Submetido para TEMA

Feature selection in Alzheimer's biomarkers using kNN algorithm with SMOTE oversampling [1]

2 3

4

5 6

Abstract. Biomarkers are clinical measures related to disease progress, such quantities combined allow better diagnosis prediction. In order to maximize the prediction rate, feature selection methods seek for suitable subspaces to represent the patterns. Although high dimensional feature spaces demand inaccessible data volume leading then to biased models and time consuming training. In this work we present a comparison of prediction models for Alzheimer's disease obtained by solving a classification problem. In order to this we use the k-nearest neighbors (kNN) rule along with the SMOTE (Synthetic Minority Oversampling TEchnique) preprocessing in a wrapped based scheme to feature search. The effectiveness of these non-parametric techniques are validated in this work for unbalanced datasets that are a challenge in medical applications and machine learning. In the validation process we use confusion matrix combined with 10-fold cross validation. Our results agrees with neurologists hypothesis about biomarkers relevance, identifying potentially discriminant subsets.

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

1. Introduction

Alzheimer's disease is a frequent dementia that affects mainly elderly, and is characterized as the results between neuropathologies which imply in complex conformities [8]. Its spectrum extends from cognitively normal to severe compromised cognition faculties, causing memory and coordination loss and ultimately leading to death [12]. Currently, despite the negative economic impact there is no treatment to revert the neuronal loss, making early stages identification essential to palliative cares

1
2
3
4
5
6

[13]. In the last years, biomarkers have been developed to allow tracking of disease progression and improvement of prediction [8]. Motived by that we propose to find among several subsets of features (biomarkers) those that maximizes the prediction rate, more specifically we want to assign as best as possible test samples to one of following classes, control normal (CN), middle cognitive impairment (MCI) and Alzheimer's disease (AD).

Lately several classifiers and other experiments in Alzheimer's prediction have obtained successful results. Khedher et al. apply support vector machines and principal component analysis in tissue segmented MRI images to classification of the same three class problem [10]. The main difficulty in their work is that images have high dimensionality compared with the available number of training samples, this can potentially degrade the prediction. Using graph theory to analyze the connectivity of different brain regions obtained by fMRI, Khazaee et al. [9] differentiate between AD and CN classes perfectly. Although, separation between extreme cases are straightforward, obstacles arise when overlapping classes, such as MCI, are considered. There is the *Alzheimer's Disease Big Data DREAM Challenge*, divided in sub-challenges and released in June of 2014, which includes, more classes such as Early MCI and Late MCI, and a great variety of features, turning the problem really hard due to the class overlapping and the feature space dimensionality.

In this work, we will use k-nearest neighbors (kNN) classifier that is a nonparametric supervised algorithm. It means there is no assumptions about the distributions of the different classes. kNN is a simple metric based algorithm but competitive when the constrains for efficacy are followed. Here we deal with the unbalanced data using SMOTE (synthetic minority over-sampling technique) to avoid information loss caused by undersampling [4], this technique enforce the decision region of the minor represented data and improve classification. We compare the global prediction rate with greedy fashioned search algorithms for feature selection. The rest of paper is organized as follow: Section 2 we present the data and the methods kNN and SMOTE following by the feature selection and validation procedures, in Section 3 the experiment and results and then conclusions.

2. Methods

2.1. Dataset

In this study we use a sample from ADNI (*Alzheimer's Disease Neuroimaging Initiative*) database. Our features space consist in five neuropsychological tests and three proteomic biomarkers [14]: LM (Logical Memory), ADAS-COG (Alzheimer's Disease Assessment Scale-cognitive), MMSE (Mini Mental State Examination), REY (Rey Auditory Verbal Learning Test), TAU (Total τ protein), ABETA142 (Amyloid Beta 1-42) and PTAU181P (Hyperphosphorylated τ protein). In the table 1 we describe the dataset demographics.

Quite often, clinical datasets come with unbalanced classes due to different probability occurrences in pathologies stages. This is a critical issue that generates unfair separation in decision surface, causing overfitting and worse prediction rates in validation stage. We will discuss further the two main strategies to adjust the proportion between classes without assume any distribution model, namely oversampling and undersampling techniques.

