MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-1069

Impact of data quality assessment on development of clinical predictive models

Jitendra Jonnagaddala [a,b,c], Siaw-Teng Liaw [a] and Pradeep Ray [d]

[a] *School of Public Health and Community Medicine, UNSW Australia*
[b] *Asia-Pacific Ubiquitous Healthcare Research Centre, UNSW Australia*
[c] *Prince of Wales Clinical School, UNSW Australia.*
[d] *Asia-Pacific Ubiquitous Healthcare Research Centre, UNSW Australia.*

Abstract

Data quality plays a very important role in predicting clinical outcomes. Data quality is multi dimensional and most relevant studies consider just one or two dimensions. In this study a systematic data quality assessment is performed using four data dimensions. The results demonstrate that performance of predictive models improves when the quality of the data is assessed and addressed systematically.

Keywords

predictive modeling; data quality assessment; data preprocessing

Introduction

Low data quality can be a serious issue in predictive modeling. The typical workflow of predictive model development includes data preprocessing, feature selection, model development and evaluation steps. There are various studies about data preprocessing and the effects of techniques employed on overall prediction problem [1, 2, 3]. However, these studies did not preprocess data from a systematic data quality assessment (DQA) perspective. Thus, the objectives of this study are i) to systematically assess data quality using four different data quality dimensions ii) develop predictive models using four algorithms to predict in-hospital mortality. The algorithms used are Naïve bayes (NB), Random forest (RF), Support Vector machines (SVM) and Multi-layer perceptron (MLP) and iii) present the effects of the systematic DQA on the performance of predictive models.

Methods

The authors used the dataset which was provided as part of the 2012 Physionet Computing in Cardiology (CinC2012) challenge [4]. Four different predictive models were developed to predict whether an ICU patient survives hospitalization. The selected data quality dimensions include completeness, consistency, correctness and contextual accuracy. Data quality metrics were calculated in percentages on each column of the dataset using the selected data quality dimensions. The calculated metrics were used to systematically profile the dataset to select relevant techniques (imputation or deletion or classification) to address the data quality issues [5]. All available variables in the dataset were used as features. The performance of the developed predictive models were measured using accuracy (represented in percentage), positive prediction value (PPV) and the F-score.

Results

DQA was performed on both training and test sets using completeness, consistency, correctness and contextual accuracy dimensions. DQA Results show that a significant amount of variables had more than 50% of missing data. The final results suggest that performing DQA systematically improves the performance of predictive models. The results are consistent with results reported in studies where just one or two data quality dimensions are used [3]. The RF based predictive model was the one which was most affected by the DQA and it showed marked improvement when all the data quality issues identified using data quality dimensions were rectified.

Conclusion

In this study, the authors explored the impacts of data quality on clinical predictive modelling by performing a systematic DQA. The authors observed that NB based model performance remained consistent but in the end the RF based model outperformed the rest of the models after DQA. The results also demonstrate that performance of predictive models improve when the quality of the data is assessed and addressed systematically.

References

[1] Bellazzi, R. and B. Zupan, *Predictive data mining in clinical medicine: current issues and guidelines.* International journal of medical informatics, 2008. 77(2): p. 81-97.

[2] Lin, J.-H. and P.J. Haug. *Data preparation framework for preprocessing clinical data in data mining*. in *AMIA Annual Symposium Proceedings*. 2006.

[3] Johnson, A.E., A.A. Kramer, and G.D. Clifford, *Data preprocessing and mortality prediction: the Physionet/CinC 2012 challenge revisited.* 2014.

[4] Silva, I., et al., *Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012.* Computing in cardiology, 2012. 39: p. 245.

[5] Cismondi, F., et al., *Missing data in medical databases: Impute, delete or classify?* Artificial intelligence in medicine, 2013. 58(1): p. 63-72.