methXGB package: user manual

Alberto J. Leon[1]

[1]Ontario Institute for Cancer Research, Toronto, Canada

10 December 2018

Package

BiocStyle 2.10.0

Contents

The package methXGB contains trained XGBoost models (Chen and Guestrin 2016) that can be used to infer tumour purity and extent of immune infiltrate (immunescore) in the DNA methylation data from tumour samples analyzed with Illumina's 450k and EPIC microarrays.

1 Introduction

A machine learning framework to infer biological variables in DNA methylation data from Illumina 450K and EPC arrays from different cancer types. The methXGB package contains machine learning models based on the XGBoost algorithm, that can be used interpret DNA methylation data from different cancer types. The XGBoost models were trained using DNA methylation data expressed as M-values. Nonetheless, the package also accepts beta-values as input (converted internally).

2 Installation

The package methXGB is distributed through Bioconductor, and the standard method to install its packages apply.

```
install.packages("BiocManager")
BiocManager::install("methXGB")
```

3 Load Data

The methXGB package takes as input a data frame containing M-values, where the column names correspond to the sample identifiers and where each row name is an Illumina probe ID. To load the example data included in the package:

```
library(methXGB)
data(mvalGBM)
view(mvalGBM) #in RStudio, a screenshot of the top 7 rows is provided below
```

	DIAM_351	DIAM_590	DIAM_753	DIAM_765	DIAM_778
cg00000029	-0.4953237	0.3201540	0.7392343	1.9537716	0.3951609
cg00000165	-2.3335115	0.6271613	1.7999966	2.4114230	1.1438947
cg00000236	1.5179594	2.8923605	2.7768141	3.3835264	1.5143376
cg00000289	1.2063658	2.6997941	3.0955713	2.0756709	1.6513726
cg00000292	-1.6901414	0.0308565	0.9696212	1.3498444	-0.4786496
cg00000321	1.1847025	1.6399148	2.0902622	2.1663696	1.2297973
cg00000363	-2.3986025	-1.1075816	-1.2505528	1.3730515	-2.0590474

table_view_mvalGBM

A typical workflow starts with the idat files being processed with the minfi package (Aryee et al. 2014) for normalization (i.e. Illumina method) and extraction of M-values. We recommend the use of native save and load functions to store and retrieve data frames containing M-values. At the end of a minfi workflow, store the M-values in a data frame called “mval”, column names samples and row names Illumina probe IDs.

```
save(mval, file="M_myCancerStudy.RData")
To load the data frame:
load(file="M_myCancerStudy.RData")
```

the data frame named “mval” is loaded into the workspace Beta-values can be transformed into M-values as follows: beta.values is a data frame where column names are sample names and row names Illumina probe IDs.

```
mval<-betaToMval(beta.values)
```

This transformation is described in the paper xxx (Du et al. 2010).

3.1 M-values extraction from microarray data (idat files)

The methXGB package incorporates methods to produce readily usable M-values from raw microarray data (idat files). This workflow uses GEOquery to fetch DNA methylation data from the Gene Expression Omnibus, and minfi to normalize the microarray data with the Illumina method, followed by M-value extraction.

```
idatParse(in="GSE60274",task="GEOquery") #downloads and extracts idats
idatParse(in="GSE60274",task="generate_samplesheet")
idatParse(in="GSE60274",task="extract_mval")
#This results in a file called M_ GSE60274.RData.

#output of the last process (summary)
#Finally, to load the data frame mval contained in the RData file:
load(file="M_myCancerStudy.RData")
```

The `idatParse()` functon is designed to provide users readily access to data in M-values format, but it is not a replacement for cumstom-built minfi workflows.

4 Estimation of biological variables in DNA methylation data

Different models (families) have been trained. To view the models that are included in the package: (displays model_summary.txt) methXGB.summary() Family: purity Description: .. Family: immunescore Description: . For family-level details, use methXGB.summary(family=".")

4.1 Infer tumour purity

```
methXGB.summary(family="purity")
```

Something goes here.Absolute?? To infer the tumour purity in DNA methylation data:

```
pred.purity<-methXGB.pred(m.values=mval, model.family="tumour_purity",training.
set="TCGA-LGG")
```

4.2 Infer immunescore

The following XGBoost models were trained using the TCGA 450k data from different tumour types, and ESTIMATE immunescore (Yoshihara et al. 2013) which was derived from an mRNA expression pan-cancer signature. See the ESTIMATE paper for details. Only tumour types with more than 100 with 450 data and ESTIMATE.immunescore (RNA-seq_v2).

```
methXGB.summary(family="immunescore")
```

To infer the immunescore in DNA methylation data:

```
pred.immunefscore<-methXGB.pred(m.values=mval, model.family="immunescore",train
ing.set="TCGA-LGG")
```

5 Dataset-level result cross-validation

We have included a cross-validation method that employs a different methodology. There are situations were we want to obtain validity of the results. This can be applied to situations were a tumour type of interest was not trained due to limited data availability.

Session info

Output of `sessionInfo()` :

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] BiocStyle_2.10.0
##
## loaded via a namespace (and not attached):
##  [1] BiocManager_1.30.4 compiler_3.5.1     magrittr_1.5
##  [4] bookdown_0.8       tools_3.5.1        htmltools_0.3.6
##  [7] yaml_2.2.0         Rcpp_1.0.0         stringi_1.2.4
## [10] rmarkdown_1.11     knitr_1.20         stringr_1.3.1
## [13] xfun_0.4           digest_0.6.18      evaluate_0.12
```

The trunk…

Let's see what happens… (Aryee et al. 2014). Let's see what happens… (Du et al. 2010). Let's see what happens… (Chen and Guestrin 2016). Let's see what happens… (Yoshihara et al. 2013).

Table 1: **A knitr kable**

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

References

Aryee, Martin J, Andrew E Jaffe, Hector Corrada Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. 2014. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.” *Bioinformatics* 30 (10): 1363–9. doi:10.1093/bioinformatics/btu049 (https://doi.org/10.1093/bioinformatics/btu049).

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *ArXiv*. doi:10.1145/2939672.2939785 (https://doi.org/10.1145/2939672.2939785).

Du, Pan, Xiao Zhang, Chiang-Chiang Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. 2010. “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.” *BMC Bioinformatics* 11: 587. doi:10.1186/1471-2105-11-587 (https://doi.org/10.1186/1471-2105-11-587).

Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martinez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Trevino, et al. 2013. “Inferring tumour purity and stromal and immune cell admixture from expression data.” *Nat Commun* 4: 2612. doi:10.1038/ncomms3612 (2013) (https://doi.org/10.1038/ncomms3612 (2013)).