A three class treatment of the FHR classification problem using latent class analysis labeling

George Georgoulas[1], Jiří Spilka[2], Petros Karvelis[1], Václav Chudáček[2], Chrysostomos Stylios[1] *Member, IEEE*, and Lenka Lhotská[2] *Member, IEEE*

***Abstract*— Electronic Fetal Monitoring in the form of cardiotocography is routinely used for fetal assessment both during pregnancy and delivery. However its interpretation requires a high level of expertise and even then the assessment is somewhat subjective as it has been proven by the high inter and intra-observer variability. Therefore the scientific community seeks for more objective methods for its interpretation. Along this path, presented work proposes a classification approach, which is based on a latent class analysis method that attempts to produce more objective labeling of the training cases, a step which is vital in a classification problem. The method is combined with a simple logistic regression approach under two different schemes: a standard multi-class classification formulation and an ordinal classification one. The results are promising suggesting that more effort should be put in this proposed approach.**

I. Introduction

Electronic Fetal Monitoring is used for fetal surveillance during antepartum and intrapatrum periods. It is predominately performed by means of cardiotocography (CTG) – simultaneous recording of fetal heart rate (FHR) and uterine contractions. The CTG monitoring has been around for over 40 years with the aim to provide better information about fetal behavior and status compared to intermittent auscultation. The rationale of the monitoring is that it gives hints to clinicians for timely and appropriate intervention to prevent adverse long term consequences caused by intrapartum asphyxia. Since its introduction however, the initial enthusiasm seems to have fainted and the CTG is blamed for an increased rate of cesarean sections [1]. Moreover, interpretation of CTG is difficult resulting in high inter and intra-observer variability among clinicians [2]. Nevertheless CTG is the method of choice for intrapartum fetal surveillance [3] with its interpretation relying primarily on visual assessment of CTG trace that follows the clinical guidelines issued by the International Federation of Gynecology and Obstetrics (FIGO) guidelines [4].

This work was partially supported by the research project "Intelligent System for Automatic CardioTocoGraphic Data Analysis and Evaluation using State of the Art Computational Intelligence Techniques" by the program "Greece-Czech Joint Research and Technology projects 2011-2013" of the General Secretariat for Research & Technology, Greek Ministry of Education and Religious Affairs, co-financed by Greece, National Strategic Reference Framework (NSRF) and the European Union, European Regional Development Fund and by Czech Grant Agency project number 14-28462P entitled Statistical methods of intrapartum CTG signal processing in the context of clinical information

[1] Laboratory of Knowledge and Intelligent Computing, Department of Computer Engineering, TEI of Epirus, 47100 Artas, Kostakioi, Greece (emails: georgoul@kic.teiep.gr, pkarvelis@kic.teiep.gr, stylios@teiep.gr)

[2] Dept. of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague (emails: {spilka.jiri, chudacv, lhotska}@fel.cvut.cz)

In order to tackle the persistent inter and intra-observer variability, research efforts were devoted to incorporate the domain knowledge of clinicians into an automatic decision systems. In order to design such a system three integral parts are necessary: i) features used for FHR characterization ii) a method for collection and aggregation of expert annotations, and iii) a learning method (classifier) to map the features to expert annotations. The features were initially based upon clinical guidelines such as basal heart rate, variability, and decelerations [5]. In addition many other features, inspired by adult heart rate variability, were employed, e.g. time-domain, frequency-domain, and nonlinear features [6], [7], [8], [9]. Among the methods most commonly used for aggregation of expert annotation were: simple majority voting [5], [10], consensus achieved in a panel discussion [11], [12], combination of majority voting and one clinician as an oracle [5], or annotation by one clinician [13]. The works [11], [12] used the UCI Cardiotocography data set [14]. For the classification many machine learning paradigms were studied and employed such as Support Vector Machines (SVMs) [6], [11] and artificial neural networks (ANNs) [7] to name just a few.

