Volume-4, Issue-5, May-2015 • ISSN No 2277 - 8160

Research Paper | Technology

Statistical Methods over Data Mining on the Reemerging Dengue D2 Virus

A. Shameem Fathima — Research Fellow, Computer science and Engineering, Manonmaniam Sundaranar University (MSU),India

Dr.D.Manimegalai — Head of Department, Department of Information Technology, National Engineering College, India

ABSTRACT

Dengue is a life threatening disease prevalent in several developed as well as developing countries like India. This is a virus born disease caused by breeding of Aedes mosquito. Datasets that are available for this study describe information about the patients suffering with dengue disease and patients suffering from other febrile illnesses, though the symptoms seemed to be like dengue positive. Data mining is a well known technique used by health organizations for classification of diseases such as dengue, diabetes and cancer in bioinformatics research. Prior to the Data mining, a detailed statistical approach were applied to process the dengue infection data. In this paper we discuss about the various statistical tools such as Discriminant analysis, Independent samples t-test, Chi-square analysis using Statistical package for social sciences for dengue disease prediction.

KEYWORDS : Statistical, models, Discriminant analysis, Independent samples t-test, Chi-square analysis

INTRODUCTION

Data mining roots are traced back along three family lines which are Classical Statistics, Artificial Intelligence and Machine learning. The longest of three lines is classical statistics. [1] Without statistics there would be no data mining as statistics is the foundation of most technologies on which data mining is built.

Our research objective is to apply and evaluate the statistical tools such as Discriminant analysis, Independent samples t-test, Chi-square analysis using Statistical package for social sciences (SPSS v 16.0) on the Arboviral dataset under study. We are interested in the association between the independent variables and our categorical variable. Specifically, we would like to know how many dimensions we would need to express this relationship. Using this relationship, we can predict a classification based on the independent variables or assess how well the independent variables separate the categories in the classification. This paper explores the two key questions about statistical models developed to describe the recent past and future of vector-borne diseases, with special emphasis on dengue:

(1) How many variables should be used to make predictions about the future of vector-borne diseases?

(2) Which are the key predictor variables involved in determining the distributions of vector-borne diseases in the present and future?

The paper is organized as follows: a description including employed datasets and statistical techniques analyzed and the findings obtained after analysis is provided in Section 2. Results and discussions are presented in Section 3 and in Section 4, conclusions and directions for future work are given.

MATERIALS AND METHODS

This study aims to identify clinical and hematological features that could be useful to discriminate dengue from other febrile illnesses (OFI) from the day of admission up to the day of discharge. From the 5000 patient disease diagnosis, immunological data was available for dengue positive cases and patients suffering from other febrile illnesses, though the symptoms seemed to be like dengue positive. Using a classifier for analysis of all clinical, hematological .[3] and virology data we obtained classification. The validity, reliability and analysis of the data in this study was analysed using Discriminant analysis, Independent samples t-test and Chi-square analysis with Statistical package for social sciences (SPSSv 16.0).

In our study of interest, Twenty six independent variables factors/ symptoms causes for infection (Age, Gender, Fever, Chills, Coryza, Systolic pressure, Diastolic pressure, Shock, Myalgia, Malaise, Arthralgia, Hallucinations, Confusion, Altered consciousness, Unconscious, Convulsion, Neck rigidity, Motor weakness, Paralysis, Lymphadenopathy, Skin rash, Hemorrhagic symptoms, Pleural effusion, Hb, RBC and Platelet count at admission) are the scaled numeric variables and Infection is the categorical variable. The chi-square statistic compares the observed count in each table cell to the count which would be expected under the assumption of no association between the row and column classifications. To assess the association [4] between infection and platelet count at the time of admission, Chi-square test is performed to identify the association between infection and platelet count at the time of admission. The cross tabulation between infection and platelet count at the time of admission is presented in the table 1.

