DOI:10.13718/j.cnki.xsxb.2021.05.019

基于鲸鱼优化和深度学习的不平衡大数据分类算法①

孙二华[1], 胡云冰[2]

1. 重庆房地产职业学院 信息工程学院,重庆 401331;2. 厦门大学 信息科学与技术学院,福建 厦门 361005

摘要:针对当前不平衡数据分类算法中存在的分类精度低和容易陷入局部最优状态的问题,提出一种基于鲸鱼优化和深度学习的不平衡大数据分类算法.该算法由特征选择、预处理和分类 3 个阶段组成:①为了提高分类精度,使用鲸鱼优化算法(whale optimization algorithm, WOA)在不平衡数据中寻找最优特征子集,消除不相关和多余的特征;②采用局部敏感哈希的合成少数类过采样技术(locality sensitive hashing synthetic minority oversampling technique, LSH-SMOT)对数据集进行预处理,解决类不平衡问题;③使用基于 WOA 算法优化的双向递归神经网络(bidirectional recurrent neural networks, BRNN)对预处理后的数据集进行分类.实验结果表明:本文算法能够有效解决不平衡数据集的分类问题,相比于其他算法,本文算法在分类精度和局部最优避免率方面具有明显优势.

关 键 词:不平衡大数据分类;鲸鱼优化算法;深度学习;合成少数类过采样技术

中图分类号:TP393 **文献标志码**:A **文章编号**:1000-5471(2021)05-0127-07

近年来随着网络信息的迅猛发展,在医疗、金融、生物等多个行业积累了海量的数据.大数据在信息分析和行为预测方面起着重要的作用[1].由于数据分布不平衡,从如此大的数据集中分析和提取关键点是一项十分困难的任务.基于最大化分类准确率的传统分类算法无法处理此类数据集,可能会导致分类结果有偏差和决策错误[2].因此,研究有效处理不平衡数据中少数类识别率的技术成为科研人员关注的热点问题.

当一种类别的样本数远小于其他类的样本数时,该数据集被认为是不平衡的.由于不平衡大数据会导致分类算法的学习性能下降,并且存在小样本、重叠或小间断的问题,使得不平衡数据的分类面临巨大的挑战[3].当前大多数分类方法都是基于大数据平衡理论,而用于解决数据集类不平衡情况的方法还较少.Abdel 等[4]提出一种基于动态 Spark 的不平衡大数据分类框架,通过对少数类样本进行重采样和增强处理,提高分类性能.Jing 等[5]提出一种不相关的成本敏感多集学习方法用于解决高度不平衡的数据分类问题,该方法首先将不平衡数据随机分区构造出多个平衡子集,然后利用深度度量学习和多集特征学习技术解决每个子集中存在的非线性问题,实现更好的算法性能.Lu 等[6]提出一种改进的加权极限学习机,通过将投票方法引入到加权极限学习机中解决不平衡数据分类问题.Zhang 等[7]提出内核修改的最佳余量分配机分类方法,通过引入具有两个自由参数的共形函数来缩放 ODM 的核矩阵,提高特征空间中训练数据的可分离性以及消除数据不平衡的影响,使得改进后的分类器具有更平衡的检测率和更好的泛化性能.Li 等[8]提出一种基于粒子群优化的自适应提升算法,该算法可以重新初始化参数以避免陷入局部最优,防止冗余或无用的弱分类器消耗过多的系统资源,能够更好地处理不平衡度较高的数据.虽然上述算法能够在一定程度上处理不平衡数据的分类问题,但是在大多数情况下,模型的复杂性和数据分类的精度之间存在着不可避免的权衡,而且没有细致考虑数据分布问题及类之间的不平衡问题,因此容易产生分类精度低的效果.

① 收稿日期:2020-03-16
基金项目:重庆市教委高职教育双基地建设项目(20180310).
作者简介:孙二华,副教授,主要从事大数据及物联网研究.

针对上述存在的问题及分类优化算法容易陷入局部最优状态的现象，本文提出一种基于鲸鱼优化和深度学习的不平衡大数据分类算法，该算法分为特征选择、预处理和分类3个阶段. 在特征选择阶段，将使用鲸鱼优化算法消除不相关和多余的特征，以提高分类精度；在预处理阶段，通过预处理数据集来解决类不平衡问题，获取数据集类之间的平衡；在分类阶段，基于双向递归神经网络的深度学习方法对预处理后的数据集进行分类.

