Multiomics integration in PANS

Acciaro Gennaro Daniele, Morisco Michele

May 2022

Abstract

Integrate metabolomics and proteomics data from PANS. Perform classification of patients based on the two types of data, and extract biomarkers that maximize classification performance.

Github Repo: https://github.com/gdacciaro/CHL_PANS_Project

1 Introduction

PANS is a pediatric disorder that involves symptoms such as obsessive-compulsive disorder (OCD), tics, anxiety attacks, depression, and sleep problems. The causes of the onset of this disease are probably related to an immune response to a bacterial or viral infection; therefore, current treatments include antibiotics, anti-inflammatories, and antidepressants.

Our project is based on two data sets: metabolomic data and proteomic data. Most of the patients they refer to are in common. Then, as we described in the Method section, we merged these two datasets into a new one to try the a-priori methodology. Hence, our aim is to integrate omics data to maximize the classification performance extracting the biomarkers.

This report is organized as following: In section **2** we introduce datasets and our analysis method. In section **3** we introduce our preprocessing tools. In section **4** we describe our model selection. In section **5** we show the results. In section **6** we conclude our work.

2 Data and analysis

We have used two datasets for our analysis: metabolomics, proteomics. Specifically, the metabolomics dataset contains 931 features on 61 patients, while the proteomics dataset contains 1317 features on 52 patients. Firstly, we have checked the presence of possible missing values on datasets, but fortunately, our datasets don't have any missing values, although the proteomics dataset has fewer patients with respect to the metabolomics set. Each patient is associated with a disease state that can be Flare, Remission or Healthy, and within the same dataset, it is possible to find some patients multiple times, with different stages of the disease. Our analysis involved the use of two datasets for integration. To do this, we used both a-posteriori and a-priori methodology. For both cases we performed Feature Selection, comparing different scalers, then performed Model Selection and finally obtained the results by applying Ensemble methods.

3 Methods

3.1 A-Priori and A-Posteriori methodologies

In our experiment, we used two data integration methodologies: a-priori and a-posteriori data integration. The first one provides to merge the two datasets before performing the classification task. Indeed, with this method, we will obtain a unique merged dataset. On the other hand, the a-posteriori data integration provides to perform separate classification on two datasets and to do an ensemble technique to merge the separate models obtained by model selection.

3.2 Nested Cross-Validation

In order to avoid bias in the choice of development and test set, we divided each of the three datasets used for this analysis (metabolomics, proteomics, and their union for the a-priori methodology) into three parts: alternately, two of these parts are used as the development set and the remaining part is used as the test set. We then performed our feature and model selection pathway for each of these parts and then merged them at the end of the process using an Ensemble method, the Voting Classifier. The effect obtained therefore at the Model Selection phase, as we shall see will be that of a Nested Cross-Validation.

3.3 Feature Selection

In this part, we have chosen three different methods for feature selection: Isomap, PCA, and RFE. RFE is effective at selecting the features in a training dataset that is more relevant in predicting the target variable, then we have chosen this algorithm for our task. In addition, we have wanted to compare other methods with RFE so we have chosen the other two feature selection algorithms mentioned early.

Isomap Isomap stands for Isometric Mapping. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbors on the manifold.

Principal Component Analysis Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

Recursive Feature Elimination Recursive Feature Elimination (RFE) is a feature selection algorithm. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

This is achieved by fitting the estimator used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains. Using in a similar context of [**coviddet**], for this method, we have created a small grid-search to select the best estimator to find a subset of features.

3.4 Normalization

To converge faster and achieve better performance, we normalized the data. Therefore, after choosing the feature selection method, we compared 3 different scalers: MinMaxScaler, StandardScaler and Yeo-Johnson Transformer. In addition, we compared these scalers with unscaled data.

MinMaxScaler Given a feature, the MinMaxScaler subtracts the minimum value of the feature from every possible value. After that, it divides each feature value by an interval which is the difference between the original maximum and minimum. The peculiarity of this scaler is that it preserves the shape of the original distribution. Mathematically we can see the MinMaxScaler as follows: let X be the feature:

$$X_{new} = \frac{(X - \min X)}{\max X - \min X}$$
$$X_{scaled} = X_{new} * (\min X - \min X) + \min X$$

StandardScaler This technique standardizes a feature by subtracting the mean and dividing all values by the standard deviation. Mathematically:

$$X_{scaled} = \frac{(X - u)}{s}$$

where u is the mean and s is the standard deviation.