Null hypothesis H_0 *(*): There is no significant association between infection and platelet count at the time of admission

			Platelet count at the time of admission		Total	Chi-square value
			Abnormal	Normal		
Infection	Dengue	N	676	60	736	1901.32** (p < .001)
		%	13.6%	1.3%	14.9%	
	Viral	N	633	3580	4213	
		%	12.8%	72.3%	85.1	
Total		N	1309	3640	4949	
		%	26.4%	73.6%	100.0%	

** Significant at 1% level

TABLE 1- ASSOCIATION BETWEEN INFECTION AND PLATELET COUNT AT THE TIME OF ADMISSION

From the Table 1 , it is observed that there is significant association between infection and platelet count at the time of admission. Chi-square value (1901.32) shows that the null hypothesis is rejected at 1% level. Hence it is concluded from the analysis that infection and platelet count at the time of admission are well associated. From the table----- it is evident that most of patients (13.6%) affected by dengue are having abnormal platelet count at the time of admission. Similarly, chi square test is carried out to assess the association between infection and other Twenty six independent variables factors/ symptoms that seems to be the cause for infection.

Discriminant analysis is used to analyze relationships between a

non-metric dependent variable and metric or dichotomous independent variables.. The interest is to identify how many dimensions are needed to express this relationship. Using this relationship, prediction of classification based on the independent variables or assesses how well the independent variables separate the categories in the classification .Discriminant analysis works by creating a new variable called the discriminant function score which is used to predict to which group a case belongs. Discriminant function scores are computed similarly to factor scores, i.e. using eigenvalues. The computations find the coefficients for the independent variables that maximize the measure of distance between the groups defined by the dependent variable. The discriminant function is similar to a regression equation in which the independent variables are multiplied by coefficients and

	Wilks' Lambda	F	p-value
Platelet count at admission	0.908	4697.50**	< .001
Coryza	0.391	2851.54**	< .001
RBC	0.389	2192.01**	< .001
Myalgia	0.389	1702.05**	< .001
Gender	0.393	1398.23**	< .001
Heamorrhagic Symptoms	0.396	1210.65**	< .001
Pleural Effusion	0.387	1064.64**	< .001
Malaise	0.385	949.38**	< .001
Confusion	0.385	853.83**	< .001
Arthralgia	0.385	774.72**	< .001
Fever	0.384	708.48**	< .001
Neck rigidity	0.384	652.81**	< .001
Shock	0.384	605.44**	< .001
Chills	0.384	564.39**	< .001
Diastolic pressure	0.384	528.16**	< .001
Systolic pressure	0.383	495.94*	.020
Skin Rash	0.383	467.35*	.032
Convulsion	0.383	441.95*	.035

*Significant at 5% level ** Significant at 1% level

TABLE -2 TESTS OF EQUALITY OF GROUP MEANS

From the table 2, it is observed that Wilks's lambda of eighteen independent variables included in the discriminant function is less than 1. All the F-values are significant at 1% level except Systolic pressure, Skin Rash and Convulsion. This shows that all the eighteen factors shown in the above table significantly predicts and contributes discriminant function to separate the groups.

Function	Eigen value	% of Variance	Cumulative variance %	Canonical Correlation	Variance explained
1	1.614	100.0	100.0	0.786	0.618(61.8%)

TABLE -3 EIGEN VALUES

From table 3, The Eigen value is 1.614 which is larger and it shows that it explains more of the variance in the dependent variable in the discriminant function. Since the dependent variable Infection has two categories, only one discriminant function will exist. The canonical correlation 0.786 which is good is the measure of association between the discriminant function and the dependent variable. The square of canonical correlation coefficient 0.618 (61.8%) is the percentage of variance explained in the dependent variable Infection.

The standardized discriminant function coefficients in the table 4 -indicate the relative importance of the independent variables in predicting the dependent variable Infection. They allow us to compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. The Coefficient of Platelet count at admission is 1.141 which is greater among the coefficients of independent variables and it is discriminating the function more. The coefficient of is 0.042 which implies that Skin rash is discriminating the function less.

