Artificial Intelligence in Medicine (2005) **33**, 261–280

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Applying spatial distribution analysis techniques to classification of 3D medical images

Dragoljub Pokrajac[a,b,*], Vasileios Megalooikonomou[c,d], Aleksandar Lazarevic[e], Despina Kontos[d], Zoran Obradovic[c,d]

[a]*Computer and Information Science Department, Delaware State University, 1200N Dupont Hwy, Science Center North, Dover, DE 19901, USA*
[b]*Applied Mathematics Research Center, Delaware State University, 1200N Dupont Hwy, ETV Building, Dover, DE 19901, USA*
[c]*Center for Information Science and Technology, 303 Wachman Hall (038-24), Temple University, 1805 N. Broad St., Philadelphia, PA 19122-6094, USA*
[d]*Department of Computer and Information Sciences, Temple University, 303 Wachman Hall, 1805 N. Broad St., Philadelphia, PA 19122-6094, USA*
[e]*Computer Science Department, University of Minnesota, 1100 South Washington Ave., Suite 101, Minneapolis, MN 55415, USA*

Received 29 July 2003; received in revised form 19 May 2004; accepted 9 July 2004

KEYWORDS
Classification;
Medical images;
Regions of interest;
Similarity measures;
Probability distributions;
Spatial data mining

Summary

Objective: The objective of this paper is to classify 3D medical images by analyzing spatial distributions to model and characterize the arrangement of the regions of interest (ROIs) in 3D space.
Methods and material: Two methods are proposed for facilitating such classification. The first method uses measures of similarity, such as the Mahalanobis distance and the Kullback–Leibler (KL) divergence, to compute the difference between spatial probability distributions of ROIs in an image of a new subject and each of the considered classes represented by historical data (e.g., normal versus disease class). A new subject is predicted to belong to the class corresponding to the most similar dataset. The second method employs the maximum likelihood (ML) principle to predict the class that most likely produced the dataset of the new subject.
Results: The proposed methods have been experimentally evaluated on three datasets: synthetic data (mixtures of Gaussian distributions), realistic lesion-deficit data (generated by a simulator conforming to a clinical study), and functional MRI activation data obtained from a study designed to explore neuroanatomical correlates of semantic processing in Alzheimer's disease (AD).

* Corresponding author. Tel.: +1 302 857 7053; fax: +1 302 857 6552.
E-mail address: dragoljub.pokrajac@verizon.net (D. Pokrajac).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2004.07.001

Conclusion: Performed experiments demonstrated that the approaches based on the KL divergence and the ML method provide superior accuracy compared to the Mahalanobis distance. The later technique could still be a method of choice when the distributions differ significantly, since it is faster and less complex. The obtained classification accuracy with errors smaller than 1% supports that useful diagnosis assistance could be achieved assuming sufficiently informative historic data and sufficient information on the new subject.

1. Introduction

Current advances in medical image acquisition techniques have made available enormous amounts of remarkable high-resolution three-dimensional (3D) image data. In particular, the wide availability of non-invasive methods for capturing structural (e.g., magnetic resonance imaging (MRI), computed tomography (CT)) and functional/physiological (e.g., positron emission tomography (PET), functional MRI (fMRI)) information that complement clinical assessment, have opened new horizons towards a deeper understanding of the human body and its functionality. In addition to the continuous development of improved imaging techniques, greater computer capabilities and improvements in analysis techniques are leading to the creation of large repositories of medical image data. The work presented in this paper addresses problems related to the classification of 3D medical images.

Although significant research has been done in content-based image retrieval and classification for general types of images (see [1,2] for comparative surveys), progress in this type of analysis for medical images has been very slow. Global signatures [3—5] that are usually employed in the content-based image retrieval and classification do not work well in the medical imaging domain where the regions of interest (ROIs) occupy a small portion of the image. In this case, usually a distinction between important and unimportant features or among multiple objects in an image has to be made. We propose to overcome these problems by performing analysis focusing on the ROIs and their spatial distribution. Characterization of an image based only on regions that are of interest to an expert seems to be more meaningful in applications dealing with medical decision making [6—8]. The 3D images or volumes we consider here consist of *region data* that can be defined as sets of (often connected) voxels (volume elements) in three-dimensional space that form 3D structures (or objects). We actually focus on 3D binary volumes, where information is provided only with respect to whether a particular voxel is part of a certain ROI or not (voxel values $\in \{0, 1\}$). This assumption is often made in medical image analysis applications, since it simplifies the processing without being very restrictive. Examples of binary ROIs in medical images are lesions, tumors, areas of brain activity, etc.

In this study, we are given a set of 3D medical image data and an assignment of these images to a number of classes based on certain non-spatial attributes (e.g., normal versus disease states). The objective is to derive a classification scheme that will correctly assign a new 3D image to a particular class (e.g., normal or disease) according to spatial information only. A specific example of this task from the brain imaging domain is the following: Given an MR image of the brain of a new subject that contains lesions, determine whether it belongs to a group of subjects who did or did not develop a particular disorder (e.g., attention-deficit hyperactivity disorder (ADHD) after closed head injury).

We suggest two approaches for automatic classification of ROIs and quantitative measurement of their levels of similarity. Unlike existing techniques where ROIs are considered individually (see Section 2 for an overview), we propose methods that classify ROIs based on their global spatial arrangement taking into account the co-existence of multiple ROIs. The methods presented here are based on measures of similarity between 3D spatial probability distributions. In particular, we suggest applying distance based techniques and maximum likelihood (ML) criteria to facilitate the classification of 3D ROI distributions. One of the main advantages of these approaches is that they can be applied directly on the 3D space preserving the spatial locality of the ROIs. Hence, we avoid the loss of information and complexity encountered in approaches originally developed for 2D slices (applied to pixels instead of voxels) that are repeated for each slice of a 3D volume.

We perform an evaluation of the proposed classification framework based on synthetic and realistic datasets. The realistic datasets conform to MRI studies and have been used in lesion-deficit analyses. In addition, we include experiments on clinical data to demonstrate the applicability of the proposed methodology in real-world problems. The clinical data are obtained from a study [9] on Alzheimer's disease (AD), consisting of fMRI contrast

activation maps and other associated clinical assessment. Part of the motivation for this work comes from the analysis performed as part of the Human Brain Project [10] and other initiatives, for the purpose of meta-analysis of data pooled from multiple studies and the detection of relationships between human brain structures and brain functions (i.e., human brain mapping).

The rest of the paper is organized as follows. In Section 2, background and related work in the area of classification of ROIs in medical images is summarized. This is followed by a description of the performed preprocessing and the proposed methodology based on the ML and distributional distances in Section 3. Experimental results are presented in Section 4. Discussion of the methods and the experimental results are presented in Section 5 while conclusions and future directions are presented in Section 6.

2. Background and related work

Necessary pre-processing steps, prior to any analysis of region data in medical images, are the segmentation and registration procedures. Image segmentation is required to delineate the particular regions (that are of interest) ensuring that image data are labeled consistently across samples. It can be performed manually, automatically, or semi-automatically. Extensive image segmentation work has been done in the medical imaging domain. Proposed methods can be divided into two broad groups: methods that incorporate prior spatial information and methods that incorporate solely signal-intensity based methods (see [11—13] for a review). Image registration deals with the existing morphological variability among samples and is vital to ensure that images are comparable across samples. The image registration is performed to bring the sample's image data into register, i.e., spatial coincidence, with a common spatial standard. The registration is done using normalization to a particular template and is necessary in order to determine whether two samples have ROIs in the same location. The methods employed for image segmentation and registration are often domain specific. In the rest of this study we assume that, prior to the analysis, the region data have already been segmented and normalized.

Applications of ROI classification in medical images range from the detection of electromagnetic field sources [14,15] to the analysis of fMRI activations [16—18]. For example, to select among texture and morphological ROI features, genetic algorithms were used and these were inputs to neural networks trained to classify mass and normal breast tissue [19]. Another example is the use of shape derived ROI features, such as compactness, Fourier descriptors, moments, and chord-length statistics to distinguish between circumscribed and speculated tumors [20].

In the clinician-in-the-loop approach to content based retrieval, after segmentation, the pathology bearing ROIs are characterized by a set of attributes (shape, texture, and other grey-level attributes) or by perceptual categories that domain experts rely upon for disease detection [21]. The challenges that image retrieval engines are confronted with when dealing with medical image collections are numerous [22]. To overcome some of these problems, a method was proposed that efficiently extracts a d-dimensional feature vector using concentric hyperspheres (spheres in 3D or circles in 2D) radiating out of the ROI's center of mass [23]. This results in obtaining a unique characterization signature for each ROI that can be used for further analysis.

In functional brain imaging, ROIs are usually delineated by using a thresholding approach. Voxels having values above a certain threshold are found and merged in order to construct informative activation regions. For example, in [24], analysis of functional image data was done by paying attention to their functional and spatial characteristics, such as ROIs of high activity and their relative positions to each other and to the rest of the image. A neural-network classifier based on coarse ROI analyses was used in [25] to classify normal versus abnormal PET scans. A neural network was also employed to analyze and classify single photon emission computerized tomography (SPECT) datasets from healthy and patient subjects with AD [26].

Most of the previous approaches deal with ROIs individually, without considering the distribution of their spatial arrangement. Efforts that consider this information include computing the divergence between probability distributions based on the Kullback—Leibler (KL) approach [18]. In this case, a brain activation map was constructed after the analysis of fMRI signal. Also, the likelihood that particular voxels exhibit significant changes between conditions has been estimated, using statistical tests [27].

Methods such as the statistical parametric mapping [28—30] are of a great value in the analysis of fMRI activations but they do not automatically classify or compare the activation ROIs. Data mining methods have been recently applied to brain images to discover associations between binary lesion ROIs and deficits [31]. However, little work has been done in brain image data classification.

In summary, previous work in classification of ROIs in medical images mostly focuses on shape, texture

and other attributes of ROIs. ROIs have been considered individually, i.e., not in connection with other ROIs that may be present in the same image. Moreover, the various statistical techniques that have been applied to the analysis of medical images mostly focus on detecting voxels that exhibit significant changes between conditions, without actually performing classification based on the ROI information. We seek to overcome these problems by proposing a unified framework for 3D medical image classification based on ROI analysis. We propose classifying ROIs in medical images utilizing the properties of their 3D spatial distribution. We apply statistical techniques based on distributional distances as well as on ML methods. These approaches preserve the spatial locality of the voxels in the 3D space, avoiding the loss of information encountered in 2D per slice analysis of 3D image data. Also, by focusing on ROI distributions, we avoid the multiple comparison problem (which occurs when computing a statistic for many pairwise tests) that voxel-based statistical analysis introduces. Finally, by proposing a unified classification framework we introduce medical image analysis tools that can assist medical decision making and diagnosis.

