35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Feature Selection for Computerized Fetal Heart Rate Analysis using Genetic Algorithms*

Liang Xu, Antoniya Georgieva, Christopher W. G. Redman, Stephen J. Payne

Abstract— **During birth, timely and accurate diagnosis is needed in order to prevent severe conditions such as birth asphyxia. The fetal heart rate (FHR) is often monitored during labor to assess the condition of fetal health. Computerized FHR analysis is needed to help clinicians identify abnormal patterns and to intervene when necessary. The objective of this study is to apply Genetic Algorithms (GA) as a feature selection method to select a best feature subset from 64 FHR features and to integrate these best features to recognize unfavorable FHR patterns. The GA was trained on 408 cases and tested on 102 cases (both balanced datasets) using a linear SVM as classifier. 100 best feature subsets were selected according to different splits of data; a committee was formed using these best classifiers to test their classification performance. Fair classification performance was shown on the testing set (Cohen's kappa 0.47, proportion of agreement 73.58%). To our knowledge, this is the first time that a feature selection method has been tested for FHR analysis on a database of this size.**

I. Introduction

During labour, a baby's oxygen supply can be reduced due to the stress caused by uterine contractions. Unable to cope with such situation, some babies suffer from birth asphyxia (suffocation), which may lead to seizures, permanent brain damage or even death in severe conditions. Cerebral palsy occurs in approximately 2 cases per 1,000 births, of which birth asphyxia accounts for 10-30% [1]. In clinical practice, to prevent birth asphyxia, it is crucial to carry out timely intervention to assist delivery immediately. On the other hand, interventions such as Caesarean sections, forceps and ventouse deliveries may cause complications, thus these interventions are best avoided when possible. Therefore, timely and accurate diagnosis of birth asphyxia is essential to minimize damage while avoiding unnecessary interventions.

In order to monitor the condition of fetal health, the fetal heart rate (FHR) and uterus contracts are electronically recorded during labor on a paper strip called a cardiotocogram (CTG). The complicated CTG patterns are usually assessed by eye, which is tedious, error-prone and associated with low reproducibility due to high inter- and intra-observer variability [2, 3]. It has long been recognized that computerized analysis of the CTG in fetal monitoring has potential for improving decision making for interventions. Therefore, computerized analysis for CTG patterns has become a significant and pressing need. Currently, our group is developing a computerized FHR analysis system (OxSys) to automatically recognize unfavorable intrapartum FHR patterns.

At present, the OxSys system contains 64 features [4-7]. Previous studies have shown that the combination of different FHR features can be better than using the features independently [6, 7], revealing the potential of selection and integration of these features to provide a better predictor than using individual features. On the other hand, the number of all possible combination of the 64 features used here is $2^{64}=1.8\times10^{18}$, which is too big for an exhaustive search. Thus feature selection methods need be applied to the task of selecting a subset of FHR features.

Amongst the different approaches, Genetic Algorithms (GA) are known for their competitive exploration ability, i.e. the ability to explore the feature space as widely as possible. They can be powerful and efficient global optimisers in various fields of data analysis [8].

The aim of this study is to apply Genetic Algorithms (GA) as a feature selection method to select a best feature subset and to integrate these best features to recognize unfavorable FHR patterns. In this way, the system could be able to automatically predict adverse labor outcome, in order to help the clinicians make decisions on interventions during labor.

II. Data

A. Data selection criteria

The process of labor is divided into three stages according to different physiological activities. The first stage is identified as frequent regular contractions with less than 10 cm cervix dilation. The second stage is identified as descent of the baby's head through the mother's pelvis, with cervical dilation of 10 cm. The third stage is identified as the delivery of the placenta. To ensure that all cases are equally selected at comparable stages of labor, only the last 30 minutes of second labor stage (before birth) were examined, since the second stage has more drastic changes in FHR due to uterus contractions. The assumption of this study is that, in the last 30 minutes of second labor stage, adverse labor outcome related to fetal heart rate are detectable using CTG. Therefore, in this study, included were only CTG records taken directly after the onset of pushing with fair signal quality. From 107,614 deliveries in John Radcliffe hospital between 20 Apr 93 - 28 Feb 08 (currently world's largest FHR database), 7,568 recordings were selected using these criterions. The details of selection criterions can be found in a previous study [6].

B. GA training set and testing set

Adverse outcome in this study was defined as acidosis at birth (clinically defined as umbilical arterial pH < 7.05) [9].

*Resrach supported by Clarendon Fund.

