OPPOSITION-BASED FIREFLY ALGORITHM OPTIMIZED FEATURE SUBSET SELECTION APPROACH FOR FETAL RISK ANTICIPATION

V.Subha[1] and D.Murugan[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India.

[2]Associate Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India.

ABSTRACT

Recently huge amount of data is available in the field of medicine that helps the doctors in diagnosing diseases when analysed. Data mining techniques can be applied to these medical data to extract knowledge so that disease prediction becomes accurate and easier. In this work, cardiotocogram (CTG) data is analysed using Support Vector Machine (SVM) for predicting fetal risk. Opposition based firefly algorithm (OBFA) is proposed to extract the relevant features that maximise the classification performance of SVM. The obtained results show that opposition based firefly algorithm outperforms the standard firefly algorithm (FA).

KEYWORDS

Cardiotocography, SVM Classifier, Feature Selection, Opposition-based firefly algorithm

1. INTRODUCTION

Cardiotocography (CTG) is a commonly used technique to monitor and assess fetal state during pregnancy and delivery. It is a combination of two signals: fetal heart rate (FHR) and uterine contractions (UC). This CTG signal is used by obstetricians to monitor babies having either acute or chronic hypoxia. Visual analysis of CTG often leads to incorrect interpretations and hence computer aided systems are needed for classifying CTG which helps the obstetricians to decide if the baby can be given a natural birth or caesarean section.

Numerous methods have been reported in literature for analyzing the CTG data. A SVM classifier to classify the fetal state in to two classes [1]. Additionally, Genetic Algorithm has been used for selecting the most relevant features and thereby the performance of the classifier has been improved. Least squares-SVM, Particle Swarm Optimization and binary decision tree have been used to classify the CTG data [2]. An adaptive neuro fuzzy inference system has been presented to classify the CTG data of fetal state into two classes [3]. Classification of CTG data using Random forest classifier combined with feature reduction technique has been presented in [4]. Improved accuracy has been achieved using discriminant analysis, decision tree and artificial neural network for fetal distress prediction in [5]. A classifier for CTG data which uses neural network and simple logistics is proposed in [6]. Naïve Bayes Classifier has been used for classification of CTG data along with feature selection approaches in [7]. A classifier which classifies the data into three classes by applying modular neural network is proposed in [8]. A neural network based classifier is proposed in [9] to improve the performance of other clustering algorithms in CTG classification. Naïve Bayes Classifier has been used in [10] to classify the CTG data in to three classes. In [11], a feature selection method based on Artificial Bee Colony algorithm is reported. Further, in [12] feature selection method based on Artificial Bee Colony algorithm and Support Vector Machines for medical datasets classification is proposed. A fetal state classifier using SVM and Firefly algorithm has been proposed in [13] to improve the classification accuracy of CTG. In [14], a genetic algorithm based feature subset selection is proposed to find the relevant features for CTG classification.

In this paper, opposition based firefly algorithm (OBFA) together with Support Vector Machine (SVM) classifier has been proposed for classification of CTG data. OBFA has been used to produce optimal and reduced feature set which results in improvement of performance of the SVM classifier. Initially, CTG data are classified using the full feature set. Then, optimal feature set has been produced using FA and OBFA along with SVM classification. The experimental results reveal that the use of optimal feature set generated with OBFA improves the accuracy of classification.

This paper has been organized as follows; Section 2 describes the CTG data set. The Support Vector Machine classifier has been explained in section 3. Section 4 describes firefly algorithm and opposition based firefly algorithm (OBFA) is described in section 5. The proposed method of finding the optimal and reduced data set has been explained in the section 6 followed by the results and discussion in the section 7. Finally, the section 8 concludes the proposed method.

2. CTG DATA SET

The CTG dataset of UCI Machine Learning Repository [15] has been used for experiment. In this dataset, there are totally 2126 fetal cardiotocograms belonging to three different classes with 21 attributes and 1 class attribute. Three expert obstetricians have classified this data set consisting measurements of fetal heart rate and uterine contractions and assigned classification labels to them based on the fetal heart rate class codes (N-Normal, S-Suspect and P-Pathologic).

