An ordinal classification approach for CTG categorization

George Georgoulas, Peter Karvelis, Dimitris Gavrilis, Chrysostomos D. Stylios, *Senior Member, IEEE*, George Nikolakopoulos

Abstract— **The analysis of cardiotocogram (CTG) is the standard approach employed during pregnancy and delivery. Its interpretation requires however high level of expertise in order to decide whether the recording is Normal, Suspicious or Pathological, which is not always available. Therefore, a number of attempts have been carried out over the past three decades for development of automated systems that can mimic the human expert. These systems are usually (multiclass) classification systems that assign a category to the respective CTG. However most of these systems usually do not take into consideration the natural ordering of the categories associated with CTG recordings. In this work we examine the utility of using an algorithm which explicitly takes into consideration the ordering of CTG categories. The results suggest that the investigated ordinal classification approach is marginally better than the traditional multiclass classification approach.**

I. Introduction

Since its introduction, electronic fetal monitoring has become the prevalent measure of fetal surveillance. Electronic Fetal monitoring is usually synonymous to the evaluation of the cardiotocogram (CTG) (the fetal heart rate (FHR) and the uterine contractions (UC) printed on a single stripe of paper) by an expert clinician using specific guidelines [1]. However, such an expertise is not always available, making CTG evaluation a difficult task.

With the advances both in computer technology and machine learning, a number of systems [2],[3] were proposed with the ultimate goal to replicate if not surpass human performance. All these systems follow the same pattern recognition principal: a set of features/indicators is extracted from the CTG (primarily from the FHR) and then is fed to a classifier to perform the categorization / decision making. To this end numerous features [4]-[6] and machine learning algorithms [7]-[11] have been proposed and tested over the past three decades.

Nevertheless, these approaches do not explicitly take into consideration the natural ordering of the main three categories used to characterize a CTG recording (Normal, Suspicious and Pathological). In our previous work this ordering was investigated for the first time through the use of an ordinal classification scheme and the results showed that the ordinal classifier achieves better performance compared to its nominal counterpart [12]. In this work we further exploit that using a different and bigger data set [2],[13],[14]. The achieved results show that the ordinal classification paradigm should be given more attention.

The rest of the paper is structured as follows. Section II summarizes all the involved methods. In Section III the experiments conducted and the achieved results are presented while section IV concludes the paper offering also some hints for future research directions.

II. Methods

The most common machine learning problems are classification and regression ones. In the former the machine learning algorithm tries to predict a categorical variable while in the latter the targeted variable takes real values. While in the former the assumption is that the categorical variables have not ordering and that taking the difference between them has no actual meaning, in the latter the predicted outputs, being real valued, can be compared and conclusions can be drawn by assessing for example their absolute different. In between these two paradigms, lies the ordinal classification/regression, where the output is still fixed to a number of predefined classes, but the ordering of those classes has actual meaning. For example, in the medical field the health status is routinely assessed using terms such as normal, suspicious, pathological or severely deteriorating. In other words, class labels contain order information, e.g. a sample vector associated with class label “average” has a higher rating (or better) than another from the “poor” class, but “good” class is better than both.

Using mathematical notation, the ordinal classification problem consists of predicting the label y of an input vector x, where $x \in X \subseteq \Re^K$ and $y \in Y = \{C_1, C_2, \ldots, C_Q\}$ with $C_1 \prec C_2 \prec \ldots \prec C_l$, where $\prec$ defines the ordering relation between the labels. This ordering relation differentiates.

There a number of ordinal classification approaches [15] which are the ordinal counterparts of conventional nominal classifiers. In this work since the aim is to validate the utility of the ordinal approach to the specific CTG classification problem and based on previous results that showed that simpler approaches can be competitive to more elaborated ones [16], the simple ordinal classification scheme described in [17], was selected. One of the advantages of this approach is that conventional classifiers can be used with no need of modification.