3. Methodology

Our goal is classification of 3D medical images based on the analysis of ROIs' spatial distribution. Classification in this case refers actually to the process of characterizing the spatial arrangement of ROIs of a new subject and comparing it to the distributions of labeled historical data. An example is comparing the 3D image of a new subject to the ones of subjects in the normal and in the diseased group. The two approaches we consider are applicable directly on the 3D domain (voxel-based). The first one utilizes distributional metrics such as the Mahalanobis distance and the KL divergence. The second approach employs maximum likelihood.

Formally, the ROI distribution classification task is stated as follows. Let r_{xyz} denote the value of a *voxel* (volume element) of a 3D medical image. A voxel has a value $r_{xyz} = 1$ if it belongs to a ROI (such voxels are subsequently referred to as "ROI voxels") and $r_{xyz} = 0$ otherwise. We consider voxels that belong to the ROI by observing the distributions of the voxel coordinates. Given two sets S_X and S_Y that contain coordinates of ROI voxels for N subjects that belong to either one of two distinct classes (i.e., normal and disease states), the task is to identify whether a dataset $S_Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_{n_Z}\}$, that contains coordinates of n_Z ROI voxels of a new subject, comes from the same distribution p_Y as the set S_Y or the distribution p_X of the set S_X. Let n_X and n_Y stand for the numbers of ROI voxels in S_X and S_Y, respectively.

Prior to applying distribution characterization and classification algorithms, we perform segmentation to delineate the ROIs and registration to a standard spatial template to bring homologous regions in spatial coincidence, i.e., we spatially normalize the images to make them comparable to each other. We propose using SPM99 [28] for this purpose. The templates supplied by SPM99 conform to the space defined by the ICBM, NIH P-20 project. An approximation of this space is described by the atlas of Talairach and Tournoux [32]. The basic spatial registration technique employs resampling of the image voxels, while minimizing the sum of squares between the image and the chosen template. This is performed with affine and quadratic automated algorithms. In the rest of this study we assume that, prior to the analysis, the region data have already been segmented and normalized.

The proposed 3D image classification framework employs statistical methods for classifying the 3D spatial distributions of voxel-based ROIs. The steps of the proposed methodology are the following:

(1) Estimate the probability distribution of ROIs in the 3D space for each labeled class of historical data (e.g., control versus patient). For this purpose we estimate the mean and covariance matrix of data, or apply the expectation-maximization (EM) algorithm [27,29] and its variant, the *k*-means algorithm [33].
(2) Given a 3D image belonging to an unknown sample, characterize the corresponding distribution of ROIs. For this process we apply either distributional distance-based methods (such as the Mahalanobis distance [27] and the KL divergence [34]), or maximum likelihood methods.
(3) Classify the new 3D image according to the characterization of the ROI distribution. When using distributional distance metrics assign the label of the class being closer to the ROI distribution of the new sample. When using ML methods assign the label of the class whose distribution is most likely to have generated the new sample.

Details of these steps are described in the rest of this section.

3.1. Estimation of ROI distribution densities

In general, a probability density function of a random variable can be estimated using parametric,

non-parametric or semi-parametric techniques [35]. In this study, we use semi-parametric techniques [36,37] that can provide flexibility (by allowing a very general class of functional forms for the estimated probability density function) and controllable complexity of the chosen functional form to avoid overfitting. We consider the use of the EM algorithm and its variant, the *k*-means algorithm [36,37], to estimate the distributions in form of Gaussian mixtures (see Appendix A for the definition).

3.1.1. The EM approach

The EM approach presented in Fig. 1, computes maximum likelihood estimates of mixture parameters [24,29]. The algorithm maintains probabilities P_{ij} that a data example $\mathbf{z}_i$ belongs to the j-th distributional component (j = 1, ..., k). In each iteration, the expectation (E) step is applied to compute expected probabilities P_{ij} based on the values of the distribution parameters (prior probabilities π_j, component means $\boldsymbol{\mu}_j$ and component covariance matrices Σ_j) from the previous iteration. The E step is followed by the maximization (M) phase, where new values of the parameters are computed to maximize the likelihood, based on previous parameter values and estimated values of P_{ij}. The time complexity of the EM algorithm is $O(Lkn)$ where L is the number of iterations, n is the number of ROI voxels in dataset S (where S can stand for S_X, S_Y or S_Z) k is the number of distributional components.

3.1.2. The *k*-means approach

The *k*-means approach presented in Fig. 2, avoids the costly computation of probabilities P_{ij} in the E step by assigning each data example $\mathbf{z}_i$ to a Gaussian component, whose mean $\boldsymbol{\mu}_j$ is the closest to $\mathbf{z}_i$ based on the Euclidean distance (see Eq. (A.8) in Appendix A for definition) [33]. This is equivalent to the partitioning of a dataset S (containing the examples $\mathbf{z}_i$) into k clusters. In the M phase of the *k*-means, the means $\boldsymbol{\mu}_j$ are recomputed by averaging the examples assigned to each cluster. After the convergence, the *k*-means algorithm minimizes the average Euclidean distance $(1/n)\sum_{\mathbf{z}_i \in S} \min_j d_E^2(\mathbf{z}_i, \boldsymbol{\mu}_j)$ between the vectors $\mathbf{z}_i$ and the cluster means closest to them (d_E stands for the Euclidean distance, n stands for n_X, n_Y, n_Z). The computational cost of *k*-means is $O(Lkn)$ as well.

Both proposed methods require specifying the number of mixture components k. In this paper, we determine the number of mixture components using the log-likelihood principle [38]. Other advanced techniques exist for the selection of k, (e.g., based on minimum-description length [39], I2 criterion function [40]); however, a more complete study is outside the scope of this paper.

3.2. Classification with statistical distance based metrics

Distance-based methods rely on an appropriately defined distance measure between two distributions. In this study, the distances are computed

Given: Examples (data vectors) $z_i, i = 1,...,n$; the number of distribution components k; convergence criterion; the number of algorithm iterations

Initialization: Initialize k component means $\boldsymbol{\mu}_j$, j = 1,…,k to take values of randomly chosen examples from the given data vectors; Initialize prior probabilities as π_j=1/k. Initialize covariance matrices to be equal to the covariance matrix of the given dataset.

Repeat:

Expectation step:

$$P_{ij} = \frac{\pi_j f_j(z_i)}{\sum_{j=1}^{k} \pi_j f_j(z)}, j = 1,...,k; i = 1,...,n$$

Maximization step:

$$\boldsymbol{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} P_{ij}\mathbf{z}_i, j = 1,...,k;$$

$$\Sigma_j = \frac{\frac{1}{n}\sum_{i=1}^{n} P_{ij}(\mathbf{z}_i - \boldsymbol{\mu}_j)(\mathbf{z}_i - \boldsymbol{\mu}_j)^T}{\pi_j}, j = 1,...,k;$$

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} P_{ij}, j = 1,...,k;$$

Until pre-specified maximal number of iterations is reached.

Figure 1 The outline of the EM algorithm.

```
Given:   Examples (data vectors);
         the number of clusters k;
         a convergence criterion;

Initialization: Initialize k seed vectors (e.g. by taking a random subset of
examples)

Repeat

     Assign each example to the nearest seed (creating k distinct subsets).

     Compute a new seed for each subset by averaging examples in the subset.

Until max number of repetitions.

Estimate covariance matrices of mixture components as sample covariance matrix
corresponding to each subset.

Estimate distribution prior probabilities as fractions of examples that belong
to each subset.

Return parameters of the estimated Gaussian mixture (means, covariance matrices
and priors).
```

Figure 2 The scheme of the *k*-means algorithm.

between a new dataset (subject) S_z and each of datasets that correspond to considered classes (distributions) S_X or S_Y in order to determine to which existing distribution a new subject S_z is closer to. Here, we consider the Mahalanobis distance [41], and the KL divergence [39].

3.2.1. The Mahalanobis distance approach

We employ the Mahalanobis distance (see Appendix A) to quantify the similarity between the new subject S_z and an existing dataset S (S_X or S_Y). Given datasets S_z and S, the Mahalanobis distance, d_M, between them is computed as:

$$d_M = \sqrt{(\boldsymbol{\mu}_{S_z} - \boldsymbol{\mu}_S)^T \Sigma^{-1} (\boldsymbol{\mu}_{S_z} - \boldsymbol{\mu}_S)}, \tag{1}$$

where $\boldsymbol{\mu}_{S_z}$ and μ_S are mean vectors of the datasets S_z and S, respectively, and Σ is the sample covariance matrix [42]:

$$\Sigma = \frac{(n_z - 1)\Sigma_{S_z} + (n - 1)\Sigma_S}{(n_z + n - 2)}, \tag{2}$$

with Σ_{S_z} and Σ_S denoting estimated covariance matrices of the datasets S_z and S, respectively, and n_z and n denoting the size of the datasets S_z and S, respectively.

