34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Learning dependencies among fetal heart rate features using Bayesian networks

Shishir Dash,[1] J. Gerald Quirk,[2] and Petar M. Djurić[1]

Abstract— **We present preliminary results on the use of Bayesian-network (BN) structure learning algorithms for deciphering dependencies from amongst different fetal heart rate (FHR) features collected from a real database. We used a greedy search-and-score procedure known as the K2 algorithm for the estimation of the BN structure. The database consists of a collection of discrete-valued features quantifying presence of morphological changes as prescribed by expert guidelines (updated by the National Institute of Child Health and Human Development (NICHD)) and implemented as rule-based programs. We compare the results of structure learning to the expert-guided structure and use decision functions for classification using posterior probabilities. It was found that the BN estimated from structure learning algorithms had comparable classification performance, but fewer edges, leading to more efficient characterization of conditional probability tables (CPD's). Moreover, structure learning algorithms offer a method of learning novel correlations between FHR features that may be exploited for automatic categorization.**

I. Introduction

Computer-aided categorization of fetal heart rate (FHR) records has yet to be translated effectively from the realm of academic research to clinical applications. In the labor room, visual assessment of FHR tracings is very important since fetal oxygen inadequacy has a direct effect on the FHR. Purely visual assessment of FHR segments suffers from high intra- and inter-observer variability [1], leading to higher rates of caesarian sections and increasing litigation costs. Several attempts have been made to standardize the human interpretation of FHR and uterine-pressure (UP) signals, such as those in [2]. Computer-aided diagnosis is an attempt to solve such problems using techniques of machine learning.

In a previously published paper [3], we demonstrated the use of simple statistical metrics to evaluate relevant feature values from FHR/UP data. This was followed by coarse graining of the continuous-valued features into qualitative categories such as, say, "Absent variability" or "Presence of recurrent decelerations," and categorization of the feature set using NICHD rules [2]. This is a rule-based categorization approach similar in principle to the Oxford Sonicaid system [4] and the SisPorto system [5]. As such, it precludes the calculation of confidence measures for making decisions.

Other sophisticated methods for FHR classification include the use of neural networks [6], particle swarm optimization [7], and support vector machines [8]. In previous studies, we developed a Bayesian network (BN) formulation to integrate the features from our expert system (ES) into a probabilistic framework. A BN [9] is a specific type of graphical model in which known (or hypothesized) causal relationships between nodes can be represented as conditional probability relationships. Edges between the nodes can be endowed with directions representing the flow of information from one node to the other. One well known use of BNs for medical diagnosis is the Quick Medical Reference (QMR) system [10], which has more than 4000 observable nodes and 600 unobservable nodes representing the presence or absence of specific diseases and their symptoms.

Our original BN structure was learnt entirely from expert guidance and can be interpreted easily. However, expert guided structure may not be the best choice in terms of classification performance or elicitation of the most relevant causal dependencies. There exists a very rich literature on the problem of efficient Bayesian structure learning methods, such as those developed in [11] (K2 algorithm) and in [12] (Markov chain Monte Carlo methods). For the current problem, we have used the K2 method to get an accurate representation of the probabilistic dependencies between FHR features described earlier, based on real-data evidence. Using this learnt structure, one can learn conditional probability table (CPT) parameters using traditional Bayesian or maximum likelihood techniques. Then, one can also get the posterior probabilities of the "class" variable conditioned on the instantiations of the attribute variables. This can be used as classifier decision function. Though there exist many other sophisticated structure learning techniques for such applications, our purpose here is not a comparison of these techniques, but a demonstration that using such algorithms can tell us more about real FHR data than just expert guidance.

In the sequel, in Section II-A we present the K2 structure learning algorithm, in Section II-B, the ES features, and in Section II-C, the classification procedure. Results of classification performance using Leave-One-Out (LOO) procedure are provided in Section IV. We conclude the paper with a discussion of the results, and possible future work in Section V.

