Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1649-1654
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5851

Early diagnosis of stroke disorder using homogenous logistic regression ensemble classifier

C.D. Anisha[a], K.G. Saranya[1,*]

[a] *Department of CSE, PSG College of Technology, Coimbatore, India*

(Communicated by Madjid Eshaghi Gordji)

Abstract

A stroke occurs in the scenario wherein the blood supply to the brain is blocked, leading to a lack of oxygen to the blood. There is a need for the early diagnosis of the stroke to handle the emergency situations of stroke in an efficient manner. Integration of Artificial Intelligence (AI) in the early diagnosis of stroke provides efficiency and flexibility. Artificial Intelligence (AI), which is a mimic of human intelligence has a wide range of applications from small scale systems to high-end enterprise systems. Artificial Intelligence has emerged as an efficient and accurate decision-making system in healthcare systems. Machine Learning (ML) is a subset of Artificial Intelligence (AI). The incorporation of machine learning techniques in stroke diagnosis systems provides faster and precise decisions. The proposed system aims to develop an early diagnosis of stroke disorder using a homogenous logistic regression ensemble classifier. Logistic regression is a linear algorithm that uses maximum likelihood methodology for predictions and a standard machine learning model for two-class problems. The prediction is improved by accumulating the predictions of two or more logistic regression using a bagging ensemble classifier thereby increasing the accuracy of the stroke diagnosis system. The accumulation of prediction of two or more same models is known as a homogenous ensemble classifier. The results obtained show that the proposed homogenous logistic regression ensemble model has higher accuracy than single logistic regression.

Keywords: Index Terms—Stroke, machine learning, logistic regression, Homogenous logistic regression Ensemble classifier.

[*] Corresponding author

Email addresses: ani.c.dass@gmail.com (C.D. Anisha), kgs.cse@psgtech.ac.in (K.G. Saranya)

Received: July 2021 *Accepted:* November 2021

1. Introduction

Stroke is a disorder which largely affects adults and elderly people thereby resulting in their social and economic issues. The two main types of strokes are blocked artery and leaking o blood vessels. The main reasons for stroke are heart disorder, glucose level changes, increase Body Mass Index (BMI). Early diagnosis of stroke with integration of Artificial Intelligence (AI) can enhance decision making and increase accuracy.

Artificial Intelligence is prevalent in all fields to provide higher accuracy and efficient systems. Artificial Intelligence is the technology which mimics the human intelligence. Healthcare systems started integrating their systems with AI for faster and accurate decision making.

Machine Learning is the subfield of Artificial Intelligence (AI). Machine Learning consists of Unsupervised Learning, Supervised Learning and Reinforcement Learning. There are single classifiers and ensemble classifiers. The ensemble classifiers are of two types, they are homogenous classifier and heterogenous classifier.

2. Related Work

Jaehak et al, [9] presents a stroke prediction system based on physiological signals or bio-signals namely surface Electromyography (EMG) signals.

Kunder Akash Mahesh et al [**?**] presents a stroke prediction system by incorporating machine learning techniques to the retrieved dataset.

Revanth S et al, [7] provides stroke prediction using Machine Learning (ML) algorithms. The algorithms namely Support Vector Machine (SVM), Multi- Layer Perceptron (MLP), Decision Tree and Random Forest are used as models for stroke prediction.

JoonNyung Heo et al [3] presents machine learning technique namely deep neural network, random forest and logistic regression model. The deep neural network helps in the long-term prediction of ischemic stroke patients.

Maihul Rajora et al [6] provides a machine learning based approach for stroke prediction in distributed environment. The machine learning technique used are Naïve Bayes, Logistic Regression, decision tree, Random Forest and Gradient Boosting.

M. S. Singh et al [8] presents a stroke prediction system based on Artificial Intelligence (AI) wherein decision tree is used for feature selection, Principal Component Analysis (PCA) for dimensionality reduction and neural network for classification.

Kuo-Liong Chien et al [2] constructs the prediction for stroke prediction in Chinese population using Multi-variate Cox Model.

Leila Amini et al [1] presents the prediction and control stroke model using K-Nearest Neighbor classifier (KNN) and C4.5 decision tree.

Min SN et al [4] develops a stroke pre-diagnosis based on logistic regression model focused on modifiable risk factors.

Moons KGM et al [5] presents a multi-variable logistic regression model for prediction of stroke in Europe based on cerebrovascular and cardiovascular correlation factors and Electro Cardio Gram (ECG) signals.

3. Methodologies

Figure 1 presents the framework of the proposed system.

