6.872
Contemporary Approaches to Machine Learning

Tristan Naumann
tjn@mit.edu

Goals & Outline

- Review considerations for machine learning
  - *A Few Useful Things to Know about Machine Learning*
  - Additional clinical considerations
- Highlight some contemporary methods
  - Regularization: *Simultaneous Modeling of Multiple Diseases for Mortality Prediction in Acute Hospital Care*
  - Tensor Factorization: *Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics*

Pedro Domingos, CACM 2012

A FEW USEFUL THINGS TO KNOW ABOUT MACHINE LEARNING

Useful Things to Know about ML

1. Learning = representation + evaluation + optimization
2. It's generalization that counts
3. Data alone is not enough
4. Overfitting has many faces
5. Intuition fails in high dimensions
6. Theoretical guarantees are not what they seem
7. Feature engineering is key
8. More data beats cleverer algorithm
9. Learn many models, not just one
10. Simplicity does not imply accuracy
11. Representable does not imply learnable
12. Correlation does not imply causation

Useful Things to Know about ML

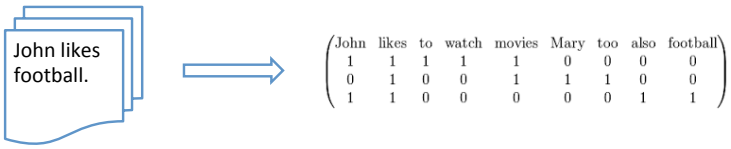
1. Learning = representation + evaluation + optimization
2. It's generalization that counts
3. Data alone is not enough
4. Overfitting has many faces
5. Intuition Fails in high dimensions
6. Theoretical guarantees are not what they seem
7. Feature engineering is key
8. More data beats cleverer algorithm
9. Learn many models, not just one
10. Simplicity does not imply accuracy
11. Representable does not imply learnable
12. Correlation does not imply causation

1. Learning = representation + evaluation + optimization

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

2. It's generalization that counts

- Goal: generalize beyond training examples
- Large, sparse feature spaces mean
  possible inputs >> observed inputs
- Training error is surrogate for test error



3. Data alone is not enough

- Features space grows more rapidly than examples
  - We'll never have enough data
- Real world *not* drawn uniformly at random
- Need assumptions, even simple ones
  - Smoothness
  - Similar examples have similar classes
  - Limited dependeces
  - Limited complexity
- Regularization

7. Feature engineering is key

- Many independent features that correlate well with class? Learning is easy!
- Complex function of features? Lots harder
- Machine learning is fast, but data
  - Gathering
  - Integrating
  - Cleaning
  - Pre-processing
  - Etc.

8. More data beats a cleverer algorithm

- Three constraints
  - Compute: CPU
  - Memory: RAM
  - Data: training
- Ironically, more data affords more complex models, but simple ones are often chosen due to scalability constraints

Additional Clinical Considerations

- Data
  - Availability
    - ICD codes (post-hoc billing)
    - Aggregate statistics (max/min SAPS)
  - Representation
    - Granularity (e.g. in ontologies)
    - Reasonable (respects underlying physiology)

Additional Clinical Considerations

- Learning
  - Evaluation
    - Clinicians do not operate across full ROC curve
    - Tradeoff between early warning and false alarms
  - Knowledge
    - Incorporate medical domain knowledge

Additional Clinical Considerations

- Outcome
  - Impact
    - Dollars saved
    - Lives preserved
    - Time conserved
    - Effort reduced
    - Quality of life increased
  - Actionable

Nozomi Nori et al., KDD 2015

SIMULTANEOUS MODELING OF MULTIPLE DISEASES FOR MORTALITY PREDICTION IN ACUTE HOSPITAL CARE

Overview

- Goal: mortality prediction
  - EHR data -> in-hospital mortality
  - Supervised
- Method: **regularization**
  - Similarities among diseases
  - Similarities among features
  - Ridge regression
- Evaluation: quantitative performance
  - Internal consistency and "make sense" results
  - Comparison of AUCs
- Thanks to Nozomi et al. for sharing slides!

Problem: Mortality risk prediction in ICU

- In ICU, clinicians have to make decisions in a very limited time.
- Accurate assessment of patient severity is crucial.
- Use mortality as a surrogate for the patient severity.
- The accurate prediction of the mortality risk could assist clinicians to pay more attentions to patients with a higher mortality risk.
  - Could lead to reducing "preventable deaths"

Our contribution

- We incorporated disease-specific contexts into ICU mortality modeling by multi-task learning where a task corresponds to a disease.
- We incorporated medical/clinical domain knowledge on the categorization of both the diseases and EHRs by two graph Laplacians, thereby alleviating data sparsity.
- We showed that our disease-based multi-task learning with medical domain knowledge worked better than conventional methods using a real world dataset.

A salient feature of ICU: diversity of diseases

- Patients with a wide variety of diseases in ICU.
- But the prediction is typically made by constructing one common predictive model for all the diseases.

