Tema

Tendências em Matemática Aplicada e Computacional, **18**, N. 1 (2017), 15-34

www.scielo.br/tema
doi: 10.5540/tema.2017.018.01.0015

Wrappers Feature Selection in Alzheimer's Biomarkers Using kNN and SMOTE Oversampling†

Y.E. RODRIGUES[1*], E. MANICA[1], E.R. ZIMMER[2], T.A. PASCOAL[3], S.S. MATHOTAARACHCHI[3] and P. ROSA-NETO[3]

Received on December 20, 2015 / Accepted on February 7, 2017

— For the Alzheimer's Disease Neuroimaging Initiative (ADNI)**

ABSTRACT. Biomarkers are characteristics that are objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention. The combination of different biomarker modalities often allows an accurate diagnosis classification. In Alzheimer's disease (AD), biomarkers are indispensable to identify cognitively normal individuals destined to develop dementia symptoms. However, using the combination of canonical AD biomarkers, studies have repeatedly shown poor classification rates to differentiate between AD, mild cognitive impairment and control individuals. Furthermore, the design of classifiers to access multiple biomarker combinations includes issues such as imbalance classes and missing data. Due to the number of biomarkers combinations wrappers are used to avoid multiple comparisons. Here, we compare the ability of three wrappers feature selection methods to obtain biomarker combinations which maximize classification rates. Also, as the criterion to the wrappers feature selection we use the k-nearest neighbor classifier with balance aids, random undersampling and SMOTE oversampling. Overall, our analyses showed how biomarkers combinations affect the classifier precision and how imbalance strategy improve it. We show that non-defining and non-cognitive biomarkers have less precision than cognitive measures when classifying AD. Our approach surpasses in average the support vector machine and the weighted k-nearest neighbor classifiers and reaches $94.34 \pm 3.91\%$ of precision reproducing class definitions.

Keywords: k-nearest neighbor, SMOTE, feature selection, Alzheimer's biomarkers, Alzheimer's disease classification.

---

† Work presented at Congress of Applied and Computational Mathematics – CMAC SE 2015.

*Corresponding author: Yuri Elias Rodrigues – E-mail: yuri.rodrigues@acad.pucrs.br

[1] Instituto de Matemática e Estatística – IME, Universidade Federal do Rio Grande do Sul – UFRGS, 91509-900 Porto Alegre, RS, Brasil. E-mail: evandro.manica@ufrgs.br

[2] Instituto do Cérebro do Rio Grande do Sul – InsCer, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, 91410-000 Porto Alegre, RS, Brasil. E-mail: eduardo.zimmer@ufrgs.br

[3] Translational Neuroimaging Laboratory – TNL, McGill University Research Centre for Studies in Aging, H4H 1R3, Montreal, QC, Canadá. E-mails: tharick.alipascoal@mail.mcgill.ca; sulantha.s@gmail.com; pedro.rosa@mcgill.ca

**Data used in the preparation of this article were obtained from the ADNI database. As such, the investigators within the ADNI contributed to the provided data but did not participate in analysis or writing of this report.

1 INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia worldwide, posing enormous economic and social costs for the society [1]. AD [2] is pathophysiologically characterized by the gradual brain deposition of amyloid plaques, neurofibrillary tangles, and eventual neuronal depletion [2]. The AD spectrum is composed by preclinical (CN), mild cognitive impairment (MCI) and AD dementia phases [2]. Preclinical AD individuals are those cognitively normal with amyloid plaques and tangles, individuals with MCI have cognitive symptoms without meeting clinical criteria for dementia, and AD dementia individuals present severely compromised cognitive faculties [3]. In recent years, a plethora of biomarkers has been developed in order to track AD progression, such as biomarkers for beta peptide 1-42 ($A\beta_{1-42}$) and tau proteins that indicate the presence of the hallmark pathological features of AD, amyloid plaques, and neurofibrillary tangles, respectively [1, 2].

It is a well-established fact that combined biomarkers provide higher classification rates than single biomarkers [4]. In this regard, neuropsychological tests associated with different biomarker modalities have been used to classify AD [5]. These studies have combined positron emission tomography (PET), magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI) as well as cerebrospinal fluid (CSF) and blood biomarkers to perform binary classifications of AD [1], e.g. healthy versus unhealthy individuals. Despite AD's classification problem being an inherently multiclass binary classification multiclass driven by binary strategies are the rule [1]. This happens since some classifiers are naturally binary and must be adapted to be multiclass by means of one-versus-all and one-versus-one strategies, e.g. support vector machine (SVM) [6]. Thus, to solve an n-class problem using binary classifiers, $\frac{n(n-1)}{2}$ rules are required to build a multiclass classifier. As benefit, binary classifiers are well suited for the receiver operator characteristic (ROC) analysis [7] which has been largely applied in AD comparative studies, biomarkers model selection and conversion diagnosis prediction.

Recently, various approaches used for AD's identification have achieved successful results and satisfactory classification rates. For instance, Khedher et al. [8] were able to accurately differentiating the three clinical classes of the AD spectrum reaching the maximum sensitivity (85.11%), specificity (91.27%), and accuracy (88.49%) values by implementing binary strategy and reduction of input space with SVM and principal component analysis (PCA) techniques [6]. Khazaee et al. [9] were able to perfectly differentiate between cognitively healthy and AD classes in a small dataset of 40 individuals, using graph theory applied to brain connectivity assessed with fMRI they reach an accuracy of 100% for linear SVM and 87.5% for k-nearest neighbor (kNN). Although the separation between extreme cases is straightforward, difficulties are expected when considering the overlapped intermediary classes. Classifiers performance can be potentially affected by data issues, such as class overlapping, feature space with high dimensionality, missing-data, class imbalance, etc [10]. Particularly, imbalanced datasets [11] are considered one of the 10 most challenging problems in machine learning [10]. When imbalanced data issues are disregarded, it could lead to a decreased classification rate for the minor represented class and in globally averaged scores [12]. Solutions that address this problem are based on re-sampling

methods when the distribution mechanisms are known. Alternatively, the methods are mainly based on the creation of synthetic data for minor classes or/and pruning data for major classes [11].

There is a need to identify biomarker combinations that maximize the classification and understand how much they contribute to differentiate between AD classes [13, 1]. However, this goal faces multiple classifier's comparisons when assessing biomarker combinations. In order to avoid the excessive number of comparisons, feature selection techniques are able to find a set of biomarkers that meet defined criteria [14]. For a given task (e.g. classification) examples of criteria are: to identify the most cost-effective biomarkers, with higher precision and low false-positive, find a subspace of reduced dimensionality with the same or enhanced discriminant properties; extract/build relevant features from raw data [15].

