Genetic Algorithm based feature subset selection for fetal state classification

V.Subha, D.Murugan and S.Prabha
Department of Computer Science and Engineering
Manonmaniam Sundaranar University
Tirunelveli, India.

A.Manivanna Boopathi,
Department of Electrical and Electronics Engineering,
PSN College of Engineering & Technology,
Tirunelveli, India.

***Abstract*—Huge amount of data are available in the field of medicine which are used for diagnosing the diseases by analysing them. Presently, prediction of diseases are made easier and accurate by employing various data mining techniques to extract information from these medical data. This paper presents an improved method of classifying the cardiotocogram data using Multiclass Support Vector Machine (MSVM) through an optimized feature subset produced by Genetic Algorithm (GA). Various performance metrics have been evaluated and the experimental results exhibit improved performance when using optimized feature set comparing to the full feature set.**

Keywords—Feature Selection; SVM Classifier; Cardiotocogram; Genetic Algorithm, Multiclass SVM.

I. INTRODUCTION

Since 1960s, Cardiotocography (CTG) has been in use to find and assess the fetal state during pregnancy and delivery [1]. The fetal state comprises of two signals viz. Fetal Heart Rate (FHR) and Uterine Contractions (UC). Obstetricians use CTG to determine the oxygen level of babies and to decide if the baby can be given a natural birth or caesarean section. To get rid of incorrect interpretations by manual analysis of CTG data, computer assisted decision making systems are used by the obstetricians to classify the recorded CTG data and make a decision [2].

Variety of methods have been proposed in literature for analysing the CTG data. A SVM classifier has been proposed to classify the fetal state in to two classes [1]. Further, GA has been used for selecting the most relevant features and thereby the performance of the classifier has been improved. In [2], a Least squares-SVM, Particle Swarm Optimization and binary decision tree have been used to classify the CTG data.

Using soft computing techniques, an adaptive neuro fuzzy inference system has been presented in [3] to classify the CTG data of fetal state into two classes. CTG data have been classified using Random forest classifier combined with feature reduction technique in [4]. Improved classification performance has been achieved in [5] by using discriminant analysis, decision tree and artificial neural network for fetal distress prediction. In [6], a classifier has been proposed for CTG data which uses neural network and simple logistics. Naïve Bayes Classifier has been used for classification of CTG data along with feature selection approaches in [7]. In [8], a classifier has been proposed which classifies the data into three classes by applying modular neural network. A neural network based classifier has been presented in [9], to improve the performance of 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 has been presented. Further, in [12], a feature selection method has been reported which is based on Artificial Bee Colony algorithm and Support Vector Machines for medical datasets classification. Recently in [13], a fetal state classifier using SVM has been proposed. Firefly algorithm has been used for producing optimal feature set and thereby the performance of classification has been improved.

In this paper, a Multiclass Support Vector Machine (MSVM) classifier has been proposed for classification of CTG data. GA has been used to produce optimal feature set which results in improvement of performance of the proposed classifier. Initially, CTG data are classified using the full feature set. Then, optimal and reduced feature set has been produced using GA with MSVM classification. The experimental results reveal that the use of optimal feature set improves the performance of classification.

This paper has been organized as follows; Section 2 describes the CTG data set. The Multiclass Support Vector Machine has been explained in section 3. The method of finding the optimal and reduced data set has been explained in section 4 followed by the results and discussion in the section 5. Finally, the section 6 concludes the proposed method.

II. CTG DATA SET

The CTG dataset of UCI Machine Learning Repository [14] has been used for experiment. There are totally 2126 fetal

cardiotocograms with 21 attributes and 1 class attribute in this data. Three expert obstetricians have classified this data set consisting of measurements of fetal heart rate (FHR) and uterine contractions (UC) and assigned classification labels to them based on the fetal heart rate class codes (N-Normal, S-Suspect and P-Pathologic).