Disease-specific context:

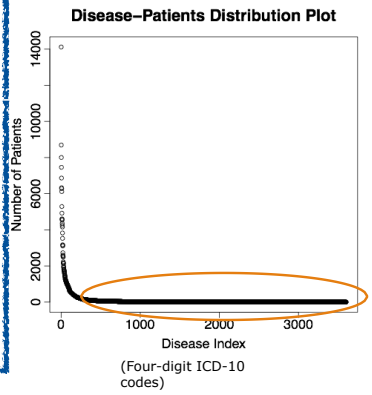
- Yet, each disease has a specific prediction rule that explains the mortality risk.

Hypothesis: customizing the model for each disease would improve the predictive modeling.

Challenge: data scarcity and sparsity

1. Data scarcity resulting from the customization

Attempts to build a customized model for each disease are complicated by the limited availability of sufficiently large datasets, because many diseases only have a small number of patients.



Challenge: data scarcity and sparsity

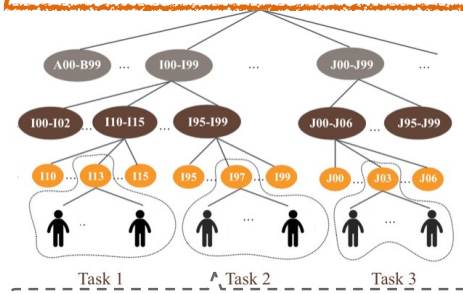
2. Data sparseness associated with electronic health records (EHRs)

- Raw EHRs are extremely sparse.
- One reason behind this sparsity: a significant number of EHRs are subject to medical classification, which categorizes medical information from multiple viewpoints, producing highly fine-grained features.

code:
1143001X1015
Efficacy Ingredients …

Proposed solution: multi-task learning with medical domain knowledge

A task corresponds to a disease and prediction tasks for different types of diseases are jointly solved by sharing information across the diseases.



Idea: exploit more information from more similar diseases in terms of medical classification of diseases; similarly for EHRs

ICD10 hierarchy: diseases are categorized in terms of cause, symptoms, morphological disparity, etc., encoding important information that might affect the mortality risk.

Integrate medical domain knowledge by graph Laplacians encoding the similarities among diseases and EHRs into the regularization term in an optimization problem

Optimization Problem

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) + \Omega(\boldsymbol{W})$$

Loss function (log loss) — Regularization

$$\mathcal{L}(\boldsymbol{W}) \equiv -\sum_{t}^{T}\sum_{n}^{N_t}\{y_{t,n}\log\sigma({\boldsymbol{w}^{(t)}}^{\top}\boldsymbol{\phi}_n^{(t)}) + (1-y_{t,n})\log(1-\sigma({\boldsymbol{w}^{(t)}}^{\top}\boldsymbol{\phi}_n^{(t)}))\}$$

$$\Omega(\boldsymbol{W}) \equiv \lambda^{\mathrm{dz}}\Omega^{\mathrm{dz}}(\boldsymbol{W}) + \lambda^{\mathrm{feat}}\Omega^{\mathrm{feat}}(\boldsymbol{W}) + \lambda^{\mathrm{rid}}\Omega^{\mathrm{rid}}(\boldsymbol{W})$$

Incorporation of Domain Knowledge — Avoid overfitting

$\boldsymbol{S}^{\mathrm{dz}}$: Similarity matrix for diseases

$\boldsymbol{S}^{\mathrm{feat}}$: Similarity matrix for EHRs

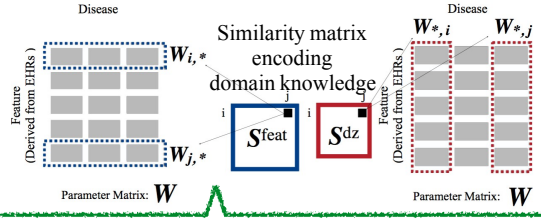
Disease Similarity

$$\Omega^{\mathrm{dz}}(\boldsymbol{W}) \equiv \frac{1}{4}\sum_{i=1}^{T}\sum_{j=1}^{T}\boldsymbol{S}^{\mathrm{dz}}_{i,j}\left\|\frac{\boldsymbol{W}_{*,i}}{\sqrt{\boldsymbol{D}^{\mathrm{dz}}_{i,i}}} - \frac{\boldsymbol{W}_{*,j}}{\sqrt{\boldsymbol{D}^{\mathrm{dz}}_{j,j}}}\right\|^2 = \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}\boldsymbol{\mathcal{L}}^{\mathrm{dz}}\boldsymbol{W}^{\top})$$

EHRs Similarity

$$\Omega^{\mathrm{feat}}(\boldsymbol{W}) \equiv \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1}^{M}\boldsymbol{S}^{\mathrm{feat}}_{i,j}\left\|\frac{\boldsymbol{W}_{i,*}}{\sqrt{\boldsymbol{D}^{\mathrm{feat}}_{i,i}}} - \frac{\boldsymbol{W}_{j,*}}{\sqrt{\boldsymbol{D}^{\mathrm{feat}}_{j,j}}}\right\|^2 = \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}^{\top}\boldsymbol{\mathcal{L}}^{\mathrm{feat}}\boldsymbol{W})$$

Integrate medical domain knowledge by graph Laplacians encoding the similarities among diseases and EHRs into the regularization term in an optimization problem (cont'd)



Make two model parameters for two diseases similar if the two diseases are similar in terms of the medical classifications given as the similarity matrix; similarly for EHRs.

