

Machine Learning Final Report

Team22

Abstract

Epigenetics is a mystery and an important part of biology; it makes influences how we behave, how we think, and even how we die. Lots of data are generated from the complex system of our body. Thus, machine learning is ideal for finding the pattern of both cause and effect in our epigenetic system.

The main reason we generate the idea of analysis of high dimensional data is that Horvath makes an age predictor (1). He found that age is highly related to the methylation level of our body. However, the feature number compared to the number of data set is far from enough. So he made a feature selection is then found not the most ideal. Thus having a proper feature preprocess becomes a good question to ask.

To solve the problem Horvath met, we set our aim of surveying paper to discuss the method of how and when to make a preprocess of the feature so as to make machine learning applicable for epigenetic analysis. We also discuss some methods of how the processed feature is trained.

Introduction

The main reason for these papers is to have a proper discussion of how the feature is fed into the machine train. For example, among the 151 possible features for death risk predictor, IMPACT [2] defines a distance to measure the relation, then use hierarchical clustering the filter out features that are too closely related in order to make a prediction with certain bias on features.

In the comparison, the others [3] do not have enough discussion on feature selection. It uses medical reasons to filter out less related features. No numerical reason is provided to discuss the quality of selection.

The most important feature selection for methylation is discussed in [4]. He uses a correlation ratio to make 250,000 accessible methylation loci to select the feature and uses matrices to make loci on the same gene reduce to one same feature, thus reducing to 7000 features. Compared to the Horvath method, [4] prediction accuracy is better even using the same data.

Background

Methylation

DNA methylation alters the activities of DNA segments without changing the sequence, which thus yields a wide variety of roles in the cellular processes across organisms or tissues. In addition, DNA methylation appears essential for normal development.

6mA

The 6mA refers to a biological process where the methyl group is attached to the 6-th nitrogen atom of adenine by the enzyme of DNA methyltransferase. The 6mA is a type of non-canonical DNA modification because it might occur in other nucleotide molecules, including mRNA, tRNA, rRNA, small nuclear RNA (snRNA), as well as long non-coding RNA.

T-value, p-value and significant difference

P-value is a statistical index that can be used to determine whether the class result induces a significant difference in a feature. To understand how the p-value is calculated, we should first introduce the t-value. The equation for the t-value is provided below:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In brief, the t-value is the difference of a feature between two classes of objects normalized with respect to their standard deviation. A larger t-value implies that the difference in the feature between the two classes is more significant. P-value is simply the probability the variable is larger than the t-value under Student's t-distribution. A lower p-value implies that it is more likely there are differences between the two sets. We can set a threshold p-value (normally 0.05). When a feature has a lower p-value than the threshold, we say that there is a significant difference and select the feature.

Bonferroni correction

Using the p-value to select a hypothesis gives an upper bound on type 1 error (wrongly rejecting the null hypothesis). However, when there are multiple hypotheses, the type 1 error rate increases. Bonferroni correction is a conservative way to protect the model from such an effect. If there are n hypotheses, the threshold p-value is divided by n (harder to show a significant difference). However, the method is not perfect. Because the correction is too conservative on type 1 errors, it makes the model vulnerable to type 2 errors.

SMOTE

Smote is a method to enlarge the minority data to deal with an unbalanced dataset. First, for a minority data point, find its neighbors. Then, randomly choose one from the neighbor set and create an artificial data point between them. This makes the cost of making mistakes in the minority class higher.

ICC

In brief, the interclass correlation coefficient (ICC) is a number between 0 and 1, used to represent how similar members in a small group are, compared with the entire dataset.

In brief, the coefficient is the variance of interest divided by the total variance. For example, if we want to evaluate how reliable a test is, we can repeat the test on the same sample to create replicates. The variance of interest is the variance between different samples, and the total variance is the variance of interest plus the variance among replicates.

Attention

Attention is initially applied to language-related tasks. Rather than doing training sequentially (like RNN), attention gives the output directly based on the entire input. However, for outputs at different positions, the weight of each input is different. This solves the problem that RNN cannot be trained parallelly and have a similar effect with RNN structure.

Simulated Annealing:

Simulated annealing is a non-deterministic approach to optimize the given problem. The algorithm perturbs the current solution and evaluates the result. If the result

becomes better, the new state is accepted. If not, the algorithm accepts the solution with a probability function as below:

$$p = e^{-\frac{\Delta_c}{T}}$$

Where $\Delta_c = C_{next} - C_{previous}$ and T is a parameter decreases over iterations. The property of accepting a worse solution gives the algorithm ability to escape from the local minimum. When used as a feature selection method, the neighborhood structure used in perturbation is the features chosen, and the cost is the training loss.

Discussion

In the below section, we will go through the result of each paper and provide discussion about their methods and models.

Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species

Related Work

2 major ways to detect DNA 6mA sites:

- Wet experiments
 - mass spectrometry / methylation-specific polymerase chain reaction / single-molecule real-time sequencing
 - Times consuming / expensive / labor-intensive
 - Dry experiments - computational methods
 - learning a classifier: feature-based and the deep learning-based methods
1. MM-6mAPred: identifying DNA N6- methyladenine sites based on Markov model
 2. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome
 3. i6mA-DNCP: computational identification of DNA N6- methyladenine sites in

the rice genome using optimized dinucleotide-based features

4. iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice
5. 6mA-RicePred: a method for identifying DNA N6-methyladenine sites in the rice genome based on feature fusion
6. csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule
7. 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes
8. 6mA-Pred: identifying DNA N6- methyladenine sites based on deep learning

Summary

The result below is the experience on 6mA-rice-LV(rice) dataset, and this paper method is Deep6mAPred. They used 5-fold cross validation on this data(6mA-rice-LV). In the original context, they said:

The Deep6mAPred reached better S_n than three baseline methods (Deep6mA , SNNRice6mA-large and Deep6mAPred), and achieved competitive SP, ACC and MCC in contrast with the Deep6mA, which completely outperformed the SNNRice6mA-large and MM-6mAPred.

However, the fun fact is the performance of **S_p , ACC, MCC, AUC** is not good enough in this dataset.

Performances by 5-fold cross validation over the **6mA-rice-Lv**

Method	S_n	S_p	ACC	MCC	AUC
Deep6mAPred	0.9538	0.9255	0.9397	0.8798	0.9793
Deep6mA*	0.9506	0.9296	0.9401	0.8800	0.9800
SNNRice6mA-large*	0.9347	0.8975	0.9204	0.8400	0.9700
MM-6mAPred *	0.9347	0.8951	0.9149	0.8300	0.9600

The asterisk (*) indicated that the results were from the literature [83].

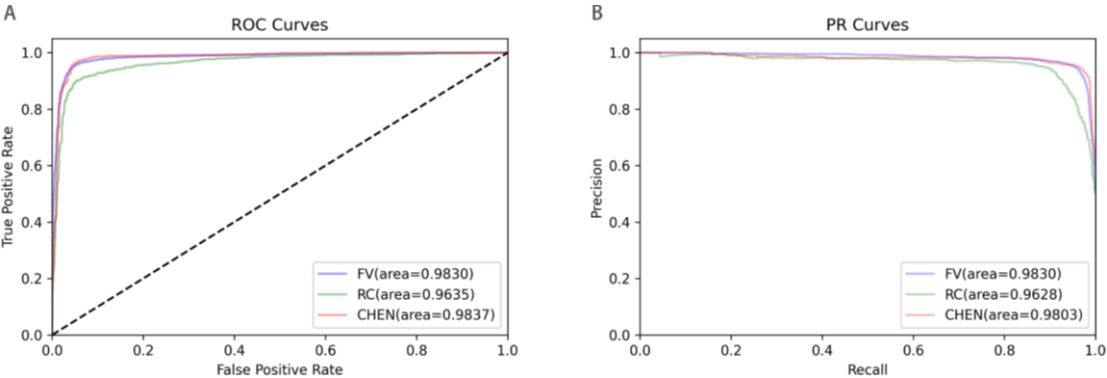
The result below is for the 6mA-rice-chen dataset. Compared with Deep6mA, Deep6mAPred increased **S_n** by 0.1572, **ACC** by 0.0750, **MCC** by 0.1436, and **AUC of the ROC curve** by 0.0237, completely superior to the other two methods. The **S_p** of Deep6mAPred is slightly lower than that of Deep6mA but much higher than that of the other two methods.

Performances over the 6mA-rice-chen dataset.

Method	S_n	S_p	ACC	MCC	AUC
Deep6mAPred	0.9545	0.9568	0.9556	0.9136	0.9837
Deep6mA*	0.7973	0.9640	0.8806	0.7700	0.9600
SNNRice6mAlarge*	0.7790	0.8742	0.8267	0.6500	0.8900
MM-6mApred*	0.7682	0.9170	0.8426	0.6800	0.9100

The asterisk (*) indicated that the results were from the literature [83].

