Team Meeting

01 APRIL 2022 / 11:00 AM / Zoom

Attendees

Zexi, Stefan, Xavier, Ate, Nancy, Jiadong, Shuai, John

Agenda

Introduction of the team members and Jiadong — 5mins

1. Zexi (Team Leader)
2. Ate
3. Nancy
4. Xavier
5. Stefan

Ask for client's general understanding of this project — 10mins

1. Aim: to train an epigenetic predictor to recognize MZ twins and show that this tool can predict with reasonable accuracy if an individual is an MZ twin based on a blood or buccal DNA methylation profile, without the need to obtain DNA from both co-twins.
2. Methods: The paper first checks if there are differences between MZ and DZ' s methylation levels by using various statistical test methods, then they decide to use a classification model to predict
3. The Classification Model used: penalized regression model (elastic net)

   The alpha parameter of glmnet was set to 0.5 (elastic net regression) and the lambda value was selected by taking the minimum lambda using 10-fold cross-validation on the training data with the AUC method (R command: cv.glmnet(x = methylation, y = zygosity, alpha = 0.5, nfolds = 10, family = "binomial", type.measure = "auc")).

4. Results : The area under the curve (AUC) of the best- performing predictors were 0.77 and 0.80, respectively, in an independent blood data set from NTR (N ~ 1000) and in blood data from a second independent twin cohort (N=606, BSGS). The predictor performed similarly on methylation data from buccal (N = 1237) and on methylation data from monochorionic or dichorionic MZ twins.

Questions for client — 40mins

1. Verified the classification model in the main nature paper
   - Penalized regression model, elastic net?

  - But the codes provided by the nature paper used Partial Least Square regression, can not find the glmnet model claimed in the paper
2. Ask if they are working on the same data as the main paper or a different one, if a different one?
   - General: specific things about the dataset that we need to know if we don’ t know biology
   - Sampling methods: From the MZ twin, they take one person, from the DZ twin, they take both of the twins. Are the sampling methods the same if we use different datasets? (All MZ twins and DZ twins should be included in this study)
   - Data Cleaning: Different filtering methods, p-value, responding to the missing values (only need to remove the given variables)
3. Ask if we are testing the same classifier (used in the paper) on different datasets or building a different but better classifier?
4. Ask them to explain the data in the sample one they gave us (and the corresponding data dictionary if applicable)?
5. What is the state of art standard data processing procedures for their dataset?
6. How accurate would a classifier suffice to prove the hypothesis? What is the client’ s expectation of this? What level of accuracy and robustness is considered good in this domain?
7. What programming language should be used? Python? R?
8. Clarify the expected outcomes? Is this project a small part of a bigger project? If so, What would be the appropriate timeline for this project so that it can fit into this bigger project?
9. How many datasets (quantity) are available for us to use? Will these datasets come from different centres like what’ s on the paper? (Small datasets should be used as validation instead of training)
10. Do they also have another research group that is doing similar things?

Communication Tool

Slack

Next meeting schedule — 5mins

Fortnightly

Notes

- Data integration?

- Heterogeneity => Batch effects, one batch measure 96 rows, dealt with already
- Statistical view is 0.8 auc , but not easy to to achieve
- Consistency of the predictors, train from big datasets first, to smaller dataset

Action Items

Meeting closes at 12 PM

Next Meeting Agenda