A Competent Approach to Predict Dengue Diseases using a Hybrid Approach in Machine Learning algorithm

R. Vijay Sai
Assistant Professor
Department of Computer Science and Engineering
K S Rangasamy College of Technology,
Tiruchengode, India

S Madhiarasi
Department of Computer Science and Engineering
K S Rangasamy College of Technology,
Tiruchengode, India.

S Keerthana
Department of Computer Science and Engineering
K S Rangasamy College of Technology,
Tiruchengode, India

S Preethi
Department.of Computer Science and Engineering
K S Rangasamy College of Technology,
Tiruchengode, India.

***Abstract*:- In this paper, we have a tendency to describe regarding the unwellness disease prediction by means that of feature choice for the info inheritable from the University of California Irvine repository. The aim of the work is to look at the performance of various classification techniques. A unwellness disease will cause severe damages to the society. Hence, it's crucial to predict a unwellness sickness ahead to minimalize the injury and loss caused by the disease. The clinical documents maintained area unit a pool of data relating to the infected patients. By keeping the voluminous knowledge will predict the long run occurrences of the sickness earlier and safe guard the folks. breakbone fever the world drawback is common in additional than a hundred and ten countries. breakbone fever infection has vulnerable a pair of.5 billion populations all round the world. per annum there area unit fifty million folks that suffer from it globally breakbone fever infectious sickness could be a vector borne disease caused by the feminine yellow-fever mosquito and Aedes albopictus mosquitoes that adapt well to human atmosphere. data processing could be a well-known technique employed by health organizations for classification and prediction of diseases. methodology ordered smallest improvement that will accurately predict unwellness sickness area unit greatly required and smart prediction techniques can facilitate to predict breakbone fever disease additional accurately. It uses 2 feature choice ways, forward choice and backward choice, to get rid of tangential options for up the results of unwellness disease prediction.**

INTRODUCATION

DATA MINING CONCEPT

Data Mining is an analytic process designed to explore data usually large amounts of data - typically business or market related in search of consistent patterns. the final word goal of information mining is prediction - and prognostic data processing is that the most typical variety of data processing and one that has the foremost direct business applications. the method of information mining consists of 3 stages:

(1) The initial exploration,

(2) Model building or pattern identification with validation/verification,

(3) Preparation.

1. **Exploration:** This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records. The process of data Mining may involve anywhere between a simple choice of straightforward predictors for a regression model.
2. **Model building and validation:** This stage involves considering various models and choosing the best one based on their predictive performance.
3. **Preparation:** That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

OVERVIEW OF DATA MINING

Data mining also called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

DATA MINING TECHNIQUES

The foremost unremarkably used techniques in data processing are:

- Artificial neural networks: Non-linear prognostic models that learn through coaching and fit biological neural networks in structure
- Decision trees: Dendroid structures that represent sets of selections. These choices generate rules for the classification of a dataset. Specific call tree ways embrace Classification and Regression Trees (CART) and Chi sq. Automatic Interaction Detection (CHAID)
- Genetic algorithms: Improvement techniques that use processes, like genetic combination,

mutation, and survival of the fittest in a very style supported the ideas of evolution

- Nearest neighbour method: A method that classifies every record in a very dataset supported a mixture of the categories of the k record most kind of like it in a very historical dataset (where k1). generally referred to as the k-nearest neighbour technique
- Rule induction: The extraction of helpful if-then rules from knowledge supported applied mathematics significance

DENGUE SICKNESS MINING

Disease ranks second as a reason for cancer death in ladies, following closely behind carcinoma. Statistics recommend the likelihood of diagnosis nearly a pair of.5 lakhs new cases in Asian nation by the year 2015. Prognosis therefore takes up a big role in predicting the course of the sickness even in ladies World Health Organization haven't succumbed to the sickness however area unit at a larger risk to. Classification of the character of the sickness supported the predictor options can change oncologists to predict the likelihood of incidence of breakbone fever sickness for a brand new case. The dismal state of affairs wherever additional folks area unit assent to the sway of breakbone fever sickness, in spite of outstanding advancement in clinical science and medical aid is definitely troubling. The motivation for analysis on classification, to accurately predict the character of unwellness disease.
This projected work primarily focuses on building associate degree economical classifier for the Wisconsin Prognostic breakbone fever sickness (WPDC) knowledge set from the UCI machine learning repository.

UNSUPERVISED OPTIONS

The third class extracts unattended options from distributions in large-scale unlabelled corpora. Such studies embrace Riloff (1996), Yangarberetal (2000). These options area unit used once there's not abundant coaching knowledge, or the coaching and testing knowledge has totally different distribution.
During this thesis, the primary investigate a way to extract supervised and unattended options to enhance a supervised baseline system. Compared to a supervised multi-label classifier, the unattended approach can do comparable, even higher.

AUTOMATIC CONTENT EXTRACTION (ACE) ANALYSIS

ACE began in 2000 when MUC. the target of the ACE program is to develop automatic content extraction technology to support automatic process of human language in text type from a spread of sources like newswire, broadcast voice communication, and weblogs. ACE technology R&D is aimed toward supporting numerous classifications, filtering, and choice applications by extracting and representing language content the that means sent by the info.
Entity Detection and Recognition (EDR) is that the core annotation task of ACE, providing the muse for all remaining tasks.
This ACE task identifies seven forms of entities: Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity (GPS).
Problem Formulation: Typically, the particular refinement is split into 2 phases, referred to as the match and therefore the refine phases. In NMS, such a natural evolution method of nuggets can even be controlled by users.
Match phase: During this phase, the aim is to match the known nuggets with patterns around them among the info house.
Refinement Phase: The match part reveals to U.S. what variety of patterns that a user was finding out. With this information, it will end hunk refinement victimisation the 2 steps of rending and modification. These 2 steps can build every hunk an ideal representative of one pattern. the info utilized in the study area unit provided by the UCI Machine learning repository settled in breakbone fever Wisconsin sub-directory, filenames root: Dengue-Wisconsin having 699 instances, a pair of categories (malignant and benign), and nine integer-valued attributes. The program offers a well outlined framework for experimenters and developers to make and assess their models. The results show clearly that the projected methodology performs well compared to alternative similar ways within the literature, taking into the actual fact that the attributes taken for analysis don't seem to be direct indicators of breakbone fever sickness within the patients.

EXISTING SYSTEM:

Undesirable impact of fixing a breakbone fever patients existing check knowledge teams, probably undoing the patients own manual efforts in organizing her history. it involves a high process value. got to repeat an outsized range of attribute check knowledge cluster similarity computations for each new check knowledge.

As existing approaches to extract unwellness disease prediction suffer from measurability. it's imperative to deal with the measurability issue. connections in breakbone fever prediction don't seem to be same. Dengue could be a threatening sickness caused by feminine mosquitos.from long periods of your time, specialists try to seek out out a number of options on unwellness disease in order that they will justly reason patients as a result of totally different patients need differing types of treatment. pakistan has been target of unwellness disease from previous
few years. dandy fever is employed in classification techniques to guage and compare their performance. the dataset was collected from district headquarter hospital (dhq) jhelum. for correctly categorizing our dataset, totally different classification techniques area unit used. these techniques area unit nave bayesian, rep tree, random tree, j48 and smo. rail was used as data processing tool for classification of information. first of all we'll assess the performance of all the techniques individually with the assistance of tables and graphs relying upon dataset and second we'll compare the performance of all the techniques.

Dengue infection has vulnerable a pair of 2.5 billion populations all round the world. per annum there area unit fifty million folks that suffer from it globally. pakistan has been victim of this quickly growing illness from previous few years. since 2007 in pakistan, sizable amount of cases was marked particularly in lahore. in 1994 at city pakistans initial case of breakbone fever was appeared and dengues happening in 2011, that was preceding years and 1400 folks were affected

PROPOSED SYSTEM

Methods that will accurately predict breakbone fever sickness area unit greatly required and smart prediction techniques can facilitate to predict breakbone fever sickness additional accurately. during this system, it used 2 feature choice ways, forward choice (FS) and backward choice (BS), to get rid of tangential options for up the results of breakbone fever sickness prediction. The results show that feature reduction is helpful for up the prognostic accuracy and density is tangential feature within the dataset wherever the info had been known on full field digital mammograms collected at the UCI Repository. additionally, call tree (DT), support vector machine-sequential smallest improvement (SVM-SMO) and their ensembles were applied to resolve the breakbone fever sickness diagnostic drawback in a shot to predict results with higher performance. The results demonstrate that ensemble classifiers area unit additional correct than one classifier.

The projected framework SMO supported sickness prediction is shown to be effective in addressing this prediction. The framework suggests a completely unique means of network classification: initial, capture the latent affiliations of actors by extracting sickness prediction supported network property, and next, apply extant data processing techniques to classification supported the extracted prediction. within the initial study, modularity maximization was utilized to extract sickness prediction. the prevalence of this framework over alternative representative relative learning ways has been verified with breakbone fever prediction breakbone fever knowledge.Prove that with this projected approach, insufficiency of sickness prediction is warranted.

SMO-Sequential Minimal Optimization

Classification is that the variety of data processing, that deals with the problematic things by recognizing and detective work options of infection, among patients and forecast that that technique shows prime performance, on the bottom of WEKAs outcome. 5 techniques are utilized in this paper. These techniques uses individual interface and it depends on dissimilar techniques NB, REP Tree, RT, J48 and SMO. All techniques, that we have a tendency to used, were applied on a dataset of dandy fever, as enlightened on top of. Classification and accuracy used was mentioned.

MODULE DESCRIPTION

DATA VISUALIZATION AND PRE-PROCESSING

The Wisconsin Prognostic Dengue Disease dataset is downloaded from the UCI Machine Learning Repository website and saved as a text file. This file is then imported into Excel spread-sheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values. The ID of the patient cases does not contribute to the classifier performance. Hence it is removed and the outcome attribute defines the target or dependent variable thus reducing the feature set size to 33 attributes. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented in the following sections.

SMO FEATURE SELECTION ALGORITHMS

The generic problem of supervised feature selection can be outlined as follows. Given a data set {(xi, yi)} ni = 1 where xi ∈ Rd and yi ∈ {1, 2…c}, Our aim is to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the WPDC dataset are briefly outlined below.

Mean and STD Score Filtering:

It is termed Univariate Mean and STD Score's ANOVA ranking. It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the Dengue Disease (recurrent/non- recurrent) and the predictor attributes. After computing the Mean and STD Score for each feature, it selects the top-m ranked features with large scores. The next subsection directs focus on another technique of feature selection based on logistic regression.

LEVERAGE BACKWARD LOGISTIC REGRESSION RISK ANALYSIS

When the number of descriptors is very large for a given problem domain, a learning algorithm is faced with the problem of selecting a relevant subset of features backward regression includes regression models in which the choice of predictor variables is carried out by an automatic procedure. The iterations of the algorithm for logistic regression are given in steps as stated as follows.

* The feature set with all 'ALL' predictors.
* Eliminate predictors one by one.
*'ALL' models are learnt containing 'ALL-1' descriptor each.

These iterations are further continued till either a pre-specified target size is reached or the desired performance statistics (classification accuracy) is obtained. After feature relevance, it classifies the nature of the Dengue Disease cases in the Wisconsin Prognostic Dengue Disease dataset using twenty classification algorithms.

FEATURE REDUCTION BY SMO

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as pre-processing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly Contributes to classification that cuts pre-processing costs and minimizes the effects of the 'peaking phenomenon' in classification. Thereby improving the overall performance of classifier based intrusion detection systems. The commonly used dimensionality reduction methods include supervised approaches such as Linear Discriminant Analysis (LDA), unsupervised ones such as SMO, and additional spectral and manifold learning methods. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. Consider the two dimensional cases then the basic principle of transformation.

CONCLUSION

The analysis check totally different algorithms. The results of the analysis targeted on correctness of the algorithms within the coaching. It trusted WDBC knowledge set. The check result shows that the SMO is that the best formula. the most effective means was once the analysis removed the sample for missing worth in coaching for SMO. However, Random Tree result was keep higher correctness once keeps the sample for missing worth. The analysis undertook associate degree experiment on application of assorted data processing algorithms to predict the breakbone fever and to check the most effective methodology of prediction. The analysis results don't gift dramatic variations within the prediction once victimisation totally different classification algorithms in data processing.The experiment will function a vital tool for physicians to predict risky cases within the observe and advise consequently. The model from the classification are ready to answer additional complicated queries within the prediction of breakbone fever diseases. The prognostic accuracy determined by SMO formula suggests that parameters used area unit reliable indicators to predict the presence of breakbone fever diseases.

ACKNOWLEDGEMENT

We Acknowledge DST- File No.368. DST – FIST (SR/FIST/College-235/2014 dated 21-11-2014) for financial support and DBT – STAR College – Scheme - ref.no: BT/HRD/11/09/2018 for providing infrastructure support.

REFERENCES

[1] Aziz MM, Hasan KN, Hasanat MA, Siddiqui MA, Salimullah 30. M, Chowdhury AK, et al. Predominance of the DEN 3 genotype during the recent dengue outbreak in Bangladesh. Southeast Asian J Trop Med Public Health 2015; 33 : 428.

[2] Blackburn G L, Wang K A (2013) "Dietary fat reduction and Dengue Disease outcome: results from the Women's Intervention Nutrition Study (WINS)", IEEE International Journal of Computer Science and Engineering, vol. 32, pp.512

[3] Boffetta P, Hashibe M (2011). "The burden of cancer attributable to alcohol drinking", IEEE Journal of Cancer Research and Treatmentvol.90, pp.119.

[4] Boris Pasche (2010). "Cancer Genetics", (Cancer Treatment and Research). Berlin: Springer. pp. 19–20.

[5] Collaborative Group on Hormonal Factors in Dengue Disease (2012), "Dengue Disease and breastfeeding", IEEE International Journal for Cancer Research and Treatment, vol.15.

[6] Cummings DAT, Irizarry RA, Huang NE, Endy TP, Nisalak 27. A, Ungchusak K, et al. travelling waves the occurrence of dengue haemorrhagic feverinThailand.Nature2014; 427 : 344-7.

[7] Delen D, Walker G, (2015), "Predicting Dengue Disease survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, vol. 34, pp. 113-127.

[8] Ferro, Roberto (2012), "Pesticides and Dengue Disease ", IEEE International Journal for Cancer Research and Treatment, vol.76.

[9] Gage M, Wattendorf, D (2012), "Translational advances regarding hereditary Dengue Disease syndromes". IEEE International Journal of Computer Science and Engineering, vol. 90.

[10] Grey N and Sener S. (2013), "Reducing the global cancer burden", IEEE Journal of Computer Science and Engineering. Vol.45, pages 201-202.