Below are ROC curves and PR curves in 6mA-Fuse-R(Rosa chinensis) and 6mA-Fuse-F(Fragaria vesca, a kind of wild strawberry), respectively. In order to show how robust their model is, they tried to test different species, such as rose and wild strawberries, without training, and the result is quite significant, which is similar to rice data.



Following is a self-created table that I wanna show the AUC of two curves with different species.

The original context said:

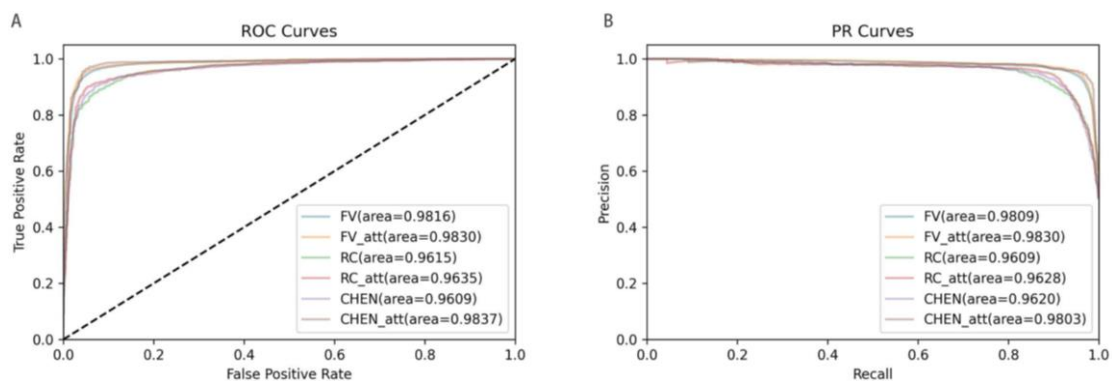
As for the 6mA-Fuse-R, the Deep6mAPred outperformed three baseline methods in terms of the AUCs of ROC curves, while in terms of the AUCs of the PR curves, it was equivalent to the Deep6mA but superior to the SNNRice6mA-large and MM-6mAPred a bit

Following the description above, we can know that the result of 6mA-Fuse-R is better than the three baseline methods but without any table or figure to prove that, and this is not rigorous enough for this information.

Model	AUC of ROC Curves		AUC of PR Curves	
	6mA-Fuse-F	6mA-Fuse-R	6mA-Fuse-F	6mA-Fuse-R
Deep6mAPred	0.9830	0.9635	0.9830	0.9628
Deep6mA*	0.9820	N/A	0.9820	N/A
SNNRice6mA-large*	0.9640		0.9630	
MM-6mAPred*	0.9600		0.9590	

They also do some ablation experiments to prove that the attention mechanism they choose is quite valid and useful in this project.

We can see that in each experiment of different species, with attention mechanism is generally better than the experiment without attention.



Other Issue

- Why can wild rose and rice use the same architecture or we can ask how to process input data so that they can be applicable at the same model structure.
- There is no extra explanation for the selected attention mechanism method.

Ensemble Learning of Convolutional Neural Network, Support Vector Machine, and Best Linear Unbiased Predictor for Brain Age Prediction: ARAMIS Contribution to the Predictive Analytics Competition 2019 Challenge

Related Work

1. Brain age and other bodily “ages”: implications for neuropsychiatry
2. DNA methylation-based biomarkers and the epigenetic clock theory of ageing.
3. Age prediction based on brain MRI image: a survey. J Med Syst
4. Biomarkers of aging. Exp Gerontol
5. Brain age predicts mortality. Mol Psychiatr
6. Longitudinal changes in individual brainAGE in healthy aging, mild cognitive impairment, alzheimer’s disease
7. BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimer’s disease
8. Gray matter age prediction as a biomarker for risk of dementia
9. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophrenia Bull
10. Quantitative neurobiological evidence for accelerated brain aging in alcohol dependence
11. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage
12. The association between “Brain- Age Score” (BAS) and traditional neuropsychological screening tools in Alzheimer’s disease

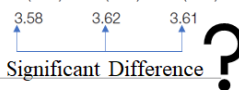
Summary

- The * symbol represents a significant reduction in MAEMAE by Ensemble Learning compared to Inception alone (p value <0.05 p value <0.05)
 - For the objective of minimizing MAE, the way of deep learning is better than BLUP and SVM (value of paired t-test $<3.1e-4$ value of paired t-test $<3.1e-4$)
 - There was no significant difference in the performance of the deep learning algorithms ($p>0.027$ $p>0.027$)
 - In contrast, Ensemble Learning’s MAE=3.46MAE=3.46, there is a significant difference ($p=1.3e-4$)
 - Taking challenge two as an example, the author uses median and mean absolute deviation per site to rescale the prediction. The results show that MAEMAE will increase by one year compared to the original one but will reduce the bias. The same ensemble learning has a significant improvement compared to Inception($p=0.010$ $p=0.010$).

		BLUP-mean	BLUP-quantiles	SVM	6-layer CNN	Age spe. 6-layer CNN	ResNet	Inception V1	Ensemble prediction	PAC results
First challenge	MAE (SE)	5.32 (0.19)	4.90 (0.19)	5.31 (0.18)	4.18 (0.16)	4.01 (0.15)	4.02 (0.15)	3.82 (0.14)	3.46 (0.13)*	3.33
	ρ	0.32	0.37	0.58	0.25	0.30	0.24	0.41	0.32	0.21
Second challenge	MAE (SE)	6.15 (0.23)	5.96 (0.23)	6.14 (0.23)	5.27 (0.21)	5.17 (0.20)	5.25 (0.20)	4.97 (0.19)	4.69 (0.19)*	4.83
	ρ	0.14	0.15	0.15	0.084	0.068	0.11	0.058	0.058	0.021

- They also tried to evaluate whether their conclusions depending on the train/test split used in the previous section by performing a 5-fold cross-validation experiment.
 - Within each fold, they found a nominally significant difference in MAE between BLUP/SVM and ResNet ($p < 5.5E-3$ and $p < 5.5E-3$)
 - In each fold, the composite age score using linear regression outperformed Inception V1's predictions ($p < 0.0022$ and $p < 0.0022$). For folds 2 and 3, ensemble learning via random trees significantly outperforms Inception V1 alone ($p = 4.0E-3$ and $3.4E-4$ and $p = 4.0E-3$ and $3.4E-4$)
 - Note that the MAE obtained using Random Forest is very close to the MAE obtained by taking the mean or median score for each person. We cannot conclude that there is a significant difference between **linear model combinations** and **random forests**.

	Individual algorithms							Ensemble learning			
	BLUP-mean	BLUP-quantiles	SVM	6-layer CNN	Age spe. 6-layer CNN	ResNet	Inception V1	LM	RF	Mean	Median
Fold 1	5.32 (0.19)	4.90 (0.19)	5.31 (0.18)	4.18 (0.16)	4.01 (0.15)	4.02 (0.15)	3.82 (0.14)	3.46 (0.13)*	3.62 (0.15)	3.74 (0.13)	3.67 (0.14)
Fold 2	5.05 (0.18)	4.79 (0.19)	5.34 (0.18)	4.47 (0.15)	4.12 (0.13)	4.01 (0.14)	3.97 (0.15)	3.53 (0.13)*	3.60 (0.15)*	3.69 (0.13)	3.74 (0.13)
Fold 3	4.90 (0.18)	4.37 (0.16)	4.84 (0.17)	4.41 (0.16)	4.27 (0.15)	3.88 (0.14)	4.00 (0.16)	3.33 (0.13)*	3.46 (0.15)*	3.46 (0.12)*	3.45 (0.13)*
Fold 4	5.07 (0.18)	4.71 (0.18)	5.06 (0.18)	4.55 (0.17)	4.27 (0.16)	4.11 (0.15)	3.85 (0.15)	3.57 (0.13)*	3.72 (0.14)	3.68 (0.14)	3.74 (0.15)
Fold 5	5.22 (0.19)	4.69 (0.18)	5.20 (0.18)	4.02 (0.16)	3.89 (0.15)	3.99 (0.16)	3.75 (0.15)	3.34 (0.13)*	3.51 (0.14)	3.56 (0.13)	3.47 (0.13)
5-fold combined MAE	5.11	4.69	5.15	4.33	4.11	4.00	3.88	3.44	3.58	3.62	3.61



Significant Difference ?