Table 1: Dataset demographics

Features	CN	MCI	AD	label
Male	78	220	58	-
Female	78	181	44	-
Education Years	16.55 ± 2.47	16.18 ± 2.67	15.82 ± 2.62	-
Age	77.48 ± 6.42	75.27 ± 7.52	78.29 ± 8.29	-
ADAs-cog	5.87 ± 3.14	9.28 ± 4.48	20.37 ± 7.45	1
LM	14.52 ± 2.98	9.55 ± 4.05	4.61 ± 2.93	2
REY	12.78 ± 2.27	11.26 ± 3.18	6.50 ± 3.95	3
MoCA	25.87 ± 2.47	23.68 ± 3.30	17.92 ± 4.45	4
MMSE	28.98 ± 1.30	28.05 ± 1.73	23.14 ± 2.14	5
ABETA142	194.96 ± 50.74	176.20 ± 51.78	135.72 ± 38.95	6
TAU	67.28 ± 33.20	86.04 ± 52.64	133.83 ± 65.25	7
PTAU181P	34.14 ± 18.78	40.68 ± 23.49	55.44 ± 28.94	8

2.2. k-Nearest neighbor

Classification problem can be defined using discriminant functions [11]. For now on, consider a classifier $\mathcal{C}(x)$ that assign a pattern $x \in \mathbb{R}^n$ from a vector space to a class into the class space $w \in \Omega := \{1, ..., c\}$,

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

The maximum *a posteriori* classifier (MAP) uses a sequence of discriminant functions to assign to the most probably class. Let $\{f_w\}_{w=1}^c$ be a sequence of discriminant functions. A classifier $\mathcal{C}$ is well defined when for all values in pattern space its possible to assign a class, the MAP classifier is given by

$$\hat{w} := \arg\max_{w \in \Omega} f_w(x) \tag{2.1}$$

The sequence of discriminant functions divide the pattern space in decision regions $\{R_1, ..., R_c\}$. An example of classification problem is depicted in the figure 1 along with two class distributions of the opposite classes.

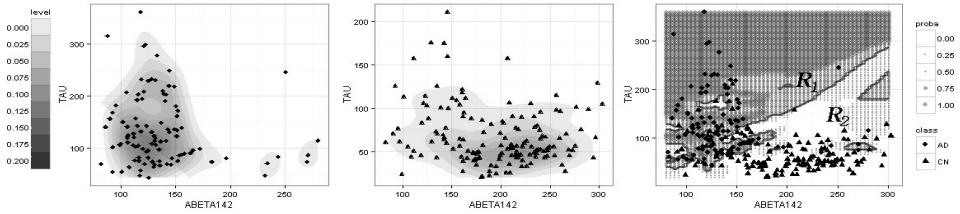

Figure 1: Left and middle - plots of class densities, respectively, the CN and AD, on the right - plot of decision boundaries generated by the 3NN method.

Binary classifiers use the one-against-all strategy in order to be adapted to the multiclass problems. The nearest neighbors strategy can deal with that naturally. Here we define the kNN rule, that uses a neighborhood of k training instances around a test instance for classification. Let $T = \{(x_i, w_i)\}_{i=1}^{m}$ be a training set, with tuples $x_i \in \mathbb{R}^n$ and $w_i \in \Omega$ as training pattern with its known classes and let x_* be a test instance that we want assign to a class in Ω. The MAP classifier of equation 2.1 for the kNN is given by

$$\hat{w}_* = \arg\max_{w \in \Omega} \sum_{x_j \in N(x_*,k)} \delta(w_j, w) \tag{2.2}$$

that is

$$\hat{w}_* = \arg\max \left\{ \sum_{x_j \in N(x_*,k)} \delta(w_j, w_1), ..., \sum_{x_j \in N(x_*,k)} \delta(w_j, w_c) \right\} \tag{2.3}$$

using the discriminant functions monotonicity to equation 2.3 we have

$$\hat{w}_* = \arg\max \left\{ \sum_{x_j \in N(x_*,k)} \frac{\delta(w_j, w_1)}{k}, ..., \sum_{x_j \in N(x_*,k)} \frac{\delta(w_j, w_c)}{k} \right\} \tag{2.4}$$

