Pak.j.stat.oper.res. Vol.19 No.4 2023 pp 623-635 DOI: http://dx.doi.org/10.18187/pjsor.v19i4.4211

Pakistan Journal of Statistics and Operation Research

Black hole algorithm as a heuristic approach for rare event classification problem

Elif Yıldırım[1,2*]

*Corresponding author

1. Department of Statistics and Quality Coordinator, Konya Technical University, Konya, Turkey
2. Department of Statistics, Hacettepe University, Ankara, Turkey, edil@ktun.edu.tr

Abstract

The logistic regression is generally preferred when there is no big difference in the occurrence frequencies of two possible results for the considered event. However, for the events occurring rarely such as wars, economic crisis and natural disasters, namely having relatively small occurrence frequency when compared to the general events, the logistic regression gives biased parameter estimations. Therefore, the logistic regression underestimates the occurrence probability of the rare events. In this study, a modification of the black hole algorithm (BHA) is proposed as an alternative to the classical logistic regression method in order to obtain more reliable and unbiased rare event parameter estimates. To examine the performance of the proposed approach, we calculate bias and root mean square errors based on Monte Carlo (MC) simulations. We used logistic regression to generate data for the rare event in the simulations and gave values to the β_0 parameter to obtain different rarity levels. The performance of the methods was examined in different scenarios using comprehensive MC simulations under different conditions for the rarity level and number of subjects. In addition, real-life data was used to examine the classification performance of the proposed approach and the precision, sensitivity and specificity values of the two methods were compared. As a result, we obtained that the proposed BHA gives less biased predictions than logistic regression in simulation and real-life data and has higher classification performance. Additionally, rareness levels have a significant impact on the parameter estimates of the methods.

Key Words: Black hole algorithm, Meta heuristics algorithm, Rare events, Logistic regression, Simulation study, Bias

Mathematical Subject Classification: 60E05, 62E15.

1. Background

The logistic regression determines the effect of the independent variables on the explained (dependent) variable by constructing a model between independent variables and explained variable. Unlike the standard regression in which the dependent variable takes continuous values, here a categorical and dependent variable is considered, such that the variable consists of two-level 0 and 1, or more than two levels (Hosmer and Lemeshow, 2000).

In large samples, maximum likelihood estimators of the model parameters have asymptotic unbiased and efficiency properties, and therefore, the logistic regression is used as an efficient model for classification problems of groups with homogeneous (balanced) distribution. However, for some situations which emerge depending on the internal structures

of independent and dependent variables, the constructed model may not give viable and reliable results (Bilgin, 2012). Mehta and Patel 1995 showed that the logistic regression does not give unbiased, consistent, efficient and sufficient parameter estimations with minimum variance in small or rare data sets of maximum likelihood function (Santner and Duffy, 1986; Elizabeth and Thomas, 2002) .

In this paper, we propose a novel heuristic method which can be used for the rare events where logistic regression encounters problems in parameter estimations, and mostly observed in real life data. The BHA is a meta-heuristic algorithm that mimics the star-attracting behavior of the black hole event, and various applications, modifications and hybrid versions have been developed in the literature. The BHA is firstly proposed by Hatamlou 2013 as imitating the events in nature, and being one of the heuristic methods. Inspired by the very large gravitational attraction of black holes, the developed BHA is used as an alternative to the clustering problem. Azizipanah-Abarghooee et al. 2014 proposed a new optimization approach known as gradient-based modified teach-learning-based optimization combined with black hole (MTLBO-BH) algorithm to search for optimum operating cost. Kumar et al. 2015 explained that BHA can provide a model for solving wireless sensor network optimization problems because it is independent of parameter tuning problems. Gao et al. 2016 have proposed the BHA and Spencer method to examine the stability of the high set slope of an airport. As a result, they showed that their proposed approach is very efficient for measuring the slope stability. Caio et al. 2018 used a new meta-heuristic optimization technique called BHA to examine the parallelism between the problem of commercial losses in Brazil and the irregular consumer characterization and applied it on two data sets. Wu et al. 2017 developed an adaptive multiobjective black hole algorithm (AMOBH) based on the BHA for cell density. They showed that AMOBH has a good performance to the SPEA-II, PESA-II, NSGA-II, and MOEA/D approaches in terms of convergence rate, population diversity, and population convergence. Pashaei and Aydin 2017 proposed a binary version of the BHA to solve the feature selection problem in biological data and showed that their proposed approach has better performance than other approaches in the literature. Pashaei et al. 2019 developed a hybrid meta-heuristic approach for gene selection using Binary Black Hole Algorithm (BBHA) and Binary Particle Swarm Optimization (BPSO) model. As a result, they showed that the BPSO-BBHA model could successfully define biologically and statistically significant genes known from clinical datasets.

