International Journal of Engineering & Technology, 7 (1.3) (2018) 13-15

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Detection of dengue disease using artificial neural network based classification technique

K. Balasaravanan [1]*, M. Prakash [2]

[1] *Associate Professor/CSE, Karpagam College of Engineering, Coimbatore, Tamilnadu , India*
[2] *Professor/CSE, Karpagam College of Engineering, Coimbatore, Tamilnadu , Indi*
**Corresponding author E-mail: kbalasaravanancse@gmail.com*

Abstract

The information about the patients can be maintained with clinical documents. By keeping huge volume of clinical documents we can easily predict the occurrence of any disease in the patients. Dengue is considered to be one of the vital disease which are spreading in more than 110 countries. It is a vector borne disease caused by the mosquito's of female Aedes Albopictus and Aedes Aegypti which are well suited human environment. We have implemented a data mining technique called ANN which is a well-known technique for classification of data used here to classify diseases. We have analyzed the patients' dataset for the occurrence of dengue and experimented with Weka and Netbeans IDE and the result is proved to be more accurate than the CART algorithm.

Keywords: *ANN algorithm, Dengue, ARM, Classification*

1. Introduction

Medical data mining are used to discover the hidden patterns in the medical datasets. This patterns can be utilized for disease diagnosis. However the medical data are highly voluminous, distributed and heterogeneous in nature. So these data should be arranged to make it more usable for hospital information system. Dengue fever is a contiguous and deadly disease occurs every year during monsoon days in tropical and sub-tropical areas around the world. At present there is no medicine for dengue and the patient should be given full rest and should take plenty of fluids to get rid of this toxins. Time is an essence in dengue medication. Once the patient is detected to be suffered from dengue he have to be undergone two stages of tests, the preliminary CBC and the confirmatory serology tests. The problem in this test is that the CBC test for the dengue is similar to the preliminary test of other diseases too. Dengue serology test may take around 10 days to diagnose the infection in patient's blood, and the health can be progressed within 10 days of time.

The objective of this work is to find the possibility of detecting the occurrence of dengue to the patient by implementing the Artificial Neural Networks (ANN's) technique. In uses some real parameters like mean temperature, mean relative humidity, total rainfall and the reported dengue cases with respect to the outputs of the above three parameters. In this work we used ANN because it is the perfect tool to learn the problem from given examples and it don't demand any mathematical modelling of any cases. This work utilize recognition of dengue confirmed cases.

The real data received from a medical agency has been utilized here to model the dengue behaviour confirmed cases based on the statistics of the above three parameters.

This work also use a predictive or descriptive data mining technique for the disease prediction. Tanagra is a tool used here to classify the data and the data is assessed with many cross validations to test the output.

The objective of Tanagra is to improve the usability of the data mining software and to let them to analyse the data either really of synthetically. Tanagra supports all data mining feature implementations like clustering, classification, feature selection, supervised learning, Meta supervised learning, data visualization and feature construction algorithms.

The ANN algorithm is very easy and simple to implement and needs some knowledge or parameter settings to handle high dimensional data. The results received from this algorithm can be easily interpreted. The result can be received using decision tree output. It will try to improve the posterior probability to determine the result. An another advantage of using this technique is that it generates probability based output so that when greater volume of data is given as input the accuracy will be high.

It uses machine learning algorithms to engage the supervised learning technique. It degraded with the presence of noise features which should be removed. We performed our experiment with 4500 instances of 10 difference attributes of dengue data and determined the disease occurrence.

2. Literature review

N. Aditya Sundar et al [5] proposed a system which contributes the detection of dengue disease using blood pressure, viral infection, sex and age factors. It used Naïve Bayesian classification and WAC 55 to train the model on existing data. Even patient and nurse can use this model to supply features and get the prediction on disease occurrence.

Oona Frunza et al [6] had defined a machine learning technique to identify the semantic relationship of treatment and the diseases and it mainly focussed on three semantic relationships i.e.,

prevent, cure and side effects. Then certain features were extracted from unstructured medical data and were utilized to determine the relationship between the disease and the treatments.

Jyoti Soni et al have introduced an algorithm which predicts the dengue using 15attributes. It creates a decision tree to produce the result. Classification based on clustering has been used in this model. It proposed a medical diagnosis system for predicting the risk of fever. It used Genetic Algorithm to determine the weightage for neural network. This neural network can be used for prediction and classification of dengue.

