Do Data Mining Methods Support the Three-Group Diagnostic Model of Primary Progressive Aphasia?

Telma Pereira*
tpereira@kdbio.inesc-id.pt

Carolina Maruta†
carolmaruta@gmail.com

Alexandre de Mendonça†
mendonca@medicina.ulisboa.pt

Manuela Guerreiro†
mmgguerreio@gmail.com

Sara C. Madeira*
sara.madeira@tecnico.ulisboa.pt

*INESC-ID and Instituto Superior Técnico
Universidade de Lisboa, Portugal

† Laboratório de Neurociências, Instituto de Medicina Molecular and Faculdade de Medicina
Universidade de Lisboa, Portugal

ABSTRACT
Primary Progressive Aphasia (PPA) is a neurodegenerative disease characterized by a gradual dissolution of language abilities, with higher risk to evolve to dementia. For that reason, discovering the different subtypes of PPA patients is fundamental to timely introduce pharmacological or other therapeuthical interventions in order to improve patient's quality of life. Although three variants of PPA (logopenic, agrammatic and semantic) are currently accepted, uncertainties still persist regarding this classification as many patients don't seem to fit in this subdivision. In this context, we applied data mining techniques to a clinical series of PPA patients in order to test whether the outcome would support the three-group diagnostic model of PPA. We used an unsupervised approach to test the three-group diagnostic model of PPA versus the existence of distinct number of groups. A supervised learning approach was also tried, to test the feasibility of using an automatic procedure to classify patients in one of the three variants considered. Our experimental results do not support a clear distinction of the three PPA variants. Clustering results pointed to the existence of two main groups, where most agrammatic cases were clustered separately from the logopenic and semantic cases. The logopenic variant was the most difficult class to individualize. The classification outcome achieved good performances only when a restrict set of patients was analyzed.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data mining; J.3 [**Life and Medical sciences**]

General Terms
Algorithms, Experimentation.

Keywords
Primary progressive Aphasia (PPA), neuropsychological data, data-mining, clustering, classification.

PH-KDD'14, 8 24-27, 2014, New York, NY, USA
Copyright 2014 ACM 1-58113-000-0/00/0010 …$15.00.

1. INTRODUCTION
Primary Progressive Aphasia is a clinical condition characterized by a gradual, progressive loss of specific language functions resulting from neurodegeneration of the left-hemisphere language network [1][2]. Although the language impairment should be the most salient feature during at least two years, other cognitive deficits may emerge during the course of the disease [3]. For decades, only two disease presentations were accepted: Non-fluent Aphasia and Semantic Dementia. Later on, it was acknowledged that this reflected an oversimplification of the clinical entity of PPA since several patients seemed to display distinct clinical features not entirely concordant with the previous variants. A third, logopenic, phenotype was then defined [3]. In 2011, an international criteria established three variants (agrammatic or nfvPPA, semantic or svPPA and logopenic or lvPPA), which were formerly recognized as the main PPA presentations [1]. Despite this effort, uncertainties still persist regarding this classification. Not only a significant proportion of cases remain unclassifiable nowadays [4], [5], but also recent evidence argue in favor of the existence of other linguistic profiles [6]–[10]. On the other hand, lvPPA was the last clinical syndrome to be described and thus it may suffer from a lack of specific and distinctive features, which in turn, may lead to an erroneous classification and induce a delayed diagnosis [7][18]. Distinguishing different disease presentations in PPA from a neuropsychological standpoint is important to effectively tackle the progression and conversion to dementia, improve diagnostic accuracy and lead to adequate pharmacological and other therapeutic intervention.

Data mining driven analysis applied to large datasets, in particular, in the context of large multicentric studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) project [13], is becoming increasingly relevant in the field of neurodegenerative diseases. Moreover, our group has already shown that these methods can improve accuracy, sensitivity and specificity of classification and predictions through neuropsychological testing in Mild Cognitive Impairment (MCI) and AD patients [14], [15]. To our knowledge, and with respect to PPA, few (if any) clustering/classification data mining based techniques have been systematically tried on extensive neuropsychological datasets, representing this work, a novel and promising contribution to this research field. In this context, the present study aims to analyze the applicability of several unsupervised and supervised learning algorithms in a clinical series of PPA patients. Specifically, we used an unsupervised approach to test the three-group diagnostic model of PPA versus

the existence of two classic groups, as well as detect the existence of additional disease presentation patterns. We also used supervised learning to verify the feasibility of using an automatic procedure (classification model) to classify patients in one of the three variants considered.

This paper is organized as follows: in section 2 we present the data under study along with the preprocessing and feature selection performed in the dataset. Still in the same section, we describe the algorithms and strategies used, either concerning the unsupervised or the supervised learning approach. Following we present and discuss the main results, conclusions and future work.

2. METHODS

This section describes the dataset, its preprocessing, feature selection; and the unsupervised and supervised strategies.

2.1 Dataset description

The dataset used in this work is composed by a clinically-based cohort of patients with the diagnosis of PPA, recruited at the Dementia Outpatient Clinic of Hospital Santa Maria and a private memory clinic in Lisbon (Memoclínica), between 1983 and 2012. It comprises demographic and clinical data, the results of neuropsychological and language tests, and the judgment of medical doctors, regarding the patients' cognitive condition in one of the three new variants of PPA [1].

The original dataset included 155 patients, 104 (67%) of which were clinically classified into one of the three PPA subtypes (31 nfvPPA, 35 svPPA, 38 lvPPA). The remaining 51 cases (33%) could not be classified (unPPA). Since we were interested in testing the true existence of three PPA variants under optimal conditions (grouping typical cases of each variant), we decided to define a "model" subset. Thus, within the overall sample, we defined a subgroup of 36 "*model patients*" (14 nfvPPA, 12 svPPA, 10 lvPPA), which comprised patients whose classification was performed with a high degree of confidence by clinical experts, and are therefore good representatives of each subtype. Since the main goal of this analysis was to infer whether, either the unsupervised or supervised algorithms, were able to distinguish the patients by their predefined PPA variants, we decided to discard the unclassifiable cases (and also to remove noise from the dataset) and proceed to analyze in more detail the (clinically) classified cases. Table 1 summarizes the proportion of PPA cases encompassed in each set of patients under study.

