Journal of Science-FAS-SEUSL (2020) 01(02) 10-17
ISSN:2738-2184

Decision tree based automated prediction of infant low birth weight

MAC Akmal Jahan and AM Razmy

Faculty of Applied Sciences, South Eastern University of Sri Lanka

Abstract

Data engineering with decision trees plays a vital part in the field of health and medical diagnosis. Low birth weight (LBW) is the single most important factor determining the survival chances of an infant and predicting this LBW based on the maternal factors through the data engineering can be another contribution to the medical diagnosis. In this research, decision tree classifiers are used to predict the incident of low birth weight of newborns based on the maternal factors. A set of decision tree classifiers such as C4.5, Random Tree, Random Forest, Decision Stump, Logistic Model Tree, REP Tree (reduced error pruning tree) and BF (best-first) tree are assessed for this classification purpose. These classifiers are evaluated for their accuracy and time complexity on classifying the child birth weight. A set of data collected from pregnant mothers in the hospitals located in the Eastern part of Sri Lankan have been used for this study. From the experimental analysis, Random Forest produces highest accuracy as 79.71 % while BF Tree and C4.5 shows 79.23%. However, for the time complexity, Random Tree, REP Tree and C4.5 perform less than 1 second whereas the Random Forest utilizes around 5 seconds. On the other hand, BF Tree and C4.5 show precision and recall as 0.792 and 1.0 respectively. In overall, C4.5 could be a best classifier to construct a decision tree model for the prediction of LBW with the acquired date set.

Keywords: low birth weight, decision tree, random forest, C4.5, classification

1. Introduction

The birth weight of an infant is the first weight recorded after birth, ideally measured within the first hours after birth, before significant postnatal weight loss has occurred [1]. Low birth weight (LBW) is defined as a birth weight of less than 2500 g (up to and including 2499 g), as per the World Health Organization (WHO) [2]. LBW is the single most important factor determining the survival chances of an infant and the LBW babies become more susceptible to the protein energy malnutrition and future infections [3]. There are many maternal factors that leads to the outcomes of LBW babies. Some known maternal factors reported in the literature are weight, age, parity, body mass index (BMI), income, education level, consultancy care, history of LBW infants, birth interval, hemoglobin level, disease records and mode of delivery. In addition, there may be other genetic factors which may influence the birth weight of a child.

Early detection or prediction of output of LBW babies in the pregnancies can help to prevent the LBW or can increase the survival chance of the LBW babies by providing extra antenatal care for mother. This predication can be implemented through an automated health care system which has become an active research area in the world with the support of Artificial Intelligence (AI) and Machine learning (ML) where a series of certain algorithms can be used for pattern recognition based on computer models with existing medical and health related data. Classification and regression are the process in machine learning where classifiers are used without much human intervention.

Data engineering with decision trees play a vital part in the field of health and medical diagnosis. In this research, decision tree classifiers are evaluated for their accuracy and time complexity on child birth weight data set acquitted from pregnant mothers in a regional part of Sri Lankan hospitals. To predict child birth weight and construct a decision model, a decision tree algorithm with good performance in terms of accuracy, precision, recall and time complexity is to be designated. The objective of this research is to forecast or classify the infant

Corresponding Author: akmaljahan@seu.ac.lk

weight category with the assistance of decision tree-based models. This will enhance the survival chance of the LBW babies by providing extra antenatal care for mother.

In medical decision making, machine learning methods can be used where they perform prediction based on classification and regression. There are several machine learning algorithms used in medical decision making, and a decision tree is one of the important classification algorithms in current use of medical applications as it is very popular method for pattern classification. Decision tree technique does a major role in clinical diagnosis of diseases such as diabetes, breast and ovarian cancer, thyroid, chest pain and post-operative recovery decisions [4][6]. In the prediction of the child birth weight, few works had been reported in classifying the weight category of the infants using decision classifiers. Moreover, the prediction pattern may vary region to region [12] and can depend on the selected attributes used for the evaluations. Therefore, this work focuses on a prediction using some decision tree based algorithmic models

which utilize a regional child birth related database with a set of attributes collected in Sri Lanka.

