EFFICIENT SELECTION OF NON-REDUNDANT FEATURES FOR THE DIAGNOSIS OF ALZHEIMER'S DISEASE

Pedro M. Morgado, Margarida Silveira and Jorge S. Marques

Instituto Superior Técnico, Institute for Systems and Robotics, Lisbon, Portugal

ABSTRACT

Recently, a large research effort has been made on the development of discriminative techniques for the computer-aided diagnosis (CAD) of both Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI) using neuroimages as the main source of information. Often, such systems use the Voxel Intensities (VI) directly as features, and a feature selection procedure is needed in order to tackle the *curse of dimensionality*. In this paper, we will propose an efficient selection algorithm based on Mutual Information which, unlike the procedures typically used within this research field, is able to avoid the redundancy existing between brain voxels that are typically highly dependent. The proposed approach was able to join a higher amount of relevant information in a feature vector of fixed dimension and, therefore, was able to improve the classification performance attained when using a typical selection procedure.

Index Terms— Computer-Aided Diagnosis, Alzheimer's Disease, Mild Cognitive Impairment, Minimal Redundancy Maximal Relevance, Positron Emission Tomography, Support Vector Machine

1. INTRODUCTION

Alzheimer's disease is the most common cause of dementia for which no cure is currently available. Consequently, and also due to the demographic ageing and population growth, the number of deaths related to AD is still experiencing a marked increase. Early detection, still at the stage of MCI, is essential for an effective treatment, slowing down the progress of symptoms and improving patients' life quality.

Neuroimages have been extensively explored within the AD research field. They have been used for the automated diagnosis of AD and MCI [1, 2] and even to predict the transition from the latter to the former [3]. In this paper, we will focus on diagnosis (i.e. on the AD vs. CN, MCI vs. CN and AD vs. MCI classification tasks) using the voxel intensities of FDG-PET scans directly as features. However, a problem common to such CAD systems typically appears at the learning stage. In fact, if all voxel intensities were used, the high number of features would probably deteriorate the performance of the diagnostic system due to the *curse of dimensionality* [4]. Consequently, a dimensionality reduction procedure is often explored, which tries to reduce the number of voxels while retaining as much information as possible.

Two categories of dimensionality reduction techniques are typically explored for the diagnosis of AD. The first class, medically driven procedures, uses prior knowledge about the disease. The large majority of these techniques segment the brain into Regions of Interest (ROIs) that are typically associated with atrophy caused by the disease and then use the voxel intensities of each ROI as features [5, 6]. Recently, an innovative selection technique was proposed [7] where the medical expertise was captured by recording the movement of a physician's eyes while performing the diagnosis of several patients in order to subsequently select the voxels that captured most of the physician's attention. The second class restricts itself to the information that can be extracted from the whole brain pattern. Several procedures were already tested for the CAD of AD, such as Principal Component Analysis (PCA) [8, 2], Linear Discriminant Analysis (LDA) [8] and Partial Least Squares (PLS) [9] that reduce the input space dimensionality through a linear combination of the input features, or feature selection techniques [7, 10, 11] which are univariate methods that rank each feature based on some criterion, such as its correlation or mutual information with the class label, and then select the highest ranking ones. Other techniques that do not fall into any of these categories can also be explored. For instance, in [9] a Gaussian Mixture Model was used to model automatically ROIs and, from each Gaussian function, a feature was extracted.

A problem related with the feature selection approach is that selected voxels, or features, may have a lot of redundancy between them. Although several techniques have been proposed to avoid this redundancy, such as Mutual Information Feature Selection (MIFS) [12], MIFS-Uniform (MIFS-U) [13] or minimal Redundancy Maximal Relevance (mRMR) [14], the associated computational burden does not allow their use with neuroimaging data. To deal with this issue, our approach (which is based on mRMR) exploits the fact that neighboring brain voxels are the main source of dependency in order to perform an efficient selection of non-redundant features.

The remainder of this paper is organized as follows: First, the proposed selection procedure (minimal Neighborhood Redundancy Maximal Relevance (mNRMR)) is described in section 2, preceded by a brief review of the mRMR algorithm, for completeness. The system's performance is then presented and discussed in section 3. Section 4 concludes the paper.

