



Available online at <http://jeasiq.uobaghdad.edu.iq>

Selection of variables Affecting Red Blood Cell by Firefly Algorithm

Rehab Hamza Obeid

Department of Statistics / College of
Administration and Economics /University of
Baghdad
Baghdad, Iraq
rehab.hamza1201a@coadec.uobaghdad.edu.iq

Nazik J. Sadik

Department of Statistics / College of
Administration and Economics /University of
Baghdad
Baghdad, Iraq
dr.nazik@coadec.uobaghdad.edu.iq

Received: 20/12/2022

Accepted: 22/1/2023

Published: June / 2023



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

Abstract

Some maps of the chaotic firefly algorithm were selected to select variables for data on blood diseases and blood vessels obtained from Nasiriyah General Hospital where the data were tested and tracking the distribution of Gamma and it was concluded that a Chebyshevmap method is more efficient than a Sinusoidal map method through mean square error criterion.

Paper type: Research paper.

Keywords: Gamma regression model, Firefly algorithm, variables selection.

1. Introduction

Selecting or identifying variables is important in machine learning and data mining. Algorithms are one of the best modern methods in selecting independent variables that affect the independent variable in regression models and this paper data were selected for blood diseases, where it was found that they follow the distribution of gamma where the dependent variable is distributed continuously with a positive value, where this type of regression is applied in health, economic and other data.

In this paper, data related to blood and vascular diseases were selected, where the effect of 49 independent variables affecting the dependent variant, which is red blood cells (RBC), was studied.

1.1 Literature review

Yang (2013) presented a new algorithm, the multiobjective firefly algorithm (MOFA) for variable selection which was compared with a set of optimization algorithms such as the multiobjective bee algorithm and others, where the proposed algorithm proved its superiority.

Bossio and Cuervo (2015) suggested the linking functions (Identity link function) and (Log link function) to link the gamma regression model and the parameters of the model were estimated in maximum likelihood method through simulation. It was applied to real data, which is the effect of temperature on the growth of an insect (fruit fly). The simulation results showed that the parameter estimates are close to the assumed parameters of the model values and for the real data that temperature affects the stages of fruit fly development.

Zhang et al (2016) introduced the new Binary firefly (BFA) algorithm to remove redundant variables and determine optimal parameters of the model by identifying DNA actually active proteins.

Amin and Qasim (2019) proposed the shrinkage method to estimate the parameters of the gamma regression model. It was compared with the maximum likelihood method, and simulations were conducted and applied to real data, and the comparison criterion was the mean square error (MSE), the results showed that the method of minimization is better than the method of maximum likelihood.

While Zakariya ALgamal et al (2020) proposed the Gray Wolf algorithm (GWO) to select a variable for the gamma regression model. The results showed the efficiency of the proposed algorithm compared to other common methods.

Ahmed ALkhateeb and Zakariya ALgamal (2021) presented a paper that dealt with the variable selection of the Kama regression model using the firefly algorithm, and it was compared with known statistical methods. The results showed the efficiency of the proposed methods. This method was applied to real data related to chemical measurements

2. Materials and Methods

The data on red blood cells were tested using the statistical program (EasyFit) to test for good conformity and it was found that the red blood cells follow the distribution of gamma distribution as shown in Figure (1).

[#6] Gamma					
Kolmogorov-Smirnov					
Sample Size	157				
Statistic	0.05253				
Rank	8				
α	0.2	0.15	0.1	0.05	0.01
Critical Value	0.0854	0.09098	0.09737	0.10854	0.13009
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size	157				
Statistic	0.59018				
Rank	8				
α	0.2	0.15	0.1	0.05	0.01
Critical Value	1.3749	1.6024	1.9286	2.5018	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom	7				
Statistic	7.3086				
Rank	8				
α	0.2	0.15	0.1	0.05	0.01
Critical Value	9.8032	10.748	12.017	14.067	18.475
Reject?	No	No	No	No	No

Figure (1): Test for good conformity of gamma distribution

2.1 Gamma Regression Model

A two- Parameters gamma distribution can be taken by a random variable (y) (Adekanmbi, 2017).

