LEARNING RICH GEOGRAPHICAL REPRESENTATIONS: PREDICTING COLORECTAL CANCER SURVIVAL IN THE STATE OF IOWA

BIBM 2017

Michael T. Lash¹, Yuqi Sun¹, Xun Zhou², Charles F. Lynch³, and W. Nick Street²

¹Department of Computer Science, ²Department of Management Sciences, ³Department of Epidemiology

www.michaeltlash.com
Colorectal Cancer in Iowa

Colorectal Cancer Mortality Rate by ZCTA in Iowa: 1989 to 2013

Legend
- Iowa Mortality
- 0.0 - 0.145
- 0.145 - 0.307
- 0.307 - 0.447
- 0.447 - 0.714
- 0.714 - 1.0

created by: Michael T. Lash
Colorectal Cancer in Iowa

Colorectal Cancer Mortality Rate by ZCTA in Iowa: 1989 to 2013

Legend:
- Iowa Mortality
  - 0.0 - 0.145
  - 0.145 - 0.307
  - 0.307 - 0.447
  - 0.447 - 0.714
  - 0.714 - 1.0

created by: Michael T. Lash
Colorectal Cancer in Iowa

Colorectal Cancer Mortality Rate by ZCTA in Iowa: 1989 to 2013

Legend
- Iowa Mortality
  - 0.0 - 0.145
  - 0.145 - 0.307
  - 0.307 - 0.447
  - 0.447 - 0.714
  - 0.714 - 1.0

created by: Michael T. Lash
Colorectal Cancer in Iowa

Colorectal Cancer Mortality Rate by ZCTA in Iowa: 1989 to 2013

Legend

Iowa Mortality

- 0.0 - 0.145
- 0.145 - 0.307
- 0.307 - 0.447
- 0.447 - 0.714
- 0.714 - 1.0

created by: Michael T. Lash
Colorectal Cancer in Iowa

- **Takeaway**: Geography appears to be indicative of colorectal cancer survivability.
Colorectal Cancer in Iowa

- **Takeaway**: Geography appears to be indicative of colorectal cancer survivability.
- **Furthermore**:
  - Lead-based paint in pre-1970’s housing.
  - Rural healthcare availability.
  - Attitudes surrounding healthcare.
Problem: Predict patient-specific colorectal cancer survival curves (KMSC).
Colorectal Cancer in Iowa

- **Problem**: Predict patient-specific colorectal cancer survival curves (KMSC).
- **This work**: Does geography aid in the prediction of survival curves and do richer geographical representations produce more accurate predictions?
Censored Data & Survival Curve Re-representation

- **Data:** \( \{x^{(i)}, e^{(i)}, t^{(i)}\}_{i=1}^{n} \)
- \( e^{(i)} \in \{0, 1\} \leftarrow \text{Event indicator} \)
- \( t^{(i)} \in \{1, \ldots, T\} \leftarrow \text{Discrete time} \)
Data: \( \{x^{(i)}, e^{(i)}, t^{(i)}\}_{i=1}^n \)

\( e^{(i)} \in \{0, 1\} \leftarrow \text{Event indicator} \)

\( t^{(i)} \in \{1, \ldots, T\} \leftarrow \text{Discrete time} \)

Three different “scenarios” w.r.t. \( e^{(i)} \) and \( t^{(i)} \):

1. \( e^{(i)} = 1 \)
2. \( e^{(i)} = 0, t^{(i)} < T \)
3. \( e^{(i)} = 0, t^{(i)} = T \)
- **Re-represent** $e^{(i)}$ and $t^{(i)}$ as a vector $y^{(i)}$.
- Where $\tilde{p}_i = 1 - P \left( e^{(i)}_{\tilde{t}} = 0 | e^{(i)}_{\tilde{t}-1} = 1 \right)$

| $e^{(i)} = 1$ | $y^{(i)} = \begin{bmatrix} 1 \cdots 1 0 0 \cdots 0 \end{bmatrix}$ |
| $e^{(i)} = 0$, $t^{(i)} < T$ | $y^{(i)} = \begin{bmatrix} 1 \cdots 1 \tilde{p}_{\tilde{t}=t^{(i)}} \cdots \tilde{p}_{\tilde{t}=T} \end{bmatrix}$ |
| $e^{(i)} = 0$, $t^{(i)} = T$ | $y^{(i)} = \begin{bmatrix} 1 \cdots 1 1 1 \cdots 1 \end{bmatrix}$ |
A result

Average Actual vs Average Predicted KMSC: No Geo Feats

- ABC
- Five years
- Avg. $y \pm \frac{1}{4}$ st. dev.
- Avg. $\hat{y} \pm \frac{1}{4}$ st. dev.

