International Journal of Trend in Research and Development, Volume 4(1), ISSN: 2394-9333
www.ijtrd.com

Cardiotocography - A Comparative Study between Support Vector Machine and Decision Tree Algorithms

Jagannathan D
M.Phil. (PG Scholar), Department of Computer Science, Dr. C. V. Raman University, Chhattisgarh, India

Abstract-- Cardiotocography (CTG) is a simultaneous recording of Fetal Heart Rate (FHR) and Uterine Contractions (UC). It is one of the most common diagnostic techniques to evaluate maternal and fetal well-being during pregnancy and before delivery. By observing the Cardiotocography trace patterns doctors can understand the state of the fetus. There are several signal processing and computer programming based techniques for interpreting a typical Cardiotocography data. Even few decades after the introduction of Cardiotocography into clinical practice, the predictive capacity of the these methods remains controversial and still inaccurate. In this paper, we implement a model based CTG data classification system using a supervised SVM and Decision Tree which can classify the CTG data based on its training data. According to the arrived results, the performance of the supervised machine learning based classification approach provided significant performance. We used Accuracy, Specificity, NPV, Precision, Recall and ROC as the metric to evaluate the performance. It was found that, the SVM based classifier was capable of identifying Normal, Suspicious and Pathologic condition, from the nature of CTG data with very good accuracy.

Keywords-- CTG, Data mining, Classification, Support Vector Machine and Decision Tree.

I. INTRODUCTION

Data mining refers to a collection of techniques that provide the necessary actions to retrieve and gather knowledge from an exhaustive collection of data and facts. Data is available in enormous magnitude, but the knowledge that can be inferred from the data is still negligible. Data mining concepts are focused on discovering knowledge, predicting trends and eradicating superfluous data. Discovering knowledge in medical systems and health care scenarios is a herculean yet critical task. Knowledge discovery describes the process of automatically searching large volumes of data for patterns that can be considered additional knowledge about the data. The knowledge obtained through the process may become additional data that can be used for further manipulation and discovery. Application of data mining concepts to the medical arena has undeniably made remarkable strides in the sphere of medical research and clinical practice saving time, money and life. Clinical data mining is the application of data mining techniques using clinical data. Clinical Data-Mining (CDM) involves the conceptualization, extraction, analysis, and interpretation of available clinical data for practical knowledge-building, clinical decision-making and practitioner reflection. The main objective of clinical data mining is to haul new and previously unknown clinical solutions and patterns to aid the clinicians in diagnosis, prognosis and therapy. Moreover application of software solutions to store patient records in an electronic form is expected to make mining knowledge from clinical data less stressful.

Cardiotocography (CTG) is a simultaneous recording of Fetal Heart Rate (FHR) and Uterine Contractions (UC). It is one of the most common diagnostic techniques to evaluate maternal and fetal well-being during pregnancy and before delivery. FHR patterns are observed manually by obstetricians during the process of CTG analyses. For the last three decades, great interest has been paid to the fetal heart rate baseline and its frequency analysis. Fetal Heart Rate (FHR) monitoring remains widely used as a method for detecting changes in fetal oxygenation that can occur during labor. Yet, deaths and long-term disablement from intrapartum hypoxia remain an important cause of suffering for parents and families, even in industrialized countries. Confidential inquiries have highlighted that as much as 50% of these deaths could have been avoided because they were caused by non-recognition of abnormal FHR patterns, poor communication between staff, or delay in taking appropriate action. Computation and other data mining techniques can be used to analyze and classify the CTG data to avoid human mistakes and to assist doctors to take a decision.

