Application of Data Mining in Health Domain

CS 591 Seminar
Sheng Wang
swang141@illinois.edu
Outline

- Dataset in health domain
- Recent study of data mining in health domain
  - Disease and behavior prediction
  - Social networks
  - Other fundamental data mining problem
- Discussion
Dataset in Health Domain

- Electronic Health Record (EHR)
  - High quality, private data
- Online social networks
  - Large amounts of user generated data
- Health && medical forum
  - Good feedback for doctors and drug producers
- ProMED-mail, World Health Organization
  - Official public healthcare information source
- Search logs
  - Users search health information through search engine
- Pubmed
- TREC medical record
Heterogeneous Information on Electronic Health Record (EHR)

Numerical Data

Text Data

Image Data
EHR VS. Tradition Paper Record

- More organized and complete
- Accessible anytime anywhere
- Faster information retrieval
- Large amount of heterogeneous, high-quality data but not public

Potential research problems on EHR?
- Disease prediction [KDD10, ICDM12]
- Privacy-preserving and efficient access retrieval system [CIKM11]
- Finding similar symptoms or patients [KDD12, SDM12]
- Mining heterogeneous information [KDD12]
Online Social Networks

- People cares about health on Twitter, Facebook, YouTube
- Large amount of user-generated data
- Noisy, fake information, rumor....
- Potential research problem on SNS?
  - Truth analysis on medical information
  - Role discovery and user recommendation
  - Advertisement for healthcare products and drugs
Health and medical forum

Health Forum
- Recommendation
- Advertisement
- Summarization
- QA community

Patients
- Share experience
- Search for medical concern

Doctors
- Receive feedback
- Find similar symptom
Health and medical forum

- Vertical search: more professional
- Users have a stronger medical concern and anxiety in comparison with SNS
- Beneficial for doctors and drug producers
- Potential research problems on health forum?
  - QA community
  - Auto summarization
  - Similar symptom search
  - Advertisement & recommendation
Recent Study

- **Disease and behavior prediction**
  - Stroke Prediction
  - Human Behavior Prediction in a Health Social Network
  - HIV-1 Protease Drug Resistance Prediction

- Social networks

- Other fundamental data mining problem
An integrated machine learning approach for Stroke Prediction

Accurate prediction of stroke is highly valuable for early intervention and treatment.

Desirable framework for general disease prediction problem

<table>
<thead>
<tr>
<th>Input survey data collected between 1989 to 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>796 features</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Missing data estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean, median</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward feature selection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification for prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
</tr>
</tbody>
</table>

A. Khosla et al. KDD10
### Top 12 features of stroke prediction

<table>
<thead>
<tr>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Number of symbols correctly coded</td>
</tr>
<tr>
<td>Maximal inflation level</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>Calculated 100 point score</td>
</tr>
<tr>
<td>Total medications</td>
</tr>
<tr>
<td>Isolated systolic hypertension</td>
</tr>
<tr>
<td>General health</td>
</tr>
<tr>
<td>Calculated hypertension status</td>
</tr>
<tr>
<td>Time to walk 15 feet</td>
</tr>
<tr>
<td>Any ECG abnormality</td>
</tr>
<tr>
<td>Right/Left stenosis</td>
</tr>
</tbody>
</table>

**Red features are not found in previous medical study**

A. Khosla et al. KDD10

More important
## Experiment Result

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCR + CM feature selection</td>
<td>0.777</td>
</tr>
<tr>
<td>SVM + CM feature selection</td>
<td>0.774</td>
</tr>
<tr>
<td>SVM + 16 manually features</td>
<td>0.753</td>
</tr>
<tr>
<td>SVM + forward selection</td>
<td>0.751</td>
</tr>
<tr>
<td>Cox + L1 feature selection</td>
<td>0.747</td>
</tr>
<tr>
<td>Cox + 16 manually features</td>
<td>0.734</td>
</tr>
</tbody>
</table>

AUC is a good metric for imbalance class problem (i.e. disease prediction)

A. Khosla et al. KDD10
Human Behavior Prediction in a Health Social Network

- Could social affiliations affect individual behaviors?
- How can we leverage social network to help predict individual’s behaviors?

\[
\begin{align*}
\text{Past Personal Behavior} & \quad + \quad \text{Social correlation Factors} \\
\text{Time series data } X = (x_1, x_2, \ldots, x_t) & \quad + \quad \text{Social network } G = (V, E) \\
\text{Future Behavior} & = \quad \text{User behavior } x_{t+1} \text{ at time } t+1
\end{align*}
\]

Y. Shen et al. ICDM12
Experiment

- Dataset “YesIWill”
  - 185 users’ daily physical activities in 10 months
  - 684 connections, on average each user has 4 friends
- Predict whether user will do exercise or not in the following days
- Achieve 0.7902 prediction accuracy

