Healthcare Data Mining with Matrix Models

KDD 2016 Tutorial Part II
August 13th, 2016

Ping Zhang
Center for Computational Health
IBM T.J. Watson Research Center
pzhang@us.ibm.com
Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping
Omics technologies in biomedicine

The Cancer Genome Atlas Pan-Cancer analysis project

Data integration from multiple heterogeneous sources

How to combine different measurements?

Issues:
• Large number of measurements, small sample sizes (p>>n)
• Need to integrate common and complementary information
• Not all measurements can be normalized and mapped to the same unit
Similarity network fusion

Step 1. Construct a similarity network for each data source

Step 2. Integrate networks using data fusion method

Construct similarity networks (1)

Patient similarity:

\[ W(i, j) = \exp\left( \frac{\rho(x_i, x_j)^2}{\eta \xi_{ij}^2} \right) \]

Adjacency matrix:

\[ P(i, j) = \frac{W(i, j)}{\sum_{k \in V} W(i, k)} \]

Sparsification

1) \( W(i, j) = \begin{cases} W(i, j) & \text{if } x_j \in KNN(x_i) \\ 0 & \text{otherwise} \end{cases} \)

2) \[ S(i, j) = \frac{W(i, j)}{\sum_{x_k \in KNN(x_i)} W(i, k)} \]
Construct similarity networks (2)
Combine networks (1)

Sample Similarity Networks

Fusion

\[ \mathbf{p}_{t+1}^{(1)} = \mathbf{s}^{(1)} \times \mathbf{p}_{t}^{(2)} \times (\mathbf{s}^{(1)})^T \]

\[ \mathbf{p}_{t+1}^{(2)} = \mathbf{s}^{(2)} \times \mathbf{p}_{t}^{(1)} \times (\mathbf{s}^{(2)})^T \]

Can also be extended to more than 2 data types
Combine networks (2)

Sample Similarity Networks → Fusion → Fused Similarity Network

![Diagram showing the process of combining networks](image)

- **Patient similarity:**
  - mRNA-based
  - DNA Methylation-based
  - Supported by all data

\[ \frac{||W_{t+1} - W_t||}{||W_t||} \leq 10^{-6} \]
Case study: glioblastoma multiforme (GBM)

1491 genes

12042 message genes

534 miRNA
Clinical properties of the subtypes
Biological characterization of the subtypes
From subtype-based to network-based outcome prediction
Comparisons on an METABRIC breast cancer data

### Cox objective

\[
l_p(z) = \sum_{i=1}^{n} \delta_i \left( x_i^T z - \log \left( \sum_{j \in R(t_i)} \exp(x_j^T z) \right) \right)
\]

### Network-regularized objective

\[
l_p(z) = \sum_{i=1}^{n} \delta_i \left( x_i^T z - \log \left( \sum_{j \in R(t_i)} \exp(x_j^T z) \right) \right) - \lambda \sum_{i,j} (x_i^T z - x_j^T z)^2 w_{ij}
\]

Incorporate fused patient network structure

CNV and expression data

**Discovery:** 997 patients, **Validation:** 995 patients

<table>
<thead>
<tr>
<th></th>
<th>PAM50 (5 clusters)</th>
<th>iCluster (10 clusters)</th>
<th>SNF (5 clusters)</th>
<th>SNF (10 clusters)</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value discovery cohort</td>
<td>3.0 x 10^{-9}</td>
<td>1.2 x 10^{-14}</td>
<td>6.10 x 10^{-11}</td>
<td>3.31 x 10^{-12}</td>
<td>–</td>
</tr>
<tr>
<td>P value validation cohort</td>
<td>1.7 x 10^{-9}</td>
<td>2.9 x 10^{-11}</td>
<td>5.12 x 10^{-13}</td>
<td>7.86 x 10^{-12}</td>
<td>–</td>
</tr>
<tr>
<td>CI discovery cohort</td>
<td>0.560</td>
<td>0.621</td>
<td>0.638</td>
<td>0.638</td>
<td>0.720</td>
</tr>
<tr>
<td>CI validation cohort</td>
<td>0.551</td>
<td>0.605</td>
<td>0.633</td>
<td>0.633</td>
<td>0.706</td>
</tr>
</tbody>
</table>
Summary of patient networks framework

