J Intell Inf Syst (2008) 31:217–242
DOI 10.1007/s10844-007-0043-2

An adaptive partitioning approach for mining discriminant regions in 3D image data

Vasileios Megalooikonomou · Despina Kontos · Dragoljub Pokrajac · Aleksandar Lazarevic · Zoran Obradovic

Received: 18 July 2005 / Revised: 29 November 2006 /
Accepted: 19 June 2007 / Published online: 10 August 2007

Abstract Mining discriminative spatial patterns in image data is an emerging subject of interest in medical imaging, meteorology, engineering, biology, and other fields. In this paper, we propose a novel approach for detecting spatial regions that are highly discriminative among different classes of three dimensional (3D) image data. The main idea of our approach is to treat the initial 3D image as a hyper-rectangle and search for discriminative regions by adaptively partitioning the space into progressively smaller hyper-rectangles (sub-regions). We use statistical information about each hyper-rectangle to guide

V. Megalooikonomou · D. Kontos
Data Engineering Laborarory (DEnLab), Temple University,
319 Wachman Hall, 1805 N. Broad St., Philadelphia, PA 19122-6094, USA

V. Megalooikonomou · D. Kontos · Z. Obradovic
Center for Information Science and Technology, Temple University,
303 Wachman Hall (038-24), 1805 N. Broad St., Philadelphia, PA 19122-6094, USA

D. Pokrajac
Computer and Information Science Department, Delaware State University,
1200N Dupont Hwy, Science Center North 305D, Dover, DE 19901, USA

D. Pokrajac
Applied Mathematics Research Center (AMRC), Delaware State University,
1200N Dupont Hwy, ETV Building, Dover, DE 19901, USA

A. Lazarevic
United Technologies Research Center,
411 Silver Lane, MS 129-15, East Hartford, CT 06108, USA

V. Megalooikonomou (✉) · D. Kontos · Z. Obradovic
Department of Computer and Information Sciences, Temple University,
314 Wachman Hall, 1805 N. Broad St., Philadelphia, PA 19122-6094, USA
e-mail: vasilis@temple.edu

the selectivity of the partitioning. A hyper-rectangle is partitioned only if its attribute cannot adequately discriminate among the distinct labeled classes, and it is sufficiently large for further splitting. To evaluate the discriminative power of the attributes corresponding to the detected regions, we performed classification experiments on artificial and real datasets. Our results show that the proposed method outperforms major competitors, achieving 30% and 15% better classification accuracy on synthetic and real data respectively while reducing by two orders of magnitude the number of statistical tests required by voxel-based approaches.

Keywords Data mining · Image databases · 3D images · Adaptive partitioning · Statistical pattern analysis · Classification · Intelligent knowledge discovery

1 Introduction

Mining multidimensional data has been one of the core aspects of knowledge discovery in several fields and applications. In the traditional data mining literature (Han et al. 1996, Agrawal et al. 1992), multidimensional data have been most often encountered in the framework of relational databases and data warehouses. Within these frameworks, they have been defined as vector/tuple instances constructed by features (attributes) and most of the mining techniques have focused on extracting strong association rules and dependencies among these attributes (Agrawal et al. 1993, Agrawal and Srikant 1994). However, these attribute-oriented induction approaches can only handle this form of explicitly (tuple-formatted) stored data, thereby limiting the ability to discover patterns in datasets that may not involve this type of attributes. Recent advancements in data acquisition technology have made it possible to collect several different types of multidimensional data that require more elaborate analysis techniques (Faloutsos 1996).

A special case of multidimensional data that are encountered in many applications are spatial data (Guting 1994, Gaede and Gunter 1998). In most of the reported research, spatial data have been analyzed in the context of Geographical Information Systems (GIS) and have been treated as representations of certain geographic entities, such as points, lines, and polygons (Fotheringham and Rogerson 1994). Mining these spatial entities usually involves extracting interesting knowledge that is implicitly present in the data, such as spatial relations or patterns engaging spatial and non-spatial attributes (Ester et al. 1977, Koperski and Han 1995). Recently, modern automated data collection tools-such as high-resolution medical image scanners, remote sensing technology, and satellite imagery devices-have allowed vast amounts of spatial data to be captured in a wide variety of image formats. Examples of such 3D volume data repositories are geographic information systems (GIS) storing 3D surface data obtained from geophysical studies (Kriegel and Seidl 1998); brain image databases including volumes obtained from MRI[1] fMRI[2] or CT[3] scanners (Megalooikonomou et al. 2000b); and 3D protein structure databases (Bernstein et al. 1977). Challenges for these types of data include developing efficient frameworks for organizing, storing, retrieving and analyzing such datasets.

[1] Magnetic Resonance Imaging: shows soft-tissue structural information.

[2] Functional-Magnetic Resonance Imaging: shows physiological activity in the brain.

[3] Computed Tomography: shows hard-tissue structural information.

Considerable effort has been devoted to developing tools for visual exploration of 3D image data because they reflect real-world 3D spatial distributions (Kaufman et al. 1993), but little work has been performed on developing efficient and effective methods for modeling and mining these 3D datasets. Traditional knowledge discovery tasks, such as classification, extraction of discriminative patterns, and detection of associations, need to be extended to support efficient decision making at emerging 3D volume data collections. Although many techniques have been developed in the case of content-based retrieval and classification for general types of images (Smeulders et al. 2000, Flickner et al. 1995), the analysis is based on extracted features that treat the image in a holistic manner: i.e., they are based on extracted attributes that refer to the entire image space. These global attributes may be less useful than attributes that focus on specific sub-regions that are of interest in certain application domains, such as medicine, geography, and meteorology. The specific sub-regions that are of interest may occupy a small portion of the image data, such as temperature and precipitation contour maps in meteorological images (Kontos and Megalooikonomou 2005) or lesions and tumor structures visualized by medical images (Megalooikonomou et al. 2000b, Euripides et al. 1997).

For this type of analysis there is a need to detect highly informative image sub-regions of interest, extract local informative features and then perform specialized sub-region focused pattern analysis.

In this work we are interested in developing intelligent tools for mining discriminative patterns among groups of 3D image data. The data we consider are 3D volumes (or spatial data) that represent the spatial distribution of certain quantities in the 3D space. Examples of such image data are spatial distributions of brain lesion centroids (Fig. 1a), regions with high levels of precipitation or temperature-relative humidity in meteorological maps (Fig. 1b), brain activation areas during a particular task in fMRI contrast maps (see Fig. 1c), and 3D molecular structures studied in biological sciences (Fig. 1d). We propose a novel Dynamic Recursive Partitioning (DRP) approach for distinguishing between 3D volume data generated by highly non-uniform distributions, which is based on detecting discriminative sub-regions and extracting informative quantitative features. This approach facilitates the mining of associations between the spatial distribution of 3D regions of interest and other non-spatial characteristics, such as class membership. We propose to perform an adaptive partitioning of the 3D space into progressively smaller sub-regions (hyper-rectangles) until discriminative areas are discovered in the initial volume. The selectivity of the splitting is guided by statistical tests that determine the discriminative power for each sub-region. In order to evaluate the discriminative power of the quantitative

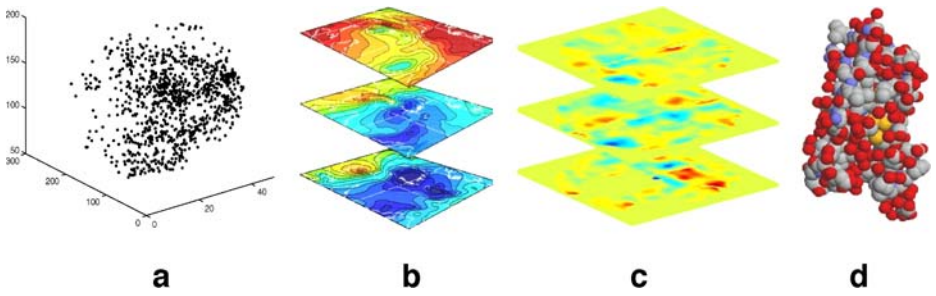

Fig. 1 Examples of 3D image data: **a** 3D binary data illustrating the distribution of brain lesion centroids, **b** 3D geophysical spatial data, **c** a 3D fMRI brain contrast activation map, and **d** a 3D molecular structure

features extracted from the detected areas, we classify the initial 3D volumes using these features. In our experiments we use artificial and real datasets, including mixtures of Gaussian distributions, synthetic fractal data, real 3D fMRI data, and realistic lesion data generated by a simulator conforming to a clinical study. To classify the data, we employ neural networks.

The rest of this paper is organized as follows. In section 2 we present background information on spatial data analysis. In section 3 we provide a comprehensive description of the proposed methodology. In section 4 we present a description of the datasets used for the experimental evaluation. In section 5 we describe the experiments performed and present the corresponding results, including comparative experiments with the alternative techniques described in the Background section. Finally, in section 6 we summarize our concluding remarks and give directions for future research.

2 Background

Spatial data mining refers to the process of extracting significant implicit knowledge from large amounts of spatial data. To facilitate this process, traditional data mining techniques (Agrawal et al. 1993, Agrawal et al. 1996) have been extended and modified in order to be applied to spatial data entities (Ester et al. 1996, Koperski and Han 1995, Lazarevic and Obradovic 2002, Ng and Han 1994, Sheikholeslami et al. 1998, Son et al. 1998). In general, spatial data mining techniques extract patterns of three different types (Koperski and Han 1995): *spatial characteristics* (general characteristics and patterns of a spatial-entities set, such as precipitation patterns in meteorological maps), *spatial associations* (implications and associations among spatial features), and *spatial discriminant patterns* (contrasting discriminative characteristics of distinct spatial entity classes). Classification in the context of spatial data is defined as the process of assigning non-spatially related labels to classes of spatial entities. For this purpose, discovering spatial patterns that are discriminative among classes is very useful. Discovering such patterns is the main focus of this paper.

