Scalable Predictive Modeling Methods for Healthcare Applications

Jimeng Sun

jsun@cc.gatech.edu
Jimeng Sun

1 Research Scientist: M. Bahadori

8 PhD students: R. Chen, I. Perros, H. Su, E. Choi, Z. Wu, K. Malhotra, Y. Zhang, S. An

many MS/undergrad students

Collaborators & Sponsors

Government

Provider

University

Company

jsun@cc.gatech.edu  sunlab.org
Research in SunLab

Predictive Modeling using Electronic health records (EHR)

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
PREDICTIVE MODELING using ELECTRONIC HEALTH RECORDS (EHR)

Explosion in interest

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Requirements: Healthcare Predictive Modeling

- Parallel predictive modeling pipeline
- Nonnegative tensor factorization for phenotyping
- Deep learning for heart failure onset prediction

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
RESEARCH PORTFOLIO

Application
- Predictive Modeling
- Computational Phenotyping
- Patient Similarity
- Disease progression
- Public health informatics

Algorithm
- Big data system
- Tensor factorization
- Distance metric learning
- Deep learning
- Search & recommendation

Jimeng Sun, jsun@cc.gatech.edu, sunlab.org
SpaceShip
PARALLEL PREDICTIVE MODELING PLATFORM

Hang Su  Yuyu Zhang  Sungtae An  Kunal Malhotra

Jimeng Sun jsun@cc.gatech.edu,  sunlab.org
Epilepsy Treatment Prediction

Can you predict which treatment will work?

Early Responders – 32%
Achieved seizure freedom with first drug

Late Responders – 24%
Achieved seizure freedom after 2 to 5 years

Non-Responders – 44%
Continued seizures despite treatment


Chris Clark

Recommend the right treatment earlier

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Performance Comparison in Treatment Recommendation

10% to 160% improvement over human

- Accuracy in terms of positive predictive value (PPV) aka precision
- Drug works = treatment stability

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Challenges for Epilepsy Treatment Recommendation

**Big Data**

- # of patients: 30 million
- Diagnosis Claims: (2010-2013), 12 billion records
- Medication Claims: (2010-2013), 2.3 billion records
- Inpatient claims: 2.7 million
- Outpatient claims: 8.3 million
- Number of anti-epilepsy drugs = 27
- 300 GB data

**Many models**

- Over 100 different treatment combinations
- Potentially multiple outcomes
  - Stable treatment
  - Hospitalization
- Many modeling choices

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Predictive Modeling Pipeline

Many models to be built and evaluated
- Different patient cohorts
- Different targets
- Different features
- Different algorithms
- Multiple training and testing splits in cross-validation

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
**Keys:** Efficient Computation over Pipelines

- Parallel computing
  - dependency graph
  - parallel execution
  - cloud computing engine

- Smart Sampling
  - Two-level Bayesian optimization
    - algorithm selection
    - parameter tuning

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Parallel Computation on Dependency Graph

Start -> Cohort Construction

- Feature Construction: Medication
- Feature Construction: Diagnosis
- Feature Construction: Demographics
- Feature Construction: Hospitalization
- Feature Construction: Procedures

Cross Validation

Feature Selection: InfoGain
Feature Selection: Fisher Score

Feature Construction

Classification: Random Forest
Classification: Log Regression
Classification: Naive Bayes

Output

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Runtime for Predictive modeling (38K patients)

Parallel: 10 minutes
Sequential: 60x faster

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Smart Sampling via Bayesian Optimization

Feature Construction: Medication, Diagnosis, Demographics, Hospitalization, Procedures

Feature Selection: InfoGain, Fisher Score

Classification: Random Forest, Log Regression, Naïve Bayes

Output

Start → Cohort Construction → Feature Construction → Feature Selection → Classification → Output

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Smart Sampling via Bayesian Optimization

