Prediction of Dengue Fever using Neuro Hybrid Algorithm

Mr. Mani Deepak Choudhry, Ms. B. Renuga Devi, Ms. N.Roshna, Ms. R. Sruthi

Abstract— Dengue is a threatening disease caused by female mosquitoes. It is typically found in widespread hot regions. From long periods of time, experts are trying to find out some of features on Dengue disease so that they can rightly categorize patients because different patients require different types of dengue treatment. India has been targeting of Dengue disease from last few years. Dengue fever is utilized in characterization strategies to assess and look at their exhibition. The datasets were collected from repositories to predict dengue fever .MATLAB was used as Data mining tool for classification of data. Firstly, evaluate the performance of classification technique such as Support Vector Machine and Radial Basis Function Neural Network along with the help of graph depending upon datasets and secondly, compare the performance of the technique using graphs. The proposed system describes the accuracy of predicting the dengue fever and possible occurrences of fever irrespective of ages and climatic conditions.

Index Terms—Dengue fever, datasets, Support Vector Machine, Radial Basis Function Artificial Neural Network, MATLAB.

I INTRODUCTION

The Dengue fever can be predicted to help and improve the patient health. Collecting datasets required for prediction and perform pre-processing on the data. Applying classification techniques to find whether the patient is affected by dengue or not by using parameters associated with it. To train the machine learning model by using training data and feed the testing data to predict it. On the basis of results, it shows the accuracy of classification techniques. It predicts accurate result related to numerical value from the datasets collected along with a predefined threshold values. Support Vector Machine uses hyper plane to classify the dengue classes whereas Radial Basis Function Artificial Neural Network proposed, which is based on a collection of connected units or nodes referred to as artificial neurons, which loosely model the neurons in a organic brain. Each connection, just like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron that receives a signal then processes it with several hidden layers to provide result.

Manuscript received March, 2020.

B. Renuga Devi*, Information Technology, Dr NGP Institute of Technology, (e-mail: renugadevikrishnan2014@gamil.com). Tirupur, India,9865842517.*

N.Roshna*, Information Technology, Dr NGP Institute of Technology, Namakkal ,India, 7339332877, (email: roshna2404@gmail.com).*

R. Sruthi*, Information Technology, Dr NGP Institute of Technology, Coimbatore, India,9487984980 (e-mail: rsruthi98@gmail.com).*

II LITERATURE SURVEY

The paper proposes Long Short Term Memory - Recurrent Neural Network (LSTM RNN) [1] for predicting dengue fever. Data used for prediction will include the climatic conditions, pollution and the statistics of patients diagnosed with dengue in the previous years. The data has been taken from various government websites as well as healthcare centers with 94% accuracy.

This paper apply extreme learning machine (ELM)[2] method to predict the risk of outbreak based on weather condition. We Develop ELM architecture with weather variables as input nodes and risk level of DHF outbreak as the target. We use binary sigmoid activation function and bipolar sigmoid with a number of hidden neurons between 5 - 200 nodes.

K-Means Clustering as preprocessing and Support Vector Machine Algorithm for classification with weather data presents the research about the prediction of Dengue Hemorrhagic Fever in Bandung Regency with weather data from BMKG[3] (Meteorological, Climatologically, and Geophysical Agency) in Bandung Regency from 2009 until 2016 using the dot and radial kernels on the SVM algorithm. The radial kernel obtains testing accuracy up to 93%, while the kernel dot obtains average of testing accuracy 62%.

The task targets performing Named Entity Recognition[4] to extricate scatter makes reference to, time articulations and other significant highlights from clinical information.. These can be utilized to assemble a model, which can thus be utilized to foresee the nearness or nonappearance of the illness, dengue.. The frequency analysis and frequent occurrences of word in the dataset with 94% accuracy.

All Rights Reserved © 2020 IJARCET

Classification techniques were used for ordering of dataset such as Naive Bayesian, REP Tree, Random tree, J48, SMO, SVM, Decision Tree Approach, and Spatial Data Analysis[5].Data set for dengue prediction is DNA microarray data which have information of gene's expression responsible for dengue virus. Correlation among the methods presumed that Naive Bayes is noticeable among all others as it conveys an exactness of 92% with high likelihood and viability.

III EXISTING SYSTEM

Classification is the process of categorizing a given set of data into the different classes. Classification can be done for both structured and unstructured data. There are seven levels of Classification, they are Kingdom, Phylum, Class, Order, Family, Genus, Species. There are several parameters to predict with climatic conditions.

