**A picture containing text

Description automatically generated**

**Optimization of Logistic Regression Algorithm with Genetic Algorithm for Breast Cancer Classification**

Marzieh Khalili Chachaki

Asia Pacific University of Technology & Innovation (APU), Technology Park Malaysia, Kuala Lumpur, Malaysia

[TP067265@mail.apu.edu.my](mailto:TP067265@mail.apu.edu.my)

[marzieh.khaliliy@gmail.com](mailto:marzieh.khaliliy@gmail.com)

**Abstract**

**An increasing number of women are diagnosed annually with breast cancer. If a problem is detected early, there are more treatment choices and a higher likelihood of a cure. This study uses feature selection to enhance machine learning. In this work, genetic algorithms (GA) are employed to enhance the logistic regression approach. The primary purpose of this method is to improve the dataset's ability to detect and classify breast cancer. Various classifiers were used to assess the performance of subgroups with certain characteristics. In these experiments, the logistic regression method performed superiorly and improved the accuracy with genetic algorithm optimizer.**

***Key words\_\_ Breast cancer, Logistic Regression, Machine Learning, Genetic Algorithm***

1.Introduction

Cancer has been the top cause of mortality and the most important global public health issue for decades. Breast cancer has been the most common kind of cancer in women for many years. – Who In 2018, invasive breast cancer was responsible for the deaths of an estimated 627 000 women, or around 15% of all female cancer-related fatalities. Moreover, according to WHO projections, the incidence of breast cancer is growing in practically every country around the world. When breast tissue contains a malignant tumour, it is evident that early detection and diagnosis are essential for optimum treatment planning and patient care(Kassani *et al.*, 2020).

Machine learning is a study to design computer programs that can access the data. It is a subset of Artificial Intelligence in the field of computer science. With the help of machine learning, it may be able to independently learn and adapt to new conditions(Sethi, 2018). Several data mining and machine learning approaches have been developed in recent decades for the diagnosis and classification of breast cancer. These techniques may be broken down into three primary stages: pre-processing, feature extraction, and classification. In recent years, interest in the application of linear algebra and statistical methods to data mining, machine learning, bioinformatics, and other fields has increased. According to many, logistic regression, the most prominent approach for disease prediction utilising gene expression data, gives great results for cancer classification. Among these methods, the logistic regression method is particularly relevant(Bazzoli and Lambert-Lacroix, 2018).

Genetic algorithms are a metaheuristic optimization technique based on the Darwinian idea of evolution. By executing crossover, mutation, and selection processes, the approach advances a population of emerging candidate solutions. Using well-designed operators and optimal parameters, GAs have shown a high level of resiliency in identifying optimal solutions to difficult optimization problems. It does this by selecting a suitable group member from a wide selection of applicants. In contrast, fit members may readily endure until a better set of solutions is found. In recent years, interest in population-based optimization strategies has increased. For bilevel optimization, several population-based optimization algorithms have been proposed(Jennings *et al.*, 2019).

The remaining research is structured as follows. Section 2 describes the Logistic Regression Algorithm and the Genetic Algorithm, whereas Section 3 describes how these algorithms are implemented. Data and analysis from the experiment are given in Section 4, while conclusions and future work are discussed in Section 5.

2.Methodology

In this section, we justify our proposed method of action. The Logistic Regression and Genetic Algorithm (GA) methods applied in this work are presented in detail.

(I)Logistic Regression Model

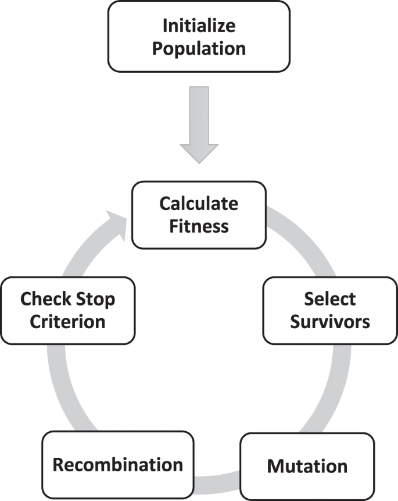
Classification is the most popular use of logistic regression in machine learning. When the outcome is binary, logistic regression is the optimal choice. The 'logit' (log odds) function is what the term 'logistic' refers to when applied to binary outcomes (Khairunnahar et al., 2019). A logistic regression classification model employs a random set of inputs to categorise data. Using the output 0 or 1, it is possible to categorise data in two categories, for example (Zou et al., 2019). Any patient in the dataset may potentially have a continuous result ranging from a negative infinity to a positive infinity; the model predicts the continuous outcome for this patient by putting the covariate values into the linear regression equation that estimates the continuous outcome (Zabor et al., 2022). Following is the algebraic form of the logistic regression equation:

