Stratification of Dengue Fever using SMO and NSGA-II Optimization Algorithms

Suresh Limkar, Preet Dalsania,Swarada Deshpande,Aishwarya Dixit,Poonam Doddamani
Department of Computer Engineering, AISSMS IOIT, Pune-01

Abstract. In today's world, millions of cases of dengue are reported ever year. The number of cases has increased, afflicting many individuals. For prediction of dengue clinical methodology comprises of antigens and anti-bodies tests. 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. The dataset [GDS5093] being referred in this proposed model are of acute dengue 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 Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) and to increase the optimality of the model, we have also used Probabilistic Neural Network (PNN). PNN uses feedforward technique for classification.

Keywords: SMO, NSGA-II, PNN, DF, DHF.

1 Introduction

Dengue is a mosquito borne viral disease which is mainly transmitted by the species of female mosquitoes named "Aedes aegypti". The "Aedes aegypti" mosquito lives in urban habitats and breeds on a large scale. Symptoms of dengue include high fever, pain behind the eyes, muscle and joint pain, severe headache.[1]

The primary task is to find out whether the person is suffering from dengue or is he a healthy person. After this, the more challenging part is to find whether he is infected from Dengue fever (DF) or dengue Hemorrhagic fever (DHF)[4].There is a need for research in this field.

Our proposed model gives an architecture of stratification of dengue disease using the PNN model along with optimization algorithms. A probabilistic neural network is widely used in classification and pattern recognition problems. This type of neural network has four layers in it the input layer, the hidden layer, pattern layer, output layer. Dataset of acute dengue patients [GDS5093][3]. These layers compute the result and compare it with the training set and produce the desired output. Probabilistic neural network sometimes converge at the local optima, hence to avoid this we will apply two algorithms that are NSGA-II which is the advancement over Genetic Algorithm [2] and SMO. These algorithms will help us to converge at the global optima rather than the local optima. These algorithms have their independent features.

The current work proposes NSGA-II and SMO trained Neural Network using greedy feedforward selection algorithm for selection of the specific prominent gene [4]. Only the useful data must be extracted from the dataset as there are 54715 genes for 56 homo-sapiens subjects [4]. We will be using PNN as we are mainly considering the probability of achieving an optimal accuracy [5] as compared with the previous PSO trained NN which also uses the greedy forward selection algorithm for gene selection. The NN-PSO model gave the accuracy of 90.91% [4].Our proposed model is to find whether we achieve a better accuracy than NN-PSO

Rest of the paper is organized as section II discusses literature survey, section III discusses the proposed system, section IV discusses about the conclusion followed by references.

2 Literature Survey

P. Manivannan et al [6] provided K-Mediod Clustering Algorithm for prediction of dengue fever. Research work done on predicting the people who are affected by dengue depending upon categorization of age using the K-medoid clustering algorithm. The K-mediod clustering algorithm was applied on the dengue dataset. The result obtained by using K-mediod clustering algorithm has increased the efficiency of output. This is the most effective technique to predict the dengue patients with serotypes. It lacks in scalability of large datasets. It also has high time and space complexity.

Sankhadeep Chatterjee et al [4] has provided a PSO approach for classification of dengue virus into DF (Dengue fever) and DHF (Dengue Hemorrhagic Fever). The greedy forward methodology is used to select the significant genes through the pre-processing stage from the available blood samples. Further PSO trained Neural Network (PSO-NN) was implemented for detecting and classifying the fever into DF and DHF. The accuracy obtained by this method is 90.91%. Different nature based algorithms like NSGA-II and SMO can be used to increase the accuracy of the same.

Tarmizi et al [7] has referred the weather of Thailand, Indonesia and Malaysia where the climate is more humid , and hence water borne diseases like dengue is more prone in that area. The study proposes different machine learning techniques like Data mining (DM), Artificial neural network (ANN). It also includes rough set theory (RS). The classifications algorithms which are used to predict dengue disease. The data set referred is of public hospital at Selangor state. 10 cross fold validation and Percentage split are the two tests used with the simulator WEKA tool. The accuracy obtained with 10 cross fold validation was 99.5% with DT, 99.8% with ANN, 100% accuracy with RS. Using percentage split 99.2% of accuracy was achieved with DT and ANN, whereas 99.72% of accuracy was obtained using RS.

