡数据分类的最优特征子集出发，寻求最优的分类模型。通过遗传算法种群的优胜劣汰和多层神经网络每次训练时训练集和测试集随机选择的方式，使模型特征组合的空间变大了，整体的数据使用效率提高了。但也是由于这个原因，本文模型在时间效率上都偏低。

其次，由于高维度医学数据很难获取，本文选择的是 UCI 医学公开数据集，数据集的特征个数较少，但实际上很多癌症的影响因子往往多达上百种，而且数据集极度不平衡。实验室目前参与的乳腺癌相关研究项目，仅仅乳腺癌的影响因子就至少有 200 个，从身高、体重、腰围、臀围到吸烟、饮食、身体活动、心理活动、血液信息、疾病用药等，对于如此高维度的特征集来说，传统的朴素贝叶斯等可能无法建立起一个合适的模型。本文所提方法由于遗传算法初始种群的生成并不是随机的，而是按照单类 F-score 值从大到小的概率生成，在不平衡数据集维度越高的情况下，其初始种群选出对于不平衡数据分类更有利的特征集合的概率越大。对于这种高维度不平衡数据，本文所提方法的处理能力相比于“遗传算法+神经网络”的处理能力可能更好。但遗憾的是，实验室参与的乳腺癌相关项目正在进行过程中，一期数据尚未收集完成，无法进行实验验证。

此外，本文主要基于医学数据集进行考虑，只分析了不平衡数据的二分类问题，在使用单类 F-score 方法时偏重于对 2 类问题的特征处理，因而缺乏对多数类类间不平衡数据的进一步研究与分析。

下一步的工作可能会选择在高维度医学数据集上进行实验，考虑加强对多数类不平衡问题的处理分析，同时在提升时间效率方面，可能会采用并行计算、算法优化等方式减少算法运行时间，提升算法的计算效率。

6　结束语

本文通过单类特征选择和遗传算法建立多层神经网络模型，以 AUC 值为适应度指标，选出对于不平衡数据分类更有利的特征集合，从而建立更适用于不平衡数据分类的深度模型。使用单类 F-score 特征选择替代传统 F-score 对于特征的评价方法，以此选出偏向少数类分类的特征集合，在使用多层神经网络进行适应度函数计算时，对于每一个个体网络输入都会以 7:3 的概率重设训练集和测试集，同时遗传算法每次通过交叉变异生成新种群时都具有一定随机性，从而确保本文模型能探索的空间尽可能大，且能在更大的空间里选出更偏向于少数类分类的特征子集。最终，通过对 4 组不同 UCI 数据集进行测试，并与传统的机器学习算法进行对比验证。实验表明，本文模型评价指标 AUC 值更大，表明本文模型在处理不平衡数据分类问题上的表现更优秀。

参考文献：

[1] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: special issue on learning from imbalanced data sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1):1-6.

[2] EZAWA K J, SINGH M, NORTON S W. Learning goal oriented Bayesian networks for telecommunications risk management[C]// Thirteenth International Conference on International Conference on Machine Learning. 1996:139-147.

[3] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1):20-29.

[4] JAPKOWICZ N, STEPHEN S. The class imbalance problem: a systematic study[M]. Amsterdam: IOS Press, 2002.

[5] WEISS G M. Mining with rarity: a unifying framework[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 7-19.

[6] AKBANI R, KWEK S, JAPKOWICZ N. Applying support vector machines to imbalanced datasets[J]. Lecture Notes in Computer Science, 2001, 3201: 39-50.

[7] RASKUTTI, BHAVANI, KOWALCZYK. Extreme re-balancing for SVMs: a case study[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 60-69.

[8] WU G, CHANG E Y. Class-boundary alignment for imbalanced dataset learning[J]. ICML Workshop on Learning from Imbalanced Data Sets, 2003: 49-56.

[9] ZHANG J, MANI I. KNN approach to unbalanced data distributions: a case study involving information extraction[C]// The ICML 2003 Workshop on Learning from Imbalanced Datasets. 2003.

[10] PATCHA A, PARK J M. An overview of anomaly detection techniques: existing solutions and latest technological trends[J]. Computer Networks, 2007, 51(12): 3448-3470.

[11] FAWCETT T, PROVOST F. Adaptive fraud detection[J]. Data Mining & Knowledge Discovery, 1997, 1(3): 291-316.

[12] CARDIE C, NOWE N. Improving minority class prediction using case-specific feature weights[C]// Fourteenth International Conference on Machine Learning. 1997: 57-65.

[13] BLAKE C. UCI repository of machine learning databases[J]. Department of Information and Computer Science, 1998.

[14] MALOOF M A. Learning when data sets are imbalanced and when costs are unequal and unknown[J]. ICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2003.

[15] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//International Conference on Machine Learning. 2012: 179-186.

[16] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[17] JOSHI M V, KUMAR V, AGARWAL R C. Evaluating boosting algorithms to classify rare classes: comparison and improvements[C]// IEEE International Conference on Data Mining. 2001: 257-264.

[18] 王和勇, 樊泓坤, 姚正安, 等. 不平衡数据集的分类方法研究[J]. 计算机应用研究, 2008, 25(5): 1301-1303.
WANG H Y, FAN H K, YAO Z A, et al. Research on the classification method of unbalanced dataset[J]. Computer Application Research, 2008, 25(5): 1301-1303.

[19] LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J]. Expert Systems with Applications, 2009, 36(8): 10896-10904.

[20] MALDONADO S, WEBER R. A wrapper method for feature selection using support vector machines[J]. Information Sciences, 2008, 179(13): 2208-2217.

[21] LIU Y, ZHENG Y F. FS_SFS: a novel feature selection method for support vector machines[J]. IEEE International Conference on Acoustics, 2006, 39(7): 1333-1345.

[22] RAMARAJ N, RAMARAJ N. A hybrid prediction model with F-score feature selection for type II Diabetes databases[C]// Amrita ACM-W Celebration on Women in Computing in India. 2010: 13.

[23] LIN X, WEI H, WANG F, et al. A breast cancer risk classification model based on the features selected by novel f-score index for the imbalanced multi-feature dataset[C]//International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. 2017.

[24] HOLLAND J H. Adaption in natural and artificial systems[J]. Quarterly Review of Biology, 1975, 6(2): 126-137.

[25] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313: 504-507.

[26] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[27] 沈崇圣. 遗传算法中常用选择算子在 MATLAB 中的实现[J]. 上海应用技术学院学报(自然科学版), 2003, 3(3):199-202.
SHEN C S. The implementation of commonly used selection operators in MATLAB in genetic algorithm[J]. Journal of Shanghai Institute of Technology (Natural Science Edition), 2003, 3(3): 199-202.

[28] 林晓丽. 复杂高维医学数据挖掘与疾病风险分类研究[D]. 北京: 北京科技大学, 2016.
LIN X L. Research on complex high-dimensional medical data mining and disease risk classification[D]. Beijing: University of Science and Technology Beijing, 2016.

[作者简介]

张雪（1995-），女，北京科技大学硕士生，主要研究方向为医疗数据分析、算法设计与分析。

石志国（1978-），男，博士，北京科技大学教授，主要研究方向为智能系统与物联网技术。

刘璇（1993-），女，北京科技大学硕士生，主要研究方向为医疗数据分析、算法设计与分析。