• Creates a unified view of patients based on multiple heterogeneous sources
• Integrates gene and non-gene based data
• Robust to different types of noise
• Obtain superior results on regular tasks such as subtyping and outcome prediction
• Scalable

Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping
The Challenge of Drug Discovery

Drug repositioning

Drug repositioning (also known as Drug repurposing, Drug re-profiling, Therapeutic Switching and Drug re-tasking) is the application of known drugs and compounds to new indications (i.e., new diseases).

<table>
<thead>
<tr>
<th>Drug</th>
<th>Original indication</th>
<th>New indication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viagra</td>
<td>Hypertension</td>
<td>Erectile dysfunction</td>
</tr>
<tr>
<td>Wellbutrin</td>
<td>Depression</td>
<td>Smoking cessation</td>
</tr>
<tr>
<td>Thalidomide</td>
<td>Antiemetic</td>
<td>Multiple Myeloma</td>
</tr>
</tbody>
</table>

The repositioned drug has already passed a significant number of toxicity and other tests, its safety is known and the risk of failure for reasons of adverse toxicology are reduced.
Shorter timelines & less risk

Drug Resources and Disease Resources

Drug

Chemical Structure
Target Proteins
Side-effect Keywords

Calculate drug/disease similarities

Disease

Phenotype/Symptom
Ontology
Disease Gene
Algorithm Flowchart of JMF

Inputs:
- drug chemical structure similarity network
- drug target protein similarity network
- drug side effect similarity network

Outputs:
- known drug-disease associations

A unified computational framework for drug repositioning hypothesis generation

Outputs:
1. predicted additional drug-disease associations
2. interpretable importance of different information sources
3. latent drug and disease groups as by-products
JMF as an optimization problem

Notations and symbols of the methodology

- $D_k$: n×n, The $k$-th drug similarity matrix
- $S_l$: m×m, The $l$-th disease similarity matrix
- $U$: n×C_D, Drug cluster assignment matrix
- $V$: m×C_S, Disease cluster assignment matrix
- $\Lambda$: C_D×C_S, Drug-disease cluster relationship matrix
- $R$: n×m, Observed drug-disease association matrix
- $\Theta$: n×m, Densified estimation of $R$
- $\omega$: K_d×1, Drug similarity weight vector
- $\pi$: K_s×1, Disease similarity weight vector

- We aim to analyze the drug-disease network by minimizing the following objective:

  $$J = J_0 + \lambda_1 J_1 + \lambda_2 J_2$$

- The reconstruction loss of observed drug-disease associations:

  $$J_0 = \| \Theta - U\Lambda V^T \|_F^2$$

  Similar Drugs/diseases (latent groups) have similar behaviors

- The reconstruction loss of drug similarities:

  $$J_1 = \sum_{k=1}^{K_d} \omega_k \| D_k - UU^T \|_F^2 + \delta_1 \| \omega \|_2^2$$

- The reconstruction loss of disease similarities:

  $$J_2 = \sum_{l=1}^{K_s} \pi_l \| S_l - VV^T \|_F^2 + \delta_2 \| \pi \|_2^2$$

- Putting everything together, we obtained the optimization problem to be resolved:

  $$\min_{U,V,\Lambda,\Theta,\omega,\pi} J, \text{ subject to } U \geq 0, V \geq 0, \Lambda \geq 0, \omega \geq 0, \omega^T 1 = 1, \pi \geq 0, \pi^T 1 = 1, P_\Theta(\Theta) = P_\Omega(R)$$
BCD approach for solving the problem

- **Block Coordinate Descent (BCD) strategy:** The BCD approach works by solving the different groups of variables alternatively until convergence. At each iteration, it solves the optimization problem with respect to one group of variables with all other groups of variables fixed.