Y. Zhang, M. Bahadori, H. Su, J. Sun. FLASH: Fast Bayesian Optimization for Data Analytic Pipelines. KDD’16

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Limestone

How to handle high-dimensionality of EHR data

UNSUPERVISED PHENOTYPE GENERATION VIA TENSOR FACTORIZATION

COMPUTATIONAL PHENOTYPING

Demographics
Diagnoses
Medications
Clinical notes
Procedures
Lab result

Raw Data → Phenotyping → Medical Concepts (Phenotypes)

A phenotype = a cluster of patients that share similar characteristics

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
PHENOTYPING ALGORITHM

Type 1 Diagnoses

Type 2 Diagnoses

Type 2 Rx

Abnormal Lab

EHR

Type 1 Rx

Type 2 Rx

Type 2 Rx

Type 2 Rx Precedes Type 1 Rx

Case

Type 2 Diabetes Cases

Type 2 Diabetes Cases by phys. >= 2

YES

NO

YES

NO

YES

NO

YES

YES

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
USE TENSOR TO MODEL EHR DATA

- Tensor is a generalization of matrix

**Data element types:**
- Binary
- Count (integer)
- Continuous (numeric)

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
PHENOTYPING THROUGH NONNEGATIVE TENSOR FACTORIZATION

Phenotype importance $\approx \lambda_1$

Medication factor

Diagnosis factor

Patients factor

Phenotype 1

$\lambda_R$

Factor elements sum to 1

Elements sum to 1

Phenotype $R$

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
A phenotype = a group of patients that share common characteristics

Diagnosis factor

Medication factor

Candidate Phenotype $k$
(40% of patients)

<table>
<thead>
<tr>
<th>Hypertension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta Blockers Cardio-Selective</td>
</tr>
<tr>
<td>Thiazides and Thiazide-Like Diuretics</td>
</tr>
<tr>
<td>HMG CoA Reductase Inhibitors</td>
</tr>
</tbody>
</table>

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
**CP ALTERNATING POISSON REGRESSION (CP-APR)**

- Nonnegative input tensor
- Nonnegative constraints
- Stochastic column constraints on factor matrices

\[
\min f(\mathcal{M}) \equiv \sum_{\hat{i}} m_{\hat{i}} - x_{\hat{i}} \log m_{\hat{i}}
\]

s.t \( \mathcal{M} = [\lambda; A^{(1)}; \ldots; A^{(N)}] \in \Omega \)

\[\Omega = \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N\]

\[\Omega_\lambda = [0, +\infty)^R\]

\[\Omega_n = \{ A \in [0, 1]^{I_n \times R} \mid \| a_r \|_1 = 1 \ \forall r \}\]


Jimeng Sun {jsun@cc.gatech.edu, sunlab.org}
Limestone

- Nonnegative input tensor
- Nonnegative constraints
- Stochastic column constraints on factor matrices
- Hard thresholding on elements in factor matrices

\[
\min f(\mathcal{M}) = \min \sum_i [m_i - x_i \log m_i] + \gamma \sum_{j,n,r} 1\{a^{(n)}_{jr} > 0\}
\]

s.t. \( \mathcal{M} = [\lambda; A^{(1)}; \ldots; A^{(N)}] \in \Omega \)

\( \Omega = \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N \)

\( \Omega_\lambda = [0, +\infty)^R \)

\( \Omega_n = \{A \in [0, 1]^{I_n \times R} \mid \|a_r\|_1 = 1 \ \forall r\} \)

EXPERIMENTS OF LIMESTONE

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
**Tensor Construction**

- Medication orders from Geisinger dataset
- 31,816 patients \( \times \) 169 diagnose categories \( \times \) 471 medication classes

---

**Diagnosis**
- Hypertension
- Diabetes
- Sulfonylureas
- Coronary Atherosclerosis
- Beta Blockers
- Cardio-selective
- Loop Diuretics
- Congestive Heart Failure

**Medication**
- ACE Inhibitors
- Nitrates
- Beta Blockers
- Cardio-selective
- Loop Diuretics

**Index Date**

**Time**
- \( t_0 \)
- \( t_1 \)
- \( t_2 \)
- \( t_3 \)
- \( t_4 \)

**Observation Window = 2 years**

---

Jimeng Sun [jsun@cc.gatech.edu](mailto:jsun@cc.gatech.edu), [sunlab.org](http://sunlab.org)
HEART FAILURE (HF) PREDICTION