Discrete Time (one unit=6 months)

Probability
Problem Formulation

\[ g^* = \arg\min_{g \in \mathcal{G}} \left\{ \mathcal{L} \left( y^{(i)}, g(x^{(i)}) \right) : i = 1, \ldots, n \right\} \]  

- \( \mathcal{G} \) is defined to be a neural network hypothesis set.
- **Goal:**
  - See whether the elicited \( g^* \) is better when geographical features are added.
  - See whether the elicited \( g^* \) is better when richer geographical representations are used.
Geographic Representations

- A simple binary representation (SBR).
- A rich, spectral analysis-elicited representation (RR-SA).
- Assume: Can compute discrete geographic entity-membership using original geographic features ($x_z$).
Geographic Representations: SBR
Geographic Representations: SBR
Geographic Representations: SBR

Input node
Hidden node
Output node
(logistic)

$X^{(i)}_{Z}$

Bin

Smooth

$\hat{y}^{(i)}$
A Quick Aside: Output Smoothing

- **Key insight**: Probability of survival never increases from $\tilde{t}$ to $\tilde{t} + 1$.

$$\hat{y}_{\tilde{t}+1} = \min\{output_{\tilde{t}}, output_{\tilde{t}+1}\} \text{ for } \tilde{t} = 1, \ldots, T \quad (2)$$
Geographic Representations: RR-SA
Geographic Representations: RR-SA
Geographic Representations: RR-SA

- $Q_{spec} \in \mathbb{R}^{k \times p}$, where $k$ is user-specified.
Geographic Representations: RR-SA

- Q_{spec} \in \mathbb{R}^{k \times p}, where k is user-specified.
Geographic Representations: RR-SA

\[ X_{Z_i} \]

Input node
RR-SA Feats
Hidden node
Output node (logistic)

Adj
Top_k
Enrich

Smooth
\[ \hat{y}(i) \]
Experiments

- Compare average cross-validation $\hat{y}$ with average $y$ using two different measures:
  - Mean Absolute Error (MAE)
  - Area Between Curves (ABC) — A new measure
Results: Predictions

(a) No geo feats 
(ABC=14.32, MAE=0.467).

(b) SBR
(ABC=12.60, MAE=0.4512).

(c) RR-SA, \(k = 10\)
(ABC=11.41, MAE=0.446).

(d) RR-SA, \(k = 20\)
(ABC=12.31, MAE=0.453).

(e) RR-SA, \(k = 30\)
(ABC=11.65, MAE=0.445).

(f) RR-SA, \(k = 40\)
(ABC=10.77, MAE=0.442).
Results: Spectral Clustering

Spectral Clustering: k=10

Spectral Clustering: k=20

Spectral Clustering: k=30

Spectral Clustering: k=40
Conclusions

- Geographical features improve colorectal cancer survival curve predictions.
- Richer, spectral analysis-elicited features provide better predictions than simple, binary representations.
  * Predictive performance deviates at, approximately, the five-year mark:
    - Future work: Improve on these predictions by exploring other geographical representations.
LEARNING RICH GEOGRAPHICAL REPRESENTATIONS: PREDICTING COLORECTAL CANCER SURVIVAL IN THE STATE OF IOWA

BIBM 2017

Michael T. Lash$^1$, Yuqi Sun$^1$, Xun Zhou$^2$, Charles F. Lynch$^3$, and W. Nick Street$^2$

$^1$Department of Computer Science, $^2$Department of Management Sciences, $^3$Department of Epidemiology

www.michaeltlash.com