



Exploring the risk factors of preterm birth using data mining

Hsiang-Yang Chen^{a,1}, Chao-Hua Chuang^{b,*,1}, Yao-Jung Yang^a, Tung-Pi Wu^c

^a Department of Applied Information, Hsing Kuo University of Management, Tainan, Taiwan

^b Department of Nursing, Chang Jung Christian University, Tainan County, Taiwan

^c Department of Obstetrics and Gynecology, SinLau Hospital, Tainan, Taiwan

ARTICLE INFO

Keywords:

Preterm birth
Data mining
Neural network
Decision tree

ABSTRACT

Preterm birth is the leading cause of perinatal morbidity and mortality, but a precise mechanism is still unknown. Hence, the goal of this study is to explore the risk factors of preterm using data mining with neural network and decision tree C5.0. The original medical data were collected from a prospective pregnancy cohort by a professional research group in National Taiwan University. Using the nest case-control study design, a total of 910 mother–child dyads were recruited from 14,551 in the original data. Thousands of variables are examined in this data including basic characteristics, medical history, environment, and occupation factors of parents, and variables related to infants. The results indicate that multiple birth, hemorrhage during pregnancy, age, disease, previous preterm history, body weight before pregnancy and height of pregnant women, and paternal life style risk factors related to drinking and smoking are the important risk factors of preterm birth. Hence, the findings of our study will be useful for parents, medical staff, and public health workers in attempting to detect high risk pregnant women and provide intervention early to reduce and prevent preterm birth.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Preterm birth, the birth of an infant prior to 37 completed weeks of gestation, is the leading cause of perinatal morbidity and mortality (Goldenberg, Culhane, Dlams, & Romero, 2008; McCormick, 1985). The prevalence rate for such birth is about 12–13% in the USA, and 5–9% in Europe, other developed countries and Taiwan (Chuang, Chang, Hsieh, et al., 2007; MacDorman, Martin, Mathews, Hoyert, & Ventura, 2005; Slattery & Morrison, 2002). The reasons for preterm birth remain unclear, although data mining is a promising approach to explore potential factors from large amount of data (Chang, 2007; Chen, Hou, Chuang, & TBPS Research Group, 2009; Courtney, Stewart, Popescu, & Goodwin, 2008; Liao, Hsieh, & Huang, 2008). Hence, the purpose of this work, based on the nest case-control study design, is to explore the risk factors of preterm by neural network and decision tree in data mining, to find more potential information.

2. Literature review

2.1. Preterm birth

Preterm birth is the birth of an infant within 37 weeks of gestation, which accounts for 75% of perinatal mortality and half the long-term morbidity (McCormick, 1985). Studies show that maternal race, age, weight, income, previous preterm history, weight gain, infection, stress during pregnancy and other immunologically mediated processes are the risks factors for such birth (Goldenberg et al., 2008; Moore, 2003; Romero et al., 2006). However, despite the identification of such factors, a precise mechanism cannot be established in most cases, and only about half of women who experience preterm birth have an identifiable risk factor (Moore, 2003). Consequently, other factors that may be associated with preterm birth are currently being explored by data mining (Courtney et al., 2008).

2.2. Data mining

Data mining is the process of extracting hidden patterns from a huge amount of data (Kantardzic, 2003). It is commonly used in a wide range of profiling practices, such as marketing, fraud detection, performance, scientific discovery and medicine (Chen et al., 2009; Huang, Chang, & Wu, 2009; Lin, Shiue, Chen, & Cheng, 2009; Çakır, Çalış, & Küçüksille, 2009). Preterm birth is now

* Corresponding author.

E-mail addresses: i14248@mail.hku.edu.tw (H.Y. Chen), chchuang@mail.cjcu.edu.tw (C.H. Chuang).

¹ These authors contributed equally to the work.

thought to be a syndrome initiated by multiple mechanisms, most of which still can not be established. Thus, this works uses data mining to uncover potentially related factors, and our methodology is shown in Fig. 1. Firstly, we use neural network to find the top 15 impact factors from thousands of variables in our database. Then, we use decision tree C5.0 to classify these factors by weight. The results of our study can provide the information for medical staff or pregnant women to prevent the incidence of preterm.

2.2.1. Neural network

A neural network model is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information (Liu, Yuan, & Liao, 2009; Rajan, Ramalingam, Ganesan, Palanivel, & Palaniappan, 2009). Neural network is composed of artificial neurons or nodes. A single neuron may be connected to other neurons, and the total number of neurons and connections in a network may be extensive. Such neural network can be make predictions. For example, a credit card company may use a neural network to quickly identify transactions which have a high probability of being fraudulent. In this work, we use a neural network to find the top 15 impact factors from thousands of variables in our database. These factors are than analyzed by our secondary strategy, a decision tree, as explain below.