Table 1. Proportion of PPA variants of each set of patients.

		Variants of PPA		
	n	lvPPA	nfvPPA	svPPA
All classified patients	104	31 (30%)	35 (34%)	38 (36%)
Model patients	36	10 (28%)	14 (39%)	12 (33%)

2.1.1 Preprocessing

Attributes with percentage of missing values greater then 30% were removed. The dataset was then composed by 154 attributes (against the original 193 attributes). Moreover, we imputed missing values beforehand using the average value or mode (whether the attribute is numerical or nominal, respectively) for algorithms that cannot deal with missings or for which we considered favourable to handle missing values *a priori* (Expectation Maximization - EM, K-Means and X-Means). We did not perform imputation when algorithms were prepared to handle missing values (Hierarchical Clustering). Z-scores variables were discretized (into 4 classes) following a subjective division method: classes were defined by the judgement of experts in the application domain. Numerical and ordinal variables were normalized following the min-max normalization [16].

2.1.2 Feature Selection

We performed a feature selection based on different domains. Apart from the original set of (154) attributes, we used the following sets of features: *Language attributes* (96 attributes; features whose purpose is to examine the language defects hold by the patient) and *Model attributes* (46 attributes; features selected on the basis of being necessary to classify the subgroup of "model patients"). We also defined a set of *Operational Criteria Attributes* (OCA): nine qualitative language dimensions that were operationally defined after the core features specified in the consensus guidelines and defined by other authors [12]. The latter attributes were only defined for the "model patients".

2.2 Unsupervised Learning

We followed several clustering approaches. We used standard clustering algorithms, available in WEKA® [17], such as K-means and EM. Different number of clusters (*k* set as 2, 3, 4 and 5) were predefined with the purpose of testing the possible existence of more (or less) than three PPA classes or possible groups corresponding to intersections between them (Figure 1). We also used bottom up hierarchical clustering. In this case, contrary to the previous clustering algorithms, we do not specify a fixed number of clusters (*k*) *a priori*. We used the complete linkage method to determine the distance between groups of patients. The distance between two observations followed a distance metric for mixed composition attributes [16]. These analyses were performed in Matlab®. In addition, we used the X-Means algorithm [18] (variant of K-Means estimating the number of clusters by optimizing the Bayesian Information Criteria).

While applying the standard clustering algorithms, we followed different strategies, motivated by the evolving results. Therefore, besides clustering the entire dataset (with different parameters and set of attributes) we also ran the algorithms in datasets composed by only two PPA classes. In addition, in another strategy, we started by clustering the set of patients with *k*=2 and then, ran the algorithms in cases located in clusters obtained from the previous grouping. The idea was to test whether patients were more easily separable (regarding the three PPA classes) when some patients' division was already performed, and therefore, some noise removed.

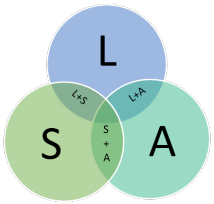

Figure 1. Problem formulation. Note: A - agrammatic variant; S - semantic variant; L - logopenic variant

Since, in some situations, individual clustering algorithms may not be capable of correctly find the underlying structure for all datasets, we used consensus clustering [19] as an alternative approach. The goal was to identify stable clusters, given that clusters discovered by several algorithms tend to be more reliable. The first step of our consensus clustering consisted on building the ensemble by running alternatively K-Means and EM several times with the parameter *k* set as 2, 3 and 4, and alternating the set of attributes used in each analysis (original set of attributes,

language and attributes of model cases). A novel dataset was then created, where columns depicted the cluster assigned to each case, according to different clustering algorithms and/or type of attributes entered in the analyses. In a second step, the EM clustering algorithm combined the clustering results from the first step to generate a representative consensus clustering, with the global parameter *k* set as 2, 3, 4 and 5 (*k* global). This task was performed with WEKA®.

We performed an external clustering evaluation, where the results were evaluated based on data that was not used for clustering, that is, using the PPA classes clinically classified by experts (gold-standard). Moreover, by running the algorithms with different number of clusters and set of variables, we analyzed which clusters were more stable and consistent (groups that remained almost unchanged whenever the number of clusters and/or the set of variables was changed).

2.3 Supervised learning

Supervised learning algorithms (available in WEKA® [17]) were applied to infer whether they would support the three-group based diagnostic model for PPA. We tested both eager algorithms (*Näive Bayes* [16]*, Decision Trees* [20]*, Multilayer Perceptron* (MLP) [16], *Support Vector Machines* (SVMs, SMO) [21]) and lazy learning algorithms (*K-Nearest-Neighbour* (KNN) [20] and the *Lazy Bayesian Rules* (LBR) [22]).

We followed different strategies while applying the classifiers. At first, we ran the classifiers in different datasets: dataset composed by all the patients with the original PPA variants (nfvPPA, lvPPA and svPPA); and datasets composed by all the instances, but in which one class is maintained (positive class) and the two remaining are merged together to form the negative class ("dataset A/¬A", "dataset S/¬S" and "dataset L/¬L") since we decided to reduce the ternary problem to binary. Table 2 shows the proportion of PPA cases belonging to each generated dataset. This study was motivated by our interest in verifying whether the classifiers presented a good performance when learning a classification model for only one PPA class at a time. Furthermore, multiclass classification is in general harder than binary classification. Since the problem under study decreases its complexity, it should be easier to learn the relationships hidden in the data and create more efficient models. Moreover, with few learning examples, as in this case, the multiclass learning is even harder.

Table 2. Proportion of PPA variants for each binary dataset.

class/¬ class	lvPPA (L)		nfvPPA (A)		svPPA (S)	
	L	¬ L	A	¬ A	S	¬ S
All classified patients	30%	70%	34%	66%	37%	63%
Model patients	39%	61%	28%	72%	33%	67%

Finally, we integrated the results from the unsupervised study within a supervised learning approach: we ran the classifiers in the set of patients composing the emergent clusters obtained by the clustering algorithms (*k*=2, 3, 4). This analysis was motivated by our interest in inferring whether the classifiers' performance would improve when a previous division of patients was already made, and thus, some noise could have been removed. Additionally, the classifiers were applied in the entire set of patients, using as target class the clusters' labels (instead of the PPA classes), to test the meaning of the obtained clusters. Once again, such analyses were performed for both the set of model patients (36 instances) and the entire set of classified patients (104 instances) as well as the distinct set of features (section 2.1.2).

