THE UNIVERSITY OF MELBOURNE

Data Science Project:

Epigenetic Classifier for Twin Zygosity

Zexi Liu (813212),

Haoze Xia (1131343),

Atefeh Zamani (1129712),

Ni Zhang (1081143),

Tianyu Zhou (1199306)

Supervisors:

Michael Kirley and Jiadong Mao

Host Organisation:

Twins Research Centre of Epidemiology and Biostatistics UoM

Contents

1 Introduction

Twins refer to two offsprings produced by the same pregnancy, and they can be categorized as dizygotic (DZ) and monozygotic (MZ) twins. DZ twins, also known as fraternal twins or nonidentical twins, are two siblings coming from separate eggs, released at the same time from an ovary and are fertilized by separate sperms. On the other hand, if early in development, a single fertilized egg cell, the zygote, is divided into two or more embryos, MZ or identical twins are developed. The process which results in this type of twinning is still elusive and is a long-standing enigma of human developmental biology [van Dongen et al., 2021]. Although MZ twinning is a rare event in families, research shows that the chance of occurrence (prevalence) is similar across the world (3–4 per 1000 births), Segal [2017]. Besides, MZ twinning is stable with the mother's age, Hoekstra et al. [2008]. Consequently, with no clear genetic or environmental cause, the prevailing hypothesis is that identical twins arise at random, Lambert [2021].

Determining if someone is a MZ twin or not is vital since it helps us to find out if an individual is a twin who was separated at birth or they lost one of their twins during a multiple pregnancy, a phenomenon known as vanishing twin syndrome. Additionally, MZ twins are predisposed to a range of conditions from left-handedness to certain congenital disorders, such as spina-bifida. Therefore, for some people who suffer from these conditions, the stem might be an unknown identical twin. Identifying the root of these kinds of disorders might be beneficial for finding the proper treatment or cure.

A new aspect that attracts researchers' attention is classifying MZ and DZ using machine learning techniques on epigenetic data. As one of the most recent researches in this area, in 2021, van Dongen et al. [2021] demonstrated that while monozygotic twins have the same DNA sequence, they could have different epigenetic modifications. Moreover, they observed that MZ twins have a robust DNA methylation signature in somatic tissues, and DNA methylation differences in MZ twins are not randomly distributed across the genome. Their research clarified that these differences are enriched in certain parts of particular chromosomes and genes. These uniquely common shared marks among MZ twins were applied in training two epigenetic predictors to classify MZ from DZ twins or MZ twins from DZ twins and other family members. The accuracy of prediction was reported to be up to 80 percent. This project aims to study and improve these classifiers.

The goal of this project is to develop an epigenetic classifier for twin zygosity, trained from

high-throughput DNA methylation data using machine-learning techniques. In this project, we collaborate with researchers from the Twins Research Centre of Epidemiology and Biostatistics UoM, a national research centre dedicated to twin studies. The project has the following 2 aims:

- Aim 1: Validate the classifiers provided by van Dongen et al. [2021] on new datasets to evaluate their performance.
- Aim 2: Train new classifiers using different methods to obtain better performance compared to classifiers provided by van Dongen et al. [2021].

From a data science perspective, the main challenges when dealing with DNA methylation datasets include:

(i) Understanding the problem and the terminology. To be able to analyse the DNA methylation data and produce a meaningful interpretation of the results requires us to have certain knowledge about the biological background and the biomedical techniques producing the data. DNA methylation datasets used in this project are produced by HumanMethylation450 or HumanMethylationEPIC arrays.

(ii) Data preparation. Data cleaning and preparation is another important step before data analysis. For biological datasets, this step is often time-consuming, but it is essential for improving the quality of the data.

(iii) Sharing and Process Data. First, datasets in van Dongen et al. [2021] are not accessible to the public except for one, BSGS in Table 1. Second, The DNA methylation datasets are big datasets with extremely high dimensionality. The datasets we process have 450,000 covariates. Methods developed in this project have to be computationally efficient to be able to handle these datasets.