1 鲸鱼优化算法

鲸鱼优化算法(whale optimization algorithm，WOA)是根据座头鲸捕猎行为提出的一种新的群智能优化算法[9]. 在海洋活动中，座头鲸采用一种bubble-net捕食策略，通过向上螺旋和双循环等觅食行为狩猎食物. Bubble-net捕食策略大致可以分为包围猎物、狩猎行为和搜索猎物3步. 在WOA算法中，假设鲸鱼种群规模为N，搜索空间为d维，第i只鲸鱼在搜索空间内的位置表示为$K_i=(k_i^1, k_i^2, \cdots, k_i^d)$，猎物的位置表示全局最优解决方案.

WOA算法在模拟座头鲸包围猎物时，假定目标猎物为当前群体在搜索空间中的全局最优位置，其他个体可以通过最优解更新位置.

$$K(t+1) = K^*(t) - \boldsymbol{A} \cdot D \tag{1}$$

$$D = | \boldsymbol{C} \cdot K^*(t) - K(t) | \tag{2}$$

式(1)、式(2)中：t为当前迭代次数，$K(t)$和$K^*(t)$分别为当前鲸鱼位置和最优解位置，D是搜索个体与猎物的距离，$\boldsymbol{A}$和$\boldsymbol{C}$是系数向量，可以定义为

$$\boldsymbol{A} = 2a \cdot r_1 - a \tag{3}$$

$$\boldsymbol{C} = 2r_2 \tag{4}$$

式(3)、式(4)中：r_1和r_2是[0，1]之间的随机数，a为随迭代次数增加而递减的控制参数.

$$a = 2 - \frac{2t}{t_{\max}} \tag{5}$$

采用下列模型模拟座头鲸螺旋向上的狩猎行为

$$K(t+1) = D' \cdot e^{ql} \cdot \cos(2\pi l) + K^*(t) \tag{6}$$

式(6)中：$D' = | K^*(t) - K(t) |$，l是[-1，1]之间的随机数，q是对数螺旋线形函数的常数. 根据bubble-net捕食策略，座头鲸沿着螺旋形路径移动的同时还会缩小包围圈. 在优化过程中，选择在收缩圆圈机制中移动或选择螺旋模型来更新鲸鱼位置的概率被认为是0.5，即

$$K(t+1) = \begin{cases} K^*(t) - A \cdot D & P < 0.5 \\ D' \cdot e^{ql} \cdot \cos(2\pi l) + K^*(t) & P \geqslant 0.5 \end{cases} \tag{7}$$

式(7)中：P是[0，1]之间的随机数，控制螺旋或圆形运动之间的转换.

在搜索猎物时，其数学模型为

$$D = | C \cdot K_{rand}(t) - K(t) | \tag{8}$$

$$K(t+1) = K_{rand}(t) - A \cdot D \tag{9}$$

式(8)、式(9)中：$K_{rand}(t)$表示从种群中随机选择的个体位置. 参数A控制探索和开发之间的转换，为了保证算法的探索和收敛，当$|A|<1$时，选择当前迭代中最优解位置来更新其他个体；当$|A|\geqslant 1$时，随机选择一个个体位置来更新其他鲸鱼的位置.

2 基于鲸鱼优化和深度学习的分类算法

针对不平衡数据的分类问题，本文提出一种基于鲸鱼优化和深度学习的不平衡大数据分类算法，该算法由特征选择、预处理和分类3个阶段组成.

2.1 特征选择

数据集中冗余或不相关特征对分类精度有很大的影响. 特征选择阶段的主要目标是寻找最佳特征子集，以提高分类精度. 在这个阶段考虑到每个特征子集是鲸鱼的一个位置，使用WOA算法进行特征选择.

每个子集将包含小于或等于 na 的随机特征数，其中 na 是原始数据集中的特征总数. 具有最少特征数和较高分类精度的鲸鱼将被视为最优解.

