Improving Patient Safety via minimizing Adverse Effects

GA Data Science - Final Project Outline

Ari Kamlani

November 19, 2014

1. Problem to be solved

Consider the following scenario: A patient is taking multiple drugs and develops multiple adverse reactions. The patient and/or doctor does not know which prescription drug caused a particular onset reaction. The goal is to predict which drug is the likely contributor to a reaction, and provide some insight and confidence to why this reaction occurred.

2. Description of dataset

* What dataset will you be using for your project?
  + openFDA ([http://open.fda.gov](http://open.fda.gov/))
* How will you access this data? API? Scraping? Other?
  + Restful Web API Service via Queries, returned in JSON format
* What sort of information is in the dataset? In other words, what features are available?
* Adverse report Info {id, date, duplicate, reported by/source type}
* Patient Id Information {Sex, Weight, Age}, Country of Occurrence
* Level of Seriousness, Death Occurrence
* Time Series data from 2004-2014 (Date Format)
  + Number of Drugs being taken by Patient
  + Drug info {Characterization, Dosage, Start /End Date, Indication, Administration}
  + Harmonization: {Generic Name, Brand Name, Manufacturer}
* Reactions per [MedDRA](http://www.meddra.org/), Reaction Outcome type
* Drug Recall Information (via different endpoint)
* How will you turn that data into a training set? (If using a supervised approach)
* The dataset will be retrieved via Restful web API service queries via identified endpoints and selected parameter terms. As the data will include multiple embedded dictionaries and lists to a record, a set of relation tables will need to exist. However this does not necessarily need to be in the context of a database. Imputation of data where no voluntary information is given will need to be performed.
* The data may need to be scaled and/or normalized based on the model and/or feature ranges, before being split into train/test vectors.
* How do you anticipate processing that data to get it into a form to use for your modeling?
* Further Feature Engineering and sub-selection to determine the best correlated features and possibly transforming the reduced dataset to a lower dimensional space.

3. Hypothesis

What is your hypothesis? In other words, what do you hope to predict or otherwise learn as the outcome of your project?

* Confidence in which a prescribed drug has led to a select reaction. In some cases, it may be determined there is no single drug as the culprit, but when administered in combination with others, this has led to the adverse reaction. It is possible the model may help predict that this was not the proper drug to prescribe based on the indication and active prescriptions in the first place.

What are some of the features you might use?

* De-identified Patient Information
* Seriousness of report
* Reactions, Indications
* Correlation of Drugs of Adverse Reactions and Recall based Drugs
* Drug Characterization, Dosage Threshold, Active Prescription, Type of Drug
* Harmonized Drug Names and/or Ingredients

4. Statistical methods I plan to use and why

Think back to our 2x2 matrix slide: Is your problem a classification problem or a regression problem? Will your approach use supervised approaches or unsupervised approaches?

* Classification problem.
  + A combination of supervised and unsupervised methods will be utilized.
  + The given target is a supervised multinomial value.

Which of the machine learning algorithms that we have learned do you plan to use for your final project and why? Which do you explicitly NOT plan to use and why?

* Supervised
  + Multinomial Logistical Regression classification
  + Ensemble Methods: Random Forest classification
* Unsupervised
  + Kmeans to find structure within the groups (drugs, ingredients, patients)
  + Pre-Processing: Dimensionality Reduction – e.g. PCA (or derived variant)
* ML Algorithms not to be used
  + Algorithms based on Binary Classification alone
    - Require to turn categorical target values each into a binary classification
  + Linear Regression
    - Not targeting a continuous problem. There is a continuous aspect in terms of duration, but this may not be a concentration.

5. Applications the finding may have

Once you have completed your project, what are some of the applications of your findings? In other words, how might those findings be applied? What is the “practical” value of the model you will have built?

* When test results do not show a definitive trigger source, doctors too often prescribe 'preventive' drugs for the patient. Unfortunately these 'preventive' drugs can cause adverse effects. Therefore the real value will be improving upon patient safety and improving quality of life. The next phase could turn this into a personalized recommendation engine for prescriptions to patients.

Also, what will your deliverable be in addition to your code and data? Will you write a report in the style of CS229? Will you create a visualization? (NOTE: Do NOT attempt to learn D3 on top of everything else unless you are already a javascript ninja! Seriously.)

* Visualization with iteration through different filters of patients, drugs, and reactions to see how the model performs and can it generalize to some samples it has not seen.

REFERENCES:

[Optional] links to relevant sources

Other sources of information that have launched using openFDA, primarily offering a service to search for information:

* ResearchAE (<http://www.researchae.com/>)
* AdverseEvents (<http://www.adverseevents.com/>)
* IODINE (<http://www.iodine.com/>)

Recall Articles

* <http://www.raps.org/Regulatory-Focus/News/2014/09/02/20197/Massive-Recall-of-Medical-Devices-Largest-Ever-Recorded-by-FDA/>