Researchers face the following challenges when mining spatial data: (a) a very large number of dimensions (in considered domains proportional to the number of pixels or voxels (volume elements) that are observed); (b) difficulties extending existing mining techniques for non-spatial data or designing novel algorithms for learning from spatial objects directly; (c) errors introduced by preprocessing that involves spatial normalization (registration) and intensity normalization (to deal with different acquisition mechanisms); (d) a high correlation among data; and (e) large heterogeneity. In particular, the statistical dependence that exists among neighboring spatial objects needs to be taken into account. In addition, the granularity of the space considered (i.e., coarse or fine) determines the number of spatial regions examined and affects the computational cost of an approach.

Most of the approaches for pattern analysis and classification of 3D image data proposed in the literature have treated these data in a holistic manner without focusing on certain discriminative sub-regions and their spatial distribution within the volume that might be of particular interest. In several cases, volume data have been treated as multivariate variables. The work of (Lazarevic et al. 2001, Pokrajac et al. 2005) assumes that each dimension is represented by a component attribute and the voxel values are generated by probability distributions. The probability density function is estimated using parametric, non-parametric, or semi-parametric techniques (Devore 2000). Histograms and kernel-based methods (Bishop 1995) have been applied to estimate the probability density function from the available volume data. Under specific assumptions on the form of the distributions

(number of Gaussian components, etc.) the expectation-maximization (EM; Lloyd 1982, McLachan and Krishnan 1996) and the *k*-means (Lloyd 1982) algorithms have also been employed to estimate the underlying 3D distribution models. Based on the estimated probability density functions and the corresponding representation of volume data as multivariate variables generated from multidimensional distributions, (Lazarevic et al. 2001, Pokrajac et al. 2005) have used distance based methods and maximum likelihood approaches to distinguish between different classes of 3D image data.

Distance based methods define a distance measure to estimate to what extent distributions diverge. Distance metrics frequently used in this context[4] are the Euclidean distance, the Mahalanobis distance (Fukunaga 1990) for normally distributed data and the Kullback–Leibler (KL) divergence (Duda et al. 2000; see the Appendix for definitions of those distances). When employing a maximum likelihood (ML) approach (Duda et al. 2000, Aladjem 1998, Bhattacharyya 1943, Mitchell 1997), the similarity (or equivalently dissimilarity) between volumes can be assessed by the likelihood that a new distribution is the same as one of the existing distributions, assuming that the probability densities of existing distributions have been estimated. The probability that the particular data are observed is calculated on the condition that a pre-determined hypothesis holds for the class distribution. ML selects the hypothesis that maximizes this conditional probability, defining similarity between a 3D sample and labeled volume data classes. Distance based and maximum-likelihood techniques have been mainly applied to binary (homogeneous) volumes (Lazarevic et al. 2001, Pokrajac et al. 2001), where voxel values are either 0 or 1. In many recent applications though, such as volume data analysis (Megalooikonomou et al. 2000b, Kontos et al. 2004), there is a need for mining non-homogeneous (non-binary) 3D image datasets. In addition, when using distributional distance or ML techniques to facilitate volume data classification, we cannot obtain any information about the spatial sub-domains where the distributions are different.

In contrast to approaches that measure distances between spatial distributions in 3D image data by looking at the space as a whole, voxel-wise statistical analysis has been enthusiastically applied to distinguish among 3D spatial distributions (in particular in the neuroimaging community). In this type of analysis, each voxel's changes are analyzed independently among the available samples and a map of statistical significance is built for each voxel. To ascertain the discriminatory significance of each voxel, a statistical test such as the *t*-test, ranksum test, or the F-test is applied (details about these tests are provided in section 3.2). For example, when applying the *t*-test on a voxel-by-voxel basis a t-value is obtained for each voxel that indicates how likely it is that the voxel's variability across classes is observed by chance. Clustering is often employed in the process to construct highly informative regions with respect to classification. A detailed review of this and other analysis techniques when applied to neuroimage data is provided by (Megalooikonomou et al. 2000b). *Statistical Parametric Mapping* SPM (Friston et al. 1995) is a widely used implementation of the voxel-wise analysis, which is mostly applied to the analysis of brain image data such as fMRI data. In contrast to distributional distance or ML techniques

[4] Other applicable distance measures that incorporate distributional information are the Bayesian distance, the Patrick–Fisher distance (Aladjem, M. (1998) Nonparametric discriminant analysis via recursive optimization of Patrick–Fisher distance. IEEE Transactions on Syst., Man, Cybern., 28B, 292–299 and the Bhattacharyya distance (Bhattacharyya (1943) On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society, 35, 99–110).

applied on the whole image space, voxel-wise analysis can easily identify spatial sub-regions that are discriminative among classes. However, voxel-wise approaches appear to be significantly biased towards distribution differences that are highly localized in space and linear in nature (Davatzikos 2004). Also, when applying voxel-wise analysis, a significant computational overhead is added because computations that consider each voxel individually are involved. Moreover, the computation of a statistic for many pairwise tests introduces the *multiple comparison problem*: a certain portion of the tests becomes positive just by chance. Approaches to overcome this problem include the standard Bonferroni correction (Andersen 1997), which overestimates the number of independent tests performed in this case due to the spatial correlation among neighboring voxels, and heuristic modifications such as the sequential Bonferroni correction (Andersen 1997). Clustering is also usually applied to detect and discard outlier voxels when constructing discriminative regions.

In order to distinguish among volume data distributions while reducing the number of statistical tests performed by voxel-wise analysis, a static partitioning approach has been proposed (Lazarevic et al. 2001). This approach partitions the volume into a predefined number of 3D hyper-rectangles; then for each hyper-rectangle, certain statistical information is extracted (such as mean, median, or sum). This statistical information is used for training a classification model to assist in distinguishing between groups of volume data. One problem of this approach is to determine the splitting resolution, i.e., the optimal size of hyper-rectangles.

In this paper we propose an approach that effectively classifies 3D spatial distributions by detecting highly discriminative regions among different classes and extracting informative quantitative features that can be employed for classification. Rather than treating the spatial domain in a holistic manner by extracting features that reflect the information content of the entire image space, our approach focuses on specific regions that are of interest. In addition to performing classification, our technique can also indicate discriminative spatial sub-domains (where the distributions differ between the classes) for possible further examination by domain scientists. The approach has been designed to be applicable both to homogeneous (binary) and to non-homogeneous (non-binary) data. To moderate the extent of the *multiple comparison problem* observed in voxel-wise analysis, it reduces the number of statistical tests performed by applying statistical tests on groups of voxels rather than on individual voxels. To deal with the problem of determining the right splitting resolution of the static partitioning approach, it *dynamically* and *adaptively* partitions the space by using statistical tests to guide the splitting. For this reason we call the proposed approach Dynamic Recursive Partitioning (DRP). Preliminary results of the application of DRP to binary volume data classification have been reported in (Megalooikonomou et al. 2002). Here, we present a comprehensive description of the proposed methodology for both homogeneous (binary) and non-homogeneous (non-binary) 3D image data. In our experiments, we evaluate the proposed method both on artificial and real data and show that it outperforms major competitors by 30 and 15% respectively.

DRP treats the 3D space as a hyper-rectangle and searches for discriminative sub-regions (sub-domains) by partitioning the space into progressively smaller hyper-rectangles (cuboids) in an adaptive way. Although a few recently proposed methods (Wang et al. 1997, Keim 1999, Castelli and Kontoyiannis 2002, Castelli et al. 1996) are based on a similar idea, these methods differ significantly from ours. In (Wang et al. 1997), a recursive partitioning method has been proposed for clustering 2-dimensional data, while in (Keim 1999) oct-trees are used for hierarchical approximation of connected 3D shapes to provide efficient search and querying. Unlike these approaches, our approach uses recursive

partitioning for supervised learning (classification) and dynamically applies statistical tests to decide the depth of the constructed oct-tree. In Castelli and Kontoyiannis (2002), Castelli et al. (1996), a recursive partitioning algorithm is proposed which generates a tree-structured subdivision of the image data using an adaptive rule based on the wavelet transform. The authors deal with the problem of classification that consists of observing an image with known pixel values and unknown labels and assigning a class label to each pixel. The pixel classification is performed by considering the labels of neighboring pixels while looking at lower resolution representation of the data. DRP is different from progressive classification in several terms: (a) the main objective of DRP is to find discriminative areas among groups of images, (b) it focuses on image classification rather than on pixel classification, (c) it allows features to be extracted from the hyper-rectangles at different resolutions, rather than being computed for each of the hyper-rectangles as a function of the features of their partitions and in this respect it is superior to the progressive classification approach, (d) it uses statistical tests for partitioning rather than relying on classification performance, (e) it does not rely on images being transformed with a pyramidal decomposition based on wavelets. In the following section we describe the proposed methodology in detail.

3 Methodology

This section describes our approach for detecting highly discriminative patterns of the distribution of voxel values in 3D image data. In addition to the partitioning method, we discuss the statistical tests that are used, implementation details, and the classification model that is used for evaluating the proposed technique.

3.1 Dynamic recursive partitioning

The main idea of the approach is to perform an adaptive partitioning of the space into progressively smaller regions until areas are detected that reflect significant differences among the distinct classes. The selectivity of the splitting is guided by statistical tests. We use features extracted from the highly discriminative regions detected by the proposed algorithm to perform classification of the 3D image data. To ensure a clear description of the methodology, we consider a two-class problem here; an extension to more than two classes can be easily obtained.

