International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **05** Issue **02** | **2021**

Optimization of lung cancer classification using Machine learning and AWS

Aishwarya R Gowri,
MCA scholar, School of CS &IT, Dept. of MCA, Jain (Deemed-to-be) University, Bengaluru

ABSTRACT

Lung cancer is considered as one of the deadliest disease and it has higher mortality and takes up 7th position in the mortality rate index and causing 1.5% total mortality rate in the world. On February 4th 2020, World Health Organization has released two global reports on occasion of World Cancer Day. The report uncovers that one of every 10 Indians will develop cancer disease during their lifetime and one out of 15 will die from the infection. There are an expected 1.16 million new disease cases enlisted every year in India and around 7,84,800 die from it every year. As indicated by the report, of 5.70 lakh new cancer growth cases in men, the most common is oral disease, lung cancer, stomach cancer, colorectal cancer and esophageal cancer represent 45% of enrolled cases. In this paper an approach is made to effectively classify the lung cancer related attributes using machine learning algorithm. 15 lung cancer related attributes are considered and the classification is based on logistic regression without and shallow neural network. The machine learning code is executed in AWS Sage Maker

Keywords: cancer, lung, AWS, Machine learning, logistic regression, neural network

1. INTRODUCTION

Lung cancer is one of the main causes for cancer-related death. Lung cancer can begin in the windpipe, lungs or main airway. [7]It is caused due to unidentified growth and spread of some cells from the lungs. People with lung disease such as emphysema and dealing with chest problems have higher chance to be diagnosed with lung cancer. Overconsumption of tobacco, beedis and cigarettes, are the major risk factor which leads to lung cancer in Indian men; however, among Indian women, smoking is not so common, which also indicate that there are other factors leading to the cause of lung cancer. Other risk factors which include, air-pollutions, exposure to radon gas and chemicals in the workplace. cancer is categorized into two types such as primary lung cancer and secondary lung cancer. A primary lung cancer which initiates in lung is called as primary lung cancer whereas those that initiates in lung and spread to other parts of body is secondary lung cancer. the stage of the cancer is determined by the size of the cancer and how far the infection has spread. An initial stage of cancer is a small cancer which is diagnosed in lung and the advanced lung cancer is the one that has spread into other part of body or surrounding tissue. in order to prevent the lung cancer A better understanding of risk factors can help. The key to improve the survival rate is early.

2. LITERATURE REVIEW

[1] This paper mainly deals with comparative study of the classification algorithm to detect a tumour in Brain. Using location and volumetric features and overall accuracy rate has been calculated based on two classification classes such as Quadratic Discriminant and logistic regressionand three classification classes such as Coarse Gaussian SVM, Linear SVM, Cosine KNN and Complex and median tree.

[2] In this paper, various results are produced for each classifier on the lung cancer dataset that was obtained. The classifiers such as NN, SVM, KNN and Logistic Regression were implemented and also their corresponding accuracy rates were obtained. the highest accuracy was produced by Support Vector Machine with 99.3%. This proposed method was applied to medical dataset which helped doctors in better decision making.

[3] This paper examines the accomplishment of support logistic regression (LR) and support vector machine (SVM) algorithmsby predicting the survival rate of lung cancer patients and compares also by identifies the effectiveness of these two algorithmsthrough accuracy, recall, F1, precision, score and confusion matrix. These techniques have been applied in order to detect the survival possibilitiesof lung cancerpatientsand help the physicians to take decisions based on the forecast of the disease.

[4] Different division calculations were discussedwhich incorporates Naïve Bayes, Hidden Markov Model and so on Appropriate clarification is given about how and why different division calculations are utilized in discovery of Lung tumour.

[5] The classification algorithms such as Naive Bayes, SVM, Decision tree and Logistic Regression were used to predict lung cancer The main objective of this paper is to diagnose lung cancer at the earlier stage by examining the performance of classification algorithms.

