Learning with Block-wise Missing Data to Diagnose Psychiatric Disorders

Pedro Sampaio Lindeza
pedro.lindeza@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisbon, Portugal

November 2019

Abstract

The proposed thesis aims to assess whether the effective handling of block-wise missing data leads to an improved description and classification of psychiatric disorders. Answering this hypothesis can further support the delineation of the borders between psychotic and affective disorders, as well as within psychotic disorders (such as strict versus broad schizophrenia) and affective disorders (such as schizoaffective versus bipolar disorders). To this end, the SchizConnect initiative consolidates different cohort studies, being associated with heterogeneous and highly sparse data. Despite the relevance of this initiative, the data sparsity, heterogeneity, and high-dimensionality brings challenges not settled by the related work on the domain. Along with the applicational matter of highlighting the most discriminative features for the diagnosis of specific classes of disorders, this work also proposes an approach based on finding and weighting subspaces of data for an effective handling, when compared to value estimation, of heterogeneous data with block-wise missings. The proposed partition-wise approach is an embedding that can be applied with classic classifiers. The gathered results confirm its relevance compared to mean, median and nearest neighbours imputation.

Keywords: Classification, Psychiatric Disorders, Schizophrenia, Bipolar Disorder, Disorder Boundaries, Block-wise Missing Data

1. Introduction

A psychiatric disorder is considered to be any mental disorder where a person’s personality is severed, causing mild to heightened mood changes and, sometimes, an inability to distinguish what is real from what is not [4]. In what concerns diagnosing those disorders such as schizophrenia, schizoaffective disorder, manic depression or bipolar disorder, their characteristics seem to not have well-defined boundaries [6]. This compromises the ability to classify each disorder, leading to wrong diagnoses which in turn leads to an incorrect subject’s treatment.

The main purpose of this work is, with the help of Data Mining techniques, to identify the best classification approach of psychiatric disorders on a dataset with severe characteristics that stand as the main problems to overcome. To this end, we had aim to establish the boundaries between such disorders, as finding the assessment which suits better to classify each one.

A better delimitation of the previously referred boundaries between disorders can provide a substantial contribution for two major reasons. First, clear boundaries can help on a more precise clinical description along time. By discovering the relevant discriminative features of some mental disorders, we can also prevent misled treatments on poorly diagnosed patients. Second, their correct identification can increase the ability to perform correct diagnostics at an early stage of the disorder.

However, studies in this domain are hampered by lack of cohort studies. In this work, we consider one of the largest population of case individuals monitored in six major studies conducted in the United Kingdom: BrainGluSch, COBRE, MCIC-Share, NMorphCH, NUSDAST and fBIRNPhaseII. Despite the relevance of this cohort study, it is limited by some aspects. The current monitoring level of individuals with such disorders leads to a high rate of missing values. Not every hospital or institution uses the same assessment on patients, and the patients usually are only tested in one single hospital. This conditioning requires a more careful analysis since it prevents simple comparisons between all the patients. In the context of this work, the target data collected from individuals with psychological disorders is challenged by four major aspects:

1. Sparsity (block-wise missing data),
2. Heterogeneity,
3. High-dimensionality,
4. Class multiplicity and imbalance.

To address the listed challenges we propose a methodology for block-wise missing data, which consists on partitioning the dataset into blocks of data without missing values, weight those partitions according to their relevance and apply known classifiers featuring a voting strategy. The gathered results confirm the relevance of these contributions. In particular, we see that the performance of some classifiers improved from 2 percentage points to 20 percentage points when compared to traditional imputation methods, commonly used to estimate missing values.

2. Background
Psychiatric disorders are a type of mental illness diagnosed based on symptoms like disturbances on rational thinking or severe mood deregulation which can cause disability on day-to-day tasks, loss of freedom, pain and even death. From now onwards, these kinds of disorders are classified in two major groups: Affective Disorders and Psychotic Disorders. Despite the fact that mental illnesses are not circumscribed to these two groups, we will focus on them since they are the focus of the SchizConnect study.

Affective disorders, also known as mood disorders, are characterised by persistent feelings of sadness combined with periods of overly happy emotions. An example of this kind of disorder is Bipolar Disorder previously known as Manic Depression. In addition, there is the psychotic disorders’ group, disorders that cause more severe symptoms like distorted awareness and thinking, hallucinations or delusions, being Schizophrenia the most known and common example of psychotic illness.

The dataset built and used in the context of our work is a result of merging several assessments monitored from a group of individuals from different data sources (studies).

The raw data gathered from the different sources can be found on an online database called SchizConnect. The information from disparate, heterogeneous databases are queried and integrated in a uniform, semantically-consistent structured manner. To a user or a client program, the system appears as a single (virtual) database with a uniform schema/model of the domain, but the data remains at the repositories and under the control of the data providers [5].