In the literature, there are some meta-heuristic approaches as well as different classical approaches for solving rare event classification problems. Li et al., 2016 proposed a new optimization model using different swarm strategies (Bat-inspired algorithm and PSO) to increase classification performance in imbalanced data sets. They applied their proposed approach on five different unbalanced datasets consisting of lung surgery and bioanalysis scanning data and discovered that their proposed approach gave better results than other classes of balancing problems. Vergé et al., 2016 used the island particle algorithm called Monte Carlo square in their study to accurately predict rare event probabilities and discussed their results on different aviation test scenarios. Li et al., 2017 created two different methods for the unbalanced data classification problem, consisting of the effects of synthetic minority over-sampling technique (SMOTE) and the meta-heuristic method. Since the Swarm Balancing Algorithms they first used were not effective in relatively small and unbalanced data sets, they proposed the Adaptive Swarm Balancing Algorithms, which they created by updating the parameter estimates, and achieved more effective results. Ling and Zhenzhou, 2021 proposed a novel two-stage meta-model importance sampling based on the support vector machine approach for multiple failure regions and rare events. In their study, they tested the applicability of this two-stage method on different examples.

In this article, we propose a new meta-heuristic algorithm that includes BHA modification as an alternative to the classification and prediction problems encountered in rare event datasets consisting of unbalanced data sets. Unlike other studies, we test our proposed approach with a simulation study consisting of different scenarios. In the simulation study, we compare the parameter estimates, bias values, root mean square error (RMSE), classification table, and accuracy rates of the proposed BHA and the logistic regression model at different sample sizes and rareness degrees. Consequently, in small samples and cases where the imbalance is large between the groups, it is found that the proposed BHA gives less biased and reliable results compared to the logistic regression. Moreover, we compare the classification performance of the methods on the Militarized Interstate Disputes (MID) Data which is a literature data as for the rare events. As a result, we observe that the proposed BHA gives better results than the logistic regression for the specificity ratios.

1.1. Rare Events

The rare events are defined as the events occur less frequently, but make a wide effect and may give a big damage through the society (King and Zeng, 001a). Because the rare events encountered less frequently they are statistically

unexpected and non-continuous events. Therefore, the calculations of the probabilities the events statistically less probable, namely non-continuous for the occurrence in the future, and directly their prediction is a quite difficult problem. The credit card fraud, improper pronunciation of words, hurricanes, telecommunication tool crashes, militarized interstate disputes (King and Zeng, 001b) , landslides, derailments of trains, earthquakes, forest fires, diagnose of rare diseases in medicine and etc. can be given as an example for the rare events (Paal, 2014) . While the experimental data often have very large event/sample number for one of the groups, the other group is represented by only a few sample in rare event applications.

The rareness can be given in two forms. The first one is called as the relative rareness or imbalanced data, and it is a data set in which one of the considered group is much much smaller than the majority of the other group. The classification algorithm experiences problems when the data are skewed through one of the groups. The second one is called as the absolute rareness, and it is basically described as the small sample problem. Because the sampling size is very small maximum likelihood estimation starts to give worse predictions. Therefore, Allison 2012 claims that only absolute rareness becomes problematic in the framework of logistic regression, while Agresti 2002 states that the relative rareness must be considered with the comparison of number of estimators (Paal, 2014).

1.2. Binary logistic regression

In order to predict the relation between two or more variables having a cause-effect relation, to summarize the data, to make coefficient prediction among other variables, to determine and investigate the important variables effecting the dependent variable, the regression analysis is commonly used (Alpar, 2011). In linear regression analysis as one of the most important regression analysis, the dependent variable takes continuous values while the independent variable can take continuous and discrete values. If k is the independent variable and N is the number of observation, the general form of the linear regression model for the i-th observation is given by

$$y_i = \beta_0 + \beta_1 X_{1i} + \ \ldots\ldots\cdots + \beta_k X_{ki} + \varepsilon_i. \tag{1}$$

However, if the dependent variable takes two categorized values such as 0 and 1, the binary logistic regression is used instead of the linear regression. In binary logistic regression, the dependent variable is given by $y=1$ for the considered event, while the other event is given by $y=0$. Accordingly, the general form of the logistic regression is given as;

$$\eta = P(y_i/X_1, X_2, ..., X_k) = X_i^{'}\beta. \tag{2}$$

The logistic regression model for expressing the probability of considered event occurrence is given by

$$\eta = \ln\left(\frac{p_i}{1-p_i}\right) = X_i^{'}\beta$$

$$P(y_i = 1/X_i) = \frac{e^{X_i^{'}\beta}}{1+e^{X_i^{'}\beta}} \tag{3}$$

Similarly, the probability of same event not occurring is

$$1 - P(y_i = 1/X_i) = 1 - p_i = 1 - \frac{e^{X_i^{'}\beta}}{1+e^{X_i^{'}\beta}} = \frac{1}{1+e^{X_i^{'}\beta}}. \tag{4}$$

Because the dependent variable takes categorical values in the logistic model, constant-variance assumption breaks and the least square method used for parameter estimation in linear regression does not give unbiased and consistent estimations. Because the constant-variance assumption is not satisfied, the maximum likelihood method is generally used for regression coefficient estimation of logistic model (Pampel, 2000). This method enables us to estimate the parameters which make the probability of reaching observed data set maximum. In order to use the maximum likelihood method, it is necessary to construct the likelihood function. This function presents the probabilities of observed data for the unknown parameters (Santner and Duffy, 1986). Other methods used in parameter estimation are the weighted least square method and minimum logit chi-squared method (Şahin, 1999). When the parameters are estimated by using the maximum likelihood method, some other iterative methods like Newton Raphson method is used in order to obtain parameter estimations because a linear form of the parameters cannot be obtained.