II. RELATED WORK

Many typical findings are included in a CTG and obstetricians make clinical decisions about the state of the fetus considering these findings. However the interpretation of the information provided by CTG is not standardized. The deficient interpretation of CTG leaded to unnecessary surgical intervention, e.g. increase in cesarean births [1]. Therefore, computer-based approaches are presented recently. Huang and Hsu [2] proposed discriminant analysis (DA), decision tree (DT), and artificial neural network (ANN) in their study to evaluate fetal distress by the same CTG data used in this study. They reached the results showing that the accuracies of DA, DT and ANN are 82.1%, 86.36% and 97.78% respectively, and 80%, 10%, and the remaining 10% of the whole dataset were randomly used for training, testing, and validation respectively. Sundar et al. [3] implemented a supervised ANN which can classify the CTG data, the results are evaluated with respect to rand index, precision, recall and f-Score. The authors presented another related work in which neural network based classification model has been compared with the most commonly used unsupervised clustering methods; Fuzzy C-mean and k-mean clustering [4]. The arrived results show that the performance of the supervised ANN approach provided outperformed the other compared unsupervised clustering methods significantly. In a study, least squares support vector machine (LS-SVM) is employed utilizing a binary decision tree is for classification of the same cardiotocogram data to determine the fetal state [5]. Particle swarm optimization (PSO) is used for the optimization of parameters of LS-SVM, they reached a classification accuracy rate of 91.62%.

III. MOTIVATION AND JUSTIFICATION

Cardiotocography (CTG), consisting of fetal heart rate (FHR) and tocographic (TOCO) measurements, is used to evaluate fetal well-being during the delivery. Since 1970, many researchers have worked different mining methods to help the doctors that interpret the CTG trace pattern from the field of signal processing and computer programming [15]. With the help of CTG trace pattern analysis the doctors with interpretations in order to reach a satisfactory level of reliability. So, they act as a decision support system in obstetrics. For everyday practice, none of them has been adapted worldwide. Baseline estimation in computer analysis of cardiotocographs , which is currently no consensus on the best methodology. More than 30 years after the introduction of antepartum cardiotocography into clinical practice, the predictive capacity of the method remains controversial. In a review of lot of articles published on this subject, it was found that its reported sensitivity varies between 2 and 100%, and its specificity between 37 and 100% [16]. So, in this work, we are going to evaluate two clustering algorithms for clustering CTG data.

IV. MATERIAL AND METHODOLOGY

A. Dataset Description

The Cardiotocography data set used in this study is publicly available at The Data Mining Repository of University of California Irvine (UCI). By using 21 given attributes data can be classified according to FHR pattern class or fetal state class code. In this study, fetal state class code is used as target attribute instead of FHR pattern class code and each sample is classified into one of three groups Normal, Suspicious or Pathologic. The dataset includes a total of 2126 samples of which is 1655 normal, 295 suspicious and 176 pathologic samples which indicate the existing of fetal distress.

Attribute information is given as:

LB—FHR baseline (beats per minute)
AC—# of accelerations per second
FM—# of fetal movements per second
UC—# of uterine contractions per second
DL—# of light decelerations per second
DS—# of severe decelerations per second
DP—# of prolongued decelerations per second
ASTV—percentage of time with abnormal short term variability
MSTV—mean value of short term variability
ALTV—percentage of time with abnormal long term variability
MLTV—mean value of long term variability
Width—width of FHR histogram
Min—minimum of FHR histogram
Max—Maximum of FHR histogram
Nmax—# of histogram peaks
Nzeros—# of histogram zeros
Mode—histogram mode
Mean—histogram mean
Median—histogram median
Variance—histogram variance
Tendency—histogram tendency
CLASS—FHR pattern class code (1 to 10)
NSP—fetal state class code (N = normal; S = suspect; P = pathologic)

B. Classification

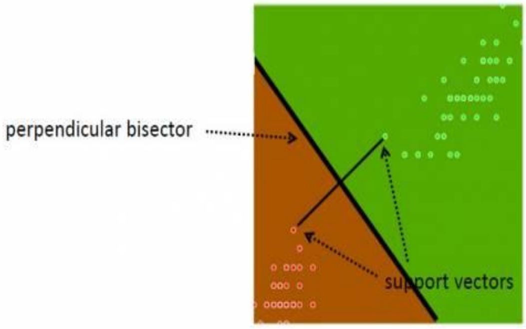
Classification process may be applied in different areas of research and practice, e.g., farms, military, medicine, Earth remote sensing. The classical classification techniques use statistical approach, which typically assumes the normal multidimensional distribution of probability in the experimental data set. Data classification may be supervised and unsupervised.

