An MCMC Feature Selection Technique for Characterizing and Classifying Spatial Region Data

Despina Kontos[1], Vasileios Megalooikonomou[1], Marc J. Sobel[2], and Qiang Wang[1]

[1] Department of Computer and Information Sciences
Temple University, 1805 N.Broad St., Philadelphia, PA 19122, USA
{dkontos,vasilis,wang32}@temple.edu
[2] Department of Statistics, Temple University, 1810 N.13th Street
Philadelphia, PA 19122, USA
sobel@sbm.temple.edu

Abstract. We focus on characterizing spatial region data when distinct classes of structural patterns are present. We propose a novel statistical approach based on a supervised framework for reducing the dimensionality of the initial feature space, selecting the most discriminative features. The method employs the statistical techniques of Bootstrapping simulation, Bayesian Inference and Markov Chain Monte Carlo (MCMC), to indicate the most informative features, according to their discriminative power across the distinct classes of data. The technique assigns to each feature a weight proportional to its significance. We evaluate the proposed technique with classification experiments, using both synthetic and real datasets of 2D and 3D spatial ROIs and established classifiers (Neural Networks). Finally, we compare our method with other dimensionality reduction techniques.

1 Introduction

Feature selection is a very important process for analyzing patterns in spatial data. In certain application domains, such as in geography, meteorology or medicine, we seek to focus on specific Regions of Interest (ROIs) that occupy a small portion of the data and extract informative features [1]. Examples of such ROIs are areas with high levels of precipitation in meteorological maps and brain regions of high activity in fMRI[1] (see Figure 1). A well-known characterization technique is to map data using the extracted features into points in a *K*-dimensional (*K*-d) space [2].

When dealing with spatial patterns, shape is one of the main characteristics that needs to be represented. Several approaches have been used for this purpose [3]. To obtain the initial characterization vectors here, we use an approach initially presented in [4] that considers properties of internal value of ROIs in addition to their shape. This method works particularly well for non-homogeneous as well as for homogeneous ROIs. It efficiently forms a *K*-d feature vector using concentric spheres in 3D (or circles in 2D) radiating out of the ROI's center of mass and extracting quantitative information regarding both its structure and content. In several cases though, the number of features extracted is too large to support efficient pattern analysis and classification.

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

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 379–387, 2004.

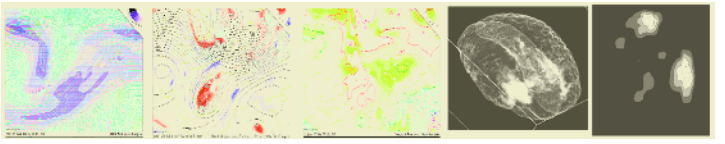

Fig. 1. Examples of geographical / meteorological and medical 2D and 3D spatial ROIs.

Several techniques have been proposed for reducing the dimensionality of data [5]. These approaches can be separated into two categories: (i) those having the property of transforming the initial features introducing a completely new subspace and, (ii) those that attempt to find an optimal subset of the initial features that are considered to be more significant.

Principal Component Analysis (PCA), also known as Singular Value Decomposition (SVD), is the most widely used technique from the first category due to its conceptual simplicity and efficient computation. It has been extensively used in many applications, such as medical image pattern analysis [6]. Multidimensional Scaling (MDS) is another dimensionality reduction technique with wide applicability [7]. The Discrete Fourier Transform (DFT) [8] and Wavelet Transform [9] have also been applied. Algorithms in the second class search for an optimal subset of the initial vector attributes, rather than a transformation. A well-known technique is forward feature selection that seeks to find an optimal subset of features [10]. Other approaches [11] combine the process of attribute selection with the induction algorithm used for classification. Statistical pattern recognition techniques have also been proposed [12]. Although proven effective, the first class of techniques fails to preserve the initial attribute values; the new feature vectors do not correspond to real data measurements. This introduces greater difficulty in interpreting the conceptual representation of the new feature space.

Our approach shares mostly characteristics of the second class of feature selection techniques and is based on a statistical framework that employs Bootstrapping, Bayesian inference and the Markov Chain Monte Carlo (MCMC) techniques. Our method applies to cases where distinct classes of data are present and a training set of labeled instances is available. In the particular case examined here, the level of the discriminatory significance of features varies across the classes of the observed data. These statistical techniques are used to select the most significant features, according to their discriminative power across the distinct classes of data, giving rise to a significant reduction in dimensionality.

