Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 9, September 2018, pg.13 – 17

Predicting Dengue Using Bayes Net Classifier

Anitha A[1], Freeda Jebamalar S[2]

[1]Department of Information Technology, Francis Xavier Engineering College, India

[2]Department of Information Technology, Francis Xavier Engineering College, India

[1] dr.aanitha@yahoo.com; [2] sfreeda26@gmail.com

Abstract— Dengue is a mosquito-borne fever in the southernmost part of India. It is caused by female mosquitoes, grown in stagnant water .The major symptoms for dengue are fever, bleeding, pain behind eyes, abdominal pain, fatigue, loss of appetite etc., Early diagnosis is the most important, in order to save the human from this deadly disease. Classification techniques helps to predict the disease at an early stage. In this research, Bayes belief network is classification technique is used to predict the probability for various disease occurrence using the probability distribution.

Keywords— Prediction, Diagnosis, Bayes belief network, Probability distribution, Classification

I. INTRODUCTION

Data mining is a process of finding new patterns, which is more beneficial and help us to understand the data efficiently and evolve the new patterns associated to the rules. It is used in many fields like Medical, education, telecommunication, etc.,.It is most important in medical field for predicting the disease in early stage. Because there are number. of case sheets available in paper-based format, which is difficult to handle huge dataset everywhere and anytime. It leads to the loss of history of patient's data which leads to the critical situation for human life .In order to avoid this situation classification techniques will be more helpful to predict the disease early. It can store large amount of data efficiently. The Quick diagnosis of disease also means quick recovery. Because every minute wasted in predicting the disease, the human's chance of survival is reduced. Especially in case of emergency and availability of doctor, the disease prediction can be a life saver.

Bayes belief Network is a probabilistic directed acyclic graphical model[1].It consists of set of variables and their conditional dependencies. It represents the probabilistic relationships between the diseases and the symptoms. Each node denotes the variable knows as observable quantities and the edges denotes the conditional dependencies. The edges which are not connected are conditionally independent.

II. LITERATURE SURVEY

In this paper the authors took the data from different hospitals of Lahore[2],[3].The dataset consists of 19 attributes and 654 instances (records).Out of 19 ,14 are categorical and 5 are numerical data. The fever attribute has value of 'yes' throughout the data. The data which are split into 10 fold cross validation. Out of 10 partitions,9 partition are given as the training data and the remaining one partition is used for testing purpose. The label attribute is classified as target class which hold the nominal values known as DHF(Dengue Hemorrahic fever and DF(Dengue fever).The dataset included the patient's history are fever, headache, retro orbital pain, muscle pain, joint pain, and rash, Examination values are blood pressure, haemraghic manifestation and tourniquet test, and lab test are WBC, Platelets count, hematocrit test, IgM, IgG, ultrasound and chest x-ray. Applying preprocessing technique to remove the noisy inconsistent, useless data. On the given dataset. Performance is measured based on accuracy, precision, sensitivity, specificity and false positive against each class DF and DHF. Sensitivity increases upto 14 order attributes and implies that NB identifies high true

positive and 14 attributes are correctly classified as DHF. The Multilayered perception algorithm have higher precision than all other classifier. Decision tree and the Multilayered perception have low false negative. The accuracy is defined as the ratio of the sum of true positive and true negative or correctly classified instances. Decision tree has accuracy level among the other classifiers.

In this paper the authors used the data from puskesmas pandanran semarang in excel format[3].The attributes includes patient's name, age, address, kelurahan, gender etc.,.It has 184 records. To perform the experiment by preprocessing the above given dataset was most useful for prediction. Out of 14 attributes,8 attributes are selected for prediction They achieved the accuracy of 77.17% level.