Algorithm 1: A BCD Approach for Solving Problem (11)

1. **Require:** $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\delta_1 \geq 0$, $\delta_2 \geq 0$, $K_d > 0$, $K_s > 0$, $\{D_k\}_{k=1}^{K_d}$, $\{S_i\}_{i=1}^{K_s}$, $R$

2. Initialize $\omega = (1/K_d) \mathbf{1} \in \mathbb{R}^{K_d \times 1}$, $\pi = (1/K_s) \mathbf{1} \in \mathbb{R}^{K_s \times 1}$

3. Initialize $U$ and $V$ by performing Symmetric Nonnegative Matrix Factorization on $\tilde{D} = \sum_{k=1}^{K_d} \omega_k D_k$ and $\tilde{S} = \sum_{i=1}^{K_s} \pi_i S_i$.

4. **while** Not Converge **do**

5. Solve $\Theta$ as described in section 2 (as a **constrained Euclidean projection**)

6. Solve $\omega$ and $\pi$ as described in section 3 (as a **standard Euclidean projection onto a simplex**)

7. Solve $\Lambda$ as described in section 4 (as a **nonnegative quadratic optimization problem**)

8. Solve $U$ as described in section 5 (as a **nonnegative quadratic optimization problem**)

9. Solve $V$ as described in section 6 (as a **nonnegative quadratic optimization problem**)

**Closed-form solution**

**Solved by Projected Gradient Descent (PGD) method**

Computational complexity is $O(Rrmn)$, where $R$ is the number of BCD iterations, and $r$ is the average PGD iterations when updating $\Lambda$, $U$, and $V$. 
Data Description

• Benchmark dataset was extracted from NDF-RT, spanning 3,250 treatment associations between 799 drugs and 719 diseases

• **Three** $799 \times 799$ matrices were used to represent drug similarities between 799 drugs from different perspectives

• **Three** $719 \times 719$ matrices were used to represent disease similarities between 719 human diseases from different perspectives
ROC comparisons of five drug repositioning approaches
Distribution of weights of the similarity weight vectors obtained by JMF

(a) Drug similarity weight vector
(b) Disease similarity weight vector
Top 10 drugs for diseases Alzheimer's Disease (AD) and Systemic Lupus Erythematosus (SLE)

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction Score</th>
<th>Clinical Evidence?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selegiline*</td>
<td>0.7091</td>
<td>—</td>
</tr>
<tr>
<td>Carbidopa</td>
<td>0.6924</td>
<td>No</td>
</tr>
<tr>
<td>Amantadine</td>
<td>0.6897</td>
<td>No</td>
</tr>
<tr>
<td>Procyclidine</td>
<td>0.6826</td>
<td>No</td>
</tr>
<tr>
<td>Valproic Acid*</td>
<td>0.6745</td>
<td>—</td>
</tr>
<tr>
<td>Metformin</td>
<td>0.6543</td>
<td>Yes</td>
</tr>
<tr>
<td>Bexarotene</td>
<td>0.6426</td>
<td>Yes</td>
</tr>
<tr>
<td>Neostigmine</td>
<td>0.6385</td>
<td>No</td>
</tr>
<tr>
<td>Galantamine*</td>
<td>0.6348</td>
<td>—</td>
</tr>
<tr>
<td>Nilvadipine</td>
<td>0.6159</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Repositioning candidates