Techniques of feature selection have been largely applied to AD-related problems, intending to provide a better understanding of biomarkers relationship [13] and achieve defined criteria of usefulness [14]. Interesting applications of feature selection techniques contributed to the understanding of AD, like the construction of potential biomarkers for enhanced classification. For instance, Lopez-de-Ipiña et al. owing to determine preclinical biomarkers for AD apply feature selection techniques on spontaneous speech to extract discriminant features [16]. They also were able to correctly classify AD subjects using kNN and multilayer perceptron (MLP) classifiers obtaining accuracy of 87.30% and 90.90%, respectively to each classifier. Feature selection AD-related also is found in the gene microarray analysis [15] and neuroimaging both with high-dimensional feature spaces and its own big data challenges. These fields have been provoking adaptation of feature selection techniques to deal with high dimensionality (tens of thousands features) and small sample size in the case of microarray datasets [15]. In neuroimaging, the feature selection methods in 3D matrices are able to mitigate performances issues and improve the classification accuracy [17].

Here, we propose to find subsets of features among several feature combinations which maximize classification rates between three AD classes. Specifically, we solve a multiclass classification problem in which test patterns are assigned to one of following classes: control normal (CN), mild cognitive impairment (MCI) or AD. To do that, we compare three feature selection techniques that depend on the classifier's outcome as a measure of usefulness [14]. This requirement characterizes the feature selection techniques called wrappers which select features based on the classifier's performance. However, instead of widely applied binary strategies, here we will use the all-versus-all strategy naturally achieved by the kNN classifier. The misclassification and the comparison between the biomarker combinations will be done by scalar measures of confusion matrices [18]. In order to observe the effect of training set size, we compare two validation processes, 10-fold cross-validation (10-fold CV) and leave-one-out cross-validation (LOOCV) [6]. Our analysis shows how the imbalanced dataset affects classification rates and shows a comparison of the feature's probability to reach higher precision. Two techniques to aid the class balances are compared with the imbalanced dataset: an oversampling technique called synthetic

minority oversampling technique (SMOTE) and the random undersampling. All algorithms and plot codes are available on-line `https://github.com/yurier/TEMA-R-CODES`.

2 METHODS

2.1 Dataset

Data used in the preparation of this article were obtained from the ADNI database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see `www.adni-info.org`. Features in this work consist of two neuroimaging biomarkers (labels 1,2), four neuropsychological tests (labels 3,4,5,6) and two proteomic biomarkers (labels 7,8) [19], respectively: 2-[18F]fluoro-2-Deoxy-D-glucose (FDG) PET, florbetapir-fluorine-18 (18F-AV-45) PET, clinical dementia rating sum of boxes (CDRSB), Alzheimer's disease assessment scale-cognitive with 11 items (ADAS11), mini mental state examination (MMSE), Ray auditory verbal learning test percent forgetting (RAVLT), $A\beta_{1-42}$ CSF, phosphorylated tau protein (p-tau$_{181}$) CSF. Table 1 depicts dataset demographics.

Table 1: Dataset demographics described by mean and standard deviation.

Feature	CN	MCI	AD	Label
Male	73	227	59	–
Female	79	187	41	–
Age	73.31 ± 6.35	71.39 ± 7.44	74.88 ± 8.19	–
Education	16.53 ± 2.50	16.18 ± 2.65	15.72 ± 2.55	–
FDG *	6.59 ± 0.51	6.33 ± 0.65	5.28 ± 0.76	1
18-F-AV-45 **	1.10 ± 0.17	1.20 ± 0.22	1.39 ± 0.20	2
CDRSB	0	1.44 ± 0.86	4.70 ± 1.63	3
ADAS11	5.85 ± 3.13	9.25 ± 4.45	20.96 ± 7.13	4
MMSE	29.05 ± 1.18	28.07 ± 1.73	22.96 ± 1.98	5
RAVLT	35.22 ± 26.69	55.41 ± 31.37	89.17 ± 20.72	6
$A\beta_{1-42}$ pg/mL	196.67 ± 49.96	174.79 ± 51.55	133.20 ± 35.84	7
p-tau$_{181}$ pg/mL	33.52 ± 16.40	41.26 ± 24.30	58.06 ± 29.39	8

*Average of FDG-PET of angular, temporal, and posterior cingulated with pons as reference region [20]. ** Average of standardized 18-F-AV45 uptake value ratio (SUVR) of frontal, anterior cingulate, precuneus, and parietal cortex relative to whole cerebellum as reference region [20].

Due to different probability in pathologies stages, clinical datasets are subject to imbalanced classes. Also, the data imbalance is a critical issue that generates an unfair separation between classes, when the imbalance is extreme [12] it will lead to decreased prediction rates for validation stages. The present study evaluates two strategies to adjust the class imbalance by assuming any class posterior distribution. Since we are using the kNN that is a distance based algorithm features were scaled by min-max normalization in all experiments [21] except for figures. Importantly, since features CDRSB and MMSE are employed to define the diagnosis or are similar to categorization protocol they will be used as contrast. The feature-wise comparison will be performed only for non-defining and non-cognitive features since it is well known that cognitive measures have more discriminative power [1]. Moreover, feature selection will be applied for non-defining features which include cognitive measures.

2.2 k-nearest neighbors (kNN) algorithm

Classifiers can be defined by discriminant functions [18], which are a set of functions to predict categorical dependent patterns. Here, we define a classifier C as a function that assigns a pattern $x \in \mathbb{R}^n$ to a class into the class space $\omega \in \Omega := \{\omega_1, \ldots, \omega_c\}$,

$$C(x) : \mathbb{R}^n \to \Omega.$$

The *maximum a posteriori* (MAP) classifier uses a set of discriminant functions to assign the most probable class. Let $\{f_{\omega_i}(x)\}_{i=1}^{c}$ be a set of discriminant functions, a classifier C is said to be well-defined if, for all patterns, is possible to assign a class. Let $\tilde{\omega}$ be the calculated class, the MAP classifier is given by,

$$\tilde{w} := \arg\max_{\omega \in \Omega} f_{\omega_i}(x) = C(x). \tag{2.1}$$

The discriminant functions set is monotonous, since for a given x associated to a class ω_j we have that $f_{\omega_i}(x) \geq f_{\omega_j}(x)\ \forall i \neq j$. An example of a set of discriminant functions is the class conditional probability $\{p(x|\omega_i)\}_{i=1}^{c}$ [18]. The classifiers taxonomy is based on how they approximate its discriminant functions [6]. *Generative* models parametrically approximate the posterior class probabilities, $p(x|\omega_i)$, through the class conditional probability $p(\omega_i|x)$ and the class prior probability, $p(\omega_i)$ (e.g. gaussian, gamma, etc). Alternatively, *discriminative* models directly approximate posterior class probabilities, $p(\omega_i|x)$, without assuming a distribution for it (e.g. kNN, MLP, etc). Aside from generative and discriminative models, there are the non-probabilistic models in which the discriminant functions are not required to be a distribution (e.g. SVM). Generally, the goal of discriminant functions is to divide the pattern space into decision regions $\{R_1, \ldots, R_c\}$. Class densities are depicted in Figure 1 and the classification mapping example is depicted in Figure 2. Also, Figure 2 shows a binary classification problem for feature combination labels 7 and 8 in which overlapped regions are subject to higher misclassification.