$$f(y; \theta, \tau) = \frac{\theta^\tau}{\Gamma(\tau)} y^{\tau-1} \exp(-\theta y) I_{(0, \infty)} y \quad \tau, \theta > 0 \quad (1)$$

Where:

(τ) is the Shape paramete

$\Gamma(\cdot)$ is the Scale parameter

$$E(Y_i) = \frac{\tau}{\theta}$$

$$Var(Y_i) = \frac{\tau}{\theta^2} = \mu^2 \left(\frac{1}{\tau}\right) = \sigma^2 (E(Y_i))^2$$

CDF IS given by:

$$F(y) = \frac{1}{\Gamma(\tau)} \int_0^y u^{\tau-1} e^{-u} du$$

$$Y_i \sim \text{Gamma}(\mu_j, \tau_j),$$

where $j = 1, 2, \dots, n$, (Adekanmbi, 2017)

2.2 Maximum Likelihood Method

Cuervo (2001) introduced the maximum likelihood estimation method as follows:

$$\frac{\partial L}{\partial \gamma_k} = \sum_{i=1}^n -\tau_i \left[\frac{d}{d\tau_i} \log \log \Gamma(\tau_i) - \log \log \left(\frac{\tau_i y_i}{\mu_i} \right) - 1 + \frac{y_i}{\mu_i} \right] z_{ik}$$

$$k = 1, \dots, r \quad (j \geq k)$$

$$p \geq r$$

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^n \frac{\tau_i}{\mu_i^2} \left(1 - \frac{2y_i}{\mu_i} \right) x_{ij} x_{ik} \quad ; j, k = 1, \dots, p$$

$$\frac{\partial^2 L}{\partial \gamma_k \partial \beta_j} = \sum_{i=1}^n -\frac{\tau_i}{\mu_i} \left(1 - \frac{y_i}{\mu_i} \right) x_{ij} z_{ik}$$

$$\begin{aligned} \frac{\partial^2 L}{\partial \gamma_k \partial \gamma_j} &= \sum_{i=1}^n -\tau_i \left[\frac{d}{d\tau_i} \log \log \Gamma(\tau_i) - \log \log \left(\frac{\tau_i y_i}{\mu_i} \right) - 1 + \frac{y_i}{\mu_i} \right] z_{ij} z_{ik} \\ &\quad - \sum_{i=1}^n \tau_i \left[\tau_i \frac{d^2}{d\tau_i^2} \log \log \Gamma(\tau_i) - 1 \right] z_{ij} z_{ik} \end{aligned}$$

Where: (Adekanmbi, 2017)

$$I(B) = \left[-E \left(\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} \right) - E \left(\frac{\partial^2 L}{\partial \gamma_k \partial \beta_j} \right) - E \left(\frac{\partial^2 L}{\partial \gamma_k \partial \gamma_j} \right) - E \left(\frac{\partial^2 L}{\partial \gamma_k \partial \gamma_j} \right) \right]$$

$$-E \left(\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} \right) = \sum_{i=1}^n \frac{\tau_i}{\mu_i^2} x_{ij} x_{ik}$$

$$-E \left(\frac{\partial^2 L}{\partial \gamma_k \partial \beta_j} \right) = 0 \quad ; k = 1, 2, \dots, r ; j = 1, 2, \dots, p$$

$$-E \left(\frac{\partial^2 L}{\partial \gamma_k \partial \gamma_j} \right) = \sum_{i=1}^n \tau_i^2 \left[\frac{d^2}{d\tau_i^2} \log \log \Gamma(\tau_i) - \frac{1}{\tau_i} \right] z_{ij} z_{ik} \quad ; j, k = 1, 2, \dots, r$$

$$I(\beta) = \left[\sum_{i=1}^n \frac{\tau_i}{\mu_i^2} x_{ij} x_{ik} \quad 0 \quad 0 \quad \sum_{i=1}^n \tau_i^2 \left[\frac{d^2}{d\tau_i^2} \log \log \Gamma(\tau_i) - \frac{1}{\tau_i} \right] z_{ij} z_{ik} \right]$$

