CLASSIFICATION OF DENGUE FEVER USING DECISION TREE

WAJEEHA FAROOQI¹, SADAF ALI² AND ABDUL WAHAB³
¹Department of Computer Science, The university of Lahore
  wajeeha.farooqi@cs.uol.edu.pk
²Department of Computer Science, The university of Lahore
  sadafkamboh66@gmail.com
³Department of Computer Science, The university of Lahore
  abdul.wahab@cs.uol.edu.pk

Received october 2013

ABSTRACT. Dengue fever is widespread disease in the tropical areas caused by bite of female Eddie mosquito. Pakistan has been victim of this rapidly growing disease since last few years. The world health organization identified the four types of dengue fever. The experts are facing the problem of misdiagnosis of dengue fever. The Tests needed for empirical Classification of Dengue Fever takes a lot of time and money especially in epidemic situation in a country with limited resources. Therefore, we have used data mining techniques for the efficient classification of the dengue fever Type. The decision tree learning algorithm has been used as a classification Model. We performed two experiments using decision tree. The first general experiment shows the accuracy of 99.44%. It prunes the attributes which classify the dengue fever on the basis of the values in the dataset. The Second experiment classifies the dengue fever on the basis of expert weighted attributes, which are used in the classification on the basis of Minimum Cost and resource availability. The accuracy of this Model is still high 98.62%. We compared the performance in term of Type II error. We found that the Type II error is very low in second experiment.

Keywords: Data mining; Supervised Machine Learning; Decision tree; Weighted attributes.

1. Introduction. Dengue is mosquito borne infection which has endangered 2.5 billion populations all over the world. It is transferred in human by bite of female Eddie mosquito. It is also known as Bone Breaking Fever presenting with symptoms of headache, retro orbital pain, joint pain, muscular pain and rash evidence [1]. The symptoms of dengue fever are too much fierce and cause rapid changes in patient. According to World Health Organization dengue is divided into two types one is Classical Dengue also called Dengue Fever and the other is Dengue hemorrhagic Fever (DHF). Dengue hemorrhagic Fever further have four types called DHF1, DHF2, DHF3 and DHF4 [2]. DHF is marked by the onset of fever which lasts for 2 to 7 days with number of symptoms like leakage of plasma, shock, weak pulse. In critical condition it is hard to distinguish Dengue Fever from Dengue Hemorrhagic Fever [3].

Dengue fever remained very serious epidemic problem in Punjab-Pakistan from last few years. The large numbers of cases were diagnosed since 2007 in Pakistan especially in Lahore [1]. First case of dengue in Pakistan was reported in 1994 in Karachi Pakistan. Dengue’s outbreak in 2011 was more critical than previous years causing 300 casualties and 14000 people were affected. In 2010 more than 21,204 reportedly infected and 170 were died majority of them from Punjab Pakistan. The Previous studies show 10 confirmed cases with loss of four lives in 2003 and 4000 confirmed cases in 2007 [4].
In this paper, interesting attributes are identified which classify the dengue fever into DF, DHF1, DHF2, and DHF3. The dataset has been collected from the different hospitals of Lahore- Pakistan. In order to mine most interesting attribute, data mining techniques are applied on the dengue dataset. We perform two experiments by using the decision tree algorithm to classify the dengue fever on the basis of values and weightage of attributes. We calculate the performance measures of two experiments. We compare the experiments result on the basis of Performance Measure and Type II error.

2. Related Work. Lot of work has been done on classification of dengue data by using different data mining techniques. F. Ibrahim, M.N. Taib, W.A.B.W. Abbas, C.C. Guan, S. Sulaiman [5] worked on classification of DF and DHF by using Artificial Neural Network and they predict the day before defervescence of Fever in DF and DHF patients. They generated very promising results with prediction error 10% and 90% prediction accuracy by using multilayer feed forward neural network. They used SPSS for data analysis and MATLAB for Artificial neural network.