The data used in this study was collected from the paper entitled proteomic analyses of membranes enriched proteins of Leishmania donovani[4].The author of this paper took four features including Molecular mass ,PI function, and cellular localization for training for the identification of unknown protein as DT/VC [7].Visceral Leishmaniasis or Kala-azar is a fatal disease for human and it is caused by a protozoan parasite Leshmania donovani. Naïve Bayes,C4.5 ,Random tree and Support vector machine classifiers were applied on the membrane proteomics dataset and performed 10 fold cross validation was performed. Among the above algorithm Naïve Bayes outperformed in its accuracy level (76.17%) .It was trained with 28 known proteins and test with 37 unknown sample .The accuracy level was calculated using given formula

$$\text{Accuracy}=TP+TN/TP+TN+FP+FN \quad (1)$$

where,
TP-True Positive
TN-True Negative
FP-False Positive
FN-False Negative

In this paper the author took the dataset, which contains 18 attributes, and 108 instances[5].The attributes consists of PID(Patient ID),Date of fever(Month),Residence(city),Days(No.of days),current Temperature(Fever),WBC(No.of WBC),severe Headache(Yes or no),joint /Muscle pain(yes or no),Mettalic taste(yes or no),Appetite(yes or no),Abdominal pain(yes or no),Nausea/vomiting(yes or no),Diarrhea(Yes or no),Hemoglobin(Range),Hematocrit(Range)platelets(No.of platelets),Dengue(yes or no),with the help of Explorer Interface preprocess the data and filter the data.10 cross validation was performed. Naive bayes achieved the classification accuracy of 100% for 99 correctly classified instances. Mean Absolute Error is 0.0011 and ROC area is 1.J48 tree is used to predict the target value based on various attributes of dataset and achieved the 100% accuracy the mean absolute error obtained is 0,time taken to build is 0 seconds and ROC is 0.958.SVM (Support Vector Machine) is the algorithm used in this paper split the data on the given dataset resulted in 100% accuracy. Mean Absolute Error and time taken to build is 0 seconds and ROC area is 0.875.REP tree used to build a decision used to reduce error by sorted values of numeric attribute and split the instances into piece and achieve 74.74% are classified correctly for 74 instances with the Mean Absolute Error is 0.3655 and ROC is 0.544.For Random tree randomly choosing K attributes at each nodes allow the estimation of class probabilities with 87.87887% accuracy level. Experiment Interface is used to analyse through the algorithms such as Naïve bayes,J48,REP tree ,and Random tree to classify the dataset with the training and test set. Each algorithm is repeated for 10 times 'v' stands for best accuracy and '*' stands for worst accuracy. Naive bayes and j48 achieved the best accuracy with least error and Maximum ROC.

In this paper the author took real time hospital data. Inpatient department data included 31919 hospitalized patients with 20320848 records in detail from china[6].It consists of structured and unstructured data. This paper mainly focus on the prediction of cerebral infraction because it is a fatal disease. The input value is the attribute value of the patient.

$$X=(x_1, x_2, \ldots\ldots x_n) \quad (2)$$

It denotes the age,gender,the prevalence symptoms and living habits[Smoking or not].The output value can be consider as

$$C=\{c_0,c_1\} \quad (3)$$

There are three dataset. The S data(structured) is used to predict the high risk of cerebral infraction. T data is an unstructured is used to predict the high risk of cerebral infraction. S and T data is a multidimensionally fuse the both data that helps in prediction of high risk of cerebral infraction. Out of 706 patients records, 606 used for the training dataset and 100 dataset was used for testing. For T data, extracting 815073 words to learn word embedding. In S data, the three conventional machine learning algorithms(NB,KNN,DT) was used. T data used the CNN unimodal Disease risk prediction algorithm consists of 5 steps: The text can be represented in vector including n words. Then output of convolution layer as taken as a input of pooling layer for converting fixed Pooling layer is connected into fully connected neural network

$$h^3 = w^3 h^2 + b^3 \quad (4)$$

S and T data CNN-MDRP algorithm was used for training into two parts using ICTACIAS word segmentation tool word .vector dimensional is set to 50 after training get 52100 words in the word vector. For the structured data is applied with three algorithm NB,DT,KNN.NB is used to estimate the discrete feature attributes KNN is used the Euclidean distance with K=10 .DT has achieved 63% high accuracy. The recall of NB 0.86 for DT and KNN .By combining both S and T(CNN-MDPR) achieved 94.80% for the prediction than CNN-UDRP.

III. Proposed system

The proposed system includes three modules namely about data collection, bayes belief network, prediction.

