Journal of
Data Mining in Genomics & Proteomics

Dar et al., J Data Mining Genomics Proteomics 2015, 6:3
DOI: 10.4172/2153-0602.1000181

Research Article Open Access

Dengue Fever Prediction: A Data Mining Problem

Kamran Shaukat[1]*, Nayyer Masood[2], Sundas Mehreen[1] and Ulya Azmeen[1]

[1]IT Department, University of the Punjab, Jhelum Campus, Pakistan
[2]Mohammad Ali Jinnah University, Islamabad Campus. Pakistan

Abstract

Dengue is a threatening disease caused by female mosquitos. 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 treatment. Pakistan has been target of Dengue disease from last few years. Dengue fever is used in classification techniques to evaluate and compare their performance. The dataset was collected from District Headquarter Hospital (DHQ) Jhelum. For properly categorizing our dataset, different classification techniques are used. These techniques are Naïve Bayesian, REP Tree, Random tree, J48 and SMO.

WEKA was used as Data mining tool for classification of data. Firstly we will evaluate the performance of all the techniques separately with the help of tables and graphs depending upon dataset and secondly we will compare the performance of all the techniques.

Keywords: Dengue fever classification; Naïve bayes; J48; SMO; REP

Introduction

Dengue infection is vital disease caused by dengue germ, which extent in body of human by female mosquito [1]. With indications of headache, retro orbital pain, joint-pain, muscular pain and rush evidence [2]. It is also known as bone breaking illness [3].

Dengue infection has endangered 2.5 billion populations all around the world. Every year there are 50 million people who suffer from it globally [1]. Pakistan has been victim of this rapidly growing sickness from last few years. Since 2007 in Pakistan, large number of cases was marked especially in Lahore. In 1994 at Karachi Pakistan's first case of dengue was appeared and Dengue's outbreak in 2011, that was more life-threatening than preceding years and 1400 people were affected [3].

Dengue is divided into two types, i.e., type 1 and type 2, according to world health organization [3]. First one is classical dengue called dengue fever and the other is dengue hemorrhagic fever. DHF1, DHF2, DHF3 and DHF4 are further four types of dengue hemorrhagic fever. DHF is revealed by start of fever which continues for 2 to 7 days with number of signs like leakage of plasma, shock and weak pulse. In earliest cases it's hard to differentiate dengue fever from dengue hemorrhagic fever.

Different techniques for dengue fever classification can be used such as NB classifier; decision tree, KNN Technique, multilayered Technique and SVM [1,4,5]. These techniques are evaluated based on five measures accuracy, precision, sensitivity, specificity and negative rate.

Some researchers worked on dengue (fever) classification such as Tanner et al. and Tarig et al. Tanner's team used Decision tree approach and they classified 1200 patients and found 6 remarkable features. They got 84% accurateness [6]. Tarig's team used Self Organizing MAP (SOM) and ML feed-forward neural networks (MFNN). They clustered patients into two sets and got only 70% correctness [7]. Fatimah Ibrahim et.al used ML perceptron's (MLP) and got 90% accuracy [8]. Daranee et al. suggested using decision tree method to classify dengue patients from two data sets [9]. They got 97.6% and 96.6% accuracy from first and second experiment respectively. The accuracy of both experimentations in unseen test set were more than 90% But in experiment of day0 correctness was very low and tree was found to be over fitted. So, experimental results shown that decision tree approach did not counterpart this task very much.

Wajeeha Farooqi et al. categorized Dengue fever by using one of classification technique Decision Tree [3]. They used Data Mining techniques for the efficient classification of the dengue fever type. They performed two experimentations using Decision tree. The first general experiment demonstrates the accuracy of 99.44%.The Second experiment classifies dengue fever on the base of expert weighted attributes, which are used in classification on the base of Minimum Cost and source availability. Correctness of this model's still high 98.62%. We matched performance in term of Type II error. It was found that Type II error is very little in second experimentation.