Fathima et al [8] The work proposed is prediction of Arbovirus- Dengue disease. The data mining algorithm that is used by them are support vector machine (SVM). The reference being used in the implementation was taken from the surveys of hospitals and laboratories located in Chennai, India. The data set they referred had 29 attributes and 5000 samples. T R project version tool was used for examination and achieved the accuracy of 90.42%.The only disadvantage we can talk about is accuracy achieved by using SVM and hence the accuracy achieved by rough set theory which was 100% is referred.

Ibrahim et al [9] The model was suggested which used Artificial neural network with multi-layer feedforward neural network. It is used for forecasting the defervescence fever in patients of dengue disease. The data is gathered from 252 hospitalized patients, in which 4 patients are having Dengue Fever and 252 patients had Dengue hemorrhagic fever. MATLAB neural network tool box is used and achieved the accuracy of 90%. Accuracy can be increased using different methodologies.

The earlier methods used were K-medoid clustering, support vector machine and PSO algorithm for classification of dengue fever. A better method for classification of dengue fever has been deployed considering the literature survey of the above papers.

3 Proposed System Flow

3.1 Algorithms Used

Spider Monkey Optimization (SMO)

SMO is an algorithm which falls under swarm intelligence. It is based on the foraging behavior of spider monkeys. The spider monkeys' behavior is categorized as fission-fusion as their foraging behavior involves a unit-group of individuals dividing themselves into subgroups (fission) in order to look for food. Later in the day when they sit to eat, these members come together with the other subgroups (fusion).[10]

In order to reduce the competition while finding food, the swarm divides itself into smaller subgroups.

The algorithm of spider monkey basically has 7 following steps

1. Initialization of population

The initial population is a uniformly distributed population which consists of N number of spider monkeys. Each spider monkey SM_x (x = 1, 2, …, N) is a D-dimensional vector. D represents the no. Of variables in the problem. The potential solution to the problem is represented by each spider monkey SM.

Each SM_x is initialized as follows:-

$$SM_{xy} = SM_{miny} + R(0,1) \times (SM_{maxy} - SM_{miny})$$

2.Local Leader Phase (LLP)

In this phase, each spider monkey updates its position using experience of local leader. The fitness value of the new position is calculated. The SM updates its position

only if the newly calculated fitness value is higher than the fitness value of the old position. The equation for the position update for xth SM is

$$SMnew_{xy} = SM_{xy} + R(0,1) \times (LL_{ky} - SM_{xy}) + R(-1,1) \times (SM_{ry} - SM_{xy})$$

where SM_{xy} is the yth dimension of the kth local group leader position. SM_{ry} is the yth dimension of the rth SM, it is chosen randomly within the kth group and $r \neq x$, R(0,1) is any random number between 0 and 1.

3. Global Leader Phase (GLP)

Next comes the Global Leader Phase. In this phase, all the positions of the spider monkeys are updated using experience of the global leader. The position update equation is as follows

$$SMnew_{xy} = SM_{xy} + R(0,1) \times (GL_y - SM_{xy}) + R(-1,1) \times (SM_{ry} - SM_{xy})$$

where Gly is the yth dimension of the global leader position and y ϵ to {1, 2, …, D} where y is the index that is randomly chosen.

Probabilities P_i are calculated using fitness of position of spider monkeys (SM_x) and the positions are updated based on the P_i.

The probability P_i can be calculated using

$$P_x = 0.9 \times \frac{fitness}{max_{fitness}} + 0.1$$

here, fitness means the fitness of the ith spider monkey and max_fitness means the maximum fitness of the subgroup. Further, the fitness value of the new position is generated and the new position is adopted only if this value is greater than the fitness value of the old position.

4. Global Leader Learning Phase (GLL)

This phase updates the position of the global leader with the spider monkey having the best fitness in the entire population, by making use of greedy selection. If the Global leader's position is not updated then GlobalLimitCount, a predetermined counter, is incremented by 1.

5.Local Leader Learning Phase (LLL)

Similar to the previous phase, this updates the position of the local leader with the spider monkey having the best fitness in that group. If the local leader's position is not updated then LocalLimitCount, a predetermined counter, is incremented by 1.

6. Local Leader Decision Phase (LLD)

If any of the local leaders' position is not updated to a predetermined value LocalLeaderLimit, then positions of all members of that group are updated through following equation

$$SMnew_{xy} = SM_{xy} + R(0,1) \times (GL_y - SM_{xy}) + R(0,1) \times (SM_{xy} - LL_{ky})$$

7.Global Leader Decision Phase (GLD)

If the position of global leader is not updated up to a predetermined value GlobalLeaderLimit then the population is divided into smaller subgroups. Whenever new subgroups are formed, the new local leader in the newly formed subgroup has to be elected, and that is initiated by the LLL process. If such a situation occurs where number of groups formed is maximum, but yet the global leader's position is not updated, then all the small subgroups are combined to form one single group, thus following the fission-fusion behavior of spider monkeys.