In order to assess the performance of each classifier and compare it with that of the other classifiers, we evaluated the following metrics: prediction accuracy, sensitivity and specificity. Confusion matrices were also analyzed. We used the stratified *5-fold* [16] and *Leave-one-out* (LOO) [23] cross-validation methods and performed permutation tests.

3. RESULTS AND DISCUSSION

In this section we present and discuss the main results obtained with the unsupervised and supervised analysis. The supervised results are preliminary. Due to space limitation, we present only part of the achieved results.

3.1 Unsupervised learning approach

3.1.1 EM and K-Means (X-Means)

Clustering with standard algorithms (K-Means and EM) and with the predefined number of clusters set as 3 (*k*=3) produced, in general, clusters composed by cases from all the PPA variants, when applied to the original set of attributes and either to the set of classified patients or model patients (Table 3, *first* and *second panel*). In the case of model patients the results are slightly better given the existence of one group with a few and isolated nfvPPA cases (Table 3, *second panel,* cluster $\mathbf{C0_{k3}}$). However, most of the lvPPA and svPPA cases remains inseparable (Table 3, *second panel,* cluster $\mathbf{C1_{k3}}$) as well as the remaining nfvPPA cases which were also grouped with some logopenic and semantic cases (Table 3, *second panel,* cluster $\mathbf{C2_{k3}}$). This means that gold-standard (clinical judgment) was not concordant with the generated clusters. Little improvement was observed when we used different sets of features (Table 3, *third and fourth panel*). Although, in general, we obtained a clearer separation of some nfvPPA cases, a high dispersion of cases from different PPA variants was still present in the clusters. This shows the non-triviality on clustering patients reproducing the criteria of Gorno-Tempini *et. al.* [1], either using the entire population or a restrict group of selected model cases.

Due to the inconsistency between the clusters produced by the algorithms (*k*=3) and the gold-standard, we performed a detailed evaluation of the emergent clusters with different number of predefined clusters (*k*=2, 4, 5). We aimed to find a pattern in the distribution of PPA cases along the clusters and/or to find the real number of clusters detected by this study. We noticed that the composition of the clusters generated with *k*=2 was very consistent along all the analysis: with different clustering algorithms, set of features and set of patients (Table 4). In general, one of those clusters comprised mainly nfvPPA cases (or a majority of nfvPPA and some lvPPA cases) whereas the other was mostly composed by svPPA and lvPPA cases. Basically, most of the nfvPPA cases were well separated by clustering the dataset with *k*=2, meaning that this PPA variant must be an easier class to individualize. We noticed that, when the entire set of patients is used, the nfvPPA cases which were isolated in the smaller cluster are mainly composed by model cases. Thus, the unsupervised learning algorithms were able to identify the typical group of agrammatic patients. In the case of model patients, we can also say that agrammatic and semantic cases are easily separated in two different groups while lopopenic cases may be found on both groups (although most are more frequently grouped with semantic then with agrammatic cases).

When the predefined number of clusters was higher than two (*k*= 3, 4 and 5), although the proportion of PPA cases in each cluster was not as coherent across all the analysis as it was with

k=2, it was possible to compile some conclusions. As aforementioned, in general, when *k* was increased to 3, the logopenic and semantic variants remained inseparable, being most of those PPA cases grouped. On the other hand, a smaller group composed mainly by nfvPPA cases emerged. The increase in the *k* value, to 4 and 5, in the case of model patients, allowed the formation of a smaller group including almost uniquely svPPA cases. The remaining emergent clusters comprised a mixture of patients from all the PPA classes, excepting one small mainly agrammatic group. When all classified patients are used, it is rare to find a cluster of isolated svPPA cases (even for large values of *k)*, given that many svPPA and lvPPA cases remained attached. Moreover, even for large values of *k*, no group comprised almost exclusively logopenic cases, since lvPPA cases were spread over the remaining groups. This was verified both in the entire population (classified patients) and model patients. In brief, for larger values of *k,* three groups were usually found: one group including many lvPPA and svPPA cases, one cohort of nfvPPA and lvPPA cases and a smaller cluster with almost uniquely nfvPPA cases.

As aforementioned, clustering in two groups (*k*=2) produces two cohorts: one mainly agrammatic and other mainly semantic and logopenic (Table 4). Concerning the analysis where we clustered the dataset with *k*=2 and then, we ran the clustering algorithms in the set of patients which composed each group, we were especially interested in the division of the cluster of the group that included many lvPPA and svPPA cases. In fact, we wanted to test if svPPA and lvPPA cases were more easily separable in the absence of most of the nfvPPA cases. We observed that, in the case model patients, in general, most of the svPPA cases were practically isolated on a group while the remaining lvPPA cases along with some nvfPPA and svPPA cases constituted the other group. Consequently, svPPA cases were easily separable from the lvPPA cases in the absence of the nfvPPA. Notwithstanding, the parallel study with the entire set of patients revealed that, although for large values of *k*, some small groups with mainly semantic or, less frequently, logopenic cases emerged, there were also larger groups including many svPPA and lvPPA cases. Therefore, in this case, removing most of nfvPPA cases does not solve the difficulty of separating svPPA and lvPPA cases.

By running the algorithms in datasets containing only two PPA variants each, we observed that svPPA and nfvPPA cases are well separated in distinct clusters. Clustering the datasets with lvPPA cases produces groups with a mixture of either lvPPA and nfvPPA or lvPPA and svPPA cases. The logopenic variant represents the most difficult class to individualize, being harder to separate it from the semantic then from the agrammatic variant.

It is well known that finding the ideal number of clusters is a difficult task in unsupervised learning. In this context, we ran different sets of attributes with different features, the X-Means algorithm (able to estimate the number of clusters). This algorithm produced successively two clusters in all the datasets, meaning that *k*=2 was the ideal number of clusters found by this algorithm.