- The low performance of BLUP/SVM shown above compared to deep learning algorithms motivated the authors to test whether it could be attributed to the input data or the algorithm itself. Therefore, the author retrains BLUP and SVM (trained on gray matter maps)

- † Symbols represent: the algorithm trained with the gray matter map is significantly **better than** the algorithm trained with surface-based vertices ($p < 0.05/15$).
- The * symbol indicates: the performance of the algorithm trained on the gray matter image is significantly **lower than** that of Inception v1 ($p < 0.05/15$).
- Despite the reduction in MAE, BLUP-mean and SVM trained on gray matter still performed **worse than** Inception V1 ($p < 0.0033$), although the difference between Inception V1 and BLUP-quantile became not significant.

	BLUP-mean	BLUP-quantiles	SVM	Ensemble learning
Fold 1	4.51 (0.16) ^{†*}	3.91 (0.14) [†]	4.64 (0.17) ^{†*}	3.39 (0.13)
Fold 2	4.45 (0.16) ^{†*}	4.06 (0.15) [†]	4.75 (0.16) ^{†*}	3.46 (0.13)
Fold 3	4.67 (0.17)*	4.02 (0.16)	4.62 (0.17)*	3.26 (0.13)
Fold 4	4.59 (0.16)*	4.16 (0.16) [†]	4.52 (0.16)*	3.55 (0.14)
Fold 5	4.86 (0.18)*	4.21 (0.17)	4.78 (0.17)*	3.35 (0.14)
5-fold MAE	4.61	4.07	4.66	3.42

- The participant is older; the prediction error is larger. → Therefore, the predictor will tend to underestimate the age of older participants and overestimate the age of younger participants.
- We did not observe significant associations of prediction errors with gender or location.

	BLUP-mean	BLUP-quantiles	SVM	6-layer CNN	Age spe. 6-layer CNN	ResNet	Inception V1
Age	2.9E-10*	5.8E-13*	5.8E-46*	7.3E-10*	2.2E-13*	9.1E-05*	7.7E-20*
Site	3.7E-01	4.4E-02	4.5E-03	2.8E-02	4.3E-02	2.3E-02	5.0E-02
Sex	7.1E-02	1.4E-01	3.6E-02	1.0E+00	8.5E-01	1.0E+00	5.4E-01

Other Issue

- They didn't explain why they used two 6-Layers CNN to combine and the effect in detail.
- They also didn't explain the gray/white matter map difference and the properties of these maps in detail.

Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data

Related Work

- High multiclass & unbalanced classification problem
 1. Development of biomarker classifiers from high-dimensional data
 2. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting
 3. The Elements of Statistical Learning: Data Mining, Inference and Prediction
 4. An extensive comparison of recent classification tools applied to microarray data
 5. Roadmap for developing and validating therapeutically relevant genomic classifiers
- DNA methylation application
 6. DNA methylation-based classification of central nervous system tumours
 7. Cancer epigenetics reaches mainstream oncology. Nat. Med. 17,330–339 (2011).
 8. A DNA methylation fingerprint of 1628 human samples
 9. Assessing CpG island methylator phenotype, 1p/19q codeletion, and MGMT promoter methylation from epigenome-wide data in the biomarker cohort of the NOA-04 trial
- The number of features p vastly outnumbers the sample size (n)
 10. The Elements of Statistical Learning: Data Mining, Inference and Prediction
 11. Class probability estimation for medical studies

Summary

- **Random Forest (RFs)**
 - Vanilla RF (vRF)
 - The ME of vRF was 4.8%, the AUC was 99.9%, and the corresponding BS and LL were 0.32 and 0.78, respectively
 - Platt scaling with LR and FLR improves BS and LL by a factor of 2-4, furthermore, FLR is better than LR

- MR slightly outperformed Platt's two variants and achieved very low 10th and 9th overall BS (0.073) and LL (0.155) metrics respectively
- tuned RF(tRF)
 - RF tuned for ME (tRFME) showed 10th overall error rate (3.5%) and 4th AUC (99.9%), while it had relatively high BS (0.35) and LL (0.86) similar to vRF
 - Both tRFBS and tRFLL have higher error rates, about 5.5%
 - After calibration with **MR**, almost all versions of tRF get the biggest performance improvement

Workflow	Top 10 BS	Classifier	Run-time 5 × 5 CV (/fold)	No. of CPU threads [hardware]	R package	Calibrator	Optimized metric	Hyperparameters	ME	AUC	BS	LL
vRF		RF	38 min	1 [2]	randomForest	raw	ME	ntree = 500, mtry = 100	0.048	0.999	0.320	0.780
vRF + LR		RF	+30 s (LR)	1 [2]	randomForest	Platt LR (us)	ME	pvarsel = 200	0.052	–	0.106	0.289
vRF + LR		RF	+30 s (LR)	1 [2]	randomForest	Platt LR	ME	pvarsel = 200	0.052	0.994	0.081	0.262
vRF + FLR		RF	+8-9 min (FLR)	1 [2]	randomForest	Platt Firth (us)	ME	pvarsel = 200	0.048	–	0.105	0.193
vRF + FLR		RF	+8-9 min (FLR)	1 [2]	randomForest	Platt Firth	ME	pvarsel = 200	0.048	0.999	0.081	0.193
vRF + MR	10	RF	+7-8 min (MR)	11 [2]	randomForest	MR	ME	pvarsel = 200	0.043	0.999	0.073	0.155
tRF _{BS}		RF	12-13 h (16-25 min/fold)	72 [5]	randomForest	raw	BS	ntree = (500, 1,000, 1,500, 2,000)	0.055	0.999	0.272	0.673
tRF _{ME}		RF			randomForest	raw	ME	mtry = (80, 90, 100, 110)	0.035	0.999	0.351	0.855
tRF _{LL}		RF			randomForest	raw	LL	pvarsel = (100, 200, 500, 1,000, 2,000, 5,000, 7,500, 10,000)	0.055	0.999	0.273	0.672
tRF _{BS} + LR		RF	+30 s (LR)	1 [2]	randomForest	Platt LR	BS		0.056	0.997	0.086	0.266
tRF _{ME} + LR	9	RF	+30 s (LR)	1 [2]	randomForest	Platt LR	ME	nodesize = 1	0.042	0.998	0.062	0.156
tRF _{LL} + LR		RF	+30 s (LR)	1 [2]	randomForest	Platt LR	LL	nodesize = 1	0.058	0.995	0.089	0.291
tRF _{BS} + FLR		RF	+8-9 min (FLR)	1 [2]	randomForest	Platt Firth	BS	nodesize = 1	0.054	0.997	0.086	0.194
tRF _{ME} + FLR	8	RF	+8-9 min (FLR)	1 [2]	randomForest	Platt Firth	ME	nodesize = 1	0.037	0.999	0.062	0.150
tRF _{LL} + FLR		RF	+8-9 min (FLR)	1 [2]	randomForest	Platt Firth	LL	nodesize = 1	0.056	0.999	0.089	0.205
tRF _{BS} + MR		RF	+7-8 min (MR)	11 [2]	randomForest	MR	BS	nodesize = 1	0.051	0.997	0.082	0.176
tRF _{ME} + MR	4	RF	+7-8 min (MR)	11 [2]	randomForest	MR	ME	nodesize = 1	0.027	0.999	0.046	0.095
tRF _{LL} + MR		RF	+7-8 min (MR)	11 [2]	randomForest	MR	LL	nodesize = 1	0.055	0.999	0.086	0.188

• ELNET

- It used 1,000 most variable CpG probes
- ME ranked 8th, AUC ranked 5th
- ME (2.7%), BS (0.048) and LL (0.109) and negligibly low AUC (99.9 %)

ELNET (1k)	7	ELNET	-7.5 h (12-15 min/fold)	31 [4]	glmnet	raw	ME	$\alpha = 0 \mid 0.025 ; \lambda = (0.0010-0.0036)$	0.032	0.999	0.059	0.131
ELNET (10k)	5	ELNET	-72 h (2-2.25 h/fold)	31 [4]	glmnet	raw	ME	$\alpha = 0 ; \lambda = (0.012-0.038)$	0.027	0.999	0.048	0.109

• SVM

- More effective ME = 2.1% (lowest overall) with Platt scaling with Firth regression
- While simple LR can be more effective in improving BS (second) and LL (fourth) by 8-9 times respectively
- MR (**SVM-LK+MR**) achieves the most comprehensive improvement across all metrics. It reduced BS by a factor of 9.5 and LL by a factor of 11.5,

resulting in the second lowest ME (2.1%) and AUC (99.9%), lowest BS (0.039), and lowest LL (0.085)

