Dengue Disease Prediction Using Decision Tree and Support Vector Machine

Dr. Arun Kumar P.M, Associate Professor, 1 Chitra Devi B, 2 Karthick P, 2 Ganesan M and 3 Madhan A.S

Department of Information Technology, Velalar College of Engineering and Technology, Thindal.

ABSTRACT

The clinical documents maintained are a pool of information regarding the infected patients. By keeping this voluminous data we can predict the future occurrences of the disease earlier and safeguard the people. Dengue the global problem is common in more than 110 countries. Dengue fever is a vector borne disease caused by the female Aedes Aegypti and Aedes Albopictus mosquitoes which adapt well to human environments. Data mining is a well known technique used by health organizations for classification of diseases. The collected dataset was experimented with Weka and Net Beans IDE and a decision tree is generated using Fisher Filtering and prediction done effectively with Support Vector Machine.

KEYWORDS

Data mining, Feature extraction, Decision tree, Support Vector Machine.

INTRODUCTION

Dengue is the life threatening disease, caused by the mosquito extent in the body of humans and leads to mortality. Dengue is also known as bone breaking illness. Dengue infection has endangered nearly two billion populations throughout the world. Dengue is divided into two types as type 1 and type 2 namely Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) by the World Health Organization.

Dengue Hemorrhagic Fever (DHF) is again classified into DHF 1, DHF 2, DHF 3, and DHF 4. It causes abdominal pain, hemorrhage, circulatory collapse, acute platelet deficiency. The symptoms of dengue include bleeding, low levels of blood platelets, low blood pressure and metallic taste in mouth, headache, joint pain and rashes. It is difficult to differentiate dengue fever and dengue hemorrhagic fever.

The disease transmission occurs when Aedes Aegypti mosquito bites a healthy person; the virus enters into the body fluids of that person. Then it starts reproducing inside the white blood cells and initiates the dengue virus cycle.

The dataset was loaded in Net Beans IDE, preprocessed to remove missing values, Fisher filtering algorithm employed with feature extraction and a decision tree is generated based on the Aegypti rate then Support Vector Machine used to improve the prediction system.

LITERATURE REVIEW

[1] Dengue being a life threatening disease also known as break bone fever was predominant over all parts of the world. The female mosquito Aedes Aegypti causes the disease and more prevalent during the rainy season.

Data mining provides an efficient way to resolve many problems in the field of bioinformatics. This paper focuses on finding the best algorithm with high accuracy for the prediction of dengue among several classification algorithms. [8]

A dengue dataset with 18 attributes and 108 instances used to predict the disease and their accuracy compared with different classification algorithms to find out the best algorithm. The classification accuracy found using 10 cross validation with interfaces Explorer, Experimenter, Knowledge Flow and Simple CLI.
The algorithms such as Naive Bayes, J48, SMO, REP tree and Random tree employed in this paper to obtain better accuracy.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time taken to build the model (sec)</th>
<th>Correctly classified instances</th>
<th>Incorrectly classified instances</th>
<th>Mean Absolute Error</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>J48</td>
<td>0</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
<td>0.958</td>
</tr>
<tr>
<td>SMO</td>
<td>0</td>
<td>100%</td>
<td>0%</td>
<td>0</td>
<td>0.875</td>
</tr>
<tr>
<td>REP Tree</td>
<td>0.02</td>
<td>74.74%</td>
<td>25.25%</td>
<td>0.365</td>
<td>0.544</td>
</tr>
<tr>
<td>Random Tree</td>
<td>0</td>
<td>87.87%</td>
<td>12.12%</td>
<td>0.185</td>
<td>0.876</td>
</tr>
</tbody>
</table>

Based on the results obtained, explorer interface and knowledge flow interface data mining techniques it is clear that Naive Bayes and J48 algorithms have maximum accuracy.

[2] Fumigation used to reduce the population of mosquitoes and temephos employed. Dengue Hemorrhagic fever (DHF) occurs very frequently and its epidemic seasons must be predicted earlier.

Dengue outbreak tracked, monitored and predicted to help the local authorities for future prediction. The less developed areas of the world such as Iloilo city of Philippines more prone to hospitalization of children. This paper brings a web based Dengue tracking system (DTS) wherein environmental factors considered for future prediction. [5]

The climatic factors such as amount of rainfall, relative humidity, mean temperature and monthly-recorded cases considered. The proposed web based system includes two users. They are the Admin and the User.

The developed system works with cloud computing. Google applications play a vital role in the system working.

Google Map utilized to plot the cases infected with dengue. Google Chart utilized for dengue rates and prediction. Google Heat Map utilized for determining the hotspot districts at periodic intervals.