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BPS</td>
<td>0.6435</td>
<td>0.6347</td>
<td>0.6965</td>
<td>0.6993</td>
<td>0.6982</td>
<td>0.7058</td>
<td>0.6808</td>
</tr>
<tr>
<td>LAR</td>
<td>0.6946</td>
<td>0.6583</td>
<td>0.7435</td>
<td>0.7367</td>
<td>0.7388</td>
<td>0.7421</td>
<td>0.7096</td>
</tr>
<tr>
<td>SLAR</td>
<td>0.6837</td>
<td>0.6646</td>
<td>0.7421</td>
<td>0.7343</td>
<td>0.7371</td>
<td>0.7432</td>
<td>0.7084</td>
</tr>
<tr>
<td>PGP</td>
<td>0.6707</td>
<td>0.7080</td>
<td>0.7717</td>
<td>0.7428</td>
<td>0.7509</td>
<td>0.7640</td>
<td>0.7402</td>
</tr>
<tr>
<td>SGP</td>
<td>0.6983</td>
<td>0.7376</td>
<td>0.7902</td>
<td>0.7524</td>
<td>0.7583</td>
<td>0.7628</td>
<td>0.7431</td>
</tr>
</tbody>
</table>

Using week 1~ week 20 to predict week 21 and week 22
HIV-1 Protease Drug Resistance Prediction

- Estimate the specific resistance of a given strain of HIV to individual drugs.
- A classification method based on the sparse representation theory.
- Experiment on HIV dataset achieved 0.97 classification accuracy
Prediction framework

Input new genotype v0

drug-resistance strain v1

none-resistance strain v2

none-resistance strain v3

drug-resistance strain v4

drug-resistance dictionary

none-resistance dictionary

Nearer to drug-resistance dictionary?

Yes

No

X. Yu et al. SDM12
Experimenter result

- 10000 HIV structure vectors, half are resistant and half are non-resistant.
- 2000 testing vectors for each drug.
- Four drugs: SQV, TPV, IDV and LPV.
Recent Study

- Disease and behavior prediction
  - Stroke Prediction
  - Human Behavior Prediction in a Health Social Network
  - HIV-1 Protease Drug Resistance Prediction

- Social networks
  - Modeling the lifestyle on Health on Scale
  - Real-time Disease Surveillance using Twitter Data
  - Google Flu Trend

- Other fundamental data mining problem
Modeling the lifestyle on Health through Social Networks

- What did the authors do by using Twitter dataset?
  - Inference health state
  - Model associations among environment, lifestyle and health
  - Health prediction
- 6237 users, each posted more than 100 GPS-tagged tweets during one month.
- All the users located in New York City metropolitan area

A. Sadilek et al. WSDM12
Dataset

Purple circle: majored pollution
White outlines: ZIP-Code boundary
Red spot: sick people
Green spot: healthy people

Web application: [http://fount.in/](http://fount.in/)
No longer available 😞

A. Sadilek et al. WSDM12
Model associations among environment, lifestyle and health

High correlation with health

A. Sadilek et al. WSDM12
Real-time Disease Surveillance using Twitter Data

- Traditional detection of disease outbreak based on collection of patents’ data from sentinel medical practices.
  - **Drawbacks:** 1-2 weeks time lag

- The authors use social network to detect real-time disease outbreak.
Experiment

- Use 6 million flu-related tweets by 3.3 million users in the past 5.5 months to build cancer dataset and flu dataset

K. Lee et al. KDD13
Time series analysis

K. Lee et al. KDD13
Google Flu Trend

- Certain search terms are good indicators of flu activity.
- Google Flu Trends uses aggregated Google search data to estimate current flu activity around the world in near real-time.
Recent Study

- Disease and behavior prediction
  - Human Behavior Prediction in a Health Social Network
  - HIV-1 Protease Drug Resistance Prediction
  - Stroke Prediction

- Social networks
  - Modeling the lifestyle on Health on Scale
  - Real-time Disease Surveillance using Twitter Data
  - Google Flu Trend

- Other fundamental data mining problem
  - Real-time Clinic Monitoring and Deterioration Warning
  - Patient Similarity Assessment
  - Studies of the Onset and Persistence of Medical Concerns in Search Logs
Real-time Clinic Monitoring and Deterioration Warning

• Lots of patients in hospitals could be saved if warning of serious clinical events could be provided early.
• The author use wireless sensing devices to collect real-time vital sign data for patients, then apply data mining approach to real-time clinical monitoring.
• Features for time series data
  • Detrended fluctuation analysis (DFA), Approximate entropy (ApEn), Spectral analysis, First order features, Second order features
  • Linear correlation, coherence
• Exploratory undersampling
  • Iteratively remove those samples that can be correctly classified by a large margin to the class boundary by the existing model

Y. Mao et al. KDD12
An integrated data mining framework

772 patients’ vital signal time series: heart rate & oxygen saturation rate

Preprocess data:
Remove abnormal value

Extract features:
DFA, spectral analysis, first order and second order features, and multi-sign features

