Dengue Prediction Using (MLP) Multilayer Perceptron - A Machine Learning Approach

V. Janani, N. Maadhuryaa, D. Pavithra and S. Ramya Sree

January 22, 2020

DENGUE PREDICTION USING MLP (MULTILAYER PERCEPTRON) – A MACHINE LEARNING APPROACH

#Dr. V. Janani M.E., Ph.D. *N. Maadhuryaa *D. Pavithra *S. Ramya Sree
#Associate Professor *Final Year B.E (CSE)
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ADHIYAMAAN COLLEGE OF ENGINEERING
HOSUR, TAMILNADU, INDIA
vajjiram.janani@gmail.com
maadhuryaanaganathan@gmail.com
saipavithra1998@gmail.com
ramyasubramanyam98@gmail.com

ABSTRACT:

Machine learning (ML) is the application of Artificial Intelligence (AI) that provides systems the ability to learn automatically from experience. The primary aim of ML is to allow computers learn without human intervention or assistance. It easily identifies trends and patterns, continuous improvement, handling multidimensional and multi-variety data. Dengue is the most virally occurring mosquito-borne disease in recent days. The Multilayer Perceptron (MLP) algorithm in Machine Learning is used to achieve accuracy in analysing and predicting dengue disease. The parameters of the dengue integrated model are identified using an optimization-based methodology in multiple stages. The prediction of dengue using the MLP algorithm is carried out by three phases. The initial phase of dengue prediction is data visualization and pre-processing which is implemented by SMO (Sequential Minimal Optimization).SMO is an algorithm for solving the quadratic programming problem that arises during the training of Support Vector Machines (SVM).The second phase is the MLP feature selection algorithm which includes a leverage backward logistic regression risk analysis. The final phase includes feature reduction by MLP is a part of dimensionality reduction in the dataset. The implementation of dengue prediction is done by the WEKA tool. WEKA tool is a collection of machine learning algorithms for data mining tasks. This tool can be applied to a dataset directly or called from the java code and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Keywords *– Data Mining (DM), Machine Learning (ML), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Sequential Mining Optimization (SMO), WEKA tool.*

I. INTRODUCTION TO DATA MINING CONCEPT

Data Mining is an analytic process. It is designed to explore large amount of dataset typically business or market-related in search of consistent patterns. The concept of data mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques. Data mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modelling and it shares with them both some components of its general approaches and specific techniques Accordingly the ultimate goal of data mining is prediction - and predictive is the most common type of data mining and one that has the most direct business applications. The process of DM consists of three stages:

(1) The initial exploration,

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

(3) Deployment.

II. LITERATURE REVIEW

Natural Medical Data Processing has attracted substantial attention in many applications and research areas. Many of the existing approaches are based on calculating distances among the points in the dataset. Besides, current datasets usually have a large number of dimensions. These datasets tend to be sparse, and traditional concepts such as Euclidean distance or nearest neighbor become unsuitable.

DIETARY FAT REDUCTION AND DENGUE DISEASE OUTCOME: RESULTS FROM THE WOMEN'S INTERVENTION NUTRITION STUDY (WINS)

In this paperwork [1] Blackburn G L has proposed An algorithm to perform outlier detection on time-series data is developed, the intelligent outlier detection algorithm (IODA). This treats a time series as an image and segments the image into clusters of interest, such as ‘‘nominal data’’ and ‘‘failure mode’’ clusters. The algorithm uses density clustering techniques to identify sequences of coincident clusters in both the time domain and delay space, where the delay space representation of the time series consists of ordered pairs of consecutive data points taken from the time series. ‘‘Optimal’’ clusters that contain either mostly nominal or mostly failure-mode data are identified in both the time domain and delay space.