II. Bayesian network formulation

The BN consists of two sets: a set of nodes U and a set of edges E. For the current application, we denote the N features extracted from the data set as random variables X_i, with $i \in \{1, 2, \ldots, N-1\}$. The set $U = \{X_1, \ldots, X_N\}$,

[1]Shishir Dash (sdash at ic.sunysb.edu) and Petar M. Djurić (djuric at ece.sunysb.edu) are with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY - 11790, USA

[2] J. Gerald Quirk is with the Department of Obstetrics and Gynecology, Stony Brook University Medical Center, Stony Brook, NY - 11790, USA.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

contains the random variables representing the features and an additional variable representing the "true" fetal state (equivalent to the "class" variable) as labeled by a physician or some other objective diagnostic procedure. We refer to these variables as *nodes* of the graph. For full description of the graph, we also need the set of directed *edges* E, which represent conditional dependencies between the nodes. Thus, the graph is formally denoted by $G = (U, E)$. For every pair of variables connected as $Y \rightarrow X$, Y is called the *parent* of X, and X is the *child* of Y. Each feature X_i has a range of instantiations $\{x_{i,1}, \ldots, x_{i,k}, \ldots, x_{i,r_i}\}$, where r_i is the number of possible instantiations for that variable. For graph structure G, the set of parents of node X_i is denoted $\mathbf{Y}_i^G$ and this collection of variables can take values from the set $\{\mathbf{y}_{i,1}^G, \ldots, \mathbf{y}_{i,j}^G, \ldots, \mathbf{y}_{i,q_i^G}^G\}$, where q_i^G is the product of the cardinalities of all the variables in $\mathbf{Y}_i^G$.

The advantage of using BNs stems from the fact that given a specific BN structure, the joint probability distribution over all the nodes of the graph factorizes as

$$P(X_1, \ldots, X_N|G) = \prod_{i=1}^{N} P(X_i|\mathbf{Y}_i^G). \quad (1)$$

This enables very efficient characterization of the probability distribution of the features since the number of parameters can be greatly reduced when the conditional independencies encoded in the BN structure are taken into account. In addition, structure elicitation presents a precise method of detection of (possibly) causal dependencies between the various variables represented on the graph.

A. BN structure learning

It is well known that learning the structure of any general BN is NP-hard since the number of possible structures increases super exponentially with the number of nodes in the network. Thus, a variety of search heuristics have been developed to address this problem. In the current work, we focus on the well-known K2 algorithm from [11].

The K2 method is a conceptually simple greedy hill climbing method for searching the space of directed acyclic graphs (DAGs). The method is constrained by a user-input ordering of the nodes, and it maximizes a chosen scoring metric γ (described below) that captures how well the DAGs represent the observed datasets. The input node-ordering Φ reflects the user knowledge of the total node ordering and not individual subgraph structure.

Initially the algorithm assumes that none of the nodes have any parents and calculates an "empty" score $\gamma(G)$ for the graph G having no edges. Thereafter, for every node $X_i \in \Phi$, the algorithm searches for the single best parent Y from the set Φ_{-i} (consisting of all nodes preceding X_i in the total ordering) that, when connected to the node X_i, provides the greatest increase in γ. If it finds no such parent, it stops and goes to the next node in Φ. Otherwise, it (a) updates the graph G to have the edge $Y \rightarrow X_i$ and (b) restarts the search for other possible X_i parents. The procedure is repeated until all nodes have been explored. Since the total ordering is provided, the algorithm does not need to check for graph acyclicity at each step. We note that the cost of using this efficient heuristic is that the learnt BN structure is significantly influenced by the choice of topological ordering, and thus susceptible to bias. However, since the purpose is not to replace expert guidance but merely to refine it, we feel the benefits of the algorithm outweigh the costs.

There are several choices for the score function as reviewed in [12]. Since we used Dirichlet priors for the discrete features, we employed the Bayesian scoring criterion $\gamma(G) = P(\mathbb{D}|G)$ [13]:

$$P(\mathbb{D}|G) = \prod_{i=1}^{N} \prod_{j=1}^{q_i^G} \frac{\Gamma(N_{ij}^G)}{\Gamma(N_{ij}^G + M_{ij}^G)} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}^G + c_{ijk}^G)}{\Gamma(\alpha_{ijk}^G)}, \quad (2)$$

where $\mathbb{D}$ is the set of observed data values, α_{ijk}^G is the Dirichlet parameter associated with the event $\{X_i = x_{i,k}|\mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\}$, $N_{ij}^G = \sum_{k=1}^{r_i} \alpha_{ijk}^G$, c_{ijk}^G is the number of data cases in which node X_i takes the value $x_{i,k}$ and X_i's parents take the value $\mathbf{y}_{i,j}^G$, and $M_{ij}^G = \sum_{k=1}^{r_i} c_{ijk}^G$. Since this scoring criterion decomposes into local frequency computations for each node, it is computationally quite efficient.