The Artificial Neural Network model gets five inputs such as amount of rainfall, humidity rate, temperature rate, number of dengue cases and previous month cases. The output values from Artificial Neural Network been integrated into the database. The monthly rainfall rate, monthly temperature rate, monthly relative humidity and the number of cases and the values of each category integrated into the system and used for future predictions. [4]

The user interface has three major components such as the Case Entry, the Analysis and the Case Reports. The Admin interface has four major components. They are the Case Entry, the Analysis, Trends Prediction and the Case Reports.

The proposed web system was useful in many aspects to the Iloilo City. It helps to reduce the dengue cases and knowledge about it to the local surveillance. [7]

In addition to process effectively, offline tools needed since Google applications may not be available often. The Artificial Neural Network system needs additional features to work better.

Moreover, to handle missing data, another algorithm required.

[3] The research article titled Dengue Fever Prediction; A Data Mining Problem focuses on the severe dengue outbreak at Pakistan where many victims suffer. According to the World Health Organization, dengue disease divided into two types as type1 and type2. These types named as Classical dengue and dengue hemorrhagic fever respectively.

The dataset used for classification collected from the District Headquarter Hospital Jhelum consisting of ninety-five entries. The classification techniques such as Naive Bayes, REP, Random, J48, SMO were employed. [6]

The attributes used in the dataset are id, fever, bleeding, myalgia, flu, fatigue, other symptoms and results. There are four types of precision employed in it.

TN = case was negative and predicted negative, TP = case was positive and predicted positive.
FN = case was positive but predicted negative, FP = case was negative but predicted positive.

<table>
<thead>
<tr>
<th>Technique</th>
<th>TP Rate</th>
<th>ROC Rate</th>
<th>Error Rate</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>0.92</td>
<td>0.815</td>
<td>0.08</td>
<td>0.92</td>
</tr>
<tr>
<td>REP Tree</td>
<td>0.76</td>
<td>0.999</td>
<td>0.24</td>
<td>0.76</td>
</tr>
<tr>
<td>Random Tree</td>
<td>0.76</td>
<td>0.999</td>
<td>0.24</td>
<td>0.76</td>
</tr>
<tr>
<td>J48</td>
<td>0.88</td>
<td>0.596</td>
<td>0.24</td>
<td>0.88</td>
</tr>
<tr>
<td>SMO</td>
<td>0.76</td>
<td>0.494</td>
<td>0.24</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Based on the above results, the ROC of Naive Bayes found to be 0.815 and has the smallest error rate. And it is revealed that Naive Bayes and J48 had accuracy 92% and 88% respectively. [8]

**IMPLEMENTATION**

The proposed framework based on disease prediction is shown to be effective in addressing this prediction.

The framework suggests a novel way of network classification: first, capture the latent affiliations of actors by extracting disease prediction based on network connectivity and next apply extant data mining techniques to classification based on the extracted prediction. In the initial study, modularity maximization was employed to extract disease prediction. The superiority of this framework over other representative relational learning methods has been verified with dengue prediction dengue data. We propose an effective edge-centric approach to extract sparse disease prediction. [9]

We prove that with our proposed approach, sparsity of disease prediction is guaranteed.

**DATA VISUALIZATION AND PREPROCESSING**

The file containing dataset is saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. In this preprocessing state the missing values are replaced. The ID of the patient cases does not contribute to the classifier performance. Hence it is removed and the outcome attribute defines the target or dependant variable thus reducing the feature set size to desired attributes. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented in the following sections.

**FEATURE SELECTION ALGORITHMS**

We aim to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the dataset are briefly outlined below.

**FISHER FILTERING**

This module is used to filter that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the dengue and the predictor attributes. After computing the Fisher score for each feature, it selects the top-m ranked features with large scores. The next subsection directs focus on another technique of feature selection based on logistic regression.

**FEATURE REDUCTION BY SVM**

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as preprocessing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly contributes to classification that cuts preprocessing costs and minimizes the effects of the ‘peaking phenomenon’ in classification. Thereby improving the overall performance of classifier based intrusion detection systems.

SVM is a linear transformation with linear orthogonal basis vectors; it can be expressed by a translation and rotation.

The below figures mention the snapshots of implementation process where first image represents the preprocessing stage and third diagram denotes the decision tree generated with Fisher filtering.
CONCLUSION

The infection rates of Aedes Aegypti mosquitoes increase morbidity rate hence the decision tree is generated with the Aegypti rate as the root node and prevent further occurrences. The prediction of dengue infection carried out using Weka data mining tool and data mining techniques such as Decision tree and Support Vector Machine. Thus the model helps to predict the dengue cases earlier and reduce mortality rate.

REFERENCES

[1] “Data Mining of Dengue Infection Using Decision Tree” by Draraneethitprarayonwongse, Prapat suriyaphol, Department of Computer Science, Faculty of Science, Kasettsart University, Kasettsart University Research and Development Institute under grant no.130.55 and ISBN:978-1-61804-092-312