The team of Lukas Tanner et al. [6] classified 1200 patient’s data using decision tree. First they classified that patient suffering from fever having Dengue fever or non-dengue fever. The accuracy of distinguishing DF and non DF was 84.7%. In second experiment they worked on severity level of dengue fever which is known as DHF. They also worked on severity prediction of DF patient going into DHF or DSS (Dengue Shock Syndrome) on basis of few attributes.

D. Thitiprayoonwongse, P. Suriyaphol, N. Soonthornphisaj [7] applied decision tree approach to classify dengue patients’ data from Srinagarindra hospital and Songkalagarind hospital using 48 attributes. They used Entropy equation for multiclass Problem. They preprocess data first by excluding noisy data and replace missing values by taking mean of same numeric attributes. They performed four experiments by dividing data into two datasets. In first experiment they found six significant attributes for classification. They found three rules for DF; two rule for DHF I, two rules for DHF II and one rule for DHF III. In second experiment they used second data set with four classes. They got nine significant attributes for classification. Rules they found for DF, DHF I, DHF II and DHF III were one, two, six and one respectively. Accuracy for first and second experiment was 97.6% and 96.6%. In third experiment they merged both data sets for complete knowledge by applying decision tree. They again found eight significant attributes and found three rules for DF, three rule for DHF I and one rule for DHF III. For day 0 experiment they labeled training set into five classes and low accuracy with value of 67.8% was found.

D. Thitiprayoonwongse, P. Suriyaphol, N. Soonthornphisaj [8] applied decision tree and fuzzy logic on dengue dataset to compare performance of algorithms. They found eight informative attributes. They derived three rules for DF; three for DHF I, three for DHF II and only one rule for DHF III. Overall accuracy of this approach is 96.7%. In order to compare 8 significant attribute were input in Fuzzy Logic approach of 1001 patients. Fuzzy logic approach gives accuracy of 97.3% for 0 day detection problem by using informative attributes found from decision tree and assigned them into fuzzy logic. It correctly identifies day 1 with 71.11% accuracy.

F. Ibrahim, T. Faisal, M. I. M. Salim, M. N. Taib [9] diagnosed and classified risk in dengue patients using bioelectrical impedance analysis (BIA) and artificial neural network (ANN). They studied blood tests of healthy people and the people suffering from dengue fever. Dengue patients were classified according to risk group and corresponding BIA parameter. Training and testing data for ANN comprises four attributes fever, reactance, gender and risk group. Prediction accuracy of system for diagnosis was 96.27%.

F. Ibrahim, M. I. Mohammad, S. N. Makhtar, J. Ibrahim [10] designed a rule based expert system using bioelectrical Impedance to diagnose risk of dengue infection. This system classified risks in three types called Lower risk, Higher risk and healthy group using values of Gender, Reactance, and day of fever. Dataset has values of two groups in which 209 are confirmed dengue cases and 223 records of healthy people. System was designed and programmed in VB 6.0 with classification accuracy 66.7%.

D. Thitiprayoonwongse, P. Suriyaphol, N. Soonthornphisaj [11] applied decision tree to classify dengue fever into DHF1, DHF2, DHF3 and DF with average accuracy of 96.50%. Then they performed experiments to classify Dengue fever and dengue Hemorrhagic fever by using Decision tree. This experiment gives accuracy of 96.00%. Then they compared their performance in TYPE 1 error according to World health Organization.
3. **Approach.** The decision tree is a supervised learning approach[12]. The objective is to create a model that predicts the value of a target variable based on values of input variables. The structure of decision tree consists of leaves and branches. The leaves show the target class. The branches of tree show the input variables. The branches also show the conjunction of input variables leads to the target class as leaves.

The Input of the decision tree is 1) the dataset having class or target class. 2) The values of the variables can be numeric or nominal data 3) informative attributes 4) the size of training data.

The concept of Decision tree algorithm is based on the informative attributes. The useless attributes must be avoided. The informative attributes are extracted by the splitting criteria of decision tree to reduce the entropy of the class. There are three main splitting criteria of data in decision tree. Information gain is the best criteria of splitting data[13].