George Georgoulas and George Nikolakopoulos are with the Robotic Team, at the Control Engineering Group, Department of Computer Science, Electrical and Space Engineering Luleâ University of Technology, Luleâ, 97187, Sweden (email: geogeo@ltu.se; geonik@ltu.se).

Petros Karvelis and Chrysostomos D. Stylios are with the Laboratory of Knowledge and Intelligent Computing, Department of Computer Engineering, Technological Educational Institute of Epirus, Arta, 47100 Greece (0030-26810- 50330; e-mail: pkarvelis@gmail.com; stylios@teiep.gr).

Dimitris Gavrilis is with the Department of Electrical Engineering and Computer Technology, University of Patras, Kato Kastritsi, Patras, Greece, P.O. Box 26500 (email: gavrilis@upatras.gr).

Preprint submitted to 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received February 21, 2017.

A. Binary Decomposition

The method in [17], decomposes the original Q class ordinal problem into $Q-1$ binary class problems and uses the estimated probabilistic values of the binary classifiers to perform an overall prediction. The method works as follows: a model is built to predict what is the probability of a given instance to belong at any of the classes that are located higher than C_1, higher than C_2 etc. up to the probability of the instance belonging to the "highest" class C_Q. Then taking into consideration the monotonic ordering of classes the probability of each category can be easily estimated:

$$\Pr\left(C_1 \mid x\right) = 1 - \Pr\left(class \succ C_1 \mid x\right) \tag{1}$$

$$\Pr\left(C_n \mid x\right) = \Pr\left(class \succ C_{n-1} \mid x\right) - \Pr\left(class \succ C_n \mid x\right), \quad \text{for } 1 < n < Q \tag{2}$$

$$\Pr\left(C_Q \mid x\right) = \Pr\left(class \succ C_{Q-1} \mid x\right) \tag{3}$$

Obviously the class with the highest probability defines the outcome of the algorithm.

$$C_i = argmax_i\left(Pr(C_i)\right) \tag{4}$$

In this work the C4.5 is involved as the base binary classifier as in the original publication [17], considering also the results of [16].

B. The C4.5 algorithm

The C4.5 algorithm builds a Decision Tree using a greedy algorithm. During the training phase each node of the tree is assigned a number of samples which are weighted to take into account unknown attribute values. Given that the set of samples associated with the node t is denoted as V, the weighted frequency $freq(i,V)$ of cases in V whose class is $i, i \in [1,..,Q]$ with Q being the number of classes, is computed. Since usually V contains cases belonging to two or more classes the information gain must be computed for each attribute

$$gain = \text{info}(V) - \sum_i \frac{|V_i|}{|V|} \times \text{info}(\text{V}_i) \tag{5}$$

where $V_i, i = 1,\ldots,s$ is the set of the splitting produced by the test on the selected attribute and s is the number of splitting of node t. After that the entropy $\text{info}(V)$ of the set V is given by:

$$\text{info}(V) = \sum_{j=1}^{Q} \frac{freq(j,\text{V})}{|V|} \times \log_2\left(\frac{f(j,V)}{|V|}\right) \tag{6}$$

If V_i is not empty, the divide and conquer approach consists of recursively applying the same operations on the set consisting of V_i plus those cases in V with unknown value of the selected attribute.

C4.5 can also provide the probability of an input vector belonging to a specific class based on a count of the training samples belonging to each leaf.

C. Synthetic minority oversampling TEchnique (SMOTE)

As it is shown in the following section, the dataset is heavily imbalanced towards the normal class. This is a common situation in the biomedical field, where non-patients outnumber patients. This from an algorithmic point of view makes training difficult and without proper consideration models that can only predict the majority category can be created with very high, though useless, overall accuracy. One way to deal with this kind of data imbalance is to alter the number of training samples either by downsampling the majority class or by oversampling the minority class(es). In this work the Synthetic Minority Oversampling TEchnique (SMOTE) is adopted [18]. This technique instead of simply oversampling the minority class, it creates synthetic examples "filling" the space between the original instances that belong to the minority class.