In this work we extend on the idea of machine learning method driven by expert opinions. The main obstacle, often neglected in many works, represents aggregation and interpretation of the high clinician's variability. To that point, the inter-individual variability is diminished here by using latent class analysis (LCA) [15] for labeling FHR records and by an ordinal classifier casting the assessment problem into a classification one. The LCA offers a natural way to combine different, possibly noisy, annotations from multiple experts. The classification is performed by the so called three step approach: i) the LCA is used to estimate posterior probabilities of individual examples, ii) labels are determined by maximum posterior probability, and iii) a classifier is trained using these labels. The results are promising showing that this is probably a better way to assess FHR recordings instead of relying on pH values.

The paper is structured as follows: Section II presents in brief all the employed technologies: FHR preprocessing, feature extraction, LCA and ordinal classification. Section III summarizes the results achieved while section IV presents the conclusions and provides directions for future work.

II. Methods

A. Preprocessing

Fetal heart rate, recorded either by ultrasound Doppler probe or by a scalp electrode, contains a lot of artifacts.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

Therefore it is necessary to preprocess the FHR signal before applying any feature extraction method. The values outside interval 50-220 beats per minute (bpm) were considered as artifacts and treated as missing data. Then, missing data were interpolated using a Matlab implementation of Hermite spline interpolation. After that a number of features was extracted in order to condense the most relevant information.

B. Feature Extraction

Since the initial attempts to quantify the morphological quantities described in the FIGO's guidelines, a number of features from various domains have been proposed trying to design a comprehensive set reflecting condition of the fetus. In this work we use the FIGO based morphological features (baseline, decelerations, and accelerations), statistical time domain features (short and long term variability), frequency domain features (energy in different spectral bands), and non-linear features (entropy and complexity). Due to space limitations we refer the interested reader to a thorough description of used features presented in our previous works [6], [8] or to works of others [5], [7], [9], [16].

C. Latent Class Analysis

In this study we use clinical annotations from nine expert clinicians – obstetricians. All clinicians were working in delivery practice with experience ranging from 10 to 33 years. Clinicians categorized/assigned the CTG recordings into three classes: normal, suspicious, and pathological (FIGO classes [4]). Since there is a large inter-observer variability in evaluation the simple majority voting among clinicians might lead to wrong aggregation of annotation, especially when a large number of clinicians provides annotations [17]. Therefore a more powerful approach – the latent class analysis [15] was employed. The LCA is used to estimate the true (unknown) evaluation of CTG and to infer weights of individual clinicians' evaluation. The LCA and its advantages over majority voting were described in [17]. For other examples on LCA in machine learning see, e.g. [18]. The clinical evaluation for the i-th example, obtained from the j-th clinician were considered as coming from a mixture of multinomial distributions with an unknown multinomial parameter, α, and unknown mixing proportions p_c. For the model the likelihood function of $\boldsymbol{\theta}$ given evaluation y_i^j can be formulated as:

$$p(y_i^1, \ldots, y_i^j | \boldsymbol{\theta}) = \prod_{i=1}^{M} \left[\sum_{c=1}^{C} p_c \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j, k)} \right], \quad (1)$$

where C is number of classes, J is number of clinicians, $\delta(y_i^j, k)$ is an indicator function that equals 1 when the j-th clinician evaluates $y_i^j = c$ and 0 otherwise and α_{ck}^j is a multinomial parameter that represents probabilities that the c-th class corresponds to an evaluation in the k-th class, $k \in C$, assigned by the j-th clinician. The Expectation Maximization (EM) algorithm [19] was used to estimate the unknown parameters. The EM algorithm was restarted several times with different initialization to verify the convergence to the same solution. The limit of log-likelihood convergence was set to 10e-3. The resulting class for individual examples was determined by the largest posterior probability.

D. Classification

FHR categorization is usually treated as a standard multiclass classification problem. However in this problem as in many real life problems there is a natural ordering of the classes. To be more specific the three categories, normal, suspicious, and pathological are ranked according to their severity and by slight abuse of notation we can write Pathological>Suspicious>Normal (in terms of severity). As a result an ordinal classification scheme was tested along with a standard multiclass classification procedure.

The ordinal classification approach adopts the proposal of [20] which can be used with standard classification learners producing a probabilistic output by transforming the original C-class ordinal problem into $C-1$ binary class problems. In our case the original three class problem with the ordered values Pathological>Suspicious>Normal is transformed into two two-class problems, creating two new datasets: the first one having a (binary) class attribute representing Target>Normal and the other one having a class attribute representing Target>Suspicious. The process is depicted in Figure 1.

