Journal of Neuroscience Methods 299 (2018) 45–54

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Clustering fMRI data with a robust unsupervised learning algorithm for neuroscience data mining

Hadeel K. Aljobouri[a,b,*], Hussain A. Jaber[a], Orhan M. Koçak[c], Oktay Algin[d,e], Ilyas Çankaya[a]

[a] Electrical and Electronics Engineering Department, Graduate School of Natural Science, Ankara Yıldırım Beyazıt University, Ankara, Turkey
[b] Biomedical Engineering Department, College of Engineering, Al-Nahrain University, Baghdad, Iraq
[c] Psychiatry Department, School of Medicine, Kırıkkale University, Kırıkkale, Turkey
[d] Department of Radiology, Ataturk Training and Research Hospital, Ankara Yıldırım Beyazıt University, Ankara, Turkey
[e] National MR Research Center, Bilkent University, Ankara, Turkey

HIGHLIGHTS

- A novel application of the robust unsupervised learning approach is proposed in the current study. Robust growing neural gas (RGNG) algorithm was fed into fMRI data and compared with growing neural gas (GNG) algorithm, which has not been used for this purpose or any other medical application.
- Learning algorithms proposed in the current study are fed with real and free auditory fMRI datasets.
- Another comparison was conducted with the model-based (hypothesis) data analysis approach using the statistical parametric mapping (SPM) package, which is based on the general linear model.
- The fMRI result obtained by running RGNG was within the expected outcome and is similar to those found with the hypothesis method in detecting active areas within the expected auditory cortices.
- Results show that the fMRI application of the presented RGNG approach is clearly superior to other approaches in terms of its insensitivity to different initializations and the presence of outliers, as well as its ability to determine the actual number of clusters successfully, as indicated by its performance measured by minimum description length (MDL) and receiver operating characteristic (ROC) analysis.

ARTICLE INFO

Article history:
Received 17 December 2017
Received in revised form 13 February 2018
Accepted 14 February 2018
Available online 20 February 2018

Keywords:
Clustering technique
Data mining
Growing neural gas (GNG)
Robust growing neural gas (RGNG)

ABSTRACT

Background: Clustering approaches used in functional magnetic resonance imaging (fMRI) research use brain activity to divide the brain into various parcels with some degree of homogeneous characteristics, but choosing the appropriate clustering algorithms remains a problem.
New method: A novel application of the robust unsupervised learning approach is proposed in the current study. Robust growing neural gas (RGNG) algorithm was fed into fMRI data and compared with growing neural gas (GNG) algorithm, which has not been used for this purpose or any other medical application. Learning algorithms proposed in the current study are fed with real and free auditory fMRI datasets.
Results: The fMRI result obtained by running RGNG was within the expected outcome and is similar to those found with the hypothesis method in detecting active areas within the expected auditory cortices.
Comparison with existing method(s): The fMRI application of the presented RGNG approach is clearly superior to other approaches in terms of its insensitivity to different initializations and the presence of outliers, as well as its ability to determine the actual number of clusters successfully, as indicated by its performance measured by minimum description length (MDL) and receiver operating characteristic (ROC) analysis.
Conclusions: The RGNG can detect the active zones in the brain, analyze brain function, and determine the optimal number of underlying clusters in fMRI datasets. This algorithm can define the positions of the center of an output cluster corresponding to the minimal MDL value.
© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Functional magnetic resonance imaging (fMRI) is a powerful tool used by neuroscientists to examine brain activity by calculating the levels of oxygen in the blood. Blood oxygenation level dependent

* Corresponding author at: Biomedical Engineering Department, College of Engineering, Al-Nahrain University, Baghdad, Iraq.
E-mail address: hadeel_bme77@eng.nahrainuniv.edu.iq (H.K. Aljobouri).

https://doi.org/10.1016/j.jneumeth.2018.02.007
0165-0270/© 2018 Elsevier B.V. All rights reserved.

(BOLD) signal represents the ratio of oxygenated to deoxygenated hemoglobin measurements in the blood and is closely related to neural activity. FMRI considers metabolic function in measuring neural activity because it determines the hemodynamic response function (HRF) or metabolic demands (oxygen consumption) in the brain or spinal cord (Aljobouri et al., 2015).

FMRI is used to understand neuronal mechanisms behind many disorders, such as bipolar disorder, schizophrenia, Parkinson's disease, autism spectrum disorders, and Alzheimer's disease.

The fMRI dataset is acquired from a scanner machine in the form of raw data as sequences of 3D images because of the variations of voxel intensities over time. With different experimental conditions, the acquired fMRI data are formed as a combination of BOLD signal changes and noises or artifacts. These artifacts are attributed to hardware systems (the MRI scanner itself), individuals themselves (e.g., head motion), or physiological effects.

Clustering techniques in fMRI research are considered model-free or exploratory data analysis approaches. These techniques can define the active zones and find structures in the brain and fMRI data competently without the need for prior knowledge about activation patterns or experiments. However, choosing the appropriate clustering algorithms remains a problem. Independent component analysis (ICA) and principal component analysis (PCA) algorithms are regarded as fine methods to separate fMRI signals into a group of defined components. These algorithms cannot easily predict occurrences during acquisition and have limitations in terms of independence and orthogonality, respectively (Korczak, 2012). Various clustering algorithms are applied in fMRI for data mining instead of the previous classical methods, which cannot easily predict occurrences during acquisition. The classical methods include K-means, fuzzy classification, hierarchical classifications, Linde–Buzo–Gray (LBG), clustering using representatives (CURE), neural models Kohonen's self-organizing map (SOM), neural gas (NG), and Fritzke's growing neural gas (GNG) algorithms. However, one of the main problems of fMRI clustering algorithms is deciding the number of clusters as an input (Dimitriadou et al., 2004; Wismuller et al., 2004). Results with a high level of interpretation were obtained using clustering approaches, but these approaches are associated with high cost in terms of computing time and memory space (Bock and Diday, 2000; Lindquist, 2008; Goutte et al., 1999; Baumgartner et al., 1998; Liao et al., 2008; Katwal, 2011; Pereira et al., 2009).

The GNG algorithm exhibits the best clustering performance and produce robustness; however, this algorithm has limitations associated with the sensitivity for initialization (choosing a set of neuron vectors), the order of input vectors, and the existence of many outliers (Qin and Suganthan, 2004). Therefore, a novel application, which relies on using the robust growing neural gas (RGNG) algorithm with fMRI datasets, is proposed to detect the active zones in the brain. This algorithm was compared with the GNG algorithm, which has not yet been used for this purpose.

RGNG was proposed to identify activated regions in the brain of various fMRI datasets with different and important features unlike other clustering approaches. Different robustness properties are associated with the RGNG network because it is insensitive to initialization, input sequence ordering, and outliers, determines the optimal number of underlying clusters during different growth stages, and deals with multimodal datasets effectively.

The approach of using RGNG with fMRI dataset is the first attempt in the literature. The current study is organized as follows: Section 2 provides the most important packages used with fMRI data analysis in comparison with the clustering and especially the proposed RGNG approach. Section 3 describes the proposed work and algorithms using simple flowcharts and tables. Section 4 describes the preprocessing and performance measures. Section 5 presents the experimental output results. Finally, Section 6 concludes the paper and introduces future research directions.