Regarding data sources, this dataset is an integration of six online databases corresponding to different studies where individuals took part of. The name of the studies are: BrainGluSch, COBRE, MCICShare, NMorphCH, NUSDAST and fBIRN-PhaseII.

Our data consists in a set of 1293 individuals, where each one is categorized in one of seven different classes: Bipolar Disorder, No Known Disorder, Schizoaffective, Schizophrenia Broad, Schizophrenia Strict, Sibling of No Known Disorder and Sibling of Schizophrenia Strict. Comorbidity does not exist in this collection of data. On Fig. 1 we can see the distribution of individuals per class is not equal, leading to a Class Imbalance setting.

![Figure 1: Class distribution in SchizConnect](image)

The features of the dataset correspond to the assessments made on individuals across the different studies. Summed up, we have 3284 features, divided into two categories, Neuropsychological and Clinical. The first category, Neuropsychological, offers measures pertaining to the cognition of individuals, such as their attention vigilance, intelligence, processing speed or episodic memory. The second group of assessments, Clinical, includes measures associated with the individuals’ medical history, demographics, depression, nicotine dependence, among others. The resulting data of all cognitive assessments is, in the context of our target dataset, heterogeneous, since is formed by categorical values in coexistence with discrete values as well.

A major problem associated with the target dataset is the large presence of missing values.

In our target dataset (1293 individuals × 3284 features), approximately 73% of the values are missing due to two reasons: a) an individual is only assessed by a single study and b) studies not always have features in common among them, in fact, only 584 out of the 3284 can be found in more than one study.

Having said that, when we integrate all the studies’ data into one single dataset, a large number of missing values is expected. Fig. 2 represents a model schema of the final dataset. As previously mentioned, data mining techniques that replace missing occurrences by estimating their val-
ues in this highly sparse dataset can hamper the learning.

Another problem this work aims at addressing is to find the most suitable features – features showing greater discriminative power – to guide the classification of each individual class (Schizophrenia Broad, Bipolar Disorder, etc.) or helping set the boundaries between them. However, if we group the individuals according to their classes and count the features with, at least, one value not missing, the total number of features per class is, once again, imbalanced. The feature selection method to be used shall handle the arbitrarily-high level of missing values.

3. Related Work

Syndromal, genetic and environmental causal studies often place statistical tests to assess the discriminative power of certain symptoms. Nevertheless, with the increasing availability of psychiatric data from case-and-control populations, data mining techniques allow us to take some other conclusions.

Inside the Magnetic Resonance Imaging (MRI) area, a research conducted Honghui Yang et al. [2] proposes a classification model trained on schizophrenia patients and healthy controls, combining functional Magnetic Resonance Imaging (fMRI) and Single Nucleotide Polymorphisms (SNPs) from DNA samples with 222 genes with functions related with six major physiological systems: neurobiology, metabolism, cell proliferation, cardiovascular, inflammation, and cholesterol biochemistry [2]. A collection of 367 SNPs was generated and used in this experiment as features. A Feature Selection method was applied to find the most interesting features from all the 367 SNPs to guide the classification of individuals on either schizophrenic or healthy, based on Feature Selective AdaBoost proposed by Freund and Schapire [1]. Overall, the classifier achieved 87% accuracy on diagnostics, suggesting that combining genetic and brain functional information best represents the majority of symptomatic information used currently to arrive at a clinical diagnosis [2].

Focusing now on bipolar disorder, a very simple study orchestrated by Imanol Perez Arribas et al. [3] sought to diagnose of participants on the basis of their evolving mood. The participants who were not from the healthy group, suffered from one of two affective disorders: Borderline Personality Disorder (BPD) and Bipolar Disorder (BD). Participants rated their mood daily for minimum of 3 months across six different categories (anxiety, elation, sadness, anger, irritability and energy) using a 7-point Likert scale with values from 1 (not at all) to 7 (very much) [3].

By using those values as features, and Naïve Bayes classifier, it was possible to correctly diagnose 75% of the participants [3]. Since we are talking about a balanced dataset, the accuracy value of the experiment is 75%.

Recent advances in data collection technologies have made it possible to collect large amounts of data from many application domains. A study guided by Xiang et al. [7], different types of measurements were used to predict Alzheimer’s Disease. Those types were Magnetic Resonance Imaging (MRI), positron emission tomography (PET), cerebrospinal fluid (CSF), blood test, protein expression data, and genetic data [7]. Since most patients do not have complete data collected from every data source, the result of having multi-source data and high-dimensionality is the existence of (block-wise) missing data [7]. To learn from this data is the major challenge to overcome within this work.

