Application of Correlation & Regression Tree (CART) for management of Malaria in Arunachal Pradesh, India

U Murty, N Arora

Citation

U Murty, N Arora. *Application of Correlation & Regression Tree (CART) for management of Malaria in Arunachal Pradesh, India*. The Internet Journal of Tropical Medicine. 2007 Volume 5 Number 1.

Abstract

Malaria is a focal disease with multitudinous variations in its epidemiological pattern in relation to topographical features. The present paper demonstrates the application of CART (Classification & Regression Trees) for control of malaria in Arunachal Pradesh, India. Baseline epidemiological data of 12 districts of Arunachal Pradesh was employed for deriving prediction rules. The data was categorized into 2 different aspects, namely (1) Epidemiological (2) Meteorological. The intricate and complex interactions that exist between diverse input data sets, as they relate to the target features, are learned and modeled through exhaustive analysis. Predictor variables (maximum temperature, minimum temperature, rainfall, relative humidity, number of rainy days and month) were ranked by CART according to their influence on the target variable (MPI). Application of these easily conceptualized rules, rather than more abstract epidemiological principles, enables even non-specialists to gain an understanding of the malaria problem and in forecasting the malaria transmission dynamics to formulate the intervention strategies to combat malaria effectively.

INTRODUCTION

Malaria, the third leading cause of death attributable to an infectious disease worldwide, has plagued mankind for countless generations. The problem of Malaria is deeply entrenched in more than 90 countries of the world (WHO, 1998) and result in approximately 300 million acute illnesses and at least one million deaths annually (WHO, 1999). India being a tropical country is a malarial paradise with annual burden estimated to be nearly 2 to 2.5 million cases. North-Eastern region of India is in the Indo-Chinese hill zone of Macdonald's classification of stable malaria (MacDonald, 1957) and contributes nearly 9% of total malaria cases in India (Shiv Lal et al, 2000). In this region, perennial transmission of malaria slashes potential economic growth and thus is a major impediment to the overall development and progress of these areas. Despite several anti-malaria programmes, this region has seen little tangible progress in alleviating the burden of malaria (Mohapatra et al, 1998; Sen et al, 1994). Apparently, there are definite inadequacies that continue to dampen the spirit of public health specialists even since the halcyon days of malaria eradication. On closer scrutiny, operational difficulties stemming from the financial constraints and lack of definite knowledge about the malaria transmission trends are hampering the effective malaria control in the North-Eastern region (Mohapatra et al, 2003). Inaccessible areas owing to floods bear the maximum brunt of malaria. Main factors leading to failures in combating malaria in such regions are predominance of Plasmodium falciparum (Sharma et al, 1999, Dev et al, 2003), difficult terrain (Yadava & Sharma, 1995), favorable eco-climatic conditions (Sharma et al, 1996), lack of proper execution of control operations and wide communication gap between health researchers and policy makers. The problem of drug resistance (Kondrashin et al, 1987; Satyanarayana et al, 1991; Mohapatra et al, 2003), exophilic and exophagic vector behavior and high efficiency of vectors (Sharma et al, 1996, Dev et al, 2003) further aggravate the gravity of complex situation. Due to these various factors encountered in the North-Eastern region, malaria continues to present health services with an immensely difficult and complex challenge. Despite committed attempts for widespread implementation of conventional control methods of recognized efficacy, persistent malaria transmission in this region has underlined the need for alternate strategies to tackle the problem. The highly focal nature of malaria requires targeting of interventions to specific regions at appropriate time. Data mining applications have been successfully used in the past for the spatial clustering of endemic zones (Murty and Neelima Arora, 2007 a, b) prediction of disease outbreak (Kumar et al, 2005). CART is an acronym for Classification and Regression Trees, a