- **Task**: predict patients with HF
- **Model**: logistic regression with $\ell_1$ regularization
- **Evaluation**: Cross-validation

- **Methods for feature construction**:
  1. Baseline using source independence matrix
  2. Principal Component Analysis (PCA)
  3. Nonnegative Matrix Factorization (NMF)
  4. Limestone
**PREDICTIVE PERFORMANCE**

*Small # of features outperforms 640 features*

![Graph showing predictive performance with different features and number of phenotypes.]

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
## MAJOR DISEASE PHENOTYPES

<table>
<thead>
<tr>
<th>Uncomplicated Diabetes</th>
<th>Mild Hypertension</th>
<th>Chronic Respiratory Inflammation/Infection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phenotype 3</strong></td>
<td><strong>Phenotype 4</strong></td>
<td><strong>Phenotype 5</strong></td>
</tr>
<tr>
<td>(17.6% of patients)</td>
<td>(31.1% of patients)</td>
<td>(36.7% of patients)</td>
</tr>
<tr>
<td>Diabetes with No or Unspecified Complications</td>
<td>Hypertension</td>
<td>Other Ear, Nose, Throat, and Mouth Disorders</td>
</tr>
<tr>
<td>Sulfonylureas</td>
<td>ACE Inhibitors</td>
<td>Viral and Unspecified Pneumonia, Pleurisy</td>
</tr>
<tr>
<td>Biguanides</td>
<td>Thiazides and Thiazide-Like Diuretics</td>
<td>Significant Ear, Nose, and Throat Disorders</td>
</tr>
<tr>
<td>Diagnostic Tests</td>
<td></td>
<td>Cough/Cold/Allergy Combinations</td>
</tr>
<tr>
<td>Insulin Sensitizing Agents</td>
<td></td>
<td>Azithromycin</td>
</tr>
<tr>
<td>Diabetic Supplies</td>
<td></td>
<td>Fluoroquinolones</td>
</tr>
<tr>
<td>Meglitinide Analogues</td>
<td></td>
<td>Sympathomimetics</td>
</tr>
<tr>
<td>Antidiabetic Combinations</td>
<td></td>
<td>Penicillin Combinations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antitussives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glucocorticosteroids</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tetracyclines</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anti-infective Misc. - Combinations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clarithromycin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cephalosporins - 2nd Generation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cephalosporins - 1st Generation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expectorants</td>
</tr>
</tbody>
</table>

Jimeng Sun [jsun@cc.gatech.edu](mailto:jsun@cc.gatech.edu), [sunlab.org](http://sunlab.org)
DISEASE SUBTYPES CAN BE IDENTIFIED

<table>
<thead>
<tr>
<th>Mild Hypertension</th>
<th>Moderate Hypertension</th>
<th>Severe Hypertension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype 4</td>
<td>Phenotype 2</td>
<td>Phenotype 6</td>
</tr>
<tr>
<td>(31.1% of patients)</td>
<td>(31.5% of patients)</td>
<td>(24.3% of patients)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Hypertension</td>
<td>Hypertension</td>
</tr>
<tr>
<td>ACE Inhibitors</td>
<td>Beta Blockers</td>
<td>Calcium Channel</td>
</tr>
<tr>
<td>Thiazides and</td>
<td>Cardio-Selective</td>
<td>Blockers</td>
</tr>
<tr>
<td>Thiazide-Like</td>
<td>Antihypertensive</td>
<td>Antihypertensive</td>
</tr>
<tr>
<td>Diuretics</td>
<td>Combinations</td>
<td>Combinations</td>
</tr>
<tr>
<td></td>
<td>Angiotensin II</td>
<td>Antiadrenergic</td>
</tr>
<tr>
<td></td>
<td>Receptor Antagonists</td>
<td>Antihypertensives</td>
</tr>
<tr>
<td></td>
<td>Loop Diuretics</td>
<td>Potassium Sparing</td>
</tr>
<tr>
<td></td>
<td>Potassium</td>
<td>Diuretics</td>
</tr>
<tr>
<td></td>
<td>Nitrates</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alpha-Beta Blockers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vasodilators</td>
<td></td>
</tr>
</tbody>
</table>