2.1. Attribute Information

LB - FHR baseline (beats per minute)
AC – No. of accelerations/second
FM – No. of fetal movements/second
UC – No. of uterine contractions/second
DL – No. of light decelerations/second
DS – No. of severe decelerations/second
DP – No. of prolonged decelerations/second
Width - Width of FHR histogram
Min - Minimum of FHR histogram
ASTV - Percentage of time with abnormal short term variability

Max - Maximum of FHR histogram
Nmax - Number of histogram peaks
Nzeros - Number of histogram zeros
Mode - Histogram mode
Mean - Histogram mean
Median - Histogram median
Variance - Histogram variance
Tendency - Histogram tendency
MSTV - Mean value of short term variability
MLTV - Mean value of long term variability

ALTV - Percentage of time with abnormal long term variability

CLASS- Fetal state class code (Normal=1; Suspect=2; Pathologic=3)

3. SUPPORT VECTOR MACHINE

Support vector machine (SVM) has been widely used for solving classification problems [16]. SVM separates the classes with an optimal hyperplane that increases the margin between the classes. The data points closest to this hyperplane are called support vectors. The nonlinear data used in this work is subjected to nonlinear kernel functions to transform the data into a new feature space where a hyperplane separates the data. Radial Basis Function (RBF) has been used widely by researchers for its better generalisation capability and hence RBF kernel has been adopted in this work. A One-Against-All SVM classifier is used here to classify the data into three classes.

4. FIREFLY ALGORITHM

Firefly algorithm is one of the efficient optimization algorithms [17]. Fireflies are insects producing a flashing light. Firefly algorithm makes use of three idealised rules. First, all fireflies are considered unisex which means that one firefly will be attracted to other fireflies regardless of their sex. Secondly, the degree of the attractiveness of a firefly is proportion to its light intensity, thus for any two flashing fireflies, the less brighter one will move towards the more brighter one. Finally, the light intensity of a firefly is somehow related with the analytical form of the fitness function. The basic steps of the FA are summarized as the pseudo code shown in figure 1.

```
begin
   Fitness function f(M), M=(m1, ...,md )^c
   Generate initial population of fireflies Mi(i=1, ...,n)
   Brightness Li at Mi is determined by f(Mi)
   Define light absorption coefficient α
   s=1
   while (s< smax)
   for i=1:n all n fireflies
     for j=1:n all n fireflies
        if (Lj>Li)
              Move firefly i towards j in d-dimension
        end if
        Attractiveness varies with distance via exp[-α p^2]
        Evaluate new solutions and update brightness
    end for j
  end for i
 Rank the fireflies and find the current best
 s=s+1
 end while
 Post process results and visualization
end
```

Figure 1. Pseudocode of the standard firefly algorithm

The dimension of the function to be optimized is given by d, n is the number of fireflies, s_{max} is the maximum number of generations, α is the light absorption coefficient, L_i is the light intensity and the distance p between any two fireflies i and j located at positions M_i and M_j can be evaluated as follows.

$$p_{ij} = Distance\left(M_i, M_j\right) = \sqrt{\sum_{k=1}^{d}\left(m_{i,k} - m_{j,k}\right)^2} \tag{1}$$

The light intensity (L) decreases as the square of the distance increases (p^2). It can be approximated using the following form.

$$L\ (p) = L_0 e^{-\alpha p^2} \tag{2}$$

where, L_0 is the light intensity at source. As the firefly's attractiveness is proportional to the light intensity, we can define the attractiveness σ as follows;

$$\sigma\ (p) = \sigma_0\, e^{-\alpha p^2} \tag{3}$$

Here, σ_0 is the attractiveness at p = 0. Now the movement of a firefly i attracted to another more attractive firefly j is given by,

$$m_{i+1} = m_i + \sigma_0\, e^{-\alpha p_{ij}^2}\left(m_j - m_i\right) + \lambda\left(rand() - 0.5\right) \tag{4}$$

where, λ is the randomization parameter and rand ( ) is a random number generator.

Even though the standard FA outperforms the other evolutionary algorithms like genetic algorithm it faces some difficulties like premature convergence and obtaining better solutions.

5. OPPOSITION-BASED FIREFLY ALGORITHM

In order to overcome the above mentioned problems of FA, a novel approach called opposition-based learning (OBL) suggested by Tizhoosh [18] has been applied with FA. It has been successfully applied with several optimisation algorithms like genetic algorithm, differential evolution algorithm, ant colony optimisation and gravitational search algorithm. In OBL the candidate solution and its corresponding opposite solution are considered simultaneously. Let z ∈[x, y] be a real number, the opposite number of z is denoted as z` and is defined as:

$$z` = x + y - z \tag{5}$$

The above concept can be extended to the case of higher dimensions. Let Q (z_1, z_2,..., z_m) be a m-dimensional vector, where $z_i \in [x_i, y_i]$ and i = 1, 2,..., m. The opposite vector of Q is defined by Q` = ($z`_1$, $z`_2$,..., $z`_m$), where $z`_i = x_i + y_i - z_i$.

The proposed algorithm applies OBL concept in two phases of optimisation namely initialising the population and producing new generations.

Initially a population of n fireflies is generated. The opposite position of each firefly is computed using equation 5. Each firefly is evaluated using the fitness function and the n fittest individuals are selected from the total of 2n individuals based on the fitness value. The basic steps of the OBFA are shown as pseudo code in figure 2.

Additionally, the OBFA uses OBL technique for producing new generations and updating the firefly's positions. In this method, e fireflies yielding the worst fitness values are replaced by their opposite fireflies at each iteration of the optimization process. At the start, variable e should possess a larger value to provide an effective global search. As the iteration increases, the value of e should be reduced to provide a local exploitation. Therefore, the value of e is given as follows:

$$e = \text{Round}\left[\frac{n}{3}\left(1 - \frac{S}{S_{max}}\right)\right] \tag{6}$$

where, Round (x) rounds the value of x to the nearest integer and s_{max} is the maximum number of generations.

```
begin
   Fitness function f(M), M=(m1,...,md)^c
   Generate initial population of fireflies Mi(i=1,...,n)
   Generate opposite population of fireflies OMi(i=1,...,n)
   Find the fittest population n from {Mi, OMi} as initial population
   Brightness Li at Mi is determined by f(Mi)
   Define light absorption coefficient α
   s=1
   while (s< smax)
   for i=1:n all n fireflies
     for j=1:n all n fireflies
        if (Lj>Li)
              Move firefly i towards j in d-dimension
        end if
        Attractiveness varies with distance via exp [-αp^2]
        Evaluate new solutions and update brightness
        Calculate e based on equation 6
        Change e worst fireflies with their opposite using equation 5
     end for j
   end for i
  Rank the fireflies and find the current best
 s=s+1
  end while
  Post process results and visualization
end
```

Figure 2. Pseudocode of proposed OBFA algorithm

6. PROPOSED OBFA OPTIMIZED FEATURE SUBSET SELECTION USING SVM

Feature selection is the process of excluding irrelevant features which may otherwise degrade the performance of the classifier. Feature selection is performed either as wrapper based or filter based. Wrapper based methods make use of the performance of a classifier to evaluate the feature subsets. On the other hand, the filter based methods use feature evaluation techniques.

In this work, SVM classifier is used to classify the CTG data with the complete set of 21 features. In addition, FA and OBFA has been used with SVM for finding the optimal feature subset. The features of the dataset are represented as a binary string of 0's and 1's. The value of 1 (one) represents the presence of a particular feature and 0 (zero) represents its absence. The whole data set is divided in to 75% (1594 instances) and 25% (532 instances) and used for training and testing the classifier respectively. 10 fold cross validation is applied to the training set and the testing set is completely hidden from the classifier during training.

The fitness function (F) is given by,

$$F = E_c + w_1 T_f \tag{7}$$

where, E_c is average accuracy rate of SVM classifier and T_f is number of zeros (absence of feature) in the feature subset and w_1 is the weight which is equal to 0.1. The parameters of Firefly algorithm are listed in Table 1.

Table 1. Firefly algorithm parameters

Number of fireflies	30
Number of generations	100
Randomisation parameter (λ)	0.5
Attractiveness (σ)	0.2
Light absorption coefficient (α)	1

7. RESULTS & DISCUSSION

Experiments have been performed using the original dataset and the optimal reduced data subset. The various performance measures being considered and their expressions are listed from equations (8) to (15).

$$\text{Accuracy} = \left[\frac{TP+TN}{TP+TN+FP+FN}\right] \tag{8}$$

$$\text{Sensitivity} = \left[\frac{TP}{TP+FN}\right] \tag{9}$$

$$\text{Specificity} = \left[\frac{TN}{TN+FP}\right] \tag{10}$$

$$\text{Positive Predictive Value (Precision): } PPV = \left[\frac{TP}{TP+FP}\right] \tag{11}$$

$$\text{Negative Predictive Value: } NPV = \left[\frac{TN}{TN+FN}\right] \tag{12}$$

Geometric mean: $\text{Gmean} = \sqrt{\text{specificity} \times \text{sensitivity}}$ (13)

$$\text{F-measure} = 2 \times \left[\frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \right] \quad (14)$$

$$\text{Area under ROC} = \frac{\text{specificity} + \text{sensitivity}}{2} \quad (15)$$

where, TP - True Positives, TN - True Negatives, FP - False Positives and FN - False Negatives The results are presented in Table 2.

