Team Meeting

06 MAY 2022 / 14:15 AM / ZOOM

Attendees

Ate, Xavier, Jesse, Ni, Stefan, Jiadong

Agenda

Last Meeting Follow-up

1. Result report
    a. Missing data “NA”, (194 in 1658)
    b. 0.5971 AUC for classifier of supplementary 13 (MZ vs DZ), 0 AUC for classifier of supplementary 14 (MZ vs (DZ & family))
    c. Confusion matrix for classifier 13

	mdz	
Estimate	DZ	MZ
DZ	177	81
MZ	435	771

    d. Excel file

New Business

1. Machine learning Classifiers (ideas need to search in google scholar)

Classifier	Pros	Cons
SVM	- The dimensionality reduction process can be rejected to achieve high classification accuracy. - Insensitive to overfitting.	- Not suitable for large dataset (time consuming)
Logistic Regression	- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.	- High dimension data tend to overfit the model - feature selection / dimensionality reduction process is necessary.
Bayesian Algorithm	- Good performance on small-scale data, suitable for incremental training - Not sensitive to missing data and the algorithm is relatively simple, which is easy to interpret	- Need to calculate the prior probability - Classification decision has error rate - Not good if the sample attribute is related
Adaboosting	- High-precision classifier that can be used to construct sub-classifiers using various methods - Simple to implement - Overfitting is not easy to occur	- The number of iterations is not easy to set - Data imbalance leads to the decline of classification accuracy - The training is time-consuming
Stochastic Gradient boosting	- Avoid the problems arising from overfitting of its base classifier - Inherent variable selection and assigning variable amount of degrees of freedom to the selected variables by boosting algorithms could be a	- High computational complexity (due to low learning rate in shrinkage process)

	reason for high performance in high dimensional problems. - Boosting yields consistent function approximations even when the number of predictors grows fast to infinity, where the underlying true function is sparse.	
Random Forests	- Not computationally intensive - Limits correlation among trees can help in building an ensemble classifier with high generalization accuracy for high dimensional data problems - Low classification error rates (compare with boosting and svm) - 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

References:

http://plaza.ufl.edu/psnvijay/Site/Publications_files/Classification_HDD.pdf

https://www.usu.edu/math/jrstevens/bioinf/8.Forests.pdf - Random forests

https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-015-0723-9.pdf - Stochastic gradient boosting

2. Report

Notes

- Possible reasons for low AUC:
    - Different normalization methods
    - Something wrong in the code
    - Different distribution of the dataset
    - Not transform the response value (link function)
- Cut-off influences the confusion matrix, but the AUC and ROC

Action Items

- Double check the code for aim 1
- Ask shuai if the test data is normalized the same as the nature paper
- PCA - data visualization（scatterplot - different colors for different datasets)
- Ask access to the original dataset from the nature paper, replicate the testing step. (for debugging)
- Split our dataset to training and testing, retrain the model, then see whether AUC would be improved.

Next Meeting Agenda