Effective Classification of 3D Image Data using Partitioning Methods

Vasileios Megalooikonomou*, Dragoljub Pokrajac, Aleksandar Lazarevic, Zoran Obradovic

Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 N. Broad St., Philadelphia, PA 19122

ABSTRACT

We propose partitioning-based methods to facilitate the classification of 3-D binary image data sets of regions of interest (ROIs) with highly non-uniform distributions. The first method is based on recursive dynamic partitioning of a 3-D volume into a number of 3-D hyper-rectangles. For each hyper-rectangle, we consider, as a potential attribute, the number of voxels (volume elements) that belong to ROIs. A hyper-rectangle is partitioned only if the corresponding attribute does not have high discriminative power, determined by statistical tests, but it is still sufficiently large for further splitting. The final discriminative hyper-rectangles form new attributes that are further employed in neural network classification models. The second method is based on maximum likelihood employing non-spatial (k-means) and spatial DBSCAN clustering algorithms to estimate the parameters of the underlying distributions. The proposed methods were experimentally evaluated on mixtures of Gaussian distributions, on realistic lesion-deficit data generated by a simulator conforming to a clinical study, and on synthetic fractal data. Both proposed methods have provided good classification on Gaussian mixtures and on realistic data. However, the experimental results on fractal data indicated that the clustering-based methods were only slightly better than random guess, while the recursive partitioning provided significantly better classification accuracy.

Keywords: 3D image, classification, clustering, maximum likelihood, dynamic recursive partitioning

1. Introduction

A lot of research has been done in the field of content-based retrieval and classification for general types of images (see [1, 2] for comparative surveys). In most cases the extracted features (usually color-based [3-5]) characterize the entire image rather than image regions and there is no distinction between important and unimportant features or between multiple objects in an image. In certain cases these features do not seem to be useful. Rather, characterization of an image on the basis of only those regions that are of interest to an expert seems to be more meaningful [6-8]. The 3-D 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 3-D structures (or objects). We focus on 3-D volumes that are binary, i.e., only information about the presence or absence of a particular voxel in a certain region is available. Examples of such binary volumes are regions of interest (ROIs) in medical images, i.e., regions that differ from the norm, e.g. due to the presence of lesions, tumors, or gene expressions, etc. Focusing on the ROIs is very important for the characterization and classification of images.

Necessary pre-processing steps prior to any data analysis of region data are the segmentation and registration procedures of the 3-D volumes. 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. In the medical imaging domain extensive image segmentation work has been done. Proposed methods can be divided into two broad groups: those that incorporate prior spatial information and those that are solely signal-intensity based (see [9-11] for review). Image registration deals with the existing morphological variability among samples and is required 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. It is done using normalization to a particular template and it is used to determine whether two samples have ROIs in the same location. The methods used for image segmentation and registration are often domain specific. In the following we assume that the region data have already been segmented and normalized.

* vasilis@ist.temple.edu; phone: 1- 215-204-5774; fax: 1-215-204-5082; www.cis.temple.edu/~vasilis; Center for Information Science and Technology, 303 Wachman Hall (038-24), 1805 N. Broad St., Philadelphia, PA 19122

The problem we focus on is the following: Given a set of region data and an assignment of these data to a number of classes based on certain non-spatial attributes, derive a classification scheme that will correctly classify a new sample of region data (predicting this way the non-spatial data) based only on spatial information. An example from the medical imaging domain is the following: Given a magnetic resonance (MR) image of a new subject that contains lesions, the goal is to 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). In this case the image data have resulted from scanning of a patient at multiple layers and then combining the images into a voxel-based 3-D representation.

Here we are proposing methods for the automatic classification of ROIs and quantitative measurement of their levels of similarity. After the review of related work in Section 2, in Section 3 we introduce the proposed methodology based on clustering and dynamic recursive partitioning, followed by a survey of experimental results on data with various complexities in Section 4 and conclusive remarks in Section 5.

2. Background and Related Work

Statistical distance based methods are very often used for distinguishing among distributions. A new sample s is predicted to belong to the class that corresponds to one of the datasets S_Y or S_N (corresponding to two classes, "Y" and "N") that is closer (in terms of some "distance") to s.

To compute the distance, the Mahalanobis distance [12] and the Kullback-Leibler (KL) divergence [13] are most often employed. Given data s corresponding to a new sample, the Mahalanobis distance between the new sample s and an existing data set S (S_Y or S_N) is computed as:

$$d_M = \sqrt{(\boldsymbol{\mu}_s - \boldsymbol{\mu}_S)^T \cdot \Sigma^{-1} \cdot (\boldsymbol{\mu}_s - \boldsymbol{\mu}_S)}$$

Here, $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_S$ are mean vectors of the data sets s and S respectively, and Σ is the pooled sample covariance matrix [14]. The KL distance for distinguishing between the new sample s and an existing data set S (S_Y or S_N) is defined as relative entropy between corresponding distributions:

$$d_{KL}(s,S) = \int_D p_s(\mathbf{x}) \ln \frac{p_s(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

where $p_s(\boldsymbol{x})$ and $p(\boldsymbol{x})$ are probability densities corresponding to the distributions from which data sets s and S are drawn, respectively. In this paper, we use a discrete approximation of the Kullback-Leibler distance on histograms estimated using a technique described in [15].

Static partitioning [15] may also be used for distinguishing among distributions. This static partitioning method first partitions the volume into a prespecified number of 3-D hyper-rectangles. The number of voxels at region data inside the small 3-D hyper-rectangles are averaged over the total number of voxels at region data inside the whole 3-D domain and treated as new attributes for training a classification model. The main problem in this approach is to determine the splitting resolution (the optimal size of hyper-rectangles), an one of possible solutions is to gradually increase the number of hyper-rectangles until a satisfactory classification accuracy is achieved.

