Optimizing The Cost Function Of The Logistic Regression Parameters Of A Diabetes Classification Problem Using Hybrid Chaotic Pattern Search

Adithya Nair
AID23002
Amrita School Of Engineering
Bengaluru, India
bl.en.u4aid23002@bl.students.amrita.edu

Avighna Reddy Katipally
AID23005
Amrita School Of Engineering
Bengaluru, India
bl.en.u4aid23005@bl.students.amrita.edu

P Ananthapadmanabhan Nair
AID23036
Amrita School Of Engineering
Bengaluru, India
bl.en.u4aid23005@bl.students.amrita.edu

***Abstract*—This project aims to demonstrate the usage of the Hybrid Chaotic Pattern Search Algorithm [1] and its efficacy in finding the global optimum point of a function. This paper proposes a novel approach for cancer risk prediction using a Hybrid Chaotic Pattern Search Algorithm (HCPSA). HCPSA combines the strengths of chaotic maps for global search and pattern search for local refinement. This hybrid approach effectively navigates the complex, high-dimensional space of cancer risk factors to identify optimal parameter configurations that maximise prediction accuracy. The algorithm comprises two stages: The first stage utilizes chaotic maps (Cubic, ICMIC, Neuron, or Sine) to rapidly explore the solution space and converge towards a near-optimal region. The second stage employs the pattern search algorithm to fine-tune the search and pinpoint the global optimum, representing the parameter configuration that yields the most accurate cancer risk prediction. The performance of HCPSA is evaluated using a comprehensive dataset of patient medical records, encompassing various risk factors associated with cancer development. The algorithm is trained to predict the likelihood of cancer occurrence based on individual patient profiles. Comparative analysis against established prediction models, such as logistic regression and Particle Swarm Optimization (PSO), demonstrates the superior accuracy and computational efficiency of HCPSA.**

***Index Terms*—optimization, hcpsa, cancer risk reduction**

I. Introduction

Optimization is a field with applications spanning all disciplines. Numerical techniques that leverage high speed computing effectively are woefully insufficient at finding global solutions to these problems. Classical techniques are effective at finding local optimum points [2]. This paper implements HCPSA (Hybrid Chaotic Pattern Search Algorithm) [1] to demonstrate its efficacy in obtaining the global solution and usage in the medical field in preventive care as well as creating awareness on the significant indicators of health risk due to life-threatening diseases such as cancer.

This paper goes into detail on the implementation details of the algorithm and then applying it to a dataset of cancer health factors.

Chaotic Optimization Algorithms [3] are very effective at searching for solutions without getting trapped in local optima and significantly reduce computational complexity [4]. They employ the use of functions known as chaotic maps, which generate chaotic sequences or sequences of numbers. Due to the chaotic nature of these sequences, they approach the neighborhood of the global minima very quickly. It takes a significantly larger number of computations to move towards the global optima while in the neighborhood. HCPSA overcomes this limitation by switching to a Hooke and Jeeves pattern search to quickly shoot to the nearest local optima (which inadvertently is the global optima).

Optimization problems are abound in the field of healthcare [5] [6]

We will be examining the health factors that leads to a risk of cancer for a given patient and the significance of these factors. We will use this algorithm to minimize cancer risk and examine the inputs that provide the least risk.

II. Proposed Methodology

A. Chaotic Maps

Chaotic Maps are functions that generate chaotic sequences. These sequences are used to generate pseudorandom numbers, which are mapped to the bounds of the function's output. These maps have points in their input from which no new numbers can be generated in the sequence. These points are known as fixed points. We will go over the fixed points of each map. When the input variable is equal to one of these fixed points, the remedy is to add $(0.01)r$ to the point obtained where $r \in (0, 1)$.

A metric used to evaluate the chaotic nature of a system is the Lyapunov Exponent. It's a measure of the average rate of exponential convergence or divergence of nearby paths. A high Lyapunov exponent indicates that two nearby points as initial conditions for points will yield vastly different chaotic sequences for a given map.

1) Cubic Map:

$$x_{n+1} = \beta x_n(1 - x_n^2), x_n \in (0, 1)$$

The fixed point of this map is 0.7835 and it was found in [1] that the best value for β is 2.59 for generating chaotic sequences in $(0, 1)$. It was also noted in [1] that this map is

effective at minimizing functions with exponential terms when used in the algorithm.

