Joint Learning of Medication-Diagnosis Correspondence and Phenotypes via Hidden Interaction Tensor Factorization

Kejing Yin[1], William K. Cheung[1], Yang Liu[1], Benjamin C. M. Fung[2], Jonathan Poon[3]

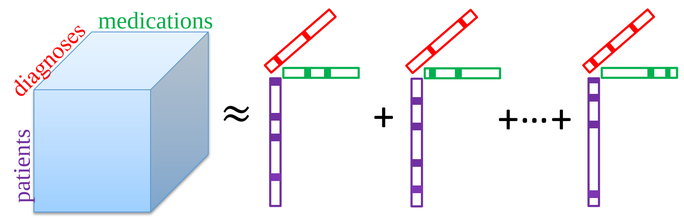
[1]Hong Kong Baptist University [2] McGill University [3]Hong Kong Hospital Authority

{cskjyin, william, csygliu}@comp.hkbu.edu.hk, ben.fung@mcgill.ca, jonathan@ha.org.hk

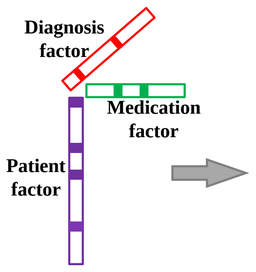
Introduction

- **Computational Phenotyping** is the process of converting the raw EHR data into clinically relevant features with minimum human supervision.
- **Tensor Factorization** has been proven an effective method for the task, with the capability of preserving and modelling the interaction structures.

❖ Phenotyping using Tensor CP Factorization



- **CP factorization** approximates the interaction tensor with the sum of rank-one tensors.



Diagnosis factor

Diagnosis entry	value
Cardiac dysrhythmias	**0.7**
Diabetes mellitus	0
Heart failure	**0.3**
Asthma	0
Pneumonitis	0

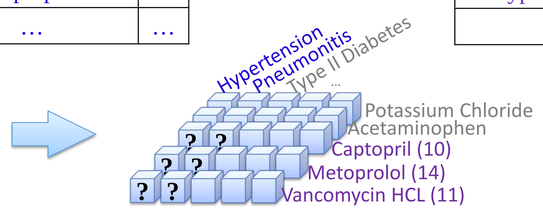
Medication factor

Medication entry	value
Insulin	0
Metoprolol	**0.55**
Furosemide	**0.45**
Albuterol	0
Diltiazem	0

- Each rank-one tensor is defined as a candidate of the resulting phenotypes.

❖ Challenge: Missing Interactions

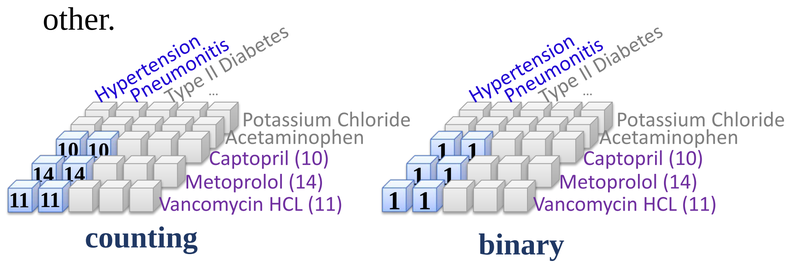
- For example, only a list of diagnoses and a list of medications are recorded, with their correspondence totally missing.



How to construct the interaction tensor?

❖ Existing Solutions: Equal Correspondence

- Existing models adopt the "equal-correspondence" strategy, assuming that all diagnoses and medications recorded per visit to be equally corresponding to each other.



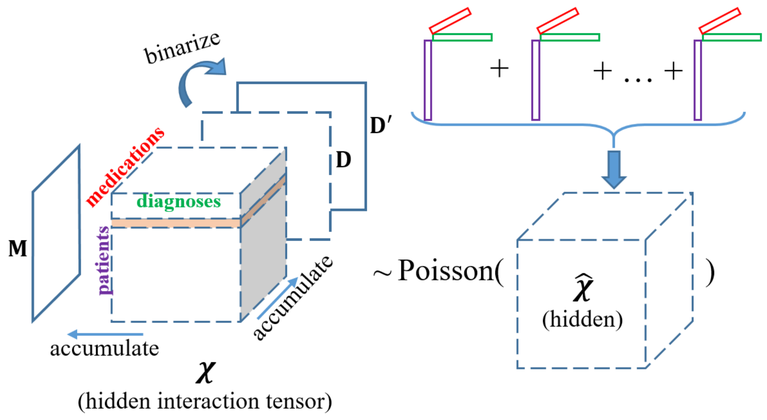
- Sometimes not realistic, e.g. vancomycin HCL is for Pneumonitis, but not Hypertension in clinical practice.
- The objective of CP factorization is to approximate the interactions as much as possible.

