Predicting conversion from Mild Cognitive Impairment to Alzheimers Disease using ensemble learning

Ana Rita da Costa Pereira
ana.costa.pereira@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2016

Abstract

Alzheimer's disease (AD) is one of the most frequent types of dementia. It is characterized by the progressive deterioration of cognitive functions. Mild Cognitive Impairment (MCI) represents the transitional period between normal ageing and dementia. Currently, there is no cure for Alzheimer's, but it is believed that early diagnosis and treatment can slow down its progression. In order to study individuals with this disease, machine learning methods have became widely used due to their good results in diagnosis and prognosis. This project consists in the proposal of an ensemble learning approach to improve the performance of classifiers in predicting the conversion from MCI to AD. Instead of building a single classifier, we will train multiple individual classifiers and combine them to achieve a more accurate and robust classification. Two approaches will be followed: one the prediction of the prognosis is made considering all patients as similar, and other the prediction is made considering different groups of patients, evolving differently in the prognosis. In both approaches temporal windows will be considered in the classification process. The patients involved in this work were, firstly, diagnosed using neuropsychological tests.
Keywords: Alzheimer's Disease, ensemble learning, prognosis, temporal windows, clustering

1. Introduction

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder that leads to progressive loss of memory and cognition function, and it is the most common cause of dementia in the elderly [17]. Mild Cognitive Impairment (MCI) is a transitional state from normal ageing to dementia, and when associated with memory loss, it is believed to be precursor of AD [9]. Currently there is no cure for the AD due to its complexity and, despite the numerous researches about this issue, it is hard to know its first causes and the precise reasons for its progression. Current studies try to predict the likelihood of MCI patients developing AD, which can enrich clinical trials of disease-modifying therapies, that aim to slow or prevent AD. Moreover, the sooner this studies are conducted, using people with possible symptoms of this disease, the sooner the disease can be detected and/or prevented.

In order to have a better understanding of AD, some approaches using neuroimaging biomarkers, along with machine learning algorithms, have became popular, such as Single-photon Emission Computed Tomography (SPECT) [15], structural Magnetic Resonance Imaging (MRI) [8] and Position Emission Tomography (PET images) [3, 7]. This neuroimaging biomarkers provide insights into the disease biology, as well as staging and prognosis. These classification methods consist mainly in: 1) Extraction and selection of discriminative features from the original neuroimaging data, and 2) create a learning model that can assign a class label into a subject, in a high dimensional feature space. In (1), the extraction and selection of the most discriminative features is achieved: by removing the redundant ones, and choosing only the features that play an important role in classification. The most used techniques, to solve this problem, are PCA (Principal Components Analysis) [10] or selecting the best ranking features according to some criteria such as t-test [6], Mutual Information [3] and random sampling [4]. In the learning phase (2), supervised learning is applied for the classification. The most used classification algorithm is Support Vector Machines (SVM) due to its good performance classifying in high dimensional data [3, 8].

In our case, it makes sense to understand more about AD by looking at other possible factors, such as neuropsychological, to explain the appearance and progression of the disease. Considering neuropsychological tests, we decided to take a different approach that will not be a classification of individ-

uals but an analysis of the conversion of patients from MCI to AD using supervised models. People diagnosed with MCI can be divided into two subgroups: people with MCI who convert to AD (MCI converters, Evol), and people with MCI who do not convert to AD (MCI non-converters, noEvol). As such, our proposal consists in the prediction of the progression of patients with MCI to AD by looking at their first and their last medical appointment, according to some temporal windows. The considered temporal windows are divided in groups of two, three, four and five years. This proposal will be divided into two approaches: first, all patients are assumed to evolve similarly and in the second, the patients profiles are considered, or using unsupervised learning methods or by considering important characteristics of the patients. In both approaches, the classification process, is conducted using supervised learning methods. We decided not to use simple methods, as SVM, but ensemble methods. The idea, of using these methods, instead of the others, is to avoid possible overfitting problems and to achieve a good generalization, since these methods combine a classification made by several classifiers. Nowadays, ensemble methods as Bagging and Boosting are still very used with clinical data, thus we will explore these methods and even more [13, 18].

The dataset used was provided by the Dementia Group at IMM. The data was collected from individuals evaluated in Lisbon, with a total number of patients of 616 and a number of evaluations of 1604.