In order to observe how the distributions of CN and AD are overlapped, one can use the Bhattacharyya distance [22] which is a metric to measure how two distributions differ and also provides a bound for the probability of classification error using the Bayes optimum classifier. By

assuming Gaussian shapes the Bhattacharyya coefficient coincides with Mahalanobis measure. The Bhattacharyya coefficient [22] derived from Bhattacharyya distance ranges between 0 (distribution non-overlapped) and 1 (distribution overlapped). Assuming normal distributions between CN and AD distribution the Bhattacharyya coefficient is 0.6398824. Further this measure will be used to show how oversampling modifies the original distribution.

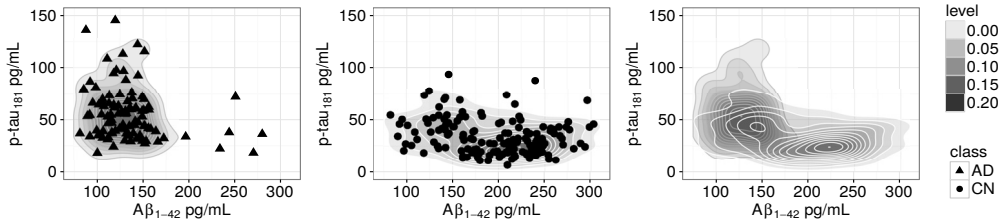

Figure 1: On the left and middle, class distributions for CN and AD classes, respectively. On the right, AD and CN classes distribution overlapped in some regions.

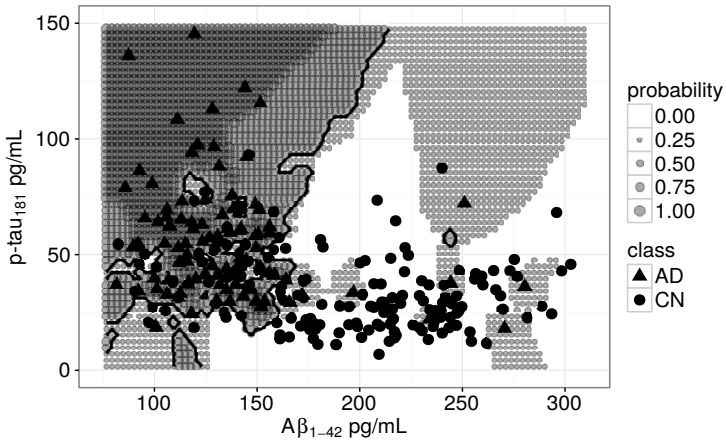

Figure 2: Decision boundaries created by the 3NN classifier.

The kNN is a non-parametric classifier that estimates the class posterior probability assuming that nearest patterns (here assumed to be vectors in a metric space) have similar properties (classes) [23]. This assumption was proved for infinite sample case as shown in [23]. There are preprocessing techniques, which depend on kNN, for instance: in data imbalance, completing missing values, dimensionality reduction, and metric learning. Here we focus specifically on data imbalance challenge and a feasible solution; SMOTE algorithm.

The kNN classifier deals with multiclass problems straightforwardly by means of all-versus-all strategy. To achieve this, it uses a neighborhood defined by k training instances nearest to a query instance to be classified. Let $T = \{(x_i, \omega_i)\}_{i=1}^{m}$ be a training set, with tuples $x_i \in \mathbb{R}^n$ and $\omega_i \in \Omega$ as labeled patterns. Also, let x_* be a test instance with an unknown label to be assigned into

a class $\omega_* \in \Omega$. Using the MAP framework given in equation (2.1) the kNN classifier can be written as,

$$\tilde{\omega}_* = \arg\max_{\omega \in \Omega} \sum_{x_j \in N(x_*,k)} \delta(\omega_j, \omega), \tag{2.2}$$

where $N(x_*, k)$ is a neighborhood with k training instances around x_* for a given metric and $\delta(.,.)$ is the Kronecker delta function. Rewriting equation (2.2) we have,

$$\tilde{\omega}_* = \arg\max \left\{ \sum_{x_j \in N(x_*,k)} \delta(\omega_j, \omega_1), \ldots, \sum_{x_j \in N(x_*,k)} \delta(\omega_j, \omega_c) \right\}. \tag{2.3}$$

The monotonicity of discriminant functions is applied to equation (2.3) and dividing it by k,

$$\tilde{\omega}_* = \arg\max \left\{ \sum_{x_j \in N(x_*,k)} \frac{\delta(\omega_j, \omega_1)}{k}, \ldots, \sum_{x_j \in N(x_*,k)} \frac{\delta(\omega_j, \omega_c)}{k} \right\}. \tag{2.4}$$

Thus, each term in equation (2.4) is the posterior class probability,

$$p(\omega_i | x_*, k, T) = \sum_{x_j \in N(x_*,k)} \frac{\delta(w_j, w_i)}{k}, \qquad \text{from} \quad i = 1, \ldots, c. \tag{2.5}$$

From equation (2.5) we have that the most probable class for the training instance x_* is given by,

$$\tilde{\omega}_* = \arg\max_{\omega \in \Omega} p(\omega | x_*, k, T). \tag{2.6}$$

The parameter k that adjusts the k-neighborhood $N(x_*, k)$ is searched empirically since there are no guidelines for its optimality. However, Bhattacharyya [24] proposes a bound to the optimal k, that is $k < \sqrt{m}$. In binary classification, one can restrain the range of k by using only odd values in order to avoid ties in equation (2.1). The class posterior distribution, $p(\omega_j | x)$, is an alternative to the optimum Bayesian classification approach, which requires the complete knowledge of data generation underlying mechanisms. As showed by Cover & Hart [25], the error of the nearest neighbor is limited between the optimal error and twice the optimal error for the infinite sample case. This means that the more data available the closer the optimum error will be. This was supported by Stone on the existence of a universally consistent classifier [23]. Let's show how the parameters affect the classifier properties. Figure 3 depicts the role of parameter k in the decision regions as well as the probability of class assignment described by the equation (2.6). Also, in Figure 3, note that there are only two probability values for $k = 1$ (1NN classifier). This happens since there are no misclassification errors or ties for training set in 1NN. A perfect score in training set would be an indicator of overfitting that generally leads to poor performance issues, that is, the classifier is unable to generalize the training results. Contrasting to 1NN, when parameter k is increased, more data is needed in order to evaluate the pattern assignment leading to more stratified probability values.