$$\hat{\beta}^{(h+1)} = (XW_1^{(h)} X)^{-1} XW_1^{(h)} Y$$

$$W_1^h = \frac{(\mu_i^2)^{(h)}}{\tau_i^{(h)}}$$

$$\hat{\gamma}^{(h+1)} = (ZW_2^{(h)} Z)^{-1} XW_2^{(h)} Y$$

$$W_2^h = \frac{1}{\tau_i^{(h)}}$$

$$di = \tau_i^{-2} \left[\frac{d^2}{d\tau_i^2} \log \log \Gamma(\tau_i) - \frac{1}{\tau_i} \right]^{-1}$$

2.3 Firefly Algorithm

The firefly algorithm is an improvement algorithm inspired by the collective behavior of fireflies through bright lighting.

There are three rules in the FA (Yu et al., 2015):

- Regardless of the gender of a firefly, one firefly can be attracted to another
- The attraction of the firefly increases with the brightness of the brighter fireflies as they are attracted to them, and if there is no sufficiently illuminated firefly, they will move randomly.
- That the intensity of the firefly brightness depends on the fitness function, where the intensity of the firefly brightness is directly relevant to the value of the fitness function.
- The Cartesian distance can be described as follows:

$$r(s_i, s_j) = \sqrt{\sum_{c=1}^d (s_{i,c} - s_{j,c})^2} \quad (2)$$

Applications I_I can be rounded up as follows:

$$I(z) = I_0 e^{-\delta z^2} \quad (3)$$

I_0 : is the (original intensity of light) .

$$\varphi(z) = \varphi_0 e^{-\delta z^2}$$

Where the φ_0 represents gravity when $z = 0$

The best place for a firefly to move location is:

$$s_i^{(t+1)} = s_i^{(t)} + \varphi_0 e^{-\delta z_{i,j}^2} (s_j^{(t)} - s_i^{(t)}) + \alpha(k_1 - 0.5) \quad (4)$$

Zhan et al (2016) suggested a BFA.

The location is binary, and the result of choosing variables is expressed as yes or no, the variable that is chosen is denoted by the number 1, while the variable that is not chosen is denoted by the symbol 0 and therefore it is expressed by a binary vector BFA.

$$\varphi_0 e^{-\delta z_{i,j}^2} (s_j^{(t)} - s_i^{(t)}) + \alpha(k_1 - 0.5)$$

$$\text{Sig} = \frac{1}{1 + \exp[\varphi_0 e^{-\delta r_{i,j}^2} (s_j^{(t)} - s_i^{(t)}) + \alpha(h_1 - 0.5)]} \quad (5)$$

Fire flies position in $s_i^{(t+1)} = \{1 \text{ if } \text{sigm} \geq k_2 \text{ 0 otherwise}$

2.4 The suggested Maps

Chaotic maps are conceived and can be transmitted as particles in a limited range of nonlinear, Dynamic, and Nonlinear systems (Sayed et al., 2018).

Table (1): Explanation of the Two Maps

Maps	Function	Range
Chebyshef	$x_{h+1} = \cos\left(h \cos^{-1}\left(x_h\right)\right)$	(-1, 1)
Sinusoidal	$x_{h+1} = 2.3x_h \sin\left(\pi x_h\right)$	(0, 1)

3. Discussion of Results

Cardiovascular disease is a general term for conditions that affect the heart or blood vessels and is usually associated with atherosclerosis and an increased risk of blood clots. It is considered one of the most dangerous deadly diseases in humans.

