Chapter 1
Predictive Dynamics: Modeling for Virological Surveillance and Clinical Management of Dengue

V. Sree Hari Rao and M. Naresh Kumar

1 Introduction

Dengue fever (DF) is a mosquito-borne infectious disease caused by the viruses of the genus *Togaviridae* subgenus *Flavirus*. The disease has first appeared in the Phillipines in 1953, and from then on it has become the most important anthropod-borne viral disease due to its spread among humans (Monath 1994). The reemergence of this disease worldwide is causing larger, more frequent epidemics especially in cities and in the tropics. Dengue virus infection has been reported in more than 100 countries, with 2.5 billion people living in areas where dengue is endemic (CDC 2000; Guzman and Kouri 2002; PAHO 2007) (see Fig. 1.1). Dengue is one of the major international public health concerns of World Health Organization (WHO) because of the growing geographic distribution of virus and mosquito vectors, co-circulation of multiple virus serotypes and higher frequency of the epidemics.

The disease is caused by four distinct, but closely related viruse serotypes DEN1, DEN2, DEN3, and DEN4, which are transmitted to humans through the bites of infective female *Aedes* mosquitoes (Gubler 1998). A person who recovers from the infection due to one of the virus serotypes would have life long immunity against that serotype but he is susceptible to subsequent infection by the other three serotypes. There is strong evidence (De Paula and Fonseca 2004; Gubler 1998; Halstead 2007; Harris et al. 2000; Monath 1994; Nimmannitya 1997; Ooi et al. 2007; Wilder-Smith and Schwartz 2005) that subsequent infections would increase the risk of more acute

V. Sree Hari Rao (✉)
Department of Mathematics, Jawaharlal Nehru Technological University,
Hyderabad, Andhra Pradesh 500 085, India
e-mail: vshrao@gmail.com

M. Naresh Kumar
Software and Database Systems Group, National Remote Sensing Center (ISRO),
Hyderabad, Andhra Pradesh 500 625, India
e-mail: nareshkumar_m@nrsc.gov.in

V. Sree Hari Rao and R. Durvasula (eds.), *Dynamic Models of Infectious Diseases: Volume 1: Vector-Borne Diseases*, DOI 10.1007/978-1-4614-3961-5_1,

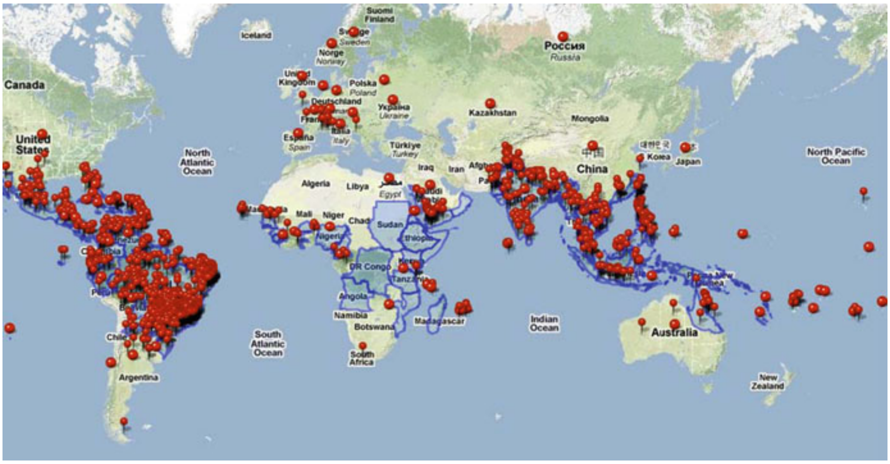

Fig. 1.1 Worldwide spread of dengue from 2007 to 2010 (CDC 2011)

forms of the disease known as dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) which could be fatal and may even lead to death. The annual occurrence is estimated to be around 100 million cases of DF and 250,000 cases of DHF. The mortality rate is around 25,000 per year (Gibbons 2002). The mortality rate is most common in children. The main pathophysiology of DHF and DSS is the development of plasma leakage from the capillaries, resulting in hemoconcentration, ascites, and pleural effusion that may lead to shock (Halstead 1998).

The clinical symptoms of dengue illness overlap with other illnesses (George and Lum 1997; Harris et al. 2000; Wilder-Smith and Schwartz 2005) causing a confounding problem in disease surveillance and management (Ooi et al. 2007). Definitive laboratory diagnosis requires isolation of the virus ribonucleic acid (RNA) by polymerase chain reaction (PCR) test, immunofluorescence, or immunohistochemistry (De Paula and Fonseca 2004; Halstead 1998; Vaughn et al. 2000). Further, the places where dengue is endemic may not have the necessary infrastructure to carry out these tests (Ooi et al. 2007). Thus, a scheme for a reliable clinical diagnosis based on the data would be useful for early recognition of dengue fever.

WHO (2009) has evolved a scheme for classifying dengue infection based on the symptoms of the disease (see Table 1.1). Halstead (Halstead 2007) reviewed the clinical diagnosis and pathophysiology of vascular permeability and coagulopathy, parenteral treatment of DHF/DSS, and suggested new laboratory tests.

Recent mathematical models both deterministic (Derouich et al. 2003; Esteva and Vargas 1998, 1999; Pongsumpun and Tang 2001) and stochastic (Grassly and Fraser 2008; Medeiros et al. 2011; Paula et al. 2003; Wearing and Rohani 2006) provide an insight into the dynamics of the dengue disease. In most of the studies the incidence rates and age structure play a vital role in understanding the transmission of the virus. The rate of spread of an infectious disease which is an important aspect for disease management is estimated using a neural network technology (Sree Hari Rao and Naresh Kumar 2010). Statistical analysis based on the χ^2 tests

Table 1.1 WHO characteristics of dengue fever

Dengue fever: Headache; retro-orbital pain; myalgia; arthralgia; rashes; hemorrhagic manifestations; leukopenia and supportive dengue fever serology or occurrence at the same location and time as other confirmed cases of dengue

Dengue hemorrhagic fever. (a) fever or history of acute fever, lasting 27 days, occasionally biphasic; (b) bleeding (hemorrhagic tendencies), evidenced by at least one of the following; a positive tourniquest test (TT); petechiae, ecchymosis, or purpura; bleeding from the mucosa; gastrointestinal tract; injection sites or other locations; hemotemesis or melena; thrombocytopenia (100,000 cells/mm^3 or less). (c) Evidence of plasma leakage due to increased vascular permeability, manifested by at least one of the following: a rise in the hematocrit equal or greater than 20% above average for age, sex and population; a drop in the hemotocrit following volume-replacement treatment equal to or greater than 20% of baseline; signs of plasma leakage such as pleural effusion; ascites, and hypoproteinemia

Dengue shock syndrome: Fever; hemorrhagic tendencies; thrombocytopenia, and plasma leakage must all be present plus evidence of circulatory failure manifested as: rapid and weak pulse; narrow pulse pressure (<20 mmHg) or hypotension for age (this is defined as systolic; pressure <80 mmHg for those less than 5 years of age, or <90 mmHg for those 5 years of age and older); cold clammy skin and restlessness

for discrete attributes, logistic regression and Mann–Whitney *U* test for continuous attributes are applied on the clinical data sets for classifying issues related to the diagnosis (Chadwick et al. 2006; Kalayanarooj et al. 1997; Ramos et al. 2009). Decision tree-based algorithms such as C4.5 have been used to differentiate dengue from non-dengue illness and predict the outcome of the disease. We have examined these issues critically and have established that our methodology yields more positive predictions when compared with those obtained by using C4.5 decision tree approach (Tanner et al. 2008).

Strategies to identify individuals likely to be in the early phase of dengue infection based on clinical features alone using the evidences or rules generated from the data would be of great help to the public health officials in prioritizing and directing patient stratification for clinical investigations and management. The authors have developed a new alternating decision tree (RNIADT for short) (Sree Hari Rao and Naresh Kumar 2011c) methodology which generates more accurate decisions rules as compared to the C4.5 decision tree (Tanner et al. 2008) and logistic regression (Chadwick et al. 2006; Ramos et al. 2009) for identifying the early clinical features that predict the diagnosis of dengue. Tanner et al. (2008) have applied C4.5 decision tree algorithm on acute febrile illness affected individuals using simple clinical and hematological parameters. Further, this study also requires laboratory features such as platelet count, crossover threshold value of a real-time PCR (RT-PCR) for dengue viral ribonucleic acid (RNA) and the presence of preexisting anti-dengue immunoglobulin G (IgG) antibodies. It is known that administration of these laboratory tests require 2–12 days (Sa-Ngasang et al. 2006; Vaughn et al. 1997) and in some cases the condition of the patient may not allow such a long wait. However, the research in Tanner et al. (2008) provides more insight into the scientific understanding of the disease prevalence among the infected individuals. From the effective clinical management point of view, it is desirable to have a methodology that helps one to identify the suspected dengue individuals from simple clinical features. This helps to reduce the spread of the disease in the community.

The main emphasis in this chapter is to present methods other than those followed conventionally by clinicians. The following are the principal objectives of the present study:

(a) To define the early clinical features of suspected dengue in children and adults which helps reduce the dengue virus transmission in a community
(b) To develop a new alternating decision tree methodology for predicting the diagnosis of dengue utilizing both clinical and laboratory features and to compare with other approaches based on statistical methods, logistic regression, and decision tree algorithms such as C4.5
(c) To examine the conformability of the WHO definitions of dengue fever on the realistic clinical and laboratory data
(d) To develop an accurate model which can predict the diagnosis of dengue based on clinical and laboratory features

In order to achieve this, we have used the data sets having 1,044 data records of dengue affected populations consisting of both children and adults from central and western States of India.

2 Dengue Virus Biology

The following details concerning the dengue virus and Dengue virus biology may be found in Net DV (2011). For the sake of brevity we present the following details (Net DV (2011)).

The size of the dengue virus is around 50 nm and is enveloped with a lipid membrane (Fig. 1.2). The total genome is approximately 10.6 kb in length. A short transmembrane segment attaches the viral membrane with 180 identical copies of the envelope (E) protein. The genome of the virus has about 11,000 bases that encode a single large polyprotein that is subsequently cleaved into several structural and nonstructural mature peptides. The polyprotein is divided into three structural proteins, C, prM, E; seven nonstructural proteins, $NS1$, $NS2a$, $NS2b$, $NS3$, $NS4a$, $NS4b$, $NS5$; and short noncoding regions on both the $5'$ and $3'$ ends (Fig. 1.3). The structural proteins are the capsid (C) protein, the envelope (E) glycoprotein and the membrane (M) protein, derived by furine-mediated cleavage from a prM precursor. The E glycoprotein is responsible for virion attachment to receptor and fusion of the virus envelope with the target cell membrane and bears the virus neutralization epitopes. In addition to the E glycoprotein, only one other viral protein, $NS1$, has been associated with a role in protective immunity. $NS3$ is a protease and a helicase, whereas $NS5$ is the RNA polymerase in charge of viral RNA replication.

2.1 *Life Cycle of Dengue Virus*

The life cycle of dengue virus involves endocytosis via a cell surface receptor (Fig. 1.4). The virus uncoats intracellularly via a specific process. In the infectious form of the virus, the envelope protein lays flat on the surface of the virus,

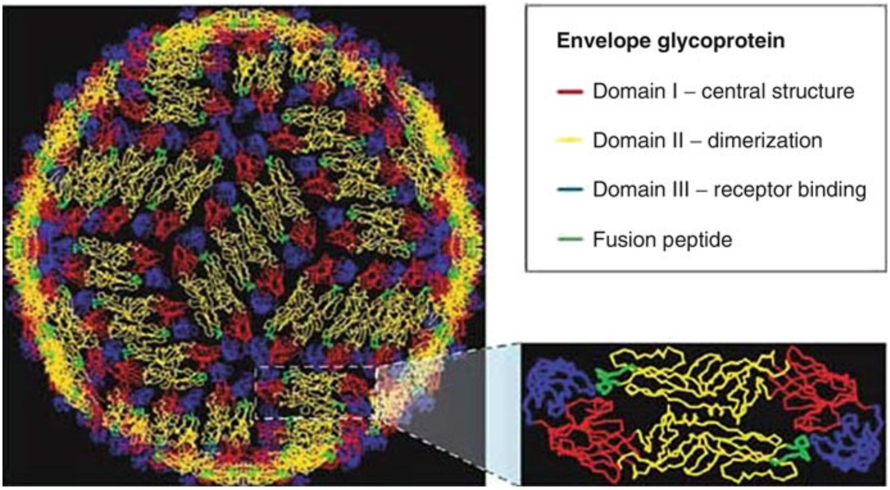

Fig. 1.2 Dengue virus particle (Stephen et al. 2007)

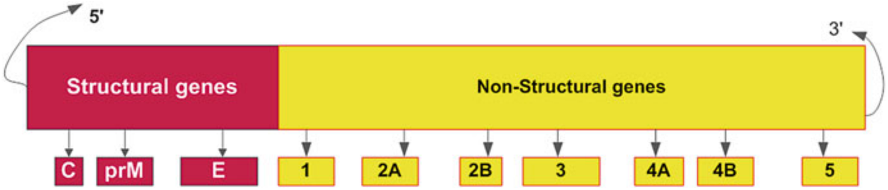

Fig. 1.3 Dengue virus genome

forming a smooth coat with icosahedral symmetry. However, when the virus is carried into the cell and into lysozomes, an acidic environment causes the protein to snap into a different shape, assembling into trimeric spike. Several hydrophobic amino acids at the tip of this spike inserts into the lysozomal membrane and causes the virus membrane to fuse with lysozome. This releases the RNA into the cell and infection starts.