The computational time $\text{Time}_{\text{total}}$ of the Mahalanobis distance approach (as well as other methods proposed in this paper) consists of *learning* time, $\text{Time}_{\text{learning}}$ necessary for distribution analysis and *query* time $\text{Time}_{\text{query}}$, needed for classification of a new dataset. The learning time $\text{Time}_{\text{learning}}$ is directly proportional to the time needed for computing the covariance matrices of datasets, which is in turn proportional to the sizes of the data sets S_X and S_Y. In general, the time complexity of the Mahalanobis distance method also depends on the number of dimensions, but in our case this number is always three and does not significantly influence the total computational time. Since the classification involves estimation of covariance matrix Σ_{S_z}, the query time $\text{Time}_{\text{query}}$ is linearly proportional to the size of the dataset S_z. Hence the computational complexity of the method is:

$$\begin{aligned}\text{Time}_{\text{total}} &= \text{Time}_{\text{learning}} + \text{Time}_{\text{query}} \\ &= O(n_X + n_Y) + O(n_z).\end{aligned}$$

3.2.2. The KL divergence approach

Let $p_z(\mathbf{x})$ and $p(\mathbf{x})$ be probability densities corresponding to the distributions intrinsic to the datasets S_z and S, respectively (hence, $p(\mathbf{x})$ can be $p_X(\mathbf{x})$ or $p_Y(\mathbf{x})$). Unlike the Mahalanobis distance, the KL divergence $d_{\text{KL}}(p(\mathbf{x}), q(\mathbf{x}))$ [43] is computed directly between the estimated probability densities of the distributions corresponding to the new subject S_z and to the existing data distribution S (corresponding to datasets S_X or S_Y) as:

$$d_{\text{KL}}(S_z, S) = \int_D p_z(\boldsymbol{x}) \log \frac{p_z(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}. \tag{3}$$

Since the datasets S_X, S_Y and S_z obtained from medical imaging or simulation contain coordinates of discrete volumes—voxels, we use a discrete

approximation to compute the KL divergence (Eq. (3)) as

$$d_{\mathrm{KL}}(S_z, S) \approx \sum_{\text{all voxels } v_{i_1,i_2,i_3}} p_z(\boldsymbol{x}_{i_1,i_2,i_3}) \log \frac{p_z(\boldsymbol{x}_{i_1,i_2,i_3})}{p(\boldsymbol{x}_{i_1,i_2,i_3})} \Delta \boldsymbol{x}. \tag{4}$$

here, $p_z(\boldsymbol{x}_{i_1,i_2,i_3})$ and $p(\boldsymbol{x}_{i_1,i_2,i_3})$ are estimated probability densities of the involved distributions at the voxel centers $\boldsymbol{x}_{i_1,i_2,i_3}$, and $\Delta\boldsymbol{x}$ is the product of corresponding discretization intervals (see Appendix A for more details). The learning time of the KL divergence approach depends on the complexity of the algorithm used to estimate distribution density (see Section 3.2.1). When using L iterations of EM or k-means (as proposed in this paper), the learning time is $\mathrm{Time}_{\mathrm{learning}} = O(Lk(n_X + n_Y))$. The query time, when using Eq. (4) to compute the KL divergence, is directly proportional to the number n_z of ROI voxels in the dataset S_z (resulting in $O(Lkn_z)$ to estimate the distribution $p_z(\mathbf{x})$) and to the number B of total voxels in the volume D (including voxels belonging to ROIs). Hence, the total computational complexity is:

$$\begin{aligned}\mathrm{Time}_{\mathrm{total}} &= \mathrm{Time}_{\mathrm{learning}} + \mathrm{Time}_{\mathrm{query}} \\ &= O(Lk(n_X + n_Y)) + O(Lkn_z + B).\end{aligned} \tag{5}$$

Similar to that in Section 3.1, here L is the number of iterations for the distribution estimation algorithm and k the number of predefined clusters.

3.3. Classification with maximum likelihood (ML) methods

The ML methods are based on the computation of the likelihood-probability that the particular dataset is observed upon condition that a pre-determined hypothesis is satisfied [39]. The ML methods result in the hypothesis that maximizes this conditional probability. In the observed case of distribution classification, the hypothesis states that the new dataset belongs to one of the observed distributions. Hence, given an observed dataset $\mathcal{D}$ and a collection of hypotheses $H = \{h_1, h_2, \ldots, h_H\}$, ML chooses the hypothesis h_{ML} such that:

$$h_{\mathrm{ML}} = \arg\max_{h \in H} P(\mathcal{D}|H). \tag{6}$$

The maximization of likelihood is equivalent to the maximization of its logarithm—a log-likelihood—such that:

$$h_{\mathrm{ML}} = \arg\max_{h \in H} \log P(\mathcal{D}|H). \tag{7}$$

Given a new dataset S_z and estimated probability densities of the existing distributions $p_X(\mathbf{x})$ and $p_Y(\mathbf{x})$, we estimate a likelihood that a probability distribution $p_z(\mathbf{x})$ of S_z is the same as one of the existing distributions S_X or S_Y. The new dataset S_z is assigned the class label of the distribution that maximizes the likelihood. More formally,

$$S_{\mathrm{classified}} = \arg\max_{S} P(S_z|S),$$

$$S_{\mathrm{classified}} \in \{S_X, S_Y\}.$$

We assume that the positions of ROI voxels, observed as random variables through the coordinates $\mathbf{z}_i$, are independent of each other such that the following holds:

$$P(S_z|S) = \prod_{\mathbf{z}_i \in S_z} P(\mathbf{z}_i|S). \tag{8}$$

In this case, it is suitable to apply the maximum log-likelihood criterion (7) which combined with Eq. (8) leads to the following computationally more convenient equation:

$$S_{\mathrm{classified}} = \arg\max_{S} \sum_{\mathbf{z}_i \in S_z} \log P(\mathbf{z}_i|S). \tag{9}$$

Similar as for the KL method, the learning time of the ML approach depends on the complexity of the algorithm used to estimate distribution density—$O(Lk(n_X + n_Y))$ when EM (or k-means) is employed. The query time is equal to the time to compute the likelihood, which is linearly proportional to the size of the dataset S_z (see Eq. (9)) but also to the number of distributional components of S_X and S_Y (that need be evaluated to compute $P(\mathbf{z}_i|S)$). Hence, the total computational complexity in this case is:

$$\begin{aligned}\mathrm{Time}_{\mathrm{total}} &= \mathrm{Time}_{\mathrm{learning}} + \mathrm{Time}_{\mathrm{query}} \\ &= O(Lk(n_X + n_Y)) + O(kn_z).\end{aligned} \tag{10}$$

4. Results

The proposed classification framework was experimentally evaluated on synthetic data (mixtures of Gaussian distributions), on realistic brain lesion-deficit data generated by a simulator [33] conforming to a clinical study [44], and on real fMRI brain activation distributions obtained from a study that explores neuroanatomical correlates of semantic processing in Alzheimer's disease [45]. These datasets as well as the experimental results are described below.

4.1. Experiments with synthetic data

Synthetic data used in our experiments contained samples from two mixtures of nine normal distributions (the data is available at http://denlab.temple.edu/data_repository). We varied the

parameters (means and variances) of mixture components, constructing different distributions S_X and S_Y (see Figs. 3 and 4). As described in our Section 3, we initially estimated the class distributions (i.e., we computed the distributions means and covariance matrices for the Mahalanobis approach and estimated the probability densities for the other proposed techniques). We randomly chose one of the distributions (S_X or S_Y), generated a new 3D sample S_z according to the chosen distribution, and predicted the class of the dataset using each of the proposed classification approaches. The classification error was estimated as the number of incorrectly predicted new samples divided by the total number of the S_z samples

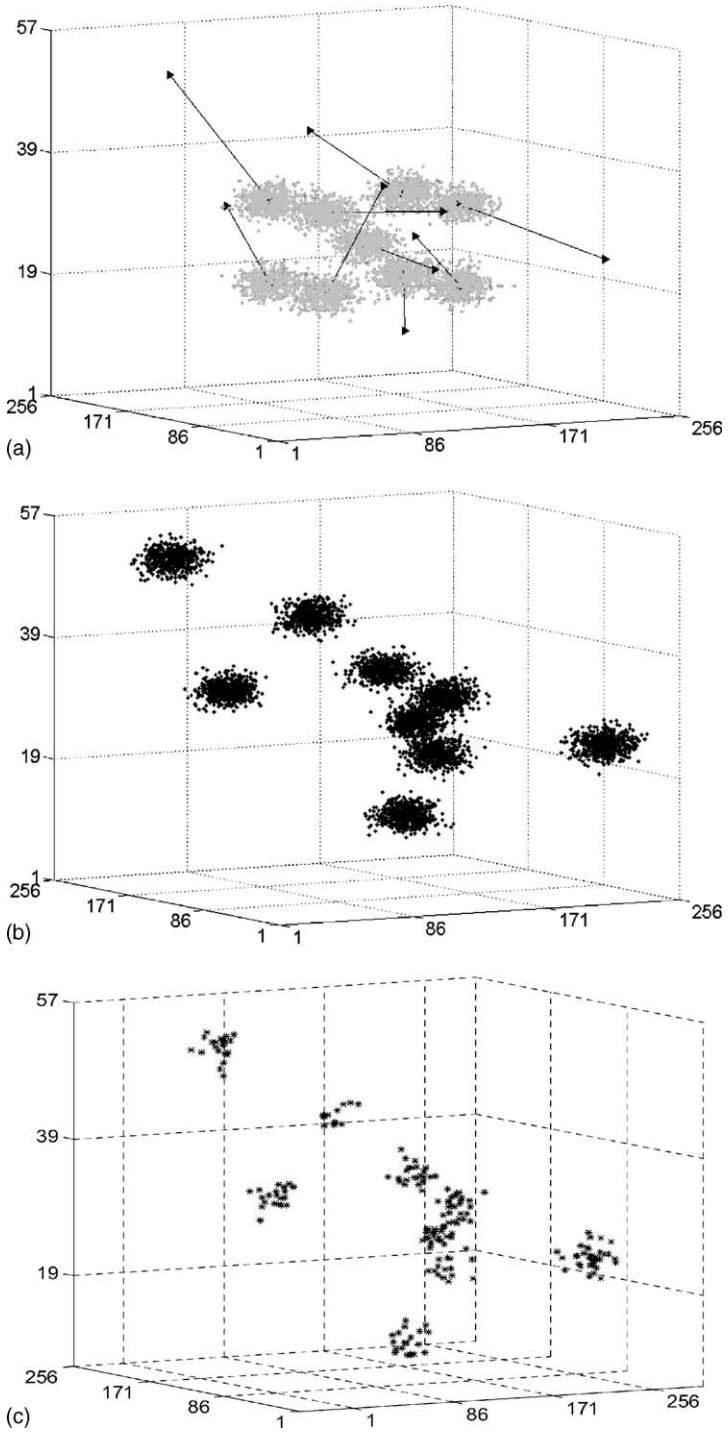

Figure 3 Mixtures of distributions that differ in means of components with random shifts of the distributional means: (a) distribution S_X; (b) distribution S_Y; and (c) New examples S_z consisting of 200 voxels drawn from the distribution S_Y.