(iv) Analytical Challenges: In light of the ultra-high dimensionality of the datasets, one of the main analytical challenges is dimension reduction. In machine learning, there are a considerable number of methods for building classifiers for high-dimensional data. Comparing the performances of these methods and making practical recommendations for biologists are some of the main tasks of our project.

2 Related works

Researchers on classifying twins date back to the early 20-th century. At that point of time, the close resemblance of twins in physiognomy to such a degree that even near relatives could not tell them apart was considered as the easiest and best criterion for distinguishing MZ from DZ twins, Dahlberg [1926]. This method, however, has a drawback in that th ae judgment of the degree of resemblance depends largely on the subjective sense, and is naturally more or less arbitrary.

Another method to classify twins into MZ and DZ is based on their genetic markers, such as polymorphic blood markers, Smith and Penrose [1955] and Juel-Nielsen et al. [1958]. As suggested by Race et al. [1968], if the blood groups are different, the twins are certainly DZ, but if they are the same, there is a chance that they are MZ or DZ twins, which can be calculated. Although being accurate, studying blood markers is expensive and, when dealing with large numbers of twins, is hard to obtain. Besides, there are ethical considerations in obtaining blood specimens.

Along with these methods, questionnaires, Nichols and Bilbro Jr [1966], Cohen et al. [1973], Cohen et al. [1975] and Sarna et al. [1978], and logistic regression, Spitz et al. [1996] are also applied as methods for classifying twins into MZ and DZ groups.

As mentioned in Section 1, van Dongen et al. [2021] is one of the most recent researches on classifying the zygosity of twins. Compared to the previous studies, this research is a great improvement in studying twins and can be considered as the groundwork for further research in this field. In this research, the focus is on the methylation signature of twins, and it is claimed that because identical twins keep a lifelong molecular signature, they can retrospectively diagnose if a person was conceived as a monozygotic twin or not. The research covered a large number of samples from different parts of the world, and they investigated epigenetic differences at over 450k sites along the genomes. Their findings demonstrated that at 833 spots along the genome, identical twins were strikingly similar. They applied penalized regression models (elastic net) for classification, and the reported accuracy of the proposed method was reported to be up to 80 percent.

In the following sections, the provided method in van Dongen et al. [2021] will be studied in more details and will be implemented to classify some new datasets.

3 Data analysis and preliminary model development

The focus of the project in the first semester was on Aim 1. To gain a clear picture of what has been done in this part of the project, we describe the existing classifiers trained in van Dongen et al. [2021], provided the client's data sets, applied testing methods on the client's datasets and the obtained results for this part of the project (Aim 1).

3.1 Classifiers Introduced by van Dongen et al. [2021]

van Dongen et al. [2021] trained two classifiers, referred to as classifiers 13 and 14, from the preprocessed data. In this section, we describe the datasets they used, how classifiers 13 and 14 were trained and their performance on test datasets.

3.1.1 Datasets and Data

In the mentioned research van Dongen et al. [2021] one dataset (NTR) in Table 1,was split 70 % for training and cross-validation, 30 % for testing. Five others in the Table 1 were used for testing. As can be seen in Table 1, these datasets consist of more than 6000 samples, which included MZ twins, DZ twins and family members. The average dimension of the datasets is around 1000 * 450,000. A row is a person, and a column is one CpG site on DNA sequence.

Figure 1: The CpG sequence of one DNA strand

A CpG site is a subsequence structure on DNA sequence shown in Figure 1 and DNA methylation data, or called β value, which range from 0 to 1, indicates the proportion of DNA that is methylated at a specific CpG in a sample. Data was measured on each site by machine chips and then standardized. The machine used for measurement is called Illumina BeadChips according to the manufacturer's protocol: the Illumina Infinium HumanMethylation450 BeadChip (450k array), which measures more than 450,000 methylation sites (majority of cohorts), or the Illumina

MethylationEPIC BeadChip (EPIC array), which measures more than 850,000 methylation sites van Dongen et al. [2021].