The data we consider consist of three-dimensional volumes of voxels (volume elements). We focus on regions that can be defined as sets of (often connected) voxels with certain characteristics. We consider both binary and non-binary data. In the case of binary data, the voxels that form the regions of interest have the same value (e.g., “1” corresponding to “black”) which is different from the background (e.g., “0” corresponding to “white”; see Fig. 1a). For non-binary volumes, the voxel values can range while they are still different from the background (see Fig. 1b, c). Another way to view the variability in non-binary volumes is that voxel values reflect a probability generated by a 3D distribution. In some applications this is a highly non-uniform distribution. We will refer to this value as the *density* of a voxel.

The classification problem we address here can be defined as follows: Given two sets $\mathbf{S}_X$ and $\mathbf{S}_Y$ of 3D spatial distributions (each spatial distribution representing the spatial arrangement of voxel values in one instance of a spatial object as in Fig. 1a) that belong to one of two distinct classes, the task is to detect discriminative patterns, extract information

from these patterns, and then use this information to identify whether a new data sample s_z comes from the same distribution as the set $\mathbf{S}_X$ or the set $\mathbf{S}_Y$. In order to facilitate this classification, we identify spatial discriminative aspects of the class distributions by detecting highly informative (discriminative) spatial sub-regions. Hence, the overall goal of our work may be stated as developing a framework for mining discriminative patterns of spatial distributions in 3D image data.

The proposed method operates directly on the 3D image and does not necessarily require any prior processing of the voxel values. In specific applications (e.g. medical image analysis), preprocessing steps such as segmentation and registration to a standard template (i.e., spatial normalization) may be required. Image segmentation is used to delineate the boundaries of regions that are of interest, and it is performed using manual, automated, or semi-automated methods (Worth et al. 1997). Image registration is performed to bring the sample's image data into register (i.e., spatial coincidence) with a common spatial standard. This process is performed using normalization to a particular template and is necessary to determine whether two samples have specific sub-regions of interest in the same location. Throughout the description of our proposed methodology, we assume that the 3D image data has been normalized prior to the analysis.

The DRP algorithm treats the 3D space as a hyper-rectangle and searches for discriminative sub-regions (sub-domains) by partitioning the space into progressively smaller hyper-rectangles (cuboids) in an adaptive way. Although a few recently proposed methods (Wang et al. 1997, Keim 1999) are similar to our approach, these approaches differ significantly from ours. In (Wang et al. 1997), a recursive partitioning method has been proposed for clustering 2-dimensional data, while in (Keim 1999) oct-trees are used for hierarchical approximation of connected 3D shapes to provide efficient search and querying. Unlike these approaches, our approach uses recursive partitioning for supervised learning (classification) and dynamically applies statistical tests to decide the depth of the constructed oct-tree.

Our proposed algorithm begins by considering the total volume as one initial hyper-rectangle. Statistical information reflecting the distribution of voxel values in the hyper-rectangle (under consideration) is collected and used as a candidate feature (attribute) representing the corresponding sub-region. In our experiments in section 5, the mean (V_{mean}) of all voxel values of a hyper-rectangle is used; other measurements, however, such as the median or sum of the voxel values, can also be used as candidate features (attributes). The adaptive partitioning of the 3D space continues by splitting every hyper-rectangle whose attribute does not have the discriminative power to determine the class of samples. This is determined by applying a statistical test and assigning a significance cutoff threshold that determines the discriminative power of the specific sub-region. The procedure progresses recursively until all remaining sub-regions are discriminative or a sub-region becomes so small (given the resolution of the original data) that it cannot be further partitioned. Figure 2a illustrates the main idea of the process. The maximum number of partitioning steps (depth) is predefined to avoid excessive fragmentation of the volume. Given the resolution of the original data, the partitioning obviously cannot proceed beyond the voxel level. In specific applications, such as in medical imaging, going down to the voxel level is not suggested due to the errors caused by non-perfect registration techniques.

3.2 Use of statistical tests

As described above, the adaptive partitioning of the 3D space is guided by a statistical test and a significance cutoff threshold that determines the discriminative power of a given sub-

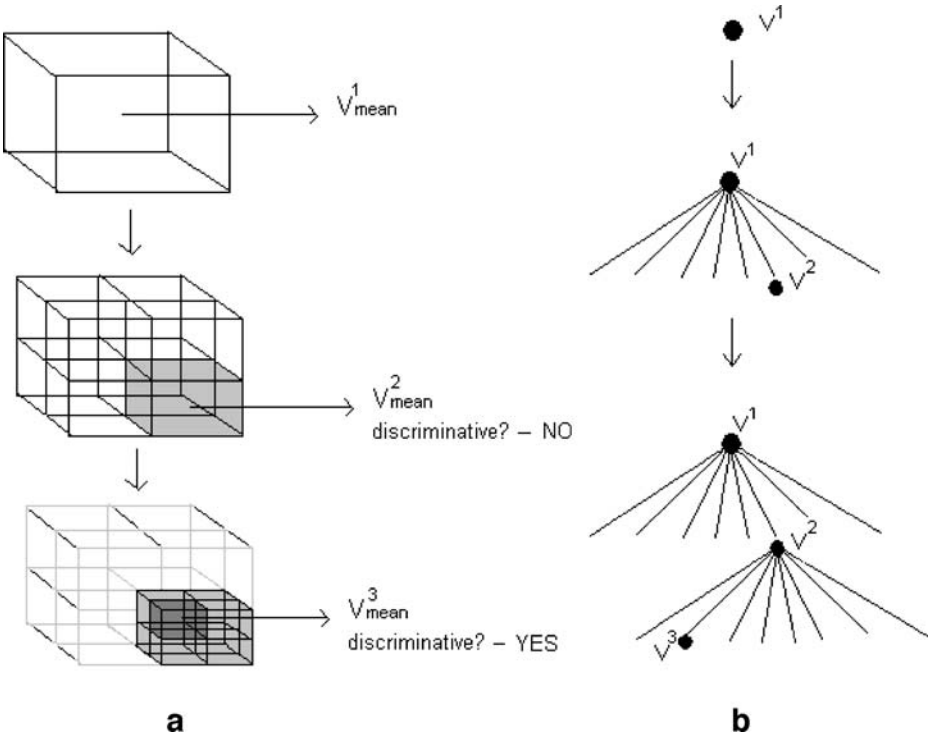

Fig. 2 Illustration of the adaptive partitioning of **a** an initial domain into spatial sub-domains and **b** splitting of an oct-tree node

region. Depending on the nature of the dataset, several statistical tests can be employed for this purpose. In our experiments, we have used the Pearson correlation coefficient, the *t*-test, and the ranksum test.

Based on the Pearson correlation coefficient (Devore 2000) between the class label (considered as a binary numeric value) and the attribute value for each sample (V_{mean}), we consider an attribute significant if the correlation coefficient is larger than a pre-determined threshold. In a sample of n observations the correlation coefficient r between two random variables X and Y can be estimated as

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n\sigma_X \sigma_Y}$$

where x_i, y_i are the observed values of the random variables X, Y and $\overline{x}$, $\overline{y}$, σ_x, σ_Y are respectively the empirical means and standard deviations of X and Y calculated from the sample.

Alternatively, we can assess the significance of a candidate attribute by deciding whether the distributions of attribute values, which correspond to the classes, differ substantially using parametric (e.g., *t*-test (Devore 2000)) or non-parametric tests (e.g., Wilcoxon rank sum (Conover 1999)). Here, we present in more detail these tests and how they are being used. Let us consider again the two distinct classes $\mathbf{S}_X$ of size N_{Sx} and $\mathbf{S}_Y$ of size N_{Sy}. The initial assumption that holds for the two classes is that the observed samples from the two

populations are generated by two independent (unrelated) processes. Under this assumption, the following hypothesis testing scenario is examined at each step of the recursive partitioning process:

- H_0 (*null hypothesis*): The E[V_{mean}] is the same for the two classes; the hyper-rectangle under consideration is not discriminative among the populations.
- H_1 (*alternative hypothesis*): The E[V_{mean}] is not the same for the two classes; the hyper-rectangle under consideration is discriminative among the populations.

Prior to performing the statistical test, experimental data is used to determine the distribution of the test statistic under the null hypothesis. Based on the value of the test statistic, the null hypothesis may be rejected in favor of the alternative hypothesis. The rejection region (the set of values for the test statistic when the null hypothesis H_0 is rejected) is determined based on a pre-specified *significance level* α—the probability of the Type I error (that the null hypothesis H_0 is wrongly rejected by the test in the case that it is in fact true). The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis H_0, is true. A p-value is a measure of evidence against the null hypothesis; the smaller the p-value, the more evidence we have against H_0.

Depending on the characteristics of the datasets we employ parametric or non-parametric statistical tests:

1. The V_{mean} attributes have normal (Gaussian) distribution for each class, and their variances are the same for the two groups.

In order to verify the normality and equal variance assumptions, we need to have a reasonable sample size. If these assumptions are satisfied, to evaluate the *null hypothesis* H_0 that the means of attributes in the two classes are the same, we use the unpaired t-test. We compute the following test statistic:

$$t = \frac{\left(\overline{V_{\text{mean}_{Sy}}} - \overline{V_{\text{mean}_{Sx}}}\right)}{s\sqrt{\frac{1}{N_{Sy}} + \frac{1}{N_{Sx}}}} \tag{1}$$

where $\overline{V_{\text{mean}_{Sy}}}, \overline{V_{\text{mean}_{Sx}}}$ are the empirical means of V_{mean} for class $\mathbf{S_Y}$ and $\mathbf{S_x}$ respectively calculated from the available sample and s is the pooled standard deviation of the two classes given by

$$s = \sqrt{\frac{(N_{Sy}-1)s_1^2 + (N_{Sx}-1)s_2^2}{N_{Sy}+N_{Sx}-2}}$$

Under H_0, t from Eq. 1 follows the t-distribution with $\left(N_{Sy}+N_{Sx}-2\right)$ degrees of freedom (Devore 2000).