Inevitable error may be caused.

Related Works

- **Marble**[2]: Poisson non-negative tensor factorization model for computational phenotyping task, with a bias tensor to capture the global information.
- **Rubik**[3]: Incorporating guidance of existing medical knowledge, and adding pairwise constraint to make phenotypes distinct.
- **SiCNMF**[5]: Collective matrix factorization based method, jointly factorize the *patient-by-medication* matrix and the *patient-by-diagnoses* matrix, with the patient loading factor matrix shared.
- However, this model does not explicitly consider the interaction between the diagnoses and medications.

Proposed Method



- Do not observe the interactions directly, but instead the accumulations of the interaction tensor are observed.
- **Observations:**
  a *patient-by-medication* matrix **M** and
  a binarized *patient-by-diagnosis* matrix **D′**
- **Poisson distribution:** Assume the interaction entries follow Poisson distributions, with their means being reconstructed from the CP factors.

❖ Maximum Likelihood Estimation

Sum of independent Poisson yields another Poisson: $\mathbf{M} \sim \text{Poisson}(\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T})$

$\mathbf{D}'$ is binary, thus following a Bernoulli distribution: $\mathbf{D}' \sim \text{Ber}\left(1-\exp\left(-\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T}\right)\right)$

Solve for the factors via maximizing the joint likelihood:

$$\mathcal{L}=\mathbf{1}^T(-\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T}+\mathbf{M}*\log(\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(2)})\mathbf{U}^{(3)^T}))\mathbf{1}+\mathbf{1}^T(\mathbf{D}'*\log(\exp(\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T})-\mathbf{E})-\mathbf{U}^{(1)}\,\text{diag}(\mathbf{1}^T\mathbf{U}^{(3)})\mathbf{U}^{(2)^T})\mathbf{1}.$$

Experiments

- **Data Set**: MIMIC-III, an open-source, large-scale, de-identified ICU patients related EHR data set.
- Extract a subset containing 7,652 adult patients, each of which has 11 diagnoses per visit on average.
- **Baselines**: **Rubik** (one of the state-of-the-art NTF computational phenotyping models), **CP-APR** (widely used Poisson non-negative tensor factorization model) and **SiCNMF** (collective non-negative matrix factorization model).

Cardiac dysrhythmias(39.0%)		Diabetes mellitus(25.3%)		Asthma(5.5%)	
HITF	**Rubik**	**HITF**	**Rubik**	**HITF**	**Rubik**
Furosemide(0.08)	Potassium Chloride(0.08)	Insulin(0.64)	Insulin(0.09)	Albuterol 0.083% Neb Soln(0.46)	Potassium Chloride(0.08)
Potassium Chloride(0.07)	Insulin(0.06)	Insulin Human Regular(0.05)	Potassium Chloride(0.07)	Ipratropium Bromide Neb(0.39)	Insulin(0.06)
Metoprolol(0.06)	Furosemide(0.06)	Aspirin(0.05)	Furosemide(0.06)	Furosemide(0.08)	Furosemide(0.05)
Amiodarone HCl(0.05)	Magnesium Sulfate(0.04)	Furosemide(0.03)	Magnesium Sulfate(0.03)	Heparin(0.06)	Magnesium Sulfate(0.04)
Heparin Sodium(0.04)	Acetaminophen(0.03)	Atorvastatin(0.03)	Acetaminophen(0.03)		Acetaminophen(0.03)

Table 1: Top Five Corresponding Medications for Three Diagnoses Inferred by HITF and Rubik.

Diagnoses	Medications
Diabetes mellitus Other diseases of lung Acute kidney failure Essential hypertension …	Insulin Insulin Human Regular
Cardiac dysrhythmias Heart failure Other diseases of lung	Amiodarone HCl Metoprolol Furosemide …
Other diseases of lung Cardiac dysrhythmias Heart failure Chronic airway obstruction, not elsewhere classified	Albuterol Diltiazem Ipratropium Bromide MDI Fluticasone Propionate …

Table 2: Three Examples of Inferred Phenotypes.