Clustering in machine learning represents grouping of data points. In a given set of data points, we can use clustering algorithm to represent the data in a specific group. Data points with similar properties and are grouped in similar clustering group where as the unsimilar are grouped in other group.

IV DRAWBACKS OF EXISTING SYSTEM

The existing system uses several classification and clustering machine learning algorithms such as Support Vector Machine, Recurrent Neural Network, SMO and so on. The major restrictions is parameters decided for prediction and the accuracy of the algorithm in predicting the dengue fever.

V PROPOSED SYSTEM

The Proposed System describes the prediction of dengue fever using hybrid machine learning algorithms such as Support Vector Machine and Radial Basis Function Artificial Neural Network based on the weights associated with the function and results in better accuracy of prediction.

It also depicts which algorithm produces better accuracy in predicting dengue fever either Support vector Machine or Radial Basis Function Neural Network.

BLOCK DIAGRAM

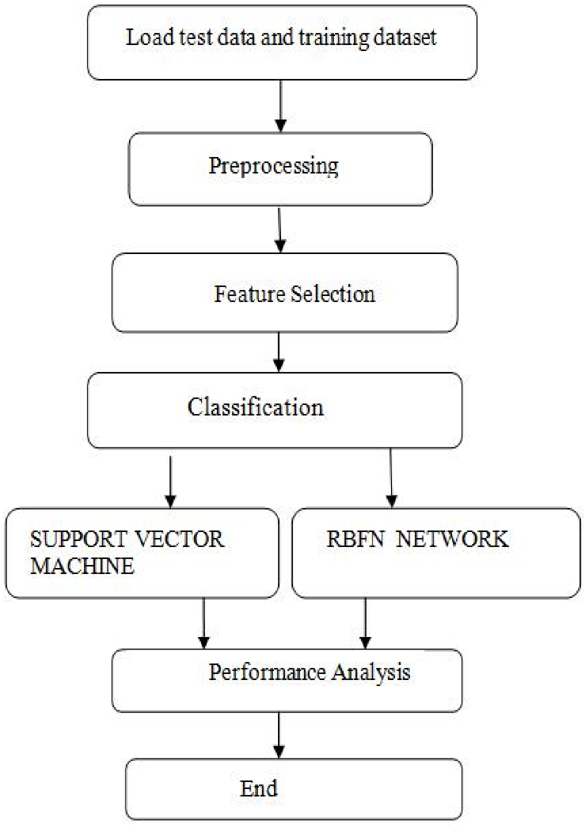

Fig 5.1 Overview of System

SVM algorithm is to find a hyper plane in an N-dimensional space that distinctly classifies the data points. Most extreme edge is the greatest separation between information purposes of both classes. Support Vector Machines (SVMs, likewise support-vector systems) are managed learning models with related learning calculations that dissect informationused for classification and regression analysis. The optimum hyper plane is the linear classifier with minimum margin for a given finite set of learning patterns.

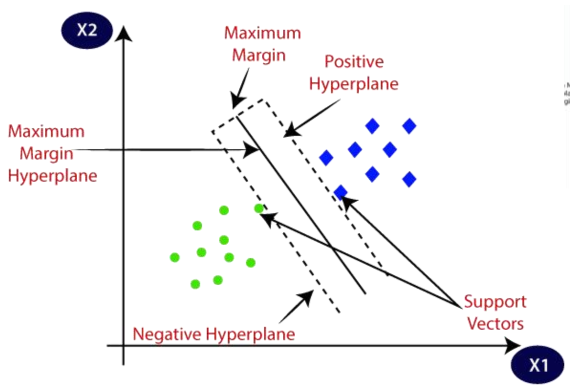

Fig 5.2 Support Vector Machine

Algorithm for RBFN:

1.The RBF Mapping

2.The RBF Network Architecture

3.Computational Power of RBF Networks

4.Training an RBF Network

5.Unsupervised Optimization of the Basis Functions

6.Computing the Output Weights and Supervised RBF Network Training.

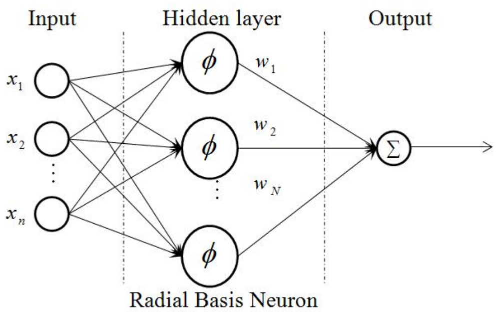

Fig 5.3 Radial Basis Function Neural Network

VI IMPLEMENTATION

1)FORMULATING THE PROBLEM

The first and foremost step in machine learning is to decide what to predict. There are several means to define the problem by using machine learning. Selecting the problem to be defined depends on the use case or business need. If the target is numeric, regression problem needs to be solved. If the target is binary, binary classification problem needs to be solved. It is important to avoid over-complicating the problem and to frame the simplest solution that meets your needs. The Gaussian basis equation as follows,

$$\varphi(x) = \exp\left(-||x-\mu||x^2\right)/2\sigma x^2 \qquad \text{----- (1)}$$

Where μ and σ is $centre$ and $widths$.

2)COLLECTING LABELED DATA

Machine learning problems start with collecting labeled data(the data which has a result). Information for which the objective answer is realized called marked information. In supervised machine learning, the algorithm learns itself from the labeled data. Each data must contain two elements:

The target – The result to be predicted. Provide data that is labeled with the target to the algorithm to learn. Then, the trained machine learning model to predict the result on unlabelled data .

Variables/features – These are attributes that can be used to identify patterns to predict the end result.

Once data labeled, convert data to a format that is acceptable to the algorithm or software. Use machine learning that need to convert the data to comma-separated value (CSV) format with each data making up one row of the CSV file, each column containing one input variable, and one column containing the target result. The project uses the protein levels as parameters to predict results.

3)ANALYZING DATA

Before feeding the labeled data to a machine learning model, it is necessary to inspect data to identify issues and gain insights. The prescient intensity of the model may be tantamount to the information fed. While breaking down the information, the accompanying contemplations:

Variables and target data – Variables is useful to understand the values and identify which is dominant in data. It could show these synopses to a topic master for the issue that need to tackle. Check whether there more missing values or invalid data than you expect.

Variable-target correlations – Variable-target connections – Knowing the connection between every factor and the objective class is useful on the grounds that a high relationship suggests that there is a connection between the variable and the target class. In general, include variables with high correlation because they are the ones with higher predictive power (signal), and leave out variables with low correlation because they are likely irrelevant.

4)FEATURE PROCESSING

Subsequent to finding a good pace through information rundowns and representations, change factors further to make the data more meaningful. This is known as highlight preparing.

Feature processing to form more generalized data points to learn from that can provide significant improvements to the predictive models. Other examples of common feature processing:

Replacing missing or invalid data with more meaningful values. A common strategy used to impute missing values is to replace missing values with the mean or median value. It is critical to comprehend your information before picking a methodology for supplanting missing qualities.

Non-direct changes such as binning numeric factors to classifications. As a rule, the connection between a numeric component and the objective isn't direct (the element esteem doesn't increment or reduction monotonically with the objective). In such cases, it may be helpful to canister the numeric component into unmitigated highlights speaking to various scopes of the numeric element. Each clear cut element (canister) would then be able to be displayed as having its own direct relationship with the objective.

the continuous numeric feature like age is not linearly correlated with the likelihood to having a disease. You can canister age into absolute highlights that may have the option to catch the relationship with the objective all the more

All Rights Reserved © 2020 IJARCET

precisely. The ideal number of containers for a numeric variable is reliant on qualities of the variable and its relationship to the objective, and this is best decided through experimentation through Space explicit highlights and Variable-explicit highlights.

Counting progressively pertinent highlights assists with improving expectation power. Obviously, it isn't constantly conceivable to know the highlights with "signal" or prescient impact ahead of time. So it is a great idea to incorporate all highlights that can conceivably be identified with the objective name and let the model preparing calculation pick the highlights with the most grounded relationships.

5)SPLITTING THE DATA INTO TRAINING AND EVALUATION DATA

The principal objective of machine learning is to sum up past the information cases used to prepare models. Assess the model to gauge the nature of its speculation for information the model has not been prepared on.

A typical procedure is to take all accessible marked information, and split it into preparing and assessment subsets, normally with a proportion of 70-80 percent for preparing and 20-30 percent for assessment. The machine learning system uses the training data to train models to see patterns, and uses the evaluation data to evaluate the predictive quality of the trained model. The machine learning system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics. The system uses 70 percent of dataset for training and 30 percent dataset for evaluation.

VII COMPONENTS

1)NEURONS

ANNs held the natural idea of counterfeit neurons which get input, join the contribution with their inward state (actuation) and a discretionary edge utilizing an enactment capacity, and produce yield utilizing a yield work. The underlying sources of info are outside information, A definitive yields achieve the undertaking.

Neurons are the basic unit of RBFN, while performing evaluation there are three layers they are,

Input layer-the layer in which the datasets fed.

Hidden layer- The subsequent layer is the shrouded layer which is made out of nonlinear units that are associated legitimately to the entirety of the hubs in the information layer. It is of sufficiently high measurement, which serves a distinctive reason from that in a multilayer perceptron.

Each concealed unit takes its contribution from all the hubs at the parts at the information layer. As referenced over the shrouded units contains a premise work, which has the parameters place. The focal point of the premise work for a hub I at the shrouded layer is a vector whose size is the same as the information vector and there is ordinarily an alternate community for every unit in the system, the radial distance d_i between the input vector u and the center of the basis function c is computed for each unit i in the hidden layer

$$d = ||u - c|| \quad \text{--------(2)}$$

Output layer- The change from the info space to the shrouded unit space is nonlinear, though the change to the concealed unit space to the yield space is straight. It the layer in which it describes whether the patient is affected by dengue or not and it also reveals the performance of both algorithm.

2) CONNECTIONS AND WEIGHTS

The system comprises of associations, every association giving the yield of one neuron as a contribution to another neuron. Every association is doled out a weight that speaks to its relative significance. A given neuron can have numerous information and yield associations. Assigning weights to neurons are related to major parameters.

3) PROPAGATION FUNCTION

The proliferation work figures the contribution to a neuron from the yields of its antecedent neurons and their associations as a weighted whole. An inclination term can be added to the consequence of the spread. The propagation has feed forward and feed backward.

$$x = f(u) = \sum_{i=1}^{l} wi\ G\ i\ (||u - ci||) \quad \text{---------(3)}$$

4) ORGANIZATION

The neurons are ordinarily sorted out into different layers, particularly in profound learning. Neurons of one layer interface just to neurons of the promptly going before and quickly following layers. The layer that gets outer information is the information layer. The layer that creates a definitive outcome is the yield layer. In the middle of them are at least zero shrouded layers. Single layer and unlevered systems are likewise utilized. Between two layers, various association designs are conceivable.

They can be completely associated, with each neuron in one layer interfacing with each neuron in the following layer. They can be pooling, where a gathering of neurons in a single layer interface with a solitary neuron in the following layer, in this way diminishing the quantity of neurons in that layer. Neurons with just such associations structure a coordinated non-cyclic diagram and are known as systems. On the other hand, organizes that permit associations between neurons in the equivalent or past layers are known as repetitive systems.

5) HYPERPARAMETER

A hyper parameter is a parameter whose worth is set before the learning procedure starts.

Learning Rate: The learning rate is a constant value used in the Stochastic Gradient Descent (SGD) algorithm. The SGD algorithm makes updates to the weights of the linear model for every data.

Model size: In case of many input features, the number of possible patterns in the data can result in a large model. Enormous models have useful ramifications, such as requiring more RAM to hold the model while preparing and while creating forecasts.

Number of passes: The SGD calculation makes consecutive disregards the preparation information. The Number of passes parameter controls the quantity of passes that the calculation makes over the preparation information. More passes bring about a model that fits the information better.

Data rearranging: Rearranging information on the grounds that the SGD calculation is impacted by the request for the lines in the preparation information. Rearranging the preparation information brings about better ML models since it enables the SGD calculation to stay away from arrangements that are ideal for the main sort of information, however not for the full scope of information. Rearranging stirs up the request for information with the goal that the SGD calculation doesn't experience one sort of information for an excessive number of perceptions in progression.

Regularization: It keeps direct models from over fitting preparing information models by punishing extraordinary weight esteems. L1 regularization has the impact of diminishing the quantity of highlights utilized in the model by pushing to zero the loads of highlight that would some way or another have little loads.

VIII. CONCLUSION

The detailed design and implementation of Support Vector Machine & Radial Basis Function Neural Network algorithms used for a system to identify dengue fever and feature extraction algorithms are used to get the highlights significant to the infection, Dengue. The model generated further aids in the prediction of the disease. Graphs are then used to succinctly represent their performance and correlation. In spite of the correlation of training samples with time frame was compared with the correlation obtained from predicted results. RBFNs are considered as mathematical models to enhance existing data analysis technologies. Although it is not comparable with the power of the human brain, still it is the basic building block of the Artificial intelligence. The proposed system restricted to one epidemic species dengue, that can be extended to more than one epidemic species with better accuracy and better machine learning algorithms with multidimensional datasets. The support vector machine produces 44% accuracy and Radial Basis Function Neural Network with 100% accuracy. Hence RBFN is more efficiency to SVM.

REFERENCES

[1] Megha Chovatiya, Anushka Dhameliya, JayitaDeokar, Jessica Gonsalves,Amrita Mathur Prediction of Dengue using Recurrent Neural Network B.E ,Information Technology ,SFIT, Mumbai, IndiaProfessor, Department Of Information Technology, SFIT, Mumbai, India ,2019.

[2] Abdul Mahatir Najar Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction Department of Mathematics Insitute Teknology Sepuluh November Surabaya, Indonesia,2018.

[3] Computational Science Telkom University Bandung 40257 muflimuzakki The Spreading Prediction of Dengue HemorrhagicFever (DHF) In Bandung Regency Using K-MeansClustering and Support Vector Machine Algorithm M MufliMuzakkiFhiraNhita and Computational Science Telkom University Bandung 40257,2018.

[4] Dave Kaveri Atulbhai and Shilpa Serasiya, "A Survey: Prediction & Detection of Dengue – Mining Methods & Techniques, IJARIIE–ISSN (O)-2395-4396, Vol-3, Issue-2, 2017.

[5] Dengue detection and prediction system using data mining with frequency analysis T. P3 Department of Computer Science and Engineering,SSN College of Engineering, Kalavakkam1nandin,2018.

[6] Henny Pramoedyo Tri Wulandari Kesetyaningsih, "Determination of environmental factors affecting dengue incidence in sleman district, yogyakarta, indonesia", 2018.

[7] Min Chen, Yixue Hao, Kai Hawang, Lu Wang, Lin Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities",IEEE Access,2017.

[8] Chuang T W, Chaves L F, Chen P J. Effects of local and regional climatic fluctuations on dengue outbreaks in southern Taiwan:[J]. Plos One, 2017, 12(6):e0178698.

[9] Tian H, Huang S, Zhou S, et al. Surface water areas significantly impacted 2014 dengue outbreaks in Guangzhou, China[J]. Environmental Research, 2016, 150(150):299-305.

[10] Youngjo C, Tang C S, Lachlan M I, et al. Effects of weather factors on dengue fever incidence and implications for interventions in Cambodia:[J]. Bmc Public Health, 2016.

[11] Guo P, Liu T, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China[J]. Plos Neglected Tropical Diseases, 2017.

[12] Cheng Q, Jing Q, Spear R C, et al. The interplay of climate, intervention and imported cases as determinants of the 2014 dengue outbreak in Guangzhou[J]. Plos Neglected Tropical Diseases, 2017. [13] Marimuthu, T., and Balamurugan, V., "A Novel BioComputational Model for Mining the Dengue Gene Sequences", International Journal of Computer Engineering & Technology, Oct 2015.

[13] Shaukat, K., Masood, N., Shafaat, B, A., Jabbar, K, Shabbir, H., and Shabbir, S., "Dengue Fever in Perspective of Clustering Algorithms",Data Mining in Genomics & Proteomics, 2015.

[14] Sudsom, N., Thammapalo, S., Pengsakul, T., and Techato, K., "A Spatial Clustering Approach to Identify Risk Areas of Dengue Infection After Insecticide Spraying", Journal Technology, Dec 2015.

[15] Shweta., Bhatia, R., Jindal, A., and Sood, M., "Prediction of Dengue

Outbreak using Environmental Factors", 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences, 2016.

B. Renuga Devi pursuing B.Tech Information Technology in Dr NGP Institute of Technology under Anna University. Participated in codiecon competition.Particpated in bigdata workshop in IIT Delhi.

N.Roshna pursuing B.Tech Information Technology in Dr NGP Institute of Technology under Anna University. Participated in codiecon competition.

R.Sruthi pursuing B.Tech Information Technology in Dr NGP Institute of Technology under Anna University. Participated in Hack an Android world Record organized by wikitechy.