with this each term in 2.4 is the probability of class assignment

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

given a k-neighborhood and by equation 2.5 we have that the most probable class for the training instance x_* is given by

$$\hat{w}_* = \arg\max_{w \in \Omega} p(w)_{(x_*,k)} \tag{2.6}$$

where $N(x_*, k)$ is a neighborhood with k training instances around x_* and $\delta(.,.)$ is a Kronecker delta. The parameter k that adjusts the neighborhood $N(x_*, k)$ must be searched empirically, Bhattacharyya [2] proposes a bound to the optimal k, that is, $k < \sqrt{m}$, and searching only odd values its possible to avoid ties in (2.1). The

class posterior distribution, $p(\omega_j|x_*)$, that is the probability of class assignment w_i given a pattern x_*, is often used as discriminant function in equation 2.1 within a bayesian framework. The approximation of class posterior distribution obtained locally by equation 2.6, was used to plot the pattern space probability assignment along with the decision boundaries for the knn plots. In the figure 2 we depicted the role of k in the decision boundaries as well the probabilities of class assignments.

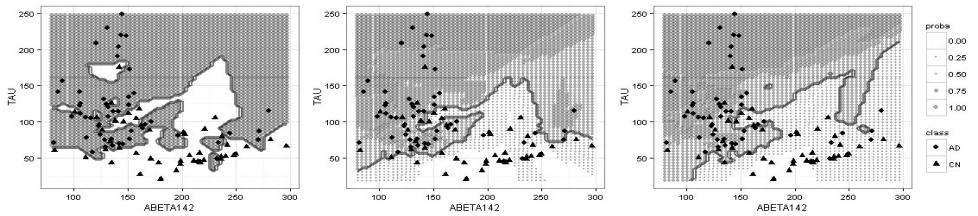

Figure 2: Left to right, the influence of the k-neighborhood in the decision boundaries, for $k = 1$, $k = 5$ and $k = 9$. These classification problems have 50 samples to each class. Also, note that there is only two probabilities in the 1NN pattern space. This happens because there is no misclassification for the training instances causing overfitting.

As any other nonparametric method for classification, the functional in **??** is data depended. In order to aid the effect of data removal, we apply leave-one-out cross validation (LOOCV) to kNN on each search for the optimal k. In the figure 3 we show the shrink effect of unbalanced classes in decision boundaries and data removal.

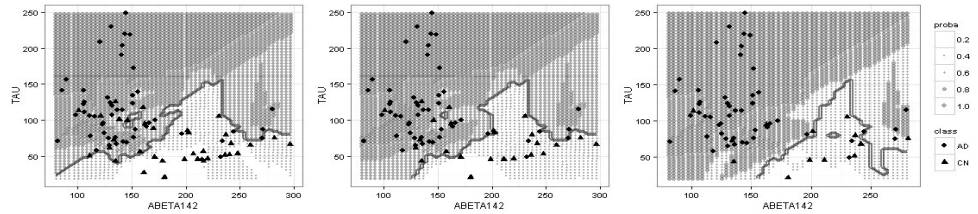

Figure 3: On each plot it was applied 5NN with unbalanced datasets. From the CN class, it was removed respectively 75 %, 50 % and 25 % of the data increasing the decision region of the majority class.

The kNN method does not rely on any assumption about the data structure and distribution. Thus, being employed in many different applications. Although its local nature can suffer by the curse of dimensionality. This is happens when the training patterns become sparse in high dimensions requiring more data to fill out the whole space. There are other formulations to the neighborhood that can overcome the kNN problems in high dimensional spaces, such as the weighted kNN [5]

that account with a weighted scheme $W(.,.)$ as an argument of a kernel function $K(.)$ in order to adjust relative distances between the patterns and avoiding the bias caused by the relative distances.

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

where

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

Instead using a wkNN defined by the equation 2.7 we will use the kNN combined with a preprocessing algorithm called SMOTE to deal with the unbalanced class. SMOTE is a sampling preprocessing step that make the minority class more representative and improve the decision boundaries.