statistical procedure introduced by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone (1984). CART has been exploited in genomics (Michailidis, and Shedden, 2003; Garcia et al., 2002), proteomics (Clarke et al., 2003), microarray studies (Boulesteix et al., 2003), ecological (De'ath & Fabricius, 2000; Moisen & Frescino, 2002; Koh & Sodhi, 2004;), risk prediction (Gottschalk, et al., 1998; Paul and Munkvold, 1995), epidemiology (Chang et al., 1999) and social studies(Özge et al., 2004). The CART method has also been used to winnow phenotypes (Hermanek 1994; Masic, 1998), symptoms and prognoses for diagnostic characteristics with which to create decision trees to aid in medical diagnoses(Lopez et al.,1999) and therapeutics(Wolfe et al., 2003;.Smolle and Kahofer, 2001). In the current study, we describe association rules derived using Correlation & Regression tree for prediction of Malaria transmission. This approach will translate into most important attributable benefit of reduced disease burden by providing a better understanding of the variability of malaria to program managers and public health workers.

MATERIALS AND METHODS

STUDY AREA

Arunachal Pradesh is the largest state area-wise situated in the North-East region of India, sharing a long international border with Bhutan, China and Myanmar. This state is situated between latitude 26° 30' N and 29° 30 ' N and longitude 91° 30' E and 97° 30' E. The climate of the state is dominated by the Himalayan system and variations in altitude. The climate is highly hot and humid at the lower altitudes and in the valleys covered by swampy dense forest particularly in the eastern section, while it becomes exceedingly cold in the higher altitudes. Average temperature during the winter months ranges from 150 C to 210C and 220C to 300C during monsoon. Forested terrain and perennial streams are congenial for rapid multiplication and longevity of malaria vectors. Population of state is estimated to be 1091117 according to 2001 census. The state has a major population of 20 scheduled tribes and numerous sub-tribes. Agriculture is the primary driver of the economy. Nearly 80% of the population is engaged in. agriculture. The traditional method of agriculture is Jhumming, a kind of shifting cultivation. The main crops are rice, maize, millet, wheat and mustard.

12 districts of Arunachal Pradesh were randomly selected for the study. Epidemiological and meteorological data from 1999-2004 was collected from Directorate of Health, State Government of Arunachal Pradesh.

DATASET

A dataset consisting of meteorological and epidemiological parameters from 12 districts of Arunachal Pradesh with 12 attributes was used.

Monthly parasite incidence (MPI) expressed as positive blood smears for malaria/total population*1000 was considered as the malariometric index in this study.

DATA MINING TOOL

CART version 5.0 from Salford Systems, California, USA, was used for the current analysis (http://www.salford-systems.com). CART automatically searches for important patterns and relationship uncovering hidden structure even in highly complex data, which can then be used to generate highly accurate and reliable predictive models for various applications (Breiman et al, 1984). CART can be used to analyze either categorical (classification) or continuous data (regression) using the same technology. The CART methodology is technically known as “binary recursive partitioning” as a branching factor of 2 is used for the entire tree in a repeated manner until the tree terminates. In this process, each terminal node is assigned to a class outcome. CART contains sound statistical tool that enables the development of fast and accurate models. CART methodology is characterized by a reliable pruning strategy, automatic self validation procedures and its inherent simplicity.

METHODOLOGY

The steps used in the analyses are summarized as follow:

1. Preprocessing of Data: Conversion of *.xls to *.csv format

2. Variable selection: The data consists of several fields describing each attribute. The attributes include (1) Name of Primary Health Centre (PHC) (2) locality (3) district (4) state (5) country (6) month (7)maximum temperature, (8) minimum temperature, (9) total rainfall, (10) relative humidity, (11) Number of rainy days (12) Monthly Parasite Incidence(MPI). Seven of the twelve attributes were further used for developing association rules. These include (1) maximum temperature, (2) minimum temperature, (3) total rainfall, (4) relative humidity, (5) number of rainy days and (6) month. These attributes form the independent (predictor) variables. The dependent (predictive) variable is MPI. All variables except month are continuous; hence, regression

model was selected for this analysis.