1.2.1. Cases that logistic regression estimations give bad results

We generally encounter a case in which the quasi or complete separation occurs (Elizabeth and Thomas, 2002). For instance, in binary logistic regression model with single independent variable, when $Y = 1$ let $10 \leq X \leq 20$, when $Y = 0$ let $30 \leq X \leq 50$. We call this case as complete separation because the values of independent variable are separated with respect to the dependent variable. If the independent variable is categorical, only the corresponding cross cells contain data in 2×2 table ($X = 0$ values correspond to $Y = 0$, or $X = 1$ values correspond to $Y = 1$, and vice versa) (Boyle, 1996). Moreover, if the dependent variable is $Y = 1$ the independent variable is given as $10 \leq X \leq 30$, then we call this case as quasi separation. For the cases in which the independent variable is categorical this corresponds to only one empty cell in 2×2 table (Zorn, 2005). When there occurs a complete or quasi separation maximum likelihood function cannot estimate the parameters (Elizabeth and Thomas, 2002).
When the dependent variable is heterogeneous, the category of the event that we consider occurs less frequently compared to other category. When the parameter estimations of the logistic regression model for these type of data sets are conducted by maximum likelihood function, it may not give unbiased, consistent, efficient, sufficient and minimum-variance parameter estimations (Croux and Haesbroeck, 2003).

1.3. Black Hole Event

John Michell and Pierre Laplace are the first scientists who proposed the black hole concept in 18^{th} century. Although they formulate the unseen stellar theory by using Newton's gravitational law, this phenomenon is not considered as a black hole. Firstly, the American physicist John Wheeler (1967) called the mass collapse events as the black hole.
A black hole is formed due to the collapse of giant stars in space. The magnitude of a black hole's gravitational field is very large as much that even the light cannot escape from its attraction because the mass of collapsed star is concentrated in a very small region. Everything going into the effective boundary of a black hole is swallowed by the black hole and nothing can escape from its extreme gravitational force. The spherical boundary of a black hole is called as the event horizon of the black hole. The radius of the event horizon is called as the Schwarzschild radius. The minimum escape velocity from this radius is greater than the maximum universal speed, as the speed of light, and therefore even the light cannot escape from this radius. Consequently, because nothing can exceed the speed of light nothing can escape from the event horizon once it goes into the horizon. Schwarzschild radius is given by

$$R = \frac{2GM}{c^2}, \tag{5}$$

where G is gravitational constant and its value is $6.67 \times 10^{-11} N(m/kg)^2$, M is the mass of black hole and c is the speed of light.
If an object gets close to the event horizon, then it is attracted to the inside of the event horizon, then it can no more escape from the event horizon, and terminated forever according to the outer observer.

1.3.1. Black Hole Algorithm

Black hole concept is firstly introduced by Zhang et al., 2008 in order to bring a new mechanism to the PSO. This method is developed as an extension of the PSO and a new produced particle named as the black hole attracts all other particles under certain conditions. These conditions are used for accelerating the converging speed of PSO, and for preventing the local optimum problems. The event horizons of black holes, and the termination of candidates (stars) concepts are not used in this method. The better positions found by the candidates are assigned as the best positions. However, the new BHA proposed by Hatamlou, 2013 introduced these concepts, and while the best candidate is assigned as the black hole, the other candidates become the stars which get close to the best candidate - black hole. Moreover, the candidate stars falling into the event horizon vanish and some other new candidate stars are produced instead of them in the search space. By doing so the BHA is improved for the whole population. Consequently, this new proposed BHA imitates the nature of astrophysical black hole phenomenon more than the first version, and it completely differs from the previous PSO algorithms (Hatamlou, 2013).
BHA is an optimization algorithm having some common properties with other population based algorithms, such as Genetic algorithm (GA) and PSO. In the population based algorithms, a population of candidate solutions is arbitrarily produced and distributed in the search space for a given problem. Later on, the produced population is improved through the optimum solution by using certain mechanisms. For example, while the improvement is conducted by the

mutation and crossing mechanisms on genes in GA, it is improved by using the best positions obtained in the search space, and by transferring the candidate solutions through the best position in PSO.

As the core part of the study, we propose the BHA, as a novel version of the standard BHA, in which the population (stars) of arbitrarily produced candidate solutions is placed into the search space of logistic regression model. Later on, the fitness values of the population are determined and the best candidate having the best fit value are described as a black hole. Remaining candidates other than the black hole is considered as the stars.