Classification:
SVM and logistic regression

Class imbalance:
Exploratory undersampling method

Feature selection:
Forward feature selection

Y. Mao et al. KDD12
### Experiment

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVM</td>
<td>DFA of heart rate (HR)</td>
<td>0.5759</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>DFA of HR</td>
<td>0.4742</td>
</tr>
<tr>
<td>Kernel SVM</td>
<td>DFA of HR</td>
<td>0.5897</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>DFA of oxygen saturation (OS)</td>
<td>0.4473</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>DFA of OS</td>
<td>0.4902</td>
</tr>
<tr>
<td>Kernel SVM</td>
<td>DFA of OS</td>
<td>0.5016</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>DFA of HR+OS</td>
<td>0.5757</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>DFA of HR+OS</td>
<td>0.5370</td>
</tr>
<tr>
<td>Kernel SVM</td>
<td>DFA of HR+OS</td>
<td>0.6332</td>
</tr>
</tbody>
</table>

Kernel SVM > linear SVM: Sparse feature clinic data set is better separated by a nonlinear classifier. The combination of both time series greatly improves the performance
Identifying leading risk factors

<table>
<thead>
<tr>
<th>Feature selection</th>
<th>Logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of heart rate</td>
<td>Local homogeneity of heart rate</td>
</tr>
<tr>
<td>Apen of heart rate</td>
<td>Standard deviation of oxygen saturation</td>
</tr>
<tr>
<td>Energy of oxygen saturation</td>
<td>Entropy of oxygen saturation</td>
</tr>
<tr>
<td>LF of oxygen saturation in SPA</td>
<td>Low frequency of heart rate</td>
</tr>
<tr>
<td>LF of heart rate in SPA</td>
<td>Local homogeneity of oxygen saturation</td>
</tr>
<tr>
<td>DFA of oxygen saturation</td>
<td>Low frequency of oxygen saturation</td>
</tr>
<tr>
<td>Mean of heart rate</td>
<td>Mean of oxygen saturation</td>
</tr>
<tr>
<td>Inertia of heart rate</td>
<td></td>
</tr>
<tr>
<td>Homogeneity of heart rate</td>
<td></td>
</tr>
</tbody>
</table>

Y. Mao et al. KDD12
Patient Similarity Assessment

- Each doctor has his individual patient similarity metrics based on his own experience
- Integrate distance metrics obtained from individual expert into a single optimal objective metric
  - Knowledge sharing
  - Efficient collaboration
  - Lower the risk of disclosing private data
- Tow case study on real-world clinic data

F. Wang et al. SDM11
Composite distance integration

- Neighborhood construction
  - **Homogenous neighborhood:** Nearest patient with the same labels.
  - **Heterogeneous neighborhood:** Nearest patient with different labels.

F. Wang et al. SDM11
Composite distance integration

- Loss function: Make the data set in same class compact while the data in different class diverse locally.

Yellow circle: patient vector
Red square: heterogeneous neighbor
Blue circle: homogeneous neighbor

(a) Neighborhood in the original space  (b) Neighborhood in the projected space

F. Wang et al. SDM11
Case study on clinical data

- A real health network containing 135K patients over one year.
- All the patients are assigned to 247 doctors (metrics).
- Whether or not each patient has diabetes as the patient labels.

F. Wang et al. SDM11
Figure 9: Classification performance comparison with different measurements on our data set with HCC019.

F. Wang et al. SDM13
Studies of the Onset and Persistence of Medical Concerns in Search Logs

- Interaction with web content may **heighten anxiety** and stimulate healthcare utilization.
- The authors use search logs to characterize how users focuses on particular medical concerns, how concerns persist and influence future behavior.
- Dataset: 170,000 users, 25 million search sessions, over three month from Google, Yahoo! and Bing.
- On average, each user issued 10-15 queries per day. 3% are medically related.

R. White et al. SIGIR12
Distributions of medical queries

- Serious illnesses and symptoms are the most popular concern types.
- However, the benign explanations of symptoms are typically much more likely given prevalence statistics.

Figure 1. Distribution of medical queries per concern type. Queries are labeled using the wordlist method.

R. White et al. SIGIR12
Distributions of medical queries

- Prior work examined within-session transition, where users were observed to transition from searches on common symptoms to serious illness.
- This work examined between session transition.

Figure 2. Examples of different escalation types.

R. White et al. SIGIR12
Prediction task

- Onset: transition from symptom searching to condition searching
- Escalation: escalate between-session to a particular condition
- Interruption: interrupted by the condition post-onset
- Accurate prediction capabilities afford search engine to better help medical searchers

Figure 4. ROC curves for each prediction task. Performance of the marginal model shown as dotted line.

R. White et al. SIGIR12


Harrison, Robert; Weber, Irene; Yu, Xiaxia. Sparse Representation for HIV-1 Protease Drug Resistance Prediction. SDM 2012


Discussion

- Challenge in health domain
  - Data and privacy
  - Finding a specific research problem (both solvable and challenge)

- Future of medical health domain
  - Can we learn innovative knowledge from the data?
    - It is hard to fight with certain disease only through mining the data.
  - Can we achieve personalized service?
    - People may use search engine lots of time, but they don’t go to hospital and provide health information so many times.
  - How to fight with less common disease?
    - Common disease (i.e. heart failure) has more dataset, but less common disease has less dataset and hard to generate rules and knowledge from them.