Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease

Chao Gao [1,2], Hanbo Sun [1,3], Tuo Wang [1,3], Ming Tang [1,2], Nicolaas Bohnen [4,5], Martijn Muller [4,5], Talia Herman [6], Keren Rosenberg-Katz [6], William Dauer [5], Jeffrey Hausdorff [6], Ivo D. Dinov [1,7,8]*

[1] Statistics Online Computational Resource, Department of Health Behavior and Biological Sciences, [2] Department of Biostatistics, [3] Department of Statistics, [4] Department of Radiology, [5] Department of Neurology, [7] Department of Computational Medicine and Bioinformatics, [8] Michigan Institute for Data Science, University of Michigan, [6] Sackler Faculty of Medicine, Tel Aviv University,

Abstract

Parkinson's disease (PD) is an age-related neurodegenerative disorder affecting over 10 million people worldwide. PD itself is not fatal, however, people with PD suffer from disrupted balance, tremor, rigidity, slowness in movement, etc. Patients are at high risk of falling due to the motor impairment or freeze of gait. Serious falls, especially among elderly patients, can lead to a variety of injuries including head and bone fractures. Complications from these injuries may further lead to hospitalization or long-term rehabilitation or death.

In this study, we holistically investigated falls in PD patients using clinical, demographic and neuroimaging data from two independent initiatives (University of Michigan and Tel Aviv University). Using machine learning techniques, we constructed predictive models classifying fallers and non-fallers. Through controlled feature selection, we identified the most impactful predictors of patient falls including gait speed, Hoehn and Yahr scale, postural instability and gait difficulty-related measurements. The model-based and model-free analytical methods we employed include logistic regression, random forests, support vector machines, and XGboost. The reliability of the forecast was assessed by internal statistical (5-fold) cross validation as well as by external validation. Our findings suggest that model-free machine learning techniques provide the most reliable forecasting of falls with classification accuracy about 70-80%.

Keywords: *Parkinson's disease, machine learning, Big Data, clinical decision support, predictive analytics, falls*

Introduction

Model-based methods vs model-free methods

Both model-based and model-free techniques may be employed for prediction of specific clinical outcomes or diagnostic phenotypes. The application of model-based approaches heavily depend on the a priori statistical statements. On the contrary, the model-free methods adapt to the intrinsic data characteristics without the use of a priori model. Given complicated information, model-free techniques are able to construct non-parametric representations, which may also be referred as (non-parametric) models, using machine learning algorithms or ensemble of multiple base learners without simplification of the problem.

Predictive Analytics Strategy

(Top) Identify critical features and build predictive models independently on the Udall and the Tel-Aviv datasets, respectively. (Bottom) Harmonize and merge the two data archives and perform the same analytics on the aggregate data. The bottom-right branch of the diagram illustrates the process of training the models on one of the datasets and (externally) validating their accuracy on the other complementary dataset (**Figure 1**).

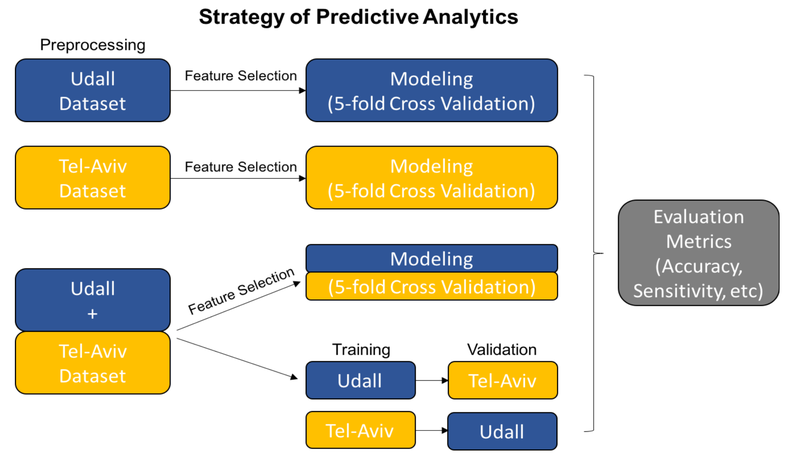

Figure 1. Predictive Analytics Strategy

Methods

Data Preprocessing

Data were provided by Udall Center (U-M) and Sackler Faculty of Medicine (Tel-Aviv U) independently. Despite the slight discrepancy in features collected, they both contain clinical measurements (e.g. Unified Parkinson Disease Rating Scale (UPDRS)), demographic information (e.g. age, gender), and MRI-based metrics. After imputation using *mice* (R package) and harmonization, we achieve the complete datasets (**Table 1**).

Cohort	Original Size(n)	Effective Size (m)	#Features
Udall	225(48)	148(45)	179
Tel-Aviv	105(41)	103(41)	165
Aggregate	330(89)	251(86)	129

Table 1. Basic summary statistics for the three datasets (n/m represents # fallers).