2 Methodology

The general idea of the proposed feature selection technique is based on the assumption that the classes are generated by distinct structural pattern distributions reflected by the characterization vectors for each class. After learning a model/distribution for each class (in fact, a posterior over the models) using probabilistic modeling, bootstrapping and Bayesian inference, we find the features that are generated significantly differently under each model/class. We observe a training set T consisting of a number n_j of objects (ROIs) $O_{i,j}$, $i=\{1,\ldots, n_j\}$ of class j, $j=\{1,\ldots,M\}$. Each object is characterized by a feature vector of size K. That is, $O_{i,j}=(f_{i,j}[1],\ldots,f_{i,j}[K])$. We would

Table 1. Symbol Table.

$r_1,\ldots,r_K$	Radii
$O_{i,j}$	Object $i=1,\ldots,n_j$ of class $j=1,\ldots,M$
$f_{i,j}[k]$	Feature $k=1,\ldots,K$, of object $O_{i,j}$
$\Delta f_{i,j}=\left(\Delta f_{i,j}[2],...,\Delta f_{i,j}[K]\right)$	Observed probability vectors for object $O_{i,j}$
$\Delta f_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\Delta f_{i,j}$, $j=1,...,M$	Observed probability vector for objects of class j
$p_j=(p_j[2],\ldots,p_j[K])$, $j=1,...,M$	Parametric probability vector for objects of class j
$num_j=(num_j[2],\ldots,num_j[K])$, $j=1,...,M$	Combined bootstrap count vector for objects of class j
N_j, $j=1,...,M$	Combined bootstrap sample sizes for class j

like to use these features to determine appropriate ways to distinguish between different classes. In the presentation of the proposed feature selection technique we assume that these features are given by the characterization procedure [4] described briefly in Section 1, although with appropriate preprocessing any other characterization procedure can be used instead. In this case, feature $f_{i,j}[k]$ corresponds to the fraction of the object $O_{i,j}$ occupied by a sphere of radius r_k (the fraction of the sphere occupied by the object $O_{i,j}$ can be used as well). Let us consider consecutive features $f_{i,j}[k]$, $f_{i,j}[k\text{-}1]$ corresponding to radii r_k, r_{k-1} respectively. The difference between such features, calculated as a proportion of the total feature difference, is $\Delta f_{i,j}[k]=(f_{i,j}[k]-f_{i,j}[k-1])/(f_{i,j}[K]-f_{i,j}[1])$, where k=2,...,K. Note that after this normalization, the fractional proportions satisfy the relationship: $\sum_{k=2,..,K}\Delta f_{i,j}$ and the components of $\Delta f_{i,j}$ can be treated as probabilities. Let $\Delta f_{i,j}$ = ($\Delta f_{i,j}[2],\ldots,\Delta f_{i,j}[K]$) be the *observed probability vector* attached to object i of class j (i=1,…,n_j; j=1,…,M). We will consider these vectors to be the vectors representing (characterizing) the initial ROIs. We employ also the notation $\Delta f_j = {}^{1}\!/_{n_j}\sum_{i=1}^{n_j}\Delta f_{i,j}$, (j=1,.., M) for the *observed probability vector* for objects of class j.

We assume that the observed probability vectors Δf_j arising from class j are generated by a (parametric) probability vector: p_j=($p_j[2],\ldots,p_j[\mathrm{K}]$), j=1,…,M. These are the "true" apriori values for the observed probabilities atattached to objects of class j, capturing the spatial pattern of the corresponding class. The procedure for selecting the most discriminative features is as follows:

1. Bootstrapping is done by 'sampling' a large number, B, of instances from each observed probability vector $\Delta f_{i,j}$ of class j. Each sample consists of different features, selected according to their component probabilities, which are actually equal to the

feature values after the normalization step. Under the assumptions that (i) the feature vectors for different objects are mutually independent, and (ii) the observed probability vectors $\Delta f_{i,j}$ arising from each class are well characterized by parameters, the count vectors obtained from this sampling are easily combined to form an approximate probability distribution for each class. We use the notation $num_j[k]$ to denote the number of times feature f_k, k=1,…,K is chosen by the sample for class j, j=1,…,M.