L. X. and S. J. P. are with Institute of Biomedical Engineering, Department of Engineering Science, Old Road Campus Research Building, University of Oxford, Headington, Oxford, OX3 7DQ, UK (phone: (+44) 757-003-8893 fax: (+44) 1865 617702 email: liang.xu@wadh.ox.ac.uk).

A. G. and C. W. G. R. are with Nuffield Department of Obstetrics & Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK (email: antoniya.georgieva@obs-gyn.ox.ac.uk)

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

Acidosis at birth is one of the clinical diagnosis conditions of birth asphyxia [6]. The dataset is heavily imbalanced: only 255 out of 7,568 cases have an adverse outcome (3.37%). Training a classifier using the entire dataset will result in a high classification performance with low prediction power. Therefore, a balanced dataset was created to train the classifier.

From the total set of 7,568 cases, 255 cases were adverse outcome cases. Normal outcome was defined as 7.27 < arterial pH < 7.33 and no form of compromise (959 cases). To create a balanced dataset of 50% normal outcomes and 50% adverse outcomes, 255 cases were randomly selected out of the 959 normal outcomes. Therefore, the dataset used in this study consists of 510 cases, with 255 normal outcomes and 255 adverse outcomes.

In these 510 cases, 80% (204 normal cases and 204 adverse cases) were selected randomly for the feature selection process using GA. The remaining 20% (51 normal cases and 51 adverse cases) were used as a testing set to evaluate the performance of the features selected by the GA. In addition, in each GA run, a training set itself was separated into a training set and a validation set (70%-30%). To avoid confusion, the data used in GA were referred to as the GA training set, and the separated testing set used to evaluate the performance of GA was referred to as the testing set (Fig. 1). The widely accepted 'rule of thumb' [10] was followed that at least 10 training samples per input feature of each class are needed. Therefore, the maximal number of features that could be selected is 14.

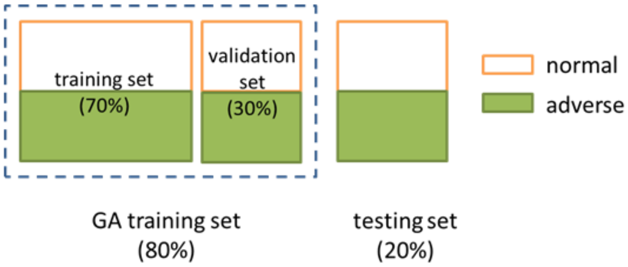

Figure 1. Conformation of all 510 cases used for GA.

III. Methods

A. Basic framework of GA

The framework of the method is adapted from the traditional GA [8, 11]. The GA method consists of five major steps:

(1) Generation of the population

(2) Selection based on fitness value

(3) Reproduction (crossover and mutation)

(4) Accepting of the new generation

(5) Checking the stopping criterion

The population size was set to 100. Ranking strategy was used as the selection strategy, where the individuals ranking in the best 50% fitness were selected to create offspring. single-point crossover with a proportion of 0.8 and single-point mutation with a proportion of 0.2 were applied to every generation to generate the next population. The elitist selection was set to 2, i.e. the best two individuals of the current generation were included in the next population. The maximum number of generations with the same best fitness value was set to 20. The maximum number of generations was set to 200. The number GA run with different initial conditions (same data splits) was set to 100. Each of these parameters was optimized based on preliminary tests to ensure that the output of GA was consistent at the minimum cost of computation time.

B. Fitness function

In the GA, each genome (individual) was given a fitness value by the fitness function. In this study Cohen's kappa value [12] was used as the fitness evaluation of the classifier. Kappa is a statistical measure of agreement between predicted and actual results. It is a more robust measure than simple proportion of agreement, since kappa takes into account that agreement occurs by chance.

To evaluate the agreement between predicted and actual results, the data in the GA training set were randomly split into a training set and a validation set as mentioned before. A classifier was then trained from the training data using the feature subset, and the predictions were compared to actual results on validation set.

The performance of the classifier will vary depending on how the training set is drawn from the GA training set. In order to improve the classifier's ability of generalization, cross-validation is necessary. Due to the limited size of the data, repeated random sub-sampling strategy is used. Before each run of the GA, the data are randomly split into ten 70% training-30% validation sets. For each genome, the classifier was trained respectively by each of the ten training sets. The performance on each set was recorded as the kappa value comparing predictions and actual results of its testing set. The median of these ten kappa values was then recorded as the fitness value of the genome. By doing this, it is ensured that the convergence of GA is less associated with the splitting of data.