After assigning the black hole and stars, the black ho starts to attract the stars around itself and all the stars move toward the black hole. The gravitational attraction of stars by the black hole is formulated as

$$x_i(t+1) = x_i(t) \ + rand \times (x_{BH} - x_i(t)\ ) i = 1,2,\ldots,N \tag{6}$$

where $x_i(t)$ and $x_i(t+1)$ are the position of the i-th star in t-th and $(t+1)$-th iteration, respectively. Also, x_{KD} is the position of the black hole in the search space, *rand* is an arbitrary number in $[0,1]$ interval, and N is the number of candidate solutions (stars) (Hatamlou, 2013).

When a star moves toward the black hole, it may find a better position with less cost than the black hole. In such a situation, the black hole replaces its position with that star. The BH algorithm proceeds with its new position and the remaining stars start to move toward this new black hole position.

In addition, there is the probability of stars falling into the event horizon of the black hole during their movement toward the black hole. Accordingly, each star (candidate solution) falling into the event horizon of the black hole is terminated by the black hole. In our BHA, the radius of the event horizon is obtained by

$$R = \frac{f_{KD}}{\sum_{i=1}^{N} f_i} \tag{7}$$

where f_{KD} and f_i are the logistic regression fitness values of the black hole and i-th star, respectively. In our logistic regression model, if the obtained candidate solution is less than R this candidate collapses and a new candidate is produced and arbitrarily distributed in the search space.

According to the above descriptions, main steps of the BHA can be summarized as:

Input
Population size (star size), Maximum number of iterations
Initial values of θ values ($\theta \sim$ Uniform)
Start
Start a population of stars with their arbitrary positions in the search space ($x_i \sim$ Uniform)
For j=1:number of star
Calculate the objective function by using the likelihood function of logistic regression for each star(j) and save in fitness array(f)
Next
j Assign the best star with best fitness value as the black hole
While a ¡ max_{iter} or $conv_{criteria}$
For i=1:number of star
$x_i(t+1)$=$x_i(t)$ +$rand\times$ $(x_{BH} - x_i(t))$
Evaluate fitness value of the star x_i
If fitness of $x_i >$ fitness of x_{BH}
x_{BH}=x_i
End if
Replace the new fitness value of the x_i with the previous value
Update fitness array f and calculate R=$\frac{f_{KD}}{\sum_{i=1}^{N} f_i}$
If $\sqrt{(x_{BH} - x_i)^2}$ ¡ R
Replace x_i with a new star in the search scope
End if
End for
set $a = a + 1$
End while
Return best solution
End

Here, we identify the termination criterion as 5000 iteration because of the capacity of computers.

2. Simulation study

The main assumption of logistic regression requires the both group to be balanced. However, if there exists a group imbalance in the sample this assumption is not satisfied, and the maximum likelihood estimations therefore become biased. We then compare the performance of the modification BHA and logistic regression for these imbalanced cases. For this purpose, we obtain the mean bias and RMSE values in different rareness degrees of the considered group (group "1") by performing 1000 MC simulation studies, and we investigate that for 3 different cases.

Simulation Algorithm

1. In order to determine the effect of sample size on the methods in rare events, $n = 100$, $n = 500$ and $n = 1000$ are assigned.

2. By generating the independent variable from $X \sim N(0,1)$ for Case 1, and Case 2, and from binary categorical variable for Case 3, we determine the performance of BHA for the complete separation case.

3. Regression coefficient leads to different rareness degrees depending on the values of β_0 as given in Table 1 $\beta_1 = 0$ for Case 1, and $\beta_1 = 1$ for Case 2 and Case 3 are assigned.

4. In order to generate the binary dependent variable Y, probability values are obtained from $p_i = P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$, and then by comparing a dummy value u generated from the uniform distribution it is found as

$$y_i = \begin{cases} u \leftarrow runif(1) \\ 0 \quad p_i < u \\ 1 \quad p_i \geq u \end{cases} \tag{8}$$

5. After obtaining the dependent and independent variables, by using these variables BHA and logistic regression

methods are applied to the same data. Accordingly, the mean bias and RMSE $\sqrt{E(\hat{\theta}-\theta)^2}$ obtained from the methods are investigated and the performance of the methods are compared.

Table 1: Different rareness percentages with respect to the values of β_0

β_0	-7	-6	-5	-4	-3	-2	-1	0
%	0.16	0.33	1.11	2.75	6.95	15.24	29.80	50

In Case 1, the mean bias values of two methods have become very close to the real parameter value for $\beta_0 = 0$ case with no group imbalance. When the percentage ratio of the considered rare event decreases, namely when the rareness degree increases by -2, -3, -4 for β_0, the bias of the logistic regression increases. However, in the BHA, mean bias values are approximately close to zero despite the occurrence of rareness. The occurring bias is found to be in intercept parameter rather than the slope parameter. When the rareness increases in both methods, the RMSE value increases. On the other hand, the RMSE values in logistic regression are relatively higher than the ones in BHA.