The increasing prevalence of cardiovascular disease with high mortality rates represents a major risk and burden to healthcare systems around the world. Two methods of selecting variables, Chebyshev and Sinusoidal method, were applied to the study data, which includes 49 independent variables affecting the RBC dependent variable for 157 people, the study data are:

- V₁ Gender (Male, Female)
- V₂ Age
- V₃ Weight
- V₄ Social status (Single, Married, Divorced, Widowed)
- V₅ Referral body (Hospital, Clinic)
- V₆ Housing environment (Rural, City)
- V₇ Entry type (First, Second, Repeated)
- V₈ patient's stay (1, 2, 3, days)
- V₉ Review status (As per appointment, urgent)
- V₁₀ Exit status (improved, recovered, the same condition, the patient is responsible, dead)
- V₁₁ Have surgery or not
- V₁₂ Employment (unemployed, employee, retired, housewife, disabled, student)
- V₁₃ Employment Sector (Public Sector, Private Sector)
- V₁₄ Blood Group (O, A, B, AB)
- V₁₅ Rhesus analyzes RH (RH+, RH-)
- V₁₆ Random blood sugar (R B sugar) [mg/dl]
- V₁₇ Hemoglobin Test (HGB) [g/dl]
- V₁₈ Packed cell volume (PVC) [L/L].
- V₁₉ analysis of census of white blood cells (WBC) [Cu.mm].
- V₂₀ Creatinine blood test (S C) [mg/dl].
- V₂₁ IgM test (Positive, Negative)
- V₂₂ IGg test (Positive, Negative)
- V₂₃ Troponin (Positive, Negative)
- V₂₄ HCV (Positive, Negative)
- V₂₅ B-Urea (Blood urea) [mg/L]
- V₂₆ blood pressure low (BPL)
- V₂₇ Blood pressure high (BPH)
- V₂₈ prolactin analysis (PR) [b/m]
- V₂₉ oxygen saturation (SPO2) (%)
- V₃₀ Temperature degree (Temp.)
- V₃₁ Heart rate (HR) (heartbeats per unit of time]
- V₃₂ RBC
- V₃₃ Hematocrit analysis (HCT) (%)
- V₃₄ mean corpuscular volume (MCV) [FL]
- V₃₅ mean corpuscular hemoglobin (MCH) [g/dL]
- V₃₆ Mean corpuscular hemoglobin concentration (MCHC) [g/dL]
- V₃₇ platelet count analysis (PLT) [x10³/mL]
- V₃₈ Lymphocytes (Lym%)
- V₃₉ Mixed Cells Absolute Count (MDX %)
- V₄₀ S Cholesterol [mg/dl]
- V₄₁ Neutrophils (NEUT) (%)

- V_{42} Lymphocyte Absolute Count (LYM#) [$\times 10^3/\text{mL}$]
- V_{43} Mixed Cells Absolute Count (MXD#) [$\times 10^3/\text{mL}$]
- V_{44} Neutrophil Absolute Count (NEUT#) [$\times 10^3/\text{mL}$]
- V_{45} Red Cell Distribution Width (RDW-SD) [fL]
- V_{46} Red Cell Distribution Width (RDW-CV) %
- V_{47} Platelet Distribution Width (PDW) [fL]
- V_{48} Mean Platelet Volume (MPV) [fL]
- V_{49} P-LCR% platelet large cell ratio (P-LCR) %
- V_{50} procalcitonin (PCT%)

Furthermore, the algorithm was implemented using the R program, and the following results were obtained:

Table(2)

Variables	Solution Chebyshe v map	Variables	Solution Chebyshe v map
V1	1	V27	1
V2	0	V28	0
V3	0	V29	0
V4	1	V30	1
V5	1	V31	1
V6	1	V32	0
V7	1	V33	1
V8	0	V34	0
V9	0	V35	1
V10	0	V36	1
V11	0	V37	1
V12	0	V38	0
V13	0	V39	1
V14	1	V40	0
V15	1	V41	1
V16	0	V42	1
V17	1	V43	1
V18	0	V44	0
V19	1	V45	1
V20	0	V46	1
V21	0	V47	0
V22	0	V48	0
V23	1	V49	0
V24	1	MSE=0.2 407996	
V25	1		
V26	0		

