Towards an efficient Prediction Model of Malaria Cases in Senegal

Ousseynou Mbaye, Mouhamadou Lamine Ba

Université Alioune Diop de Bambey, Bambey, Senegal
{ousseynou.mbaye,mouhamadoulamine.ba}@uadb.edu.sn

Abstract. One of the most deadly disease in the world, Malaria remains a real flail in Sub-saharan Africa in particular. In countries like Senegal, such a situation is acute due to the lack of high quality healthcare services and well-formed staffs able to perform accurate diagnosis of diseases that patients suffer from. This calls for the need of finding automated tools to help medical actors in their decision making process. In this paper, we present first steps towards an efficient way to automatically diagnosis Malaria occurence or not based on patient signs and symptoms, and the outcome from the quick diagnosis test. Our prediction approach is built on the logistic regression function. First expermients on a real world patient dataset, as well as a semi-synthetic dataset, show promising performance results regarding the effectiveness of the proposed approach.

Keywords: Malaria · Diagnosis · Data imputation · Prediction Model.

1 Introduction

Malaria is one amongst the most deadly disease in the world, especially in sub-saharan Africa countries such as Senegal. Malaria is caused by parasitic single-celled microorganisms belonging to the Plasmodium group; it is an infectious disease which is transmitted to human being through bites from infected female Anopheles mosquitoes. Someone who suffers from Malaria may present symptoms that typically include fever, tiredness, vomiting, and headaches. In its severe form, the disease can cause yellow skin, seizures, coma or death.

Studied problem and motivations. According to the last report [5] about the propagation of Malaria disease around the world, published in November 2017 by the World international Health Organization (WHO in short), 216 millions of cases have been reported in 2016. As a result, the number of cases has significantly increased when compared to the 211 millions of reported Malaria patients in 2016. As for the number of death due to Malaria, it does not decrease between 2016 and 2017 (446.000 vs. 445.000) despite the huge effort made by governements and non-governmental organization to improve healthcare services and the awareness strategies, especially in critical areas. When analyzing the statistics above in details, one can easily notice that the burden of the Africa region of the World international Health Organization is colossal. Indeed, 90% of Malaria

cases and 90% of deaths due to the disease were located in this area in 2016. More specifically, 80% of the burden in terms of morbidity is distributed in fifteen countries, all located in Sub-saharan Africa except India. This demonstrates that Malaria is a real flail in Sub-saharan Africa states and Senegal is not spared at all. We investigate in this study an efficient approach to predict, using machine learning, the occurence or not of Malaria when a patient has to be diagnosed. Given the patient signs and symptoms, as well as the result from the quick diagnosis test, our solution should be to automatically tell if she suffers from Malaria or not with a high accuracy.

Malaria is an acute problem in Senegal due mainly to the lack of high quality healthcare services and well-formed staffs able to perform accurate diagnosis of diseases that patients suffer from. Over the past years, the government with the help of international organizations have tried to eradicate Malaria by implementing various proactive and reactive solutions to fill the gap in terms of services and human resources. However, the mortality rate is still very high, e.g. in underserved areas, areas without required healthcare needs, uneducated people, population with low income, etc. Most of these deaths cases are reported to be caused by inaccurate diagnosis, sometimes incomplete leading to a bad prediction of the exact type of Malaria. On the other hand, Malaria occurence or complication can often occur during popular events (for instance religious events) which group thousands of persons from everywhere in the country during a short time period. During those popular events, non-permanent medical points are set in order to assist and treat ill persons; the staff in a given health point can consist sometimes of only volunteers with no medical skills. Every medical point can have to receive hundreds of patients each day with some of them potentially suffering from Malaria. This appeals for the need of finding automated tools to help medical actors in their decision making process, and thereby to improve provided healthcare services.

Proposed diagnosis approach. In this paper we present first steps towards an efficient manner to automatically diagnosis Malaria occurence or not based on patient signs and symptoms, and the outcome from the quick diagnosis test. We define our diagnosis task as a classical binary classification problem by considering two classes: “Malaria” and “not-Malaria”. Given a patient data, our main goal is to properly find to which class the patient belongs. To solve this classification problem we rely on machine learning and use the logistic regression function as the basis of our prediction approach. Machine learning has been largely used in several domains (e.g. Health Informatics [2]) for various purposes whereas logistic regression has demonstrated its efficiency when dealing with a binary classification problem.

As an application scenario, we focus on predicting Malaria cases in Senegal. At this end, we use a large volume of patient record dataset collected during the most popular religious event in Senegal from the different installed health points, namely more than twenty points that deadly receive hundreds of patients. As an immediate result of this work, we introduce a data preparation pipeline in order to (i) explore the dataset for profiling purpose; (ii) only retain records

related to Malaria; (iii) clean and transform attributes, as well data values, into the extracted Malaria dataset; and (iv) impute missing values (there were lot of missing values in the collected health dataset as reported in Section3). Such a data preparation pipeline has been realized using *OpenRefine* (formely Google Refine) to perform various cleaning and profiling tasks on our raw atient dataset and *missForest*, a robust algorithm for imputing missing data of diverses types; see Section3 for more details. Expermients on the real world patient dataset, augmenting with a semi-synthetic dataset, show promising performance results regarding the effectiveness of the proposed approach.

Paper organization. The remaining of the paper is organized as follows. We summarize the related work on data imputation and binary classification methods in Section 2. In Section 3 we introduce a data preparation pipeline on the raw collected patient records for the prediction phase. We then present our prediction model for Malaria cases in Section 4. Experiments and performance analysis on the collected real-world dataset, as well as a semi-synthetic dataset, are detailled in Section 5 before we conclude in Section 6.

2 Related work

In this section, we summarize the sate-of-the-art research on Malaria in general, and in particular the use of machine learning technique in health Informatics to deal with the different aspects related to the major healthcare problems worldwide such as Malaria.

Studies on Malaria. As it is well-known, Malaria is caused by the bite of the *female Anopheles*, the most dangerous of which is *Plasmodium falciparum*. Many early works have been consequently focused on the study of the evolution and the distribution of the responsible mosquito, mainly with the goal to detect or diagnosis the severity of the disease given an infected patient [3,1]. Recent research on Malaria have largely adopted machine learning and showed its ability to solve various aspects of the disease. Most oth these machine learning based techniques are based on the analysis of blood data obtained from high-definition microscopic screenshots as in [4]. The authors in [4] propose an unsupervised learning algorithm that detects and determines the types of infected blood cells. Used prediction approach consists of quantifying the amount of plasmodium parasites in a blood smear.

Other ML works in HI.

3 Data imputation

4 Prediction Model

5 Experimentation and results

5.1 Experimentation setting

5.2 Performance measures

5.3 Analysis of the results

6 Conclusion

References

1. AS, A., AM, V., SH., K.: Malaria parasite development in the mosquito and infection of the mammalian host pp. 195–221 (2009)
2. Dua, S., Acharya, U.R., Dua, P.: Machine Learning in Healthcare Informatics. Springer Publishing Company, Incorporated (2013)
3. Ferguson, H.M., Mackinnon, M.J., Chan, B.H., Read, A.F.: Mosquito mortality and the evolution of malaria virulence. Evolution **57**(12), 2792–2804 (2003)
4. Kunwar, S.: Malaria Detection Using Image Processing and Machine Learning. ArXiv e-prints (Jan 2018)
5. WHO: World malaria report in 2017 (2017)