2. fMRI data analysis techniques

FMRI data analysis methods can be divided mainly into two categories, namely, model-driven or model-based (hypothesis) and data-driven or model-free (exploratory) approaches. The model-driven methods deal with definite activation patterns, response functions, or experiments. These models require previous knowledge and statistically test the analyzed data on the presence or absence of a response. The methods related to this category differ either by statistical method or signal estimation procedure in performing the activation. An example is the commonly used general linear model (GLM), which is the most fundamental and basic approach used for fMRI data analysis with statistical parametric mapping (SPM) (SPM, 1991).

Data-driven methods, in contrast, have the ability to count all of the voxels simultaneously, define the active zones and find structures in the brain and fMRI data competently without previous knowledge about activation patterns or experimental paradigms. These methods can be divided mainly into two groups, namely, blind source separation (BSS) and clustering approach.

BSS attempts to find unobserved signals or "sources" from several observed mixtures and generate a model of the data. Various methods are used for BSS: PCA (Friston et al., 1993; Friston et al., 1996), ICA (Hyvarinen et al., 2001; McKeown et al., 1998; Calhoun et al., 2001; Mckeown, 2000), and canonical correlation analysis (CCA) (Friman et al., 2002) methods are used to separate these mixtures to obtain source signals. The FMRIB Software Library (FSL) package (The Analysis Group, 2012) uses melodic ICA, which is a data-driven (model-free) approach, but is insufficient for most fMRI datasets because ICA has some limitations. ICA attempts to find maximally independent maps and split the wide activation areas into a number of maps, which have a strong correlation between time courses (TCs) of different components. The independent components (ICs) from ICA decomposition are unordered, that is, this feature is associated with the model order selection for linear model-based region extraction, which remains an open problem. Thus, determining whether or not ICs are correlated with nonlinear activation is difficult.

Clustering (Chen et al., 2006; Seghier et al., 2007) analysis is based on group voxels according to their TC signals in a similar HDR (Hemodynamic response) over time. The advantages presented for the proposed clustering RGNG algorithm will be explained. This approach is mainly a data-driven or model-free (exploratory) approach. Table 1 compares statistical, transformation and clustering methods.

Clustering analysis is widely used for fMRI data processing to detect the brain active area effectively. In the following paragraph, the data mining idea is identified based on the GNG network.

Lachiche et al. (2005) introduced a new interactive data mining approach to fMRI images, which has not been used for the purpose of the current study, and showed that GNG successfully recognized the active areas in the fMRI images of the brain (Lachiche et al., 2005). The idea of defining a distance between voxels of fMRI images was argued, and this distance is proposed to be based on the signal only.

Korczak (2007) introduced a new interactive data mining technique to fMRI images to observe cerebral activity; this technique is based on a data-driven approach (Korczak, 2007). Different unsupervised clustering algorithms were presented, developed, and tested on sequences of fMRI images. Five clustering algorithms (GNG, SOM, LBG, K-means, and CURE) were applied to synthetic and real data. The experimental results showed that the perfor-

Table 1
Comparison among statistical, transformation and clustering.

Based Approach	fMRI Analysis Methods	Approach Properties
Statistical method	Model-Driven/Model-Based/Hypothesis Used with SPM package and based on GLM	The most fundamental, basic and commonly used approach for fMRI data analysis, but it needs previous knowledge about activation patterns or experiments.
Transformation method	Data-Driven/Model Free/Exploratory Used with FSL package and based on melodic IC analysis	It is based on linear mixing and is unordered. Thus, it must deal with independent data.
Clustering method	Data-Driven/Model Free/Exploratory RGNG is an example which were used in this study	It can define active zones and identify structures in the brain and fMRI data competently without the need for previous knowledge about activation patterns or experiments.

mance of the GNG algorithm was the best among all other clustering methods, with acceptable robustness.

Heydar et al. (2009) developed the algorithm of the GNG network, which can run the optimal number of clusters automatically (Heydar et al., 2009). The experimental results used artificial and real fMRI datasets with the proposed algorithm, which is an improved version of the GNG algorithm. They compared the Jaccard coefficient of the proposed algorithm with some well-known clustering algorithms, such as K-means, NG, GNG, and fuzzy C-means (FCM); the results showed that the proposed algorithm outperformed the other algorithms.

The GNG originates from the NG algorithm by Fritzke (Fritzke, 1995; Fritzke, 1997), and the RGNG algorithm was introduced by Qin and Suganthan (2004) within the GNG structure. The work presented in this paper using RGNG with fMRI data will be the first attempt in the literature. The current study presented how to feed RGNG with real and free auditory fMRI datasets.

GNG and RGNG, both artificial neural network approaches based on unsupervised clustering for fMRI analysis, are compared in Table 2. This table presents the researchers who introduced these approaches, the researchers who used these approaches in fMRI research, and the advantages and limitations of each approach (AlJobouri et al., 2017).

3. Methodology and proposed work

The GNG algorithm is reviewed before introducing the proposed RGNG algorithm for feeding with fMRI data. The GNG and RGNG algorithms are extensive and complex. Thus, flowcharts and a mathematical model were developed for convenience and easier writing of the related codes.

3.1. GNG algorithm

The GNG algorithm was developed by Fritzke (1995, 1997); he proposed changing the unit numbers (mostly increased) in a SOM network with a variable topological structure. The GNG is a growing soft competitive learning algorithm, which combines the topology formation rules of the competitive Hebbian learning (Martinetz and Schulten, 1991) with the growing cell structures (Fritzke, 1994) into a new model.

Before feeding the GNG algorithm, the following parameters must be defined:

N	maximal number of neurons
$\varepsilon_b, \varepsilon_n$	constant learning rate for the winner and its topological neighbors, respectively
λ	an iteration of a new neuron will be created
α	reduction of the error counter by inserting a new neuron
β	reduces the overall value of the error counter every iteration step
Max_iter	maximal number of iterations

In the subsequent experiments, the parameter settings are fixed for each algorithm, with typical values proposed in the literature. The GNG algorithm was set with typical values as in (Fritzke, 1997): $\varepsilon_b = 0.05$, $\varepsilon_n = 0.006$, $\alpha_{max} = 100$, $\beta = 0.0005$, and $\lambda = 300$.