Their solution was to design a model extension of an unified feature learning model for complete data able to handle missing data in blocks. They called it Incomplete Source-Feature Selection (iSFS) [7]. This method 1) partitions the data into multiple groups according to the availability of data sources, 2) learns a consist model for each data source across different data source combinations and 3) in every data source combination, the weights $\alpha$ of each source from the data was calculated.

In their experiment, illustrated in 3, the dataset was partitioned into four subspaces in accordance with the availability of data sources, as highlighted by the red boxes. The goal is to learn three models $\beta_1$, $\beta_2$ and $\beta_3$ for each data source as well as the coefficient $\alpha$ that combines them. Notice that, for the i-th data source, $\beta_i$ remains identical while $\alpha$ may vary across different groups [7]. The outcome of this experiment can be found on Fig. 3.

The experiment to examine the efficacy of iSFS was based on the classification between individuals with Alzheimer disease, normal cognition (con-
4. Solution

In the context of the thesis problem, and inspired by the work of Xiang et al. [7], we propose an alternative approach. The solution for the problem requires finding subspaces within the dataset where there are no missing values in block, or quite few, and consider them as partitions, properly weighting them and applying several known classification algorithms, e.g. Naïve Bayes or Decision Trees, with a voting strategy which takes into account each partition vote and weight on reaching an agreement for the classification drill.

4.1. Preprocessing

The preprocessing started by integrating the data from the six study sources, BrainGluSch, COBRE, MCICShare, NMorphCH, NUSDAST and fBIRN-PhaseII. As a result, a total of 3284 features for the different assessments run on individuals are produced, plus the “subjectid” representing the individual’s ID and one created feature “study” to identify the individuals’ data source origin. For the few cases where we have two or more assessments for a single individual with different values, corresponding to different rehearsals at different periods, the average value of the repeated set of observations replaces all them and stands as the only value for that individual regarding that test.

Secondly, from an additional data source (metadata), we extracted the diagnostics of the 1293 individuals, our target class for classification.

After having all the gathered data, it was possible to observe the existence of some outliers. First, values out-of-range were detected due to disparity between their value and column’s average value. In addition, it was possible to find some values in form of annotations (e.g. “Patient refused to answer”), along with other observations from type “DK” (Do not Know), “MV” (Missing Value) or “Nan” (Not a number). All the previously mentioned and similar observations were removed from the dataset and treated like missing values. As a result, some assessments ended up having no observations, being in this way discarded from the set of features.

4.2. Partition Discovering

The first approach to find the least possible number of partitions maps the target dataset into a bipartite graph and discovers bicliques. First, the data space is seen as a bipartite graph, $G = (U, V; E)$, being the index of the dataset, “subjectid”, along with all features, two disjoint and independent sets $U$ and $V$ respectively. $E$ the edges representing the presence of an observation of an individual in $U$ from test in $V$.

Second, and given such bipartite graph $G$, we find bicliques in $G$. In the context of the solution, a biclique would correspond to a partition. By finding the maximum edge bicliques it is possible to assure the minimum number of partitions possible. However, this method has shown to be unsuitable given its NP-hard complexity and the large dimensions of the generated graph.

The second approach consisted on dividing the target dataset space into two different types of subspaces. First, a subspace where the number of observations is more conglomerate and its rate to missing values is high, which we call Partition, and second, a subspace mostly compound of missing values, named Block of Missing Data.

We aimed on finding the overlaps of features among studies and the first conclusion taken was that the study MCICShare had no overlaps with any other study. As a matter fact, the presence of this source can hamper performance. Without overlaps, this can be treated like a separated dataset and would not bring any contributions to this block-wise approach. Having said that, all features from MCICShare were discarded along with the subjects from the mentioned study. Our target dataset was now reduced to 1080 individuals and 2864 features.

Calculating feature overlaps allows us to find partitions as shown on Fig.4. Basically, each overlap

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSFS</td>
<td>0.8103</td>
<td>0.8077</td>
<td>0.8124</td>
</tr>
<tr>
<td>SVD</td>
<td>0.7756</td>
<td>0.7770</td>
<td>0.7746</td>
</tr>
<tr>
<td>KNN</td>
<td>0.7668</td>
<td>0.7161</td>
<td>0.8072</td>
</tr>
<tr>
<td>Mean</td>
<td>0.7789</td>
<td>0.7845</td>
<td>0.7744</td>
</tr>
<tr>
<td>EM</td>
<td>0.8089</td>
<td>0.7963</td>
<td>0.8189</td>
</tr>
<tr>
<td>MC</td>
<td>0.5957</td>
<td>0.5710</td>
<td>0.6155</td>
</tr>
</tbody>
</table>

Table 1: Classification results of individuals with Alzheimer disease, normal cognition, and mild-cognitive impairment.
of study features stand as a partition, with group of features from the same study standing as well.