Table 2: Different rareness percentages with respect to the values of β_0 ($\beta_1 = 0$ and $x\sim$normal)(n=1000)

β_0		LOGISTIC		BHA	
		β_0	β_1	β_0	β_1
0	Bias	-0.0026	0.0079	-0.0006	0.0006
	RMSE	0.0699	0.0837	0.0144	0.0119
-1	Bias	-0.0026	0.0036	-0.0019	-0.0015
	RMSE	0.0804	0.0938	0.0134	0.0126
-2	Bias	-0.0047	0.0063	-0.0034	0.0054
	RMSE	0.1117	0.1095	0.0108	0.0113
-3	Bias	-0.0301	0.0137	0.0223	0.0053
	RMSE	0.1643	0.1415	0.0621	0.1315
-4	Bias	-0.0686	0.0173	-0.0598	0.0117
	RMSE	0.3859	0.2825	0.3685	0.2743
-5	Bias	-0.1449	0.0237	-0.0949	-0.0060
	RMSE	0.4988	0.3377	0.4119	0.3083
-6	Bias	-2.6796	0.0083	0.1065	-0.2310
	RMSE	15.0232	0.7307	0.4609	0.5064

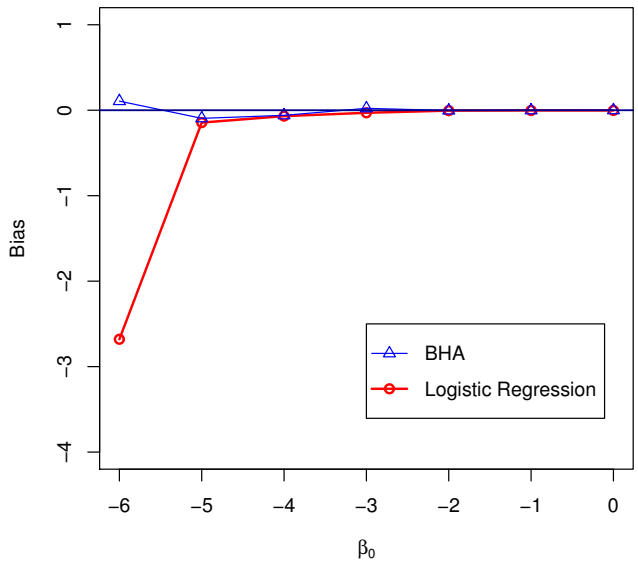

Figure 1: Comparison of bias values for both methods in Case 1

In Case 2, when the two groups are balanced with $\beta_0 = 0$ value, mean bias values of two methods are approximately close to zero, similar to Case 1. When the percentage ratio of the considered rare event decreases 0.16% from %30, the mean bias values of logistic regression increases higher than the BHA.

Table 3: Different rareness percentages with respect to the values of β_0 ($\beta_1 = 1$ and $x\sim$normal) (n=1000)

β_0		LOGISTIC		BHA	
		β_0	β_1	β_0	β_1
0	Bias	-0.0026	0.0008	0.0005	0.0006
	RMSE	0.0699	0.0837	0.0144	0.0119
-1	Bias	0.0026	0.0036	-0.0019	-0.0015
	RMSE	0.0804	0.0938	0.0134	0.0126
-2	Bias	-0.0047	-0.0063	-0.0034	-0.0054
	RMSE	0.1117	0.1095	0.0108	0.0113
-3	Bias	-0.0186	0.0126	-0.0176	0.0105
	RMSE	0.1740	0.1413	0.1730	0.1387
-4	Bias	-0.0686	0.0173	-0.0574	0.0117
	RMSE	0.3859	0.2825	0.3685	0.2743
-5	Bias	-0.1548	-0.0070	-0.0581	0.1195
	RMSE	0.5046	0.4517	0.4805	0.4664
-6	Bias	-7.5336	0.0156	-0.2909	0.1782
	RMSE	26.6786	0.5943	1.4262	0.6248

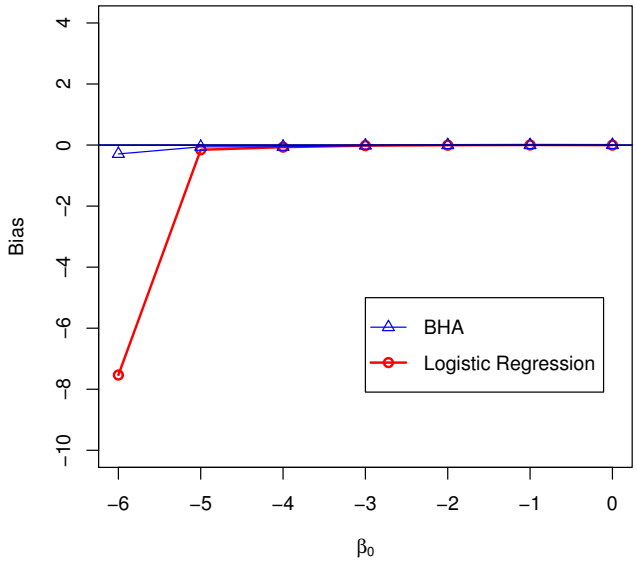

Figure 2: Comparison of bias values for both methods in Case 2

For the cases the independent variable is categorical (or it has a complete separation), we obtain much biased values because the logistic regression fails parameter estimation (In Table 3). Correspondingly, we find that the BHA has much better mean bias and RMSE values for the complete separation case.