PREDICTING DENGUE DISEASE SURVIVABILITY: A COMPARISON OF THREE DATA MINING METHODS

In this paperwork [5] Delen D has proposed it proves an upper bound for the memory consumption which permits the discovery of all outliers by scanning the dataset 3 times. The upper bound turns out to be extremely low in practice. Since the actual memory capacity of a realistic DBMS is typically larger, we develop a novel algorithm, which integrates our theoretical findings with carefully designed heuristics that leverage the additional memory to improve I/O efficiency. The technique reports all outliers by scanning the dataset at most twice and significantly outperforms the existing solutions by a factor up to an order of magnitude.

TRANSLATIONAL ADVANCES REGARDING HEREDITARY DENGUE DISEASE SYNDROMES

In this paperwork [7] Gage M has proposed Support Vector Machines (SVMs) suffer from an O (n2) training cost, where n denotes the number of instances. In this system, propose an algorithm to select boundary instances as training data to substantially reduce n. The algorithm eliminates instances that are likely to be non-support vectors. In the concept independent pre-processing step of our algorithm, we prepare nearest-neighbor lists for training instances. In the concept-specific sampling step, then effectively select useful training data for each target concept. Empirical studies show our algorithm to be effective in reducing n, outperforming other competing down sampling algorithms without significantly compromising testing accuracy.

DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR DENGUE DISEASE DIAGNOSIS AND PROGNOSIS

In this paperwork [9] Gupta has proposed the Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge of observed data. Let X is a data sample containing instances, Xi where each instance is the Dengue Disease risk factors (modifiable and non-modifiable). Let H be a hypothesis that X belongs to class C which contains (unlikely, likely and benign cases). Classification is to determine P (Hj|X), (i.e., posterior probability): the probability that the hypothesis, Hj (unlikely, benign or likely) holds given the observed data sample X.

III. EXISTING SYSTEM

The classification is based on only the undesirable effect of changing a dengue patient's existing test data groups, potentially undoing the patient's manual efforts in organizing her history. It involves a high computational cost, have to repeat a large number of attribute test group.

Dengue fever is used in classification techniques to evaluate and compare their performance. The dataset was collected from District Headquarter Hospital (DHQ) Jhelum. For properly categorizing our dataset, different classification techniques are used. These techniques are Naïve Bayesian, REP Tree, Random tree, J48, and SMO. WEKA was used as Data mining tool for classification of data

Motivate and propose a method to perform test data grouping dynamically. Our goal is to ensure good performance while avoiding disruption of existing patient-defined test data groups

- Improper classification may provide wrong results
- Poor performance
- Complex data processing to find dengue prediction
- The data retrieval based on user requirement is not done
- This relation-type information, however, is often not readily available in dengue prediction.

IV. PROPOSED SYSTEM

In this system, it used two feature selection methods, Forward Selection (FS) and Backward Selection (BS), to remove irrelevant features for improving the results of Dengue Disease prediction. The results show that feature reduction is useful for improving the predictive accuracy and density is an irrelevant feature in the dataset where the data had been identified on full-field digital mammograms collected at the UCI (Unique Client Identifier) Repository. Also, Decision Tree (DT), support vector machine—sequential minimal optimization (SVM-SMO) and their ensembles were applied to solve the Dengue Disease diagnostic problem in an attempt to predict results with better performance.

The proposed framework SMO based on disease prediction is shown to be effective in addressing this prediction. The framework suggests a novel way of network classification: first, capture the latent affiliations of actors by extracting disease prediction based on network connectivity, and next, apply extant data mining techniques to classification based on the extracted prediction.

The superiority of this framework over other representative relational learning methods has been verified with dengue prediction dengue data. Prove that with this proposed approach, the sparsity of disease prediction is guaranteed.

SVM

The two well-performing feature selection algorithms on the dataset are briefly outlined below.

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. SVM is a linear transformation with linear orthonormal basis vectors; it can be expressed by a translation and rotation.