2.2　预处理

在该阶段将使用基于局部敏感哈希的合成少数类过采样技术(locality sensitive hashing synthetic minority oversampling technique，LSH-SMOTE) 进行采样. 为了控制过拟合问题提出了合成少数类过采样技术(synthetic minority oversampling technique，SMOTE)[10]. SMOTE 算法的基本思想是通过创建少数类的合成实例来处理属性域而不是实例域，从而生成少数类的合成样本. 通过沿 k 个少数类最近邻居来创建合成实例，少数类中的每个实例都被过采样. LSH-SMOTE 技术是将局部敏感哈希技术用于快速高效地找到少数群体的最近邻居，而不是 SMOTE 算法中使用的标准技术 K -最近邻居[11]. 这种过采样技术将数据集进行哈希处理并划分为多个存储桶，将具有相似哈希码的相似项分配在同一桶内，这会增加相似项之间的碰撞概率，并简化了每个桶中 k 邻近的搜索. 从每个桶中选择碰撞次数最多的实例，然后使用欧几里德距离对实例进行排序，仅选择 K 个最近邻的碰撞实例作为查询实例. 最后，将包含 k 邻近实例的列表返回到主SMOTE类以生成合成实例. 由于LSH复杂度低，因此将LSH技术与SMOTE算法相结合会大大节省运行时间.

2.3　分类

在分类阶段，采用双向递归神经网络(bidirectional recurrent neural networks，BRNN) 的深度学习方法，用于对预处理后的数据集进行分类. BRNN 分类器具有权重和偏差两个重要参数，它们对 BRNN 分类器的性能产生巨大的影响. 为了提高 BRNN 的性能，本文采用鲸鱼优化算法来寻找两种类型参数的最佳值，从而解决 BRNN 不稳定梯度问题，提高分类的精度和准确性.

2.3.1　双向递归神经网络

双向递归神经网络[12]，其主要优点是增加了网络可用的输入信息量，弥补了标准递归神经网络、多层感知器和时延神经网络存在固定输入数据和没有未来信息的局限性. BRNN 的基本思想是将隐藏层从相反方向连接到输出. 由于这种独特的结构，过去和将来的信息可随时输出.

递归神经网络可视为普通前馈多层感知器网络的扩展，其中输入值和输出值为向量而不是离散值. 假定 $\boldsymbol{X}=\{\boldsymbol{x}_t\}$ 和 $\boldsymbol{Y}=\{\boldsymbol{y}_t\}$ 分别表示递归神经网络的输入和输出，其中 $\boldsymbol{x}_t\in R^n$ 和 $\boldsymbol{y}_t\in R^n$ 是每个时间步长 t 下的向量. 通过建立 $P(\boldsymbol{Y}\mid\boldsymbol{X})$ 分布模型来求取最终解，在此过程中，不仅利用递归神经网络将输入向量映射到输出向量，同时还需要使用 RNN 来预测下一个输入.

单向递归神经网络的输出 $\boldsymbol{y}_t$ 可以表示为

$$P(\boldsymbol{y}_t\mid(x_i)_{i=1}^{t})=\sigma(\boldsymbol{W}_y\boldsymbol{h}_t+\boldsymbol{b}_y) \tag{10}$$

$$\boldsymbol{h}_t=\tanh(\boldsymbol{W}_h\boldsymbol{h}_{t-1}+\boldsymbol{W}_x\boldsymbol{x}_t+\boldsymbol{b}_h) \tag{11}$$

式(10)、式(11) 中：$\boldsymbol{W}_y$ 是将隐含层连接到输出层的权重矩阵，$\boldsymbol{W}_h$ 是隐藏到隐含层的权重矩阵，$\boldsymbol{W}_x$ 是将输入层与隐含层连接的权重矩阵. $\boldsymbol{b}_y$ 为输出层偏差向量，$\boldsymbol{b}_h$ 为隐含层偏差向量. 对于最终的非线性 σ，可以使用 sigmoid，tanh 和 relu 作为分类的激活函数. RNN 将根据隐含层传播的信息来评估输出 $\boldsymbol{y}_t$，而不管它是否直接或间接地依赖于值 $(\boldsymbol{x}_i)_{i=1}^{t}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t\}$.

