Classification and Clustering on Schizophrenic Patients data

Axelle Piguet, Adham Elwakil and Melissa Faggella
Mentored by Simona Garobbio and Aline Cretenoud from the EPFL's Laboratory of Psychophysics

***Abstract*—Current diagnosis and treatment methods for some psychiatric conditions like schizophrenia have shown their limits. Indeed, those procedures are still symptom-based as the biological mechanisms that underlie the disorder are not well understood yet. Machine learning is a field that could help scientists find what causes schizophrenia. In the present article, we try classification algorithms on schizophrenic patients, their first-degree relatives and healthy controls to learn more about the diagnosis aspect. We also apply clustering methods on schizophrenic patients data to try to transcend the current schizophrenia categories with new suspected subtypes of the disease.**

I. Introduction

When a patient presents fever, the possible pathologies underlying that symptom are countless. Fortunately, with modern medical knowledge and technologies like imaging devices and laboratory analysis machines, physicians are usually able to precisely diagnose the patient and thus to prescribe him an adapted treatment. However, there is still one medical field in which diagnosis is mainly based on symptoms and it's psychiatry. As there are several symptoms overlaps between conditions, notably between schizophrenia and bipolar disorder, many patients are misdiagnosed and given inadequate medication [1]. This is due to the fact that so far, we haven't found what biological mechanisms dysfunction in those diseases and thus, we can only treat the symptoms and not the true cause.

Approaches that are currently used to try transcend the 100 years old outdated classification criterias for schizophrenia [2] are statistics and machine learning. In particular, unsupervised learning is used on genetic, neurophysiological and cognitive features in order to unravel new biologically meaningful categories [3], the so-called biotypes or endophenotypes. Indeed, it may well be that the current schizophrenia label that psychiatrists attribute to some patients and not to some others is completely invalid as it still hasn't been backed up by any neuroscientific study. Maybe, the biological reality will lead us to rethink that disease diagnosis either by showing the existence of distinct subtypes of schizophrenia, either by showing that schizophrenia is a spectrum-like disorder rather than an entity with clear boundaries [4].

In our search for biotypes, we had access to two data sets: one "variable" data set that contains measures from cognitive and behavioral tasks and one "demographics" data set that contains basic information about the patient's health and demographic group. Those two data sets contains 524 examples either labelled as SCH : schizophrenic patients (227 persons), as REL : schizophrenic patient's relatives (119 persons) or as CON : controls who are healthy people (178 persons). The relatives group has been included to observe the inheritable genetic and epigenetic characteristics of schizophrenia.

Our main data set, the variable data set contains 8 features derived from 3 cognitive and behavioral tasks: the visual backward masking (VBM), the continuous performance task (CPT) and the Wisconsin card sorting task (WCST). All those three tasks have been chosen for their previously known ability to discriminate healthy controls from schizphrenic patients [5], [6], in particular the VBM task that is also able to discriminate relatives from controls [7], [8].

Statistical and computational approaches are relevant in that context because they might help us to better comprehend what neurophysiological mechanisms dysfunction in schizophrenia and thus, could help us choosing more precise and efficient pharmacological targets to treat patients.

II. Methods and Results

A. Exploratory Data Analysis

The data set that is the most relevant in our research for biotypes is the variable data set because we do not want our model to learn from symptoms-related data but rather from cognitive and neurophysiological data. While exploring the data set's properties, we noticed that it consists in 8 features arising from 3 distinct cognitive tasks. The features coming from the same task were usually correlated with one another (SOA25 with SOA5 and WCSTCAT with WCSTPERS, WCSTKORR and WCSTERR). Even though two features (WCSTCAT and WCSTPERS) were discrete we assumed that they are continuous as they have respectively 7 and 17 distinct values. Using pairplots (seaborn [9]), we observed that the distribution of most features are noticeably different between schizophrenic patients and controls. On the other hand, the differences between relatives and controls were almost not perceptible. We thus hypothesised that it would be relatively hard to correctly classify the relatives in supervised learning. We also explored the relative features importance after classification. The three approaches (impurity-based feature importance, permutation feature importance and lasso regularization) on three conditions (SCH vs REL vs

CON, SCH vs REL + CON, SCH vs CON), made us conclude that the most important feature is SOA25.