3. PROBLEM STATEMENT

Achieving a better accuracy rate was challenging. As the estimation of the efficiency of any algorithm completely depends on the dataset, if the dataset is lost then it hinders the whole process.

4. IMPLEMENTATION

In this model 15 lung cancer related attribute were considered for the classification purpose. the 15 attributes that were considered are age, gender, smoking, yellow finger, Anxiety, Peer pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Coughing Shortness of Breath, Swallowing Difficulty, Chest pain, Lung Cancer. The dataset is collected from Data World platform.

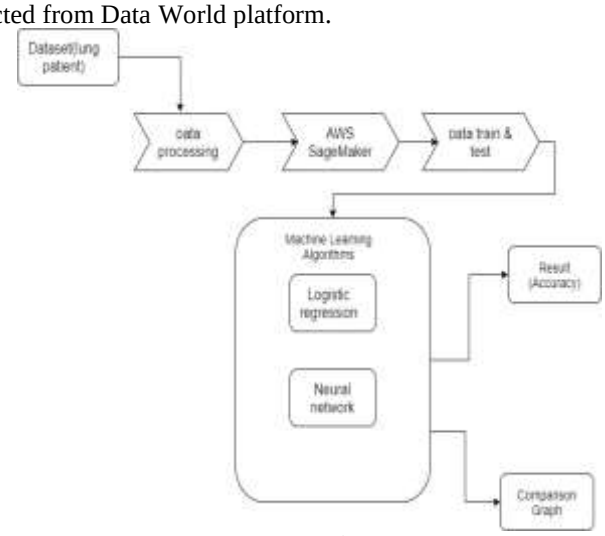

Fig1: Architectural flow diagram

The gathered data is processed and uploaded in AWS sage makerstudio. The dataset is divided into training dataset and testing dataset. Machinelearning algorithms such as logistic regression and Shallow neural network is applied on both datasets. If the attribute specified is true, then it is initialized to 2 and if the attributes specified if false then it is initialized to 1. The accuracy along with required graph was generated. And compare output of both these classifiers

4.1 Advantages of proposed model

- The evaluated result was generated within seconds
- Produced better accuracy
- with the help of AWS sage maker studio users can generate sharable link.
- Sage maker studio consists of pre-installed Amazon Sage Maker python SDK which reduces the complexity of importing certain python packages.

5. METHODOLOGY

5.1 Logistic regression

Logistic regression is a machine learning technique for the function that is used at the core of the method, the logistic function. The logistic function is also called as the sigmoid function and this was developed by statisticians in to describe the properties of population growth in rising quickly, ecology and maxing out the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but it will never be exactly at those limits. The logistic function equation is given as:
The lung cancer data set consists of 15 attributes; hence the hypothesis for lung cancer detection is of the form

$$h_\theta(x) = \frac{1}{(1+e^{-\theta^T x})} \text{---------- (1)}$$

Where $\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots\ldots + \theta_{15} x_{15}$ ..... (2)
Here x1, x2, x3, … x15 are 15 attributes and these 15 attributes will facilitate lung cancer. Total number weights in this case is 16 including θ_0.
The cost function for logistic regression is

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log\left(h_\theta(x^i) + (1 - y^i)(1 - \log\left(h_\theta(x^i)\right)\right) \text{----------(4)}$$

Where m is total number of input instances.
$\min_\theta J(\theta)$, where θ_j is computed as follows

$$\theta_j := \theta_j - \alpha \frac{\delta J}{\delta \theta_j} \qquad 0<=j<=7 \qquad \text{--------(5)}$$

Here partial derivatives of cost function parameterized by θ is estimated as follows

$$\frac{\delta J(\theta)}{\delta \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i) \quad \text{----------(6)}$$

$$\frac{\delta J(\theta)}{\delta \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i \quad 1<=j<=7 \quad \text{----------(7)}$$

The main objective is to minimize the cost function parameterized by θ, using the gradient descent rule ie.