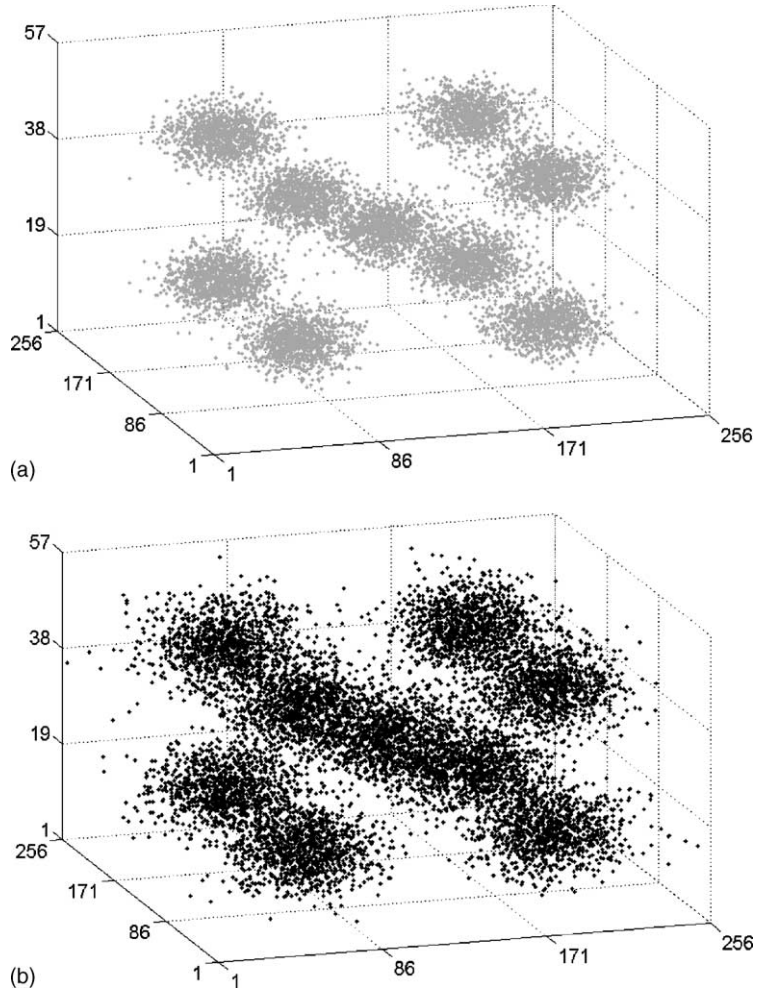

Figure 4 Two distributions mixtures that differ only in variance of the distribution components: (a) distribution S_X and (b) distribution S_Y.

generated from the same distribution. We repeated the classification task 200 times. During these experiments, we varied the number of examples in distributions S_X and S_Y as well as the number of voxels n_z of the dataset S_z.

4.1.1. Gaussian mixtures with different means

In the first series of experiments, the distribution components had the same variances but different means for each class. As demonstrated in Fig. 3, each of the nine component means of the distribution S_Y (Fig. 3b) was shifted for a random value as well as in a random direction with respect to the means of the distribution S_X components (Fig. 3a). The standard deviation of this shift was 63% (corresponding to 40% variance) of each linear dimension of the domain D (length, width, and depth). The number of examples in distributions S_X and S_Y was varied from 1 to 50 (with 200 ROI voxels per example), while the dataset S_z size n_z was varied from 50 to 1000. Fig. 3c illustrates a dataset S_z of the size n_z = 200 that corresponds to the distribution S_Y.

The prediction error of all considered classification methods decreased with the size n_z of the dataset S_z and with the number N of examples in distributions S_X and S_Y. Since the class distributions of experimental data were clearly distinguishable (see Fig. 3a and b), the ML technique for estimating Gaussian mixtures could provide very accurate results (e.g., prediction error less than 1% when mixtures of k = 9 distributions were estimated, regardless of the algorithm used to estimate the underlying distributions—EM or less complex k-means). Similarly, the technique employing the KL divergence achieved almost perfect classification for all considered sizes N and n_z (the prediction error was less than 2% for mixtures of nine distributions). The Mahalanobis distance method provided smaller classification accuracy in comparison to other methods (classification error between 1 and 10%).

4.1.2. Gaussian mixtures with same means and different variances

Another group of experiments on synthetic data involved mixtures that had the same component means but different component variances for each class. Fig. 4 illustrates an example of such distributions S_X and S_Y, where the component variances of S_X (Fig. 4a) were twice smaller than the component variances of the second mixture S_Y (Fig. 4b). As we can see, the smaller variance, the higher concentration of samples around the component means (and the smaller dispersion), but the distributions were generally very similar. Therefore, in these experiments, classification was typically more chal-

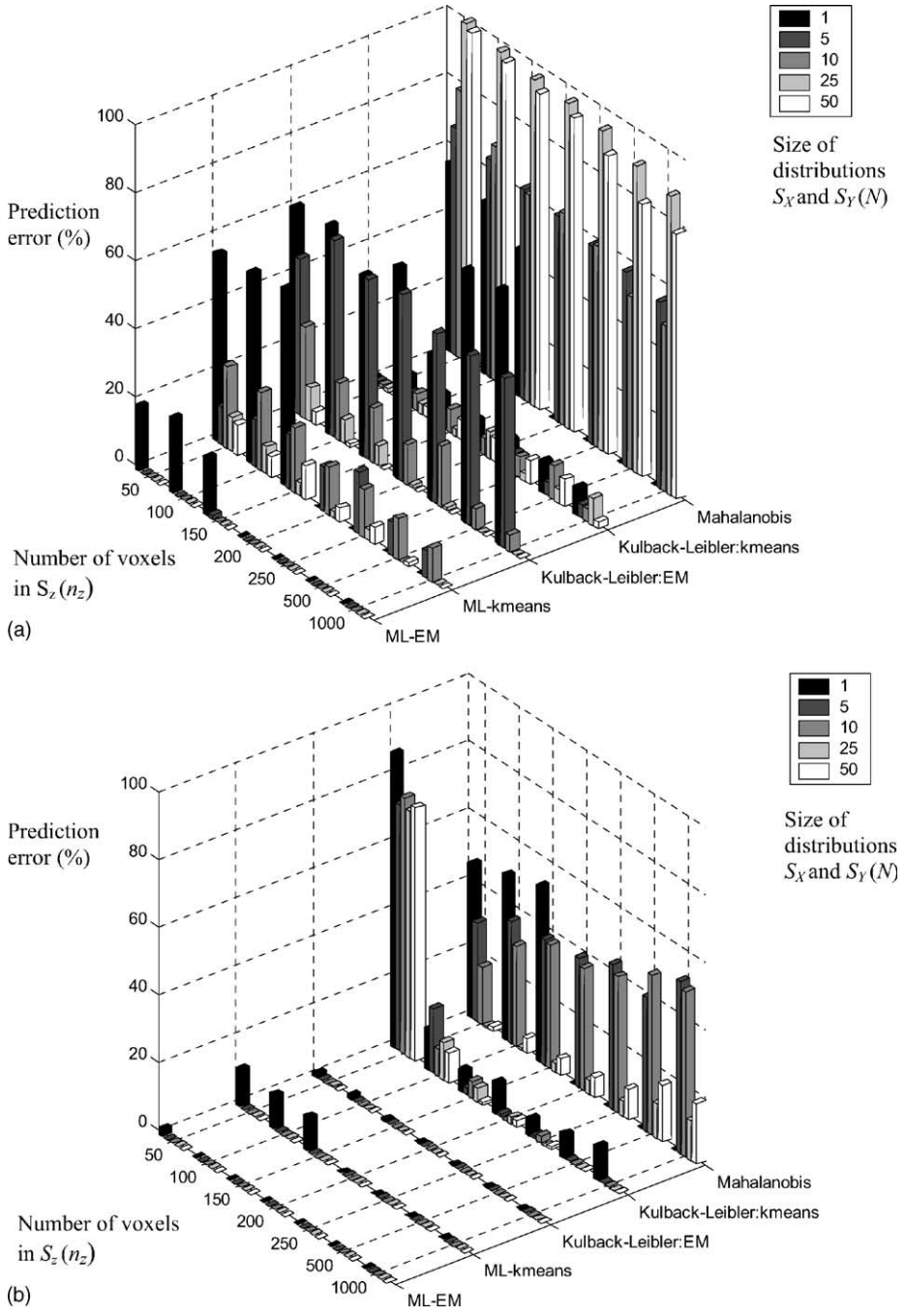

Figure 5 The prediction error when classifying new examples S_z from two distributions with different component variances using the proposed techniques. In ML and KL methods, mixtures of nine distributions were estimated. The variance of distribution components was 0.01 and 0.1. (a) S_z belongs to distributions p_X with the smaller variance and (b) S_z belongs to distributions p_Y with the larger variance.

lenging since the distributional components with the smaller variances (corresponding to the probability distribution p_X) were often overshadowed by the mixture components with the larger variances (corresponding to p_Y). We used $k = 9$ mixture components for probability distribution estimation. All the experimental results were obtained when standard deviations of the distributions components of p_X and p_Y were 10 and 31% of the linear dimensions of the domain D. These results are shown in Fig. 5.

Applying the Mahalanobis distance in this dataset did not perform very well. The error in general was very high when predicting the mixture with the smaller variances, and significantly lower when predicting the one with the larger variances. This is because with S_X and S_Y having the same means, the Mahalanobis distance (Eq. (1)), becomes predominantly dependent on the sample covariance matrix Σ. Consequently, the method tends to predict the example as belonging to the distribution with the larger variance regardless of the actual class the example belongs to and the accuracy did not increase with the training set size N.

Similarly to the Mahalanobis distance techniques, the methods based on the KL divergence were more successful in predicting new examples S_z when they belonged to the distributions p_Y with the larger variances (prediction error less than 1%—see Fig. 5). However, when predicting examples from the distribution with the smaller variance the performance was much better than that of the Mahalanobis (see Fig. 5). For larger sizes of the training sets similar performance was achieved regardless of whether k-means or EM was used for distribution estimation.