III. EXPERIMENTS – RESULTS

A. Data set

For this study the Cardiotocogram data set [13] from UCI repository is used [14] which contains 1655 Normal, 295 Suspicious and 176 Pathological cases with the labels assigned by three experts following a consensus procedure. The data set was automatically created using the SisPorto 2.0 software [2]. 21 features were extracted in total, which are summarized in Table I.

TABLE I. EXTRACTED FEATURES

Feature Abbreviation	Feature description
LB	baseline value
AC	accelerations
FM	foetal movement
UC	uterine contractions
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
DL	light decelerations
DS	severe decelerations
DP	prolongued decelerations
Width	histogram width
Min	low freq. of the histogram
Max	high freq. of the histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency

B. Experimental procedure

For evaluating the conventional and the ordinal classification schemes, 2 fold stratified cross validation (CV) was repeated 20 times (20x2 CV). In other words, each time the original data was divided randomly into a training and a

Preprint submitted to 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received February 21, 2017.

testing set, retaining the ratios of the different classes. Then SMOTE was applied to the Suspicious and the Pathological population of (only) the training set in such a way as the resulting training samples for each class to be approximately equal. After that the conventional as well as the ordinal classifiers were trained and tested using the test set and their classification performance was assessed using the measures described in the following section. After that the training (without the synthetically created examples) and the testing sets would be switched and the aforementioned procedure was performed once more. This procedure was repeated 20 times and the results are summarized in the form of confusion matrices in the Appendix

C. *Results*

For conventional classification problems, overall accuracy (the percentage of correctly classified instances) is the most common measure of performance. However, for imbalance data sets, other measures which are not affected by the class imbalance are preferred like the geometric mean of the individual recall values:

$$gmean = (\prod recall_i)^{1/3} \tag{6}$$

However, both of these measure do not take into account the ordering of the classes and the corresponding implication of confusing non-adjacent categories (misclassifying a normal instance as pathological is more severe an error than misclassifying a normal instance as suspicious). Therefore other measures could be considered [15], [19] when it comes to the assessment of different ordinal classifiers such as the Mean Absolute Error (MAE) and the Average MAE (AMAE).

MAE as its name implies, is the average deviation of the prediction from the true class

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|s(\hat{y}_i) - s(y_i)\right| \tag{7}$$

where s is a score function which in its simplest form- also considered here- is given by $s(y_i) = i$, $1 \le i \le Q$

As in the case of overall performance, in the case of imbalanced classes MAE is also not the most appropriate measure. Therefore for that case, the Average MAE (AMAE) was proposed [20] which takes into account the differences in size of the involved classes:

$$AMAE = \frac{1}{Q}\sum_{j=1}^{Q}\frac{1}{N_j}\sum_{i=1}^{N_j}\left|s(\hat{y}_i) - s(y_i)\right| \tag{8}$$

where N_j is the number of cases belonging to class j. This way the “contribution” of each class is conversely proportional to its cardinality.

Given these measures, the performance of the two algorithms is summarized in Table II and graphically depicted in Fig. 1-4. As it can be seen the ordinal approach is better even though the difference is not statistically significant.

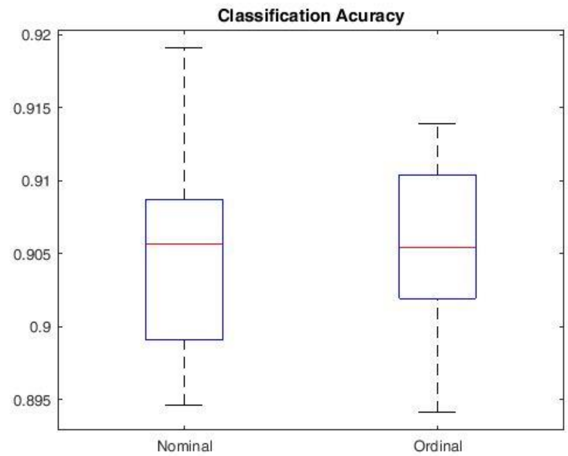

Figure 1. Box plots depicting the overal accuracy

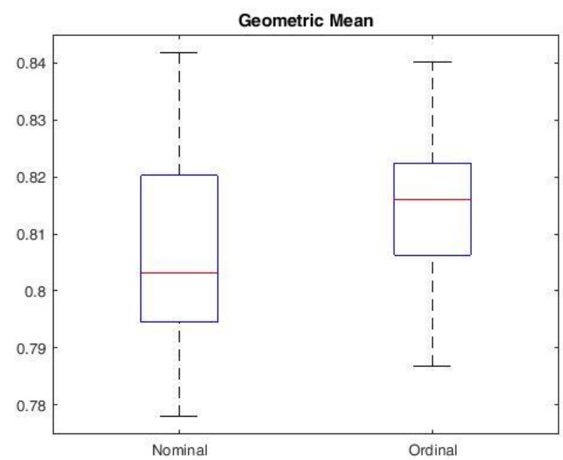

Figure 2. Box plots depicting the geometric mean

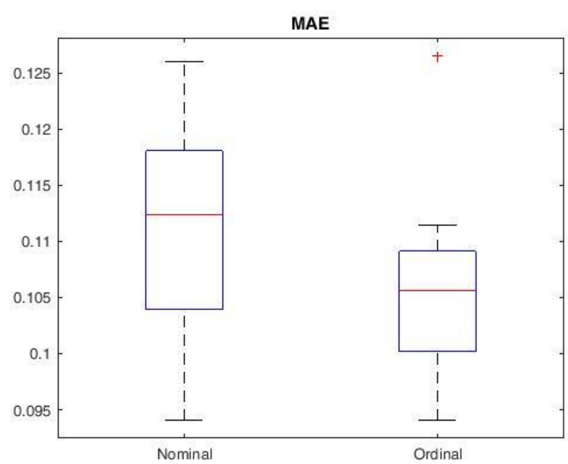

Figure 3. Box plots depicting the MAE

Preprint submitted to 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received February 21, 2017.

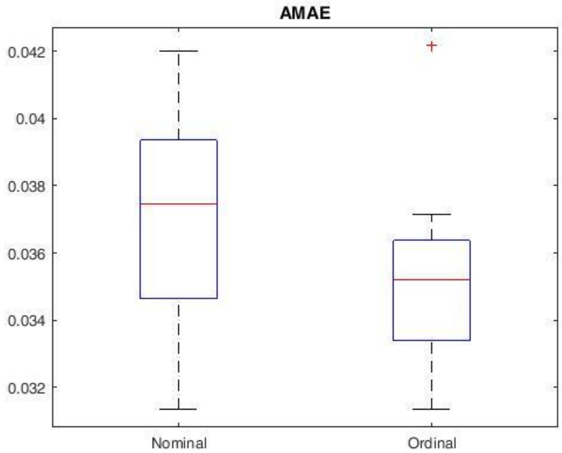

Figure 4. Box plots depicting the AMAE

TABLE II. AVERAGE CLASSIFICATION MEASURES

	Accuracy	GMean	MAE	AMAE
Nominal	0.9041	0.8066	0.1105	0.0368
Ordinal	0.09055	0.8151	0.1053	0.0351

IV. CONCLUSIONS

In this work an ordinal classification approach was investigated for the assessment of CTG recordings. The method was compared to its nominal counterpart and the results are promising indicating that this specific approaches is suitable for this specific application since it is able to exploit the intrinsic ordering of the assigned classes. In future work more advanced ordinal schemes will be tested to fully exploit the applicability of the proposed method.