Devendra at al proposed a prediction system using Naïve Bayes approach. They developed a web interface to aid healthcare practitioners to assess the probability of occurrence of dengue. A similar work has been done combining both decision tree and Naïve Bayes to predict the disease.

Some techniques in data mining that can be used to model and predict the dengue include SVM, decision tree and neural network. Detection of dengue in the early stage can be more helpful in preventing the dengue outbreaks. Making a more accurate prediction on epidemic seasons can provide sufficient time to take measures to protect patient from more serious results.

In a study, they used a technique called Vector Error Correction Model (VECM) to define the dengue patients by noting the climatic factors. It used vector correction method to predict the disease outbreak. It considered only the humidity and temperature but not the rainfall into account. They also proved that the above mentioned scheme is more accurate in prediction with the above stated factors.

The prediction served by the above model can be used by the local authorities to make preventive measures during epidemic seasons so that the disease can be controlled at the earliest. The study in [12] have demonstrated that the weather variables play a vital role in dengue occurrence. It can be used to develop a simple, precise and low cost measure for dengue early warning. They created a weather based forecasting model with evidence for scientific data which shows that temperature and rainfall has great impact in vectors and dengue viruses. The factors like ecology, environment, vectors, virus factors and human are the most important influencers of dengue detection. 16 weeks forecasting shown by them provides an ample time for the local authorities to make an awareness for dengue.

The work in [12] used ARM technique to make extraction of relationship between clinical, climatic, meteorological, and sociopolitical data. These relationships are used as rules and the best rules are chosen automatically and the classifier is formed. This classifiers can be used to analyse the dengue for its severity as either HIGH or LOW in which those values can be used as the above and below the means of the previous dengue incidence plus two standard deviations respectively. To perform the spatiotemporal predictions, the entire variables should be fit to the same spatiotemporal scale. This work have chosen the spatiotemporal scale using the distribution of dengue data. The temporal scale taken was 1 week and the spatial distribution was considered for one district.

That classifier is then used to predict future dengue incidence as either HIGH (outbreak) or LOW (no outbreak), where these values are defined as being above and below the mean previous dengue incidence plus two standard deviations, respectively. In order to perform spatiotemporal predictions, all the variables need to fit the same spatiotemporal scale. The spatiotemporal scale used in this work was selected based on the distribution of the dengue data: the chosen temporal scale was one week and the chosen spatial distribution was one district.

In [8], they examined the meteorological factors on dengue occurrence over three geographical areas of Sri Lanka using a weekly data. A week's average maximum temperature, humidity and rainfall has been noted for time series data in the study. Also, this time based analyses are done with least squares regression analysis. Then vector based autoregressive model are used. They successfully forecasted the dengue persistence in Singapore to make a warning of dengue outbreak there.

In [12], they created a model with absolute shrinkage and selection operator methods to forecast weekly dengue notification for a three month time span. Machine learning methods like LASSO has been utilized to forecast the outbreak of dengue at the earliest.

3. Methodology

Generally ANN algorithm is a three layered feed forward technique. As it uses one step ahead forecasting, one output will be utilized. Many engineering problems has been resolved using ANN model from business forecasting to computer vision using available data. Neural networks has huge volume of interconnected units with input, output and hidden nodes.

The processing unit of the ANN combines the weighted activation on the inputs, transform the sum to an activation function and send the result to the output. So the information processing in ANN can be done by transforming the input into some output and then it is modulated by the weight of connections as input to all other units.

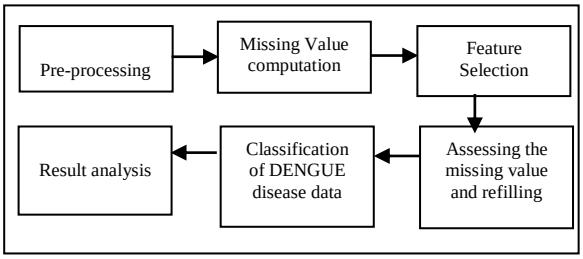

Fig. 1: Architecture of Dengue Detection using ANN Algorithm

After pre-processing which learn the set of inputs being given, missing value computation will be performed. In this steps if any data is missing in the given dataset, it will be refilled here. It is very common for the dataset to miss any of the data while consider for analysis. Such kind of data will be either eliminated or refilled here. After this steps follows the feature selection. The feature selection will be made to consider very few of the data that are need for data prediction. The sensitivity of dataset will be assessed here and only certain attributes will be considered here. After the feature selection is made classification using ANN is deployed. The main benefit of using ANN is their behaviour in which in don't need any knowledge and rules for learning and implementing the technique. Instead the ANN itself generate its own rules to by learning form the data sets. This is used in our dataset to make pattern recognition and classification of dengue.

