Performance Evaluation of Tree Based Classifiers Using ZIKA Virus Dataset

J Uma Mahesh (&), Dr S Viswanadha Raju
Department of Computer Science & Engineering,
Geethanjali College of Engineering & Technology, Telangana,

Dr Siddhartha Ghosh, H.O.D, C.S.E,
Vidya Jyothi Institute of Technology

Dr S Viswanadha Raju, Professor, C.S.E
J.N.T.U Jagtial.

Abstract : *Data mining have been used in real time applications due to its artificial intelligence nature. Data mining is highly used in medical domain as it helps in making better predictions and supports in decision making. It also supports physicians in developing better diagnostic treatments. We have used Data mining to analyze Zika virus disease which leads to many deaths in South Africa & America. Zika virus is very fatal and spreads due to virus transmitted primarily by Aedes Mosquito. In this research work we have worked on tree based mining algorithms and further improvement is done by using filters which removes noise from the dataset. In this we worked on J48, decision tree, SVM & Random forest algorithms and indicate Experimental results.*

Keywords Data mining; classification; tree based classifier; WEKA, Zika virus dataset, Decision tree

1 Introduction

Data mining is the process of getting interesting and relevant information from huge data which is stored in repositories. Data mining process is used to discover, examine and mine useful data using various algorithms [3].

In healthcare organizations, data mining is highly used so as to extract useful information from patient's raw data which helps in intelli gent discussion making and helps the physicians in diagnosis of disease[1,12] likewise in this paper we have depicted the spread of Zika virus & its symptoms for the disease, generated few graphs for purpose of vizualisation & the statistics obtained serves the purpose of obtaining knowledge from the processed data by applying the concepts of Data miningData mining can be consequently branching off into sub processes that consist of selecting data preprocessing ,transformation, data mining and finally interpreting data. Data Mining is also known as Knowledge Discovery of Data (KDD)[8].

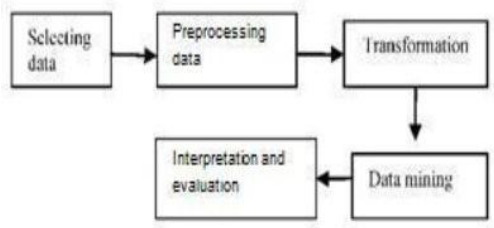


fig 1: Data mining process

Steps involved in the Data mining process can be depicted by

Steps followed in Data mining process:

Step 1: Data selection in this data which is required for our application domain is selected. Relevant data is retrieved from data repositories. Medical data can gathered from various health records of patients also it can be frequently obtained from various health care centers [7]

Step 2: After selecting the data, Data preprocessing is done in which ambiguous data is handled; data is converted into specific formats and deals with missing values [5]

Step 3: Transformation of data in which unstructured data is handled and data is converted into structured and in numeric form.

And data is mined using various functions & algorithms to extract hidden information.

Step 4: After mining results are evaluated, and visualized in form of graphs which helps in making better interpretations [6] In health care, mining is all about extracting and analyzing

The patient's conditions which helps in making assured predictions so as to raise the accuracy of diagnosis [9]

In medical domain, classification technique is widely used. Classification gives step by step guide to build a classifier model using on training data, and the model is tested using the test data and helps in making predictions [2]. In this we compare the classification of various tree based algorithms(SVM, Random forest and J48).

Classification is process examining the attributes of each instance and assigning it to one of a predefined class label [4] In this classifIcation is done based on the symptoms and other factors and predict either patient is died of Zika virus ,it has widely spread in areas around the rain forests of Central Africa. Zika virus has been the leading cause of death in South European countries. In 2014, there was 90% death rate due to this virus. There is great need of mining, as handling patients in these situations is very difficult and to get efficient results mining is very helpful. The virus transmits primarily through the Aedes mosquitoes infecting persons.

Vaccines for the cure of Zika virus are under progress but variety of blood, immunological and therapies are given to infected person. Also supportive cares with rehydration, symptoma t ic treatments are given [10,11] Data mining methods applied to the datasets to find out relations and patterns that are useful in understanding the evolution of disease [16] In this paper, we evaluated the performance measures using tree based classifIcation algorithms. The aim is to evaluate the performance of various classifiers and improvement is made by using unsupervised filters and further fusion of algorithm is done so as to make better prediction. In this research work we worked on machine learning WEKA tool we generated the decision tree.

fig 2: Causes for zika virus

2. RELATED WORKS

Rahman et al.[1]classified Zika affected patients into various categories according to their health conditions. And various decision tree models are developed and compared and predictions are made using the efficient decision tree model. Robu et a1[2] have analyzed medical data using various data mining algorithms on 4 differe nt datasets and improvements are made so as to make better predictions.