2) ICMIC Map (Iterative Chaotic Map With Infinite Collapses):

$$x_{n+1} = \sin\left(\frac{a}{x_n}\right), a \in (0, \infty), x_n \in (-1, 1)$$

It was found in [1] that the best value for a is 2.5,3 for generating chaotic sequences. This map, like the cubic map, is effective at minimizing functions with exponential terms. It has the following fixed points: ±0.0952, ±0.1065, ±0.1188, ±0.1373, ±0.1578, ±0.1934, ±0.2343, ±0.3301, ±0.4448 and there are an infinite number of fixed points in (-0.1, 0.1)

3) Neuron Map:

$$x_{n+1} = \alpha - 2\tanh(\beta)e^{-3x_n^2}$$

It was found in [1] that the best values for alpha and beta are 0.9, 5 respectively. The fixed point for this map is -0.3842.

4) Sine Map: [7]

$$x_{n+1} = \frac{\alpha}{4}\sin(\pi x_n), x_n \in (0, 1)$$

It was found in [1] that the best value for alpha is 4. The fixed point for this map is 0.7365.

B. Hooke-Jeeves Pattern Search

The Hooke-Jeeves pattern search algorithm is a very effective algorithm for finding the nearest local optima of a given initial point. It does this with the following steps:

C. Hooke-Jeeves Optimization Method

The Hooke-Jeeves method is a derivative-free optimization method.

1) *Exploratory Move:* From the current base point x_k, the algorithm explores along each coordinate direction to find a better point:

$$x_k^{\text{new}} = x_k \pm \Delta x \cdot e_i.$$

If $f(x_k^{\text{new}}) < f(x_k)$, update x_k.
2) *Pattern Move:* If exploratory moves succeed, extrapolate the trend:

$$x_k^{\text{pattern}} = x_k + (x_k - x_{k-1}).$$

Accept x_k^{pattern} if it improves $f(x)$; otherwise, retain x_k.
3) *step size reduction:* if no improvement occurs, reduce the step size:

$$\Delta x = \Delta x/2.$$

The process terminates when Δx is below a threshold or a maximum iteration limit is reached.

D. Hybrid Chaotic Pattern Search Algorithm

We can now move to elaborating the main algorithm in this article. This search algorithm is, as mentioned earlier, used for finding the global optima of a function. It combines the use of chaotic maps and the Hooke-Jeeves optimization method. It consists of two stages:

1) Stage 1: Generate chaotic sequences using a chaotic map and shifting them to the bounds specified. A variable c^k is moved to within the bounds specified using the following transformations, where U and L are the upper and lower bounds:
   - For Logistic and Cubic:
   $$L + (U - L)c^k, i = 1, 2, \ldots, n$$
   - For ICMIC:
   $$x_i^k = \left(\frac{U+L}{2}\right) + \left(\frac{U-L}{2}\right)c^k, i = 1, 2, \ldots, n$$

   Compute $f(x^k) < f_{\text{min}}$, if true, then $f_{min} = f(x^{(k)})$. Perform this for N iterations.
2) Stage 2: Using the output of stage 1 as an initial point, use the pattern search algorithm to find the global optimum.

E. Logistic Regression

We use logistic regression to create a function that can model the relations between the input (health indicators) and the output (risk of cancer).

Logistic regression is often used in healthcare as it's good at classifying [8] [9] [7] [10]. is a statistical method used for binary classification tasks, where the output variable y takes values in $\{0, 1\}$. It models the relationship between the input features $\mathbf{x}$ and the probability $P(y = 1|\mathbf{x})$ using the logistic function. **[?]**

1) Model: The probability of the positive class is modeled as:
$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b),$$
where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, $\mathbf{w}$ is the weight vector, and b is the bias term.
2) Loss Function: Logistic regression minimizes the log-loss (cross-entropy):
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right],$$
where $\hat{y}_i = P(y = 1|\mathbf{x}_i)$ is the predicted probability for sample i.
3) Optimization: Parameters $\mathbf{w}$ and b are optimized using gradient descent or related methods by computing:
$$\nabla\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\mathbf{w}}, \quad \frac{\partial\mathcal{L}}{\partial b}.$$
4) Output: Predictions are made by thresholding $\hat{y}$:
$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1|\mathbf{x}) \geq 0.5, \\ 0 & \text{if } P(y = 1|x) \leq 0.5 \end{cases}$$