Uniform histograms [16] are standardized methods for non-parametric modeling of probability distributions. Recently, importance-sampling method [17] has been developed to enhance estimation of highly non-uniform distributions. Using this technique, the histograms are estimated through a recursive procedure where in each step a hyper-rectangle with the highest frequency of discrete objects is subsequently partitioned. Importance sampling employs oct-trees (e.g. [18,19]) to maintain the spatial structure of hyper-rectangles and priority queues [20] for ordering the hyper-rectangles according to objects frequencies.

3. Methodology

Without loss of generality we assume that the dataset consist of a number of samples assigned to one of two classes (in this paper, we will denote the classes "Yes" and "No"). For each sample, the data contains a finite number of region data specified by their coordinates and positioned in the same 3-dimensional domain—a hyper-rectangle. Given two data sets S_Y and S_N containing coordinates of region data that belong to samples (patterns) from two classes, the objective is to determine the class of a new sample specified by a corresponding set s of region data, based on a classification model

trained on provided data. In this paper, we propose two techniques for such a classification. The first is based on partitioning using clustering algorithms and use of maximum likelihood. The second is a novel method based on dynamic recursive partitioning.

3.1. Clustering-based Partitioning for Maximum Likelihood Methods

In the maximum likelihood method [13,21], the underlying distributions corresponding to each class are estimated and a new sample is classified according to the likelihood that it belongs to each of the distributions. Formally $c = \max_i p_i(s)$, where $p_i(s)$ is estimated probability that a new sample s belongs to a class i. To apply maximum likelihood methods, it is necessary to estimate the probability density of data distributions that correspond to each class. Distributions can be estimated using parametric or non-parametric methods [16]. In this study we apply the non-spatial (k-means) [22] and the spatial DBSCAN clustering algorithm [23] to estimate the distribution parameters. Consequently, the obtained clusters are employed to estimate covariance matrices and priors of Gausian mixture components.

The standard k-means algorithm [22] is a variant of expectation-maximization (EM) method aimed to determine the means of Gausian mixture components. Through an iterative procedure, a dataset S containing n vectors x_i is partitioned into k clusters with a goal to find cluster means $\{m_j\}_{j=1}^{k}$ that minimize the average Euclidean distance $\frac{1}{n}\sum_{x_i \in X} \min_j d_E^2(x_i, m_j)$, where $d_E = \sqrt{(x_i - m_j)^T (x_i - m_j)}$.

The second considered clustering algorithm, DBSCAN, relies on a density-based notion of clusters and was designed to discover clusters of an arbitrary shape [23]. The key idea of a density-based cluster is that for each point of a cluster its *Eps*-neighborhood (for a given *Eps* > 0) should contain at least a minimum number of points (*MinPts*), (i.e. the density in the *Eps*-neighborhood of points has to exceed some threshold), since the typical density of points inside clusters is considerably higher than outside of clusters. Unlike the cluster centroids in the k-means, here the centers of the clusters can be outside of the clusters due to their arbitrary shapes.

3.2. Dynamic Recursive Partitioning

When performing classification of 3-D binary image data sets of region data with highly non-uniform distributions that can not be distinguished very well, standard statistical, static partitioning and clustering-based maximum likelihood methods may not achieve satisfactory performance. Therefore, in order to facilitate the classification, we propose the dynamic recursive partitioning (DRP) technique. It is aimed to determine a proper set of attributes corresponding to spatial sub-regions of the considered 3D domain and consists of the following three steps:

- generation of candidate attributes,
- attribute selection, and
- learning classification models.

At the beginning of the generation of candidate attributes, the voxels belonging to region data (ROI) within the global hyper-rectangle D are counted for each sample, and this number becomes the first candidate attribute. A hyper-rectangle is partitioned only if the corresponding attribute does not have a sufficient discriminative power to determine the classes of samples. The procedure continues recursively and stops when all remaining hyper-rectangles are discriminative or when there is an insufficient number of voxels of region data inside a hyper-rectangle.

For attributes generation we use oct-trees [18, 19] augmented to satisfy requirements for an efficient data representation and manipulation. Here, each tree node stores the boundaries of a corresponding 3D hyper-rectangle and the number of voxels at region data in the hyper-rectangle corresponding to each sample. In addition, to accomplish an efficient access to candidate attributes, we maintain a dynamic array [20], containing pointers to the leaf nodes of the tree.

The outline of the DYNAMIC RECURSIVE PARTITIONING (DPR) algorithm is shown in Figure 1. Each recursion call results in further splitting of a spatial sub-domain, represented by a node of the oct-tree, and it is performed if and only if a suitable splitting criterion is satisfied. A generic splitting criterion is to continue with partitioning if the average number of voxels at region data in a sub-domain is larger than a pre-specified threshold but the corresponding candidate attribute is not discriminative in distinguishing the sample classes. Observe that unlike techniques proposed in [24] here

significance tests are typically performed on larger spatial regions instead of on each voxel reducing the multiple comparison problem.

One of the simplest significance criteria for splitting is based on the computation of the Pearson correlation coefficient [25] between the class label (considered as a binary numeric value) and the attribute value for each sample. Here, an attribute is considered significant if the correlation coefficient is larger than a pre-determined threshold. Although this technique may be useful in practice, the major difficulty is the absence of a formal procedure to establish a threshold value.

Another criterion is based on discretization of the candidate attribute and evaluation of the class/attribute contingency matrix (see Figure 2b and 2c) using statistical tests (chi-square or the Fisher exact test [26]) with pre-determined maximal type I errors. A suitable value for the discretization threshold can be set ad-hoc or by using discretization techniques that maximize class/attribute mutual information [27].

Finally, the significance of a candidate attribute can be assessed by deciding whether the distributions of attribute values corresponding to the two classes differ substantially (see Figure 2d). To determine this, both parametric tests (t-test [25]) and non-parametric tests (Wilcoxon rank sum [28]) can be applied, with confidence pre-set to a specified value (usually 0.05 or 0.01).

If the splitting criterion is satisfied, the spatial sub-domain corresponding to the node of the oct-tree is partitioned into 8 smaller sub-domains. The tree node itself becomes the parent of eight children nodes, each corresponding to a smaller sub-domain (see Figure 3), and the number of voxels of region data in a subdomain per each sample becomes a new candidate attribute.