Table (2) represents the Chebyshev map method for selecting variables where (25) independent variables that affect the dependent variable (RBC) have been selected, namely Gender, marital status, referrer, housing environment, blood type, rhesian factor (RH)), cell size (PVC)), creatine blood test (SC), hepatitis virus type (, (blood urea , hypotension BPL)), analysis of righteousness and lactin PH)) , heart rate (HR), average body size (MCV), average concentration of somatic hemoglobin (MCHC)), platelet count analysis (PLT), lymphocytes (LYM)), cholesterol, absolute lymphocyte count (LYM)), Absolute number of mixed cells (MXD), absolute neutrophils #NEUT, red cell distribution display (RDW–CV), platelet distribution display (PDW)

Table (3)

Variables	Solution Sinusoidal map	Variables	Solution Sinusoidal map
V1	1	V26	1
V2	0	V27	0
V3	0	V28	1
V4	1	V29	0
V5	0	V30	1
V6	0	V31	1
V7	0	V32	0
V8	0	V33	0
V9	1	V34	1
V10	0	V35	0
V11	0	V36	1
V12	0	V37	1
V13	1	V38	0
V14	1	V39	0
V15	0	V40	1
V16	0	V41	1
V17	0	V42	0
V18	0	V43	0
V19	0	V44	1
V20	1	V45	0
V21	1	V46	1
V22	1	V47	1
V23	1	V48	1
V24	0	V49	1
V25	1	MSE=0.24 94499	

Table (3) shows the results of the (Sinusoidal) method of selecting variables, where 24 independent variables were selected that affect the RBC dependent variable, namely: Gender, marital status review status, employment sector, employment species, IgM test, (Igg test), troponin, hepatitis virus type(c), (Clow blood pressure BPL), hypertension (BPH), oxygen saturation (SPO2), heart rate (HR), hemoglobin test (HGB), somatic hemoglobin (MCH), lymphocytes (LYM)%, neutrophils (NEUT%) , absolute lymphocyte number (LYM)#, red cell distribution display (RDW-SD)), platelet distribution display PDW , Average platelet size (MPV), large platelet ratio (P-LCP), procalstyone PCT.

Table (4): Variables Selection using maps

Method	Variables Selection	MSE
Chebyshev map	25	0.2407996
Sinusoidal map	24	0.2494499

Regression equation according to the variables selected according to the method:

$$\begin{aligned}
 \hat{Y}_i &= E(Y_i) = \hat{\mu}_i \\
 &= \beta_0 + 2.548V_1 - 5.117V_4 - 2.438V_5 - 1.355V_6 \\
 &\quad + 2.070V_7 - 3.477V_{14} + 1.939V_{15} + 2.197V_{17} \\
 &\quad - 2.108V_{19} + 1.341V_{23} - 1.482V_{24} - 1.519V_{25} \\
 &\quad - 6.898V_{27} + 1.990V_{30} + 3.056V_{31} - 7.633V_{33} \\
 &\quad - 2.531V_{35} - 1.521V_{36} + 9.751V_{37} - 1.314V_{39} \\
 &\quad + 2.012V_{41} - 2.567V_{42} - 1.099V_{43} + 2.550V_{45} \\
 &\quad + 6.047V_{46} \quad \dots (6)
 \end{aligned}$$

4. Conclusion

- The results obtained by applying some chaotic maps of the binary firefly algorithm showed Chebyshev map the efficiency of a comparative Sinusoidal map method of selecting variables for the Gamma regression model in table (3)
- In equation (8), we note that each of the independent variables $V_1, V_7, V_{15}, V_{17}, V_{23}, V_{30}, V_{31}, V_{37}, V_{41}, V_{45}, V_{46}$ variables have a direct effect on the RBC (response variable).
- we note that each of the independent variables $V_4, V_5, V_6, V_{14}, V_{19}, V_{24}, V_{25}, V_{27}, V_{33}, V_{35}, V_{36}, V_{39}, V_{42}, V_{43}$ It has the opposite effect on the RBC(response variable).