SMO

Classification is the type of Data mining, which deals with the problematic things by recognizing and detecting features of infection, among patients and forecast that which technique shows top performance, on the base of WEKA's outcome. Five techniques have been used in this paper. These techniques use Explorer interface and it depends on dissimilar techniques NB, REP Tree, RT, J48, and SMO.

- Test data reformulation graph and the click graph into a single graph that it refers to as the test data fusion graph, and by expanding the test data set when classification relevance occur
- High Relevance Measure
- Good classification accuracy
- The online test data grouping process. High Similarity function provides types of dengue data classification
- It supports multiple disease classification with better accuracy.

V. ARCHITECTURE DIAGRAM

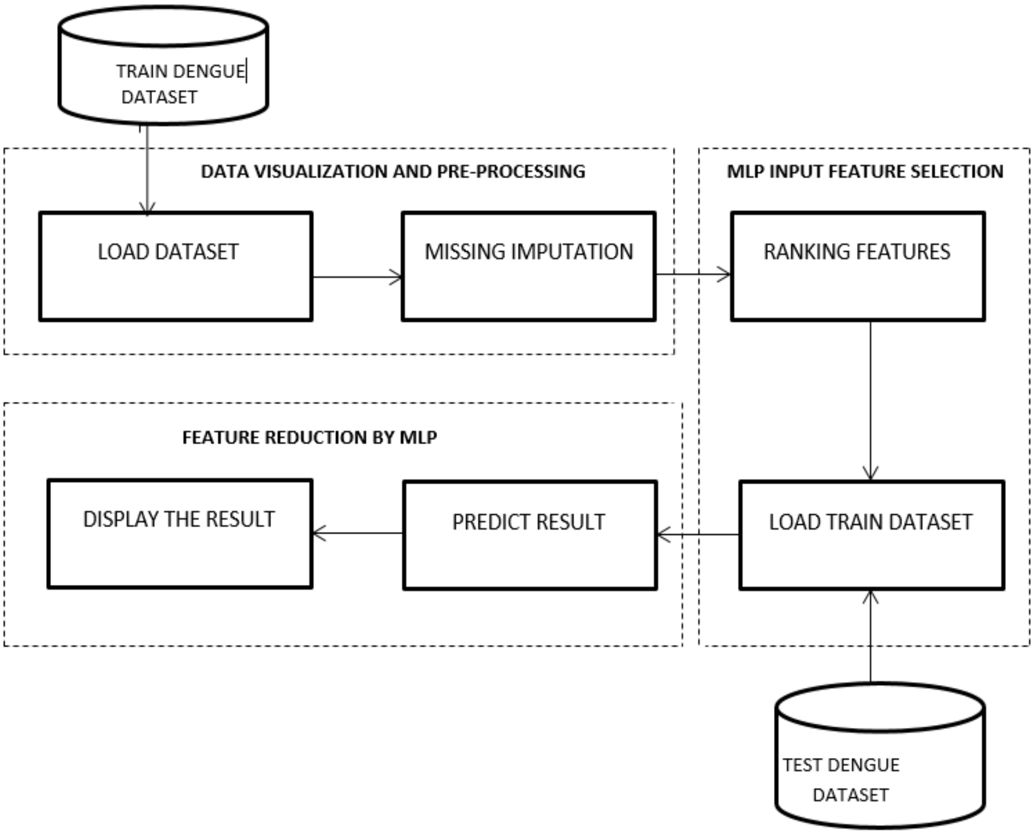

Figure 1. Dengue Prediction Using Multilayer Perceptron (MLP)

In Figure 1, the dengue dataset is trained and loaded. The missing data are analyzed and filled with data visualization and pre-processing. Using the MLP input feature selection algorithm the ranking feature of the dataset is loaded and tested. The result is displayed using feature reduction using MLP.

VI. MODULES

- Data visualization and pre-processing.
- SMO feature selection algorithms.
- Feature reduction by SMO.