A. Data Collection

Fever is the most common illness that strikes the human's immunity. Fever are diverse from a simple viral fever to the life costing dengue, pneumonia etc., These disease also tend to have a similar symptoms and differ in a very minute level. The dataset regarding dengue fever in TamilNadu. The dataset consists of 150 dengue cases with 28 attributes were collected and stored in excel file format for future analysis.

B. Bayesian Belief Network

It is a Probabilistic model or statistical model, which consists of set of variables and their Conditional dependencies through Directed Acyclic Graph. It can also used for representating the relationship between the disease and symptoms. For given Symptoms, it compute the probabilities of the presence of various diseases. It consists of observable quantities,Latent variables,edges and nodes. Edges denotes the conditional dependencies. There are two types of nodes. Conditionally independent ,which means they are independent to each other. If the node is dependent ,the associated to the probability function is taken as input and the Probability distribution as output. They are used in the field of modeling the sequence variables. It can solve the decision problems under certainty. This technique is also called as Bayesian Decision Networks. There are two ways of building network. One is to design the DAG manually with the expert available knowledge and with respect to corresponding probability distributions. The second way to analyse the given data through the machine learning .The application of bayes networks are Computational biology bioinformatics, data fusion, Information retrieval, etc.,. Feed the input file for disease prediction. Major symptoms for the disease are given as input to prediction module. Hill climbing approach is used in these technique for the purpose of evaluating the candidate subset of features iteratively. The Bayesian Belief Network denotes diseases with associated attributes as shown in fig.1

1) Bayesian Belief Network Figure:

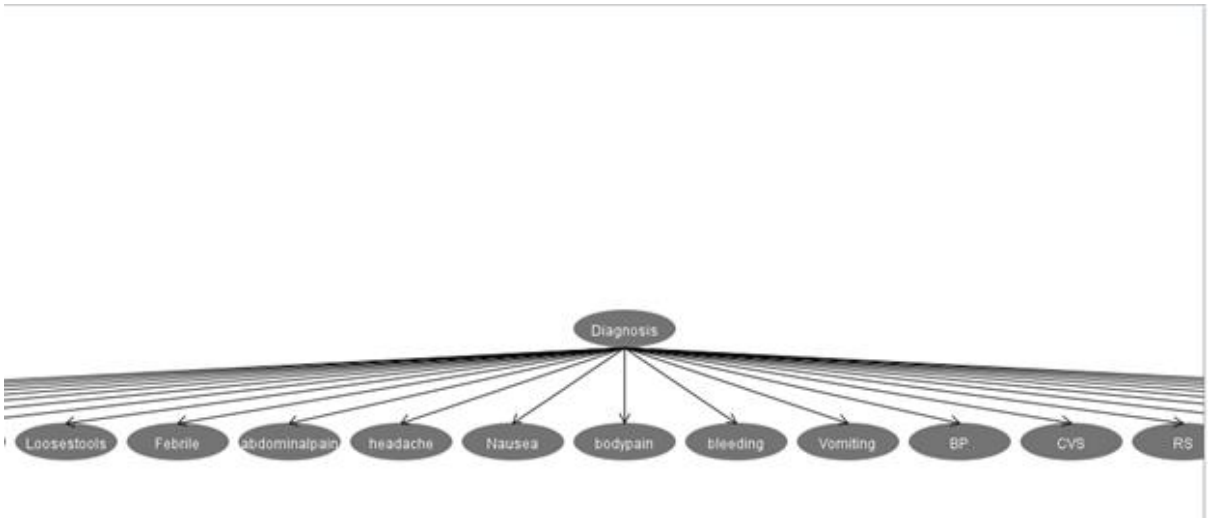

Fig. 1 Bayes belief network graph

C. Predictions

The probability distribution table is used for prediction of occurrence of disease with their possible probability for all given attributes. The probability distribution for some attributes are shown in Fig.2.