2. We also employ the notation N_j; j=1,…,M for the total bootstrap sample size used for sampling from class j; j=1,…,M.

3. We assume that the components of the observed probability vectors $\Delta f_{i,j}$ (i=1,…,n_j; j=1,…,M) are mutually independent, conditional on the true probability vectors p_j=(p_j[2],…, p_j[K]), j=1,…,M i.e.,

$$P(\Delta f_j \mid p_j) \approx \prod_{i=1}^{n_j} P(\Delta f_{i,j} \mid p_j); \quad P(\Delta f \mid p) = \prod_{j=1}^{M} P(\Delta f_j \mid p_j) \tag{2.1}$$

We note that, although it is tempting to assume that the observed probability vectors have a multinomial (or multinomial-type) distribution in the parameters $p_j[k]$; k=2,...,K , j=1,…,M, this is not possible in view of the fact that the components of the observed probability vectors are not integers. An alternative approach, which we propose here, involves constructing an approximate multinomial likelihood for the observed probability vectors using the bootstrap counts, $num_j[k]$ (k=1,..,p; j=1,…,M) and basing inference on this likelihood. We make the assumption that the likelihood of the observed probability vectors takes this form. This assumption is tantamount to assuming that the bootstrap sample sizes N_j (j=1,..,M) are 'large enough' to have the property that the vector, N_j * Δ_j, has components all of which are positive integers. The multinomial probability density function (see Equation 2.1) in the bootstrap count takes the form:

$$P(num_j \mid p) = \prod_{k=1,..,K} \binom{N_j}{num_j[1]...num_j[K]} p_j[k]^{num_j[k]} \quad , j=1,..,\mathrm{M} \ .$$

$$P(num \mid p) = \prod_{j=1}^{M} P(num_j \mid p) \ , \quad j=1,..,\mathrm{M} \ . \tag{2.2}$$

3. We postulate a parameter λ_j with (apriori) mean value $1/N_j$ having the following property: the observed probability vectors $\Delta f_{1,j},.., \Delta f_{n_j,j}$ are mutually independent with a likelihood similar to that of λ_j*num_j (j=1,…,M). This is easily done by assuming an exponential prior with density λ_j*exp{-λ_j*N_j}. After inserting this parameter, the multinomial likelihood is transformed into:

$$P(num_j \mid p, \lambda_j) = \left(1 \Big/ \sum_{j=1}^{M} p_j^{\lambda_j}[k] \right)^{N_j} \prod_{k=1,..,K} \binom{N_j}{num_j[1]...num_j[K]} p_j[k]^{\lambda_j * num_j[k]} \ . \tag{2.3}$$

where j=1,..,M. We assume therefore that :

$$P(\Delta f_j \mid p, \lambda_j) \approx P(num_j \mid p, \lambda_j); \quad P(\Delta f \mid p) = \prod_{j=1}^{M} P(\Delta f_j \mid p) \quad . \tag{2.4}$$

The posterior distribution of the parametric probability vectors $p_1,..,p_M$ and scaling parameters, $\lambda_1,..,\lambda_M$ (calculated using the approximate likelihood (2.2)) is complicated in this case. We have evaluated it using Markov Chain Monte Carlo (MCMC). We note that the variability in estimates of the true probability values arises in Equation (2.4) from the variability in the (normalized) bootstrap counts. This variability decreases as the bootstrap sample sizes increase (by the law of large numbers). Equation (2.4) is used in MCMC simulations to update the p's. We also note that as the bootstrap sample size increases, the distribution of the (scaled) combined count statistics approach that of the mean observed probability attribute vectors, making our model asymptotically correct.

4. We distinguish the best features (corresponding to radii) by evaluating a measure of variation $Var[k]$ for the p's at each radius k;

$$Var[k] = \sum_{j=1}^{M} (p_j[k] - \overline{p}[k])^2; \quad \overline{p}[k] = 1/M \sum_{j=1}^{M} p_j[k] \quad . \tag{2.5}$$

We distinguish the best one over each posterior simulation by choosing that feature $f_{k'}$ corresponding to radius $r_{k'}$ having the property that:

$$Var[k'] = \max\{Var[k]; k = 1,.., K\} \quad . \tag{2.6}$$

The feature having the maximum variation over the greatest number of posterior simulations (calculated using MCMC) is deemed the best. This is justified by the fact that a high degree of variance for a specific attribute across the distinct labeled classes (inter-class variance) indicates a high degree of dissimilarity in the spatial pattern at the specific radius increment. Hence, the attribute can be considered to be highly informative with respect to class membership. Also, employing a large number of bootstrap samples B reduces high variance that might exist due to noise.