2. Background

2.1. Alzheimer's Disease and Mild Cognitive Impairment

Alzheimer's disease (AD) is the most common form of dementia worldwide, among elders, and the number of affected patients is expected to double in the next 20 years [19]. This type of dementia causes memory, thinking and behavior problems. Symptoms usually develop slowly and get worse over time, becoming severe enough to interfere with daily tasks. People even lose the ability to carry on a conversation and to respond to their environment. Even though AD does not take part on normal aging, the majority of people with this disease are indeed 65 and older, which lead studies to conclude that the greatest known risk factor of it, is actually age increase. But not all cases of AD appearance is in older ages, it also happens when someone is in between their 40's and 50's, this phenomenon is called younger-onset. It is extremely rare, just about 5% of the people that have the disease have symptoms at an early age. The longevity with the disease is around eight years after the symptoms start to become noticeable to others, but survival can range from four to twenty years, depending on age, health conditions, and the time passed since people have got diagnosed [1]. In the past years, the early clinical signs of AD have been extensively investigated, leading to the concept of Mild Cognitive Impairment (MCI). Patients with MCI have cognitive deficits but are capable of independent living. The diagnosis of MCI is established by (1) evidence of memory impairment, (2) preservation of general cognitive and functional abilities, and (3) absence of diagnosed dementia. Due to doctors' necessity to share a standard way of analysing their patients, a criteria named Clinical Dementia Rating (CDR) was created. In this case, MCI is staged clinically at the level 0.5 and AD at level 1. To classify patients in this scale is necessary to run tests. Those tests are called biomarkers, which can be neuroimaging techniques, biological tests or genetic tests [19].

We decide to use neuropsychological tests. These tests are specifically designed tasks prepared to diagnose, determine the stage, assess and monitor AD, MCI and other dementia. These tests aim to identify and quantify cognitive, functional and behavioral symptom. The most important batteries are: Mini Mental State Examination (MMSE), California Verbal Learning Test (CVLT), and Bateria de Lisboa para Avaliao de Demncia (Lisbon Test Battery for Dementia Evaluation in English) (BLAD). MMSE is the most commonly used test for complaints of memory problems or when a diagnosis of dementia is being considered. It consists of a series of questions and tests, each of which scores points if answered correctly. The MMSE tests a number of different mental abilities, including a person's memory, attention and language [2]. The CVLT battery have tests which are used to assess an individual's verbal memory abilities [5]. BLAD tests have been evaluating multiple cognitive domains and they have been validated for the Portuguese population. This battery includes tests for the following cognitive domains: attention (Cancellation Task); verbal, motor and graphomotor initiatives (Verbal Semantic Fluency, Motor Initiative and Graphomotor Initiative); verbal comprehension (a modified version of the Token Test); verbal and non-verbal abstraction (Interpretation of Proverbs and the Raven Progressive Matrices); visual- constructional abilities (Cube Copy) and executive functions (Clock Draw); calculation (Basic Written Calculation); immediate memory (Digit Span forward); working memory (Digit Span backward); learning and verbal memory (Verbal Paired-associate Learning, Logical Memory and Word Recall). The data used in this work was obtained using MMSE, BLAD, CVLT, Toulouse Piron, Trail Making Test, Geriatric Depression Scale and other subjective tests. Each evaluation of a patient corresponds to an instance identify by the date and the

ID of the patient.

2.2. Feature selection
One of the big challenges of machine learning is identify the set of features that best represent the data that can be used in a classification model for a particular problem. This is a big problem when the dataset is high in dimensionality, for that reason the procedure of Feature Selection was created. This procedure consists in removing irrelevant and/or redundant features from the dataset, in order to allow learning algorithms to operate faster and more effectively, may even improve the accuracy of classification models [10]. The algorithms can be classified into three main categories: filter methods, wrapper methods and embedded methods.

2.3. Unsupervised Learning
Unsupervised Learning is aimed at discovering patterns and affinities in a dataset without an attribute class. Most of the unsupervised learning analyses are performed to discover clusters of instances that are similar within each cluster and different from members of the other clusters [16]. Regarding to the similarities, clusters are constructed taking into account the attributes values of each instance. One of the most used, and applied in this thesis, cluster algorithms is Expectation Maximization (EM).

2.4. Ensemble Learning
Supervised Learning is one of the tasks of machine learning, oriented towards prediction and interpretation with respect to a target attribute [16]. It seeks to find the perfect model that is accomplished by creating a classifier through a learning algorithm and a training data with the respective class labels. Given a testing data, without class labels, the model should predict the classification for each instance. The most common learning techniques are Decision Trees, Naive Bayes, k-Nearest Neighbors and Support Vector Machines.

In contrast with the supervised methods, the ensemble methods construct a set of learners, also known as classifiers, and combine them to classify new data, as shown in Figure 1. The methods of ensemble learning used in this project were Bagging, Boosting, Random Forests, Random SubSpace, Stacking and Vote.

2.5. Class Imbalance
The common understanding in the community is that imbalanced data corresponds to data where number of instances of one class is significantly higher than the number of instances of any other class. This issue causes a poor performance in most of the machine learning algorithms, since they are biased toward the majority class. A possible solution for this problem consists in the modification of an imbalanced dataset, in order to obtain a balanced distribution. One of these examples is SMOTE.