Consider the entropy in (1). For the multiclass problem, entropy equation is defined as shown in (2).

$$E = - \frac{P}{P+N} \log_2 \frac{P}{P+N} - \frac{N}{P+N} \log_2 \frac{N}{P+N}$$  \hspace{1cm} (1)

Where, $P$ is the number of positive class and $N$ is the number of negative class in data set

$$E(D) = \sum_{i=1}^{c} -p_i \log p_i$$  \hspace{1cm} (2)

Note that $D$ is the training data $p_i$ is a ratio of class i compare with all data, and $c$ is the number of class.

4. **Methodology.** We have taken dengue dataset from different hospital of Lahore in Pakistan. We performed two experiments using RapidMiner[14]. We followed five steps as shown in Fig. 1, 1) Pre-processing 2) Weight attribute 3) Select attribute 4) Train Model 5) Test model

The decision tree is applied on experiment I and experiment II, after preprocessing. The first experiment classifies the dengue fever type by following three steps. 1) Pre-processing, 2) Train Model, 3) Test model. The second experiment also classifies the dengue fever type by following five steps where, 1) Pre-processing 2) Weight attribute 3) Select attribute 4) Train model 5) Test model. We calculated the performance measure by using confusion matrix.

4.1 **Dataset.** We collected data of dengue patients from different hospitals of Lahore, Pakistan. Dataset consist of 20 attributes, in which 5 attributes have Numerical values, 14 are categorical attributes and one is class attribute. The class attribute is the target class which has nominal values. There are four values of class DF, DHF1, DHF2, and DHF3. There are 376 instances in data. There are 101 instances of DF, 73 are DHF’s instances, 97 of DHF1 and 99 of DHF3. This dataset have values from patient’s history, examination, and laboratory test reports. Patient’s history includes Fever, Headache, Retro orbital pain, Muscle pain, Joint pain and rash evidence. Examinations include values of Pulse, Blood pressure, Haematocrit test, IgM, IgG, Ultrasound and X-ray.
TABLE I  Description of the Dengue attributes

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever</td>
<td>High grade temperature lasts for 2-7 days</td>
</tr>
<tr>
<td>Headache</td>
<td>Severe headache</td>
</tr>
<tr>
<td>Retro orbital pain</td>
<td>Red eyes or pain behind eye-ball</td>
</tr>
<tr>
<td>Muscle pain</td>
<td>Deep Muscular pain during illness/Myalgia</td>
</tr>
<tr>
<td>Joint Pain</td>
<td>Pain in all joints /Arthralgia</td>
</tr>
<tr>
<td>Rash</td>
<td>Rash evidence on skin</td>
</tr>
<tr>
<td>Pulse</td>
<td>Pulse can be weak or normal</td>
</tr>
<tr>
<td>B.P</td>
<td>Blood pressure may be weak or normal</td>
</tr>
<tr>
<td>Hemorrhagic</td>
<td>Manifestation small blotches of redness</td>
</tr>
<tr>
<td>Tourniquet Test</td>
<td>Tourniquet test absent or positive</td>
</tr>
<tr>
<td>WBC</td>
<td>Leucopenia/Normal</td>
</tr>
<tr>
<td>Platelette</td>
<td>Normal Platelet count/Thrombocytopenia</td>
</tr>
<tr>
<td>Heamatocrit Test</td>
<td>Blood test that measures the percentage of red blood cells</td>
</tr>
<tr>
<td>Rise in Hct</td>
<td>Rise value of % of Red blood cells</td>
</tr>
<tr>
<td>IgG</td>
<td>Measure the level of immunoglobulin G/</td>
</tr>
<tr>
<td>IgM</td>
<td>Positive/Negative</td>
</tr>
<tr>
<td>Chest X-ray</td>
<td>For pleural effusion</td>
</tr>
<tr>
<td>UltraSound</td>
<td>Evidence of Ascites/size of liver &amp; spleen</td>
</tr>
<tr>
<td>Bleeding</td>
<td>Spontaneous bleeding</td>
</tr>
<tr>
<td>Classification</td>
<td>DF, DHF1, DHF2, DHF3</td>
</tr>
</tbody>
</table>

4.2 Preprocessing. The preprocessing plays very important role in the quality of result. The preprocessing includes the cleaning of noisy, missing, inconsistent and useless data [15]. The noisy data is incorrect entry in dataset. Useless attributes have no significant role in the classification like Fever.