A. Attribute Information

1. LB - FHR baseline (beats per minute)
2. AC - Number of accelerations per second
3. FM - Number of fetal movements per second
4. UC - Number of uterine contractions per second
5. DL - Number of light decelerations per second
6. DS – Number of severe decelerations per second
7. DP - Number of prolonged decelerations per second
8. ASTV - percentage of time with abnormal short term variability
9. MSTV - mean value of short term variability
10. ALTV - percentage of time with abnormal long term variability
11. MLTV - mean value of long term variability
12. Width - width of FHR histogram
13. Min - minimum of FHR histogram
14. Max - Maximum of FHR histogram
15. Nmax - Number of histogram peaks
16. Nzeros - Number of histogram zeros
17. Mode - histogram mode
18. Mean - histogram mean
19. Median - histogram median
20. Variance - histogram variance
21. Tendency - histogram tendency
22. CLASS- fetal state class code (Normal=1; Suspect=2; Pathologic=3)

III. Multiclass Support Vector Machines

Even though there are variety of data classification methods used, Support Vector Machine (SVM) is still being widely used for classification of data [15]. Consider a set of m training samples $\{(p_i, q_i), \ \forall i = 1,2,....m\}$, where pi is the input feature vector for the i^{th} sample and q^i is corresponding value of target class output.

The aim is to determine the optimal values of the weight vector w and bias b by satisfying the constraint,

$$\begin{aligned} &d_i\left(w^T p_i + b\right) \geq 1 - \varepsilon_i \qquad for i = 1,2,3,..m \\ &\varepsilon_i \geq 0 \qquad\qquad\qquad\qquad for\ all\ i \end{aligned} \tag{1}$$

and weight vector w and the variables ε_i minimize the cost function

$$\Phi(w, \varepsilon) = \frac{1}{2} w^t w + D\sum_{i=1}^{m} \varepsilon_i \tag{2}$$

Where, D > 0 is a user-specified regularization parameter.

This can be also written as; for the given training sample $\{(p_i, d_i)\}_{i=1}^{m}$, the Lagrange multiplier can be calculated as $\{\beta_i\}_{i=1}^{m}$ to maximize the objective function

$$R(\beta) = \sum_{i=1}^{m} \beta_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m} \beta_i \beta_j d_i d_j p^t{}_i p_j \tag{3}$$

subject to the constraints

$$\begin{aligned} &\sum_{i=1}^{M} \beta_i d_i = 0 \\ &0 \leq \beta_i \leq D \quad for\ i = 1,2,...m \end{aligned} \tag{4}$$

After finding the Lagrange multipliers, the optimal solution for the weight vector w can be found as,

$$W_o = \sum_{i=1}^{m} \beta_{0,i} d_i p_i \tag{5}$$

and the optimal bias b_0 can be obtained using the following equation

$$b_o = 1 - w_0{}^t p \tag{6}$$

If the patterns are not linearly separable in the current input space, the SVM can do a nonlinear transformation through the inner-product kernel L (p_i, p_j) to map the current space to a new feature space where the patterns cane be linearly separable. This kernel function will lead to a decision function which is non-linear in the input space but its image will be linearly separable in the high dimensional feature space. This can be stated as

Maximize:

$$R(\beta) = \sum_{i=1}^{m} \beta_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m} \beta_i \beta_j d_i d_j L\left(p_i, p_j\right) \tag{7}$$

Subject to the constraints

$$\sum_{i=1}^{M} \beta_i d_i = 0$$

$$0 \le \beta_i \le D \qquad for\ \ i = 1,2,.....,m \qquad (8)$$

It is a well-known fact that SVM is meant for only classification into two classes [16]. By constructing and combining several binary classifiers, an n-class SVM can be designed. The n-class SVM can either be a One-Against-All or One-Against-one type. In this work, a One-Against-All type has been used. Here, 3 binary classifiers are trained to distinguish one class from other 2 classes.

IV. Genetic algorithm

Since the early 70s, Genetic Algorithms (GAs) are being popular in many engineering applications. Particularly, GA is popularly and widely used for finding feature subset selection algorithms [15]. GAs are stochastic search algorithms based on the phenomena of natural selection and evolutionary process. Initially a set of candidate solutions is taken which is called as population. Each individuals of this population are traditionally encoded as a binary bit string. These bit strings are named as chromosomes or genotypes. Fitness of each of these individuals is found by evaluating the objective function. The more fit individuals are retained in the population and the bit strings of other individuals are modified to form new individuals which will replace the unfit individuals of current population thereby making a next generation. The modifications in the bit strings are performed by means of genetic operations such as crossover and mutation. The fitness of the individuals of this new generation is found and based on the fitness of individuals, the next generation is created. After a number of iterations, the population will contain the individuals of better fitness. This iterative process is terminated if the maximum number of generations is reached or the required fitness is reached.

A. Genetic Algorithm optimized feature subset selection using SVM evaluation

Feature selection is a process in which the irrelevant features are identified and excluded from the full set, thereby producing an optimal feature subset. This process improves the classification performance considerably. 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, GA has been used with MSVM to find the optimal feature subset. The features of the dataset are represented as a binary strings 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. Cross validation is applied to the training dataset and finally the classifier is trained with the optimal feature subset. The training set accuracy is determined and then the test set is applied to the trained classifier to find the accuracy.