Disease Similarity

$$\Omega^{\mathrm{dz}}(\boldsymbol{W}) \equiv \frac{1}{4}\sum_{i=1}^{T}\sum_{j=1}^{T}\boldsymbol{S}^{\mathrm{dz}}_{i,j}\left\|\frac{\boldsymbol{W}_{*,i}}{\sqrt{\boldsymbol{D}^{\mathrm{dz}}_{i,i}}} - \frac{\boldsymbol{W}_{*,j}}{\sqrt{\boldsymbol{D}^{\mathrm{dz}}_{j,j}}}\right\|^2 = \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}\boldsymbol{\mathcal{L}}^{\mathrm{dz}}\boldsymbol{W}^{\top})$$

EHRs Similarity

$$\Omega^{\mathrm{feat}}(\boldsymbol{W}) \equiv \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1}^{M}\boldsymbol{S}^{\mathrm{feat}}_{i,j}\left\|\frac{\boldsymbol{W}_{i,*}}{\sqrt{\boldsymbol{D}^{\mathrm{feat}}_{i,i}}} - \frac{\boldsymbol{W}_{j,*}}{\sqrt{\boldsymbol{D}^{\mathrm{feat}}_{j,j}}}\right\|^2 = \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}^{\top}\boldsymbol{\mathcal{L}}^{\mathrm{feat}}\boldsymbol{W})$$

Experiments

Experimental Condition1:

- Dataset from a hospital
- In-hospital mortality risk prediction
- Features: gender, age (< 65/>=65), cormobidities, interventions; 2,000~2,500 features
- Three settings: prediction before ICU discharge (2 days before / 1day before) and retrospective prediction (1 day before hospital discharge)

Disease code	Disease name
A41	Other septicaemia
C15	Malignant neoplasm of oesophagus
C16	Malignant neoplasm of stomach
C18	Malignant neoplasm of colon
C20	Malignant neoplasm of rectum
C22	Malignant neoplasm of liver and intrahepatic bile ducts
C34	Malignant neoplasm of bronchus and lung
G93	Other disorders of brain
I20	Angina pectoris
I21	Acute myocardial infarction
I35	Nonrheumatic aortic valve disorders
I50	Heart failure
I70	Atherosclerosis
I71	Aortic aneurysm and dissection
K52	Other noninfective gastroenteritis and colitis
K56	Paralytic ileus and intestinal obstruction without hernia
K65	Peritonitis
K76	Other diseases of liver
K91	Postprocedural disorders of digestive system, not elsewhere classified
N18	Chronic renal failure

Experimental Condition2:

Rule for creating similarity matrix

- Similarity between two diseases: the number of shared levels in the ICD hierarchy.
- Similarity between two features: if the feature is medication and if two medicines have the same drug efficacy, then the similarity between them is set to 1, otherwise 0.

Table: Comparison of various methods used in our experiment

	Regularization	Domain Knowledge	Disease-based Customization
Proposed	Task, Feature, l_2	Disease, EHRs	✔
Proposed-feat	Feature, l_2	EHRs	✔
Proposed-dz	Task, l_2	Disease	✔
non-MTL-1 (separate)	l_2		✔
non-MTL-2 (common)	l_2		
MTL-1 (l_{21})	$l_{2,1}$, l_2		✔
MTL-2 (Trace)	Trace		✔

Result: Multi-task learning with medical domain knowledge worked best

Table: Comparison of averaged AUCs among various methods with Wilcoxon signed rank test

	Pre-ICU discharge prediction (2 days before)	Pre-ICU discharge prediction (1 day before)	Retrospective prediction
Proposed	**0.763**	**0.795**	**0.914**
Proposed-feat	0.692	0.717	0.837
Proposed-dz	0.758	0.793	0.913
non-MTL-1 (separate)	0.692	0.714	0.836
non-MTL-2 (common)	0.760	0.783	0.886
MTL-1 (l_{21})	0.717	0.723	0.819
MTL-2 (Trace)	0.724	0.739	0.837

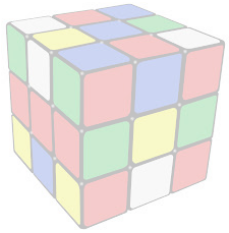
※Proposed method outperforms other methods significantly ($p<0.05$) except for the gray colored ones.

There is no other method that performs equally well as our proposed method throughout the prediction settings.

Both the domain knowledge on medical classification of diseases and clinical classification of EHRs can improve the predictive performance.