Concerning unsupervised learning, as we are interested in using clustering algorithms on the schizophrenic patients samples only, we explored the density of distribution within the schizophrenic samples. We did that by adding a kernel density estimate on top of a pairplot and noticed the absence of any interesting patterns or emerging clusters. From this we hypothesised that density based clustering algorithms are likely to perform poorly on our data set.

B. Supervised Learning : classification

One of our two main goals was to try to classify the schizophrenic patients, patient's relatives, and controls using the variable data set to see if the different groups can be distinguished according to the cognitive tasks. This may be useful to improve the diagnosis criterias, currently used by psychiatrists. Thus, it could be interesting to use statistical approaches to help us implementing better diagnosis techniques.

The baseline classification technique applied was the multi-class logistic regression model [10]. The model was trained, optimized, and tested as follow:

1) Create a pipeline to merge the two steps of Feature Transformation and the application of the Classification model [10].
2) Create a parameter grid having a range of values for hyper-parameters (usually C (regularization parameter) and polynomial degree)
3) Use GridSearchCV [10] to fit the model, find optimum hyperparameters, and calculate the cross-validated accuracy (k =10 folds)

Those steps are applied to all the classification models we used in this project. For the remainder of the report, these 3 steps will be referred to as the "system". The choice of k for cross-validation is important and must be chosen carefully depending on the data set. Nevertheless, it was shown through experiments that k=10 generally results in a model skill estimate with low bias and a modest variance [11]. The accuracy of the linear logistic regression model was equal to 60.7% (Note that we did not split the data so we do not have a validation set). With the aim of trying to understand why the model had a low accuracy, a decision boundary plot was visualised using a scatter plot and principal component analysis (PCA) to reduce the dimensionality of the data from 8 to 2. It is shown in Figure 1 that the data is not linearly separable, at least in this configuration.

Since the data is linearly inseparable, a linear decision boundary will not be able to efficiently classify the data. Hence, to improve accuracy, we decided to increase the model complexity by augmenting the features using PolynomialFeatures [10] to the pipeline. Again, we used the system, where GridSearchCV was used to optimize the

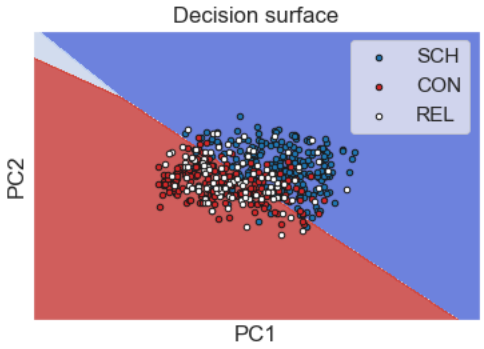


Figure 1: Decision boundary for Linear Logistic Regression

hyper-parameters (C and polynomial degree) and the cross-validated accuracy was calculated. The accuracy finally only improved by 2 %.

We hypothesised that applying a support vector classifier (SVC) could improve accuracy by minimising margin violations. We started off using Linear SVC and followed the same steps stated in the system. The linear SVC accuracy was still rather low with 61.1%. Again, since the data is not linearly separable, non-linear decision boundaries could increase the accuracy, therefore, data augmentation using PolynomialFeatures was done. Moreover, we also tried using an Radial basis function (RBF) kernel for a non-linear decision boundary. The RBF kernel was chosen as it is the most generalized form of kernelization. In both cases (data augmentation and kernelization), accuracy only improved by approximately 1% compared to linear logistic regression.