Non-Dominated Sorting Genetic Algorithm (NSGA-II)

From the past few years, many multi-objective evolutionary algorithms (MOEAs) are used because of their ability to produce the optimal results in a single run. One other kind of MOEA is Non-Dominated Sorting Genetic Algorithm (NSGA-II). NSGA the predecessor of NSGA-II has been criticized a lot due to its computational complexity, lack of elitism and need for specifying the sharing parameter. To alleviate these drawbacks NSGA-II was proposed by Prof. Srinivas. The main features of NSGA-II are low computational complexity, less diversity prevention, elitism and real valued representation.[11]

NSGA-II implements elitism using elitism-preserving approach. Beginning from the initial population all non-dominated solutions are sorted. NSGA-II is implemented using these following steps:-

1. Random parent population P is created, of size Z
2. Non-domination sort is applied on this random population
3. For each solution a fitness value is assigned according to its non-domination level.
4. Offspring population Q is created using binary tournament selection, recombination and mutation operators.
5. In this step offspring and the parent population are mated. A mating pool R is created, and then again non-domination sorting is applied to this new population, which is of size 2Z. Fitness value is assigned and some of the lower value solutions are rejected. Again perform the crossover, perform selection and mutation operations.
6. Step 5 is until maximum number of iterations are reached. [11] **Probabilistic Neural Network (PNN)**

Probabilistic neural network (PNN) is a feed forward neural network. Basically used for classification problems. PNN offers a great speed compared to back propagation. And according to studies it has showed that for a PNN paradigm was 20,000 times faster than back propagation. It also provides high accuracy. It is derived from Bayesian network. PNN is closely related to Parzen window function or Gaussian functions.[12][5]

It usually has 4 layers, more layers can be added when required.

When input is available, the first layer computes the distance from the input value to the trained value. This produces a value where it shows how close the input value is to the trained value.

The second layer consists of Gaussian functions or parzen window estimator which is formed using the given set of data points as centers. The Gaussian function for PNN for N classes is defined as follows:-

$$y_j(X) = \frac{1}{n_j} \sum_{i=1}^{n_j} \exp\left(\frac{-(X_{y,i} - X^2}{2\sigma^2}\right)$$

Where, j = 1, 2,…,N

$n_j$denotes the number of data points in class j

n_j denotes the number of data points in class j

$(X_{y,i} - X^2)$ is calculated as the sum of squares

σ = Gaussian window function

The third layer performs summation of the outputs from the second layer for each class

The fourth layer selects the largest value from the summed value. And hence decide a label for the value..[12][5]

Flow of the proposed model is seen in the figure 1, which depicts the complete flow of the inputs and outputs. Initially in the model the data is pre-processed, where selection of prominent genes is computed using the greedy forward selection algorithm. Random weights are been assigned to the PNN, and if the output doesn't converges to the global optima SMO and NSGA-II optimization algorithms are used to optimize the output. And this loop is carried until the output converges to the global optima.

Pseudo code

```
begin
    initialize data
    assign random values to PNN
    if(convergence==local optima)
        use SMO
        use NSGA-II
    else
        print output
    end
end
```