Text

Description automatically generated

(II)Genetic Algorithm Model

Darwin's evolutionary theory serves as the structure's foundation in GA. As the process unfolds, the solution evolves over time. The GA chromosomes may have the solution to this issue. At each generational stage, the genetic algorithm randomly selects individuals of the current population to act as parents for the children of the following generation. Using Crossover and Mutation, the next generation of GA is produced. Crossbreeding is the process of combining the genes of two parents in order to generate offspring, while mutation is the random alteration of the genes of two parents in order to produce offspring. Over time, individuals "evolved" toward a more acceptable answer. The GA is a bio-inspired optimization method that parallels the evolutionary process of natural selection. The assessment, selection, crossover, and mutation processes of this methodology are well-suited for offering accurate answers to global optimization and search issues (Davoudi and Thulasiraman,2021).



In the GA, binary, real-coded, and integer formats are just a handful of the gene representation alternatives. The fitness function, an evaluative operator, is used to assess people. We utilise the fitness function to assess a person's ability to come up with the optimal solution. During this phase, each participant's fitness level is assessed and graded. Future generations will be more likely to produce offspring from individuals with greater fitness levels. Depending on the problem of application and search, the fitness qualities vary. There are a number of common fitness procedures for classification difficulties, such as the classification error.

1. Selection (encoding of a chromosome)

Once the fitness values have been determined, the stage of selection begins. Before reproduction, the selection stage determines which individuals should be coupled together. If a person has a high degree of physical fitness, their genes may be handed on to future generations. It is often considered as one of the GA's most crucial components. This genetic process creates two new chromosomes from the DNA of both parents for the kid. It is possible for chromosomes to hold information about the solution they represent due to their structure. A binary string is the most typical encoding format. This will be the last appearance of the chromosome.Text

Description automatically generated with low confidence

1. Crossover

As soon as we are assured of the code to be used, we may commence crossover operations. In crossover, a portion of the genes from the parents' chromosomes are utilised to produce an offspring. Choosing a random intersection point between the first parent point and this point is the simplest way to do this (Lambora, Gupta and Chopra, 2019). The illustration below illustrates an example of a crossing point:Graphical user interface, text, application

Description automatically generated

There are a variety of crossing techniques, as well as a selection of crossing locations. It may be more complicated and subtle to cross over. The most significant element is the chromosomal coding. To increase the performance of the genetic algorithm, choose the appropriate crossover for the specific issues.

Chart, box and whisker chart

Description automatically generated

1. Mutation

This results in the birth of a new generation of offspring. In order to increase the genetic diversity of the population, newborns' genes are modified. Individuals and solutions may benefit from a population with more diversity. Binary data is often represented in the GA by flipping bits, inverting them, or experiencing random mutations. As soon as the crossover is complete, the mutation stage begins. Mutation is a deliberate attempt to prevent the whole population from reaching a local optimum. Mutation is a random process that influences the progeny of cross-breeding. We may choose bits at random in binary encoding and flip them from 0 to 1 or 1 to 0. The following are some potential additions to a mutation:

Text

Description automatically generated

A picture containing icon

Description automatically generated

Encoding of chromosomes is entirely reliant on the mutation process (as well as crossover).

Crossover and Mutation Probability:

Crossover probability and mutation probability are the two most significant factors in genetic algorithms.

