

DENGUE FEVER PREDICTION USING CLASSIFICATION TECHNIQUES

R. Sanjudevi¹, D. Savitha²

¹Research scholar, M.Phil. Computer Science, Vellalar College for Women, Erode 12.

²Assistant Professor, Department of Computer Applications, Vellalar College for Women, Tamilnadu, India.

Abstract - Data mining is the process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics, and systems. Data mining is the analysis stage of the "knowledge discovery in databases" process, or KDD. The dengue risk disease accounts to be the leading cause of death worldwide. It is problematic for medical consultants to predict the dengue risk attack as it is an intricate task that requires experience and knowledge. The health sector today contains concealed information that can be important in making decisions. Dengue fever dataset is used in classification techniques to assess and compare their performance. For properly categorizing our dataset, the proposed method is compared with the decision tree. WEKA was used as Data mining tool for organization of data. This thesis implemented a feature model construction and comparative analysis for improving prediction accuracy of dengue disease in three phases. In first phase, dengue disease data sets are collected from UCI repository. In second phase, feature selection is done by forward and backward step wise regression methods. Third phase, SVM and decision tree classification algorithms are applied on the data set. In the fourth phase, the accuracy determination be calculated using Sensitivity (SE), Specificity (SP), and Area under the curve (AUC). Finally, the evaluation is done based on accuracy values. Thus outputs shows that proposed classification algorithm SVM gives better accuracy than decision tree algorithm with the help of feature selection.

Key words: Classification, SVM-SMO, Decision Tree.

1. INTRODUCTION

Dengue fever is diseases caused by a family of that are transmitted by mosquitoes. Includes severe joint and, exhaustion. The presence of fever is characteristic of dengue fever. Dengue is prevalent throughout the tropics and subtropics. Dengue fever is caused by a virus, and there is no specific medicine or antibiotic to treat it. For typical dengue fever, the treatment is directed toward relief of the symptoms (symptomatic treatment). Leaf extract can be used to treat dengue fever. Dengue fever (DF) is a specific syndrome that tends to affect children under 10 years of age. This complication of dengue causes, hemorrhage (bleeding), and circulatory collapse. The prevention of dengue fever requires control or eradication of the mosquitoes carrying the virus that causes dengue. The dengue fever was approved in April 2016 for use in dengue endemic areas.

1.1 Characteristics of Data Mining

Data Mining is an logical process designed to explore data usually large amounts of data typically business or market related in search of consistent patterns and systematic relationships between variables, and then to authorize the findings by applying the detected patterns to new subsets of data. The final goal of data mining is prediction and predictive data mining is the most common type of data mining and one that has the most direct business applications.

The procedure of data mining consists of three stages:

- (1) The initial exploration,
- (2) Model building or pattern identification with validation/verification,
- (3) Deployment.

1.2 NEED FOR DENGUE DISEASE MINING

Statistics suggest the possibility of diagnosing nearly 2.5 lakhs new cases in India by the year 2015. Prognosis thus takes up a significant role in predicting the course of the disease even in women who have not succumbed to the disease but are at a greater risk. Classification of the nature of the disease based on the predictor features will enable oncologists to predict the possibility of occurrence of dengue disease for a new case. The dismal state of affairs where more people are conceding to the influence of dengue disease, in spite of remarkable advancement in clinical science and thereby is certainly worrying. The motivation for research on classification, to accurately predict the nature of dengue disease.

1.3 PROBLEMS IN DENGUE FEVER PREDICTION

The problem of this kind of study can help researchers make a good prediction of the number of dengue cases at real-time even though the people who chirped did not go to hospitals or get the work should focus the effectiveness of applying machine learning techniques in the field of public health. Artificial intelligence based applications could improve health outcomes and the quality of life for millions of people in the coming years. The general objective of this research is to use two classification techniques to determine the dengue fever. Objective of this study also include the comparison of two classification algorithms based on the dataset [4].

1.4 Objective

The objective of the dengue fever prediction is to

- Pre-processing the dataset.
- Applying classification techniques to find whether the patient is affected by dengue or not by using some attributes.
- On the basis of results, it shows the accuracy of classification techniques.

2. LITERATURE SURVEY

BuchadeOmkar, et al [2] 2017 described the tests are conducted on the blood samples collected from the patients. In our proposed system we are stratifying dengue into Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF) and healthy patients. Existing work uses PSO approach which achieved the accuracy of 90.91%, in order to achieve high accuracy, we are using optimization algorithms like Spider Monkey Optimization (SMO) and to increase the optimality of the model, we have also used Probabilistic Neural Network (PNN). PNN uses feed forward technique for classification.