3. Specification of the Tree type: Two tree types available in the CART version are classification and regression. Regression tree type was applied because of the predictive variable "MPI" is in continuous in nature in this study.

4. Splitting method selection: In choosing the best splitter, the program seeks to maximize the average "purity" of the two child nodes. A number of different measures of purity can be selected, loosely called "splitting criteria" or "splitting functions." 8 splitting methods were incorporated in the CART interface i.e GINI (default), Splitting GINI, Entropy, Class Probability, Twoing, Ordered Twoing, Least square and LAD. Since Least Square method is the preferred method for regression trees, it was selected for the generation of trees in this study.This approach resulted in generation of 22 trees with different relative error and complexity (Table 1). Out of the 22 trees generated, LS splitting model gave optimal tree with 23 nodes with the minimum Resubstitution relative error and minimum complexity (Table1).

5. Selection of testing criteria: As the target class was having more distinct values (i.e. 10) than the folds specified, V fold cross validation method was selected with a value of 10 for testing the data. Default parameters i.e. search intensity and threshold level for enabling intelligent search were selected at 200 and 15 respectively

RESULTS

CART generated 22 trees having different number of terminal nodes with different relative error (Table 1).

Figure 1

Table 1: Details of trees generated by CART

Tree Number	Terminal Nodes	Cross-Validated Relative Error	Resubstitution Relative Error	Complexity
1	23	0.57295 ± 0.06086	0.37070	0.00000
2	22	0.57256 ± 0.06083	0.37078	1.37283
3	21	0.57480 ± 0.06111	0.37126	7.91728
4	20	0.57917 ± 0.06191	0.37221	15.72451
5	19	0.58290 ± 0.06246	0.37478	42.59352
6	18	0.58370 ± 0.06246	0.37751	45.32642
7	17	0.58236 ± 0.06235	0.38051	49.69166
8	16	0.58760 ± 0.06283	0.38609	92.54084
9	14	0.58722 ± 0.06226	0.39905	107.45406
10	13	0.56743 ± 0.05412	0.40771	143.47861
11	12	0.58490 ± 0.05567	0.41922	190.86327
12	11	0.58115 ± 0.05498	0.43127	199.66040
13	10	0.59184 ± 0.05427	0.44518	230.58461
14	9	0.59959 ± 0.05369	0.45990	243.98184
15	8	0.60732 ± 0.05098	0.47712	285.52481
16	7	0.60538 ± 0.05135	0.49627	317.47940
17	6	0.66022 ± 0.05057	0.52255	435.60928
18	5	0.66841 ± 0.04877	0.55605	555.30310
19	4	0.67910 ± 0.04213	0.59896	711.37830
20	3	0.74772 ± 0.04251	0.65741	968.88464
21	2	0.73586 ± 0.04533	0.72333	1092.68298
22	1	1.00004 ± 0.00017	1.00000	4586.32275

The optimal tree obtained using LS method possessed 13 terminal nodes with a cross-validated error of 0.56743 ± 0.05412. A node is partitioned in such a way that left child node gets all cases with lower value of the splitting variable. Each decision rule is represented as a terminal node in the tree. The tree was further grown elevating each level at a time for comparison of rules and relative cost. The maximal grown tree showed 23 nodes with a relative cost of 0.567.

The decision rules (IF – THEN) used in this analysis are given in Table 2.