The objective function (F) is given by,

$$F = w_1 E_c + w_2 T_f \qquad (9)$$

where, w_1 and w_2 are weights, E_c is average misclassification rate of MSVM classifier and T_f is the number of features in the selected feature subset. The values of weights w_1 and w_2 are set to 1 and 0.1 respectively.

In GA, the size of initial population is taken as 30 and the maximum number of generations as 100. In each chromosome a gene value of '1' means the particular feature is selected for evaluation. If it is '0', the feature is not selected. The subset of features which maximizes the objective function till the termination condition (i.e. maximum number of iterations) is reached is chosen as the optimal feature subset. The parameters of GA are listed in Table 1.

TABLE I. Genetic algorithm parameters

Population size	30
No. of genes per individual	21
Elite count	2
Crossover percentage	70%
Mutation percentage	30%
Max. no. of generations	100

V. Results & discussion

Experiments have been performed using the original dataset and the optimal reduced data subset. The results are presented in Table 2. It is found that the average accuracy is 88.75% with full feature set and the same is achieved as 91.35% with optimal feature set.

TABLE II. Comparison of MSVM accuracy with and without feature selection

Data set	**Accuracy (average)**
Full feature set	88.75
Training set	94.85
Testing set	91.35

The results given in Table 2 are shown in graphical form in Fig. 1 for a better illustration.

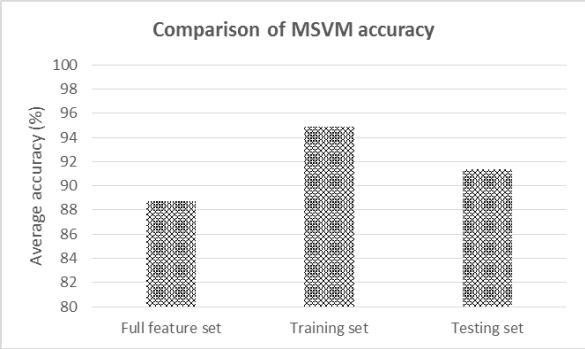

Fig. 1. Comparison of MSVM accuracy with and without feature selection

Following are the features selected by GA;

- LB- FHR baseline (beats per minute)
- AC- Number of accelerations per second
- FM - Number of fetal movements per second
- DL - Number of light decelerations per second
- DS – Number of severe decelerations per second
- DP - Number of prolonged decelerations per second
- ASTV-Percentage of time with abnormal short term variability
- ALTV-Percentage of time with abnormal long term variability
- Min-Minimum of FHR histogram
- NZeros- Number of histogram zeros
- Median- Histogram median
- Variance-Histogram variance

The various performance measures being considered and their expressions are listed from equations (10) to (17) and the results have been presented in Table 3.

$$\text{Accuracy} = \left[\frac{TP+TN}{TP+TN+FP+FN}\right] \quad (10)$$

$$\text{Misclassification rate= 1- accuracy} \quad (11)$$

$$\text{Sensitivity} = \left[\frac{TP}{TP+FN}\right] \quad (12)$$

$$\text{Specificity} = \left[\frac{TN}{TN+FP}\right] \quad (13)$$

$$\text{Positive Predictive Value: } PPV = \left[\frac{TP}{TP+FP}\right] \quad (14)$$

$$\text{Negative Predictive Value: } NPV = \left[\frac{TN}{TN+FN}\right] \quad (15)$$

$$\text{Geometric mean: } Gmean = \sqrt{speificity \times sensitivity} \quad (16)$$

$$\text{F1-measure} = 2 \times \left[\frac{precision \times sensitivity}{precision \times sensitivity}\right] \quad (17)$$

where, TP - True Positives, TN - True Negatives, FP - False Positives and FN - False Negatives

TABLE III. PERFORMANCE METRICS (AVERAGE) OF MSVM

Performance Metrics (%)	**Without FS**	**With FS**
Sensitivity	77.30	80.71
Specificity	90.22	92.50
PPV	78.56	83.06
NPV	90.70	93.77
Gmean	82.92	85.92
F1-measure	77.92	81.87

The performance metrics of MSVM given in Table 3 are presented graphically in Fig. 2 for a better illustration.