Yeo-Johnson Transformer This technique was introducted by I.K. Yeo and R.A. Johnson in [2]. It is parametric, monotonic transformations that make data more Gaussian-like. In particular, the Yeo-Johnson method can be used for both negative and positive feature values.

4 Model Selection

After defining all the Feature Selection methods and all the ways to normalize the data, we were able to apply Model Selection, comparing the performance of seven classifiers, about these classifiers we relied on [1] that tested different models using the microarray gene expression data:

Naive Bayes Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Bayes theorem provides a way of calculating the posterior probability $p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$ This assembly is like a directed acyclic graph where each node resemblance to state variables and edges corresponds to dependencies between variables.

Decision Tree A classification tree is a technique that describes the data in the form of n-ary order through each node and branch have a certain convoying consequence, likelihood and weights.Root of the tree is selected usually by calculating the entropy or by its contrary, the information gain.

K-Nearest Neighbor The K-NN works on the concept that examples in nearby spaces are likely to fit in the alike class. A K-NN assigns samples to the class that is most persistent among K neighboring. K is a control for fine-tuning the classification algorithm.

Logistic Regression Logistic regression works on real-valued input and is a discriminative classification technique. The dimension of the input vector is known as features or predictors. In logistic regression, probability P of a dichotomous result can be considered as expanding from Bernoulli trial and can be related with investigative event.

Multi-layer perceptron In MLP, neurons are organized into layers like a multistage graph. Each node at each layer obtains an input from the associated node of the previous layer, and it computes value of a function and provides input to the connected node in the next layer.

Random Forest A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Support Vector Machine SVM uses the concept of structural risk minimization to avoid the over-fitting problems of machine learning. The classifiers inspect for the optimum separating hyperplane which is in between of two classes. This hyper plane has numerous statistical characteristics.

4.0.1 Grid-search

In order to obtain the best results, we performed an exhaustive grid search on all possible hyperparameters of these models. In total, we tested a *113,292* different models, as we tried every possible combination of feature elimination methods, scaler and fold (for Nested Cross-Validation).

Distributed Grid-search In order to perform all the calculations at feasible times, we had to speed up the whole process by distributing the calculation through a simple MySQL database shared among all the workers. In particular, we first loaded an encoding of each trial, a JSON file with the name of the model, and the hyper-parameters to be used on the database. At each iteration, a thread retrieves a trial flagged as 'Available', sets it as 'Reserved', and starts the computation. After that, it saves the accuracy according to validation data on the database and retrieves another trial. It is worth noting that only when the results of all the models were computed, we took the best models (based on the *validation set*) and tried all of them again using the *test set*.

4.0.2 Ensemble methods

We performed several ensembles to combine the results obtained from the different folds of the Nested Cross-Validation. In particular:

- After obtaining the Model Selection results, we selected the folds that had the best accuracy calculated on the validation set, for the metabolomics, proteomics and a-priori datasets.

  Then for each dataset, we used a VotingClassifier to merge the three folds and get one combined model. In this way, we obtained the results on the test set of A-priori and the intermediate results of metabolomics and proteomics.

- Once we had the intermediate results of metabolomics and proteomics, we did another ensemble, this time, in manual mode to get the final results of the a-posteriori methodology.

5 Results

Feature Selection Results Referring to Figure 1 in the appendix, these are the results of the RFE used by the models that won the various Grid Searches. The numbers 1,2 and 3 refer to the 3 folds of the Nested Cross-Validation and the numbers within these Venn diagrams are the number of features selected.

Features Extracted Table 5 and Table 6 summarize the feature selected by our models, describing them. Each fold of the dataset selected different features, then we considered only those that appeared in all folds, indeed the Venn diagrams defined previously we helped to find the features more present. Between a-priori and a-posteriori methodology, the features shared are a few. Moreover, the feature extraction phase has always selected more metabolomics than proteomics.

