Precision Medicine on Cancer Treatment: A Joint Matrix Factorization Approach

Ping Zhang, Filippo Utro, Fei Wang, Jianying Hu

IBM T.J. Watson Research Center
Outline

- Introduction
- Methodology
- Experimental Results
- Future Works
Moving towards Precision Medicine (PM)

- PM: the right patient with the right drug at the right dose at the right time.
  - In his 2015 State of the Union address, President Obama stated his intention to infuse funds into a United States national "precision medicine initiative"
  - For patients, safer and more effective treatments
  - For doctors, reduce wasted time and resources with futile treatments
  - For pharms, lower cost marketing due to targeted patients, faster clinical trials, less focus on animal trials
Patient similarity and drug similarity analytics

- Patient Similarity analytics: Find patients who display similar clinical characteristic to the patient of interest
- Resulting insights: medical prognosis, risk stratification, care planning (especially for patients has multiple diseases)

- Drug Similarity analytics: Find drugs which display similar pharmacological characteristic to the drug of interest
- Resulting insights: drug repositioning, side-effect prediction, drug-drug interaction prediction

How to leverage both patient similarity and drug similarity for personalized medicine?
Outline

- Introduction
- Methodology
- Experimental Results
- Future Works
Multiple data sources of drug and patient information

Drug

Chemical Structure

Target Proteins

Side-effect Keywords

Calculate drug/patient similarities

Patient

Demographic

SNP

Copy-number variations

- weight loss
- impotence
- dizziness
- blurred vision
- .......

© 2014 IBM Corporation
A graphical illustration of the main idea

And more drug similarity networks
......

And more patient similarity networks
......

$D_1$, $D_2$, $D_3$

$UU^T$, $VV^T$, $R$, $\Theta$

$\omega_1$, $\omega_2$, $\omega_3$

$\pi_1$, $\pi_2$, $\pi_3$

$S_1$, $S_2$, $S_3$
A unified computational framework for PM treatment recommendation

Outputs:
1. predicted additional drug-patient associations
2. interpretable importance of different information sources
3. latent drug and patient groups as by-products

**Algorithm Flowchart (Input and Output)**

- drug chemical structure similarity network
- drug target protein similarity network
- drug side effect similarity network
- known drug-patient associations
- patient demographic similarity
- patient SNPs similarity
- patient CNV similarity

© 2014 IBM Corporation
Solution: A Joint Matrix Factorization Approach

Notations and symbols of the methodology:

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

- We aim to analyze the drug-patient network by minimizing the following objective:
  \[ J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \]

- The reconstruction loss of observed drug-patient associations:
  \[ J_0 = \| \Theta - U \Lambda V^T \|^2_F \]
  Similar drugs/patient (latent groups) have similar behaviors

- The reconstruction loss of drug similarities:
  \[ J_1 = \sum_{k=1}^{K_d} \omega_k \sqrt{\| D_k - U U^T \|^2_F + \delta_1 \sqrt{\| \omega \|^2_2}} \]
  \[ \text{Reconstruct integrated drug/patient networks} \]

- The reconstruction loss of patient similarities:
  \[ J_2 = \sum_{l=1}^{K_s} \pi_l \sqrt{\| S_l - V V^T \|^2_F + \delta_2 \sqrt{\| \pi \|^2_2}} \]

- Putting everything together, we obtained the optimization problem to be resolved:
  \[ \min_{U,V,\Lambda,\Theta,\omega,\pi} J, \text{ subject to } U \geq 0, V \geq 0, \Lambda \geq 0, \omega \geq 0, \omega^T 1 = 1, \pi \geq 0, \pi^T 1 = 1, P_{\Omega}(\Theta) = P_{\Omega}(R) \]
BCD approach for solving the non-convex optimization 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)

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

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

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

3. **while** Not Converge **do**

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

5. Solve $\omega$ and $\pi$ as described in section 3 (as a standard Euclidean projection onto a simplex) **closed-form solution**

6. Solve $\Lambda$ as described in section 4 (as a nonnegative quadratic optimization problem) **Solved by Projected Gradient Descent (PGD) method**

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

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

9. **end while**

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 $V$. 

© 2014 IBM Corporation
Outline

- Introduction
- Methodology
- Experimental Results
- Future Works
Problem description

- **Background:** Glioblastoma multiforme (GBM) is the most common and most aggressive malignant primary brain tumor in humans.

- **Raw data:** GBM data from The Cancer Genome Atlas (TCGA) website.

- **How to define an “effective” treatment?**
  - Median survival without treatment is 4½ months. Median survival with standard-of-care radiation and chemotherapy is 15 months.
  - We define treatments for patients who live for more than 15 months (i.e., 450 days) are effective.

- **Final data:** 118 patients, 41 distinct drugs and 261 known effective patient-drug associations.
  - by matching patients who have clinical information (e.g., age, race, gender), treatment information, genomic information (i.e., SNPs), and a positive response to treatment (i.e., live for more than 15 months), we got 118 patients.

- **Task:** For a given patient, predict a personalized treatment of GBM.
Measurements of patient similarities and drug similarities

- **Drug similarity evaluation**
  - Chemical structure: each drug was represented by an 881-dimensional PubChem fingerprint. Tanimoto coefficient (TC) of two fingerprints as chemical structure similarity. \( TC(A,B) = \frac{|A \cap B|}{|A \cup B|}. \)

- **Patient similarity evaluation**
  - Demographics: based on patient’s risk factors of GBM, derive demographic vectors based on gender (male, female), race (Asian, African American, Caucasian), and age (<41, 41~50, 51~60, 61~70, >70).
  - SNPs: 9507 mutated genes in the data, each patient was represented by a 9507-dimensional binary profile - elements encode for the presence or absence of each gene by 1 or 0 respectively.
  - Use Jaccard similarity to measure both patient demographics and SNPs similarities.
Alternative Method: Label Propagation (LP) methods

\[ F^* = (1 - \mu)(I - \mu W)^{-1} Y \]

Label Propagation on patient similarity

\[ F, Y \]
\[ W \]
Patient similarity

patients propagate their known effective treatments to other patients based on the patient network.

Label Propagation on drug similarity

\[ F, Y \]
\[ W \]
Drug Similarity

drugs propagate their target effective patients to other drugs based on the drug network.

LP method doesn’t take the consideration of both drug and patient at the same time.
LOOCV comparison of four treatment recommendation methods

- **LP-Drug ROC Curve (AUC=0.5848)**
- **LP-Patient Demographic ROC Curve (AUC=0.7127)**
- **LP-Patient SNPs ROC Curve (AUC=0.8335)**
- **JMF-All information ROC Curve (AUC=0.8505)**
Outline

- Introduction
- Methodology
- Experimental Results
- Future Works
To support precision medicine, we apply joint matrix factorization methodology by leveraging patient similarity and drug similarity analytics.

For further investigation

− Include more genomic data (e.g. CNV, RNA in TCGA) and drug data (e.g., target information)
− Explore more sophisticated drug and patient similarity measures
− Consider drug combinations
Thank you! | Questions?

Ping Zhang: pzhang@us.ibm.com