Table 1: The datasets and the number of MZ twins, DZ twins and family members analysed by van Dongen et al. [2021]

	N Total	N MZ Twins	N DZ Twins	N Family Members
NTR	1957	924	1033	237
E-Risk	1164	470	694	-
FTC	1708	559	1149	-
TwinksKU	492	395	97	-
BSGS	356	134	222	257
NTR (Buccal)	765	564	201	-

Preprocessing the CpGs cites (columns) and samples (rows) is in the following. In the training data (NTR), zygosity was regressed on all methylation CpGs sites (N = 381,376) that (1) were present both on the Illumina 450K and EPIC array (2) survived quality control in the training set (NTR) and in the test data sets (NTR-buccal, Australia-blood). van Dongen et al. [2021] also dropped samples in NTR dataset with missing data in more than 5% of the CpGs and also excluded CpGs with missing data in more than 5% of the samples. Last, using some meta-analysis, they identified a subset of epigenome-wide significant CpGs, which is consisted of 833 CpGs.

3.1.2 Classifier 13, 14 training and performances

Using the 833 CpGs obtained from the meta-analysis, two classifiers, 13 and 14 were trained using the penalized logistic regression model. It was reported to outperform when classifiers were trained on N = 381,376 CpGs cites data in Supplementary Data 12. f Classifiers 13, 14 were 10-th folded cross-validated in the training dataset (NTR), the hyperparameters α was fixed to be 0.5, the λ was picked to be minimal. The differences between these two classifiers are in the following :

- **Classifier 13**
  - This classifier was trained on data from twins only, to classify their zygosity. In the final model 251 variables have nonzero coefficients/251 variables were selected. We have access to the model's coefficient's value in link,Supplementary Data 13.
- **Classifier 14**

– This classifier was trained on data from twins and a small group of family members of twins to distinguish MZ twins from the rest (dizygotic twins and family members). In the final model 251 variables have nonzero coefficients/251 variables were selected. We have access to the model's coefficient's value in Supplementary Data 14.

The area under the curve (AUC) of the receiver operating characteristic (ROC) curve, was applied to measure the accuracy of the classifiers. Based on the information provided in the paper and its supplementary documents, the AUC of the best performing predictors were reported to be 0.77 and 0.80, respectively, in an independent blood data set from NTR (N about 1000) and in blood data from a second independent twin cohort (N=606, BSGS), more test results conducted by the paper, is in Supplementary Data 12.

3.2 Data Sets Client provided for testing

For the present research, the clients provided 5 datasets to be analysed. Table 2 provides some information about these datasets. These datasets provide DNA methylation data and are quantified using the Illumina Infinium HumanMethylation450 BeadChip ("450K array") in DNA samples.

We still consider the problem as binary classification even if the client's provided dataset has more than 2 classes in Table 2. This is to be consistent with the van Dongen et al. [2021]'s results and classifiers. The third class is the family member in BSG and AMDTSS. We combine the samples of family class into the DZ twin, so that it would be a binary classification of MZ vs. everyone else (ie. MZ vs. (DZ and family members)). And for E-risk, XZ is unlabeled data, we removed these samples in testing.

Table 2: The datasets and the number of MZ twins, DZ twins, XZ twins and family members provided by the client.

	N Total	N MZ Twins	N DZ Twins	N XZ Twins	N Family Members
E-Risk	1658	860	608	190	-
AMDTSS	479	132	132	-	215
BSGS	614	135	223	-	256
Denmark	180	94	86	-	-
E-MTAB	648	240	408	-	-

3.3 Applied testing methods

We have access to the trained weights for van Dongen et al. [2021] penalized logistic regression model in Supplementary Data 13 and Supplementary Data 14. Then we filtered the 450K to match the exact 251 covariates for classifier 13 and 231 covariates for classifier 14. Next, we use the matched covariates in data and the coefficients to compute the logistic regression response and last calculate the confusion matrix and AUC scores.

