Rural Health Data Analysis of Epidemic Disease

Shridevi D Kadlaskar
Dept of Computer Science and Engineering
KLE Institute of Technology,
HUBBALLI (INDIA)

Yerriswamy T.
Dept of Computer Science and Engineering
KLE Institute of Technology,
HUBBALLI (INDIA)

***Abstract*:- The paper, considers classification of two epidemic diseases dengue and malaria. Accurate prediction could result in the appropriate hospital triage of dengue and malaria risk patients. The data is collected from government hospital located in a town which is surrounded by many villages. People often visit government hospital near to their villages once they fall sick. We conduct multivariable analysis to construct a predictive model for dengue and malaria using algorithms like Naive Bayes, Random Forest, Logistic Regression and ANN and the result from them are compared for the accuracy, it is observed that each algorithm produce different classification result, hence the ensemble of these result is carried out. The effectiveness of the methods is verified in Anaconda platform.**

Keywords:- Classification, Epidemic disease, Naive Bayes, Random Forest (RF), Neural networks (ANN), Logistic Regression (LR), Ensemble method.

1. INTRODUCTION

From historical days till present epidemic disease affect the people demographically, economically, financially making epidemic disease a global threat. Epidemic disease is the infectious disease which will spread to a wide population within small time span leading to public health at risk, if not treated at proper time. Examples of epidemic disease are malaria dengue, cholera etc. This study considers two epidemic diseases like malaria and dengue. Malaria is an epidemic disease which spreads by protozoan parasite 'plasmodium' as a vector of malaria transmission. Malaria is spread by mosquito anopheles lateen. Malaria incidence is increasing every year all over the world. India may become epidemic country, as it shares border with other already epidemic countries. Dengue is mosquito borne disease which is spread by Aedes Egyptae. In India dengue is a serious problem from fast few years, leading to which India has highest number of dengue cases of about 33 million apparent cases and 100 million asymptotic cases occurring annually leading to half of the population at risk. Both disease pose serious problems to the country as the lack of accurate medicines and proper treatment for the diseases [Pentapati et al, 2018]. Hence it is necessary for us to build a predictive model for both of these diseases that help in identifying disease at its initial stage. In this work we are building a machine learning model that will predict the disease at its initial stage. To achieve this goal we are using Naive Bayes, RF, ANN and LR algorithms which will work by dividing the collected data into training set and test set. Where in machine will learn from training data based on that data model will predict the disease when we apply test set which was not known by the model earlier. Already many works are carried out in this regard for predicting the dengue and malaria [Adimi et al. 2010], [Anwar et al. 2016], [Briet et al, 2008], [Pi Guo et al, 2017] where in they have

considered climatic conditions for building a predictive model. But this study relies on hospital data which includes radiology reports, patient demographics, past medical history and Laboratory reports. We are using CBC (Complete Blood Count) report, which will be different for normal persons to that of diseased persons. As of which we can predict dengue and malaria suspected patient once we use CBC reports for training the machines.

2. OBJECTIVES

The main objective of this work is to develop a model for analyzing hospital data and to obtain results by implementation of machine learning techniques. And to show that it is possible to create a model that predicts dengue and malaria at earlier stage by providing a great ally to government in combating this disease or any other diseases.

3. STUDY AREA AND DATASET

We have collected CBC reports of dengue and malaria patients from the year 2014 to 2017. Data is divided into training data and test data, with amount of training data is 80% and test data as much as 20%.

3.1 Naive Bayes Technique

Gaussian naive Bayes (GNB) classification is a supervised learning algorithm that uses Baye's theorem as a structure for classifying observations into one of a pre-defined set of classes based on information provided by predictor variables. GNB classifier estimate the conditional probabilities that an observation belongs to a particular class given the values of the predictor variables under the hypothesis that the predictor variables are class-conditionally independent, and thus (naively) do not take into consideration the covariance among the predictor variables. Thus, the posterior probability that an observation Y has class index k given the values of predictor variables X_1…. X_P is modeled according to Bayes theorem as in equation (1):

$$P\hat{}(Y = k/X1, \dots \dots, Xp) = \frac{\pi\left(Y = k \prod_{j=1}^{P} P\left(\frac{Xj}{Y} = k\right)\right)}{\sum_{K=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P\left(\frac{Xj}{Y} = k\right)} \quad (1)$$

Where (Y = k) is the prior probability that the class index is k. For each predictor $X_1, \dots, X_P$, the algorithm estimates a separate Gaussian distribution for each class, and observations are assigned to the class with the maximum posterior probability given the predictor values. We verified our training set by applying Gaussian NB Technique; the results are as in the Table 1**.**