2.3. SMOTE

Sampling techniques are required when there are unbalanced classes in a classification problem. Since the class of interest has few training patterns it turns out to be misrepresented and, thus, has the decision region shrunken.In order to avoid this problem that affects the final outcome, one can remove instances from the major class randomly until achieve the same proportion, or oversample the minor class, literally creating synthetic patterns based on the existing ones. The SMOTE algorithm is a oversampling technique that uses nearest neighbor strategy to add synthetics patterns. The algorithm works as follows [3]: given the minority training patterns, for each pattern, select randomly other pattern in a k-neighborhood and add a synthetic pattern between then, repeat this procedure until achieve the desired proportion.

SMOTE encumbers the parameter space that must be searched to obtain the model with highest prediction rate. To avoid this setting we will the k-neighborhood at $k = 7$. SMOTE when applied in spurious points can generate more of them. There are some moderns versions of this algorithm that is called SMOTE borderline [7], that chooses a secure subsample of the minority class to oversample, avoiding the spurious points propagation. In the figure 4 and 5 we depict respectively, the effect of the parameter k and the proportion size and compare it using Bhattacharyya distance [1] to measure how two distributions differ. Given two normal distributions, $q_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $q_2 = \mathcal{N}(\mu_2, \Sigma_2)$, the Bhattacharyya distance between q_1 and q_2 is given by the following expression

$$d_b(q_1, q_2) = \frac{1}{8}(\mu_1 - \mu 2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\left(\frac{\det(\Sigma)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}}\right)$$

where $\Sigma = (\Sigma_1 + \Sigma_2)/2$.

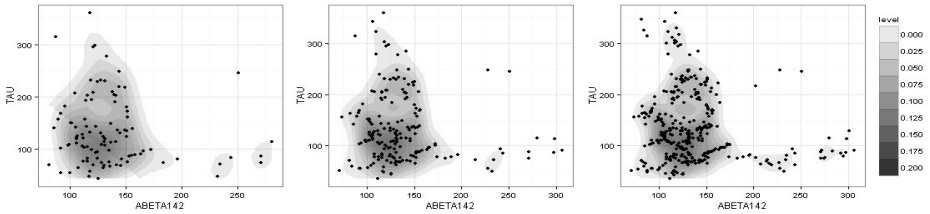

Figure 4: On the left - the original data; on the middle - original data composed with SMOTE for $k = 5$ oversampled 100%; on the right - original data composed with SMOTE for $k = 5$ oversampled 200%. The Bhattacharyya distances to the original distribution are 0.0013571 and 0.0013571.

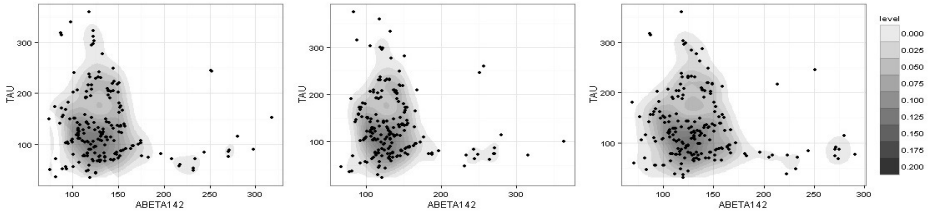

Figure 5: All plots here are compositions of the original data and SMOTE oversampled 100% for differing values of k, namely, $k = 7$, $k = 9$ and $k = 11$, theirs Bhattacharyya's distances from original sample are respectively, 0.0025587, 0.0106067 and 0.0022428.

2.4. Feature selection

Is not uncommon that real problems have hundreds of features, such as in genetics, natural language processing, Parkinson disease, Alzheimer's disease among others, to have hundreds of features to be considered in order to well understanding the problem. Feature selection methods try to answer how many features are needed to make a secure classification. For this purpose, these techniques seek the subset with the most relevant features reducing the dimensionality and overcoming overfitting. There are many notions of relevance, such as informativeness, correlation, statistical significance and others. Here we use *usefulness* for the task, that is those features that maximizes the prediction rate [6]. here are three main strategies for feature selection. Filtering methods that, regardless of the classifier, are suitable for large set of features; wrapper methods are based on search strategies and use the classifier performance and the earlier embedded methods that balance two opposite strategies [6]. We will compare three search strategies for wrappers with global results, backward elimination, forward selection and hill-climbing selection [6]. In figure 6 we depict a example of search graph stages.