M Naresh Kumar used alternating Decision Tree Approach for early diagnosis of Dengue fever and accorded its performance with C4.5 algorithm [10]. An alternating Decision Tree technique was able to distinguish the dengue fever using the clinical and laboratory data with number of correctly classified occurrences as F-measure, and (ROC) as compared to C4.5, h F-measure. Alternating Decision tree based approach with boosting has been able to foresee dengue fever with a greater degree of correctness than C4.5 based Decision tree using simple clinical and laboratory features.

Noor Diana et al. presented Malaysian dengue outbreak detection model using three classification methods [11]. They presented a collection of dissimilar dengue, data attributes are used for classification modeling and performances are matched with previous related work. Experimental results show that suggested classifiers improve performance of other methodologies. Significant selection of attributes in dengue dataset supports to good results. The Decision tree and Nearest Neighbor models were generally used methods in this problem, while RS was a rule based method which provides significant knowledge to be further well-thought-out by professionals.

***Corresponding author:** Kamran Shaukat, IT Department, University of the Punjab, Jhelum Campus, Pakistan, Tel: 0544-448770; E-mail: dfbxff@gmail.com

Received June 04, 2015; **Accepted** October 19, 2015; **Published** October 25, 2015

Citation: Shaukat K, Masood N, Mehreen S, Azmeen U (2015) Dengue Fever Prediction: A Data Mining Problem. J Data Mining Genomics Proteomics 6: 181. doi:10.4172/2153-0602.1000181

Weka data mining tool was used by Kashish Ara et al. for Dengue Disease prediction. Dengue data was firstly classified and then equated the different Data Mining techniques in WEKA through different interfaces as mentioned in (Figure 1) [12]. A dengue dataset with 107 illustrations was used to validate approach additionally but weka Data mining tool used 99 rows and 18 attributes to discover best performance and to conclude forecast of disease and their correctness using classifications of different techniques. For categorizing data and to support manipulators in mining useful info from data and effortlessly recognize an appropriate technique for precision of analytical exemplary from that, as it was core objective of their research. The conclusion is that NB and J48 are efficient techniques for accuracy as less time was consumed for constructing this model through WEKA applications outcomes and they attained maximum accuracy=100% with 99 correctly categorized instances ,maximum ROC=1, had minimum mean absolute error [12].

Objective

The general objective of this research is to use few of the classification techniques to determine the population of Dengue fever infected cases in Jhelum district and in surrounding areas geographically. So, that we can compare performance of different classification techniques. Objective of this study also includes the comparison of different classification algorithms with the help of graphs, based on our dataset. We have implemented all the techniques by using weka tool and all the procedure of implementation is within it.

Methodology

We used WEKA as the DM tool for testing and execution. WEKA is a popular set for machine learning software carved in JAVA developed at the University of Waikato, New Zealand [13]. We are using some basic techniques of classification from ML method. Our main focus is on dengue testing that whether a patient is affected by dengue or not by using some attributes. On the basis of results, we will show accuracy of classification techniques and then compare them. It is very good Data Mining tool for the classification of accurateness, by using the different techniques.

Classification

Classification is the type of Data mining, which deals with the problematic things by recognizing and detecting features of infection, among patients and forecast that which technique shows top performance, on the base of WEKA's outcome.

Five techniques have been used in this paper. These techniques uses Explorer interface and it depends on dissimilar techniques NB, REP Tree, RT, J48 and SMO.

All techniques, which we used, were applied on a Dataset of Dengue fever, as enlightened above. Classification and accuracy used was mentioned in (Table 1).

Dataset

The Dataset is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data has almost 95 entries but we are using 25 random entries.

This dataset was taken from District Headquarter Hospital Jhelum.