3.1.2 Hierarchical Clustering

The outcome of hierarchical clustering revealed the formation of two main groups, either using the dataset with model patients (Figure 2 - *left*) or with all classified patients (Figure 2 - right). As in the previous analysis, although both groups tended to be composed by a mixture of patients from all variants, one of the groups was composed mainly by svPPA and lvPPA cases [Figure 2 – *left* (*left cluster*) - Table 5, *first panel*; and Figure 2 - *right* (*left cluster*), Table 5, *second panel*] whereas the other included most of the nfvPPA and a few cases of the remaining subtypes [Figure 2 - *left* (*right cluster*) - Table 5, *first panel*; and Figure 2 - *right* (*right cluster*), Table 5, *second panel*].

Since almost all lvPPA and svPPA cases were included in the same cluster, we decided to explore its corresponding sub-clusters (at a lower level of the dendogram), in order to inspect whether those sub-clusters were composed by patients coming from only one PPA variant. Therefore, a new cut to the dendogram of Figure 2 (*right*), specifically to cluster 2, at a distance d'=0.28, was done. The emergent sub-clusters did not allow a clear separation of lvPPA and svPPA. Indeed, sub-cluster 1.1 comprised 14 logopenic, 5 agrammatic and 19 semantic patients while sub-cluster 1.2 included 15 logopenic, 7 agrammatic and 8 semantic cases. The remaining 3 lvPPA, 1 nfvPPA and 1 svPPA cases belonged to the two remaining small sub-clusters. Only at a (even more) deeper level of the dendogram, and thus by cutting again the first sub-cluster (d''=0.25), we obtained a group with a majority of semantic patients (7 lvPPA, 4 nfvPPA and 19 svPPA). Similar results were achieved when we particularized the study to the set of model patients. We cut the dendogram of Figure 2 (*left*) at a distance of (*d'*=0.35). We observed that sub-cluster 1.1 aggregated 9 svPPA patients and only 3 lvPPA cases. The sub-cluster 1.2 included 5 logopenic and 3 semantic cases. Therefore, at a lower level of the dendogram, a cohort with a majority of semantic cases emerged. Logopenic patients were, again, dispersed in different groups.

In a word, most of the patients from the agrammatic subtype were easily separated from the remaining PPA variants. On the contrary, svPPA and lvPPA cases were only considerably separable at a deep level of the dendogram. In addition, the results obtained with this algorithm did not evidence the three groups of patients distributed according to their predefined PPA classes.

3.1.3 Consensus Clustering

The results obtained through consensus clustering corroborate the outcome of the previous analysis. Indeed, when *k* (global) is set as 3, the emergent clusters did not allow a clear separation of the PPA variants according to their predefined classification, despite the fact that, usually, one of these clusters was composed mainly by patients from a unique PPA variant (typically nfvPPA). When *k* (global) was set as 2, the composition of the two emergent clusters was similar to the ones obtained with previous analysis (one of the groups included mainly lvPPA and svPPA cases while the other comprised mainly nfvPPA and some lvPPA cases). In general, most of the lvPPA and svPPA cases remained inseparable when *k* (global) was increased to 3. However, a large part of the analysis with *k* (global) higher than 3, showed the emergence of clusters composed by a majority of cases from each of the PPA variants (more frequently nfvPPA and svPPA), being the remaining clusters represented by intersections from various subtypes (data not shown).

3.2 Supervised learning approach

As described in the section 3.1, the unsupervised analysis did not allow a clear distinction of the three-group diagnostic model for PPA. Notwithstanding, we decided to perform a supervised analysis to infer whether a classification model was able to classify correctly patients in one of the three PPA classes.

Several standard classifiers (available in WEKA®) were run in datasets composed either by the entire set of classified patients or by model patients and by different set of features. Table 6 show the results of prediction accuracy, sensitivity and specificity obtained for the different set of attributes/classifier, in respect to

the entire set of patients and model patients, using LOO. According to these results, we may notice that, some classifiers reached good performances with the set of model patients (maximum accuracy of 0.94 with the SVM and the OCA attributes - high values of sensitivity and specificity). Consequently, for this restrict set of typical PPA patients, the three-group classification diagnosis of PPA is supported by the supervised learning approach. The results of the permutation tests showed that the outcome obtained with these classifiers are not random (example: model patients and OCA attributes; SVM; p-value<0.004). Moreover, in general, we noticed that the sensitivity of the logopenic variant was lower than the sensitivity of the remaining classes (on average, sensitivity of lvPPA cases was 0.60±0.19 against 0.88±0.09 and 0.80±0.09 for the agrammatic and semantic classes, respectively). Therefore, the lower accuracy evidenced in some classifiers/set of attributes was probably due to the difficulty in learning a model that classifies these instances accurately [example: Naïve Bayes and set of model attributes; accuracy of 0.72; sensitivity: 0.5 (lvPPA), 0.79 (nfvPPA) and 0.83 (svPPA)]. The specificity values were also good, being the higher values associated with the agrammatic variant [specificity (lvPPA)=0.86±0.07; specificity (nfvPPA)=0.91±0.18; specificity (svPPA)=0.87±0.06, on average]. On the other hand, the performances obtained with the entire set of classified patients were much lower (average accuracy of 0.59±0.07, against 0.78±0.09 for the model patients; maximum accuracy reached was 0.70 for the set of model attributes and with the SVM classifier). Sensitivity was also low for all PPA classes [sensitivity (lvPPA)=0.59±0.09; sensitivity (nfvPPA)=0.59±0.05; sensitivity (svPPA)=0.58±0.11, on average]. However, in this case, the sensitivity of the logopenic variant was not consistently the lowest, as happened with the set of model attributes. Regarding specificity, the agrammatic variant reached again the highest values [specificity (lvPPA)=0.67±0.09; specificity (nfvPPA)= 0.91±0.05; specificity (svPPA)=0.78±0.05, on average]. A further examination of the set of correctly classified agrammatic cases showed that, it included most of the agrammatic model patients. This means that, similarly to unsupervised learning, supervised algorithms classified correctly a restricted group (mostly model cases) from the overall group of agrammatic patients.