Table 1: Bayesian Technique

Attributes name	**Measure**
Correctly Classified	84%
Incorrectly Classified	16%
Precision	0.84
Recall	0.89
F-measure	0.86
Support	19

3.2 Random Forest

Random forest (RF) is a supervised learning method. This means that each instance or sample is labeled with the outcome (RESULT). RF consists of an ensemble of k classifiers $h_1(x), h_2(x) \dots\dots h_k(x)$.With h(x) being the joint classifiers. Each classifier $h_1(x)$ consists of a decision tree, in which nodes are features as in Table 5. The selection of the feature is collocated in a node n,which is performed as follows.(1) A subset of attributes is randomly selected.(2) An evaluation measure is applied to the selected attributes according to their capability for providing homogeneity partitions of the samples, and (3) The attributes with the highest score is chosen[Beatriz López, 2017]. In particular, we use the change of the Gini impurity to compute the score as described in equation (2)

$$\mathrm{G(gi,n)} = -\sum_{\mathrm{Ck} \in \mathrm{C}} \mathrm{p}^2(\mathrm{Ck}) + \sum_{\mathrm{j=1}}^{\mathrm{NVAi}} \mathrm{p}\left(\mathrm{v_{i,j}}\right) \sum_{\mathrm{Ck} \in \mathrm{C}} \mathrm{p}^2\left(\frac{\mathrm{Ck}}{\mathrm{v_{i,j}}}\right) \quad (2)$$

Where $V_{i,j}$ is the j value of the i nodes. Probabilities are estimated according to the instances that reach the n nodes. Once a node is assigned to the attributes gi, the data is split in to sets values. The tree is grown with the new nodes in each branch. These are obtained by repeating the attribute selection process. The stopping condition is defined according to the number of instances that remain in the node. If this number is lower than a given threshold τ, the algorithm stops. Samples used to build each tree are also selected randomly with replacements. Once the RF is built, it can be tested for predicting dengue or malaria suspected patients, which will be the final prediction. The result with the highest prediction is assigned to q. We verified our training set by applying Random Forest technique; the results are as in the Table 2**.**

Table 2: Random Forest Technique.

Attributes name	**Measure**
Correctly Classified	84%
Incorrectly Classified	16%
Precision	0.85
Recall	0.95
F-measure	0.89
Support	19

3.3 Neural Networks

Artificial Neural Network (ANN) takes its motivation from human brain which has incredible processing ability because of having webs of interconnected neurons. ANNs are designed by using basic processing unit called Perceptron. Perceptron has only one layer and solves linearly separable problems. The problems which are not linearly separable can be solved by Multilayer Perceptron Neural Network (MLP). MLP has multiple layers, including input, hidden and output layers. The proposed epidemic disease prediction system was designed as a multilayer Perceptron neural network. The designed ANN has three layers: namely an input layer, a hidden layer and an output layer.

• Input Layer was designed to contain 13 neurons. Number of neurons was decided to be equal to the number of attributes in the data set.
• Hidden Layer was designed to contain 3 neurons. This number was decided as a startup point. The number was changed increasing one by one until it reached to the number of neurons of the input layer by comparing performance of them and then selecting the best one. This approach is based on one of machine learning best practices that the number of neurons of hidden layer should be the mean of the number of the neurons of input and output layers.
• Output Layer was designed to contain 3 neurons. The designed Neural Network is a classifier running in Machine Mode which means returning a class label (e.g., "Dengue Positive"/"Dengue Negative and Malaria"). Deciding 3 neurons is based on idea that the output layer has one node per class label in model. The ANN requires the learning rate, number of nodes in a single hidden layer, and maximum number of training epochs are specified[Ozkan K1l1ç, 2017].The percentages influence of input variable on the output value, Qik (%), indicating the importance of input variables were determined as in equation (3).

$$\mathrm{Qik}(\%) = \frac{\sum_{j=1}^{n}\left(\frac{|Wij|}{\sum_{i=1}^{m}|Wij|}|Vjk|\right)}{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\left(\frac{|Wij|}{\sum_{i=1}^{m}|Wij|}|Vjk|\right)\right)} * 100 \quad (3)$$

We verified our training set by applying Artificial Neural network (ANN) Technique; the results are obtained as in Table 3.

Table 3: ANN Technique.

