Computer Methods and Programs in Biomedicine 140 (2017) 45–51

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach

Brook Tesfaye [a,*], Suleman Atique [b], Noah Elias [a], Legesse Dibaba [c], Syed-Abdul Shabbir [b], Mihiretu Kebede [d,e,f]

[a] Health Policy and Planning Directorate, Ethiopian Federal Ministry of Health, Addis Ababa, Ethiopia
[b] Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan
[c] Health Information Technology Directorate, Ethiopian Federal Ministry of Health, Addis Ababa, Ethiopia
[d] University of Gondar, College of Medicine and Health Science, Institute of Public Health, Gondar, Ethiopia
[e] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstraße 30, Bremen, Germany
[f] University of Bremen, Health Sciences, Bremen, Germany

ARTICLE INFO

Article history:
Received 15 November 2016
Revised 24 November 2016
Accepted 25 November 2016

Keywords:
Child mortality
Data mining
Sustainable development goals
Developing country
Ethiopia

ABSTRACT

Background: Improving child health and reducing child mortality rate are key health priorities in developing countries. This study aimed to identify determinant sand develop, a web-based child mortality prediction model in Ethiopian local language using classification data mining algorithm.

Methods: Decision tree (using J48 algorithm) and rule induction (using PART algorithm) techniques were applied on 11,654 records of Ethiopian demographic and health survey data. Waikato Environment for Knowledge Analysis (WEKA) for windows version 3.6.8 was used to develop optimal models. 8157 (70%) records were randomly allocated to training group for model building while; the remaining 3496 (30%) records were allocated as the test group for model validation. The validation of the model was assessed using accuracy, sensitivity, specificity and area under Receiver Operating Characteristics (ROC) curve. Using Statistical Package for Social Sciences (SPSS) version 20.0; logistic regressions and Odds Ratio (OR) with 95% Confidence Interval (CI) was used to identify determinants of child mortality.

Results: The child mortality rate was 72 deaths per 1000 live births. Breast-feeding (AOR = 1.46, (95% CI [1.22. 1.75]), maternal education (AOR = 1.40, 95% CI [1.11, 1.81]), family planning (AOR = 1.21, [1.08, 1.43]), preceding birth interval (AOR = 4.90, [2.94, 8.15]), presence of diarrhea (AOR = 1.54, 95% CI [1.32, 1.66]), father's education (AOR = 1.4, 95% CI [1.04, 1.78]), low birth weight (AOR = 1.2, 95% CI [0.98, 1.51]) and, age of the mother at first birth (AOR = 1.42, [1.01–1.89]) were found to be determinants for child mortality. The J48 model had better performance, accuracy (94.3%), sensitivity (93.8%), specificity (94.3%), Positive Predictive Value (PPV) (92.2%), Negative Predictive Value (NPV) (94.5%) and, the area under ROC (94.8%). Subsequent to developing an optimal prediction model, we relied on this model to develop a web-based application system for child mortality prediction.

Conclusion: In this study, nearly accurate results were obtained by employing decision tree and rule induction techniques. Determinants are identified and a web-based child mortality prediction model in Ethiopian local language is developed. Thus, the result obtained could support child health intervention programs in Ethiopia where trained human resource for health is limited. Advanced classification algorithms need to be tested to come up with optimal models.

1. Introduction

Child mortality is a core indicator for child health and well-being. In 2000, world leaders agreed on the Millennium Development Goals (MDGs) and called for reducing the Child mortality rate by two thirds between 1990 and 2015. This particular goal is known as the MDG 4 [1]. It refers to the death of infants and children under the age of five or between the ages of one month to four years. The global Child mortality rate dropped 53(50, 55)%, from 91 (89, 92) deaths per 1000 live births in 1990 to 43 (41, 46) in 2015. Over the same period, the annual number of under-five deaths dropped from 12.7 million to 5.9 million [1].

The world as a whole has achieved an accelerating progress in reducing the child mortality rate. Promisingly, sub-Saharan Africa,

* Corresponding author.
E-mail address: balagerusew7@gmail.com (B. Tesfaye).

http://dx.doi.org/10.1016/j.cmpb.2016.11.013
0169-2607/

the region with the highest child mortality rate in the world, has also registered a substantive reduction. Its annual rate of reduction increased from 1.6% in 1990s to 4.1% in 2000–2015. This remarkable decline in child mortality since 2000 has saved the lives of 48 million children [1].