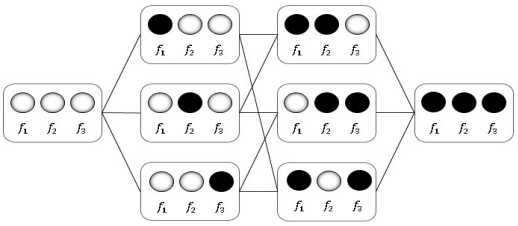

Figure 6: With only 3 features that allows 4 combinations in this simplified example each node represent a subset of features. In the first node at left there is no features and in the last node all features are chosen.

2.5. Validation

Overfitting happens when a model gives reasonable prediction in the training phase but gives poor performance on the testing stage. Validation after the training phase is appropriate to observe if the model can avoid overfitting, that is, how general and applicable the model is. The confusion matrix P provides a way to select the best results and interpreting the classification accuracy. The matrix P is defined bytge probability of classify x_* in the class w_j given that was generated by class w_i , that is, $P(x_* \in w_i | x_* \in w_j)$. With respect to all test samples we denote P only as $P(w_i|w_j)$. Furthermore P is a stochastic matrix,

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

The average of the trace of P is the probability of correct classifications for all classes. We will define this metric in order to have a scalar magnitude to compare to.

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

Here we'll use the 10-fold cross validation in the confusion matrices to obtain the deviations of the models. In this process we take apart 20% of the training sample to use as test sample, the remaining data is preprocessed by SMOTE then is generated a new model, this process is repeated 10 times then is averaged. The data was converted by decimal normalization to improve the prediction as suggested in [4].

3. Results

Figure 7 shows confusion matrices for the best and worse models with the same number of features. When the number of features is increased the difference between the worse and the best is less evident. This effect can be better noticed with more features.

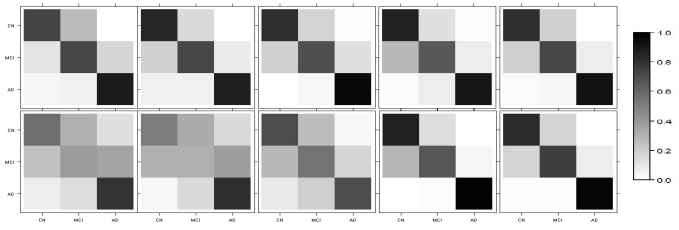

Figure 7: Confusion matrices, in the first line models with the best combinations for 2,3,4,5 and 6 features, in the second line the worst combinations for the same numbers of features.

In table 2, it is shown the prediction rates obtained in the training phase and validation for the best models for each number of features. On this table, there is few significant improvement (0.1958 %) from the best model with 2 features from the best with 6, meaning that use of the 2 features model can save data volume to make a decision. All the wrappers feature selection strategies achieve the highest prediction rate combination with the following scores: backward elimination hits 84.7215 ± 0.0897 %, forward selection hits 84.0648 ± 0.0963 % and hill-climbing hits 83.9983 ± 0.1191 %. Altogether, was optimized 247 models to compose the ranking, with the training phase varying from 52.3857 ± 0.3587 % to 84.2394± 0.0700. As expected, the validation process increases the standard deviation but with more computational resources and data the standard deviation can be reduced. Furthermore the prediction rate can be increased. In the classification problem, some features show no improvement in some combinations, showing that some features are almost completely redundant. It was noted that when the models have few features the size of k-neighborhood tends to be nearly of the limit defined to the search.

Table 2: Dataset demographics

Feature set labels	Rank	Training accuracy (%)	k-neighbors	$val(P)$ (%)
2,1,5,4,7,6	1	84.2394± 0.0700	3	82.2798 ± 1.5906
2,1,5,4,3,7,6	2	83.9650± 0.1325	3	82.7690 ± 2.2888
2,1,5,4,3,7,6,8	5	83.9069± 0.0893	3	82.2664 ± 1.5332
2,1,5,7,6	7	83.6741± 0.1122	3	83.4542 ± 2.0988
2,1,7,6	25	82.3275± 0.1251	5	81.6872 ± 1.7498
2,1,6	49	80.6317± 0.1239	13	79.9043 ± 1.0984
2,1	87	78.4289± 0.1579	25	78.9773 ± 2.1823

3.1. Conclusion

When applying pattern recognition methods, we refer all possibilities of missing features that would contribute for the final outcome. In this sense the relative eval-

uations of relevance are restrict to these features. An interesting finding is that proportionally the neuropsychological tests have higher position in the ranks than proteomical biomarkers. This contrasts with the fact that biomarkers are more objectively measurable than the neuropsychological tests and, in theory, should give better results.