1) Probability Table for IgM:

Probability Distribution Table For IgM

Diagnosis	negative	positive	mildpositive	mildnegat...
Dengue	0.933	0.043	0.012	0.012
Dengueenterictever	0.625	0.125	0.125	0.125
Dengueoldseizuredisorder	0.5	0.167	0.167	0.167
Denguewithappendicitcs#EOAnemia	0.5	0.167	0.167	0.167
Denguewiththrombocytopenia	0.417	0.417	0.083	0.083
DengueNS1Positive	0.844	0.094	0.031	0.031
DenguewithNS1andIgMPositive	0.167	0.5	0.167	0.167
DenguewithNS1andI	0.5	0.167	0.167	0.167
PostDenguegastritis	0.5	0.167	0.167	0.167
Denguelikeillness	0.5	0.167	0.167	0.167
DenguelikeillnessNS1+ve	0.625	0.125	0.125	0.125
DengueShockSyndrome	0.5	0.167	0.167	0.167
DengueleverNS1	0.375	0.375	0.125	0.125
DenguewithThrombocytopenia	0.3	0.5	0.1	0.1
DenguewithThromb	0.5	0.167	0.167	0.167
DengueIgMPositive	0.167	0.5	0.167	0.167
DenguewithThrmbocytopenia	0.012	0.965	0.012	0.012

Fig. 2 Probability distribution for IgM

2) Accuracy Level:

Then applying the bayes net classifier to the given dataset and obtained the 83% accuracy level

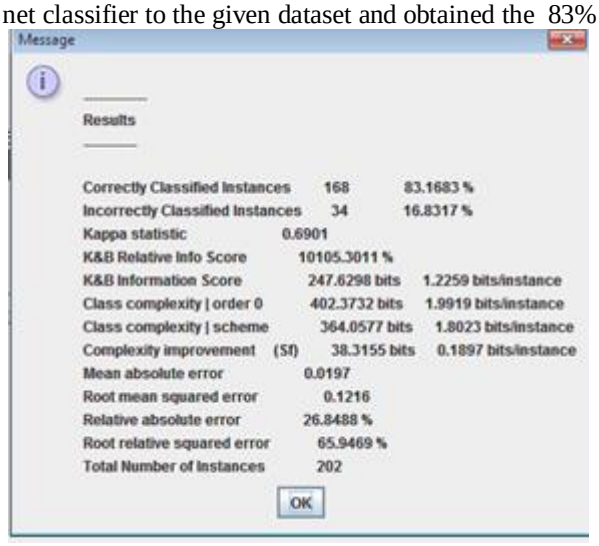

Fig. 3 Accuracy for Bayes Belief Network

IV. CONCLUSIONS

From the proposed system, it has been concluded that bayesian belief network classification technique with probability distribution table will helps to predicting the probability of various attributes with respect to the corresponding disease occurrence. In future, it is proposed to improve the accuracy level with more latest techniques.

REFERENCES

[1] C. Subrata, M. Kerrie, F. Colin, M. Lin, and L. David,"A bayesian Network-based customer satisfication model :a tool for management decision in railway transport,"Available:http://doi.org/10.1186/s40165-016-0021-2,Sep.2016.

[2] F. Wajeeha and A. Sadaf , "A Crictical study of selected classification algorithms for dengue fever and dengue haemraghic fever,"11th International Conference on Frontiers of Information Technology,IEEE,2013.

[3] D.J.Gubler,"Dengue and dengue haemorrhagic fever,"Clin.Microbiology Rev.,vol.11,pp.480-496,Jul.1998.

[4] B. Maharditya Restu ,"Dengue hemorrhagic fever (DHF) classification for patient in puskesmas using naïve bayes algorithm".

[5] S. Arvind Kumar, S. Pradeep ,P Anand, P. Dharm, D. Anuradha, and K. Awanish,"Putative drug and vaccine target identification in leishmania donovani membrane proteins using naïve bayes probalistic classifier",IEEE/ACMtran on computional biology and bioinformatics,vol.14,no.1,Jan/Feb.2017.

[6] Kashish Ara Shakil, Shadma Anis, and Manasaf Alam,"Dengue disease prediction using weka data mining tool".

[7] Min Chen , Yixue Hae, Kai Hwang, and Lu Wang,"Disease prediction by machine learning over big data from healthcare communities,"IEEE Access,2017.

[8] A. Kumar, P.Misra, B. Sisodha, A. K. Shasany, S. Sundar, and A. Dube, "Proteomic analyses of membrane enriched proteins of Leishmania donovani Indian clinical isolate by mass spectrometry," *Parasitol.Int.*, vol. 64,no. 4,pp. 36-42, 2015.