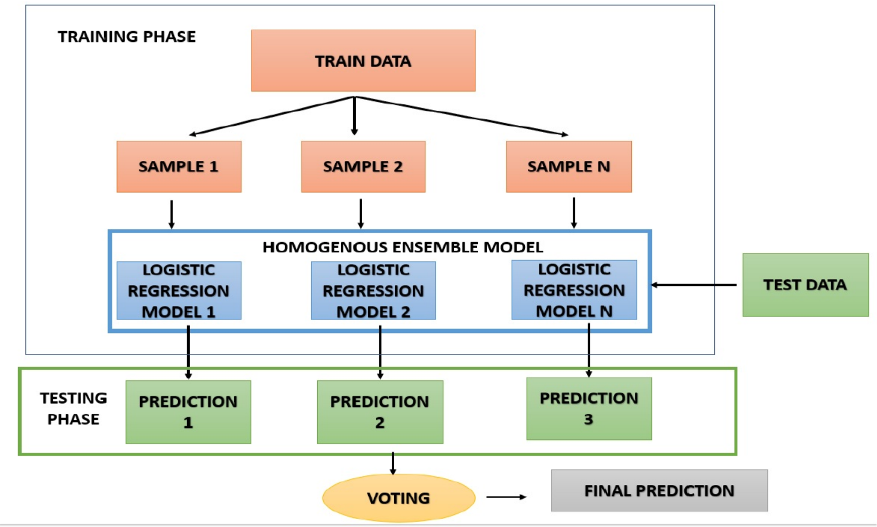

Figure 1: Framework of Homogenous Logistic Regression Ensemble Model for Stroke Diagnosis

3.1. Dataset Description

The dataset has been retrieved from Kaggle repository. The dataset consists of 5110 rows and 11 columns (attributes). The attributes present in the dataset are id, gender, age, hypertension, heart disease, marriage status, work type, residence type, average glucose level, Body Mass Index (BMI) and the class label indicating the presence and absence of stroke.

3.2. Label Encoding

The categorical attributes are encoded into integers for making the data into acceptable format for model training. The attributes gender and smoking-status is encoded.

- Gender Attribute: 0 for Male and 1 for Female
- Smoking-status:
  - 0 for unknown,
  - 1 for formerly smoked
  - 2 for never smoked
  - 3 for smokes

3.3. Splitting of Train and Test Set

The data is split into train and test set in the ratio 80:20 which means 80% of data is allocated as train set and 20% of data is allocated as test set.

3.4. Homogenous Logistic Regression Model Training

Logistic regression uses logistic function or sigmoid function as the core method. The logistic function is an S shaped curve graphically. The estimation of coefficients for logistic function is done using maximum likelihood technique.

The train data is sampled into various samples and fed into 1 to N logistic regression models present in the ensemble. The bagging classifier provides parameter for creation of homogenous logistic regression model.

- Initialization of Bagging Classifier Parameters
  - n_estimators is initialized to 200.
    This parameter indicates the number of base estimators to be created. The base estimator is logistic regression.
  - oob_score: This parameter is initialized to True which indicates that out of bag samples to be used for sampling data.
  - Random_state: This parameter is set to 100 to control the resampling procedure of data in training model.
- Initialization of Logistic Regression Parameters
  - Penalty: This parameter is set to L2.
    This parameter is used as regularization parameter.
  - Solver: This parameter is set to Liblinear.
    This is the solver used for fitting the model.
  - Class_weight: This parameter is set with the dictionary class 0: 0.10, class 1: 0.90.

3.4.1. Homogenous Logistic Regression Ensemble Classifier Testing

The Homogenous Logistic Regression Ensemble Classifier is tested with test data wherein each logistic regression model in the ensemble provides predictions and the prediction are aggregated using voting or averaging which pertains to the final prediction. The evaluation metrics used are “Confusion Matrix” and “Accuracy”.

4. Results and Discussions

The implementation and comparison of single logistic regression model and proposed homogenous logistic regression model is presented. The evaluation metrics used are confusion matrix and accuracy.

The confusion matrix consists of True Positive (correctly predicted positive samples), False Positive (wrongly predicted as positive), False Negative (wrongly predicted as negative), True Negative (correctly predicted negative samples).

Accuracy is evaluated using the formula: Addition of True Positive (TP) and True Negative (TN) divided by total number of samples.

4.1. Insights from the results

- Figure 2,3 – presents the confusion matrix of single logistic regression and proposed homogenous logistic regression ensemble model. From the confusion matrix, it is understood that True Positive prediction has been improved in proposed system.
- Figure 4- presents the accuracy graph. The accuracy of proposed system is higher (91%) than single logistic regression model (90%).

