Identifying Unproven Cancer Treatments on the Health Web: Addressing Accuracy, Generalizability, and Scalability

Center for Health Informatics and Bioinformatics
NYU Langone Medical Center

Yin Aphinyanaphongs MD, PhD
Lawrence Fu PhD
Constantin Aliferis MD, PhD
Medinfo, Copenhagen, Denmark 08/19/2013
Our Vision: A Safer Health Web

• We envision an health web that provides **safe, reliable, validated** medical information to health consumers.
• Focus on the **lowest quality websites** and building classifiers to identify those.
• Our initial pilot targeted unproven treatments in **cancer**.
Cancer, Health Consumers, and the Health Web

- In the US, 65% of cancer pts searched unproven treatments and 12% purchased at least one unproven treatment.
- 83% of cancer patients used at least one unproven treatment.
- Patients with dire prognosis are more susceptible to using these unproven treatments.

- Routine cancer search topics return sites with non-conventional treatments.
- These sites were of variable quality. 23% discouraged the use of conventional medicine, 15% discouraged adherence to physician advice, and 26% provided anecdotal experiences.
- Schmidt and Ernst found that nearly 10% of their sample had potentially harmful or definitely harmful material.

- Impact of bad advice:
  - Death
  - Financial Costs with No Benefit.
  - Reduction in Efficacy of Known Treatments
  - Delayed Access to Proven Therapies
How is the Health Web policed now?

- Manual efforts
  - Rating agencies (HONCode)
  - Self ratings
  - Operation Cure.all
    - Effort led by Federal Trade Commission in 1997 and 1999 in collaboration with 80 agencies and organizations from 25 countries.
    - Identified approximately 800 to 1200 web sites purporting to cure a variety of diseases.
How Do We Get to a Safer Health Web?

• Automation

• In our prior work, we built models that successfully identified known unproven cancer treatments on the Internet. (where an unproven treatment is known and the goal is to identify new websites that may market the known unproven treatment, Area Under the Curve = 0.93).

• Some questions left to consider:
  • What happens with unknown unproven treatments. (i.e treatments that are not known a priori.) Is it possible to build models that will identify these unknown unproven treatments?
  • What methods may scale a high performing machine learning model to billions of web pages?

• This paper addresses the limitations of this prior work.
Where Machine Learning May Generalize and Scale?

1. Preprocessing

2. Generalizability

3. Model Building

4. Model Application
Solution 1: Map Reduce for Pre-processing

Input data

Intermediate data

Output data

DVD  Blu-ray
CD  DVD
CD  CD  CD
CD  Video
Blu-ray  DVD

Map

Map

Blu-ray : 1  DVD : 1
CD : 1  DVD : 1
CD : 1  CD : 1  CD : 1
CD : 1  Video : 1
Blu-ray : 1  DVD : 1

Reduce

Reduce

Blu-ray : 2
CD : 5
DVD : 3
Video : 1

Image from: http://www.infosun.fim.uni-passau.de/cl/
MapReduceFoundation/
Solution 2: Feature Selection

- Identifying a subset of relevant features for use in model construction that ideally maintain or improve predictivity from a dataset that includes all features.

- Generalized Local Learning Parents Children is one such algorithm that learns a subset of features for use in model construction.
  - Will identify the smallest subset of features that gives optimal classification performance under several assumptions.
Area Under the Curve

- Receive Operating Curves
- Area Under the Curve

Area Under the Curve as a Measure of Performance

Perfect Separation
Area Under the Curve ~ 1.0

Okay Separation
Area Under the Curve ~ 0.8

Random Separation
Area Under the Curve ~ 0.5
Gold Standard

• 8 quack treatments identified by quackwatch.org.
• Applied to Google appending "cancer" and "treatment."
• Top 30 results for each treatment labeled by the authors.

• Two authors labeled 191 out of 240 web pages as making unproven claims or not (Inter-rater Reliability - Kappa 0.76)
• Excluded
  • not found (404 response code) error pages
  • no content pages
  • non-English pages
  • password-protected pages
  • pdf pages
  • redirect pages
  • pages where the actual treatment text does not appear in the document
Build 8 Independent Training/ Testing Sets