C. Classifier

The classifier used in this study is a linear SVM. The support vector machine (SVM) is widely used in data analysis due to its intuitive definition and simple implementation [13]. The SVM constructs a hyperplane or a set of hyperplanes to separate data points, in order to achieve the largest distance to the nearest training points of any class. Intuitively, a larger functional margin means lower generalization error of the classifier. The detailed principle of the linear SVM can be found in [14]. The method used to find the separating hyperplane is the Least Squares method. Linear SVM algorithms were taken from LIBSVM [15].

IV. Results

A. Classification performance

The GA was run 100 times, and the best classifier for each run of the GA was applied independently on the testing set. The agreement of the output for each classifier was measured by a kappa value comparing the prediction of these classifiers and the actual result. The classification performances on both the GA training set and the testing set are shown in Fig. 2.

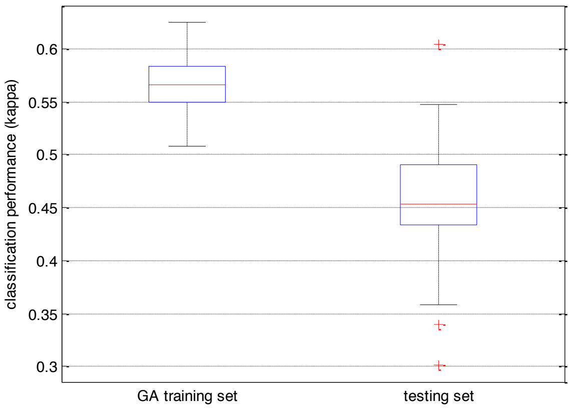

Figure 2. The classification performance (kappa) for different classifiers over 100 runs on GA training set and testing set.

The 100 best classifiers described above were then used to form a voting committee. For each case in the testing set, the mode of the committee voting was taken as the prediction. The classification performance is shown in Table I, compared with three other frequently used feature selection methods: Random Forest using classification method (RF-C), Random Forest using regression method (RF-R) and Least Absolute Shrinkage and Selection Operator (LASSO) [16, 17]. These methods are known for their classification accuracy, but they either can't give a specific best set of feature (Random Forest), or is hard for interpretation (LASSO). It is shown here that the classification performances of the GA are comparable, or slightly better than, the performances of these classical feature selection methods. In addition, the selected feature subset performs better than using all 64 features.

TABLE I. CLASSIFICATION PERFORMANCE (TESTING SET) AND COMPARISON WITH OTHER FEATURE SELECTION METHODS

Method	Sensitivity	Specificity	Kappa	Proportion of agreement
All 64 features	63.20%	66.83%	0.30	65.09%
GA	66.83%	81.13%	0.47	73.58%
RF-C	67.92%	77.36%	0.45	72.64%
RF-R	64.15%	73.58%	0.38	68.87%
LASSO	66.83%	78.25%	0.45	72.64%

B. *Feature frequency*

GA was run 100 times with different splits of training-validation data, thus there are 100 sets of best features. The importance of each feature can be assessed by the frequency of the feature being selected in the 100 best feature subsets. Table II shows the feature importance ranking of the ten most frequently selected features. It shows that some features could be important regardless of different splits of training-validation sets. For example, Feature No.61 (Phase Rectify Signal Averaging – DC component) was selected for almost all best feature subsets. The feature frequency, indicating the importance of different features, could be very useful for future reference. It will also be very helpful for other feature selection methods such as forward selection and backward elimination.

TABLE II. MOST FREQUENTLY SELECTED FEATURES IN 100 RUNS OF GA USING LINEAR SVM

Ranking	Feature index	Feature name	Feature frequency (%)
1	61	Phase Rectify Signal Averaging –DC component	98
2	10	Median of short term variability change tracker	60
3	48	Mutual information	51
4	16	Median of contraction duration	46
5	2	Zero difference between neighbor points (%)	41
6	51	Standard deviation of Sample Entropy	37
7	9	Signal Stability Index of the residual FHR signal	18
8	50	Mean of local Approximate Entropy	18
9	63	Bivariate Phase Rectify Signal Averaging –DC component	17
10	58	Interquartile range of the smoothed signal	15

V. DISCUSSION AND CONCLUSION

Fetal Heart Rate (FHR) is used during labor to assess the condition of fetal health and assist diagnosis of birth asphyxia. The objective of this study was to find a best feature subset using feature selection methods. Clear and intuitive clinical interpretation is needed to assist the clinicians in decision making, thus a GA was chosen for its ability to explore the whole feature space and give a clear best feature subset. In addition, a linear SVM was chosen as the classifiers for GA to investigate the linear relationship between features and adverse outcome. To our knowledge, this is the first time a feature selection method was used in this large scale of FHR data (510 balanced cases).