The ML based methods that used the EM algorithm for estimating the underlying Gaussian distributions were capable of providing useful classification as well. The prediction error dropped to less than 1%, for sufficiently large sizes of the distributions S_X and S_Y. Similarly to the other techniques, the classification task was particularly challenging when predicting samples from the smaller variance distribution. The prediction error varied from 50% to 2% for the larger sizes of the training set N and the data sample n_z. However, when classifying examples belonging to

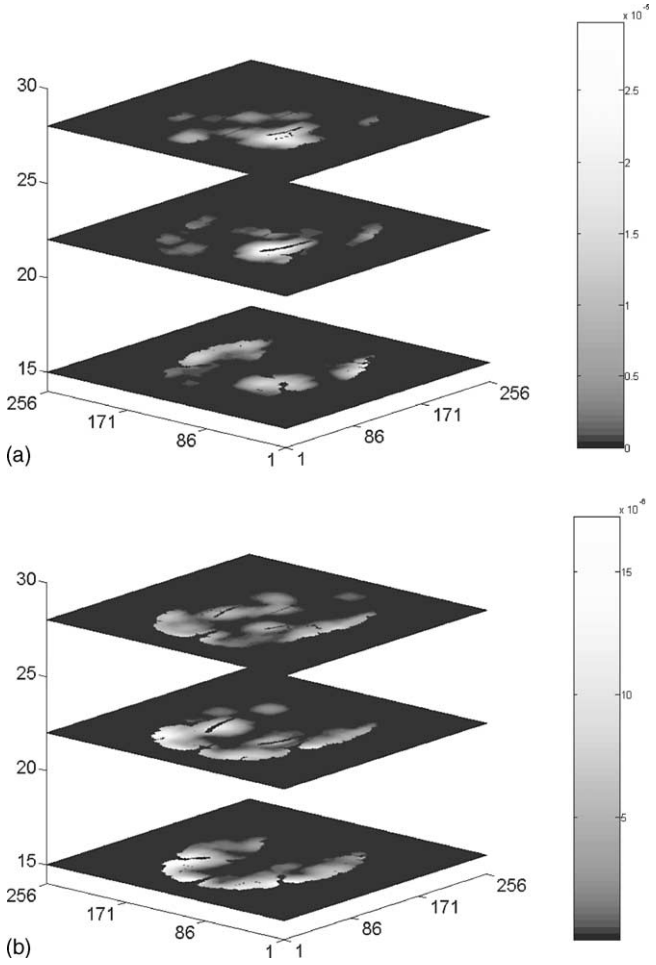

Figure 6 Distributions of (a) "Yes ADHD" and (b) "No ADHD" classes.

the distribution p_Y with the larger variances, the ML methods were more successful; the prediction error dropped to less than 1% for new samples with large numbers of voxels.

4.2. Experiments with realistic data

We performed classification of realistic brain lesion distributions (available at http://denlab.temple.edu/data_repository) that were generated using a lesion-deficit simulator (LDS) [33] with the spatial statistical model conforming to the Frontal Lobe Injury in Childhood (FLIC) study [44]. In implementing the LDS we used probability distributions to model the number, size, and spatial distribution of lesions, as well as registration error and structure–function associations. From these parameters, the simulator generates a complete dataset, including spatial lesions and clinical variables. The LDS has been successfully used in evaluation and scalability testing of data mining techniques and to provide an approximate number of subjects needed to discover certain associations so that medical experiments can be planned accordingly [33]. The segmentation of ROIs in the FLIC study was performed manually by a neuroradiologist using thresholding. A non-linear method based on a 3D elastically deformable model [46] was used to register the ROIs to the Tailarach anatomical atlas [32]. After a normalization of image data to a common coordinate system with a resolution of $256 \times 256 \times 57$ voxels, we applied the classification methods proposed in Section 3 to lesion-deficit analysis and magnetic resonance imaging datasets. The samples (subjects in this case) were classified into two classes according to subsequent development of ADHD after closed head injury. Therefore, there were two distributions corresponding to subjects who developed ADHD (''Yes ADHD'' class) and who did not develop ADHD (''No ADHD'' class) (Fig. 6). Given a new subject with a set of lesioned voxels, the goal was to determine the more plausible class. The subjects contained a number of lesioned voxels that varied from 50 to 500, although in the specific FLIC study [44,47] approximately 200 voxels of region data were present on average per a 3D brain image (i.e., per subject).

The experimental design used here was similar to that described in Section 4.1. We varied both the size N of datasets for the classes and the number n_z of ROI voxels belonging to a new subject. For each combination of these parameters, we performed the experiments through a predetermined number of repetition rounds (200 in our experiments). Each round consisted of random drawing of a new subject from one of the classes. The classification performance was again measured by computing average accuracy, as done in Section 4.1 for synthetic data. In reported experimental results, the number of mixture components k in the ML and the KL divergence methods was 3.

The experimental results in this case showed that Mahalanobis distance could provide more reliable and more accurate classification for the ADHD data, compared to the samples from synthetic distributions described in Section 4.1, as Fig. 7 demonstrates. Classification with prediction error less than 10% was possible both for the subjects who did and who did not develop ADHD when sufficient knowledge of the distributions corresponding to the classes was available (sets S_Y, S_X large enough). This was apparent especially when 150 or more voxels of region data were available for a new subject. The prediction was perfect ($<$1% error) when the number of voxels of region data for a new subject was larger than 1000. It is interesting to notice that the classification accuracy was slightly better when predicting subjects in the ''Yes ADHD'' class than in the ''No ADHD'' class.

The method based on the KL divergence was again more successful than the Mahalanobis distance in classification of new subjects (Fig. 7). When the EM algorithm was used, the prediction error was less than 2% for subjects with number of lesioned voxels n_z small as compared to the size N of training datasets. Perfect classification was achieved for sufficiently large number of lesioned voxels ($<$1% prediction error). Interestingly, when k-means was used to estimate distributions, the accuracy was even better than when EM was used.

Finally, when applying the ML based methods on realistic data, for all the combinations of the datasets size and the number of voxels of region data in a new subject, the achieved prediction error was less than 1% regardless of the algorithm employed for estimating the underlying distributions (k-means or EM) (Fig. 7). The difference in prediction capability when employing k-means or EM was negligible.

4.3. Experiments with clinical data

To evaluate the applicability of the proposed approach on clinical data, we experimented with real fMRI activation datasets. The ROIs in this case are brain areas that are being activated when a certain task is performed. We analyze the spatial arrangement of ROIs corresponding to high activation levels in 3D fMRI scans. These were obtained from a study designed to explore neuroanatomical correlates of semantic processing in Alzheimer's disease. More specifically, the dataset consists of 3D activation contrast maps of nine controls and nine Alzheimer's disease patients on a category-

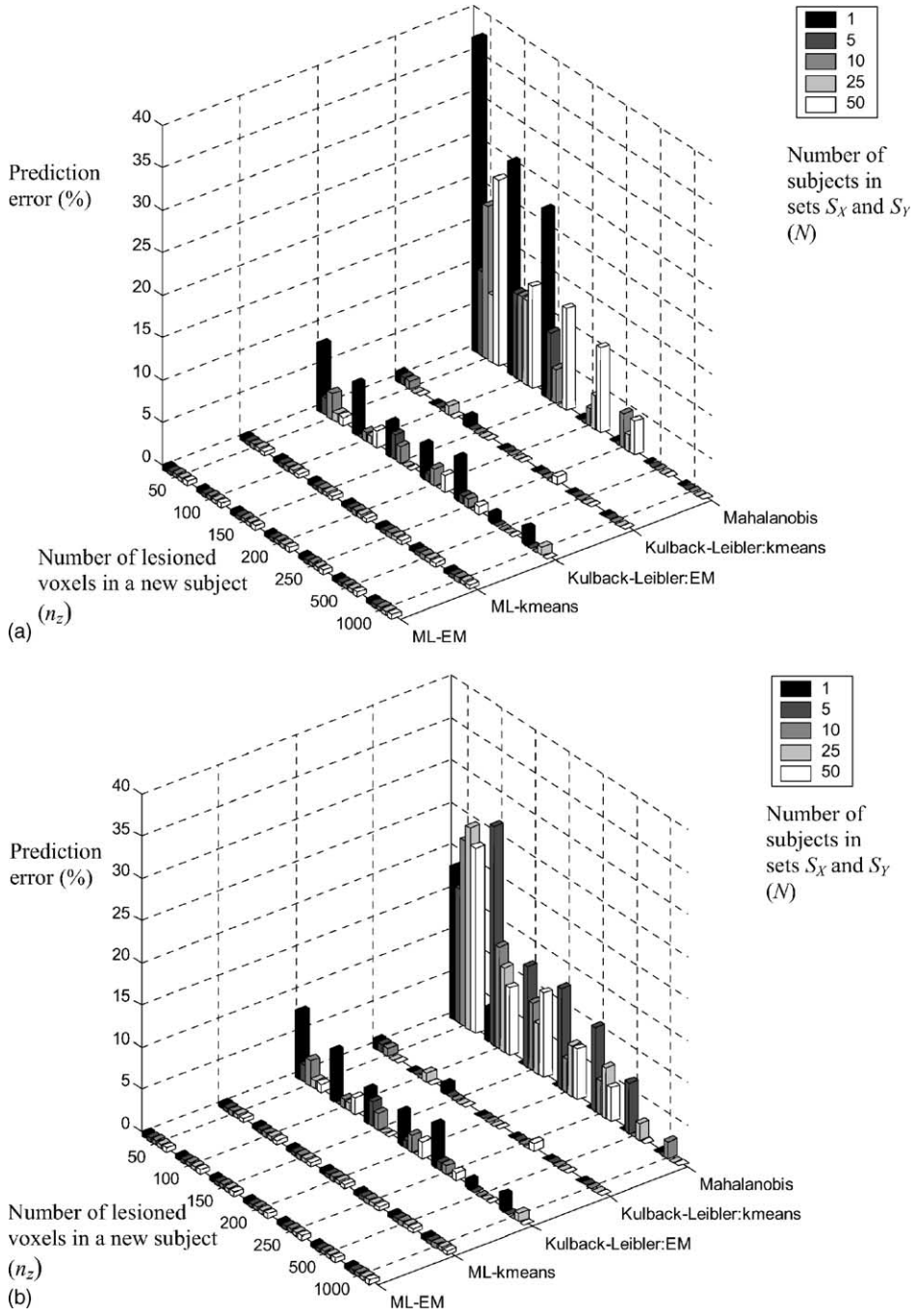

Figure 7 The prediction error when classifying new subjects from ADHD realistic data. (a) Subjects who belong to the class ''Yes ADHD'' and (b) subjects who belong to the class ''No ADHD''.

exemplar word pair (the preprocessed dataset is available at http://denlab.temple.edu/data_repository). A contrast map is a 3D activation map that measures the difference in activation observed in two different states—usually between rest and activity while performing a certain task.

