

Performance comparison approach of different disease prediction studies using quantitative index

Halima ElHamdaoui

Universite Sidi Mohamed Ben Abdellah Ecole Nationale des Sciences Appliquees de Fes

Said Boujraf (✉ sboujraf@gmail.com)

Universite Sidi Mohamed Ben Abdellah Ecole Nationale des Sciences Appliquees de Fes

Short report

Keywords: Performance, Comparison, prediction of diseases, accuracy

Posted Date: May 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-20623/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background and aim Comparing databases performances is quite challenging, especially they are presenting different characteristics including the quantitative and quantitative aspects acquired. The primary aim of this paper is to introduce a quantitative method for performance comparison of prediction of diseases studies using measurable parameters.

Results The proposed method focuses on the following parameters: the accuracy and the number of attributes, to compute the performance index. We outline the major obstacles in terms of the construction of a comparative reliable approach, and a demonstrative case regarding the prediction of heart disease is given to validate the effectiveness of the proposed method.

Conclusions Comparing performance of databases used in computer aided diagnosis should be based on objective quantitative basis, inhere presented index allows to achieve this goal, and overcome the variability found through databases. The presented methods should constitute a major tool for normalizing the comparison of databases performances, hence allowing more transparent and better evidencing databases qualities.

Introduction

The integration of machine learning (ML) in the medical field is helping to improve health care delivery and its quality [1]. Indeed, ML techniques are evolving towards solving real health problems. Indeed, among these problems we find the detection and the prediction of diseases such as diabetes, heart disease, cancer, liver disease, and brain diseases [2–6]. The ML techniques usually include neural network, support vector machine, Naïve Bayes, Decision Tree, and genetic algorithm. In order to evaluate the predictive model, researchers compared their achieved results with other studies in literature using applied metrics. Typically, the most commonly employed metrics are those based on the confusion matrix [7]. However, comparing methods in a measurable and informative way is very complex, since each study used different techniques under different conditions with various and different parameters whereas homogeneity is required the same for all researchers to compare different applied techniques, so that a reasonable comparison can be conducted.

Hence, this paper aims to provide measurable and quantitative methods to perform a reasonable comparison of used methods, so that it could be possible to evaluate whether their proposed models were reliable comparing to others studies, such approach allows minimizing accurately biases. Special consideration is given to scientific literature for the prediction of heart disease to appraise the proposed methods using many parameters.

Methods

The assessment of the prediction of performance is based on the analysis of the confusion matrix method. This method has binary class, one is indexed by the actual class of an object, and the other is

indexed by the predictive class. Various well-known measurement indices can be derived, such as sensitivity, specificity, accuracy and precision [8]. These measurements are describing the achieved results in order to evaluate the predictive model. Generally, researchers are comparing their achieved results with other studies in literature using the accuracy which was being the most known and used metric. In fact, each model takes into account effectively different parameters such as the ML algorithm, number of output class, number of subjects and the number of used attributes, besides it is more useful and convenient to identify and introduce the contribution of a new model of comparison where more than one parameter of studied models is different. Thus, we suggested a new performance metric for reliable comparison. The introduced performance metric is defined as a ratio of the achieved accuracy to the number of used attributes as follows:

[Please see the supplementary files section to view the equation.] (1)

The introduced index would enhance methods comparison based on ratio considering both the accuracy and the number of used attributes.

Discussion

We have tested the suggested comparison measurable methodology described above for predicting of heart disease. The parameters used of reviewed studies were summarized in the Table 1 [9-15]. In the literature, studies employed different data sets which is impacting the comparison process and the objectivity and measurability. In addition, whenever using the same data set, the numbers of attributes are varying and depending of the conducted study. Indeed, the number of attributes remains a critical parameter for the prediction of heart disease; subsequently it is related to the risk factors to develop such disease. In this case, the known University of California Irvine (UCI) heart disease data set has used larger number of attributes to predict the disease that the total number of attributes 76 attributes. Furthermore, each study used different number of attributes among 76 attributes suggested. Thus, we used the metric index ρ_i (Performance index) in our suggested method to improve the performance comparison. The findings of the performed comparison using the introduced indices are shown in Figure 1. Indeed, the finding showed significant correlation between the accuracy and the number of attributes. Therefore, the result is well illustrating the interest of used metric index and demonstrated the validity of our proposed method which is taking the form of measurable performance designed specifically for this particular application.

Conclusion

We introduced new quantitative metric standard to evaluate the performance of a predictive model. The suggested approach using Performance index ρ_i is reliable, repetitive, allowing quantifying the performance and the ability to compare new model to reported models in the literature.

This proposed method allows advancing the research methodology and helps researchers to evaluate their models in most quantitative and reliable manner.

Declarations

- **Availability of data and materials**

The used materials and data are both available upon request

- **Competing interests**

Authors declare to not have any competing interests

- **Funding**

This manuscript production processes did not benefit of any funding

- **Authors' contributions**

SB: Design and concept, editing, final approval

HE: concept discussion and validation, First drafting, editing

- **Acknowledgements**

'Not applicable'

Abbreviations

Machine learning: ML

Performance index: ρ_i

Number of attributes: NofAttributes

University of California Irvine: UCI

References

[1] K.Y. Ngiam, I.W. Khor, "Big data and machine learning algorithms for health-care delivery", *The Lancet Oncology*(2019), **20**(5) e262-e273.

