RESEARCH QUESTION, HYPOTHESIS AND PRELIMINARY DESIGN

Maks Drzezdzon |C15311966

TU060/2 |Data Science

DOMAIN AND SCOPE – ACM 2012

- Computing Methodologies => Machine Leaning => Machine Learning Approaches = > Learning Linear Models (UYSAL, L et al. 1999) (Lee, K.-Y et al. 2017) (Singh Suri et al. 2021)
- Computing methodologies => Machine Leaning => Machine Learning Approaches (Sadeghi, D et al. 2021)
- Computing Methodologies => Modeling and Simulation => Model Development and Analysis => Modeling Methodologies (Castanon, J. 2019, March 19) (Lin E et al. 2021)
- Computing Methodologies => Modeling and Simulation => Model Development and Analysis => Modeling Verification and Validation (Wang, H et al. 2013) (Colyer, A. 2019, June 5) (Vadavalasa, Rammohan et al. 2021)
- Computing Methodologies => Machine Leaning => Machine Learning Algorithms = > Feature Selection (Chen, R. 2020, July 23) (Hasan, M. A et al. 2015) (Miao, J et al. 2016)
- **SCOPE:** Investigate the difference in performance of regression techniques on fMRI and sMRI modalities presented in a HDLSS dataset
- **ASSUMPTIONS:** Distance weight discrimination will outperform SVM because it doesn't depend on feature selection
- **LIMITATIONS:** The complexity of the dataset used or unknown knowledge gaps can be limiting factors
- **DELIMITATIONS:** There are many different approaches to schizophrenia classification that are out of scope for regression techniques that proved to be effective such as clustering, deep learning among ensemble appraoches

GAPS IN THE LITERATURE REVIEW AND RESEARCH QUESTION

- There is an application gap because of a limited amount of data making it even more difficult when working with an already complex and elusive condition. Deep learning models seem to attain the best performance, however its short lived as it plumets by ~0.25 on an AUC when used to classify new datasets especially from younger cohorts with less extreme or early symptoms. This routes back to the issue of overfitting due to a lack of data, deep learning is prone to overfitting when used on HDLSS data. GANs and CADs could be investigated to see the effectiveness of synthesized data. Required literature exists, most techniques used for HDLSS datasets are in microbiology where researchers work on gene arrays, I think these methodologies could be leveraged for rare mental health disorders, this sparked the investigation to use more traditional approaches such as SVM. (Sadeghi, D et al. 2021) (Cortes-Briones, J. A. et al. 2021) (Oh, J. et al. 2020)

Research Question

- What are the differences in classification accuracy between different implementations of regression techniques when classifying Schizophrenia using HDLSS data through sMRI and fMRI modalities?

HYPOTHESIS

Null Hypothesis

- There is no statistically significant difference between SVM and DWD classification accuracy

Alternate Hypothesis

- There is statistically significant difference between SVM and DWD classification accuracy

FEASIBILITY OF THE STUDY

- Select regression methods (SVM, Partial Least Square Regression, Distance Weighted Discrimination, LASSO Regression, Multivariate Regression)
- Construct baseline for each machine learning algorithm used
- Apply each regression method to available dataset/s
- Tune models appropriately to each HDLSS method document and explore feature selection techniques for tunning
- Record and compare results based on MSE, RMSE, MAE
- Document potential for future work

BIBLIOGRAPHY

- Sadeghi, D., Shoeibi, A., Ghassemi, N., Moridian, P., Khadem, A., Alizadehsani, R., Teshnehlab, M., Gorriz, J. M., & Nahavandi, S. (2021). An Overview on Artificial Intelligence Techniques for Diagnosis of Schizophrenia Based on Magnetic Resonance Imaging Modalities: Methods, Challenges, and Future Works. *Advanced Researches In Biomedical Engineering Lab.* Published. https://arxiv.org/abs/2103.03081

- Castanon, J. (2019, March 19). *10 Machine Learning Methods that Every Data Scientist Should Know*. Towardsdatascience.Com. Retrieved October 28, 2021, from https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

- Wang, H., & Zheng, H. (2013). Model Validation, Machine Learning. *Encyclopedia of Systems Biology*, 1406–1407. https://doi.org/10.1007/978-1-4419-9863-7_233

- Riccio, V. (2020, September 15). *Testing machine learning based systems: a. . .* Empirical Software Engineering. Retrieved October 28, 2021, from https://link.springer.com/article/10.1007/s10664-020-09881-0?error=cookies_not_supported&code=a9b11f32-dc9a-4091-8237-a8c50e2637c3

- Colyer, A. (2019, June 5). *Data validation for machine learning | the morning paper*. Blog.Acolyer.Org. Retrieved October 28, 2021, from https://blog.acolyer.org/2019/06/05/data-validation-for-machine-learning/

- Vadavalasa, Rammohan. (2021). Data Validation Process in Machine Learning Pipeline. https://www.researchgate.net/publication/351022721_Data_Validation_Process_in_Machine_Learning_Pipeline

- Oh, J., Oh, B. L., Lee, K. U., Chae, J. H., & Yun, K. (2020). Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00016

BIBLIOGRAPHY

- Chen, R. (2020, July 23). *Selecting critical features for data classification based on machine learning methods*. Journal of Big Data. Retrieved October 28, 2021, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00327-4

- Hasan, M. A., Hasan, M. K., & Mottalib, M. A. (2015). Linear regression-based feature selection for microarray data classification. *International Journal of Data Mining and Bioinformatics*, *11*(2), 167. https://doi.org/10.1504/ijdmb.2015.066776

- Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

- UYSAL, L., & GÜVENIR, H. A. (1999). An overview of regression techniques for knowledge discovery. *The Knowledge Engineering Review*, *14*(4), 319–340. https://doi.org/10.1017/s026988899900404x

- Lee, K.-Y & Kim, K.-H & Kang, J.-J & Choi, S.-J & Im, Y.-S & Lee, Y.-D & Lim, Y.-S. (2017). Comparison and analysis of linear regression & artificial neural network. International Journal of Applied Engineering Research. 12. 9820-9825. https://www.researchgate.net/publication/328827642_Comparison_and_analysis_of_linear_regression_artificial_neural_network

- Singh Suri, G., Kaur, G., & Moein, S. (2021). Machine Learning in Detecting Schizophrenia: An Overview. *Intelligent Automation & Soft Computing*, *27*(3), 723–735. https://doi.org/10.32604/iasc.2021.015049

- Lin, E., Lin, C. H., & Lane, H. Y. (2021). Prediction of functional outcomes of schizophrenia with genetic biomarkers using a bagging ensemble machine learning method with feature selection. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-89540-6

- Cortes-Briones, J. A., Tapia-Rivas, N. I., D’Souza, D. C., & Estevez, P. A. (2021). Going deep into schizophrenia with artificial intelligence. *Schizophrenia Research*. Published. https://doi.org/10.1016/j.schres.2021.05.018

DATASET

- **Name**: Mind Research Network's Schizophrenia Dataset
- **Source**: https://www.kaggle.com/c/mlsp-2014-mri **&** http://schizconnect.org/
- This dataset consists of functional connectivity values and source based morphometry loadings, the latter is a collection of similar datasets (request for access pending)
- Data format is numeric with column names mapping to each loading in the format of SBM_xx or FNC_xx
- These neuroimaging modalities are extracted from fMRI and sMRI images and are very difficult and time consuming to parse, data like this has to undergo an 8 step cleaning process by an expert before being made available in the format above