Attributes name	Measure
Correctly Classified	73%
Incorrectly Classified	27%
Precision	0.45
Recall	0.64
F-measure	0.53
Support	28

3.4 Logistic Regression

Logistic regression is a nonlinear regression technique for prediction of dichotomous (binary) dependent variables in terms of the independent variables (covariates). The dependent variable can represent the status of the patient (e.g., Dengue Positive,Y=1:Dengue Negative,Y=0: and Malaria, Y = 2). The expected probability of a positive outcome P(Y = 1) for the dependent variable is modeled as in equation (4):

$$P(Y=1)\frac{1}{1+e^{-(B_o+\sum_{i=1}^{n}B_i x_i)}} \qquad (4)$$

Where xi, i = 1,…, n are the independent variables (covariates), B_i are the corresponding regression coefficients and B_0 is a constant, these

all will contribute to the probability. Eq. (3) reduces to a linear regression model for the

logarithm of odds ratio (OR) of positive outcome, i.e.

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=B_o+\sum_{i+1}^{n}B_i x_i \qquad (5)$$

In producing the LR model, the maximum-likelihood ratio is used to determine the statistical significance of the independent variables [Biswanath Samanta, 2009]. Verified our training set by applying Logistic Regression (LR) Technique; the results are obtained as in Table 4.

Table 4: Logistic Regression Technique.

Attributes name	Measure
Correctly Classified	72%
Incorrectly Classified	28%
Precision	0.75
Recall	0.84
F-measure	0.79
Support	19

4. COMPARISON OF ACCURACIES FOR NAIVE BAYES, RF, ANN AND LR

Results from Naïve Bayes, Random Forest ANN and LR are compared for their accuracy. Continues variables (examples AGE, HB, WBC, RBC, and HCT etc) were treated continuous, while discrete variables (example GENDER, RESULT) were treated categorically. We used performance measures like precision, recall, F1 score are considered from all four algorithms. We also tested for significant interactions among variables. At the output, numerical values were used with one category representing patients with dengue positive outcomes (1), other with dengue negative outcomes (0), malaria outcomes as (2). Table (5) shows the clinical features that we have considered for this work. It can be noticed that our proposed system are very effective and robust for predicting dengue and malaria using Anaconda 3.6 software.

Table 5: clinical features

CLINICAL FEATURES	DESCRIPTION
AGE	Age
GENDER	Male=1,Female=0
HB	Haemoglobin
WBC	White Blood Cells
NEUTROPHILS	Neutrophils
LYMPHOCYTES	Lymphocytes
MIXEDCELLS	Mixed cells
RBC	Red Blood Cells
PCV	Packed Cell Volume
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
PLATELETCOUNT	Platelet count
RESULT	0=Dengue Negative,1=Dengue Positive,2=Malaria

5. CONCLUSION

This work presents an automatic system for both prediction and identification of dengue and malaria. The system is based on machine learning technique wherein our data is divided into training set and testing set. Machine will learn from training dataset and will predict disease when applied with testing set. The model can accurately categorize diseased person from non diseased person. And an average accuracy of above 70 % is obtained from Naïve Bayes, RF, ANN and LR. Results from them are compared for their accuracies and which is represented graphically as in fig1. We represented the accuracies of Naive Bayes, Random Forest, ANN and Logistic Regression using bar graph because it’s very effective to understand and the fig 1

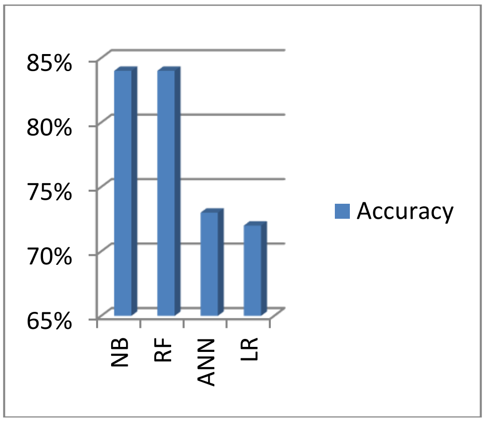

Fig 1: Comparison of accuracies of Naïve Bayes, RF, ANN and LR

Depicts that, Naive Bayes and Random Forest yielding better results compared to Logistic Regression and ANN with this it shows the potential to be applied to a range of medical analysis. Finally, it is hoped that model will be successfully applied to other problem domains

6. REFERENCES

[1] Ajith TA et al (2007) Ascorbic acid and alpha-tocopherol protect anticancer drug cisplatininduced nephrotoxicity in mice: a comparative study. Clin Chim Acta 375(1–2):82–86

[2] Arima H et al (2001) Comparative studies of the enhancing effects of cyclodextrins on the solubility and oral bioavailability of tacrolimus in rats. J Pharm Sci 90(6):690–701

[3] Pritesh Mistry1 · Daniel Neagu1 · Paul R. Trundle1 · Jonathan D. Vessey Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology Published online: 20 November 2015 © The Author(s) 2015. This article is published with open access at Springerlink.com)

[4] Basavaraj S, Betageri GV (2014) Can formulation and drug delivery reduce attrition during drug discovery and development–review of feasibility benefits and challenges. Acta Pharm Sin B 4(1):3–17

[5] Berman JJ ,Principles of big data, (2013)

[6] Werbos PJ: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. In PhD Thesis Harvard University, Cambridge, MA; 1974.