The supervised classification method requires the presence of training data set typically defined by the expert-the teacher. Each class of objects is characterised by the basic statistical parameters (mean values vector, covariance matrix), which are values vector, covariance matrix), which are computed from the training set. These parameters guide the discrimination process. The Bayesian classifiers are typical representatives (Bayes classifier, Fisher, Wald sequential).

The unsupervised classification is also known as classification without the teacher. This classification uses, in most cases, the methods of cluster analysis. The device that performs the function of classification is called classifier. The classifier is the system containing several inputs that are transported with signals carrying information about the objects. The system generates information about the competence of objects into a particular class on the output.

C. Support Vector Machine

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. The black line that separates the two cloud of class is right down the middle of a channel. The separation is in 2D, a line, in 3D, a plane, in four or more dimensions and a hyperplane. Mathematically, the separation can be found by taking the two critical members, one for each class. These points are called support vectors. These are the critical points (members) that define the channel. The separation is then the perpendicular bisector of the line joining these two support vectors. That's the idea of support vector machine.



$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

Subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right)\geq 1-\xi_i \; and \; \xi_i \geq 0, i=1,\ldots,N$$

where, C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and xi represents the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

D. Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Entropy

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Information Gain

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

V. EXPERIMENTATION RESULT

A. Performance Evaluation

This is a measurement tool to calculate the performance

Accuracy = $\left[\dfrac{TP+TN}{TP+TN+FP+FN}\right]$

Sensitivity = $\left[\dfrac{TP}{TP+FN}\right]$

Specificity = $\left[\dfrac{TN}{TN+FP}\right]$

Positive Predictive Value: $PPV = \left[\dfrac{TP}{TP+FP}\right]$

Negative Predictive Value: $NPV = \left[\dfrac{TN}{TN+FN}\right]$

ROC = $\dfrac{sensitivity + speificity}{2}$

where,

1. The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified
2. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive
3. The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly
4. The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative
5. The accuracy (AC) is the proportion of the total number of predictions that were correct.
6. The Sensitivity or Recall the proportion of actual positive cases which are correctly identified.
7. The Specificity the proportion of actual negative cases which are correctly identified.
8. The Positive Predictive Value or Precision the proportion of positive cases that were correctly identified.
9. The Negative Predictive Value the proportion of negative cases that were correctly identified.

Table 1: Performance analysis for two classifiers using Cross Validation

	Support Vector Machine	Decision Tree
Accuracy	97.9304	97.4130
Sensitivity	96.9135	95.4520
Specificity	99.2001	97.7919
PPV	95.5561	95.8897
NPV	97.7207	97.6064
ROC	98.0568	89.9895

CONCLUSION

This work has evaluated the performance of the four methods with respect to confusion matrix and accuracy. The performance neural network based classification model has been compared with SVM and DT. According to the arrived results, the performance of the supervised machine learning based classification approach provided significant performance. It was found that the SVM classifier was capable of identifying Normal, Suspicious and Pathologic condition, from the nature of CTG data with very good accuracy. This work trains the system with all the classes of samples, there is a chance by which the trained system may be incapable of identifying suspicious record. That is why we are getting comparatively poor average performance while classifying suspicious records. It is a major weakness of the system and it should be overcome in future design. One may address the way to improve the system for getting proper training with different classes of CTG patterns.

References

[1] Steer, P.J. (2008) Has Electronic Fetal Heart Rate Monitoring Made a Difference? Seminars in Fetal and Neonatal Medicine, 13, WB Saunders, 2-7.

