Preterm Birth Prediction using Hybrid Ant Colony-Genetic Optimization Algorithm in Data Mining

M. Varusai Mohamed, P. Mayilvahanan

Abstract : Many data mining (DM) methods are used to explore the risk factors of Preterm birth (PTB) and to predict preterm birth. High rates of infant mortality, preterm births and maternal mobility and continuous variation in pregnancy outcomes are an important public health issue in India and worldwide. In this paper, aims to develop and evaluate prevent factors of preterm birth using hybrid algorithm which is optimized the model with genetic algorithm and ant bee colony algorithm based also analysis of risk factor of preterm birth prediction. It is identified that variables which were highly influenced to forecast less weight child birth are Mother's weight (pounds) before pregnant, age of Mother, during first three months the number of physician meet, number of early premature labors. The results of this work have improved prediction accuracy when compared with other optimization techniques. Maximum accuracy of 0.9629 is produced in proposed method.

Keywords : Preterm birth, optimization, classification, accuracy, iteration

I. INTRODUCTION

PTB is important public long lasting health issue having huge sensitive and economical weak to societies and family. It mainly leads to long duration disability and neonatal mortality. Moreover, annually 26 billion dollars are spent care and delivery of 12 to 13% of child who born early i.e before pregnancy due date in United States [1]. The main threat is to find the women with major risk for early premature labor and creates interference. As much of importance is given to find the women with low risk to avoid insignificant and interference cost. Due to lack of early pregnancy history it will mostly occurs in first time mothers (nulliparous women) in the particular challenging population [2].

A compelling application from machine learning viewpoint are shown by analyzes of premature labor. It has increasingly challenging issues majorly due to genetic complication of heterogeneous multifactorial etiology, pregnancy physical dynamics and reduction in strategies ability in interpreting and integrating huge multidimensional information. Its major risk is heterogenous and has the history like age, bacterial vaginosis, race, urinary tract infection, parity, bleeding, smoking, cervix length of the mother [3]. Using analyses of coincidence with PTB the major studies concluded individual risks independently of each other. This study led to observation of PTB issue, current methods insufficiently good prediction used clinically. Using multivariate logistic regression method the past results gives sensitivity of 18.2% & 24.2% and specificity 28.6% & 33.3% for first time mother and borne having more than one child [4].

For health and economy are the important factors for predicting the premature birth and analyzing premature labor. The effects of the premature birth are reduced by involvement of major efforts. The suggestion of prevention gives good outcome this is because this method remains costly [5] [6]. The major finding of premature births gives increasing prevention by lifestyle interference and correct medical. The important hopeful technique is use of EHG. It analyze the electrical movement of uterus as well as specific form of EMG, and also in muscular tissue. This analyses represents that EHG record varies from woman to woman either true or false labor and premature or deliver term. It gives a strong support for premature birth diagnosis and objective prediction. The detection and prediction of true labor uses EHG in many research studies.

On comparing this research target on EHG usage to find the delivery is term or preterm. This is obtained by using new optimization technique detailed in this research and it is calculated against many existing machine learning classifiers by open dataset [7]. The classifying term and preterm records are suited by pre-selected features and signal filter, it is used to obtain the feature set from coarse signal and it is used by all classifiers. The results represents the selected classifiers gives numerous methods and it is used in many literatures.

II. LITERATURE SURVEY

Tuyen Tran et.al have suggested techniques for premature birth prediction consist of risk factors, discovering, derivation and qualifying of illustratable prediction rules and use of stabilized sparse logistic regression to determine linear prediction models. For the estimation of upper-bound accuracy for data[8] a hybrid method called Randomized Gradient Boosting used by the authors.

Batoulet.al created an ANN method to determine the preterm birth on PPPESO database based on baby's gender, current pregnancy number of babies, preterm babies, smoking,

Revised Manuscript Received on January 30, 2020.