SVM-LK		SVM	-28 h (50-70 min/fold)	11 [3]	e1071	raw	ME	C = 0.001 0.01	0.032	0.999	0.372	0.978
SVM-LK+LR	2	SVM	+30 s (LR)	1 [2]	e1071	Platt LR	ME	C = 0.001 0.01	0.025	0.999	0.043	0.112
SVM-LK+FLR	3	SVM	+8-9 min (FLR)	1 [2]	e1071	Platt Firth	ME	C = 0.001 0.01	0.021	0.999	0.044	0.135
SVM-LK+MR	1	SVM	+7-8 min (MR)	11 [2]	e1071	MR	ME	C = 0.001 0.01	0.021	0.999	0.039	0.085
SVM-LK (GPU)	6	SVM	-5 h	1080Ti	Rgtsvm-GPU	global softmax	ME	C = 0.01 0.001 ; n.SV= 1,300-1,600	0.033	0.998	0.056	0.144
SVM-CS ⁶		SVM	-6 h (13-15 min/fold)	7 [1]	Liblinear	-	ME	C ≥ 0.001	0.028	-	-	-

- **Boost Tree**

- Boosted model using ME as evaluation metric outperforms model using LL
- Overall ME of 5.1% and AUC of 99.9%, with the second lowest BS (0.15) and LL (0.43) among the base ML classifiers studied

XGBoost		BT	-65-70 h (110-130 min/fold)	72 [5]	xgboost	raw	ME	Tables 3 and 4	0.051	0.999	0.150	0.430
XGBoost+LR		BT	+30 s (LR)	1 [2]	xgboost	Platt LR	ME	Tables 3 and 4	0.055	0.991	0.087	0.452
XGBoost+FLR		BT	+8-9 min (FLR)	1 [2]	xgboost	Platt Firth	ME	Tables 3 and 4	0.053	0.993	0.089	0.384
XGBoost+MR		BT	+7-8 min (MR)	11 [2]	xgboost	MR	ME	Tables 3 and 4	0.046	0.999	0.092	0.247

Machine Learning-Based DNA Methylation Score for Fetal Exposure to Maternal Smoking: Development and Validation in Samples Collected from Adolescents and Adults

Related Work

1. DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking during Pregnancy
2. 450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy
3. Identification of DNA Methylation Changes in Newborns Related to Maternal Smoking during Pregnancy

Summary

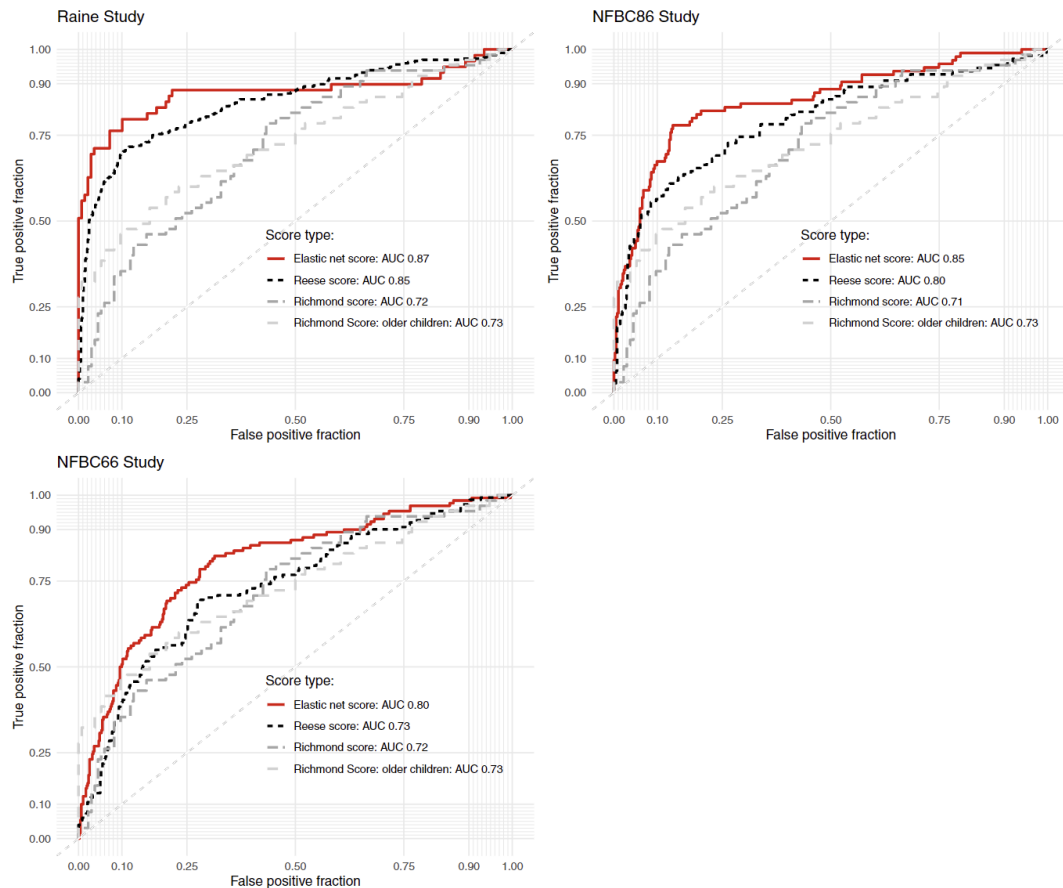
The paper aims to find an effective way to predict if a person is exposed to maternal smoking using DNA methylation data. The datasets used for training, testing, and validating are three datasets with sample numbers smaller than 1000. Before finding

suitable models, they do data pre-selecting by statistical method, which makes the computation less expensive without harming the performance. To solve the unbalance dataset problem, they use SMOTE to create ‘artificial’ data in the minority class. After pre-processing, they try about ten network structures and fine-tune four of them with better performance. Their result shows that elastic net performs best on this problem, and their model achieves better sensitivity and specificity than existing models.

This work does not contain any fancy network structure. All models they used are standard ones provided in suites. The interesting part is the data pre-processing. Because of the high dimension of DNA methylation data, feature selection becomes an important process. Using p-value helps select data truly causes the difference, which makes the computational cost affordable and avoids overfitting. For the unbalance data problem, SMOTE generates more data by hand, which increase the cost of misclassifying minority class samples.

- Result shows that, after fine-tuning, elastic net regression had the best overall performance on the task.
- Reese: On Raine set, the performance shows no significant difference between elastic net and Reese score both on the ROC curve and statistical metrics. While on NFBC1986 and NFBC1966, the model provided outperformed on nearly all metrics.
- Richmond: On the Raine set, the performance is much better than the Richmond score, which can be observed in the ROC curve. While in NFBC1986, the specificity is slightly lower, while other metrics show significant advantages. For NFBC1966, elastic net outperforms Richmond in all metrics.

	Sensitivity	Specificity	Cohen's κ	Accuracy	AUC	Brier score	# CpGs required
Raine Study test data set							
Elastic net score	0.91	0.76	0.68	0.83	0.87	0.13	204
Gradient boosting machine	0.91	0.82	0.72	0.88	0.88	0.1	1,511
Random forest	0.87	0.73	0.58	0.83	0.83	0.17	1,511
Support vector machine	0.87	0.73	0.6	0.83	0.85	0.13	1,511
Reese score	0.88	0.72	0.6	0.83	0.85	0.21	28
Richmond score 568 CpGs	0.7	0.68	0.34	0.69	0.72	0.22	568
Richmond score 19 CpGs	0.79	0.58	0.37	0.72	0.73	0.22	19
NFBC1986							
Elastic net score	0.87	0.75	0.56	0.84	0.85	0.13	204
Gradient boosting machine	0.95	0.29	0.19	0.54	0.74	0.39	1,511
Random forest	0.79	0.16	0.06	0.64	0.54	0.24	1,511
Support vector machine	0.87	0.44	0.33	0.77	0.79	0.16	1,511
Reese score	0.87	0.61	0.46	0.82	0.8	0.18	28
Richmond score 568 CpGs	0.65	0.76	0.34	0.74	0.71	0.22	568
Richmond score 19 CpGs	0.65	0.77	0.31	0.68	0.73	0.22	19
NFBC1966							
Elastic net score	0.72	0.78	0.39	0.73	0.8	0.19	204
Gradient boosting machine	0.88	0.26	0.1	0.45	0.68	0.48	1,511
Random forest	0.77	0.18	0.05	0.64	0.48	0.24	1,511
Support vector machine	0.88	0.45	0.33	0.76	0.75	0.2	1,511
Reese score	0.72	0.7	0.32	0.71	0.73	0.18	28
Richmond score 568 CpGs	0.66	0.63	0.22	0.69	0.72	0.22	568
Richmond score 19 CpGs	0.61	0.72	0.23	0.63	0.73	0.22	19



Other Issue

However, there are still some parts ambiguous in these processes. First, for the pre-selecting algorithm, they introduce Bonferroni correction to control type 0 error. While in the end, the level they choose for the significant difference is between the value with and without correction. Since the given threshold has no statistical significance, I don't think introducing Bonferroni correction is needed. I think the threshold value is simply a fine-tuned result. Next, I doubt if using SMOTE to generate artificial data in the minority class is always feasible. This method strongly relies on how data points are distributed in the space. If data points are not linearly separable, this method might cause trouble since a point between two in the same class does not necessarily belong to the same set.