1.1. Decision Tree Induction

Decision tree is a prediction structure in which nodes represent the attributes of the dataset and leaves represents the class distribution. Decision tree learning algorithms use *divide and conquer* strategy in top-down recursive manner in which root node represents the whole training data which splits into subsets depending on the selected attribute values based on the splitting condition [6]. A decision tree consists of internal nodes similar to flow chart and follows two steps [7]:

i) Growing phase

ii) Pruning phase

Initially, the growth phase builds the decision tree based on the optimal criteria. Tree is constructed by splitting the training set recursively based on local optimal criteria until most of the records belonging to each of the partitions bearing the same class label [8]. Since the tree may over fit the data, there is a need to handle this over fitting problem. Pruning is the fundamental process in decision trees, which handles over fitting of data by reducing the size of the tree to avoid complexity [13]. The noise and outliers are removed in pruning and the tree is generalized to improve the accuracy of the classification [10][14]. The pruning process supports to handle the problem of overfitting the data in the decision and removing noise and outliers to generalize the tree. Comparatively, time consumption for pruning phase is lesser than the building phase.

There are several algorithms related to decision tree such as C4.5, Random Tree, Random Forest, Decision Stump, Logistic Model Tree, REP Tree (reduced error pruning tree) and BF Tree. They can mostly handle nominal and numeric attributes of the dataset and some of them can handle only numeric attributes. C4.5 partitions the attributes value using a threshold, which handle missing values, and uses Gain ratio as an attribute selection measure to build a decision tree. Decision trees try and split the dataset based on the similarity of the data presents in each group. The idea of *Entropy* and *Information Gain* are used in the decision tree. The former refers to a measure of disorder in a given system and the later is a measure of the decrease in the amount of disorder.

Entropy can be computed as:

$$E = \sum_{i=1}^{n} -p \log_2 p_i$$

where n is number of classes and p_i is probability of a data point in a group of class i.

Rather than using one decision tree, Random Forest uses a set of decision trees. Each tree is trained in the forest using a different, random sampling of data, called Bagging. Random forests combine many individual trees, which is diverse by using random samples to build each tree and designed to operate quickly over large datasets. On the other hand, REP Tree is also used in this experiment. It is a fast decision tree learner where information gain/variance is used to build decision tree and REP tree is used for pruning.

2. Methodology

2.1 Dataset

The dataset used in this work comes from hospitals in Eastern region of Sri Lanka, which has the maternal factors such as age, parity, hemoglobin, initial and last weight, disease control, BMI. There were 2702 observations in the data set. The outcome variable in this study had 3 categories such as Normal birth weight (NB) Low birth weight (LB) and High birth weight (HB).

2.2 Experimental Setup

The work was carried out using machine learning algorithms in Weka 3.9 data mining and machine learning platform. In this experimental setup, three major processes such as pre-processing and attribute selection and construction of decision tree were performed.

A. Pre-Processing

The dataset consists of continuous and discrete attributes. Therefore, the uniformity of the values should be maintained throughout the data set. Discretization is an important method for the evaluation of the performance of the classifier without any bias. It converts continuous attributes into categorical attribute for the uniformity to avoid bias. Unsupervised discretization method was used in pre-processing as supervised discretization produces complex search space [5, 14]. Attribute selection is an important pre-processing step where suitable attributes that can be used for the classification are selected and ranked. Attribute selection process can enhance the performance of the prediction [11]. Figure 1 shows the selected attributes for the evaluation using information Gain and Ranking filter.