Our dataset had noisy data and useless attributes. Like the spelling Mistakes or different conventions followed in different data sources e.g. weak, week is entered in the pulse attribute. The DHF3 patient is diagnose with weak and week values. To avoid these results, we check the data and correct it. In our dataset fever and IgM have the same value in data. The fever attribute has yes value, and IgM is positive in all the type of dengue fever shown in fig. 2.

The entropy of fever and IgM attribute is high in the data. The useless attribute fever and IgM is removed by using useless attribute operator in rapid miner. The uncertainty of data is decreased and the information gain is increased to make better decision of dengue type fever.
4.3 Performance Measure. The Accuracy, sensitivity, specificity, error rate and type II error are calculated against each class DF, DHF1, DHF2, DHF3.

\[
\text{Accuracy} = \frac{TP + TN}{P + N} \tag{3}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{5}
\]

\[
\text{Error rate} = \frac{FP + FN}{P + N} \tag{6}
\]

Where the TP is the true positive, TN is the true negative. The P is the total number of true positive and total number of false positive; N is the total number of true negative plus the total number of false negative.

Type II error is the number of false negative instances in the data. It is very important in the medicine field. It identifies the dangerous misclassification of disease. In our study, the type II error will be, the DHF2 patient is diagnose as DF.

5. Experiment I. The DF patients are classified as DF. When the system predicts sensitivity as 100%, it means that entire patients are correctly classified. When the system predicts 100% specificity it means all the patients who are not DHF3 patient are not classified as DHF3. The error rate (4) is the number of instances which are incorrectly classified. The DHF3 patient is classified as DF. It is the dangerous miss classification as compared to the DHF1 patient as DF.

We load the dengue data in rapid miner. The IgM and fever attributes are removed from the dataset. We used cross validation with k-folds using stratified sampling. The k is set as 10. The k part of data is run as test data and other k-1 part run as train data [16]. We train and test the data on the model. The gain ratio is set for the splitting criteria of decision tree. The confidence of decision tree is 0.25.

5.1 Result. There is one rule for DF, DHF1, DHF2 and DHF3 in our first experiment. The rule for DF stated that “if Pulse is normal, Tourniquet test value is less 19.500 then patient is diagnosed as DF”. The rule for DHF1 stated that “If pulse is normal, Tourniquet test value is greater than 19.500 and there is no bleeding evidence then patient is diagnosed as DHF1”. Rule for DHF2 stated that “If pulse is normal, Tourniquet test value is greater than 19.500 and there is bleeding evidence then patient is diagnosed as DHF2”. Rule for DHF3 is that “if Pulse is weak then patient diagnosed as DHF3”. All rules are extracted with the value of Fever is yes and IgM is Positive every time in all instances. We assigned weight to those attributes which are important to conform the type of fever by analyzing doctor’s way of diagnosing Fever type. The experiment II base on the weighted attributes to compare performance with without weighted attributes in experiment I.
6. **Experiment II.** The attributes are weighted in descending order. According to the expert opinion, highest weighted attributes are platelet, x-ray, and ultrasound used to confirm the presence of DF and DHF types. Highest weighted attributes are 0.5 of laboratory tests. Platelet, bleeding, pulse has 4.0 weightage and rest attributes have 3.0. We select those attributes whose weights were greater than or equal to 4.0. Through these weighted attributes we can generate the decision tree which fulfill expert diagnose. The confidence of decision tree is 0.25. We test model and calculate the performance measure from the confusion matrix.

