www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Predictive models of missing data in multi-view dataset

C. Santhosh[1], S.Saranya[2], K. Harini[2], T.Dharshini[2]

[1]Assistant Professor, Department of Information Technology, Nandha Engineering

[2]UG Students, Department of Information Technology, Nandha Engineering College (Autonomous)

ABSTRACT

Dengue the global problem is common in more than 110 countries. Dengue fever is a vector borne disease caused by the female Aides Egypt and Aides Albopictus mosquitoes which adapt well to human environments. Dengue disease can cause severe damages to the society. Hence, it is critical to predict a dengue disease in advance to minimize the damage and loss caused by the disease. By keeping this voluminous data we can predict the future occurrences of the disease earlier and safe guard the people. The collected dataset was experimented with Weak and Net Beans IDE and preprocessed to remove missing values in multi-view dataset, feature selection is done and classification is done effectively with Support Vector Machine and Sequential minimal optimization applied in this research for predicting dengue disease.

Keywords: Data mining, Support Vector Machine, Feature extraction, Predictive Analysis

INTRODUCTION

With the increase of data modality in representing real-world objects, more and more multi-view data become available in medical diagnosis. These data have multiple views that generally correspond to distinct sets of feature representations for the same set of underlying objects. A provocation in learning from multi-view data is that not all data, which are stored in particular moment, fully represented in all views depicted as missing view data. The missing view issues in multi-view learning is not similar when compared with the missing data problem in single view learning, as the missing of a view results missing of all attributes in the same view.

More notably, since each view of multi-view data may contain some common and consistent information, multi-view learning can be employed to reduce the noise, as well as to learn the correlations between different views to obtain higher-level information. Nevertheless, missing view data are directly discarded in general, resulting in a severe loss of available information.

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. 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.

The disease transmission occurs when Aides Egypt 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.

In this work, we develop a set of methods and algorithms to address the above challenges. The dataset was loaded in Net Beans IDE, preprocessed to remove missing values in multi-view dataset, feature selection is done and classification is done

Author for correspondence:
Department of Information Technology, Nandha Engineering College (Autonomous)

effectively with Support Vector Machine and Sequential minimal optimization applied in this research for predicting dengue disease. The research result of SMO shows better prediction accuracy. [1-5]

EXISTING SYSTEM

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 user can rightly categorize patients because different patients require different types of treatment. For properly categorizing the dataset, different classification techniques are used. The algorithm C4.5 has classified only undesirable effect of changing a dengue patient's existing test data groups, potentially undoing the patient's own manual efforts in organizing the history. It involves a high computational cost have to repeat a large number of attribute test data group similarity computations for every new test data. C4.5 helps to extract dengue disease prediction suffer from scalability. It is imperative to address the scalability issue. Connections in dengue prediction are not homogeneous.

PROPOSED SYSTEM

Methods that can accurately predict dengue disease are greatly needed and good prediction techniques can help to predict dengue disease more accurately. In this system, it used feature selection method, to choose relevant features for improving the results of dengue disease prediction. The results show that feature selection is useful for improving the predictive accuracy and density is irrelevant feature in the dataset where the data had been identified on full field digital mammograms collected at the UCI Repository. In addition, support vector machine and sequential minimal optimization (SVM-SMO) were applied to solve the dengue disease diagnostic problem in an attempt to predict results with better performance. The results establish that ensemble classifiers are more accurate than a single classifier. The proposed SMO based on disease prediction is shown to be effective in addressing this prediction.

DATASET

The Dataset is a collection of data. Often, every dataset is depends on tables in database, where each column represents variable and row represents content. The database holds almost several entries but few of those random entries are used. The projected data set is derived from UCI machine learning repository.

ATTRIBUTES

The Attributes that we have chosen for the testing of dengue are fever, fever duration, body temperature, headache, eye pain, nausea, joint pain, vomiting, skin rashes, HB, WBC, RBC, platelets, body pains, swollen lymph and other indications. The attributes description is given in the following table. [6-10]

ATTRIBUTES	DESCRIPTION
Fever	Yes or No
Fever Duration	No. of days
Body Temperature	Temperature of body
Headache	Yes or No
Eye pain	Yes or No
Nausea	Yes or No
Joint pain	Yes or No
Vomiting	Yes or No
Skin Rashes	Yes or No
HB	HB count
WBC	WBC count

RBC	RBC count
Platelets	Platelets count
Body Pain	Yes or No
Swollen Lymph	Yes or No

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 algorithms are described below to predict, analyze and classify the future data streams.

SUPPORT VECTOR MACHINE

To find a feature of subset of size m, this contains the most informative features. The two well performing feature selection algorithms on the dataset are briefly outlined below.

Feature reduction results in converting the multidimensional space into lower dimensions. Feature extraction uses features construction, space dimensionality reduction, sparse representations, and feature selection where all are commonly used for early processing and pattern acceptance can be included.

Feature space techniques may reduce the predictable data in dataset which is used for classification and minimize the preprocessing cost. So performance of classifiers can be enhanced. SVM is a linear transformation with linear ortho normal basis vectors; it can be expressed by a translation and rotation.

