A Comparative Study between Navie Bayes and Decision Tree Algorithms for classification of Cardiotocograms

D. Jagannathan
Department of Computer Science
M.Phil Scholar,Dr. C. V. Raman University,Chhattisgarh – 495113.

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 these methods remains controversial and still inaccurate. In this paper, we implement a model based CTG data classification system using a supervised Decision Tree and Navie Bayes 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 DT 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, Decision Tree and Navie Bayes

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. 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)

III. CLASSIFICATION

Classification process may be applied in different areas of research and practice, e.g., farms, military, medicine, remote Earth 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.

3.1 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

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

3.2 NAVIE BAYES

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem: Probability of **event** A given **evidence** B

Prob(A given B)=Prob(A and B)Prob(A)Prob(A given B)=Prob(A and B)Prob(A)

where:

- A (Class) represents the dependent event: A target attribute and

- B (Instance) represents the prior event: A predictors attribute

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- P(A) is a priori probability of A (The prior probability) Probability of event before evidence is seen. The evidence is an attribute value of an unknown instance.
- P (A|B) is a posteriori probability of B. Probability of event after evidence is seen. Posteriori = afterwards, after the evidence.

IV. EXPERIMENTATION RESULT

4.1Performance Evaluation

This is a measurement tool to calculate the performance

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

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

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

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

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

$$\text{ROC} = \frac{sensitivity + speificity}{2}$$

where,

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

	Decision Tree	Navie Bayes
Accuracy	97.4130	84.8542
Sensitivity	95.4520	70.9042
Specificity	97.7919	85.5353
PPV	95.8897	72.5203
NPV	97.6064	87.8121
ROC	89.9895	78.2197

Table 1:Performance analysis for two classifiers using Cross Validation

V. 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 DT and NB. According to the arrived results, the performance of the supervised machine learning based classification approach provided significant performance. It was found that the DT 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 overcomes in future design. One may address the way to improve the system for getting proper training with different classes of CTG patterns.

REFERENCES

[1] 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.

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

[3] M. Cesarelli, M. Romano, P. Bifulco, Comparison of Short Term Variability Indexes in Cardiotocographic

Fetal Monitoring, Comput. Biol. Med. 39 (2009) 106–118.

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

[5] 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.

[6] 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.

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

[8] E. Yilmaz, C. Kilikcier. Determination of Fetal State from Cardiotocogram Using LS-SVM with Particle Swarm Optimization and Binary Decision Tree, Comput. Math. Methods Med. 2013 (2013) 487179.

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

[10] 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. FrancuskeRevolucije b.b., Ilidza, Sarajevo 71000, Bosnia and Herzegovina, Applied Soft Computing 33 (2015) 231–238.