Short report

Hsin

Abstract

Introduction

In neuroscience research, behaviour analysis plays an important role in validating novel treatment. Depending on animal model and therapy, researchers choose various behaviour assessments to investigate effectiveness of medical interventions. Gait is one of the essential behaviour tests to examine mobility alteration in lysosomal storage disorders, such as Niemann-Pick diseases. There are automated systems, such as Catwalk (Noldus), that are able to collect and analyze gait data with minimized human effort. However, Catwalk system generates numerous variables which are around 300. Some scientists filter these variables via hypothesis-driven approach. This only investigates certain variables that are selected by researchers' assumption. Others merely reported significant results. Neither can provide objective judgment of the gait result, because selected variables were not represented the animal model. Here, we proposed a method of variable selection to avoid subjective conclusion. In order to distinguish relevant variables, the mixed ANOVA analysis is introduced. The method is considered individual subjects, age and group of a animal model effects on a study. This can construct growth curve in wild type group and disease progression in individual variable. In other words, the mixed ANOVA model identifies variables which are significant different between growth curve and disease progression. Thereafter, animal model will have its unique variable profile. The resulted profile is according to animal model, which means the selection process can provide a set of variables tailoring to specific disease studies. The treatment can be inspected in these variables which are affected by a disease. Instead, researchers utilized biased variable selection. Interpretation of behavioral study for inexperienced researchers is a challenging task and involved largely personal judgment. A summarized result of assessments can help scientist to determine the effect of treatment on animal model. We proposed a consultation group of machine learning algorithms in order to provide a unbiased conclusion. In this project, five classic classifiers, such as K-Nearest Neighbors(KNN), Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machines(SVM) were demonstrated the consultation group concept. These algorithms indicate the likelihood of behaviours between treated group and neither wild-type or disease groups. These results will be the reference for researchers to make a conclusion.

#20/09/2021 meeting In tanganil treatment study, 9 week of NPC mice, negative control group, lacked of gait data because of poor paw detection. This

brought a challenging task to analyze the effectiveness of one week treatment from 8 week to 9 week. A method of two stages of analysis was proposed to deal with this dataset. First Mixed anova is to select relevant variables and followed by t-test. With selected variables, t-test is conducted to compare treated group and wild-type group at 9 week. In the result, the non-significant variables indicates that the one week treatment benefits in these behavioural aspects and vice versa. To rephrase it, this analysis is to investigate whether the treatment can alter disease progression towards wild-type feathers while by-passing the negative group comparison.

Methods

Tanganil treament

There were three groups, such as wild-type, NPC and treated groups. The behaviour data were collected one week apart between age of 6 week old and 8 week. The treatment was introduced at 8 week of homozygous mice.

Programming environment

R version : 4.0.3 library : ANOVA analysis : tydiverse, nlme packages Machine learning algorithms : caret, DoMc packages

Mixed ANOVA model

In order to determine the relevant variables which are affected by disease, data with wild type and disease groups in different time points was analyzed by mixed ANOVA model. This can indicate the difference between growth curve and disease procession. nlme package in R was applied in here.

```
library(nlme)
library(tidyverse)
gait_data=nlme::groupedData(from=Y~1|ID,data=gait_data)
aov(data=gait_data,Y~Age*Group+Error(ID))
# mixed aov : repeated measures and various time points
```

- gait_data= the data generated from Catwalk system
- Y= one of variables from gait_data

T-test analysis

In the second stage, t-test is for comparing treated and wild-type groups in resulted variables from mixed ANOVA model. The programming code is described below :

```
f <- var.test('wild-type data','NPC data',
              ratio = 1,alternative = 'two.sided',
              na.action=na.omit())
# variance test before t-test
```

```r
if(!is.na('result_var.test')){
  ifelse('result_var.test'>=0.05,
         t <- t.test('wild-type data','NPC data',
                     var.equal = T,alternative = 'two.sided',
                     mu=0,na.action=na.omit()),
         t <-t.test('wild-type data','NPC data',
                     var.equal = F,alternative = 'two.sided',
                     mu=0,na.action=na.omit()))
# use t-test according to the result of variance test.
```

