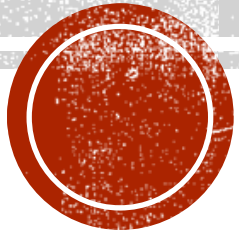
RESEARCH QUESTION, HYPOTHESIS AND PRELIMINARY DESIGN



Maks Drzezdzon | C15311966

TU060/2 | Data Science

DOMAIN AND SCOPE – ACM 2012

- Computing Methodologies => Machine Leaning => Machine Learning Approaches = > Learning Linear Models
- Computing methodologies => Machine Leaning => Machine Learning Approaches
- Computing Methodologies => Modeling and Simulation => Model Development and Analysis => Modeling Methodologies
- Computing Methodologies => Modeling and Simulation => Model Development and Analysis => Modeling Verification and Validation
- Computing Methodologies => Machine Leaning => Machine Learning Algorithms = > Feature Selection
- **SCOPE** is limited to examining classification techniques such as SVM, DWD and their implementations over a period of ~13 weeks applied to FNC/SBM correlation values gathered from HDLSS data from the Mind Research Network's Schizophrenia Dataset consisting of 35,432 observations gathered from 162 patients and 169 healthy controls. The Aim of this study is to derive differences between implementations and their classification accuracy via the F1 score.
- **ASSUMPTION** is that the data from Mind Research Network was properly prepared and filtered when extracted from MRI images as it was overseen by now a distinguished professor in his discipline among other experts.
- **LIMITATIONS** are lack of data and getting access to existing data from schizconnect a collection of HDLSS schizophrenia datasets, no information about the stage, age or severity of schizophrenia of patients which this data was gathered from, this is a limitation because models trained on HDLSS data tend to overfit such as deep learning models acquiring a 0.9 AUC but drop to 0.6 or 0.7 when younger cohorts or vice versa are introduced for testing
- **DELIMITATIONS:** SVM and DWD were chosen over other techniques such as deep learning or regression because DWD and its implementations are state of the art, DWD was designed to address short comings of SVMs and regardless of other limitations DWD has when classifying HDLSS data its still currently state of the art which is why it was picked for examination in this study

GAPS IN THE LITERATURE REVIEW AND RESEARCH QUESTION

- There is an application and methodological gap. This is because the state of the art relies on SVM and DWD. DWD was designed to address SVMs limitations when working with HDLSS data, SVMs shortcomings stem from data-pilling where training data vectors from each class project to a single point for classification which causes it to overfit where as DWD is sensitive to the sample size ratio between classes denoted by the intercept term β, this is a problem because when taking into consideration the differences between cohorts age, stage of schizophrenia, type of schizophrenia among other intricacies that make it hard to diagnose and distinguish between, once accounted for, this can cause/causes a class imbalance. In order to better understand the methodological gap an investigation between both methods and their subsequent implementations that follow will be undergone. Most techniques used for HDLSS datasets are in microbiology where researchers work on gene micro-arrays. There is very little variability in methods used at the top end. They range from different implementations of SVM and DWD, meaning research is still being conducted on how to tackle limitations of both methods when used for classifying HDLSS data and their target class. *(Sadeghi, D et al. 2021; Cortes-Briones, J. A. et al. 2021; Oh, J. et al. 2020; UYSAL, L et al. 1999; Lee, K.-Y et al. 2017; Singh Suri et al. 2021; Sadeghi, D et al. 2021; Castanon, J. 2019, March 19; Lin E et al. 2021; Wang, H et al. 2013; Colyer, A. 2019, June 5; Vadavalasa, Rammohan et al. 2021; Chen, R. 2020, July 23; Hasan, M. A et al. 2015; Miao, J et al. 2016; Marron, J. S et al. 2007; Qiao, X et al. 2015; Liu, Y et al. 2011; Randall, H et al. 2020; Marron, J. et al. 2007; Lui, Y. et al. 2011; Randall, H et al. 2020; Wang, B. et al. 2016; Wang, B. et al. 2017; Ahn, J. et al. 2015; Zahoor, J. et al. 2020)*

Research Question

- What are the differences between implementations of SVM and DWD and their performance when classifying Schizophrenia using HDLSS data through fMRI/FNC features and sMRI/SBM loadings?

HYPOTHESIS + RESEARCH METHODS

Null Hypothesis

- There is no statistically significant difference in F1 score, Log Loss, Categorical Cross entropy or AUC when classifying the class of schizophrenic patients vs healthy controls using fMRI/FNC features (correlation values that summarize connection between brain maps over time) and sMRI/SBM loadings (weights of brain maps derived from gray matter concentration of all subjects) with Support Vector Machine compared to Distance Weighted Discrimination.