APPENDIX

TABLE III. NOMINAL CLASSIFICATION – AGGREGATE CONFUSION MATRIX

Confusion Matrix		**Predicted Class**		
		Normal	***Suspicious***	***Pathological***
True Class	***Normal***	30879	1078	190
	Suspicious	1787	4404	169
	Pathological	434	418	3161

TABLE IV. CLASSIFICATION – AGGREGATE CONFUSION MATRIX

Confusion Matrix		**Predicted Class**		
		Normal	***Suspicious***	***Pathological***
True Class	***Normal***	30790	920	199
	Suspicious	2048	4648	256
	Pathological	262	332	3065

REFERENCES

[1] FIGO, "Guidelines for the Use of Fetal Monitoring," *Int. J. Gynaecol. Obstet.,* vol. 25, pp. 159–167, 1986.

[2] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. P. M. de Sa, and L. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," J. *Matern.. Fetal Med*., vol. 9, pp. 311–318, 2000

[3] R. Czabanski, J. Jezewski, J. Wrobel, K. Horoba, Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and-insensitive learning, IEEE Transactions on Information Technology in Biomedicine, 14 (4) (2010) 1062-1074.

[4] Task-Force, Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, Eur. Heart J., 17 (3) (1996) 354–381.

[5] L. Xu, C.W. Redman, S.J. Payne, A. Georgieva, Feature selection using genetic algorithms for fetal heart rate analysis, Physiological Measurement, 35 (7) (2014) 1357-1371.

[6] J. Spilka, V. Chudáček, M. Koucký, L. Lhotská, M. Huptych, P. Janků, G. Georgoulas, C. Stylios, Using nonlinear features for fetal heart rate classification. Biomedical Signal Processing and Control, 7 (4) (2012) 350–357.

[7] G. Georgoulas, C. Stylios, 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., 53 (5) (2006) 875-884.

[8] A. Georgieva, S.J. Payne, M. Moulden, C.W. Redman, Artificial neural networks applied to fetal monitoring in labour, Neural Computing and Applications, 22 (1) (2013) 85-93

[9] H. Sahin, A. Subasi, Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques, Applied Soft Computing, 33 (2015) 231-238.

[10] S. Dash, J.G. Quirk, P.M. Djuric, Fetal heart rate classification using generative Models, IEEE Trans. Biomed. Eng., 61(11) (2014) 2796-2805.

[11] R. Czabanski, J. Jezewski, A. Matonia, M. Jezewski, Computerized analysis of Fetal Heart Rate signals as the predictor of neonatal acidemia, Expert Systems with Applications, 39 (15) (2012) 11846–11860.

[12] P. Karvelis, J. Spilka, G. Georgoulas, V. Chudáček, C.D. Stylios, and L. Lhotská, "Combining latent class analysis labeling with multiclass approach for fetal heart rate categorization," Physiological measurement, vol. 36, no. 5, pp. 1001-1024, 2015.

[13] https://archive.ics.uci.edu/ml/datasets/Cardiotocography#

[14] Frank A, Asuncion A (2010) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine. http://archive.ics.uci.edu/ml

[15] A.P. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martınez, "Ordinal regression methods: survey and experimental study," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 127-146, 2016.

[16] Georgoulas, G., Kolios, S., Karvelis, P. and Stylios, C., 2016, September. Examining nominal and ordinal classifiers for forecasting wind speed. In Intelligent Systems (IS), 2016 IEEE 8th International Conference on (pp. 504-509). IEEE.

[17] E. Frank, and M. Hall, "A simple approach to ordinal classification", Lecture Notes in Computer Science, vol. 2167, pp. 145-156, 2001.

[18] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, pp. 321-357, 2002.

[19] S. Janitza, G. Tutz, and A.L. Boulesteix, "Random forest for ordinal responses: Prediction and variable selection," Computational Statistics & Data Analysis, vol. 96, pp.57-73, 2016.

[20] J. Sánchez-Monedero, P. Gutiérrez, P. Tiňo, and C. Hervás-Martínez, "Exploitation of pairwise class distances for ordinal classification," Neural computation, vol. 25, no. 9, pp. 2450-2485, 2013.

Preprint submitted to 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received February 21, 2017.