Forecasting Potential Diabetes Complications

Yang Yang, Jie Tang, Juanzi Li
Tsinghua University

Walter Luyten, Marie-Francine Moens
Katholieke Universiteit Leuven

Lu Liu
Northwestern University
Diabetes Complications

- **Life-Threatening**
  - Over **4.8 million** people died in 2012 due to diabetes[^1].
  - Over **68%** of diabetes-related mortality is caused by diabetes complications[^2].
  - **471 billion** USD, while **185 million** patients remain undiagnosed[^1].

- **Need to be diagnosed in time**

Forecasting Diabetes Complication

Input: a patient’s lab test results

- Bilirubin example
- Routine urine analysis

Output: diabetes complications

- Heart
- Stroke
- PAD (Peripheral Arterial Disease)
- Eye
- CRF (Chronic Kidney Failure)
- Peripheral Neuropathy

- Leading cause
- Increased risk

- Coronary heart disease
- Diabetic retinopathy
Data Set

- A collection of real clinical records from a hospital in Beijing, China over one year.

<table>
<thead>
<tr>
<th>Item</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical records</td>
<td>181,933</td>
</tr>
<tr>
<td>Patient</td>
<td>35,525</td>
</tr>
<tr>
<td>Lab tests</td>
<td>1,945</td>
</tr>
</tbody>
</table>

**Challenge: feature sparseness**

- Each clinical record only contains 24.43 different lab tests
- 65.5% of lab tests exist in < 10 clinical records (0.00054%).
Our Approach
Baseline Model I

Learning task: \( f(x_i) \rightarrow y_i \)

Limitations:
1. Cannot model correlations between \( y \)
2. Cannot handle sparse features
Baseline Model II

Objective function: \[ P(Y|X) = \prod_n P(y_n|x_n) \prod_c P(y_{c1}, y_{c2}) \]

Still cannot handle sparse features!
Proposed Model

Objective function:

\[ O(\lambda) = \sum_n \log P(y_n | \theta_n, x_n) + \sum_c \log P(y_{c1}, y_{c2}) \]

\[ = \sum_n \alpha f(\theta_n, y_n) + \sum_c \beta g(y_{c1}, y_{c2}) \]

\[ + \sum_n \sum_l \log \sum_k \theta_{nk} \Omega_{k,l,x_{nl}} - \log Z \]
**Learning Algorithm**

**Input:** a feature matrix $X$, learning rate $\eta$

**Output:** estimated parameters $\lambda$

Initialize $\alpha$, $\beta$, $\theta$, $\mu$, $\phi$ randomly;

Initialize $\delta \leftarrow 1$;

**repeat**

1. Calculate $P(k_{nl}|x_n, \lambda_n)$ according to Eq. 8;
2. Update $\theta$, $\mu$, $\delta$, $\phi$ according to Eq. 9-12;

Call LBP to calculate $E[\sum_n f(\theta_n, y_n)]$ and $E[\sum_c g(y_{c1}, y_{c2})]$;

Call LBP to calculate $E_{P\alpha}(y|\theta)[\sum_n f(\theta_n, y_n)]$ and $E_{P\beta}(y)[\sum_c g(y_{c1}, y_{c2})]$;

Calculate $\frac{\partial O(\alpha, \beta)}{\partial \alpha}$ and $\frac{\partial O(\alpha, \beta)}{\partial \beta}$ according to Eq. 14;

$\alpha_{new} = \alpha_{old} + \eta \frac{\partial O(\alpha, \beta)}{\partial \alpha}$;

$\beta_{new} = \beta_{old} + \eta \frac{\partial O(\alpha, \beta)}{\partial \beta}$

until **Convergence**;

**Algorithm 1:** Learning algorithm of SparseFGM.
Learning Algorithm (cont.)

• Update the dimensional reduction parameters
  – The remaining part of SparseFGM could be regarded as a mixture generative model, with the log-likelihood

  \[ O(\theta, \phi, \mu, \delta) = \sum_n \sum_{l_1} \log \sum_k \theta_{nk} \frac{\exp\left\{ \frac{-(r_{nl1} - \mu_{kl})^2}{2\delta_k^2} \right\}}{\delta_k \sqrt{2\pi}} \]

  \[ + \sum_n \sum_{l_2} \log \sum_k \theta_{nk} \phi_{kl2} r_{nl2} \]