Alternate Hypothesis

- If DWD is used to classify the class a patient belongs to using fMRI/FNC features and sMRI/SBM loadings, then on average a lower statistically significant F1 score, Log Loss, Categorical Cross entropy or AUC is expected compared to Support Vector Machine

Testing Hypothesis

- **Type**: Secondary research, using Mind Research Networks dataset supported by a systematic review of existing research on SVM use cases for mental illness classification along with state of the art HDLSS data analysis methodologies such as DWD to create a statistical model to compare performance among other differences between SVM and DWD when examining HDLSS data
- **Objective**: Quantitative research, via the development of classification models evaluated by F1 score, Log Loss, Categorical Cross entropy or AUC on top of investigating the causation of differences in accuracy between a specialized method such as DWD
- **Form**: Empirical research, accept or reject the null hypothesis based on results gathered from model evaluation once the experiment is concluded and evaluate the differences between a method that is more suited for HDLSS data
- **Reasoning**: Deductive approach, comparing SVM with DWD to form a hypothesis that will lead to an experiment from which metrics can be gathered that will either confirm or refute the null hypothesis

GENERAL AND SPECIFIC RESEARCH OBJECTIVES FOR EXPERIMENTAL PURPOSES TOWARDS HYPOTHESIS TESTING USING STATISTICAL TOOLS

- AIM: Derive the differences in classification performance and examine the differences between SVM and DWD implementations

Objectives

- **O1**: Quickly review gathered materials such as "support scripts" appended to the Mind Research Networks dataset and literature on SVM & DWD in order to gather sufficient notes
- **O2**: Box time for each method of SVM and DWD and their subsequent implementations, ~4 weeks each ~8 in total leaving ~4 weeks for documentation and write up in order to attempt as many implementations as possible
- **O3**: Develop a basic prototype of SVM and DWD with default out of the box params to act as benchmark, collect samples of size 15 each for each identified performance metric described in **O6**
- **O4**: Repeat for each implementation of SVM and DWD, rerun experiment 15 times to gather a sample of performance metrics that will be saved to a csv file and used for hypothesis testing and analysis
- **O5**: Prepare data to best suit SVM using feature selection/sampling suited for each method identified in **O1**, document steps taken then using notes from **O1**
- **O6**: Build SVM model/s, tune hyper parameters and record evaluation metrics (F1 score, Log Loss, Categorical Cross entropy and AUC) refer to **O4** - use k fold validation for sampling in order to have models to gather metrics from
- **O7**: After models have been built and metrics have been gathered into samples of x size, collect summary statistics using a box plot or a programmatic way to better understand distributions, using these descriptive statistics assess the distribution of each metric to pick the correct statistical test of significance – save data into .csv format
- The type of statistical test is also dependant on the sampling technique used, for each model trained, in this scenario, using k fold validation will mean that estimated metrics are dependant - this requires a solid understanding of the sampling techniques used in order to pick the correct hypothesis test - when HDLSS data is used to train a model from which metrics are gathered, it further limits the possibility for truly independent samples (Brownlee, J. 2019; Brownlee, J. 2019a)
  - Using a Nonparametric tests such as a paired t-test - Wilcoxon singed-rank test to test the hypothesis is preferred albeit it holds less statistical power
  - Alternatively estimation statistics can be used such as effect size, interval estimation, confidence intervals or meta analysis
- **O8:** There are a few options
  - **O8.1**: Use PCA to lower the amount of features by their highest eigen values for dimension reduction or factor analysis followed by a MANOVA for a global hypothesis test. This is preferred if:
    - Data is of Gaussian distribution – validated or otherwise in **O8** studies have relaxed this and the requirement below this (Muller, K. E et al. 2006)
    - More observations than variables – this seems to be the case here, general way HDLSS data used to be referred to before the term HDLSS was coined
    - For PCA to be successful it requires a "simple covariance structure, at least asymptotically" (Chi, Y. et al. 2013)
    - Factor analysis is sensitive to an unequal ratio of observations to variables which also holds true for PCA
  - There are other specialized alternatives, however more time needs to be allocated to identify a more suitable method/s
  - **O8.2:** More understanding of the data and its innerworkings such as mapping values to the brain and dividing it by regions (Chi, Y. et al. 2013), this information/approach is not available to me
- **O9:** Reject the null hypothesis (Ho) if $p < 0.05$ from the above hypothesis test MANOVA with the use of PCA/factor analysis
  - Potentially another dataset may be used if access to schizconnect is granted to find a dataset with better documented features, more info on brain regions or any other information that could allow for better hypothesis testing