Machine Learning Algorithms

The data from wild-type and disease groups were fed to machine learning algorithms. The mixed ANOVA analysis is regarded as features selection process. The selected variables were to train the classifiers. The training and 5 fold cross validation were executed by CARET package in R.

```r
library(tidyverse)
library(caret)
library(doMC)

fitControl <- trainControl(
  method = 'repeatedcv',number = 5,
  savePredictions = 'final',
  # save prediction for optimal tuning parameter
  classProbs=T,
  # class probabilities
  summaryFunction=twoClassSummary,
  # results binary classes summary function
  allowParallel = T
  # multiple cores paralleling calculation
                           )
model_selected <- c('regLogistic','naive_bayes',
                    'glm','xgbTree','knn')
  # five classifiers selected
ml_list <- list()
  # a list to store model result
for (i in 1:length(model_selected)) {
  ml_list[[i]] <-
    caret::train(Group~.,
    data=trainData,method=model_selected[[i]],
    trControl=fitControl,tuneLength=5,metric='ROC')
```

- fitControl = the setting of 5 folds cross validation
- model_selected = the classifier chosen for this study
- ml_list = the result of the trained classifiers
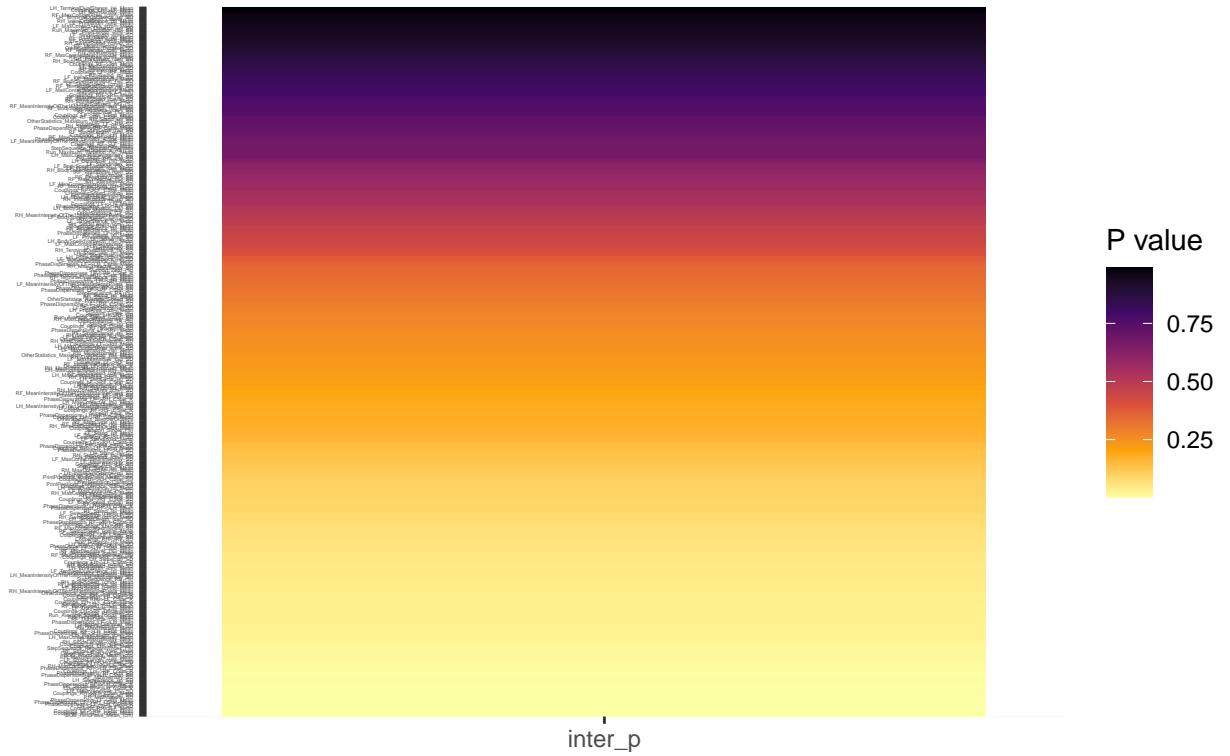
Figure 1: The heapmap of Mixed AOV result

Results

The two-stage method was demonstrated by tanganil one week treatment study. This was proposed to solve the problems of variable selection and absence of NPC data at age of 9 week. Wild-type and disease groups firstly was analyzed by mixed anova model. The relevant variables from Catwalk system were selected, Figure 1. These were represented the disease affected variables. Afterward, t-test between wild-type and treated groups at 9 week old was shown the benefits of tanganil in the NPC animal model, Figure 2.

