Archives of Gynecology and Obstetrics (2019) 300:1565–1582
https://doi.org/10.1007/s00404-019-05325-3

MATERNAL-FETAL MEDICINE

Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features

Toktam Khatibi[1,2] · **Naghme Kheyrikoochaksarayee**[3] · **Mohammad Mehdi Sepehri**[1,2]

Received: 5 February 2019 / Accepted: 9 October 2019 / Published online: 24 October 2019

Abstract

Purpose High rate of preterm birth (birth before 37 weeks of gestation) in the world, its negative outcomes for pregnant women and newborns necessitate to predict preterm birth and identify its main risk factors. Premature deliveries have been divided into provider-initiated (with medical intervention for early terminating the pregnancy) and spontaneous preterm birth (without any intervention) categories in the previous studies. The main aim of this study is proposing methods for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features.

Methods Data from national databank of Maternal and neonatal records (IMAN registry) is used in the study. The collected data have information about more than 1,400,000 deliveries with 112 features. Among them, 116,080 preterm births have occurred (from which 11,799 and 104,281 cases belong to provider-initiated preterm birth and spontaneous premature delivery, respectively). The data can be considered as big data due to its large number of data records, large number of the features and unbalanced distribution of the data between three classes of term, provider-initiated and spontaneous preterm birth. Therefore, we need to analyze data based on big data algorithms. In this paper, Map Reduce-based machine learning algorithms named MR-PB-PFS are proposed for this purpose. Map phase use parallel feature selection and classification methods to score the features. Reduce phase aggregates the feature scores obtained in Map phase and assign final scores to the features. Moreover, the classifiers trained in Map phase are aggregated based on two different ensemble rules in Reduce phase.

Results Experimental results show that the best performance of the proposed models for preterm birth prediction is accuracy of 81% and the area under the receiver operating characteristic curve (AUC) of 68%. Top features for predicting term, provider-initiated preterm and spontaneous premature birth identified in this study are having pregnancy risk factors, having gestational diabetes, having cardiovascular disease, maternal underlying diseases, and mother age. Chronic blood pressure is a high rank feature for preterm birth prediction and father nationality is highly important for discriminating provider-initiated from spontaneous premature delivery.

Conclusions Identifying the pregnant women with high risk of spontaneous premature or therapeutic preterm delivery in our proposed model can help them to: (1) reduce the probability of premature birth with monitoring and management of the main risk factors and/or (2) educate them to care from the premature newborn. Management and monitoring top features discriminating term, provider-initiated preterm and spontaneous premature birth or their associated factors can reduce preterm labor or its negative outcomes.

Keywords Preterm birth prediction · Big data · Map-reduce · Feature selection · Ensemble classifier

Introduction

Preterm birth occurs before 37 weeks of gestation [1]. Rate of preterm labor ranges from 5 to 18% in different countries [2]. Every year, about 15 million premature delivery is occurred all over the world [1].

The preterm birth has substantial and increasing burden [3]. Premature delivery is the leading cause of infant

✉ Toktam Khatibi
toktamk.khatibi@modares.ac.ir; toktamk@gmail.com; khatibi.t@iums.ac.ir

Extended author information available on the last page of the article