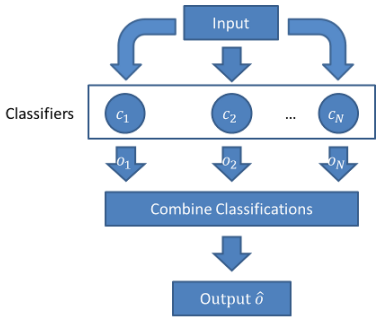

Figure 1: Common ensemble architecture

2.6. Model Validation
The assessment of the effectiveness of the learning algorithms is composed by two parts: create models with the training and testing dataset, and rate the model using the effective metrics, for example accuracy. One of the techniques used to create models is k-fold cross-validation. In this technique, the original data is randomly divided into k subsets of equal size. Of the k samples, a single sample is used for the testing phase, and the remaining (k 1) samples are used as training data. The cross-validation procedure is then repeated k times, with each of the k subsets used exactly once in the testing phase. After this procedure, metrics are applied to do the adequate assessment. Some of those metrics are:

- Accuracy is the proportion of correctly classified instances (both positivity and negativity) among the total number of instances examined.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Sensitivity is the proportion of positive instances that were correctly labeled.

$$Sensitivity = TruePositiveRate = \frac{TP}{TP + FN} \quad (2)$$

- Specificity measures the proportion of negatives instances which are correctly identified as such.

$$Specificity = TrueNegativeRate = \frac{TN}{TN + FP} \quad (3)$$

- F-Measure metric is a trade-off between sensitivity and specificity.

$$F - measure = (\frac{TP + FN}{TP + TN + FP + FN})(\frac{2TP}{2TP + FP + FN}) + (\frac{FP + TN}{TP + TN + FP + FN})(\frac{2TN}{2TN + FP + FN}) \quad (4)$$

- Receiver Operating Characteristics (ROC) space is defined by False Positive Rate (FPR) and True Positive Rate (TPR), as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). FPR corresponds to (1 - specificity), and TPR is sensitivity. In the case of probabilistic classifiers, as Naive Bayes, a ROC curve is produced by a series of points varying the threshold of the continuous numeric values that represent the confidence of an instance belonging to the predicted class. In these cases, the Area Under Curve (AUC) is generally used to provide the average performance of the classifier. When working with non-probabilistic classifiers that return only the predictive class, he AUC is given by (5).

$$\frac{(TPR + TNR)}{2} \tag{5}$$

3. Experimental Methodology

3.1. Data Description

he Dementia Group at Instituto de Medicina Molecular (IMM), in Lisbon, conducted a study, the Cognitive Complaints Cohort (OOC) [11], to investigate cognitive stability and evolution to dementia of subjects with cognitive complaints. The dataset resulted from the application of neuropsychological tests, taking into consideration the inclusion and exclusion criteria specified by Silva [14] and medical doctors decision. The raw dataset used for this experiment, provided by IMM, is composed by 1604 instances, representing the number of evaluations for total patients, taking into consideration that each patient has several evaluations. The patients analyzed were 616 and were characterized as Normal, Pre-MCI, MCI and AD, considering the patient's cognitive condition. The total number of features used was 93, both categorical and numerical.

Data suffered to several steps to become clean, as such: First, the group of pre-MCI subjects were grouped with the MCI subjects, since those subjects only differ in the results of one neuropsychological test. Second, the patients whose diagnostic was unstable, i.e., patients that regress from MCI or Dementia to Normal, were removed. Meanwhile, Normal evaluations were removed since they were not pertinent for the context of the problem. Lastly, evaluations corresponding to a patient with one evaluation were also removed. Data end up reduced to 1355 evaluation, 1134 from MCI patients and 221 from patients with Dementia (AD).

3.2. Tools description

The computational tool used in this thesis was Waikato Environment for Knowledge Analysis (WEKA), version 3.7.0. During all the work, Java programming language was used to unify the different tasks applied in this thesis.

3.3. Creating learning examples

We focused on creating learning examples for temporal windows, so we could predict if patient will ever convert to Dementia in certain period of years. The learning examples are labeled as evolution (Evol) if within the specific temporal window the MCI patient converts to AD, and as no evolution (noEvol) if after the specified temporal window the MCI patient remains MCI. However, there are instances which is not possible to assign a class label: instances with second evaluation categorized as MCI still within the temporal window; instances without knowing the state of the patient in the next evaluation; and instances with only one observation in the specified temporal window. These instances are then removed. These assignments and exceptions are simulated in the Figure 2.

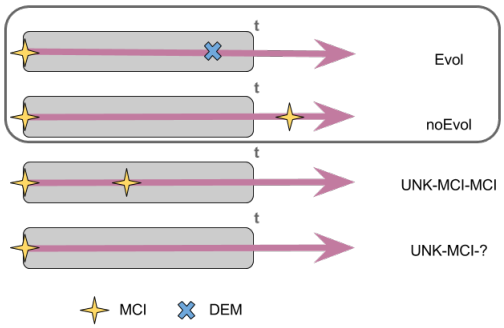

Figure 2: New class labels created for the Temporal Window prognosis problem

3.4. Handling Missing Values

Missing values can correspond about one or more attributes values missing from each instance. o overcome this difficulty we decided to use the WEKAs method for replacing missing values, that uses the modes and means from the training data to address new values to the missing ones. In our initial data, the window of 2 years had 53%, the window of 3 years had 56%, the window of 4 years had 55% and the window of 5 years had 57% of missing values.

3.5. Feature Selection

Attribute Selection was performed for each dataset individually, as each temporal window, using a supervised filter. This kind of filters has two parameters: the type of subset evaluator and the search method. We used 4 versions of feature selection. The first one with correlation based feature subset selection with the BestFirst search, which we called FS1; the second using the Information Gain evalua-

tor with the Ranker searcher, called FS2; the third one, FS3, was with the Pearson Correlation evaluator with, also, the Ranker searcher; and the last one, the FS4 version, was created by us not exactly as new Feature Selection method but as a new way of combining the classification obtained after the application, in a dataset, of the previous three Feature selection methods together, as an ensemble of Feature Selection algorithms.

In Table 3.5, we show the datasets distributions of all four Temporal Windows, after all preprocessing steps in the original dataset.

Total Instances (%)	***Evol***	***NoEvol***
2 Years Temporal Window	86 (28%)	220 (72%)
3 Years Temporal Window	122 (44%)	155 (56%)
4 Years Temporal Window	148 (60%)	97 (40%)
5 Years Temporal Window	169 (73%)	62 (27%)

Table 1: Temporal Window datasets distribution

3.6. Classification

In the methodology followed, six main ensemble classifiers were used: Bagging, Boosting, Random Forest, Random SubSpace, Stacking Generalization and Vote, all implemented in WEKA. For most of these methods were needed base classifiers, for that, we decided to use the most known procedures: Naïve Bayes, SVM kernel, SVM polynomial, K-Nearest Neighbors, C4.5 Decision Tree, Logistic Regression and Decision Stump. Since, most classifiers are highly sensitive to imbalanced distribution in class, the SMOTE, technique was applied.

During the training phase, since the best percentage of SMOTE depends on each problem, subset of features, classifiers parameters and the classifier itself, we decided to use an automated model tool, a grid search. The main idea of this grid search was to determine the best SMOTE percentage, together with the best classifiers parameters, finding an optimal triple Classifier, Parameters, SMOTE percentage.

The metric used to compare the different models, generated by the variance of SMOTE percentages and the parameters' values, was the AUC. This search was preformed in a 3x5-fold CV. After finding the best triple, one for each ensemble classifier, it were tested in a 10x5-fold CV. The automated tool was applied, firstly, to the base classifiers individually, in order to find out the best triple for each one. So that they could be used by the ensemble classifiers in their optimal state. The grid search flow, in the Figure **??** is described as: the dataset, after preprocessing, start by suffering the feature selection and the missing value imputation (optional) phase, and only then the cross-validation is initiated. The 5 folds are created, and after training, the respective best classifier's parameters and SMOTE value are derived. After found the best parameters values and if the MVi and FS procedure are useful to the classifier good results, the data is again trained, in a 10x5 CV, and the performance results are driven.

4. Prediction conversion from MCI to AD

In this research, our goal is to improve the technique of prognosis of conversion to AD, so that clinicians might do an adequate supervision, care and medication, when it appears, to patients. Our proposal is to use a supervised learning analysis, based on ensemble classifiers, in order to find the one which best classifies this dataset. At this stage, we already know, for all temporal windows and for all ensemble classifiers, which are the best SMOTE percentage, the best classifiers' parameters, if MVi procedure should be used and which Feature Selection algorithm should be applied, due to a previous extensive grid search performed.

As baseline of this work, we decided that should be the classification performed by the base classifiers. Thus, in this case, all ensemble methods should outperform the classification produced by this baseline.

All base classifiers were applied to the ensemble classifiers and only one was selected, the one which allowed the ensemble learner to achieve the best performance. After we found the four best ensemble classifiers, in a temporal window, we present an ensemble of experts. For this ensemble, we choose from the four ensemble classifiers, three, that are used to create a classification model with them, using the Vote method. Then, the mixture of experts is compared with the best ensemble classifier, per temporal window.

4.1. Four Years temporal window

A paired t-test, using AUC, revealed that Bagging was the classifier which better performed, from the other five ensemble classifiers. Although, there were no significant difference between the classifiers, Random Forest, Vote, RSS and Naïve Bayes, Bagging outperformed all of them.

Table 2 shows the final metrics values for the 4 best ensemble classifiers. All classifiers had very similar results between them, with a good balancing between sensitivity and specificity.

4.1.1 Ensemble of the best models

The Table 3 shows the final results in the four years temporal window. In this table, it is visible that the mixture of experts did not improve any metric value of Bagging, moreover it revealed even lower results. Bagging is slightly better than the mixture of experts in the 4 years temporal window. The expert is

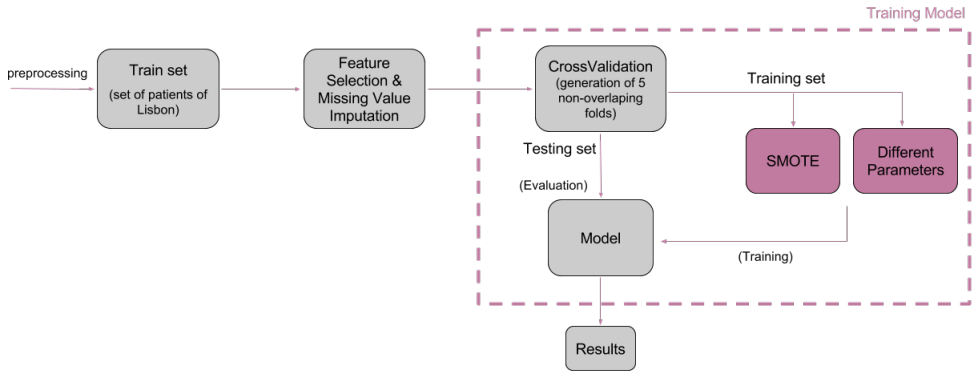

Figure 3: Grid Search Flow on Training Model

Classifier	***Accuracy***	***Sensitivity***	***Specificity***	***AUC***
Bagging	0,83	0,79	0,86	0,88
Random Forest	0,82	0,79	0,83	0,87
Vote	0,82	0,81	0,84	0,88
Naïve Bayes	0,82	0,81	0,83	0,88

Table 2: Evaluation metrics of each ensemble model for four years temporal window

composed by the classifiers: Random Forest, Vote and Bagging.

The mixture of experts was accomplished by making the product of all best ensemble classifiers predictions.

4.2. Summary

In all temporal windows the best results were always obtained by the same classifiers, such as Bagging, Random Forest, Vote and the base classifier Naïve Bayes. The results showed that with the increase of the temporal window, it was noticed a higher prediction capability of the models.

The biggest detachment from sensitivity and specificity it is felt in the window of 2 years, where the results are, also, the lowest, even this temporal window have the biggest number of instances. These conclusions shows the difficulty in the conversion of the disease in just a short time.

In relation with the baseline of this work, we were not perfectly successful since Naïve Bayes had, several times, the best results. However, the good predictability of the Bagging method was noticeable in this project, as a promising classifier in this type of data. In a final phase of this work, we tried to outperform the results obtained by the best ensemble methods at each temporal window, using a mixture of experts. In general, we were not successful since the results showed that the mixture of experts achieved, only, the average of all 4 best ensemble learners.

5. Prediction conversion from MCI to AD based on different MCI characteristics

Nowadays is extensively explored the idea of creating MCI patients profiles based on their symptoms to infer about the development of the disease and its prognosis [12]. For that reason, the purpose of the following studies is to evaluate the performance of the ensemble classifiers taking into account the differences between the MCI patients. So our proposal is to create different groups of patients for evaluation, and it consists in two phases: the first one creating the groups according to clinical criteria, and the other one using clustering methods. We decided that in this classification procedure should only be applied the classifiers which best performance had in the previous chapter, and they were: Bagging with the base classifier Naïve Bayes, Vote with the combination rule product, Random Forest and the base classifier Naïve Bayes. The grid search was, again, applied so that could be found the best triple: SMOTE percentage, classifier and parameters.

5.1. Prognosis prediction based on clinical criteria: depressed/ not depressed

In the first phase of this work, we decided to use the attribute that measures the state of depression of a patient named GDS, Geriatric Depression Scale. We aim to use this attribute, so we can study the influence of depressive symptoms in MCI patients in the prognosis prediction. This approach consists in two steps: in the first one, we execute a clustering procedure where we separate the patients according to their GDS score values. GDS values go

Classifier	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*
Bagging	0,83	0,79	0,86	0,88
Mixture of experts	0.81	0,78	0,84	0,87

Table 3: Comparison between mixture of experts and best ensemble classifier, on 4 years temporal window

from 0 to 14, so we divided in almost half, creating two groups: 0-4 (not depressed patients) and 5-14 (depressed patients). So the clustering step implies fitting the patients in one of the two groups according to their personal GDS score. The final step is the classification, when we can effectively infer something about the impact of depression in two different level groups.

