Weight of Evidence Inference and Bayesian Nomograms Using Genomic Feature Pairs

Sebastian Gomez
Mentor: Dr. Pablo Tamayo
Why do we care?

• Despite significant progress in recent years clinical prediction and stratification of risk in patients remains a challenge.

• The focus has shifted from clinical parameters to molecular markers. e.g. expression of specific genes and selected genomic abnormalities.
Purpose of the research

• Develop a way of predicting a particular outcome based on several input features (risk factors)

Examples:

• Clinical Prognosis: Predicting the probability of responding to treatment in different cancer types
Our predictions are based on evidence provided by features

• Over-expression of gene or pathway
• Genomic abnormality: amplification
• Genomic abnormality: deletion

Types of outcomes we want to predict
• Response to treatment (e.g. chemo)
• Platinum Status (platinum sensitive or resistant)
Pathway rather than gene features are used in model

Gene 1
Gene 2
........
Gene N

Input Dataset

GENE FEATURES

Single-sample
GSEA
Signature
Projection

Projected Dataset

PATHWAY FEATURES

Signature 1
Signature 2
........
Signature M

Molecular Signatures
database (MSigDB)
(www.broadinstitute.org/msigdb/)

GENE FEATURES

GSEA Signature Projection
Survival of platinum resistant patients versus sensitive patients in ovarian cancer

70 Resistant patients vs 156 Sensitive patients
Likelihood ratio 51.61, p-value 0

Platinum Resistant: recurrence of cancer within the first 6 months after platinum therapy.

Platinum Sensitive: No recurrence of cancer within the first 6 months after treatment.
Nomograms are a graphical way of representing a classification model.
Ovarian cancer Bayesian nomogram for a specific patient. Showing 5 out of 12 total single features (current model used)

- Regulated by LIF treatment
- Involved in focal adhesion
- Down-regulated by AKT
- WNT target genes from literatures

Cumulative Log Odds (CLO) = -3.2 -> 0.36%

Sensitive
Resistant
Current model for ovarian cancer has two main issues: features are strongly correlated and there are a lot of errors made even when sufficient evidence is available.

<table>
<thead>
<tr>
<th>Catégories</th>
<th>Sensitive</th>
<th>Resistant</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

**Diagram**

- **Cumulative Log Odds (CLO)**
- **Amount of Absolute Evidence**

- Sensitive
- Resistant
One way of improving these results is using pairs of features instead of single features

• Important things to know about dealing with pairs of features:

  • The evidence of a pair is not the same as the sum of the individual evidences.

  • Before, we had n features, now we have \((n^2 - n)/2\) possible pairs of features.

  • Pairs not only perform better, they also have the potential to uncover interesting
How did I approach this problem?

Find the top 100 features (out of ~3 million) based on subset based on their quality and unambiguous annotation.

The longer the arms of the nomogram, the better the feature.

Create the nomogram based on selected features.
Ovarian cancer pairs work well because they correct each other’s mistakes and they reinforce each other when correct.

281 Patients

Feature 1: genes down regulated by AKT (oncogene)
Feature 2: genes up regulated by KRAS (oncogene)

Intensity of color represents the certainty of the prediction. Red represents sensitive prediction and blue represents resistant prediction.
Here’s how we visualize pairs of features in a nomogram.

Feature 1
- down-regulated by AKT
- regulated by KRAS

Feature 2

State of feature 1

State of feature 2

-0.86
Ovarian cancer Bayesian nomogram - pairs

MMS (drug) induced lymph genes + genes involved in focal adhesion
B cell genes regulated by CD44 + Age-regulated genes in the human frontal cortex
down-regulated by AKT + up-regulated by KRAS
up-regulated by mercaptopurine + sarcoma associated genes
A-KAP13 pathway + genes up-regulated by 4NQO (drug)
down-regulated by loss of LKB1 + down-regulated by infection with cytomegalovirus
Ovarian cancer pairs show a lesser degree of correlation than single features
Comparison of models for ovarian cancer. Both ROC and Error rate improvement from using pairs