B. FHR Features

In order to stay close to physician guidelines, we restricted our feature set to those features recommended by the standard guidelines [2]. We have described the software implementation of these feature extraction techniques in a previous paper [3], and provide only brief descriptions here. In the following list, we represent the FHR features as random variables, and provide the range of possible instantiations for each of them:

1) Baseline FHR $B \in \{$Bradycardia, Normal, Tachycardia$\}$,
2) Baseline Variability $V \in \{$Absent, Minimal, Moderate, Marked$\}$,
3) Presence of Accelerations $A \in \{$No, Yes$\}$,
4) Presence of Decelerations $D \in \{$No, Yes$\}$,
5) Presence of Recurrent Decelerations $D_r \in \{$No, Yes$\}$,
6) Presence of Early Decelerations $D_e \in \{$No, Yes$\}$,
7) Presence of Late Decelerations $D_l \in \{$No, Yes$\}$,
8) Presence of Variable Decelerations $D_v \in \{$No, Yes$\}$,
9) Presence of Prolonged Decelerations $D_p \in \{$No, Yes$\}$,
10) Presence of Recurrent Early Decelerations $D_{re} \in \{$No, Yes$\}$,
11) Presence of Recurrent Late Decelerations $D_{rl} \in \{$No, Yes$\}$, and
12) Presence of Recurrent Variable Decelerations $D_{rv} \in \{$No, Yes$\}$.

The fetal state random variable S can take values from the set $\{1, 2, 3\}$, which correspond to the subjective assessments $\{Normal, Indeterminate, Abnormal\}$.

The *expert-guided* BN structure is provided in Figure 1. The fetal status S is assumed to have a direct causal effect on *all* the "symptom" variables $\{B, V, A, \ldots, D_{rl}, D_{rv}\}$.

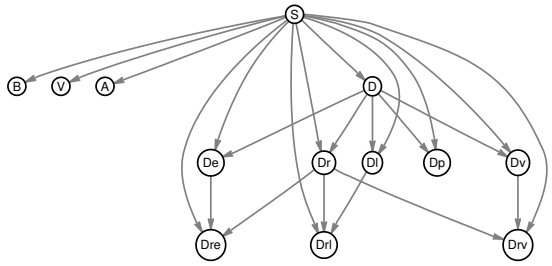

Fig. 1. Expert-guided BN structure for categorization of FHR features. The fetal status S has a direct causal effect on all the other variables.

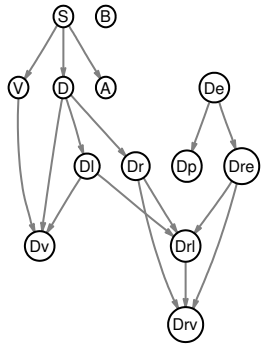

Fig. 2. BN structure learnt from the K2 algorithm. The fetal status S has a direct causal effect on only $\{V, D, A\}$.

The features B, V, A and D are pairwise conditionally independent given knowledge of S. In addition, we have 8 more variables representing deceleration types that are directly dependent on the presence of decelerations. For instance, if D takes the value "No", then all the other deceleration variables have to take the value "No"; however, if $D =$ "Yes", then it is not necessary that, say, recurrent decelerations are also present. The directed edge $D \rightarrow D_r$ encodes this intuitive notion.

C. FHR classification using BN

The CPTs for each parent-child pair are learnt from the training set of feature instantiations using maximum likelihood frequency updates. For any test data set d_i whose status variable S is unknown, we can derive the marginal posterior distribution for the S_i conditioned on knowledge of the features using simple sum-product techniques [14]. A MAP criterion is used to find the S_i instantiation $s_i \in \{1, 2, 3\}$ having the highest marginal probability mass. This is taken to be the classifier output for d_i.