The unique variables profile was provided by mixed ANOVA, which mean the profile depends on the disease model. It was shown in heatmap. The colour codes are according to p-values of mixed ANOVA analysis. Classifiers which were trained by resulted data from mixed ANOVA model were K-Nearest Neighbors(KNN), Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machines(SVM). The
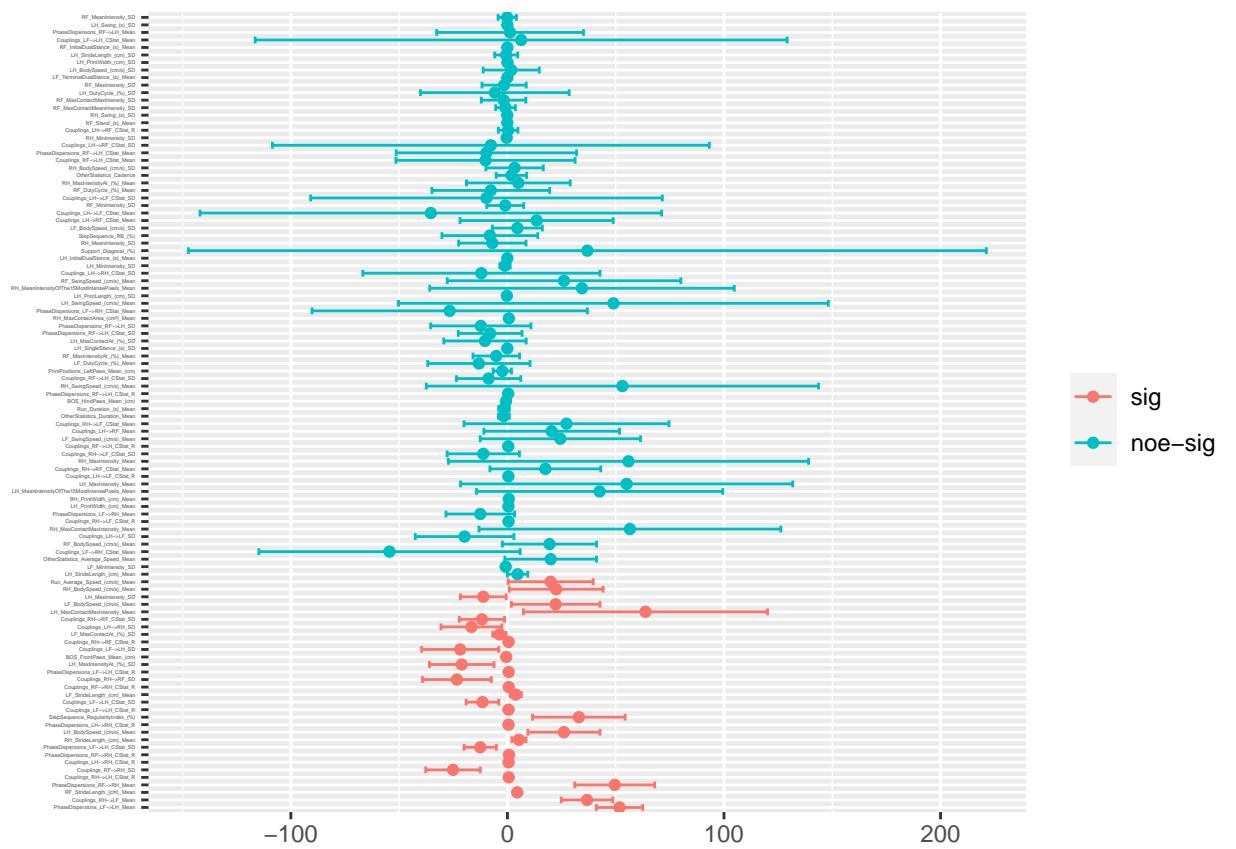
Figure 2: The result of t-test : red colour indicates the significant different between wild-type and treated group at ago of 9 week. blue colour shows opposite result. In here, the therapy benefits in blue variables