2. The V_{mean} attributes do not satisfy the normality and/or equal variance assumptions.

In the case where the normality and/or equal variance assumptions do not hold or cannot be validated due to a small dataset size, we use the non-parametric *Wilcoxon rank sum test*, which makes no distributional assumptions and can be considered the alternative of the unpaired t-test for this type of analysis. The *Wilcoxon rank sum test* is based on the sum of ranks of the V_{mean} attributes in each of the two classes. More specifically, the V_{mean} attributes are ranked according to their value, and equal V_{mean} observations are assigned the average of the corresponding rank values. The sum of the ranks T is calculated for the group

with the smaller sample size N_{Ss}=min(N_{Sx},N_{Sy}). For datasets with relatively small size (N_{Ss}<10, N_{Sx}+N_{Sy}<20), we consider as test statistic the value of T. Otherwise, (as is the case in our experiments) the test statistic is calculated as

$$z = \left(\frac{T - \mu_T}{\sigma_T} \right) \tag{2}$$

$$\text{where } \mu_T = \frac{N_{Ss}(N_{Sx} + N_{Sy} + 1)}{2}, \text{ and}$$

$$\sigma_T = \sqrt{\frac{N_{Sx} N_{Sy}(N_{Sx} + N_{Sy} + 1)}{12}}.$$

Under the null hypothesis H_0 (the attributes in two classes have the same medians), z from Eq. 2 follows a normalized Gaussian univariate distribution.

Another criterion that can be used is based on discretization of the candidate attribute and evaluation of the class/attribute contingency matrix using statistical tests (chi-square or the Fisher exact test (Agresti 1996)) with pre-determined maximal type I errors. In this case, a suitable value for the discretization threshold is set ad-hoc or by using discretization techniques that maximize class/attribute mutual information (Ching and Wong 1995).

The DRP algorithm effectively reduces the number of statistical tests applied, because the statistical tests are applied selectively on groups of voxels (cuboids) instead of on individual voxels. This also helps alleviate the effect of the *multiple comparison problem*, although a type of Bonferroni correction may still be employed (see section 2).

3.3 Implementation

For the implementation of this procedure, efficient data representation and manipulation is done using augmented oct-trees (Fujimura et al. 1983) and dynamic arrays (Cormen et al. 2001) for storing pointers to the leaf nodes. (Similarly, quadtrees are used for the implementation in two dimensions (2D)). If the splitting criterion is satisfied, the spatial sub-domain (or cuboid) corresponding to the node of the oct-tree is partitioned into eight smaller sub-domains using the SPLIT8 routine similar to the splitting algorithm in Quad trees (Samet 1984; see Fig. 3). The corresponding tree node becomes the parent of eight children nodes, each of which represents a new sub-domain (new cuboid). Figure 2b illustrates the concept of splitting a tree node V^2 into smaller sub-domains (children nodes). The new measurements V^3_{mean} corresponding to the children node V^3 becomes a new candidate attribute. Figure 3 shows the outline of the DRP algorithm.

3.4 Classification model

To verify the discriminative power of the patterns detected by DRP, we propose to build classifiers using the extracted highly informative attributes. In our experiments, we used linear and non-linear neural networks as classification models. These are universal approximators and have been often reported to outperform the alternatives for classifying real life linear and non-linear phenomena (Haykin 1999). Depending on the size of the dataset and the number of discriminative regions detected by DRP, different network

Given: Oct-tree *T* (a global variable) corresponding to the spatial domain D; Two sets $S_Y = \{S_{1,Y},...S_{n1,Y}\}$, $S_N = \{S_{1,N},...S_{n2,N}\}$ containing region data for samples belonging to classes Y and N respectively; Splitting criterion *s_c* to decide whether to split further the oct-tree node according to the *threshold* value.

```
DYNAMIC RECURSIVE PARTITIONING (node, SY, SN, s_c, threshold )
If SPLITTING_DECISION(node, SY, SN, s_c, threshold)==’yes’
    CHILDREN=SPLIT8(node)
    for node_c in CHILDREN
    DYNAMIC RECURSIVE PARTITIONING (node_c, SY, SN, s_c, threshold )


SPLITTING_DECISION(node, SY, SN, s_c, threshold)
switch s_c
        case ‘Pearson’
                Compute Pearson Correlation Coefficient ρ between the class label and the attribute values;

                If ρ >threshold
                        return ‘yes’
                else
                        return ‘no’.

        case ‘Contigency test’
                Discretize the attribute values;
                Construct contingency table containing discretized attribute values and the class labels;
                Apply Fisher/t-test to determine association between class and attribute;
                If p_value<threshold
                        return ‘yes’
                otherwise
                        return ‘no’;
        case ‘t-test’
        Use t-test to determine whether the means of attributes significantly differ between classes.
                If p_value<threshold
                        return ‘yes’
                otherwise
                        return ‘no’;
        case ‘Wilcoxon’
       Use rank sum test to determine whether the means of attributes significantly differ between classes.
                If p_value<threshold
                        return ‘yes’
                otherwise
                        return ‘no’;
```

Fig. 3 The outline of the DRP algorithm in pseudocode

architectures and training algorithms were implemented to achieve optimal performance. For example, to avoid overfitting due to the small size of real-world datasets, we employed one-layer perceptron networks trained by the Pocket algorithm (Gallant 1990). We used sigmoidal neural networks in all other cases. This is discussed in detail in section 5.

As inputs to the classifier we employ the attributes V_{mean} of the discovered regions, after being standardized to have zero mean and unit standard deviation. The output of the model is represented by binary class label (in the case of two class dataset) indicating the class of the samples. A statistically valid estimate of the classification accuracy on out-of-sample data is obtained by the leave-one method that is explained in section 5.2.

In the case that a large number of attributes are extracted by DRP, techniques for attribute selection can be further employed to reduce the dimensionality and eliminate highly correlated attributes. In addition to sequential backward and forward searches for

dimensionality reduction, the branch and bound search can be applied for iterative reduction of the attribute set.

To demonstrate the ability of this approach to effectively detect discriminative regions of the 3D image data and to evaluate the discriminative power of the attributes corresponding to these regions, we present classification experiments on various datasets. We first introduce these datasets.

4 The datasets

We evaluated the proposed approach on both synthetic and real datasets. We used two synthetic datasets, a mixture of Gaussians, and a fractal dataset. We also used a real dataset of 3D fMRI brain image scans acquired from a study on Alzheimer's disease (AD) and a realistic dataset of brain lesions. All datasets are publicly available at http://denlab.temple.edu/data_repository.

4.1 Synthetic data

Gaussian Mixtures This set consists of two distinct classes of homogeneous (binary) volume data composed of mixtures of nine Gaussian distributions. These mixtures had the same component means for each class but different component variances, namely 0.1 and 0.01. (In the case of generating volume data with different component means, classification was trivial (Pokrajac et al. 2005).) We experimented with 50 and 200 3D volume data samples for each class. Figure 4 illustrates an example from each distribution.

Fractal Data In order to evaluate our proposed approach on highly non-uniform data, we included in our experiments a set of binary fractal volume data with varied levels of density within different spatial sub-domains. These data were generated according to the probabilistic structure of a non-complete oct-tree with a maximal depth $L_{\max}$ and probability $p_l(L)$ that a node of the tree at level L is a leaf [i.e., the node will branch further with the probability of $1-p_l(L)$]. The set consisted of two distinct classes with different voxel densities on tree leaves. Each class included 100 samples. The probability that a leaf node at level L is discriminative, was specified as $p_{\text{dis}}(L)$. The probabilities $p_l(L)$ and $p_{\text{dis}}(L)$ were modeled to have fractal-like exponential dependence on the node level L according to $p_l(L) = 1/2^{(3-d_l)L}$ and $p_{\text{dis}}(L) = p_{\text{dis}}(L_{\max})/2^{(3-d_{\text{dis}})\cdot(L_{\max}-L)}$. All the volume data were generated with parameter values $L_{\max}$=8, d_l=2.1, d_{dis}=2.5 and $p_{dis}(L_{\max}) = 20\%$.

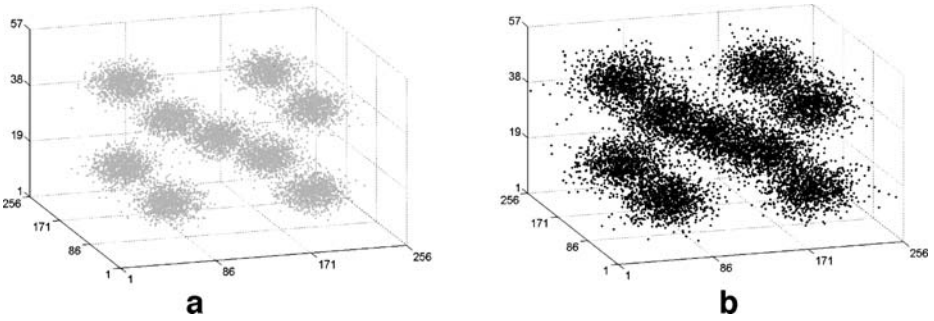

Fig. 4 Volume data samples from each class; sample with variance **a** 0.01 and **b** 0.1

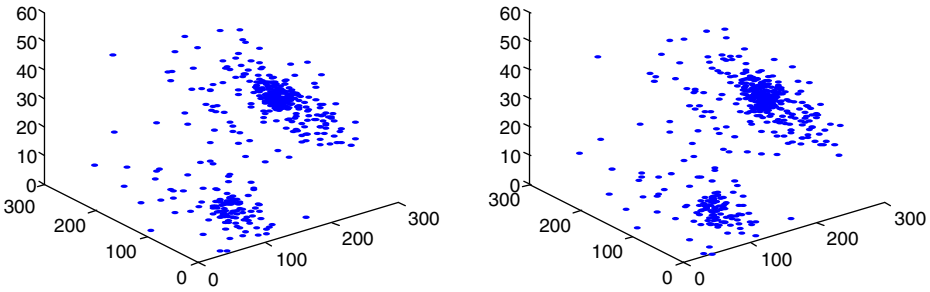

Fig. 5 Two distinct classes of fractal data (that at first glance look very similar)

To make the experimental evaluation more challenging, the fractal data distributions we created were very similar to each other. Figure 5 shows 3D fractal volume samples from each class.