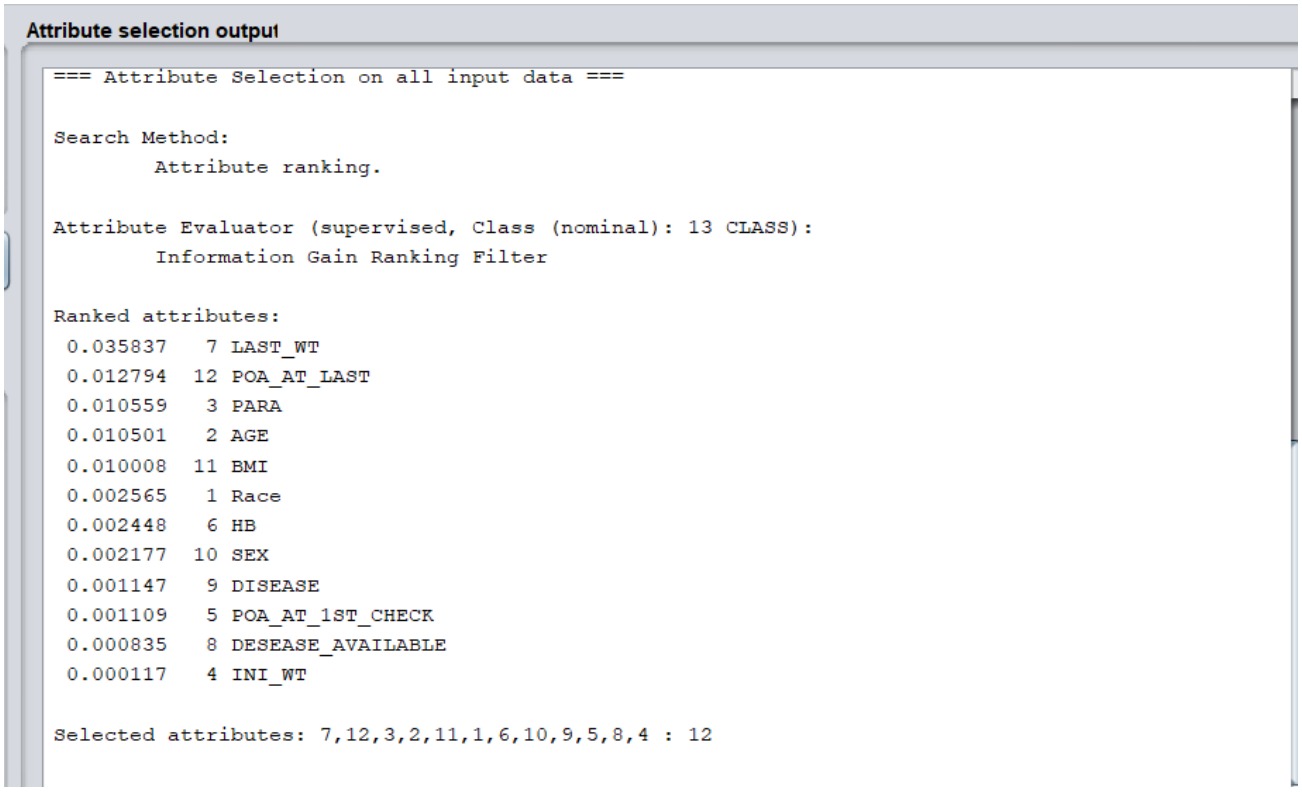

Figure 1: Attribute selection using information Gain and Ranking filter

B. Decision tree Construction

Decision tree classifiers do a major role in data mining and machine learning since it provides user friendly rules for classification, quick construction of decision trees and provide interpretation for better accuracy [7]. Six prominent decision tree algorithms are used in this evaluation where C4.5 provides most frequency usage in construction of decision trees among various decision tree algorithms [9]. WEKA implements C4.5 algorithm using "J48 decision tree classifier". The accuracy and time complexity for the six proposed tree construction were measured. Ten-fold cross-validation was adopted in the evaluation process of the classifiers.