EPID #	Fever	Bleeding	Myalgia	Flu	Fatigue	Results
1	Yes	No	Yes	No	Yes	Positive
2	Yes	No	Yes	Yes	Yes	Positive
3	Yes	No	No	No	Yes	Positive
4	Yes	No	Yes	No	Yes	Positive
5	Yes	No	No	No	No	Negative
6	Yes	No	No	No	Yes	Negative
7	Yes	No	No	No	Yes	Positive
8	Yes	No	Yes	No	No	Negative
9	Yes	Yes	No	No	No	Negative
10	Yes	No	Yes	No	No	Positive
11	Yes	Yes	No	No	Yes	Positive
12	Yes	No	Yes	No	Yes	Positive
13	Yes	Yes	Yes	No	No	Positive
14	Yes	No	Yes	No	Yes	Negative
15	Yes	No	No	No	Yes	Positive
16	Yes	No	No	No	No	Negative
17	Yes	Yes	Yes	No	Yes	Negative
18	Yes	No	Yes	No	Yes	Positive
19	Yes	No	Yes	No	No	Negative
20	Yes	No	No	No	Yes	Positive
21	Yes	No	Yes	No	Yes	Positive
22	Yes	No	Yes	No	No	Positive
23	No	No	Yes	No	Yes	Positive
24	Yes	No	Yes	No	Yes	Negative
25	Yes	No	Yes	No	No	Negative

Figure 1: Chunk of dataset.

Chunk was selected from this dataset which was treated as Training set and tested this dataset on WEKA Data Mining tool. Some data was classified and rest was tested to check accuracy of data.

Attributes

CSV is the file format of datasets which is taken by weka tool. The Attributes that we have chosen for the testing of dengue are fever, bleeding, myalgia, flu, fatigue and other indications with class label of results with positive and negative consequences (Figure 2). The attributes description is given in (Table 2).

Data mining techniques

Different DM techniques have been used for predicting Dengue virus. These predictions have been done for the purpose of classification and accuracy by using different techniques. The edge used for this objective in paper is Explorer Interface. Accuracy can be observed by selecting the following procedures: NB, REP tree, RT, J48 and SMO.

The techniques we are using are following:

- NB
- REP Tree
- RT
- J48
- SOM

Naïve bayes technique: It performs arithmetical prediction, i.e., forecasts class membership possibilities. It is based on Bayes formula. A simple NB classifier; ensures comparable performance with ID3 and selected neural system classifiers. We verified our training set on Weka Data Mining tool with NB Technique, we got the outcomes mentioned in the (Table 3).

REP tree: Rep Tree uses a regression tree reason and creates several trees in different reiterations. After that it picks best one from all produced trees. That will be measured as the illustrative (Figure 3). In pruning the tree an amount used is a mean square error on the

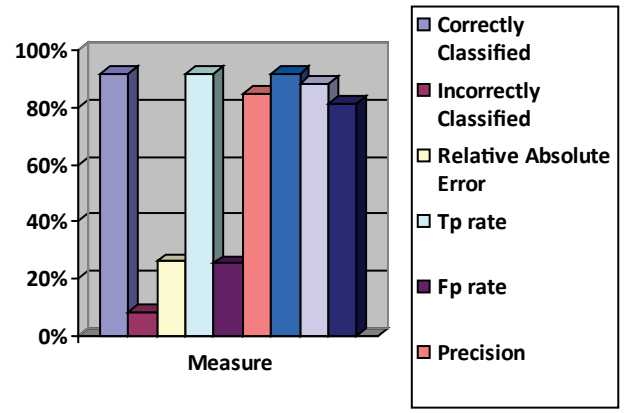

Figure 2: Baysian graph.