3 Experimental Results

All the experiments were implemented in Matlab using the Statistics v.3 toolbox of Mathworks. For classifiers we used Neural Networks implemented by the *PRTools v.3.1* toolbox for Matlab [13].

3.1 Artificial Data

We used artificial data sets that were generated using a parametric growth model for spatial ROIs introduced in [4]. The main idea is that the growth process begins with one initial voxel (or cell) at time t=0 and progresses using an "infection" procedure (see Figure 3), where each infected cell may infect its non-diagonal neighbors with some probability. The datasets are the following: (i) 2DHom: 100 2D Homogeneous ROIs, 50 spherical and 50 elongated to two opposite directions (north-south), with 14- feature characterization vectors (see Figure 2(b)-(c)), (ii) 2DNonHom: 100 2D Non-Homogeneous ROIs, 50 elongated to the one direction (north) and 50 elongated

to two opposite directions (north-south), with 14 - feature characterization vectors (see Figure 2(d)-(e)), (iii) 3DHom: 100 3D Homogeneous ROIs, 50 spherical and 50 elongated to two opposite directions (north-south) , with 7 - feature characterization vectors and (iv) 3DNonHom: 100 3D Non-Homogeneous ROIs, 50 spherical and 50 elongated to two opposite directions (north-south), with 14 - feature characterization vectors. Each dataset consists of two distinct labeled classes of spatial pattern.

(a) (b) (c) (d) (e)

Fig. 2. (a) a sample of the growth process, 2D samples of (b),(c) Homogeneous, and (d),(e) Non-Homogeneous ROIs used in our experiments.

Using these artificial datasets we run a set of basic experiments. We tested 4 different combinations of bootstrapping sample size *B* and number of MCMC posterior simulations. For each of the combinations, we discovered the most discriminative attributes for the ROIs of each set and assigned a weight to them in the range [0...1] that indicates their discriminative power. Figure 3 (a)-(d) illustrates these results. A basic observation from this first set of experiments is that, by increasing the number of bootstrap sample *B* the method discovers less attributes each with a more significant weight. On the other hand, reducing the number of bootstrap sample *B* and increasing the number of MCMC simulations tends to spread the weights to more attributes, with a less significant weight factor to each individual attribute.

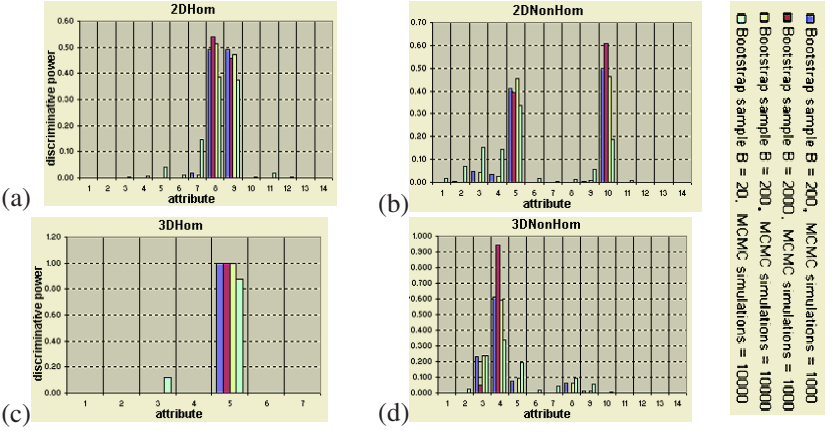

Fig. 3. Discriminative power of attributes discovered by the proposed method for (a) **2DHom**, (b) **2DNonHom**, (c) **3DHom** and (d) **3DNonHom** datasets and for different combinations of Bootstrap sample size B and number of MCMC simulations.