Classification Results In the following tables, there are the Accuracy and the F1-Score calculated for each Feature Selection method and for each Scaler on the test set. First of all, we calculated the results for metabolomics and proteomics data independently (tables 1 and 2), specifically the bold results in these two tables are the models that, on average on folds, performed best on the validation test. Then we ensemble them getting the results described in Table3, these are the results for the A-posteriori methodology. At the end, in Table 4 there are the result for the A-priori methodology.

Feature Selection	**Scaler**	**F1-Score**	**Accuracy**
Isomap	Min-max-scaling	0.39933	0.481481
Isomap	Standard-scaling	0.18832	0.333333
Isomap	Yeo-johnson-transformation	0.345595	0.462963
Isomap	Unscaled	0.307324	0.388889
Pca	Min-max-scaling	0.260521	0.388889
Pca	Standard-scaling	0.269296	0.407407
Pca	Yeo-johnson-transformation	0.33488	0.407407
Pca	Unscaled	0.323368	0.388889
Rfe	Min-max-scaling	0.407346	0.526316
Rfe	Standard-scaling	0.372238	0.45614
Rfe	**Yeo-johnson-transformation**	**0.584775**	**0.666667**
Rfe	Unscaled	0.359813	0.491228

Table 1: Metabolomics results

Feature Selection	**Scaler**	**F1-Score**	**Accuracy**
Isomap	Min-max-scaling	0.270899	0.3125
Isomap	Standard-scaling	0.287497	0.3125
Isomap	Yeo-johnson-transformation	0.263492	0.416667
Isomap	Unscaled	0.318965	0.395833
Pca	Min-max-scaling	0.315331	0.395833
Pca	Standard-scaling	0.19727	0.270833
Pca	Yeo-johnson-transformation	0.212367	0.229167
Pca	Unscaled	0.293557	0.354167
Rfe	Min-max-scaling	0.325712	0.395833
Rfe	Standard-scaling	0.476946	0.541667
Rfe	**Yeo-johnson-transformation**	**0.462027**	**0.479167**
Rfe	Unscaled	0.365333	0.395833

Table 2: Proteomics results

Feature Selection method	**Scaler**	**F1-Score**	**Accuracy**
RFE	Yeo-Johnson	0.465624	0.537036

Table 3: A-Posteriori results

Feature Selection method	Scaler	F1-Score	Accuracy
Isomap	Min-max-scaling	0.248859	0.354167
Isomap	Standard-scaling	0.244337	0.291667
Isomap	Yeo-johnson-transformation	0.262533	0.291667
Isomap	Unscaled	0.349415	0.4375
Pca	Min-max-scaling	0.293633	0.354167
Pca	Standard-scaling	0.34111	0.395833
Pca	Yeo-johnson-transformation	0.286257	0.375
Pca	Unscaled	0.20899	0.270833
Rfe	Min-max-scaling	0.318893	0.354167
Rfe	Standard-scaling	0.492415	0.5
Rfe	**Yeo-johnson-transformation**	**0.587381**	**0.645833**
Rfe	Unscaled	0.287756	0.333333

Table 4: A-Priori results

6 Conclusions

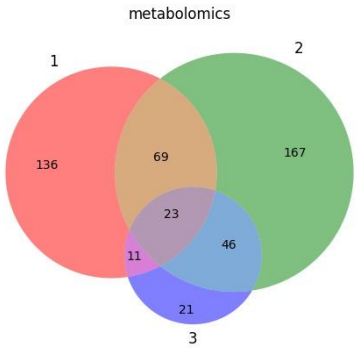
PANS is a pediatric disorder that involves symptoms such as obsessive-compulsive disorder, tics, anxiety attacks, depression, and sleep problems. In this paper, therefore, we have seen methodologies of Feature Selection first and Model Selection later to perform disease classification using metabolomics and proteomics data. Despite the many methodologies and models we have explored our results still do not allow us to generalize for new patients. Possible reasons for this is could be that the patients, first of all, are very few and also each patient has had different treatments of the treatments over the course of the disease, compared to all others. We expect, as future developments, the collection of new data and the use of more advanced frameworks such as PyTorch or Tensorflow to be able to reduce computation time even more. In addition, we could develop a genetic algorithm to compare with RFE and to check if that could outperform it, especially for this kind of task.