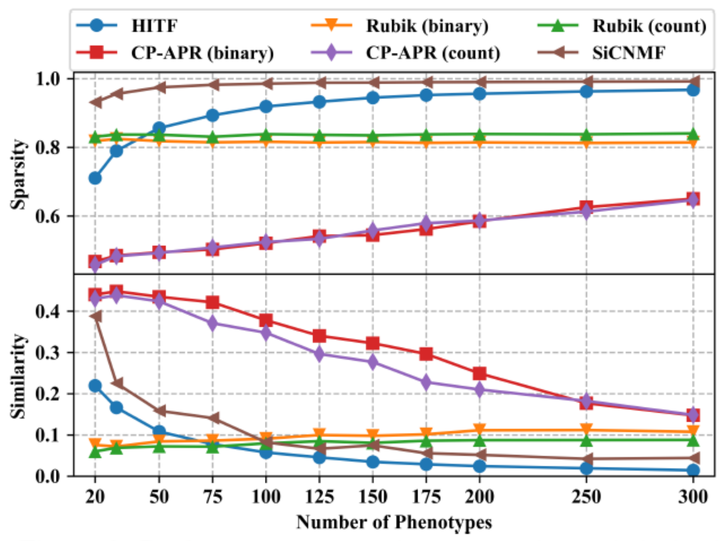


Figure 3: Performance comparison - sparsity and similarity.

Conclusion

- Evaluated by a medical expert, the diagnosis-medication correspondence and phenotypes inferred are **clinically meaningful**.
- HITF infers **more sparse and distinct** phenotypes than baselines.
- HITF inferred phenotypes **significantly improves the prediction performance**.

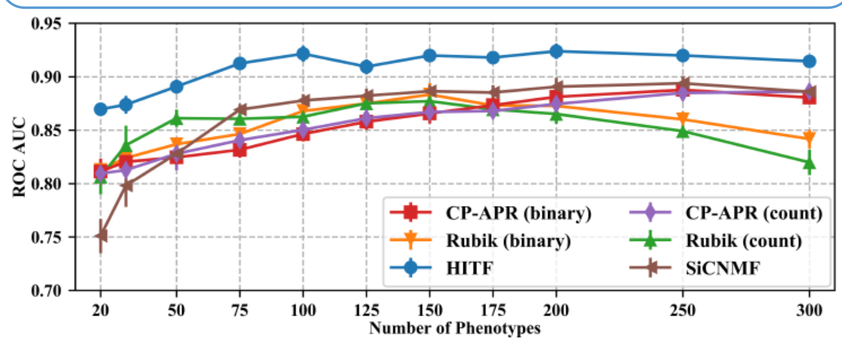


Figure 4: Prediction accuracy given different numbers of phenotypes.

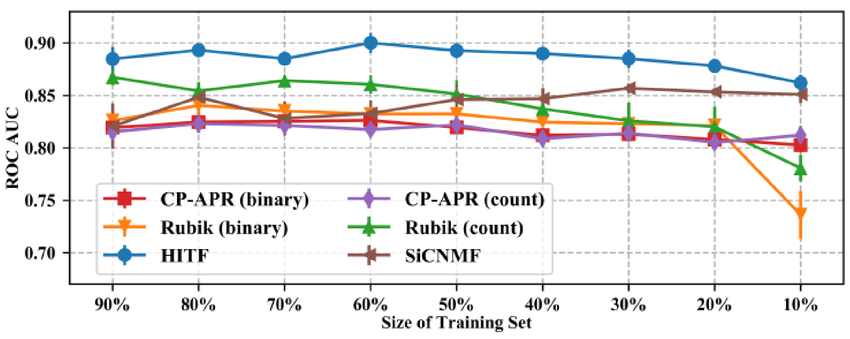


Figure 5: Prediction accuracy given different training sets.

References

[1] Chi, Eric C., and Tamara G. Kolda. "On tensors, sparsity, and nonnegative factorizations." *SIAM Journal on Matrix Analysis and Applications* 33.4 (2012): 1272-1299.

[2] Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization." *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014.

[3] Wang, Yichen, et al. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[4] Kim, Yejin, et al. "Discriminative and distinct phenotyping by constrained tensor factorization." *Scientific Reports* 7.1 (2017): 1114.

[5] Gunasekar, Suriya, et al. "Phenotyping using Structured Collective Matrix Factorization of Multi-source EHR Data." *arXiv preprint* arXiv:1609.04466 (2016).