In result, we obtained eight partitions, from which five correspond to the remaining studies. Two of them are a group of common features between BrainGluSchi and COBRE, NMorphCH and NUS-DAST, and the third partition is the interception of features among BrainGluSchi, COBRE and fBIRN-PhaseII. Every time we split the dataset into training and testing sets for classification purposes, features will be grouped within the training set according to the groups previously mentioned.

4.3. Classification

The classification methodology of the solution consists in setting a classification model for each discovered partition so we can test every individual from the testing set on all partitions separately. After doing so, we assign a weight to each partition in accordance with the classification performance on that partition along with some of the partition’s properties. A partition is relevant when its model’s performance is above a defined threshold. Finally, the classification result of an individual is given by a weighted voting from all the relevant partitions.

4.3.1 Formulation

It is important to first introduce some notations which will simplify the demonstration.

The accuracy $acc$,

$$acc = \frac{\sum True}{\sum False},$$

and class Sensitivity,

$$S_A = \frac{True A}{True A + False B_A + False C_A + False D_A}$$

of a classification model are based on a confusion matrix, where rows correspond to predicted class, columns to expected class, as seen on table ??, and consists of the ratio between true and false predictions.

Suppose $I$ represents the group of all partitions discovered within the training set. The model prediction’s accuracy of the partition $i$ is represented by $acc_i$, where $i \in I$, $0 \leq acc_i \leq 1$.

A partition is relevant if its accuracy is greater or equal to 0.50. $E$ is the subset of relevant partitions, therefore $E \subset I$. For,

$$\forall i \in I, i \in E \iff acc_i \geq 0.50.$$

Let $C$ stand for the group of different classes represented in the target variable of the target dataset, and $T$ the group of individuals from testing set. Regarding predictions, the classification model’s prediction of an individual $x \in T$ on partition $i$ results on an array $\delta_{x,i}$,

$$\delta_{x,i} = (P_{x,i,1}, \ldots, P_{x,i,c}), \ 1 \leq c \leq |C|,$$

of size $|C|$ where the $c^{th}$ position means the probability of the individual $x$ of belonging to class $c$, given by the partition $i$’s classification model. Where,

$$1 = \sum_c P_{x,i,c}, \ 1 \leq c \leq |C|.$$  

The weighted voting method towards calculating the global prediction, $\delta_x$, of an individual $x$ is given by the formula:

$$\delta_x = \arg \max_{P_{x,c}} \sum_{i \in E} \alpha_i \delta_{x,i}, \ 1 \leq c \leq |C|,$$

where $\alpha_i \in \mathbb{R}$, is the weight of the $i^{th}$ partition. Nevertheless, partition weights are also constrained to,

$$1 = \sum_{i \in E} \alpha_i.$$  

although, if a partition $i$ is not relevant then,

$$\alpha_i = 0.$$

By defining the total number of predictions made with the $i^{th}$ classification model as $\delta_i$, therefore the accuracy of the $i^{th}$ partition is deducted by:

$$acc_i = \frac{\sum_i True \ \delta_i}{\sum_i False \ \delta_i}, \ i \in I,$$

therefore, the global accuracy is given as:

$$acc_{global} = \frac{\sum_x True \ \delta_x}{\sum_x False \ \delta_x}, \ x \in T.$$  

4.3.2 Partition-wise Classification

Once the target dataset is properly preprocessed, we split the dataset into 10-fold subsets using a stratified cross-validation technique, where each k-fold is balanced by having the same number of same class individuals. The process iterates 10 times, although each time the testing set is a different k-fold and the remaining nine compound the training set.

The following steps take place ten times:

1. Discover the partitions $I$ within the training set by featuring overlapping,
2. Find estimator of missing values for each partition from training data using the estimator learned from training data for the corresponding partition,
3. Estimate testing set missing values by feature’s mean imputation from the training set (corresponded partition),
4. Learn a classification model for each partition $i \in I$ based on its observations,
5. Predict the class, $\delta_{x,i}$, of every individual $x \in T$ on each partition $i \in I$, by discarding observations of $x$ from features not present in $i$,
6. Calculate every $acc_i$, $i \in I$ and find the relevant partitions $E (acc_i \geq 0.50)$,
7. Calculate $\alpha_i$ for each $i \in E$,
8. For each individual $x \in T$, compute the global prediction $\delta_x$,
9. Reckon the global accuracy, $acc_{\text{global}}$, of the drill.