双向 RNN(BRNN) 是单向 RNN 的扩展，它引入了第 2 个隐含层，其中隐含层与隐含层的连接按相反的时间顺序排列. 因此，该模型能够利用过去和将来两个方向的数据，其输出 $\boldsymbol{y}_t$ 可以表示为

$$P(\boldsymbol{y}_t\mid(x_i)_{i\neq t})=\sigma(\boldsymbol{W}_y^f\boldsymbol{h}_t^f+\boldsymbol{W}_y^{bac}\boldsymbol{h}_t^{bac}+\boldsymbol{b}_y) \tag{12}$$

$$\boldsymbol{h}_t^f=\tanh(\boldsymbol{W}_h^f\boldsymbol{h}_{t-1}^f+\boldsymbol{W}_x^f\boldsymbol{x}_t+\boldsymbol{b}_h^f) \tag{13}$$

$$\boldsymbol{h}_t^{bac}=\tanh(\boldsymbol{W}_h^{bac}\boldsymbol{h}_{t+1}^{bac}+\boldsymbol{W}_x^{bac}\boldsymbol{x}_t+\boldsymbol{b}_h^{bac}) \tag{14}$$

式(12)、式(13)、式(14) 中：上标 f 所表示的前进方向具有独立的非束缚重量，上标 bac 表示后向隐藏的激活. 在式(12)、式(13)、式(14) 中，由于非循环连接，$\boldsymbol{y}_t$ 不会从 $\boldsymbol{x}_t$ 中获取数据. 因此，可以通过设置 $\boldsymbol{Y}=\boldsymbol{X}$ 以无监督的方式使用该模型，用于预测输入序列中给定时间的下一步长输出值. 在 BRNN 的向后传递中，有两个反向传播时间阶段，该阶段采用最小化均方误差来更改权重值.

BRNN 中的权重是决定输入对输出影响程度和控制隐藏层学习率的最重要因素. 在线性回归中，输出

是通过将权重乘以输入然后相加而形成的. 权重是控制神经元相互影响程度的数值，对于任何神经元，输出可以表示为权重与输入乘积的和的形式.

$$\boldsymbol{y} = f(\boldsymbol{x}) = \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{w}_j + b \tag{15}$$

式(15)中：b是偏差，用于调整输出与神经元输入间的平衡.

2.3.2 基于鲸鱼优化的BRNN

采用WOA算法对BRNN中最重要因素的权重和偏差进行优化，通过找到权重和偏差的最佳值来实现BRNN的最佳训练，进而达到更好的分类精度. WOA算法的优化变量可以表示为

$$\boldsymbol{V} = \{W, b\} = \{W_{11}, W_{12}, \cdots, W_{ij}, b_1, b_2, \cdots, b_j\} \tag{16}$$

式(16)中：输入节点的数目为n，从第i个节点到第j个节点的连接权重为W_{ij}，偏移量为b_j，其中$i,j \in \{1, 2, \cdots, n\}$.

由于BRNN训练的目标是能够达到最高分类精度，而均方误差MSE是度量BRNN分类效果的常用指标，因此采用MSE作为WOA算法优化的目标函数. 该度量的工作原理：通过对BRNN应用一组训练实例，计算期望输出和BRNN输出之间的差异.

$$MSE = \sum_{i=1}^{m} (o_i^k d_i^k)^2 \tag{17}$$

式(17)中：m是输出数，o_i^k和d_i^k分别表示使用第k个训练实例时第i个输入神经元的实际输出和最佳输出. 为了使BRNN更有效，采用所有训练实例的MSE平均值来评估BRNN的性能.

$$MSE_{avg} = \sum_{k=1}^{s} \frac{\sum_{i=1}^{m} (o_i^k d_i^k)^2}{s} \tag{18}$$

式(18)中：s是训练实例数. 最后，可以用WOA算法的变量和平均MSE来表示BRNN的训练问题，即

$$Minimize: f(V) = MSE_{avg} \tag{19}$$

在使用WOA训练BRNN的整个过程中，WOA算法采用最小化所有训练实例平均均方误差的方式为BRNN优化了权值和偏差. 因此，当迭代次数足够多时，WOA可以收敛到最优解.

3 实验与结果分析

为了评价本文算法的性能，所有实验在一台配置为i7-6700HQ CPU @2.60 GHz，12 GB RAM的电脑上运行. 在3.9.1版Weka学习环境和Matlab2018a软件上进行测试实验，并在相同条件下与基于粒子群优化的自适应提升算法(PSO+AdaBoost)[8]、基于遗传算法改进的随机森林分类器(GA-RF)[13]和基于蚁群-遗传优化的支持向量机(ACOGA+SVM)[14]等现有算法相比较.

3.1 数据集和评估指标

本文选择KEEL数据集[15]中类别不平衡比(imbalance ratio，IR)大于9的8个数据子集和ECBDL′14数据集[16]作为测试数据(表1).