Datasets include:

1. Spread of Zika virus across countries of North America
2. Symptoms of Zika virus affected cases.

A min et al.[3]have discussed data mining and analyzed the Zika affected patients. In, this paper we have discussed and compared various data mining algorithms and performances of these algorithms are evaluated.Nookala et al. [8] used various classification t ion algorithms to predict Zika on basis of gene

expression data. And performance is evaluated when worked against 2 different Zika virus datasets. Bahramirad et al. [9] worked on real datasets using various classification algorithms and performance is evaluated on basis of various parameters

like accuracy, precision and recall and various techniques in order to improve the results.Zandi et al.[12] presented decision support framework in health care domain.Farid et al.[15] has] has proposed hybrid algorithm in order to enhance the accuracy of classifiers and worked on 10 real datasets. Mohamed et al. worked on various classification algorithms and performance is evaluated. Zhao et al. [16] has presented general data mining framework.

3. METHODOLOGY:

Step 1: Information gathering(ZIKA virus)

Zika virus disease (Zika) is a disease caused by Zika virus that is spread to people primarily through the bite of an infected *Aedes* species mosquito. The most common symptoms of Zika are fever, rash, joint pain, and conjunctivitis (red eyes). The illness is usually mild with symptoms lasting for several days to a week after being bitten by an infected mosquito. People usually don't get sick enough to go to the hospital, and they very rarely die of Zika. For this reason, many people might not realize they have been infected. Once a person has been infected, he or she is likely to be protected from future infections. Zika virus was first discovered in 1947 and is named after the Zika forest in Uganda. In 1952, the first human cases of Zika were detected and since then, outbreaks of Zika have been reported in tropical Africa, Southeast Asia, and the Pacific Islands. Zika outbreaks have probably occurred in many locations. Before 2007, at least 14 cases of Zika had been documented, although other cases were likely to have occurred and were not reported. Because the symptoms of Zika are similar to those of many other diseases, many cases may not have been recognized.

Symptoms:

Most people infected with Zika virus won't even know they have the disease because they won't have symptoms. The most common symptoms of Zika are fever, rash, joint pain, or conjunctivitis (red eyes). Other common symptoms include muscle pain and headache. The incubation period (the time from exposure to symptoms) for Zika virus disease is not known, but is likely to be a few days to a week.Areas with active mosquito-borne transmis s ion of Zika virus:Prior to 2015, Zika virus outbreaks occurred in areas of Africa, Southeast Asia, and the Pacific Islands.

In May 2015, the Pan American Health Organization (PAHO) issued an alert regarding the first confirmed Zika virus infections in Brazil.

Currently, outbreaks are occurring in many countries. Zika virus will continue to spread andit will be difficult to determine how and where the virus will spread over time.

Step 2: Framing of dataset (ZIKA virus)

Symptoms and countries of spread of ZIKA virus are taken as attributes and data is collected then designed data in excel sheet by retrieving as(.csv extension) pie charts, box plots are designed.

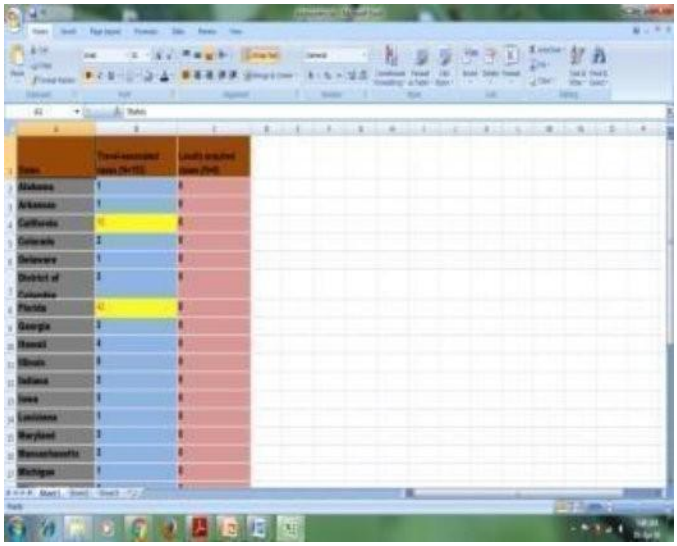


fig 3: Spread of "ZIKA" virus across countries of North USA

The above dataset comprises of data regarding spread of zika virus across various countries of North America that happens by 2 modes, they are:

1. Locally transmitted cases
2. Travel associated cases

Step 3: Applying algorithms on ZIKA dataset

Various algorithms like Random forest is been applied on the Zika virus dataset collected in this paper. The outcomes obtained are

:

Generated Confusion matrix

Time taken for generating the Random forest.