Random forest was primarily used with the aim of calculating feature importance, but we also tried to use it on the data to see the performance of an ensemble method. The model was less accurate than linear logistic regression, with an accuracy off 58%.

The performance of the classifiers seemed to be low, so to further analyze the problem, confusion matrix were calculated for each model using the confusion-matrix implementation in scikit metrics [10]. It was shown that almost all relatives were systematically misclassified. Moreover, Figure 1 shows that the relatives (white points) appear to overlap both schizophrenic patients and controls groups and have no clear boundaries. As a result, we decided to modify the data set twice, once by combining relatives and controls together (SCHvsREL&CON condition), and the other by removing the relatives group from the data set (SCHvsCON).

The six previously mentioned models were then applied to the two new data sets and the same steps stated in the system were repeated. The results of all the models are summarized in Table I. Combining or removing the relatives from the data set increased the accuracy to 78.7% and 81.6% respectively.

Model	SCHvsRELvsCON	SCHvsREL&CON	SCHvsCON
Logistic Regression	0.607 +/- 0.106	0.784 +/- 0.104	0.798 +/- 0.118
Poly LR	0.628 +/- 0.088	0.769 +/- 0.108	0.812 +/- 0.126
Linear SVC	0.611 +/- 0.098	0.784 +/- 0.104	0.795 +/- 0.122
Poly SVC	0.621 +/- 0.09	0.787 +/- 0.104	0.802 +/- 0.126
SVC with RBF	0.62 +/- 0.09	0.782 +/- 0.104	0.816 +/- 0.124
Random forest	0.584 +/- 0.116	0.761 +/- 0.106	0.785 +/- 0.126

Table I: Summary of the performances of our models

C. *Unsupervised Learning : clustering*

In order for the misdiagnosed and the many patients with an atypical clinical picture to be more efficiently detected and more precisely treated, it is crucial to better understand the true nature of schizophrenia. It is still uncertain if some subtypes of the disorder may exist or if the conditions is a continuous spectrum without definite borders between pathological forms and non-pathological ones. It could be useful to search for clusters within the schizophrenic population, that could represent different variations of schizophrenia. Moreover, clustering on neurophysiological and cognitive data might be a promising approach to discover more meaningful subentities than the current limitating psychiatric diagnosis.

We began by scaling the schizophrenic population variables data using sklearn's standardScaler [10]. We know that some of the variables are correlated; we thus applied a Principal Component analysis (PCA) in order to decorrelate the data and reduce the dimension. We tried two different methods (percentage of explained information and Kaiser rule) to choose the optimal number of components. The percentage of explained information was high for the first three components. In total, they were losing only 23 % of information. Additional components carried less than 10 % of total information, which made us neglect them. In the end, we chose to use 3 components as the Kaiser rule also advocated (see Figure 2). To add interpretability to our components, we also visualised the PCA loadings (see Figure 3).

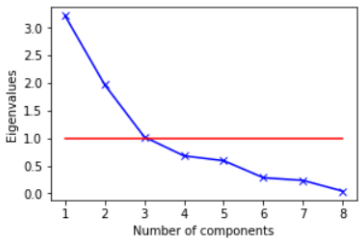


Figure 2: The Kaiser rule suggests that the optimal number of components for the PCA is 3

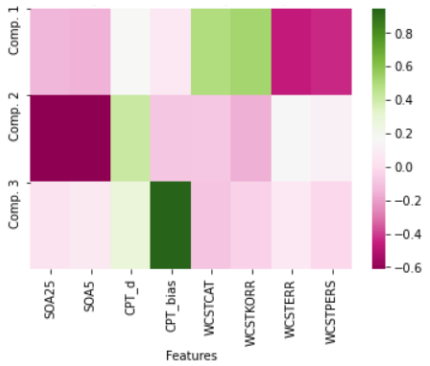


Figure 3: PCA loadings on schizophrenic patients

As a baseline clustering algorithm, we used the centroid-based K-means algorithm from the sklearn library [10]. We applied the Elbow method as well as the Silhouette method (sklearn metrics [10]) in order to determine which number of cluster would be optimal for our dataset. The interpretation of the results of those two methods were not completely conclusive and hard to interpret. However, the silhouette method seemed to indicate that 2 or 7 clusters could be the best (highest score) as seen in Table II:

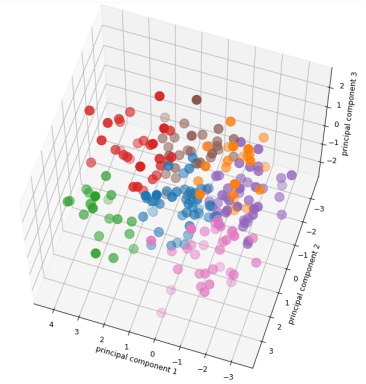
Number of clusters	2	3	4	5	6	7	8
Silhouette Score	0.33	0.27	0.26	0.25	0.26	0.28	0.27

Table II: K-means Silhouette scores

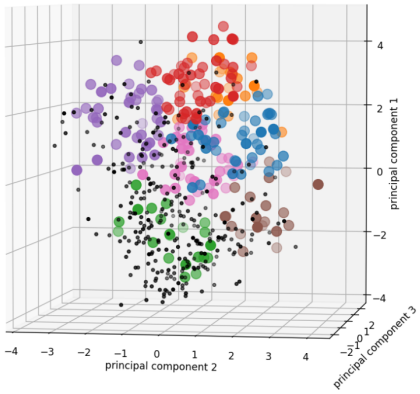
K-means was run on the schizophrenic population variable data. The seven clusters obtained are visualised (alone and within the whole population) in Figure 4.

The resulting clusters are not overlapping much and more importantly, the K-means algorithm is consistent in the assignment of the examples. Indeed, the assignment sequence of each point showed that most of them always get assigned to the same group.

In addition to K-means, we also tried Distribution-based Clustering algorithm Gaussian Mixture Model (sklearn.mixture library [10]) with and without Dirichlet process. Mixed membership to different clusters as allowed by this type of method can be interesting for the spectrum-like distribution of schizophrenia. However, silhouette method was unstable for this algorithm, which may indicate its inappropriateness. This was confirmed by trying Gaussian Mixture with several different number of clusters. The cluster assignment was not consistent for this type of method, thus not very relevant for our task. Finally, we also tried to use the density-based algorithm DBSCAN from sklearn [10] but the results were not conclusive. Indeed, as stated before, we hypothetized that density-based algorithms would not perform well on our data set.



a) 7 clusters in schizophrenic population



b) Schizophrenic population 7 clusters (colors) with controls and relatives (black)

Figure 4: Visualisation of schizophrenic patients clusters

D. Technical aspects

For our experiments, we used the following programming tools, libraries and their corresponding versions : for coding Jupyter Notebook (6.0.3), numpy (1.18.5), pandas (1.0.5), for machine learning algorithms scipy (1.5.0), for visualisation purposes seaborn (0.10.1), matplotlib (3.2.2) and finally for organisation Overleaf and Github.

III. Discussion

A. Supervised Learning : classification

It was not possible to properly classify schizophrenic patients, patient's relatives, and controls together as the patient's relatives data set was too noisy. By combining the relatives and the controls groups or by removing the relatives group, the accuracy improved drastically to respectively 78.2% and 81.6% using the support vector classifier with an RBF kernel. We estimate that it would be rather difficult to reach higher accuracy while trying to classify the 3 labels. Maybe this could be due to the spectrum nature of schizophrenia: just like for autism, there might exist light, mild and severe forms of schizophrenia with no clear separation between the pathological forms that require medication and the forms that allow to have a "normal" life.