We continue with classification experiments using these selected discriminative features. The neural network consisted of one hidden layer with 5 neurons, number of inputs equal to the number of attributes used in each case, and one output indicating

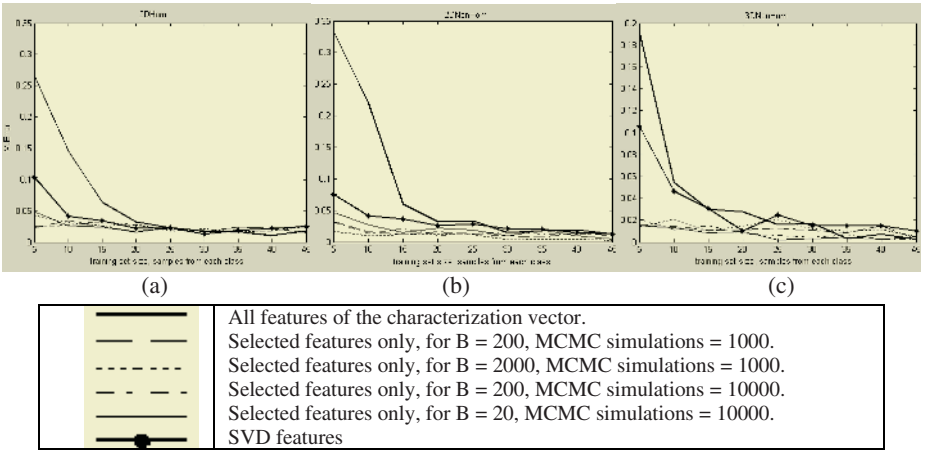

Fig. 4. Neural network classification performance (mean error) when using all features, features obtained by SVD and features obtained by the proposed method for (a) 2DHom, (b) 2DNonHom and (c) 3DNonHom datasets.

the class. The training was performed using the Levenberg-Marquardt optimization and the training set size ranged from 5 to 45 samples from each class (two classes of 50 ROIs each in every set). We report the curves of mean classification error after 40 repetitions for all different sizes of the training data set. We also include the comparative classification performance using (i) all the attributes of the $\Delta f_{i,j}$ characterization vectors and (ii) the first 5 (2D data) and 3 (3D data) most significant components of the SVD transformation applied on the $\Delta f_{i,j}$ characterization vectors. Figure 4 (a)-(c) shows the classification performance for the various cases. A first observation is that in any case the classification error is very small. The classification performance when using only the discriminative features selected by the proposed approach is, in almost all cases, comparable to (or better than) that of using all the features of the characterization vectors or those features obtained by SVD.

3.2 Real Data

We experimented using ROIs extracted from 3D fMRI brain activation contrast maps. The fMRI scans were obtained from a study designed to explore neuroanatomical correlates of semantic processing in Alzheimer's disease [14]. For the experiments pre sented here, we focused only on a specific region of the brain that has been shown to be highly associated with the development of Alzheimer's disease [15]. The dataset consisted of 9 control and 9 patient 3D ROIs and the characterization vectors were constructed using 40 features. Figure 5 shows this ROI in consecutive 2D slices of the 3D volume.

We applied the proposed feature selection technique for sample size $B = 400$ and number of MCMC posterior simulations = 1000. We perform classification experiments, using the discriminative features selected. To avoid overfitting due to a small training dataset (9 controls vs. 9 patient samples) we applied one-layer perceptron networks trained by the Pocket algorithm and leave-one-out cross validation. We

repeated the process of training and testing for 10 times and report the average accuracy. The first row of Table 2 includes the classification results when using all the discriminative features selected, as well as subsets of them based on the significance

Fig. 5. The ROI used for applying the proposed feature selection technique. It is shown in consecutive 2D slices after being overlaid on a canonical brain atlas.

Table 2. Classification performances for the fMRI ROIs.

	The best features (13) ($w_k > 0$)	The best 7 features ($w_k > 0.005$)	The best 5 features ($w_k > 0.02$)	The best 2 features ($w_k > 0.2$)	INITIAL All 40 features
MCMC	85.56 %	**84.44 %**	**87.22 %**	**82.78 %**	82.22 %
FFS	**86.67** %	**84.44** %	82.22 %	81.11 %	

weights, w_k, obtained by the proposed approach. The second row of Table 2 shows the comparative results when using the forward feature selection (FFS) approach. In all the experiments we used the initial characterization attribute values. The accuracy obtained when using all the initial 40 features is also reported. It is interesting to observe that the proposed MCMC approach behaves better than the forward feature selection technique (greedy approach), especially when using only the highly discriminative features ($w_k > 0.02$); it is comparable to the greedy approach in all other cases.