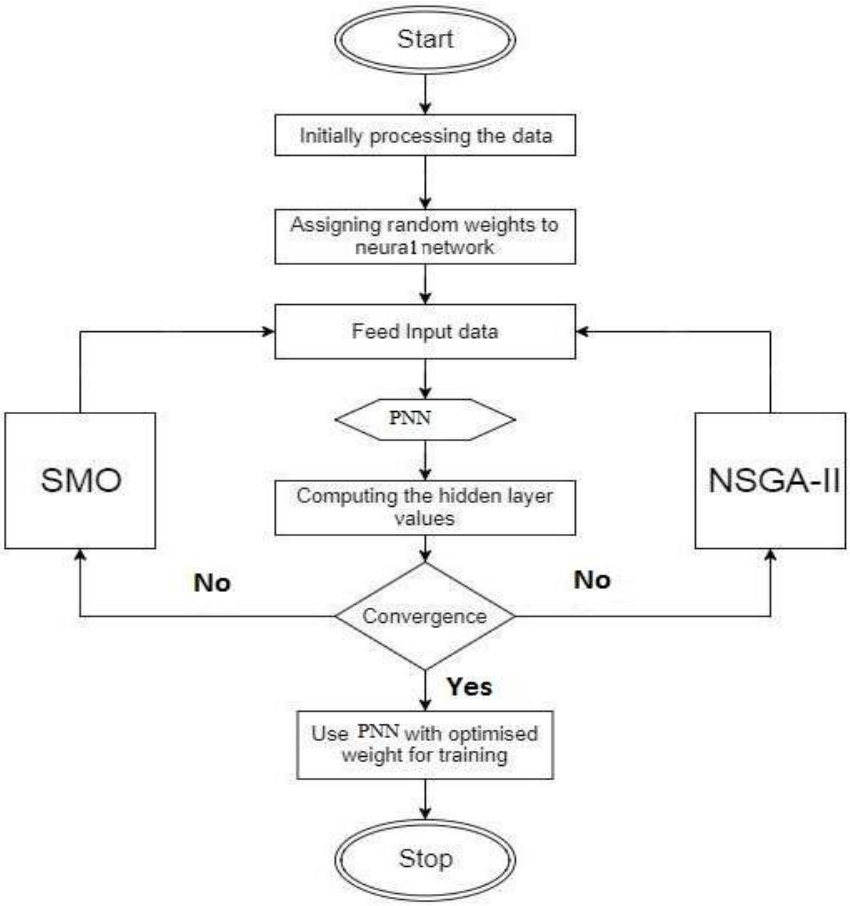

Fig. 1.Process flowchart

4 Conclusions

The paper gives a comparative study of PSO with the new model proposed which includes NSGA-II and SMO algorithm. According to previous studies, SMO algorithm is used as it is been proved better than the PSO algorithm. NSGA-II is used because of its genetic behavior. Hence the proposed model may increase the optimality of classification of dengue fever because of the algorithm.

5 References

[1] http://www.who.int/mediacentre/factsheets/fs117/en/

[2] Suresh Limkar et al "Genetic Algorithm:Paradigm Shift over a Traditional Approach of TimeTable Scheduling" ,Proceedings of the 3rd International Conferences on Frontiers of Intelligent Computing:Theory and Applications (FICTA)2014,vol 32,pp772-78

[3] https://www.ncbi.nlm.nih.gov/geoprofiles?term=GDS5093[ACCN

[4] Sankhadeep Chatterjee, Sirshendu Hore, Nilanjan Dey, Sayan Chakraborty, Amira S.Ashour, "Dengue Fever Classification using Gene Expression Data:A PSO based Artificial Neural Network Approach", Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications : FICTA 2016, Volume 2, Springer Singapore, 2017

[5] https://en.wikipedia.org/wiki/Probabilistic_neural_network

[6] P.Manivannan, Dr. P. Isakki, "Dengue Fever Prediction using K-Medoid Clustering Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Special Issue 1, March 2017

[7] Tarmizi, N.D.A., Jamaluddin, F., Abu Bakar, A., Othman, Z.A., Zainudin, S. and Hamdan, A.R. (2013) Malaysia "Dengue Outbreak Detection Using Data Mining Models", Journal of Next Generation Information Technology (JNIT), 4, 96-107.

[8] Fathima, A.S. and Manimeglai, D. (2012) "Predictive Analysis for the Arbovirus-Dengue using SVM Classification", International Journal of Engineering and Technology, 2, 521-527.

[9] Ibrahim, F., Taib, M.N., Abas, W.A.B.W., Guan, C.C. and Sulaiman, S. (2005) "A Novel Dengue Fever (DF) and Dengue Haemorrhagic Fever (DHF) Analysis Using Artificial Neural Network (ANN)", Computer Methods and Programs in Biomedicine, 79, 273-281

[10] Jagdish Chand Bansal, Harish Sharma, Shimpi Singh Jadon and Maurice Clerc, "Spider Monkey Optimization algorithm for numerical optimization," Springer, Memetic Computing, pp. 31-47, 2014

[11] Kalyanmoy Deb, Associate Member, IEEE, Amrit Pratap, Sameer Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 6, NO. 2, APRIL 2002

[12] Donald F. Specht, "Probabilistic Neural Networks", 1990 Pergamon Press, Neural Networks, Vol. 3. pp. 109 118, 1990