[2] I. Kavakiotis et al. "Machine learning and data mining methods in diabetes research", *Computational and Structural Biotechnology journal*, 15(2017) 104-116.

[3] M. [Tayefi et al.](#)"hs-CRP is strongly associated with coronary heart disease (CHD): A datamining approach using decision tree algorithm". *Comput. Methods Programs Biomed.* 141(2017) 105–109.

- [4] K. Kourou et al. "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology journal, 13(2015) 8-17.
- [5] C.C. Wu et al. "Prediction of fatty liver disease using machine learning algorithms", Computer Methods and Programs in Biomedicine, 170 (2019) 23-29.
- [6] H.E Hamdaoui, M .Maaroufi, B. Alami, N.E. Chaoui, S. Boujraf. "Computer-aided diagnosis systems for detecting intracranial aneurysms using 3D angiographic data sets: Review". Proceedings - 3rd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2017 8075568.
- [7] A. Luque et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix", Pattern Recognition, 91(2019) 216-231.
- [8] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks", Inf. Process. Manag. 45 (2009) 427–437.
- [9] M. Shouman, T. Turner, R. Stocker, "Using decision tree for diagnosing heart disease patients", Proceedings of the Ninth Australasian Data Mining Conference (2011) 121.
- [10] E.M. Karabulut, T.Ibrikçi, "Effective diagnosis of coronary artery disease using the rotation forest ensemble method". J. Med. Syst (2012), 36 (5) 3011–3018.
- [11] M. Abdar et al. "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases", International Journal of Electrical and Computer Engineering (2015), 5(6) 1569-1576.
- [12] Purushottam et al. "Efficient Heart Disease Prediction System", Procedia Computer Science, 85(2016) 962-969.
- [13] K. H. Miao et al. "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning", International Journal of Advanced Computer Science and Applications (2016), 7(10).
- [14] O.W. Samuel, G.M. Asogbon, A.K. Sangaiah, P. Fang, G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction". Expert Syst. Appl, 68(2017) 163–172.
- [15] L. M. Hung et al. "Automatic Heart Disease Prediction Using Feature Selection and Data Mining Technique" Journal of Computer Science and Cybernetics(2018), 34(1) 33-47.

Table

Table1. Summary of reviewed studies for the prediction of heart disease

| Author | Year | ML algorithms | Datasets | Number of subjects | Number of Attribute | Best Accuracy % |
|--------------------------------|------|-----------------------------|--------------------------------------|--------------------|---------------------|-----------------|
| M. Shouman et al. [9] | 2011 | Decision Tree | UCI(Cleveland) | 297 | 13 | 84.10 |
| E.M. Karabulut, T.Ibrikçi [10] | 2012 | ANN | UCI(Cleveland) | 297 | 13 | 91.2 |
| M. Abdar et al. [11] | 2015 | C5.0 | Statlog | 270 | 13 | 93.02 |
| Purushottam et al. [12] | 2016 | Decision rules | UCI(Cleveland) | 303 | 13 | 86.70 |
| K. H. Miao et al. [13] | 2016 | Adaptive Boosting Algorithm | UCI(Cleveland) | 303 | 29 | 80.14 |
| O.W. Samuel et al. [14] | 2017 | ANN | UCI(Cleveland) | 297 | 13 | 91.1 |
| L. M. Hung et al. [15] | 2018 | Linear SVM | UCI(Cleveland+Hungarian+Switzerland) | 699 | 58 | 89.93 |

Figures

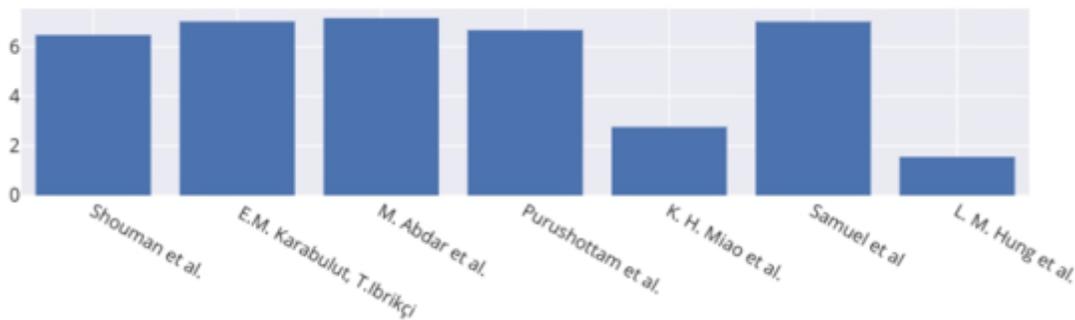


Figure 1

Comparison of performance index of reviewed studies for prediction of heart disease

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Equation.docx](#)