[文章编号] 1003－4684(2021)05-0051-04

基于分布式混合灰狼蝗虫优化算法航班延误预测

涂胜红，陈宏伟，杨威威，杨智慧

（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 针对航班延误成因复杂、数据量大，传统模型预测准确率差、效率低的问题，提出一种基于灰狼优化算法和蝗虫优化算法的混合算法建立航班延误预测模型。将灰狼优化算法的等级机制引入到蝗虫优化算法中，在种群迭代时生成不同层级狼群共同指导种群的进化方法，避免单个体对种群进化的绝对控制。采用 Spark 大数据框架设计分布式混合灰狼蝗虫优化算法提高模型运行效率。仿真实验结果表明：分布式混合灰狼蝗虫算法能够提高航班延误预测准确率和运行效率。

[关键词] 蝗虫优化算法；灰狼优化算法；分布式混合算法模型；航班延误预测

[中图分类号] TP391 [文献标识码] A

世界航空货运预测：航空市场将保持 4.7％的增速[1]。随着航班数量增加，航班延误和取消的现象变得越来越普遍。航班延误和取消不仅影响个人出行，也严重损害航空公司的声誉和利益[2]，对我国民航业的发展造成阻碍。近年来国内外研究者对航班延误预测进行了一些研究，Khanmohammadi 采用多层级 ANN 挖掘航班数据集输入变量与输出变量关系[3]；Farshchian 等人提出了一种基于深度学习的航班延误预测模型，结合堆栈的自动去噪编码技术提高模型预测准确率[4]；Yu 等采用深度信念网络与支持向量机融合方法在大型数据集捕获特征方面具有很好的效果[5]；李华峰采用贝叶斯网络拓扑结构相结合的方法建立预测模型，另外加入影响航班延误因素，进一步建立航班延误波及传递的贝叶斯网络模型[6]；张兆宁等人基于 SEIR 思想将航班分为正常、延误、延误传播及恢复四种状态，通过计算基本再生的值来预测下一时段的航班延误发生[7]；王语桐等人采用支持向量机和多元线性回归方法建立组合预测模型，对航班延误进行预测[8]。

本文使用基于机器学习的分类方法，通过对航班延误数据集的各个特征分析，确定与航班延误相关的特征因素，运用随机森林分类算法，得到航班延误预测等级。为确定最优的随机森林分类模型，使用分布式灰狼蝗虫优化算法对随机森林的参数进行调优。数据集实验结果显示，优化后的航班延误预测的准确率更高，运行时间更少。

1 航班延误预测

由于航空系统的复杂性，航班延误的成因也具有复杂性和随机性，其影响因素包括人为因素如旅客影响、空管影响、机场管理影响等，还有不可控的因素如天气、军事原因导致的航班延误或取消。研究使用的航班延误数据来自 Kaggle 网站上美国交通运输部的（BTS）2009－2018 年全美航空公司航班运行信息数据，该数据集包含超过五千万条航班运行信息，采集航班运行过程中多种特征数据，同时航空系统的运行具有相似性，其数据特征的研究具有相通性，因此采用 BTS 的信息数据研究对我国航空发展仍然具有意义。

2009－2018 年航班延误数据集包含 28 个特征，分别是航班日期、航空公司识别码、航班号等等。将上述数据进行特征分析保留对航班延误具有相关性的特征，一方面能够提高预测准确率，另一方面减少无关特性对预测结果带来干扰，缺失值采用随机取值然后计算均值的方法插入。字符属性数值化处理包括航空公司识别码、机场代码都是字符类型数据，按照在数据集中出现顺序，依次使用不同的数值代替。在构建航班延误预测模型时，按照航班到达时间将航班延误进行分级预测，在延误分级之后将标签数值化，分为五个类别。在下面的实验部分将对数据集标签进行训练和预测。分级范围和标签设置见表 1。

[收稿日期] 2021－03－23

[基金项目] 国家自然科学基金(61772180)；湖北省大学生创新创业项目(S201910500048)

[第一作者] 涂胜红(1996－)，男，湖北广水人，湖北工业大学硕士研究生，研究方向为大数据，云计算

[通信作者] 陈宏伟(1975－)，男，湖北武汉人，工学博士，湖北工业大学教授，研究方为大数据，云计算

表 1 延误分级

等级标签	延误等级	延误时间范围/min
0	正常航班	0
1	轻度延误	0～15
2	中度延误	15～30
3	较长延误	30～60
4	重度延误	≥60

在数据集的设计中考虑到后续研究实验将在 Spark 大数据平台对模型进行训练，为了后续对比算法运行效率，将航班延误数据集按照数据条数分别划分为不同的数据集，分别包含的数据条数为 500 万条、1000 万条、2000 万条和 5000 万条数据。

2 混合灰狼蝗虫优化算法

灰狼优化算法(Grey Wolf Optimizer，GWO)是一种进化元启发式算法，其仿真行为是灰狼的领导等级和狩猎食物的行为，蝗虫优化算法(Grasshopper Optimization Algorithm，GOA)是 2017 年提出的模仿蝗虫运动行为的群智能优化算法，与大多数算法相比，GWO 算法的主要优势是：元启发式算法不需要特定的输入参数，同时具有较强的局部寻优能力。GOA 的优势在于处理复杂、高维数据时全局寻优能力强。考虑到 GWO 和 GOA 各自的优势，这两种算法非常适合杂交混合，混合算法由 GWO 和 GOA 组成，称为混合灰狼蝗虫优化算法(Hybrid Gray Wolf Optimizer Grasshopper Optimization Algorithm，GWOGOA)，将 GWO 等级机制引入到 GOA 中，每一次迭代过程中，种群按照适应度大小依次选出 α 狼、β 狼和 δ 狼，根据它们的位置共同指导种群的进化方法，而不是仅仅根据单一的最优个体进行更新，避免了单一个体对种群进化的绝对控制，有效避免算法陷入局部最优，GWOGOA 算法的蝗虫种群位置更新依据公式(1)进行。

$$x_i^d(t+1)=c\left(\sum_N c\,\frac{ub^d-lb^d}{2}s(|x_j(t)-x_i(t)|)\,\frac{x_j(t)-x_i(t)}{d_{ij}(t)}\right)+S^d(t) \tag{1}$$

$$S(t+1)=\frac{X_\alpha+X_\beta+X_\delta}{3}$$

$$c=c_{\max}-t\,\frac{c_{\max}-c_{\min}}{\text{Max\_iter}}$$

其中，t 为蝗虫迭代的次数，N 为蝗虫的种群规模，参数 c 是缩小系数，用以线性减小舒适空间、排斥空间和吸引空间，$x_i^d(t)$ 为第 t 次迭代的第 i 只蝗虫的第 d 维，$d_{ij}(t)$ 为蝗虫群在第 t 次迭代时第 i 个蝗虫和第 j 个蝗虫之间的距离，ub^d、lb^d 分别表示蝗虫位置在 d 维的上边界和下边界，S 函数定义为一个函数，表示蝗虫间的社会作用力：

$$S(r)=fe^{\frac{-r}{l}}-e^{-r}$$

其中：f 表示吸引力的强度，l 是吸引力的大小范围，在文献中取值 $f=0.5$，$l=1.5$。

启发式算法还使用交叉和变异操作增强智能优化算法的勘探和开发能力，它有助于 GWOGOA 避免过早的收敛影响算法性能。交叉策略将种群个体根据适应度大小分为两部分，对于优秀种群部分予以保留，对于非优种群使用公式(2)交叉策略。变异策略是为了增强种群的随机性，以保留算法跳出局部收敛点的可能性，保持种群多样性以避免算法陷入局部最优，变异率是变异个体占种群总数量的比例，过大的变异率使种群波动过大难以收敛，变异率太小起不到保持种群多样性，跳出局部最优的目的，对所有个体按照公式(3)进行变异。

$$x_i(t)=x_m(1)+\cdots,x_m(k)+x_n(k+1),\cdots+x_n(d),\quad[k]=\text{rand}*\dim \tag{2}$$

$$x_i^{[k]=\text{rand}*\dim}(t)=\begin{cases}\text{rand}()*(\text{up}-\text{down})+\text{down},\text{if}(\text{rand}()<\text{p})\\ x_i^d(t),\text{else}\end{cases} \tag{3}$$

公式(2)中 k 为随机选择的交叉位置，m、n 表示选中的两个蝗虫；公式(3)中 rand()为[0，1]之间的随机数，k 表示变异的位置，变异概率 p 为 0.2，交叉变异之后的种群继续计算适应度迭代更新 α 狼、β 狼和 δ 狼的位置。具体算法实现步骤如下：

算法 1 GWOGOA 算法伪代码

Input：蝗虫搜索个数 N，搜索空间范围 up、down，个体维度 dim，迭代次数 Max_iter

Output：α 狼个体

```
1：初始化蝗虫群 position←InitPosition(N,up,down,dim)
2：for all grasshopper in position
3：    do getFitness(grasshopper, trainRDD)
4：end for
5：α、β、δ wolf ← getWolfs(grasshoper)
6：while iter < Max_iter
7：    for all grasshopper in position
8：    c←公式 3
9：    x(t+1) ←公式 1、2
10：   newPosition ←position.cross()，变异交叉策略
11：end for
12：   α、β、δ wolf←getWolfs(grasshoper) //更新 α、β、δ 狼的位置
13：   iter += 1
14：end while
```

