Social Data Integration and Analytics for Health Intelligence

Xiang Ji
Advisors: James Geller (NJIT), Soon Ae Chun (CUNY)
New Jersey Institute of Technology, Newark, NJ, 07102, USA

Abstract
The aim of this Ph.D. work is to develop a big-social health data analytics framework for addressing the following interdisciplinary problems:

- Online social health data are distributed in various forms, it is hard to grasp collective intelligence for patients.
- Monitoring public health concerns is critical for timely response of epidemics, but the current surveillance system is expensive, has limited coverage and time delay.
- Patients’ conditions are correlated with each other to a measureable degree (“co-morbidities”). The relationships are usually too complex to model.

Methods and Results

Social Infobuttons Prototype

Information of the patients with the condition: PTSD
How many patients? 73
who are in the hospital
how are they distributed over six county levels
where is the individual patient
what is the patient gender distribution

Treatments of the condition: PTSD

Individually therapy
(a) physical: a medical
(b) psychological: a mental

Information of the patients with the condition: Fibromyalgia
How many patients? 73
who are in the hospital
how are they distributed over six county levels
where is the individual patient
what is the patient gender distribution

Treatments of the condition: Fibromyalgia

Individually therapy
(a) medical: a physical
(b) psychological: a mental

Results:
1.Ranked list of treatments generated by Social Infobuttons reflects the officially-reported treatments well.
2. Social Infobuttons complemented the treatments or symptoms that are not in authoritative source.


Architectural overview of two-step sentiment classification and opinion summary

Paths of research threads

Problems
Hard to derive collective intelligence from online health data

Technical Challenges
Data in different formats, how to organize?
Classify sentiment tweets? Correlation Between sentiment and news?

The patients’ conditions are related, but hard to model

What features can be used to predict future conditions? How to evaluate predictions?

Technical Approaches
Triple Representation, Semantic Model
Two-Step Classification Algorithm
Recommendation Engine

Applications
Social Infobuttons
Epidemic Outbreak and Spread Detection System
Personal Health Assistant

Social Health Analytics Framework

Result 1: 2Step-SVM performed better than MNB and NB in most of datasets

<table>
<thead>
<tr>
<th>Dataset Id</th>
<th>TS-MNB</th>
<th>TS-NB</th>
<th>TS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidemic</td>
<td>0.95</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>Mental</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Crisis</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Result 2: Good Qualitative correlation between sentiments and news broadcasts.

Result 3: Negative sentiment correlated better with news broadcasts. (not shown)

3. A Recommendation-based Approach to Assess Individual Condition Risk

Co-morbidity describes conditions that are correlated with each other, for example, in patients with type-2 diabetes, chronic nephropathy often results from diabetic nephropathy. Prediction steps are:

1. Select the neighborhood of patients
2. Aggregate the neighborhood’s likelihood
3. Rank the diseases by their likelihood

Preliminary Results

<table>
<thead>
<tr>
<th>Patient Id</th>
<th>Diagnosed Conditions</th>
<th>Top 2 Predicted Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>296</td>
<td>Migraine, Fibromyalgia</td>
<td>Chronic Fatigue Syndrome, Generalized Anxiety Disorder</td>
</tr>
<tr>
<td>42</td>
<td>Eating Disorder, Pudic disorders</td>
<td>Social Anxiety Disorder, PTSD</td>
</tr>
<tr>
<td>50</td>
<td>HIV, Schizophrenia</td>
<td>Bipolar Disorder, Late Inherence</td>
</tr>
</tbody>
</table>

Conclusions and Working Directions
We presented different components of the Social Health Analytics framework. The working directions are:

- Expand data collection process into batch procedure or real-time.
- Build a pipeline to process twitter sentiment and topic sentiments.
- Explore better similarity measures between patients,
- Use neighborhood pickup function to improve predictions.

Acknowledgment
The work is partially supported by Intelligent Automation Inc. for their summer internship training and by CUNY-PSC grant.

Publications