The task consisted of an auditory presentation of word pairs (categories and possible exemplars) requiring a semantic decision (match—mismatch) [45]. Each subject was tested with the same timing and word set with a blocked design. The word pairs were presented in groups of four at 7.0 s intervals, with each 28.0 s block of decision followed by a 10.5 s period of rest. Scans were conducted at 1.5 T using a single shot, gradient echo, echo planar functional scan sequence (TR = 3500 ms, TE = 40 ms, interleaved, FOV = 24 cm, slice thickness = 6 mm, NEX = 1, flip angle = 90) on a General Electric Signa scanner with a multi-axial local gradient head coil system (Medical Advances, Inc., Milwaukee,

WI). Scans consisted of 20—23 contiguous sagittal slices in a 64 × 64 matrix with in-plane resolution of 3.75 mm^2 (total slice acquisitions per run = 1920 scans) with anatomical reference images in the same slice locations using a T1-weighted spin-echo pulse sequence (TR = 450 ms; TE = 17 ms; interleaved; matrix = 256 × 192; NEX = 1; the same FOV, slice thickness, and locations as the functional scans). All scans for each subject were acquired in the same session.

4.3.1. Data preprocessing

Prior to the application of the proposed technique, we applied preprocessing to bring homologous regions into spatial coincidence through spatial normalization. The spatial normalization of the scans to a standard template brain using the anatomical reference images was carried out in SPM99 [28], resulting in resampling of the data to 2 mm^3 isotropic voxels. The resampled data were smoothed with a Gaussian filter (FWHM 15 mm^3). Each subject's task-related activation was analyzed individually versus the subject's rest condition, resulting in individual contrast maps giving a measurement of fMRI signal change at each voxel. To reduce the effect of noise and sensor fluctuations in the original functional data we applied the following steps. First, we removed the effect of the background noise by subtracting the signal value measured in representative background voxels from all the voxels of the 3D volume. Second, we masked the data using a binary mask extracted from the T1 anatomical atlas used as the template the data were spatially registered to. Only the signal within the binary mask was included in the analysis.

To segment ROIs that reflect the distribution of highly activated regions we followed a procedure of utilizing an activation cut-off threshold [9]. The threshold was based on t-values measurement, indicating each voxel's significance of activation. For this purpose we used the t-maps referring to the contrast dataset (output of SPM99). The cut-off value of t = 2.12 was used, which corresponds to the p-value of 0.025 for one-tailed significance. The result of this final step was to obtain the 3D binary volumes used in our experiments, consisting of ROIs that reflect highly significant activation regions. Fig. 8 shows a sample slice of a contrast map along with the corresponding t-map and the binary image that is obtained after the thresholding. The aim of applying the proposed distribution estimation techniques to this dataset was to distinguish the spatial arrangement of ROIs of healthy subjects compared to that of Alzheimer's. Distributions of peak activation obtained after thresholding of contrast maps

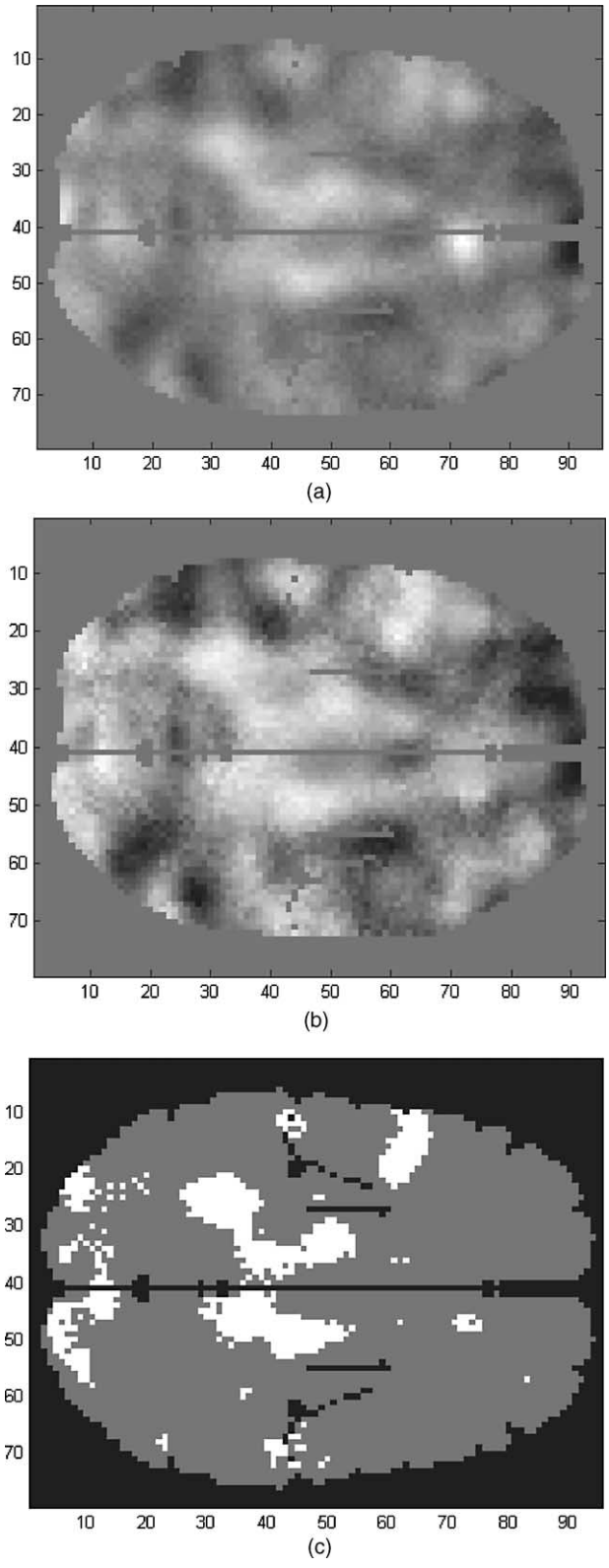

Figure 8 A sample slice of (a) a contrast map; (b) a corresponding t-map; and (c) a binary image obtained after thresholding.

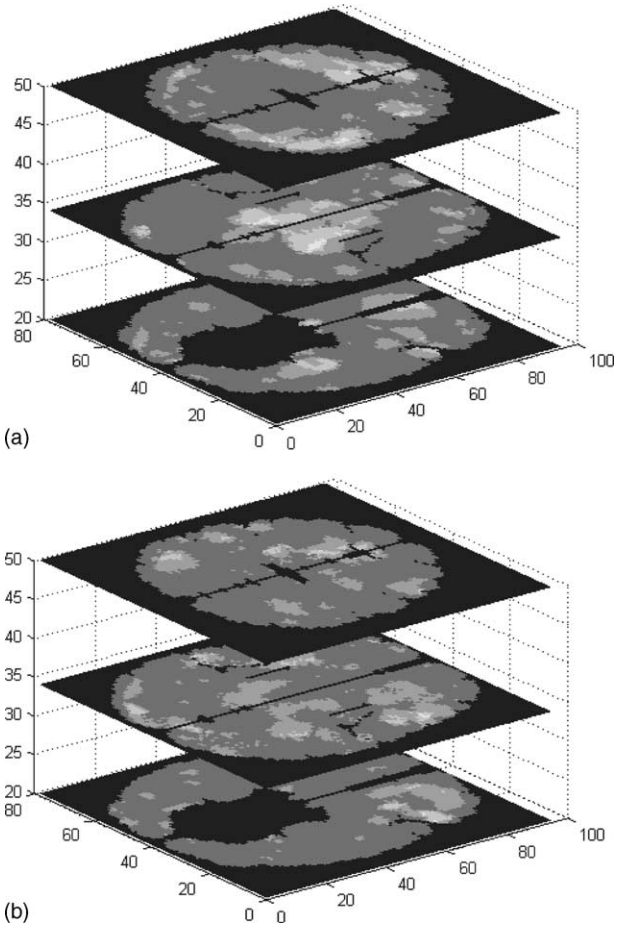

Figure 9 Distributions of peak activation obtained after thresholding of contrast maps for (a) controls and (b) Alzheimer's patients.

for both controls and Alzheimer's patients are shown in Fig. 9.

4.3.2. Classification results

The leave-one-out approach was employed to evaluate out-of-sample classification performance [39,41]. More specifically, the training set consisted of patients and controls with indices 1, 2, 3, ..., $i-1$, $i+1$, ..., 9 and the method was tested on patient and control with an index i, where $i = 1, \ldots, 9$. Due to the stochastic nature of the distribution estimation techniques (EM, k-means) for each leave-one-out loop we performed 30 repetitions, in order to obtain statistically significant averaged classification measure. For EM and k-means algorithms, when employed for distribution estimation, we used $k = 3$ clusters. Accuracy was evaluated on the average over all repetitions for each sample and over all the leave-one-out loops for the whole dataset.

All examined techniques, except the Mahalanobis distance, provided useful prediction (accuracy better than a random guess). The Mahalanobis distance technique could not work presumably due to similarity of distributional means (recall from Section 4.1 that in such cases this method does not work well).

Table 1 Prediction errors from experiments on fMRI data for maximum likelihood and Kullback–Leibler divergence-based techniques[a]

	Classification error (%)		
	Controls	Patients	Average
Maximum likelihood			
EM	22.96	32.96	27.96
k-means	22.96	17.04	20.00
Kullback–Leibler			
EM	20.74	42.96	31.85
k-means	23.33	34.07	28.70

[a] For each technique, the errors on controls, patients and the average error are given in percents.

In Table 1 the estimated classification error for each of the useful algorithms is shown for controls, patients and averaged. Both ML and KL divergence methods resulted in similar specificity (classification errors when predicting controls in range 20–23.5%). However, ML methods provided better selectivity (smaller classification errors of patients). The ML method based on k-means resulted with the smallest error when classifying patients (17.04%) and the best overall performance (the average classification error (controls and patients) of 20.00%). The average classification error of other useful methods varied within 27–32% range.

5. Discussion

Results presented in this study clearly demonstrate the ability of the proposed methodology to provide an efficient framework for the classification of 3D medical images based on spatial distribution analysis of ROIs. Experiments also showed that the accuracy of this classification scheme crucially depends on the quality and availability of data and the intrinsic complexity of underlying data generation processes.

The classification methods examined in the study varied from the Mahalanobis approach, based on statistical distance and implicitly assuming the single-component Gaussian distribution of data, to more complex semi-parametric techniques based on probabilistic divergence and maximization of the likelihood that a new subject corresponds to one of the two distributions learned from the training datasets. While the Mahalanobis distance was capable of providing useful classification results when the data distributions corresponding to the

observed two classes (e.g., presence and absence of ADHD) differed significantly, this technique is generally inferior in comparison to the other proposed methods. However, the Mahalanobis approach could still be a method of choice when the requirements for the classification accuracy are not extremely strict while the major emphasis is put on efficiency and simplicity of the technique.

