Hidden Markov Model for the Prediction of Copycat Suicide

Rifat Zahan

PhD Student, CEPHIL
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, CANADA

December 6, 2016

Motivation

- Suicide is one of the leading causes of death worldwide.
- Suicide is associated with significant social, economic and health system cost.
- Suicide can be *contagious* (copycat).
- Finding the transition to "*copycat effect*" state can help health-care providers to deliver early-stage counselling support (person or community-based).
- Prediction of such state can help in stopping further suicide in a particular community.

PC: Suicide Prevention Day, Men of Hope

Objective

We aim to predict the copycat suicide state in a particular geographic region using machine learning algorithm.

Dateset

Data Source

Individual mortality record in the U.S. was obtained from CDC (2016) for year 1968-2014.

Information Retrieved

- Date (1972-1988) and Place (5-digit FIPS code) of the suicide.
- Age and Sex of the individual.
- Method used to commit suicide [using ICD8* and ICD9* code].

Data Summary

- Los-Angeles (LA) had the highest counts of suicide among all the counties in California.
- Females, who used drugs/medication to commit suicide exhibited an evidence of "copycat effect". Therefore, these cases will be considered as study sample.

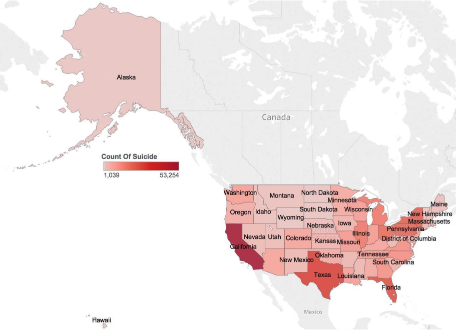

Figure 1: Geo-map of the counts of suicide in different states of U.S. during 1972-1978.

* CDC = Center for Disease Control and Prevention
** ICD = International Code for Disease

Algorithm: Hidden Markov Model (HMM)

- Given the counts of total suicide for a given day, HMM can be used to predict whether it is a copycat state or not.
- **Observed sequence**: counts of suicide in each day.
- **Hidden states**: Copycat or No-Copycat (limited information).
- **Poisson distributed likelihood** naturally characterizes arrival process; therefore, we will use Poisson distribution for the HMM.
- **Unsupervised learning**: *Baumen-Welch (BW)* algorithm.

Results

Training HMM

Real data was split into 67% training data and 33% test data.

Transition Probability Matrix

		Next State	
		Copycat	No-Copycat
Current State	Copycat	0.456	0.544
	No-Copycat	0.643	0.357

Initial Probabilities

$$\begin{array}{r} Copycat \\ No-Copycat \end{array} \begin{pmatrix} 0.001 \\ 0.999 \end{pmatrix}$$

Emission Probabilities

$$\lambda = \left[0.608, 4.926 \times 10^{-05}\right]$$

Model Adequacy Checking

Confusion Matrix: Real Data

		Observed	
		Copycat	No-Copycat
Predicted	Copycat	590	18
	No-Copycat	540	57

Sensitivity: 0.52; Specificity: 0.76

Confusion Matrix: Synthetic Data

		Observed	
		No-Copycat	Copycat
Predicted	No-Copycat	224222	21649
	Copycat	232969	21161

Sensitivity: 0.49; Specificity: 0.49

Discussion

Strength of the study

- The sensitivity and specificity calculated based on the real data indicates that the model is able to predict the transition of the states moderately well.

Limitation of the study

- Both the real data and the synthetic data exhibited moderate predictive ability.
- The seasonality of the time-series data was not considered here.
- Diversity of Los-Angeles may not represent the true scenario of copycat and non-copycat states.

Future work

- Data will be available in future with reliable information on the occurrence of copycat suicide in a particular community.
- Other Machine Learning algorithm (e.g, Neural Network), can be applied to predict the copycat states.