***** Correspondence Author

M. Varusai Mohamed*, Research Scholar, School of Computing Science, VISTAS (Vels University), Chennai, India. Email: varusai76@yahoo.com.

Dr. P. Mayilvahanan, Research Supervisor, Department of Computer Application, VISTAS (Vels University), Chennai, India. Email: mayilkadir@yahoo.com

physician input, parity, eight obstetrical variables i.e Age, previous term and smoking on the basis of BP feed forward ANN and for increasing accuracy [9] it uses the effectiveness of weight elimination cost function.

M. M. Van Dyne et.al suggested LERS in which the rules are obtained from the data of preterm birth prediction. This method can give 78% and 90% for full term and preterm and 73% for combined cases [10].

Ryan W Loftin, every year worldwide there are 15 million babies born prematurely and cause disability. The telemedicine monitoring is important for caring the preterm babies immediately after birth. An EMG uterine monitoring transmission solution EUMTS is proposed in this research the issues by efficient EMG data loss compression is solved by using lossless efficient real time EMG transmission solution. The mother's EMG signals are sensed by the mo0nitoring system and sending it to the health device by wireless communication[11]

S. Karger AG discovered that the frequent and sever disease caused for preterm baby is Respiratory disease. For a preterm respiratory disease the clinical impact for both antioxidant and neurotoxin is Bilirubin. In respiratory the oxidative stress is involved. For the prevention of genotypes the antioxidants and preterm babies disease are considered as important and it is combined with high requirement of oxygen supply and BPD risk in preterm newborns [12].

Colin Joseph Brown., (2017) suggested "Modeling and Prediction of Neurodevelopment in Preterm Infants using Structural Connective Data." He found that ever year millions of infants born before 32 weeks postmenstral age (very preterm). The altered and neurodevelopment[13] are the major risks for very preterm birth babies. The dMRI studies shows that very preterm birth can affect the brain white matter maturation.

Masaki Ogawa et al. suggested for the identification of combination between obstetric complication and causative determinant uses the univariate analysis and he used multivariate analysis for unconditional logistic regression [14].

Nynke R. van den Broek et al. for the discovery of factors involved in early and late preterm birth he uses the multivariate logistic regression. The major risks in preterm birth is indentified maternal underweight, anemia, women's pregnancy history and malaria and it does not affected by the HIV status [15].

III. PROPOSED METHODOLOGY

The clinical databases contain huge number of data's about patient's clinical history. The effective decision making and patterns and relationships of data are provided by the Data mining process. For supporting and improving the capacity of medical diagnosis in healthcare applications the major data mining method involved is classification analysis. In this method it can automatically updates the ensemble method and uses traditional mining classifier since it shows the new ideas in data streams. For novel class detection this method recognize idea that data points of same class are close to each other and far away for different classes. If data point is separated from existing data clusters, it is identified as novel class instance.

A. Attribute Measurement

Let us consider set of N labeled data: {xr;i=1 ,N}, where .r, k} is class label and every xr is vector with n numerical components, each component being related to attribute value. We also consider that each class Cr contains nr data. Intraclass distance is a measure which reflects density of data within class and can be defined as average distances between data and center of their corresponding classes:

D (1) = 1/N $\sum_{r=0}^{n} d(x(i),m(r))$ (1)

Where m(r) = ($\sum_{i=0}^{n} y(i)/nr$ denotes center of class Cr and d is standard Euclidean distance.

Separability between classes is measured by interclass separation which is defined as average distance between centers of classes and entire data set.

