21/09/2020

Detection and Visualization of Patterns in Medical Data

Zeineb Sahnoun

In italics: comments from me

* Infectious diseases: sub-Saharan Africa = 30% of global deaths
  + HIV/AIDS, malaria, diarrheal diseases, tuberculosis, neglected tropical diseases
  + Rapid identification to limit spread
  + Lacking infrastructure
  + E.g. Ebola: not detected until it had reached major cities!
    - Factor: “inability to detect a new pattern of symptoms with the characteristic spatio-temporal clustering of an infectious outbreak”
* Surveillance process flow:
  + Recognize case
  + Classify case as unusual
  + Report case
  + If there are enough cases so that they are statistically significant: subsequent investigation…
* Recognition:
  + difficult in resource-limited settings
  + political instability, limited reporting and information
  + lack of infrastructure
  + difficulty in recognizing new diseases!
  + Geographically localized
* Mobile applications
  + Guide clinicians 🡪 standardization
  + ePOCT: electronic clinical decision algorithm, Tanzania, guide clinicians for management of childhood infections.
    - Clinical decision algorithm (CDA)
    - Point of care testing (POCT)
    - eCDAs: collect patient data, but algorithms are static
  + *Use data to both: improve diagnostic AND detect / monitor potential outbreaks*
* Unsupervised: helpful for pattern recognition and anomaly detection, however “black box” so no interpretability of the results is possible.
  + Make the results more interpretable, notably using visualization tools
  + Necessary if clinicians want to use this
  + Model-agnostic interpretation method

AIM: platform with interactive data visualization

* Feature engineering, choice of visualization techniques
* Comparing unsupervised learning models for pattern detection
* Build “interpretability” layer on top of clustering algorithm

METHODS:

DATASET:

* Study goal: perform non-inferiority trial of the ePOCT vs. a validated electronic algorithm derived from IMCI (ALMANACH)
* 9 different clinics, 3192 consultations, 2 year period, children 2-59 months old, 1028 features (from administrative to demographic data, diagnosis, lab test results, medical history, etc)

FEATURE SELECTION:

* remove all irrelevant features and those with cardinality = 1
* systematic missing data notably because only a number of parameters within a medical decision tree are explored for a given patient. Dropped features with missing rate that was too high
* 145 features left
* Soft-impute for numerical features
* One-hot encoding for categorical features with a limited number of values (so as to not add too many new features), with missing values being a category 🡪 missing values are considered as one, but are also seen as “unseen” data, vs. non-missing and seen data.
* Spatial features: location names to GPS coordinates (important for disease spreading)
* Temporal features: consultation number per month varies a lot 🡪 keep a “date” feature, which is cyclic (time relative to start of year). *Would it be interesting to keep an indication of the year? For a different purpose: see epidemic evolution over years (nothing to do with the per patient diagnostic)*.

CLUSTERING METHODS:

* Numerical features only: k-means; categorical features only: k-modes
* Mixed features: still a challenge. Can discretize continuous features, or dummy-encode categorical features. Not great though.
  + Need to define measures of similarity between mixed features
  + Need to adapt existing algorithms / develop new ones
  + Gower distance metric
  + K-prototypes algorithm
* K-prototypes:
  + Cost function: trace of the intra-cluster dispersion matrix
  + Similarity measure: combine Hamming distance and Euclidean distance
* Gaussian Mixture Models (GMM):
  + Assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters; mixture component weights and means / covariances. Use with log-likelihood and Expectation-Maximization algorithm (EM)
  + Not great: most features are binary 🡪 Bernoulli would be more appropriate
* Gower distance metric: average of partial dissimilarities across all dimensions
  + Complete linkage algorithm
* Choosing the number of clusters
  + Elbow method: total loss: choose the number of clusters at which point adding a new clusters would not help the total loss
  + Likelihood of data under GMM.

MAP VISUALIZATION:

Use chloropleth map

CLUSTERS INTERPRETABILITY

* Represent cluster by its centroid: but, what if cluster is non-isotropic, or elongated
* Use PCA (2D): limited interpretability of dimensions
* Logistic regression for multi-class classification: define class A and all other classes as B, train classifier predicting class membership, and check out coefficients to determine which features are most important in determining clusters.
* Question that need to be answered
  + When: consultations over time
  + Where: per district
  + What: which symptoms are most useful to discriminate (using logistic regression)
  + Who: demographic details

RESULTS:

* No ground truth labels
* Use GMM although it is not the best fitted to the data

CLUSTERS MAPPING

CLUSTERS REPORTING

* Select cluster data and check it out over time?

DISCSSION:

* Complete pipeline for detection and visualization of patterns in medical data