2.2.2. Decision tree

A decision tree is a predictive model, which is used in classification, clustering, and prediction (Duman, Erdamar, Eroglu, Telatar, & Yetkin, 2009; Wu, Lee, Huang, Liu, & Horng, 2009). A decision tree is a diagram, which uses a tree-like graph or model of decisions and their possible consequences as a visual and analytical decision support tool. The expected values of competing alternatives in each node are calculated, and a decision tree is thus a mapping from observations of an item to conclusions of its target value. Commonly used decision tree models (algorithms) include ID3 (Iterative Dichotomiser 3), C4.5, C5.0, CART (Classification and Regression Tree), CHAID (Chi-squared Automatic Interaction detection) and QUEST (Quick, Unbiased and Efficient Statistical Tree).

ID3 and its successors were developed by Ross Quinlan, who discovered it while working in the 1970s (Quinlan, 1986). ID3 is a heuristic method for providing a decision tree, which it generates by employing a top-down, greedy search through the training data

set at each of its tree nodes, seeking the attribute that best separates the instances. ID3 later evolved into C4.5 (Quinlan, 1993), and this was an important with regard to the splitting rule and the calculation method. C5.0 is a commercial version of C4.5, and is available as a closed-source product, such as Clementine and RuleQuest (Han & Kamber, 2007). C5.0 improves the rule generation of C4.5, and can obtain similar results with considerably smaller decision trees (Quinlan, 1997). Other decision tree methods include CART (Breiman, Friedman, Olshen, & Stone, 1984) and CHAID (Loh & Shih, 1997), which provide a set of rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating two-way splits, while CHAID creates multi-way splits. CART typically requires less data preparation than CHAID. QUEST, another type of decision tree, is similar to the CART algorithm, but is designed to reduce the processing time required for large CART analyses (Agrawal, Mehta, Shafer, & Srikant, 1996).

A decision tree is exploratory in nature, identifying clusters or segments of interest. We thus try use a decision tree to identify the 15 most important impact factors for preterm birth. We used the C5.0 algorithm, which obtained considerably more results than the other decision tree methods.

3. Empirical study

3.1. Research structure

Clementine 10.0 is a commercial data mining tool (SPSS, 2005) that supports various analyses, such as neural network, decision tree, regression, logistic, and so on. To analyze the nominal variables, we used the neural network and C5.0 of Clementine 10.0 to mine the data.

The research procedures were as follow:

1. Problem definition: Preterm birth is one of the leading causes of diseases and death among newborns. In addition, preterm infants often suffer long-term health problems, including lung diseases, vision and hearing impairments, and learning disabilities. However, the mechanism that causes preterm birth remains unclear. Hence, it is very important to investigate potential risk factors and offer early intervention when necessary.
2. Data collection: The medical data were collected from a prospective pregnancy cohort, which was established between 1984 and 1987. Pregnant women with 26 or more weeks' gestation who came to one Hospital in Taiwan for prenatal care were enrolled in the study and interviewed using a structured

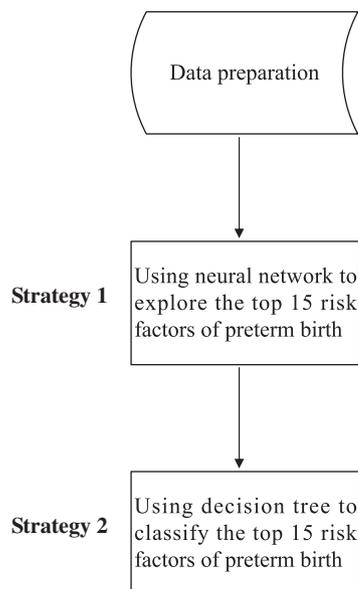


Fig. 1. The process of data mining to explore risk factors of preterm birth.

Table 1
Relative importance of inputs found by neural network.

Number	Factor	Coefficient
1	Number of birth	0.3597
2	Paternal smoking	0.1086
3	Hemorrhage during pregnancy	0.1077
4	Parity	0.0899
5	Maternal age	0.0659
6	Paternal occupation	0.0636
7	Maternal hypertension	0.0620
8	Medicines taken during pregnancy	0.0600
9	Maternal gynecological diseases	0.0599
10	Maternal body height	0.0567
11	Maternal body weight before pregnancy	0.0500
12	Paternal age	0.0499
13	Paternal drinking	0.0492
14	Previous preterm birth	0.0415
15	Vitamins taken during pregnancy	0.0350

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات