Medical Coding Classification by Leveraging Inter-Code Relationships

Y. Yan[+], G. Fung[*], J. Dy[+], R. Rosales[*]

[+]Northeastern University, [*]Siemens Healthcare

July 18th 2010

Medical Coding

-is the process of transforming information contained in patient medical records into standard pre-defined medical codes.

Sample Medical Transcript

Description: A 23-year-old white female presents with complaint of allergies.
(Medical Transcription Sample Report)

SUBJECTIVE: This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have asthma but doest not require daily medication for this and does not think it is flaring up.

MEDICATIONS: Her only medication currently is Ortho Tri-Cycle and the Allegra.
ALLERGIES: She has no known medicine allergies.

OBJECTIVE:
Vitals: Weight was 130 pounds and blood pressure 124/78.
HEENT: Her throat was mildly erythematous without exudate. Nasal mucosa was erythematous and swollen. Only clear drainage was seen. TMs were clear.
Neck: Supple without adenopathy.
Lungs: Clear.

ASSESSMENT: Allergic rhinitis.

PLAN:
1. She will try Zyrtec instead of Allegra again. Another option will be to use loratadine. She does not think she has prescription coverage so that might be cheaper.
2. Samples of Nasonex two sprays in each nostril given for three weeks. A prescription was written as well.

ICD-9, the Ninth Revision of ICD (**I**nternational Statistical **C**lassification of **D**iseases and Related Health Problems), provides the standard for coding clinical records in the US.

- The codes include diagnosis information, classifications for signs, symptoms, abnormal findings, complaints, social circumstances, and causes of injury or disease.
- Useful for clinical care, research, education and billing purposes.

Example Codes:

493.1	Intrinsic asthma
786.2	Cough
98.83	Domestic tasks therapy
970.1	Nalorphine
...	.....

Characteristics for Problem

- Sparsity of Classes
- High Dimensionality
- Multi-Label Classification Problem

Typical Approaches to Multi-Label Classification

- 1-vs-Rest
  - 1-vs-Rest SVM
  - K-Nearest Neighbors
- Voting
  - *rank*-SVM
- Probabilistic Modeling
  - BPMLL

Formulation

Objective Function

Proximal SVM with l1-norm

$$\min_{W,M} \mu \sum_{i=1}^{D} \|\mathbf{e} - \mathbf{diag}(\mathbf{y}_i)\hat{\mathbf{y}}_i'\|^2 + \|W\|_1$$

where

1-vs-Rest

$$\hat{\mathbf{y}}_i \triangleq (\mathbf{x}_i W \ \mathrm{I} \ + \boldsymbol{\gamma})$$

Notations

Inputs: $\mathbf{x}$, X	Labels: $\mathbf{y}$, Y	Classifiers: W	Relation Matrix: M
Prior: $\tilde{\mathcal{M}}_{\mathtt{sim}}$	Shifts: $\boldsymbol{\gamma}$	One Vector: $\mathbf{e}$	Trade Off: μ, ν

Formulation

Objective Function

Proximal SVM with l1-norm | **Regularization on Class Relationship Matrix**

$$\min_{W,M} \mu \sum_{i=1}^{D} \|\mathbf{e} - \mathbf{diag}(\mathbf{y}_i)\hat{\mathbf{y}}_i'\|^2 + \|W\|_1 + \nu\|M - \tilde{\mathcal{M}}_{\mathtt{sim}}\|^2_{\mathtt{frob}}$$

$$\mathbf{s.t.} \qquad -1 \leqslant M(i,j) \leqslant 1 \qquad i,j = \{1,2,\ldots,D\}$$

where

Class Inter-Related

$$\hat{\mathbf{y}}_i \triangleq (\mathbf{x}_i W M + \boldsymbol{\gamma})$$

Notations

Inputs: $\mathbf{x}$, X	Labels: $\mathbf{y}$, Y	Classifiers: W	Relation Matrix: M
Prior: $\tilde{\mathcal{M}}_{\mathtt{sim}}$	Shifts: $\boldsymbol{\gamma}$	One Vector: $\mathbf{e}$	Trade Off: μ, ν

Class Relationship Matrix Prior, $\tilde{\mathcal{M}}_{\mathtt{sim}}$

$$\mathcal{M}_{\mathtt{sim}} = [\omega_{ij}]_{L\times L}$$

$$\omega_{ij} = \sum_{q=1}^{D} \mathbf{y}_q(i)\mathbf{y}_q(j) \leftarrow \text{Similarity between Classes (Codes)}$$

$$D_{\mathcal{M}} = [d_{ij}]_{L\times L} \leftarrow \text{Degree Matrix}$$

$$d_{ii} = \sum_j |\omega_{ij}| \qquad d_{ij} = 0 \ (i \neq j)$$

$$\boxed{\tilde{\mathcal{M}}_{\mathtt{sim}} = \mathcal{M}_{\mathtt{sim}} D_{\mathcal{M}}^{-1}}$$

samples	**word-1**	**word-2**	...	**class**	**code-1**	**code-2**	...
$\mathbf{x}_1$	-1.2	231	...	$\mathbf{y}_1$	-1	-1	...
$\mathbf{x}_2$	2.4	50	...	$\mathbf{y}_2$	-1	1	...
$\mathbf{x}_3$	0.5	873	...	$\mathbf{y}_3$	-1	-1	...

Large Scale Formulation

- For large scale problems, it is not computationally efficient to solve the original form of the objective function.
- We re-write the objective function in vector form (factorize both W and M), and update them one vector at a time.

$$\min_{W,M} \mu \sum_{i=1}^{D} \|\mathbf{y}_i - \mathbf{x}_i \sum_{j=1}^{L} \mathbf{w}_j \mathbf{m}_j^{r\prime}\|^2 + \sum_{j=1}^{L} \|\mathbf{w}_j\|_1 + \nu \|M - \tilde{\mathcal{M}}_{\mathtt{sim}}\|_{\mathtt{frob}}^2$$

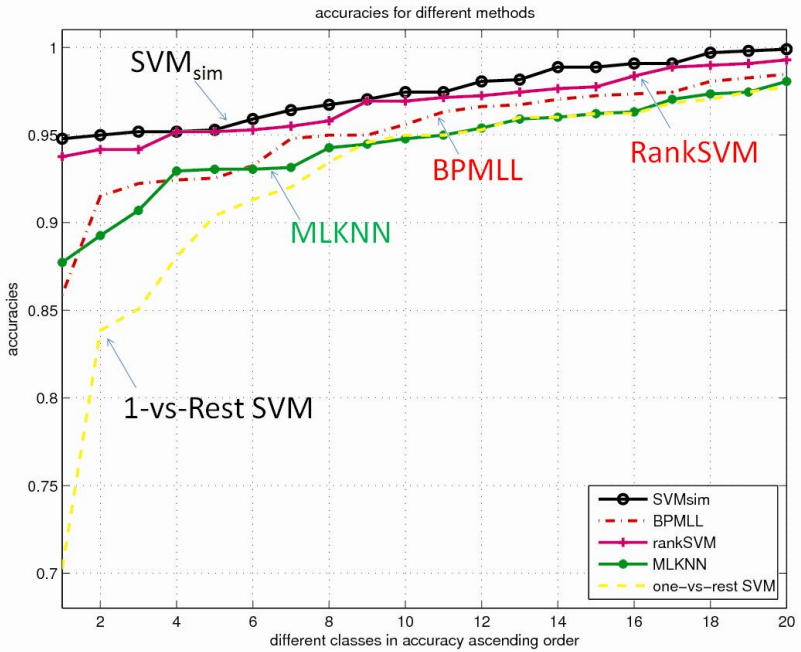
$$\mathtt{s.t.} \qquad -1 \leqslant M(i,j) \leqslant 1 \qquad i,j = \{1,2,\ldots,D\}$$

where:

