Analytics, Machine Learning & NLP – use in BioSurveillance and Public Health practice

Mujitha B. K B*[1], Ajil Jalal[2], Vishnuprasad V[1] and Nishad K A[1]

[1]Informatics, LongRiver Infotech, Bangalore, India; [2]IIT Madras, Chennai, India

Objective

To summarize ways in which Analytics, Machine Learning (ML) and Natural Language Processing (NLP) can improve accuracy and efficiency in bio surveillance and public health practices. We also discuss the use of this framework in typical surveillance applications (Integration with Devices/Sensors, Web/Mobile, Clinical Records, Internet queries, Social/News media).

Introduction

Currently, there is an abundance of data coming from most of the surveillance environments and applications. Identification and filtering of responsive messages from this big data ocean and then processing these informative datasets to gain knowledge are the two real challenges in today's applications.

Use of Analytics has revolutionized many areas. At LongRiver Infotech, we have used various Machine Learning techniques (Regression, Classification, Text Analytics, Decision Trees, Clustering etc.) in different types of applications. These methodologies are abstracted in a generic platform, which can be put to use in many public health and surveillance applications, which are enumerated here.

Methods

In this generic ML platform, we brought together modules covering each of the ML and NLP areas. This platform was then evaluated in a simulated environment – interfacing with a Web/mobile surveillance data capture application, medical devices/sensors over RFID, and social feeds. Surveillance data in both streaming and batch modes have been used for this test environment. 'R' was used for ML algorithms and Infrastructure tools like NoSQL database (Apache Spark), Map Reduce (Spark/Hadoop) and Visual tools (R/Tableau) were integrated in this pilot study.

Results

Each of the independent modules included in this environment had been evaluated in separate projects (Precision, Recall rates, R-squared values, AUC etc. for respective algorithms). Scaling capabilities (input data, ML processing) of the platform was evaluated in an Apache spark cluster.

Conclusions

This framework can be plugged into any surveillance application, which has the required IT infrastructure in place – for efficient and scalable distributed processing and big data handling.

From our evaluation so far, there is an increased interest from various stakeholders in using these Machine Learning algorithms and NLP technology on Surveillance data.

Further enhancements in NLP include:

1) Speech recognition, which enables users to tell their problems (which can first be converted to text and then NLP can act upon it)

2) Support for multiple languages (which enables public to tell in their own local language)

3) Question-Answering (which enables machine processing of user stories and responding with the findings/solutions)

A primary motivation for presentation at this conference is to solicit feedback from public health practitioners on this idea and its potential / challenges for use in existing surveillance systems.

Analytics use in Bio Surveillance

Classify surveillance messages based on symptoms (e.g. classify Malaria, Dengue using symptoms mentioned by citizens)	Classification (Supervised Learning, Logistic Regression, Naïve Bayes classifiers)
Identify clusters of similar problems (e.g. Skin lesions and similar problems reported from neighborhood areas, ILI in school start plotted against space and time)	Clustering (Unsupervised Learning, kNN algorithm)
Identify responsive messages from streaming (or batch) social feeds, news and clinical messages	Text Analytics
Outbreaks (e.g. Severe breathing problem for almost all people in an area - due to Air pollution)	Anomaly Detection
Users can tell (or type in free text) whatever stories they want to - on any particular disease/health problem. This will result in a) Increased participation - mostly from rural people b) More detailed explanations from citizens on their actual problems/observations/queries	Natural Language Processing (NLP)

Analytics use in Public Health practice

Optimal allocation of public health operation funds among available skilled resources, at a desired location	Optimization
During outbreaks, match demands and supplies - against space and time (e.g. Diagnosis needs v/s Lab capacities, Research needs v/s Research Lab facilities)	Recommendation + Location based Analytics
Optical Character Recognition (OCR) (e.g. for scanning immunization reports and extract digitized data)	Machine Learning

Keywords

responsive messages; clustering; classification; machine learning algorithm; text analytics

*Mujitha B. K B
E-mail: mujitha@longriverinfotech.com