Fig. 1 presents the flowchart of the GNG algorithm and shows that the inactive neurons that do not win during a long time interval may be detected through the GNG algorithm by tracing the

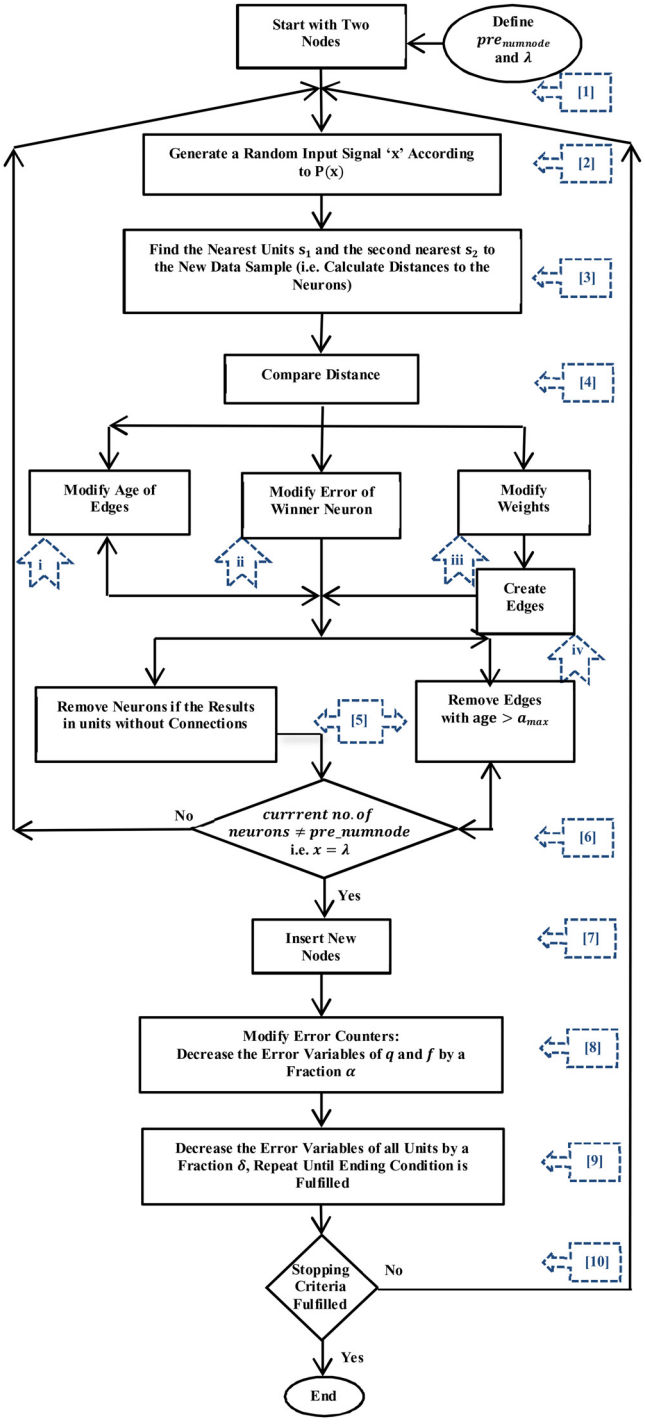

Fig. 1. Flowchart design of the GNG algorithm.

changes of an age variable associated with each edge. The proposed flowchart can be summarized in the following steps:

1. Initialization
2. Select the input vector
3. Calculate the distances by determining winner s_1 and the second nearest node s_2 to the new data
4. Compare the distances:
   - Modify the age of the edges
   - Update the local error of the winner neuron
   - Modify the weights
   - Create the edges
5. Dead node removal procedure
6. Selection of the number of prototypes of the current number of neurons #$pre_{numnode}$; return to step (2)
7. Insertion procedure:
   - Determine prototype q
   - Determine prototype f of neighbors to q
   - Create a new prototype r between q and f
   - Create edges connecting r with q and f, and remove the original edge between q and f
8. Modify the error counters
   - Decrease the error of q and f by multiplying with α
   - Initialize the error of r with q and f
9. Decrease the error of all neurons by multiplying with β
10. Select the stopping criterion. If finished, then stop or repeat step (2)

The GNG starts with a minimal network size, and a few numbers of new neurons and connections are inserted into a growing structure using vector quantization until the desired quality of the model is achieved (e.g., net size, time limit, predefined numbers of neurons inserted, or some performance measure).

The GNG starts with a minimal network size, and a few numbers of new neurons and connections are inserted into a growing structure using vector quantization until the desired quality of the model is achieved (e.g., net size, time limit, predefined numbers of neurons inserted, or some performance measure).

3.2. Robust growing neural gas (RGNG) algorithm

The "dead node" problem occurs in the GNG algorithm because of the growth scheme associated with the GNG algorithm. Dead node problems occur because of inappropriate initializations, which cause some prototypes to never win through the training process. Even with the initialization insensitive clustering methods, good clustering results may not be obtained if the order of the input sequence is not chosen properly.

Aside from problems related to the sensitivity for initialization and the order of input vector sorting, other problems related to the presence and position of various outliers occur. Thus, the GNG network may fail to differentiate the outliers from the inliers through the original prototype updating rule when various outliers exist in a dataset.

A novel RGNG was presented because of the limitations of the GNG algorithm (Qin and Suganthan, 2004) within the GNG structure. The robustness of RGNG toward initialization, input vector sorting, and the existence and position of various outliers improved, as well as its ability to find the optimal number of neurons during runtime dynamically.

Fig. 2 presents the flowchart of the RGNG algorithm. The proposed flowchart can be summarized in the following steps:

1. Initialization
2. Select the input vector
   - Calculate the harmonic average distance $d_m^i(0)$ with respect to its current position
   - Calculate the learning rates $\varepsilon_b, \varepsilon_n$ for current neuron l
3. Calculate the distances by determining winner s_1 and second nearest node s_2 to the new data
4. Compare the distances:
   - Modify the age of the edges
   - Modify the weights
   - Create the edges
5. Dead node removal procedure
6. Select the number of prototypes of the current number of neurons #$pre_{numnode}$; return to step (2)
7. Find the minimum description length (MDL) value
8. Select the MDL value and save the smallest value
9. Find the neuron with the largest local accumulated error
10. Node insertion procedure:
    - Determine prototype q
    - Determine prototype f neighboring q
    - Create a new prototype r between q and f
    - Create edges connecting r with q and f, and remove the original edge between q and f
11. Set the ranking counter $prenode^r$ by inserting a new node
12. Select the stopping criterion. If completed, then stop or return to step (2)

Before feeding the RGNG algorithm, the following parameters must be defined:

Table 2
Comparison between Two Artificial Neural Network Approaches based on the Unsupervised Clustering for fMRI Analysis.

Methods	GNG	RGNG
Introduced by	• Fritzke (1995)	• Lachiche et al. (2005) • Korczak (2007) • Heydar et al. (2009)
Used with fMRI	• Qin and Suganthan (2004)	It was not previously proposed
Advantages	• Its ability to modify the network topology by removing edges with its age variable • The neighborhood sorting step is unnecessary • It can find a network size and structure automatically, continue learning, and add units and connections until a performance criterion is fulfilled • The number of classes is not fixed in advance as in most clustering algorithms	• Insensitive to initialization, input sequence ordering, and the presence of outliers during different growth stages • Can automatically determine the optimal number of clusters • Deals with multimodal datasets effectively
Limitations	Its sensitivity for: • Initialization • The order of input vectors • Existence of many outliers	Undetected limitations yet

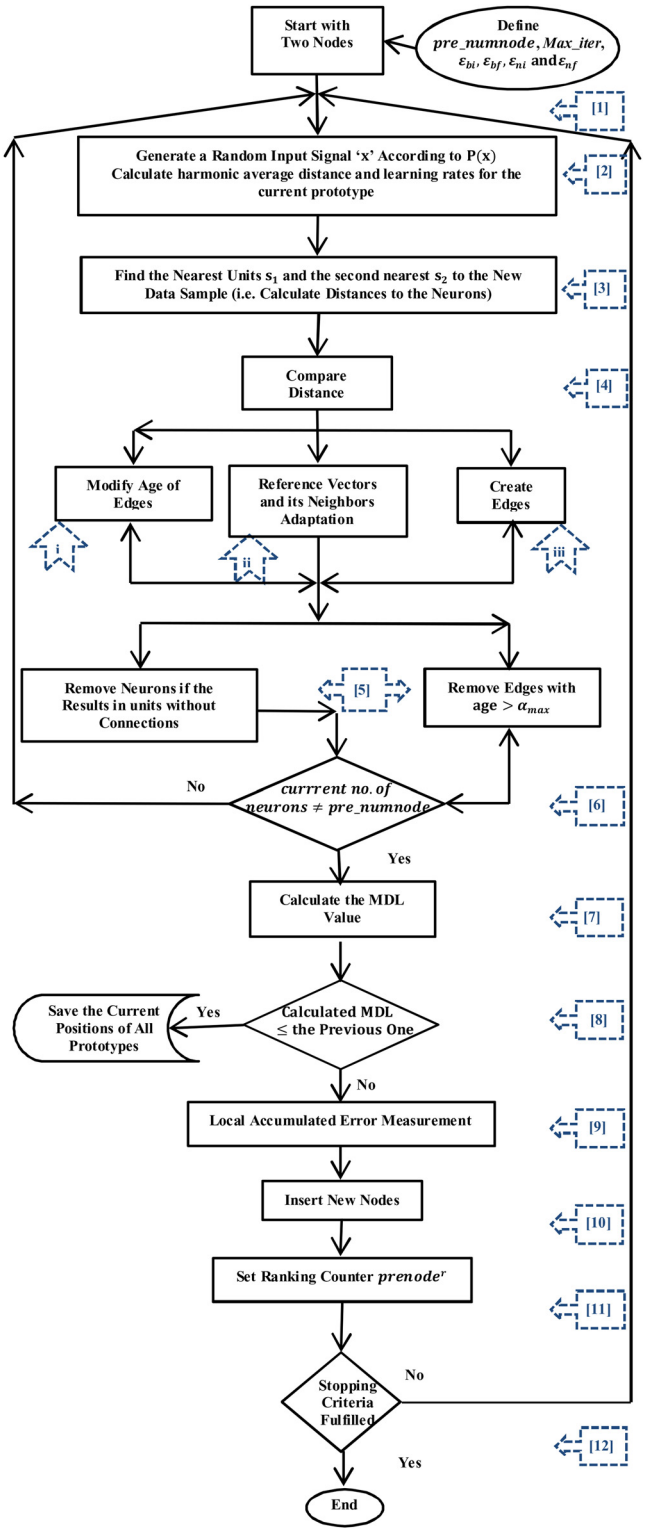

Fig. 2. Flowchart design of the RGNG algorithm.

In the subsequent experiments, the parameter settings were fixed for each algorithm, with typical values proposed in the literature. The RGNG algorithm was set with typical values as in [12]: $\varepsilon_{bi} = 0.1$, $\varepsilon_{bf} = 0.01$, $\varepsilon_{ni} = 0.005$, $\varepsilon_{nf} = 0.0005$, $\alpha_{\max} = 100$, $k = 1.3$, and $\eta = 1 \times 10^{-4}$.

For each reference vector w_i, $i = 1, 2, \ldots, c$, a series of edges emerged from its location to a joint with its direct topological neighbors. Similar to the GNG, the RGNG algorithm starts in step 1 with the initialization of a few prototype vectors (usually two), $W = \{w_1, w_2\}$.

In fMRI, W represents the TC of the fMRI dataset (see Fig. 3), w_i denotes the TC of exemplar i, and w_c is the TC of the closest exemplar c. Prototype vectors w_1, w_2 are randomly chosen with reference vectors from the TC of all voxels $P(x)$, and a data voxel x is generated as an input signal from the fMRI dataset used for training, $X = \{x_1, x_2, \ldots, x_N\}$.

The maximum number of neurons to grow is defined as $pre_{numnode}$ and the maximum predefined training epoch is defined as Max_iter during each growth stage with a certain prototype number. The initial or current training epoch number is set as $m = 0$.

The iteration point in the training epoch (task periods) m, $t = 0$. Thus, the full iteration step $iter$ over each growth step is expressed as:

$$iter = m \cdot N + t, \quad (1)$$

where N is the length of the fMRI TC.

Clustering algorithms attempt to classify the TC signals of the voxels into different groups according to the similarity among the groups. The temporal information is ordered in clusters and is independent of its spatial neighborhood. These clusters are described by an average TC or a cluster center obtained by averaging all of the TCs of the cluster. The fMRI data are transformed into a TC of voxel intensity variations proportional to its average, as follows:

$$I_{av}^x = \frac{1}{N} \sum I_i^x, \quad (2)$$

where I_{av}^x is the average intensity of voxel a of a series of N images;

$$W_i = I_{av}^x - I_i^x. \quad (3)$$

The distances between two fMRI signals W_a and W_b may be computed as a Euclidian distance:

$$d_E = \sqrt{(W_{ai} - W_{bi})^2}. \quad (4)$$

The activity level of the dataset is generally based on the distance between input vectors x compared with all of the exemplar TCs W_i. In RGNG, the smallest Euclidean distance $x - w_i$ can be made to define the best matching node.

The RGNG algorithm used the principle of the MDL value as the clustering validity index (to find the optimal number of clusters and their center positions) corresponding to the smallest MDL value. Thus, the optimal number of clusters is determined automatically by searching the extreme minimum value of the MDL measure through the network-growing process. The RGNG approach has the smallest MDL value recorded with respect to the GNG combined with the MDL principle, as explained in the results. Thus, the RGNG

N	maximal number of neurons
ε_b^l	learning rate of the winner
ε_n^l	learning rate of its topological neighbors
$\varepsilon_{bf}^l, \varepsilon_{bi}^l, \varepsilon_{nf}^l, \varepsilon_{ni}^l$	initial and final values of ε_b^l and ε_n^l
α_{max}	maximal age of a connection
β	mobility of the winner's neighborhood toward the input vector
k, η	used to determine the MDL value
Max_iter	maximal number of iterations

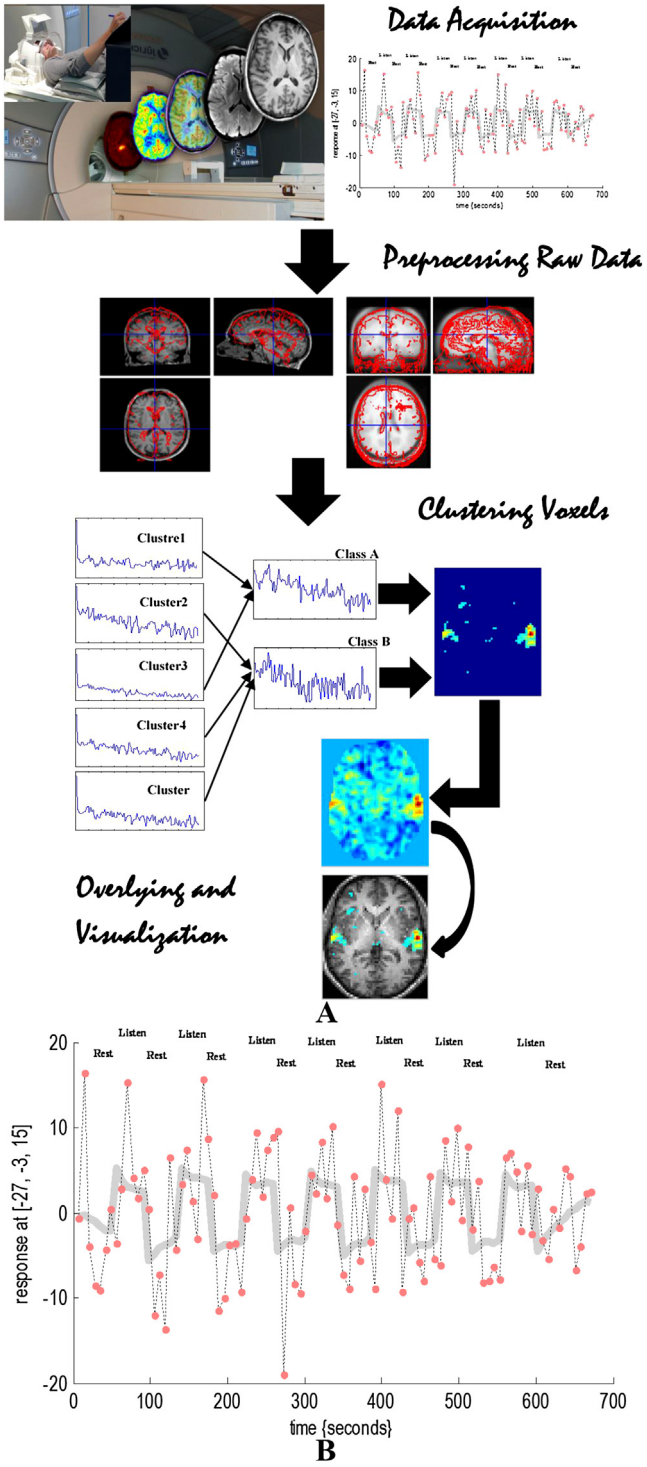

Fig. 3. Proposed data mining system architecture. (a) Main block diagram. (b) Experimental paradigm "silence" and "talk".

approach can find the optimal number of clusters and their center positions corresponding to the smallest MDL value.

4. Simulation design

The block diagram shows the process of brain function data analysis, which is performed in the current study. The process is composed of five stages (see Fig. 3B):

- preprocessing of the raw data;
- clustering voxels together based on the similarity of their intensity profile in the TCs of the image;
- overlay with the structural image;
- visual fMRI image;
- validation.

4.1. Image spatial preprocessing

The experiments in the present work were performed in MATLAB 2016a and SPM12 package for the preprocessing stage. Various noise factors interfere with the fMRI signals of interest. The subject is typically never completely motionless. Thus, the preprocessing steps must be adapted to each identified artifact before the clustering phase. The present work used SPM for the auditory fMRI data spatial preprocessing stages. The fMRI dataset is preprocessed by applying the following steps:

- Realignment;
- Coregistration;
- Segmentation;
- Normalize;
- Smoothing using FWHM=6.

The functional images were reoriented to MNI space, which is standard brain formed by using a large series of MRI scans on normal controls developed at the Montreal Neurological Institute. Then the functional raw data were realigned to correct for the head movements. The high-resolution anatomical T1 images were coregistered with the realigned functional images to enable anatomical localization of the activations. Segmentation process is not mandatory. SPM12 uses MNI template image, which are the most common templates used for fMRI spatial normalization. In this step, the anatomical and functional images were spatially normalized into MNI space. Finally, the functional raw data were spatially smoothed with a Gaussian smoothing kernel of 6.

4.2. FMRI dataset

Quantitative performance assessment uses an auditory fMRI dataset. Auditory data is composed of entire brain BOLD/EPI images acquired on a modified 2T Siemens MAGNETOM Vision system (John et al., 2013). Each acquisition consists of 64 contiguous slices ($64 \times 64 \times 64$ $3 \times 3 \times 3$ mm voxels). Acquisition took 6.05 s, with a scan to scan repetition time (TR) set arbitrarily to 7 s. A total of 96 acquisitions were made from a single subject in blocks of 6 scans (acquired during the same condition as a stimulant or rest), yielding 16 blocks and each block for 42 s.

The experimental paradigm for successive blocks alternated between rest and auditory stimulation, starting with rest (see Fig. 3B). The functional data start at acquisition 4, functional image (fM4). Auditory stimulation was composed of bisyllabic words (e.g., "mother," "house," "weather," and "movie") presented binaurally at a rate of 60/min. The first few scans must be discarded ("dummy" lead did not exist in scans) because of T1 effects.

4.3. Performance measure

A novel application of the RGNG algorithm was compared with GNG and SPM using two performance measures, namely, the MDL value and receiver operating characteristic (ROC) analysis.

In the RGNG algorithm, the MDL value is one of the well-known information theory evaluation measures, which has been used as the clustering validity index (Rissanen, 1983). The average MDL values during the growth stages have been plotted versus the length of fMRI TC (N). Fig. 4 shows the curves for the RGNG and GNG

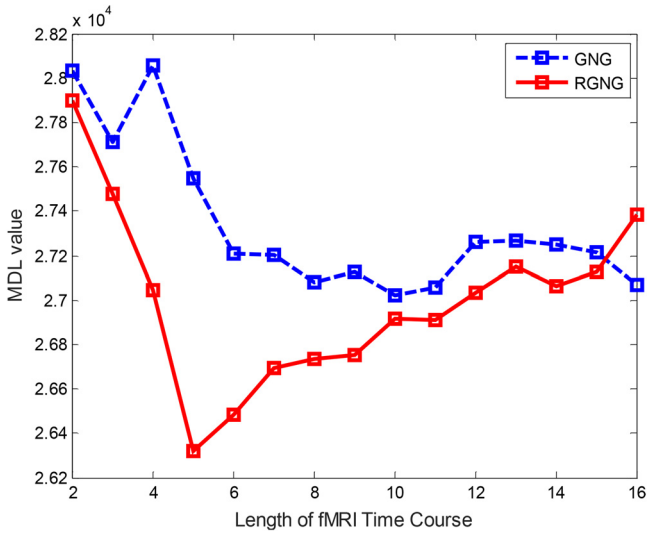

Fig. 4. MDL values versus *N*.

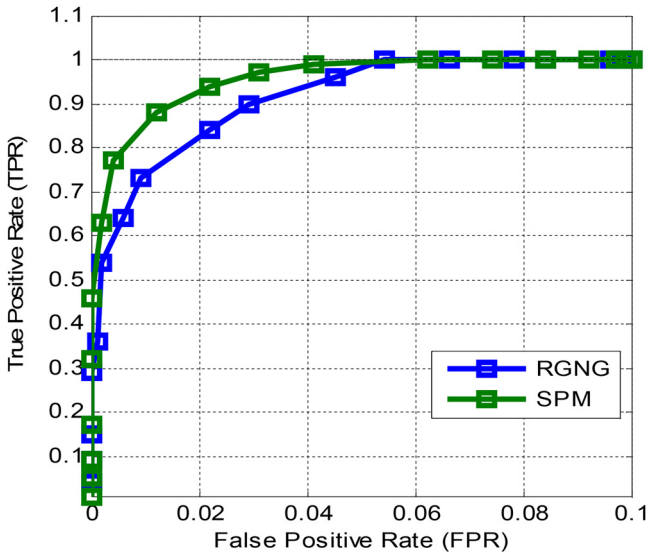

Fig. 5. ROC curves analyses of the auditory fMRI dataset.

approaches combined with the MDL criterion; the length of the fMRI TC is selected randomly as $N = 16$. Each detected cluster number corresponded to the MDL value. The RGNG approach has the smallest MDL value recorded with respect to GNG combined with the MDL principle; thus, it can successfully determine the actual number of clusters.

The ROC analysis is another index of the performance of RGNG in comparison with SPM (Skudlarski et al., 1999). The ROC is well-known in medical imaging and machine learning applications; the ROC space consists of the false positive ratio (FPR) on the *x*-axis and the true positive ratio (TPR) on the *y*-axis (Sun and Xu, 2014). The good classifier space is indicated by a high TPR and a low FPR, whereas the bad classifier space is indicated by a low TPR and a high FPR.

The curves in Fig. 5, generally indicated that the two methods work as good classifiers with a high TPR and a low FPR. The RGNG method can detect real activations under the same FPR ratio.

In fMRI, the FPR is calculated by dividing the number of misclassified inactivated voxels by the total number of voxels considered, whereas the TPR is calculated by dividing the number of correct classifications of activated voxels by the total number of voxels considered (Lange et al., 1999). In the same situation, the ROC curves for the RGNG and SPM methods are compared, as shown in Fig. 5.

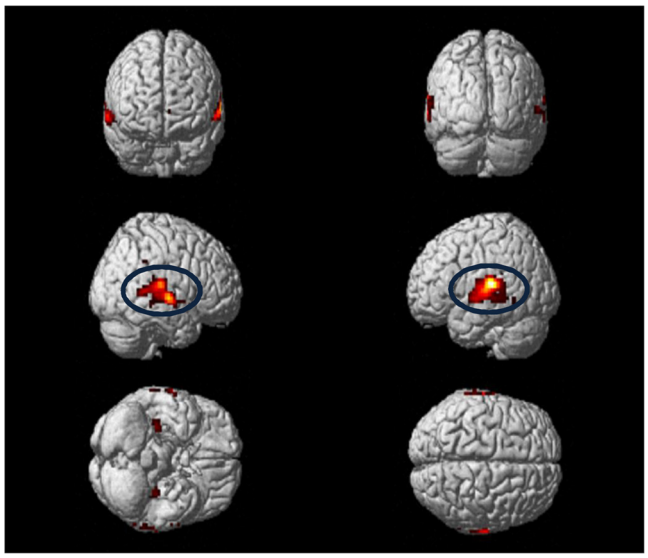

Fig. 6. Active areas in the brain auditory cortex area within the SPM package.

5. fMRI results

The principles behind the prototype-based clustering algorithms were introduced in this work. The validity of the performance of the RGNG was analyzed and verified with fMRI experiments. FMRI analysis involves known areas and functions of the brain. Thus, the common and expected results must be used in the experiments. One of these areas is the auditory cortex. Real auditory fMRI data, which are freely available for education and evaluation purposes, were used in the experiments [http://www.fil.ion.ucl.ac.uk/spm/data/auditory/]. These data were utilized by previous works (Lachiche et al., 2005; Korczak, 2007; Heydar et al., 2009).

One of the decisive advantages of fMRI is that fMRI studies do not require the analysis of a group of volunteers, but can produce valuable results at the level of single individuals. The analysis of single volunteers is crucial in analyzing small structures, which exhibit strong interindividual variation (Campain and Minckler, 1976; Francesco et al., 2003), similar to the auditory cortex, as shown in Fig. 6.

5.1. Comparing auditory data running RGNG with that of GNG

A block design experiment was conducted using auditory stimulus. Figs. 7 and 8 A and B show the active areas in the auditory cortex of the entire brain when running the GNG, and RGNG algorithms. Although auditory cortex regions were found by GNG and RGNG algorithms, in GNG other areas are also activated outside this cortex. In RGNG, these areas are less or approximately disappeared under the same experiment and the auditory stimulus of the whole brain.

Fig. 7 shows the clusters in a transparent or glass brain image which is a more flexible approach by specifying a real RGB (red-green-blue) color value for every voxel in the image. Fig. 8A and B show the alignment of the obtained clusters into a structural space of the brain when running the GNG and RGNG, respectively. With regard to the output results obtained by running the three unsupervised clustering algorithms, spatial information is visualized as fine clusters in the auditory cortex area. The GNG algorithm was used in previous works (Lachiche et al., 2005; Korczak, 2007; Heydar et al.,

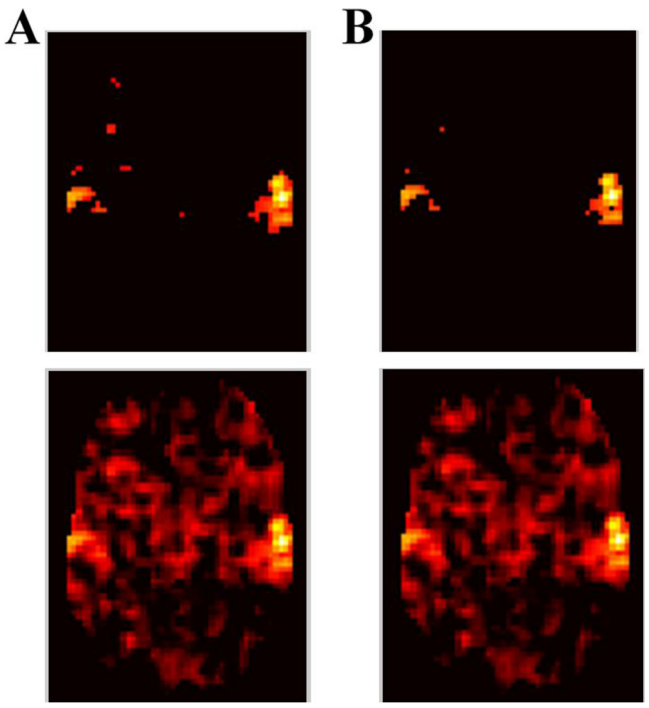

Fig. 7. Clusters in a transparent brain image when running the (a) GNG and (b) RGNG clustering techniques.

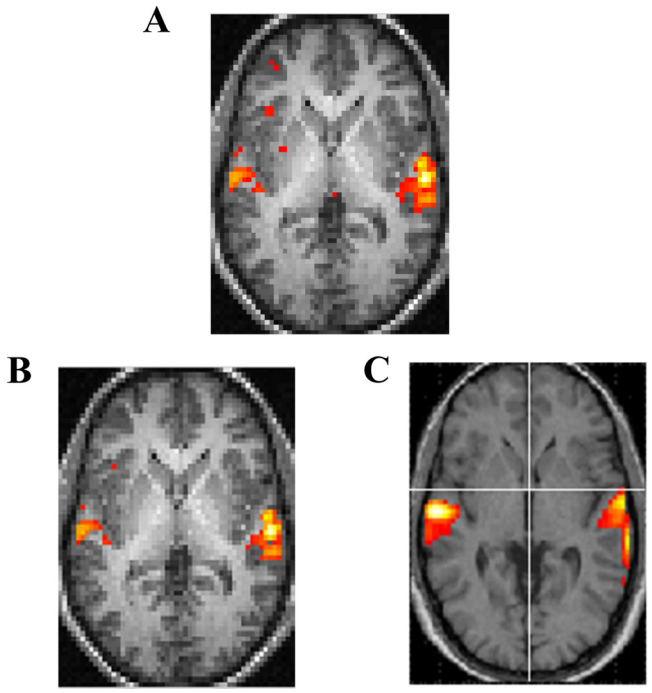

Fig. 8. Clusters overlaid onto the anatomical image when running the (a) GNG (b) RGNG and (c) SPM.

2009). The ROI obtained within the auditory cortex when running the GNG algorithm (Fig. 8A) is similar to that obtained by the same approach in the literature. In general, a cluster corresponds to a group of voxels with a similar HDR over a TC.

The block design experiment was conducted by running the proposed RGNG approach using auditory data. The activation shown in Fig. 8B is located in the temporal lobe. The spatial information shows that the areas of activation obtained are similar to those expected from the auditory cortex experiments, which are detected as variations of voxel intensity over time. In the current study, the data obtained by the RGNG approach was separated according to the TC signals of voxel intensity variations relative to its average. Similar to all clustering algorithms, the RGNG attempted to portion homogeneous areas of activation in the brain that were comparable to those areas located using other approaches and found in the recognized cortices related to the experiment. These areas or clusters are described by an average TC or a cluster center obtained by averaging all of the TCs of the cluster.

The novel application of RGNG output clustering results can be recognized as the best with respect to the GNG approach because the cluster results defined the specific auditory cortex area. Moreover, the fMRI output results obtained by running the RGNG was the same as the outcome obtained by running the SPM using the same dataset and the same paradigm, as will be discussed in the next subsection.

5.2. Comparing auditory data running RGNG with that of SPM

The paradigm of the block design experiment alternates two conditions, namely, without the stimulus and with auditory stimuli, which consist of repetitions of two-syllable words, such as "mother," "house," "weather," and "movie". Fig. 8B shows the ability of the RGNG clustering technique to identify winner nodes, determine the optimal number of underlying clusters, and produce a TC for activation detection in an auditory dataset. Fig. 8**C** shows the area of activation in the auditory cortex of whole brain running SPM with *f*-contrast test results with family-wise error (FWE) threshold, with no masking, the FWE-corrected p value = 0.05.

The results of SPM based on GLM using the paradigm as a reference signal introduced bias in the experiment. By contrast, the RGNG approach did not use the paradigm as the reference signal because it works as a model-free method. In summary, the RGNG results were within the expected outputs and have similar results to those found with the hypothesis method in detecting active areas within the expected auditory cortices. The RGNG signal changes over a TC in auditory fMRI datasets which can be calculated by labeling the pixels of the same cluster (membership TC) or by plotting the distance of the TCs to a given cluster center (distance TC).

Novel and extensive simulation studies on real fMRI datasets were conducted using the RGNG unsupervised clustering algorithm. A potential problem associated with GLM model is the requirement of an accurate estimate of the fMRI paradigm design. In different cases, it is difficult to provide precise model designs; either the problem from the subjects who did the task incorrectly (also the same subject may give a different response for the same paradigm at a different time) or different subjects may still give different BOLD signals during the same paradigm. The result in Fig. 8B shows that this method can complement the model-based method to cope with the difficulties and challenges in fMRI data analysis. The findings can improve the recognition of the nature of the fMRI data and the underlying mechanisms.

6. Conclusions

The major objective of this study is to detect and classify the activated areas of the brain using a robust and efficient algorithm. This type of study has not yet been conducted, and the current study, which uses RGNG with fMRI, is the first attempt to do so.

In conclusion, the RGNG can detect the active zones in the brain, analyze brain function, and determine the optimal number of underlying clusters in fMRI datasets. This algorithm can define the positions of the center of an output cluster corresponding to the minimal MDL value. The validity of the performance of the RGNG algorithm was tested using real auditory fMRI data, which are based on the stimulation of the auditory cortex.

Some difficulties were addressed by using the conventional clustering algorithms. For example, the number of clusters must be defined earlier and the cluster detection problem has different dimensions within the same dataset. The RGNG merges the GNG structure with robust properties and uses MDL to define the problems of optimal network representations and parameters, which made the RGNG insensitive to the initializations, input sequence ordering, and outliers and more robust toward noisy input data. During the network-growing process, the RGNG can effectively determine the optimal number of clusters and their corresponding positions, which are closer to the actual cluster centers (with the smallest MDL value) with minimal influence from the outliers.

The experimental output results showed the superior performance of the RGNG over model-based approaches and one of the prototype-based clustering algorithms on real fMRI datasets as revealed by their performance measured by MDL and ROC analysis. This work proposed novel and powerful methods for fMRI data analysis, which integrate the advantages of the hypothesis and exploratory analysis methods.

Two types of fMRI analysis methods were compared, namely, GLM and data-driven analyses using machine learning classifiers. The GLM is the most common method for fMRI data analysis but is based heavily on a priori BOLD model design. In some cases, the GLM cannot be used for brain activation detection when previous information about the data is unavailable. An example is a research involving mental subject or during daydreaming and mind-wandering (default mode of brain function) (Yongnan, 2010). Thus, effective alternative approaches using data-driven analysis were introduced to detect brain activity based on the data structure. The proposed application of RGNG on a real fMRI dataset was reviewed on a single-subject auditory fMRI data. This method can be also extended to multi-subject data-driven analysis (multiple subject data) of fMRI dataset. RGNG approach may be preferable for multiple subject studies instead of analyses data from single-subject as the used auditory data.

The paradigm of the auditory dataset used in the experiment was a block-type data design. For future, this work can be extended toward experiments with event-related data design.

The RGNG can deal well with fMRI, which is composed of multimodal datasets. Thus, the approach can be applied to other real multimodal datasets, such as MRI image segmentation in the brain and other regions of the body. Thus, clusters of different organ shapes in the body can be detected using other distance metrics because the Euclidean distance metric used with the RGNG can detect the clusters of the brain, which is an approximately spherical or ellipsoidal region with minimal differences in the variance in each dimension (Frigui and Krishnapuram, 1999).

In future studies, cluster validity measures other than the MDL criterion can be used with RGNG. Minimum message length, Bayesian information criterion, and Akaike's information criterion can be applied to tackle the use of the common MDL validity index used in this work. The findings from this work can help address the various difficulties that neurologists and psychologists encounter during analysis to improve the interpretation of fMRI data.

Acknowledgment

The authors would like to thank the reviewers whose comments greatly improved the quality of the manuscript.

References

AlJobouri, H.K., Jaber, H.A., Çankaya, I., 2017. Performance evaluation of prototype-Based clustering algorithms combined MDL index. Comput. Appl. Eng. Educ. 25 (4), 642–654 (Wiley Inc.).
Aljobouri, H.K., Çankaya, I., Karal, O., 2015. From biomedical signal processing techniques to fMRI parcellation. Biosci. Biotechnol. Res. Asia 12, 1115–1138.
Baumgartner, R., Windischberger, C., Moser, E., 1998. Quantification in functional magnetic resonance imaging: fuzzy clustering vs correlation analysis. Magn. Reson. Imaging 16, 115–125.
Bock, H.H., Diday, E., 2000. Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data Studies in Classification. Data Analysis and Knowledge Organization, Springer-Verlag.
Calhoun, V., Adali, T., Pearlson, G., Pekar, J., 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum. Brain Mapp. 13, 43–53.
Campain, R., Minckler, J., 1976. A note on the gross configurations of the human auditory cortex. Brain Lang. 3, 318–323.
Chen, H., Yuan, H., Yao, D., Chen, L., Chen, W., 2006. An integrated neighborhood correlation and hierarchical clustering approach of functional MRI. IEEE Trans. Biomed. Eng. 53, 452–458.
Dimitriadou, E., Barth, M., Windischberger, C., Hornika, K., Moser, E., 2004. A quantitative comparison of functional cluster analysis. Artif. Intell. Med. 31, 57–71.
Francesco, D.S., Fabrizio, E., Tommaso, S., Elia, F., Elio, M., Claudio, S., Sossio, C., Raffaele, E., Klaus, S., Erich, S., 2003. fMRI of the auditory system: understanding the neural basis of auditory gestalt. Magn. Reson. Imaging 21, 1213–1224.
Frigui, H., Krishnapuram, R., 1999. A robust competitive clustering algorithm with applications in computer vision. IEEE Trans. Patt. Anal. Mach. Intell. 21, 450–465.
Friman, O., Borga, M., Lundberg, P., Knutsson, H., 2002. Exploratory fMRI analysis by autocorrelation maximization. Neuroimage 16, 454–464.
Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S., 1993. Functional connectivity: the principal component analysis of large PET data sets. J. Cereb. Blood Flow Metab. 13, 5–14.
Friston, K.J., Poline, J.B., Strother, S., Holmes, A.P., Frith, C.D., Frackowiak, R.S., 1996. A multivariate analysis of PET activation studies. Hum. Brain Mapp. 4, 140–151.
Fritzke, B., 1994. Growing cells structures—a self-organizing network for unsupervised and supervised learning. Neural Netw. 7, 1441–1460.
Fritzke, B., 1995. A Growing Neural Gas Network Learns Topologies, Advances in Neural Information Processing Systems 7. MIT Press, Cambridge, pp. 625–632.
Fritzke, B., 1997. Some Competitive Learning Methods (draft), Technique Report. Institute for Neural Computation, Ruhr-University, Bochum.
Goutte, C., Toft, P., Rostrup, E., Nielsen, E.F., Hansen, L., 1999. On clustering fMRI time series. Neuroimage 9, 298–310.
Heydar, D., Ali, T., Emad, F., 2009. Extracting activated regions of fMRI data using unsupervised learning. In: Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia USA, pp. 641–645.
Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons.
John, A., Gareth, B., Chun-Chuan, C., Jean, D., Guillaume, F., Karl, F., Stefan, K., James, K., Vladimir, L., Rosalyn, M., Will, P., Maria, R., Klaas, S., Darren, G., Rik, H., Chloe, H., Volkmar, G., Jeremie, M., Christophe, P., 2013. SPM8 Manual, Functional Imaging Laboratory, Trust Centre for Neuroimaging. Institute of Neurology, London UK.
Katwal, S.B., 2011. Analyzing fMRI data with graph-based visualizations of self-Organizing maps. In: IEEE International Symposium on Biomedical Imaging, Chicago, pp. 1577–1580.
Korczak, J., 2007. Interactive Mining of Functional MRI Data, Signal-Image Technologies and Internet-Based System, (SITIS '07). IEEE Computer Society, Washington, DC USA, pp. 912–917.
Korczak, J., 2012. Visual exploration of functional MRI data. In: Karahoca, A., INTECH (Eds.), Data Mining Applications in Engineering and Medicine. , pp. 249–264.
Lachiche, N., Hommet, J., Korczak, J., Braud, A., 2005. Neuronal clustering of brain fMRI images. Proceeding of Pattern Recognition and Machine Inference, 300–305.
Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.A., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis. Neuroimage 10, 1999.
Liao, W., Chen, H., Yang, Q., Lei, X., 2008. Analysis of fMRI data using improved self-Organizing mapping and spatio-temporal metric hierarchical clustering. IEEE Trans. Med. Imaging 27, 1472–1483.
Lindquist, M.A., 2008. The statistical analysis of fMRI data. Stat. Sci. 23, 439–464.
Martinetz, T., Schulten, K., 1991. A. Neural Gas Network Learns Topologies, Artificial Neural Networks. Elsevier, pp. 397–402.
McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., Sejnowski, T., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp. 6, 160–188.
Mckeown, M.J., 2000. Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. Neuroimage 11, 24–35.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45.
Qin, A.K., Suganthan, P.N., 2004. Robust growing neural gas algorithm with application in cluster analysis. Neural Netw. 17, 1135–1148.
Rissanen, J., 1983. A universal prior for integers and estimation by minimum description length. Ann. Stat. 11, 416–431.
SPM, 1991. Statistical Parametric Mapping. http://www.fil.ion.ucl.ac.uk/spm/.
Seghier, M.L., Friston, K.J., Price, C.J., 2007. Detecting subject-Specific activations using fuzzy clustering. Neuroimage 36, 594–605.
Skudlarski, P., Constable, R.T., Gore, J.C., 1999. ROC analysis of statistical methods used in functional MRI: Individual subjects. Neuroimage 9, 311–329.

Sun, X., Xu, W., 2014. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 21, 1389–1393.
The Analysis Group, 2012. FMRIB (Oxford, UK) http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/.
Wismuller, A., Meyer-Base, A., Lange, O., Auer, D., Reiser, M.F., Sumners, D., 2004. Model-free functional MRI analysis based on unsupervised clustering. J. Biomed. Inform. 37, 10–18.
Yongnan, J., 2010. Data-driven fMRI Data Analysis Based on Parcellation, Ph.D Thesis. University of Nottingham (October).