4.2 Real data

The real datasets include 3D volumes of brain image scans. More specifically, the samples consist of 3D fMRI contrast activation maps[5] acquired from a study (Saykin et al. 1999) on Alzheimer's disease (AD). The study involved subjects participating in cognitive tests specifically designed to differentially probe semantic knowledge of categorical, functional, and phonological congruence between word pairs, in order to explore neuroanatomical correlates in AD. The task was to identify word pairs with correct category exemplar (CATX) relationships. The 3D volumes were registered to a common spatial template in order to become comparable across subjects, and background noise from sensor fluctuations was removed (Saykin et al. 1999, Kontos et al. 2004). The set consists of nine control and nine patient fMRI contrast activation volumes. Figure 6 shows examples of these fMRI volumes.

Although the Alzheimer's disease dataset size is small compared to our artificial data, it is important to mention that this comparably small number of samples is usual for fMRI studies because this kind of imaging is costly to acquire. In addition, by including real data in our evaluation, we intend to explore the potential of applying our proposed approach to real world datasets and problems. Notice that fMRI data are usually highly heterogeneous and non-uniform distributions and therefore make good candidates for such an evaluation.

4.3 Realistic data

To experiment with a larger dataset that closely approximates real dataset, we generated a realistic dataset using an established lesion-deficit simulator (LDS; Megalooikonomou et al. 2000a, Megalooikonomou 2002). In implementing the LDS, probability distributions are used to model the number, size, spatial distribution, and shape of lesions, as well as other parameters, such as registration error and structure-function associations. This dataset consists of simulated 3D volumes that comply with a statistical model for brain lesions

[5] Contrast activation maps show the difference between rest and active conditions for activation/deactivation levels in the human brain.

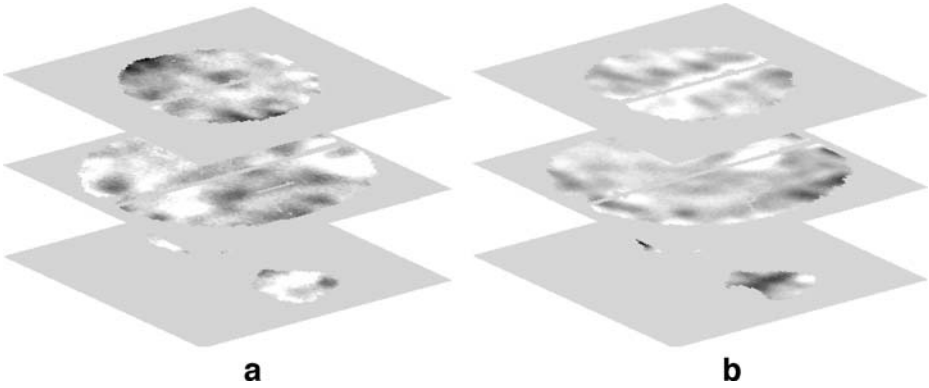

Fig. 6 Sample 2D slice views of 3D fMRI contrast activation volume data for **a** a control and **b** patient subject of the CATX dataset

derived from a medical study. More specifically, the medical study that these data conforms to is the Frontal Lobe Injury in Childhood (FLIC) study for Attention Deficit Hyperactivity Disorder (ADHD) patients (Gerring et al. 1998). The dataset includes two classes corresponding to two distributions. The first class simulates the subjects who developed ADHD, ("yes ADHD" class) while the second class simulates the subjects who did not develop ADHD ("no ADHD" class) after a closed head injury (see Fig. 7). Each class consists of 50 samples with approximately 200 lesion voxels per 3D volume (i.e., per subject).

5 Experimental evaluation

Here we report the results of applying our proposed approach to the four datasets described in section 4. We explored the effectiveness of DRP by performing classification experiments utilizing Neural Networks (Haykin 1999). Given a training set with class labeled 3D volume data, our goal was to detect discriminative areas and then use these areas to successfully determine class membership for new unseen 3D volume data. In

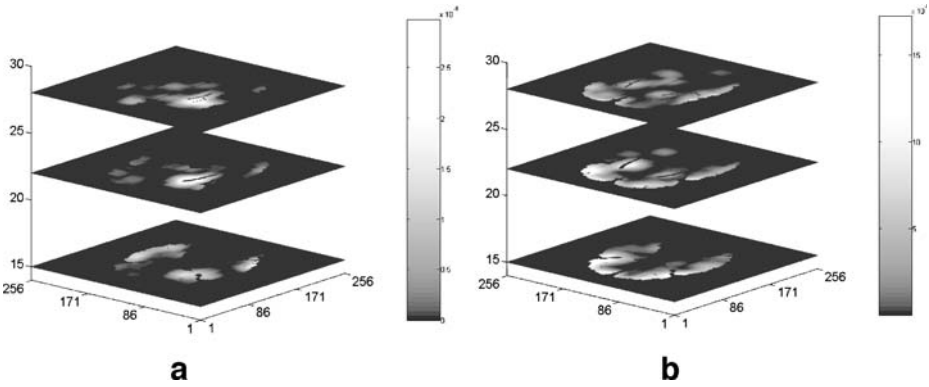

Fig. 7 Distributions for **a** "yes ADHD" and **b** "no ADHD" classes

general, we evaluated the performance by calculating the classification accuracy defined as the ratio of the number of rounds when the classification of a new sample was successful to the total number of rounds. In the experiments we tested a wide range of parameters for DRP (statistical stopping criterion, maximum splitting levels, etc.) in different experimental settings. We also compared our proposed approach with other methods for 3D volume classifications, such as distributional distance techniques, maximum likelihood approaches, and voxel-based statistical techniques (see section 2). For distributional based techniques that are deterministic in nature (repetition of the experiments on the same data will give the same result) we reported only the accuracy. For the stochastic techniques (those based on neural networks training with random parameter initialization) we reported the mean and average accuracy through the repeated experiments.

We experimented with two statistical tests: the *t*-test and the ranksum test. The *t*-test was employed under the normality and equal variance assumption for the V_{mean} attributes in each class. To effectively evaluate cases with relatively small datasets (e.g., fMRI data), we also employed the Wilcoxon rank-sum test to assess the statistical significance of the V_{mean} attributes divergence among classes. A threshold value on the *p*-value that is typically used in statistical analysis is 0.05. Another typically used but stricter *p*-value threshold is 0.01. We experimented with both values. The tree depth limit that we used in the fMRI experiments is justified, taking into consideration the resolution of our data. The dimensions of the initial fMRI contrast activation maps are $79\times95\times68$. Hence a tree depth limit of 3 equals a voxel neighborhood of approximately $10\times12\times9$ voxels. Also, a tree depth limit of 4 equals a voxel neighborhood of approximately $5\times6\times4$ voxels.

5.1 Results on synthetic data

Gaussian Mixtures We applied DRP with the non-parametric ranksum test as a stopping criterion and a significance threshold set to 0.05. The maximum splitting levels (tree length) that the algorithm was allowed to proceed was 4. For the classification, we used the best eight attributes extracted by DRP according to their correlation with the class label. We trained feed forward sigmoidal neural networks with one hidden layer and a number of neurons equal to the number of input attributes. The number of output nodes was equal to the number of classes, and classification was decided according to the *winner-takes-all* principle (class corresponds to the output with the larger response). Training and parameter estimation for the classifiers was performed using the Levenberg–Marquardt optimization (Hagan and Menhaj 1994). We performed ten classification trials for DRP and Maximum likelihood method, while experiments were not repeated for Kullback–Leibler and Mahalanobis based methods since they are deterministic in nature. For each trial, we generated a new random sample from the underlying mixtures of distributions and repeated the training and testing process ten times with a random initialization of the network parameters. Table 1 shows the classification performance achieved by DRP, as well as the comparative accuracies obtained when using the approaches reviewed in section 2. DRP was able to almost perfectly classify samples drawn from either one of the classes. In general, DRP was comparable to the Maximum Likelihood parametric approach. To verify this, we applied an unpaired *t*-test which could not reject hypothesis of the same mean accuracies at the 0.05 significance level. However, in this case DRP still has additional advantage of detecting the sub-areas in the 3D volumes where the class distributions differ substantially. The other approaches were not able to provide good classification, especially for samples belonging to the class with smaller variance; compared to them, Both DRP and Maximum Likelihood obtained significantly higher classification accuracy (p-value$<10^{-4}$).

Table 1 Comparative classification performances on Gaussian mixture data after applying *DRP* and neural networks, as well as other volume data classification techniques

Method	Classification accuracy (%)		
	0.01 Variance class	0.1 Variance class	Total
DRP	99	99	99±10
Maximum likelihood	95	100	98±7
Kullback–Leibler	58	100	79
Mahalanobis	57	48	53

Fractal Data Since the highly non-uniform distributions (see section 4.1) were very difficult to discriminate, we applied DRP with two different experimental settings. In the first set of experiments, we used a *t*-test as a stopping criterion with a significance threshold level set to 0.01. For the rest of the experiments, we employed the ranksum test with the significance threshold set to 0.05. In both cases, the maximum splitting levels (tree length) that the algorithm was allowed to proceed was 5. We trained sigmoidal feed forward neural networks with one hidden layer and neurons equal to the number of input attributes. The number of output nodes was again equal to the number of classes, and classification was performed according to the *winner-takes-all* principle (class corresponds to the output with the larger response). The training was performed using 50 samples from each class (100 samples in total). The remaining 100 samples were used for testing. Each training–testing trial was repeated ten times. The model coefficients of the network were estimated using the Levenberg–Marquardt (Hagan and Menhaj 1994) optimization algorithm. Table 2 illustrates the classification accuracy and standard deviation achieved by DRP and other approaches (as in previous experiments, standard deviation for Kullback–Leibler and Mahalanobis based methods were not computed since these methods are deterministic in nature).