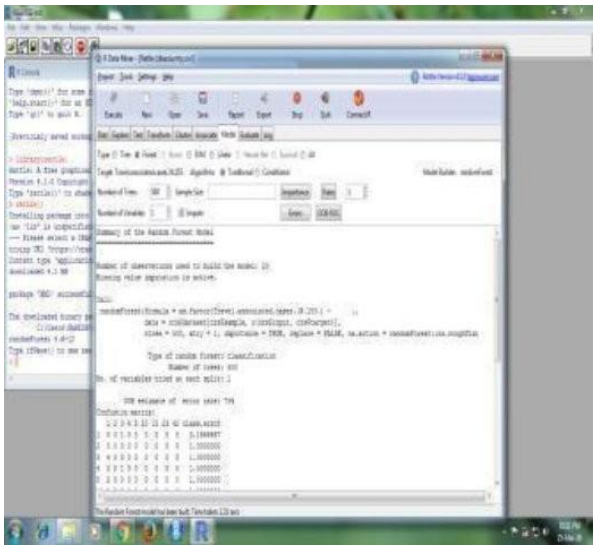


fig 4: Generated Confusion matrix

To create confusion matrix following steps are followed:

#create ZIKA dataset obs=c(sample(c(0,1),20,replace=TRUE),NA); obs = obs[order(obs)]
pred = runif(length(obs),0,1); pred = pred[order(pred)]
#calculate the confusion matrix
confusion.matrix(obs,pred,threshold=0.5)
where the parameters are described as:

obs a vector of observed values which must be 0 for absences and 1 for occurrences pred a vector of the same length as obs representing the predicted values. Values must be between 0 & 1 prepresentinga likelihood.

Value

Returns a confusion matrix (table) of class 'confusion.matrix' representing counts of true & false presences and absences

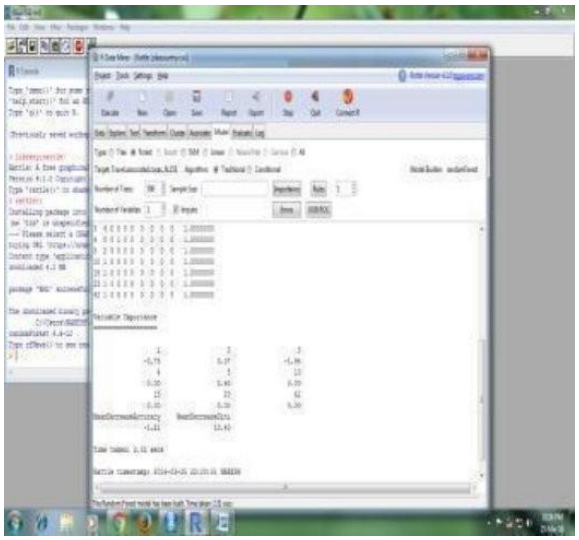


fig 4: Computing Confusion matrix

Confusion matrix with a 2x2 table with notation:

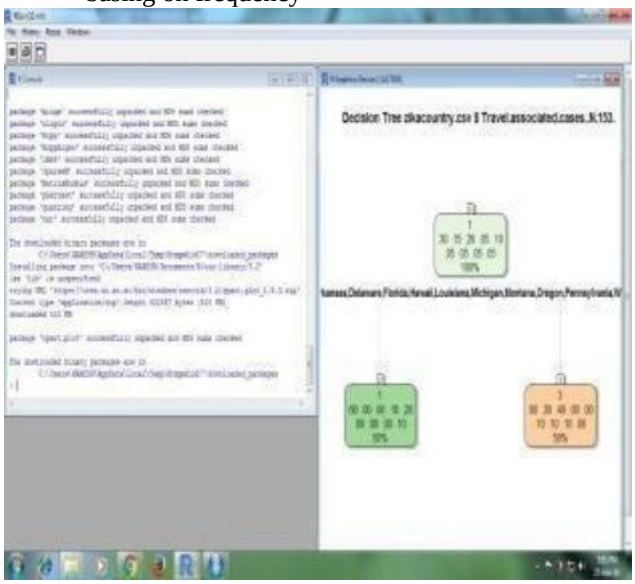
	Reference	
Predicted	Event	No Event
Event	A	B
No Event	C	D