Conclusions

- Authors claim to “take a step toward personalized medicine in ICU”
- Cool method!
  - MTL
  - Cross-regularization



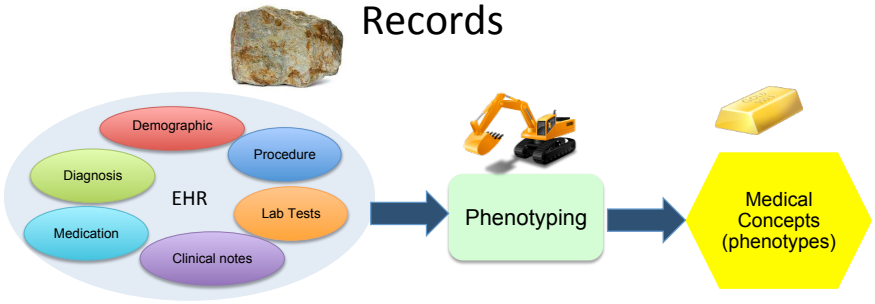
Yichen Wang et al., KDD 2015

RUBIK: KNOWLEDGE GUIDED TENSOR FACTORIZATION AND COMPLETION FOR HEALTH DATA ANALYTICS

Overview

- Goal: computational phenotyping
  - EHR data -> meaningful clinical concepts
  - Unsupervised
- Method: **tensor factorization**
  - Guidance constraints align medical knowledge
  - Pairwise constraints for distinct phenotypes
  - Completion for missing and noisy data
- Evaluation: algorithmic characteristics
  - Previous work: Limestone, Marble
  - Internal consistency and "make sense" results
  - Scalability
- Thanks to Yichen et al. for sharing slides!

Phenotyping from Electronic Health Records



- Limitations of existing phenotyping methods
  - Unable to leverage existing knowledge
  - High overlapping between discovered phenotypes
  - Not robust to missing and noisy data
  - Not scalable

Ideal Phenotyping Algorithms

- Guidance: incorporate medical knowledge
- Non-overlap: discover distinct and meaningful phenotypes
- Robust: handle noisy and missing data
- Scalable

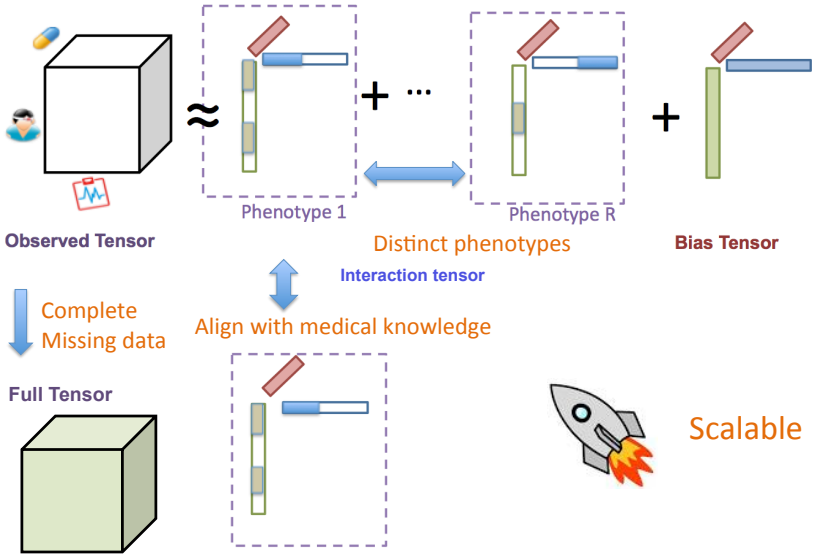
Algorithms Comparison

Property	Marble [1]	FaLRTC [2]	CTMF [3]	TF-BPP [4]	WCP [5]	NETWORK [6]	**Rubik**
Guidance							✓
Non-overlapping			✓			✓	✓
Robustness	✓	✓		✓	✓	✓	✓
Scalability							✓

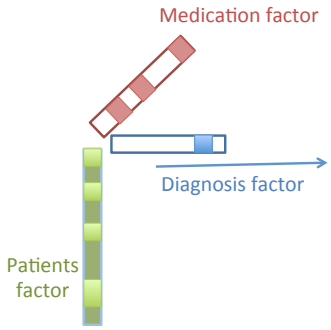
Table 2: A comparison of different tensor models

[1] 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.
[2] Liu, Ji, et al. "Tensor completion for estimating missing values in visual data." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.1 (2013): 208-220.
[3] Acar, Evrim, Tamara G. Kolda, and Daniel M. Dunlavy. "All-at-once optimization for coupled matrix and tensor factorizations." *arXiv preprint arXiv:1105.3422* (2011).
[4] Kim, Jingu, and Haesun Park. "Fast nonnegative tensor factorization with an active-set-like method." *High-Performance Scientific Computing*. Springer London, 2012. 311-326.
[5] Acar, Evrim, et al. "Scalable tensor factorizations for incomplete data." *Chemometrics and Intelligent Laboratory Systems* 106.1 (2011): 41-56.
[6] Davidson, Ian, et al. "Network discovery via constrained tensor analysis of fmri data." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