5.2. Shallow neural network

In a shallow neural network, the values of the feature vector of a data that has to be classified (the input layer) will be passed to a layer of nodes (also known as neurons or units) (the hidden layer) each of these nodes generates a response according to some activation function, gg, acting on the weighted sum of those values, zz. The responses of each of these unit in a hidden layer is then passed to an outputlayer whose activation will produce the classification prediction output.

6. RESULT ANALYSIS

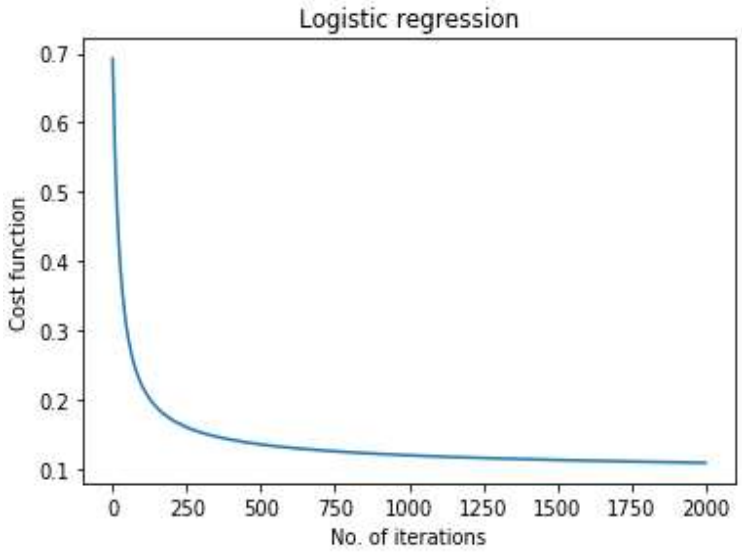

Fig 2: Logistic regression cost function graph

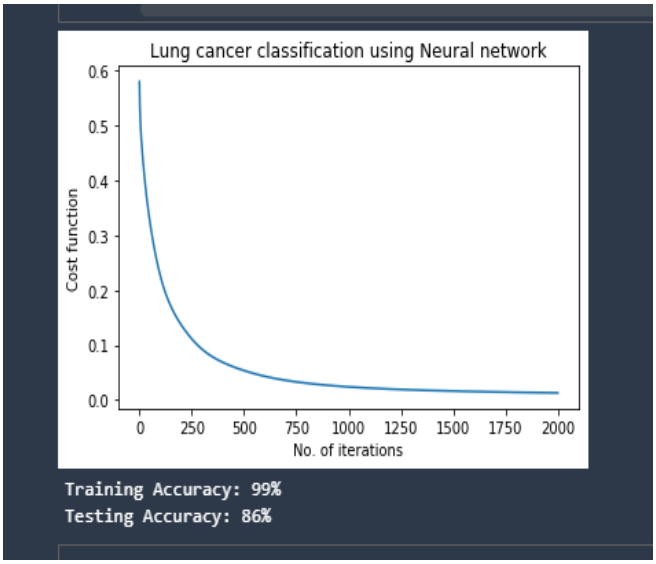

Fig 3: Neural network cost function graph

7. CONCLUSION

Various learning algorithms for effective classification of lung cancer were used in this paper. upon applying logistic regression and shallow neural network algorithms it was observed that shallow neural network produced better accuracy and performance with 86% accuracy rate when compared to logistic regression.