The dengue virus (DENV) RNA genome in the infected cell is translated by the host ribosomes. The resulting polyprotein is subsequently cleaved by cellular and viral proteases at specific recognition sites. The viral nonstructural proteins use a negative-sense intermediate to replicate the positive-sense RNA genome, which then associates with the capsid protein and is packaged into individual virions. Replication of all positive-stranded RNA viruses occurs in close association with virus-induced intracellular membrane structures. DENV also induces such extensive rearrangements of intracellular membranes, called replication complex (RC). These RCs seem to contain viral proteins, viral RNA, and host cell factors. The subsequently formed immature virions are assembled by budding of newly formed nucleocapsids into the lumen of the endoplasmic reticulum (ER), thereby

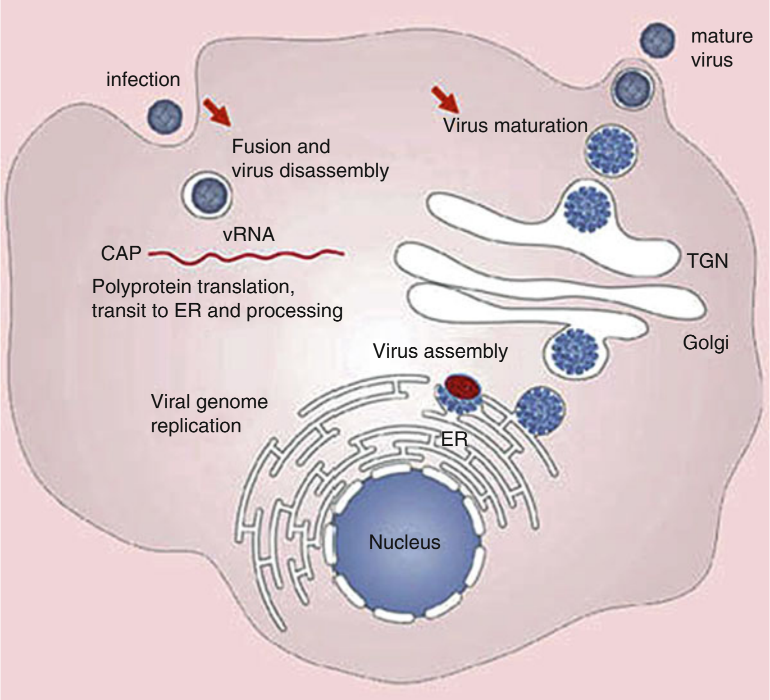

Fig. 1.4 Dengue virus life cycle (Net 2011)

acquiring a lipid bilayer envelope with the structural proteins *prM* and *E*. The virions mature during transport through the acidic trans-golgi network, where the *prM* proteins stabilize the *E* proteins to prevent conformational changes. Before release of the virions from the host cell, the maturation process is completed when *prM* is cleaved into a soluble *pr* peptides and virion-associated *M* by the cellular protease furin. Outside the cell, the virus particles encounter a neutral *pH*, which promotes dissociation of the *pr* peptides from the virus particles and generates mature, infectious virions. At this point the cycle repeats itself (Net DV, (2011).

3 Transmission of Dengue Virus

The dengue virus is transmitted mainly by the mosquitoes belonging to *Aedes* species. Among them the most prevalent species are *Aedes aegypti* and *Aedes albopictus*. In some of the regions in Pacific Islands and New Guinea *Aedes polynesiensis*, *Aedes scutellaris* and *Aedes pseudoscutallaris* transmit the disease. The *A. polynesiensis* in Society Islands and *Aedes niveus* in the Philippines are the other mosquitoes belonging to this species that transmit the virus (http://www.nathnac.org/pro/factsheets/dengue.htm). These mosquitoes prefer to breed close to human habitation where water-filled receptacles, small pools that collect in discarded human waste are found. They are active during the daylight hours and they feed throughout the day indoors and during overcast weather.

The *A. aegypti* being a holometabolous insect undergoes a complete metamorphosis with an egg, larvae, pupae, and adult stage in its life cycle. The life cycle of *A. aegypti* can be completed within one-and-a-half to 3 weeks. The environmental conditions play a crucial role in deciding the adult lifespan which may range anywhere from 2 weeks to a month.

The bites of the infective female *Aedes* mosquitoes transmit the disease to humans. The main source of virus for the uninfected mosquitoes is the infected humans. The virus is acquired by the mosquitoes while probing and feeding on the blood of an infected person. The infected mosquito is capable of spreading the disease after 8–10 days of incubation. During the incubation period the virus replicates within the mosquito's salivary gland. Once the mosquito acquires the infection it is capable of spreading the disease to the end of its life. The mosquito's eggs, however, can survive for as long as 1 year and at temperatures as low as 10°C (50°F). The mosquitoes transmit the disease to a susceptible human during probing and blood feeding. There is no definitive theory to say whether a particular mosquito carries the dengue virus or not. The infected female mosquitoes through the transovarial process may also transmit the virus to their offsprings, but the role of this in sustained transmission of the virus to humans has not yet been defined.

Clinical symptoms in humans indicate the circulation of the virus, and this condition would prevail approximately around 2–7 days.

4 Clinical Epidemiology

The clinical symptoms such as malaise and headache, followed by sudden onset of fever, intense backache and generalized pains, mainly in the orbital and periarticular areas are manifested within 6 days of infection (http://www.histopathology-india.net/Dengue.htm). There would be a recurrence of fever for a day or two (saddleback fever) after a nonfebrile interval of 24–48 h. During this time skin rashes and lymphadenopathy appear in the infected humans. There is a greater risk to persons who are previously exposed to this virus as an enhanced uptake of the virus into the host cells by the antiviral antibodies which may lead to disseminated intravascular coagulation and death due to shock (hemorrhagic dengue).

4.1 Pathological Features

Biopsy studies of the rashes reveal that in the cases of nonfatal dengue, lymphocytic vasculitis is found in the dermis whereas in the cases of fatal DHF the gross findings are petechial hemorrhages in the skin, hemorrhagic effusions in the pleural, pericardial, and abdominal cavities. In many organs hemorrhage and congestion are seen. Histopathological examinations reveal hemorrhage, perivascular edema, and focal necrosis but no evidence of vasculitis or endothelial lesions. It is observed that most of the morphologic abnormalities are due to disseminated intravascular coagulation and shock.

4.2 Serotypes

The dengue infection may spread due to any of the four known serotypes of the flavivirus. Based on the serotype of the virus spreading the infection, the dengue fever is termed DEN-1, DEN-2, DEN-3, and DEN-4. Even though the viral subtypes are closely related, they are antigenetically distinct. Therefore, a person already infected by one specific dengue serotype has lifelong homotypic immunity against a reinfection by the same serotype. In addition there will be a brief period of some partial heterotypic immunity but it does not provide permanent immunity or protection against the potential infection by any of the other serotypes. There is a possibility of having several serotypes circulating concurrently within an exposed population during epidemics. This is of vital importance in view of the fact that, dengue fever that produces some minor nonspecific viral symptoms, may also progress towards its more aggressive and often fatal form known as DHF.

Once a human being becomes infected by the bite of the *Aedes* mosquito, the incubation period is anywhere between 3 and 14 days (with an average lag time of 4–7 days), during which the viral replication takes place. The virus primarily targets the reticuloendothelial system, including dendritic cells, endothelial cells and hepatocytes (http://www.medicinemd.com/Med_articles/Dengue_fever_en.html). After 5–7 days of acute febrile illness, recovery is usually complete within 1–2 weeks.

4.3 Symptoms

The initial dengue infection may be asymptomatic and results in a nonspecific febrile illness, or it may produce complex manifestations of the classic dengue fever. A characteristic presentation of the symptoms includes sudden onset of fever, accompanied by severe frontal headaches, and joint (arthralgia), and muscle pains (myalgia). Some patients also experience nausea or vomiting and develop rashes on skin. The rashes would appear 3–5 days after the initial infection, and spreads from the torso to the extremities and the face.

Some patients, who have previously been infected by one of the dengue serotypes, may also develop bleeding and endothelial leakage upon infection with another dengue serotype. This syndrome is termed DHF. Subsequently, some patients with DHF may also develop shock (DSS), which is lethal and may lead to death of the infected person.

The symptoms of DHF and/or DSS are much more severe than in dengue fever, and usually occur within 3–7 days of the illness, coinciding with the time of decline or interruption of the phase of fever. The primary symptoms of DHF and DSS consist of plasma leakage and bleeding. The plasma leakage is caused by an increased capillary permeability, often resulting in hemoconcentration, pleural effusions, and ascites. Bleeding is caused by capillary fragility and thrombocytopenia (a marked decrease of platelets) which may result in bleeding incidents into the skin (petechial skin hemorrhages), or even life-threatening bleeding into the gastrointestinal tract.

The DHF or DSS symptoms appear only in patients who are earlier infected by one or more of the dengue serotypes. Typically, the basic dengue fever lasts for about 6–7 days, with a trailing end of the fever curve after a small peak (biphasic fever pattern). The patient's thrombocytes (platelets) keep dropping until the patient's temperature has returned to normal.

It is found that dengue clinical symptoms share a commonality with those of others illnesses such as malaria, typhoid fever, leptospirosis, West Nile virus infection, measles, rubella, acute human immunodeficiency (AIDS) virus conversion disease, viral hemorrhagic fevers, rickettsial diseases, early severe acute respiratory syndrome (SARS), and any other disease that can manifest in the acute phase as an undifferentiated febrile syndrome.

4.4 Diagnosis

A confirmed diagnosis is established by culture of the virus, PCR tests, or serologic assays. The diagnosis of DHF is made on the basis of the following symptoms and signs: hemorrhagic manifestations; a platelet count of less than 100,000 per mm^3; and an objective evidence of plasma leakage, shown either by fluctuation of packed cell volume (greater than 20% during the course of the illness) or by clinical signs of plasma leakage, such as pleural effusion, ascites, or hypoproteinemia. Hemorrhagic manifestations without capillary leakage do not constitute DHF. Additional laboratory criteria for a positive diagnosis include one or more of the following:

- Demonstration of a fourfold or more increase in reciprocal IgG or immunoglobulin M (IgM) antibody titers to one or more dengue virus serotype antigens
- Isolation of the dengue virus from serum, plasma, or leukocytes
- Demonstration of dengue virus antigens or viral genomic sequences, derived from autopsy tissues

4.5 WHO Guidelines for Diagnosis of Dengue

WHO in 1975 established the following guidelines for the diagnosis of dengue fever:

- Fever
- Hemorrhages positive tourniquet test, spontaneous bruising, mucosal bleeding, vomiting blood or bloody diarrhea

- Thrombocytopenia less than 100,000 platelets/mm
- Plasma leakage evident by a hematocrit level of more than 20% higher than expected, or a drop of the hematocrit level by 20% or more, following IV fluid therapy; hypoproteinemia, pleural effusion and ascites (collection of fluids in the thoracic cavity and/or abdominal cavity)

In addition to the symptoms of dengue fever, DSS is defined as including the following:

- A rapid and weak pulse
- A narrow pulse pressure (<20 mmHg)
- Hypotension
- An altered mental status
- Cool and clammy skin

Dengue fever being a viral disease, there is no direct therapy available. The treatment is usually limited to supportive care. To maintain an adequate blood pressure and to prevent dehydration oral and intravenous fluids are provided. Platelet transfusions are indicated, if the platelet count falls below 20,000 per μl (normal level: 200,000–400,000 per μl), or if significant episodes of bleeding occur. Blood in the stool (melena) may indicate gastrointestinal bleeding and requires platelet and/or red blood cell transfusions. To manage the febrile episodes, acetaminophen containing drugs are preferred over aspirin, nonsteroidal anti-inflammatory drugs (NSAIDs) or corticosteroids. Patients with DHF or DSS require close observation, including intravenous (IV) fluids, such as Ringer's lactate solution, starch, dextran 40 or albumin 5%, all of which may be of value to the patient. Blood transfusions to replace blood loss or fresh frozen plasma for patients with a coagulopathy may be necessary in individual cases.

For more details we refer our readers to URL http://www.medicinemd.com/Med_articles/Dengue_fever_en.html

5 Knowledge Extraction Methods

Our notations and terminology are fairly consistent and may be understood by referring to WHO (2009) and other earlier works. Standard definitions are used to compute the specificity, sensitivity, predictive positive value, predictive negative value, and area under the curve (AUC).

5.1 Missing Values: Concerns

The missing values in databases may arise due to various reasons such as value being lost (erased or deleted) or not recorded, incorrect measurements, equipment errors, or possibly due to an expert not attaching any importance to a particular procedure. The incomplete data can be identified by looking for null values in the data set. However, this is not always true, since missing values can appear in the form of outliers or even wrong data (i.e., out of boundaries) (Pearson 2005). Especially in medical databases, most data are collected as a by-product of patient care activities rather than from an organized research point of view (Cios and Mooree 2002). There are three main strategies for handling missing data situations. The first consists in eliminating incomplete observations, which has major limitations namely loss of substantial information, if many of the attributes have missing values in the data records (Kim and Curry 1977) and this renders introduction of biases in the data (Little and Rubin 1987). The second strategy is to treat the missing values during the data mining process of knowledge discovery and data mining (KDD) as envisaged in C4.5. The third method of handling missing values is through imputation, replacing each instance of the missing value with a probable or predicted value (Dixon 1979), which is most suitable for KDD applications, since the completed data can be used for any data mining activity.

There are numerous methods for predicting or approximating missing values. Single imputation strategies involve using the mean, median, or mode (Schafer 1997) or regression-based methods (Horton and Lipsitz 2001) to impute the missing values. Traditional approaches of handling missing values like complete case analysis, overall mean imputation and missing-indicator method (Heijden et al. 2006) can lead to biased estimates and may either reduce or exaggerate the statistical power. Each of these distortions can lead to invalid conclusions. Statistical methods of handling missing values consist of using maximum likelihood and expectation maximization algorithms (Allison 2002; Roderick and Donald 2002; Schafer 1997). Some of these methods would work only for certain types of attributes either nominal or numeric. Machine learning approaches like neural networks with genetic algorithms (Mussa and Tshilidzi 2006), neural networks with particle swarm optimization (Qiao et al. 2005) have been used to approximate the missing values. The use of neural networks comes with a greater cost in terms of computation and training. Methods like radial basis function networks, support vector machines, and principal component analysis have been utilized for estimating the missing values.