The classification procedure in this problem consists in answer the question: Given that we know the state of depression of a patient, it is better to predict the conversion to AD using a general model containing all patients, or using a specific model considering their state? In other words, we want to find out if there is a higher capability to predict conversion based on the characteristics of the patients instead of assuming the same profile for the entire set of patients, using ensemble classifiers. To do so, we organized this problem into two approaches. Firstly, the datasets used to create the models were 3, considering the GDS information. One containing the instances of all patients that had a GDS score value (D_{all}), and other two built from D_{all}, one for depressed patients (D_{5-14}) and other for not depressed patients (D_{0-4}).

As illustrated in Figure 4, the first approach consists in training the models using the entire dataset, D_{all}, to understand how general models effect specific groups of patients. The best classifier under the cross-validation evaluation in M_{all}, is then applied to the distinct datasets (D_{0-4}, D_{5-14}) and the respective results are generated for each model created $M_{0-4(all)}$, $M_{5-14(all)}$. The second approach, the initials datasets are already the D_{0-4}, D_{5-14}, and they are used separately to train the classifiers under a cross-validation procedure, find the best one, and generate the specific models, M_{0-4}, M_{5-14}, respectively. Finally, a output result is presented.

5.1.1 Two Years Temporal Window

According to the first approach, it is important to understand how the general model classifies the patients according to their state of depression. For that reason the best classifier for D_{all} was the classifier applied to the models $M_{0-4(all)}$ and $M_{5-14(all)}$, in this case was Bagging.

In the Table 4, it is noticeable that the general model Mall has a good predictive capability independent of the score of GDS of the patient. And the predictability is even better in the specific groups of patients.

In the second approach, the goal is to understand the predictability of specific models considering different characteristics of the patients, when compare to the performance of general models. The Table 5 shows the final results for the specific models M_{0-4} and M_{5-14}.

If we look at the accuracy and sensitivity levels from the Table 4 and the Table 5, it is visible that there is no significant difference between the models M_{all} that predict the different characteristics of the patients without learning with it, and M_{0-4} and M_{5-14} that are built knowing the specific groups of patients. However, in terms of specificity, it is important to report that the models M_{all} performed better than the specific models. And the models which represent the patients with values of GDS score from 0 to 4 had a visible positive distinct performance when compared with the models representing the patients with GDS score values from 5 to 14.

5.1.2 Summary

The results presented in this study corroborate the hypothesis that the separation of the patients according to their state of depression influences positively the prognosis results. This can be seen, at all temporal windows, at any table of any model if compared with the graphs of the previous chapter, where the AD prediction was made with all patients together in a dataset. In this chapter, most of the performance results, in AUC metric were high above the 90%, while in the previous chapter, the best case went to 90%,. That shows that the GDS score attribute have a positive effect in the AD prognosis. Beside this, we can conclude that the usage of ensemble classifiers in these models had a positive impact, since most of the best results, here demonstrated, were originated by the ensemble classifiers.

The two models trained proved that a general model, trained with all patients with GDS information, is better predicting the AD prognosis in a depressed group of patient and not depressed group of patients, than models trained with these specific groups of patients according to the GDS information.

For all these reasons, it is so important the medical clinicians to collect the depression level when the diagnosis of a AD patient is happening.

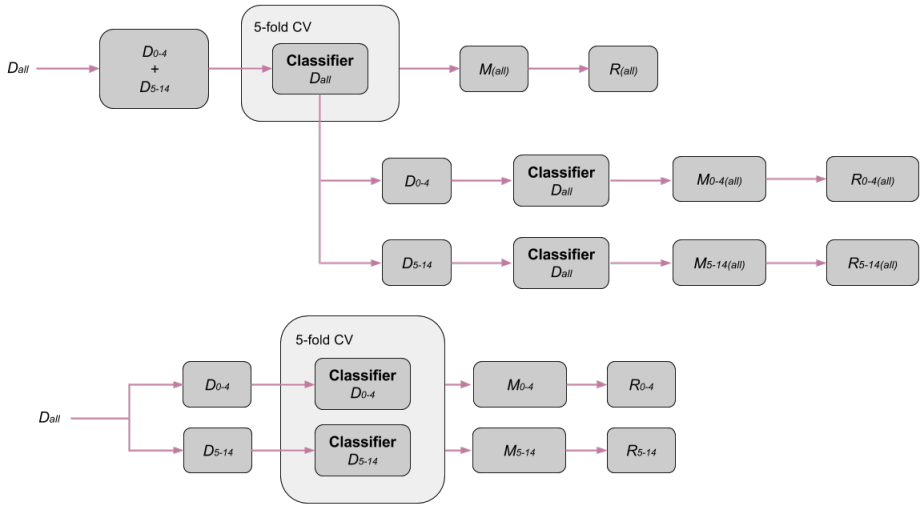

Figure 4: Workflow of the two approaches performed on GDS dataset

Model	*Classifier*	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*
M_{all}		0,83	0,85	0,77	0,87
$M_{0-4(all)}$	Bagging	0,89	0,88	0,94	0,95
$M_{5-14(all)}$		0,89	0,89	0,88	0,94