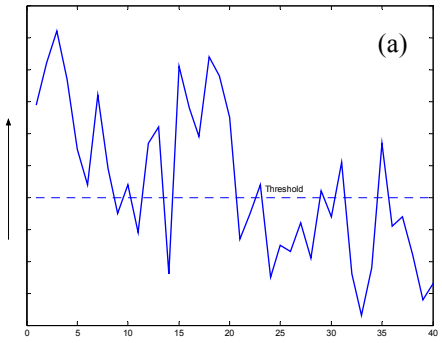
Given: Oct-tree T corresponding to the spatial domain D; Two sets $S_Y = \{S_{1,Y}, \ldots, S_{n_Y,Y}\}$, $S_N = \{S_{1,N}, \ldots, S_{n_N,N}\}$ containing region data for samples belonging to classes Y and N, respectively.

```
DYNAMIC RECURSIVE PARTITIONING (T,node, S_Y, S_N)
If SPLITTING_CRITERION(T,node, S_Y, S_N)=='yes'
        T=SPLIT(T,node)
        for node_c in CHILDREN (T,node)
                T=DYNAMIC RECURSIVE PARTITIONING (T,node_c, S_Y, S_N)
Else
        ADD_TO_LEAF_LIST (node)
Return T
```

Figure 1. A generic procedure for dynamic recursive partitioning (DRP).

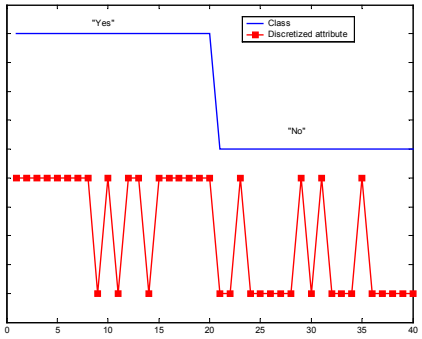
Techniques for attribute selection are employed to eliminate irrelevant and highly correlated attributes and to reduce the total number of attributes. To perform these techniques, it is desirable to have a quick access to candidate attributes, which is accomplished by maintaining dynamic array pointing to the leaves of the oct-tree. The applied classification-based selection algorithms involved inter-class and probabilistic selection criteria using Euclidean and Mahalanobis distance [12]. In addition to sequential backward and forward searches, the branch and bound search can be applied for iterative reduction of the attribute set.

For classification model trained on selected attributes, in this paper we propose the application of neural networks, universal approximators that were often reported to outperform the alternatives for classification of real life non-linear phenomena [29]. In addition, other classification techniques, including decision trees [30] can also be applied. We trained feedforward sigmoidal neural network classification models [29] with one hidden layer having the number of neurons equal to the number of input attributes. The neural network classification models had the number of output nodes equal to the number of classes, where the predicted class was decided according to the *winner-take-all* principle (class corresponds to the output with the larger response). The Levenberg-Marquardt [31] learning algorithm is applied to estimate coefficients of the neural network.

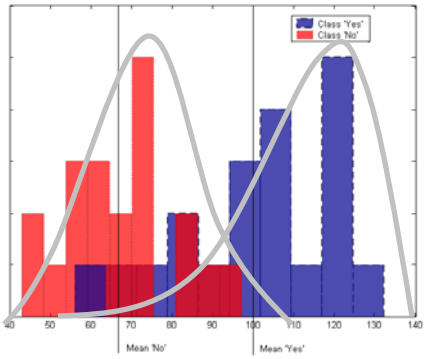


		True class	
		Yes	No
Discretized attribute	"high"	17	4
	"low"	3	16

(c)

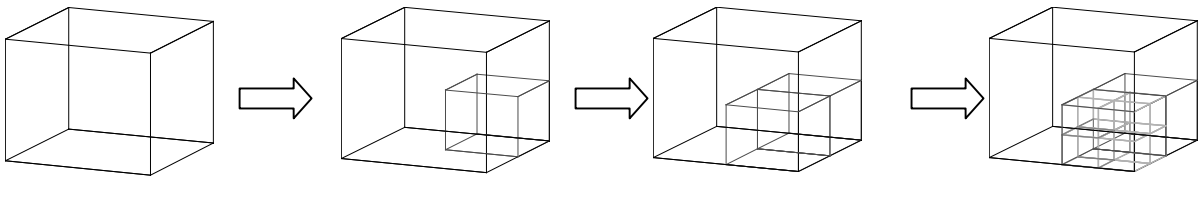


(b)

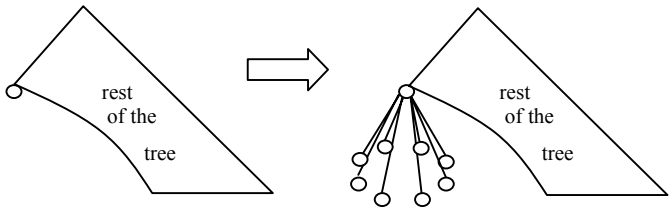


(d)

Figure 2. a) The values of a candidate attribute for a synthetic case of 40 samples. b) The true sample's class ('Yes' and 'No') and the discretized attribute using the threshold from a). c) True class/discretized attribute contingency table. d) Histograms and estimated distributions of the values of a candidate attribute that correspond to each class.



a)



b)

Figure 3. a) Illustration of partitioning of initial domain into spatial sub-domains. (b) Splitting of an oct-tree node.

4. Experimental Results

The proposed methods were experimentally evaluated on three data sets. The first data set contained synthetic data representing two mixtures of nine Gaussian distributions. The second data set represented realistic data generated using a lesion-deficit simulator with a spatial statistical model conforming to the Frontal Lobe Injury in Childhood study [32] where the subjects were classified into two classes according to subsequent development of ADHD (attention deficit and hyperactivity disorder) after a closed head injury. Finally, the third data set corresponded to highly heterogeneous fractal data designed to mimic situations where different subregions of a 3D image have diverse discriminative power.