3 分布式混合灰狼蝗虫优化算法

Apache Spark 是一种基于内存计算的大数据计算框架，能将数据加载至内存后重复使用，减少数

据写磁盘,有效提高大数据迭代计算运行效率。传统优化算法只能单机串行计算,面对计算复杂度高和数据量大的场景,算法运算速度受限于单机配置,运行效率低。因此基于 Spark 大数据平台,使用 Spark 提供的算子将混合灰狼蝗虫优化算法做分布式改进。算法初始化在 driver 进行,进一步抽象为 RDD 分布到不同的计算节点并行计算,算法具体步骤描述如下:

Step1 算法初始化,在搜索空间范围内随机生成蝗虫群;

Step2 根据目标函数计算蝗虫个体适应度,生成 α、β 和 δ 狼并使用 Spark 广播变量;

Step 3 使用 parallelize()函数将种群转为 positionRDD;

Step4 将 positionRDD 使用 mapToPair()算子将种群映射为(个体,种群)的格式生成 pairRDD;

Step5 使用公式 3 更新参数 c、使用公式 1 对 pairRDD 使用 map 算子转换计算,使用公式 5、6 使用交叉和变异策略得到 newPairRDD;

Step6 使用 collect()算子将种群更新后的个体回收得到更新后的蝗虫种群;

Step7 根据目标函数计算蝗虫适应度,更新 α、β 和 δ 狼并使用 Spark 广播变量;

Step8 迭代次数加 1,如果达到停止条件则输出当前 α 狼的位置,否则返回 Step4。

随机森林(Random Forest,RF)是一种基于 Bagging 的集成机器学习方法,具有准确率高、抗噪能力强的优点,因此使用随机森林对航班延误做分类预测。随机森林的性能受不同参数影响较大,为了找到性能最优的随机森林模型,将分布式混合灰狼蝗虫算法用于随机森林参数的调优。选择随机森林主要的参数 n_estimators、max_features、max_depth 作为蝗虫个体的编码,进行迭代计算寻优。蝗虫个体解码出参数建立随机森林,输入航班延误数据集做分类预测,最后分类预测的准确率作为该个体的适应度。分布式 GWOGOA 预测模型的伪代码如下:

算法 2　基于 Spark 的分布式 GWOGOA 预测模型伪代码

Input:蝗虫搜索个数 N,搜索空间范围 up、down,个体维度 dim,迭代次数 Max_iter

Output:α 狼个体

```
1: spark←SparkSession.builder().appName("SparkGOA").getOrCreate() // Spark 集群入口
2: trainRDD←spark.read().format("libsvm").load(datapath)
3: positions←InitPosition(N,up,down,dim)
4: for all position in positions
5:    α、β、δ←getWolfs(grasshopper, trainRDD)
6:    positionRDD←spark.parallelize(position)
7:    pairRDD←positionRDD.mapToPair( _ , position) // 映射为(个体,种群)
8: end for
9: while iter < Max_iter
10:  c←公式 3
11:   for all grasshopper in position
12:   newPositionRDD←pairRDD.map()//公式 1、2 计算位置
13:   newPosition←newPositionRDD.cross().collect() //个体回收到 driver
14:   α、β、δ← getWolfs(newTargetPosition) //更新 α、β、δ 狼的位置
15:   end for
16:   iter +=1
17: end while
```

4　实验分析

为评估分布式 GWOGOA 算法的性能,使用四种测试函数来判断算法的收敛性,为体现分布式 GWOGOA 算法的寻优能力,采用海鸥优化算法(Seagull Optimization Algorithm,SOA)和混合灰狼蝗虫优化算法作为对比,海鸥优化算法是 2018 年提出的新型智能优化算法,具有收敛快、精度高、算法新的特点,因此使用 SOA 作为智能优化算法的代表进行对比实验。测试函数分别为 Sphere、Schwefel2.22、Schwefel1.2、Schwefel2.21 实验结果如图 1 所示,可以看出分布式 GWOGOA 算法具有较好的寻优能力。