Bhavani, M., et al [3] 2016 has used the dataset consist of attribute like fever, bleeding, metallic taste, Fatigue. The main objective of this study is to calculate the performance of various classification Techniques and compare their performance. The classification techniques used in this study are REP Tree, J48, SMO, ZeroR and Random Tree. The performance of classification techniques were compared by plotting graphs and table. Weka the data mining tool is used for the classification.

RosstichaAnjarKesuma Tazkia, et al [4] 2015 has analyzed the performance of DEWS by testing its accuracy of predictions, using data of environmental factors, climate and surveillance in District CempakaPutih. Naïve Bayes was chosen as Dengue outbreak predictor. Through the process of selecting a subset of attributes (Feature Subset Selection) with exhaustive search approach and Naïve Bayes accuracy as feature subset quality evaluation criteria, as the result we identified four attributes that contributed significantly to the prediction accuracy. The four attributes are house density, free larvae index, container potential nest larvae, and average rainfall in the last 2 months. The system achieved an accuracy of 97.05% in term of Geometric Mean. Further error analysis revealed that the sensitivity, specificity, Positive Predicted Value, and F1 of the system were 94.52%, 99.65%, 98.57% and 96.50%, respectively.

Kamran Shaukat, et al [5] 2015 has used 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. Firstly we will evaluate the

performance of all the techniques separately with the help of tables and graphs depending upon dataset and secondly we will compare the performance of all the techniques.

Daranee Thitiprayoonwongse, et al [6] 2014 described the experiments. It is divided into 4 parts. The first two experimental results show the useful knowledge to classify dengue infection from Srinagarindra Hospital's dataset and Songklanagarind Hospital's dataset, respectively. Each set of knowledge is tested by different dataset to make sure that the test data was a real unseen data. The third experimental results show the useful knowledge when we integrated 2 datasets. Another objective of this research is to detect the day of defervescence of fever which is called day0. The day0 date is the critical date of dengue patients that some patients face the fatal condition. Therefore the physicians need to predict day0 in order to treat the patients. They expect to have an intelligent system that can trigger the day0 date of each patient.

Subitha, N., et al [7] 2013 has used k-means algorithm randomly selects k number of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then it computes new mean for each cluster. This process iterates until the criterion function converges. The above concept is applied in the area of image segmentation where to apply the microscopic blood image as input and signals are filtered with the help of neural network to predict the best result about dengue fever.

Gage M, et al [8] 2012 has applied Support Vector Machines (SVMs) algorithm with $O(n^2)$ training cost, where n denotes the number of training instances. In this system, proposed an algorithm to select boundary instances as training data to substantially reduce n. The algorithm is motivated by the result of that, removing non-support vectors from the training set does not change SVM training results. The algorithm eliminates instances that are likely to be non-support vectors. In the concept independent preprocessing 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.

3. SYSTEM DESIGN

3.1 OVERVIEW OF DENGUE FEVER PREDICTION MODEL

Methods that can accurately predict Dengue Disease are greatly needed and good prediction techniques can help to predict Dengue Disease more accurately. 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 irrelevant feature in the dataset where the data had been identified on full field digital dengue collected at the UCI Repository. In addition, 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 accuracy. The results demonstrate that collective classifiers are more accurate than a single classifier [4].

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 the data mining techniques to classification based on the extracted prediction. In the initial study, modularity maximization was employed to extract disease 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, sparsely of disease prediction is guaranteed.

3.2 SMO ALGORITHM

SMO is techniques used for QP problem that rises during the training of SVM. This techniques mainly used for the training of SVM [2]. This techniques used on the dataset for splitting the data. The SMO techniques applied on training set by using WEKA tool and all the procedures of implementation is within it.

3.3 SVM CLASSIFICATION ALGORITHM

Sequential Minimal Optimization (SMO) is a new algorithm for training Support Vector Machines (SVMs). The SMO algorithm proposed by John Platt in 1998, is a simple and fast method for training a SVM. The main idea is derived from solving dual quadratic optimization problem by optimizing the minimal subset including two elements at each iteration. The advantage of SMO is that it can be implemented simply and analytically. Training a support vector machine requires the solution of a very large quadratic programming optimization problem. SMO breaks this large quadratic programming problem into a series of smallest possible quadratic programming problems [7].