4.1. Experiments with Synthetic Data

Synthetic data used in our experiments contained samples from two mixtures of nine normal distributions. We were varying the parameters (means and variances) of mixture components, thus constructing different mixtures of distributions (see Figure 4).

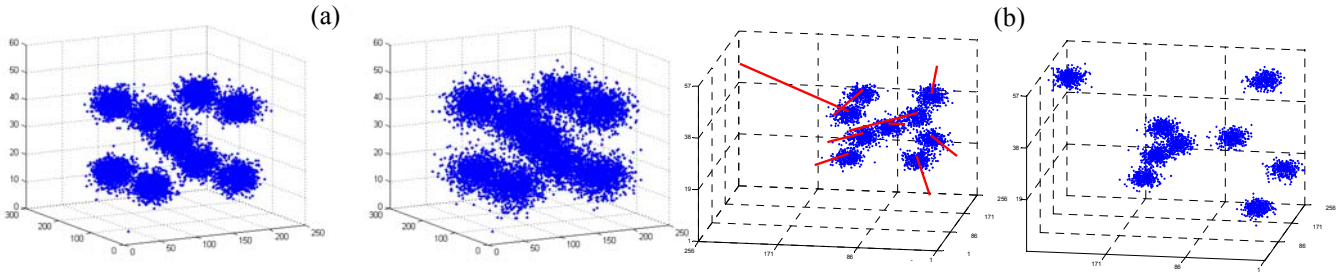
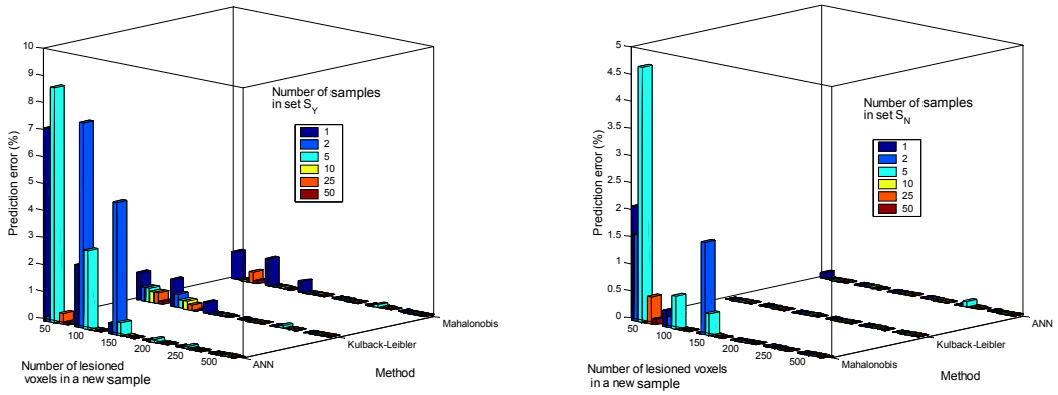


Figure 4. Two mixtures of distributions that differ only in (a) variance of the distribution components; (b) means of components

In the first series of experiments, the distribution components had the same variances but different means for each class (Figure 4a). We have repeated the experiments through 200 rounds, and each round consisted of random drawing of a new sample from one of the classes. The classification performance was monitored by measuring accuracy rate. The rate was computed as the ratio of the number of rounds when the classification of a new sample was successful and the total number of rounds. The samples contained a number of voxels of region data that varied from 50 to 500.

When using the Mahalanobis distance, we were able to adequately classify a new set of samples that belonged to one of two mixtures in 90% to 99% of cases, depending on the size of sets S_Y, S_N and the number of voxels of region data in a new sample (Figure 5). Analyzing the charts from Figure 5, it can be noticed that the prediction error of considered classification methods decreased when the size of sets S_Y, S_N increased and when the number of region data increased too.



a) Samples who belong to the first distribution b) Samples who belong to the second distribution

Figure 5. The prediction error when classifying new samples from two distributions with different means using Mahalanobis distance. Distribution means difference was 0.6.

Unlike the method using the Mahalanobis distance, the method using the Kullback-Leibler (KL) distance and static partitioning methods achieved almost perfect classification, for all considered sizes of data sets and numbers of voxels of region data in a new sample (the prediction error was less then 2%) (see Figure 5).

Another group of experiments on synthetic data involved mixtures that had the same component means but different component variances for each class (See Figure 4b). In this case, classification was typically more challenging since the mixture of distributions with smaller variances is often overshadowed by the mixture with larger variances. When using Mahalanobis distance in this scenario, the achieved classification accuracy was very low when predicting the mixture with smaller variances (from 0% to 50%), and significantly higher when predicting the mixture with larger variances (from 50% to 99%) [15].

The method based on computing the KL distance was more successful in predicting new samples when they belonged to the distribution with larger variance. When predicting the mixture of distributions with smaller variance, the accuracy varied from 14% for the small size of the set S_Y to 99% for the larger sizes of set S_Y [15], which is much better than using Mahalanobis distance. When predicting the mixture of distributions with larger variances, the method with KL distance was able to perform almost-perfect classification in all cases (error less than 1%).

When predicting the mixture of distributions with smaller variance using static partitioning methods, the accuracy varied from 55% to 99% with sizes of set S_Y (the larger size, the better accuracy) [15]. When predicting the mixture of distributions with larger variances, the static partitioning methods persistently provided almost perfect classification (error less than 1%).

Experiments performed on the synthetic data when the distribution components had different means but same variances demonstrated that the maximum likelihood based parametric methods were also almost perfectly accurate (prediction error less than 1%) regardless of what clustering algorithm was used to estimate the underlying distributions (*k*-means or DBSCAN). On the other hand, when performing experiments on the synthetic data with mixtures having the same component means but different component variances, the maximum likelihood based parametric methods were not successful in classification (the distribution with smaller deviation had tendency to be miss-classified, see Figure 6). When predicting the mixture of distributions with smaller variance, the prediction error varied from 50% for the smaller data set S_Y and for the smaller number of voxels in a new sample to 2% for the larger size of data set S_Y and for the larger number of voxels in a new sample. However, when predicting the mixture of distributions with larger variances, the maximum likelihood parametric methods were more successful (Figure 6), since the prediction error was ranging from 12% for the smaller number of voxels in a new sample to the perfect classification (prediction error less than 1%) for the larger number of voxels in a new sample.

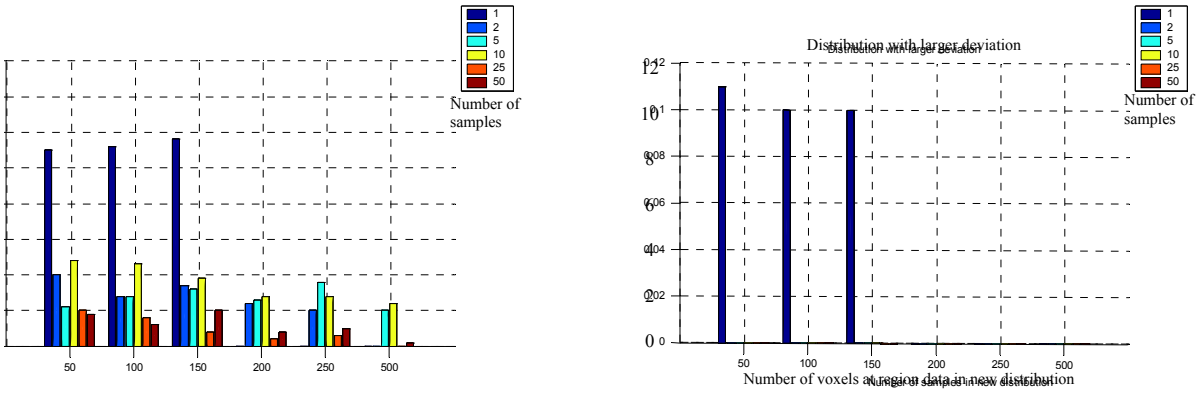


Figure 6. The prediction error (%) when classifying new samples from two distributions with different variances. The variances of distributions were 0.01 and 0.02. Maximum likelihood parametric classification was used.

Similarly to the maximum likelihood parametric based methods, dynamic recursive partitioning was able to classify new samples almost perfectly (prediction error less than 1%) when the experiments were performed on the synthetic data with the distribution components having different means but same variances. However, when predicting on the synthetic data with mixtures having the same component means but different component variances, the classification was not

always exquisite (Table 1). When the difference between the component variances was larger, the prediction was more successful, and vice versa.

Component Variances →		0.3	0.2	0.1	0.05	0.02
DRP	Prediction error (smaller variance)	0	0	1.4	6.4 – 2.5	21.7 – 10.3
	Prediction error (larger variance)	0	0	1.5	9.1 – 1.8	19.8 – 8.3
DBSCAN based maximum likelihood	Prediction error (smaller variance)	0	0	0	1	4
	Prediction error (larger variance)	0	0	0	0	1

Table 1. The prediction error (%) when predicting new samples from two distributions with different variances using clustering based maximum likelihood and DRP (dynamic recursive partitioning) methods. Variance of components in one mixture was 0.01 while the variances in the other mixture were varied.

Results presented in Table 1 were ranging since we employed different stopping criteria, different thresholds of the statistical tests and different structures of the neural networks. Table 1 also compares the prediction performance of the DRP method to the DBSCAN clustering based maximum likelihood method. It is apparent that for the smaller difference between component variances (0.02 and 0.05 for the mixture with larger variance), the maximum likelihood parametric methods achieved slightly better prediction accuracy than the DRP, since the former method naturally fits to Gaussian mixtures.

4.2. Experiments with Realistic Data

We performed classification of realistic brain lesion distributions that were generated using a lesion-deficit simulator [33] with the spatial statistical model conforming to the Frontal Lobe Injury in Childhood (FLIC) study [32]. 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 did not develop ADHD ("no ADHD" class) (Figure 7). Given a new subject with a set of region data, the goal was to determine the more plausible class. The subjects contained a number of voxels of region data that varied from 50 to 500, although in the specific FLIC study [32] approximately 200 voxels of region data are present on average per a 3-D brain image (i.e. per subject).

In experiments, we varied both the size of data sets for the classes and the number of voxels of region data belonging to a new subject. For each combination of these parameters, we performed the experiments through a predetermined number of 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 accuracy, as in Section 4.1 for synthetic data.

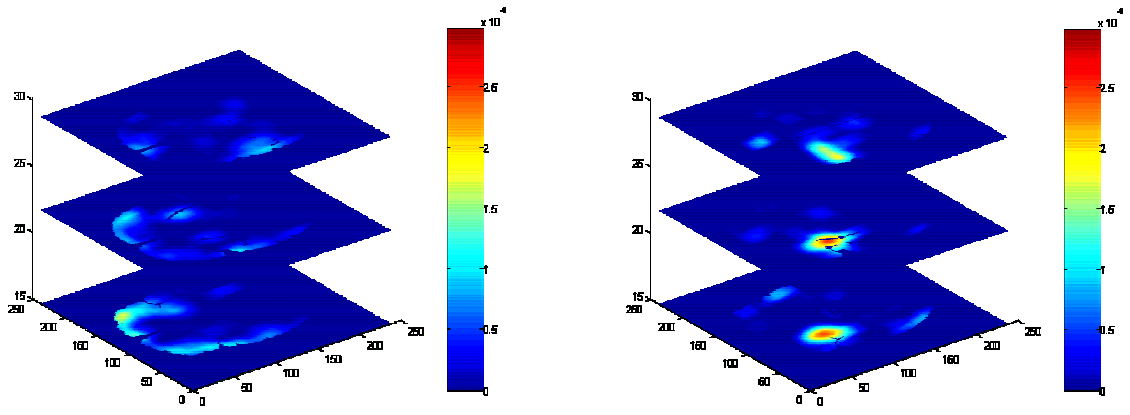
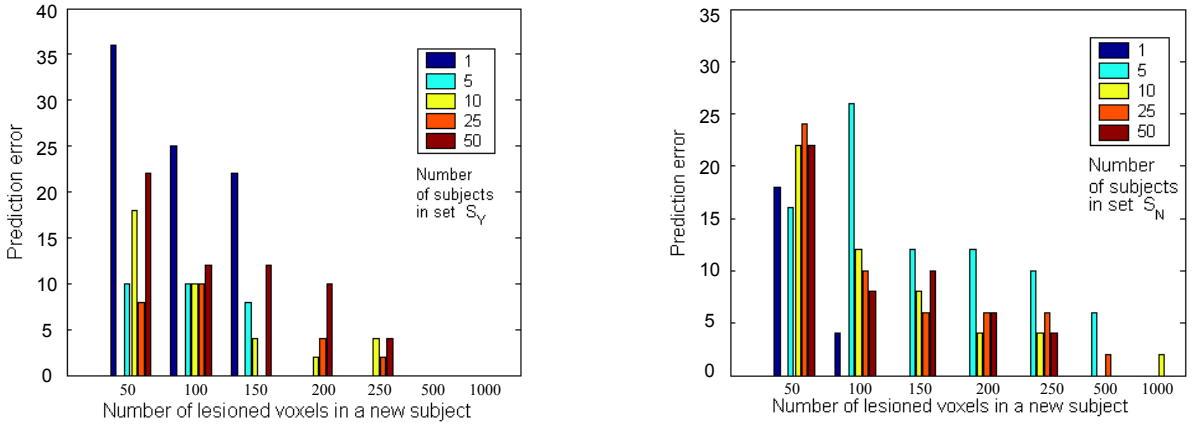