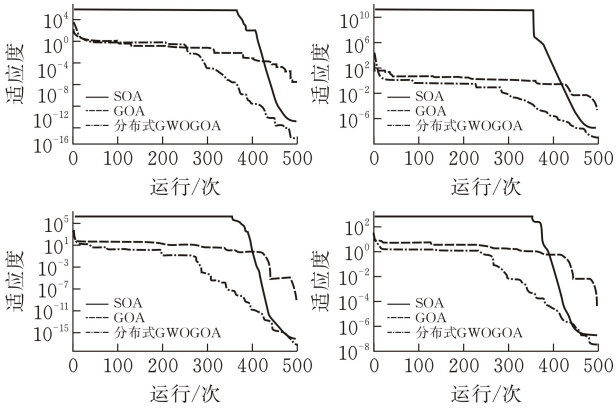

图 1　不同算法迭代曲线

集群实验环境由四台虚拟机组成,虚拟机搭建在 Win10 系统上,处理器为 9400,内存为 24G,使用 Hadoop3.2.1,Spark3.0 搭建 4 节点的分布式集群。航班延误预测模型的参数设置种群个体 N=10、迭代次数为 20 次、分布式 GWOGOA 与 GOA 的准确率迭代结果如图 2 所示,可以看到基于 Spark 的分布式 GWOGOA 算法有效提高航班延误预测准确率。在 Spark 集群上使用 Yarn 作为集群资源管理器,将 GOA 算法模型作为对比,算法的运行效率对

比见图 3,分布式 GWOGOA 算法解决了模型运行效率低的问题,显著减少了模型训练运行时间。

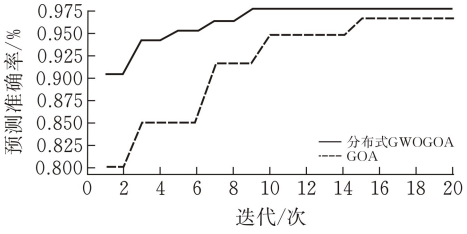

图 2 分布式 GWOGOA 对比 GOA 预测准确率

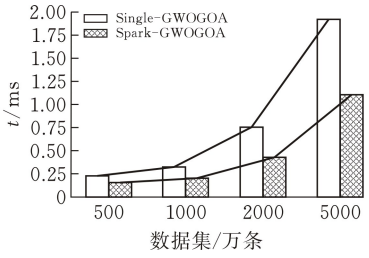

图 3 对比单机和 Spark 平台运行时间

5 结论

将混合算法思想与分布式思想结合提出分布式混合灰狼蝗虫优化算法,具有更强的寻优性能,将分布式混合灰狼蝗虫优化算法用于随机森林参数的调优,选择更合适的参数,构建性能更优的随机森林分类模型,提高了航班延误预测准确率,且模型运行时间缩短了 42%。

[参 考 文 献]

[1] NIU B,DAI Z,ZHUO X.Co-opetition effect of promised-delivery-time sensitive demand on air cargo carriers' big data investment and demand signal sharing decisions[J]. Transportation Research Part E: Logistics and Transportation Review,2019,23(MAR):29-44.

[2] 杜亚倩,张聊东.我国航班延误的现状及应对措施[J].中国科技信息,2020(21):35-37.

[3] KHANMOHAMMADI S, TUTUN S, KUCUK Y.A new multilevel input layer artificial neural network for predicting flight delays at JFK airport[J]. Procedia Computer Science, 2016, 95:237-244.

[4] YAZDI M F, KAMEL S R, CHABOK S, et al. Flight delay prediction based on deep learning and Levenberg-marquart algorithm[J]. Journal of Big Data, 2020,7(1):51-63.

[5] YU B,GUO Z,ASIAN S,et al.Flight delay prediction for commercial air transport:A deep learning approach[J]. Transportation Research Part E: Logistics and Transportation Review,2019,125(MAY):203-221.

[6] 李华峰.枢纽机场大面积航班延误波及传递模型的研究[D].天津:中国民航大学, 2012.

[7] 张兆宁,张佳.大面积航班延误发生的预测方法[J].系统工程,2020,38(4):115-121.

[8] 王语桐,朱金福,马思思.基于支持向量回归和线性回归的航班延误组合预测[J].武汉理工大学学报(交通科学与工程版),2019,43(3):426-431.

Flight Delay Prediction Based on Distributed Hybrid Gray Wolf Grasshopper Optimization Algorithm

TU Shenghong,CHEN Hongwei, YANG Weiwei, YANG Zhihui

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Aiming at the problems of complicated flight delays, large amount of data and poor prediction accuracy and low efficiency of traditional models, a hybrid algorithm based on gray wolf optimization algorithm and grasshopper optimization algorithm is proposed to establish a flight delay prediction model. The level mechanism of gray wolf optimization algorithm is introduced into the grasshopper optimization algorithm to generate different levels of wolves to jointly guide the evolution of the population when the population is iterated, so as to avoid the absolute control of the evolution of the population by a single body. The Spark big data framework is used to design a distributed hybrid gray wolf grasshopper optimization algorithm to improve model operation efficiency. The simulation experiment results show that the distributed hybrid gray wolf grasshopper algorithm can improve the accuracy of flight delay prediction and operational efficiency.

Keywords: grasshopper optimization algorithm; grey wolf optimization; distributed hybrid algorithm model; flight delay prediction

[责任编校: 张岩芳]