A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking

Related Work

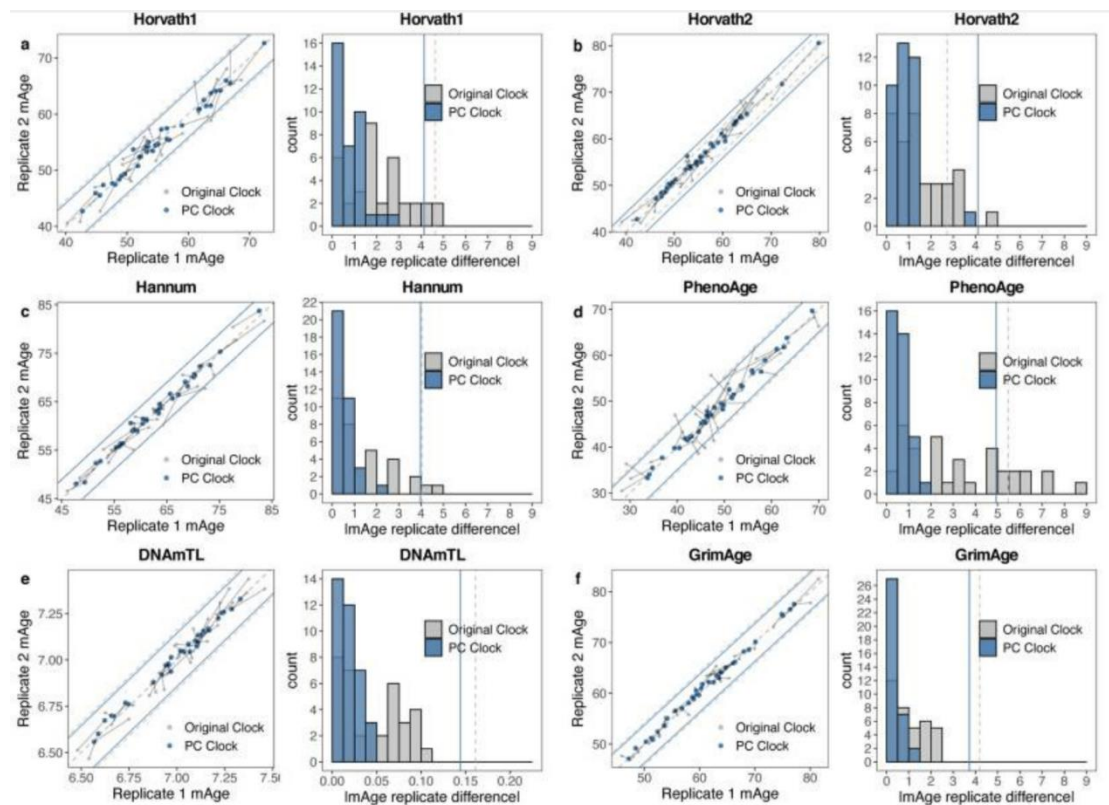
1. Epigenetic clock: A promising biomarker and practical tool in aging
2. Many chronological aging clocks can be found throughout the epigenome: Implications for quantifying biological aging

Summary

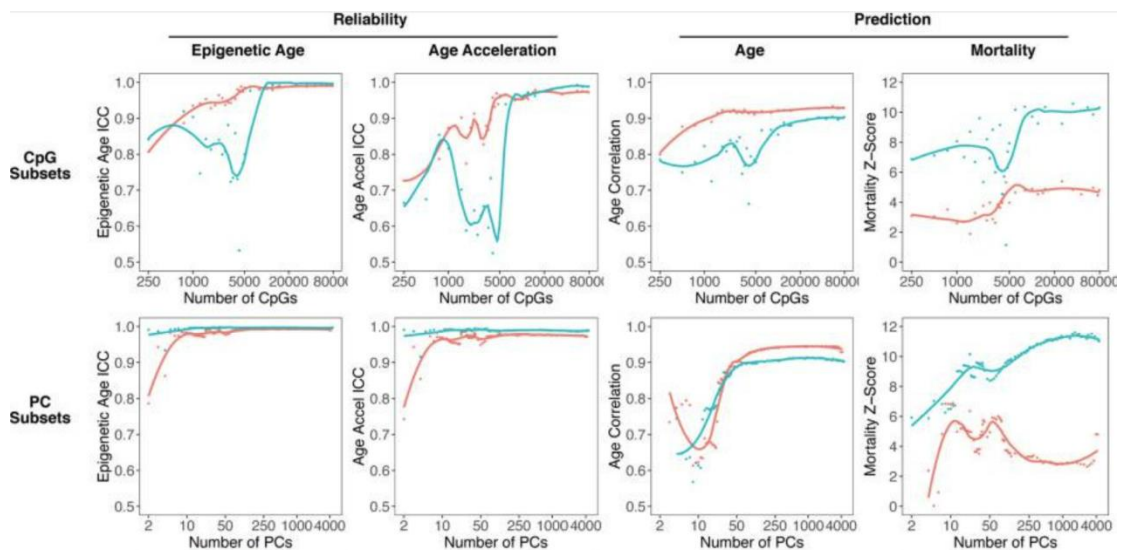
The epigenetic clock can be helpful in clinical diagnosis. However, epigenetic clocks have limited application because of a serious problem: reliability. It is found that existing epigenetic clocks give predictions with large variances between technical replicates, which should be, ideally, exactly the same. Thus, enhancing reliability becomes an important issue. In work, it is found that simply filtering features using ICC does not lead to a good solution, and performing weighting still suffers from noise. Moreover, the authors found that based on statistical analysis, there is much more gene related to aging than those considered in existing clocks. Thus, they choose to apply PCA to the data given to include more information and reduce the noise effect. Their result shows that principal components indeed have better reliability than original CpGs, and using CpGs for training can reduce the sample size and dimension needed to achieve convergence.

This work shows that sometimes a simple structure can lead to good results. PCA allows the model to include more information (since they observe that there are much more CpGs related to aging than those already included in the existing clock) while keeping the dimension the same. Moreover, the noise is reduced effectively, and the reliability of the clock has increased significantly. To be mentioned their data shows that sometimes PC with low variance contain important information. Thus, after doing PCA, they didn't filter out PCs by hand. Instead, they apply elastic net regression, using the regularization term to select features automatically. They claim that this method can get rid of terms representing noise effectively.

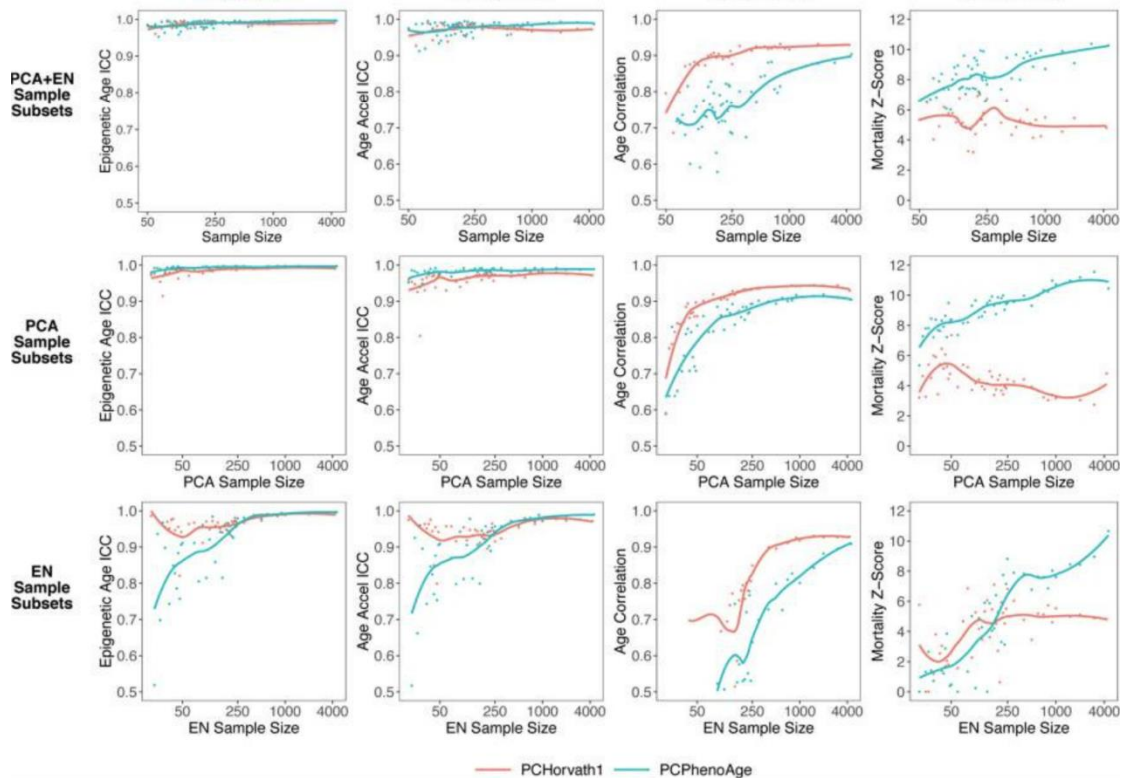
- The graph gives the relation between technical replicates. It is shown that using PCA, the ICC of features (CpGs for original clocks, PC for new method) has increased significantly, which implies that the new clock should have better reliability.