表1 实验数据集特征参数

数据集名称	样本	属性	IR
yeast-2_vs_4	514	8	9.08
yeast-1_vs_7	459	7	14.3
yeast-1-4-5-8_vs_7	693	8	22.1
yeast-2_vs_8	482	8	23.1
yeast4	1 484	8	28.1
yeast-1-2-8-9_vs_7	947	8	30.57
yeast5	1 484	8	32.73
yeast6	1 484	8	41.4
ECBDL 14 datase	2 897 917	23	58.58

针对不平衡分类性能的评估问题，研究人员提出了许多算法，本文采用受试者工作特征曲线(ROC)下

的面积 AUC 和均方误差 MSE 来评估不平衡数据集上的分类性能，其中 AUC 的定义为

$$AUC = \frac{Sensitivity + Specificity}{2} \tag{20}$$

式(20)中：$Sensitivity$ 和 $Specificity$ 分别表示召回率和特异性，其定义为

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

式(21)、式(22)中，TP，FP，TN 和 FN 是二分类问题中的术语，分别表示真阳性、假阳性、真阴性和假阴性.

3.2　实验结果

比较本文算法与其他 3 种算法在 9 个数据集的测试结果，并利用 AUC 和 MSE 两个指标进行评估. 表 2和表 3 分别给出了不同算法在九个数据集的测试对比结果. 表 3 的结果表明，本文算法在所有数据集上都取得了最好的 AUC 分数，这有力地证明了本文算法具有更高的效率.

表 2　不同算法在 *AUC* 指标的对比结果

数据集	PSO+AdaBoost	GA-RF	ACOGA+SVM	本文方法
yeast-2_vs_4	0.988	0.967	0.989	0.992
yeast-1_vs_7	0.915	0.893	0.947	0.963
yeast-1-4-5-8_vs_7	0.819	0.906	0.962	0.972
yeast-2_vs_8	0.910	0.943	0.986	0.993
yeast4	0.949	0.954	0.983	0.994
yeast-1-2-8-9_vs_7	0.901	0.932	0.984	0.989
yeast5	0.991	0.992	0.996	0.999
yeast6	0.970	0.961	0.986	0.995
ECBDL 14 datase	0.733	0.964	0.991	0.994
平均	0.908	0.945	0.980	0.988

从表 3 结果可知，本文算法在所有数据集上的 MSE 得分最低，从而说明本文算法在大的不平衡数据集上具有很高的局部最优回避能力. 总的来说，本文算法能够有效处理极不平衡的大数据集，获得高精度的分类结果.

表 3　不同算法在 *MSE* 指标的对比结果　　10^{-3}

数据集	PSO+AdaBoost	GA-RF	ACOGA+SVM	本文方法
yeast-2_vs_4	0.14	1.09	0.12	0.081
yeast-1_vs_7	7.23	13.69	2.82	1.37
yeast-1-4-5-8_vs_7	32.76	8.83	1.45	0.78
yeast-2_vs_8	8.10	3.25	0.20	0.064
yeast4	2.60	2.12	0.29	0.049
yeast-1-2-8-9_vs_7	9.81	4.62	0.26	0.12
yeast5	0.081	0.064	0.016	0.001
yeast6	0.9	1.52	0.21	0.025
ECBDL 14 datase	71.29	1.31	0.081	0.049
平均	8.46	3.03	0.4	0.17

4　结　　语

本文提出了一种基于鲸鱼优化和深度学习的不平衡大数据分类算法，用于解决当前大多数不平衡数据分类算法中存在的分类精度低和容易陷入局部最优状态的问题. 本文算法包括 3 个阶段：在第一阶段利用

鲸鱼优化算法对不平衡数据进行特征优化，消除冗余特征，寻找最优特征子集；第二阶段采用LSH-SMOT技术对数据集进行预处理，获得大多数类和少数类之间的平衡；第三阶段通过WOA算法优化双向递归神经网络，使用优化后的BRNN对预处理后的数据集进行分类. 实验结果表明，本文算法能够处理极不平衡的大数据集，获得高精度的分类结果.

参考文献：

[1] ARIYALURAN HABEEB R A, NASARUDDIN F, GANI A, et al. Real-time Big Data Processing for Anomaly Detection: a Survey [J]. International Journal of Information Management, 2019, 45: 289-307.

[2] HASANIN T, KHOSHGOFTAAR T M, LEEVY J L, et al. Examining Characteristics of Predictive Models with Imbalanced Big Data [J]. Journal of Big Data, 2019, 6(1): 1-21.