Table 2. Comparison of SVM accuracy with and without feature selection

	Data set	**Average accuracy (%)**
Without FS	Full feature set	88.75
With FS	FA	91.92
	OBFA	92.85

It is found that the average accuracy is 88.75% with full feature set and the same is achieved as 91.92% with optimal feature set produced by FA and as 92.85% with optimal feature set produced by OBFA.

Table 3. Performance metrics of SVM with and without feature selection

Performance Metrics (%)	**Without FS**	**With FS**	
		FA	**OBFA**
Sensitivity	77.30	84.83	83.81
Specificity	90.22	93.78	93.72
PPV	78.56	83.14	85.45
NPV	90.70	93.26	95.02
G-mean	82.92	89.19	88.62
F-measure	77.92	83.94	84.62
Area under ROC	83.76	89.30	88.76

The results given in tables 2 and 3 are depicted in graphical form in figures 3 and 4 respectively for a better illustration.

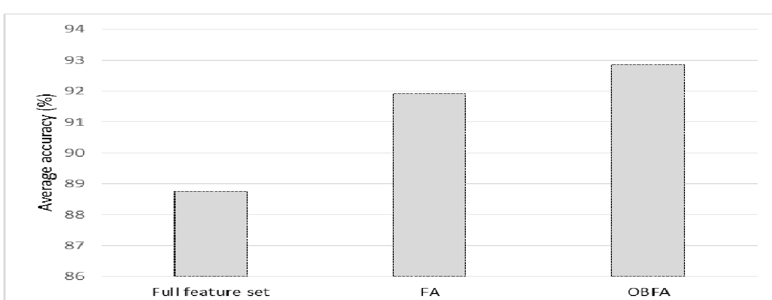

Figure 3. Comparison of SVM accuracy with and without feature selection

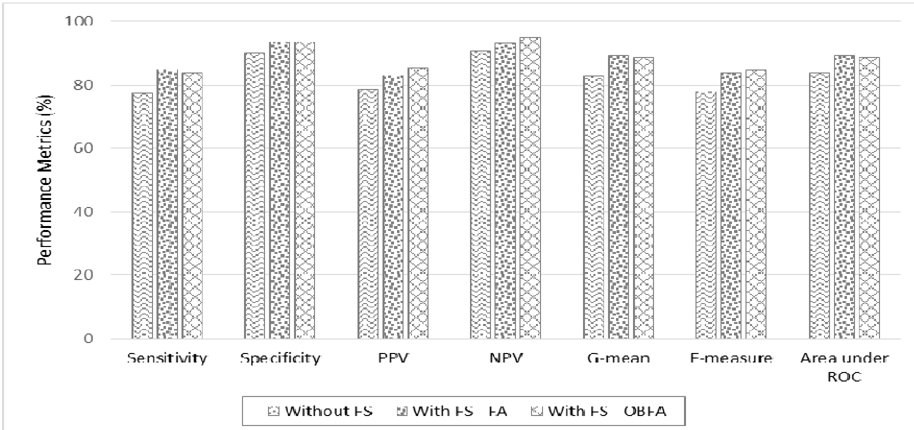

Figure 4. Performance metrics of SVM with and without feature selection

Tables 2 and 3 and figures 4 and 5 show the performance measures of SVM using the full feature set, optimal feature set using FA and OBFA. The results show that OBFA performs better than FA and full feature set.

8. CONCLUSION

In this paper, Opposition based Firefly algorithm is proposed for producing optimal feature set for CTG classification. CTG dataset from UCI Machine Learning Repository has been taken for experimentation. The classification results are presented in terms of Accuracy, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, Geometric Mean, F-measure and Area under ROC. The results of experiments show that there is a marginal improvement in the performance of proposed OBFA optimized classifier than the existing FA optimized classifier. However, there is a significant improvement in the performance of the proposed classifier when compared to the classifier with full feature set (without feature selection). This improvement in performance will ensure that the obstetricians can make more accurate decisions from CTG recordings.

REFERENCES

[1] Hasan Ocak, (2013) “A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being”, *Journal of Medical Systems*, Vol. 37, No.2, Article ID: 9913.

[2] Ersen Yılmaz & Çağlar Kılıkçıer, (2013) “Determination of fetal state from using LS-SVM with particle swarm optimization and binary decision tree”, *Computational and Mathematical Methods in Medicine,* Article ID: 487179.