4 Conclusions

We presented a novel dimensionality reduction technique which employs the statistical framework of Bootstrapping simulation, Bayesian inference and Markov Chain Monte Carlo (MCMC). The method applies when labeled distinct classes of spatial ROIs are available, aiming to select the most informative features with respect to class membership. The proposed approach assigns a weight to each selected feature revealing its discriminative power. We experimented both with synthetic and real data performing classification experiments using both all the initial characterization attributes and only the selected ones by the proposed method). We compared the proposed approach with SVD and forward feature selection. We concluded, on the data we experimented with, that the proposed approach always outperforms SVD. Also, it is better than forward feature selection as the number of selected features is reduced, making it a better alternative over the greedy approach. Finally, the proposed technique was shown to be effective when applied on real data. In this case the proposed technique performed better than the forward feature selection approach, especially when using highly discriminative attributes, while being comparable in other cases.

Acknowledgement

The authors would like to thank A. Saykin for providing the fMRI dataset and clinical expertise and H. Dutta for working on preliminary experiments. This work was supported, in part, by NSF research grants IIS-0083423 and IIS-0237921 and by NIH Research Grant #1 R01 MH68066-01A1 funded by the National Institute of Mental Health and the National Institute of Neurological Disorders and Stroke and the National Institute on Aging.

References

1. Megalooikonomou, V., Ford, J., Shen, L., Makedon, F., Saykin, A.: Data mining in brain imaging, Statistical Methods in Medical Research, Vol. 9 (4) (2000) 359-394
2. Guting, R.H.: An Introduction to Spatial Database Systems, VLDB Journal 3 (4) (1994) 357-399
3. Loncaric, S.: A Survey of Shape Analysis Techniques. Pattern Recognition, Vol. 31 (8) (1998) 983-1001
4. Megalooikonomou, V., Dutta, H., Kontos, D.: Fast and Effective Characterization of 3D Region Data, In Proc. of the IEEE International Conference on Image Processing (ICIP) 2002, Rochester, NY (2002) 421-424
5. Carrerira-Peripinan, M.A.: A review of Dimension Reduction Techniques, Technical Report CS-96-09, Dept. of Computer Science, University of Sheffield (1997)
6. Petrakis, E.G.M., Faloutsos, C.: Similarity Searching in Medical Image DataBases, IEEE Trans. on Knowledge and Data Engineering, Vol. 9 (3) (1997) 435-447
7. Kruskal, J.B., Wish, M.: Multidimensional scaling, SAGE publications, Beverly Hills (1978)
8. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-series databases, In Proc. of the ACM SIGMOD Int.Conf. on Management of Data, Minneapolis, MN (1994) 419-429
9. Chan, K.P., Fu. A.C.: Efficient time series matching by wavelets, In Proc. of the Intl. Conference on Data Engineering ICDE , Sydney, Australia (1999) 126-133
10. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd Ed., Academic Press, New York (1990)
11. Kohavi, R., John, G.: Wrappers for Feature Subset Selection, Artificial Intelligence, Vol. 97 (1-2) (1997)273-324
12. Jain, A., Duin, P., Mao, J.: Statistical pattern recognition: A review, IEEE Transactions on PAMI, Vol. 22 (1) (2000) 4-37
13. Duin, R.P.W.: A Matlab Toolbox for Pattern Recognition, PRTools Version 3.0 (2000)
14. Saykin, A.J., Flashman, L.A., Frutiger, S.A., Johnson, S.C., Mamourian, A.C., Moritz, C.H., O'Jile, J.R., Riordan, H.J., Santulli, R.B., Smith, C.A., Weaver, J.B.: Neuroanatomic substrates of semantic memory impairment in Alzheimer's disease: Patterns of functional MRI activation, Journal of the International Neuropsychological Society, Vol. 5 (1999) 377-392
15. Megalooikonomou, V., Kontos, D., Pokrajac, D., Lazarevic, A., Obradovic, Z., Boyko, O., Saykin, A., Ford, J., Makedon, F.: Classification and Mining of Brain Image Data Using Adaptive Recursive Partitioning Methods: Application to Alzheimer Disease and Brain Activation Patterns, Human Brain Mapping Conference (OHBM'03), New York, NY (2003)