Rubik



Examples of Phenotype

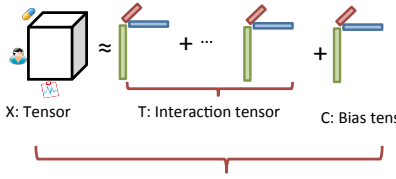


Diagnoses	Medications
Hypertension	Statins
Ischemic heart disease	Angiotensin receptor blockers
Hyperlipidemia	ACE inhibitors
Obesity	Loop diuretics
	Cardioselective beta blockers

Phenotype with Guidance info on heart failure

Diagnoses	Medications
Cardiac dysrhythmias	first generation cephalosporins
Heart failure	calcium channel blocking agents
Atrial fibrillation and flutter	antianginal agents
Cardiomyopathy	mvi

Problem Overview



$$\min_{\mathcal{X},\mathcal{T},\mathcal{C}} \quad \Psi = \|\mathcal{X} - \mathcal{C} - \mathcal{T}\|_F^2 + \frac{\lambda_a}{2}\|(\mathbf{A}^{(p)} - \hat{\mathbf{A}}^{(p)})\mathbf{W}\|_F^2 + \frac{\lambda_q}{2}\|\mathbf{Q} - \mathbf{A}^{(k)T}\mathbf{A}^{(k)}\|_F^2$$

Factorization error — Guidance — Nonoverlap constraint

$$\text{s.t.} \quad \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{O})$$

O: Observed tensor

$$\mathcal{T} = [\![\mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!] \in \Omega_\mathcal{T}, \quad \mathcal{C} = [\![\mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!] \in \Omega_\mathcal{C}$$

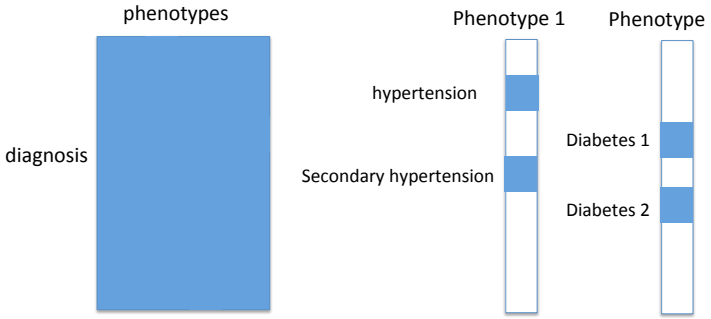
$$\Omega_\mathcal{T} = \Omega_{A_1} \times \cdots \times \Omega_{A_N}, \ \Omega_{A_n} = \{\mathbf{A} \in \{0\} \cup [\gamma_n, +\infty)^{I_n \times R}\}$$

Sparsity

$$\Omega_\mathcal{C} = \Omega_{u_1} \times \cdots \times \Omega_{u_N}, \ \Omega_{u_n} = \{\mathbf{u} \in [0, +\infty)^{I_n \times 1}\}$$

Nonnegativity

Guidance Information



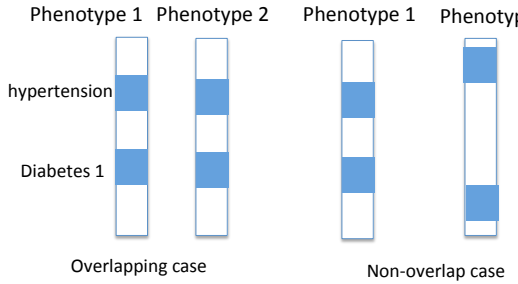
$$\mathbf{W} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

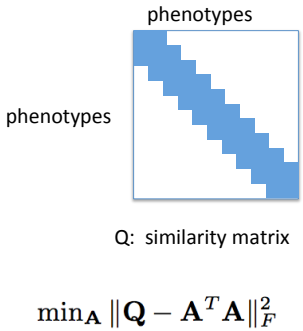
Guidance is limited

$$\min_{\mathbf{A}} \|(\mathbf{A} - \hat{\mathbf{A}})\mathbf{W}\|_F^2$$

Pairwise Constraints

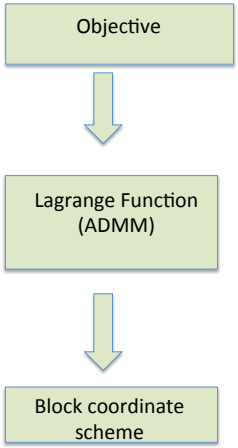
- We can penalize the cases where phenotypes have overlapping dimensions.