Non-parametric methods as kNN do not rely on distribution assumptions [18] (e.g. Gaussian shape). As an example see that the kNN classifier given in equation (2.1) depends only on the

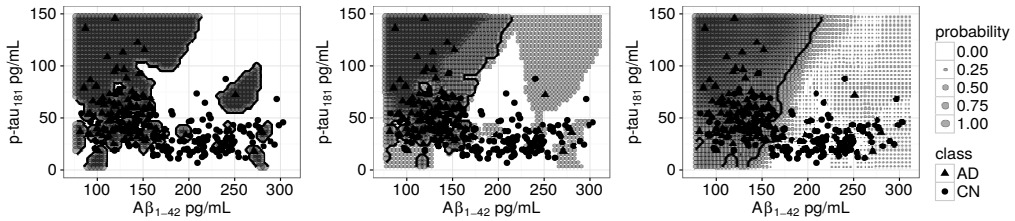

Figure 3: Left to right, the influence of parameter k in the decision boundaries, for $k = 1$, $k = 5$ and $k = 15$, respectively.

training data T and k to approximate the class posterior distribution. That is to say, data modification or removals imply in different classifiers since T is modified. By that, the biomarker combinations are validated using 10-fold CV and LOOCV in order to observe the variation between different training data sizes. Although, variations using the same dataset also can occur due to ties in kNN. For instance, with 5NN some query pattern class could be evaluated as "AD+AD+CN+ CN+MCI". The random tie-breaking is adopted in this work because is computationally more efficient than tie-breaking strategies described in [26]. Figure 4 shows the effect of random tie breaking in the decision regions for the three class problem proposed.

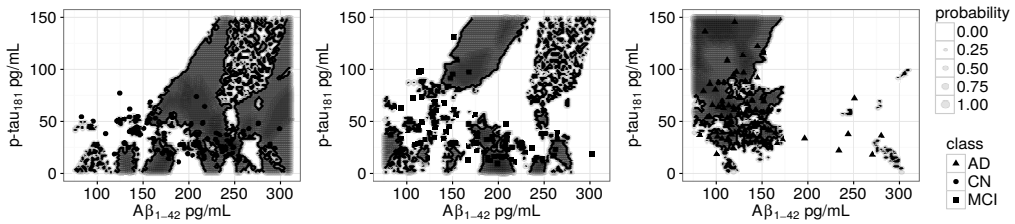

Figure 4: Decision region for a three class problem solved with 5NN using the same number of samples for each class. From left to right, decision region of CN, MCI and AD, respectively.

One can observe in Figure 4, apart from noisy region created by the random tie breaking, the decision regions are complementary. Furthermore, unlike the binary case that requires only one class posterior probability to describe the classification mapping, the multiclass problem needs all the posterior probabilities to describe the classification mapping. For instance, binary mapping needs only to evaluate $p(\omega_2|x) = 1 - p(\omega_1|x)$ and the decision border can be described with $p(\omega_1|x) = p(\omega_2|x) = 0.5$. In turn, the multiclass problems need all posteriors $\{p(\omega_i|x)\}_{i=1}^{c}$ to describe the classification mapping and the decision borders are drawn between the class transitions. In equation (2.2) we suppose that every nearest neighbor contributes equally to calculate the class, independently from the distance to the query point. The wkNN is a kNN's extension to handle this issue. The wkNN attributes weights for each voting pattern reducing the high dimensionality effects. In high-dimensional feature spaces, the training patterns become sparse requiring more data to fill out the decision region. The wkNN overcomes the kNN issues in

high-dimensional spaces using a weighted scheme $W(.,.)$ as an argument of a kernel function $K(.)$. This is to adjust relative distances between patterns avoiding the sparsity [27]. Furthermore, fractional metric modifications in kNN (or metric based classifiers) also can grant some level of reliability in high-dimensional feature spaces [28]. The wkNN can be written as,

$$\tilde{\omega}_* = \arg\max_{\omega \in \Omega} \sum_{x_j \in N(x_*,k)} \delta(\omega_j, \omega) K(W(x_*, x_j)), \tag{2.7}$$

where the weighted scheme [27] can be defined as follows,

$$W(x_*, x_j) = \begin{cases} \frac{d(x_k,x_*)-d(x_j,x_*)}{d(x_1,x_*)-d(x_k,d_*)} & \text{if} \quad d(x_k, x_*) \neq d(x_1, x_*) \\ 1 & \text{otherwise} \end{cases}$$

2.3 SMOTE

Data imbalance shows to be an adverse setup to achieve high classification rates. This happens due to the rare or less frequent instances of minor represented classes (e.g. bank fraud, cancer malignancy grading) [12]. When the minority class has few training patterns it turns out to be misrepresented leading to shrunken decision regions. For instance, in Figure 5 the minor class was generated inside the circle with uniform distribution with 8 samples while the major class follows the distribution $\mathcal{N}([2, 2], 4I)$ with 100 samples generated from it.

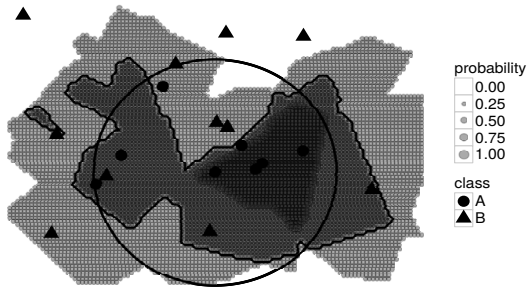

Figure 5: Decision region for an imbalanced dataset. The irregular solid line is the decision border between class A and B.

In order to avoid problems in classification mapping, one can randomly remove instances from the major classes until achieving the same prior probability (proportion). This technique is named the random undersampling. Instead to prune data one can rise the low priors of minor classes using an oversampling technique, in this case SMOTE, which creates synthetic patterns based on the existing ones [29]. Balance aid techniques developed for high-dimensional applications are also useful in situations with moderate imbalance as well [12]. In supervised learning, strategies to prevent imbalance are mainly focused in class reorganization or resampling owing to achieve the same number of training instances [11]. Methods to aid the imbalanced problems are organized into the following categories [12]: data-level, algorithm-level or hybrid. We compare the

effect of undersampling and oversampling techniques that are data-level methods since data is balanced disregarding the subsequent classifier.

SMOTE is similar to kNN and can be implemented as follows [29]: first, a pattern in the minor class is randomly selected, then, a synthetic pattern is included between a randomly chosen pattern in its k-neighborhood, repeated the procedure until achieving desired prior. SMOTE increases the parameter space that must be searched to obtain the model with highest prediction rate. In this work, we set SMOTE's k-neighborhood parameter as $k = 5$. Figure 6 depicts how the amount of oversampling changes the distribution for 50% and 100% of synthetic data added in AD class relative to MCI class, using $k = 5$. The Bhattacharyya coefficient for 50% and 100% oversampling relative to the original distribution are 0.994579 and 0.9975564 respectively. Figure 7 depicts the effect of parameter k in SMOTE algorithm. For $k = 7$, $k = 15$ and $k = 30$ the Bhattacharyya coefficient (assuming Gaussian distributions), from original sample is respectively, 0.9924127, 0.9972362 and 0.9935776, give that the value 1 means a complete distribution overlap.

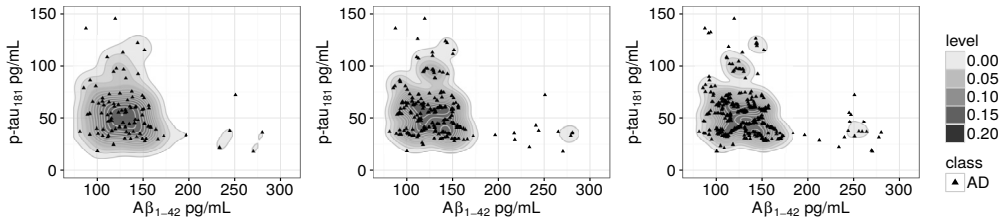

Figure 6: From left to right, the original data, synthetic oversampled 50% with original data, synthetic oversampled 100% with original data, respectively and their distributions.