2. APPROACH

2.1. Minimal Redundancy Maximal Relevance

mRMR is an established algorithm for feature selection originally proposed by Peng et al. [14]. It is an incremental algorithm, which means that it selects one feature at a time, and avoids choosing redundant features even if they have high discriminative power. Formally, mRMR can be described as follows. Consider two sets of features: the set $\boldsymbol{D}_t$ containing all the features X_i selected at time t

This work was supported by Fundação para a Ciência e Tecnologia (FCT/MCTES) through the ADIAR project (PTDC/SAU-ENB/114606/2009). Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

and the set $\boldsymbol{F}_t$ with the remaining ones. Initially, the set $\boldsymbol{D}_0$ is empty and the set $\boldsymbol{F}_0$ contains all features. Then, at each time step t, mRMR selects from $\boldsymbol{F}_t$ the feature that maximizes the utility function

$$J(X_i) = I(X_i;Y) - \frac{1}{|\boldsymbol{D}_t|} \sum_{X_j \in \boldsymbol{D}_t} I(X_i;X_j), \qquad (1)$$

where $X_i \in \boldsymbol{F}_t$, Y denotes the class label and $I(\cdot\,;\cdot)$ the mutual information between two random variables. The selected feature is then removed from the set $\boldsymbol{F}_t$ and added to $\boldsymbol{D}_t$ and the same procedure is repeated until the desired number of features, N, is reached.

Both mutual information terms in equation (1) can be calculated using the following definition:

$$I(W;Z) = \sum_{w \in \mathcal{W}} \sum_{z \in \mathcal{Z}} P(w,z) \log \frac{P(w,z)}{P(w)P(z)}, \qquad (2)$$

where $\mathcal{W}$ and $\mathcal{Z}$ are the dictionaries containing all possible events of the random variables W and Z, respectively. Density estimation was performed using an histogram approach.

As can be seen, the utility function (1) that mRMR maximizes, not only considers the mutual information between X_i and the class label Y (relevance term), but also the dependency between X_i and all the features X_j already selected (redundancy terms). This property was considered to be very relevant to the problem at hand due to the high correlation nature of neighboring voxels in the brain image. In fact, if a given VI is relevant to a diagnostic problem, some *neighboring* VI is also likely to be. However, since they hold similar information, the inclusion of both will probably not increase the discriminative power of the set of selected features.

2.2. Minimal Neighborhood Redundancy Maximal Relevance

The main setback of mRMR is its time-requirements. In fact, in order to select N features, out of a total number of K features, the information term $I(X_i;X_j)$ must be evaluated $K(N-1) - N(N-1)/2$ times (each one for a different pair of features). When K is small, such time-requirements are not problematic, but since the number of intracranial voxels in the PET image is very high, the selection of a considerable amount of features is unfeasible.

Our approach aims to reduce such timing constrains by reducing substantially the number of information terms $I(X_i;X_j)$ to estimate. It accounts essentially for the terms between features that are known to be mostly dependent, i.e. neighboring features and features in symmetric regions of both brain hemispheres. For simplicity of exposition, we will refer to all such features as *neighbors*. More precisely, the *neighborhood* condition was defined as follows:

$$\{||\mathbf{x}_i - \mathbf{x}_j||_\infty \leq d \quad \vee \quad ||\mathrm{sym}(\mathbf{x}_i) - \mathbf{x}_j||_\infty \leq d\}\,, \qquad (3)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the features' coordinates, $\mathrm{sym}(\mathbf{x}_i)$ is the reflection of $\mathbf{x}_i$ on the opposite hemisphere, and d is a parameter that controls the range of the *neighborhood*. Three examples are given in Fig. 1. Now, if we assume that the mutual information between each feature and all its non-*neighboring* features is constant, we can rewrite the utility function (1) as

$$J'(X_i) = I(X_i;Y) - \frac{1}{|\boldsymbol{D}_t|} \left(P\hat{I}_{\mathrm{nn}} + \sum_{X_j \in \{\boldsymbol{N}_i \cap \boldsymbol{D}_t\}} I(X_i;X_j) \right), \qquad (4)$$

where $\boldsymbol{N}_i$ is the set of all *neighbors* of the feature X_i, $\hat{I}_{\mathrm{nn}}$ is an estimation of the mutual information between two non-*neighboring* voxels and P is the number of features currently selected which are not *neighbors* of X_i, i.e. $P = |\{\neg \boldsymbol{N}_i \cap \boldsymbol{D}_t\}|$. The estimation of $\hat{I}_{\mathrm{nn}}$ was conducted beforehand as the mean mutual information obtained from 100000 pairs of non-*neighboring* feature locations randomly selected throughout the 3D brain image.

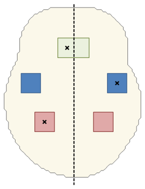


Fig. 1. *Neighbor* regions of three features. The three regions are differentiated by color and contain their respective feature. Despite the 2D representation, the *neighborhood* is three-dimensional.