Crossover probability:

This may be the result of the crossover frequency's wide frequency range. If children are identical to their parents, then there is no crossover. The offspring is the outcome of two components of the parent's crossover. There is a 100% likelihood that offspring from a hybridization will exist. In contrast, if all future generations are generated from carbon copies of chromosomes from a past population, the percentage will be zero percent (although this does not necessarily imply that all future generations would be identical). Crossover is undertaken to improve the possibility of new chromosomes possessing favourable features from the old chromosomes, resulting in a better new crossover than the old).

Mutation probability:

It may be seen as the likelihood that chromosomal segments will undergo mutation. As long as there is no mutation, the offspring will be identical to their parents prior to crossing over.

As the number of chromosomes varies, the likelihood of mutation rises. At 100 percent, there is no chromosomal change; but, at 0 percent, there is always one. The mutation process prevents local extremes from occurring in GA. Mutation should occur at a modest frequency to prevent GA from devolving into a random search.

The initial step is to concentrate on a subset of the original population before analysing the fitness factor of each chromosome using a fitness function. After the first step, parents for the next generation will be selected from chromosomes with fitness values near to the threshold. From this point forward, there are only two options: recheck the star function criteria or continue searching until we discover a solution that fulfils our needs. If it satisfies these criteria, we should quit; otherwise, we should continue searching until we discover one that matches our requirements.

3.Implementation

Dataset description:

The Python computer language's sklearn package gives access to a breast cancer dataset. From a digitised picture of a breast mass fine needle aspirate, features are created (FNA). Using this method, it is possible to identify the nuclei of imaged cells. Eleven real-valued properties are present in every cell's nucleus. We determined 30 characteristics for each photograph by calculating the 'worst' or 'worst' feature (the mean of the three biggest values). Examples of radius measures include the "mean radius," "radius standard error," and "worst radius." To achieve four significant figures of accuracy, the value of each characteristic is recalculated. Two goal classes indicate both positive and negative results (Malignant).

Table

Description automatically generated

Python's code is easy and simple to implement, but the compilation procedure is time-consuming. Despite its small size, this research makes full use of the available information. Python is used to implement every algorithm described in this article.

Text, letter

Description automatically generated

Table

Description automatically generated

The description of feature statistics is shown below.

Table

Description automatically generated

Data pre-processing techniques

It is the process of converting unstructured data into a format that machine learning algorithms can use to gain new insights or predict future events. The data must be cleansed in order to accelerate data processing and analysis and improve the accuracy of the prediction. I will use train test split to divide our data into training and testing sets this time. This will be followed by the loading and testing of any Machine Learning models I've developed using data.

Text

Description automatically generated with medium confidence

Correlation between variables

Using the plot correlation function, the correlation between two discrete variables may be seen. A correlation heatmap depicts the two-dimensional correlation matrix between two distinct dimensions, whilst coloured pixels reflect data from a single monochromatic scale. The values of the first dimension are represented as rows, while the values of the second dimension are represented as columns. Consider the cell colour to determine how many measurements correspond to each dimension value. Correlation heatmaps are advantageous for data analysis because they facilitate the identification of patterns and anomalies within the same data. Using colour bars, the data in a correlation heatmap may be more readily seen and understood, similar to a regular heatmap.

Chart, treemap chart

Description automatically generated

Utilize these correlation values as a reference when searching for correlations between variables that are directly related.

Using ml algorithms, this study categorised malignant and benign tumour cells. In this study, parametric approaches are used to evaluate logistic regression machine learning algorithm. Due to the fact that there are only two groups in this investigation, logistic regression was determined to be appropriate.

4.Experiments and results

The experiment, which needed 64GB of RAM and an Intel Core i7 8th processor, was conducted using a Python 3.9 Jupyter notebook. Among the scikit-learn library functions used in the experiment were Numpy, Pandas, matplotlib, Random and etc.

Based on following equation, false positive rate (FP) and false negative rate (FN) are acronyms for the TP and TN values, respectively, of the confusion matrix.

Text, letter

Description automatically generated

Statistical criteria like as accuracy, precision, and recall were used to examine the relative benefits of this model are shown below.