Attribute Name	Definition
Correctly Classified	Displays the percentage of correctness test that how many instances are categorized accurately.
Incorrectly classified	Displays the percentage of incorrectness test that how many instances are categorized accurately.
TP Rate	Those which were true and classified as True.
FP Rate	Those which were false but classified as True.
ROC Rate	ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. ROC graphs have long been used in signal detection theory.
Precision	Calculating precision and recall is actually quite easy. When you get the actual results you sum up how many times you were right or wrong
Types of Precision TN TP FN FP	There are four ways of being right or wrong: case was negative and predicted negative case was positive and predicted positive case was positive but predicted negative case was negative but predicted positive
Accuracy	A measure of a predictive model that reflects the proportionate number of times that the model is correct when applied to data.
Error Rate	A number that reflects the rate of errors made by a predictive model. It is one minus the accuracy.

Table 1: Attributes definition.

Attributes	Description
Epid	id of Patient
Fever	Yes or no
Bleeding	Yes or no
Flu	Yes or no
Myalgia	Yes or no
Others	Other symptoms
Results	Positive or negative

Table 2: Atrribute description.

Attributes name	Measure
Correctly Classified	92%
Incorrectly Classified	8%
Relative Absolute Error	26%
Tp rate	0.92
Fp rate	0.253
Precision	0.848
Recall	0.92
F-measure	0.882
Roc Area	0.815

Table 3: Bayesian technique.

estimations made by the tree. We tested our training set on weka Data Mining tool with REP tree technique; we got the outcomes mentioned in the (Table 4).

RT: Random Tree is the supervised Classifier; it was a collective learning technique which generates many single learners. It employs a catching idea to create a set of random data for building an ID3 (Figure 4). In standard tree near each node is divided using the best split amongst all variables. In the random forest, every node is split using a best amongst the subset of predicators arbitrarily chosen at that node [14]. We tested our training set on weka Data Mining tool with Random tree technique; we got the outcomes mentioned in the (Table 5).

J48: C4.5 is the technique used to create a decision ID3 developed by Ross Quinlan. C4.5 is an addition of Quinlan's earlier ID3 Technique. The decision trees created by C4.5 can also be used for classification, and for this purpose, C4.5 is often stated to as an arithmetical classifier. C4.5 constructs decision trees from the set of training data in the identical way as ID3, with the concept of the information entropy (Figure 5). We tested our training set on weka Data mining tool with J48 Technique; we got the outcomes mentioned in the (Table 6).

SMO: SMO is abbreviation of Sequential minimal optimization, which is a technique for answering the QP problem that rises during the training of SVM. SMO is widely used for the training of SVM [15]. We are using this technique on the base of dataset, for splitting our data (Figure 6). After running this technique we assessed the output of classifier by altered measurements to create prediction for each and every occurrence of Dengue dataset. We tested our training set on Weka Data Mining tool with SMO technique; we got the outcomes mentioned in the (Table 7).

Comparison

With 5 techniques of Data Mining, We have completed classification on our dataset. After analysis of our dataset with each technique we are paralleling them in the conclusion. When we have done the comparison among all of them we concluded that naïve Bayes Technique is greatest among all others. As the accuracy of Naive bayes is 92% which was biggest of all. Naïve Bayes is the best also for the aim that it gives the probability and efficiency while Random Tree and REP Tree don't give us probability. The below gives the comparison of all the techniques (Table 8). The graph comparison is given in (Figure 7).

Conclusion

The main Objective of this paper is toward prediction of dengue infection using WEKA Data Mining tool. Basically it has four edges. Out of these four edges we are consuming only one edge which is Explorer. We are using five techniques of classification, i.e., NB, SMO, J48, RT and REP tree. These techniques were applied using Weka Data Mining tool to evaluate the accuracy which was gained after analysis of these techniques. After testing these techniques the outcome were compared on the base of accuracy. These techniques match classifier accuracy with each other on base of correctly classified instances, a precision, error rate, TP rate, FP rate and ROC Area.

Over Explorer technique it has concluded that NB and J48 are the top performance classifier techniques by way that, they has achieved an accuracy of 92% and 88%, takes fewer time to run and shows ROC area=0.815, and had smallest error rate.