We also conducted some experiments where we used the set of model patients to train the classifier, and then, tested with the non-model patients. This strategy did not improve the results (similar to those depicted in Table 6). This fact (and the discrepancy between the results obtained with the model and entire set of patients) is probably due to two possible scenarios: 1) the set of model patients comprises a low number of instances, which may be insufficient to obtain a sufficiently generalist classification model (we note that this group is composed only by 36 instances and we intended to solve a multiclass classification problem); 2) the set of model patients might not be representative of the general population (classified patients who are not considered as typical PPA cases). In fact, model patients correspond to cases where clinicians were able to reach, with high confidence, a subtype according to Gorno-Tempini *et. al*. [1] criteria.

3.2.1 Conversion to a Binary Classification Problem

Motivated by the previous results and, in order to check if we would be able to classify accurately at least one of the PPA variants, we decided to convert the original ternary classification problem into a binary one. More specifically, we created three new datasets, in which, one PPA class was maintained (positive class) whereas the other two were merged (negative class) (Table 2). Due to space limitation, Table 7 depicts results regarding only the best outcome (regarding the corresponding accuracy, sensitivity, specificity and other metrics derived from the confusion matrix). Comparing the accuracies obtained concerning each of the three PPA classes, we noticed that nfvPPA was the class that reached better performances, followed by svPPA and finally lvPPA, in both set of patients. Regarding sensitivity and specificity, in general, we did not have good results simultaneously on both. This was more evident when all classified patients were used. In fact, we were especially interested in experimenting this binary classification strategy with the entire set of classified patients (n=104) given the low performances obtained with the ternary classifiers for this dataset. Comparing the best outcome for the ternary classification approach (Table 6, entire set of patients; model attributes; SVM classifier) with the results of Table 7, we observed that, although some improvements were achieved in one of the performance metrics (individually) at least with one set of attributes/classifier (for all the three PPA classes) on average, considering the combination of both sensitivity and specificity values, the results did not improve much. For instance, running the SVM classifier with the set of model attributes and entire set of classified patients in the "dataset A/¬A", produced a maximum accuracy of 0.87 (against the 0.70 of the best ternary outcome). However, the sensitivity (0.65 against the 0.61 of the best ternary outcome) and the specificity (0.96 against 0.97 of the best ternary outcome) did not improved much. Notwithstanding, we believe that these results might be enhanced with an oversampling method (such as SMOTE) since data is imbalanced (Table 2). On the other hand, the low sensitivity and high specificity values showed that, despite the fact that the classifier could not recognize all the nfvPPA cases as being in fact from this class, there were only a few non-agrammatic cases incorrectly classified as agrammatic. This was already true for the ternary analysis. A further examination of the results showed that, agrammatic model patients, were correctly classified as being from the agrammatic variant. This means that, the "A/¬A classifier" was able to find some classification rules which classify, with good performance, the typical agrammatic cases (even when they are mixed with non-model cases). This did not happened neither for the semantic nor logopenic PPA variants. This may be useful for designing new approaches for PPA classification, as discussed in section 4.

The results of the ternary classifier were already good in the model patients. Only the agrammatic and semantic variants had a higher performance with the binary classifier (Table 7). The agrammatic variant showed, again, the highest performance on binary classification. This suggests that nfvPPA cases are easily distinguishable from the remaining PPA variants and is thus easier to define an accurate classifier for this variant. Although the logopenic variant presents a lower performance (in relation to the other variants), by comparing these results with those of the ternary classifier, we noticed an improvement both in its sensitivity and (in some cases) in its specificity.

3.2.2 Integrating Clustering Results in Classification

We decided to test the performance of PPA classification using the groups obtained with unsupervised learning. Since the classifiers did not perform well in the entire population, we inferred whether the results would improve by classifying patients that were previously separated in clusters, and thus, removing some noise. Once again, we were especially interested to perform the study using the entire set of classified patients.

As shown in section 3.1, when k=2, one of the emergent clusters was composed mainly by nfvPPA cases whereas the other included most of the lvPPA and svPPA cases. When we classified the set of patients which composed the "mainly agrammatic" cluster, we obtained good performances for the agrammatic variant, in terms of sensitivity (between 0.85 and 1.00) and specificity (between 0.74 and 0.933). On the other hand, many lvPPA and svPPA cases were misclassified. This was expected due to the low number of instances from these variants in that cluster. Regarding the classification outcome of the "mainly logopenic and semantic" cluster, the values of sensitivity and specificity from those classes were similar to the ones reached when classifying the entire set of patients (without dividing the patients by clustering). This fact evidenced that, the difficulty on classifying accurately lvPPA and svPPA cases was not diminished by removing nfvPPA cases. When k was increased to 3 or 4, some smaller clusters (of variable composition) arose. Those small groups were frequently composed either by nfvPPA cases, or by svPPA cases or even by a mixture of two or three PPA variants. By running the classifiers in the latter clusters, we observed that, usually only the nfvPPA cases achieved good values of sensitivity and specificity, being evident a high percentage of svPPA misclassified as lvPPA and vice versa. When smaller groups were composed by almost uniquely one PPA variant, that variant was successfully classified while the other variants had many instances misclassified due to its low cardinality. In addition, in the clusters composed by lvPPA and svPPA cases, performance was similar when no patients' subdivision was made before running the classifiers. The classification models developed were effective in classifying nfvPPA cases, when the clusters were composed mainly by that type of patients or by a mixture from all PPA variants (with approximately same proportion of PPA cases). But, many svPPA were misclassified as lvPPA cases and conversely, in most clusters. Comparing the performance of the classifiers either running the algorithm in the entire set of patients (Table 6) or in groups obtained by clustering, we concluded that the sensitivity and specificity values were similar on both analysis, with little improvement in the agrammatic PPA variant.

Concerning the analogous analysis with binary classifiers instead of multiclass, we observed again that the performance related with svPPA and lvPPA cases did not improve, comparing with the results using the entire set of patients together (Table 6). On the other hand, the results regarding the sensitivity and specificity of the agrammatic variant improved, compared with the analysis that uses all cases together, both in the clusters constituted mainly by nfvPPA cases and in the clusters with a mixture (approximately in the same proportion) of cases from all the PPA variants. Good performances were achieved with the set of model cases (n=36), but the result did not enhance the one obtained with the entire set of patients (not divided by clustering), which was already good.