[3] Hasan Ocak & Huseyin Metin Ertunc, (2013) “Prediction of Fetal State From the Cardiotocogram Recordings using Adaptive Neuro-Fuzzy Inference Systems”, *Neural Computing and Applications*, Vol.23, No.6, pp1583–1589.

[4] Peterek Tomas, Jana Krohova, Pavel Dohnalek & Petr Gajdos, (2013) “Classification of Cardiotocography Records by Random Forest”, *IEEE 36th International conference on telecommunications and signal processing*, pp620-623.

[5] Mei-Ling Huang & Yung-Yan Hsu, "Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network", *Journal of Biomedical Science and Engineering*, Vol. 5, No. 9, pp526-533.

[6] Hakan Sahin & Abdulhamit Subasi, (2012) "Classification of Fetal State from the Cardiotocogram Recordings using ANN and Simple Logistic", *In: 3rd International Symposium on Sustainable Development*, Sarajevo, pp499-505.

[7] Mohamed El Bachir Menai, Fatimah J. Mohder & Fayha Al-mutairi, (2013) "Influence of Feature Selection on Naïve Bayes Classifier for Recognizing Patterns in Cardiotocograms", *Journal of Medical and Bioengineering*, Vol. 2, No.1, pp66-70.

[8] Shivajirao Jadhav, Sanjay Nalbalwar & Ashok Ghatol, (2011) "Modular Neural Network Model Based Foetal State Classification", *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp915-917.

[9] Sundar.C., M. Chitradevi & G. Geetharamani, (2013) "An Overview of Research Challenges for Classification of Cardiotocogram Data", *Journal of Computer Science*, Vol. 9, No. 2, pp198-206.

[10] Sundar.C, M.Chitradevi & G.Geetharamani, (2013) "An Analysis on the Performance of Naïve Bayes Probabilistic Model Based Classifier for Cardiotocogram Data Classification", *International Journal on Computational Sciences & Applications*, Vol. 3, No. 1, pp17-26.

[11] Mauricio Schiezaro & Helio Pedrini, (2013) "Data feature selection based on Artificial Bee Colony algorithm", *EURASIP Journal on Image and Video Processing*, Vol. 47, pp1-8.

[12] Mustafa Serter Uzer, Nihat Yilmaz & Onur Inan, (2013) "Feature Selection Method based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", *The Scientific World Journal*, Article ID 419187.

[13] V.Subha & D.Murugan, (2014) "Fetal State Determination using Support Vector Machine and Firefly Optimization", *International Journal of Knowledge Based Computer Systems*, Vol. 2, No. 2, pp7-12.

[14] V.Subha, D.Murugan, S.Prabha & A.Manivanna Boopathi, (2015) "Genetic Algorithm based feature subset selection for fetal state classification", *Journal of communications technology, Electronics and Computer Science*, Vol. 2, pp13-17.

[15] K.Bache & M.Lichman (2013) *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[16] Sindhu Ravindran, Asral Bahari Jambek, Hariharan Muthusamy & Siewchin Neoh, (2015) "A Novel Clinical Decision Support System Using Improved Adaptive Genetic Algorithm for the Assessment of Fetal Well-Being", *Computational and Mathematical Methods in Medicine*, Article ID: 283532.

[17] Shuhao Yu, Shenglong Zhu, Yan Ma & Demei Mao, (2015) "Enhancing firefly algorithm using generalized opposition-based learning", *Computing*, Vol. 97, No. 7, pp741-754.

[18] Tizhoosh HR, (2006) "Opposition-based learning: a new scheme for machine intelligence", *International Conference on Computational Intelligence for Modelling, Control and Automation Jointly with International Conference on Intelligence Agents, Web Technologies and Internet Commerce*, Vol. 1, pp695-701.

Authors

V. Subha received B.E. (Electronics and Communication Engineering) from Bharathiyar University Coimbatore in 2000. She received her M.E. (Computer Science and Engineering) from Manonmaniam Sundaranar University, Tirunelveli in 2002. She is presently working as Assistant Professor and pursuing research towards Ph.D. degree in Manonmaniam Sundaranar University. Her areas of interest are Data Mining and Pattern recognition.

Dr.D.Murugan received B.E. (Electronics and Communication Engineering) and M.E. (Computer Science and Engineering) from Madurai Kamaraj University respectively. He received his Ph.D. from Manonmaniam Sundaranar University, Tirunelveli. He is presently working as Associate Professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University. Presently he is a Principal Investigator of the twofunded research projects granted by UGC and DST. His areas of interest are Image processing, Software engineering, Data mining and Pattern recognition.