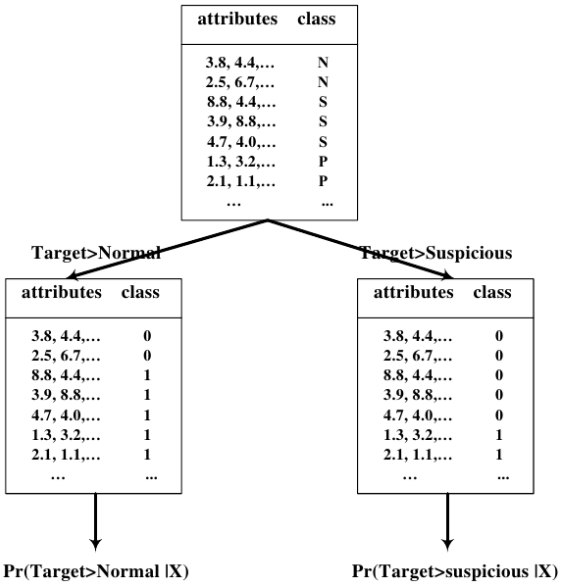

Fig. 1. The process of breaking down the original data set into two datasets with binary labels (N:Normal, S:Suspicious, and P:Pathological).

The prediction of the probabilities of a new instance/case belonging to one of the C original classes relies on the probabilities produced by each one of the $C-1$ models. In our case:

$$\Pr(N) = 1 - \Pr(Target > N), \quad (2)$$

$$\Pr(S) = \Pr(Target > N) - \Pr(Target > S), \quad (3)$$

$$\Pr(P) = \Pr(Target > S), \quad (4)$$

where N stands for Normal, S for Suspicious, and P for Pathological. For the estimation of the probabilities a logistic regression model was used [19], which is given by the following set of equations for a problem with C classes:

$$\log \frac{\Pr(c=1|X)}{\Pr(c=k|X)} = b_{10} + b_1^T X \quad (5)$$

$$\log \frac{\Pr(c=2|X)}{\Pr(c=k|X)} = b_{20} + b_2^T X \quad (6)$$

$$\dots \quad (7)$$

$$\log \frac{\Pr(c=k|X)}{\Pr(c=k|X)} = b_{k0} + b_k^T X, \quad (8)$$

which can easily be shown that corresponds to the following probability estimates:

$$\Pr(c=n|X) = \frac{\exp(b_{n0} + b_n^T X)}{1 + \sum_{m=1}^{C-1} \exp(b_{m0} + b_m^T X)}, \quad (9)$$

for $n = 1, \dots, k-1$

$$\Pr(c=k|X) = \frac{1}{1 + \sum_{m=1}^{C-1} \exp(b_{m0} + b_m^T X)}, \quad (10)$$

where $n = 1, \dots, C$ is the class label, (b_{n0}, b_n^T) are linear coefficients for each class, X is the attribute input vector, and $\Pr(c = n|X)$ is the conditional probability of class n given the attribute input vector.

III. Experimental Results

Data set. The proposed approach was tested using the recently released CTU-UHB database [21]. The database consists of 552 records selected from more than 9164 intrapartum CTG records that were acquired between years 2009 and 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. The CTG signals were carefully chosen with clinical as well as technical criteria in mind. The CTGs were recorded using STAN and Avalon devices. The acquisition was done either by scalp electrode (FECG 102 records), ultrasound probe (412 records), or combination of both (35 records). For three records the information was not available. All recordings were sampled at 4Hz. The majority of babies were delivered vaginally (506) and the rest using cesarean section (46). A more detailed description of the CTU-UHB is provided in [21]. In this work the features were systematically extracted on 60 minutes long FHR signals at the end of first stage of labor.

A. Results

First, the labels were estimated using the LCA. The EM algorithm was iterated until converge. Figure 2 shows fast convergence of the logarithm of the likelihood function defined in (1).