3.3.1 Linear Separability of the Pre-processed Datasets

For the E-Risk dataset in Figure 2, we plotted PCA for the filtered 251 covariates data for classifier 13 and 231 covariates data for classifier 14 to check the linearity. PC1 and PC2 explain 11.6 % and 5.7 % of the 251 covariates, classifier 13 on the Left. Similarly, for classifier 14, 231 covariates dataset, PC1 and PC2 explains 11.3 % and 5.7 %. The percentages of PC1 and PC2 to explain the data variance is very low, and we did not find a clear linear boundary to separate classes.

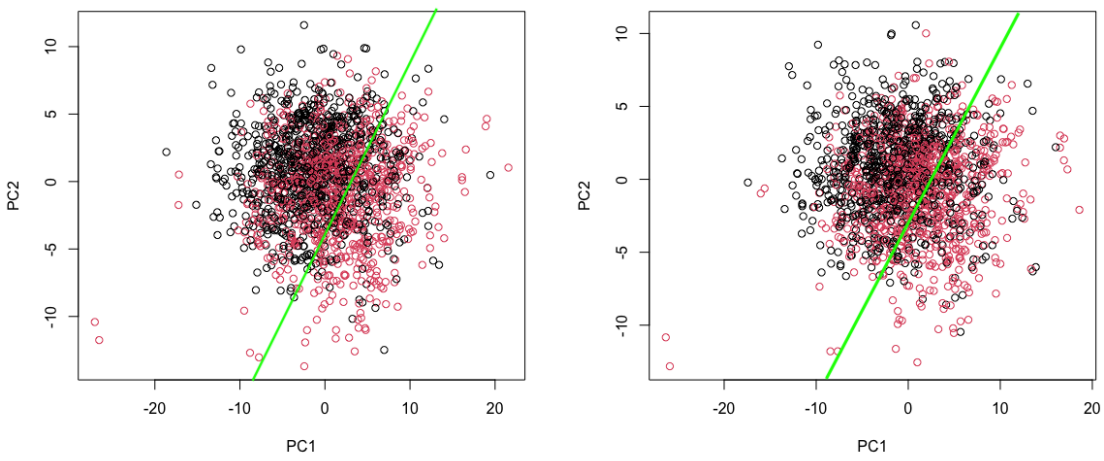

Figure 2: PCA on dataset NCBI-GSE105018 matching the coefficients of classifier 13 (left hand side) and 14's coefficients(right hand side)

3.4 Main Results

van Dongen et al. [2021] claimed that classifier 13 and 14 have similar prediction results. However, we may find different conclusions from the client's datasets. In Table 3 and Table 4's last column,

we can see that classifier 14 has a better performance when predicting non-twins (family members) than classifier 13. We suspect this may be due to differences in training datasets, as classifier 14 was trained from the dataset containing family members while classifier 13 was not.

Table 3: The performance of classifier 13 on different test datasets

Classifier 13	AUC	Proportion MZ twins correctly predicted	Proportion DZ twins correctly predicted	Proportion non-twins correctly predicted
NTR - testing	0.751	0.843	0.469	0.444
BSGS - nature	0.796	0.916	0.450	0.302
E-Risk	0.728	0.684	0.683	NaN
BSGS	0.784	0.904	0.404	0.313
Denmark	0.544	0.596	0.570	NaN
AMDTSS	0.648	0.742	0.545	0.567
E-MTAB	0.522	0.882	0.201	NaN

Table 4: The performance of classifier 14 on different test datasets

Classifier 14	AUC	Proportion MZ twins correctly predicted	Proportion DZ twins correctly predicted	Proportion non-twins correctly predicted
NTR - testing	0.766	0.808	0.572	0.625
BSGS - nature	0.799	0.901	0.468	0.451
E-Risk	0.739	0.621	0.740	NaN
BSGS	0.774	0.637	0.798	0.848
Denmark	0.563	0.755	0.395	NaN
AMDTSS	0.667	0.750	0.530	0.600
E-MTAB	0.492	0.564	0.390	NaN

Besides, Classifier 13 and Classifier 14 may not be generalizable to the datasets provided by the client. As can be seen in Table 5, the average AUC over all test datasets for both classifier 13 and 14 is around 0.65. This is a low number because AUC 0.5 means classifying randomly. Besides, for both classifiers, the dataset E-MTAB has an extremely low AUC around 0.5. Moreover, the Denmark dataset also has a low AUC around 0.55. These differences in the performance compared to the ones mentioned in van Dongen et al. [2021] can put extendability of the classifiers to new datasets in doubt.