DRP was significantly more accurate than the other classification approaches (*p*-value for *t*-test to compare accuracies was smaller than 10^{-4}). The tree structures representing the partitioning performed by DRP in this case are shown in Fig. 8. We would like to add that the accuracy of the two variants of DRP were also similar (the null hypothesis that they have the same accuracies could not be rejected even at 0.05 significance level).

5.2 Results on real data

We applied DRP to the fMRI dataset and tested a range of experimental settings. We employed the *t*-test and parametric ranksum test as stopping criteria with the *p*-value threshold set to 0.05 and 0.01. The maximum splitting steps that the algorithm was allowed to proceed (tree depth) was 3 and 4 (this choice was justified earlier given the resolution of the data). For the classification task, we implemented one-layer perceptron networks trained by the Pocket algorithm (Gallant 1990) to avoid overfitting due to the small size of the

Table 2 Comparative classification performances on the *Fractal* data after applying *DRP* and neural networks as well as other volume data classification techniques

Method	Classification accuracy (%)
DRP (*t*-test)	92±11
DRP (ranksum test)	95±12
Kullback–Leibler	57
Mahalanobis	66

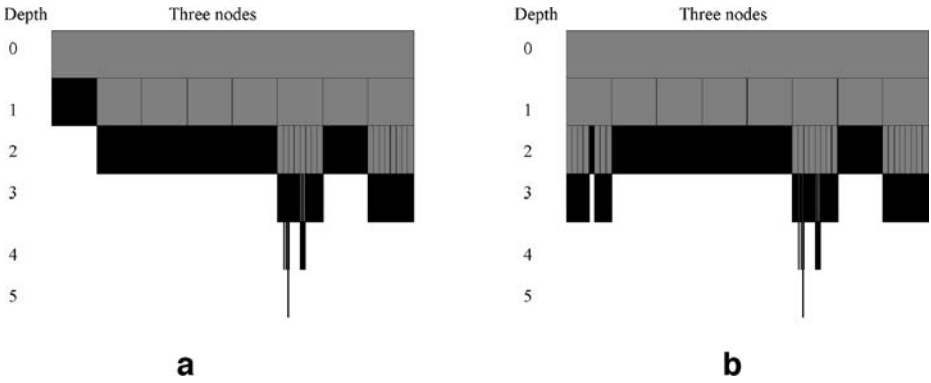

Fig. 8 The tree structures discovered when applying DRP to the fractal data using **a** the correlation criterion and **b** rank-sum test. Gray and black rectangles denote non-leaf and leaf nodes respectively. Children nodes are represented below its parent corresponding to a successive tree level (at the next depth)

datasets. As inputs to the classifier model, we used all the V_{mean} attributes of the detected highly discriminative hyper-rectangles, which were standardized to have zero mean and unit standard deviation. In this particular application of DRP on brain imaging 3D volumes, these highly informative sub-regions reflect discriminative functional activation patterns. They also reveal associations between spatial regions and non-spatial characteristics, such as the presence of a disease (AD). Figure 9 illustrates sample views of discriminative sub-regions detected by DRP within the 3D fMRI brain imaging domain. As output, we used a binary class label indicating the class of the samples (control vs. patient). The goal of these experiments was, given an fMRI image of a new subject, to determine the group to which it belongs (i.e., controls vs. patients).

For training and testing, leave-one-out cross-validation was employed to evaluate out-of-sample classification performance (Fukunaga 1990, Duda et al. 2000). 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 the patient and control with index i, where $i=1,...,9$. Taking into consideration the stochastic nature of the Pocket algorithm, we repeated the process of training and testing the model in each of the leave-one-out loops for five times and averaged the percentage of the correct predictions to obtain the reported accuracy. Table 3 shows the classification performances for all the experimental settings, as well as

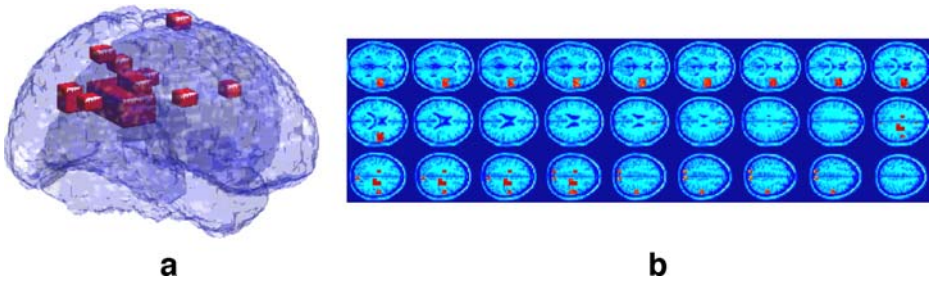

Fig. 9 Sample views for the discriminative sub-areas detected by DRP for the *CATX* data, when applied with *t*-test, significance threshold=0.01, and tree depth=4. A 3D view (**a**) and a 2D slice view of results overlaid on the brain atlas template (**b**)

Table 3 Classification accuracies (± standard deviations) when applying DRP and neural network classifiers on the real fMRI contrast activation volume dataset, as well as the comparative results when applying maximum likelihood and distance-based techniques

Method				Classification accuracy (%)		
	Criterion	Threshold	Tree depth	Controls	Patients	Total
DRP	Correlation	0.4	3	82	93	88±5
	t-Test	0.05	3	89	100	94±0
		0.05	4	84	100	92±3
		0.01	4	87	100	93±2
	Ranksum	0.05	3	87	100	93±6
		0.05	4	80	100	90±5
		0.01	4	87	96	91±3
Maximum likelihood/EM				77	67	72±7
Maximum likelihood/*k*-means				77	83	80±5
Kullback–Leibler/EM				79	57	68±8
Kullback–Leibler/*k*-means				77	66	71±6

comparative results obtained using other techniques. Using DRP we were able to achieve a prediction accuracy of 90% or more, improving on results previously published in the literature for the same dataset that used principal components analysis and a Fisher linear discriminant classifier (Ford et al. 2003). In addition to improving classification accuracy, the obtained results also support the argument that the hyper-rectangles indicated by DRP in the specific study are indeed discriminative with respect to class membership (control vs. patient).

To further evaluate the effectiveness of our proposed approach, we present comparative experimental results obtained when applying other volume classification techniques. We compare the DRP algorithm to maximum likelihood and distance-based techniques, voxel-wise statistical processing, and a static partitioning approach (see section 2).

DRP vs. distance-based and maximum likelihood techniques We compared the performance of DRP to that of maximum likelihood and distance-based classification techniques. These techniques have been previously evaluated on the same Alzheimer's dataset (Pokrajac et al. 2005). EM and *k*-means were considered to estimate the distribution of the discriminative functional activity within the fMRI volume. The classification process was repeated 30 times over each leave-one-out cross-validation loop, due of the stochastic nature of the algorithms and the small size of dataset. From Table 3 that reports achieved averaged accuracy and standard deviation of evaluated techniques, it is apparent that DRP clearly outperforms other proposed methods. On average, DRP outperforms the other methods by 18% (p-value$=1.0125e-005<0.0001$ using unpaired t-test).

DRP vs. voxel-wise statistical analysis In order to compare DRP to voxel-wise statistical volume analysis (as used for example in SPM), we applied the same statistical tests and criteria employed in the DRP algorithm on a voxel-by-voxel level. We performed a computational complexity comparison and compared the detected areas to the corresponding sub-regions formed by voxels that are indicated as highly discriminative. Recall that one of the disadvantages of voxel-wise analysis is the *multiple comparison problem* (see section 2). As Table 4 illustrates, DRP has the ability to alleviate this effect by applying statistical tests selectively on groups of voxels thus reducing the number of

Table 4 Comparison of statistical tests performed by DRP and voxel-wise processing in different experimental settings

Method	Experimental settings			Number of statistical tests performed
	Statistical test	Threshold	Tree depth	
DRP	Correlation	0.4	3	465
	t-Test	0.05	3	569
		0.05	4	4,425
		0.01	4	4,665
	Ranksum	0.05	3	553
		0.05	4	4,297
		0.01	4	4,585
Voxel wise analysis	Correlation	0.05/0.01	–	201,774
	t-Test	0.05/0.01	–	201,774
	Ranksum	0.05/0.01	–	201,774

statistical tests performed. Although DRP and voxel-wise analysis indicate almost similar discriminative sub-regions for the same experimental settings (see Fig. 10), DRP is 2–3 orders of magnitude faster.