Table 4: Different rareness percentages with respect to the values of β_0 ($\beta_1 = 1$ and x categorical)(n=1000)

β_0		LOGISTIC		BHA	
		β_0	β_1	β_0	β_1
0	Bias	-99.5661	204.1321	-0.7407	2.9072
	RMSE	99.5661	204.1321	0.9656	3.2203
-1	Bias	-100.5661	204.1321	-1.3371	3.1023
	RMSE	100.5661	204.1321	1.4852	3.3370
-2	Bias	-101.5661	204.1321	-1.7731	3.1946
	RMSE	101.5661	204.1321	1.8798	3.4421
-3	Bias	-102.5661	204.1321	-2.8172	8.0436
	RMSE	102.5661	204.1321	2.8172	8.0447

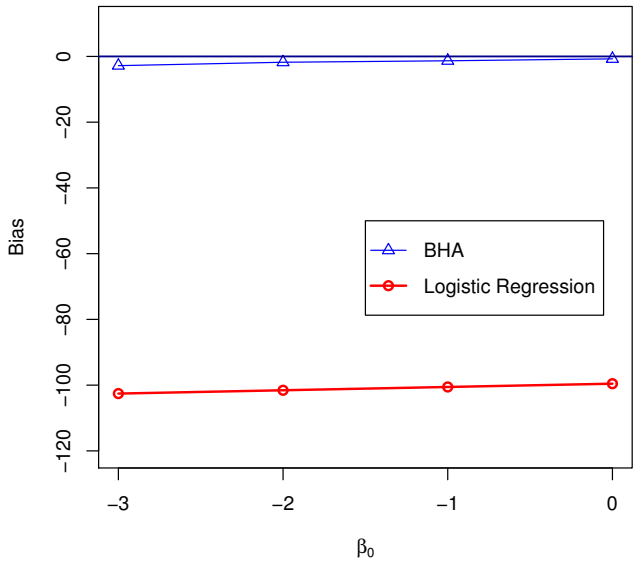

Figure 3: Comparison of bias values for both methods in Case 3

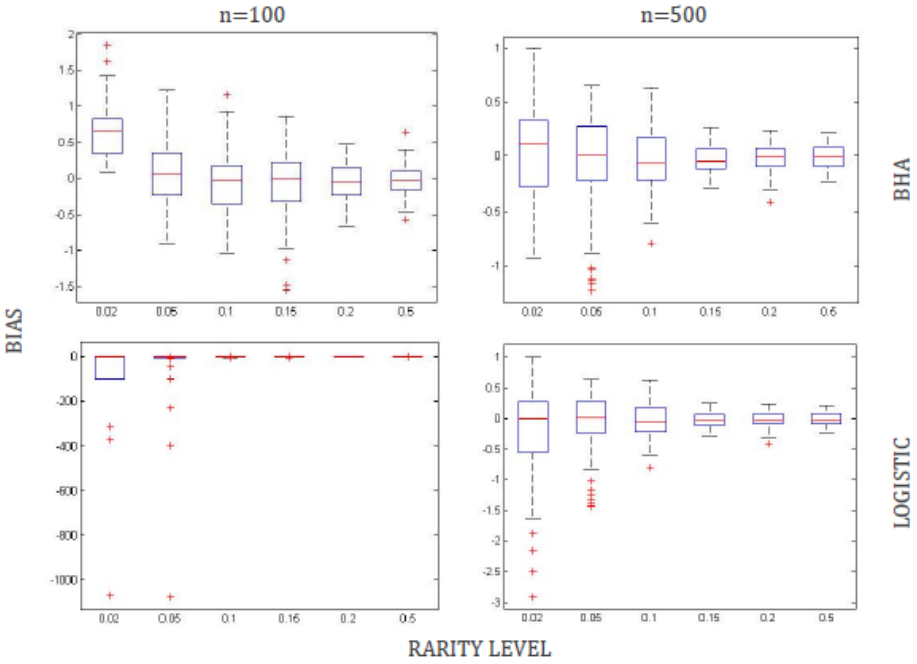

Figure 4: Box-plot graph of mean bias values for both methods in Case 2

In Figure 4, we investigate the mean bias values of both methods with respect to the given sample sizes and conclude that the BHA method gives better results compared to the logistic regression methods. Moreover, we observe that the bias values of both method decreases when the sample size increases.