[2] Huang, M. and Hsu, Y. (2012) Fetal Distress Prediction Using Discriminant Analysis, Decision Tree, and Artificial Neural Network. Journal of Biomedical Science & Engineering, 5, 526-533.http://dx.doi.org/10.4236/jbise.2012.59065

[3] Sundar, C., Chitradevi, M. and Geetharamani, G. (2012) Classification of Cardiotocogram Data Using Neural Network Based Machine Learning Technique. International Journal of Computer Applications, 47, 19-25. http://dx.doi.org/10.5120/7256-0279

[4] Sundar, C., Chitradevi, M. and Geetharamani, G. (2013) An Overview of Research Challenges for Classification of Cardiotocogram Data. Journal of Computer Science, 9, 198-206. http://dx.doi.org/10.3844/jcssp.2013.198.206

[5] Yılmaz, E. and Kılıkçıer, Ç. (2013) Determination of Fetal State from Cardiotocogram Using LS-SVM with Particle Swarm Optimization and Binary Decision Tree. Computational and Mathematical Methods in Medicine, 2013, 2013. http://dx.doi.org/10.1155/2013/487179

[6] G. Georgoulas, D. Stylios and P. Groumpos. Predicting the Risk of Metabolic Acidosis for New Borns Based on Fetal Heart Rate Signal Classification Using Support Vector Machines. IEEE Trans. Biomed. Eng. 53 (2006) 875–884.

[7] S.L. Salzberg. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Min. Knowl. Discov. (2007) 317–328.

[8] M. Cesarelli, M. Romano, P. Bifulco, Comparison of Short Term Variability Indexes in Cardiotocographic Fetal Monitoring, Comput. Biol. Med. 39 (2009) 106–118.

[9] K. Bache, M. Lichman. Cardiotocography dataset in: UCI Machine Learning Repository, 2010. Web.http://archive.ics.uci.edu/ml/datasets/cardiotocography, 14 Nov. 2016.

[10] N. Krupa, M. Ali, E. Zahedi, S. Ahmed, F.M. Hassan. Antepartum Fetal Heart Rate Feature Extraction and Classification Using Empirical Mode Decomposition and Support Vector Machine, Biomed. Eng. Online 10 (2011) 6.

[11] R. Czabanski, J. Jezewski, A. Matonia, M. Jezewski. Computerized Analysis of Fetal Heart Rate Signals as the Predictor of Neonatal Academia. Expert Syst. Appl.39 (2012) 11846–11860.

[12] C. Sundar. “Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time”. International Journal Computer Science Application (2012).

[13] Tomas Peterek, Peter Gajdos, Pavel Dohnalek and Jana Krohova. Human Fetus Health Classification on Cardiotocographic Data Using Random Forests, Intelligent Data Analysis and its Applications. Volume II, pp:189-198,(2014).

[14] Hakan Sahin and Abdulhamit Subasi. Classification of the Cardiotocogram Data for Anticipation of Fetal Risks Using Machine Learning Techniques. International Burch University. Faculty of Engineering and Information Technologies. Francuske Revolucije b.b., Ilidza, Sarajevo 71000, Bosnia and Herzegovina, Applied SoftComputing 33 (2015) 231–238.

[15] Shahad Nidhal, M. A. Mohd. Ali1 and Hind Najah, 2010,“A novel cardiotocography fetal heart rate baseline estimation algorithm”, Scientific Research and Essays Vol. 5(24), p. 4002-4010

[16] Diogo Ayres-de-Camposa, Cristina Costa-Santosb, Joa˜o Bernardesa, 2005, "Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the SisPorto1 multicentre validation study”, European Journal of Obstetrics & Gynecology and Reproductive Biology 118, 52-60.

BIOGRAPHY

Jagannathan D is currently pursuing M.Phil degree in computer science in Dr. C. V. Raman University, Chhattisgarh. He has done his M.Sc degree in Computer Science from V.O.Chidambaram College,Thoothukudi and the B.Sc in Computer Science from Kamaraj College ,Thoothukudi. under Manonmaniam Sundaranar University,Tirunelveli.