We also performed a different study to test the significance of the clustering obtained. With this purpose, we ran the classifiers having as target class the clustering labels (k =2, 3, 4) instead of the PPA class. Good performances were obtained in all metrics, simultaneously. A maximum accuracy of 0.92 was achieved with the entire set of classified patients and language attributes with three classes (k=3; sensitivity of 0.90, 0.93 and 0.95 and specificity of 0.95, 0.98 and 0.95 for each class). In the case of model patients, the maximum accuracy (0.97) was reached with the set of model patients and a binary classifier (k=2; sensitivity of 1.00, 0.96 and specificity of 0.96, 1.00, for each class). The good performances showed that the groups obtained with clustering make sense. Moreover, with the set of model patients, the best outcome was consecutively achieved with k=2, which corroborates the conclusions reached with unsupervised learning. With the entire set of patients, the worst of the best results was with k=4 (k=2 or 3 were equally satisfactory).

In a word, the three-group diagnostic model for PPA were not supported through a supervised learning analysis when the entire set of (clinically) classified patients was used. However, when only the restrict set of model cases is used, the algorithms are able to classify correctly a large proportion of patients. This positive results may be due to the small number of cases in study.

4. CONCLUSIONS AND FUTURE WORK

This study aimed to use unsupervised and supervised learning in a clinical series of PPA patients to test the existence of three PPA variants (according to recent published criteria).

The results of unsupervised learning did not support the true existence of three PPA variants. In fact, using k=3 revealed that the composition of the groups obtained and the gold-standard did not match, meaning that clustering algorithms were unable to detect the three PPA variants defined in the literature, even with cases with high confidence in diagnosis (model patients). We tested different values of predefined number of clusters to test how many distinct groups of patients were present in the clinical series. Results with clustering techniques consistently revealed the emergence of two main groups that stayed practically inalterable independently of the algorithms used and either using quantitative or qualitative sets of attributes. One of these groups was composed mainly by nfvPPA (and a few lvPPA) cases whereas the other included mostly svPPA and lvPPA cases. The agrammatic variant was the most easily defined PPA class while lvPPA was undoubtedly the most difficult class to individualize. These results support previous prospective data-driven studies, where the true existence of three PPA variants defined by Gorno-Tempini has been questioned [4]–[10]. Regarding the supervised learning analysis, good performances were obtained when classifying the set of model patients into one of the three PPA variants. However, the same analysis using the entire set of classified patients achieved low performances and thus, did not support the three-group classification diagnosis, either following a ternary or a binary classification strategy. Notwithstanding, we have evidenced a superiority in the performance of the nfvPPA cases. In particular, the classification model defined for this variant has a smaller percentage of misclassified patients, compared to that of the models for other PPA variants, and correctly classified a large proportion of agrammatic cases (including most of the model agrammatic patients). In our opinion, the discrepancy between the outcome obtained with the set of model and classified patients may be justified by two scenarios: 1) the number of model patients is too small to build a general model; 2) model patients (typical from each PPA variant) are not representative of the global population.

As future work, we would like to repeat the analysis with a larger dataset and test it with an independent dataset. In addition, we would like to perform an automatic (unsupervised and supervised) feature selection and replicate the analysis. We also plan to apply semi-supervised learning. Since, according to the results obtained in this study, it is easier to classify, or even cluster, the nfvPPA cases, we would like to develop a novel classification approach, based on a pipeline in which patients would be classified following different steps. In the first step we would classify patients as agrammatic or not, and the following task would differentiate the two remaining PPA classes (lvPPA and svPPA). Moreover, since our study also showed that, the

accurate classification of PPA cases is not trivial, instead of applying standard classifiers, we plan to use a strategy based on meta-classifiers. Furthermore, due to the possible existence of more variants (not yet defined in the literature, which might correspond to intersections between variants), patients should only be assigned to a PPA class if the prediction probability of the meta-classifier is higher than a given threshold. As such, a set of patients would be classified as "unclassifiable", and would be submitted for further clinical analysis to check their putative belonging to a non standard PPA variant.

5. ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), under projects Pest-OE/EEI/LA/0021/2014 and DataStorm (EXCL/EEI-ESS/0257/2012), and individuals doctoral grants to T.P. (SFRH/BD/95846/2013) and C.M. (SFRH/BD/75710/2011), and funding to A.d.M. and M.G..

6. REFERENCES

[1] M. L. Gorno-Tempini, a E. Hillis, S. Weintraub, a Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve, F. Manes, N. F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B. L. Miller, D. S. Knopman, J. R. Hodges, M. M. Mesulam, and M. Grossman, "Classification of primary progressive aphasia and its variants.," *Neurology*, vol. 76, no. 11, pp. 1006–1014, 2011.

[2] M. Mesulam, "Primary progressive aphasia," vol. 49, no. 4, pp. 425–432, Apr. 2001.

[3] M. L. Gorno-Tempini, N. F. Dronkers, K. P. Rankin, J. M. Ogar, L. Phengrasamy, H. J. Rosen, J. K. Johnson, M. W. Weiner, and B. L. Miller, "Cognition and anatomy in three variants of primary progressive aphasia.," *Ann Neurol*, vol. 55, no. 3, pp. 335–46, Mar. 2004.

[4] J. M. Harris, C. Gall, J. C. Thompson, A. M. T. Richardson, D. Neary, D. du Plessis, P. Pal, D. M. A. Mann, J. S. Snowden, and M. Jones, "Classification and pathology of primary progressive aphasia.," *Neurology*, vol. 81, no. 21, pp. 1832–9, Nov. 2013.

[5] M. R. Wicklund, J. R. Duffy, E. a Strand, M. M. Machulda, J. L. Whitwell, and K. a Josephs, "Quantitative application of the primary progressive aphasia consensus criteria.," *Neurology*, vol. 82, no. 13, pp. 1119–26, Apr. 2014.