Results from the Ethiopian Demographic and Health Survey (DHS) data of 2011 show childhood mortality has declined significantly. Child mortality rate has declined by 47% from 166 deaths per 1000 live births to 88 deaths per 1000 live births [2]. The Ethiopian Demographic and Health Survey (EDHS) collect and accumulate a wide variety of data. There is an urgent need for a new generation of computational techniques and tools to assist health professionals in extracting useful information and knowledge from this big datasets [3].

Big data analytics is an emerging discipline that analyzes large volumes of data [4]. Data mining mainly focuses on finding meaningful information from huge dataset [5]. The algorithms of data mining are one of the methodologies of the big data analytics. These algorithms help solve questions that the traditional statistical tools fail to answer. Data mining is growing in various applications widely like analysis of organic compounds, medicals diagnosis, product design, targeted marketing, credit card fraud detection, financial forecasting, automatic abstraction, predicting shares of television audiences [6]. It uncovers previously unknown patterns from the big datasets and then uses the information to build predictive models [7]. Data mining techniques extract implicit information and knowledge which are potentially useful and unknown in advance. This extraction is from the mass, incomplete, noisy, fuzzy and random data to build predictive models [8].

Previous studies have identified predictors of child mortality that are associated significantly which helps to determine the priority areas in the intervention programs so as to achieve the policy target (reduction of child mortality and improve maternal health). Applying data mining techniques are different from traditional and widely applied statistical data analysis techniques. It mines information and extract predictors on the premise of no clear assumption beforehand [9].

While highly trained and skilled physicians are very scarce in low and middle-income countries, equipping health professionals and health facilities that had inadequate resources with decision support systems for investigating likelihood of child mortality is fundamental for reducing child mortality and improve child health in Ethiopia. Therefore, this study aimed to identify determinants and build a web-based prediction model for child mortality in Ethiopia by applying data mining techniques and algorithms.

2. Materials and methods

2.1. Study area, design and period

The study was conducted in Ethiopia. Decision tree analysis for data mining was applied on EDHS data of 2011. The study was conducted from December 2015 to June 2016.

2.2. Data collection procedures

EDHS 2011 dataset was utilized. The data were originally collected by Macro International United States of America (USA) and CSA Ethiopia. Data for this study was downloaded from the MEASURE DHS database [10]. The data was secondary; therefore detailed information about the data collection procedures is contained in the EDHS 2011 report which is available on the web platform of the data originator. Thus, relevant environmental, socioeconomic and health characteristic data were extracted from the dataset [10].

Table 1
List of predictors ranked according to information gain feature selection technique.

Predictors	Type	Measure	Information gain value
Breast-feeding	Numeric	Scale	0.089774
Family planning	Numeric	Scale	0.071407
Maternal education	Numeric	Scale	0.067241
Parental education	Numeric	Scale	0.060764
Low birth weight	Numeric	Scale	0.053356
Preceding birth interval	Numeric	Scale	0.050140
Presence of diarrhea	Numeric	Scale	0.044865
Age at first birth	Numeric	Scale	0.043173

2.3. Experimental design and predictive model building

A total of 11,654 records that met inclusion criteria were retrieved. Data was extracted from EDHS 2011 children's dataset. Extracted data were cleaned, coded, transformed and entered into Waikato Environment for Knowledge Analysis (WEKA) 3.6.4 software. The extracted dataset was stratified into "Alive" and "Dead" groups. The "Alive" group comprised mothers whose child was alive during the survey. The "Dead" group comprised mothers who had one or more dead child. Since sample sizes of 'Alive' and 'Dead' subgroups is not balanced we have applied Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset and minimize sampling errors. Pruning techniques were used to clean rules that were insignificant. The 10 fold cross validation and 95% split was done to oversee the strength of the association of determinants with the outcome variable.

Classification data mining task was performed. Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects [6].

Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. Decision tree is a classifier that use tree-like graph. The most common use of decision tree is in operations research analysis for calculating conditional probabilities [11]. Using decision tree, health professionals can choose best alternative and transversal from root to leaf indicates unique class separation based on maximum information gain. Decision tree is widely used by many researchers in healthcare field [12].

In this study, two classifiers, J48 and Pruning Rule Based Classification Tree (PART) classification algorithms were deployed for prediction model building. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [13,14].

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [15,16].

Relevant and most influential socioeconomic, demographic and health predictors were extracted from the dataset using a feature selection technique (Table 1) in data mining called information gain. Information gain in data mining helps to select variables having a strong association with the outcome variable. It ranks vari-