  – Jensen’s inequality tells us that

  \[ \geq \sum_{n,l_1,k} P(k_{nl}|x_n, \lambda_n) \left[ \log \theta_{nk} - \frac{(r_{nl1} - \mu_{kl})^2}{2\delta^2} \right. \]

  \[ - \log(\delta \sqrt{2\pi} P(k_{nl}|x_n, \lambda_n)) + \sum_{n,l_2,k} P(k_{nl}|x_n, \lambda_n) \]

  \[ \times \left. [\log \theta_{nk} + \log \phi_{kl2} - \log P(k_{nl}|x_n, \lambda_n)] \right] \]

  – Derivate with respect to each parameters, set them to zero, and get the update equations.

\[
\begin{align*}
\theta_{nk} &= \frac{\sum_l \sum_{k_{nl}} P(k_{nl}|x_n, \lambda_n)}{\sum_l \sum_{k_{nl}} P(k_{nl}|x_n, \lambda_n)} + \alpha_k y_n \\
\mu_{kl} &= \frac{\sum_n P(k_{nl}|x_n, \lambda_n)r_{nl}}{\sum_n P(k_{nl}|x_n, \lambda_n)} \\
\delta_k^2 &= \frac{\sum_n \sum_{l_1} (r_{nl1} - \mu_{kl})^2}{N \times L_1} \\
\phi_{klr} &= \frac{\sum_n P(k_{nl}|x_n, \lambda_n)}{\sum_n \sum_r P(k_{nl}|x_n, \lambda_n)}
\end{align*}
\]
Learning Algorithm (cont.)

• Update the classification parameters
  – New log-likelihood
  \[ O(\alpha, \beta) = \sum_n \alpha f(\theta_n, y_n) + \sum_c \beta g(y_{c1}, y_{c2}) - \log Z \]
  – Adopt a gradient descent method to optimize the new log-likelihood

\[
\frac{\partial O(\alpha, \beta)}{\partial \alpha} = E[\sum_n f(\theta_n, y_n)] - E_{P_\alpha(y|\theta)}[\sum_n f(\theta_n, y_n)]
\]
\[
\frac{\partial O(\alpha, \beta)}{\partial \beta} = E[\sum_c g(y_{c1}, y_{c2})] - E_{P_\beta(y)}[\sum_c g(y_{c1}, y_{c2})]
\]

**Algorithm 1:** Learning algorithm of SparseFGM.
Theoretical Analysis

**THEOREM 3.1.** The maximal log-likelihood of SparseFGM is no less than FGM when \( K \geq \max\{|t_n|\} \)

**Proof.** Assuming we have a parameter configuration of FGM \( \tilde{\lambda} = (\tilde{\alpha}, \tilde{\beta}) \), which maximizes FGM’s objective function.

\[
\tilde{O}(\lambda) = \sum_n \tilde{\alpha}_f(r_n, y_n) + \sum_c \tilde{\beta}_g(y_{c1}, y_{c2}) - \log \tilde{Z}
\]

Let \( \theta_{nk} = \frac{\tilde{\alpha}_{rk} \tau_{rk}}{Z_k} \) for \( k \leq |t_n| \) and \( \theta_{nk} = 0 \) for \( k \geq |t_n| \), where \( Z_k \) is a normalization term. Also let \( \alpha_k = Z_k \), \( \beta = \tilde{\beta} \), and for each \( 1 \leq k \leq K \), we select a particular distribution as \( \Omega_k \) such that \( \forall n \), we have \( \sum_k \theta_{nk} \Omega_k = 1 \). Thus, we have

1. \[
\sum_n \alpha f(\theta_n, y_n) = \sum_n y_n \sum_k \alpha_k \theta_{nk}
\]
   \[
   = \sum_n y_n \sum_{i=1}^{t(n)} Z_i \times \frac{\tilde{\alpha}_{rk} \tau_{rk}}{Z_k}
   = \sum_n \tilde{\alpha}_f(r_n, y_n)
\]