The small quadratic programming problems are solved analytically, which avoids using a time-consuming numerical quadratic programming optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. The two well performing feature selection algorithms on the dataset are briefly outlined below. Feature discount 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 preprocessing to machine learning and statistics tasks of prediction, including pattern recognition.

4. METHODOLOGY

The processing steps of the proposed system are summarized as,

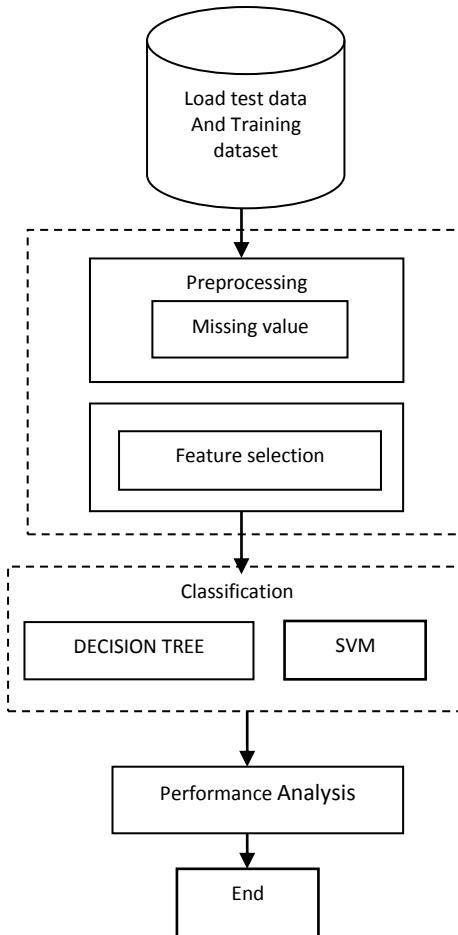
Step 1: Load test data and training dataset

Step 2: The preprocessing is done by using missing value imputation method.

Step 3: These feature selection are done using forward and backward selection method.

Step 4: The classification algorithm is used to predict the dengue disease.

Step 5: Based on the result, it will shows the classification algorithm accuracy.



4.1 IMPLEMENTATION TOOL

WEKA

WEKA is a data mining classification developed by the University of Waikato in New Zealand that implements data mining procedures. WEKA is a state of heart facility for developing machine learning (ML) methods and their application to real-world data mining problems. It is a gathering of machine learning algorithms for data mining

tasks. The procedures are applied directly to a dataset. WEKA implements procedures for data preprocessing, classification, regression, clustering, association rules; it also contains a visualization tools. The new machine learning systems can also be developed with this package. WEKA is open source software delivered under the GNU General Public License.

5. RESULTS AND DISCUSSIONS

5.1 DENGUE DISEASE DATA SET

Dataset is a collection of data or a single statistical data where every attribute of data represents variable and each instance has its own description. For prediction of dengue disease is used dengue data set for prediction and classification of algorithms in order to compare their accuracy using WEKA three interfaces: Explorer, Experimenter and Knowledge Flow. The dengue fever dataset contains 18 attributes and 108 instances for dengue disease classification and accuracy. To applied

classification algorithms using WEKA data mining tool for analysis purpose.

5.2 EVALUATION METRICS

Metrics used to evaluate the performance of the system are sensitivity, specificity, area under curve. In order to evaluate the efficiency and robustness of the system and the metrics are as follows:

- **True Positives (TP):** Predicted as yes (they have the dengue fever).
- **True Negatives (TN):** Predicted as No and they don't have the disease.
- **False Positives (FP):** predicted as yes, but they don't actually have the disease.
- **False Negatives (FN):** predicted as No, but they actually do have the disease.

This situation can be depicted as a confusion matrix also called contingency table as shown in table 5.1 below.