3. Discussion

3.1 Performance Analysis

The following equations are used to compute the performance metrics for the prediction of the infant birth weight classification.

$$Precision = TP/(TP+FP) \times 100\%$$

$$Recall = TP/(TP+FN) \times 100\%$$

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \times 100\%$$

Where, *TP*-True Positive, *FP*-False positive, *FN*-False Negative and *TN*-True Negative

Table 1: Classification accuracy of decision tree classifiers

Decision Tree Classifiers	Classification Accuracy Before Discretization (%)	Time (s)	Classification Accuracy after Discretization (%)	Time (s)
BF Tree	79.23	5.83	79.34	8.82
C4.5 (J48)	79.23	0.93	79.23	0.07
Random Forest	79.71	3.77	76.01	1.81
Random Tree	67.46	0.04	72.27	0.02
REP Tree	78.38	0.17	78.34	0.11
Logistic Model Tree	79.34	3.03	79.20	4.53

Table 1 illustrates comparison of the accuracy of the most frequently used decision tree classifiers and their time complexity. It is observed from the Table 2 that the classification accuracy after discretization is increased in Random Tree. This is due to the reason of the conversion of continuous values into discrete values where Random Tree shows a positive out toward the discretization. In contrast, the process of discretization performs a negative impact on the classification accuracy with other classifiers.

The classifiers such as Random Forest, C4.5, BF Tree and Logistic Model Tree show higher classification accuracy of 79.71%, 79.23%, 79.23% and 79.34% respectively. However, the accuracy of Random Forest is significantly reduced when the dataset is discretized.

Table 2: Precision, Recall and F1-Score of the algorithms

Decision Tree Classifiers	Precision	Recall	F-Measure
BF Tree	0.813	0.793	0.704
C4.5 (J48)	0.792	1.000	0.884
Random Forest	0.673	0.760	0.703
Random Tree	0.674	0.723	0.695
REP Tree	0.670	0.783	0.701
Logistic Model Tree	0.705	0.792	0.703

Table 2 illustrates comparison of the precision, recall and F-Measure of the most frequently used decision tree classifiers with the infant birth weight dataset.

From the experimental results on accuracy of the decision tree classifiers, Random Forest produces highest accuracy of 79.71.% while BF Tree and C4.5 result 79.23%. On the other hand, when we consider time complexity, Random Tree, REP Tree and C4.5 perform less than 1 second whereas the Random Forest utilizes around 5 seconds. When considering precision and recall, BF Tree and C4.5 work effectively where precision and recall for them are 0.792 and 1.0 respectively.

3.2 Tree Evaluation

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

BIRTH_WT = '(3.84-4.48]': HB (65.0)
BIRTH_WT != '(3.84-4.48]'
|   BIRTH_WT = '(1.28-1.92]': LB (23.0)
|   BIRTH_WT != '(1.28-1.92]'
|   |   BIRTH_WT = '(1.92-2.56]'
|   |   |   MOHH = Ninthavur: NB (2.0)
|   |   |   MOHH != Ninthavur
|   |   |   |   MOHH = Area 8: NB (2.0)
|   |   |   |   MOHH != Area 8
|   |   |   |   |   MOHH = Area 3: NB (2.0)
|   |   |   |   |   MOHH != Area 3
|   |   |   |   |   |   DISEASE = D: NB (2.0)
|   |   |   |   |   |   DISEASE != D
|   |   |   |   |   |   |   DISEASE = HD: NB (2.0)
|   |   |   |   |   |   |   DISEASE != HD
|   |   |   |   |   |   |   |   DISEASE = PIH: NB (4.0/1.0)