100 different best feature subsets were selected according to different splits of data, these different best feature subsets were used as a classifier committee. Given the lack of a gold standard and our limited ability to predict labor outcome with any tool or expert knowledge [18], the classification performance of the committee on the testing set is promising (Cohen's kappa, 0.47, proportion of agreement 73.58%). This classification performance is better than previous studies on the similar dataset using an Artificial Neural Network on similar dataset, with kappa 0.28 on the testing set [6]. The classification performance of GA is comparable, or slightly better than to other feature selection methods using the same dataset.

Feature No. 61 (Phase Rectified Signal Averaging– DC component), No. 10 (short term variability) and Feature No. 48 (Mutual information are most frequently selected. Therefore, these features appear to be useful in predicting labour outcome when used in multivariate analysis. This information will be very valuable for reference in future studies. Further investigation will be focused on the interpretation of these features in classifying labor outcomes.

One limitation of this study lies in that it took only the last 4 windows of 15 minutes length in the last 30 minutes of second stage. More information during the process of labor, especially time-series information, should be required and integrated into the classifier. There are also a number of additional clinical parameters that need to be studied, such as maternal infection, oxytocin augmentation, etc. [6]. The detailed clinical interpretation of the classifiers should be investigated with time series information, too.

In addition, how to apply the classifier trained using a balanced dataset to the whole 7,568 cases is still a question. Analytical tools such as Event Rate Estimation (EveREst) plot could be used to examine the prediction power of the classifiers in terms of clinical situations [9].

The next step of the study is to apply the classifiers throughout the duration of labor, which will provide an objective measurement of fetal health condition during different stages of labor. Further studies will be carried on to estimate the risk of compromise, based on the classifier prediction and its patient specific time-series trend.

In conclusion, Genetic Algorithms, as a feature selection method, was used for the first time to integrate and optimize the predictive power of various FHR features. The GA was trained on 408 cases and tested on 102 cases (both balanced datasets) using linear SVM as classifier. Fair classification performance was shown on the testing set (Cohen's kappa 0.47). Based on these results, it can be concluded that GA can be successfully applied to FHR features to select best feature subsets and to optimize their predictive power. More analysis and clinical interpretation of the classifiers are necessary for further work.

Acknowledgment

The authors would like to acknowledge the use of the Oxford Supercomputing Centre (OSC) in carrying out this work. A Georgieva is funded by the Henry Smith Charity and Action Medical Research.

References

[1] F. S. Alberry M, Soothill PW, "Prediction of asphyxia with fetal gas analysis," presented at the Levene MI, Chervenak FA (eds) Fetal and Neonatal Neurology and Neurosurgery, 4th edn., Churchill Livingstone, 2009.

[2] S. P. Chauhan, et al., "Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability," American Journal of Obstetrics and Gynecology, vol. 199, pp. 623.e1-623.e5, 2008.

[3] J. Westgate, "Computerizing the Cardiotocogram (CTG)." presented at the Medical Informatics in Obstetrics and Gynecology, Parry, D., & Parry, E. (Eds.), 2009.

[4] B. Fulcher, et al., "Highly comparative fetal heart rate analysis," in Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012, pp. 3135-3138.

[5] L. Xu, et al., "Detection and Analysis of Pattern Readjustment in Fetal Heart Rate Signal," presented at the MEDSIP 12, Liverpool, UK, 2012.

[6] A. Georgieva, et al., "Artificial neural networks applied to fetal monitoring in labour," Neural Computing & Applications, pp. 1-9, 2011.

[7] V. a. Chud´aˇcek, "Assessment of features for automatic CTG analysis based on expert annotation," presented at the 33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts USA,, 2011.

[8] M. Mitchell, An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press, 2001.

[9] A. Georgieva, et al., "Umbilical cord gases in relation to the neonatal condition: the EveREst plot," European Journal of Obstetrics & Gynecology and Reproductive Biology, vol. In press, 2013.

[10] T. G. Van Niel, et al., "On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification," Remote Sensing of Environment, vol. 98, pp. 468-480, 2005.

[11] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning: Addison-Wesley Professional, 1989.

[12] J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psych. Meas., vol. 20, p. 37, 1960.

[13] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[14] N. Cristianini, and Shawe-Taylor, J, An Introduction to Support Vector Machines and other kernel-based learning methods.: Cambridge University Press, 2000.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, pp. 1-27, 2011.

[16] L. Breiman, Classification and regression trees: Wadsworth International Group, 1984.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267-288, 1996.

[18] D. A. Grimes and J. F. Peipert, "Electronic fetal monitoring as a public health screening program: the arithmetic of failure," Obstetrics & Gynecology, vol. 116, p. 1397, 2010.