8. REFERENCES

[1] Vijay Suresh Gollamandala1 and Lavanya Kampa "An Additive Sparse Logistic Regularization Method for Cancer Classification in Microarray Data"
[2] Algamal Z. and Lee M., "Penalized Logistic Regression with the Adaptive LASSO for Gene Selection in High-Dimensional Cancer Classification," Expert Systems with Applications, vol. 42, no. 23, pp. 9326-9332, 2015.
[3] Hu Y. and Kasabov N., "Ontology-Based Framework for Personalized Diagnosis and Prognosis of Cancer Based on Gene Expression Data," in Proceedings of International Conference on Neural Information Processing, Kitakyushu, pp. 846-855, 2008.
[4] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1, September 2014
[5] Zehra Karhan1, Taner Tunç2, "Lung Cancer Detection and Classification with Classification Algorithms" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 22788727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77.
[6] Ada, RajneetKaur, " A Study of Detection of Lung Cancer Using Data Mining Classification Techniques " International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013
[7] Knight K. and Fu W., "Asymptotics for LASSOType Estimators," The Annals of Statistics, vol. 28, no. 5, pp. 1356-1378, 2000.
[8] Lin Y. and Zhang H., "Component Selection and Smoothing in Multivariate Nonparametric Regression," Annals of Statistics, vol. 34, no.5, pp. 2272-2297, 2006.
[9] Malioutov D., Cetin M., and Willsky A., "Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays," IEEE Transactions on Signal Processing, vol. 53, no. 8, pp. 3010-3022, 2005.
[10] Meier L., Geer S., and Bühlmann P., "The Group LASSO for Logistic Regression," Journal of the Royal Statistical Society Series B, vol. 70, pp. 53- 71, 2008.
[11] Zeng J., Lin S., Wang Y., and Xu Z., "L1/2 Regularization: Convergence of Iterative Half Thresholding Algorithm," IEEE Transactions on Signal Processing, vol. 62, no. 9, pp. 2317-2329, 2014.
[12] Zhu J. and Hastie H., "Classification of Gene Microarrays by Penalized Logistic Regression," Biostatistics, vol. 5, no. 3, pp. 427-443, 2002.
[13] Zou H. and Hastie T., "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 3, pp. 301- 320, 2005.
[14] Zou H., "The Adaptive Lasso and its Oracle Properties," Journal of the American Statistical Association, vol. 101, no. 476, pp. 1418-1429, 2006.
[15] Neha kumara and Khushi Verma, Bansal institute of science and technology, Volume 10, May-June 2019." A survey on various machine learning approaches used for breast cancer detection."
[16] Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava Department of Computer Science and Engineering, Indian Institute of Technology (BanarasHinduUniversity), Varanasi221005," Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features"
[17]. Alaá Rateb Mahmoud Al-shamash, Ph.D. Unaizah Hanum Binti Obaidellah, Ph.D. University of Malaya, Malaysia," Artificial Intelligence Techniques for Cancer Detection and Classification: Review Study"
[18] Taylan P. and Weber G., Data Science and Digital Business, Springer, 2019. [15] Tibshirani R., "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.
[19] Van-de-Geer S., "High-Dimensional Generalized Linear Models and the Lasso," Institute of Mathematical Statistics, vol. 36, no. 2, pp. 614- 645, 2008.
[20] Vincent M. and Hansen N., "Sparse Group Lasso and High Dimensional Multinomial Classification," Computational Statistics and Data Analysis, vol. 71, pp. 771-786, 2014.
[21] Ilias Maglogiannis·Elias Zafiropoulos· Ioannis AnagnostopoulosPublished online: 12 July 2007 © Springer Science+Business Media, LLC 2007," An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifier".
[22] Aik Choon tan and David Gilbert, Bioinformatics research center, Department of computing science, University of Glasgow, Glasgow, UK" Ensemble machine learning on gene expression data for cancer classification"
[23] Wang L., Chen G., and Li H., "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," Bioinformatics, vol. 23, no. 12, pp. 1486-1494, 2007.
[24] Wu S., Jiang H., Shen H., and Yang Z., "Gene Selection in Cancer Classification Using Sparse Logistic Regression with L1/2 Regularization," Applied Sciences, vol. 8, no. 9, pp. 1569, 2018.
[25] Xu Z., Chang X., Xu F., and Zhang H., "L1/2 Regularization: A Thresholding Representation Theory and A Fast Solver," IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 7, pp. 1013-27, 2012.