Table 4: Evaluation metrics of Bagging model for two years temporal window

5.2. Prognosis prediction using Clustering methods

The second phase of this work is to study how patients, clinically similar fitted together in different datasets, progress with the disease of Alzheimer through the years. To create these datasets of patients are applied unsupervised methods to the original datasets of each temporal window. Several clusters are created, in order to understand how patients and their respective class labels are divided and decide after that, how many clusters should be used in our prognosis analysis. The final step of this work, is to do the proper classification procedure, based of the previous examples and chapters.

The Expectation-maximization (EM) algorithm is applied to the MCI patients, so it can be created different groups of individuals based on their patterns of performance. For each temporal window, it is tested which is the best number of clusters that the data should be divided into to achieve the purpose, 2, 3 or 4. In general, we could see that was extremely hard to find those three subgroups of MCI patients since most of the cluster divisions were based on the classes. Which means that the clusters were mainly composed by Evol examples or noEvol examples. And this situation tended to become worse each time it increased k, the number of clusters. For that reason and because some of the datasets were too small and some of the clusters had a percentage above 90% in one of the classes, we decided that we should only use k = 2, for all temporal windows.

5.2.1 Four Years Temporal Window

In the Table 6 are presented the performance results of the two clusters on the temporal window of four years.

In the first cluster, is possible to observe that all metrics have really close score values, which allows to say that the cluster is probably a good cluster, with good performance results. While in the second cluster did not perform the same way, as can be seen by the values of sensitivity and specificity that are noticeably different. The high values of specificity can be explain by the elevate number of Evol instances (86%), and consequently the low number of sensitivity can be explains by the low number of noEvol examples (14%).

5.2.2 Summary

Although we have chosen less number of clusters to divide the dataset of each temporal window, to reduce the number of clusters with asymmetric percentages of classes, and not to verify in the classification results an accentuated division of clusters by class value, that end up to happen. Since there were clusters with low examples of one of the class values and that led to poor results in prediction for those clusters.

Model	*Classifier*	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*
M_{0-4}	Bagging	0,87	0,91	0,71	0,89
M_{5-14}	Bagging	0,83	0,86	0,77	0,87

Table 5: Evaluation metrics of each model for the two years temporal window, 10*5 CV

Model	*Classifier*	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*
$M_{cluster1}$	Bagging	0,82	0,85	0,75	0,82
$M_{cluster2}$	Naïve Bayes	0,89	0,39	0,96	0,89

Table 6: Evaluation metrics of each cluster for the four years temporal window, 10*5 CV

In this approach, there is not a correlation between the course of the years and performing results, all temporal window had singular performances results depending of the type of patients they group.

The approach of creating clusters of patients, according to their similarities, did not make the performance results to improve when compared with the performing results from the datasets with all patients.

6. Conclusions

Despite the Naïve Bayes was our baseline, it revealed statistically good results. As well as, the Bagging method which were the best ensemble method from all the others, and in half of the temporal windows outperformed the baseline. Meanwhile, the results, also, showed that with the increase of the temporal windows, the prediction capability of the models also increased. The window of two years revealed very poor results which is justified by the difficulty in the conversion of the disease in just a short time. While the temporal window of 3 and 4 years look the most reliable temporal windows in final results, considering the number of instances. Although the temporal window of 5 years produced the best prognosis results, in Accuracy and AUC, they may be influenced by the reduced number of instances. The mixture of experts did not improve the performances results, they were essentially the average by all best ensemble learners.

In the first alternative, of the second prediction of this thesis, when the depression state of a patient was taken into consideration, we could corroborate the hypothesis that the separation of the patients according to their state of depression is a good approach to predicting the conversion of AD. We could also confirm that statement when we compared the performance results of the best ensemble classifiers with the best ensemble classifier of the previous analysis that was accomplished with all patients. In all performance metrics the prognosis made considering the GDS attribute was better than the prognosis made with all patients as equals. Besides this, the best ensemble classifiers found, in these news datasets, were always performed by an ensemble method, more specifically the Bagging method. Which is very positive since our proposal was the usage of this classifiers to outperform the simpler classifiers.

In the second alternative, only two clusters at each temporal window were created to minimize the number the clusters with unbalanced class labels. However, that was not easy to accomplished, since at least one of the clusters had a significant disproportion of classes distribution. The best classification results were obtained by the clusters with the most balance classes' distribution. The other clusters' sensitivity and specificity got affected by the high variance in the classes' distribution. When the best clusters, for each temporal window, were compared with datasets with all patients, there were not an evident improvement in the classification, what was a disappointment. In the final results the usage number of ensemble classifiers were the same as the base classifiers, in the two and five years temporal window Naïve Bayes was the best classifier, and in the three and four years temporal window, Bagging and Vote was the best classifiers.