B. Unsupervised Learning : clustering

As mentioned before, clustering algorithms could be an interesting approach to find more biologically relevant entities in the case of the schizophrenia spectrum. We managed to obtain well-separated and consistent results using the K-means algorithm. However, further analysis are required to check whether those clusters are meaningful and valid. One important limitation of K-means is the typical spherical shape of the clusters and their similar size because those particularities may not reflect the reality of our data set and it may be interesting to explore other non-spherical clustering algorithms.
Managing to find meaningful clusters could be helpful to better understand the disorder in order to be able to properly treat it. The current treatments are mostly non-specific as they aim to alleviate the patient's symptoms rather than solve what causes those symptoms. Thus, they are less efficient and generate many heavy side-effects. Identifying different schizophrenia subgroups would be a step towards personalized medicine.

IV. Further Directions

Here are some suggestions of further experiments that may also be interesting to perform to improve and deepen our analysis of the data set:

1) Most importantly, it is crucial to validate the clusters that we found using K-means and the other algorithms. One could try an extrinsic approach, using the demographics data set for example to see if the different clusters correlate with specific symptoms or other demographical information at our disposition.
2) Secondly, one could explore additional clustering techniques as K-means is not well-suited for our task. The main drawback of K-means is that this approach generates similar sized clusters and we have absolutely no proof that this correctly describes our data set. In addition to that, we observed that none of the methods used to find the optimal number of clusters gave clear and objective results.
3) For classification, we did not split the data and only cross validated the accuracy, therefore, we did not have a validation test. We realized that this is not ideal as the evaluation becomes more biased because of hyperparameters tuning. Although we were limited with the data set size, splitting the data is still important for validating the results.
4) Finally, one could confirm the number of components for the PCA using a parallel analysis as it is considered as one of the most accurate method.

ACKNOWLEDGEMENTS

We want to thank Simona Garobbio, Aline Cretenoud and Michael Herzog from the Laboratory of Psychophysics (LPSY) for their exciting project proposition and for their mentoring

REFERENCES

[1] Godfrey D. Pearlson, Brett A. Clementz, John A. Sweeney, Matcheri S. Keshavan, and Carol A. Tamminga. Does biology transcend the symptom-based boundaries of psychosis? 39(2):165–174.

[2] Kenneth S. Kendler and Eric J. Engstrom. Criticisms of kraepelin's psychiatric nosology: 1896-1927. 175(4):316–326.

[3] Elena I. Ivleva, Brett A. Clementz, Anthony M. Dutcher, Sara J.M. Arnold, Haekyung Jeon-Slaughter, Sina Aslan, Bradley Witte, Gaurav Poudyal, Hanzhang Lu, Shashwath A. Meda, Godfrey D. Pearlson, John A. Sweeney, Matcheri S. Keshavan, and Carol A. Tamminga. Brain structure biomarkers in the psychosis biotypes: Findings from the bipolar-schizophrenia network for intermediate phenotypes. 82(1):26–39.

[4] S. Guloksuz and J. van Os. The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. 48(2):229–244.

[5] Sean P. Carruthers, Caroline T. Gurvich, Denny Meyer, Australian Schizophrenia Research Bank, Chad Bousman, Ian P. Everall, Erica Neill, Christos Pantelis, Philip J. Sumner, Eric J. Tan, Elizabeth H. X. Thomas, Tamsyn E. Van Rheenen, and Susan L. Rossell. Exploring heterogeneity on the wisconsin card sorting test in schizophrenia spectrum disorders: A cluster analytical investigation. 25(7):750–760.

[6] R. J. van den Bosch, R. P. Rombouts, and M. J. van Asma. What determines continuous performance task performance? 22(4):643–651.

[7] Robert K. McClure. The visual backward masking deficit in schizophrenia. 25(2):301–311.

[8] Scott R. Sponheim, Sarah M. Sass, Althea L. Noukki, and Bridget M. Hegeman. Fragile early visual percepts mark genetic liability specific to schizophrenia. 39(4):839–847.

[9] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Jason Brownlee. A gentle introduction to k-fold cross-validation, Aug 2020.