References

1. Adekanmbi, D. B. (2017), "Generalized Gamma Regression Models with Application to
2. Al-Abood, A. and Young, D. (1986). Improved deviance goodness of fit statistics for a gamma regression model. *Communications in Statistics-Theory and Methods*, 15(6):1865–1874.
3. Alkhateeb, A. N., and Algamal, Z. Y. 2021. Variable selection in gamma regression model using chaotic firefly algorithm with application in chemometrics. *Electronic Journal of Applied Statistical Analysis*, vol 14no 1, pp 266-276.
4. Al-Thanoon, N. A., Qasim, O. S., and Algamal, Z. Y. (2020). Variable selection in gamma regression model using binary gray wolf optimization algorithm. In *Journal of Physics: Conference Series*, volume 1591, page 012036. IOP Publishing.
5. Amin, M., Qasim, M., & Amanullah, M. 2019, "Performance of Asar and Genç and Huang and Yang's Two-Parameter estimation methods for the gamma regression model" , *Iranian Journal of Science and Technology, Transactions A: Science*, vol 43 no 6,pp 2951-2963
6. Bossio, M. C., & Cuervo, E. C. (2015), "Gamma regression models with the Gammareg R package", *Comunicaciones en Estadística*, 8(2), 211- 223.
7. CD4 Cell Counts Data of Aids Patients" , *International Journal of Applied Mathematics & Statistical Sciences (IJAMSS)*, 6(4), 19-36.
8. Cuervo, E. C. (2001), "Modelagem da variabilidade em modelos lineares generalizados" (Doctoral dissertation, Tese de D. Sc., IM– UFRJ, Rio de Janeiro, RJ, Brasil).
9. Qasim, O. S., Al-Thanoon, N. A., and Algamal, Z. Y. (2020). Feature selection based on chaotic binary black hole algorithm for data classification. *Chemometrics and Intelligent Laboratory Systems*, 204:104104.
10. Sayed, G. I., Darwish, A., and Hassanien, A. E. (2018). A new chaotic whale optimization algorithm for features selection. *Journal of classification*, 35(2):300–344.

11. Stacy, E. (1962), " A Generalized of the Gamma distribution " , The Annals of Mathematical Statistics" , vol.33, No.1 pp.1187-1992.
12. Yang, X.-S. (2013). Multiobjective firefly algorithm for continuous optimization. Engineering with Computers, 29(2):175–184.
13. Yu, S., Zhu, S., Ma, Y., and Mao, D. (2015). Enhancing firefly algorithm using generalized opposition.
14. Zhang, J., Gao, B., Chai, H., Ma, Z., and Yang, G. (2016). Identification of dna-binding proteins using multi-features fusion and binary firefly optimization algorithm. BMC bioinformatics, 17(1):323.

اختيار المتغيرات التي تؤثر على خلايا الدم الحمراء بواسطة خوارزمية اليراع

نازك جعفر صادق
جامعة بغداد/ كلية الادارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
dr.nazik@coadec.uobaghdad.edu.iq

رحاب حمزة عبيد
جامعة بغداد/ كلية الادارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
rehab.hamza1201a@coadec.uobaghdad.edu.iq

Received: 20/12/2022

Accepted: 22/1/2023

Published: June / 2023

هذا العمل مرخص تحت اتفاقية المشاع الابداعي نسب المصنّف - غير تجاري - الترخيص العمومي الدولي 4.0

[Attribution-NonCommercial 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

مستخلص البحث

تم اختيار بعض خرائط خوارزمية اليراع الفوضوية لاختيار المتغيرات للبيانات الخاصة بأمراض الدم والأوعية الدموية التي تم الحصول عليها من مستشفى الناصرية العام حيث تم اختبار البيانات ووجد انها تتبع توزيع جاما وتم التوصل إلى أن طريقة Chebyshev map أكثر كفاءة من طريقة Sinusoidal map بحسب معيار متوسط مربعات الخطأ (MSE).

نوع البحث: ورقة بحثية.

المصطلحات الرئيسية للبحث: انحدار كاما، خوارزمية اليراع، طريقة الامكان الاعظم، متوسط مربعات الخطأ.

*البحث مستل من رسالة ماجستير