We have also attempted to validate the correctness of our codes. The BSGS dataset has been tested and reported in van Dongen et al. [2021] and we also have access to it and can run a test on it. As can be seen in Tables 3 and 4, the AUCs are very similar, but the confusion matrix for classifier 14 is different from their results. We will bring further investigation to this issue with the

clients.

Classifiers 13 and 14 have low type 1 error but high type 2 error. In the second and third columns of Table 5, the average proportion of DZ twins correctly predicted is around 0.53, but the average proportion of MZ twins correctly predicted is very high. This pattern may be due to their unbalanced training data. In training data, the number of MZ twins is twice the amount of DZ twins. Therefore, models 13 and 14 may tend to predict more MZ twins in testing.

Table 5: The Average values of the five testing datasets

Classifier 14	AUC	Proportion MZ twins correctly predicted	Proportion DZ twins correctly predicted	Proportion non-twins correctly predicted
Classifier 13	0.645	0.726	0.481	0.440
Classifier 14	0.647	0.665	0.571	0.724

4 Plan for semester 2

In semester 2, the focus will be on achieving Aim 2. That is, we aim at training a new classifier outperforming classifiers 13 and 14 as introduced in Section 3.1. This is also the hurdle goal for the next semester.

First, We will try and apply dimension reduction methods such as PCA to get the most significant dimensions from 450k DNA methylation data. It can reduce a great amount of model calculations while retaining most of the related information.

After data cleaning and pre-processing, the dataset E-Risk will be used to train the method, and the datasets AMDTSS, BSGS, Denmark and E-MTAB will be applied to test and evaluate the performance of the obtained classifier AUC.

We will try different kinds of machine learning (ML) techniques to get the model with the best performance. The classical ML methods, such as support vector machine (SVM), logistic regression, Bayesian algorithms and random forests will be tested first. Next, we will try some boosting techniques including adaptive boosting (Adaboosting) and stochastic gradient boosting. These are integrated techniques that attempt to create strong classifiers. Table 6 pointed out some pros and cons of these methods. Finally, we will also try some deep learning methods, which create multiple hidden layers to learn intricate structures in data. In addition, several models can be combined using Stacking and Embedding techniques to produce better outputs.

Table 6: Machine learning methods applicable in classification problems

Method	Pros	Cons
SVM	• Insensitive to over-fitting	• Might not be suitable for large data-set
Logistic regression	• Good accuracy for many simple data-sets • Good performance for linearly separable data-sets	• High dimension data tend to over-fit the model • Feature selection/ dimensionality reduction is necessary
Bayesian algorithm	• Good performance on small-scale data • suitable for incremental training • Not sensitive to missing data • relatively simple algorithm	• Need to calculate the prior probability • Classification decision has error rate • Not good if the sample attribute is related
Adaboosting	• High-precision classifier • Simple to implement • Overfitting is not easy to occur	• The number of iterations is not easy to set • Sensitive to data imbalance • Training is time-consuming
Stochastic gradient boosting	• Avoid the problems from overfitting • high performance in high dimensional data • Using boosting algorithms • Consistent approximation	• High computational complexity
Random forests	• Not computationally intensive • high generalization accuracy • Low classification error rates • Little need to tune parameters • Robust and does not overfit	• For very large data sets, the size of the trees can take up a lot of memory. • Poor performance on imbalanced data

The key challenges for aim2 can be divided into 3 points. Firstly, the size of our datasets is huge. Each sample in the normalized dataset has 450k predictors. For example, the size of the E-RISK dataset for training is about 14GB. Because of this, it would be difficult for our local machines to process the datasets and build our models.