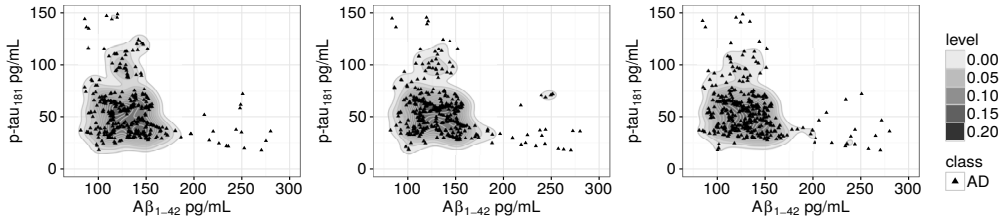

Figure 7: Original data with SMOTE oversampled 100% for values of parameter k, from left to right, $k = 7$, $k = 15$ and $k = 30$.

2.4 Wrapper feature selection

The goal of feature selection methods is to select a subset of features that is *useful* to enhance a given classifier's measure, e.g. precision. Since classifiers are induced by data with unknown underlying distributions the feature selection methods allow sub-optimal answers. There are comprehensive definitions of usefulness that would be criteria to select relevant features, e.g.

correlation and information theoretical criterion [14]. As shown in [30] the optimal choice of features does not imply the choice of relevant features. Conversely, optimality does not imply in relevancy. For instance, features that are presumably redundant may enhance the precision when combined with useful features [14].

Despite the lack of guarantees presented, feature selection methods were invaluable to deal with high-dimensional real-world problems. Feature selection methods were initially designed to deal with classification problems with no more than 40 features [30], now they are able to deal with thousands of features. For instance, in classification problems related with genetics feature space dimensionality ranges from 6000 to 60000 [14] one can expect a strong effect of the curse of dimensionality. Such extreme problems have received attention uncovering the molecular mechanisms related to AD [31] and have motivated initiatives like AlzGene focused on providing data resources for AD genetic research. Despite that, in this work, the feature selection techniques will be applied to at least 9 features in order to observe the group-wise probability of a feature being more relevant than other. Examples of feature selection techniques are feature extraction, feature construction, feature selection techniques for non-supervised learning, etc.

Feature selection methods are divided into three categories due to the relation with the classifier: filter, wrapper and embedded methods. Filters select a subset of features independently of the chosen classifier and the procedure mainly focus on ranking the features given defined criteria. Conversely, wrappers use classifiers' measures as criteria to select subsets. Lately, embedded methods use a structured model to get the set of relevant features subject to a classifier [14]. For a complete discussion on the feature selection strategies and benefits see [14].

Here we compare three wrapper methods using the kNN classifier combined with SMOTE to select the most useful subset of features. The three wrapper methods compared are defined on the following search strategies: backward elimination, forward selection and hill-climbing selection [14]. The subset obtained using the three methods will be compared to all combinations of features in order to observe if they are able to reach the optimum subset drawn from the rank with all feature combinations. Additionally, a noise feature will be included in the feature selection procedure in order to compare the features to a non-significant case. In Figure 8 we depict an example of search graph with all possibilities for three combinations. Regarding Figure 8 we have that in the far left stage, no features chosen and the far right all features chosen. Backward elimination moves right to left; forward selection left to right; hill climbing moves to any direction.

The wrappers feature selection will search in a 9 (8 + noise) features graph scheme for the more useful subset. Forward selection initializes with any feature and steps up towards completing the feature subset. Iteratively it adds features to the chosen subset using the usefulness criteria that is to rise kNN classification rate. The usefulness criteria for kNN is obtained with the LOOCV, that means to classify one pattern using all remaining patterns as a training set. The backward elimination goes in the opposite direction in the searching graph. It initializes with all features and iteratively prunes features according to with the highest usefulness defined by the kNN classification rate criteria. The hill-climbing selection can go in any direction in the searching graph

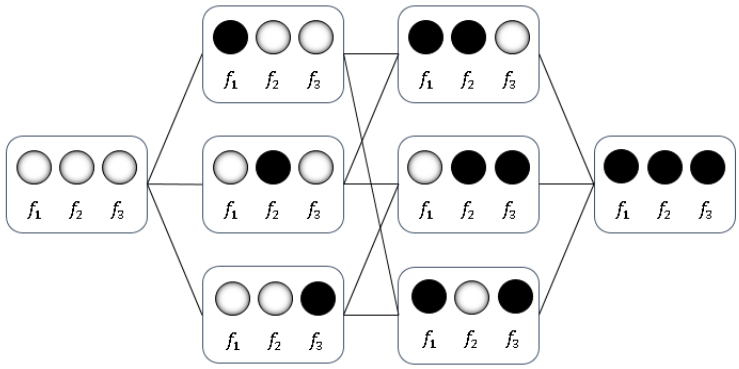

Figure 8: Scheme showing how to choose a subset of 3 features, using forward selection, backward elimination, and hill-climbing.

combining two previous approaches. Here we set up hill-climbing starting from the empty set feature. All wrappers in this work are greedy algorithms and are subject to be trapped in a local maximum [30] (relative to the classification rate). A greedy algorithm can only optimize in a short distance and do not prevent that a good choice for a given iteration can lead to losing better options. There are search strategies designed to avoid this greedy drawback, for instance, the simulated annealing and the genetic algorithms [14].

2.5 Validation

Overfitting happens when the classifier's prediction to the training phase are far better than the test phase [6]. The validation is appropriate to observe if the classifier is overfitted. An overfitted classifier cannot generalize results achieved in training phase for unseen patterns. Confusion matrix P is a tool to assess the classifier outcomes and to interpret classification precision. The confusion matrix shows the class wise probability of classifying x_* in the class ω_j given that it was generated by class ω_i, for short $P_{i,j} = p(x_* \in \omega_i | x_* \in \omega_j) = p(\omega_i|\omega_j)$. Furthermore, P is a stochastic matrix,

$$\forall i \quad \sum_{j=1}^{c} p(w_i|w_j) = 1 \quad \text{with} \quad p(w_i|w_j) \geq 0 \quad \text{for} \quad i, j = 1, \ldots, c,$$

the P trace average is the probability of correct classifications for all classes or precision [18]. This measure defines a scalar magnitude that enables us to compare the many classifiers in the feature selection process. Furthermore, the 10-fold CV [6] is applied to the confusion matrices to obtain deviations of the feature's subsets. The scalar for classifier's comparisons can be written as,

$$val(P) := \mathbb{R}^{c\times c} \to \mathbb{R} \qquad val(P) = \sum_{i=1}^{c} \frac{P(w_i|w_i)}{c}.$$