3. Data Analysis

In this section, we make a real life application on the MID Data with a rareness degree of 0.3% by choosing a sample including 1000 units with a rareness degree of 5% (King and Zeng, 001b). Here, the dependent variable "conflict" represents the international disputes and takes the values 1 if dispute exists, and 0 if dispute does not exist. The independent variables are codded as major, contig, power, maxdem, mindem and years. The descriptive variables consist of the variables used in this region. Major represents whether a country has a super power or not; contig represents whether a country has foreign policy portfolio and allies or not; power represents the military power; maxdem and mindem represent the maximum and minimum degrees of business dependencies, respectively; years represent the time passing from the last dispute (King and Zeng, 001b).

Table 5: Coefficient estimations of both methods for MID Data

Coefficient	Coefficient estimations	
	LOGISTIC	BHA
Constant	-6.3673	-5.3357
Major	2.2238	3.2381
Contig	3.9542	6.7005
Power	0.8018	-0.1920
Maxdem	0.1443	0.2039
Mindem	-0.0641	0.1838
Years	-0.0031	-2.7086
AIC	183.94	11.23776

When we investigate the coefficient estimations for the MID Data in Table 5, we observe that the BHA gives much smaller AIC value than that of the logistic regression method. Consequently, we infer that the best model is the BHA method.

Table 6: Classification performance of both methods

	LOGISTIC	BHA
Real $Y = 0$ number	970	970
Correct classification number	964	882
Real $Y = 1$ number	30	30
Correct classification number	7	25
Precision	23.33	83.33
Sensitivity	0.53	0.22
Specificity	0.976	0.994

In order to compare the correct classification percentage of both methods for the rare group, we use the Specificity percentages. According to the classification performance of both methods in Table 6, the logistic regression can assign only 7 correct groups for situations that the confliction may exists. The precision (Positive predictive value) percentage of the test is found to be 23.33 for the logistic regression. However, the specificity percentage of BHA is found to be 83.33, much better than the logistic regression.

4. Conclusions

It is very important to describe the events which have vital importance and rare frequency, and whose results may be fatal or very dangerous (earthquakes, fatal virus pandemics, wars, etc.), by using correct prediction models. The logistic regression model which is generally used for the classification, cannot correctly predict the probabilities of the rare events mostly because of the bias in intercept parameters estimation. Various classical statistical methods are available in the literature to improve the bias caused by the rare event. However, meta-heuristic algorithms were not used for the bias in parameter estimates of the rare event. In this article, a modification of the BHA, which had different applications on clustering and variable selection, was developed on rare event classification and an alternative was proposed for logistic regression parameter estimations.

In order to compare the parameter estimates obtained from the logistic regression and BHA, we conducted a simulation study with three different scenarios based on the intercept parameter and different rareness percentages. When the simulation results are examined, it is seen that BHA gives less biased parameter estimates compared to logistic

regression in all scenarios. In addition, in the simulation study conducted to examine the effect of sample size on the bias in parameter estimations, we observed that BHA gave better results compared to the logistic regression model. At the same time, we observed that the BHA was better than the logistic regression model, according to the AIC value, which was viewed to examine the relative superiority of the models to each other.

We examined the precision, sensitivity and specificity values obtained from the MID data set to examine the classification performance of the BHA and logistic regression applied for the bias in rare event parameter estimations. It is seen that the precision value obtained from the BHA is higher than the logistic regression model. As a result, in the rare event situation, it is seen that the BHA predicts actually positive values more accurately than logistic regression. Likewise, it was seen that the BHA predicted negative values more accurately than logistic regression.

As a result, we observed that the developed modification of the BHA gives better results than the logistic regression in rare events according to different rareness levels, different sample sizes and classification percentages examined on real-life data. Accordingly, we recommend the use of meta-heuristic algorithms such as black hole instead of logistic regression in order to achieve correct classification percentages in rare events such as earthquakes, deadly epidemics, and wars, and to obtain less biased parameters and to make correct interpretations.

The approach proposed in this study can be extended with different meta-heuristic algorithms and mathematical approaches to obtain a more reliable and accurate prediction of rare events. The proposed BHA approach for rare events can be applied on models simulated from different distributions (Almuqrin et al., 2022, Rasekhi et al., 2022, Althubyani et al., 2022, Alghamdi et al., 2023, Atchadé et al., 2023) and the performance of the model can be examined. Additionally, by combining the BHA with different meta-heuristic algorithm, a new hybrid algorithm can be proposed, its performance on rare events can be compared, and new approaches can be proposed for the rare event prediction and classification problem.