$$\min_{\mathbf{A}} \|\mathbf{Q} - \mathbf{A}^T\mathbf{A}\|_F^2$$

Formulation



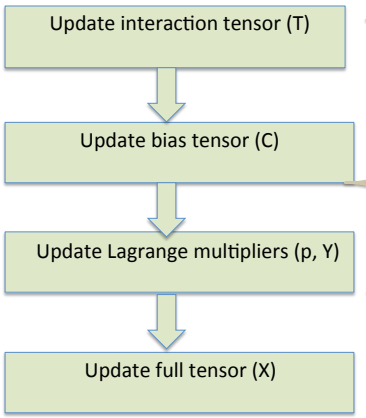
$$\Psi = \|\mathcal{X} - \mathcal{C} - \mathcal{T}\|_F^2 + \frac{\lambda_a}{2}\|(\mathbf{A}^{(p)} - \hat{\mathbf{A}}^{(p)})\mathbf{W}\|_F^2 + \frac{\lambda_q}{2}\|\mathbf{Q} - \mathbf{B}^{(k)T}\mathbf{A}^{(k)}\|_F^2$$

$$\mathbf{A}^{(k)} = \mathbf{B}^{(k)} \qquad \mathbf{v}^{(n)} = \mathbf{u}^{(n)}$$

$$\mathcal{L} = \Psi + \sum_{n=1}^{N}(\langle \mathbf{p}^{(n)}, \mathbf{v}^{(n)} - \mathbf{u}^{(n)}\rangle + \frac{\eta}{2}\|\mathbf{v}^{(n)} - \mathbf{u}^{(n)}\|_F^2)$$
$$+ \sum_{n=1}^{N}(\langle \mathbf{Y}^{(n)}, \mathbf{B}^{(n)} - \mathbf{A}^{(n)}\rangle + \frac{\mu}{2}\|\mathbf{B}^{(n)} - \mathbf{A}^{(n)}\|_F^2)$$

p and Y are Lagrange multipliers

Block coordinate scheme



$$\min_{\mathbf{A}^{(n)}} \|\mathbf{A}^{(n)}\mathbf{\Pi}^{(n)T} - \mathbf{R}_{(n)}\|_F^2$$
$$+ \frac{\lambda_q}{2}\|\mathbf{Q} - \mathbf{B}_t^{(n)T}\mathbf{A}^{(n)}\|_F^2 + \frac{\lambda_a}{2}\|(\mathbf{A}^{(n)} - \hat{\mathbf{A}}^{(n)})\mathbf{W}\|_F^2$$
$$+ \frac{\mu_t}{2}\|\mathbf{A}^{(n)} - \mathbf{B}_t^{(n)} - \mathbf{Y}_t^{(n)}/\mu_t\|_F^2$$

$$\mathbf{B}_{t+1}^{(n)} = \begin{cases} \mathbf{A}_{t+1}^{(n)} + \frac{1}{\mu_t}\mathbf{Y}_t^{(n)} & \text{if } \gamma_n \le \mathbf{A}_{t+1}^{(n)} + \frac{1}{\mu_t}\mathbf{Y}_t^{(n)} \\ 0 & \text{otherwise} \end{cases}$$

$$\min_{\mathbf{u}^{(n)}} \|\mathbf{u}^{(n)}(\mathbf{\Lambda}^{(n)})^T - \mathbf{E}_{(n)}\|_F^2 + \frac{\eta_t}{2}\|\mathbf{u}^{(n)} - \mathbf{v}^{(n)} - \mathbf{p}^{(n)}/\eta_t\|_F^2$$

$$\mathbf{v}_{t+1}^{(n)} = \max(0, \mathbf{u}_{t+1}^{(n)} + \frac{1}{\eta_t}\mathbf{p}_t^{(n)})$$

$$\mathbf{Y}_{t+1}^{(n)} = \mathbf{Y}_t^{(n)} + \mu_t(\mathbf{B}_{t+1}^{(n)} - \mathbf{A}_{t+1}^{(n)})$$

$$\mathbf{p}_{t+1}^{(n)} = \mathbf{p}_t^{(n)} + \eta_t(\mathbf{v}_{t+1}^{(n)} - \mathbf{u}_{t+1}^{(n)})$$

$$\mathcal{X}_{t+1} = \mathcal{P}_{\Omega^c}(\mathcal{T}_{t+1} + \mathcal{C}_{t+1}) + \mathcal{P}_{\Omega}(\mathcal{O})$$

Experiments

- Phenotype discovery: How Rubik discovers meaningful phenotypes?
- Phenotype discovery: How Rubik discovers fine-grained sub-phenotypes?
- Phenotype discovery: How Rubik discovers distinct phenotypes?
- Noise analysis: Is Rubik robust to noisy and missing data?
- Scalability: Is Rubik Scalable?
- Constraints analysis: Are all constraints important?

Constructing Tensor

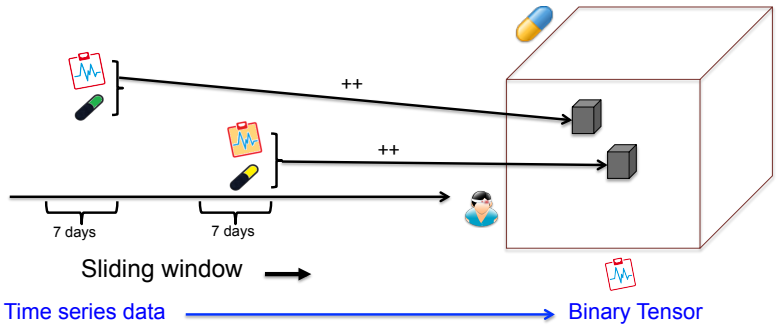


Figure 1: Co-occurrences of events within a patient's history are captured in the tensor as binary values.