3 RESULTS AND DISCUSSION

Features for specific and general cohort studies owing to identify AD spectrum come from various sources: cognitive, genetic, neuroimaging, proteomic and others [16]. These features intend to provide insights on biological factors AD which is critical to understanding the disease progression and to early prevention strategies [1]. Feature combinations provide higher classification rates than features by itself, also there are combinations more precise than others. For instance, a feature may confer a poor classification rate when many classes are included, even being highly valuable to understand AD biological processes, as depicted in Figure 4. This is, the combination $A\beta_{1-42}$ and p-tau$_{181}$, which sustain the main hypothesis for neurodegeneration [1] achieves a precision of $42.29 \pm 1.32\%$ and $43.40 \pm 4.61\%$, respectively for LOOCV and 10-fold CV for kNN with k optimized and imbalanced data. Thus, it is necessary to find additional features or combinations to better identify AD. However, for n features the number of combinations is given by $\sum_{i=1}^{n} \frac{n!}{i!(n-i)!}$, thus requiring strategies to avoid computational effort to uncover such combinations. Wrapper feature selection techniques are suitable to avoid the comparison of all features combinations while maximizing a chosen classifier's precision. In case, the kNN that allows the all-versus-all strategy to observe how the classes affect each other all at once. Also, the all-versus-all strategy contrasts to the binary adapted strategies that are widely used in AD research along with ROC analysis [1]. We compare the three techniques of wrappers feature selection to the global rank of features for each sampling strategies using confusion matrices. Moreover, Gaussian noise (mean $= 0$, sd $= 1$) was added to feature space in order to compare an irrelevant feature to the features displayed in Table 1, with label N (noise). The Table 2 shows the test performance of sorted combinations by higher classification rate among the non-defining features and by the number of features.

Excluding the defining features (3,5) most of the wrappers in this work were able to identify sub-optimal combinations given the precision of combinations available. For imbalanced dataset the combinations found are: 4,6 for hill climbing (position 2); 4,1,8 for backward elimination (position 28); 4,6,7,8 for forward selection (position 41). With random undersampling the combinations found are: 4,6 for hill climbing (position 8); 1,4,6,8 for backward elimination (position 2); 4,1 for forward selection (position 5). With oversampling (SMOTE) the combinations found are: 4,6,7,8,N for hill climbing (position 38); 1,2,4,6,7,8,N for backward elimination (position 9); 4,7,6,8 for forward selection (position 21). All strategies of sampling and wrapper feature selection found sub-optimal combinations relative to the rank position. For the complete ranking list see on-line contents. One can notice that even with hill-climbing which combines the forward and backward strategies it can be trapped in local maximum and be affect by the cross-validation components. For instance, using oversampling with backward elimination was found the position 9 and combined with hill-climbing the position 38 in the full ranking list.

The combination label 3 and 5 for the definition is depicted in Figure 9. From the selected features in training phase, the combination that provides the higher classification rate in validation phase is the (3,5). For more 2D plots see: `https://github.com/yurier/TEMA-R-CODES/tree/master/PLOTS2D`. The wrappers can be affected by the random nature of the cross-

Table 2: Rank of combinations for the three imbalanced strategies. Note the highest classification rate was bolded for each validation method.

position	LOOCV (%)	k	10-fold CV (%)	k	combination
					imbalanced
2	**60.37 ± 0.05**	13	**59.33 ± 0.35**	6	4,6
1	58.35	1	58.37 ± 2.58	4	2,4,8
3	58.11 ± 0.13	2	56.81 ± 2.56	6	1,2,4,8
7	58.01	5	56.29 ± 4.29	3	2,4,6,8,N
10	58.80	3	58.02 ± 1.09	3	1,4,6,7,8,N
38	56.94 ± 0.05	5	56.74 ± 2.25	12	1 2 4 6 7 8 N
					undersampled
5	**70.15**	24	66.78 ± 3.36	18	1,4
3	69.05	20	68.59 ± 3.40	10	4,6,8
2	69.40	12	**69.15 ± 4.46**	13	1,4,6,8
1	67.14	10	67.47 ± 4.37	23	1,4,6,8,N
7	67.08	24	66.57 ± 2.41	25	1,2,4,6,7,8
43	66.39	22	66.13 ± 2.98	24	1,4,2,6,7,8,N
					oversampled
80	**66.98 ± 0.47**	23	**69.20 ± 4.21**	24	4,6
41	63.97 ± 0.11	23	65.08 ± 5.83	21	4,7,N
12	66.35 ± 0.05	19	67.15 ± 2.81	25	1,4,6,8
1	62.16 ± 0.05	25	63.80 ± 4.63	24	1,2,4,7,N
3	64.34 ± 0.53	25	65.69 ±6.02	20	1,2,4,6,8,N
9	62.91 ± 0.11	17	64.16± 7.89	24	1,2,4,6,7,8,N

validation process and results may vary when the random generator number is unfixed. This variation ranges between the very first combinations to the middle-rank combinations. The list of 502 combinations for each technique to aid imbalance is available on-line, also the 120 combinations rank for non-defining features and the 26 combinations rank for non-cognitive features.

Comparing the three techniques of sampling (imbalanced, undersampled and oversampled) we are able to say that there is a significant improvement using the imbalance aid techniques. The subtle dominance of kNN with SMOTE for the LOOCV happens due to the randomness effect of cross-validation and tie breaking. Furthermore, oversampled data is not pruned leaving more available data for training. In order to avoid overfitting, the synthetic data was used only to generate the classifier and to validate results. In Figure 10 is depicted the confusion matrices averaged in 10-fold cross-validation for the combinations ranked between all 502 combinations (on-line contents), respectively for each method and from 8^{th} to 1^{st} combinations. For instance, the first matrix in the first row at left in Figure 10 shows the probability of the MCI pattern be

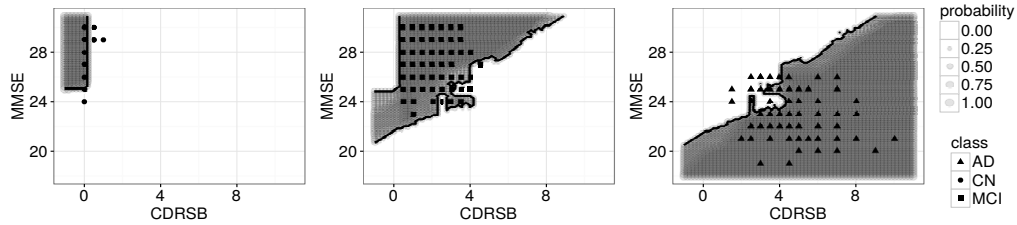

Figure 9: Decision region for balanced (undersampled) three class problem solved with 5NN. From left to right, decision regions of CN, MCI and AD, respectively. Note that even for the defining features there is some errors.

classified as CN is 35.6%. Also, in Figure 10 the confusion matrices for balanced problems (2^{nd} and 3^{rd} rows) show that the major class (MCI) have more classification errors than the minor classes (CN, AD). Conversely, in the imbalanced problem, the confusion matrices (1^{st} row) for the major class are more precise and the minor class has more misclassification errors. This result is expected and can be used to control the compromise between class resampling in order to achieve a balance that minimizes the classification error.