Figure 7. Distributions for "yes ADHD" and "no ADHD" classes.

Experiments on realistic brain lesion distributions showed that the proposed method based on Mahalanobis distance could provide more reliable and more accurate classification between the subjects regarding the development of ADHD

than when classifying samples from synthetic distributions. Figure 8 demonstrates that classification with error less than 10% was possible both for the subjects who did and who did not develop ADHD when a sufficient knowledge of the distribution corresponding to the subject was available (sets S_Y, S_N large enough). This was apparent especially when 150 or more voxels of region data were available for a new subject. The prediction was perfect (0% 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" (Figure 8).



a) Subjects who developed ADHD b) Subjects who did not develop ADHD

Figure 8. The prediction error (%) when classifying new subjects using Mahalanobis distance.

The method based on Kullback-Leibler (KL) distance was even more successful in classification of new subjects, especially when classifying subjects with a small number of voxels of region data (comparable to the size of data sets S_Y and S_N) [15]. When the number of voxels of region data per subject was 50 or 100, the prediction error was less than 2%, and when this number exceeded 200 the prediction was perfect in our experiments (0% prediction error). When using static partitioning method, the prediction error was always smaller than 5% for all explored combinations of the number of voxels of region data per subject.

When applying the maximum likelihood based parametric methods on realistic data, for all the combinations of the data sets size and the number of voxels of region data in a new subject, the achieved prediction error was less than 1% regardless of the clustering algorithm (k-means, DBSCAN) employed for estimating underlying distributions. The difference in prediction capability when employing *k*-means and DBSCAN clustering algorithm was insignificant (less than 1%), and the prediction was always almost perfect (error less than 1%). The number of clusters discovered using DBSCAN clustering algorithm was 3, and therefore the same number of clusters was used when employing *k*-means clustering algorithm in estimating the underlying distributions.

Finally, when applying dynamic recursive partitioning on realistic data, we were again able to achieve almost perfect classification (error less than 1%).

4.3. Experiments with Fractal Data

Experimental fractal data were generated based on a non-complete oct-tree that describes their structure, with a maximal depth $L_{\max}$ of the tree pre-specified. In the synthetic fractal data generation, we controlled the probability $p_l(L)$ that the node of the tree at level L is a leaf (L=0 corresponds to the root node, L=1 to the root's children, etc.). If the node is a leaf, the number of voxels of region data in the corresponding spatial sub-domain may be discriminative, i.e. useful to determine the class associated with the sample where the probability that a leaf node at level L is discriminative, was specified as $p_{dis}(L)$. The classification according to the number of voxels of region data corresponding to a discriminative leaf node may not be perfect, and therefore the classification error $p_{error}(L)$ is assumed to depend on the node depth as well. In each spatial subdomain corresponding to a leaf node, the region data were randomly displaced according to the uniform distribution, and the number of voxels of region data was n_N or n_Y, depending on the class to which the sample belongs and whether the node was discriminative or not.

The three probabilities $p_L(L)$, $p_{dis}(L)$ and $p_{error}(L)$ were modeled to have fractal-like exponential dependence on the node level L [34,35,36]. Further, classification error $p_{error}(L)$ and the probability $p_L(L)$ were assumed to decrease

with the level L, while the probability $p_{dis}(L)$ increased with L. By this construction, nodes with higher depths were less frequent in the tree, but with the increased depth, corresponding spatial subdomains were becoming more relevant for classification. Here, the probabilities were defined as $p_l(L)=1/2^{(3-d_l)L}$, $p_{dis}(L)=p_{dis}(L_{\max})/2^{(3-d_{dis})\cdot(L_{\max}-L)}$ and $p_{error}(L)=p_{error}(1)/2^{(3-d_{error})(L-1)}$, and data were generated with parameter values L_{max}=8, d_l=2.1, d_{dis}=2.5, d_{error}=1.5, $p_{dis}(L_{\max})$=20%, $p_{error}(1)$=50%, n_y=1 and n_x=0. An example of a resulting oct-tree is shown inFigure 10a.