[6] K. A. Josephs, J. R. Duffy, E. A. Strand, M. M. Machulda, M. L. Senjem, V. J. Lowe, C. R. Jack, and J. L. Whitwell, "Syndromes dominated by apraxia of speech show distinct characteristics from agrammatic PPA," *Neurology*, vol. 81, no. 4, pp. 337–45, 2013.

[7] M. M. Machulda, J. L. Whitwell, J. R. Duffy, E. A. Strand, P. M. Dean, M. L. Senjem, C. R. Jack, and K. A. Josephs, "Identification of an atypical variant of logopenic progressive aphasia.," *Brain Lang.*, Apr. 2013.

[8] M. Mesulam, C. Wieneke, E. Rogalski, D. Cobia, C. Thompson, and S. Weintraub, "Quantitative Template for Subtyping Primary Progressive Aphasia," *Arch Neurol*, vol. 66, no. 12, pp. 1545–1551, 2009.

[9] M.-M. Mesulam, C. Wieneke, C. Thompson, E. Rogalski, and S. Weintraub, "Quantitative classification of primary progressive aphasia at early and mild impairment stages.," *Brain*, vol. 135, no. Pt 5, pp. 1537–53, May 2012.

[10] S. A. Sajjadi, K. Patterson, and P. J. Nestor, "Logopenic, mixed, or Alzheimer-related aphasia?," *Neurology*, vol. 82, no. 13, pp. 1127–31, Apr. 2014.

[11] W. T. Hu, C. McMillian, D. J. Libon, and E. Al., "Multimodal predictors for Alzheimer disease in nonfluent primary progressive aphasia," *Neurology*, vol. 75, pp. 595–602, 2010.

[12] C. E. Leyton, V. L. Villemagne, S. Savage, K. E. Pike, K. J. Ballard, O. Piguet, J. R. Burrell, C. C. Rowe, and J. R. Hodges, "Subtypes of progressive aphasia: application of the international consensus criteria and validation using b-amyloid imaging," *Brain*, vol. 134, pp. 3030–3043, 2011.

[13] R. Casanova, F.-C. Hsu, K. M. Sink, S. R. Rapp, J. D. Williamson, S. M. Resnick, and M. a Espeland, "Alzheimer's disease risk assessment using large-scale machine learning methods.," *PLoS One*, vol. 8, no. 11, p. e77949, Jan. 2013.

[14] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Res Notes*, vol. 4, p. 229, 2011.

[15] L. Lemos, D. Silva, M. Guerreiro, I. Santana, A. De Mendonça, P. Tomás, and S. Madeira, "Discriminating Alzheimer's disease from mild cognitive impairment using neuropsychological data," *HI-KDD*, 2012.

[16] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*, First Edit. Italy: John Wiley & Sons Ltd., 2009, p. 417.

[17] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software : An Update," vol. 11, no. 1, pp. 10–18.

[18] D. Pellge and A. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Seventeenth International Conference on Machine Learning*, 2000, pp. 727–734.

[19] S. V. Pons and J. R. Shulcloper, "A survey of clustering ensemble algorithms," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 3, pp. 337–372, 2011.

[20] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, Second. Morgan Kaufmann, 2006, p. 743.

[21] J. C. Platt, *Fast training of support vector machines using sequential minimal optimization*. 1998.

[22] Z. Zheng, G. I. Webb, and K. M. Ting, "Lazy bayesian rules: A lazy semi-naive bayesian learning technique competitive to boosting decision trees," in *16th International Conference on Machine Learning*, 1999.

[23] A. Elisseeff and M. Pontil, "Leave-one-out error and stability of learning algorithms with applications," 2002, pp. 1–15.

Table 3. Clustering results obtained with EM and K-Means (*k*=3) and the datasets composed either by model patients or all classified patients and different set of attributes. Figures in each cell represent number of cases and C denominates the cluster.

		EM (k=3)			**EM (k=3)**			**EM (k=3)**			**K-Means (k=3)**		
Patients		*Classified*			*Model*			*Classified*			*Model*		
Attributes		*Original*			*Original*			*Language*			*Model*		
		$C0_{k3}$	**$C1_{k3}$**	**$C2_{k3}$**	**$C0_{k3}$**	**$C1_{k3}$**	**$C2_{k3}$**	**$C0_{k3}$**	**$C1_{k3}$**	**$C2_{k3}$**	**$C0_{k3}$**	**$C1_{k3}$**	**$C2_{k3}$**
Clinical class	*lvPPA*	6	13	19	0	6	4	17	18	3	5	5	0
	nfvPPA	12	12	7	4	1	9	15	5	11	1	3	10
	svPPA	6	8	21	0	10	2	9	20	6	6	6	0

Table 4. Clustering results obtained with EM (k=2) and the datasets composed either by model patients or all classified patients and different set of attributes. Figures in each cell represent number of cases and C denominates the cluster.

		EM (*k*=2)		EM (*k*=2)		EM (*k*=2)		EM (*k*=2)		EM (*k*=2)		EM (*k*=2)	
Patients		*Model*		*Model*		*Model*		*Model*		*Classified*		*Classified*	
Attributes		*Model*		*Language*		*Original*		*OCA*		*Model*		*Language*	
		$C0_{k2}$	**$C1_{k2}$**	**$C0_{k2}$**	**$C1_{k2}$**	**$C0_{k2}$**	**$C1_{k2}$**	**$C0_{k2}$**	**$C1_{k2}$**	**$C0_{k2}$**	**$C1_{k2}$**	**$C0_{k2}$**	**$C1_{k2}$**
Clinical class	*lvPPA*	0	10	3	7	3	7	6	4	30	8	3	35
	nfvPPA	11	3	12	1	10	4	12	2	13	18	17	14
	svPPA	0	12	0	12	1	11	2	10	28	7	4	31

