A Collaborative Filtering Approach
to Predict Patient Future Disease Risk
from Electronic Health Records

Riccardo Miotto, Li Li, and Joel T. Dudley

Department of Genetics and Genomic Sciences
Icahn School of Medicine at Mount Sinai
New York City
Introduction

• The increasing cost of health care has motivated the drive towards preventive medicine
  ✓ predictive approaches to protect, promote, and maintain health and to prevent diseases, disability, and death

• Secondary use of electronic health records (EHRs)
  ✓ great promise in providing tools to support physicians for identifying potentially negative events

• Automatically predict the risk that patients might develop a certain disease given their clinical status

• Approach
  ✓ collaborative filtering
• State-of-the-art
  ✓ supervised classification
    ▪ difficult to collect labels for diseases
    ▪ sparse feature vectors

• Alternative approach
  ✓ unsupervised recommendation scenario
    ▪ the patients are the users
    ▪ the items are the diseases
    ▪ we recommend diseases based on patient history

• Collaborative filtering
  ✓ analyze relationships between users and interdependencies among products to identify new user-item associations
Collaborative Filtering (1)

- Diagnosis codes in EHRs as disease indication
  ✓ e.g., ICD-9 codes
- Problem
  ✓ fill the matrix with missing codes
• Earlier approaches
  ✓ neighborhood methods
    ▪ centered on computing similarity relationships between items or users
    ▪ not working well
      ➢ data is too sparse
      ➢ time consuming

• Latent factor models
  ✓ characterize both items and users on a small number of factors inferred from the rating patterns
    ▪ **matrix factorization**
    ▪ restricted Boltzmann machines
    ▪ word2vec
    ▪ linear regression
Matrix Factorization (1)

- Patient-factor vector $\mathbf{v}_p$ and disease-factor vector $\mathbf{q}_d$
  - measure the extent to which they possess those factors

- Patient-disease interaction
  - $\mathbf{q}_d^T \mathbf{v}_p$

- Bias
  - patient and disease deviation from the diagnosis average
    - $\mu + b_p + b_d$

- Side details
  - $\mathbf{a}^T \mathbf{x}_p$

- Predicted rate
  - $\mu + b_p + b_d + \mathbf{a}^T \mathbf{x}_p + \mathbf{q}_d^T \mathbf{v}_p$
Matrix Factorization (2)

• Object function

$$\min_{b,a,v,q} \frac{1}{|D|} \sum_{(p,d) \in D} (r_{pd} - \hat{r}_{pd})^2 + \lambda_1 (||b||^2 + ||a||^2) + \lambda_2 (||v||^2 + ||q||^2)$$

• Learning Algorithm

✓ stochastic gradient descent (mini-batch)
  • for each mini-batch
    - compute the predicted rates
    - compute the associated prediction error
    - modify the parameters by a magnitude proportional to a learning rate $\gamma$ in the opposite direction of the gradient

✓ fast running time
  • semi-online algorithm
Dataset (1)

- Mount Sinai Data Warehouse
  - about 4 million patients at March 2015

- Retain patients with at least one ICD-9 code
  - remove codes starting with V and E
  - about one million patients

- Data until December 2011 used as training
  - matrix composed by 799,558 patients and 13,242 codes
    - 1 million non-zero entries (~1%)
      - number of times each ICD-9 code was assigned to each patient

- Data from January 2012 to December 2013 to test
  - 201,764 patients
• Side Details
  ✓ demographic
    ▪ race, year of birth, gender, religion
  ✓ medications, lab tests, procedures, ICD-9 codes starting with V and E
    ▪ normalize to obtain harmonized codes
    ▪ count the number of times each code was assigned to each patient during the training temporal window
  ✓ clinical notes
    ▪ parsed to extract clinical relevant concepts
    ▪ topic modeling to obtain a higher-level semantic dense representation (multinominal)
    ▪ average topic representations over all patient notes
Evaluation (1)

- Map ICD-9 codes to a disease vocabulary
  - different ICD-9 codes can refer to the same disease

- Vocabulary of 140 different diseases
  - in use at Mount Sinai Medical Center
  - the disease risk was the greatest score obtained by the ICD-9 codes associated to that disease

- Experiments
  - assign to each patient the 5-10 most likely diseases
    - evaluate precision, recall, and f-score of the annotations
  - rank patients by their score for each disease
    - evaluate mean average precision (MAP) and Area under the ROC curve (AUC-ROC) of the ranking lists
**Evaluation (2)**

- Assign top 5 diseases to patients
  - ✓ 73% of patients had at least one correct diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UpperBound</td>
<td>0.778</td>
<td>0.733</td>
<td>0.641</td>
</tr>
<tr>
<td>Random</td>
<td>0.167</td>
<td>0.091</td>
<td>0.113</td>
</tr>
<tr>
<td>DiseaseSim</td>
<td>0.195</td>
<td>0.118</td>
<td>0.139</td>
</tr>
<tr>
<td>MatrixFact-NoSide</td>
<td>0.272</td>
<td>0.149</td>
<td>0.190</td>
</tr>
<tr>
<td><strong>MatrixFact</strong></td>
<td><strong>0.381</strong></td>
<td><strong>0.222</strong></td>
<td><strong>0.259</strong></td>
</tr>
</tbody>
</table>
**Evaluation (3)**

- Assign top 10 diseases to patients
  - ✓ 83% of patients had at least one correct diagnosis

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UpperBound</td>
<td>0.588</td>
<td>0.894</td>
<td>0.641</td>
</tr>
<tr>
<td>Random</td>
<td>0.128</td>
<td>0.181</td>
<td>0.142</td>
</tr>
<tr>
<td>DiseaseSim</td>
<td>0.159</td>
<td>0.207</td>
<td>0.181</td>
</tr>
<tr>
<td>MatrixFact-NoSide</td>
<td>0.218</td>
<td>0.267</td>
<td>0.242</td>
</tr>
<tr>
<td><strong>MatrixFact</strong></td>
<td><strong>0.320</strong>*</td>
<td><strong>0.348</strong>*</td>
<td><strong>0.304</strong>*</td>
</tr>
</tbody>
</table>
Evaluation (4)

- Average AUC-ROC over all diseases = 0.695
- Average MAP over all disease = 0.190

- Top 5 Diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>AUC-ROC</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver Diseases</td>
<td>0.842</td>
<td>0.243</td>
</tr>
<tr>
<td>Diabetes Mellitus with Complications</td>
<td>0.834</td>
<td>0.250</td>
</tr>
<tr>
<td>Delirium Dementia and Other Cognitive Disorders</td>
<td>0.829</td>
<td>0.409</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.801</td>
<td>0.529</td>
</tr>
<tr>
<td>Abortion-related Disorders</td>
<td>0.789</td>
<td>0.362</td>
</tr>
</tbody>
</table>
Conclusions

• Collaborative filtering based on latent factors is promising to predict patient disease risk
  ✓ fast and scalable

• Limitations
  ✓ not portable
  ✓ if a disease is not in the matrix, it won’t be predicted
  ✓ scores need to be constantly recalculated

Applications
  • alert primary physicians if a patient is at risk to any disease
  • search across the data warehouse for patients at risks to develop a certain disease
Future Works

- Test other latent factor-based algorithms
  - latent Dirichlet allocation
  - restricted Boltzmann machine
  - word2vec

- Model the temporal sequence of the events

- Ensemble algorithms to define a more robust predictive framework

- Comparison with supervised classification algorithms
  - support vector machine
  - random forest