Figure 2

S. No	Rainfall	Relative humidity	Rainy days	Maximum temperature	Minimum temperature	Month	MPI
1.	<= 147.565	<= 89.305			<= 3.5	April, August , February January , July , June , March , May	0.257817
2.	<= 147.565	<= 89.305			> 3.5	April, August, February January, July, June, March, May	1.21218
3.	<= 147.565	> 89.305	<= 21			April, February, January, July, June, March, May, September	3.15821
4.	<= 147.565	> 89.305	> 21			April, February, January, July, June, March, May, September	1.94231
5.	<= 147.565	<= 68.335				December, November, October	4.33114
6.	<= 147.565	> 68.335 & <= 91.635				December, November, October	2.17394
7.	<= 147.565	> 91.635				December, November, October	8.63141
8.	> 147.565 & <= 533.55	<= 92.12 &&		<= 25.85	<= 15.5		0.626095
9.	> 147.565 & <= 533.55	<= 92.12		<= 25.85	> 15.5		2.69897
10.	> 147.565 & <= 533.55	<= 70.92		> 25.85			6.44491
11.	> 147.565 & <= 533.55	> 70.92 & <= 92.12		>25.85			3.90189
12.	> 147.565 & <= 253.95	> 92.12					5.15705
13.	> 253.95 &<= 385.8	> 92.12					9.80449
14.	> 385.8 &<= 533.55	> 92.12					14.0172
15.	> 533.55	<= 69.715		<= 34.5			21.0934
16.	> 533.55 & <= 69.715			> 34.5			13.7582
17.	> 533.55	> 69.715 & <= 87.75	22.5		22.28		11.6475
18.	> 533.55 & > 69.715 &&	> 69.715 & <= 87.75	<= 22.5		> 22.28		4.20769
19.	> 533.55	> 69.715 & <= 87.75	22.5 & <= 25.5				5.6125
20.	> 533.55	> 69.715 & <= 87.75	> 25.5				2.85438
21.	> 533.55	> 87.75		<= 37		April, December, February, January, May, November, October September	14.5115
22.	> 533.55	> 87.75		<= 37		August, July June	24.9901
23.	> 533.55	> 87.75		> 37			11.8499

Variable importance of different predictors is summarized in Table3

Figure 3

Table 3: Importance of Predictor variables

Parameters	Importance
Rainfall	100.00
Relative Humidity	64.92
Month	30.03
Maximum Temperature	21.68
Minimum Temperature	19.49
Rainy days	21.33

DISCUSSION

The disease transmission dynamics is modeled using the parameters such as vector (pathogen transmitting agent) surveillance, parasitic load in the human community and sudden environmental changes. We used data mining tools in CART to find relationships between epidemiological data and the Monthly Parasite Incidence (MPI). These relations are generally hidden in a large dataset. The rules in the CART system are used for the prediction of incidence of Malaria in an effective way.

These observed results could be used as predictive system and also used as a 'rules-of-thumb-guide' in controlling the transmission of Malaria in a more effective way. The interpretation of the rules is as follows:

For example:

Rule # 1. If RAINFALL= 147.565, RELATIVE HUMIDITY<=89.305, MINIMUM TEMPERATURE<=3.50 C & MONTH= APRIL, AUGUST , FEBRUARY JANUARY , JULY , JUNE , MARCH , MAY THEN MPI=0.257817.

Rule # 22. If RAINFALL >533.55, RELATIVE HUMIDITY>87.75 and MAXIMUM TEMPERATURE> 370 C & MONTH = JUNE, JULY, AUGUST, THEN MPI= 24.9901.

This is in accordance to the general trends that malaria follows, showing very high peaks of incidence in monsoon months. Lowest MPI is predicted to occur when the rainfall limits the availability of surface water required for mosquito breeding and cold temperature limits vector survival, hence showing negligible malaria incidence. While a very high MPI as reflected in rule #22 is expected when optimum temperature is accompanied with significant rainfall amount in Monsoon months.

Results indicate that the 6 predictors, namely, (1) Rainfall (2) Relative Humidity (3) Month,(4)Maximum temperature (5) Rainy Days (6) Minimum temperature influenced the target variable in descending order. This is helpful in ranking the predictor variables. Thus, decision trees play an important role in the management of vector -borne diseases. The above decision rules will be helpful in assessing the parasite incidence in the particular month in any locality. Hence, by observing above decision rules appropriate control measure can be implemented in an appropriate time to reduce the parasite load by the public health officials.