D (2) = 1/N $\sum_{r=0}^{n} d(y(i),n(r))$ (2)

The smaller and larger interclass distance are easier to divide data, the attributes are related to classification. Every subset is defined by vector weight for attribute subsets estimation. (w1,...,wn), where wi E [0,1] can be explained as attribute relevance i. If the binary values are assigned to weights then value means attribute i is neglected. The attributes weights are included to inter and intra classes distances. Euclidean distance involved in (1) and (2) should be replaced with

D (x,y) = $\sqrt{\sum_{i=1}^{n} w\hat{}2(x(i)-y(i)\hat{}2}$ (3)

B. Database attributes with risk factors

Maternal age: 15-20, 21-30,31-49,≥40
Body Mass Index: underweight multiparas, obese nulliparas
Smoking: yes/no
Parity: Low birth weight, very high birth weight
No of pregnancies: 1, 2, 3, more than 3
Hemoglobin level: 10, more than 10
Low/high red cell count: normal, abnormal
Glucose level:3.9±1.4 , 4.8±2.1
BP: Systolic-normal/abnormal, diastolic-normal, abnormal
Abdominal perimeter: circumference, mean and standard deviation
Weight gain during pregnancy: underweight, over weight
Fundus uterus-height: high, less
Types of birth: vaginal/C-Section
Child sex: male/ female
Child head perimeter: normal abnormal
Live/still birth

C. Hybrid Optimization Algorithm

This method uses GA for update initial pheromone values and to generate available results. Until the optimum is reached an ACO implementation searches. The proposed dynamic integration method is used to achieve fusion time i.e approximately equal to □□, that contains minimum iteration □□min (□□ moment), maximum iteration □□max (□□ moment), and constant □□die for their GA. The result was less than constant for □□die generation when the evolutionary loop returned, the initiate ACO search and termination for GA is due to hybrid algorithm.

For three iterations of loop (□□die = 3), evolutionary rate is low when compared with certain constant, to engage ACO and end GA loop the performance of GA is consider as low.

Determination of constant, the optimal fitness values and different constants ranging from 0.005-0.01 are compared.
Let S is given N features, and τ( ) f be pheromone level of feature f in selected subset of attributes, where s ⊂ S .
Initially, desirability of every feature is same, but desirability of those features are important and increases in every stage.
The F-Score is given by

$$\text{F-score }(x)=\frac{\sum_{i=1}^{n}(x-x(c))^2}{\frac{1}{n}\sum_{j=1}^{n} x(j)-x(c)} \qquad (1)$$

Pseudo code:

```
Input: Dataset Input for ACO: ACO's parameters [a, b, c and d] and number of ants
Input for GA: total iteration, mutation rate, crossover and number of chromosomes
Output: selected subset features
Assign same value for heuristic information and pheromone (Eq. 1)
% Chromosome generation
 Start Chromosome population randomly
Do (for each iteration)
{
Calculate population's fitness (association rule based classifier accuracy)
Parent selection with roulette wheel policy
Crossover and Mutation ( if their conditions occurs)
Delete extra Chromosomes (according classification error)
     }
 % Ant generation
A feature randomly set to each ant and calculates it
     Do (for each Ant)
{
Evaluate State Rule
Select next feature.
    Evaluate Ant (association rule based classifier accuracy)
 If three successive steps don't improve accuracy
     Continue;
    }
AcoGa←Merge the Chromosomes and Ants
Sort AcoGa according classification accuracy
Check termination condition (iteration number or accuracy)
 Return best subset
Best_AcoGa← Select the r-best of AcoGa
    Use Best_AcoGa in next population of GA beside random population
Update pheromone of all ants
End
```

The time complication of this algorithm is O(Imn), where I is iteration number, m is ants number and n is original features number. In worst case, every ant choses all attributes. After every feature is added to candidate subset, the heuristic is calculated and results in n estimation per ant. The mn evaluations can be performed after the first iteration in this method, After I iterations, the heuristic will be evaluated Imn times.