* denotes the drug is known and approved to treat the disease

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction Score</th>
<th>Clinical Evidence?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desoximetasone</td>
<td>0.7409</td>
<td>No</td>
</tr>
<tr>
<td>Azathioprine*</td>
<td>0.7269</td>
<td>—</td>
</tr>
<tr>
<td>Leflunomide</td>
<td>0.7078</td>
<td>Yes</td>
</tr>
<tr>
<td>Fluorometholone</td>
<td>0.7054</td>
<td>No</td>
</tr>
<tr>
<td>Triamcinolone*</td>
<td>0.6862</td>
<td>—</td>
</tr>
<tr>
<td>Beclomethasone</td>
<td>0.6522</td>
<td>No</td>
</tr>
<tr>
<td>Etodolac</td>
<td>0.6445</td>
<td>No</td>
</tr>
<tr>
<td>Hydroxychloroquine*</td>
<td>0.6374</td>
<td>—</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>0.6371</td>
<td>Yes</td>
</tr>
<tr>
<td>Mercaptopurine</td>
<td>0.6150</td>
<td>No</td>
</tr>
</tbody>
</table>
Summary of joint matrix factorization framework

• We proposed a general computational framework, to explore drug-disease associations from multiple drug/disease sources

• Our method could help generate drug repositioning hypotheses, which will benefit patients by offering more effective and safer treatments

• The computational framework and its solution can be used in other applications (gene-disease, drug-patient, etc.)

Next: Multi-channel detailed computational hypothesis generation
And even beyond the hypothesis generation...

Validation methods are increasingly commoditized.
Recent Applications in Biomedicine

• Similarity Network Fusion and Identification of Cancer Subtypes
• Joint Matrix Factorization and Drug Repositioning
• Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
• Tensor Factorization and Patient Phenotyping
Matrix Tri-Factorization

Simultaneous Matrix Tri-Factorization

Collective tri-factorization of matrices A-B and A-E

Reconstructed matrix A-B

Reconstructed matrix A-E
Data Fusion by Simultaneous Matrix Tri-Factorization

Input to data fusion

Simultaneous Constrained Decomposition

\[ \min_{G \geq 0} J(G; S) = \sum_{R_{ij} \in \mathcal{R}} \left( \| R_{ij} - G_i S_{ij} G^T_j \|_2^2 + \right. \]

\[ \left. + \sum_{t=1}^{\max_i t_i} \text{tr}(G^T \Theta^{(t)} G), \right\]

Repeat until convergence:
- Fix G, update S
- Fix S, update G

Zitnik M, Zupan B. Data Fusion by Matrix Factorization. PAMI 2015.
Liver and Drug-Induced Liver Injury (DILI)

- Alcohol
- Environmental Chemicals
- Foods, Nutrients, etc.
- Disease

**DRUGS**

- “Approved drugs are the most common cause of acute liver failure in the USA” - FDA
- DILI is the MOST frequent reason for drug withdrawal during drug discovery, clinical trials, and after drugs are approved for the marketplace
CAMDA 2012 Task: DILI Prediction

- CAMDA: Critical Assessment of Massive Data Analysis
- The Japanese Toxicogenomics Project (TGP) creates a gene expression database using the Affymetrix GeneChip arrays to measure the effects of 131 chemicals, mainly medical drugs, on the liver.
- DILI potential has been categorized as severe, moderate, or mild.
Data Fusion of Additional Sources

Histological and clinical chemistry data (Rat, in vivo)

Blood Chemistry
- ALP, CI, TC, Ca, TG, IP, PL, TP, TBL, RALB, DBIL, A/G GLC, AST (GOT), BUN, ALT (GPT), CRE, LDH, Na, gamma-GTP, K

Liver Weight
- Terminal body weight
- Liver weight, Relative liver weight

Drug information from DrugBank

Chemical Structure

Drug Interactions
- The metabolism of Tacrine, a CYP1A2 substrate, may be reduced by strong CYP1A2 inhibitors such as Ketoconazole. Consider modifying therapy to avoid Tacrine toxicity. Monitor the efficacy and toxicity of Tacrine if Ketoconazole is initiated, discontinued or if the dose is changed.

