IMPROVING DEVELOPING COUNTRY CLINIC RESOURCE ALLOCATION THROUGH THE APPLICATION OF MACHINE LEARNING ON IMBALANCED CLINICAL DATA

|  |  |  |
| --- | --- | --- |
|  | **Shaaheen Sacoor**  Department of Computer Science  University of Cape Town  Western Cape  South Africa  scrsha001@myuct.ac.za |  |

ABSTRACT

CCS CONCEPTS

• Outline of machine learning → Supervised Learning; • Applied Computing → Computational healthcare

KEYWORDS

Machine Learning, Clinical Resource Allocation, Electronic Health Record, Prediction modelling

1. Introduction

* Clinics are overworked in developing regions
* High patient: clinic ratio
* People in developing regions are more exposed
* Need for better efficiency so patients are better taken care (Health care) - given appropriate amount of consultation time, intervened/visited appropriately, right number of meds/right meds, LSTF Tracing i.e Clinics must allocate their limited resources in such a way that the best healthcare possible is delivered to patients
* Limited resource, Large demand = Need efficient resource allocation
* Need for more personal advice, too little time to look through every patient and their whole history when advising them
* Lots more data being recorded (OpenMRS, etc) , machine learning can help give more personal predictions for individual patients. Clinics can use these individual patient predictions to allocate and prepare its resources as precisely as possible for each patient
* This paper approaches two potential assistive machine learning models for better clinical resource allocation that could be usefully applied within the Malawian clinic context as well as other developing regions
* **Consultation defaulters**
  + Predicting if a specific patient is likely to miss their next consultation based on their history and demographics
  + Since there is generally a high doctor:patient ratio in developing nations, there is a strong need to ensure that patients always attend their consultation so as to avoid wasted resources and ensure patients who need consultations, get them
  + Intervene before miss consultation, visit them, send sms’s, warn them
* **Symptom prediction** 
  + Predicting what symptom a patient is likely to report next based on their past reported symptoms and demographics
  + Predicting the escalation of a patient’s symptoms can greatly help clinic to prepare for the patient’s future medical needs
  + set up another appointment as a check up, send chw to visit, change meds, warn patient and caretaker to prepare for specific symptom, ensure have medication stored for patient’s next symptom

1. Background
   1. Current methods for Consultation defaulters and symptom estimation
   2. Past Applications (DON’T DO ONE PAPER AT A TIME)
      1. HARM symptom prediction with statistical bayesian
         1. Clinical trial data
         2. No ML
         3. More broad – mine just HIV
         4. Learnt from association ruling system that was created. Similar to the tree classifiers set out here
      2. Masters paper
   3. Challenges with clinical data
   * Imbalanced dataset
     + Need balancing sample techniques

* Inherently temporal
* Anonymization
  + Verification needed by clinical expert
    - Variables must logically verified as sensitive problems
  + Data collected for patients not researchers
    - Inaccurate data capturing, shortcuts, gaps in data
    - Missing location data here

1. Experiment Design and Execution

* Malawi Dataset - description
* Data Analysis with visualisations - show some visuals here
* Problem exploration with consultation and data consideration
* Pipeline
  + Sklearn library with imbalance library
  + Techniques tested
  + Balancing methods used
  + Cross-validation and Unseen testing
  + Metrics
    - Roc
    - Specificity
    - Sensitivity
    - Inteperability
    - Inter-variable significance
* Problem descriptions
  + Symptom prediction
    - Features chosen/extracted
    - Multiclass
    - Time constrained
    - Which balance was chosen
  + Consultation defaulters
    - Binary classification
    - Features chosen/extracted
    - Which balance was chosen, why
  + Summary table of features chosen and result set

1. Results

* Symptom prediction
  + Multiclass - ROC values for each symptom
  + ROC, sensitivity, specificity
  + Inter Variable influence
  + Model decay
  + How much data is needed to build reasonable model i.e compare different timelines (predicting next 10 day vs 40 day, using prev 40 day vs 10 day, etc)
  + Balancing technique comparison
* Consultation defaulters
  + ROC, sensitivity, specificity
  + Balancing technique comparison
  + Feature significance

1. Ethics
2. Discussion

* How balancing the data affected everything and choosing the right balance
* Why the classifiers tended to do better here
* Is there evidence to suggest that the symptom prediction had symptoms influence each other
* Was there model decay with temporal symptoms, what is the minimum amount of data needed for reasonable result with temporal symptoms
* Why did some balancing techniques fair better than others
* Would medical professionals find these results reasonable enough to use
* Are the variables and their significance agreed by medical professional

1. Conclusion

* Which technique is best for which problem.
* What was the best accuracy reached
* Which balancing technique performed best in general
* Did the medical expert find this tool helpful

1. Further work

* List possible problems could do
* Defaulters

1. Acknowledgement