IV. Result Analysis

The statistical measures that can be considered are sensitivity, specificity, and accuracy

$$\text{Sensitivity} = \frac{TP}{TP+FN} * 100$$

$$\text{Specificity} = \frac{TN}{TN+FP} * 100$$

$$\text{Positive Prediction Value (PPV)} = \frac{TP}{TP+FP} * 100$$

$$\text{Negative Prediction Value (NPV)}, = \frac{TN}{TP+FN} * 100$$

Accuracy = TP + TN
TP +FP+FN +TN

Given two classes are considered in terms of positive Vs negative records. The positive records are referred by the True positives, simultaneously the negative records form the true negatives and both are labeled correctly by classifier. The negative records are false positives that are labeled incorrectly. Table 1 shows the comparison of proposed and existing methods cost function calculation in various level of iterations.

Table-1 Comparison of cost function with various level of iterations

Number of iterations	ACO+GA	Cuckoo Optimization Algorithm (COA)	Genetic algorithm (GA)
N=50	0.012	0.121	0.006
N=100	0.015	0.052	0.149
N=150	0.026	0.001	0.031
N=200	0.021	0.124	0.218
N=250	0.016	0.012	0.001

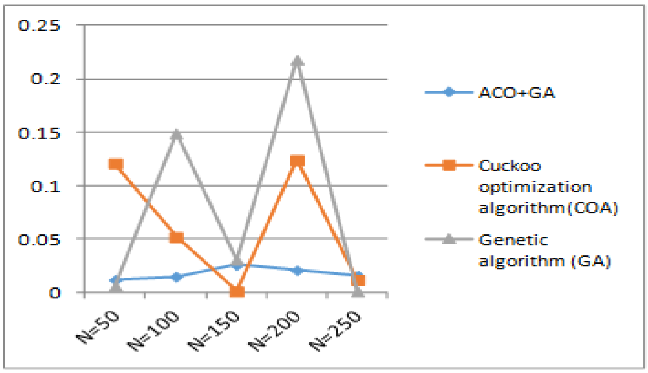

Figure-1 Cost function analysis

The above figure 1 shows the cost function analysis of proposed and existing methods with 250 iterations. For first 50 iteration, the ACO and GA method achieves the 0.02, Cuckoo optimization algorithm achieves the 0.12 also genetic algorithm achieves 0.08 respectively.

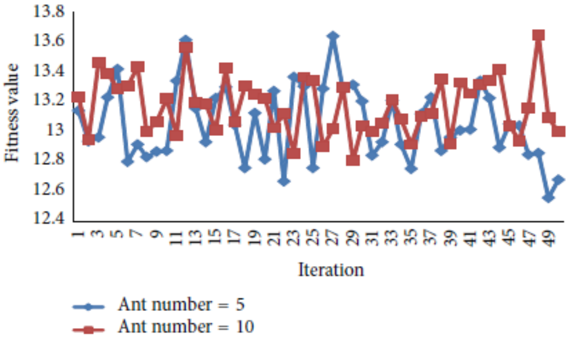

Figure-2 Calculation of fitness function

The above figure 2 shows the calculation of fitness function. The optimal values of pheromone coefficient a, heuristic coefficient b, and pheromone volatilization coefficient r were 0.4, 8, and 0.3, respectively; these outcomes are equal to traditional method, but they reached after 60 iterations.

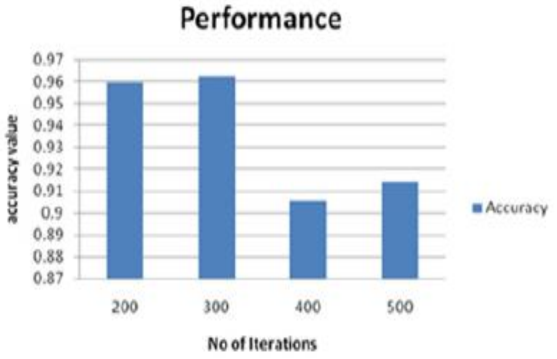

Figure-3 calculation of accuracy

In Figure 3 shows the method Maximum accuracy of 0.9629 is produced. It is very powerful in terms of classification of premature birth datasets.