Datasets

- Vanderbilt: 3rd order tensor with patient, diagnosis and medication modes of size 7,744 by 1,059 by 501, respectively.
- CMS: 472,645 patients by 11,424 diagnoses by 262,312 medication events.

Phenotype Discovery: Meaningful Interaction Tensor

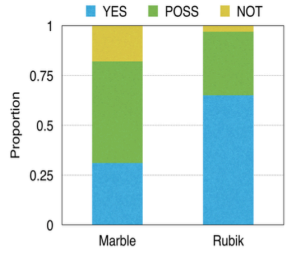


Figure 1: A comparison of the meaningfulness of the phenotypes discovered by Marble and Rubik.

- Conduct surveys with medical experts to evaluate 30 phenotypes from Marble and Rubik
- YES means clinically meaningful
- POSS means possibly meaningful
- NOT means not meaningful

Rubik generates more meaningful phenotypes

Phenotype Discovery： Meaningful Bias Tensor

Diagnoses	Medications
Hypertension	Statins
Disorders of lipoid metabolism	Loop diuretics
Heart failure	Miscellaneous analgesics
Respiratory & chest symptoms	Antihistamines
Chronic kidney disease	Vitamins
Other and unspecified anemias	Calcium channel blockers
Diabetes mellitus type 2	Beta blockers
Digestive symptoms	Salicylates
Other diseases of lung	ACE inhibitors

Table 5: Elements of the diagnosis and medication modes in the bias tensor.

- Meaningful: accurately reflects the stereotypical type of patients

- Supports the medical report:
- 80% of older adults suffer from at least one chronic condition and 50% have two or more chronic conditions

Phenotype Discovery： Meaningful Subphenotypes

Marble

Diagnoses	Medications
Chronic kidney disease	Central sympatholytics
Hypertension	Angiotensin receptor blockers
Unspecified anemias	ACE inhibitors
Fluid electrolyte imbalance	Immunosuppressants
Type 2 diabetes mellitus	Loop diuretics
Other kidney disorders	Gabapentin

Table 6: An example of a Marble-derived phenotype.

Rubik

A. Metabolic syndrome phenotype

Diagnoses	Medications
Hypertension	Calcineurin inhibitors
Chronic kidney disease	Insulin
Ischemic heart disease	Immunosuppressants
Disorders of lipoid metabolism	ACE inhibitor
Anemia of chronic disease	Cox-2 inhibitors
	Antibiotics
	Statins
	Calcium

B. Secondary hypertension phenotype

Diagnoses	Medications
Secondary hypertension	Class V antiarrhythmics
Fluid & electrolyte imbalance	Salicylates
Unspecified anemias	Antianginal agents
Hypertension	ACE inhibitors
	Calcium channel blockers
	Immunosuppressants

Table 7: Examples of Rubik-derived subphenotypes. The two tables show separate subgroups of hypertension patients: A) metabolic syndrome, and B) secondary hypertension due to renovascular disease.

Phenotype Discovery： Meaningful Subphenotypes

Experiment setup:

- Four guidance: hypertension, diabetes1, diabetes2, heart disease

- 2 sub-phenotypes for each guidance

- 8 sub-phenotypes in total

Experts evaluation:

- 62.5% are meaningful

- 37.5% are possibly meaningful

More Distinct Phenotypes

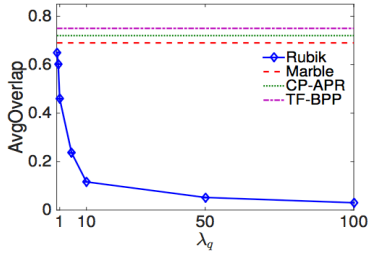


Figure 2: The average level of overlap in the phenotypes as a function of the pairwise constraint coefficient λ_q.

- Pairwise constraint leads to more distinct phenotypes

- Average similarity tend to stabilize whe λ_q is larger than 10

Missing Data Analysis

Generate missing data:
randomly set the observed values to be 0

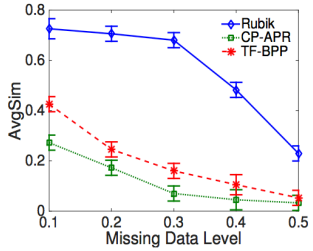


Figure 3: An average similarity comparison of different methods as a function of the missing data level.

Noise Analysis

Generate noise:
randomly set the unobserved entries to be 1

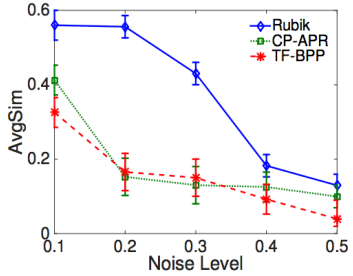


Figure 4: An average similarity comparison of different methods as a function of the noise level.