III. Data

The program was tested on a database of 830 20-minute FHR records collected from 9 subjects during the antea-partum period at the Stony Brook University Medical Center. All consent and approval guidelines were followed rigorously. The FHRs were continuously monitored using the Doppler technique via GE Corometrics devices. The usual method for extraction of FHR from the Doppler signal is to use autocorrelation functions to detect the periodic movements of the heart valves. Although this does impose some restrictions on the detection of short-term variability, for our purposes it was deemed to have sufficient resolution for effective tracing characterization.

Prior to carrying out feature extraction, FHR preprocessing was done to remove various artifacts including ones due to movement as described in [3]. Each record was independently labeled as category 1, 2 or 3 by a physician who had access only to the raw and preprocessed noise-free versions of the FHR. It was observed that for some files heavily corrupted by tracing noise and for those dominated entirely by episodic variations, our ES would not be able to get values of certain features like variability or episode locations. For this study, we ignored such data sets from the training and testing procedures. We were then left with 754 out of 830 20-min traces from the original database.

TABLE I
Confusion matrix for BN classification of 754 real data sets.

		Expert BN			K2 BN		
		1	2	3	1	2	3
Physician Labeling ↓	1	182	118	0	179	121	0
	2	91	356	1	91	356	1
	3	5	1	0	5	1	0

IV. Results

Classification performance was analysed using Leave-One-Out (LOO) procedures, i.e., for each data set d_i in the record database $\mathbb{D}$, we learn the CPTs from the database $\mathbb{D}_{-i}$ consisting of all data sets except d_i. However, structure learning was done using the entire data set $\mathbb{D}$ in order to ensure that we obtained exactly one learnt structure $\widehat{G}$ (as opposed to $|\mathbb{D}_{-i}|$ different structures) to compare against the expert-guided network G.

We first present the result of structure learning using the K2 algorithm on the 754-strong database in Fig. 2. The differences between $\widehat{G}$ and G (Fig. 1) are discussed further in Section V. The total number of edges in $\widehat{G}$ is 16, as opposed to 23 in G. It was seen that for this particular database, the S node has a causal effect only on $\{V, D, A\}$ and the node B representing the average baseline value for the record is not connected to any other node in the network. As a result of the reduction in the number of edges, the total number of independent conditional probability distribution (CPD) parameters decreased from 89 for G to 60 (or 59 if we ignore the B node from the structure entirely) in $\widehat{G}$. In Table I, we present confusion matrices for classifier performance when using posterior probabilities calculated by the expert and K2 BNs. Both networks yield similar performances ($\approx$80% sensitivity and $\approx$ 60% specificity, with classifier outputs of categories 2 or 3 treated as "positive" detections).

V. Discussion

Obstetric care providers use the Doppler ultrasound monitors to get continuous recordings of FHR. Standardized clinical guidelines are used to interpret patterns of specific morphological features in the FHR signal such as decelerations (abrupt or gradual decreases in heart rate) or loss of variability (variation around a "baseline" FHR signal). In this study we developed a method to (a) incorporate these features in a BN formulation, (b) to learn network structure from a given set of observed data, and (c) to measure classification performance using posterior probabilities. Although BN structure learning has been widely used in diverse fields such as fault diagnosis, image processing, and medical diagnosis, to our knowledge, this is its first application specific to FHR. The K2 structure learning technique reduces the redundancy in the graph and the total number of CPD parameters, while maintaining the same level of classification accuracy. This is an advantage in terms of efficiency, and it suggests that parameter learning from new data sets using the learnt structure may be more robust.

Prior to structure learning, we had prediscretized the FHR features. Although in principle, continuous features can provide better feature resolution, we worked with discrete features for several reasons including the facts that (a) our feature discretization [3] is very similar to clinical feature definitions as described in [2] and routinely used in obstetric care centers, (b) using continuous features in the BN requires the introduction of (possibly non-Gaussian) parametric continuous distributions, necessitating the learning of many more hyperparameters, and (c) structure learning with continuous features is considerably more difficult, especially when the data lack diversity.