References

1. Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Canada.
2. Şahin, M. (1999). *Lojistik regresyon ve biyolojik alanlarda kullanımı, Thesis*. Kahramanmaraş Sütçü İmam University, Turkey.
3. Alghamdi, S., Shrahili, M., Hassan, A., Gemeay, A., Elbatal, I., and Elgarhy, M. (2023). Statistical inference of the half logistic modified kies exponential model with modeling to engineering data. *Symmetry*, 15.
4. Allison, P. (2012). *Logistic Regression Using SAS: Theory and Application*. SAS Institute, USA.
5. Almuqrin, M., Gemeay, A., Abd El-Raouf, M., Kilai, M., Aldallal, R., and Hossam, E. (2022). A flexible extension of reduced kies distribution: Properties, inference, and applications in biology. *Complexity*, 2022:1–19.
6. Alpar, C. (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. Detay, Ankara, Turkey.
7. Althubyani, F., El-Bar, A., Fawzy, M. A., and Gemeay, A. (2022). A new 3-parameter bounded beta distribution: Properties, estimation, and applications. *Axioms*, 11:504.
8. Atchadé, M. N., N'bouké, M., Djibril, A., Shahzadi, S., Hossam, E., Aldallal, R., Alshanbari, H. M., Gemeay, A., and El-Bagoury, A. A.-A. (2023). A new power topp-leone distribution with applications to engineering and industry data. *PloS one*, 18:e0278225.
9. Azizipanah-Abarghooee, R., Niknam, T., Bavafa, F., and Zare, M. (2014). Short-term scheduling of thermal power systems using hybrid gradient based modified teaching–learning optimizer with black hole algorithm. *Elect Power Syst Res*, 108:16–34.
10. Bilgin, M. (2012). *Türetilmiş İkili Heterojen Veri Yapılarında Genel, Sağlam ve Kesin Lojistin Regresyon Yöntemlerinin Karşılaştırılması*. Eskişehir Osmangazi University, Eskişehir, Turkey.
11. Boyle, M. J. (1996). *Quasicomplete Separation in Logistic Regression: A Medical Example*. The SouthEast SAS Users Group Fourth Annual Conference, Charlotte, Nort Carolina.
12. Caio, C. O. R., Douglas, R., André, N. S., and João Paulo, P. (2018). On the study of commercial losses in brazil: A binary black hole algorithm for theft characterization. *IEEE Transactions on Smart Grid*, 9:676–683.
13. Croux, C. and Haesbroeck, G. (2003). Implementing the bianco and yohai estimator for logistic regression. *Computational Statistics & Data Analysis*, 44:273*295.
14. Elizabeth, N. and Thomas, P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *American Statistical Association*, 56:163–170.

15. Gao, W., Wang, X., Dai, S., and Chen, D. (2016). Study on stability of high embankment slope based on black hole algorithm. *Environ Earth Sci*, 75:1381.
16. Hatamlou, M. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information Sciences: an International Journal*, 222:175–184.
17. Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, New York, USA.
18. King, G. and Zeng, L. (2001a). Logistic regression in rare events data. *Political Analysis*, 9:137–163.
19. King, G. and Zeng, L. (2001b). Explaining rare events international relations. *International Organization*, 55:693–715.
20. Kumar, S., Datta, D., and Singh, S. (2015). Black hole algorithm and its applications. *In: Computational intelligence applications in modeling and control*, pages 147–170.
21. Li, J., Fong, S., Mohammed, S., and Fiaidhi, J. (2016). Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. *The Journal of Supercomputing*, 72:3708–3728.
22. Li, J., Liu, L.-s., Fong, S., Wong, R., Mohammed, S., Fiaidhi, J., Sung, Y., and Wong, K. (2017). Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *Computerized Medical Imaging and Graphics*, page e0180830.
23. Ling, C. and Zhenzhou, L. (2021). Support vector machine-based importance sampling for rare event estimation. *Structural and Multidisciplinary Optimization*, 63.
24. Mehta, C. and Patel, R. (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine*, 14:2143–2160.
25. Paal, B. (2014). *A Comparison of Different Methods for Modelling Rare Events Data, Master Thesis*. Gent Universiteit, Belgium.
26. Pampel, F. (2000). *Logistic regression a primer*. Sage University Papers, London.
27. Pashaei, E. and Aydin, N. (2017). Binary black hole algorithm for feature selection and classification on biological data. *Applied Soft Computing*, 56:94–106.
28. Pashaei, E., Pashaei, E., and Aydin, N. (2019). Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization. *Genomics*, 111:669–686.
29. Rasekhi, M., Saber, M. M., Hamedani, G., Abd El-Raouf, M., Aldallal, R., and Gemeay, A. (2022). Approximate maximum likelihood estimations for the parameters of the generalized gudermannian distribution and its characterizations. *Journal of Mathematics*, 2022.
30. Santner, T. and Duffy, E. (1986). A note on a. albert and j.a. anderson's conditions for the existence of maximum likelihood estimates İn logistic regression models. *Biometrica*, 73:755–758.
31. Vergé, C., Morio, J., and Del Moral, P. (2016). An island particle algorithm for rare event analysis. *Reliability Engineering System Safety*, 149:63–75.
32. Wu, C., T., W., Fu, K., Zhu, Y., Li, Y., He, W., and Tang, S. (2017). Amobh: Adaptive multiobjective black hole algorithm. *Comput Intell Neurosci*, 2017:6153951, doi: 10.1155/2017/6153951.
33. Zhang, K., Liu, Y., and He, X. (2008). *Random black hole particle swarm optimization and its application*. IEEE International Conference Neural Networks and Signal Processing.
34. Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13:1–28.