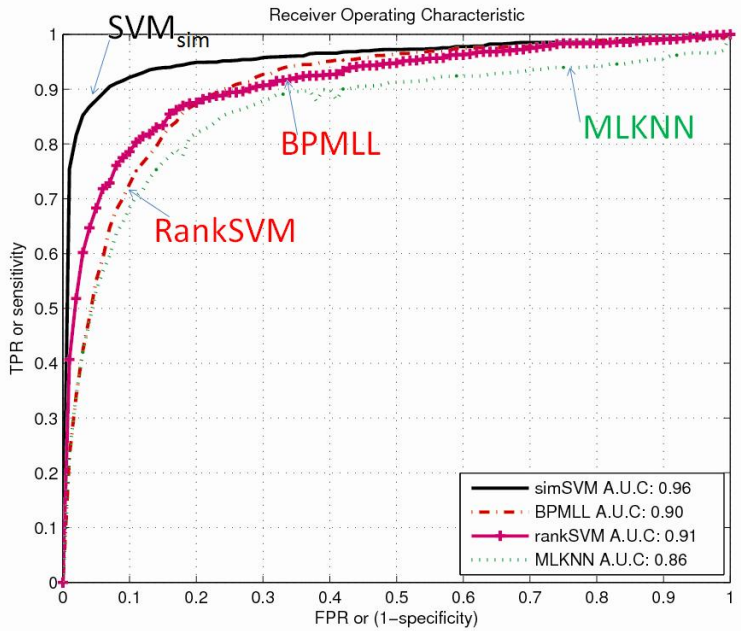
$$W = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_L]; \qquad M = [\mathbf{m}_1^{r\prime}, \mathbf{m}_2^{r\prime}, \cdots, \mathbf{m}_L^{r\prime}]'$$

Experiment Results

ACC. of ICD-9

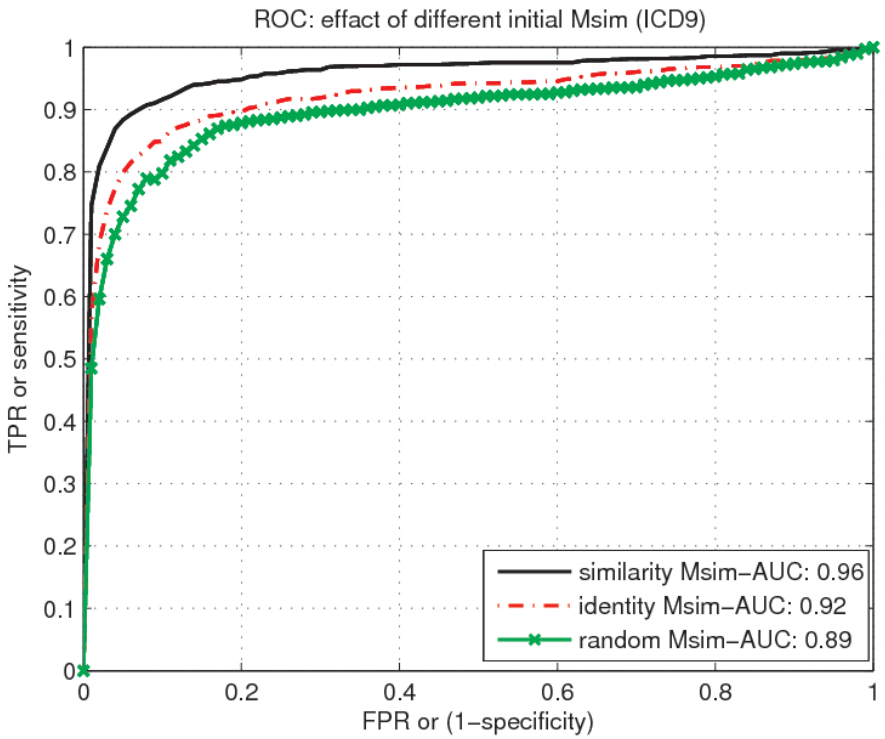


Average ROC of ICD-9



R.O.C for Different Msim

ROC for Different $\mathcal{M}_{\texttt{sim}}$

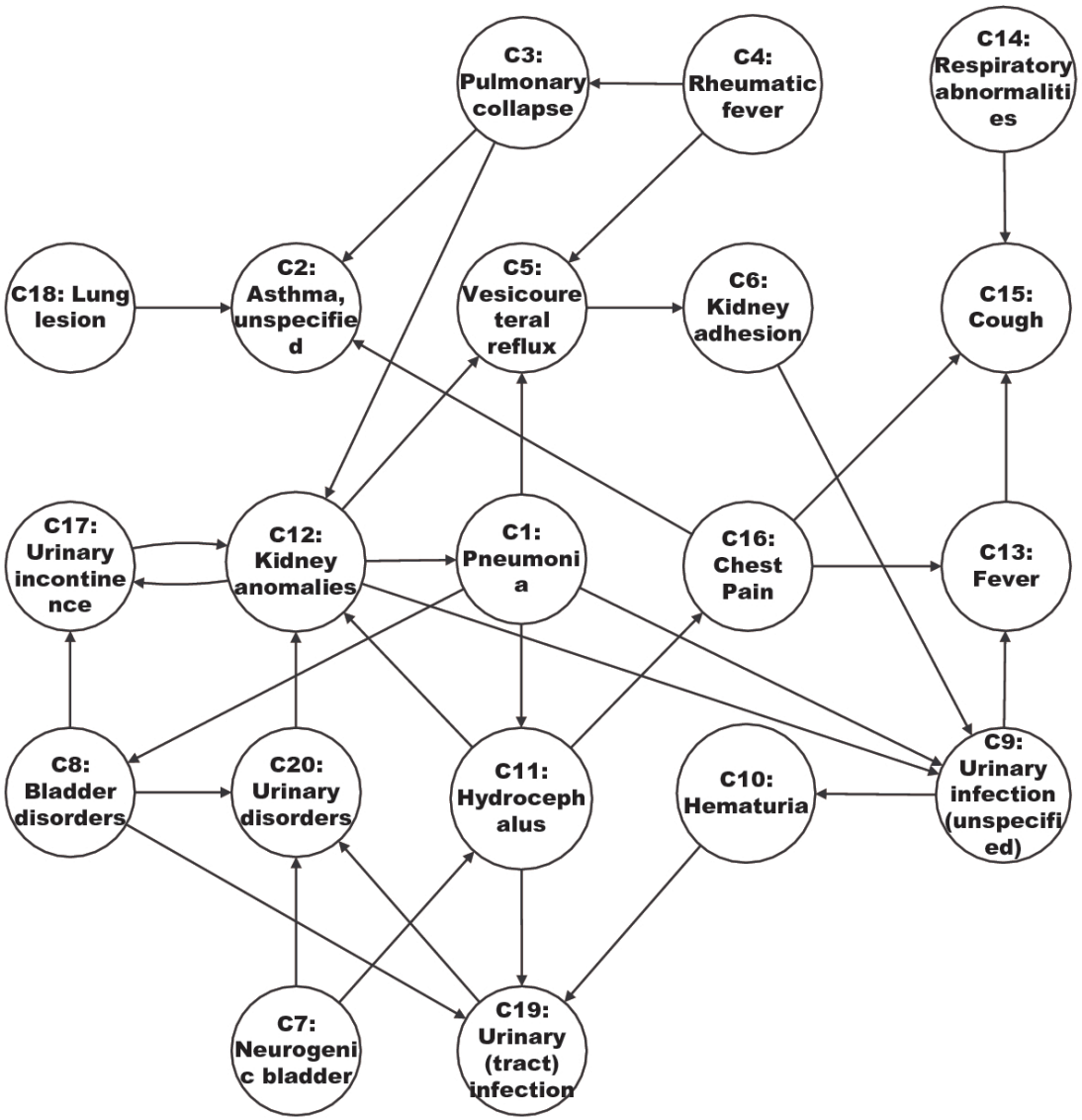


Similarity

Identity

Random

Learned Relations between Classes (Codes)



Feature Selected

Feature with the Largest ABS. Value (Up to 5) per Class

class 1 Pneu-	class 2 Asthma	class 3 Pulm.	class 4 Rheu.	... ...
postvoid	lobe	atelectasis	hydron-	...
brochial	pneumonia	collapse	ephrosis	...
ultrasound	wheezing		cough	...
pneumonia	asthma		atelectasis	...
cystitis	opacity		pyelectasis	...

Summary

Introduced a novel Multi-Label approach that...

1. Makes use of Inter-Class relations and learns the relations
2. Can work even when the samples in each class are sparse
3. Can perform feature selection