Incorrect Guidance

Generate incorrect guidance:
randomly set entries in guidance matrix to be 1

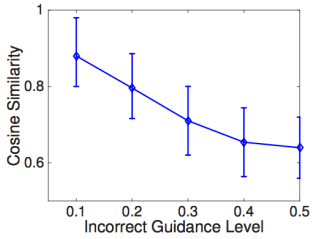


Figure 5: The similarity between the true solution and the solution under incorrect guidance as a function of the incorrect guidance level.

Scalability

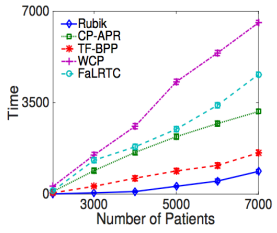


Figure 7: A runtime comparison of different methods on the *Vanderbilt* dataset as a function of the number of patients.

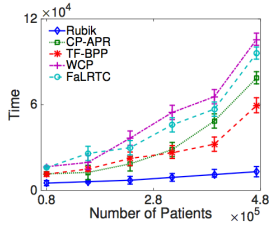


Figure 8: A runtime comparison of different methods on the *CMS* dataset as a function of the number of patients.

Rubik is around six times faster than competitors

Constraints Analysis

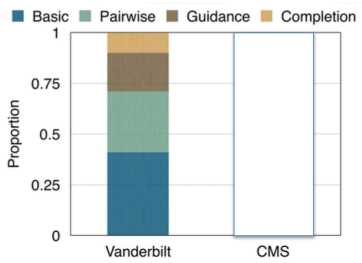


Figure 9: Proportion of contribution of each constraint.

- All constraints are important!
- Pairwise constraint provides the largest boost

Conclusions (Author)

- The resulting phenotypes are concise, distinct, and interpretable
- Rubik can also incorporate guidance from medications and patients
- Rubik is robust to noisy and missing data
- Rubik is scalable

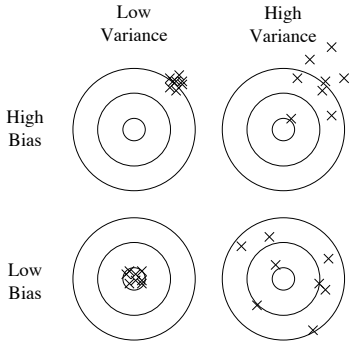
Tristan Naumann (tjn@mit.edu)

THANKS!

BACKUP

4. Overfitting has many faces

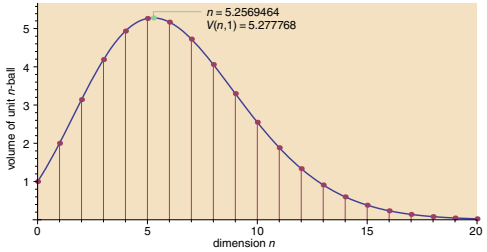
- Decompose error
  - Bias
  - Variance
- Combat with choices, e.g.
  - Linear learner has high bias because when the frontier is not hyperplane can't induce it
  - Decision tree can represent any Boolean function, but might be very different depending on training set
- Strong false assumptions often trump weak true ones



5. Intuition fails in high dimension

- Curse of dimensionality
- Doubly cursed, e.g.
  - Approx. hypersphere by inscribing in hypercube, in high dimensions nearly all volume of hypercube is outside hypersphere

Scientific American: Volume of a Unit Ball in *n* dimensions



6. Theoretical guarantees are not enough

- Guarantees on induction: really cool!
  - Unfortunately, most interesting hypothesis spaces are doubly exponential in number of features
- Asymptotic guarantees are comforting
  - Shouldn't be used to select a learning since we don't live in a world with infinite data

9. Learn many models, not just one

- Bagging
  - Generate random variations of the training set by resampling, learn a classifier for each, and combine results by voting
- Boosting
  - Training examples have weights, and these are varied so that each new classifier focuses on the examples the previous ones got wrong
- Stacking
  - Outputs of individual classifiers become inputs of a "higher-level" learning that figures out how to combine

10. Simplicity does not imply accuracy

- Occam's razor often misinterpreted
  - Results in publications which "prove" superiority of simpler models
  - Wolpert's "no free lunch" theorems reject this
- Contrary to intuition, not necessarily connection between number of parameters and tendency to overfit

11. Representable does not imply learnable

- Theory
  - "Every function can be represented, or approximated arbitrarily closely, using this representation"
- Practice
  - Great, but does not help us select appropriate learner

12. Correlation does not imply causation

- Commonly stated
- But then often seemingly ignored

Disease-specific predictive features

- We've conducted analyses about predictive features for each disease.
- The predictive features for a disease reflected the characteristics of it more adequately than when we constructed one common predictive model.
- They also contained some hypothetical suggestions.

Conclusions

- Disease-specific contexts are valuable to improve predictive performance in mortality modeling in ICU.
- Adequate incorporation of medical/clinical domain knowledge can enhance data-driven approach.
- We arguably took a step towards personalized medicine in ICU.