METCC: METric learning for Confounder Control
Making decision matter in high dimensional biological analysis

Kabir Manghmani, Adam M. Drake, Nathan Wan, Imran S. Haque
Freenome, Inc., South San Francisco, CA

BACKGROUND
- High-dimensional data acquired from biological experiments, such as next-generation sequencing of cfDNA in blood plasma, are subject to a number of confounding effects.
- These confounders pose a challenge when developing solutions to pattern recognition problems using biological data because they can obscure the biological signal of interest. Visualizations of data show institution-specific clustering/shifting.

OBJECTIVE
We analyze:
1. the extent to which data normalized with HCP and METCC retains information about the unwanted technical effects; and
2. the performance of supervised models trained on normalized data.

METHODS
1. Network Architecture
   - As an example, previous studies found that an estimated 32% of variability in biological signal of interest. Visualizations of data show institution-specific clustering/shifting.

2. EXPERIMENTAL SETUP
   - We use cross-validation of cfDNA reads that align to CHEST gene bodies from 817 sample (Wan et al., 2018). In addition to a disease label of healthy or CRC, each sample has associated an institution where it originated, age when the blood was drawn, and batch which it was processed. We grouped age into the bins: [0-5, 5-5, 55-65, 70-75, 80-85, 85+],
   - We apply all three methods to generate a set of embeddings for each overlap (CNV) cross-validation data (k=4). We subsequently trained classifiers to predict each of the 4 sets of labels using both K-Nearest Neighbor (KNN) and logistic regression (LR).

3. METHODS, cont’d
   - High-dimensional data from biological experiments, such as next-generation sequencing of cfDNA in blood plasma, are subject to a number of confounding effects.
   - We propose a Metric Learning based model to normalize out confounder signal.
   - Metric learning methods are advantageous because the loss function:
   - maximizes distance between representations of examples with different label (L ≠ L)
   - minimizes distance between representations of examples that have the same label (L = L)

Intuition:
- Penalizes variance that is not correlated to disease prediction task (e.g. batch effects) We seek to learn a distance function

\[ \text{parameterized by the map } \theta \]

\[ \text{to which data normalized with HCP and METCC retains } \]

\[ \text{unwanted information about the wanted technical effects.} \]

RESULTS

Figure 3. Comparison of sINE of HCP and METCC embeddings

- We are able to train METCC embeddings where label distance is learned. Visualized with sINE where one fold of embeddings normalized by HCP and METCC respectively.
- METCC embeddings when trained with LR achieved comparable performance.
- With KNN, disease performance for exceeded other embeddings, but performance on covariate predictions were not significantly different.

Figure 4. Disease status / Covariate Prediction Task Performance

- Performance drop in disease classification can be interpreted as a proxy measure of how much information is lost to the normalization task. Prediction of other labels can be interpreted as a measure of how much confounder information is left in the embeddings after normalization.

Table 1. Mean k-fold Prediction Task Performance (k=4)

<table>
<thead>
<tr>
<th>Normalization</th>
<th>Train ACC (HSK)</th>
<th>Test ACC (HSK)</th>
<th>Train ACC (LR)</th>
<th>Test ACC (LR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA-only</td>
<td>0.95</td>
<td>0.71</td>
<td>0.95</td>
<td>0.70</td>
</tr>
<tr>
<td>HCP</td>
<td>0.90</td>
<td>0.82</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>METCC</td>
<td>0.99</td>
<td>0.87</td>
<td>0.90</td>
<td>0.80</td>
</tr>
</tbody>
</table>

CONCLUSIONS

- METCC may not lose information about age due to the fact that the disease label is a clear confounder, which was used to train METCC embeddings.
- Future work intends to compare block box metric learning models to non-linear mixed effects model.
- We intend to do further exploration into the question of how best we can compare a supervised approach, like METCC that separates label information from age, to embeddings generated by unsupervised techniques.

ACKNOWLEDGMENTS
The authors gratefully acknowledge Koen Fransen, Ghislain Rakita and David Weinberg for their extensive suggestions, feedback, and editorial support.

REFERENCES

Presented at the 11th annual RECOMB/ISCB Conference on Regulatory & Systems Genomics, December 8-10, 2018, New York, NY USA