Feature Selection

Feature selection was carried out using two different methods: random forest (RF) [1] and Knockoff filtering (KO) [2]. We try to identify commonly selected features by both techniques that also show significant differences between "fallers" and "non-fallers" on the MWW and KS tests. (**Table 2**)

Diagnostic Prediction of Falls

For prediction (binary classification), we applied model-based (e.g., Logistic Regression) and model-free methods(e.g., Random Forest, Adaptive and gradient boosting[3,4], Support Vector Machines[5], Neural networks[6], SuperLearner[7]). The performance of each models are evaluated by accuracy(ACC), sensitivity(SENS), specificity(SPEC), Positive Predictive value (PPV), negative predictive value (NPV) and log diagnostic odds ratio (LOR).

Results

Cohort	Selected Features
Udall	PIGD_score, gaitSpeed(Off), MOT_EDL, NON_MOTOR_EDL, walking and balance, postural_stability
Tel-Aviv	gaitSpeed(Off), ABC, BMI, PIGD_score, cerebellum, getting out of bed, MOT_EDL, Attention, DGI, Tremor_score, FOG_Q, R_fusiform_gyrus, H_and_Y(Off)
Aggregate	gaitSpeed(Off), PIGD_score, MOT_EDL, BMI, getting out of bed, H_and_Y(Off), gait(Off)

Table 2. Summary of selected features

We fitted models with all features or selected features only. Given all the features at present (high dimension), logistic regression has the lowest performance among the implemented methods. Generally, the performance of the machine learning methods are increased (Results not shown). Here we present the best models fitted using selected features under each condition (**Table 3**).

Data	Method	acc	sens	spec	ppv	npv	lor
Udall*	Super Learner	0.78	0.47	0.92	0.72	0.80	2.34
Tel-Aviv*	Random Forests	0.80	0.68	0.87	0.78	0.81	2.68
Udall+Tel-Aviv*	Logistic Regression	0.77	0.43	0.95	0.82	0.76	2.696
Udall/Tel-Aviv**	Super Learner	0.76	0.56	0.89	0.77	0.75	2.310
Tel-Aviv/Udall**	SVM	0.80	0.44	0.95	0.80	0.80	2.75

Table 3. Best predictive models for each case (**5-fold cross validation;**external validation)

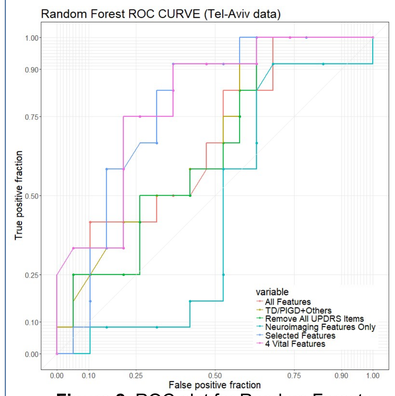

Figure 2. ROC plot for Random Forests

For Tel-Aviv data, 6 feature combinations were tried with Random Forests, ROC and AUC are shown in **Figure 2** and **Table 4**.

Training Condition	AUC
All Features	0.67
TD/PIGD*+Others	0.67
Remove UPDRS	0.64
Neuroimage Features	0.56
Selected Features	0.78
4 Features**	0.80

Table 4. Prediction performance (AUC)
*Tremor dominant/Postural instability/gait difficulty
**PIGD_score, gaitSpeed(Off), FOG_Q,H_and_Y(Off)

Discussion

In this study, model-free methods have shown better performance overall in predicting falls in PD patients. Feature selection leads to great improvement in the accuracy. Models with only MRI derived features are not able to classify "Fallers" and "non-fallers". In the future, we look forward to expanding our study on larger dataset and diving deeper in to neuroimaging data (e.g. MRI), with the hope to recognize patterns associated with falling of PD patients in the brain.

Contact
Chao Gao
Dept of Biostatistics, SOCR Lab
University of Michigan
Email: gchao@umich.edu
Phone: (734)355-4646

Acknowledgments
Colleagues at the Statistics Online Computational Resource (SOCR) and the Michigan Institute for Data Science (MIDAS) provided vital support and advice.
The analysis was conducted in R3.4.

References

[1] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.
[2] Barber, R.F. and E.J. Candes, A knockoff filter for high-dimensional selective inference. arXiv preprint arXiv:1602.03574, 2016.
[3] Rätsch, G., T. Onoda, and K.-R. Müller, Soft margins for AdaBoost. Machine learning, 2001. 42(3): p. 287-320.
[4] Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. ACM.
[5] Hearst, M.A., et al., Support vector machines. IEEE Intelligent Systems and their applications, 1998. 13(4): p. 18-28.
[6] Anagnostou, T., et al., Artificial neural networks for decision-making in urologic oncology. European urology, 2003. 43(6): p. 596-603.
[7] Van der Laan, M.J., E.C. Polley, and A.E. Hubbard, Super learner. Statistical applications in genetics and molecular biology, 2007. 6(1).