|   |   |   |   |   |   |   |   DISEASE != PIH
|   |   |   |   |   |   |   |   |   MOHH = Vellaveli: NB (9.0/3.0)
|   |   |   |   |   |   |   |   |   MOHH != Vellaveli: LB (379.0/134.0)
|   |   BIRTH_WT != '(1.92-2.56]'
|   |   |   DISEASE = HOME: LB (4.0)
|   |   |   DISEASE != HOME
|   |   |   |   BIRTH_WT = '(0.64-1.28]': LB (2.0)
|   |   |   |   BIRTH_WT != '(0.64-1.28]'
|   |   |   |   |   POA_AT_LAST = '(56-64]': HB (3.0)
|   |   |   |   |   POA_AT_LAST != '(56-64]'
|   |   |   |   |   |   BIRTH_WT = '(4.48-5.12]': HB (3.0)
|   |   |   |   |   |   BIRTH_WT != '(4.48-5.12]'
|   |   |   |   |   |   |   POA_AT_LAST = '(64-72]': HB (2.0)
|   |   |   |   |   |   |   POA_AT_LAST != '(64-72]'
|   |   |   |   |   |   |   |   BIRTH_WT = '(5.76-inf)': HB (2.0)
|   |   |   |   |   |   |   |   BIRTH_WT != '(5.76-inf)': NB (2196.0/208.0)