V. CONCLUSION

The most commonly used data mining method is classification for predicting preterm birth and also for exploring risk factors of preterm birth also considered risk factors are socio-demographic, behavioral (life style) and Pregnancy History. The classification techniques performances are measured in terms of accuracy. The best performance of this method is accuracy rate of 96% (0.9629) that is achieved from iteration 200. This method also tries to reduce number of prediction rules required for classifying under testing and training data. The classification of premature birth is processed with the help of dividing attributes into two groups. They are attributes and opponents and its values are obtained over generations using hybrid technique. Future work is to apply for the prediction of preterm birth using Electro-hysterography (EHG) signal, EHG is used to measure electrical activity in the uterus.

REFERENCES

1. V. M. Allen, R. D. Wilson, and A. Cheung. Pregnancy outcomes after assisted reproductive technology. Journal of obstetrics and gynaecology Canada, 28(3):220 – 50, 2006. ISSN 1701-2163.
2. A. Conde-Agudelo, A. Rosas-Bermudez, and A. C. Kafury-Goeta. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. JAMA : the journal of the American Medical Association, 295(15):1809 – 23, 2006. ISSN 0098-7484.
3. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010
4. R Romero, J Espinoza, JP Kusanovic, F Gotsch, S Hassan, O Erez, T Chaiworapongsa, and M Mazor. The preterm parturition syndrome. BJOG: An International Journal of Obstetrics and Gynaecology, 113:17–42, 2006.
5. M. Lucovnik, R. J. Kuon, L. R. Chambliss, W. L. Maner, S.-Q. Shi, L. Shi, J. Balducci, and R. E. Garfield, “Use of uterine electromyography to diagnose term and preterm labor,” ActaObstetricia et GynecologicaScandinavica, vol. 90, no. 2, pp. 150–157, Feb. 2011.
6. R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, “Epidemiology and causes of preterm birth,” The Lancet, vol. 371, no. 9606, pp. 75–84, 2008.
7. L. J. Muglia and M. Katz, “The Enigma of Spontaneous Preterm Birth,” N Engl J Med, vol. 362, no. 6, pp. 529–35, 2010.
8. Tran T, Luo W, Phun D, Morris J, Rickard K. [2016] Preterm Birth Prediction: Deriving Stable and Interpretable Rules from High Dimensional Data. Proceedings of Machine Learning for Healthcare 2016 JMLR W&C Track (56)
9. Batoul A, Hamid A, Soheila K, et al.[2016] Using support vector machines in predicting and classifying factors affecting preterm delivery. Journal of Paramedical Sciences. 7(3): ISSN 2008-4978
10. Van Dyne, Frize M, Walker RC, petriu DC.[2006] Predicting High-Risk Preterm Birth Using Artificial Neural Networks. IEEE Transactions on Information Technology in Biomedicine, July 2006 , 10:540-549.
11. Ryan W Loftin, MouniraHabli, Candice C Snyder, Clint M Cormier, David F Lewis and Emily A DeFranco: “Late preterm birth”, 2010
12. S. Karger AG. Winters-Miner, “Seven ways predictive analytics can improve healthcare” published in Elsevier, Oct. 2014.
13. Collin, RaghuramThiagarajan, S. M. Reza Soroushmehr, FatemehNavidi, Daniel A. Beard and KayvanNajarian, “Big Data Analytics in Healthcare” published in BioMed Research International, 10.1155/2015/370194, Jul. 2017.
14. Masaki Ogawa, Yoshio Matsuda, Eriko Kanda, Jun Konno, Minoru Mitani, Yasuo Makino and Hideo Matsui, “Survival rate of extremely low birth weight infants and its risk factors: Case-Control study in Japan,” ISRN Obstetrics and Gynecology, Hindawi Publishing Corporation, Volume 2013,pp.1-6.
15. Nynke R. van den Broek, Rachel Jean-Bapsite, James P.Nelison, “Factors Associated with preterm, early preterm and late preterm birth in Malawi,” PLOS ONE, Volume 9, Issue 3, March 2014, pp.1-8.