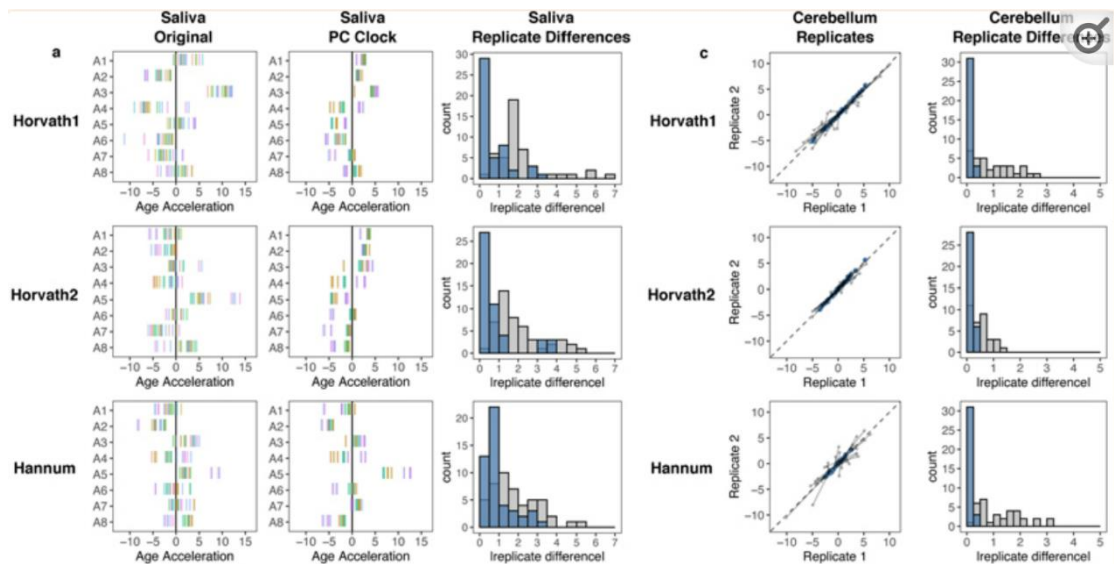
- Using PCA, less dimension is needed for achieving same performance.

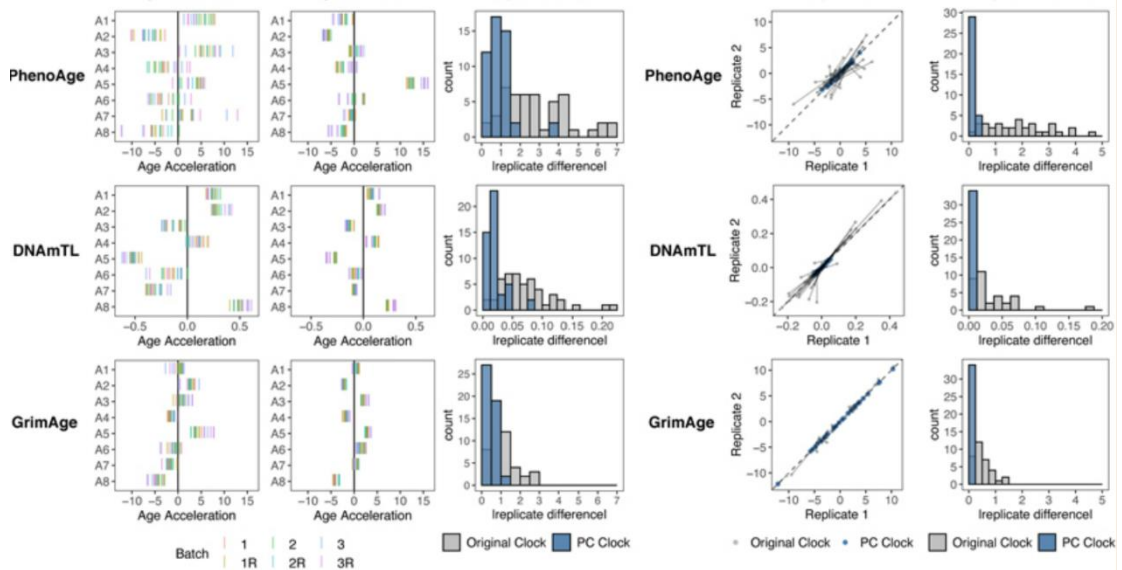


- Probably because of the higher reliability of features, PC clocks need fewer samples to achieve the same accuracy. Notice that the number of samples cannot be too low since the number of principal components that can be extracted is upper bounded by the sample number.



- Though the clock is trained using blood samples, and the team claims that, to some extent, their result can be applied to other tissue such as saliva and the brain. The graph shows that the PC clock still achieves lower variance between technical replicates when applied to those tissues.





Other Issue:

When observing that weighting doesn't help the reliability much, the team claims that it is because the noise is amplified simultaneously. We think this is not the true reason. Weighting features with higher ICC do increase the SNR mathematically. We think that the true reason for limited performance improvement needs further investigation.

Drug Response Prediction Based on 1D Convolutional Neural Network and Attention Mechanism

Related Work

1. Data Integration Using Advances in Machine Learning in Drug Discovery and Molecular Biology
2. Convolutional Neural Networks for ATC Classification

Summary

In modern clinical medicine, many seemingly identical symptoms respond very differently to drugs. To apply the most suitable treatment at the first timing, predicting the drug response on a patient using genetic information becomes an important topic. This work uses a dataset containing expression, methylation, and copy number variation to develop a drug response prediction model. They first integrate the three-dimensional data into biological pathways, then apply simulated annealing to make

feature selection. Last, they use a 1D convolutional network with attention to training the predictor. They claim that the method performs better than the random forest approach.

The team treats the prediction problem as a binary classification problem. The response profile is separated into two classes, i.e., responders (including complete and partial response) and nonresponders (including stable disease and progressive disease). After defining the problem, this work spent lots of time on data processing. They first use gene synthesis to score the training set. The score is then mapped onto the KEGG pathway to obtain the weight of subapthways on each training data. Now, the training data are integrated into functional blocks. Then they apply simulated annealing to make the feature selection. The advantage of such a process is that the result is somehow explainable since the features used for training are functional blocks. After pre-processing, they got 17x3 features on expression, methylation, and copy number variation. They then apply CNN, pooling, dropout, and softmax functions to produce the final prediction.

Result & Comparison

I don't think the result given by this work is solid. The team only provided their training curve on testing and validation and comparison between random forest and the proposed structure. No data from other related works are provided. From their result, I have no information about that if their work achieves a good solution; or if they just try two methods and claim that one is better than the other on the task.

Other Issue:

Although the attention mechanism is mentioned in the article, the team provides no information about how they integrate the attention module into their model. Attention is originally used for linguistic tasks such as translation, and the output of the classifier has dimension 1. Thus, I totally have no idea how can the mechanism can be integrated with their network structure.

Interpretable machine learning prediction of all-cause mortality

Related Work

1. Explainable artificial intelligence (XAI) in deep learning-based medical image

analysis

2. Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines
3. Overview of Explainable Artificial Intelligence for Prognostic and Health Management of Industrial Assets Based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Summary

High-dimensional embedding does not choose because the feature it chose is not representative enough for different cell types.

To make epigenetic data not only easier to analyze but also explainable, IMPACT utilizes SHAP to make feature selection very clear and easier. IMPACT tries to predict the death risk for people of different ages and habits. The data set is described below.

Feature	Type	Display_Name
Demographics_Age	Demographics	Age
Demographics_Citizenship_2.0	Demographics	Not a citizen of the US
Questionnaire_AlcoholFreqDays	Questionnaire	Avg # alcoholic drinks/day - past 12 mos
Laboratory_UrineAlbumin	Laboratory	Albumin, urine (ug/mL)
Examination_Weight	Examination	Weight (kg)

Examples of Features in Dataset, containing 47,261 samples. Collected from NHANES

Though this is not a high-dimensional case, the evaluation of risk requires the evaluation of the importance of different features. Thus, explainable feature selection is important.