4.4. Partition Weighting
Most Suitable Metric

For the purpose of fairly calculate partitions’ weights, an approach based on the relation between a certain property of a partition and the accuracy results of the classification model from that same partition was proposed. This analysis further allows us to understand how relevant is a given property of a partition to guide the classification task.

Suppose $\eta = (acc_1, \ldots, acc_n)$, $n = |I|$, is a set of each partition model prediction’s accuracy regarding a classification drill, and $\gamma = (\gamma_1, \ldots, \gamma_n)$, $n = |I|$, a set of values where the $\gamma_i$ stands for a $i$th property of partition. The target properties can be one of the following:

- Number of subjects,
- Number of features,
- Area (number of elements),
- Average number of non-missing features per subject.

To measure the proportion of the variance for a dependent variable ($\gamma$) explained by an independent variable ($\eta$), the R-squared ($R^2$) measure was used. To do so, a predictive model of linear regression was applied to identify the equation that produces the smallest difference between all of the observed values and their fitted values, with $R^2$ standing for the dependent variable variation. The most suitable metric corresponds to the property that is easier to predict through a linear regression model and in consequence, to the highest $R^2$ value’s metric.

To illustrate the stated point, a classification experiment using Naïve Bayes classifier was carried on the target dataset splitted 80% for training and 20% for testing. In Fig. 5 we can see the correlation between all the eight partitions (discovered through “Feature Overlap” method), classification's

Figure 5: Charts showing the found correlation between the number of subjects and the accuracy for the eight partitions in SchizConnect data, with a trend line showing the $R^2$ value.
accuracy, and the number of subjects. Also a trend line representing the fitted linear regression model next to the $R^2$ value obtained.

For each partition $i$ where $\gamma_i$ corresponds to the partition’s value of most suitable metric $\gamma$ (highest $R^2$), the weight $\alpha_i$ of the relevant partition $i \in \gamma$ is given by the formula (11) which normalizes $\gamma_i$ according to the sum of every value in $\gamma$ associated to a relevant partition.

$$\alpha_i = \frac{\gamma_i}{\sum_j \gamma_j}, \quad i \in \gamma, \gamma_j \in \gamma, 1 \leq j \leq |\gamma|.$$ (11)

This weighing method was not used in the experimental part of our work because it is not assured that the dimensions of a partition might be related to its accuracy.

### Accuracy Weighting

The learning model to be used for classification does a balanced prediction based on the relevance of each partition in accordance with the model’s performance on that partition. In order to perform this strategy, there is first a need to assign a weight to each partition.

Accuracy weighting is a simplistic approach. First, we consider the model's prediction accuracy for the partitions. Second, only relevant partitions are accounted for the calculation. Using the notation of section 4.3.1, if supposed $\eta = (acc_1, \ldots, acc_n)$, $n = |\gamma|$, is a set of values from each partition model’s prediction accuracy, the weight $\alpha_i$ of partition $i$ is a normalized value of $acc_i$ according to the sum of every value in $\eta$ (12):

$$\alpha_i = \frac{acc_i}{\sum_n acc_n}, \quad i \in \eta, acc_n \in \eta, 1 \leq n \leq |\gamma|.$$ (12)

### 4.5. Evaluation Methodology

To assess the performance of this methodology, we compared its effectiveness against approaches reliant on the estimation of missing values. To this end, we predicted the missing values following three commonly used imputation methods:

- Mean imputation
- Median imputation
- k-Nearest Neighbours (k-NN) imputation

For each of the four methods, partitions and the three baselines mentioned above, we considered five different and already implemented classifiers:

- Naïve Bayes (NB)
- Decision Trees
- Random Forest
- Support Vector Machine (SVM)
- k-Nearest Neighbours (k-NN)

In order to preserve a fair comparison between all the experimented methods with different classifiers, the same seed was used in all cases. In other words, the 10-folds are the same for each experiment.

Regarding each classifier, the metrics to measure the effectiveness among the partition method and the imputation techniques are the average of all 10-fold accuracy, $acc_{global}$ (10), and Cohen’s kappa coefficient of agreement. Also class sensitivity $sensitivity_e, e \in C$, to determine which combination of classifier/method better discriminates a class.

Default parameterizations were considered given the fact that our aim is to understand the relative performance impact of the alternative embedding approaches to handle missing values. Nevertheless, hyperparameterization using grid search or, more efficiently, Bayesian optimization can be used in order to optimize the performance of classifiers.

### 5. Results

The programming code developed for the experimental purposes can be found on [www.github.com/Pedrolindeza/mscthesis](www.github.com/Pedrolindeza/mscthesis).

