Mining ICDDR, B Hospital Surveillance Data and Exhibiting Strategies for Balancing Large Unbalanced Datasets

Adnan Firoze, School of Engineering and Applied Science (SEAS), Columbia University, New York City, NY, USA
Rashedur M Rahman, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ABSTRACT

This research uses a number of classifier models on Hospital Surveillance data to classify admitted patients according to their critical conditions. Three class labels were used to distinguish the criticality of admitted patients. Furthermore, set forth are two distinct approaches to address the over-fitting problem in the unbalanced dataset since the frequency of instances of the class ‘low’ is significantly higher than the other two classes. Apart from trimming the dataset to balance the classes, this work has dealt with the over-fitting problem by introducing the ‘Synthetic Minority Over-sampling Technique’ (SMOTE) algorithm coupled with Locally Linear Embedding (LLE). It has constructed three models that applied the neural, and multinomial logistic regression classifications and finally compared the performance of the work’s models with the models developed by Rahman and Hasan (2011) where they used several decision tree models to classify the same dataset using tenfold cross validation. Additionally, for a comprehensive comparative analysis, this work has compared the classification performance of the authors’ novel third model using support vector machine (SVM). After comparison, the work shows that one of the authors’ models surpasses all prior models in terms of classification performance, taking into account the performance time trade-off, giving them an efficient model that handles large scale unbalanced datasets efficiently with standard classification performance. The models developed in this research can become imperative tools to doctors when large numbers of patients arrive in a short interval especially during epidemics. Since, intervention of machines become a necessity when doctors are scarce, computer applications powered by these models are helpful to diagnose and measure the criticality of the newly arrived patients with the help of the historical data kept in the surveillance database.

Keywords: Artificial Intelligence, Classification, Data Mining, Imbalanced Data, Medical Surveillance, Multinomial Logistic Regression, Neural Network, SMOTE

DOI: 10.4018/IJHISI.2015010103
1. INTRODUCTION

Machine intervention in medicine and mining large scale medical surveillance data have caught significant attention in the recent years due to epidemics and the scarcity of physicians. We have pursued this research based on a dataset that stores patients’ data from January 1, 1996 to December 31, 2007 (which is hospital surveillance data of 12 years) that was collected at International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B, 2008). Previously, a research work using this data repository was conducted using decision-tree induction algorithms by Rahman and Hasan (2011). We have introduced several newer approaches to deal the classification problem along with a novel way of balancing the dataset.

ICDDR,B established a diarrhoeal disease surveillance system in Dhaka, Bangladesh in 1979 and later extended it to their Matlab hospital at Comilla, Bangladesh in 2003. The surveillance system collects information on clinical, epidemiological and demographic characteristics of patients. A systematic 2% sub-sample of patients attending Clinical Research and Service Centre (CRSC) and all patients from the Health and Demographic Surveillance System (HDSS) area attending the Matlab hospital are enrolled into the surveillance program. The patients and/ or their attendants supply information on socioeconomic and demographic characteristics, housing and environmental conditions, feeding practices, particularly among infants and young children, and on the use of drugs and fluid therapy at home to the interviewers. Moreover, nosocomial features e.g. clinical characteristics, anthropometric measurements, treatments received at the facility, and clinical outcomes of patients are also recorded. Extensive microbiological assessments of fecal samples (microscopy, culture, and ELISA) of patients are performed to identify diarrheal pathogens and to determine antimicrobial susceptibility of bacterial pathogens. It enables the center to detect the emergence of new pathogens and responds to early identification of outbreaks and their locations to suggest the Government of Bangladesh to take preventive measures.

Collected information is representative of the population and thus it serves as an important data repository for conducting epidemiological studies, validation of clinical studies, and it also helps develops new research ideas and study design.

1.1. Motivation

Upon arrival at hospital, an initial diagnosis is carried out by the duty physician to find out the criticality of the patient’s condition and upon completion, the duty doctor takes necessary action accordingly. This step becomes difficult yet more crucial in the event of an epidemic like that of the year when 1000 patient on an average got admitted to the hospital on daily basis due to flood. The importance of this surfaced again in 2009 after the cyclone Aila hit the southern coast of Bangladesh. It becomes increasingly difficult to diagnose every patient satisfactorily due to scarcity of duty doctors. Thus, machine intervention to diagnose and measure the criticality of the newly arrived patient with the help of the historical data kept in the surveillance database was a necessity. The application asks few questions on physical condition and history of the patient and accordingly determines the critical condition of the patient as low, medium or high.

1.2. Objective

The primary objective of this research is to create an efficient classification model that serves effectively to classify the large repository of ICDDR,B hospital surveillance data into low, mid and high criticality of patients, while taking into account the intrinsic issues of an unbalanced dataset. Instead of working with the dataset directly, for achieving a more meaningful system, we rejected incomplete data records.

The outcome field has the following values stored: 1 = Cured, 2 = Illness continued, 3 = Died, 4 = Absconded, 5 = Others, 9 = Unknown. We have considered the records of the patients with outcome = 1 rejected the others since most of those records were incomplete. Also, the
Related Content

An Unusual Association of Lung and Ovarian Malignancy in a Young Nonsmoker Female
www.igi-global.com/article/unusual-association-lung-ovarian-malignancy/75177?camid=4v1a

Evaluation of Quality of Context Information in U-Health Smart Homes
www.igi-global.com/chapter/evaluation-quality-context-information-health/64989?camid=4v1a
Examining Heterogeneous Patterns of Electronic Health Records Use: A Contingency Perspective and Assessment
www.igi-global.com/article/examining-heterogeneous-patterns-electronic-health/67366?camid=4v1a

Cloud Storage Privacy in Health Care Systems Based on IP and Geo-Location Validation Using K-Mean Clustering Technique
Mamoon Rashid, Harjeet Singh and Vishal Goyal (2019). International Journal of E-Health and Medical Communications (pp. 54-65).
www.igi-global.com/article/cloud-storage-privacy-in-health-care-systems-based-on-ip-and-geo-location-validation-using-k-mean-clustering-technique/235441?camid=4v1a