[3] HASANIN T, KHOSHGOFTAAR T M, LEEVY J L, et al. Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches [J]. Journal of Big Data, 2019, 6(1): 1-21.

[4] ABDEL-HAMID N B, ELGHAMRAWY S, DESOUKY A E, et al. A Dynamic Spark-based Classification Framework for Imbalanced Big Data [J]. Journal of Grid Computing, 2018, 16(4): 607-626.

[5] JING X Y, ZHANG X Y, ZHU X K, et al. Multiset Feature Learning for Highly Imbalanced Data Classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 29: 1-2.

[6] LU C B, KE H F, ZHANG G Y, et al. An Improved Weighted Extreme Learning Machine for Imbalanced Data Classification [J]. Memetic Computing, 2019, 11(1): 27-34.

[7] ZHANG X G, WANG D X, ZHOU Y C, et al. Kernel Modified Optimal Margin Distribution Machine for Imbalanced Data Classification [J]. Pattern Recognition Letters, 2019, 125: 325-332.

[8] LI K W, ZHOU G Y, ZHAI J N, et al. Improved PSO_AdaBoost Ensemble Algorithm for Imbalanced Data [J]. Sensors, 2019, 19(6): 1476.

[9] 闫 旭，叶春明，姚远远. 量子鲸鱼优化算法求解作业车间调度问题 [J]. 计算机应用研究，2019，36(4)：975-979.

[10] DIALLO M, XIONG S W, COULIBALY M N, et al. Synthetic Minority Oversampling Technique in Stages for Unbalanced Climate and Rice Dataset: The Office Du Niger Case Study [C]//Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering. Tokyo: ACM, 2019.

[11] HASSIB E M, ELDESOKEY A E, LABIB L M, et al. LSH-SMOTE: A Modified SMOTE Algorithm for Imbalanced Data-Sets [J]. Ciência Técnica Vitivinícola, 2018, 33(4): 50-65.

[12] SHASHIKUMAR S P, SHAH A J, CLIFFORD G D, et al. Detection of Paroxysmal Atrial Fibrillation Using Attention-based Bidirectional Recurrent Neural Networks [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Beijing: ACM, 2018.

[13] PAING M P, CHOOMCHUAY S. Improved Random Forest (RF) Classifier for Imbalanced Classification of Lung Nodules [C]//2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST). Phuket: IEEE, 2018.

[14] NURLAILY D, IRHAMAH, PURNAMI S W, et al. Support Vector Machine for Imbalanced Microarray Dataset Classification Using Ant Colony Optimization and Genetic Algorithm [C]//The 2nd International Conference on Science, Mathematics, Environment, and Education. Chongqing: AIP, 2019.

[15] TSAI C F, LIN W C, HU Y H, et al. Under-sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection [J]. Information Sciences, 2019, 477: 47-54.

[16] HASANIN T, KHOSHGOFTAAR T M, LEEVY J, et al. Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data [C]//2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). San Francisco: IEEE, 2019.

Unbalanced Big Data Classification Algorithm Based on Whale Optimization and Deep Learning

SUN Er-hua[1]，　HU Yun-bing[2]

1. School of Information Engineering，Chongqing Real Estate College，Chongqing 401331，China；

2. College of Information Science and Technology，Xiamen University，Xiamen Fujian 361005，China

Abstract：Aiming at the problems of low classification accuracy，which is easy to fall into local optimal state in the current unbalanced data classification algorithm，an unbalanced big data classification algorithm based on whale optimization and deep learning has been proposed. The algorithm consists of three parts：feature selection，preprocessing and classification. Firstly，in order to improve the classification accuracy，the whale optimization algorithm（WOA）has been used to find the optimal feature subset in the unbalanced data to eliminate the irrelevant and redundant features. Secondly，the locality sensitive hashing synthetic minority oversampling technique（LSH-SMOTE）has been used to preprocess the dataset to solve the class imbalance problem. And，finally，the bidirectional recurrent neural networks（BRNN）optimized based on WOA algorithm been used to classify the preprocessed dataset. The experimental results show that the proposed algorithm can effectively solve the classification problem of unbalanced data sets. Compared with other algorithms，it has obvious advantages in classification accuracy and local optimal avoidance rate.

Key words：unbalanced big data classification；whale optimization algorithm；deep learning；synthetic minority oversampling technique

责任编辑　夏　娟