In future work, we would like to perform a mixture of experts in each temporal window, in both alternatives, and observe its performance results. Besides this an ensemble of these mixture of experts might be, also, interesting to observe.

Acknowledgements

I would like to thank Prof. Sara Madeira, my supervisor and mentor, for taking time out from her, much, busy schedule to help me throughout the whole dissertation and for giving me opportunities and encouragement to continuously do a good job.

I would also like to thank my family for the support and inspiration they provided me through my entire life intending to encouraged me to pursue my interests. Last but not least, I would like to thank my boyfriend for the unconditional support.

References

[1] Alzheimer's disease & dementia | alzheimer's association.

[2] I. Arevalo-Rodriguez, N. Smailagic, M. Roqu i Figuls, A. Ciapponi, E. Sanchez-Perez, A. Giannakou, O. L. Pedraza, X. Bonfill Cosp, and S. Cullum. Mini-mental state examination (MMSE) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). In *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd.

[3] C. Cabral, M. Silveira, and Alzheimers Disease Neuroimaging Initiative. Classification of alzheimer's disease from FDG-PET images using favourite class ensembles. 2013:2477–2480.

[4] R. M. Chapman, M. Mapstone, J. W. McCrary, M. N. Gardner, A. Porsteinsson, T. C. Sandoval, M. D. Guillily, E. DeGrush, and L. A. Reilly. Predicting conversion from mild cognitive impairment to alzheimers disease using neuropsychological tests and multivariate methods. 33(2):187–199.

[5] J. G. Chen, C. L. Edwards, S. Vidyarthi, S. Pitchumoni, S. Tabrizi, D. Barboriak, H. C. Charles, and P. M. Doraiswamy. Learning and recall in subjects at genetic risk for alzheimer's disease. 14(1):58–63.

[6] R. Chen, K. Young, L. L. Chao, B. Miller, K. Yaffe, M. W. Weiner, and E. H. Herskovits. Prediction of conversion from mild cognitive impairment to alzheimer disease based on bayesian data mining with ensemble learning. 25(1):5–16.

[7] F. J. V. da Silva. Pattern recognition of alzheimer's disease on pet images. *Training*, 84(90.90):92.

[8] S. Farhan, M. A. Fahiem, and H. Tauseef. An ensemble-of-classifiers based approach for early diagnosis of alzheimer’s disease: Classification using structural features of brain images. 2014:e862307.

[9] A. LG, D. RA, D. ID, and et al. Conversion of mild cognitive impairment to alzheimer disease predicted by hippocampal atrophy maps. *Archives of Neurology*, 63(5):693–699, 2006.

[10] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. CRC Press.

[11] J. Maroco, D. Silva, M. Guerreiro, A. d. Mendona, and I. Santana. Prediction of dementia patients: A comparative approach using parametric vs. non parametric classifiers.

[12] B. McGuinness, S. L. Barrett, J. McIlvenna, A. P. Passmore, and G. W. Shorter. Predicting conversion to dementia in a memory clinic: A standard clinical approach compared with an empirically defined clustering method (latent profile analysis) for mild cognitive impairment subtyping. 1(4):447–454.

[13] S. A. A. Shah, W. Aziz, M. Arif, and M. S. A. Nadeem. Decision trees based classification of cardiotocograms using bagging approach. In *2015 13th International Conference on Frontiers of Information Technology (FIT)*, pages 12–17.

[14] D. Silva, M. Guerreiro, J. Maroco, I. Santana, A. Rodrigues, J. Bravo Marques, and A. de Mendona. Comparison of four verbal memory tests for the diagnosis and predictive value of mild cognitive impairment. 2(1):120–131.

[15] J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt. Classification of SPECT images of normal subjects versus images of alzheimers disease patients. In W. J. Niessen and M. A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2001*, number 2208 in Lecture Notes in Computer Science, pages 666–674. Springer Berlin Heidelberg.

[16] C. Vercellis. *Business Intelligence*. John Wiley & Sons, Ltd.

[17] F. Wang, D. Shen, P. Yan, and K. Suzuki. *Machine Learning in Medical Imaging: Third International Workshop, MLMI 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 1, 2012, Revised Selected Papers*. Springer.

[18] L. Zhan, Y. Liu, J. Zhou, J. Ye, and P. M. Thompson. Boosting classification accuracy of diffusion MRI derived brain networks for the subtypes of mild cognitive impairment using higher order singular value decomposition. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 131–135.

[19] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen. Multimodal classification of alzheimer's disease and mild cognitive impairment. 55(3):856–867.