References

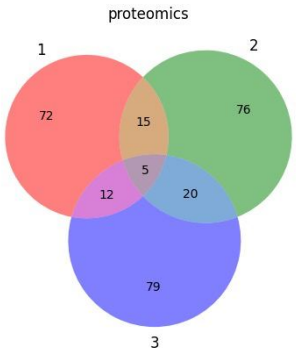
[1] Ashok Kumar Dwivedi. "Artificial neural network model for effective cancer classification using microarray gene expression data". In: (2016). URL: https://link.springer.com/article/10.1007/s00521-016-2701-1.

[2] In-Kwon Yeo and Richard A. Johnson. "A New Family of Power Transformations to Improve Normality or Symmetry". In: *Biometrika* 87.4 (2000), pp. 954–959. ISSN: 00063444. URL: http://www.jstor.org/stable/2673623.

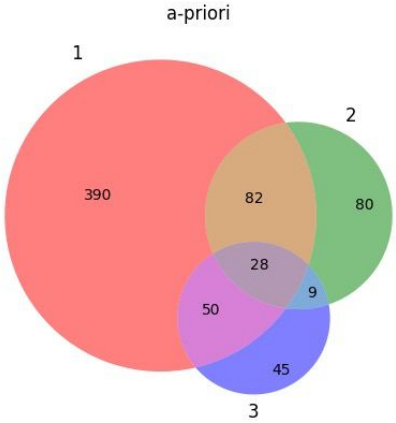
A Appendix



(a) Feature Elimination results for metabolomics



(b) Feature Elimination results for proteomics



(c) Feature Elimination results for a-priori

Figure 1: Feature Elimination results

Dataset	Feature Name	Feature Description
Metabolomics	1-stearoyl-2-arachidonoyl-GPS	Lipid, Phosphatidylserine
Metabolomics	1-stearoyl-2-docosahexaenoyl-GPC	Lipid, Phosphatidylcholine
Metabolomics	3-methylxanthine	Xenobiotics, Xanthine Metabolism
Metabolomics	5-hydroxy-2-methylpyridine sulfate	Xenobiotics, Chemical
Metabolomics	5-alpha-pregnan-3beta, 20-beta-diol monosulfate	Lipid, Progestin Steroids
Metabolomics	arachidate	Lipid, Long Chain Saturated Fatty Acid
Metabolomics	bilirubin	Cofactors, Vitamins, Hemoglobin and Porphyrin Metabolism
Metabolomics	cis-urocanate	Amino Acid, Histidine Metabolism
Metabolomics	cysteine	Amino Acid, Methionine, Cysteine, SAM and Taurine Metabolism
Metabolomics	cysteine-glutathione disulfide	Amino Acid, Glutathione Metabolism
Metabolomics	dihomolinoleate	Lipid, Long Chain Polyunsaturated Fatty Acid
Metabolomics	docosapentaenoate	Lipid, Long Chain Polyunsaturated Fatty Acid
Metabolomics	glucose	Carbohydrate, Glycolysis, Gluconeogenesis, and Pyruvate Metabolism
Metabolomics	glycerophosphoinositol	Lipid, Phospholipid Metabolism
Metabolomics	gulonate	Cofactors and Vitamins, Ascorbate and Aldarate Metabolism
Metabolomics	linoleoylcholine	Lipid,Fatty Acid Metabolism (Acyl Choline)
Metabolomics	maltose	Carbohydrate, Glycogen Metabolism
Metabolomics	methylmalonate (MMA)	Lipid ,Fatty Acid Metabolism
Metabolomics	N-stearoyl-sphingosine	Lipid, Ceramides
Metabolomics	perfluorooctanesulfonate	Xenobiotics, Chemical
Metabolomics	propionylcarnitine	Lipid, Fatty Acid Metabolism
Metabolomics	pyrraline	Xenobiotics, Food Component/Plant
Metabolomics	hydroxyproline	Amino Acid, Urea cycle Arginine and Proline Metabolism
Proteomics	60 kDa heat shock protein, mitochondrial	Chaperonin implicated in mitochondrial protein import and macromolecular assembly
Proteomics	C-type mannose receptor 2	May play a role as endocytotic lectin receptor displaying calcium-dependent lectin activity
Proteomics	Kremen protein 2	Receptor for Dickkopf proteins. Plays a role in limb development
Proteomics	Bone morphogenetic protein receptor type-1A	On ligand binding, forms a receptor complex consisting of two type II and two type I transmembrane serine/threonine kinases
Proteomics	Insulin	Insulin decreases blood glucose concentration.
Proteomics	15-hydroxyprostaglandin dehydrogenase [NAD(+)]	This enzyme is induced by androgens in hormone-sensitive human prostate cancer cells