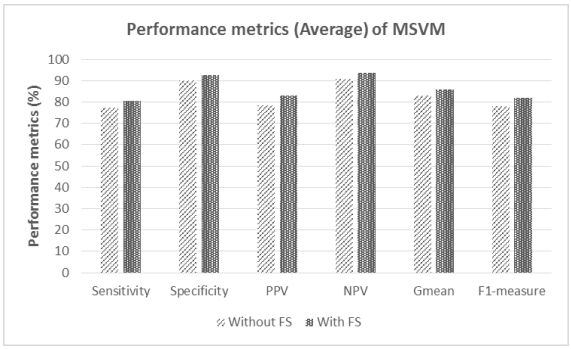

Fig. 2. Performance metrics (Average) of MSVM

The values of performance measures determined from the proposed classifier, given in Table 3, also depict that the optimal feature set improves the classification performance of MSVM than the full feature set.

This improvement in classification performance using Genetic Algorithm optimized feature subset will help the obstetricians in making more accurate decisions for fetal state anticipation.

VI. Conclusion

In this paper, a classification system for CTG data using Multiclass Support Vector Machine is proposed. Genetic Algorithm has been used to find an optimized and reduced feature set in order to improve the performance of the proposed classifier. CTG dataset from UCI Machine Learning Repository has been taken for experimentation. The results of experiments show that the performance of proposed classifier is considerably improved when using optimal feature set than the full feature set. This improvement in performance will ensure that the obstetricians can make more accurate decisions from CTG recordings.

References

[1] Hasan Ocak, "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, 2013.

[2] Ersen Yılmaz and Çağlar Kılıkçıer, "Determination of fetal state from using LS-SVM with particle swarm optimization and binary decision tree," Computational and Mathematical Methods in Medicine, 2013.

[3] Hasan Ocak and Huseyin Metin Ertunc, "Prediction of Fetal State From The Cardiotocogram Recordings using Adaptive Neuro-Fuzzy Inference Systems," Neural Comput & Applic., vol.23, no.6., pp. 1583–1589, 2013.

[4] Peterek Tomas, Jana Krohova, Pavel Dohnalek and Petr Gajdos, "Classification of Cardiotocography Records by Random Forest," IEEE 36th International conference on telecommunications and signal processing, pp.620-623, July 2013.

[5] Mei-Ling Huang and Yung-Yan Hsu, "Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network," J. Biomedical Science and Engineering, vol. 5, no. 9, pp.526-533, 2012.

[6] Hakan Sahin and Abdulhamit Subasi, "Classification of Fetal State from the Cardiotocogram Recordings using ANN and Simple Logistic," In: 3rd International Symposium on Sustainable Development, Sarajevo, pp.499-505, 2012.

[7] Mohamed El Bachir Menai, Fatimah J. Mohder and Fayha Al-mutairi, "Influence of Feature Selection on Naïve Bayes Classifier for Recognizing Patterns in Cardiotocograms," Journal of Medical and Bioengineering, vol. 2, no.1, pp.66-70, 2013.

[8] Shivajirao Jadhav, Sanjay Nalbalwar and Ashok Ghatol, "Modular Neural Network Model Based Foetal State Classification," IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp.915-917, November 2011.

[9] Sundar.C., M. Chitradevi and G. Geetharamani, "An Overview of Research Challenges for Classification of Cardiotocogram Data," Journal of Computer Science, vol. 9, no. 2, pp. 198-206, 2013.

[10] Sundar.C, M.Chitradevi and G.Geetharamani, "An Analysis on the Performance of Naïve Bayes Probabilistic Model Based Classifier for Cardiotocogram Data Classification," International Journal on Computational Sciences & Applications, vol. 3, np. 1, pp.17-26, 2013.

[11] Mauricio Schiezaro and Helio Pedrini, "Data feature selection based on Artificial Bee Colony algorithm," EURASIP Journal on Image and Video Processing, vol. 47, pp.1-8, 2013.

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

[13] V.Subha and D.Murugan, "Fetal State Determination using Support Vector Machine and Firefly Optimization," International Journal of Knowledge Based Computer Systems, vol. 2, no. 2, pp. 7-12, 2014.

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

[15] R.K.Agarwal and Rajni Bala, "A hybrid approach for selection of relevant features for microarray datasets," International Journal of Computer, Control, Quantum and Information Engineering, vol. 1, no.5, pp. 1319-1325, 2007.

[16] Yashima Ahuja and Sumit Kumar Yadav, "Multiclass Classification and Support Vector Machine," Global Journal of Computer Science and Technology Interdisciplinary, vol. 12, no. 11, pp. 14-20, 2012.