CONCLUSION

The decision rules obtained by employing CART 5.0 can be used as a prediction tool for any malaria endemic areas in India as well as abroad. The predictive model will have a vital role in estimating the parasite load in the ensuing seasons of study area. Hence, necessary precautionary measures can be undertaken for successful implementation of control strategies. Further, CART 5.0 could rank the predictor variables according to their level of influence on the target variable. From the present study it was observed that RAINFALL, RELATIVE HUMIDITY, MONTH, MAXIMUM TEMPERATURE, RAINY DAYS and MINIMUM TEMPERATURE were found to be influencing the target variable in the descending order. Therefore, it was concluded from this study that data mining tools like CART could be successfully employed for predicting the course of vector-borne diseases.

ACKNOWLEDGEMENT

Authors are grateful to Dr. J.S. Yadav, Director, IICT for his constant support and encouragement.

References

r-0. Boulesteix AL, Tutz, G and K. Strimmer. (2003). A CART-based approach to discover emerging patterns in microarray data. Bioinformatics. 19 (18): 2465-2472.
r-1. Breiman, L, Friedman J.H., Olshen R.A. and C.J. Stone. (1984).Classification and regression trees. Chapman and Hall, New York, 368 and Cardiovascular Interventions 55(3): 331 – 337.
r-2. Clarke W., Silverman B.C., Zhang Z., Chan D.W., Klein A.S. and E.P. Molmenti, (2003). Characterization of renal allograft rejection by urinary proteomic analysis, Annals of Surgery. 237: 660–664.