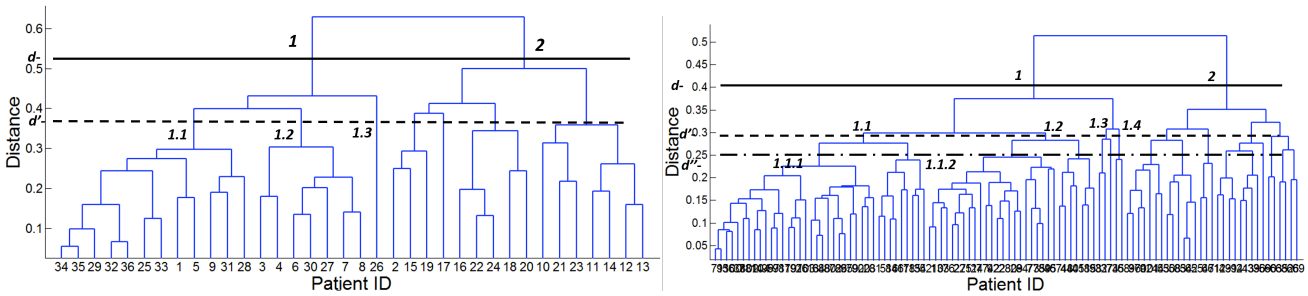

Figure 2. Dendogram obtained by applying hierarchical clustering to: (*left*) all classified patients with the original set of attributes (*right*) model patients and set of model attributes.

Table 5. Results obtained by Hierarchical Clustering (*k*=2) and with: dataset with model patients and the set of model attributes (*first panel*); dataset with classified patients and the original set of attributes (*second panel*).

Set of Patients/Attributes		Model/Model		All classified/Original	
Clusters		$C1_{k2}$	$C2_{k2}$	$C1_{k2}$	$C2_{k2}$
Clinical classification	*lvPPA*	8	2	32	6
	nfvPPA	0	14	13	18
	svPPA	12	0	28	7

Table 6. Classifier's performance (Accuracy, Sensitivity and Specificity) obtained using different set of features and different classifiers (LOO), using all classified patients (values above) or the set of model patients (values below).

	Original set of attributes				***Model attributes***				***Language attributes***				***OCA***			
	NB	**SVM**	**DT**	**RF**	**NB**	**SVM**	**DT**	**RF**	**NB**	**SVM**	**DT**	**RF**	**NB**	**SVM**	**DT**	**RF**
Accuracy	0.59	0.49	0.50	0.54	0.63	0.70	0.59	0.65	0.65	0.59	0.50	0.63	-	-	-	-
	0.67	0.69	0.72	0.58	0.72	0.86	0.78	0.81	0.75	0.75	0.81	0.78	0.92	0.94	0.86	0.83
Sensitivity	0.53	0.50	0.42	0.58	0.68	0.68	0.55	0.66	0.68	0.63	0.50	0.74	-	-	-	-
(lvPPA)	0.30	0.60	0.50	0.30	0.5	0.8	0.5	0.6	0.4	0.70	0.6	0.5	0.90	0.90	0.80	0.80
Sensitivity	0.65	0.55	0.58	0.45	0.58	0.61	0.67	0.61	0.58	0.55	0.61	0.58	-	-	-	-
(nfvPPA)	0.79	0.71	0.93	0.71	0.79	0.93	0.93	0.93	0.93	0.71	0.93	0.93	0.93	0.93	1.00	1.00
Sensitivity	0.60	0.43	0.51	0.57	0.60	0.80	0.57	0.69	0.69	0.57	0.40	0.57	-	-	-	-
(svPPA)	0.67	0.75	0.67	0.67	0.83	0.83	0.83	0.83	0.93	0.83	0.83	0.83	0.92	0.96	0,75	0.67
Specificity	0.71	0.64	0.67	0.67	0.71	0.80	0.71	0.69	0.73	0.37	0.61	0.67	-	-	-	-
(lvPPA)	0.85	0.77	0.81	0.67	0.85	0.89	0.93	0.85	0.89	0.81	0.89	0.89	0.96	0.96	0.89	0.85
Specificity	0.88	0.82	0.90	0.89	0.88	0.97	0.92	0.97	0.95	0.89	0.85	0.95	-	-	-	-
(nfvPPA)	0.86	0.91	0.96	0.23	0.86	1.00	0.96	0.96	0.91	0.96	0.96	1.00	0.96	1.00	1.00	0.96
Specificity	0.78	0.77	0.67	0.73	0.84	0.77	0.75	0.79	0.79	0.84	0.78	0.83	-	-	-	-
(svPPA)	0.79	0.88	0.83	0.80	0.88	0.92	0.79	0.88	0.83	0.88	0.88	0.79	0.96	0.96	0.92	0.96

Table 7. Classifier's performance (Accuracy, Sensitivity and Specificity) obtained to the binary datasets (described at Table 2) with different set of features and different classifiers (LOO), using all classified patients (values above) or the set of model patients (values below). Note: A - agrammatic variant; S - semantic variant; L - logopenic variant.

	Original set of attributes			***Model attributes***			***Language attributes***			***OCA***		
	L/¬L	A/¬A	S/¬S	L/¬L	A/¬A	S/¬S	L/¬L	A/¬A	S/¬S	L/¬L	A/¬A	S/¬S
	SVM	DT	NB	SVM	SVM	SVM	DT	DT	NB	NB	DT	SVM
Accuracy	0.62	0.82	0.70	0.76	0.87	0.76	0.63	0.84	0.69	-	-	-
	0.81	0.94	0.86	0.89	1.00	0.86	0.81	0.94	0.86	0.86	1.00	0.97
Sensitivity (class)	0.71	0.68	0.74	0.66	0.65	0.74	0.74	0.65	0.71	-	-	-
	0.50	0.93	0.92	0.80	1.00	0.83	0.80	0.93	0.92	0.70	1.00	1.00
Sensitivity (¬class)	0.56	0.88	0.68	0.82	0.96	0.77	0.56	0.92	0.68	-	-	-
	0.92	0.96	0.83	0.92	1.00	0.88	0.81	0.96	0.83	0.92	1.00	0.96
Specificity (class)	0.56	0.88	0.68	0.82	0.96	0.77	0.58	0.92	0.68	-	-	-
	0.92	0.96	0.83	0.92	1.00	0.89	0.81	0.96	0.83	0.92	1.00	0.96
Specificity (¬class)	0.71	0.68	0.74	0.66	0.65	0.74	0.74	0.65	0.71	-	-	-
	0.50	0.93	0.92	0.80	1.00	0.83	0.80	0.93	0.92	0.70	1.00	1.00