Feature Selection with a Genetic Algorithm for Classification of Brain Imaging Data

Annamária Szenkovits, Regina Meszlényi, Krisztian Buza, Noémi Gaskó, Rodica Ioana Lung, Mihai Suciu

Abstract Recent advances in brain imaging technology, coupled with large-scale brain research projects, such as the BRAIN initiative in the U.S. and the European Human Brain Project, allow us to capture brain activity in unprecedented details. In principle, the observed data is expected to substantially shape our knowledge about brain activity, which includes the development of new biomarkers of brain disorders. However, due to the high dimensionality, the analysis of the data is challenging, and selection of relevant features is one of the most important analytic tasks. In many cases, due to the complexity of search space, evolutionary algorithms are appropriate to solve the aforementioned task. In this chapter, we consider the feature selection task from the point of view of classification tasks related to functional magnetic resonance imaging (fMRI) data. Furthermore, we present an empirical comparison of conventional LASSO-based feature selection and a novel feature selection approach designed for fMRI data based on a simple genetic algorithm.

Key words: functional magnetic resonance imaging (fMRI), functional connectivity, classification, feature selection, mild cognitive impairment, biomarker

Annamária Szenkovits, Noémi Gaskó, Rodica Ioana Lung, Mihai Suciu
Centre for the Study of Complexity, Babeş-Bolyai University
str. Kogalniceanu, Nr.1, Cluj-Napoca, Romania, e-mail: {szenkovitsa,gaskonomi}@cs.ubbcluj.ro,rodica.lung@econ.ubbcluj.ro,mihai-suciu@cs.ubbcluj.ro

Regina Meszlényi
Department of Cognitive Science, Budapest University of Technology and Economics
Egry József utca 1, 1111 Budapest, Hungary, and
Brain Imaging Centre, Research Centre for Natural Sciences, Hungarian Academy of Sciences
Magyar tudósok krt. 2, 1117 Budapest, Hungary, e-mail: meszlenyi.regina@ttk.mta.hu

Krisztian Buza
Knowledge Discovery and Machine Learning, Institute für Informatik III, Rheinische Friedrich-Wilhelms-Universität Bonn
Römerstr. 164, 53117 Bonn, Germany, e-mail: buza@biointelligence.hu

1 Introduction

Advanced brain imaging technology allows us to capture brain activity in unprecedented details. The observed data is expected to substantially shape our knowledge about the brain, its disorders, and to contribute to the development of new biomarkers of its diseases, including *mild cognitive impairment* (MCI). MCI represents a transitional state between the cognitive changes of normal aging and very early dementia and is becoming increasingly recognized as a risk factor for Alzheimer disease (AD) [14].

The brain may be studied at various levels. At the neural level, the anatomy of neurons, their connections and spikes may be studied. For example, neurons responding to various visual stimulii (such as edges) as well as neurons recognizing the direction of audio signals have been identified [49, 50]. Despite these spectacular results, we note that *C. Elegans*, having only 302 neurons in total, is the only species for which neural level connections are fully described [43]. Furthermore, in this respect, *C. Elegans* is extremely simple compared with many other species. For example, the number of neurons in the human brain is approximately 100 billion and each of them has up to 10,000 synapses [16]. Imaging neural circuits of that size is difficult due to various reasons, such as diversity of cells and limitations of traditional light microscopy [27].

While neurons may be considered as the elementary components of the brain, at the other end, psychological studies focus on the behavior of the entire organism. However, due to the large number of neurons and their complex interactions, it is extremely difficult to establish the connection between low-level phenomena (such as the activity of neurons) and high-level observations referring to the behavior of the organism. In fact, such connections are only known in exceptional cases: for example, deprivation to visual stimuli causes functional blindness due to modified brain function, in other words: the brain does not "learn" how to see, if no visual input is provided [20, 46].

In order to understand how the brain activity is related to phenotypic conditions, many recent studies follow an "explicitly integrative perspective" resulting in the emergence of the new domain of "network neuroscience" [2]. While there are brain networks of various spatial and temporal scale, in this study we focus on networks describing *functional connectivity* between brain regions. Two regions are said to be functionally connected if their activations are synchronized. When measuring the activity of a region, usually, the aggregated activity of millions of neurons is captured as function of time. Spatial and temporal resolution (number of regions and frequency of measurements) depend on the experimental technology. In case of functional magnetic resonance imaging (fMRI), the spatial resolution is currently around fifty-thousand voxels (i.e., the brain activity is measured in approximately fifty-thousand cubic areas of the brain), while the temporal resolution is between 0.5 and 2 seconds. Roughly speaking, the raw data consists of approximately fifty-thousand time series, each one corresponding to one of the voxels in which the brain activity is measured.

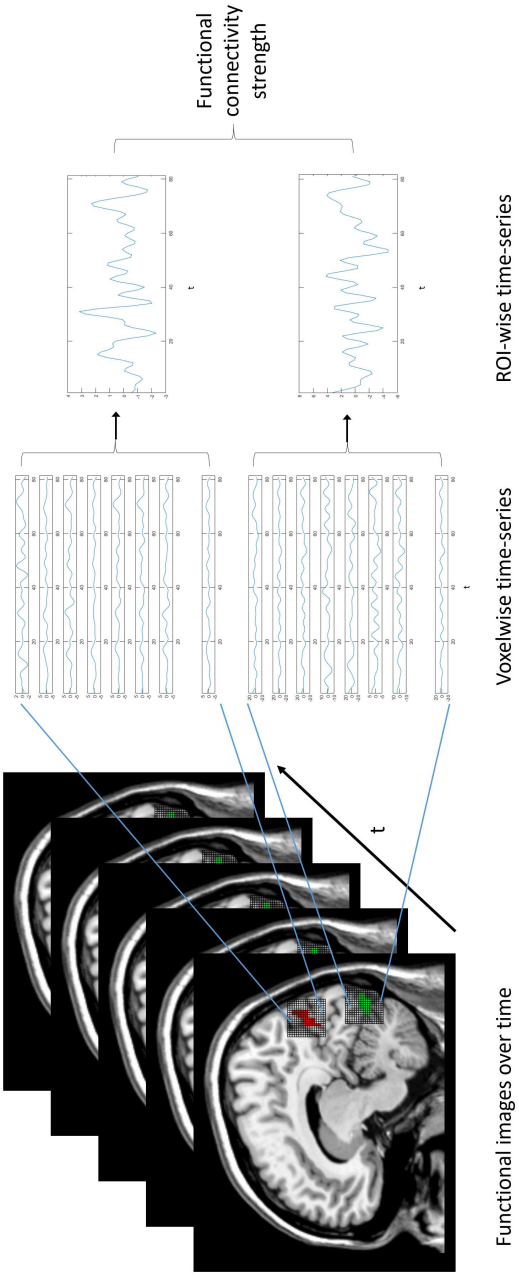

Fig. 1 Schematic illustration of data acquisition and preprocessing.

When analyzing fMRI data, after the necessary preprocessing steps, voxels are organized into regions of interest (ROIs). The time series of the voxels belonging to a ROI are aggregated into a single time series representing the activation of the ROI as function of time. Next, functional connectivity strength between ROIs can be calculated. This process is illustrated in Fig. 1. Traditionally, functional connectivity strength is described with linear correlation coefficients between the time series associated with the ROIs. However, according to recent results, more complex relationships can be represented with other time series distance measures such as dynamic time warping (DTW) [30].

The functional brain network can be represented with the connectivity matrix of the aforementioned ROIs, i.e., by the functional connectivity strength between each pair of ROIs. Functional connectivity matrices show remarkable similarity between subjects, however some connectivity patterns may be characteristic to various conditions and disorders such as gender, age, IQ, and schizophrenia, addiction or cognitive impairment, see e.g. [28, 29] and the references therein.

While the connectivity features can be seen as a compact representation of brain activity compared with the raw data, the dimensionality of the resulting data is still very high, making feature selection essential for subsequent analysis. Therefore, in this study we consider the task of feature selection, which can be described as follows: given a set of features, the goal is to select a subset of features that is appropriate for the subsequent analysis tasks, such as classification of instances. In many cases, due to the complexity of search space, evolutionary algorithms are appropriate to solve this task.

In the last decade, extensive research has been performed on evolutionary algorithms for feature selection. In this chapter, we will consider the feature selection task with special focus on its applications to functional magnetic resonance imaging (fMRI) data. Furthermore, we will present an empirical comparison of conventional feature selection based on the "Least Absolute Shrinkage and Selection Operator" (LASSO) and a novel feature selection approach designed for fMRI data based on a minimalistic genetic algorithm (mGA). Finally, we point out that feature selection is essential for the successful classification of fMRI data which is a key task in developing new biomarkers of brain disorders.

2 Materials and Methods

In this section we provide details of the dataset and preprocessing (Section 2.1), the feature selection methods we consider: LASSO (Section 2.2) and mGA (Section 2.3). Subsequently, we describe our experimental protocol in Section 2.4.

2.1 Data and Preprocessing

We used publicly available data from the Consortium for Reliability and Reproducibility (CoRR) [52], in particular, the LMU 2 and LMU 3 datasets [3, 4]. The datasets contain fourty-nine subjects (22 males, age (mean $\pm$ SD): 68.6 $\pm$ 7.3 years, 25 diagnosed with mild cognitive impairment (MCI)), each subject participated at several resting-state fMRI measurement sessions, thus the total number of measurement sessions is 146. In each measurement session, besides high resolution anatomical images, 120 functional images were collected over 366 sec, resulting in time-series of length 120 for every brain voxel. The dataset was collected at the Institute of Clinical Radiology, Ludwig-Maximilians-University, Munich, Germany. For further details on how the data was obtained we refer to the web page of the data: `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/lmu_2.html`.

Preprocessing of the raw data includes motion-correction, identification of gray matter (GM) voxels, and the application of low-pass and high-pass filters. For a detailed description of the preprocessing pipeline, see [30].

We used the Willard functional atlas of FIND Lab, consisting of 499 functional regions of interest (ROIs) [34] to obtain 499 functionally meaningful averaged blood-oxygen-level dependent (BOLD) signals in each measurement. This allows us to calculate ROI-based functional connectivity as follows: we calculate the pairwise dynamic time warping (DTW) distances [36] between the aforementioned 499 BOLD signals, resulting in $499 \times 498/2 = 124251$ connectivity features. We obtained these connectivity features for each of the 146 measurement sessions, leading to a dataset of 146 instances and 124251 features. From the publicly available phenotypic data, mild cognitive impairment (MCI) was selected as classification target.

Given the relatively low amount of instances, selection of relevant features (i.e., the ones that are characteristic for the presence or absence of mild cognitive impairment) is a highly challenging task to any of the feature selection techniques.

2.2 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator is widely used for the analysis of high-dimensional data, including brain imaging data. Therefore, we review this technique next.

We are given a dataset $\mathbf{X} \in \mathbb{R}^{N \times d}$ containing N instances with d features. Each instance x_i (the i-th row of $\mathbf{X}$) is associated with a label y_i, the vector y contains all the labels: $y = (y_1, \ldots, y_N)$. We aim at finding a function $f(x)$ in the form $f(x) = \sum_{j=1}^{d} \theta^{(j)} x^{(j)}$, where $x^{(j)}$ is the j-th component of vector x, and $\forall j : \theta^{(j)} \in \mathbb{R}$, so that the function fits the data, i.e., $f(x_i) \approx y_i$ for all (x_i, y_i) pairs. For simplicity, we use $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$ to denote the vector of all the parameters.

The above task can be considered as an ordinary least squares (OLS) regression problem, where the objective is to find the parameter vector θ^* that minimizes the sum of squared errors:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} ||y - \mathbf{X}\theta||_2^2 \tag{1}$$

In the case when $N \geq d$ (and matrix $\mathbf{X}$ has a full column rank), i.e., when we have more training instances than features, the optimal θ^* vector exists and is unique. However, if the number of features exceeds the number of available training instances, matrix $\mathbf{X}$ loses its full column rank, therefore the solution is not unique anymore. In this case, the model tends to overfit the dataset $\mathbf{X}$, in the sense that it fits not only to the "regularities" of the data, but also to measurement noise while it is unlikely to fit to unseen instances of a new dataset $\mathbf{X}'$.

To be able to choose the "correct" parameter vector from the numerous possibilities, one must assume some knowledge about the parameters, and use that in the optimization. The most common solution of this problem is to add a regularization term to the function we try to minimize, called *objective function*. In case of the well-known ridge-regression method [17] we assume that the Euclidean-norm (L_2-norm) of the θ vector is small, and the objective is to find the parameter vector θ^* that minimizes the sum of squared errors and the regularization term:

$$\theta^* = \arg\min_{\theta} \left(\frac{1}{N} ||y - \mathbf{X}\theta||_2^2 + \lambda ||\theta||_2^2 \right) \tag{2}$$

where $\lambda \in \mathbb{R}$ is a hyperparameter controlling the regularization, i.e., in case of $\lambda = 0$ the method is equivalent to the OLS, but as we increase the λ value, the L_2-norm of the θ vector has to decrease to minimize the objective function.

However, in cases when we can hypothesize that only *some* of all the available features have influence on the labels, the above regularization based on the L_2-norm of θ may not lead to an appropriate solution θ^*. In particular, ridge-regression method results in low absolute value weights for the features, while it tends to "distribute" the weights between features, i.e., in most cases almost all of the features will receive nonzero weights. In contrast, regularization with L_1-norm usually results in zero weights for many features, therefore this method can be seen as a feature selection technique: it selects those features for which the associated weights are not zero, while the others are not selected. This method is called LASSO [42].

Formally, LASSO's objective is to find the parameter vector θ^* that minimizes the sum of squared errors and the L_1 regularization term:

$$\theta^* = \arg\min_{\theta} \left(\frac{1}{N} ||y - \mathbf{X}\theta||_2^2 + \lambda ||\theta||_1 \right) \tag{3}$$

where $\lambda \in \mathbb{R}$ is a hyperparameter controlling the sparsity of the resulting model, i.e. the number of weights that are set to zero.

As mentioned above, in brain imaging studies the number of instances is usually much lower than the number of features, therefore LASSO may be used. Indeed, it has been shown to be successful for classification tasks related to brain

networks [28, 29, 35] in cases where the number of features is 10 to 50 times larger than the number of instances. In [28] and [29] the features selected by LASSO, i.e., functional connectivity strength values, did not only lead to good classification performance, but the selected connections showed remarkable stability through the rounds of cross-validation and resulted in well-interpretable networks that contain connections that differ the most between groups of subjects. The task we consider in this study is even more challenging compared with the tasks considered in [28] and [29], because the number of features is more than 800 times higher than the number of available instances.

2.3 Minimalist Genetic Algorithm for Feature Selection

Evolutionary Algorithms (EAs) represent a simple and efficient class of nature inspired optimization techniques [10, 18]. In essence, EAs are population based stochastic techniques that mimic natural evolution processes to find the solution for an optimization problem. The main advantages of these approaches are their low complexity in implementation, their adaptability to a variety of problems without having specific domain knowledge, and effectiveness [13, 31, 37].

There are many variants of EAs with different operators (e.g. selection, recombination, mutation, etc.) and control parameters such as population size, crossover and mutation probabilities [12]. A set of random solutions is evolved with the help of some variation operators and only good solutions will be kept in the population with the help of selection for survivor operators. Solutions are evaluated by using a fitness function constructed depending on the problem. In most cases it represents the objective of the optimization problem, and guides the search to optimal solutions by preserving during the selection process solutions with high (for maximization problems) fitness values. Based on the encoding and operators used, the main classes of evolutionary algorithms are considered to be: Genetic algorithms, Genetic programming (tree structure encoding), Evolution strategies (real-valued encoding), and Evolutionary programming (typically uses only mutation) [12].

Genetic algorithms (GAs) use binary encoding and specific mutation and crossover operators [12]. A potential solution evolved within a GA is referred to as an individual, encoded as a chromosome which is composed of genes - and represented in its simplest form as a string of 0 and 1. Encoding/decoding of an individual is problem dependent.

As the feature selection problem aims at selecting the relevant features from a large number of features [45], evolutionary algorithms seem to be an appropriate choice for tackling it because they can deal with the search space generated by the high number of features [21, 45]. Evaluation is performed by using either the classification accuracy or some convex combination between various indicators, such as classification accuracy, number of features, overall classification performance, class specific accuracy, and Pearson correlation [6, 39, 44, 47]. Multi-objective op-

timization algorithms have also been used to find trade-offs between the number of features and classification accuracy as two conflicting objectives [19, 25].

Genetic Algorithms are a natural choice for approaching the feature selection problem due to the encoding used: in the binary string each value shows if the corresponding feature is selected or not. Consequently, there are many studies that employ various variants of genetic algorithms for solving feature selection problems in fields such as computer science, engineering, medicine, biochemistry, genetics and molecular biology, chemistry, decision sciences, physics and astronomy, pharmacology, toxicology and pharmaceutics, chemical engineering, business, management and accounting, agricultural and biological sciences, materials science, earth and planetary sciences, social sciences and humanities. For example, in [24] a genetic algorithm is used to select the best feature subset from 200 time series features and use them to detect premature ventricular contraction - a form of cardiac arrhythmia. In [7] a genetic algorithm was used to reduce the number of features in a complex speech recognition task and to create new features on machine vision task. In [33], a genetic algorithm optimizes a weight vector used to scale the features.

Recently, genetic algorithms have been used for feature selection and classification of brain related data. Problems approached include brain tumor classification [15, 26], EEG analysis [22, 23], and seizure prediction [9]. Genetic algorithms are designed as stand-alone feature selection methods [32] or as part of a more complex analysis combined with simulated annealing [23, 26], neural networks [9, 15, 41] or support vector machines [22, 26, 41]. To the best of our knowledge, genetic algorithms have not yet been used for feature selection on fMRI data.

In the following we propose a minimalist version of a genetic algorithm for mining features in the brain data. The *Minimalist Genetic Algorithm* (mGA) evolves one binary string encoded individual by using only uniform mutation for a given number of generations in order to improve the classification accuracy while restricting the number of selected features. In what follows, the mGA is described in detail.

Encoding

mGA uses bit string representation of length $L = 124251$ – equal to the total number of features – where 1 means that a feature is selected, and 0 means that the certain feature is not selected:

$$\underbrace{010\ldots100}_{\text{length: L=124251}} \tag{4}$$

Initialization

In the first generation a random initial solution is generated by assigning each gene a value of 1 with probability p_{in}. The probability p_{in} controls number of selected features in the initial individual, it is problem dependent, and thus can be set according to domain knowledge.

Evaluation of the fitness function

The aim of the search is to maximize the classification accuracy while avoiding overfitting. Thus we construct a fitness function based on accuracy, considering also as a penalization the proportion of the selected features to the length of the chromosome (individual). By combining both accuracy and number of selected features, we hope to achieve a trade-off between them and to avoid overfitting. In particular, the fitness f of individual I is computed as:

$$f(I) = A(I) - \frac{n_I}{L} \tag{5}$$

where $A(I)$ is the classification accuracy and n_I is the number of features selected in I (the number of 1s in the binary representation). We divide the number of selected features n_I by the total length of the chromosome to keep the two terms in Eq. (5) in $[0,1]$ and maintain a balance between them.

Mutation

mGA uses the uniform random mutation for variation by which each gene is modified with a probability p_m. In detail, for each gene a random number between 0 and 1 is generated following a uniform distribution; if this value is less than the given mutation probability p_m, the value of the gene is modified (from 1 to 0 or conversely).

In the following example, the second gene of individual I in the left is modified from 0 to 1 as the second random number generated is 0.001:

$$I = 0\mathbf{1}\ldots 10 \xrightarrow[rand():0.236,\mathbf{0.001},\ldots,0.385,0.798]{} I' = 0\mathbf{0}\ldots 10. \tag{6}$$

Outline of mGA

The main steps of mGA are outlined in Algorithm 1. First, a random initial solution I is generated. Then, the following steps are repeated until the stopping criterion is met: (i) creation of an offspring I' by the application of mutation to the individual I representing the current solution, (ii) if the offspring has a higher fitness value, we keep it as the new current solution, otherwise we do not change the current solution. The process stops after a predefined number of iterations, denoted by *MaxGen*.

Parameters

mGA uses the following parameters: p_{in}: the probability of each gene being set to 1 when generating the random individual in the first generation, p_m: the probability used for uniform mutation and the number of generations (*MaxGen*).

Algorithm 1 Outline of mGA

```
Initialize random individual (I);
Evaluate I;
nrGen = 0;
while nrgen < MaxGen do
    Apply mutation to I → I' ;
    Evaluate I';
    if f(I') > f(I) (I' better than I) then
        I = I';
    end if
    nrgen++;
end while
```

2.4 Experimental set-up

The goal of our experiments was to compare the mGA with LASSO in terms of their ability to select relevant features. In order to objectively compare the selected feature sets, and to quantitatively assess their quality, we aimed to classify subjects according to the presence or absence of mild cognitive impairment (MCI) using the selected features. In other words: we evaluated both algorithms indirectly in context of a classification task. Next, we describe the details of our experimental protocol.

As measurements from the same subjects are not independent, we performed our experiments according to the leave-one-subject-out cross-validation schema. As our dataset contains measurements from 49 subjects, we had 49 rounds of cross-validation. In each round, the instances belonging to one of the subjects were used as *test data*, while the instances of the remaining 48 subjects were used as *training data*. We performed feature selection with both methods (i.e., the mGA and LASSO) using the training data.[1] Subsequently, both sets of selected features were evaluated.

For LASSO, we set the parameter λ to 0.005 in order to restrict the number of selected features to be around the number of instances (154 ± 5.1).[2] The parameter values used for mGA are presented in Tab. 1. The values of p_m and p_{in} were set empirically to limit the number of selected features (starting with approx. 500 and increasing only if better sets are generated) in order to avoid overfitting caused by selecting unnecessarily large feature sets.

Table 1 mGA Parameter settings

Parameter	p_m	p_{in}	$MaxGen$
Value	0.004	0.004	3000

[1] The accuracy for the fitness function of mGA was calculated solely on the training data. In particular we measured the accuracy of a nearest neighbor classifier in an *internal* 5-fold cross-validation on the training data.

[2] We note that $\lambda = 0.001$ and $\lambda = 0.0001$ led to very similar classification accuracy. For simplicity, we only show the results in case of $\lambda = 0.005$ in Section 3.

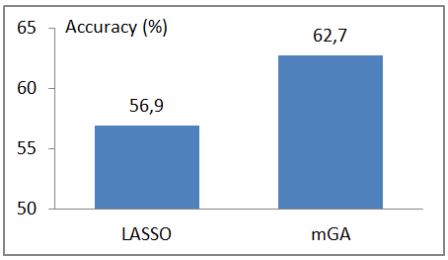

Fig. 2 Classification accuracy using features selected by LASSO and mGA

In order to assess the quality of a selected feature set, we classified test instances with a simple 1-nearest neighbor [1] classifier with Euclidean distance based on the selected features only. That is: we restrict test and training instance vectors to the selected feature set, and for every test instance x we search for its *nearest neighbor* in the training data. With nearest neighbor of the test instance x we mean the training instance that is the closest to x according to the Euclidean distance. The predicted label of the test instance x is the label of its nearest neighbor.

For the task-based evaluation of feature sets in context of the classification according to MCI described above, in principle, one could use various classifiers such as recent variants of neural networks [40] or decision rules [51], however, we decided to use the nearest neighbor classifier, because our primary goal is to compare the set of selected features and nearest neighbor relies highly on the features. Furthermore, nearest neighbor is well-understood from the theoretical point of view and works surprisingly well in many cases [5, 8, 11].

To demonstrate how our classifiers perform compared to the chance level, we generated 100 000 random labeling with "coin-flipping" (50-50% chance of generating the label corresponding to the presence or absence of MCI), and calculated the accuracy values of these random classifications. The 95th percentile of this random classifier's accuracy-distribution is 56.8%. Therefore, we treat the classification as significantly better than random "coin-flipping" if its accuracy exceeds the threshold of 56.8%.

3 Results

Classification accuracy

The classification accuracy of 1-nearest neighbour classifier based on LASSO-selected and mGA-selected feature sets are presented in Fig. 2.

For both LASSO-selected and mGA-selected feature sets, the classification accuracy is significantly better than the accuracy of "coin-flipping". Furthermore, in this task, where the number of features is extremely high compared with the number

of instances, the mGA-based feature selection clearly outperforms LASSO-based feature selection in terms of classification accuracy.

Stability of the selected features

A good feature selection algorithm should result not only in good classification performance, but an interpretable feature set as well. The two approaches we consider in this study, LASSO and mGA, differs greatly in the number of selected features (see Tab. 2). The LASSO algorithm selects about 154 features with a low standard deviation, while the mGA algorithm chooses about 5 times more features with large standard deviation.

Table 2 Mean $\pm$ standard deviation of selected feature set sizes through the 49 rounds of cross-validation with the two methods

LASSO	mGA
154.3 ± 5.1	775.4 ± 249.4

The selected feature sets can be examined from the point of view of stability, i.e., we can calculate how many times each feature was selected during the 49 rounds of cross-validation (Fig. 3. A). As features describe connections between brain ROIs, we can also calculate how many times each ROI was selected through the cross-validation (Fig. 3. B).

In terms of the stability of features, the difference between the two algorithms is undeniable (Fig. 3. A). Clearly, the feature set selected by LASSO is very stable compared with the mGA-selected features, as there are 47 connections that were selected in at least 40 rounds (about 80%) out of all the 49 rounds of cross-validation, while in case of the mGA algorithm, there is no feature that was selected more than 6 times. Interestingly, the distinction between the two algorithms almost disappear if we consider the stability of selected ROIs. In Fig. 3. B one can see that both algorithms identify a limited number (less than five) hub ROIs, that have considerably more selected connections, than the rest of the brain.

Runtime

In our experiments, the runtime of a single evaluation in mGA was $\approx$3.5 seconds on an Intel® Core™ i7-5820K CPU @ 3.30GHz $\times$ 12, 32GB RAM, with the settings from Table 1. The total runtime is influenced by the parameter settings (*MaxGen*, p_m and p_{in}) as they control the number of evaluations and the number of selected features involved in the evaluation of the classification accuracy in (5). Using the MATLAB-implementation of LASSO, on average, $\approx$4 seconds were needed for the selection of features in each round of the cross-validation. As mGA requires multiple evaluations, the total runtime of LASSO is much lower than that of mGA.

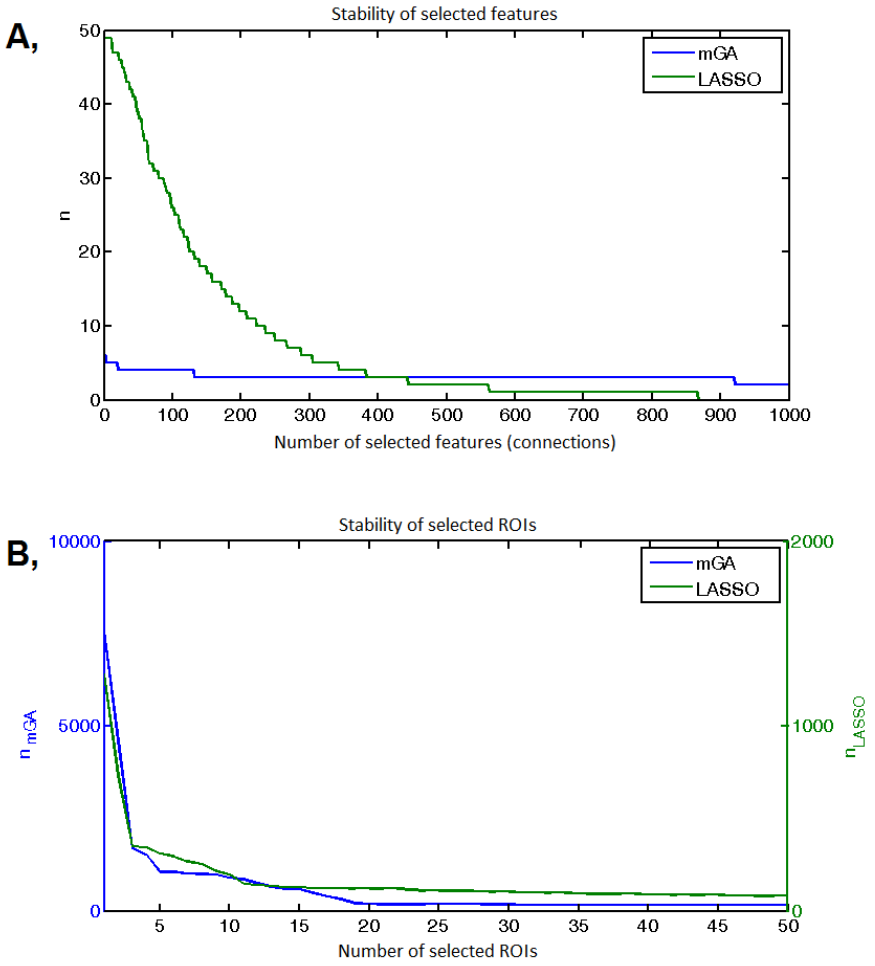

Fig. 3 **A**, The stability of selected features for LASSO and mGA. Considering the 49 rounds of leave-one-subject-out cross-validation, for each $n = 1,\ldots,49$, we count how many features appear at least n-times among the selected features. The vertical axis shows n, while the horizontal axis displays the number of features that appear at least n-times among the selected features. **B**, The stability of selected ROIs for LASSO and mGA. Considering all 49 rounds of cross-validation, we count how many times each ROIs is selected in total (the selection of a feature corresponds to the selection of two ROIs; as several features are associated with the same ROI, a ROI may be selected multiple times: e.g., if both features $f_{1,2} = (r_1, r_2)$ and $f_{1,3} = (r_1, r_3)$ were selected in all the 49 rounds of cross-validation, and no other features were selected, ROI r_1 would appear $49 \times 2 = 98$ times in total, while r_2 and r_3 would appear 49 times in total). The left and right vertical axes show n_{mGA} and n_{LASSO}, while the horizontal axis shows how many ROIs were selected at least n_{mGA}-times and n_{LASSO}-times, respectively. (As mGA selects about 5 times more features in each round of the cross-validation, there is a scaling factor of 5 between the two vertical axes.)

4 Discussion

The results show that even in the case of extremely high number of features (number of features is more than 800 times higher than the number of instances), both LASSO and mGA algorithms are able to classify subjects according to presence or absence of MCI significantly better than "coin flipping". In terms of accuracy, mGA clearly outperformed LASSO. In contrast, the set of features selected by LASSO is substantially more stable. With respect to the stability of selected ROIs, in our experiments, the two algorithms resulted in very similar characteristics.

The differences regarding stability may be attributed to inherent properties of the algorithms: if two features are similarly useful for fitting the model to the data, but one of them is slightly better than the other, due to its regularization term, LASSO tends to select the better feature, while mGA may select any of them with similar probabilities.

The most frequently selected ROIs can be important from a clinical point of view as well. The top five ROIs of the two algorithms show a significant overlap (see Tab. 3).

Table 3 The top five selected ROIs in case of LASSO and mGA

	LASSO			mGA	
	ROI ID	region		ROI ID	region
1	143	Cuneal Cortex	1	143	Cuneal Cortex
2	144	Occipital Pole, Lingual Gyrus	2	144	Occipital Pole, Lingual Gyrus
3	147	Left Lateral Occipital Cortex, superior division	3	145	Right Precentral Gyrus
			4	24	Left Frontal Pole
4	149	Cerebellum	5	146	Frontal Medial Cortex, Subcallossal Cortex
5	145	Right Precentral Gyrus			

The top 20% of ROIs are visualized in Fig. 4 A and B. One can note that while the most important ROIs i.e. the hubs of the two methods are the same, the LASSO based map is more symmetric, i.e. it respects strong homotopic (inter-hemispheric) brain connections, while the mGA based map shows clear left hemisphere dominance. Most importantly, several out of the top 20% ROIs selected by both the LASSO and the mGA, have been reported in meta-studies examining Alzheimers disease and MCI [38, 48].

5 Conclusions

The analysis of brain imaging data is a challenging task, because of the high dimensionality. Selecting the relevant features is a key task of the analytic pipeline. We considered two algorithms, the classic LASSO-based feature selection, and a novel genetic algorithm (mGA) designed for the analysis of functional magnetic

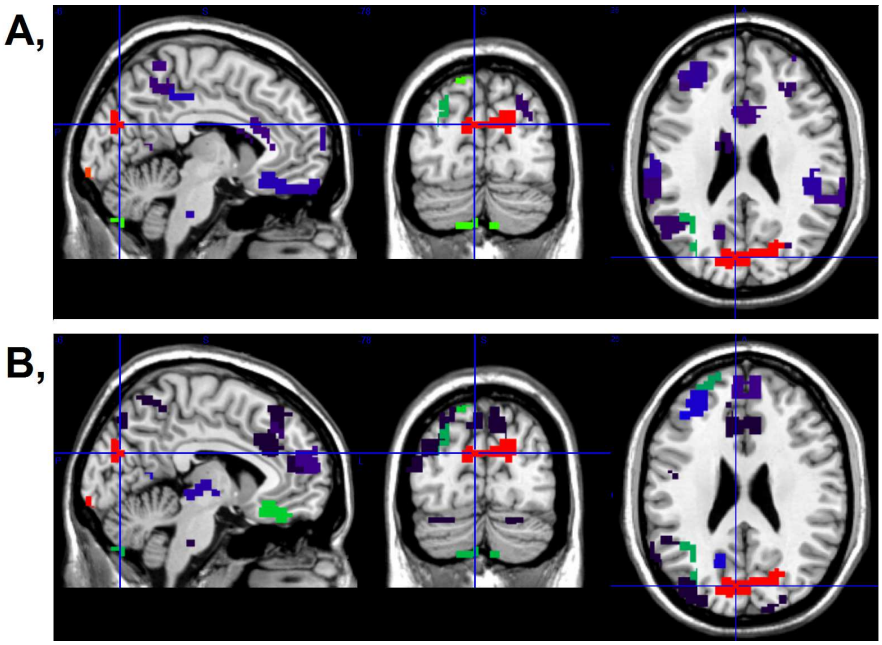

Fig. 4 The top 20% of selected ROIs using LASSO (A), and mGA (B). Warm colors represent more frequently chosen ROIs (hubs).

resonance imaging (fMRI) data. We compared them in context of the recognition of mild cognitive impairment (MCI) based on fMRI data. In terms of classification accuracy, the features selected by mGA outperformed the features selected by LASSO. According to our observations, the set of features selected by LASSO is more stable over multiple runs. Nevertheless, both methods provide meaningful information about the data, confirming the search potential of genetic algorithms and providing a starting point to further and deeper analyses of brain imaging data by heuristic methods.

Acknowledgements This work partially was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS UEFISCDI, project number PN-II-RU-TE-2014-4-2332 and the National Research, Development and Innovation Office (Hungary), project number: NKFIH 108947 K.

References

1. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician **46**(3), 175–185 (1992)
2. Bassett, D.S., Sporns, O.: Network neuroscience. Nature Neuroscience **20**(3), 353–364 (2017)

3. Blautzik, J., Keeser, D., Berman, A., Paolini, M., Kirsch, V., Mueller, S., Coates, U., Reiser, M., Teipel, S.J., Meindl, T.: Long-term test-retest reliability of resting-state networks in healthy elderly subjects and patients with amnestic mild cognitive impairment. Journal of Alzheimer's Disease **34**(3), 741–754 (2013)
4. Blautzik, J., Vetter, C., Peres, I., Gutyrchik, E., Keeser, D., Berman, A., Kirsch, V., Mueller, S., Pöppel, E., Reiser, M., et al.: Classifying fmri-derived resting-state connectivity patterns according to their daily rhythmicity. NeuroImage **71**, 298–306 (2013)
5. Buza, K., Nanopoulos, A., Schmidt-Thieme, L.: Time-series classification based on individualised error prediction. In: 13th International Conference on Computational Science and Engineering, pp. 48–54. IEEE (2010)
6. Canuto, A.M.P., Nascimento, D.S.C.: A genetic-based approach to features selection for ensembles using a hybrid and adaptive fitness function. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2012). DOI 10.1109/IJCNN.2012.6252740
7. Chang, E.I., Lippmann, R.P.: Using genetic algorithms to improve pattern classification performance. In: Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3, NIPS-3, pp. 797–803. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1990)
8. Chen, G.H., Nikolov, S., Shah, D.: A latent source model for nonparametric time series classification. In: Advances in Neural Information Processing Systems, pp. 1088–1096 (2013)
9. D'Alessandre, M., Vachtseyanos, G., Esteller, R., Echauz, J., Sewell, D., Litt, B.: A systematic approach to seizure prediction using genetic and classifier based feature selection. In: International Conference on Digital Signal Processing, DSP, vol. 2 (2002). DOI 10.1109/ICDSP.2002.1028162
10. De Jong, K.: Evolutionary Computation: A Unified Approach. Bradford Book. Mit Press (2006)
11. Devroye, L., Gyorfi, L., Krzyzak, A., Lugosi, G.: On the strong universal consistency of nearest neighbor regression function estimates. The Annals of Statistics pp. 1371–1385 (1994)
12. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing, 2nd edn. Springer Publishing Company, Incorporated (2015). DOI 10.1007/978-3-662-44874-8
13. de la Fraga, L.G., Coello Coello, C.A.: A Review of Applications of Evolutionary Algorithms in Pattern Recognition, pp. 3–28. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). DOI 10.1007/978-3-642-22407-2_1
14. Grundman, M., Petersen, R.C., Ferris, S.H., Thomas, R.G., Aisen, P.S., Bennett, D.A., Foster, N.L., Jack Jr, C.R., Galasko, D.R., Doody, R., et al.: Mild cognitive impairment can be distinguished from alzheimer disease and normal aging for clinical trials. Archives of neurology **61**(1), 59–66 (2004)
15. Gwalani, H., Mittal, N., Vidyarthi, A.: Classification of brain tumours using genetic algorithms as a feature selection method (GAFS). In: ACM International Conference Proceeding Series, vol. 25-26-Augu (2016). DOI 10.1145/2980258.2980318
16. Herculano-Houzel, S.: The human brain in numbers: a linearly scaled-up primate brain. Frontiers in human neuroscience **3**, 31 (2009)
17. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
18. Holland, J.H.: Adaptation in natural and artificial systems. MIT Press, Cambridge, MA, USA (1992)
19. de la Hoz, E., de la Hoz, E., Ortiz, A., Ortega, J., Martnez-lvarez, A.: Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps. Knowledge-Based Systems **71**, 322 – 338 (2014). DOI http://dx.doi.org/10.1016/j.knosys.2014.08.013
20. Hyvärinen, J., Carlson, S., Hyvärinen, L.: Early visual deprivation alters modality of neuronal responses in area 19 of monkey cortex. Neuroscience letters **26**(3), 239–243 (1981)
21. de la Iglesia, B.: Evolutionary computation for feature selection in classification problems. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **3**(6), 381–407 (2013). DOI 10.1002/widm.1106

22. Jalili, M.: Graph theoretical analysis of Alzheimer's disease: Discrimination of AD patients from healthy subjects. Information Sciences **384** (2017). DOI 10.1016/j.ins.2016.08.047
23. Ji, Y., Bu, X., Sun, J., Liu, Z.: An improved simulated annealing genetic algorithm of EEG feature selection in sleep stage. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016 (2017). DOI 10.1109/APSIPA.2016.7820683
24. Kaya, Y., Pehlivan, H.: Feature selection using genetic algorithms for premature ventricular contraction classification. In: 2015 9th International Conference on Electrical and Electronics Engineering (ELECO), pp. 1229–1232 (2015). DOI 10.1109/ELECO.2015.7394628
25. Khan, A., Baig, A.: Multi-objective feature subset selection using non-dominated sorting genetic algorithm. Journal of Applied Research and Technology **13**(1), 145 – 159 (2015). DOI http://dx.doi.org/10.1016/S1665-6423(15)30013-4
26. Kharrat, A., Halima, M., Ben Ayed, M.: MRI brain tumor classification using Support Vector Machines and meta-heuristic method. In: International Conference on Intelligent Systems Design and Applications, ISDA, vol. 2016-June (2016). DOI 10.1109/ISDA.2015.7489271
27. Lichtman, J.W., Denk, W.: The big and the small: challenges of imaging the brains circuits. Science **334**(6056), 618–623 (2011)
28. Meszlényi, R., Peska, L., Gál, V., Vidnyánszky, Z., Buza, K.: Classification of fmri data using dynamic time warping based functional connectivity analysis. In: Signal Processing Conference (EUSIPCO), 2016 24th European, pp. 245–249. IEEE (2016)
29. Meszlényi, R., Peska, L., Gál, V., Vidnyánszky, Z., Buza, K.A.: A model for classification based on the functional connectivity pattern dynamics of the brain. In: Third European Network Intelligence Conference, pp. 203–208 (2016)
30. Meszlényi, R.J., Hermann, P., Buza, K., Gál, V., Vidnyánszky, Z.: Resting state fmri functional connectivity analysis using dynamic time warping. Frontiers in Neuroscience **11**, 75 (2017)
31. Michalewicz, Z.: Evolutionary Algorithms in Engineering Applications, 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1997)
32. Noori, F., Qureshi, N., Khan, R., Naseer, N.: Feature selection based on modified genetic algorithm for optimization of functional near-infrared spectroscopy (fNIRS) signals for BCI. In: 2016 2nd International Conference on Robotics and Artificial Intelligence, ICRAI 2016 (2016). DOI 10.1109/ICRAI.2016.7791227
33. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation **4**(2), 164–171 (2000). DOI 10.1109/4235.850656
34. Richiardi, J., Altmann, A., Milazzo, A.C., Chang, C., Chakravarty, M.M., Banaschewski, T., Barker, G.J., Bokde, A.L., Bromberg, U., Büchel, C., et al.: Correlated gene expression supports synchronous activity in brain networks. Science **348**(6240), 1241–1244 (2015)
35. Rosa, M.J., Portugal, L., Hahn, T., Fallgatter, A.J., Garrido, M.I., Shawe-Taylor, J., Mourao-Miranda, J.: Sparse network-based models for patient classification using fmri. Neuroimage **105**, 493–506 (2015)
36. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (1978)
37. Sanchez, E., Squillero, G., Tonda, A.: Industrial Applications of Evolutionary Algorithms. Springer-Verlag Berlin Heidelberg (2012). DOI 10.1007/978-3-642-27467-1
38. Schroeter, M.L., Stein, T., Maslowski, N., Neumann, J.: Neural correlates of alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. Neuroimage **47**(4), 1196–1206 (2009)
39. da Silva, S.F., Ribeiro, M.X., do E.S. Batista Neto, J., Traina-Jr., C., Traina, A.J.: Improving the ranking quality of medical image retrieval using a genetic feature selection method. Decision Support Systems **51**(4), 810 – 820 (2011). DOI http://dx.doi.org/10.1016/j.dss.2011.01.015. Recent Advances in Data, Text, and Media Mining & Information Issues in Supply Chain and in Service System Design
40. Stańczyk, U.: On performance of drsa-ann classifier. In: International Conference on Hybrid Artificial Intelligence Systems, pp. 172–179. Springer (2011)

41. Tajik, M., Rehman, A., Khan, W., Khan, B.: Texture feature selection using GA for classification of human brain MRI scans, vol. 9713 (2016). DOI 10.1007/978-3-319-41009-8_25
42. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
43. White, J.G., Southgate, E., Thomson, J.N., Brenner, S.: The structure of the nervous system of the nematode caenorhabditis elegans. Philos Trans R Soc Lond B Biol Sci **314**(1165), 1–340 (1986)
44. Winkler, S.M., Affenzeller, M., Jacak, W., Stekel, H.: Identification of cancer diagnosis estimation models using evolutionary algorithms: A case study for breast cancer, melanoma, and cancer in the respiratory system. In: Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation, GECCO '11, pp. 503–510. ACM, New York, NY, USA (2011). DOI 10.1145/2001858.2002040
45. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2016). DOI 10.1109/TEVC.2015.2504420
46. Yaka, R., Yinon, U., Rosner, M., Wollberg, Z.: Pathological and experimentally induced blindness induces auditory activity in the cat primary visual cortex. Experimental brain research **131**(1), 144–148 (2000)
47. Yang, J., Honavar, V.G.: Feature subset selection using a genetic algorithm. IEEE Intelligent Systems **13**(2), 44–49 (1998). DOI 10.1109/5254.671091
48. Yang, J., Pan, P., Song, W., Huang, R., Li, J., Chen, K., Gong, Q., Zhong, J., Shi, H., Shang, H.: Voxelwise meta-analysis of gray matter anomalies in alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. Journal of the neurological sciences **316**(1), 21–29 (2012)
49. Ye, C.q., Poo, M.m., Dan, Y., Zhang, X.h.: Synaptic mechanisms of direction selectivity in primary auditory cortex. Journal of Neuroscience **30**(5), 1861–1868 (2010)
50. Yoshor, D., Bosking, W.H., Ghose, G.M., Maunsell, J.H.: Receptive fields in human visual cortex mapped with surface electrodes. Cerebral cortex **17**(10), 2293–2302 (2007)
51. Zielosko, B., Chikalov, I., Moshkov, M., Amin, T.: Optimization of decision rules based on dynamic programming approach. In: Innovations in Intelligent Machines-4, pp. 369–392. Springer (2014)
52. Zuo, X.N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., et al.: An open science resource for establishing reliability and reproducibility in functional connectomics. Scientific data **1** (2014)