2. \[
\log Z = \sum_n \sum_{y_n} \alpha f(\theta_n, y_n) + \sum_c \sum_{y_{c1}, y_{c2}} \beta g(y_{c1}, y_{c2})
\]
   \[
   = \sum_n \sum_{y_n} \tilde{\alpha}_f(r_n, y_n) + \sum_c \sum_{y_{c1}, y_{c2}} \tilde{\beta}_g(y_{c1}, y_{c2})
   = \log \tilde{Z}
\]

3. \[
O(\lambda) = \sum_n \alpha f(\theta_n, y_n) + \sum_c \beta g(y_{c1}, y_{c2})
\]
   \[
   + \sum_n \sum_l \log \sum_k \theta_{nk} \Omega_{k,l,r_{nl}} - \log Z
   = \sum_n \tilde{\alpha}_f(r_n, y_n) + \sum_c \tilde{\beta}_g(y_{c1}, y_{c2}) - \log \tilde{Z}
   = \tilde{O}(\tilde{\lambda})
\]

\[\max_{\lambda} \{O(\lambda)\} \geq O(\lambda) = \tilde{O}(\tilde{\lambda}) = \max_{\lambda} \{\tilde{O}(\lambda)\}\]
Experiments
Setting

• Experiments
  ▪ Is our model effective?
  ▪ How do different diabetes complications associate with each lab test?
  ▪ Can we forecast all diabetes complications well?

• Comparison Methods
  • SVM (model I)
  • FGM (model II)
  • FGM+PCA (an alternative method to handle feature sparseness)
  • SparseFGM (our approach)
Experimental Results

**HTN:** hypertension, **CHD:** coronary heart disease, **HPL:** hyperlipidemia

<table>
<thead>
<tr>
<th>Location</th>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTN</td>
<td>SVM</td>
<td>0.3804</td>
<td>0.4789</td>
<td>0.4241</td>
</tr>
<tr>
<td></td>
<td>FGM</td>
<td>0.5666</td>
<td>0.4959</td>
<td>0.5075</td>
</tr>
<tr>
<td></td>
<td>FGM+PCA</td>
<td>0.5741</td>
<td>0.3284</td>
<td>0.4178</td>
</tr>
<tr>
<td></td>
<td>SparseFGM</td>
<td>0.4714</td>
<td>0.6319</td>
<td>0.5400</td>
</tr>
<tr>
<td>CHD</td>
<td>SVM</td>
<td>0.2132</td>
<td>0.0636</td>
<td>0.0980</td>
</tr>
<tr>
<td></td>
<td>FGM</td>
<td>0.6264</td>
<td>0.1369</td>
<td>0.2247</td>
</tr>
<tr>
<td></td>
<td>FGM+PCA</td>
<td>0.2425</td>
<td>0.8367</td>
<td>0.3761</td>
</tr>
<tr>
<td></td>
<td>SparseFGM</td>
<td>0.2522</td>
<td>0.7972</td>
<td>0.3832</td>
</tr>
<tr>
<td>HPL</td>
<td>SVM</td>
<td>0.2208</td>
<td>0.0460</td>
<td>0.0761</td>
</tr>
<tr>
<td></td>
<td>FGM</td>
<td>0.6557</td>
<td>0.0591</td>
<td>0.1084</td>
</tr>
<tr>
<td></td>
<td>FGM+PCA</td>
<td>0.2047</td>
<td>0.8035</td>
<td>0.3262</td>
</tr>
<tr>
<td></td>
<td>SparseFGM</td>
<td>0.2796</td>
<td>0.8396</td>
<td>0.4195</td>
</tr>
</tbody>
</table>

- SVM and FGM suffer from feature sparseness. **-59.9%** in recall.
- FGM vs. FGM + PCA (increase **+40.3%** in recall)
- PGM+PCA vs. SparseFGM (increase **+13.5%** in F1)
WBC in the urine causes frequent voiding -> no good sleep at night

Association score: \[ AS(c, e) = \sum_k \alpha_k c \theta_{ek} \]
c: complication, e: lab test
Can We Forecast All Diabetes Complications?

HPL can be forecasted precisely based on lab test results.
Conclusion

• We study the problem of forecasting diabetes complications.

• We propose a graphical model which integrates dimensional reduction and classification into a uniform framework.

• We further study the underlying associations between different diabetes complications and lab test types.
Thanks!
Q&A?

@ Yang Yang
http://yangy.org/