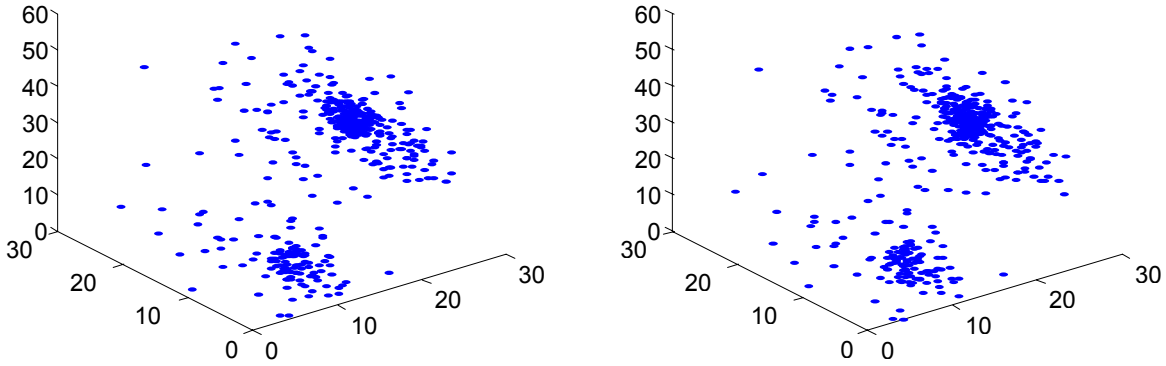


Figure 9. Distributions for "Yes" and "No" classes for fractal data.

We have simulated two classes: a "Yes" class and a "No" class, and we have generated 100 samples for each class (Figure 9). Observe that the resulting distributions were very similar and difficult for visual discrimination. The training set represented 50 random samples from the "Yes" class and 50 random samples from the "No" class. The test set represented the remaining 50 samples from "Yes" class and the remaining 50 samples from "No" class. For each sample from the test set, we computed Mahalanobis distances from the "Yes" and the "No" distributions and the obtained accuracy through 100 repeated experiments is reported at Table 2.

The standard statistical methods based on Kullback-Leibler (KL) distance had similar performance, although the prediction error was slightly larger (left half of Table 2). Maximum likelihood clustering based parametric methods on fractal data could not over-perform the methods based on standard statistical tests. The prediction error when the number of discovered clusters was nine is shown in the right half in Table 2.

Finally when DRP methods were applied to fractal data, we have achieved significantly better classification than using the previous methods, assuming that the splitting criterion was properly chosen (Table 3). As demonstrated in Figure 10, in such cases the discovered tree structure was similar to the original oct-tree employed to generate the fractal data. As we expected, the best results were obtained using the rank-sum non-parametric test, while in the absence of a proper threshold determination, the results using chi-square test were unsatisfactory.

5.Conclusions and Work in Progress

In this study, various methods to facilitate the classification of three-dimensional binary image data sets were considered. In addition to employing statistical distance-based techniques, we propose an alternative method based on partitioning by clustering combined with the maximum likelihood technique and a method based on dynamic recursive partitioning coupled with non-parametric classification algorithms.

The proposed methods were experimentally evaluated on three-dimensional binary data of various complexities, including mixtures of Gaussian distributions, realistic lesion-deficit medical data generated by a simulator conforming to a clinical study, and synthetic fractal data. All considered methods were shown to provide good classification on realistic data and on Gaussian mixtures when distributions associated with the two classes differed significantly. Methods based on Mahalonobis distance were inferior in comparison to other methods when the distribution components had the same or similar means. The experimental results on highly complex fractal data indicated the clear advantage of the recursive partitioning methods over the alternatives.

Work in progress involves comparison of the proposed and conventional classification techniques on clinical data. In particular, we are currently developing techniques that employ dynamic recursive partitioning to discover the discriminative brain regions for ADHD and other deficits. Also, we plan to thoroughly examine various splitting criteria as applied to the dynamic recursive partitioning. In addition, future work includes generalization of the proposed recursive partitioning technique to a multi-class case, the classification of multi-dimensional datasets, and classification of images where region data are not binary (i.e. present or absent on a specified location) but instead represent real-valued observation data. Finally, in the proposed dynamic partitioning technique, we did not exploit spatial relationship among constructed attributes, and we plan to consider this important aspect in our future research.

Prediction error (mean ± standard deviation)	Distance type		Clustering algorithm	
	Mahalanobis	Kullback-Leibler	k-means	DBSCAN
Classification Accuracy (%)	33.6 ± 2.5	43.4 ± 7.3	35.1 ± 11.8	30.8 ± 9.1

Table 2. The prediction error (%) on fractal data when classifying new samples using distance-based statistical techniques and clustering based maximum likelihood methods.

Splitting criterion			
Chi-square test (threshold=50)	Correlation (threshold-0.75)	t-test (significance 0.01)	Rank sum test (significance 0.05)
48.9 ± 12.1	20.3 ± 25.2	7.9 ± 9.1	5.3 ± 14.5

Table 3. The prediction error (%) (mean ± standard deviation) on fractal data when classifying new samples using DRP.

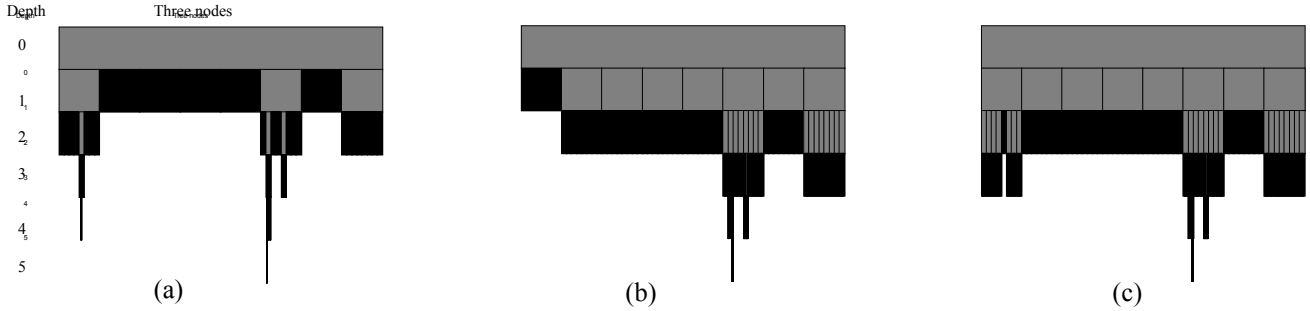


Figure 10. The structure of an oct-tree used to generate fractal data (a), and tree structures discovered using (b) the correlation criterion and (c) rank-sum test. Gray and black rectangles denote non-leaf and leaf nodes respectively. Children nodes are represented below its parent corresponding to a successive tree level (at the next depth).