#### 5.1. Explanatory Data Analysis

In the interest of characterizing the variables with high discriminative power, we performed $\chi^2$ tests ($\chi^2$ method on sklearn) along with ANOVA tests based on F-statistic ($f$test method on sklearn) and collected the resulting scores and $p$-values for the 2864 input variables. Fig. 6a contains the distribution of the scores for both tests, and Fig. 6b the $p$-values distribution, across all features of the target dataset. We observe that the cumulative $\chi^2$ score explained by the top 100 variables is higher than the cumulative score from the remaining thousands of variables. The same happens for ANOVA test. We can state that 125 features have much higher independence values compared to all dataset features.

In order to perform a deeper analysis from previous results, we selected the most interesting features by choosing the ten features with higher $\chi^2$ values, followed by plotting class condition distributions for each interesting feature in order to understand its discriminative power regarding specific classes (Fig. 7). We observe that in all top interesting features there is a noticeable difference in the distribution of values among individuals of class No Known Disorder and Schizophrenia Strict. Also, both classes Sibling of No Known Disorder
and Sibling of Schizophrenia Strict got to be discriminated from all the others in every interesting feature, while Schizoaffective class is also well differentiated from all others in the top six interesting features.

5.2. Classification Performance

The following results concern the execution of several classification experiments using different classifiers. Each classifier was used in four different approaches: 1) partition-wise approach using partition discovering of “Feature Overlaps” and “Accuracy Weighting”, 2) estimated values by mean imputation, 3) estimated values by median imputation and 4) estimated values by k-NN imputation (k=5). The seed, or k-fold division, was preserved among experiments.

Since we use a 10-fold stratified cross validation approach, the performance results values of accuracy (table 2) and Cohen’s kappa agreement coefficient (table 3) represent the average, μ, and standard deviation, σ, of the 10-fold validation under the form of μ (+/- σ).

In general, the obtained results point out to a performance improvement when the partition-wise methodology is used.

The performance of Naive Bayes classifier improves by 20pp in accuracy and kappa statistic. This improvement was achieved thanks to a delineated improved ability to classify schizophrenia strict and schizophrenia broad individuals (partitions allow to consider votes which decrease the risk of the classifier having deviations on a subset of classes from the disorders analyzed). This can also be an indirect effect of dimensionality reduction within each partition (absence of large number of poorly discriminating variables that may interfere with correct decision).

The performance of partition-wise random forests was improved by over 10pp in accuracy and kappa statistic against imputation methods. Although there is no major difference in the discrimination of the disorders, random forests show particularly good ability to discriminate the health control No Known Disorder class and Sibling of Schizophrenia Strict.

Regarding the Support Vector Machine (SVM) classifier, its performance is improved by over 10pp for both accuracy and kappa statistic when comparing partition-wise to imputation methods. The imputation of values cannot handle the classification of affective disorders (Bipolar Disorder and Schizoaffective). The partition-wise method is associated with a significant improvement in the classification of individuals belonging to Schizophrenia Strict class.

The performance of k-Nearest Neighbours (k-NN) classifier with the partition-wise approach is competitive with imputation methods. The slight improvements observed in the partition-wise method’s performance are also associated, as in SVM, with the effective discrimination of Schizophrenia Strict class. Basically, what happens is that the search for the nearest individuals for a partition is generally coincident, either the individuals selected for a given partition are the same as for the total dataset or the partition has a low weight and this has any affect on the decision.

Both k-NN and SVM are not good approaches to classify affective disorders.

In contrast with all remaining classifiers, performance with decision trees is slightly worse with partition-wise principles. From the inspection of decision trees learned using partitioning principles, we generally observed that the resulting trees are have significantly higher depths, what we believe to be the reason for the verified overfitting. Decision tree’s slight decrease in performance for the partition-wise approach is primarily driven by the classification errors experienced for Schizoaffective and Schizophrenia Broad. Broadly speaking, decision trees are substantially better on discriminating Bipolar Disorder.

Comparatively, decision trees and k-NN seem to be the best classifiers. A closer inspection on what is explaining these classifiers’ behavior leads us to the topic of local decisions. The target dataset is highly heterogeneous, with a population of very different individuals for a single disorder or even among disorders. To properly learn a model under these circumstances, a good technique is to look at the nearest individuals (k-NN) or perform test-and-splits in a way that tree decisions lead to a similar result (closer individuals).

This last observation is corroborated by the lower performance of Naive Bayes (NB) and Support Vector Machine (SVM), two classification models that in the absence of regularization (as well as proper kernels for SVM) are seen as global classification models, inherently unable to weight or remove uninformative variables.

5.3. Complementary Results

There are evident differences within a class, for the calculated sensitivities either across classifiers and across methods to handle missing values (tables 4, 5, 6, 7 and 8). The results prove these heterogeneous behaviors, being possible to verify some classifiers have better performances for a specific disorder than others. This leads one to think that an ensemble could be a good approach at managing decisions, for example an ensemble with weighting associated to observed differences.

The collected decision trees from Fig. 8, Fig. 9 and Fig. 10 correspond to the first iteration of the 10-fold stratified cross validation. The colors rep-
resent paths linked to different disorders, with the color of a rectangle being the largest number of individuals with a given disorder at a given point in the tree.

Looking at the different paths of the three trees for a given class disorder (paths with predominance of a single color), we find evident differences in each illustrated tree. This allows an easier definition of borders among disorders. This detail thus allows us to consider variables that would otherwise be overlooked due to the reduction in discriminatory power created by imputation approaches.

An analysis on the shortest path for classification of Schizophrenia Strict in each decision tree was made. The Cobre partition tree (Fig. 8) takes a minimum of four features (y289, y632, y984 e y951) to classify an individual as Schizophrenia Strict, while NMorph & Nusdast partition (Fig. 9) it takes three (y1538, y2813 e y2831). Regarding the mean imputation (Fig. 10), the shortest path only takes one feature (y3103) to classify an individual as Schizophrenia Strict.

The decision trees corresponding to partition-specific models – Cobre (Fig. 8) and (Fig. 9) – have higher depth levels than the tree with mean imputation (Fig. 10), probably due to the fact mean imputation ends up preventing / affecting the discovery of deep paths with great discriminative capacity. Nevertheless, mean imputation leads to a beneficial side effect: reduced propensity to overfitting.

6. Conclusions
This work addressed the problem of learning from heterogeneous data with block-wise missings to support the diagnosis of psychiatric disorders. Under this purpose, we compared several imputation methods with more advanced methods based on data partitioning principles where dedicated models are learned for each partition, followed by a voting strategy sensitive to the relevance of each model.

To this end, we outlined: 1) principles to find subspaces of interest, and 2) identified different criteria of weighting those same subspaces, in order to perform the voting strategy. The performance of the model learned from a partition is taken into account when weighting that partition.

The obtained results validate the introduced thesis and confirm the relevance of the proposed approach. Generally, we observe that the use of partition-wise principles – extraction of subspaces of interest from the input dataset – noticeably improves the performance of most classifiers. From the application analysis, decision trees were computed from the learning model, showing differences between the proposed approach and the comparative ones and further evidencing new associations with potential interest for the discrimination of different disorders.

6.1. Future Work
For future work we aim at pursuing a comprehensive validation of the gathered results by health professionals specialized in the targeted disorders. Second, the extended applicability of our approach on additional datasets from other domains is also an open window for further investigation.

In addition, it would be interesting to study the performance of the partition-wise methodology for prediction analysis using regression and unsupervised learning methods, such as pattern mining. In terms of the partition-wise algorithm, there is also space to explore difference approaches concerning: 1) the partition discovering methods (e.g. hierarchical partitions where groups of partitions can be

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.82 (+/- 0.11)</td>
<td>0.62 (+/- 0.08)</td>
<td>0.47 (+/- 0.13)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.88 (+/- 0.07)</td>
<td>0.98 (+/- 0.02)</td>
<td>0.93 (+/- 0.05)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.91 (+/- 0.04)</td>
<td>0.88 (+/- 0.10)</td>
<td>0.87 (+/- 0.10)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.82 (+/- 0.09)</td>
<td>0.67 (+/- 0.10)</td>
<td>0.53 (+/- 0.17)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.79 (+/- 0.09)</td>
<td>0.63 (+/- 0.05)</td>
<td>0.66 (+/- 0.07)</td>
</tr>
</tbody>
</table>

Table 2: Accuracy of 10-fold cross validation execution.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.68 (+/- 0.15)</td>
<td>0.46 (+/- 0.09)</td>
<td>0.34 (+/- 0.12)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.80 (+/- 0.10)</td>
<td>0.96 (+/- 0.04)</td>
<td>0.89 (+/- 0.08)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.85 (+/- 0.06)</td>
<td>0.82 (+/- 0.13)</td>
<td>0.79 (+/- 0.13)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.49 (+/- 0.21)</td>
<td>0.38 (+/- 0.21)</td>
<td>0.23 (+/- 0.21)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.54 (+/- 0.27)</td>
<td>0.37 (+/- 0.07)</td>
<td>0.43 (+/- 0.09)</td>
</tr>
</tbody>
</table>

Table 3: Kappa statistic of 10-fold cross validation execution.
seen as an additional partition), and 2) the weighting process by combining domain knowledge to turn the weighting schema more application-oriented.