Table no 1: events predicted from dataset

The formulas used to compute the confusion matrix are:

Sensitivity = A/(A+C) Specificity = D/(B+D) Prevalence = (A+C)/(A+B+C+D) PPV = (sensitivity * Prevalence)/((sensitivity*Prevalence) + ((1-specificity)*(1-Prevalence))) NPV = (specific ity* (1-Prevalence))/(((1-sensitivity)*Prevalence) + ((specificity)*(1-Prevalence))) Detection Rate = A/(A+B+C+D) Detection Prevalence = (A+B)/(A+B+C+D)

Balanced Accuracy = (Sensitivity+Specificity)/2 Step 4: Obtaining decision tree for ZIKA country travel associated cases basing on frequency



Creating, Validating and Pruning Decision Tree in R To create a decision tree in R, we need to make use of the functions rpart(), or tree(), party(), etc. rpart() package is used to create the tree. It allows us to grow the whole tree using all the attributes present in the data.

4. Conclusion

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce human readable classification rules and easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Zika Diagnosis. The experimental results show that SVM is the best algorithm for classification of Zika virus dataset. It is also observed that SVM performs well for classification on medical data sets of increased size

5. REFERENCES:

[I] Rahman, Rashedur M., and Fazle Rabbi

MdHasan. "Using and comparing differe nt decision tree classification techniques for mining ICDDR, B Hospital Surveilla nce data." Expert Systems with Applications 38, no. 9 (2011): 11421-11436.

[2] Robu, R., and C. Hora. "Medical data mining with extended WEKA."

Infntelligent Engineering Systems (lNES), 20i2 IEEE 16th International Conference on, pp. 347-350.IEEE, 2012. [3] Amin, Syed Umar, KavitaAgarwal, and Rizwan Beg. "Genetic neural network based data mining in

prediction of heart disease

using risk factors." Infnformation& Communication Technologies (iCT), 2013 iEEE Conference on, pp. 1227-1231. IEEE,2013.

[4] Yu, Hong, Xiaolei Huang, Xiaorong Hu, and HengwenCai. "A comparative study on data mining algorithms for individual credit risk evaluation." In Proceedings of the 2010 international Conference on Management of e - Commerce and e-Government, pp. 35-38. IEEE Computer Society, 2010.

[5] Xuexia, Dou. "Application of data mining algorithms in the analysis of financial distress early warning model of listed company. "In Computer Research and Development (ICCRD), 2011 3rd international Conference on, vol. 4, pp. 287-290.IEEE, 2011.

[6] Gupta, Shelly, Dharminder Kumar, and

Anand Sharma. "Performance analysis of various data mining classification techniques on healthcare data." International journal of computer science & Information Technolo gy (IJCSIT) 3, no. 4 (2011).

[7] Weitschek, Emanuel, Giovanni Felici, and Paola Bertolazzi. "Clinical Data Mining: Problems, Pitfalls

[8] Solutions." In DEXA Workshops, pp. 90-94. 2013

[9] Nookala, Gopala Krishna Murthy, Bharath Kumar Pottumuthu, NagarajuOrsu, and Suresh B. Mudunuri. "Performance analysis and evaluation of different data mining algorithms used for cancer classificatio n. "

[10] International Journal of Advanced Research in Artificial Intelligence(IJARAI) Vol.2,no.5,pp.49-55,(2013)

[11] http://www.who.intimediacentre/factsheets/ fsI03/en/

[12] Jarrett, Anna. "Ebola: A Practice Summary for Nurse Practitioners." The Journal for Nurse Practitioners 1\, no. 1 pp. 16-26,(2015).

[13] Zandi, Faramak. "A bi-Ievel interactive decision support framework to identify data mining-oriented electronic health record architectures."

Applied Soft Computing 18 (2014) pp: 136-145

[14] httos://github.com/mirador/ebola-dataireleases.

[15] Meng, Jianliang, and Yanyan Yang. "The application of improved decision tree algorithm in the electric power marketing." In World Automation Congress (WAC), 2012, pp. 1-4. IEEE, 2012.

[16] Farid, DewanMd, Li Zhang, ChowdhuryMofizurRahman, M. A. Hossain, and Rebecca Strachan. "Hybrid decision tree and naIve Bayes classifiers for multi-c la ss classification tasks." Expert Systems with Applications 41,vol. 4, pp: 1937-1946,2014.

[17] Zhao, Jitao, and Ting Wang. "A general framework for medical data mining." In Future Informat ion Technology and Management Engineering (FITME), 2010 International Conference on, vol. 2, pp. 163-165. IEEE, 2010.