DRP vs. static partitioning Finally, we compared DRP to the static partitioning approach. This approach is naïve (as compared to the adaptive partitioning of the space) and simply partitions the space into hyper-rectangles of equal length. Each dimension is split into

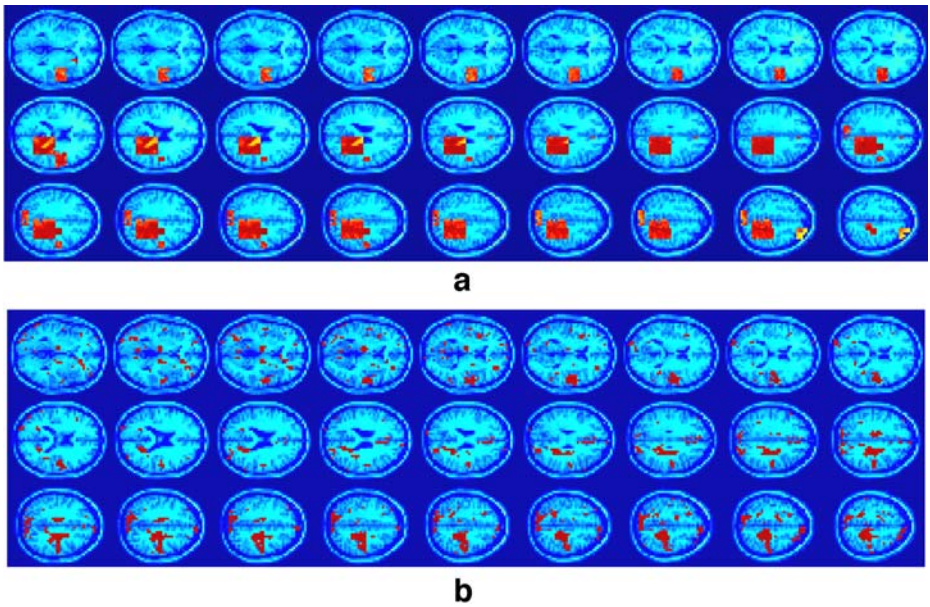

Fig. 10 Discriminative sub-regions detected when applying **a** DRP and **b** voxel-wise analysis with ranksum test and significance threshold 0.05 to the real fMRI volume data

Table 5 Classification accuracy based on static partitioning using one-layer perceptron neural networks and Pocket algorithm

k	Classification accuracy (%)		
	Controls	Patients	Total
2	59	71	65±5
3	58	79	68±4
4	100	0	50±0

l equal length bins, resulting in a total partitioning of the space of $l \times l \times l$ hyper-rectangles for the 3D domain. The V_{mean} of each cuboid, which is treated again as a representative attribute, is used as input to the classifier models. For the classification we again used one-layer perceptron networks trained by the Pocket algorithm. Training and testing was performed using the leave-one-out cross-validation approach. Table 5 illustrates these results. Observe that although for k=2, 3 the results are somewhat reasonable, for k=4 the classifier learns a decision boundary that classifies everything as one class (in all experiments performed with the specific settings). Comparing the classification results of static partitioning in Table 5 to those in Table 3 we see that, on average, the proposed method, DRP, outperforms static partitioning by 30% (p-value=2.5595e−005<0.0001 using unpaired t-test).

5.3 Results on realistic data

For this set of simulated medical data, we applied DRP with very selective settings for non-parametric ranksum test and the significance threshold set to 0.01. The maximum tree depth (i.e., number of splitting levels) that the algorithm was allowed to proceed was again 4. The classification model was implemented with feed forward two-layer neural networks with hidden neurons equal to the number of attributes and output nodes equal to the number of classes. The predicted class was again the one with the largest response. Two learning algorithms were tested, namely the resilient propagation (Riedmiller and Braun 1993) and the Levenberg–Marquardt (Hagan and Menhaj 1994). We performed 200 trials, each time drawing a new random sample from the underlying distributions.

The classification accuracy obtained using the features extracted by DRP was almost perfect, with error less than 1%. This was comparable to the Kullback–Leibler distance approach, which also achieved almost perfect classification. Finally, DRP's classification performance exceeded that of the Mahalanobis distance approach by 10%.

6 Discussion

In this section, we attempt to interpret the observed superior performance of DRP and provide reasoning behind it. We also discuss possible limitations of DRP and explain when they may happen.

Classifying the synthetic data of Gaussian mixtures is rather challenging due to the fact that the samples with smaller variance are overshadowed by the samples with larger variance. This introduces significant difficulty when deciding class membership. The experimental evaluation presented in the paper illustrates that DRP is able to almost perfectly classify samples drawn from either one of the classes. The other approaches (with

the exception of Maximum Likelihood that obtained similar classification accuracy to that of DRP) are not able to provide good classification performance, especially for samples belonging to the class with the smaller variance.

DRP is particularly successful in the analysis of fractal data as presented in Table 2. This is a very important result, because the fractal data included in this set are visually very difficult to differentiate (see Fig. 5). The proposed approach shows a potential for developing robust and efficient volume data classification tools. It is also very interesting to observe that the trees constructed by DRP during the process of the adaptive partitioning of the space (shown in Fig. 8) are very similar to each other. These results suggest that DRP partitioning trees can also be utilized to model highly non-uniform data distributions, such as fractal-like datasets. The experimental results also show that the accuracy of the two variants of DRP (one using the *t*-test and the other using the ranksum test) is similar. Further exploration is needed to differentiate between performances of these two variants.

For the fMRI dataset, the areas detected by DRP as discriminative, elucidate large hemispheric and lobar differences between Alzheimer's patients and controls for all semantic decision tasks. These findings actually comply with those obtained from other medical studies (Saykin et al. 1999, Flashman et al. 2003). DRP outperforms distance-based and maximum likelihood techniques in this analysis. In addition, DRP demonstrates an advantage of simultaneously indicating discriminative sub-regions in the 3D volumes that are of particular interest when deciding class membership.

The difficulty when performing volume distribution estimation is that usually real data conform to highly non-uniform distributions that cannot be accurately modeled by mixtures of Gaussian components. This issue is actually reflected by the comparative results of Table 3 where the other approaches we considered cannot provide highly accurate classification, performing only slightly better than a random guess in some cases.

The adaptive nature of DRP, combined with the predetermined number of levels that the algorithm is allowed to proceed (tree depth), effectively reduces the number of statistical tests and, consequently, efficiently deals with the multiple comparison problem (see section 2). Moreover, DRP avoids detecting single voxels as discriminative (outliers), which is typical with voxel-wise statistical processing. This improvement is illustrated in Fig. 10 where data processed with voxel-wise statistical analysis demonstrate fragmentation of the 3D volume. This indicates the ability of DRP to provide an efficient and effective volume data classification technique that is able to process large datasets of high resolution.

Moreover, DRP clearly outperforms the static partitioning approach. This result is expected because DRP is more flexible in partitioning the space adaptively and the number of attributes that correspond to hyper-rectangles is significantly smaller than those of static partitioning for the same granularity level (i.e., size of partitions).

The classification accuracy obtained by DRP on the realistic data is almost perfect, similarly to the Kullback–Leibler distance approach. However, the advantage of DRP is that the detected highly discriminative sub-areas in the volume domain can be utilized to form associations between class membership and spatial characteristics. This can be particularly useful in medical applications, such as the clinical study examined here, where the expert (e.g., physician) is interested mostly in specific areas within the image data that are highly discriminative.

DRP has the following limitations that can form the basis for future work in this area:

1. In datasets where the underlying distributions differ homogeneously in their entire spatial extend rather than locally, DRP can detect statistically significant differences from the very first statistical test and refrain from performing the partitioning of the

space. A solution for such a scenario is to assign a more strict p-value significant threshold (e.g., p-value<0.0001) which will force DRP to proceed with partitioning the space and detect the most significant localized distribution effects.

2. Currently DRP does not create hyper-rectangles that are overlapping. Disjoint hyper-rectangles at each level can miss discriminative areas spread over multiple adjoin hyper-rectangles. Hence, DRP may not achieve desired performance on data that express this property. More specifically, we expect that DRP will not perform well on elongated linear or complex regions of interest.
3. The algorithm does not have a pruning phase, which would merge discriminative regions and might help alleviating issue 2 above. Observe that pruning can also improve generalization of other machine learning techniques [e.g., C4.5 (Quinlan 1993) and lead to simplification of the discovered rules].

Although performed experiments have provided evidence that the proposed method can be very successful in detecting spatial regions, more work is needed to introduce new ways of allowing the algorithm to trade off accuracy for efficiency.

7 Conclusions

In this paper, we proposed and evaluated a novel Dynamic Recursive Partitioning (DRP) approach for the classification of 3D image data. We focused on efficiently detecting spatial sub-regions that are discriminative among different classes. The main idea is to apply an adaptive partitioning (guided by statistical tests) of the initial 3D domain until highly informative spatial sub-domains are detected. To validate how effective the approach is at discovering discriminative spatial sub-regions, we performed classification where we used neural networks and attributes extracted from the discriminative regions detected by DRP. We evaluated our technique on both artificial and real datasets. Our study involved a wide range of synthetic, realistic, and fractal artificial datasets. To explore the potential of our approach with real-world problems, we also used actual, 3D brain imaging datasets. Experiments demonstrated that the proposed method outperforms alternative approaches (e.g. distributional distance-based methods, maximum likelihood classification, voxel-wise analysis, and static partitioning) by achieving up to 30% better accuracy on artificial data and 15% better accuracy on real data. At the same time, the proposed method reduces by two orders of magnitude the number of statistical tests required by voxel-wise analysis; this not only reduces the computational cost but also the extent of the *multiple comparison problem* because it considers groups of voxels (spatial sub-domains) rather than individual voxels. In addition, the new method is expected to be more robust than voxel-wise analysis, which is prone to errors contributed by certain preprocessing steps such as the spatial normalization required in specific applications. We believe that DRP can be a robust and efficient approach for 3D volume data classification in other spatial data mining applications.

Acknowledgements The authors would like to thank A. Saykin for providing the fMRI dataset and clinical expertise. This work was supported in part by the National Science Foundation under grants IIS-0083423, IIS-0237921, HRD-0310163, HRD-0320991 and #HRD-0630388, the National Institutes of Health under grant R01 MH68066-03 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 and INBRE Grants (P20 RR16472, #2 P20 RR016472-04), the Pennsylvania Department of Health, and the DoD HBCU/MI Infrastructure Support Program (45395-MA-ISP Department of Army). All agencies specifically disclaim responsibility for any analyses, interpretations, or conclusions.