Drug Targets

Protein-protein interactions (PPI) from STRING

Gene Ontology (GO)
Given the aim to predict DILI potential in humans:

- Animal studies may be replaced with in vitro assays (AUC = 0.799)
- Liver injury in humans can be predicted from animal data (AUC = 0.811)
- animal in vivo > animal in vitro ≈ human in vitro

Recent Applications in Biomedicine

- Similarity Network Fusion and Identification of Cancer Subtypes
- Joint Matrix Factorization and Drug Repositioning
- Data Fusion by Simultaneous Matrix Tri-Factorization and Drug-Induced Liver Injury Prediction
- Tensor Factorization and Patient Phenotyping
Phenotyping from Electronic Medical Records (EMR)

Phenotype (American Heritage Dictionary)
• The *observable* physical or biochemical *characteristics* of an organism, as determined by both genetic makeup and environmental influences.

Why phenotyping from EMR
• Mapping *noisy, incomplete*, and potentially *inaccurate* patient representation from EMR to meaningful medical concepts Feature engineering
• Extracting clinical meaningful groups of patients from EMR Cohort generation

<table>
<thead>
<tr>
<th>Diabetes Phenotype</th>
<th>Heart Failure Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases of other endocrine glands</td>
<td>Other forms of heart disease</td>
</tr>
<tr>
<td>Complications of surgical and medical care</td>
<td>Complications of surgical and medical care</td>
</tr>
<tr>
<td></td>
<td>Symptoms</td>
</tr>
<tr>
<td></td>
<td>Cardiovascular Procedures</td>
</tr>
<tr>
<td></td>
<td>Hematology and Coagulation Procedures</td>
</tr>
<tr>
<td></td>
<td>Evaluation and Management of Other Outpatient Services</td>
</tr>
<tr>
<td></td>
<td>Surgical Procedures on the Cardiovascular System</td>
</tr>
<tr>
<td>Chemistry Pathology and Laboratory Tests</td>
<td>Chemistry Pathology and Laboratory Tests</td>
</tr>
<tr>
<td>Organ or Disease Oriented Panels</td>
<td></td>
</tr>
<tr>
<td>Hematology and Coagulation Procedures</td>
<td></td>
</tr>
<tr>
<td>Surgical Procedures on the Cardiovascular System</td>
<td></td>
</tr>
</tbody>
</table>

Tensor representation for EMR

Capture structured source interactions (e.g. group of procedures to treat a disease)

Co-occurrences of events are captured in the tensor as binary values
CP factorization for EMR

Wang Y et al. Rubik: Knowledge guided tensor factorization and completion for health data analytics. KDD 2015.
A possible application of EHR-phenotyping

Tucker factorization for pathology reports

Comparison of tensor modeling and factorization schemes

Challenges and opportunities: multiscale networks

Dynamic network: timeline of individualized genomic medicine

During an individual’s lifespan: from prewomb to tomb


Personalized multiscale networks to model dynamics of complex disease

DNA
Cell-specific RNA
Cytokines
Clinical labs
Mobile devices
Microbiome
Physiometrics

0:00 min
0:05 min
0:10 min

Dudley J. Big data in biology and medicine. Retrieved at www.aaas.org
Healthcare is really a big data industry

60% Exogenous Factors
30% Genomics Factors
10% Clinical Factors

1,100 Terabytes
Generated per lifetime

6 Terabytes
Per lifetime

0.4 Terabytes
Per lifetime

Help people live longer and feel better
“I'm telling you, our moonshot will be the impact we will have on Healthcare. It has already started. We will change and do our part to change the face of Healthcare. I am absolutely positive about it. And that, to me, while we do many other things, that will be one of the most important.”

Ginni Rometty
IBM Chairman, President and CEO
April 16, 2015
Multiple Positions Available:
• Interns
• Postdocs
• Research Engineers
• Research Staff Members

Contact: pzhang@us.ibm.com
Thank you!!!

“When you have a hammer, everything looks like a nail”