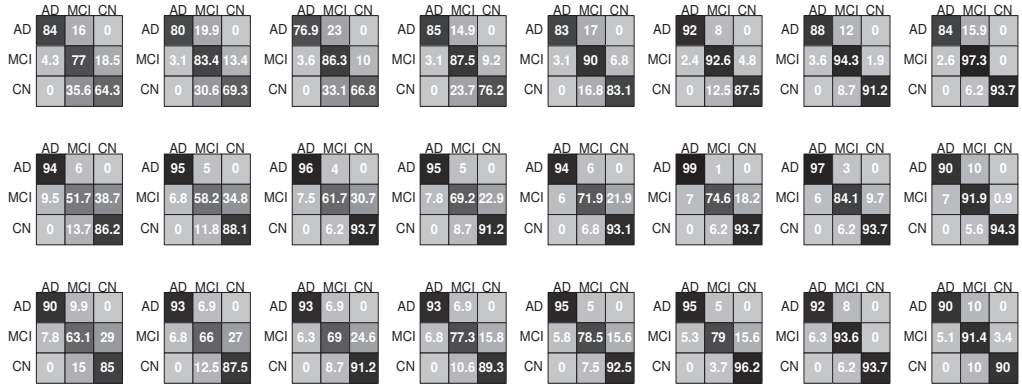

Figure 10: By row is depicted the confusion matrices for imbalanced, undersampled and oversampled combinations, respectively. Each column depicts from 8^{th} to 1^{st} ranked combinations in the 502 rank (on-line contents).

With the rank of all training results for all combinations of eight features plus noise producing 502 different combinations, one can ask the probability that a feature belonging to a given combination have higher classification rate than other. This is done by measuring how many times combinations that have feature A and not B are more precise than combinations that have B and not A divided by the number of comparisons between A and B. Figure 11 depicts this evaluation for each balance method (imbalanced, undersampling, oversampling) using only one significant digit by resolution reasons and just for the non-cognitive features (26 features). For instance, one

can be interested in the probability of combinations that contain the feature 2 and not 8 to provide higher precision than combinations containing 8 and not 2, the left matrix in Figure 11 shows that is 38.7%. However, the matrices are not symmetrical due to some combinations have the same classification rate and due to round-off error. As argued by [1] the neuropsychological tests are more precise and standardized measures to detect AD. Figure 11 shows only non-cognitive features (1,2,7,8). Despite neuropsychological tests being cost-effective biomarkers and its combinations provide a high precision, they do not provide information on the biological mechanism of AD. The suggestion of how they provide a higher precision is due to the limited possibilities of outcomes that define the neuropsychological scores, leading to overlaying patterns, as represented in Figure 9. This is, biomarkers that have fractional values are less subject to becoming an integer value grid in comparison to neuropsychological tests. In the right matrix of Figure 11 the noise has a higher probability in average to increase the classification rate than proteomic biomarker 7 however it does not mean irrelevancy [14]. As argued by [14], a feature that is supposed to be irrelevant could contribute to enhancing the classifier performance.

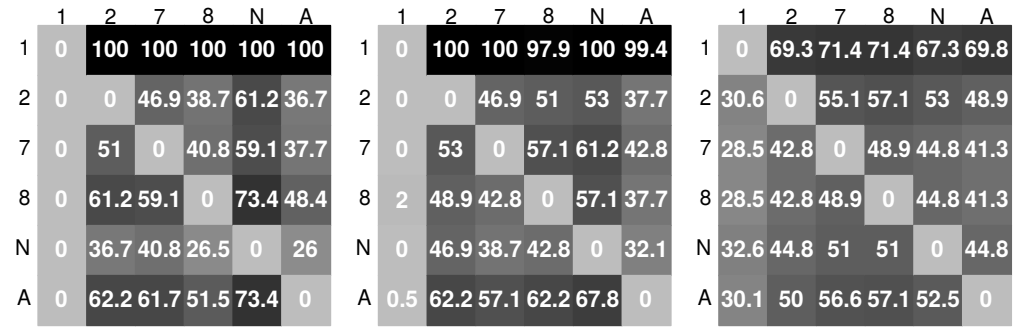

Figure 11: From right to left, probability matrix of a feature to provide higher classification rate than others for the imbalance, undersampled and oversampled, ranks respectively. Label "A" stands for column-wise and row-wise average.

The SVM classifier that uses binary adapted strategy when applied to the same multiclass classification problem achieves $92.39 \pm 4.96\%$ in 10 fold cross-validation for features 3,5, while kNN with SMOTE achieves $94.34 \pm 3.91\%$. Since SVM is not a distance based classifier as kNN there is no need for feature normalization. Using all features plus noise SVM achieves a higher precision of $89.56 \pm 6.01\%$ when compared to kNN with SMOTE that achieves $80.84 \pm 6.96\%$. This precision difference between kNN and SVM for the larger feature combination is due to the metric behavior when the dimensionality is increased [28]. The wKNN in equation (2.7) was idealized to prevent the dimensionality effect by means of assuming that not all k nearest neighbors would contribute equally to define a class assignment. The wkNN with triangular kernel function, optimized k and L_1 distance metric produces a precision of $81.03 \pm 2.12\%$ for 10-fold cross-validation using all features combination plus noise and imbalanced classes. Using same validation process the precision for the combination 3,5 reaches $92.20 \pm 1.72\%$ for

wkNN. Despite the curse of dimensionality that affects kNN more than SVM and wkNN, the kNN with SMOTE achieves higher precision in average compared to these methods without balance aids. For the canonical biomarkers combination, $A\beta_{1-42}$ and p-tau$_{181}$, the SVM reaches $33.34 \pm 0.68\%$ and the wkNN reaches $37.36 \pm 3.02\%$.

4 CONCLUSION

Wrappers techniques for feature selection have shown to be efficient to find the suboptimal combinations given by the rank for all imbalance aid strategies for the proposed problem. However, adding more features to test limits of greedy search could be less successful. Including features is challenging because it increases the number of patients who did not undergo to all examinations. This can be seen in the complete dataset available in ADNI. Fortunately, kNN inspired data missing techniques are available and would be useful to identify more precise combinations that include interpretation benefits for AD mechanisms. However, to deal with improvements to kNN which imply in non-convex optimizations will require more studies. There are options to increase kNN performance such as metric learning [28] and normalization strategies [21]. However, this will increase the parameter space to optimize and improvements in the computation performance. Even with the curse of dimensionality, the kNN with SMOTE overcome in average SVM and wKNN in reproducing the definition. However, the benefits of data balance would increment any classifier precision. The performed all-versus-all strategy requires less classifiers to be built than binary strategies. The class defining features (labels 3,5) increase artificially the class combinations for rank with 502 possibilities, however, it was useful to obtain a comparative. Comparisons using non-cognitive features reveal that FDG contributes more to increase the classification rate. However, more non-cognitive features are needed to observe if dominance for FDG holds, this is a challenge given mentioned data issues for missing values and data imbalance.

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). The author thanks CAPES for the master's scholarship provided.