The model is stable after the 10th iteration, however to reach the predefined convergence of 10e-3 more iterations are needed. Convergence of the model from the point of view of clinicians is presented in Figure 3. The S_{acc}^j represents

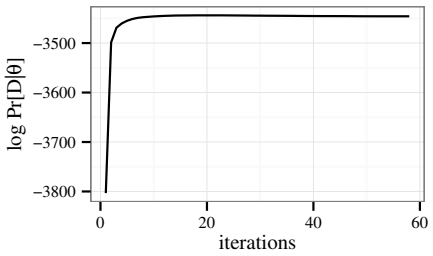

Fig. 2. Convergence of the log likelihood for the latent class analysis.

accuracy of the j-th clinician to the current estimate of latent class. In other words it represents probability of "correct" evaluation given the current estimate of latent classes. Clinicians are numbered from 1 to 9; progression of S_{acc}^j is presented for the same number of iterations as for the log likelihood. The majority voting was used for model initialization and hence the S_{acc}^j corresponding to majority voting is presented in the 0-th iteration. It can be seen that the LCA model re-weights the contribution of each clinician. When the model converge clinicians 4 and 5 are considered as the best while 3 and 6 as the worst.

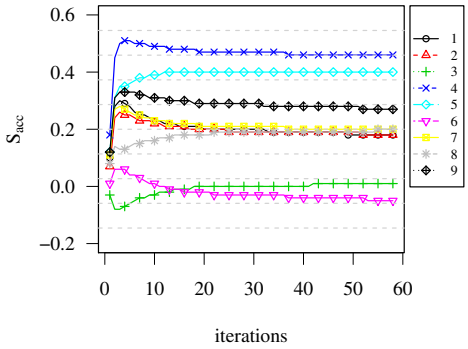

Fig. 3. Probability of correct evaluation of individual clinicians with increasing number of iterations.

In order to test the proposed approach 10 fold stratified cross validation was applied [22]. Prior to the application of the classifier the features were ranked using the chi-squared metric after discretizing the continue valued attributes using the Minimum Description Length (MDL) criterion [22] and the top 10 features were retained and fed to the classifier. The results for the case of the ordinal classification scheme and for the simple multiclass logistic regression are summarized in the following Tables I and II, respectively. All the models were developed in WEKA [23].

TABLE I
Confusion Matrix for the Ordinal Classification Scheme

		predicted class		
		normal	suspicious	pathological
true class	normal	118	55	2
	suspicious	46	170	25
	pathological	10	44	82

TABLE II

CONFUSION MATRIX FOR MULTICLASS LOGISTIC REGRESSION

		predicted class		
		normal	suspicious	pathological
true class	normal	118	54	3
	suspicious	47	166	28
	pathological	10	43	83

IV. DISCUSSION AND CONCLUSION

In this work we presented an integrated approach to FHR classification, which is a very crucial, but very difficult, task during delivery because of the "fuzzy" boundaries between the different classes as it can be seen in the Figure 4 where two of the top 10 discriminative features are depicted (along with the boundaries built using this 2-feature model).

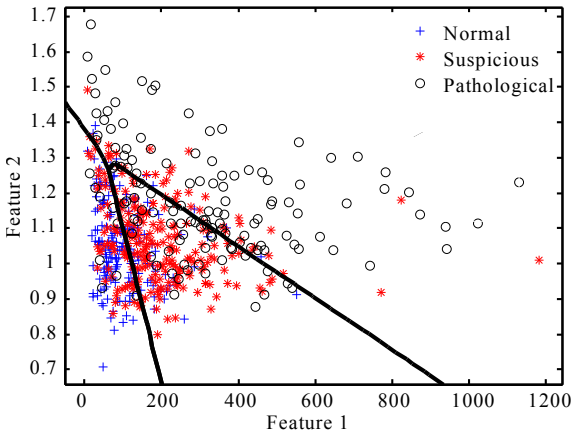

Fig. 4. Scatter plot of the dataset using two top ranked features.

Comparison to other works is rather difficult since different datasets were used in various work. The results achieved in [11], [12] on the available UCI Cardiotocography data set [14] are hardly comparable since the response of each clinician is not available. Also the CTG records are unavailable and thus new features can not be computed. Therefore, we plan to make the clinical annotations available in the future to accompany the open-access CTU-UHB database [21].