Methods based on the KL divergence and the ML methods provided superior accuracy in comparison to the Mahalanobis distance. This came with no surprise since the former methods assume more complex mixture models in contrast to one multivariate Gaussian, as in case of the applied Mahalanobis technique. Generally, the ML technique performed better than the KL divergence (the improvement in the case of mixtures with considerably different distributional means was not significant since both techniques were capable of providing classification with the errors less than 2%). There are two plausible reasons for this. First, the ML technique has more sound theoretical foundation than KL. Second, while using ML, we would need to estimate distribution parameters only on the training sets. In contrast, in case of KL, a distribution should be estimated for datasets corresponding to spatial arrangements of ROIs of each new patient. The imprecision in these additional distribution estimations could particularly play its role when the number of ROI voxels corresponding to a new subject is small, and this is exactly what happened in our experiments.

Both ML and KL methods can be combined with different techniques to estimate distributions. In this study, we compared the EM algorithm with its faster and less complex variant—the *k*-means algorithm. When the distributions differ significantly (as in the case of Gaussian distributions with different means, Section 4.1.1), both algorithms provide similar results and hence the application of the faster *k*-means can be proposed. Interestingly, in the case of examined realistic and clinical data, the *k*-means provided even better accuracy than EM, probably due to smaller complexity of the *k*-means compared to EM. In contrast, when distinguishing distributions with the same means but different variances of the components, EM clearly outperformed *k*-means in classifying the subjects coming from the distribution with the larger variances. Such results can be easily explained. Namely, unlike EM, *k*-means iteratively estimate only means of the distributional components and consider each voxel of the dataset to belong exactly to one of the components. Such assumption, although working well when the distributional components are well separated, starts to cause problems when the distributions overlap, and this is precisely what occurs when the component variances are larger.

Experimental results on realistic data suggest that excellent classification accuracy with errors smaller than 1% could be achieved assuming the proper sizes of the training set and the large enough amount of information regarding the new subject, i.e., a well-identified spatial distribution of their ROIs. However, although these results are very encouraging, we would like to accentuate that real-life clinical results may be less optimistic. For instance, the minimal average classification error on clinical Alzheimer's disease we obtained (using the ML technique with *k*-means) was 20% while other useful methods discussed here (ML approach with EM distribution estimation or KL divergence) resulted in errors within 27—32% range (see Table 1). Reasons for such difference of results on realistic and on clinical data are the greater heterogeneity of the real datasets and the small number of subjects available that lead to a difficulty in generalizing the patterns observed. In reality, image imperfections, noise, registration, and segmentation errors and other potential sources of data corruption (that are related to the imaging technology and methodology) as well as inter-subject variability (i.e., subjects belonging to different populations, age and ethnicity groups, etc.), may lead to significant heterogeneity, which could cause problems to any learning technique, including those covered in this study.

The size of the database of prelabeled cases plays a significant role in the classification of a new subject. Our study has shown that, for all the considered techniques applied on synthetic and realistic data, the classification accuracy increased with the size of the training set—the database used to learn parameters of a classification algorithm. This is intuitively expected, while our experiments quantified it. In clinical data, the total number of available subjects was rather small (nine patients and nine controls) but typical for an fMRI study and could not be increased, which in part contributed to lower accuracy, as compared to realistic data. The results were indeed impressive given the small dataset, its heterogeneity and difficulty in generalizing the patterns observed.

For all *useful* techniques on synthetic and realistic data (where the number of ROI voxels per subject—the size of a new dataset—could be varied in a controlled way), the performance significantly improved with the size of a new dataset that corresponds to a subject to be classified. The computational time for all the proposed techniques can be split into the ''learning'' time, necessary for distribution analysis and the ''query'' time, needed for

classification of a new dataset. The learning time for all proposed methods is a linear function of the size of distributions, thus making them highly suitable for large datasets. In general, the Mahalanobis distance approach provides the fastest distribution learning, since it requires only estimation of mean and covariance matrix per distribution and does not perform computations for distribution estimation or likelihood calculation. In contrast, the KL divergence-based and ML methods require semi-parametric estimation of distributions, as well as costly computation of exponential functions, in the case when the EM algorithm is employed for density estimation. Although KL divergence-based method is generally more complex, for some special cases it may still perform faster than the ML based technique (e.g., KL with *k*-means distribution estimation versus ML approach with EM distribution estimation). Since the distribution analysis is typically performed in a batch mode, query time will have major influence on the usability of the proposed methods. The query time for all proposed techniques is linearly proportional to the size of a new dataset. In our experiments, for the examined methods and sizes of the new dataset, the query time was satisfactory small (less than 10 s on a Pentium 4, 1.8 GHz computer with 256 MB memory).

6. Conclusions and future work

In this paper, we propose a framework for the classification of 3D medical images based on the analysis of the spatial distribution of ROIs. In addition, the study provides a benchmark comparison of distribution similarity-based approaches and maximum likelihood techniques. As illustrated from the experimental results in this work, the proposed methodology can assist medical image based diagnosis.

Although the individual methods proposed here have been developed for other domains, to the best of our knowledge they have not been applied to the analysis and classification of medical images based on their ROIs. Another contribution of this work is the proposed integration of various statistical techniques into the process for 3D spatial distribution analysis and classification. Finally, since the proposed methodology is voxel based and applied directly to the 3D domain, it automatically takes into account the spatial locality of the voxels in 3D during the analysis. There is no need to analyze each 2D slice separately and combine results, as in case of a 2D slice-based (pixel-based) technique. For this reason, the proposed techniques can be more accurate and faster than approaches originally designed for 2D images that are later extended to 3D images.

The methods discussed in this study analyze 3D volumes as binary objects, i.e., only information about a particular voxel being or not being part of a certain region of interest is provided. This assumption without being very restrictive simplifies the analysis and is often made. In reality, due to the uncertainty in delineating the boundaries of certain ROIs (structures or abnormalities) in medical images during segmentation, the regions usually have fuzzy boundaries. Our results on Alzheimer's disease fMRI data [17,48] employing non-binary images to represent ROIs indicate that novel advanced classification techniques might be necessary to further improve classification with such ROI representations.

The realistic data used in this study originated from MRI lesion-deficit studies of ADHD. In addition, we presented classification results on Alzheimer's disease fMRI activation data. These experiments illustrate the potential of our proposed methodology to be actually applied effectively on real-world medical imaging applications. The achieved results, demonstrated through low classification errors, are encouraging but at the same time emphasize the needs for further research in various directions. Hence, further experiments are necessary to evaluate the proposed techniques in more comprehensive experimental studies on real-life medical images related to different disorders and different organs. Also, the part of the work in progress is the application of the proposed methodology on other classes of medical images, including positron emission tomography, computed tomography, Single Photon Emission Computerized Tomography, etc. Another important direction for further study is the development and evaluation of advanced non-parametric classification techniques, as applied on medical imagery. Finally, work in progress includes the application of the proposed methods on non-binary images and in more comprehensive experimental studies on real-life medical imaging applications.

Acknowledgements

The authors would like to thank A. Saykin for providing the Alzheimer's disease dataset and clinical expertise. In addition, the authors express their gratitude to the anonymous reviewers whose comments significantly improved the structure and overall quality of the paper. This work was supported in part by the National Science Foundation under

grants IIS0083423, IIS-0237921, HRD-0310163, and HRD-0320991, the National Institutes of Health under grant R01 MH68066-01A1 funded by the National Institute of Mental Health, the National Institute of Neurological Disorders and Stroke, and the National Institute on Aging, the NIH-funded Delaware BRIN Grant (P20 RR16472) and DoD HBCU/MI Infrastructure Support Program (45395-MA-ISP Department of Army).

Appendix A. Mathematical and statistical preliminaries

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d]$ be a multivariate random variable, where $\boldsymbol{x}_i$, $i = 1, \ldots, d$ are continuous variables that can choose any value from a d-dimensional domain D. The probability $P(\mathbf{x})$ that the variable $\mathbf{x}$ belongs to a subdomain $V \subset D$ is defined as:

$$P[\boldsymbol{x} \in V] = \int_V p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{A.1}$$

where $p(\mathbf{x})$ is a probability density function satisfying non-negativity ($p(\mathbf{x}) \geq 0$) and normalization ($\int_D p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$) constraints. The probability density function $p(\mathbf{x})$ uniquely determines a distribution of the variable $\mathbf{x}$. In addition, each distribution is characterized by its histogram while parametric distributions can also be specified by values of their parameters [49].

Consider a uniform discretization of a multidimensional domain D into equally sized smaller subdomains. In a general d-dimensional case, we call these subdomains *hyper-voxels*. Thus, each hyper-voxel $v_{i_1,i_2,\ldots,i_d}$ represents a d-dimensional hyper-rectangle $[i_1\Delta x_1, (i_1+1)\Delta x_1] \times [i_2\Delta x_2, (i_2+1)\Delta x_2] \times \ldots \times [i_d\Delta x_d, (i_d+1)\Delta x_d]$ where $i_1, \ldots, i_d$ are integers and $\Delta x_1, \ldots, \Delta x_d$ are discretization intervals in each dimension. B_j, $j = 1, \ldots, d$ are the numbers of hyper-voxels in each dimension, such that $i_j = 0, \ldots, B_j - 1$. The total number of hyper-voxels in the volume D is equal to $B = B_1 B_2 \cdots B_d$. For $d = 3$, the hyper-voxels reduce to *voxels*, the three-dimensional volume elements. Using Eq. (A.1) and the mean-value theorem [50], it can be shown that there exists m such that $\min_{\boldsymbol{x} \in v_{i_1,i_2,\ldots,i_d}} p(\boldsymbol{x}) \leq m \leq \min_{\boldsymbol{x} \in v_{i_1,i_2,\ldots,i_d}} p(\boldsymbol{x})$, so that the probability that a multivariate variable $\mathbf{x}$ belongs to a hyper-voxel $v_{i_1,i_2,\ldots,i_d}$ can be expressed as:

$$P[\boldsymbol{x} \in v_{i_1,i_2,\ldots,i_d}] = m\Delta\boldsymbol{x} \tag{A.2}$$

where

$$\Delta\boldsymbol{x} = \Delta x_1 \Delta x_2 \cdots \Delta x_d, \tag{A.3}$$

is the product of discretization intervals. Assuming small discretization intervals, the probability density has an approximately constant value $p(\mathbf{x})$ for $\boldsymbol{x} \in v_{i_1,i_2,\ldots,i_d}$, such that Eq. (A.2) can be written as:

$$P[\boldsymbol{x} \in v_{i_1,i_2,\ldots,i_d}] \approx p(\boldsymbol{x}_{i_1,i_2,\ldots,i_d})\Delta\boldsymbol{x} \tag{A.4}$$

where

$$\boldsymbol{x}_{i_1,i_2,\ldots,i_d} = \left[i_1\Delta x_1 + \frac{\Delta x_1}{2}, i_2\Delta x_2 + \frac{\Delta x_2}{2}, \ldots, i_d\Delta x_d + \frac{\Delta x_d}{2}\right]^{\mathsf{T}} \tag{A.5}$$

is the center of the hyper-voxel $v_{i_1,i_2,\ldots,i_d}$.