[7] BlixHS et al Drugs with narrowtherapeutic index as indicators in the riskmanagement of hospitalised patients. Pharm Pract 8(1):50–55, (2010)

[8] Paolo Magni4, Paolo Piergiorgi4, Mark A Rubin*2,3,5 and Riccardo Bellazzi4,A hierarchical Naïve Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays Francesca Demichelis1,2,3,

[9] Data Mining, Machine Learning and Big Data Analytics Lidong Wang

[10] International Transaction of Electrical and Computer Engineers System. 2017, 4(2), 55-61. DOI: 10.12691/iteces-4-2-2Published online: July 24, 2017

[11] ALRn: accelerated higher-order logistic regression Nayyar A. Zaidi1 · Geoffrey I. Webb1 · Mark J. Carman1 · François Petitjean1 · Jesús Cerquides2 Received: 10 November 2015 / Accepted: 24 June 2016 / Published online: 22 July 2016 © The Author(s) 2016

[12] Farid E Ahmed* Address: Department of Radiation Oncology, Leo W Jenkins Cancer Center, The Brody School of Medicine at East Carolina University, Greenville,NC 27858, USA Email: Farid E Ahmed* - ahmedf@mail.ecu.edu, Artificial neural networks for diagnosis and survival prediction in colon cancer.

[13] Early warning signal for dengue outbreaks and identification of high risk areas for dengue fever in Colombia using climate and non-climate datasets Jung-Seok Lee1*, Mabel Carabali2,3, Jacqueline K. Lim3, Victor M. Herrera4, Il-Yeon Park3, Luis Villar4 and Andrew Farlow1

[14] Matthew Hamilton1,2, Guy Mahiane1,2, Elric Werst1,2, Rachel Sanders1,2, Olivier Briët3,4, Thomas Smith3,4, Richard Cibulskis5, Ewan Cameron6, Samir Bhatt6,7, Daniel J. Weiss6, Peter W. Gething6, Carel Pretorius1,2 and Eline L. Korenromp1,2*, Spectrum-Malaria: a user-friendly projection tool for health impact assessment and strategic planning by malaria control programmes in sub-Saharan Africa.

[15] Kate Zinszer1,2*, Ruth Kigozi3, Katia Charland1, Grant Dorsey4, Timothy F Brewer5, John S Brownstein2, Moses R Kamya6 and David L Buckeridge1, Forecasting malaria in a highly endemic country using environmental and clinical Predictors.

[16] Mohammad Y. Anwar1*, Joseph A. Lewnard2, Sunil Parikh2 and Virginia E. Pitzer2 Models for short term malaria prediction in Sri Lanka Olivier JT Briët*1,2, Penelope Vounatsou2, Dissanayake M Gunawardena3, Gawrie NL Galappaththy4 and Priyanie H Amerasinghe5, Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence.

[17] Steven L. Moulton1,2, Jane Mulligan2, Anon Srikiatkhachorn3, Siripen Kalayanarooj4, Greg Z. Grudic2, Sharone Green3, Robert V. Gibbons5,6, Gary W. Muniz6, Carmen Hinojosa-Laborde6, Alan L. Rothman7, Stephen J. Thomas5,8 and Victor A. Convertino6*, State-of-the-art monitoring in treatment of dengue shock syndrome: a case series.

[18] Pi Guo1, Tao Liu2, Qin Zhang3, Li Wang1, Jianpeng Xiao2, Qingying Zhang1, Ganfeng Luo1, Zhihao Li2, Jianfeng He4, Yonghui Zhang4, Wenjun Ma2*, Developing a dengue forecast model using machine learning: A case study in China

[19] Anna L. Buczak*, Benjamin Baugher, Erhan Guven, Liane C. Ramac-Thomas, Yevgeniy Elbert, Steven M. Babin and Sheri H. Lewis, Fuzzy association rule mining and classification for the prediction of malaria in South Korea.

[20] Ruchi Verma1, Ajit Tiwari2, Sukhwinder Kaur2, Grish C Varshney2 and Gajendra PS Raghava*1, Identification of Proteins Secreted by Malaria Parasite into Erythrocyte using SVM and PSSM profiles.

[21] Maquins Odhiambo Sewe1,2, Yesim Tozan3,4, Clas Ahlm5 & Joacim Rocklöv2,6, Using remote sensing environmental data to forecast malaria incidence at a rural district hospital in Western Kenya.