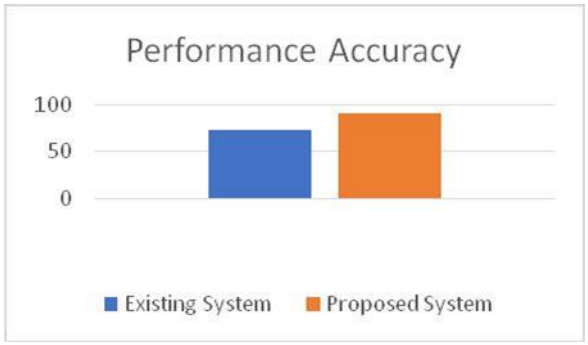
SEQUENTIAL MINIMAL OPTIMIZATION

Classification is the technique which is used to detect the infection, among patients and forecast that which technique shows top performance. Feature reduction results in converting the multidimensional space into 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 Feature space techniques may reduce the predictable data in dataset which is used for classification and minimize the preprocessing cost. The commonly used dimensionality reduction methods include supervised approaches such as Linear Discriminate Analysis (LDA), unsupervised ones such as SMO, and additional spectral and manifold learning methods. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. Consider the two dimensional cases then the basic principle of this transformation. [11-17]

RESULTS

This prediction provides the selection of minimum attributes by providing prediction and classification of data set with the support of Sequential Minimal Optimization and thus this produces better accuracy in the prediction of dengue disease



CONCLUSION

The infection rates of Aedes Aegypti mosquito's increases 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 week library and data mining techniques such as Sequential Minimal Optimization, decision tree and Support Vector Machine. This experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. This model helps to predict the dengue disease by reducing the features in the data set and classify them with better accuracy. Thus the predictive accuracy determined by SMO classification algorithm suggests that parameters used are reliable indicators to predict the presence of dengue diseases.

REFERENCES

[1]. Amin, MMM, Hussain, AMZ and Nahar K et al "Sero-diagnosis of Dengue Infections in Four Metropolitan Cities of Bangladesh", Dengue Bulletin, 24, 2000.

[2]. Aziz, KN Hasan and MA Hasanat et al. "Predominance of the DEN-3 genotype during the recent dengue outbreak in Bangladesh". Department of Immunology Immunology, Bangladesh Institute of Research and Rehabilitation in Diabetes, 33, 2002.

[3]. Blackburn G L, Wang K A ,"Dietary fat reduction and Dengue Disease outcome: results from the Women's Intervention Nutrition Study (WINS)", The American journal of clinical nutrition, PMID 18265482, 2007, 878-81.

[4]. Dana a. Focks, Eric daniels and Dan G. Haile, "A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results", Society Tropical Hygiene, 489-506.

[5]. Delen D, Walker G, "Predicting Dengue Disease survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, 34, 2005, 113127.

[6]. Derek A. T. Cummings, Rafael A. Irizarry and Norden E. Huang et al "Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand", Nature Publishing Group, 427, 2004.

[7]. Douglas M, and Watts,donald S, et al ," Effect of temperature on the vector efficiency of Aedes Aegypti for dengue 2 virus", The American Society of Tropical Medicine and Hygiene, 1987, 143152.

[8]. Eli Schwartz, Leisa H. Weld and Annelies Wilder-Smith, "Seasonality, Annual Trends, and Characteristics of Dengue among Ill Returned Travelers", Emerging Infectious Diseases, 14(7), 2008.

[9]. Glenn L. Sia Su ," Correlation of Climatic Factors and Dengue Incidence in Metro Manila, Philippines", Royal Swedish Academy of Science, 37(4), 2008.

[10]. Kanchana Nakhapakorn and Nitin Kumar Tripathi , "An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence",International Journal of Health Geographics, 2005, 1-13.

[11]. Mahbubur Rahman and Khalilur Rahman et al. "First outbreak of dengue hemorrhagic fever", Emerging Infectious Diseases, 8(7), 2002.

[12]. Michael A. Johansson, Francesca Dominici and Gregory E. Glass2, "Local and Global Effects of Climate on Dengue Transmission in Puerto Rico", 3, 2009.
[13]. Mohammad ali, Yukiko wagatsuma and Michael Emch et al, "Use of a geographic information system for defining spatial risk for dengue transmission in Bangladesh: Role for Aedes Albopictus in an urban outbreak", The American Society of Tropical Medicine and Hygiene, 2003, 634– 640.
[14]. Rico-Hessea R, and Harrisona LM et al, "Origins of dengue type 2 viruses associated with increased pathogenicity in the Americas", Department of Epidemiology and Public Health, 230, 244–251.
[15]. Rohani1, YC Wong1, I Zamre1, HL Lee1 and MN Zurainee, "The effect of Extrinsic incubation temperature on development of Dengue Serotype 2 And 4 Viruses In Aedes Aegypti" , Institute for Medical Research, 40(5).
[16]. Supawan promprou, Mullica Jaroensutasinee and Krisanadej Jaroensutasinee, "Impact of Climatic Factors on Dengue Haemorrhagic Fever Incidence in Southern Thailand", Walailak J Sci & Tech ,2005, 59-70.
[17]. Thomas W. Scott and Amy C. Morrison, "Aedes aegypti density and the risk of dengue-virus transmission", Department of Entomology, University of California, Davis, CA 95616 USA.