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation under Grant No. ISI-0083423.

REFERENCES

1. M. De Marsicoi, L. Cinque, and S. Levialdi, "Indexing pictorial documents by their content: A survey of current techniques," *Image and Vision Computing*, **15**, pp. 119-141, 1997.
2. W. M. Smeulders, M. Worring, S. Santint, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, **22**, pp. 1349-1380, 2000.
3. R. Samadani, C. Han, and L. K. Katragadda, "Content-based event selection from satellite image of the aurora," In *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases*, pp. 50-59, 1993.
4. R. Pentland, W. Picard, and S. Sclaroff, "Photobook: tools for content-based manipulation of image databases," In *Proc. of the SPIE Conf. Storage and Retrieval of Image and Video Databases* II, San Jose, CA, 1994.
5. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, b. Dom, and M. Gorkani, "Query by image and video content: the QBIC system," *IEEE Computer*, **28**, pp. 23-32, 1995.
6. F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective similarity search in Medical Tumor Databases using Morphology," In *Proc. SPIE Conf.*, Boston, MA, 1996.

7. J. G. Dy, C. E. Brodley, A. C. Kak, C. Shyu, and L. S. Broderick, "The customized-queries approach to CBIR," In *Proc. IS&T/SPIE Electronic Imaging Conf.: Storage and Retrieval for Image and Video Databases* VII, 1999.
8. T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen, "Content-based image retrieval in medical applications: a novel multi-step approach," *Proceedings SPIE*, **3972**, pp. 312-320, 2000.
9. N. Pal, and S. Pal, "A review on image segmentation techniques," *Pattern Recognition*, **26**, pp. 1277-1294, 1993.
10. Y. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, **29**, pp. 1335-1346, 1996.
11. A. Worth, N. Makris, V. Caviness, and D. Kennedy, "Neuroanatomical segmentation in MRI: technological objectives," *Int'l J. Pattern Recognition and Artificial Intelligence*, **11**, pp. 1161-1187, 1997.
12. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, 1990.
13. R Duda,., P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, New York, 2000.
14. B. Flury, A First Course in Multivariate Statistics, Springer, New York, 1997.
15. Lazarevic, D. Pokrajac, V. Megalooikonomou, and Z. Obradovic, "Distinguishing among 3-D Distributions for brain image data classification", In *Proc. 4th Int'l Conf. Neural Networks and Expert Systems in Medicine and Healthcare, NNESMED 2001,* Milos Island, Greece, pp. 389-396, 2001.
16. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.
17. A. Lomax, and A. Curtis, "Fast, probabilistic earthquake location in 3D models using oct-tree importance sampling," *In Proc.* 26th *European Geophysical Society General Assembly*, Nice*,* 2001.
18. R.A. Finkel, and J.L. Bentley. "Quad trees: a data structure for retrieval on composite keys," *Acta Informatica*, **4**, pp. 1-9, 1974.
19. K. Fujimura, H. Toriya, K. Yamaguchi, and T. L. Kunii, "Oct-tree algorithms for solid modeling," in *Computer Graphics -Theory and Applications*-, T. L. Kunii ed., pp.96-110, Springer Verlag, 1983.
20. T. H. Cormen, C. E. Leadsperson, and R. L. Rivest, *Introduction to Algorithms*, 2nd edn., MIT Press, Cambridge, 2001.
21. T. Mitchell, *Machine Learning,* McGraw-Hill, Boston, 1997.
22. S. Lloyd, "Least-squares quantization in PCM," *IEEE Trans. Inf. Theory,* **IT-2**, pp. 129-137, 1982.
23. J. Sander, M. Ester, H-P Kriegel, and X. Xu, "Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications," *Data Mining and Knowledge Discovery,* **2**, pp. 169-194, 1998.
24. V. Megalooikonomou, C. Davatzikos, and E. H. Herskovits, "Mining lesion-deficit associations in a brain image database", in *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining,* San Diego, CA, pp. 347-351, 1999.
25. J.L. Devore, *Probability and Statistics for Engineering and the Sciences*, 5th edn., International Thomson Publishing Company, Belmont, 2000.
26. A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley, New York, 1996.
27. J. Ching, and A. Wong, "Class-dependent discretisation for inductive learning from continuous and mixed-mode data", *IEEE Trans. Pattern Analysis and Machine Inteligence*, **17**, pp. 641-651, 1995.
28. W.J. Conover*, Practical Nonparametric Statistics,* 3rd edn., Wiley, New York, 1999.
29. S. Haykin, *Neural Networks, A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, 1999.
30. J. Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1983.
31. M. Hagan, and M.B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Networks*, **5**, pp. 989-993, 1994.
32. J.P. Gerring, K.D. Brady, A. Chen, R. Vasa, M. Grados, K. Bandeen-Roche, N. Bryan, and M.B. Denckla, "Premorbid prevalence of attention-deficit hyperactivity disorder an development of secondary attention-deficit hyperactivity disorder after closed-head injury," *J. Am. Acad. Children Adolescent Psychiatry*, **37,** pp. 647-654, 1998.
33. V. Megalooikonomou, C. Davatzikos, and E. Herskovits, "A simulator for evaluating methods for the detection of lesion-deficit associations," *Human Brain Mapping*, **10**, pp. 61-73, 2000.
34. B.B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman and Company, Oxford, 1977.
35. C. C. Barton, and P. R. La Pointe (ed), *Fractals in the Earth Sciences*, Plenum Press, New York, 1995.
36. C. Traina Jr., A. Traina, L. Wu, and C. Faloutsos, "Fast feature selection using the fractal dimension," In *Proc. XV Brazilian Symposium on Databases (SBBD)*, 2000; also on http://citeseer.nj.nec.com/jr00fast.html.