The wrapper algorithm (Sree Hari Rao and Naresh Kumar 2011c) presented in Appendix A checks for the presence of missing values, imputes them if they are present and then generates the decision tree. It follows from the above study that using a complete data set rather than an incomplete one results in better decision making in terms of identifying the right set of attributes that contribute to the diagnosis of the disease.

5.2 Statistical Procedures

The univariate statistical method such as χ^2 test is applied on the data sets to identify the patients with abnormal clinical findings with respect to the diagnosis of the disease. Logistic regression is used to develop a model for selecting the clinical attributes that influence the diagnosis. Those clinical attributes with $p < 0.2$ in the univariate statistic are included in the model with age and gender as potential confounders. The specificity, sensitivity, predictive value of both positives and negatives are computed using standard formulae to identify the clinical attributes that can distinguish dengue from other illnesses in children and adults. In addition to the above metrics a better measure known as area under the curve (AUC) score is being used in place of accuracies and error rate as it can represent the overall performance of a classifier (Huang and Ling 2005) in a robust manner. Based on the values (see Table 1.5) of the AUC one can categorize the performance of the classifier. The clinical attributes are selected either separately or in combination so as to have at least 70% positive and negative predictive values (Ramos et al. 2009). The statistical analysis is carried out using SPSS© software. The machine learning algorithms are developed using MATLAB© and Weka© softwares (Sree Hari Rao and Naresh Kumar 2011a, b, c, d).

5.3 What Are Decision Trees?

Decision trees are machine learning methods that can solve the problems of labeling or classifying data items out of a given finite set of classes using the features in the data items. Decision trees such as C4.5 (Quinlan 1993), classification and regression trees (CART), alternating decision trees (ADTree) (Freund and Mason 1999) have been used in computational biology, bioinformatics and clinical diagnosis (Middendorf 2004; Tanner et al. 2008; Wong et al. 2004). The C4.5 decision tree handles the missing values during the model induction phase of generating the tree.

Alternating decision trees are based on AdaBoost algorithm which generates rules based on the majority votes over simple weak rules (Freund and Mason 1999; Sree Hari Rao and Naresh Kumar 2011c). An alternating decision tree consists of decision nodes (splitter node) and prediction nodes which can be either an interior node or a leaf node. The tree generates a prediction node at the root and then alternates between decision nodes and further prediction nodes. Decision nodes specify a predicate condition and prediction nodes contain a single number denoting the predictive value. An instance can be classified by following all paths for which all decision nodes are true and summing the predictive value of the any prediction nodes that are traversed. A positive sum implies membership of one class and negative sum corresponds to the membership of the opposite class.

5.4 How to Generate and Interpret an Alternating Decision Tree?

To generate an alternating decision tree we apply the algorithm (see Appendix A) on the data set given in Table 1.2 specifically chosen for the purpose of demonstration. The data set has three attributes: Attribute1 $\in \{A, B, C\}$, Attribute2 $\in$ {True,

Table 1.2 An example data set for generating alternating decision tree

Attribute1	Attribute2	Decision
A	True	Class1
A	True	Class2
A	False	Class2
A	False	Class2
A	False	Class1
B	True	Class1
B	False	Class1
B	True	Class1
B	False	Class1
C	True	Class2
C	True	Class2
C	False	Class1
C	False	Class1
C	False	Class1

False}, and a decision attribute ∈ {Class1, Class2}. There are 14 instances out of which 9 belong to Class1 and 5 belong to Class2.

We designate Class1 as −1 and Class2 as +1. The initial sum of the weights with a precondition of the decision attribute being true is $W_+=5$ and $W_-=9$. The initial prediction value at the root node is computed as $a=\frac{1}{2}\ln\frac{5}{9}=-0.2954$. The weights associated with these instances are then updated (see Appendix A item 3 (iv)) as $w_{i,1}=e^{0.2954}=0.745$ for Class1 and $w_{i,1}=e^{-0.2954}=1.341$ for Class2. We identify a weak classifier having a rule Attribute1 = A. There are three instances in Class2 and two instances in Class1 with Attribute1 = A. Therefore, the prediction value $a=\frac{1}{2}\ln\frac{(3*1.341)+1}{(2*0.745)+1}=0.351$and $b=\frac{1}{2}\ln\frac{(2*1.341)+1}{(7*0.745)+1}=-0.2617.$ The weights are readjusted before the next boosting iteration. An alternating decision tree for the data set given in Table 1.2 is shown in Fig. 1.5. The root node indicates a predictive value of the decision tree before the splitting takes place. If the sum of all prediction values is positive then the instance belongs to the labeled Class1, otherwise it is placed in Class2. The prediction nodes are shown as ellipses and decision nodes as rectangles. The number in the ellipse indicates the boosting iteration. The dotted line connects the prediction nodes and the decision nodes, whereas a solid line connects the decision nodes with the prediction nodes. To classify an instance having attribute values Attribute1 = A and Attribute2 = true we first consider the root prediction value and based on the each instance value traverse the tree and add the prediction value of the particular node traversed. We derive the following sum by going down the appropriate path in the tree collecting all the prediction value encountered: –0.294 + (–0.2617) + (0.373) = –0.1827 indicating that the instance belongs to Class1.

The above methodology has been followed in Sree Hari Rao and Naresh Kumar (2011a, b, d) for identifying the early clinical features and assessment of laboratory features for dengue diagnosis and their results are presented in Sect. 6 of this chapter.

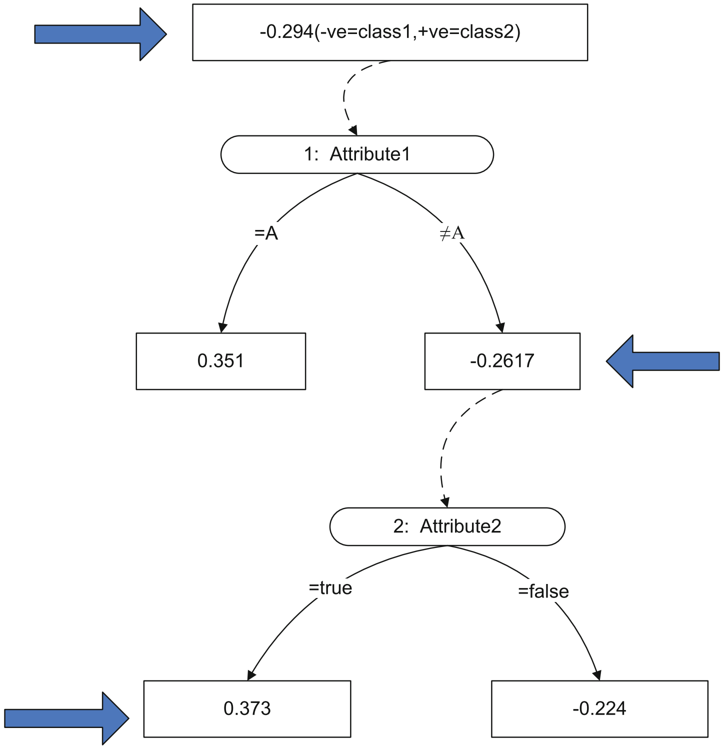

Fig. 1.5 An example alternating decision tree

5.5 *What are Influential Attributes?*

Decision making in databases is based on the attributes or features that form the data set. The set of attributes that contribute to better decision making are termed influential attributes. The presence of features that do not contribute much to the decision making degrades the performance accuracies of the supervised machine learning algorithms. The severity of this problem can be felt if one needs to search for patterns in large databases without considering the correlations between the attributes and the influence of such attributes on the decision attribute. The selection of influential features that maximizes the gain in the knowledge extracted from the data set is an important question in the field of machine learning, knowledge discovery, statistics and pattern recognition.

The machine learning algorithms including the top-down induction of decision trees such as classification and regression trees (CART), and C4.5 suffer from attributes that may not contribute much to decision making, thus affecting the performance of classifiers. A good choice of features would help reduce the dimensionality of the data set resulting in improved performance of the classifier in terms of accuracies and the size of the models, resulting in better understanding and interpretation.

5.6 *How to Extract the Influential Attributes?*

Feature selection is a popular technique to select the influential attributes as a subset of the original features. Feature selection is often used as a preprocessing step in the data mining activity. In situations presented by real world processes, influential features are often unknown a priori, hence features that are redundant or those that are weakly participating in decision making must be identified and appropriately handled.

Feature selection can be subdivided into filter-based methods and wrapper approaches. Wrapper subset evaluation models (Ron and George 1997) use the method of classification itself to measure the importance of the feature set. Wrapper methods generally result in better performance in terms of classification accuracies than filter methods because the features selected are optimized for the classification algorithm to be used. The wrapper approach (Kohavi and John 1998) defines a subset of solutions to a chosen data set and a particular induction algorithm, taking into account the inductive biases of the algorithm and its interaction with the training data set. The influential attribute selection procedure using wrapper subset evaluation is shown in Fig. 1.6. The point of concern with the wrapper method is its computational complexity as each feature set considered must be evaluated with the classification algorithm used (Dash and Liu 1997; Saeys et al. 2007).

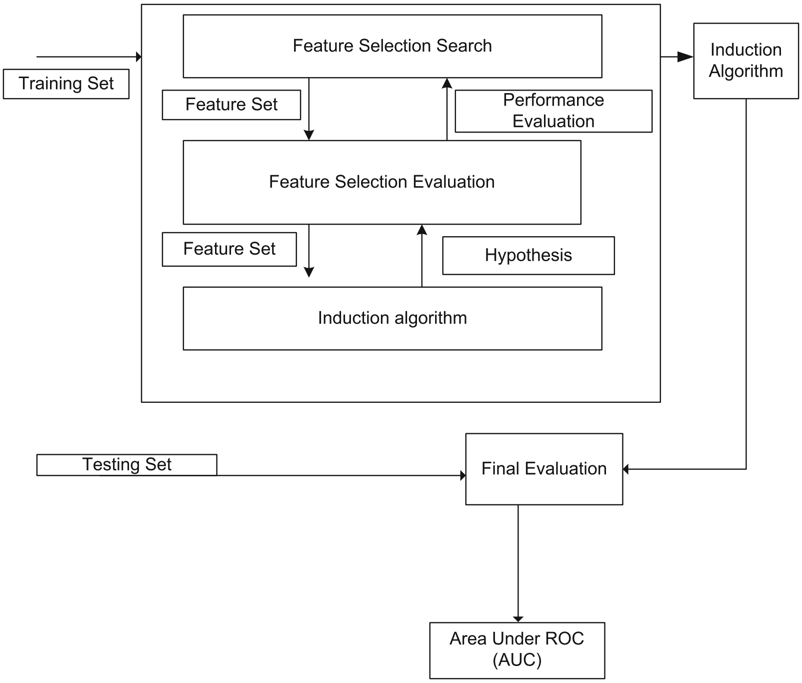

Fig. 1.6 Wrapper method of subset evaluation for selecting influencing attributes

5.7 *How to Identify Optimal Feature Subsets?*

5.7.1 Genetic Search

Genetic algorithms (GA) are stochastic optimization methods, inspired by the principle of natural selection. The search algorithms based on GA are capable of effectively exploring large search spaces (Goldberg 1989). GAs performs a global search as compared to many search algorithms, which perform a local or a greedy search.

A genetic algorithm is mainly composed of three operators: reproduction, crossover, and mutation. Reproduction selects good string; crossover combines good strings to try to generate better offsprings; mutation alters a string locally to attempt to create a better string. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the above GA operators and then reevaluated. This procedure is continued until the termination criterion is met. The default parameters for GA search (Sree Hari Rao and Naresh Kumar 2011a; Witten and Frank 2005) are given in Table 1.3. The results obtained by applying GA search (Sree Hari Rao and Naresh Kumar 2011a) for extracting influential clinical and laboratory features of dengue are discussed in Sect. 6.5 of this chapter.

5.7.2 Particle Swarm Optimization Search

The particle swarm optimization (PSO) is an evolutionary computation method which emulates the movements of flock of birds. The standard PSO consists of a randomly initialized population of size N known as particles. Each particle p_i can be viewed as a point in K dimensional space $p_i = (p_{i1}, p_{i2}, \ldots, p_{iK})$. The fitness values of the best positions of the particles at a previous time is given by $fi = (fi_1, fi_2, \ldots, fi_K)$. The index of the particle which has the best fitness value is designated as 'g_{best}'. The rate of change of position (velocity) for a particle i is represented by $V_i = (v_{i1}, v_{i2}, \ldots, v_{iK})$. The positions of the particles are updated using the following equations

$$x_{ij} = x_{ij} + v_{ij} \tag{1.1}$$

Table 1.3 Parameter values for genetic search

Attribute	Value
Start set	No attributes
Population size	20
Number of generations	20
Probability of crossover	0.6
Probability of mutation	0.033
Report frequency	20
Random number seed	1

Table 1.4 PSO search parameters

Attribute	Value
η_1	2.0
η_2	20
Max generations	50
Number of particles (*N*)	100

$$v_{ij} = w \times v_{ij} + \eta_1 \times \text{rand1}(\,) \times (f_{ij} - x_{ij}) + \eta_2 \times \text{rand2}(\,) \times (f_{gj} - x_{ij}) \tag{1.2}$$

where $j = 1, \ldots, K$, w is the inertia weight which is a positive linear function of time that changes according to the generation iteration. The parameters η_1 and η_2 represent the acceleration terms that pulls the particles towards p_{best} and g_{best}. The rand1( ) and rand2( ) are random number generation functions.

The velocities of the particles are limited by a maximum velocity $V_{\max}$. If $V_{\max}$ is too small then the particles may not explore beyond its locally good regions, i.e. they could be trapped in local optima. For the cases where $V_{\max}$ is too large the particles would fly past the good solutions.

A standard PSO search parameters are given in Table 1.4. The PSO search for extracting influential clinical and laboratory features of dengue has been utilized in Sree Hari Rao and Naresh Kumar (2011b) and their results are discussed in Sect. 6.5.

5.8 Does Descretization of Numeric Attributes Improve Decision Making?

Chadwick et al. (2006) have dichotomized all nominal laboratory features except WBC which was trichotomized to generate a user-friendly and accurate model.

5.8.1 Discretization Methods

Data discretization is the process of transforming quantitative attributes to qualitative attributes. Data attributes are either numeric or categorical. While categorical attributes are discrete, numerical attributes are either discrete or continuous. Discretization involves dividing an attribute values into a number of intervals ($\min_i \ldots \max_i$) so that each interval can be treated as one value of a discrete attribute. The choice of the intervals can be determined by a domain expert or with the help of an automatic procedure.

The discretization methods such as equal width and equal frequency discretization are unsupervised and have been used because of their simplicity and reasonable effectiveness. In equal width discretization (EWD) the attribute values are divided between $x_{\min}$ and $x_{\max}$ into k equal intervals such that each cut point is

$$x_{\min} + m * \left(\frac{(x_{\mathbf{max}} - x_{\min})}{k}\right)$$

where m takes on the values from $0,\ldots,(k-1)$. In equal frequency discretization (EFD) each subinterval in k between $x_{\min}$ and $x_{\max}$ has approximately the same number of sorted values of the attribute. Both EWD and EFD suffer from possible attribute loss on account of the predetermined value of k.

A proportional k-interval discretization (PKI) (Yang and Webb 2001, 2002) adjusts discretization bias and variance by tuning the number and size of the interval. This strategy seeks an appropriate trade-off between the bias and variance of the probability estimation by adjusting the number and size of intervals to the number of training instances.

The authors in Sree Hari Rao and Naresh Kumar (2011a, b) have implemented the PKI algorithm on a dengue data set to convert the nominal laboratory features to categorical and evaluated the accuracies of different classifiers. The results are discussed in Sect. 6.5 of this chapter.

5.9 Standard Classification Methods

Standard machine learning classifiers such as RBFNetworks (RBF) (Haykins 1994), Bayes Network (BNT) (Friedman et al. 1997), logistic regression (LOR), Naive Bayes (NIB) (George and Pat 1995), ADTree (ADT) (Freund and Mason 1999) and C4.5 (Quinlan 1993) have been utilized in Sree Hari Rao and Naresh Kumar (2011c) to benchmark the performances of RNIADT and its efficacy in extracting knowledge from dengue data set.

5.10 Performance Metrics for Comparing Machine Classifiers

To evaluate the models generated by the decision trees, we employed a k-fold cross validation algorithm ($k=10$) as it is considered a powerful methodology to overcome data over-fitting (Kothari and Dong 2000). The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all k trials is computed. To compare and evaluate the decision trees popular performance measures such as sensitivity, specificity, receiver operator characteristics (ROC), and area under ROC (AUC) (Crichton 2002; Metz 1978) have been employed. The definitions of the above measures are discussed briefly for the benefit of the readers. The classification task generates a set of rules which can be used for classifying individuals to different classes/groups. This may result in the following situations:

1. False positive (FP): the rules may predict the diagnosis of the patient as positive (presence of the disease) whereas the actual diagnosis is negative (absence of the disease).
2. False negative (FN): the rules may predict the diagnosis of the patient as negative (absence of the disease) whereas the actual diagnosis is positive (presence of the disease).

3. True positive (TP): when the prediction of the classifier matches with the actual diagnosis as positive.
4. True negative (TN): when the prediction of the classifier matches with the actual diagnosis as negative.

Based on the above situations the performance of the classifiers can be compared using the following standard measures:

(a) Sensitivity: the proportion of the people who are predicted as positive of all the people who are actually positive TP/(TP + FN).
(b) Specificity: the proportion of the people who are predicted as negative of all the people who are actually negative TN/(TN + FP).
(c) Positive predictive value: the proportion of the people whose predictions matches with the actual diagnosis as positives TP/(TP + FP).
(d) Negative predictive value: the proportion of the people whose predictions matches with the actual diagnosis as negatives TN/(TN + FN).

A theoretical, optimal prediction can achieve 100% sensitivity (i.e., predict all people from the sick group as sick) and 100% specificity (i.e., not predict anyone from the healthy group as sick).

ROC is a plot between (1 − specificity) on *x*-axis and sensitivity on *y*-axis. The AUC is a measure of overall performance of the algorithm. The accuracy of the decision tree algorithms can be evaluated using the AUC measure as given in Table 1.5.

The trade-off between the sensitivity and specificity is better captured by an ROC curve, which shows how sensitivity and specificity of a model vary with some tunable parameter, is related in a direct and natural way to cost/benefit analysis (Pepe 2003; Zweig and Campbell 1993) of diagnostic decision making. ROC curves allow one to distinguish among different models, depending on what model characteristics we need, and to determine which parameter values will give us the best performance for a given application.

By measuring the area under the ROC curve (AUC) (Hanley and McNeil 1982; Liu and Wu 2003) one can obtain the accuracy of the test. The larger the area, the better the diagnostic test is. If the area is 1.0, we have an ideal test because test achieves 100% sensitivity and 100% specificity. If the area is 0.5, we have a test

Table 1.5 AUC-based classification for assessing accuracy of the test results

Range	Class
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.7 < AUC < 0.8	Worthless
0.6 < AUC < 0.7	Not good
0.5 < AUC < 0.6	Failed

which has effectively 50% sensitivity and 50% specificity. In short the area measures the ability of the test to correctly classify those with and without the disease.

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}\ (t)dt \tag{1.3}$$

where $t = 1 - \text{specificity}$ (false positive rate) and ROC(t) is sensitivity (true positive rate). We can establish the following classification for the test.

Generally two approaches are employed for computing AUC. A nonparametric method based on constructing trapezoids under the curve as an approximation of area and a parametric method using a maximum likelihood estimator to fit a smooth curve to the data points. Huang and Ling (2005) demonstrated that AUC is a better evaluation measure than accuracy or error rate. A nonparametric method based on Mann–Whitney *U* statistic (actually the *p* statistic from the *U* statistic) has been applied for evaluating the classifiers (Sree Hari Rao and Naresh Kumar 2011d).

5.11 Data Set

We first propose to identify early clinical features in both children and adults having known clinical diagnosis. This would enable one to determine the suspected dengue individuals in the community. To accomplish this task the authors (Sree Hari Rao and Naresh Kumar 2011d) have considered clinical features from a data set (see Table 1.6) consisting of 1,044 individuals belonging to central and western States of India. The patient records were segregated into children (5–15 years) and adults

Table 1.6 Clinical and laboratory features of dengue pertaining to 1,044 individuals

Attribute	Type
Vomiting/nausea	No, yes
Myalgia	Yes, no
Rashes	No, yes
Bleeding site	No, yes
Headache	Yes, no
Restlessness	No, yes
Abdominal pain	No, yes
Retro-orbital pain	No, yes
Arthralgia	No, yes
Fever	Real
Fever duration	Integer
Pulse	Integer
Hemoglobin (Hb)	Real
White blood cell(WBC) count	Real
Platelet	Real
Packed cell volume (PCV)	Real
Diagnosis	Negative, Positive

(≥16 years) as the clinical symptoms presented by them are not similar (Pongsumpun and Tang 2001; Ramos et al. 2009). The data records included the demographic attributes age, gender in addition to clinical symptoms fever, fever duration, headache, retro-orbital pain (eye pain), myalgia (body pain), arthralgia (joint pain), nausea or vomiting, rashes, bleeding sites, restlessness, and abdominal pain.

Later, we develop a method to handle the clinical and laboratory features for more accurate diagnosis and identification of operating range of numeric attributes that can aid in detecting the severity of the infection in suspected dengue individuals (Sree Hari Rao and Naresh Kumar 2011a). The laboratory features hemoglobin (Hb), white blood cell count (WBC), packed cell volume (PCV), platelets were considered for analysis.

6 A Predictive Modeling Strategy

Our predictive modeling strategy is as follows: we have considered data records containing both clinical and laboratory features and known diagnosis of 1,044 individuals. As a first step we consider all these records with clinical features only and utilizing the known diagnosis we apply our RNIADT methodology to determine the essential clinical features that would help identify the suspected dengue individuals. In the next step we use both clinical and laboratory features and the decision to build a predictive ADTree which has the capability of yielding the decision rules that confirm the diagnosis. The machine knowledge obtained by studying these 1,044 data records will be useful to diagnose other individuals (based on clinical and laboratory features) where the clinical decision is unavailable.

Of the 1,044 individuals with suspected dengue, 398 were children and 646 were adults. Out of the 398 children, 93 (23.3%) were dengue positive and 305 (76.7%) were dengue negative. Of the 646 adults, 256 (39.6%) were dengue positive and 390 (60.4%) were dengue negative.

6.1 Predictive Clinical Features in Children

It was observed in Sree Hari Rao and Naresh Kumar (2011d) that dengue-positive children (average age 11.7 years) were likely to be younger than dengue-negative children (average age 12.9 years) ($p < 0.05$). No significant difference in the proportions of male or female children between the dengue-positive and dengue-negative children was observed. The average fever duration for dengue positive was higher by 2 days when compared to dengue-negative ($p < 0.05$) children. Arthralgia was reported as the common clinical symptom among dengue-positive children (Table 1.7). Retro-orbital pain was reported 90% among dengue-positive children and 64% among dengue-negative children. Rashes were reported 78% and 83% among dengue-positive and dengue-negative children, respectively. The attributes

Table 1.7 Reported clinical features of suspected dengue-positive children

Clinical feature	Dengue positive ($n=93$), n (%)	Dengue negative ($n=305$), n (%)	Crude odds ratio (95% CI)
Classic dengue			
Myalgia	33/93 (35.48)	114/305 (37.38)	0.92 (0.57–1.50)
Rashes	73/93 (78.49)	255/305 (83.61)	0.72 (0.40–1.28)
Bleeding site	2/93 (2.15)	0/305 (0.00)	0.72 (0.40–1.28)
Headache	11/93 (11.83)	197/305 (64.59)	0.07 (0.04–0.14)
Restlessness	2/93 (2.15)	0/305 (0.00)	0.07 (0.04–0.14)
Abdominal pain	14/93 (15.05)	93/305 (30.49)	0.40 (0.22–0.75)
Retro-orbital pain	84/93 (90.32)	196/305 (64.26)	5.19 (2.51–10.73)
Arthralgia	85/93 (91.40)	56/305 (18.36)	47.24 (21.64–103.13)
Gastrointestinal			
Nausea or Vomiting	60/93 (64.52)	245/305 (80.33)	0.45 (0.27–0.74)

Table 1.8 Early clinical features selected by RNIADT for predicting dengue

				Predictive value		
Decision attribute	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive (%)	Negative (%)	AUC
Children						
Arthralgia	83.91	91.4	81.6	60.2	96.8	0.83
Arthralgia, headache	95.22	79.5	100	100	94.14	0.95
Arthralgia, headache, retro-orbital pain, myalgia	96.48	86.02	99.67	98.77	95.9	0.98
Arthralgia, headache, retro-orbital pain, myalgia, abdominal pain	97.27	89.2	99.67	98.8	96.8	0.98
Adults						
Arthralgia	82.2	83.9	81.0	74.3	88.5	0.79
Arthralgia, myalgia, rashes, headache, vomiting or nausea, abdominal pain	84.98	75.3	91.3	85.0	84.9	0.88

bleeding site and restlessness were reported least number of times among dengue-positive and negative children; however, rashes and bleeding site have odds of 0.72 times higher in dengue-positive children than in dengue-negative children.

The multivariate analysis revealed that dengue-positive children were 47 times more likely to present with arthralgia than dengue-negative children. Children with myalgia were found to be five times more likely to have dengue positive than dengue negative.

The alternating decision tree algorithm generated a model having clinical features arthralgia, headache, retro-orbital pain, and myalgia with a predictive value of 98.8% for dengue positive and 96.8% for dengue negative with an AUC of 0.98 (Table 1.8). The alternating decision tree for children between 5 and 15 is shown in Fig. 1.7.

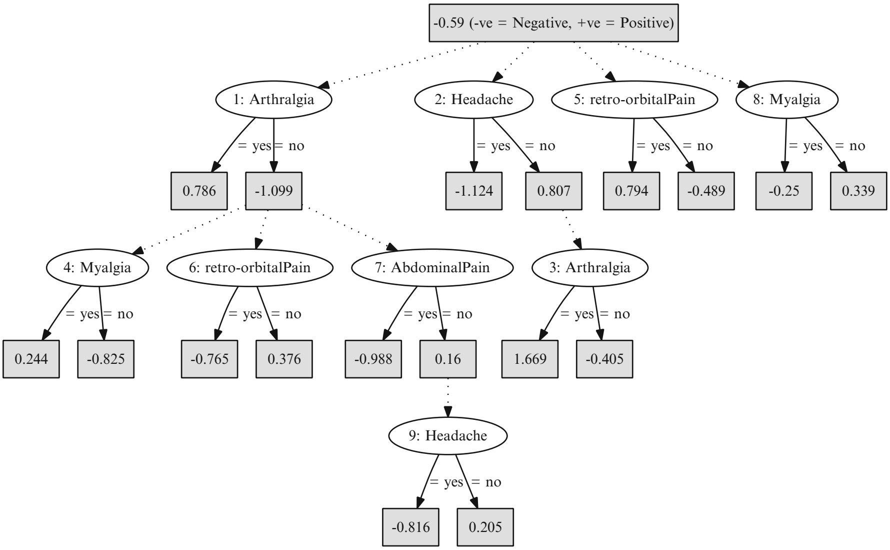

Fig. 1.7 Alternating decision tree generated based on clinical features in children

The C4.5 decision tree classifier had identified arthralgia, retro-orbital pain, headache, rashes, and abdominal pain as influential attributes with an accuracy of 90.7% and predictive positive value of 100% and negative predictive value of 89.2%. The logistic regression method when applied on the data set identified arthralgia, retro-orbital pain, bleeding site, and restlessness as having higher odds for identifying dengue positive and negative in children as compared to the other attributes. The authors have found that RNIADT has identified myalgia as an influential attribute resulting in a more accurate classifier than C4.5 and logistic regression. The authors refer the readers to Sree Hari Rao and Naresh Kumar (2011d) for a more detailed analysis and comparisons.

The decision rules extracted from an alternating decision tree for suspected dengue in children are as follows:

(a) The dominant clinical features identified are arthralgia, myalgia, retro-orbital pain.
(b) If the patient is suffering from arthralgia, retro-orbital pain, myalgia, and does not have a headache and abdominal pain then the diagnosis is positive. The predictive score is computed as (0.786 + 0.807 + 1.669 + 0.244 + 0.794 + (−0.765) + 0.16 + (−0.25) = 3.445).
(c) If the patient is not suffering from arthralgia, retro-orbital pain, myalgia and if headache and abdominal pain are present then the diagnosis is negative. The score is computed as ((−1.099) + (−1.124) + (−0.405) + (−0.825) + (−0.489) + 0.376 + (−0.988) + 0.339 = −4.215).

6.2 Predictive Clinical Features in Adults

It has been observed that the dengue-positive adults were likely older by 3 years when compared to dengue-negative adults (average of 28.99 years vs. 25.14 years respectively) ($p < 0.05$). The proportion of patients of both the male and female population did not differ between dengue-positive and dengue-negative adults. The classic dengue symptoms most commonly reported were arthralgia, retro-orbital pain followed by myalgia and rashes (Table 1.9). Arthralgia was reported most in dengue-positive patients than in dengue-negative patients.

The multivariate analysis revealed that the dengue-positive adults were more likely to report arthralgia than dengue-negative adults. They were also likely to report myalgia than dengue-negative adults. Nausea or vomiting was found to be more likely among dengue-positive than dengue-negative adults. The odds of finding bleeding sites and retro-orbital pain are 1.8 and 1.75 times, respectively, in dengue-positive adults than in dengue-negative adults.

The RNIADT generated a model with clinical attributes arthralgia, myalgia, rashes, abdominal pain, headache, and nausea or vomiting with an accuracy of 86.2% and predictive value for positive cases as 87% and for negative is 85.7% with AUC of 0.91 (Table 1.8). The RNIADT generated for adults is shown in Fig. 1.8. The Influential attributes identified by C4.5 decision tree are arthralgia, myalgia, rashes, bleeding site, vomiting or nausea, and restlessness with an accuracy of 80.2% and predictive value of 85.2% for positives and 78.2% for dengue negatives with an AUC of 0.84. The logistic regression identified clinical features arthralgia, myalgia, retro-orbital pain, restlessness, and vomiting or nausea having higher odds with an accuracy of 77.7%, predictive value of 79.2% for positives and 77.1% for dengue negatives with an AUC of 0.78.

The following decision rules were extracted from the alternating decision tree for suspected dengue in adults:

(a) The dominant clinical features identified for positive diagnosis of dengue in adults are arthralgia and myalgia.

Table 1.9 Reported clinical features of dengue in adult patients

Clinical feature	Dengue positive ($n = 256$), n (%)	Dengue negative ($n = 390$), n (%)	Crude odds ratio (95% CI)
Classic dengue			
Myalgia	197/256 (76.95)	227/390 (58.21)	2.40 (1.68–3.41)
Rashes	213/256 (83.20)	345/390 (88.46)	0.65 (0.41–1.01)
Bleeding site	15/256 (5.86)	13/390 (3.33)	1.80 (0.84–3.86)
Headache	77/256 (30.08)	136/390 (34.87)	0.80 (0.57–1.13)
Restlessness	10/256 (3.91)	32/390 (8.21)	0.45 (0.22–0.94)
Abdominal pain	18/256 (7.03)	50/390 (12.82)	0.51 (0.29–0.90)
Retro-orbital pain	231/256 (90.23)	328/390 (84.10)	1.75 (1.07–2.86)
Arthralgia	215/256 (83.98)	74/390 (18.97)	22.39 (14.73–34.05)
Gastrointestinal			
Nausea or Vomiting	87/256 (33.98)	111/390 (28.46)	1.29 (0.92–1.82)

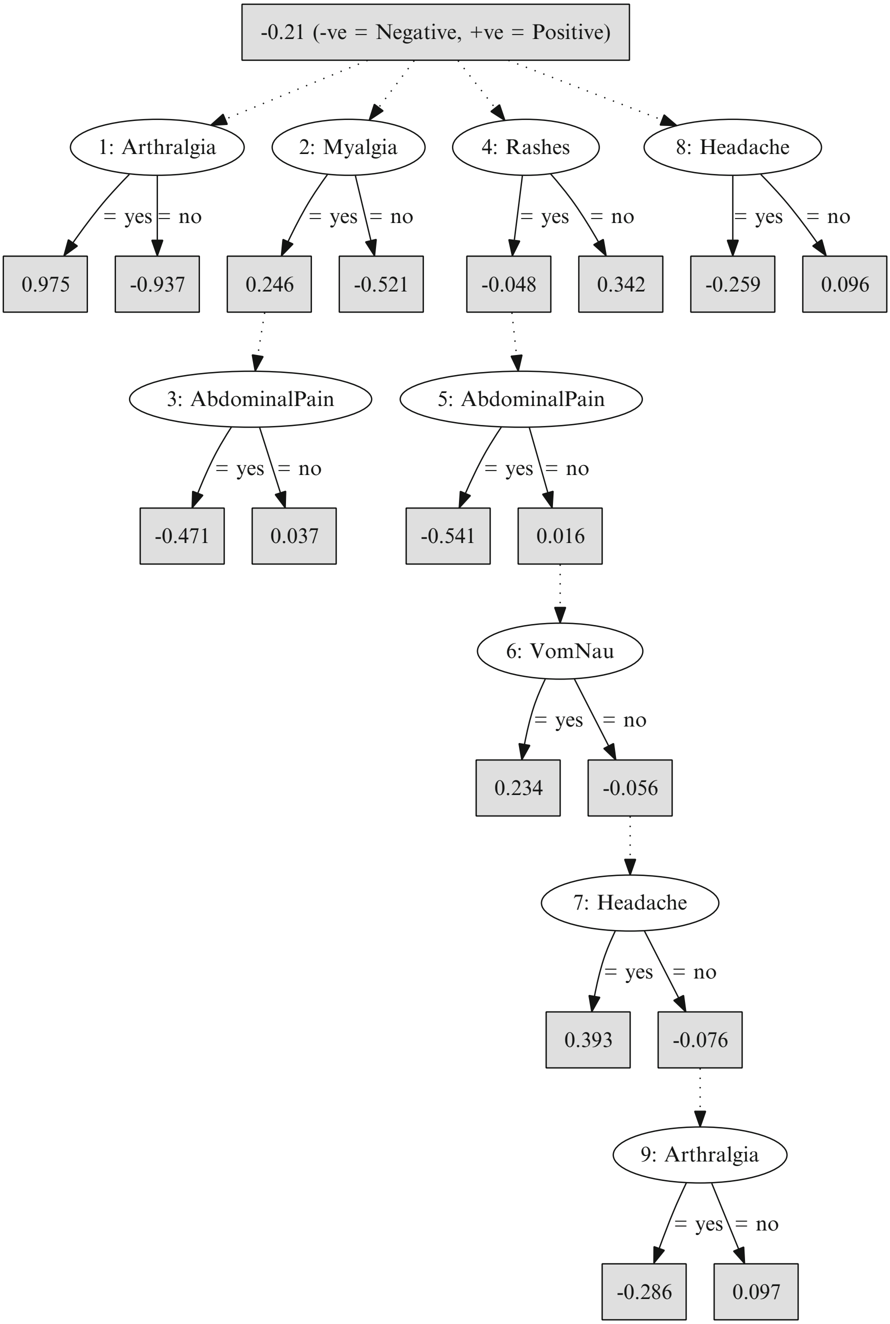

Fig. 1.8 Alternating decision tree based on clinical features in adults

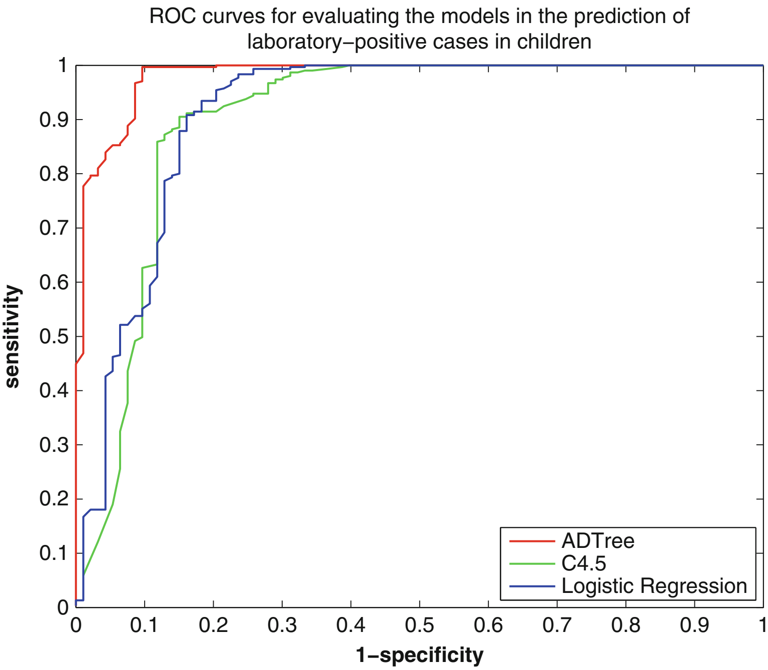

Fig. 1.9 ROC curves for evaluating the models in the prediction of dengue-positive cases in children

(b) The presence of abdominal pain is contributing for identifying negative cases of dengue.

(c) If the patient is suffering from arthralgia and myalgia and does not show signs of abdominal pain, headache, vomiting or nausea, and rashes, then he or she is dengue positive. The predictive score can be computed from the alternating decision tree as (0.975 + 0.246 + 0.037 + 0.342 + 0.016 − 0.056 –0.076 + 0.096 − 0 .286 = 1.294).

(d) If the patient is not suffering from arthralgia and myalgia but has symptoms such as abdominal pain, rashes, headache, and vomiting or nausea, then the diagnosis is negative. The predictive score is computed as (–0.937 − 0.521 − 0.4 71 − 0.048 − 0.541 + 0.234 + 0.393 − 0.259 + 0.097 = –2.053).

The receiver operator characteristic curves for RNIADT, C4.5 and logistic regression for children and adults are shown in Figs. 1.9 and 1.10, respectively. The different performance metrics suggest that RNIADT algorithm has outperformed C4.5 and logistic regression methodologies.

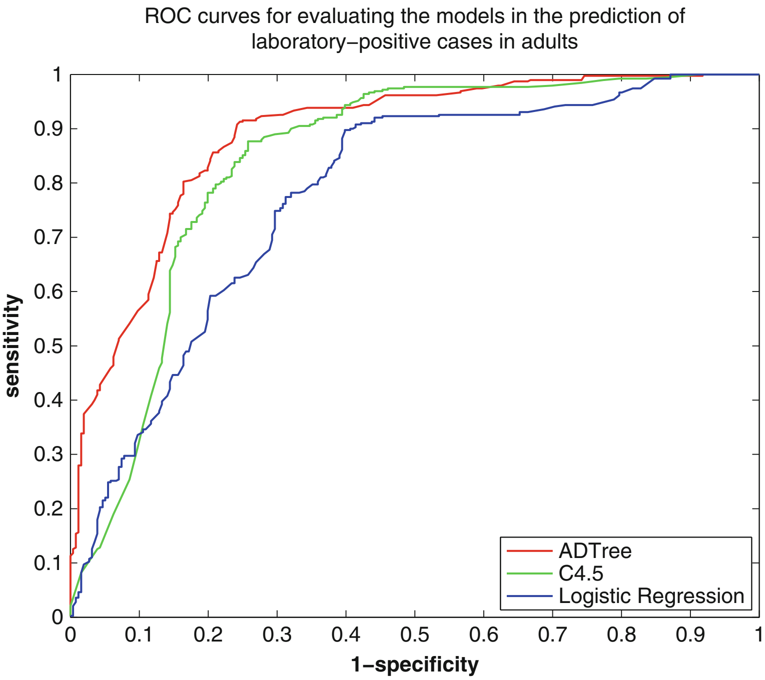

Fig. 1.10 ROC curves for evaluating the models in the prediction of dengue-positive cases in adults

6.3 Predictive Clinical and Laboratory Features in Children

The alternating decision tree identified laboratory features platelet, WBC, and Hb having 100% positive predictive value and 99.67% negative predictive value with an AUC of 0.99 (see Table 1.10). The alternating decision tree generated using the laboratory and clinical features for predicting dengue in children is shown in Fig. 1.11. Further, the laboratory attributes with platelet count less than or equal to 140, WBC over and above 8.8 and Hb less than 12.5 contributed for positive diagnosis of dengue. The clinical attributes such as fever over and above 100.5°F, pulse over and above 81.5, and the presence of arthralgia contributed for positive diagnosis.

6.4 Predictive Clinical and Laboratory Features in Adults

The alternating decision tree identified laboratory features platelet, WBC, and Hb having 100% positive predictive value and 99.24% negative predictive value with AUC of 1.0 (see Table 1.11). In adults, arthralgia (positive prediction value of 1.37)

Table 1.10 Accuracies obtained using clinical and laboratory features dengue in children

Method attribute	Accuracy (%)	Sensitivity	Specificity	Predictive value Positive	 Negative	AUC
RNIADT	99.75	98.92	100.00	100.00	99.67	0.99
ADT	97.74	91.40	99.67	98.84	97.44	0.99
BNT	98.74	94.62	100.00	100.00	98.39	0.99
C4.5	96.73	86.02	100.00	100.00	95.91	0.99
LOR	94.97	86.02	97.70	91.95	95.82	0.95
NIB	96.48	84.95	100.00	100.00	95.61	0.99
RBF	97.99	95.70	98.69	95.70	98.69	0.98

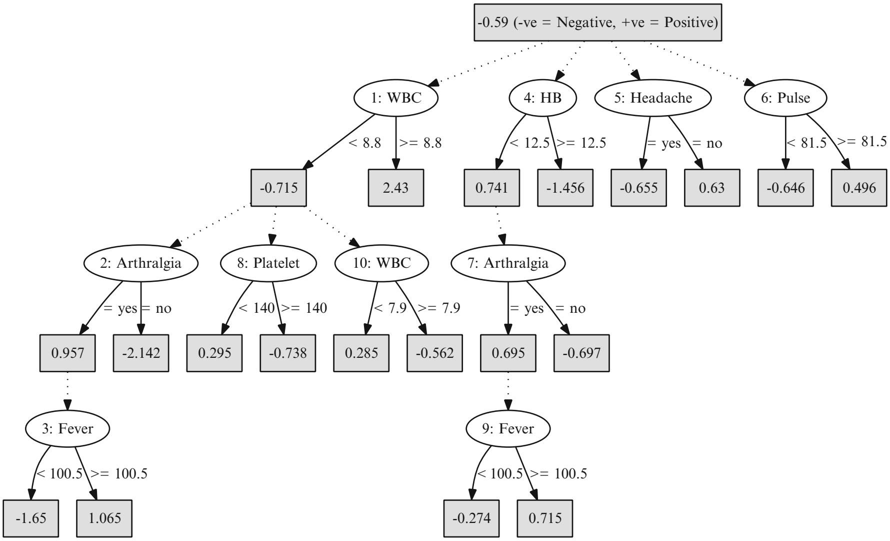

Fig. 1.11 RNIADT decision trees with predictive clinical and laboratory features of dengue in children

Table 1.11 Accuracies obtained using clinical and laboratory features dengue in adults

Method attribute	Accuracy (%)	Sensitivity	Specificity	Predictive value Positive	 Negative	AUC
RNIADT	99.54	98.83	100.00	100.00	99.24	1.00
ADT	97.99	96.88	98.72	98.02	97.96	1.00
BNT	95.67	89.45	99.74	99.57	93.51	0.99
C4.5	95.82	92.19	98.21	97.12	95.04	0.99
LOR	90.87	86.72	93.59	89.88	91.48	0.96
NIB	86.22	82.03	88.97	83.00	88.30	0.93
RBF	92.11	90.63	93.08	89.58	93.80	0.97

was found to be effective in diagnosis dengue. The alternating decision tree generated using the laboratory and clinical features for predicting dengue in adults is shown in Fig. 1.12. Further, the laboratory attributes with platelet less than 167.5, WBC over and above 8.9, and Hb less than 12.5 contributed for positive diagnosis

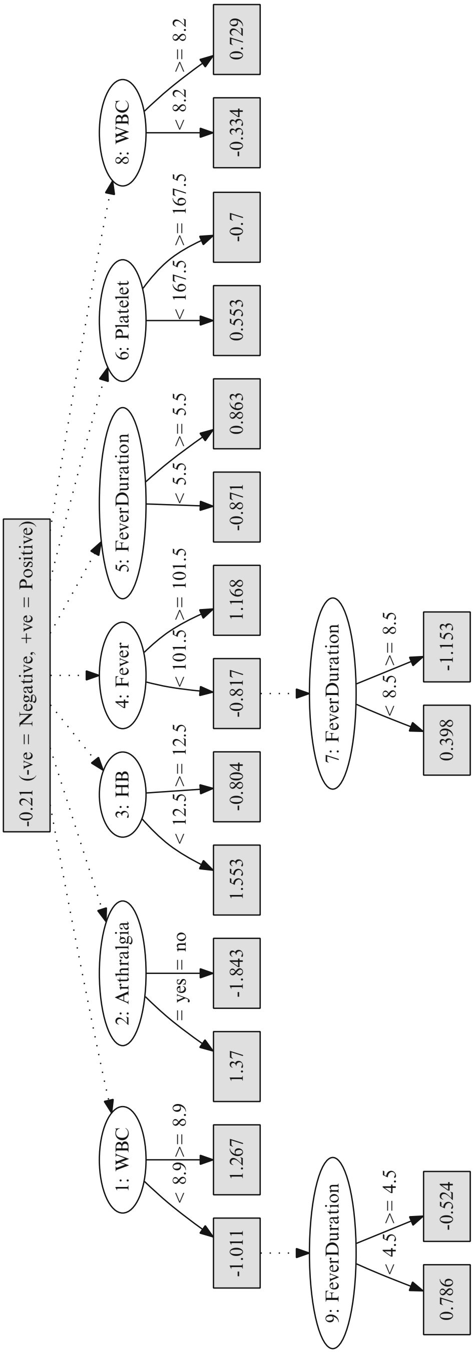

Fig. 1.12 RNIADT decision trees with predictive clinical and laboratory features of dengue in adults

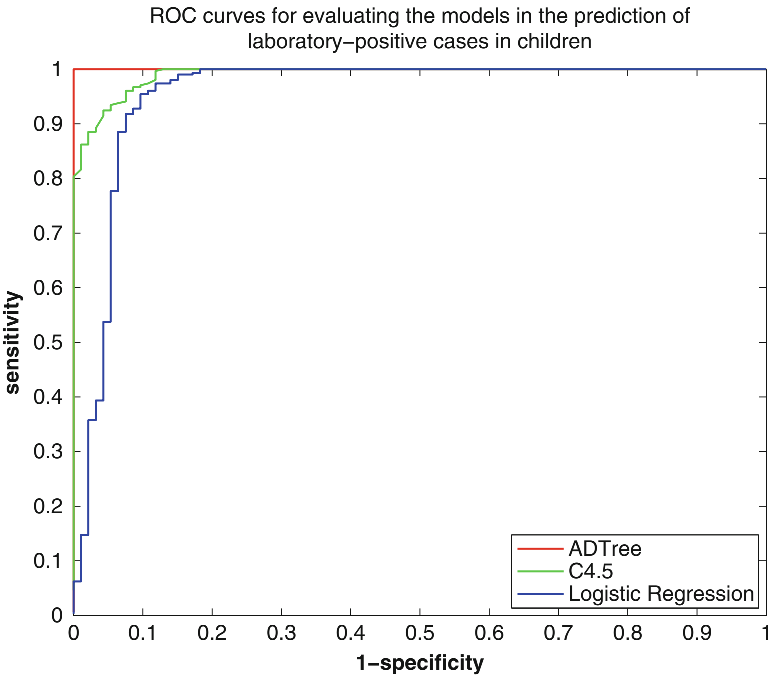

Fig. 1.13 ROC curves for evaluating the models in the prediction of dengue in children using laboratory and clinical features

of dengue. The presence of arthralgia in adults is contributed for positive predictions of dengue with a predictive value of 1.37. The clinical features such as fever over and above 101.5°F and fever duration over and above 5 days have high predictive scores for positive diagnosis of dengue.

The receiver operator characteristic curves for RNIADT, C4.5 and logistic regression for children and adults generated using clinical and laboratory features are shown in Figs. 1.13 and 1.14, respectively.

It is quite evident from ROC curves that RNIADT has outperformed C4.5 and the logistic regression methods.

6.5 Identifying Predictive Clinical and Laboratory Features Using Feature Selection Methods

A dengue data set consisting of both laboratory and clinical features has been considered in Sree Hari Rao and Naresh Kumar (2011a) (see Table 1.6) to establish more accurate and simplified decision rules. The data set had missing values up to 20% in each of the attributes. The decision tree algorithm presented in Appendix A

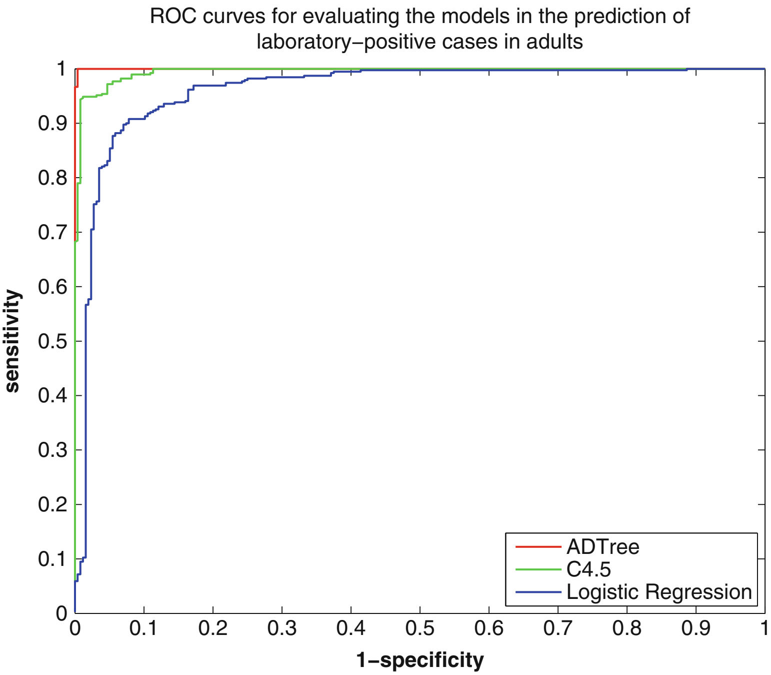

Fig. 1.14 ROC curves for evaluating the models in the prediction of dengue in adults using laboratory and clinical features

Table 1.12 Classification accuracies of different classifiers using GA search wrapper subset method

Method	RNIADT	BNT	NIB	RBF	LOR	C4.5	ADT
GA+ADT	99.71	95.40	87.07	92.34	90.23	97.03	98.75
GA+BNT	99.71	96.55	85.44	93.10	91.28	97.03	97.89
GA+NIB	99.81	93.01	86.49	91.57	90.61	96.55	97.22
GA+RBF	99.71	95.21	85.82	91.57	89.46	96.65	97.70
GA+C4.5	99.04	94.06	77.30	84.87	84.00	97.32	96.46
GA+LOR	99.90	94.16	87.07	90.13	92.05	97.03	98.37

Sree Hari Rao and Naresh Kumar (2011a, d) has been employed for generating the RNIADT and its accuracies are compared with other popular classifiers.

The authors in Sree Hari Rao and Naresh Kumar (2011a) have applied GA search algorithm for features extraction using wrapper subset evaluation procedure. These techniques were applied on dengue data set to obtain a more accurate predictive model (see Table 1.12). In Sree Hari Rao and Naresh Kumar (2011b) PSO search algorithm on dengue data set has been applied and the accuracies obtained are presented in (see Table 1.13). For a more detailed comparison of different classifiers and search algorithms the readers are referred to Sree Hari Rao and Naresh Kumar (2011a, b).

Table 1.13 Classification accuracies of different classifiers using a PSO search wrapper subset method

Method	RNIADT	BNT	NIB	RBF	LOR	C4.5	ADT
PSO+ADT	99.71	96.26	87.36	92.62	90.23	97.13	98.75
PSO+BNT	99.52	96.46	85.82	93.10	90.61	96.74	97.99
PSO+NIB	99.81	94.64	86.69	91.67	91.00	96.93	98.18
PSO+RBF	99.71	95.21	85.82	91.57	89.46	96.65	97.70
PSO+C4.5	99.71	94.16	81.23	87.16	87.26	97.32	96.74
PSO+LOR	99.71	94.25	87.36	92.15	92.24	97.03	98.37

Table 1.14 Classification accuracies of different classifiers using wrapper subset method and PKI discretization

Method	RNIADT	BNT	NIB	RBF	LOR	C4.5	ADT
PKI+GA+ADT	99.04	89.46	90.90	93.39	97.22	97.32	98.66
PKI+GA+BNT	99.90	94.06	96.17	95.79	97.32	97.70	98.66
PKI+GA+NIB	100.00	94.06	96.07	95.88	98.08	96.93	98.28
PKI+GA+RBF	99.81	93.49	94.64	95.88	96.65	97.32	98.37
PKI+GA+C4.5	99.90	88.98	91.67	90.71	97.89	97.41	95.79
PKI+GA+LOR	99.23	88.22	89.56	91.48	97.70	95.98	96.65
PKI+PSO+ADT	99.81	94.16	95.79	97.41	96.65	96.65	98.66
PKI+PSO+BNT	99.71	94.35	96.93	95.88	98.75	97.22	97.03
PKI+PSO+NIB	99.52	95.11	97.13	97.22	97.89	97.41	97.41
PKI+PSO+C4.5	100.00	92.15	93.77	95.59	97.41	97.80	97.89

Discretization method based on PKI was employed as a preprocessing step in Sree Hari Rao and Naresh Kumar (2011a) before identifying the most influential attributes. The accuracies obtained by different classifiers are shown in Table 1.14.

A comparison of the classification accuracies tabulated in Tables 1.12 and 1.13 suggests that discretization procedure improves the accuracies for the data set under consideration. It is observed in general that application of discretization method would generate user-friendly decision trees and more descriptive rules (see Fig. 1.15). The influential features identified by different methods are tabulated in Table 1.15. The RNIADT identified the attributes fever duration, pulse, WBC, and arthralgia as most influential features classified instances with a classification accuracy of 100%.

The difference in the percentage accuracy when compared with other classifiers is shown in Fig. 1.16. The RNIADT outperformed Naive Bayes, RBFNetworks, and logistic regression classifiers and the difference in accuracies were found to be greater than 7%.

The discretization method when applied on the dengue data set generated an RNIADT decision tree that outperformed Bayes Network, Naive Bayes, and RBF Network classifiers (see Fig. 1.17).

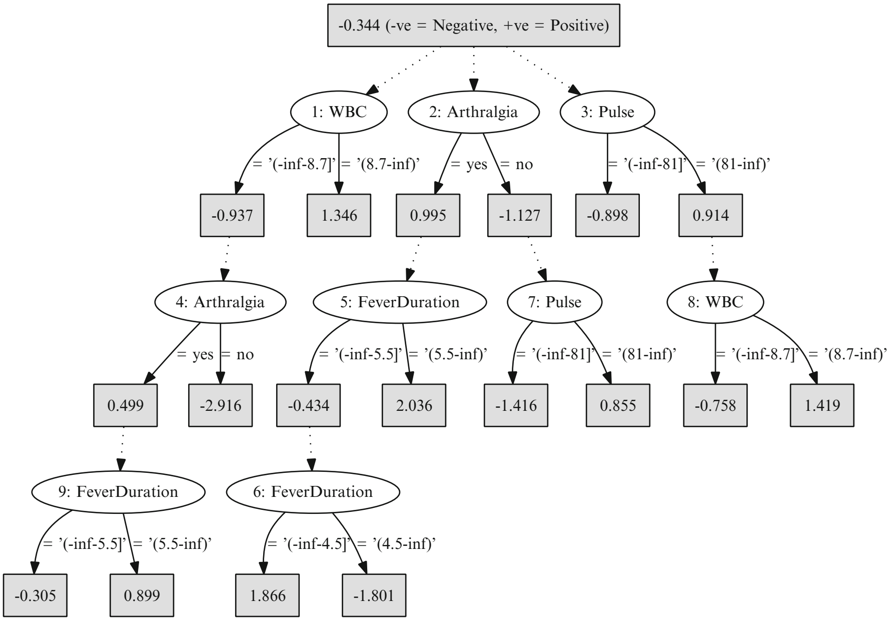

Fig. 1.15 RNIADT decision tree generated after discretization and extraction of influential attributes using a PSO search mechanism and C4.5 evaluation

Table 1.15 Influential features identified by different feature selection methods

Method	Features identified by ADT	Features identified by C4.5	Features identified by RNIADT
GA+ADT	WBC, arthralgia, Hb, fever, platelet, PCV	WBC, fever, Hb, platelets, arthralgia	Fever, platelet, arthralgia, fever duration, platelet
PKI+GA+C4.5	Hb, arthralgia, WBC, pulse, fever duration, platelet, pulse	Hb, WBC, arthralgia, platelet, pulse, fever duration, headache	WBC, arthralgia, pulse, fever duration, myalgia
PSO+C4.5	WBC, arthralgia, platelet	WBC, pulse, arthralgia, platelet, abdominalpain	WBC, arthralgia, pulse, fever duration
PKI+PSO+C4.5	WBC, arthralgia, Hb, platelet, bleeding site, pulse	WBC, Hb, arthralgia, bleeding site, platelet, pulse	WBC, arthralgia, pulse, fever duration

The ROC curves generated by different classifiers based on the dengue data set having both clinical and laboratory attributes is shown in Fig. 1.18. Figure 1.18 compares the performance of RNIADT with C4.5 and ADTree classifiers. From Fig. 1.18 we can conclude that RNIADT has outperformed the other classifiers and has a better AUC than C4.5 and ADTree.

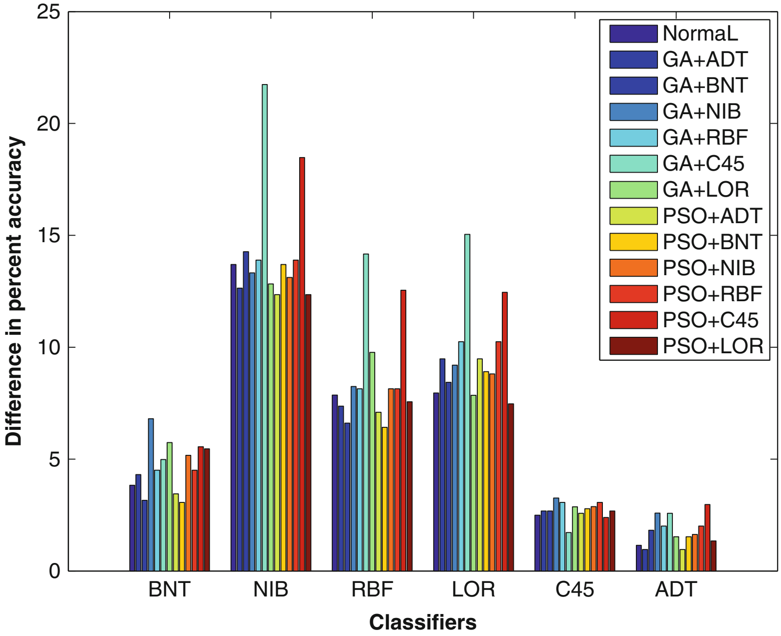

Fig. 1.16 Relative differences of other classifiers with RNIADT using different feature selection methods

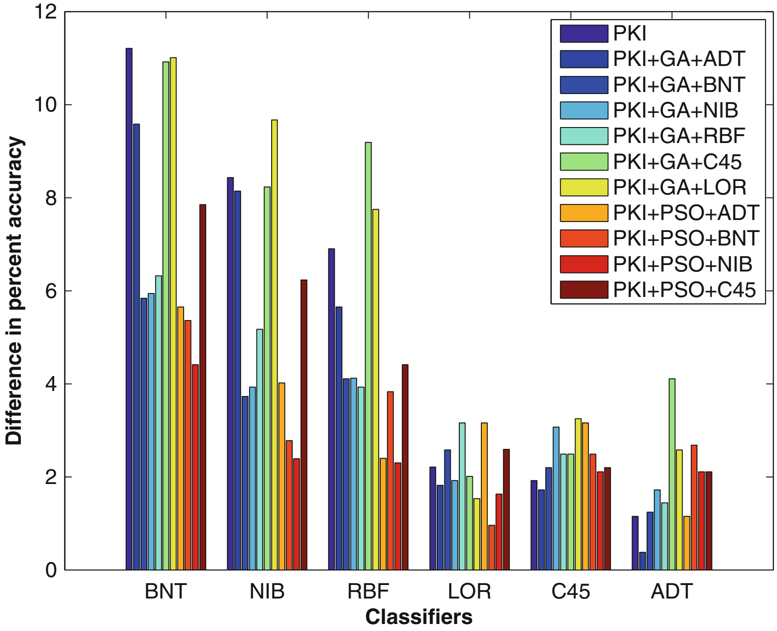

Fig. 1.17 Relative differences of other classifiers with RNIADT using different feature selection methods and PKI discretization

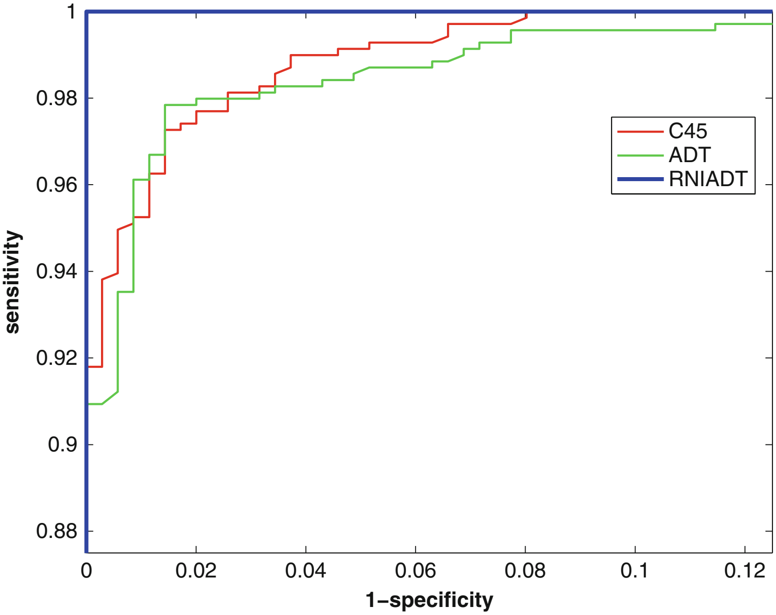

Fig. 1.18 ROC curves for classifiers trained on features extracted after discretization and using a PSO search with C4.5 evaluation procedure

7 Comparisons of Methodologies

The procedures suggested in (Chadwick et al. 2006; Ramos et al. 2009; Tanner et al. 2008) when applied on the data set (Sree Hari Rao and Naresh Kumar 2011d) (see Table 1.16) reveal the fact that the RNIADT algorithm rendered higher accuracies in terms of area under the curve and percentage predictive value for positive than those obtained by them.

Tanner et al. (2008) in their studies applied C4.5 algorithm on 1,200 patients records with data obtained in 72 h of illness. The algorithm has selected laboratory features such as platelet count, white cell count, lymphocyte, neutrophil, temperate and hematocrit as the influential attributes. The studies in Tanner et al. (2008) have suggested a WBC $\leq$6.0 × 1,000 cells with an odds ratio of 8.7 and body temperature > 37.4°C mm^3 having an odds ratio of 7.2 playing a role in splitting the decision tree. Sree Hari Rao and Naresh Kumar (2011b) have identified WBC, Hb, rashes, and fever (body temperature) as the key attributes influencing the diagnosis of dengue. The predictive value of WBC $\geq$ 8.2 × 1,000 cells was found to be 1.3, pulse $\geq$ 81 has a predictive value of 0.91 mm^3 and fever duration $\geq$5.5 has a predictive value of 2.03. The comparisons of the results are presented in Tables 1.17 and 1.18. From these observations the authors have felt that the methodologies in Sree Hari Rao and Naresh Kumar (2011a, b, d) when applied on the data set (Chadwick et al. 2006; Ramos et al. 2009; Tanner et al. 2008) would yield more accurate results.

Table 1.16 Comparison of different methods for predicting early clinical features in children and adults

Method attribute	Accuracy	Sensitivity (%)	Specificity (%)	Predictive value (%)		AUC
				Positive	Negative	
Children						
Logistic regression	92.7	74.2	98.4	93.2	92.6	0.91
C4.5	90.7	60.2	100	100	89.2	0.90
RNIADT	97.2	89.3	99.7	98.8	96.8	0.98
Adults						
Logistic regression	77.7	59.9	89.7	79.2	77.1	0.78
C4.5	80.2	60.6	93.1	85.2	78.2	0.84
RNIADT	84.98	75.3	91.3	85.0	84.9	0.88

Table 1.17 Comparison of our results with (Tanner et al. 2008)

Method	True positive	False positive	False negatives	True negatives
Tanner et al. (2008)	259	83	105	753
Sree Hari Rao and Naresh Kumar (2011a)	349	0	0	695

Table 1.18 Comparison of performance measures of our methodology with Tanner et al. (2008)

Measure	Tanner et al. (2008)	Sree Hari Rao and Naresh Kumar (2011a)
Sensitivity (%)	71	100
Specificity (%)	90	100
Positive predictive value (%)	76	100
Negative predictive value (%)	88	100
AUC	0.88	0.99

8 Conclusions and Discussion

In this chapter, we have presented several methodologies that help in the effective diagnosis of the dengue illness. A first level effort leads to the question of identifying the suspected individuals in the community, which will have the major advantage of reducing transmission risk of the disease. Laboratory investigations for the confirmation of the illness on the suspected individuals will certainly help in disease management and control by providing supportive care. A new alternate decision theoretic method designated as RNIADT (which is not followed in conventional clinical treatment procedures) developed in recent times is the subject of main discussion in this chapter. This methodology has been found extremely useful in identifying the most influential clinical and laboratory characteristics of dengue illness. Further, this analysis helps one to conclude that the WHO definitions for dengue fever hold good. To substantiate, a study has been performed on a data set consisting of 1,044 individuals both children and adults where in the original definitions of

WHO are still valid. Though the methodology discussed in this chapter may be taken as a universal tool for the effective diagnosis of this disease it remains to see whether or not this methodology is geographically dependant. Though we are certain that the RNIADT methodology is universal, we could not establish the same due to lack of clinical and laboratory data pertaining to different parts of the globe. However, we are willing to share our predictive methodologies and strategies with the researchers working on dengue illness all over the globe. We hold the view that more intensive and introspective studies of this kind will pave the way for better clinical management and virological surveillance of this illness.

Acknowledgements This research is supported by the Foundation for Scientific Research and Technological Innovation (FSRTI)—A Constituent Division of Sri Vadrevu Seshagiri Rao Memorial Charitable Trust, Hyderabad 500 035, India.

9 Algorithm 1: The RNIADT Algorithm (Sree Hari Rao and Naresh Kumar 2011c)

Input: (a) Data sets for purpose of decision making $S(m, n)$ where m and n are number of records and attributes, respectively and the members of S may have missing values in any of the attributes except in the decision attribute.

(b) The type of attribute C of the columns in the data set.

(c) The number of boosting iterations T.

(d) The number of validation folds k.

Output: (a) Classification accuracy of the RNIADT for a given data set S.

(b) RNIADT consisting of a rule that is the sign of the sum of all the base rules in

$$\text{class}(x) = \text{sign}(\sum_{t=1}^{T} rt(x))$$

Algorithm

(1) Identify and collect all records in a data set S and split them into training and testing data sets using a k fold cross validation procedure. Denote the training and testing data sets by T_k and R_k, respectively.

(2) Consider records in the training data pertaining to a particular cross fold and impute the missing values using the following procedure.

(i) Identify and collect all records in the data record set S which have missing values in one or several attributes but not those with missing values in the decision attribute. Denote this set by M i.e. $M \subseteq S$.

(ii) Pick up a record R from the set M and compute its relative distances with all members of S using the procedure given in Sree Hari Rao and Naresh Kumar (2011c). Denote this set by D.

(iii) Arrange the elements of set D in an ascending order and identify the nearest neighbors using the following procedure.

(iv) (a) Compute the score α defined as follows: $\alpha(x_k) = \frac{(x_k - median(x))}{median|x_i - median(x)|}$ where $\{x_1, x_2, \ldots, x_n\}$ denote the distances of R from R_k.

(b) Collect the data records in set S whose distances from the record R satisfies the condition $\alpha(x_k) \leq 0$. Denote this set by P.

```
    (v)    If the type of the attribute to be imputed in $R$ is nominal or categorical, then determine the frequent item set from $P$ using the following procedure:
           (a) Find the frequency of each categorical value of the categorical attribute.
           (b) The value to be imputed may be taken as the highest categorical value of the frequent item set obtained in Step (v) item (a).
    (vi)   If the type of attribute is numeric and non-integer, then determine the value to be imputed using following procedure.
           (a) Identify and collect all non-zero elements in the set $D$ computed in Step (ii). Denote this set by $B$.
           (b) For each element in set $B$ compute the quantity $\beta(j) = \frac{1}{B(j)} \forall j = 1,\ldots,\gamma$ where $\gamma$ denotes the cardinality of the set $B$.
           (c) Compute the weight matrix as $W(j) = \frac{\beta_j}{\sum_{i=1}^{\gamma} \beta(i)} \forall j = 1,\ldots,\gamma$
           (d) The value to be imputed may be taken as $\sum_{i=1}^{j} P(j) \cdot W(j) \; \forall j = 1,\ldots,\gamma$
    (vii)  If the type of attribute is numeric and integer, the procedure given in Step (v) is followed.
    (viii) Repeat Steps (2)(i)–(vi) for every record $R$ in the set $M$.
(3) Build the ADTree on the records obtained in Step (2) as follows.
    (i)    Initialize the rule set $R_1$ to consist of the single base rule whose precondition and condition are set to True $P_1$ =True. The symbols $P_t$ and $R_t$ denote the set of preconditions and rules, respectively.
    (ii)   Initialize the weights of each training sample with 1 i.e.
    (iii)  The prediction value of the root node is calculated as $a = \frac{1}{2} \ln \frac{W_+(True)}{W_-(True)}$. $W(c)$ represents the total weight of the training samples that satisfies the base condition $c$. $W_+(c)$ and $W_-(c)$ denote the weights of those examples that satisfy the condition $c$ and are labeled +1 or −1.
    (iv)   Pre-adjustment: re-weight the training instances using the formula $w_{i,1} = w_{i,0} \, e^{-ay_t}$ (for binary classification, the value of $y_t$ is either +1 or −1).
    (v)    Perform the following steps for each boosting iteration $t$.
           (a) For each base condition $c_1 \in P_t$ and each condition $c_2 \in C$ calculate
               $Z_t(c_1,c_2) = 2(\sqrt{W_+(c_1 \wedge c_2) W_-(c_1 \wedge c_2)} + \sqrt{W_+(c_1 \wedge \sim c_2) W_-(c_1 \wedge \sim c_2)} + W(\sim c_2)$. The set of base conditions (inequalities comparing a single feature and a constant) is denoted by $C$.
           (b) Select $c_1$, $c_2$ which minimizes $Z_t(c_1, c_2)$ and set $R_{t+1}$ to be $R_t$ with addition of rules $r_t$ whose precondition is $c_1$, condition $c_2$ and two prediction values are
               $a = \frac{1}{2} \ln \frac{W_+(c_1 \wedge c_2)+1}{W_-(c_1 \wedge c_2)+1}$, $b = \frac{1}{2} \ln \frac{W_+(c_1 \wedge \sim c_2)+1}{W_-(c_1 \wedge \sim c_2)+1}$
           (c) Set $P_{t+1}$ to be $P_t$ with the addition of $c_1 \wedge c_2$ and $c_1 \wedge \sim c_2$
           (d) Update the weights of each training example following the equation
               $w_{i,t+1} = w_{i,t} \exp^{r_t(x_i) yi}$
(4) Consider the records in the testing data set pertaining to that cross fold and classify using the tree built in Step (3).
(5) Compute the percentage classification accuracy for a particular cross fold by identifying the number of correctly classified instances with the total number of instances in the testing data set.
(6) Repeat the Steps (2)–(5) for each cross fold.
(7) Compute the mean accuracy $A$ by summing up the accuracies of each cross fold and dividing with the number of cross folds.
(8) RETURN $A$
(9) END
```

References

Allison P (2002) Missing data. Sage, Thousand Oaks

CDC (2000) Centers for disease control and prevention. World distribution of dengue 2000. http://www.cdc.gov/ncidod/dvbid/dengue/mapdistribution-2000.htm

CDC (2011) Centers for disease control and prevention. http://www.healthmap.org/dengue/index.php

Chadwick D, Arch B, Wilder-Smith A, Paton N (2006) Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: application of logistic regression analysis. J Clin Virol 35(2):147–153

Cios KJ, Mooree W (2002) Uniqueness of medical data mining. Artif Intell Med 26:1–24

Crichton N (2002) Receiver operating characteristic (roc) curves. J Clin Nurs 11:134–136

Dash M, Liu H (1997) Feature selection for classification, intelligent data analysis. Intell Data Anal 1:131–156

De Paula S, Fonseca B (2004) Dengue: a review of the laboratory tests a clinician must know to achieve a correct diagnosis. Braz J Infect Dis 8(6):390–398

Derouich M, Boutayeb A, Twizell E (2003) A model of dengue fever. Biomed Eng Online 2:4

Dixon J (1979) Pattern recognition with partly missing data. IEEE Trans Syst Man Cybern 9(10):617–621

Esteva L, Vargas C (1998) Analysis of a dengue disease transmission model. Math Biosci 15(2):131–151

Esteva L, Vargas C (1999) A model for dengue disease with variable human population. J Math Biol 38(3):220–240

Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Proceeding of the sixteenth international conference on machine learning bled. ACM, Slovenia

Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163

George R, Lum L (1997) Clinical spectrum of dengue infection. Dengue and dengue hemorrhagic fever. CAB International, Oxford

George HJ, Pat L (1995) Estimating continuous distributions in Bayesian classifiers. In: Eleventh conference on uncertainty in artificial intelligence, San Mateo, pp 338–345

Gibbons RV (2002) Dengue: an escalating problem. BMJ 324(7353):1563–1566

Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading

Grassly N, Fraser C (2008) Mathematical models of infectious disease transmission. Nat Rev Microbiol 6(6):477–487

Gubler D (1998) Dengue and dengue hemorrhagic fever. Clin Microbiol Rev 11:480–496

Guzman M, Kouri G (2002) Dengue: an update. Lancet Infect Dis 2:33–42

Halstead S (1998) Pathogenesis of dengue: challenges to molecular biology. Science 239(4839):476–481

Halstead SB (2007) Dengue. Lancet 370(9599):1644–1652

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143:29–36

Harris E, Videa E, Perez L, Sandoval E, Tellez Y (2000) Clinical, epidemiologic, and virologic features of dengue in the 1998 epidemic in nicaragua. Am J Trop Med Hyg 63:5–11

Haykins S (1994) Neural network: a comprehensive foundation. Prentice Hall, Upper Saddle River

Heijden G, Donders A, Stijnen T, Moons K (2006) Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 59(10):1102–1109. doi:10.1016/j.jclinepi.2006.01.015

Horton N, Lipsitz S (2001) Multiple imputation in practise: comparison of software packages for regression models with missing variables. Am Stat 55(3):244–254

Huang J, Ling C (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowledge Data Eng 17(3):299–310

Kalayanarooj S, Vaughn D, Nimmannitya S, Green S, Suntayakorn S (1997) Early clinical and laboratory indicators of acute dengue illness. J Infect Dis 176(2):313–321

Kim JO, Curry J (1977) The treatment of missing data in multivariate analysis. Sociol Methods Res 6(2):215–240. doi:10.1177/004912417700600206

Kohavi R, John GH (1998) The wrapper approach. In: Feature extraction, construction and selection: a data mining perspective. Kluwer, New York, pp 33–49

Kothari R, Dong M (2000) Decision trees for classification: a review and some new results. World Scientific, Singapore

Little R, Rubin D (1987) Statistical analysis with missing data. Wiley, New York. doi:10.1007/BF02925480

Liu H, Wu T (2003) Estimating the area under a receiver operating characteristic curve for repeated measures design. J Stat Softw 8:1–18

Medeiros CCAR, Braga C, de Souza WV, Regis L, Monteiro AMV (2011) Modeling the dynamic transmission of dengue fever: investigating disease persistence. PLoS Negl Trop Dis 5(1)

Metz C (1978) Basic principles of roc analysis. Sem Nucl Med 8:283–298

Middendorf M (2004) Predicting genetic regulatory response using classification. Bioinformatics 20:232–240

Monath TP (1994) Dengue: the risk to developed and developing countries. Proc Natl Acad Sci USA 91(7):2395–2400

Mussa A, Tshilidzi M (2006) The use of genetic algorithms and neural networks to approximate missing data in database. Comput Inform 24:1001–1013

Net DV (2011) Web site. http://denguevirusnet.com/dengue-virus.html

Nimmannitya S (1997) Dengue hemorrhagic fever: diagnosis and management. Dengue and dengue hemorrhagic fever. CAB International, Oxford

Ooi E, Gubler D, Nam V (2007) Dengue research needs related to surveillance and emergency response. Tech. rep., World Health Organization, Geneva

PAHO (2007) PAHO. Number of reported cases of dengue and dengue hemorrhagic fever (DHF) in the Americas, by country: figures for 2007 [database on the Internet]. Pan American PAHO, Washington

Paula ML, Claudia TC, Eduardo M, Jose SC (2003) Uncertainties regarding dengue modeling in Rio de Janeiro, Brazil. Mem Inst Oswaldo Cruz 98(7):871–878

Pearson R (2005) Mining imperfect data: dealing with contamination and incomplete records. SIAM, Philadelphia

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, Oxford

Pongsumpun P, Tang IM (2001) A realistic age structured transmission model for dengue hemorrhagic fever in Thailand. Southeast Asian J Trop Med Public Health 32(2):336–340

Qiao W, Gao Z, Harley R (2005) Continuous online identification of nonlinear plants in power systems with missing sensor measurements. In: IEEE international joint conference on neural networks, IEEE, Montreal, pp 1729–1734

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Francisco

Ramos MM, Tomashek KM, Arguello DF, Luxemburger C, Quiones L, Lang J, Muoz-Jordan JL (2009) Early clinical features of dengue infection in Puerto Rico. Trans R Soc Trop Med Hyg 103(9):878–884

Roderick JL, Donald BR (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Ron K, George HJ (1997) Wrappers for feature subset selection. Artif Intell 97:273–324

Saeys Y, Inza I, LarrANNaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Sa-Ngasang ASA, A-Nuegoonpipat A, Chanama S, Wibulwattanakij S, Pattanakul K, Sawanpanyalert P, Kurane I (2006) Specific IGM and IGG responses in primary and secondary dengue virus infections determined by enzyme-linked immunosorbent assay. Epidemiol Infect 134(4):820825

Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Sree Hari Rao V, Naresh Kumar M (2010) Estimation of the parameters of an infectious disease model using neural networks. Nonlinear Anal: Real World Appl 11(3):1810–1818

Sree Hari Rao V, Naresh Kumar M (2012) A new intelligence-based approach for computer-aided diagnosis of dengue Fever, IEEE Transactions on Information Technology in Biomedicine 16(1):112–118

Sree Hari Rao V, Naresh Kumar M (2011b) Novel algorithms for identification of influential features using particle swarm intelligence for effective diagnosis of dengue illness (preprint)

Sree Hari Rao V, Naresh Kumar M (2011c) Novel non-parametric algorithms for imputation of missing values and knowledge extraction in databases (preprint)

Sree Hari Rao V, Naresh Kumar M (2011d) Rule based approach for early diagnosis of dengue infection using clinical features for public health management (preprint)

Stephen SW, Joseph EB, Anna PD, Murphy BR (2007) Prospects for a dengue virus vaccine. Nat Rev Microbiol 5:518–528

Tanner L, Schreiber M, Low J, Ong A, Tolfvenstam T (2008) Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS Negl Trop Dis 2(3)

Vaughn DW, Green S, Kalayanarooj S, Innis BL, Nimmannitya S, Suntayakorn S, Rothman AL, Ennis FA, Nisalak A (1997) Dengue in the early febrile phase: viremia and antibody responses. J Infect Dis 176:322–330

Vaughn D, Green S, Kalayanarooj S, Innis B, Nimmannitya S (2000) Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. J Infect Dis 181(1):2–9

Wearing HJ, Rohani P (2006) Ecological and immunological determinants of dengue epidemics. Proc Natl Acad Sci USA 103(31):802–807

WHO (2009) Dengue-guidelines for diagnosis, treatment, prevention and control. Tech. rep., WHO, Geneva

Wilder-Smith A, Schwartz E (2005) Dengue in travelers. N Engl J Med 353:92432

Witten I, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco

Wong SL, Zhang LV, Tong AHY, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP (2004) Combining biological networks to predict genetic interactions. Proc Natl Acad Sci USA 101(44):15682–15687, http://www.pnas.org/content/101/44/15682.full.pdf+html

Yang Y, Webb GI (2001) Proportional k-interval discretization for naive-bayes classifiers. In: 12th European conference on machine learning. Springer. LNCS 2167:564–575

Yang Y, Webb IG (2002) A comparative study of discretization methods for nave Bayes classifiers. In: Proceedings of PKAW, Japan, pp 159–173

Zweig M, Campbell G (1993) Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 9(8):561–577