r-3. De'ath G. and KE Fabricius, (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology. 81: 3178– 3192.
r-4. Dev V, Bhattacharyya PC, Talukdar R. Transmission of Malaria and its Control in The Northeastern Region of India JAPI 2003; 51:1073-1076
r-5. Dev V, Hira CR, Rajkhowa Mk. Malaria-attributable morbidity in Assam, north-eastern India. Ann Trop.Med Parasitol 2001 ;95(8):789-96
r-6. Duvvuri Venkata Rama Satya Kumar, Kumarawsamy Sriram, Kadiri Madhusudhan Rao and Upadhyayula Suryanarayana Murty . Management of filariasis using prediction rules derived from data mining. Bioinformation 1(1): 8-11 (2005)
r-7. Garcia, D.K., Klimpel, K.R., Dhar, A.K.,& Hizer, S.E. (2002). RAPD markers as predictors of infectious hypodermal and hematopoietic necrosis virus (IHHNV) resistance in shrimp (Litopenaeus stylirostris). Genome.
r-8. Gottschalk, KW, Colbert, JJ, and DL Feicht, (1998). Tree mortality risk of oak due to gypsy moth. European Journal of Forest Pathology. 28(2): 121-132.
r-9. Hermanek, P. and I. Guggenmoos-Holzmann. (1994). Classification and regression trees (CART) for estimation of prognosis in patients with gastric carcinoma. J. Cancer Res. Clin. Oncol. 120:309–313.
r-10. http://www.salford-systems.com
r-11. Koh, LP. and NS Sodhi. (2005). Importance of reserves, fragments and parks for butterfly conservation in a tropical urban landscape. Ecological Applications. 14: 1695-1708.
r-12. Kondrashin AV, Rooney W and Singh N. Dynamics of P. falciparum ratio - An indication of malaria resistance or a result of control measures? Indian J Malariol 1987; 24: 89-94
r-13. López, J.A., Weilenman, C., Audran, R., Roggero, M.A., Bonelo, A., Tiercy, J.M., Spertini, F. and G.Corradin. (2001). A synthetic malaria vaccine elicits a potent CD8+ and CD4+ lymphocyte immune response in humans. Implication for vaccination strategies. Eur. J. Immunol. 31:1989–1998.
r-14. MacDonald G, The epidemiology and control of malaria. Oxford University Press, London (1957).
r-15. Masic, N., A. Gagro, S. Rabati, A. Sabioncello, G. Dai, B. Jaki and B. Vitale. (1998). Decision-tree approach to the immunophenotype-based prognosis of the B-cell chronic lymphocytic leukemia. Am. J. of Hematol. 59:143–148.
r-16. Michailidis G, and K Shedden (2003). The application of rule based methods to class prediction problems in genomics. Journal of Computational Biology; 10(5):689-98.
r-17. Mohapatra PK, Namchoom NS, Prakash A, et al. Therapeutic efficacy of anti-malarials in Plasmodium falciparum malaria in an Indo-Myanmar border area of Arunachal Pradesh. Indian J Med Res 2003 ; 118:71-6.
r-18. Mohapatra PK, Narain K, Prakash A, et al. Risk factors of malaria in the fringes of an evergreen monsoon forest of Arunachal Pradesh. J.Natl MedJ India. 2001; 14(3):139-142
r-19. Mohapatra PK, Prakash A, Bhattacharyya DR, et al. Malaria situation in northeastern region of India. ICMR Bull 1998;28: 21-30
r-20. Mohapatra PK, Prakash A, Taison K., et al. Evaluation of chloroquine (CQ) and sulphadoxine/ pyrimethamine (SP) therapy in uncomplicated falciparum malaria in Indo-Myanmar border areas. Tropical Medicine and International Health 2005; 10(5): 478-483
r-21. Moisen, GG. and TS. Frescino, (2002). Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling. 157: 202-225.
r-22. Özge C, Toros F, Bayramkaya E, Çamdeviren H and T Sasmaz (2006). Which sociodemographic factors are important on smoking behaviour of high school students? The contribution of classification and regression tree methodology in a broad epidemiological survey. Postgraduate Medical Journal. 82: 532-541.
r-23. Paul PA. and GP. Munkvold (2004)..A Model-Based Approach to Preplanting Risk Assessment for Gray Leaf Spot of Maize. Phytoptathology. 94 (12): 1350-1357.
r-24. Satyanarayana S, Sharma SK, Cheeleng PK et al. Chloroquine resistant P. falciparum malaria in Arunachal Pradesh. Indian J.Malariol 1991; 28(2):137-40.
r-25. Sen PK. Resurgence of malaria in eastern and north-eastern region of India: a critical appraisal.Indian J Public Health 1994; 38(4):155-8
r-26. Sharma VP. Re-emergence of malaria in India. Indian J.Med.Res.1996; 103:26-45
r-27. Shiv Lal, Sonal GS , Phukan PK. Status of Malaria in India. Indian Academy of Clinical.Medicine 2000; 5(1):19 -23
r-28. Smolle J and P. Kahofer. (2001) Automated detection of connective tissue by tissue counter analysis and classification and regression trees. Anal Cell Pathol. ;23 (3-4):153-8.
r-29. U. S. N. Murty, Neelima Arora.Application Of Self-Organizing Maps For Prioritization Of Malaria Control Operations In Changlang District, Arunachal Pradesh. The Internet Journal of Epidemiology. 2007. Volume 4 Number 2.
r-30. U.S.N. Murty, Neelima Arora. Prioritization of Malaria endemic zones in Arunachal Pradesh: A novel application of self organizing maps (SOM). The Internet Journal of Tropical Medicine. 2007. Volume 4 Number 1.
r-31. Wolfe F, Pincus T and J O'Dell. (2001). Evaluation and documentation of rheumatoid arthritis disease status in the clinic: which variables best predict change in therapy? J Rheumatol. 28:1712–7.
r-32. World Health Organization (WHO), Expert Committee on Malaria.WHO Expert Committee on Malaria, Twentieth Report. Geneva, WHO (1998).
r-33. World Health Organization(WHO). The World Health Report, 1999: WHO, Geneva. (1999).
r-34. Yadava RL, Sharma RS. Malaria problem and its control in north eastern states of India. J.Commun Dis. 1995; 27(4):262-6.

Author Information

U.S.N. Murty
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology

Neelima Arora
Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology