International Journal of Pure and Applied Mathematics
Volume 118 No. 18 2018, 3881-3887
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Early Detection of Dengue Using Machine Learning Algorithms

[1]N.Rajathi, [2]S.Kanagaraj, [3]R.Brahmanambika and [4]K.Manjubarkavi,,
[1,2,3,4]Kumaraguru College of Technology, Coimbatore-49.

Abstract

Dengue is an epidemic disease found in several developed as well as developing countries like India. Recent research estimates that there were 390 million dengue infections per year and predicts that dengue transmission is present throughout the tropics, with the highest risk in the American and the Asian regions. This disease is caused by breeding of a mosquito called Aedes. Dengue has become a global problem and it is common in all the countries. Early warnings of dengue outbreak will help to reduce the disease burden and to control it. In this paper, we compare various machine learning algorithms and found the accuracy of these algorithms for the early detection of dengue disease. Data mining is the technique for the classification of diseases like dengue. In this work, Weka toolbox is used to evaluate and compare the results.

Key words Machine learning – Prediction - Data mining - Dengue- Symptoms

1. Introduction

Dengue is a mosquito borne disease found in the tropical region transmitted via Aedes mosquito. Prolonged monsoon rains provide more stagnant water for mosquitoes that carry these diseases to breed. Across India, 70 people have died and more than 36,000 people have been affected by dengue since January, according to the health ministry. Most of the infections have been reported from West Bengal and Orissa in the east and Kerala and Karnataka in the south [1]. Main cause for the rise is the increased testing and reporting of fever, at least in cities and towns. So, overall the number of infections continues to rise: recorded dengue cases alone, for example, have jumped from 28,292 in 2010 to nearly 100,000 cases in 2015. The number of "official" deaths from the disease every year during this period has ranged between 110 and 242. Dengue fever affects the body by causing a high fever and flu-like symptoms. The virus infects your blood and causes headache, rash, achiness, queasiness, tiredness and fever. It can affect the blood and causes mild bleeding on the gums.[2] Dengue illness in the US was estimated to cost $2.1 billion per year on average (in 2010 US dollars), with a range of $1-4 billion in sensitivity analyses and substantial year to year variation. The results highlight the substantial economic burden from dengue in the US. The effects of dengue exceed that of the other viral illnesses, such as human papillomavirus (HPV) or rotavirus [3]. Climatic and socio economic factors influence the dengue disease. Due to the changes in weather conditions like temperature, humidity, wind direction and rainfall leads to the spearing of dengue disease.

The main objective of this paper is dengue disease prediction using various machine learning algorithms. First identify the symptoms of dengue in patients and prediction begins from this identification. The data sets are used for classification and to predict the accuracy. The various classification algorithms are discussed in the forth coming sections.

The content of this paper is organized as follows. Overview of related literatures is presented in section II. The proposed work is presented in detail in section III. In section IV, experimental results are shown. Finally, section V presents the concluding remarks.

2.Literature survey

The authors in [4] considered to employ decision tree as a data mining tool and proposed a set of meaningful attributes from the temporal data. The experiment is divided into four parts. Decision tree approach is used to classify the dengue from two different patient datasets and got the accuracy of 97.6% and 96.6% .The author in [5] aims at performing Named Entity Recognition to extract disorder mentions, time expressions and other relevant features from clinical data. They build a model to predict the presence or absence of the dengue disease and performed frequency analysis which correlates the occurrence of dengue and the manifestation of its symptoms over the months. A set of annotated discharge summaries are used as input to the proposed system. Performance metrics considered in this work are Accuracy, Kappa statistics, Mean Absolute Error, Root Mean Square Error and Relative Absolute Error. It is concluded that the performance of the SMO algorithms is better than others.

The authors [6] proposed statistics-based technique like Multivariate Poisson regression. Statistics is well established methodology of science and useful for verifying relationships among parameters when the relationships are linear while data mining techniques are useful for knowledge finding hidden in the data. They focus on the analysis of linear correlation between dengue cases and infected data of mosquito and demonstrated the role played by female mosquito, their infection season and rate in dengue outbreak prediction. This proposed model efficiently estimated the dengue incidence and assisted in dengue outbreak surveillance and control at the early stages, before outbreaks spread. The authors in [7] proposed real-time dengue risk prediction for a small area. They used risk prediction models instead of a traditional statistical model for early warning, target surveillance and intervention. The precision of the spatial and temporal units can be easily adjusted to different settings for different cities.