Table 5.1 Confusion matrix

Actual	Dengue Positive	Dengue Negative
Dengue Positive	TP	FN
Dengue Negative	FP	TN

5.3 PERFORMANCE EVALUATION

Definition of Sensitivity: Sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate. It is represented in equation as,

$$\text{Sensitivity (SE)} = [\text{TP} / (\text{TP} + \text{FN})] * 100\%$$

Definition of Specificity: Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate. It is represented in equation as,

$$\text{Specificity (SP)} = [\text{TN} / (\text{FP} + \text{TN})] * 100\%$$

Definition of Area under Curve: Area under the curve (AUC) is the definite integral in a plot of drug concentration in blood plasma vs. time. In practice, the drug concentration is measured at certain discrete points in time and the trapezoidal rule is used to estimate AUC. It is represented in equation as,

Area =

$$\int_a^b f(X) dx \int_a^b f(X) dx$$

Definition of Accuracy: Accuracy (A) is defined as ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate, False Positive Rate. It is represented in equation as,

$$\text{Accuracy (A)} = [(\text{TP} + \text{TN}) / (\text{TP} + \text{FN} + \text{FP} + \text{TN})] * 100\%$$

Thus, specificity, area under curve are useful as sensitivity to determine the accuracy of proposed algorithm in dengue fever prediction and techniques. The system produces average sensitivity 1%. Sensitivity refers to the test's ability to correctly detect ill patients who do have the condition. In the example of a medical test used to

identify a disease, the sensitivity of the test is the proportion of people who test positive for the disease among those who have the disease.

Fig 5.2.1 describes a Sensitivity (SE) Sensitivity (also called the true positive rate, or the recall in some fields)

measures the proportion of actual positives which are correctly identified.

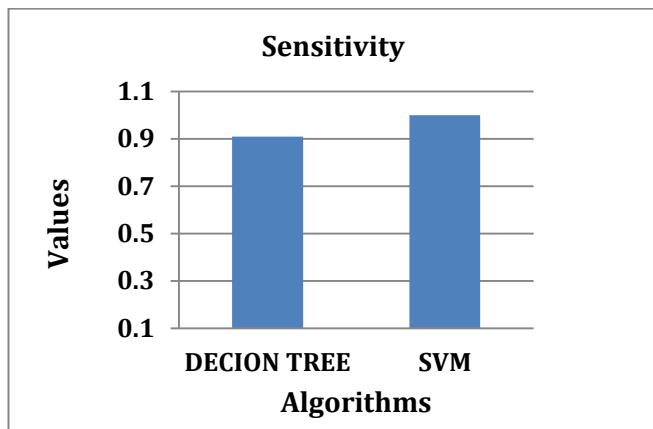


Fig 5.2.1 Sensitivity of the Proposed and Existing System

Fig 5.2.2 describes a Specificity (SP) (also called the true negative rate) measures the proportion of negatives which are correctly identified.

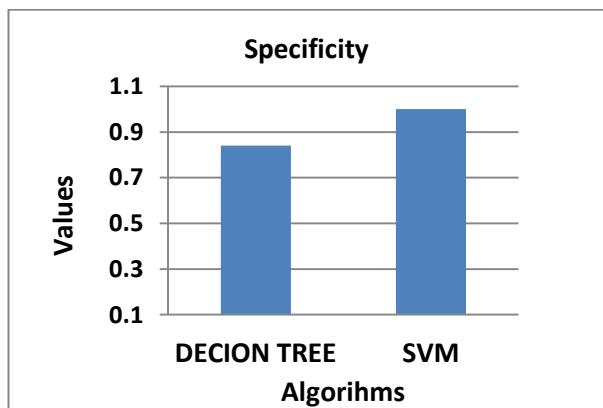


Fig 5.2.2 Specificity of the Proposed and Existing System

Fig 5.2.3 describes an Area under the curve (AUC). Area under ROC curve is used to measure of quality of the classification models.

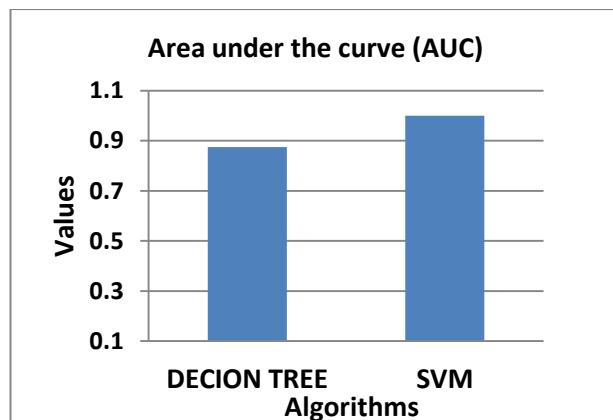


Fig 5.2.3 AUC of the Proposed and Existing System

The existing Decision Tree seems to be more efficient but this method produces a lot of false positives. This existing method increase of false positives, increase in processing time, etc. but they are easy to implement. So, the proposed method analyzed SVM classification algorithms overcome existing method. The proposed method is compared with existing method and the results shown the ability to predict dengue fever clearly than the existing method. Hence, from the results in all cases, it is shown that proposed method is found to be better than the existing method. The comparison is based on the accuracy of the results obtained with sensitivity, specificity and area under the curve. The accuracy for the proposed method is shown in the Fig 5.2.4.