MODULE DESCRIPTION

DATA VISUALIZATION AND PRE-PROCESSING

The Dengue Disease dataset is downloaded from the UCI (Unique Client Identifier) Machine Learning Repository website and saved as a text file. This file is then imported into an Excel spreadsheet 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. As a result it is removed and the outcome attribute defines the target by 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 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}, we aim to find a feature subset of size m which contains the most informative features. It is termed Uni-variate 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. 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 by k-mean determination. The algorithm for logistic regression iterations are given in steps as stated as follows.
1. The feature set with all 'ALL' predictors.
2. Eliminate predictors one by one.
3. ALL models are learned to contain 'ALL - 1' descriptor each.
These iterations are further continued till either a pre-specified target size is reached. Then the desired performance statistics is obtained. After feature relevance, it classifies the nature of the Dengue Disease cases in the Dengue Disease dataset using classification algorithms. The better performing algorithms are described in the following section.

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.

ALGORITHM

The following steps illustrate SMO as follows,
Consider the input: G = set of antigens to be recognized, n the number of worst elements to select for removal
1. Create an initially random set of attributes, A for all antigens in G do
2. Determine the predicted with each attribute in A.
3. Generate clones of a subset of the attribute in A with the highest predicted.
4. The number of clones for an attribute is proportional to its predicted.
5. Mutate attributes of these clones to the set A, and place a copy of the highest predicted
6. Attribute in A into the memory set, M.
7. Replace the n lowest predicted attributes in A with new randomly generated attributes.
8. End
Output: M = set of memory attributes capable of classifying unseen patterns.

VII. CONCLUSION

The research test different algorithms. The result of the research focused on the correctness of the algorithms in the training. It depended on the WDBC data set. The test result shows that SMO is the best algorithm. The best way was when the research removed the sample for the missing value in training for SMO. However, the Random Tree result kept better correctness when keeps the sample for the missing value. The research undertook an experiment on the application of various data mining algorithms to predict the dengue and to compare the best method of prediction. The research results do not present dramatic differences in the prediction when using different classification algorithms in data mining.

The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of dengue diseases. The predictive accuracy determined by the SMO algorithm suggests that the parameters used are reliable indicators to predict the presence of dengue diseases.

VIII. REFERENCES

[1] Blackburn G L, Wang K A (2007) "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.

[2] Boffetta P, Hashibe M (2006). "The burden of cancer attributable to alcohol drinking", IEEE Journal of Cancer Research and Treatmentvol.90, pp.119.

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

[4] Collaborative Group on Hormonal Factors in Dengue Disease (2002), "Dengue Disease and breastfeeding", IEEE International Journal for Cancer Research and Treatment, vol.15.

[5] Delen D, Walker G, (2005), “Predicting Dengue Disease survivability: a comparison of three data mining methods”, Artificial Intelligence in Medicine, vol. 34, pp. 113-127.

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

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

[8] Grey N and Sener S. (2006), “Reducing the global cancer burden”, IEEE Journal of Computer Science and Engineering. Vol.45 pages 201-202.

[9] Gupta, S.; Kumar, D., Sharma, A (2011). “Data Mining Classification Techniques Applied for Dengue Disease Diagnosis and Prognosis”. IEEE Journal of Computer Science and Engineering. Vol.23 pages 1191-1193.

[10] Hendrick, RE (2010). "Radiation doses and cancer risks from breast imaging studies”, IEEE International Journal for Cancer Research and Treatment, vol. 57.

[11] Johnson KC, Miller AB, (2009). "Active smoking and secondhand smoke increase Dengue Disease risk: the report of the Canadian Expert Panel on Tobacco Smoke and Dengue Disease Risk (2009).”IEEE Journal of Computer Science and Engineering. Vol.69, pages 198-199.

[12] Rico-Hessea R, Harrisona LM, Salasb RA, Tovarb D, Nisalak 31. A, Ramosd C, *et al*. Origins of dengue type 2 viruses associated with increased pathogenicity in the Americas. *Virology* 1997.

[13] 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 2002.