In [8], the authors used decision tree algorithm and also explores what rule can act in this area for the future prediction. The objective is to create a prediction model by using decision tree for predicting the chances of occurrences of dengue diseases in a tribal area. The database is analyzed for the creation of an unsupervised model to identify the most significant parameters of affected area and to predict the chances of hitting the disease using the supervised classifier model. The accuracy of the proposed model is 97%

3. Proposed Work

The various algorithms used for early detection of dengue are discussed in this section.

a. Symptoms

The various symptoms for dengue diseases are collected from the Karuna Medical Hospital, Kerala. Figure 1 depicts the various symptoms identified.

.

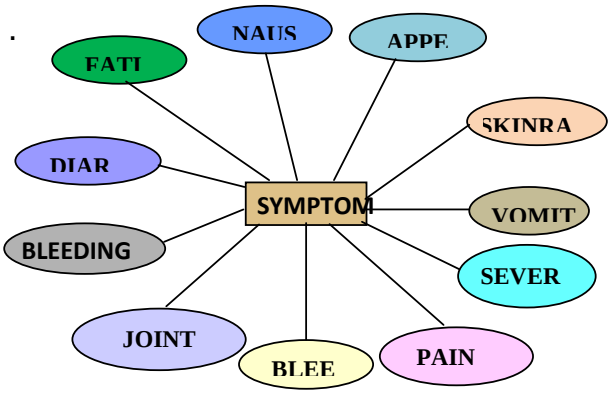

Figure 1. Symptoms for Dengue

According to these symptoms we can predict the dengue disease earlier and able to control it. Sometimes, symptoms are mild or may be mistaken to some other viral infections. However,

serious problems also develop. It includes dengue hemorrhagic fever, characterized by high fever, damage to lymph and blood vessels, bleeding from the nose and gums, enlargement of the liver, and failure of the circulatory system. The symptoms may direct to enormous shock, bleeding and death.

b. Datasets used

The data for this work were collected from health department, Karuna medical hospital, Kerala and from online sources [9]. For the early detection of dengue, totally 100 cases were used. From this 70% of dataset was used for training and the remaining for testing.

c. Algorithms used

In this work, the classification algorithms used are Naive Bayes, J48, Random forest, REP tree, SMO, LWL, AdaboostM1, and ZeroR. Brief overviews of these algorithms are given below:

- **Naive bayes:** This algorithm represents the supervised learning method as well as statistical method for classification. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. Use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.
- **J48:** It is an open source Java implementation of the C4.5 algorithm. The features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc[10]. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data.
- **Random forests**: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a huge amount of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees[11]. It runs efficiently on large data bases.
- **REP tree:** Reduces Error Pruning (REP) Tree Classifier is a fast decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance. It builds a decision/regression tree using information gain/variance reduction and prunes it using reduced error pruning. It only sorts values for numeric attributes once [12,13]. Omitted values are dealt with by splitting the corresponding instances into pieces which is same as C4.5.
- **SMO:** Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines. SMO is an iterative algorithm for solving the optimization problem. It breaks the problem into a series of smallest possible sub-problems, which are then solved analytically. This algorithm contains many optimizations designed to speed up the algorithm on large datasets and ensure that the algorithm converges even under degenerate conditions.
- **LWL:** Locally Weighted Learning (LWL) is a learning model that belongs to the category of memory based classifiers. LWL methods are non-parametric and the current prediction is done by local functions which are using only a subset of the data.
- **AdaboostM1:** AdaBoost is a popular boosting technique which helps to combine multiple “weak classifiers” into a single “strong classifier”. AdaBoost is best used to boost the performance of decision trees on binary classification problems. A weak classifier is simply a classifier that performs poorly, but performs better than random guessing.
- **ZeroR**: ZeroR is the simplest classification method which relies on the target and ignores all predictors. This classifier simply predicts the mean for a numeric class or mode for a nominal class. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

4. Results and discussion

Simulation based experiments were carried out by using the open source tool WEKA 7.0 to determine the performance of the proposed work. This toolbox contains modules for

classification algorithms. This tool was successfully applied in Bioinformatics [14,15]. To analyse the accuracy of the above algorithms, four performance metrics like TP Rate, Classification Accuracy (CA), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used.

- TP Rate: It is called as true positive rate. It is the measure of proportion of positives that are correctly identified as such.

- Classification Accuracy refers the number of correctly classified samples from the input samples.

$$\mathbf{CA= CS/n} \quad \text{--- (1)}$$

where, CS represents the correctly classified samples and n indicates the number of samples.

- Mean Absolute Error: It is the average of absolute errors

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad \text{---(2)}$$

where, f_i represents the prediction value and y_i is the true value

- Root Mean Square Error: It is the square root of sum of squares error divided by number of predictions

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2} \quad \text{--- (3)}$$

The performances of the various algorithms used are presented in table 1 and pictorially represented in Fig.2a and Fig. 2b.

Table 1. Performance of various algorithms used

Algorithms	TP Rate	Classification Accuracy	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Naive Bayes	0.633	63.3%	0.4133	0.5328
J48	0.767	76.6%	0.3260	0.4153
Random forest	0.833	83.3%	0.2233	0.3821
REP tree	0.767	76.6%	0.3400	0.4243
SMO	0.800	80%	0.2000	0.4472
LWL	0.567	56.6%	0.4633	0.5171
AdaboostM1	0.567	56.6%	0.4552	0.5167
ZeroR	0.767	76.6%	0.3444	0.4237

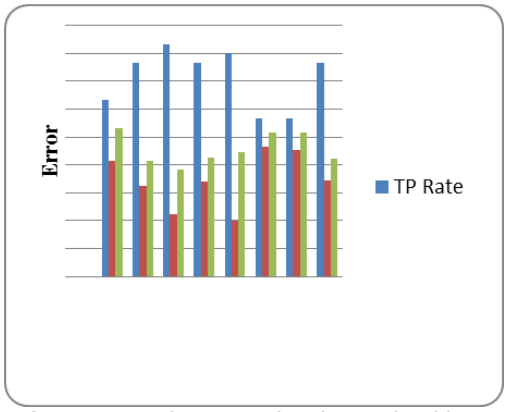

Figure.2a.Performace of various Algorithms

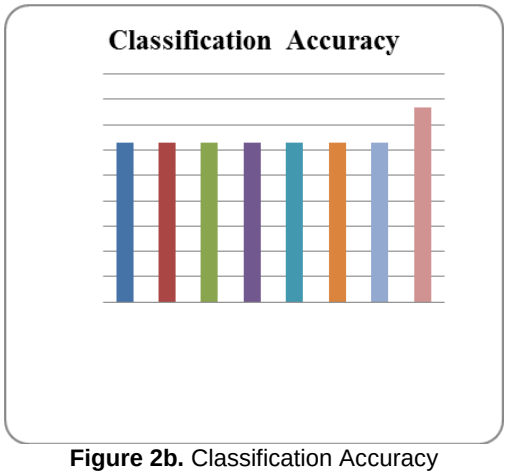

Figure 2b. Classification Accuracy

The result of this study shows that the performance of the Random forest algorithm is better when comparing to other algorithms for the application proposed.

5. Conclusion

The main focus of this paper is to detect the dengue disease at an early stage using the machine learning algorithms. The performance of the various algorithms was compared and finally it is concluded that the random forest algorithm gives better classification accuracy. Future work is towards the application of various evolutionary algorithms for the problem undertaken.

References

[1] Wikipedia-http://www.bbc.com/news/world-asia-india-37415781
[2] Wikipedia-scribol.com/science/medicine/the-horrible-effects-of-dengue-fever/
[3] Wikipedia-https://www.ncbi.nlm.nih.gov/pubmed / 21292885
[4] Thitiprayoonwongse DA, Suriyaphol PR, Soonthornphisaj NU," Data mining of Dengue Infection using Decision Tree", Entropy, 2012.
[5] Nandini V, Sriranjitha R , Yazhini TP, "Dengue detection and prediction system using data mining with frequency analysis", Computer Science & Information Technology (CS & IT), 2016.
[6] Padet Siriyasatien,Atchara Phumee,Phatsavee Ongruk,Katechan Jampac

haisri,Kraisak Kesorn, “Analysis of significant factors for dengue fever incidence prediction”, BMC Bioinformatics ,2016.

[7] Ta-Chien Chan, Tsuey-Hwa Hu, Jing-Shiang Hwang,”Daily forecast of dengue fever incidents for urban villages in a city”, International Journal of Health Geographics,2015.

[8] N K Kameswara Rao, Dr. G P Saradhi Varma, Dr.M.Nagabhushana Rao, “Classification Rules Using Decision Tree for Dengue Disease”, International Journal of Research in Computer and Communication Technology, Vol 3, Issue 3,2014.

[9] Wikipedia- www.webmd.com/a-to-z-guides/dengue-fever-reference

[10] Anshul Goyal, Rajni Mehta, “Performance Comparison of Naïve Bayes and J48 Classification Algorithms”, IJAER, Vol. 7, No. 11, 2012.

[11] N. Peter, “Enhancing random forest implementation in WEKA”, in: Machine LearningConference, 2005.

[12] Breiman, L.: Random Forests. ML Journal 45(1), 5–32, 2001.

[13] Juan Wang; Qiren Yang; Dasen Ren, "An Intrusion Detection Algorithm Based on Decision Tree Technology," Information Processing, APCIP 2009. Asia-Pacific Conference, vol.2, no.,pp.333, 335,2009.

[14] David S. K., Saeb A. T., Al Rubeaan K, “Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics”, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.

[15] Durairaj M, Ranjani V, “Data mining applications in healthcare sector a study”, Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.