<table>
<thead>
<tr>
<th>Model</th>
<th>au-ROC</th>
<th>Error Rate</th>
<th># of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singles</td>
<td>0.595</td>
<td>42.3%</td>
<td>12</td>
</tr>
<tr>
<td>Pairs</td>
<td>0.681</td>
<td>30.7%</td>
<td>6 pairs</td>
</tr>
</tbody>
</table>

**Real Outcome**

<table>
<thead>
<tr>
<th>Platinum Status</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resistant</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Sensitive</td>
<td>73</td>
<td>29</td>
</tr>
</tbody>
</table>

Prediction
Comparison of models for Medulloblastoma

<table>
<thead>
<tr>
<th>Model</th>
<th>au-ROC</th>
<th>Error Rate</th>
<th># of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.727</td>
<td>32.1%</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>0.747</td>
<td>34.6%</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0.802</td>
<td>25.6%</td>
<td>46</td>
</tr>
<tr>
<td>D</td>
<td>0.782</td>
<td>25.6%</td>
<td>53</td>
</tr>
<tr>
<td>Pairs</td>
<td>0.803</td>
<td>29.5%</td>
<td>5 pairs + 2 singles</td>
</tr>
</tbody>
</table>
Conclusion

• Pairs are better at predicting outcome

• The fact that pairs “work together” might uncover interesting interactions between the features

• Pairs dramatically reduce the correlation in the model

• Naïve Bayesian is a simple approach that physicians can understand and is as accurate as more complex models
Data Acknowledgement

Datasets:

• Medulloblastoma Dataset
  • Training Set: 94 samples multi-institutional
    • Children’s Hospital (Boston)
    • Children Oncology Group (COG)
    • U of Washington Medical Center
    • Children’s Hospital (Texas)
    • The Johns Hopkins Medical Center

• Independent test set: 78 samples multi-institutional
  • 47 samples, Pomeroy et al. 2002
  • 16 samples, Kool et al. 2008
  • 15 samples, COG 2009
Data Acknowledgement

Datasets:

- Ovarian Cancer Datasets
  - Toothill
  - TCGA
Acknowledgements

I would like to thank the following people for their contributions:
Roel Verhaak
Jill Mesirov
Todd Golub
Scott Pomeroy

I would also like to thank the following people for all their support.
Eboney Smith
Jacqueline Nkuebe
Bruce Birren
Probabilistic Model

• Conditional Probabilities & Bayes Theorem for independent features:

\[
P(r | x_1, x_2, ..., x_n) = \frac{P(r, x_1, x_2, ..., x_n)}{P(x_1, x_2, ..., x_n)} = P(r | x_1)P(r | x_2)\cdots P(r | x_n)
\]

• Weight of Evidence for target r and state x:

\[
Ev(r=\text{yes} | x=\text{female}) = \log \left( \frac{P(r = \text{yes} | x = \text{female}) / P(r = \text{no} | x = \text{female})}{P(r = \text{yes}) / P(r = \text{no})} \right)
\]

• Total evidence that feature x provides:

\[
AvEv(r | x) = \sum_{i=1}^{|X|} P(x = X_i) |Ev(r | x = X_i)| \quad \text{Where } X = \{\text{Male, Female}\}
\]
Pair Model

Computing evidence for pairs

• Very similar to computing the evidence for singles

\[
\text{Ev}(r=\text{yes} \mid x=\text{female}, y=\text{adult}) = \log \left( \frac{P(r=\text{yes} \mid x=\text{female}, y=\text{adult}) / P(r=\text{no} \mid x=\text{female}, y=\text{adult})}{P(r=\text{yes}) / P(r=\text{no})} \right)
\]

• Remember Bayes Theorem:

\[
P(r=\text{yes} \mid x=\text{female}, y=\text{adult}) = \frac{P(r=\text{yes}, x=\text{female}, y=\text{adult})}{P(x=\text{female}, y=\text{adult})}
\]

• Total evidence that pair of features x, y provides:

\[
\text{AvEv}(r \mid x, y) = \sum_{i=1}^{\mid X \mid} \sum_{j=1}^{\mid Y \mid} P(x=X_i, y=Y_j) \mid \text{Ev}(r \mid x=X_i, y=Y_j)
\]

Where \( X = \{ \text{Male, Female} \} \)

Where \( Y = \{ \text{Adult, Child} \} \)