It also possess closer to human perception than any other existing techniques. Even when the dataset have noisy data this ANN can easily produce the results. So this possibilities have the dengue assessment more easy and accurate. In this study, the ANN model gets 5 inputs x(n) from rainfall, humidity rate, temperature rate, number of dengue cases from the previous month dataset. The neuron has two units. First one combines the products of weighted coefficient with the input signals. The next units generates a nonlinear neuron activation function. After computation, the output layer will generate the desired response vector. The output will have the trends and prediction of the disease diagnosis of dengue in it. The result and the comparison of the predicted data is displayed over the graphical representation.

As we have mentioned earlier as it is based on the probability of occurrence, more number of dataset can produce more accurate result. Also, the correlation between the actual output and the predicted output is calculated after convergence and the result is cross verified. By analysing, it is assessed that the proposed work is more accurate that the normal prediction technique which is averagely 30% more accurate. The attributes used in the dataset are myalgia, flu, fatigue, id, fever, bleeding, other symptoms and results. There are four types of precision employed in it. TN =case

was negative and predicted negative, TP =case was positive and predicted positive. FN =case was positive but predicted negative, FP = case was negative but predicted positive.

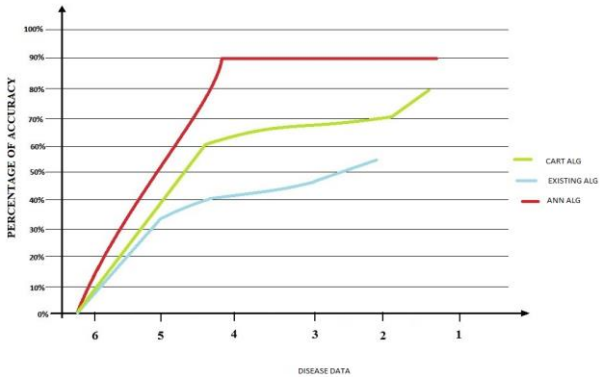

Fig. 2: Dieses vs Accuracy

Table 1: Different Techniques with accuracy

Technique	TP Rate	ROC Rate	Error Rate	Accuracy
Bayesian	0.89	0.756	0.09	0.82
REP Tree	0.86	0.979	0.34	0.72
Random Tree	0.86	0.989	0.34	0.71
J48	0.89	0.58	0.29	0.85
SMO	0.73	0.43	0.29	0.72

Based on the above results, the ROC of Naive Bayes found to be 0.874 and has the smallest error rate. And it is revealed that ANN and CART had accuracy 92% and 88% respectively.

4. Experiment and result

The feature vector is given to various supervised learning algorithms and classifiers are created from it. We implemented this in Weka tool which is the software supported for support vector classification, regression and distribution estimation. It supports multiple class classification. The ANN algorithm also uses multilayer perceptron for its classification. All the layers will be interconnected and the Naïve Bayes method is implemented by providing the supervised learning process using Bayes Theorem with the naïve assumption of independence between feature pairs. Our technique used minimal optimization technique for training the clinical data. It implements normalization as well. It can be used to build linear logistic regression models. These types of classifiers are subjected to two different types of classifications. They are percentage split and 10-fold cross validation.

Five performance metrics like mean absolute error, RMSE, Kappa statistics, Accuracy and Relative absolute error) are used to analyse our classifiers and the most effective model is selected. The result is visualized here and the count of occurrence of the disease are assess monthly using charts. These results are compared with the original training dataset and the efficiency of our algorithm is determined.

It is analysed for the month of September and its predicted that this month suffers with the maximum cases of dengue. The occurrence of carious symptoms over the months are depicted using bar charts. Further analysis are made for the complete year and its observed that August, September and October are the months which are more vulnerable to dengue compared with the original training dataset.