RESUMO. Biomarcadores são medidas biológicas que ajudam a rastrear e compreender a progressão fisiopatológica de várias doenças. A combinação de diferentes modalidades de biomarcadores muitas vezes permite uma classificação de diagnóstico preciso. Na doença de Alzheimer (DA), os biomarcadores são indispensáveis para identificar indivíduos cognitivamente normais destinados a desenvolver sintomas de demência. No entanto, usando combinações de biomarcadores canônicas DA estudos têm mostrado repetidamente que as taxas de classificação são baixas quando diferenciando entre indivíduos controle, comprometimento cognitivo leve e DA. Além disso, na avaliação de múltiplas combinações de biomarcadores os classificadores enfrentam dificuldade tais como falta de dados e dados desba-

lanceados. Uma vez que o número de combinações biomarcadores é fatorial então usamos wrappers para evitar as múltiplas comparações. Neste trabalho comparamos a capacidade de três técnicas wrapper de selecção de caracteristicas na obtenção de combinações de biomarcadores ao maximizar taxas de classificação. Além disso, como critério para os wrappers usamos o classificador *k*-vizinhos mais próximos com preprosessamento de balanço de dados, sobreamostragem aleatória e sobamostragem (SMOTE). Em geral nossa análise mostra como as combinações de biomarcadores são afetadas pela estratégia de equilíbrio de dados. Mostramos que os biomarcadores não-definidores de classe e não-cognitivos têm menos precisão do que as medidas cognitivas para classificar AD. A nossa abordagem supera em média os classificadores máquina de vetores de suporte e *k*-vizinhos mais próximos ponderado com $94,34 \pm 3,91\%$ de precisão para biomarcadores que definem a classe.

Palavras-chave: *k*-vizinhos mais próximos, SMOTE, seleção de características, biomarcadores de Alzheimer, problema de classificação.

REFERENCES

[1] M.S. Fiandaca, M.E. Mapstone, A.K. Cheema & H.J. Federoff. The critical need for defining preclinical biomarkers in Alzheimer's disease. *Alzheimer's & Dementia*, **10**(3) (2014), S196–S212.

[2] C.R. Jack, D.S. Knopman, W.J. Jagust, L.M. Shaw, P.S. Aisen, M.W. Weiner, R.C. Petersen & J.Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, **9**(1) (2010), 119–128.

[3] R.A. Sperling, P.S. Aisen, L.A. Beckett, D.A. Bennett, S. Craft, A.M. Fagan, T. Iwatsubo, C.R. Jack, J. Kaye & T.J. Montine et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, **7**(3) (2011), 280–292.

[4] A.A. Motsinger-Reif, H. Zhu, M.A. Kling, W. Matson, S. Sharma, O. Fiehn, D.M. Reif, D.H. Appleby, P.M. Doraiswamy & J.Q. Trojanowski et al. Comparing metabolomic and pathologic biomarkers alone and in combination for discriminating Alzheimer's disease from normal cognitive aging. *Acta neuropathologica communications*, **1**(1) (2013), pp. 28.

[5] S.J. Teipel, O. Sabri, M. Grothe, H. Barthel, D. Prvulovic, K. Buerger, A.L. Bokde, M. Ewers, W. Hoffmann & H. Hampel. Perspectives for multimodal neurochemical and imaging biomarkers in Alzheimer's disease. *Journal of Alzheimer's Disease*, **33**(s1) (2013), S329–S347.

[6] C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, (2007).

[7] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, **27**(8) (2006), 861–874.

[8] L. Khedher, J. Ramírez, J.M. Górriz, A. Brahim, F. Segovia & A.D.N. Initiative et al. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented mri images. *Neurocomputing*, **151** (2015), 139–150.

[9] A. Khazaee, A. Ebrahimzadeh & A. Babajani-Feremi. Identifying patients with Alzheimer's disease using resting-state fmri and graph theory. *Clinical Neurophysiology*, **126**(11) (2015), 2132–2141.

[10] Q. Yang & X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, **5**(04) (2006), 597–604.

[11] H. He & E.A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, **21**(9) (2009), 1263–1284.

[12] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, (2016), 1–12.

[13] C. Humpel. Identifying and validating biomarkers for Alzheimer's disease. *Trends in biotechnology*, **29**(1) (2011), 26–32.

[14] I. Guyon & A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, **3**(Mar) (2003), 1157–1182.

[15] Y. Saeys, I. Inza & P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19) (2007), 2507–2517.

[16] K. Lopez-de Ipiña, J.B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C.M. Travieso, M. Ecay-Torres, P. Martinez-Lage & H. Eguiraun. On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, **7**(1) (2015), 44–55.

[17] A. Sarica, G. Di Fatta, G. Smith, M. Cannataro & J.D. Saddy et al. Advanced feature selection in multinominal dementia classification from structural mri data, in *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pp. 82–91 (2014).

[18] J.S. Marques. *Reconhecimento de Padrões: métodos estatistícos e neuronais*. IST press (2005).

[19] T. Tapiola, I. Alafuzoff, S.-K. Herukka, L. Parkkinen, P. Hartikainen, H. Soininen & T. Pirttilä. Cerebrospinal fluid β-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. *Archives of neurology*, **66**(3) (2009), 382–389.

[20] A.W. Toga & K.L. Crawford. The Alzheimer's disease neuroimaging initiative informatics core: A decade in review. *Alzheimer's & Dementia*, **11**(7) (2015), 832–839.

[21] C.-M. Ma, W.-S. Yang & B.-W. Cheng. How the parameters of k-nearest neighbor algorithm impact on the best classification accuracy: In case of parkinson dataset. *Journal of Applied Sciences*, **14**(2) (2014), 171–176.

[22] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, **7**(4) (1946), 401–406.

[23] L. Devroye, L. Györfi & G. Lugosi. *A probabilistic theory of pattern recognition*, **31**. Springer Science & Business Media, **31** (2013), 638 pages.

[24] G. Bhattacharya, K. Ghosh & A.S. Chowdhury. An affinity-based new local distance function and similarity measure for knn algorithm. *Pattern Recognition Letters*, **33**(3) (2012), 356–363.

[25] T.M. Cover & P.E. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, **13**(1) (1967), 21–27.

[26] T. Bailey & A.K. Jain. A Note on Distance-Weighted k-Nearest Neighbor Rules. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-8**(4) (1978), 311–312.

[27] H. Dubey & V. Pudi. Class based weighted k-nearest neighbor over imbalance dataset, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, **7819** (2013), 305–316.

[28] C.C. Aggarwal, A. Hinneburg & D.A. Keim. On the surprising behavior of distance metrics in high dimensional space, in *International Conference on Database Theory*, Springer, (2001), 420–434.

[29] N.V. Chawla, K.W. Bowyer, L.O. Hall & W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16** (2002), 321–357.

[30] R. Kohavi & G.H. John. Wrappers for feature subset selection. *Artificial intelligence*, **97**(1) (1997), 273–324.

[31] L. Scheubert, M. Luštrek, R. Schmidt, D. Repsilber & G. Fuellen. Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. *BMC bioinformatics*, **13**(1) (2012), 266, pp 17.