We have proposed two ways to solve this challenge. First, find a method that can deal with a huge size of dataset on a local machine. In the meanwhile, we are going to learn how to deploy codes on Cloud services, such as Spartan, and use Cloud Computing technology to solve this challenge.

The second challenge is that the number of variables in the dataset is much larger than the number of samples, so we are dealing with high dimensional datasets. To address this challenge, we need to do some research and testing to find the most suitable machine learning algorithm for this data structure.

The last challenge is about our unbalanced training data. To solve this problem, we may restructure the training data to make it more balanced. For example, we may do over-sampling for the minority class or perhaps under-sampling the majority class. The random forest algorithm also works well with unbalanced data.

Figure 3 provides the suggested project plan for Semester 2.

Figure 3: The Gantt graph presenting the project plan for the second semester

References

Jenny van Dongen, Scott D Gordon, Allan F McRae, Veronika V Odintsova, Hamdi Mbarek, Charles E Breeze, Karen Sugden, Sara Lundgren, Juan E Castillo-Fernandez, Eilis Hannon, et al. Identical twins carry a persistent epigenetic signature of early genome programming. *Nature communications*, 12(1):1–14, 2021.

Nancy L Segal. *Twin mythconceptions: False beliefs, fables, and facts about twins*. Academic Press, 2017.

Chantal Hoekstra, Gonneke Willemsen, Toos CEM van Beijsterveldt, Grant W Montgomery, and Dorret I Boomsma. Familial twinning and fertility in dutch mothers of twins. *American Journal of Medical Genetics Part A*, 146(24):3147–3156, 2008.

Jonathan Lambert. All identical twins may share a common set of chemical markers on their dna. `https://www.sciencenews.org/article/identical-twin-siblings-common-set-chemical-markers-dna-epigenetics/`, 2021. Accessed: 2021-09-28.

Gunnar Dahlberg. Twin births and twins from a hereditary point of view. 1926.

Sheila Maynard Smith and LS Penrose. Monozygotic and dizygotic twin diagnosis. *Annals of Human Genetics*, 19(4):273–289, 1955.

Niels Juel-Nielsen, Arne Nielsen, and Mogens Hauge. On the diagnosis of zygosity in twins and the value of blood groups. *Acta Genetica et Statistica Medica*, pages 256–273, 1958.

Robert Russell Race, Ruth Sanger, and Ronald Fisher. *Blood groups in man*, volume 293. Blackwell Scientific Oxford, 1968.

Robert C Nichols and WC Bilbro Jr. The diagnosis of twin zygosity. *Acta genetica et statistica medica*, pages 265–275, 1966.

Donald J Cohen, Eleanor Dibble, Jane M Grawe, and William Pollin. Separating identical from fraternal twins. *Archives of General Psychiatry*, 29(4):465–469, 1973.

Donald J Cohen, Eleanor Dibble, Jane M Grawe, and William Pollin. Reliably separating identical from fraternal twins. *Archives of General Psychiatry*, 32(11):1371–1375, 1975.

Seppo Sarna, Jaakko Kaprio, Pertti Sistonen, and Markku Koskenvuo. Diagnosis of twin zygosity by mailed questionnaire. *Human heredity*, 28(4):241–254, 1978.

Elisabeth Spitz, René Moutier, Terry Reed, Marie Claire Busnel, Catherine Marchaland, Pierre L Roubertoux, and Michèle Carlier. Comparative diagnoses of twin zygosity by sslp variant analysis, questionnaire, and dermatoglyphic analysis. *Behavior genetics*, 26(1):55–63, 1996.

CpG sequence of one dna strand versus c-g base pair on complementary strands. `https://commons.wikimedia.org/wiki/File:CpG_vs_C-G_bp.svg`. Accessed: 2022-05-27.