5. Conclusion

This work is performed to identify the disorders occurred in clinical text and it's correlated with the time frame. The summary of the dataset are tagged, the feature extraction, classification algorithms are utilized to probe the disease. A feature vector is generated using the dataset and classification technique and SMO is used to produce an effective result. This model can be used to generate aids to predict the disease. We have analysed the result using Bar graph and the training samples are used to test the accuracy of our result and it's proved to be 95% more accurate. It is planned to implement this work with nonparametric iterative imputation method in future. This method can be implemented for dataset with discrete attributes and continuous data

References

[1] Ahamad I.A. and P.B. Cerrito, "Nonparametric Estimation of Joint Discrete-Continuous Probability Densities with Applications,"J. Statistical Planning and Inference, vol. 41, pp. 349-364, 1994.
[2] Allison P., Missing Data. Sage Publication, Inc., 2001.
[3] Batista G. and M. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," Applied Artificial Intelligence, vol. 17, pp. 519-533, 2003.
[4] Brown M.L., "Data Mining and the Impact of Missing Data," Industrial Management and Data Systems, vol. 103, no. 8, pp. 611-621, 2003.
[5] Cristian Molinaro, Maria Vanina Martinez, John Grant, and V.S.Subrahmanian, "Customized Policies for Handling Partial Information in Relational Databases", Knowledge and Data Engineering, 2012.
[6] Cios K. and L. Kurgan, "Knowledge Discovery in Advanced Information Systems," Trends in Data Mining and Knowledge Discovery, N. Pal, L. Jain, and N. Teoderesku, eds., Springer, 2002.
[7] Han J. and M. Kamber, Data Mining Concepts and Techniques, second ed. Morgan Kaufmann Publishers, 2006.
[8] Lakshminarayan et al.K., "Imputation of Missing Data in Industrial Databases," Applied Intelligence, vol. 11, pp. 259-275, 1999.
[9] Little R. and D. Rubin, Statistical Analysis with Missing Data, second ed. John Wiley and Sons, 2002.
[10] Xiang Lian and Lei Chen," Efficient Similarity Search Over Future Stream Time Series," Knowledge and Data Engineering,2008.
[11] Husin, N.A. 2008. "Back propagation neural network and nonlinear regression models for dengue outbreak," M. Sc. thesis, Universiti Teknologi Malaysia, Johor, Malaysia, Nov. 2008.
[12] Poovaneswari, S. 1993. "Dengue situation in Malaysia," Malaysia J Pathol, vol. 15(1), pp. 3–7, June. 1993.
[13] Gubler, J. 1997. "Epidemic Dengue/Dengue Haemorrhagic Fever: A Global Public Health Problem in the 21st Century," Dengue Bulletin, vol. 21, pp. 1–120.
[14] Er, A.C., Rosli, M.H., Asmahani A., Mohamad Naim M.R., Harsuzilawati M. 2010. "Spatial mapping of Dengue incidence: A case study in Hulu Langat district, Selangor, Malaysia," International Journal of Human and Social Sciene, vol. 15(1), pp. 410–414.
[15] Muto R. 1998. Summary of dengue situation in WHO Wester Pacific Region. Dengue Bulletin 22. [Cited 08 Feb 2011]. Available : http://www.searo.who.int/EN/Section10/ Section332/Section520.htm.
[16] Teng and Singh. 2001. "Epidemiology and new initiatives in the prevention and control of dengue in Malaysia," Dengue Bulletin, vol. 25, pp. 7–14.
[17] Qin,B., Xia, Y., Prabhakar, S. and Tu,Y. 2009. "A rule-based classification algorithm for uncertain data," in IEEE International Conference on Data Engineering. ICDE'09, pp. 1633-1640.
[18] Briem, G.J., Benediktsson, J.A. & Sveinsson, J.R. 2002. "Multiple classifiers applied to multisource remote sensing data," IDBIS, vol .40, pp. 2291-2299.
[19] Cufoglu, A., Lohi, M. & Madani, K. 2008. "Classification accuracy performance of Naïve Bayesian (NB), Bayesian Networks (BN), Lazy Learning of Bayesian Rules(LBR) and Instance-Based Learner (IB1) – comparative study," in Conf. ICCES'08, p. 210-215.
[20] Berzal, F., Cubero, J.C., Sánchez, D. & Serrano,J.M. 2002. "ART: A Hybrid Classification Model," IEEE Transactions on Geoscience and Remote Sensing, pp. 1-17.
[21] Aslandogan, Y.A. & Mahajani, G.A. 2004. "Evidence combination in medical data mining," in Proc. ITCC'04.