<table>
<thead>
<tr>
<th>Category</th>
<th>Train Size</th>
<th>Test Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin Potentiation Therapy</td>
<td>146</td>
<td>18+/ 7-</td>
</tr>
<tr>
<td>Cure for All Cancers</td>
<td>147</td>
<td>16+/ 8-</td>
</tr>
<tr>
<td>Mistletoe</td>
<td>145</td>
<td>8+/ 18-</td>
</tr>
<tr>
<td>Metabolic Therapy</td>
<td>153</td>
<td>11+/ 7-</td>
</tr>
<tr>
<td>Macrobiotic Diet</td>
<td>148</td>
<td>4+/ 19-</td>
</tr>
<tr>
<td>ICTH</td>
<td>162</td>
<td>5+/ 4-</td>
</tr>
<tr>
<td>Krebiozen</td>
<td>151</td>
<td>10+/ 10-</td>
</tr>
<tr>
<td>Cellular Health</td>
<td>154</td>
<td>9+/ 8-</td>
</tr>
</tbody>
</table>
Experimental Design

• 8 fold Leave One Treatment Out Cross Validation

• Classifiers
  • Linear Support Vector Machines
  • L1 Regularized Logistic Regression
  • L2 Regularized Logistic Regression
  • Classifiers optimized over costs and regularization coefficient respectively.

• Text pre-processed by term frequency – inverse document frequency weighting scheme.
### Results – Generalizability to unknown unproven treatments

<table>
<thead>
<tr>
<th>Category</th>
<th>Train Set</th>
<th>Test Set</th>
<th>Linear SVM</th>
<th>L1 Regularized Logistic Regression</th>
<th>L2 Regularized Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin Potentiation Therapy</td>
<td>146</td>
<td>18+/ 7-</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Cure for All Cancers</td>
<td>147</td>
<td>16+/ 8-</td>
<td>0.89</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Mistletoe</td>
<td>145</td>
<td>8+/ 18-</td>
<td>0.90</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>Metabolic Therapy</td>
<td>153</td>
<td>11+/ 7-</td>
<td>0.97</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>Macrobiotic Diet</td>
<td>148</td>
<td>4+/ 19-</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>ICH</td>
<td>162</td>
<td>5+/ 4-</td>
<td>1.0</td>
<td>0.94</td>
<td>1.0</td>
</tr>
<tr>
<td>Krebiozen</td>
<td>151</td>
<td>10+/ 10-</td>
<td>0.86</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>Cellular Health</td>
<td>154</td>
<td>9+/ 8-</td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
</tr>
</tbody>
</table>
## Results – Feature Selection

<table>
<thead>
<tr>
<th>Feature/Classifier Combination</th>
<th>Number of Features</th>
<th>Area Under the Curve Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features/Linear SVM</td>
<td>9,187</td>
<td>0.946*</td>
</tr>
<tr>
<td>Generalized Local Learning – Parents Children</td>
<td>96</td>
<td>0.974*</td>
</tr>
</tbody>
</table>

* - these performances are calculated using 8 fold nested cross validation. Example documents are held out randomly without regards to the selected treatment.
# Computational Performance Point Estimates

<table>
<thead>
<tr>
<th>Corpus Preparation</th>
<th>Number of Documents</th>
<th>Speed (Point estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Threaded</td>
<td>200,000</td>
<td>1,300 seconds</td>
</tr>
<tr>
<td>Hadoop/ Mapreduce</td>
<td>200,000</td>
<td>450 seconds</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Building Performance</th>
<th>Number of Features</th>
<th>Speed (Point estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
<td>9,187</td>
<td>0.20 seconds</td>
</tr>
<tr>
<td>Generalized Local Learning – Parents Children</td>
<td>96</td>
<td>0.0069 seconds</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Application Performance</th>
<th>Number of Documents</th>
<th>Number of Features</th>
<th>Speed (Point Estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
<td>145,000</td>
<td>9,187</td>
<td>98.6 seconds</td>
</tr>
<tr>
<td>Generalized Local Learning – Parents Children</td>
<td>145,000</td>
<td>96</td>
<td>6 seconds</td>
</tr>
</tbody>
</table>
Future Work

• Larger Sample Size.

• Address building classification models in other languages.

• Improved Computational Performance Estimate (not point).

• Address adversarial nature of this application.

• Further explore labeled sample versus ability to identify low quality websites in the full web.

• Address efficient collection of web pages.
Conclusions

• Generalization to unknown unproven treatments.

• Evidence that training in a production environment will need a small number of labeled documents.

• Scalability.
  • Mapreduce for pre-processing.
  • Feature Selection
    • Speed up model building.
    • Speed up model application.
References


Thank you.

• NYU Langone Medical Center.
• Center for Health Informatics and Bioinformatics.

• Center for Translational Sciences Institute
• Grant 1UL1RR029893 from the National Center for Research Resources, National Institutes of Health

• Collaborators
  • Lawrence Fu
  • Constantin Aliferis

• Contact: Yin Aphinyanaphongs (yin.a@nyumc.org).