2.3. Learning Machine

After selection, each one of the three diagnostic problems was learned using the SVM algorithm [15] with a linear kernel. The good generalization that this algorithm achieves in high dimensional spaces made SVM very popular within the neuroimaging based CAD research field, which was an important factor in our choice.

3. EXPERIMENTS

3.1. Dataset

All PET images were retrieved from the ADNI database, but the following restrictions were imposed to the Clinical Dementia Rating (CDR) score of each subject: 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for AD patients, resulting in a dataset composed by 59, 104 and 70 subjects, respectively. However, since the SVM algorithm can be affected by unbalanced datasets, we opted to restrict each group to only 59 subjects and selected them randomly. Table 1 summarizes important clinical and demographic information about each group.

The retrieved data had already undergone a series of pre-processing steps in order to minimize differences between images and thus allowing voxel-wise comparisons. More specifically, every PET image was co-registered, averaged, reoriented (such that the anterior-posterior axis of the subject is parallel to the AC-PC line), normalized in its intensity, and smoothed to a uniform standardized resolution[1]. An additional pre-processing step was conducted by the University of Utah Component of the ADNI PET Core, where an anatomic standardization to the Talairach brain atlas was performed using the Neurostat software [16].

Moreover, extracranial voxels were left out from all subsequent processing using a simple threshold operation over the average brain image. Note that extracranial voxels do not hold valuable information and thus they would never be chosen in the feature selection procedure, but the time required to perform the selection would increase.

3.2. Experimental Design

In this section, an experimental motivation for the proposed algorithm (mNRMR) will firstly be given and then its performance and

[1] A more detailed description of the pre-processing stage is available at http://adni.loni.ucla.edu/methods/pet-analysis/pre-processing/

Table 1. Characteristics of each group. Format: Mean (Standard Deviation). MMSE stands for Mini Mental State Exam.

Attributes	AD	MCI	CN
N^o of subjects	59	59	59
Age	78.3 (6.6)	77.7 (6.9)	77.4 (6.6)
Sex (% of Males)	57.6	67.8	64.4
MMSE	19.6 (5.1)	25.8 (3.0)	29.2 (0.9)

time-requirements will be compared with other related procedures for a varying number of selected features. Specifically, it will be compared with mRMR and with an algorithm that only considers the relevance term of the utility function (1). This procedure, which is often called Mutual Information Maximization (MIM), computes $I(X_i;Y)$ for each feature and then selects the N best ranked ones.

The three algorithms were studied for different numbers of features. Specifically, N was tested for all values in the set $\{10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000\}$, except for the mRMR procedure where the same progression was used but N was only allowed to be as high as 500 in order to be able to evaluate the system in an acceptable time. The parameter C of the SVM classifier (which controls the cost of misclassification) was tuned within the range 2^{-18} to 2^{18} using a 10×10-fold nested cross-validation procedure, guaranteeing unbiased estimates of the system's performance. The results were averaged after 10 runs. All comparisons were performed on the three binary diagnostic problems (AD vs. CN, MCI vs. CN and AD vs. MCI).

The parameter d that controls the range of the *neighborhood* was fixed to 4 voxels, i.e. 6 mm (given the 1.5 mm cubic voxels that form the PET image). In fact, a few considerations were taken into account when setting this parameter. On one hand, d needs to be large enough in order to compensate for the significant inter-hemispheric variability, and for possible errors in the localization of the symmetry plane and in the registration step. However, on the other hand, the size of the *neighborhoods* considered should also be small enough so that the proposed approach can perform the selection with acceptable execution time.

3.3. Results

Fig. 2 shows the distribution of the mutual information between *neighboring* and non-*neighboring* features, where each distribution was obtained from 100000 pairs of feature locations, randomly selected throughout the 3D brain image, and respecting each imposed spatial condition. In addition, the mutual information was discretized into intervals of 0.05. One can easily notice that the mutual information between non-*neighboring* voxels is usually lower and less spread than when *neighboring* voxels are considered, whether they are direct or symmetric *neighbors*. More precisely, the statistics $(\mu \pm \sigma) = (0.215 \pm 0.052)$ were obtained for non-*neighbors*, $(\mu \pm \sigma) = (0.408 \pm 0.201)$ for direct *neighbors* and $(\mu \pm \sigma) = (0.323 \pm 0.129)$ for symmetric *neighbors*. These distributions led us to conclude that most of the variability of $I(X_i;X_j)$ occurs between *neighbors*occurs between neigbours, which motivated the proposed simplification (i.e. assuming all terms $I(X_i;X_j)$ between non-*neighbors* to be constant and equal to its mean value: $\hat{I}_{\text{nn}} = 0.215$).

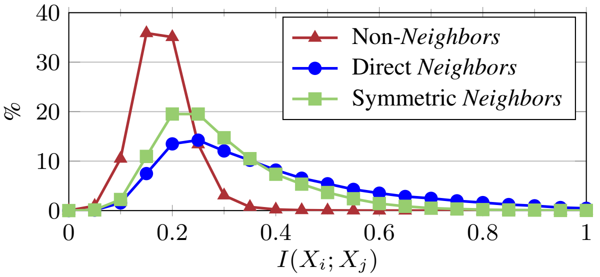
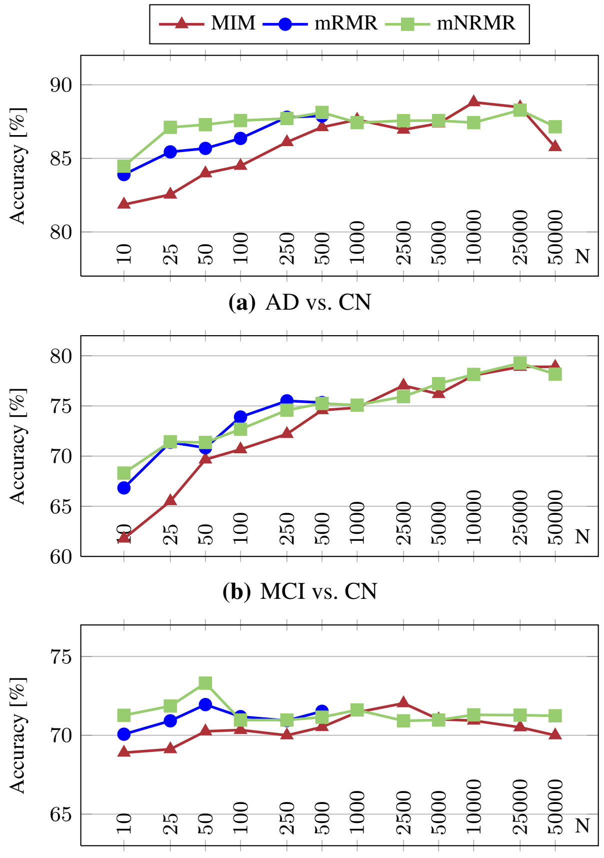


Fig. 2. Histogram of the mutual information between *neighboring* and non-*neighboring* pairs of features.



(a) AD vs. CN

(b) MCI vs. CN

(c) AD vs. MCI

Fig. 3. Comparison of feature selection algorithms for varying number of features, N.

The performance of each system applied to the three classification problems is presented in Fig. 3. First, it should be noted that mRMR outperformed MIM for corresponding number of features in the three classification problems, but since it could only be evaluated up to 500 features, its marks were always exceeded in higher dimensional spaces. This is consistent with its theoretical advantage, i.e. mRMR accounts for redundancy between selected features and, thus, it certainly joins a higher amount of information in a feature set of a given size. mNRMR also shares with mRMR this advantage over MIM and, consequently, it was also able to improve MIM's classification accuracy, especially in low dimensional spaces. In addition, the computational advantage of the novel approach over mRMR allowed us to perform classifications with larger numbers of features. This was essential in the diagnosis of MCI (Fig. 3(b)) where mNRMR achieved the best performance (79.27% acc. with $N = 25000$), although very close to MIM. However, in this classification problem, the power of selection appeared to be weaker than on the other two, since even after the selection of a large number of features, there was still a large amount of discriminative information that has not been included, and thus the results could still be improved with the inclusion of more features. As for the classification of AD (Fig. 3(a)), both mRMR and mNRMR outperformed MIM in low dimensional spaces, but the difference was attenuated with the increase of N. Most likely, this is related to the fact that, despite the different selection criteria, the methods ended up on agreeing on the best features, when large numbers of them are selected. For instance, when $N = 50000$, more than 90% of the features selected by MIM and mNRMR for the diagnosis of AD were actually the same. This phenomenon was observed in all diagnostic problems. Finally, in the AD vs. MCI classification task (Fig. 3(c)), the number of selected features did not influence the system's performance much, allowing for the new approach to achieve the best performance (73.31% acc.) using only 50 features.

The time-requirements of each procedure were also analyzed. The computation time of both mNRMR and mRMR increase approximately linearly with the number of features to select. However, mRMR took approximately $2.9\,\text{s}$ to choose every additional feature, while mNRMR took only $13\,\text{ms}$. This difference made the new approach feasible for values of N as high as 50000. Remember that the time-requirements of the feature selection stage is even more important because it has to be repeated several times due to the nested cross-validation procedure used for performance assessment. All simulations were performed under the same platform on an Intel$^{\circledR}$ Core$^{\text{TM}}$ i7-2600 processor running at 3.4 GHz.

4. CONCLUSION

This paper proposed an efficient algorithm for the selection of relevant but non-redundant features in neuroimages by taking into account the fact that most of the redundancy and most of the mutual information variability occur between *neighbors*, either local neighbors or neighbors in symmetric regions of brain hemispheres. The proposed procedure, which can be seen as a simplification of the well known mRMR selection algorithm, was able to speed up significantly the selection process and, therefore, to choose a larger number of features in the same period of time.

Regarding classification performance, mNRMR reported interesting results, outperforming MIM in low dimensional spaces (500 features or less) in all diagnostic problems tested and attaining similar results to mRMR. In fact, it was even able to improve mRMR's classification accuracy for very small numbers of selected features (50 or less) in the AD vs. CN and AD vs. MCI classification tasks, while using only a small fraction of mRMR's computing time. As for large values of N (500 or more), the proposed approach and MIM achieved always similar accuracies.

Overall, mNRMR achieved the best performance in the MCI vs. CN and AD vs. MCI classification tasks but the differences for the second best were never expressive enough to confidently claim its superiority. Nevertheless, the results presented in this paper (in low dimensional spaces) indicate that a careful selection of features, eliminating redundant voxel intensities, is capable of boosting the performance of a system for the CAD of AD and related disorders.

It should also be stressed that, although this work has focused on FDG-PET images, the proposed selection procedure has the potential to be used with any other neuroimaging modality, since the high correlation nature of neighboring voxels is common to all such brain images.

5. REFERENCES

[1] Margarida Silveira and Jorge Marques, "Boosting Alzheimer disease diagnosis using PET images," *Pattern Recognition (ICPR'10), Proceedings of the 2010 20th International Conference on*, pp. 2556–2559, 2010.

[2] Simon Duchesne, Anna Caroli, Cristina Geroldi, Christian Barillot, Giovanni B. Frisoni, and D. Louis Collins, "MRI-based automated computer classification of probable AD versus normal controls," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 4, pp. 509–520, 2008.

[3] Jieping Ye, Michael Farnum, Eric Yang, Rudi Verbeeck, Victor Lobanov, Nandini Raghavan, Gerald Novak, Allitia Dibernardo, and Vaibhav A Narayan, "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data," *BMC Neurol*, vol. 12, no. 1, pp. 46, 2012.

[4] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.

[5] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[6] Arthur Mikhno, Pablo Martinez Nuevo, Davangere P. Devanand, Ramin V. Parsey, and Andrew F. Laine, "Multimodal classification of dementia using functional data, anatomical features and 3D invariant shape descriptors," *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*, pp. 606– 609, 2012.

[7] E. Bicacro, M. Silveira, J. S. Marques, and D. C. Costa, "3D brain image-based diagnosis of Alzheimer's disease: Bringing medical vision into feature selection," *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*, pp. 134–137, 2012.

[8] M. López, J. Ramírez, J.M. Górriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia, and R. Chaves, "Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers," *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 3190–3193, 2009.

[9] F. Segovia, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, and R. Chaves, "A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database," *Neurocomputing*, vol. 75, no. 1, pp. 64–71, 2012.

[10] R. Chaves, J. Ramírez, J. M. Górriz, M. López, I. Álvarez, D. Salas-Gonzalez, F. Segovia, and P. Padilla, "SPECT image classification based on NMSE feature correlation weighting and SVM," *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, pp. 2715–2719, 2009.

[11] P. Padilla, M. López, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 207–216, 2012.

[12] Roberto Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, 1994.

[13] Nojun Kwak and Chong-Ho Choi, "Improved mutual information feature selector for neural networks in supervised learning," *Neural Networks, 1999. International Joint Conference on*, vol. 2, pp. 1313–1318, 1999.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1226–1238, 2005.

[15] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[16] Center for Alzheimer's Care Imaging and Research, "Summary of Neurostat processed images uploaded to LONI," Tech. Rep., Univ of Utah, Nov 2007, [Online] Available: `http://www.loni.ucla.edu/twiki/pub/ADNI/ADNIPostProc/UUtah_Analysis.pdf`.