	Function
Gender	0.243
Fever	-0.080
Chills	0.095
Coryza	0.269
Systolic	0.045
Diastolic	-0.069
Shock	-0.088
Myalgia	0.216
Malaise	-0.123
Arthralgia	-0.106
Confusion	0.113
Convulsion	0.050
Neck rigidity	-0.074
Skin Rash	-0.042
Heamorrhagic Symptoms	0.317
Pleural Effusion	-0.167
RBC	-0.203
Platelet count at admission	1.141

TABLE -4 Standardized Canonical Discriminant Function Coefficients

The structure matrix shows the correlations of each independent variable with each discriminant function. The correlation between discriminant function and Platelet count at admission, Coryza, Myalgia, Malaise Chills Arthralgia Skin rash Convulsion Diastolic pressure, Neck rigidity, Gender, Fever, Pleural Effusion, RBC, Confusion, Systolic pressure, Heamorrhagic Symptoms and Shock are 0.767, 0.109, 0.098, -0.094, 0.080, -0.072, -0.072, 0.063, -0.062, 0.048, 0.046, -0.043, -0.038, -0.037, 0.030, 0.026, 0.020 and -0.014 respectively.

Table 5 contains the unstandardized discriminant function coefficients used to construct the actual prediction equation which can be used to classify new cases.:

	Function
Gender	0.494
Fever	-0.017
Chills	0.215
Coryza	0.634
Systolic pressure	0.003
Diastolic pressure	-0.006
Shock	-0.356
Myalgia	0.508
Malaise	-0.467
Arthralgia	-0.352
Confusion	0.864
Convulsion	0.140
Neck rigidity	-0.523
Skin rash	-0.321
Heamorrhagic Symptoms	0.896
Pleural Effusion	-0.510
RBC	-0.0000002
Platelet count at admission	0.000027
Constant	- 4.091

TABLE- 5 CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

The t-test is probably the most commonly used Statistical Data Analysis procedure for hypothesis testing. [4] The independent t-test, also called the two sample t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. Independent samples t-test was applied to ascertain the significant difference between infections towards HB, WBC, RBC, and Systolic pressure and Diastolic pressure.

.Table 6 shows the results of difference between infections towards HB, WBC, RBC, Systolic pressure and Diastolic pressure among patients.

	Infection	N	Mean	S D	t-value
Hb	Dengue	736	13.59	2.61	2.953** (p=.003)
	Viral	4213	13.82	1.86	
WBC	Dengue	736	4447.09	487.51	2.274* (p=.023)
	Viral	4213	4949.02	5983.13	
RBC (cells/cu.mm)	Dengue	736	4680475.54	355629.81	3.275** (p=.001)
	Viral	4213	4548923.09	1079327.63	
Systolic Pressure	Dengue	736	110.31	14.63	2.352* (p=.019)
	Viral	4213	111.64	14.01	
Diastolic Pressure	Dengue	736	82.32	14.93	5.537** (p<.001)
	Viral	4213	79.83	10.52	

** Significant at 1% level * Significant at 5% level

TABLE- 6 DIFFERENCE BETWEEN INFECTIONS TOWARDS HB, WBC, RBC, SYSTOLIC PRESSURE AND DIASTOLIC PRESSURE

RESULTS AND DISCUSSIONS

Finally, we summarize the most important results of our research experiment:

There is significant association between infection and type of gender, chills, coryza, myalgia, malaise, arthralgia, altered consciousness, unconscious, convulsion, neck rigidity, motor weakness, paralysis, lymphadenopathy, skin rash, pleural effusion,platelet count at the time of admission. There is no significant association between infection caused and age, shock, hallucination, confusion, hemorrhagic symptoms.

Discriminant function for the model should be as follows

D = - 4.091 + 0.494 (Gender) - 0.017 (Fever) + 0.215 (Chills) + 0.634 (Coryza) + 0.003 (Systolic pressure) - 0.006 (Diastolic pressure) - 0.356 (Shock) + 0.508 (Myalgia) - 0.467 (Malaise) - 0.352 (Arthralgia) + 0.864 (Confusion) + 0.140 (Convulsion) - 0.523 (Neck rigidity) - 0.321 (Skin rash) + 0.896 (Heamorrhagic Symptoms) - 0.510 (Pleural Effusion) - 0.0000002 (RBC) + 0.000027 (Platelet count at admission)

The mean Hb level (13.82) of the patients affected by the Viral fever is better than the mean Hb level of the patients affected by Dengue (13.59). This shows that lower level of Hb level of the patients affected by the Dengue is low as compared with the patients affected by viral fever, lesser the HB level is one of the cause for Dengue fever.

WBC level of the patients affected by the Dengue is low as compared with the patients affected by viral fever, lesser the WBC level is one of the cause for Dengue fever.RBC level of the patients affected by the Dengue is low as compared with the patients affected by viral fever, lesser the RBC level is not considered as one of the cause for Dengue fever.

CONCLUSION AND FUTURE WORK

From the findings of this paper ,it is concluded that Wilks' lambda value of 0.383 explains the greater discriminatory ability of the function. The associated chi-square statistic 4744.123 and it is significant at 1% level tests the hypothesis that the means of the functions listed are unequal across the groups which shows that discriminant function does better than chance at separating the groups. The key predictor variables involved in determining the distributions of vector-borne diseases are Platelet count at admission, Coryza , RBC , Myalgia , Gender, Hemorrhagic symptoms, Pleural effusion, Malaise, Confusion, Arthralgia, Fever, Neck rigidity, Shock, Chills, Diastolic pressure, Systolic pressure, Skin rash, Convulsion.

		Infection	Predicted Group Membership		Total
			Dengue	Viral	
Original	Count	Dengue	722	14	736
		Viral	328	3885	4213
	%	Dengue	98.1	1.9	100.0
		Viral	7.8	92.2	100.0

TABLE- 7 Classification Results of group size

Overall 93.1% of the original grouped cases are correctly classified. Table ----shows the predicted frequencies of groups from the analysis.

These results were validated by the doctors and microbiologists who provided us with a cosmic amount of viral data needed for our research study. The analysed approach was used with Arboviral data set but we plan to extend this approach in future for prediction of other diseases such as cancer etc. with other tools such as WEKA, R language and MATLAB.

REFERENCES

[1] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, Data Mining (New York: Addison Wesley, 1996). | [2] STATISTICAL METHODS, Arnaud Delorme, Swartz Center for Computational Neuroscience, INC, University of San Diego California, CA92093-0961, La Jolla, USA. | [3]Wikipedia,http://en.m.wikipedia.org/wiki/Dengue_fever, accessed in January 2015. | | [4] David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013. | | [5] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA, 2005. | | [6] Data Mining – Decision Tree Induction in SAS Enterprise Miner and SPSS Clementine – Comparative Analysis Zulma Ramirez 2901 N Juan St. Edinburg, TX 78541 (956)802-6283 | |

AUTHORS PROFILE | | A.Shameem Fathima is a research fellow student of Manonmanium Sundaranar University, India. She is pursuing her Ph.d with specialization in computerscience and engineering under the supervision of Dr.D.Manimegalai. She has published her research results in leading international conference proceedings and journals | | Dr.D.Manimegalai is currently the Head of Department of Information Technology, National engineering College,kovilpatti,India. She had her BE & ME from Government College of Technology, Coimbatore and PhD from Manonmaniam Sundaranar University, Tirunelveli .She is working in National Engineering College in various positions. She has modest number of research publications including journals such as AMSE and Pattern Recognition letter and in National and International Conferences. Her current area of research interests include Medical Image Processing and Data Mining and Image Retrieval. |