With the proposed approach, we are able to learn new correlations (or the lack thereof) present in the data. In Fig. 2, one can see that the variable B has no parents or children. Indeed, it was seen that for this database, the vast majority of FHR recordings (741 out of 754) had Normal baseline FHR (between 110 and 160 bpm), only 12 had tachycardic baseline (greater than 160 bpm), while only one data set had bradycardic baseline. This implies that for nearly all possible instantiations of its possible child-nodes, the baseline B node takes the same value; thus, the CPD remains indifferent to the value of B. Similar arguments in the case of the nodes D_e and D_p justify their disconnection from the "class" variable S. Another prominent difference is the inclusion of new edges $D_{rl} \rightarrow D_{rv}$ and $D_l \rightarrow D_v$, which suggests that there are strong correlations between the existence of late and variable decelerations. Moreover, in $\widehat{G}$ the "class variable" S is only connected to the variables $\{V, D, A\}$. This set is also S's Markov blanket, suggesting that for classification via evidential reasoning, it may only be required to look at $\{V, D, A\}$ instead of the entire gamut of possible morphological "symptoms". However, this needs to be tested with more datasets.

Other concerns with this approach are being addressed in our ongoing research. For a small 30-strong test subset of FHR data recordings, it was found that physician labeling changed in as many as 16 cases by simply shuffling the order in which the experts were shown the recordings. One possible way to deal with this intra-observer variability would be to introduce more features representing independent tests of fetal health (e.g., umbilical pH values). Another problem is the extreme rarity of FHR tracings showing true fetal distress (category 3), which in the current database compose only 0.8% of the total observed database. This lack of diversity leads to poor performance when classifying category 3 recordings. Clearly, testing on more and more data is of paramount importance. A significant aim of future study is to increase the specificity of classification, since a 40% false-positive rate is prohibitively high for clinical use. Other goals include the integration of different statistical measures for assessing variability, information from the uterine pressure signal, efficient incorporation of continuous features into the BN, and conducting unsupervised clustering of FHR recordings.

References

[1] E. Blix, O. Sviggum, K. Koss, and P. Øian, "Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 110, no. 1, pp. 1–5, 2003.

[2] G. Macones, G. Hankins, C. Spong, J. Hauth, and T. Moore, "The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: Update on definitions, interpretation, and research guidelines," *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, vol. 37, no. 5, pp. 510–515, 2008.

[3] S. Dash, J. Muscat, J. G. Quirk, and P. M. Djurić, "Implementation of NICHD diagnostic criteria for feature extraction and classification of fetal heart rate signals," in *Proc. of the 45th IEEE Asilomar Conference on Signals, Systems, and Computers, November 2011*, 2011.

[4] J. Pardey, M. Moulden, and C. W. Redman, "A computer system for the numerical analysis of nonstress tests," *American Journal of Obstetrics and Gynecology*, vol. 186, no. 5, pp. 1095–1103, 2002.

[5] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite, "SisPorto 2.0: A program for automated analysis of cardiotocograms," *The Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[6] Y. Noguchi, F. Matsumoto, K. Maeda, and T. Nagasawa, "Neural network analysis and evaluation of the fetal heart rate," *Algorithms*, vol. 2, no. 1, pp. 19–30, 2009.

[7] G. Georgoulas, C. Stylios, V. Chudacek, M. Macas, J. Bernardes, and L. Lhotska, "Classification of fetal heart rate signals based on features selected using the binary particle swarm algorithm," in *World Congress on Medical Physics and Biomedical Engineering 2006*, 2007, pp. 1156–1159.

[8] P. Warrick, E. Hamilton, D. Precup, and R. Kearney, "Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 771 –779, 2010.

[9] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*. Springer, 2007.

[10] J. Lemaire, J. Schaefer, L. Martin, P. Faris, M. Ainslie, and R. Hull, "Effectiveness of the Quick Medical Reference as a diagnostic tool," *Canadian Medical Association Journal*, vol. 161, no. 6, pp. 725–728, 1999.

[11] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[12] N. Friedman and D. Koller, "Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks," *Machine Learning*, vol. 50, no. 1, pp. 95–125, 2003-01-01.

[13] R. E. Neapolitan, *Learning Bayesian Networks*. Prentice-Hall, 2004.

[14] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2, pp. 131–163, 1997.