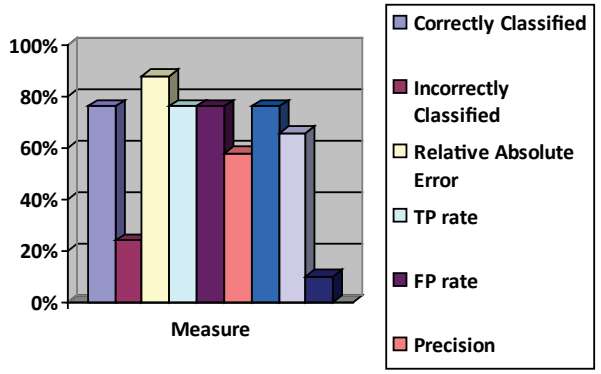

Figure 3: REP graph.

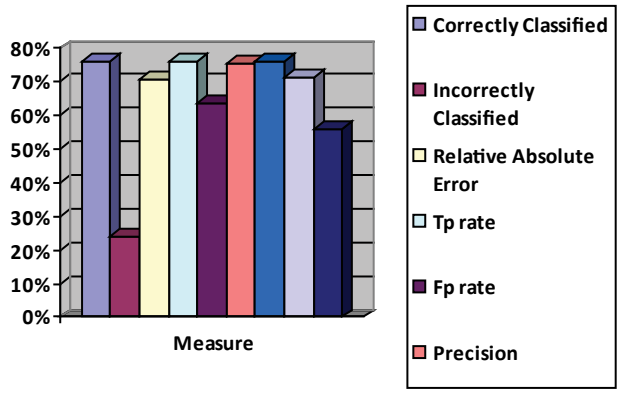

Figure 4: Random forest graph.

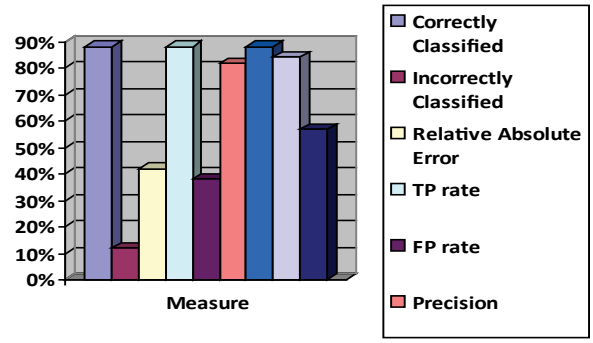

Figure 5: J48 graph.

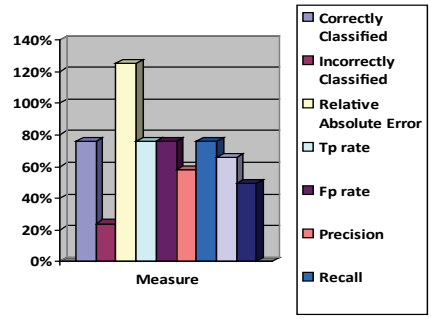

Figure 6: SMO graph.

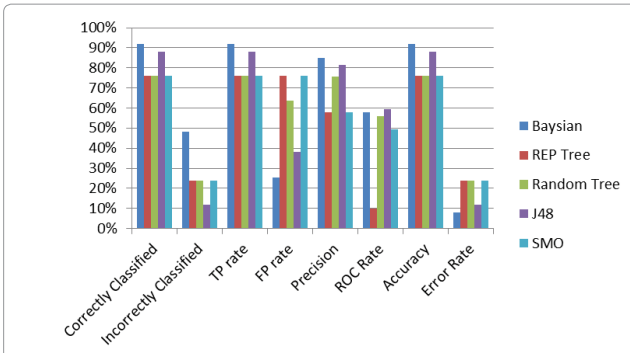

Figure 7: Comparison graph.