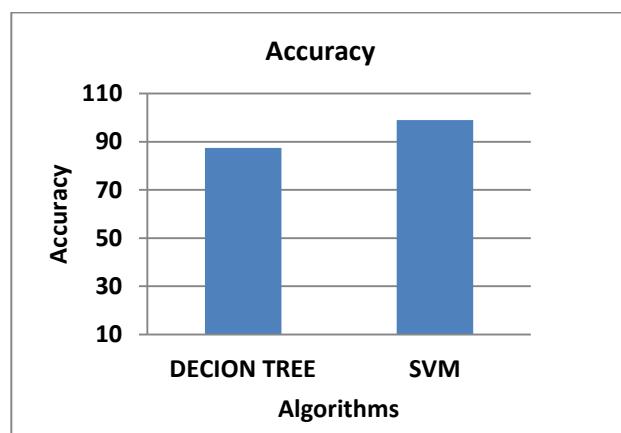


Fig 5.2.4 Comparison of Accuracy

6. CONCLUSION

This thesis has discussed the dengue disease prediction. The main objective is prediction of dengue fever using WEKA data mining tools. There are two classification techniques were used decision tree and SVM. These techniques are applied using WEKA data mining tool. To evaluate the accuracy achieved after analysis of these techniques, after testing these techniques the accuracy of proposed is compared to existing one. Forward and backward has been widely used for feature selection to improve dengue classification performance. Further, a lot of work is being done using forward and backward for feature selection to improve dengue classification performance and to reduce number of features selected well. Finally it has concluded that SVM is the top performance classifier technique by way that, it has achieved an accuracy of 99% takes fewer time to run and it has smallest error rate. Proposed improve the system the researchers recommended using more advanced tools in generating output wise dengue effect analysis the system. In future we move further to improve overall performance and computation with different set of attributes.

REFERENCE:

- 1) **P. Isakki @ Devi**, "dengue fever prediction using k-means Clustering Algorithm", 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing.
- 2) **DalsaniaPreet, Deshpande Swarada, Doddamani Poonam**, "Dengue Fever classification using SMO Optimization Algorithm", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 10 | Oct - 2017.
- 3) **S.vinodkumar**, "a data mining approach for precise Diagnosis of dengue fever", international journal of latest trends in engineering and technologyVol. (7) Issue (4), pp.352-359.
- 4) **Vanny Narita**, " Dengue Outbreak Prediction for GIS based Early Warning System", 2015 International Conference on Science in Information Technology (ICSITech).
- 5) **Nayyer Masood, SundasMehreen, UlyaAzmeen**, "Dengue Fever Prediction: A Data Mining Problem", Journal of Data Mining in Genomics & Proteomics, ISSN: 2153-0602 JDMGP, an open access journal, Volume 6 Issue 3 1000181.
- 6) **PrapatSuriyaphol, NuanwanSoonthornphisaj**, "Data Mining of Dengue Infection Using Decision Tree", Latest Advances in Information Science and Applications, ISBN: 978-1-61804-092-3.
- 7) **N.Subitha, A.Padmapriya**, "Diagnosis for Dengue Fever Using Spatial Data Mining", International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013, ISSN: 2231-2803.
- 8) **Gage M, Wattendorf, D** (2012), "Translational advances regarding hereditary Dengue Disease syndromes". IEEE International Journal of Computer Science and Engineering, vol. 90.
- 9) **D.Manimeglai**, "Predictive Analysis for the Arbovirus-Dengue using SVM Classification", International Journal of Engineering and Technology Volume 2 No. 3, March, 2012 ISSN: 2049.
- 10) **Ferro, Roberto** (2012), "Pesticides and Dengue Disease ", IEEE International Journal for Cancer Research and Treatment, vol.76.
- 11) **S.Gupta, Kumar, A. Sharma** (2011), "Data Mining Classification Techniques Applied for Dengue Disease Diagnosis and Prognosis". IEEE Journal of Computer Science and Engineering. Vol.23, pages 1191-1193.