As far as methods go and despite of restrictions with dimensionality and parameters to optimize, once we have the symmetric distance matrix, all computations are done quickly. Thus, it is worth studying the limits in modifications of local methods, for instance, changes of the metric space or weighted schemes for reduction of relative distances. The dimensionality reduction achieved for two features in this work is composed by neuropsychological features, that is more cheap and non contrasting with the proteomical tests.

Further analysis has to be done as far as methods for feature selection goes and in order to better separate the different classes and, therefore, to better classify patients en AD, increasing their way of life.

Resumo. Biomarcadores são medidas clínicas relacionadas com a evolução de doenças, tais quantidades combinadas permitir uma melhor predição do diagnóstico. A fim de maximizar a taxa de predição, os métodos de seleção de características buscam por subespaços adequados para representar os padrões. Entretanto espaços com alta dimensionalidade exigem maior volume de dados, por vezes inacessíveis, levando então a modelos tendenciosos e com treinamento demorado. Neste trabalho apresentamos uma comparação entre modelos de predição para a doença de Alzheimer obtida resolvendo um problema de classificação. Para tal, usamos a regra k-vizinhos mais próximos (*kNN*) pré-processado com SMOTE (*Synthetic técnica Minority Oversampling*) em um esquema de seleção via envólucro para realizar a busca pelas características. A eficácia dessas técnicas não-paramétricas são validados neste trabalho para conjuntos de dados desequilibradas, os quais são um desafio em aplicações médicas. No processo de validação é utilizado matrizes de confusão combinado com validação cruzada 10 vezes. Nossos resultados estão de acordo com as hipóteses dos neurologistas sobre a relevância de alguns grupos de biomarcadores e permite identificar subconjuntos características potencialmente discriminantes.

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

References

[1] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, *Bulletin of the Calcutta Mathematical Society*, **35**, (1943), 99-109

[2] G. Bhattacharyya et al., An affinity-based new local distance function and similarity measure for kNN algorithm, *Pattern Recognition Letters*, **33**, (2012), 356-363.

[3] 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

[4] M. Chih-Min et al., How the Parameter of K-Nearest Neighbour Algorithm Impact on the Best Classification Accuracy: In Case of Parkinson Dataset, *Jornal of Applied Sciences*, **14(2)**, (2014), 171-176.

[5] H. Dubey and V. Pudi, Class Based Weighted K-Nearest Neighbour over Imbalance Dataset, *Advances in Knoledge Discovery and Data Mining, Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, **7819**, (2013), 305-316.

[6] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, **3**, (2003), 1157-1182.

[7] H. Han, W. Wang and B. Mao, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *Advances in Intelligent Computing*, **3644**, (2005), 878-887

[8] C. R. Jack Jr. et al., Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, *The Lancet Neurology*, **9(1)**, (2010), 119-128.

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

[10] L. Khedher, J. Ramírez, J.M. Górrizn, A. Brahim, F. Segovia, the Alzheimer's Disease Neuroimaging Initiative, 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

[11] J. S. Marques et al., "Reconhecimento de padrões: métodos estatísticos e neuronais", IST Press, Lisboa, 12-14, 2005.

[12] G. M. McKhann et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease", *Alzheimer Dement*, **7(3)**, (2011), 263-269.

[13] R. A. Sperling 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 Dement*, **7(3):**, (2011), 280-292.

[14] T. Tapiola et al., Cerebrospinal Fluid β-Amyloid 42 and Tau Proteins as Biomarkers of Alzheimer-Type Pathologic Changes in the Brain, *Arch. Neurol*, **66(3)**, (2009), 382-389.