6.1 **Results.** In Experiment II with weighted attributes we found some interesting rules for DF, DHF1, DHF2 and DHF3. There are two rules for DHF1 according to first rule “If pulse is normal, Tourniquet test’s value is greater than 19.500, no bleeding found and platelet is greater than 40250 then patient is diagnosed as DHF1”. Second rule is stated that “If pulse is normal, Tourniquet test’s value is greater than 19.500, no bleeding
found, platelet is less than 40,250 and Haematocrit test’s value is less than 33.5 then patient is diagnosed as DHF 1”. The ratio of DHF 1 and other classes are 38:1. There are 38 instances found in DHF 1 and 1 instance is DHF 1. Two rules were identified for DHF 2 “If pulse is normal, Tourniquet test’s value is greater than 19,500, bleeding evidence found and platelet is less than 40,250 then patient is diagnosed as DHF 2”. Second rule for DHF 2 is defined as “If pulse is normal, Tourniquet test’s value is greater than 19,500, no bleeding found, platelet is less than 40,250 and Haematocrit test’s value is greater or equal to 33.5 then patient is diagnosed as DHF 2”. The ratio of DHF 2 and other class is 1:1. There is one rule found for DF which is “If pulse is normal and Tourniquet test is less than or equal to 19.500 than patient is diagnosed as DF”. One rule is found for DHF 3 “if pulse is weak then patient is diagnosed as DHF 3”. The value for Fever is yes and for IgM is Positive through all instances.

7. Discussion. The performance measure parameters are checked against both experiments. The DHF 3 has 100% accuracy, sensitivity and specificity. DHF 3 has 0 error rate in experiment I and experiment II. DF class has same accuracy, specificity and error rate in experiment I and experiment II like DHF 3. The DHF 1 and DHF 2 has different results in experiment I and experiment II, as shown in Table II and Table III. Accuracy for DHF 1 is 97.5 % in experiment I and 99.18% in experiment II Accuracy for DHF 2 in experiment I is 97.28% and 98.98% in experiment II. The error rate of DHF 1 in experiment I is 0.8 and it is 2.43% in experiment II. The DHF 2 has 1.08% error rate in experiment I and 2.71% in experiment II. The experiment I give more precise and accurate result. The experiment II has the interesting tree but accuracy is less than experiment I. The average error rate in experiment I is 1.33%. The average error rate in experiment II is 0.54%. Type II error is very important in medical type Dataset. The Type II error is the number of false negative instances in the data set. The average number of false negative instances in experiment I is 2.5. The average number of false negative instances in experiment II is 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
<td>99.7</td>
<td>100.0</td>
<td>99.6</td>
<td>0.2</td>
</tr>
<tr>
<td>DHF 1</td>
<td>97.5</td>
<td>91.78</td>
<td>98.98</td>
<td>2.43</td>
</tr>
<tr>
<td>DHF 2</td>
<td>97.28</td>
<td>95.91</td>
<td>97.78</td>
<td>2.71</td>
</tr>
<tr>
<td>DHF 3</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>98.62</td>
<td>96.9</td>
<td>99.09</td>
<td>1.33</td>
</tr>
</tbody>
</table>

8. Conclusion. This study classify the dengue fever with all interesting attributes and expert weighted attributes. This study compared the rules generated from experiment I and experiment II on the basis of accuracy, Sensitivity, Specificity, error rate and type II error. The accuracy, sensitivity, specificity and error
rate are 98.62, 96.9, 99.09 and 1.33 in experiment I. The experiment II has accuracy, sensitivity, specificity, and error rate 99.4, 98.6, 99.5, and 0.54. The average of false negative of experiment I is higher than experiment II. It means, performance of dengue fever classification through weighted attributes is higher. We can compromise on the accuracy, specificity, sensitivity if we have less average of Type II error. Our aim is to design expert system by implementing these rules.

Aknowlodge
A heart felt gratefulness is expressed by the authors of this paper to the patient and primary care position who participated in this study. We thanks Dr. Tooba Fatima and Dr. Hafiza Sidra Alifor their contribution in data accessing and entry of attributes.

REFERENCES