Attributes name	Measure
Correctly Classified	76%
Incorrectly Classified	24%
Relative Absolute Error	88%
TP rate	0.76
FP rate	0.76
Precision	0.578
Recall	0.76
F-measure	0.656
Roc Area	0.099

Table 4: REP technique.

Attributes name	Measure
Correctly Classified	76%
Incorrectly Classified	24%
Relative Absolute Error	71%
Tp rate	0.76
Fp rate	0.635
Precision	0.755
Recall	0.76
F-measure	0.715
Roc Area	0.561

Table 5: Random tree.

Attributes name	Measure
Correctly Classified	88%
Incorrectly Classified	12%
Relative Absolute Error	42%
TP rate	0.88
FP rate	0.38
Precision	0.816
Recall	0.88
F-measure	0.842
Roc Area	0.569

Table 6: J48 technique.

Attributes name	Measure
Correctly Classified	76%
Incorrectly Classified	24%
Relative Absolute Error	125%
Tp rate	0.76
Fp rate	0.76
Precision	0.578
Recall	0.76
F-measure	0.656
Roc Area	0.494

Table 7: SMO technique.

Techniques	TP rate	ROC Rate	Error Rate	Accuracy
Baysian	0.92	0.815	0.08	0.92
Rep tree	0.76	0.099	0.24	0.76
Random Tree	0.76	0.099	0.24	0.76
J48	0.88	0.596	0.24	0.76
SMO	0.76	0.494	0.24	0.76

Table 8: Comparsion table.

References

1. Farooqi W, Ali S (2013) A Critical Study of Selected Classification Algorithms for Dengue Fever and Dengue Hemorrhagic Fever. Frontiers of Information Technology (FIT), 11th International Conference on IEEE.
2. Farooqi W, Ali S, Abdul W (2014) Classification of Dengue Fever Using Decision Tree. VAWKUM Transaction on Computer Sciences 3: 15-22.
3. Rigau-Pérez JG, et.al. (1998) Dengue and dengue haemorrhagic fever. The Lancet 19: 971-977
4. Tanner L, Schreiber M, Low JG, Ong A, Tolfvenstam T, et.al. (2008) Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness. PLoS Neglected Tropical Disease 12: e196.
5. Phyu TN (2009) Survey of classification techniques in data mining. Proceedings of the International MultiConference of Engineers and Computer Scientists Vol 1.
6. Vong S, et.al. (2010) Dengue incidence in urban and rural Cambodia: results from population-based active fever surveillance, 2006–2008. PLoS neglected tropical diseases 4: e903.
7. Faisal T, Ibrahim F, Taib MN (2010) A noninvasive intelligent approach for predicting the risk in dengue patients. Expert Systems with Application 37: 2175-2181.
8. Ibrahim F, Taib MN, Abas WA, Guan CC, Sulaiman S (2005) A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). Computer Methods and Programs in Biomedicine 79: 273-281.
9. Daranee T, Prapat S, Nuanwan S (2012) Data mining of dengue infection using decision tree. Entropy 2: 2.
10. Kumar MN (2013) Alternating Decision trees for early diagnosis of dengue fever. arXiv preprint arXiv: 1305.7331.
11. Tarmizi NDA, et.al. (2013) Classification of Dengue Outbreak Using Data Mining Models. Research Notes in Information Science 12: 71-75.
12. Shakil KA, Anis S, Alam M (2015) Dengue disease prediction using weka data mining tool. arXiv preprint arXiv:1502.05167.
13. Pérez MS, et.al. (2005) Adapting the weka data mining toolkit to a grid based environment. Advances in Web Intelligence Springer, Berlin, Heidelberg, 492-497.
14. Gislason PO, Benediktsson JA, Sveinsson JR (2004) Random forest classification of multisource remote sensing and geographic data. Geoscience and Remote Sensing Symposium 2004 IGARSS'04 Proceedings 2004 IEEE International Vol 2.
15. Keerthi SS, et.al. (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13: 637-649.