We then use this on the cancer risk dataset

III. Dataset Description

The lung cancer dataset [11] under consideration contains comprehensive data on patients diagnosed with lung cancer, capturing a wide range of features related to demographic, environmental, lifestyle, and medical factors. The dataset includes 20 distinct features that encompass critical variables such as patient age, gender, and various lifestyle choices. It also collects information on factors such as air pollution exposure, alcohol consumption, dust allergies, and the presence of occupational hazards, all of which are known to influence lung cancer risk. Additionally, genetic risk factors are recorded, providing insight into heritable traits that might predispose individuals to the disease. Medical history is also well-represented, with features detailing the presence of chronic lung diseases, obesity, and smoking habits, including both active smoking and passive exposure. These variables offer a rich foundation for analyzing how lifestyle and environmental conditions interact with genetic predisposition in the development of lung cancer.

IV. Proposed Methodology

We apply logistic regression on the features, with the target variable being risk.

We then find the weights and biases of the model. This acts as the function we wish to minimize.

We then use the HCPSA algorithm to find the minima of the function. This would give us the factors that shows the least risk.

V. Results And Analysis

The algorithm and Logistic Regression model was implemented in Python on a a device with an 11th Gen Intel i5 with 16GB Ram.

The parameters obtained for the dataset is: [-0.04102102 0.4995807 0.25686717 -0.27622863 0.53424217 0.11661132 -0.21263319 1.46836467 0.07008249 1.14672606 -0.62798254 0.85400559 1.0928179 -0.4679089 0.2479672 1.34471648 0.81788685 1.15150412 0.52944402 0.63933265 1.07320183, and the bias is -33.74.

We evaluated the model from the range [0,7] since that's the range our dataset ranged in.

VI. Conclusion

This paper discusses Chaotic Maps and the Hooke-Jeeves Pattern Search method and the usage of the Hybrid Chaotic Pattern Search Algorithm to minimize the risk of lung cancer.

VII. Acknowledgements

We would thank Prof. Sarada Jayan for providing the opportunity to research and understand this topic and its application in the field of Artificial Intelligence.

References

[1] G. S. Rani, S. Jayan, and B. Alatas, "Analysis of Chaotic Maps for Global Optimization and a Hybrid Chaotic Pattern Search Algorithm for Optimizing the Reliability of a Bank," *IEEE Access*, vol. 11, pp. 24 497–24 510, 2023.

[2] "On the Local Convergence of Pattern Search | SIAM Journal on Optimization," https://epubs.siam.org/doi/abs/10.1137/S1052623400374495.

[3] D. Yang, Z. Liu, and J. Zhou, "Chaos optimization algorithms based on chaotic maps with different probability distribution and search speed for global optimization," *Communications in Nonlinear Science and Numerical Simulation*, vol. 19, no. 4, pp. 1229–1246, Apr. 2014.

[4] D. Yang, G. Li, and G. Cheng, "On the efficiency of chaos optimization algorithms for global optimization," *Chaos, Solitons & Fractals*, vol. 34, no. 4, pp. 1366–1375, Nov. 2007.

[5] "International Journal of Imaging Systems and Technology | IMA | Wiley Online Library," https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22400?casa_token=8TLyDqlwr9YAqo.

[6] "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction | IEEE Journals & Magazine | IEEE Xplore," https://ieeexplore.ieee.org/abstract/document/8371664.

[7] F. B. Demir, T. Tuncer, and A. F. Kocamaz, "A chaotic optimization method based on logistic-sine map for numerical function optimization," *Neural Computing and Applications*, vol. 32, no. 17, pp. 14 227–14 239, Sep. 2020.

[8] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," in *2018 International Conference on Robots & Intelligent System (ICRIS)*, May 2018, pp. 157–160.

[9] T. Ayer, J. Chhatwal, O. Alagoz, C. E. Kahn, R. W. Woods, and E. S. Burnside, "Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation," *RadioGraphics*, vol. 30, no. 1, pp. 13–22, Jan. 2010.

[10] X. Zhou, K.-Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 249–259, Aug. 2004.

[11] "Lung Cancer Prediction," https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link.