To understand the secret of SHAP, we first review the original paper, where this concept is unified and well organized [5]. SHAP combines the concept of Shapley value,

which is hotly discussed in the field of machine learning, with different models (i.e., linear, DNN...).

Shapley values identify the importance of a feature by evaluating different combinations of features. Thus, Shapley value won't ignore the cooperation between features. However, combination calculation is a disaster in computation, so the author [5] uses a linear model to approximate the combinatorial effect.

Then, the linear model will have weight for each feature. The weight here is the importance of the feature (in the local area). Thus, the explainable model has the additive property of all features (i.e., the combined average importance of 2 features is the sum of the weight)

In the reference [5], it ensures this is the only way to guarantee local accuracy, missingness (if the value goes 0 at the point, the weight should be 0), and consistency (as below). Thus, this system is very reliable to be used to generate explainable feature selection.

Property 3 (Consistency) Let $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ denote setting $z'_i = 0$. For any two models f and f' , if

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs $z' \in \{0, 1\}^M$, then $\phi_i(f', x) \geq \phi_i(f, x)$.

The Consistency property of SHAP

Now we get back to IMPACT. IMPACT uses gradient boosted tree as a model to work with sharply values. To generate risk evaluation, IMPACT use generated model with SHAP to calculate importance. And the risk evaluation is simply the sum of each risk in different features (by additive property of SHAP).

There is also another funny process feature when generating a gradient boosted tree. To evaluate the relation between features, in order to kill redundant features, it defined the term supervised distance.

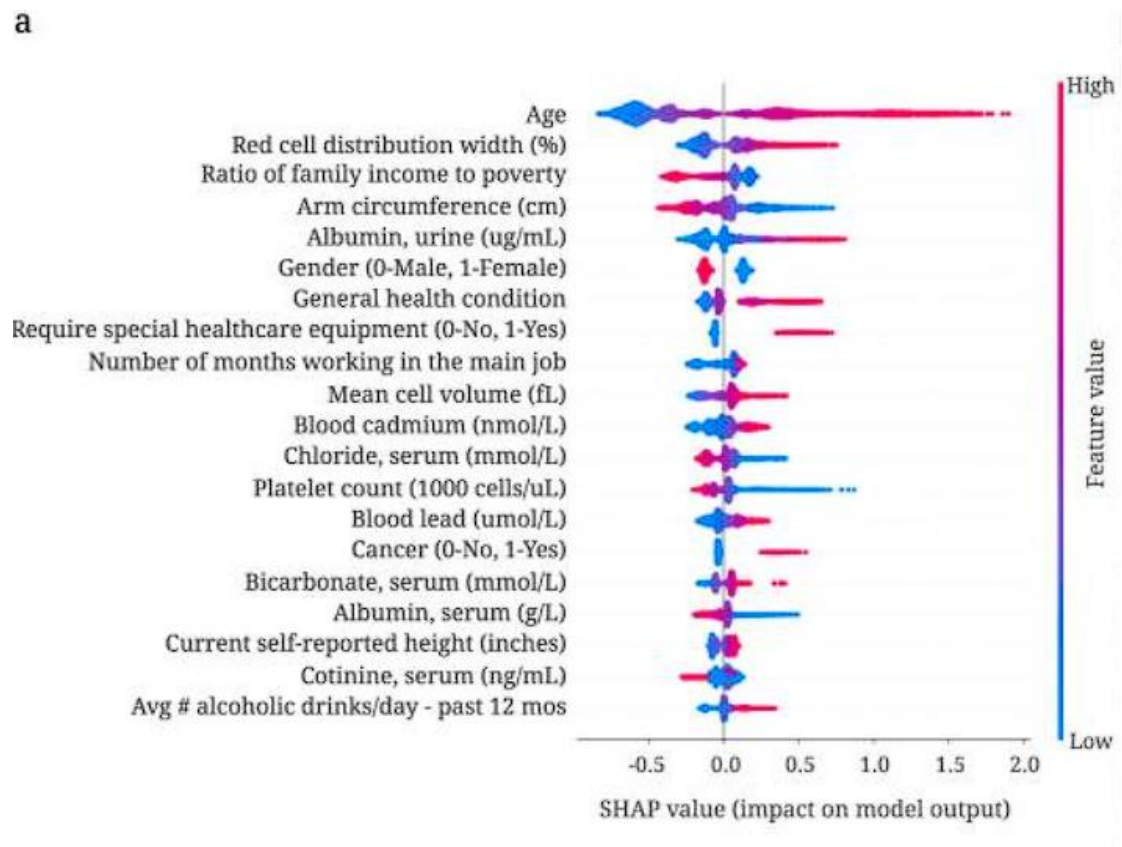
$$\begin{aligned} & \text{supervised } R^2(i, j) \\ &= \max \left(0, 1 - \text{mean} \left(\frac{\left(\text{Prediction}_i - \text{Prediction}_j \right)^2}{\text{var}(\text{Prediction}_i)} \right) \right) \end{aligned} \quad 1$$

$$\begin{aligned} & \text{supervised distance}(i, j) \\ &= \max \left(1 - \text{supervised } R^2(i, j), 1 - \text{supervised } R^2(j, i) \right) \end{aligned} \quad 2$$

Supervised distance definition. prediction_i means use *i* as 1-var model and predict with *i*; prediction_j means use *i* as 1-var model and predict with *j*. 1-var model means only 1 feature is real data; other features use random selection from other data.

After having supervised distance, it uses hierarchical clustering to find out the most redundant features, that is, the pair with low supervised distance. Then remove the one with a smaller SHAP value. By repeating clustering and removal, it can eventually find out that the feature set is predictive and less redundant.

The result is very surprising. First, high SHAP value features have been shown to have a correlation with mortality previously. For example, red cell distribution width (RDW) is identified as an indicator in the experiment [6]. Second, two of the twenty highest SHAP value features are not observed in the past medical reviews. This can be considered an important feature for medical use in the future.



The 20th features with the highest SHAP value, color, indicate the value of the feature. Length indicates the importance, and width indicates the number of data.

Hierarchical Ridge Regression for Incorporating Prior Information in Genomic Studies

Related Work

1. Incorporating prior knowledge into regularized regression.
2. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events.

Summary

This paper serves for general biological high-dimensional cases with prior information. Motivated by wanting to learn the result of classical models, the author tries to use an alternative method to regularize and tries to utilize prior knowledge while having regularization.

To utilize prior knowledge, the author invented a matrix that concentrates data into lower dimensions. And then regularize the weight of the outcoming features. The feature coming out from the matrix is called the meta-features. Like other regularizations, the author adds a penalty term to the loss function (as below). However, since there are two layers of features, there are also two layers of regularization.

$$+ \lambda \sum_{j=0}^k \beta_j^2 + \gamma \sum_{j=0}^k |\beta_j|$$

Example of a penalty in the loss function. The left is for normal regularization; the right one is for Lasso regularization.

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2,$$

$$\arg \min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda_1}{2} \|\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \frac{\lambda_2}{2} \|\boldsymbol{\gamma}\|_2^2 \right\},$$

The upper is the regular normalization loss function. The lower is the normalization author proposed. Z is a matrix from prior knowledge, B is the coefficient of subjective feature, and γ is the coefficient of meta feature.

Just like ordinary regularization, this Hierarchical Ridge Regression also has a closed-form solution. By changing the parameters, a closed-form solution can be easily derived.

$$\arg \min_{\beta, \gamma} \left\{ \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \frac{\lambda_1}{2} \|\boldsymbol{\beta} - Z\boldsymbol{\gamma}\|_2^2 + \frac{\lambda_2}{2} \|\boldsymbol{\gamma}\|_2^2 \right\},$$

$$\arg \min_{\phi, \gamma} \left\{ \frac{1}{2} \|\mathbf{y} - X(\boldsymbol{\phi} + Z\boldsymbol{\gamma})\|_2^2 + \frac{\lambda_1}{2} \|\boldsymbol{\phi}\|_2^2 + \frac{\lambda_2}{2} \|\boldsymbol{\gamma}\|_2^2 \right\}.$$

$$\begin{aligned} & \frac{1}{2} \|\mathbf{y} - X(\boldsymbol{\phi} + Z\boldsymbol{\gamma})\|_2^2 + \frac{\lambda_1}{2} \|\boldsymbol{\phi}\|_2^2 + \frac{\lambda_2}{2} \|\boldsymbol{\gamma}\|_2^2 \\ &= \frac{1}{2} \|\mathbf{y} - \tilde{X}\boldsymbol{\theta}\|_2^2 + \frac{1}{2} \boldsymbol{\theta}^T \Lambda \boldsymbol{\theta}, \end{aligned}$$

$$\hat{\boldsymbol{\theta}} = (\tilde{X}^T \tilde{X} + \Lambda)^{-1} \tilde{X}^T \mathbf{y}.$$