Text

Description automatically generated

The accuracy is evaluated using the confusion matrix and ROC/AUC (Area Under Curve). Whether a result is True Positive (TP) or False Positive (FP) is a crucial factor to consider (FP). On the Y-axis of the ROC curve, true positive (TP) and false positive (FP) values are shown, respectively. The area under the ROC curve may describe the whole correctness of the tests. The bigger the test area, the higher the exam outcomes. The number of instances that were correctly predicted and those that were not had to be computed in order to determine predictive accuracy. Consequently, the calculation is stored in the database as a confusion matrix (Derisma, Silvana and Imelda, 2018).

As shown in the following graph, the blue line represents the ROC curve with an AUC (Area Under Curve) value of 0.987. ROC curve diagnostic value criteria for the AUC are as follows:

1) 0.90 - 1.00 of accuracy = excellent classification

2) 0.80 - 0.90 of accuracy = good classification

3) 0.70 - 0.80 of accuracy = fair classification

4) 0.60 - 0.70 = poor classification 5) 0.50 - 0.60 = failure

Chart, line chart

Description automatically generated

Certain Hyperparameters will be assigned to the population of N Machine Learning models. Now as each model is evaluated, we may establish its accuracy and then keep just fifty percent of them (the ones that perform best). We can now build children with similar Hyperparameters to the top models, allowing us to construct a new population of N models. Once each model's correctness has been recalculated, the procedure may be repeated indefinitely. Consequently, at the completion of the competition, only the top models will remain.



5.Conclusion

In this study, genetic algorithm-based optimization of the Logistic Regression Classifier is shown. There is a brief review of the genetic algorithm and sufficient information on its intended use. Comparatively to all other malignancies, breast cancer is the greatest cause of death among women. Therefore, early detection of breast cancer is vital for reducing mortality rates. In this study, we effectively predicted breast cancer using data processing, feature selection, feature extraction, data discretization, and classification. Our study indicates that the Logistic Regression Classifier is the most precise data classification approach. In the near future, it may be possible to identify a particular stage of breast cancer, which might considerably improve this work.

References

Bazzoli, C. and Lambert-Lacroix, S. (2018) ‘Classification based on extensions of LS-PLS using logistic regression: Application to clinical and multiple genomic data’, *BMC Bioinformatics*, 19(1), pp. 1–13. doi: 10.1186/s12859-018-2311-2.

Davoudi, K. and Thulasiraman, P. (2021) ‘Evolving convolutional neural network parameters through the genetic algorithm for the breast cancer classification problem’, *Simulation*, 97(8), pp. 511–527. doi: 10.1177/0037549721996031.

Derisma, Silvana, M. and Imelda (2018) ‘Optimization of Neural Network with Genetic Algorithm for Breast Cancer Classification’, *2018 International Conference on Information Technology Systems and Innovation, ICITSI 2018 - Proceedings*, pp. 398–403. doi: 10.1109/ICITSI.2018.8696014.

Jennings, P. C. *et al.* (2019) ‘Genetic algorithms for computational materials discovery accelerated by machine learning’, *npj Computational Materials*, 5(1). doi: 10.1038/s41524-019-0181-4.

Kassani, S. H. *et al.* (2020) ‘Classification of histopathological biopsy images using ensemble of deep learning networks’, *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pp. 92–99.

Khairunnahar, L. *et al.* (2019) ‘Classification of malignant and benign tissue with logistic regression’, *Informatics in Medicine Unlocked*, 16(August 2018), p. 100189. doi: 10.1016/j.imu.2019.100189.

Lambora, A., Gupta, K. and Chopra, K. (2019) ‘Genetic Algorithm- A Literature Review’, *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, (1998), pp. 380–384. doi: 10.1109/COMITCon.2019.8862255.

Sethi, A. (2018) ‘Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer’, *2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018*, pp. 252–255. doi: 10.1109/ICRITO.2018.8748502.

Zabor, E. C. *et al.* (2022) ‘Logistic Regression in Clinical Studies’, *International Journal of Radiation Oncology Biology Physics*, 112(2), pp. 271–277. doi: 10.1016/j.ijrobp.2021.08.007.

Zou, X. *et al.* (2019) ‘Logistic Regression Model Optimization and Case Analysis’, *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*, pp. 135–139. doi: 10.1109/ICCSNT47585.2019.8962457.