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
- Marron, J. S., Todd, M. J., & Ahn, J. (2007). Distance-Weighted Discrimination. *Journal of the American Statistical Association*, *102*(480), 1267–1271. https://doi.org/10.1198/016214507000001120
- Qiao, X., & Zhang, L. (2015). Flexible high-dimensional classification machines and their asymptotic properties. *The Journal of Machine Learning Research*, *16*(1), 1547-1572.
- *Liu, Y., Zhang, H. H., & Wu, Y. (2011). Hard or Soft Classification? Large-Margin Unified Machines. Journal of the American Statistical Association, 106(493), 166–177. https://doi.org/10.1198/jasa.2011.tm10319*
- Randall, H., Artemiou, A., & Qiao, X. (2020). Sufficient dimension reduction based on distance-weighted discrimination. *Scandinavian Journal of Statistics*, *48*(4), 1186–1211. https://doi.org/10.1111/sjos.12484
- Brownlee, J. (2019, August 8). *Statistical Significance Tests for Comparing Machine Learning Algorithms*. Machine Learning Mastery. https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/?fbclid=IwAR331RX6HbXBnArXqhRSheTDiRWCnme5jZa5hEfPhYcebkR56HfRvlmsJEw
- Brownlee, J. (2019, August 8). *A Gentle Introduction to Estimation Statistics for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/estimation-statistics-for-machine-learning/?fbclid=IwAR0mdkFN_hFzvAtIinRey2LueMdUR8oAeQ3hX3pqz_UHHJBuc9-iwKEn_ug
- Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, *99*(3), 386–402. https://doi.org/10.1016/j.jmva.2006.11.002
- Srivastava, M. S. (2007). Multivariate Theory for Analyzing High Dimensional Data. *JOURNAL OF THE JAPAN STATISTICAL SOCIETY*, *37*(1), 53–86. https://doi.org/10.14490/jjss.37.53
- Muller, K. E., & Stewart, P. W. (2006). *Linear model theory: univariate, multivariate, and mixed models*. John Wiley & Sons.

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
- J. S. Marron, Michael J. Todd and Jeongyoun Ahn. (2004) Distance Weighted Discrimination. Journal of the American Statistical Association, (no page number). http://www.optimization-online.org/DB_FILE/2002/07/513.pdf?fbclid=IwAR19LvTVhXEcSXK0hyO1JwoXaZN0_OS0GIwiAIFq2c3z5XMROUGud--QTPo
- Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J., & Marron, J. S. (2010). Weighted Distance Weighted Discrimination and Its Asymptotic Properties. *Journal of the American Statistical Association*, *105*(489), 401–414. https://doi.org/10.1198/jasa.2010.tm08487
- Wang, B., & Zou, H. (2016). Sparse distance weighted discrimination. *Journal of Computational and Graphical Statistics*, *25*(3), 826-838. https://openreview.net/pdf?id=oVgon01wpfrlgPMRsB1E
- Wang, B., & Zou, H. (2017). Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(1), 177–198. https://doi.org/10.1111/rssb.12244
- Zahoor, J., & Zafar, K. (2020). Classification of Microarray Gene Expression Data Using an Infiltration Tactics Optimization (ITO) Algorithm. *Genes*, *11*(7), 819. https://doi.org/10.3390/genes11070819
- Ahn, J., & Jeon, Y. (2015). Sparse HDLSS discrimination with constrained data piling. *Computational Statistics & Data Analysis*, *90*, 74–83. https://doi.org/10.1016/j.csda.2015.04.006
- Chi, Y. Y., & Muller, K. E. (2013). Two-Step Hypothesis Testing When the Number of Variables Exceeds the Sample Size. *Communications in Statistics - Simulation and Computation*, *42*(5), 1113–1125. https://doi.org/10.1080/03610918.2012.659819

DATA NEEDED FOR EXPERIMENTAL PURPOSES

Mind Research Network's Schizophrenia Dataset

- Total observations in dataset: 35,432
- Trying to predict: The class/id pair indicated by a binary column called class with a range of 1 or 0 representing patients with and without schizophrenia
- Other: No differentiation between stages of schizophrenia, types of schizophrenia, age of patients and extent of their illness

FNC Features

- Shape: FNC has 86 (rows) x 379 (columns)
- Column Names: FNC column names range from **FNC1** to **378** and one **Id** column
- What is this data: It's a set of correlation values that describe the overall connection between pairs of brain maps over time

Column: Id

- Data type: Ordinal
- Other: Unique Identifier
- Range: 120,873 to 993,946

Columns: FNC1-378

- Data type: The Id column is sequential/ordinal (cate, ordinal, binary, qualitative)
- Range: -0.9871 to 0.9858

SBM Loadings

- Shape: SBM 86 (rows) x 14 (columns)
- Column Names: SBM column names range from **SBM_MAP1** to **75** and one **Id** column
- What is this data: It's a set of standardized weights of brain maps that describe the expression level of ICA brain maps derived from gray-matter concentration

Column: Id

- Data type: Ordinal
- Range: 120,873 to 993,946

Columns: SBM_MAP1-75

- Data type: The Id column is sequential/ordinal (cate, ordinal, binary, qualitative)
- Range: -8.10 to 13.07