Appendix

Mahalanobis distance:

Given datasets S_z and S, the Mahalanobis distance, d_M, between them is computed as:

$$d_M = \sqrt{\left(\mu_{s_z} - \mu_S\right)^T \cdot \Sigma^{-1} \cdot \left(\mu_{s_z} - \mu_S\right)},$$

where μ_{s_z} and μ_s are mean vectors of the datasets S_z and S respectively, and Σ is the sample covariance matrix:

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

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.

K–L divergence:

Unlike the Mahalanobis distance, the KL divergence $\mathrm{d}_{KL}(p(x),\ q(x))$ 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:

$$d_{KL}(S_z, S) = \int_D p_z(\mathbf{x}) \log \frac{p_z(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}.$$

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 as:

$$d_{KL}(S_z, S) \approx \sum_{\text{all voxels } v_{i_1,i_2,i_3}} p_z\left(x_{i_1,i_2,i_3}\right) \log \frac{p_z\left(x_{i_1,i_2,i_3}\right)}{p\left(x_{i_1,i_2,i_3}\right)} \Delta x.$$

Here, $p_z\left(x_{i_1,i_2,i_3}\right)$ and $p\left(x_{i_1,i_2,i_3}\right)$ are estimated probability densities of the involved distributions at the voxel centers $\left(x_{i_1,i_2,i_3}\right)$, and Δx is the product of corresponding discretization intervals.

References

Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., & Swami, A. (1992). *An interval classifier for database mining applications. VLDB Conference*. Vancouver, BC, Canada.

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases. ACM SIGMOD Conference on Management of Data*. Washington, DC.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). *Fast discovery of association rules. Advances in knowledge discovery and data mining*.

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. *VLDB Conference*. Santiago, Chile.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Aladjem, M. (1998). Nonparametric discriminant analysis via recursive optimization of Patrick–Fisher distance. *IEEE Transactions on Systems Man and Cybernetics, 28B*, 292–299.

Andersen, E. (1997). *Introduction to the statistical analysis of categorical data*. Berlin: Springer Verlag.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., et al. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*, 535–542.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society, 35*, 99–110.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Castelli, V., & Kontoyiannis, I. (2002). *An efficient recursive partitioning algorithm for classification using wavelets*. Brown University.

Castelli, V., Kontoyiannis, I., Li, C. S., & Turek, J. J. (1996). Progressive classification in the compressed domain for large EOS satellite databases. *IEEE ICASSP*.

Ching, J., & Wong, A. (1995). Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*, 641–651.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (2001). *Introduction to algorithms*. Cambridge: MIT Press.

Davatzikos, C. (2004). Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage, 23*, 17–20.

Devore, J. L. (2000). *Probability and statistics for engineering and the sciences*. Belmont: International Thomson Publishing Company.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification*. John Wiley and Sons.

Ester, M., Kriegel, H.-P., & Sander, J. (1977). *Spatial data mining: A database approach. Symposium on Large Spatial Databases*. Berlin, Germany: Springer.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial database with noise. International conference on knowledge discovery in databases and data mining (KDD-96)*. Portland, OR.

Euripides, G. M., Petrakis, M., & Faloutsos, C. (1997). Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering, 9*, 435–447.

Faloutsos, C. (1996). *Searching multimedia databases by content*. The Netherlands: Kluwer Academic.

Flashman, L. A., Wishart, H. A., & Saykin, A. J. (2003). Boundaries between normal aging and dementia: Perspectives from neuropsychological and neuroimaging investigations. In V. O. B. Emory, & T. E. Oxman (Eds.), *Dementia: Presentations, differential diagnosis and nosology*. Baltimore: Johns Hopkins University Press.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al. (1995). Query by image and video content: The QBIC system. *IEEE Computer*, 23–32.

Ford, J., Farid, H., Makedon, F., Flashman, L. A., Mcallister, T. W., Megalooikonomou, V., et al. (2003). *Patient classification of fMRI activation maps. 6th annual international conference on medical image computing and computer assisted intervention—MICCAI'03*. Montreal, Canada: Springer-Verlag.

Fotheringham, S., & Rogerson, P. (1994). *Spatial analysis and GIS*. Taylor and Francis.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping, 2*, 189–210.

Fujimura, K., Toriya, H., Yamaguchi, K., & Kunii, T. L. (1983). Oct-tree algorithms for solid modeling. In T. L. Kunii (Ed.), *Computer graphics—theory and applications*. Springer Verlag.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Gaede, V., & Gunter, O. (1998). Multidimensional access methods. *ACM Computing Surveys, 30*, 170–231.

Gallant, S. I. (1990). Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks, 1*, 179–191.

Gerring, J. P., Brady, K. D., Chen, A., Vasa, R., Grados, M., Bandeen-Roche, K. J., et al. (1998). Premorbid prevalence of ADHD and development of secondary ADHD after closed head injury. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*, 647–654.

Guting, R. H. (1994). An introduction to spatial database systems. *VLDB Journal, 3*, 357–400.

Hagan, M., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks, 5*, 989–993.

Han, J., Chen, M.-S., & Yu, P. S. (1996). Data mining: an overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*, 866–883.

Haykin, S. (1999). *Neural networks, A comprehensive foundation*. Prentice Hall.

Kaufman, A., Cohen, D., & Yagel, R. (1993). Volume graphics. *IEEE Computer*.

Keim, D. A. (1999). Geometry-based Similarity Search of 3D Spatial Databases. *ACM SIGMOD Conference*.

Kontos, D., & Megalooikonomou, V. (2005). Fast and effective characterization for classification and similarity searches of 2D and 3D spatial region data. *Pattern Recognition, 38*, 1831–1846.

Kontos, D., Megalooikonomou, V., Prokrajac, D., Lazarevic, A., Obradovic, Z., Ford, J., et al. (2004). *Extraction of discriminative functional MRI activation patterns and an application to Alzheimer's Disease. 7th international conference on medical image computing and computer assisted intervention—MICCAI*. Rennes-St.Malo, France: Springer-Verlag.

Koperski, K., & Han, J. (1995). *Discovery of spatial association rules in geographic information databases. 4th international symposium on large spatial databases (SSD'95)*. Portland, Maine.

Kriegel, H.-P., & Seidl, T. (1998). Approximation-based similarity search for 3-D surface segments. *Geoinformatica, 2*, 113–147.

Lazarevic, A., & Obradovic, Z. (2002). Knowledge discovery in multiple spatial databases. *Journal of Neural Computing and Applications, 10*, 339–350.

Lazarevic, A., Pokrajac, D., Megalooikonomou, V., & Obradovic, Z. (2001). *Distinguishing among 3-D distributions for brain image data classification. 4th international conference on neural networks and expert systems in medicine and healthcare*. Milos Island, Greece.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*, 129–137.

McLachan, G. J., & Krishnan, T. (1996). *The EM Algorithm and Extensions*. John Wiley & Sons.

Megalooikonomou, V., Davatzikos, C., & Herskovits, E. H. (2000a). A simulator for evaluating methods for the detection of lesion-deficit associations. *Human Brain Mapping, 10*, 61–73.

Megalooikonomou, V., Ford, J., Shen, L., Makedon, F., & Saykin, A. (2000b). Data mining in brain imaging. *Statistical Methods in Medical Research, 9*, 359–394.

Megalooikonomou, V. (2002). *Evaluating the performance of association mining methods in 3-D medical image databases. 2nd SIAM International Conference on Data Mining (SDM)*. Arlington, VA.

Megalooikonomou, V., Pokrajac, D., Lazarevic, A., & Obradovic, Z. (2002). *Effective classification of 3-D image data using partitioning methods. SPIE 14th annual symposium in electronic imaging: Conference on visualization and data analysis.* San Jose, CA.

Mitchell, T. (1997). *Machine learning*. Boston: McGraw-Hill.

Ng, R. T., & Han, J. (1994). *Efficient and effective clustering methods for spatial data mining. VLDB Conference*. New York City, NY.

Pokrajac, D., Lazarevic, A., Megalooikonomou, V., & Obradovic, Z. (2001). *Classification of brain image data using measures of distributional distance. 7th Annual meeting of the organization for human brain mapping (OHBM01)*. Brighton, UK.

Pokrajac, D., Megalooikonomou, V., Lazarevic, A., Kontos, D., & Obradovic, Z. (2005). Applying spatial distribution analysis techniques to classification of 3D medical images. *Artificial Intelligence in Medicine, 33*, 261–280.

Quinlan, J.R. (1993). C4.5: Programs for machine learning, Morgan Kaufmann, Los Altos. Palo Alto: San Francisco.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference on Neural Networks*.

Samet, H. (1984). The quadtree and related hierachical data structure. *ACM Computing Surveys, 16*, 187–260.

Saykin, A. J., Flashman, L. A., Frutiger, S. A., Johnson, S. C., Mamourian, A. C., Moritz, C. H., et al. (1999). Neuroanatomic substrates of semantic memory impairment in Alzheimer's disease: patterns of functional MRI activation. *Journal of the International Neuropsychological Society, 5*, 377–392.

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (1998). *WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB Conference*. New York City, NY.

Smeulders, A. W. M., Worring, M., Santint, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 1349–1380.

Son, E.-J., Kang, I.-S., Kim, T.-W., & Li, K.-J. (1998). *A spatial data mining method by clustering analysis. The 6th International Symposium on Advances in Geographic Information Systems, GIS'98*.

Wang, W., Yang, J., & Muntz, R. (1997). *STING: A statistical information grid approach to spatial data mining. The 23rd international conference on very large data bases*.

Worth, A., Makris, N., Caviness, V., & Kennedy, D. (1997). Neuroanatomical segmentation in MRI: Technological objectives. *International Journal of Pattern Recognition and Artificial Intelligence, 11*, 1161–1187.