The results indicate that the ordinal classification scheme is slightly better but further experimentation is needed before final conclusion can be reached. Furthermore as it was reported in [20], the advantage of ordinal classification becomes more apparent once the number of classes increases. In future work we plan to further explore that since the LCA allows investigating different number of latent classes.

REFERENCES

[1] Z. Alfirevic, D. Devane, and G. M. L. Gyte, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour.," *Cochrane Database Syst Rev*, vol. 3, no. 3, pp. CD006066, 2006.

[2] S. C. Blackwell, W. A. Grobman, L. Antoniewicz, M. Hutchinson, and C. Gyamfi Bannerman, "Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System.," *Am J Obstet Gynecol*, vol. 205, no. 4, pp. 378.e1–378.e5, Oct 2011.

[3] H. Y. Chen, S. P. Chauhan, C. V. Ananth, A. M. Vintzileos, and A. Z. Abuhamad, "Electronic fetal heart rate monitoring and its relationship to neonatal and infant mortality in the United States.," *Am J Obstet Gynecol*, vol. 204, no. 6, pp. 491.e1–491.10, Jun 2011.

[4] FIGO, "Guidelines for the use of fetal monitoring," *International Journal of Gynecology & Obstetrics*, vol. 25, pp. 159–167, 1986.

[5] B. Guijarro-Berdiñas, A. Alonso-Betanzos, and O. Fontenla-Romero, "Intelligent analysis and pattern recognition in cardiotocographic signals using a tightly coupled hybrid system," *Artif. Intell.*, vol. 136, pp. 1–27, 2002.

[6] G. Georgoulas, C. D. Stylios, and P. P. Groumpos, "Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines," *IEEE Trans Biomed Eng*, vol. 53, no. 5, pp. 875–884, May 2006.

[7] R. Czabanski, M. Jezewski, J. Wrobel, J. Jezewski, and K. Horoba, "Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and epsilon-insensitive learning.," *IEEE Trans Inf Technol Biomed*, vol. 14, no. 4, pp. 1062–1074, Jul 2010.

[8] J. Spilka, V Chudáček, M. Koucký, L. Lhotská, M. Huptych, P. Janků, G. Georgoulas, and C. Stylios, "Using nonlinear features for fetal heart rate classification," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 350–357, 2012.

[9] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings," *IEEE Trans Biomed Eng*, vol. 50, no. 3, pp. 365–374, Mar 2003.

[10] R. D. Keith, S. Beckley, J. M. Garibaldi, J. A. Westgate, E. C. Ifeachor, and K. R. Greene, "A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram.," *Br J Obstet Gynaecol*, vol. 102, no. 9, pp. 688–700, Sep 1995.

[11] H. Ocak, "A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being.," *J Med Syst*, vol. 37, no. 2, pp. 9913, Apr 2013.

[12] T. Peterek, J. Krohová, P. Dohnálek, and P. Gajdoš, "Classification of cardiotocography records by random forest," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*. IEEE, 2013, pp. 620–923.

[13] S. Dash, J.G. Quirk, and P.M. Djuric, "Learning dependencies among fetal heart rate features using bayesian networks," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 28 2012-sept. 1 2012, pp. 6204 –6207.

[14] D. Ayres de Campos, J. Bernardes, A. Garrido, J. Marques de Sa, and L. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," *J Matern Fetal Med*, vol. 9, no. 5, pp. 311–318, 2000.

[15] J. A. Hagenaars and A. L. McCutcheon, *Applied latent class analysis*, Cambridge University Press, 2002.

[16] G. Magenes, M. G. Signorini, and D. Arduini, "Classification of cardiotocographic records by neural networks," in *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000*, 2000, vol. 3, pp. 637–641.

[17] J. Spilka, V. Chudáček, P. Janků, L. Hruban, M. Burša, M. Huptych, Zach L., and L. Lhotská, "Analysis of obstetricians decision making on CTG recordings," *Journal of Biomedical Informatics*, 2014, doi: 10.1016/j.jbi.2014.04.010.

[18] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, pp. 20–28, 1979.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[20] E. Frank and M. Hall, *A simple approach to ordinal classification*, Springer, 2001.

[21] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, and L. Lhotská, "Open access intrapartum CTG database.," *BMC Pregnancy Childbirth*, vol. 14, no. 1, pp. 16, 2014.

[22] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.