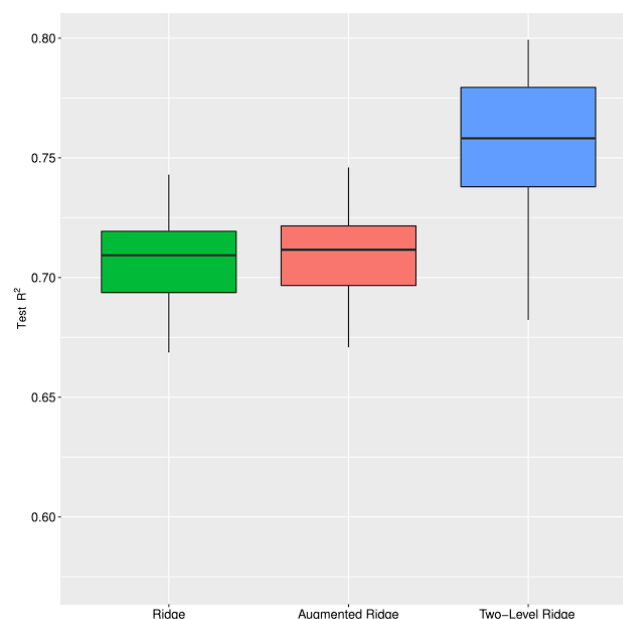
The closed-form solution for hierarchical ridge regression.

The behavior of the meta matrix can have 3 common situations: totally uncorrelated, at the same pathway, appears in 2 pathways. For a totally uncorrelated case, the meta matrix is just a diagonal matrix with different eigenvalues. For the second case, the features in the same pathway will be assigned in the same column, while each row has one nonzero value. For cases where a feature appears in multiple pathways, it is just like the previous case, besides, there might be multiple nonzero terms in the meta matrix.

$$Z = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Meta feature matrix with different behaviors. The left one shows cases in the same pathway that might be assigned in the same column. The right one shows cases that appear in multiple pathways.

The result was amazing compared to other regularizations. DNA methylation has long been believed to have a correlation with age. In 2013, Horvath used methylation patterns to build a highly predictive model for human real age, indicating methylation is indeed a reliable biomarker for age. In the analysis, they both use data with 250,000 features, which is the location of methylation. Meta matrix takes locations on DNA that belong to the same gene to be same features and find concentrated 250,000 features into 6,766. Horvath uses high dimensional regularization to enhance his predictor [8], while hierarchical ridge excels his result. This is one of the best way of showing the power of hierarchical ridge regularization.



The comparison of different regularizations.

Though hierarchical ridge regulation is easy to apply and seems to be useful, the downside is also very obvious. This method only applies to linear models, and the quality of prior knowledge is very important.

NetTIME: a multitask and base-pair resolution framework for improved transcription factor binding site prediction

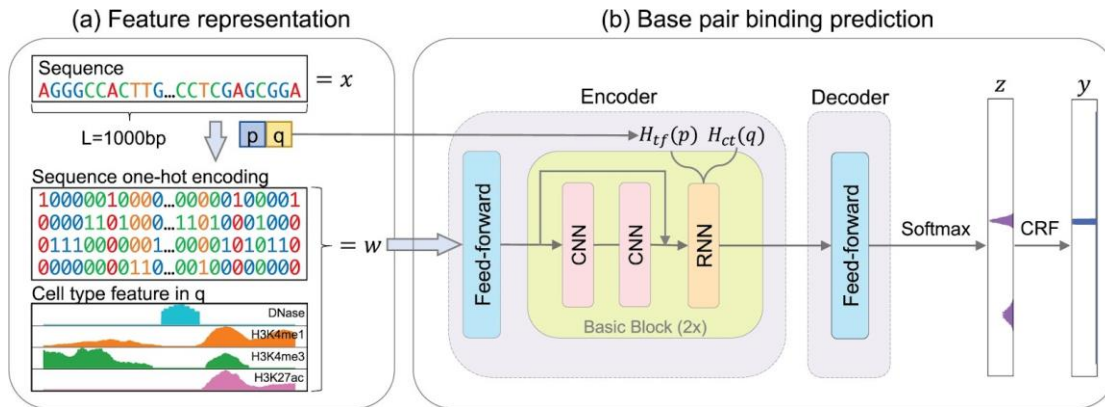
Related Work

1. Base-pair resolution detection of transcription factor binding site by deep deconvolutional network
2. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data.
3. Predicting transcription factor binding using ensemble random forest models
4. MTTFSite: cross-cell type TF binding site prediction by using multi-task learning

Summary

Nettime is quite different from other feature engineering techniques. Instead of lowering the dimension, it tries to find some new features to simplify the analysis. Nettime aims to make DNA-protein binding prediction from cell type-specific into one same question. DNA-protein binding prediction has long been researched [9]. This topic mainly discusses how to make a prediction of whether a protein will bind to given DNA sequences. While sharing the DNA sequence, different cell type has different functions, thus, behave very differently in DNA-protein binding. Thus, the problem has always been researched with specific types. In the process of organizing similarities between cell types, the author found that some of the features of a cell type might be used as a representative of cell type. By trying and prior knowledge, the author found there are 4 cell type-specific data. That is DNase, H3K4me1, H3K4me3, H3K27ac. These are modifications of DNA, which have different patterns at different loci on DNA.

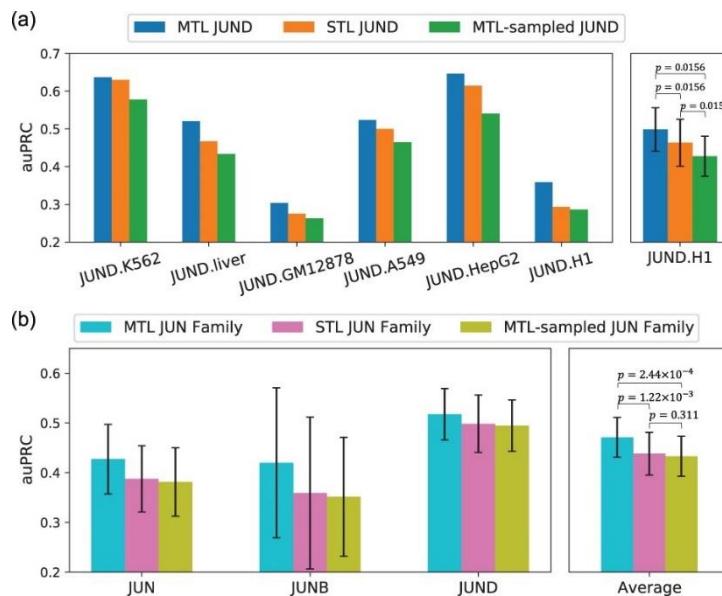
To preprocess the feature, the author turns DNA sequences into one-hot-encoded data, and each position includes 4 data mentioned above. Then the cell type represented feature is done.



One-hot-encoded DNA sequence and the 4 cell type representing data

The result is quite stunning. If all cell type is put into the training set, the performance is even better than cell type-specific training prediction.

To conclude, the feature representing different cell types is very representative. And the result shows it can make better predictions than the original method. However, due to the requirement of prior knowledge to accomplish this. This means the method is hard to improve, and the selection of this prior knowledge is also a challenge.



Result for comparing cell type specific training.

Conclusion

1. Adjusting the network structure to fit your dataset is needed to achieve a good result.
2. When dealing with biological problems, data selection is the most important task. A good data selection can lead to satisfying outcomes with a relatively simple model.
3. Prior knowledge is useful when training models. It can not only help to select suitable models but is also useful in data preprocessing.
4. After preprocessing using prior knowledge, the statistical method may help feature selection a lot (especially when a linear model is adopted).

Work Distribution

吳承軒: Abstract & Introduction writing, paper survey(x3)	33%
莊哲維: Report integration, paper survey(x3)	33%
何秉學: Video integration, paper survey(x3)	33%

References

1. DNA methylation age of human tissues and cell types
2. Interpretable machine learning prediction of all-cause mortality
3. A machine learning approach utilizing DNA methylation as an accurate classifier of COVID-19 disease severity
4. Hierarchical Ridge Regression for Incorporating Prior Information in Genomic Studies
5. A Unified Approach to Interpreting Model Predictions
6. Red cell distribution width as a novel prognostic marker in heart failure: data from the CHARM Program and the Duke Databank.
7. DNA methylation age of human tissues and cell types
8. Accelerated epigenetic aging in down syndrome
9. Convolutional neural network architectures for predicting DNA–protein binding