The Gaussian distribution for k-dimensional random vector $\boldsymbol{z}$ is defined as:

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{k/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right) \tag{A.6}$$

Multivariate data with Gaussian distribution tend to cluster about the mean vector $\boldsymbol{\mu}$, falling in an ellipsoidal cloud whose principal axes are eigenvectors of the covariance matrix Σ [39].

The Gaussian mixture has a probability density function [42]:

$$p(\boldsymbol{z}) = \sum_{j=1}^{k} \pi_j f(\boldsymbol{z}|\boldsymbol{\mu}_j, \Sigma_j) \tag{A.7}$$

where π_j is a prior probability and $f(\mathbf{z}|\boldsymbol{\mu}_j, \Sigma_j)$ is the probability density of the j-th Gaussian component specified by the mean $\boldsymbol{\mu}_j$ and the covariance matrix Σ_j.

The Euclidean distance between two multivariate variables (vectors) depends on the sum of squared differences of their components. Therefore, given two vectors $\mathbf{x}$ and $\mathbf{y}$, the Euclidean distance between them is computed as

$$d_{\mathrm{E}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y})}. \tag{A.8}$$

Observe that the Euclidean distance *does not* incorporate any information about the distribution of vectors $\mathbf{x}$ and $\mathbf{y}$. Under the assumption of data having Gaussian distribution, the distributional information can be incorporated into distance calculation by including a covariance matrix Σ into the distance formula. This results in the Mahalanobis distance [34], which is, for the two vectors $\mathbf{x}$ and $\mathbf{y}$, defined as

$$d_{\mathrm{M}} = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y})}. \tag{A.9}$$

The Mahalanobis distance is equal to the Euclidean distance only when the covariance matrix Σ is an identity matrix.

References

[1] Marsicoi MD, Cinque L, Levialdi S. Indexing pictorial documents by their content: a survey of current techniques. Image Vision Comput 1997;15:119—41.

[2] Smeulders AWM, Worring M, Santint S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Machine Intell 2000;22:1349—80.

[3] Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, et al. Query by image and video content: the QBIC system. IEEE Comput 1995;28:23—32.

[4] Pentland A, Picard RW, Sclaroff S. Photobook: tools for content-based manipulation of image databases. In: Proceedings of the SPIE Conference on Storage and Retrieval of Image and Video Databases II. 1994.

[5] Samadani R, Han C, Katragadda LK. Content-based event selection from satellite image of the aurora. In: Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases; 1993. p. 50—9.

[6] Dy JG, Brodley CE, Kak AC, Shyu C, Broderick LS. The customized-queries approach to CBIR. In: Proceedings of the IS & T/SPIE Electronic Imaging Conference on Storage and Retrieval for Image and Video Databases VII. 1999.

[7] Korn F, Sidiropoulos N, Faloutsos C, Siegel E, Protopapas Z. Fast and effective similarity search in medical tumor databases using morphology. In: Proceedings of the SPIE Conference. 1996.

[8] Lehmann T, Wein B, Dahmen J, Bredno J, Vogelsang F, Kohnen M. Content-based image retrieval in medical applications: a novel multi-step approach. Proc SPIE 2000;3972:312—20.

[9] Saykin AJ, Flashman LA, Frutiger SA, Johnson SC, Mamourian AC, Moritz CH, et al. Neuroanatomic substrates of semantic memory impairment in Alzheimer's disease: patterns of functional MRI activation. J Int Neuropsychol Soc 1999;5:377—92.

[10] Koslow SH, Huerta MF. Neuroinformatics: an overview of the human brain project. Mahway, NJ: Erlbaum; 1997.

[11] Pal N, Pal S. A review on image segmentation techniques. Pattern Recognit 1993;26:1277—94.

[12] Worth A, Makris N, Caviness V, Kennedy D. Neuroanatomical segmentation in MRI: technological objectives. Int J Pattern Recognit Artif Intell 1997;11:1161—87.

[13] Zhang Y. A survey on evaluation methods for image segmentation. Pattern Recognit 1996;29:1335—46.

[14] Ropero PJ. Towards a neural network based therapy for hallucinatory disorders. Neural Networks 2000;13:1047—61.

[15] Vigario R, Oja E. Independence: a new criterion for the analysis of the electromagnetic fields in the global brain? Neural Networks 2000;13:891—907.

[16] Ford J, Farid H, Makedon F, Flashman LA, McAllister TW, Megalooikonomou V, et al. Patient classification of fMRI activation maps. In: Proceedings of the 6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'03), 2003.

[17] Kontos D, Megalooikonomou V, Pokrajac D, Lazarevic A, Obradovic Z, Ford J, et al. Extraction of discriminant functional MRI activation patterns and an application to Alzheimer's disease. In: Proceedings of the 7th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'04); in press.

[18] Zhao Q, Principe J, Bradley M, Lang P. fMRI analysis: distribution divergence measure based on quadratic entropy. In: Proceedings of the Human Brain Mapping. 2000.

[19] Sahiner B, Chan HP, Wei D, Petrick N, Helvie MA, Adler DD, et al. Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue. Med Phys 1996;23:1671—84.

[20] Rangayyan RM, El-Faramawy NM, Desautels JE, Alim OA. Measures of acutance and shape for classification of breast tumors. IEEE Trans Med Imaging 1997;16: 799—810.

[21] Brodley CE, Kak AC, Dy JG, Shyu CR, Aisen A, Broderick L. Content-based retrieval from medical image databases: a synergy of human interaction, machine learning and computer vision. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence; 1999. p. 760—7.

[22] Tagare H, Jaffe C, Duncan J. Medical image databases: a content-based retrieval approach. J Am Med Inform Assoc 1997;4:184—98.

[23] Megalooikonomou V, Dutta H, Kontos D. Fast and effective characterization of 3D region data. In: Proceedings of the IEEE International Conference on Image Processing (ICIP); 2002. p. 421—4.

[24] Dykstra CJ. 3D contiguous volume analysis for functional imaging. Ph.D. Thesis. Canada: Simon Fraser University; 1994.

[25] Kippenham JS, Barker WW, Pascal S, Nagel J, Duara R. Evaluation of a neural network classifier for PET scans of normal and Alzheimer's disease subjects. J Nucl Med 1992;33:1459—67.

[26] Halkjaer S, Waldemar G, Lautrup B, Paulson OB. Correlation between cognitive function scores and the response of a neural network classifier for SPECT data in patients with Alzheimer's disease. Copenhagen, DK: Niels Bohr Institute; 1997.

[27] Friston KJ, Holmes AP, Price CJ, Buchel C, Worsley KJ. Multisubject fMRI studies and conjunction analyses. NeuroImage 1999;10:385—96.

[28] Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 1995;2:189—210.

[29] Missimer J, Knorr U, Maguire RP, Herzog H, Seitz RJ, Tellman L, et al. On two methods of statistical image analysis. Hum Brain Mapp 1999;8:245—58.

[30] Turner R, Friston K. SPM course 1997 notes. London: Wellcome Department of Cognitive Neurology; 1997.

[31] Megalooikonomou V, Davatzikos C, Herskovits EH. Mining lesion-deficit associations in a brain image database. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999. p. 347—51.

[32] Talairach J, Tournoux P. Co-planar stereotaxic atlas of the human brain. Stuttgart: Thieme; 1988.

[33] Megalooikonomou V, Davatzikos C, Herskovits EH. A simulator for evaluating methods for the detection of lesion-deficit associations. Hum Brain Mapp 2000;10:61—73.

[34] Mahalanobis PC. On tests and measures of groups divergence I. J Asiatic Soc Benagal 1930;26:541.

[35] Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1995.

[36] Lloyd SP. Least squares quantization in PCM. IEEE Trans Inform Theory 1982;28:129—37.

[37] McLachan GJ, Krishnan T. The EM algorithm and extensions. John Wiley and Sons; 1996.

[38] Vlassis N, Likas A. A greedy EM algorithm for Gaussian mixture learning. Neural Process Lett 2002; 15:77—87.

[39] Duda R, Hart P, Stork D. Pattern classification. 2nd ed. John Wiley and Sons; 2000.

[40] Zhao Y, Karypis G. Criterion functions for document clustering experiments and analysis. Army High Performance Com-

puting Research Center (AHPCRC) Technical Report #01-40; 2002.
[41] Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. San Diego, CA: Academic Press; 1990.
[42] Flury B. A first course in multivariate statistics. New York: Springer-Verlag; 1997.
[43] Kullback S. Information theory and statistics. Gloucester, MA: Sinauer Associates; 1968.
[44] Gerring J, Brady K, Chen A, Quinn C, Bandeen-Roche K, Denckla M, et al. Premorbid prevalence of attention-deficit hyperactivity disorder and development of secondary attention-deficit hyperactivity disorder after closed-head injury. J Am Acad Child Adolesc Psychiatry 1998;37: 647—54.
[45] Saykin AJ, Gur RC, Sussman NM, O'Connor MJ, Gur RE. Memory deficits before and after temporal lobectomy: effect of laterality and age of onset. Brain Cogn 1989; 9:191—200.
[46] Davatzikos C. Spatial transformation and registration of brain images using elastically deformable models. Comput Vision Image Understanding 1997;66:207—22.
[47] Gerring J, Brady K, Chen A, Quinn C, Bandeen-Roche K, Denckla M, et al. Neuroimaging variables related to the development of secondary attention deficit hyperactivity disorder after closed head injury in children and adolescents. Brain Injury 2000;14:205—18.
[48] Megalooikonomou V, Kontos D, Pokrajac D, Lazarevic A, Obradovic Z, Boyko OB, et al. Classification and mining of brain image data using adaptive recursive partitioning methods: application to Alzheimer disease and brain activation patterns. In: Proceedings of the 9th Annual Meeting of the Organization for Human Brain Mapping (OHBM). 2003.
[49] Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press; 1981.
[50] Apostol T. Mathematical analysis: a modern approach to advanced calculus. 2nd ed. Addison-Wesley; 1974.