Over 80% phenotype factors are clinically meaningful
Limestone

Advantages

• Unsupervised
• Intuitive phenotypes
• Predictive

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
USING RECURSIVE NEURAL NETWORK MODELS FOR EARLY DETECTION OF HEART FAILURE ONSET

How to model temporal relations in the EHR data

Edward Choi  
Andy Schuetz  
Buzz Stewart

MOTIVATIONS FOR EARLY DETECTION OF HEART FAILURE

Heart failure is a complex disease.

Improves existing clinical guidelines of HF prevention.

Reduces cost and hospitalization.

Early intervention can slow down disease progression.
PREDICTIVE MODELING PIPELINE

1. Prediction Target

2. Cohort Construction

3. Feature Construction

4. Feature Selection

5. Predictive Model

6. Performance Evaluation

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Deep Learning Pipeline

1. Prediction Target

2. Cohort Construction

3. Feature Construction

4. Feature Selection

5. Predictive Model

6. Performance Evaluation

Deep learning:

Representation learning

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Input: Longitudinal EHR

Benzonatate  Pneumonia  Amoxicillin

Cough  Fever  Chest X-ray

Time

0  0  1  0  0  0  0  0
0  0  0  0  0  0  0  0
0  0  0  0  0  1  0  0
..  ..  ..  ..  ..  ..  .. ..
..  ..  ..  ..  ..  ..  .. ..
0  0  1  0  0  0  0  1
0  1  0  0  0  0  0  0
1  0  0  0  0  0  0  0

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Representation learning: Word2vec

Input at time $t$: $x_t$

(a) One-hot encoding

- Bronchitis: $[1, 0, 0, 0, 0, ..., 0, 0, 0]$
- Pneumonia: $[0, 1, 0, 0, 0, ..., 0, 0, 0]$
- Obesity: $[0, 0, 1, 0, 0, ..., 0, 0, 0]$
- Cataract: $[0, 0, 0, 0, 0, ..., 0, 0, 1]$

(b) Word2vec vectors

- Bronchitis: $[0.4, -0.2, ..., 0.2]$
- Pneumonia: $[0.3, -0.3, ..., 0.1]$
- Obesity: $[-0.7, 1.4, ..., 1.2]$
- Cataract: $[1.2, 0.8, ..., 1.5]$
Temporal model: RNN

- $x_t$: one-hot coded Dx, Rx, Proc at time $t$
- $h_t$: hidden state at time $t$
- $y$: binary outcome of HF prediction
- $T$: total length of the medical codes
- Red box: a single unit of RNN

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
EVALUATION

- **Models**
  - Baselines: logistic regression, SVM, MLP, KNN
  - Training data: one-hot, grouped, Skip-gram

- **Performance measure**
  - AUC of 6-fold cross validation

- **Vary Observation window & prediction window**
PREDICTION PERFORMANCE OF RNN

- **RNN model achieves over 10% improvement on AUC**

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
PREDICTION PERFORMANCE OF RNN

- RNN model achieves over 10% improvement on AUC
- **Representation matters**

Jimeng Sun [jsun@cc.gatech.edu](mailto:jsun@cc.gatech.edu), [sunlab.org](http://sunlab.org)
PREDICTION PERFORMANCE OF RNN

- RNN model achieves over 10% improvement on AUC
- *Data rep. (word2vec) > knowledge rep. (medical groupers)*

Jimeng Sun jsun@cc.gatech.edu, sunlab.org
Summary: Healthcare Predictive Modeling

- Parallel predictive modeling pipeline
- Nonnegative tensor factorization for phenotyping
- Deep learning for heart failure onset prediction

Jimeng Sun jsun@cc.gatech.edu, sunlab.org