References


## Appendices

### Table 4: Sensitivity of Bipolar Disorder class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Partition-wise Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.17 (+/- 0.37)</td>
<td>0.20 (+/- 0.40)</td>
<td>0.40 (+/- 0.49)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.60 (+/- 0.49)</td>
<td>0.80 (+/- 0.40)</td>
<td>0.40 (+/- 0.49)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.00 (+/- 0.00)</td>
<td>0.10 (+/- 0.30)</td>
<td>0.00 (+/- 0.00)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.00 (+/- 0.00)</td>
<td>0.00 (+/- 0.00)</td>
<td>0.00 (+/- 0.00)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.00 (+/- 0.00)</td>
<td>0.00 (+/- 0.00)</td>
<td>0.00 (+/- 0.00)</td>
</tr>
</tbody>
</table>

### Table 5: Sensitivity of No Known Disorder class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Partition-wise Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.75 (+/- 0.17)</td>
<td>0.69 (+/- 0.22)</td>
<td>0.37 (+/- 0.18)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.97 (+/- 0.04)</td>
<td>1.00 (+/- 0.01)</td>
<td>0.99 (+/- 0.03)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.98 (+/- 0.02)</td>
<td>0.95 (+/- 0.10)</td>
<td>0.95 (+/- 0.09)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.92 (+/- 0.13)</td>
<td>0.99 (+/- 0.02)</td>
<td>0.90 (+/- 0.27)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.86 (+/- 0.09)</td>
<td>0.89 (+/- 0.05)</td>
<td>0.88 (+/- 0.06)</td>
</tr>
</tbody>
</table>

### Table 6: Sensitivity of Schizoaffective class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Partition-wise Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.33 (+/- 0.41)</td>
<td>0.36 (+/- 0.47)</td>
<td>0.38 (+/- 0.48)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.34 (+/- 0.37)</td>
<td>0.68 (+/- 0.41)</td>
<td>0.62 (+/- 0.42)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.12 (+/- 0.17)</td>
<td>0.18 (+/- 0.29)</td>
<td>0.30 (+/- 0.24)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.03 (+/- 0.05)</td>
<td>0.00 (+/- 0.00)</td>
<td>0.00 (+/- 0.00)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.12 (+/- 0.12)</td>
<td>0.00 (+/- 0.00)</td>
<td>0.01 (+/- 0.04)</td>
</tr>
</tbody>
</table>

### Table 7: Sensitivity of Schizophrenia Broad class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Partition-wise Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.92 (+/- 0.19)</td>
<td>0.74 (+/- 0.13)</td>
<td>0.89 (+/- 0.12)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.57 (+/- 0.36)</td>
<td>0.99 (+/- 0.04)</td>
<td>0.85 (+/- 0.18)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.88 (+/- 0.17)</td>
<td>1.00 (+/- 0.00)</td>
<td>1.00 (+/- 0.00)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.00 (+/- 0.00)</td>
<td>0.25 (+/- 0.16)</td>
<td>0.28 (+/- 0.18)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.67 (+/- 0.16)</td>
<td>0.54 (+/- 0.08)</td>
<td>0.56 (+/- 0.11)</td>
</tr>
</tbody>
</table>

### Table 8: Sensitivity of Schizophrenia Strict class.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Partition-wise Mean Imputation</th>
<th>Median Imputation</th>
<th>k-NN Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.85 (+/- 0.22)</td>
<td>0.51 (+/- 0.32)</td>
<td>0.53 (+/- 0.32)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.87 (+/- 0.15)</td>
<td>0.96 (+/- 0.06)</td>
<td>0.91 (+/- 0.13)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.96 (+/- 0.05)</td>
<td>0.96 (+/- 0.04)</td>
<td>0.93 (+/- 0.11)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.90 (+/- 0.20)</td>
<td>0.48 (+/- 0.29)</td>
<td>0.37 (+/- 0.35)</td>
</tr>
<tr>
<td>k-NN (k=5)</td>
<td>0.84 (+/- 0.08)</td>
<td>0.54 (+/- 0.16)</td>
<td>0.61 (+/- 0.17)</td>
</tr>
</tbody>
</table>
(a) $\chi^2$ and ANOVA tests scores.

(b) $p$-values from the $\chi^2$ test and ANOVA F-Statistic test. (Null hypothesis: there is no association between the variable and the class.)

**Figure 6**: Study on the discriminative power of the monitored clinical and neuropsychological tests recurring to the $\chi^2$ test and ANOVA test based on F-statistic.
Figure 7: Class-conditional distributions of the top 10 discriminative features (under a Normal assumption).
Figure 8: First iteration's decision tree of partition Cobre
Figure 9: First iteration’s decision tree of partition N Morph & Nudost
Figure 10: First iteration's decision tree for mean imputation method