Table 5: A-priori features selected.

Dataset	Feature Name	Feature Description
Metabolomics	13-HODE + 9-HODE	Lipid, Fatty Acid, Monohydroxy
Metabolomics	2-hydroxyhippurate	Xenobiotics, Benzoate Metabolism
Metabolomics	3-methylxanthine	Xenobiotics, Xanthine Metabolism
Metabolomics	5-alpha-androstan-3beta, 17-beta-diol disulfate	Lipid, Androgenic Steroids
Metabolomics	5-alpha-pregnan-3beta, 20-beta-diol monosulfate	Lipid, Progestin Steroids
Metabolomics	alpha-hydroxycaproate	Lipid, Fatty Acid, Monohydroxy
Metabolomics	arachidate	Lipid, Long Chain Saturated Fatty Acid
Metabolomics	carboxyethyl-GABA	Amino Acid, Glutamate Metabolism
Metabolomics	cerotoylcarnitine (C26)	Lipid, Fatty Acid Metabolism (Acyl Carnitine, Long Chain Saturated)
Metabolomics	phosphocholine	Lipid, Phospholipid Metabolism
Metabolomics	deoxycarnitine	Lipid, Carnitine Metabolism
Metabolomics	docosahexaenoylcholine	Lipid,Fatty Acid Metabolism (Acyl Choline)
Metabolomics	fructosyllysine	Amino Acid, Lysine Metabolism
Metabolomics	gamma-glutamyltryptophan	Peptide, Gamma-glutamyl Amino Acid
Metabolomics	glycerophosphoinositol	Lipid, Phospholipid Metabolism
Metabolomics	gulonate	Cofactors and Vitamins, Ascorbate and Aldarate Metabolism
Metabolomics	linoleoylcholine	Lipid, Fatty Acid Metabolism (Acyl Choline)
Metabolomics	maltose	Carbohydrate, Glycogen Metabolism
Metabolomics	N-acetylalliin	Xenobiotics, Food Component/Plant
Metabolomics	N-methylproline	Amino Acid, Urea cycle Arginine and Proline Metabolism
Metabolomics	propionylcarnitine	Lipid, Fatty Acid Metabolism
Metabolomics	theophylline	Xenobiotics, Xanthine Metabolism
Metabolomics	hydroxyproline	Amino Acid, Urea cycle Arginine and Proline Metabolism
Proteomics	Histone H2A type 3	Histones thereby play a central role in transcription regulation, DNA repair, DNA replication and chromosomal stability
Proteomics	C-type mannose receptor 2	May play a role as endocytotic lectin receptor displaying calcium-dependent lectin activity
Proteomics	Tyrosine-protein phosphatase non-receptor type 6	Modulates signaling by tyrosine phosphorylated cell surface receptors such as KIT and the EGF receptor. Plays a key role in hematopoiesis.
Proteomics	3-hydroxy-3-methylglutaryl -coenzyme A reductase	Catalyzes the conversion of (3S)-hydroxy-3-methylglutaryl-CoA to mevalonic acid and plays a critical role in cellular cholesterol homeostasis
Proteomics	Trefoil factor 1	Stabilizer of the mucous gel overlying the gastrointestinal mucosa that provides a physical barrier against various noxious agents.

Table 6: A-posteriori features selected.