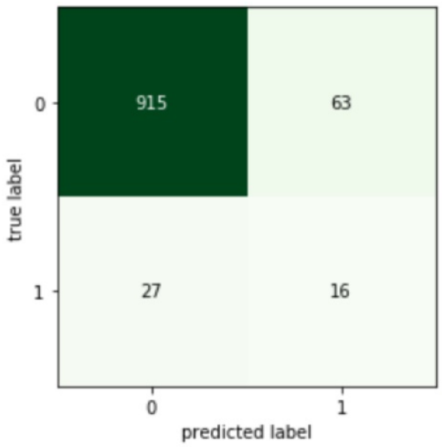

Figure 2: Confusion Matrix of Single Logistic Regression

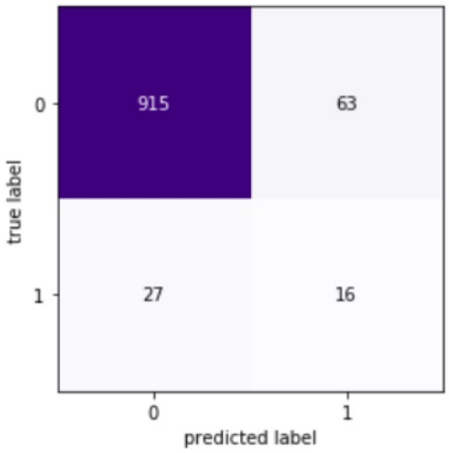

Figure 3: Confusion Matrix of Proposed Homogenous Logistic Regression Ensemble Classifier

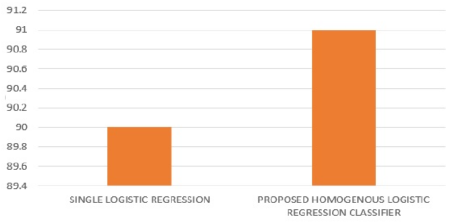

Figure 4: Accuracy Graph (Comparison)

5. Conclusion

Early diagnosis of stroke using machine learning technique is essential for quicker and precise decision making. The proposed framework – Homogenous Logistic Regression Ensemble classifier provides a robust model because it is ensembled with 1 to N logistic regression models and it forms aggregation of predictions of all logistic regression model in the ensemble. This proposed framework has the following advantages: reduced error in predictions, faster predictions and accurate predictions. The proposed framework provides an accuracy of 91% and is higher than single logistic regression model which provides and accuracy of 90%. The future work is to enhance the system to heterogenous ensemble classifier and deep learning techniques with various modalities.

References

[1] L. Amini, R. Azarpazhouh, M.T. Farzadfar, S.A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi and N. Toghianfar, *Prediction and control of stroke by data mining*, Int. J. Prev. Med. 4(Suppl. 2) (2013) 245–249.

[2] K.-L. Chien, T.-C. Su, H.-C. Hsu, W.-T. Chang, P.-C. Chen, F.-C. Sung, M.-F. Chen and Y.-T. Lee, *Constructing the prediction model for the risk of stroke in a Chinese population*, Stroke 41(9) (2010) 1858–1864.

[3] J.N. Heo, J.G. Yoon, H. Park, Y.D. Kim, H.S. Nam and J.H. Heo, *Machine learning–based model for prediction of outcomes in acute stroke*, Stroke 50(5) (2019) 1263–1265.

[4] S.N. Min, S.J. Park, D.J. Kim, M. Subramaniyam and K.S. Lee, *Development of an algorithm for stroke prediction: A national health insurance database study in Korea*, Eur. Neurol 79(3-4) (2018) 214–220.

[5] K.G.M. Moons, M.L. Bots, J.T. Salonen, P. Elwood, D. Freire, Y. Nikitin, J. Sivenius, D. Inzitari, V. Benetou, J. Tuomilehto, P. Koudstaal and D. Grobbee, *Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography?*, J. Epidemiol Community Health 56 (2002) 30–36.

[6] M. Rajora, M. Rathod and N.S. Naik, *Stroke prediction using machine learning in a distributed environment*, In: D. Goswami and T.A. Hoang (eds), Distributed Computing and Internet Technology, ICDCIT 2021, Lecture Notes in Computer Science, 12582 (2021).

[7] S. Revanth, S. Sanjay, N. Sanjay and V. Vijayaganth, *Stroke prediction using machine learning algorithms*, Int. J. Disaster Recovery and Business Contin. 11(1) (2020) 3081–308.

[8] M.S. Singh and P. Choudhary, *Stroke prediction using artificial intelligence*, 8th Annual Indust. Automa. Electromech. Engin. Conf.(2017) 158–161.

[9] J. Yu, S. Park, S.-H. Kwon, C.M.B. Ho, C.-S. Pyo and H. Lee, *AI-based stroke disease prediction system using real-time electromyography signals*, Appl. Sci. 10(19) (2020) 6791.