Number of Leaves  :     17

Size of the tree :      33
```

Figure 2: Visualizing pruned tree for C4.5 algorithm

Figure 2 illustrates the pruned tree of C4.5 algorithm. Figure 3 represents the tree structure of the REP Tree classifier where each node represents attributes of the dataset and leaves represents the class distribution.

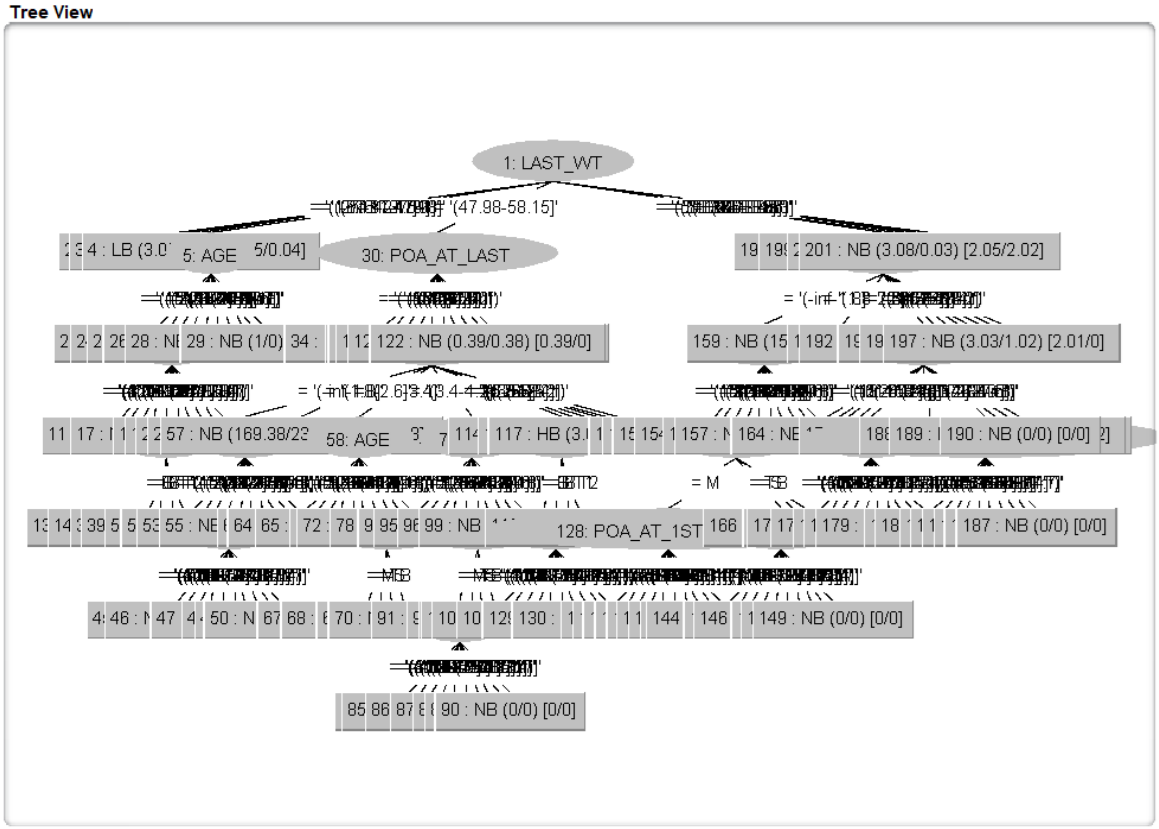

Figure 3: Visualizing REP Tree algorithm

4. Conclusion

In overall, by considering the time complexity, classification accuracy, precision and recall the algorithm C4.5 works effectively compared to the other classifiers. Therefore, C4.5 is selected as a best classifier to construct decision tree model for the prediction of newborn weight category in the regional part of Sri Lanka. Data discretization does not improve performance of the classifiers except Random Tree with this data set. Even though we have used several parameters in initial data set, some of the variables determine the decision tree with highest importance and the others are having less usage in predicting infant birth weight. The attribute selection method performs a big role for the identification of important indicators for birth weight pattern in this regional dataset.

Acknowledgement:

The authors wish to acknowledge for the support provided by Dr MMS Jazeelul Ilahi, Prof. Roshan G. Ragel and Sampath Deegalla from University of Peradeniya for advising and collecting information.

References:

[1] Clare L. Cutland, Eve M. Lackritz, Tamala Mallett-Moore, Azucena Bardají, Ravichandran Chandrasekaran, Chandrakant Lahariya, Muhammed Imran Nisar, Milagritos D. Tapia, Jayani Pathirana, Sonali Kochhar, Flor M. Muñoz (2017) Low birth weight: Case definition & guidelines for data collection, analysis, and presentation of maternal immunization safety data, 35(48Part A), pp. 6492–6500.

[2] World Health Organization (2004) International statistical classification of diseases and related health problems, tenth revision, 2nd ed.

[3] A M Razmy, M M J Ilahi. (2010). Prevalence of Low Birth Weight in Ampara Government Teaching Hospital.

[4] Antonia Vlahou, John O Schorge, Betsy W. Gregory, Robert L Coleman (2003) Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral data, Journal of Biotechnology, vol. 5, pp. 308-314.

[5] D.Lavanya, K.Usha Rani. (2011) Performance Evaluation of Decision Tree Classifiers on Medical Datasets, International Journal of Computer Applications, vol. 26(4), pp. 0975 – 8887.

[6] D.Shanthi, G.Sahoo , N.Saravanan (2008) Feature Selection using Hybrid Neuro Genetic Approach in the diagnosis of Stroke, International Journal of Computer science and Network Security, vol. 8(12), pp. 99-107.

[7] J.Han and M. Kambar (2000). Data Mining; Concepts and Techniques, Morgan Kaufman Publishers.

[8] D.Shanthi, G.Sahoo, N.Saravanan (2016) Decision Tree Classifiers to Determine the Patient's Post-Operative Recovery Decision, International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 1(4).

[9] G.Stasis, AC Loukis, EN Pavlopoulus, SA Koutsouris (2003) Using decision tree algorithms as a basis for a heart sound diagnosis decision support system, 4th International IEEE EMBS special topic conference.

[10] Ian H.Witten, Eibe Frank, Mark A. Hall (2012) Data Mining-Practical Machine Learning Tools and Techniques, Morgan Kaufman Publishers.

[11] Isabella Guyon, Andre Elisseeff (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning research, vol. 3, pp.1157-1182.

[12] J.Tao, Z. Yuan, L.Sun, K.Yu and Z. Zhang (2021) Fatal birthweight prediction with measured data by a temporal machine learning method, journal of BMC medical Informatics and Decision Making, pp. 21:26.

[13] Sam Drazin, Matt Montag, Decision Tree Analysis using Weka.

[14] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman (2002) Decision trees: an overview and their use in medicine, Journal of Medical Systems, Kluwer Academic/Plenum Press, vol. 26 (5), pp. 445-463.