On Classifying Disease-Induced Patterns in the Brain Using Diffusion Tensor Images

Peng Wang[1,2] and Ragini Verma[1]

[1] Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia PA, 19104*
[2] Department of Integrated Data Systems, Siemens Corporate Research, Princeton NJ, 08540

Abstract. Diffusion tensor imaging (DTI) provides rich information about brain tissue structure especially in the white matter, which is known to be affected in several diseases like schizophrenia. Identifying patterns of brain changes induced by pathology is therefore crucial to clinical studies. However, the high dimensionality and complex structure of DTI make it difficult to apply conventional linear statistical and pattern classification methods to identify such patterns. In this paper, we present a novel framework that uses a combination of DTI-based anisotropy and geometry features to effectively identify brain regions with pathology-induced abnormality, and to classify brains into the diseased and healthy groups. Our method first directly estimates the underlying overlap between the patient and control groups, based on a semi-parametric Bayes error estimation method. By ranking voxels based on these estimation results, the method identifies abnormal brain regions from which features are extracted through Kernel Principal Component Analysis (KPCA) for subsequent classification. Application of the method to a dataset of controls and patients with schizophrenia, demonstrates promising accuracy of this framework in identifying brain patterns to separate two groups, and hence aiding in prognosis and treatment.

1 Introduction

Diffusion tensor imaging (DTI) [1] characterizes brain tissue structure based on underlying water diffusivity, providing effective and unique characterization of the white matter. It has therefore been extensively used for studying pathology of diseases, such as schizophrenia, where the white matter is known to have been affected [2]. In clinical research, it is important to analyze group differences between patients and controls, by identifying brain abnormality patterns induced by pathology, to provide insight into the progression of the disease. There are usually two types of analysis: voxel-based analysis, which identifies regions of difference between the two groups based on statistics per voxel in the brain or region-of-interest (ROI) [3], and pattern classification that uses the whole brain to classify the patients and controls. Although much research has been done on DTI visualization, segmentation, registration [4,5,6], and group difference analysis for clinical discovery [2], little work has been done in classifying brain patterns using DTI data [7]. In this paper, we present a framework for combining the two types of

* This work was done in Section of Biomedical Image Analysis, University of Pennsylvania.

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 908–916, 2008.

analysis on DTI data, by first identifying abnormal regions at a voxel level, and then incorporating them into a pattern classification framework to differentiate diseased from healthy brains.

The tensorial structure of a 3×3 positive-definite symmetric matrix at each voxel, combined with the high dimensionality of the brain data, make classification of such data challenging. Most of current voxel-based methods apply conventional linear statistical methods, such as t-test and Fisher discriminant analysis (FDA), to single scalar measures of anisotropy and/or diffusivity computed from tensors [2,7]. However, such single scalar measurements do not fully used the rich information in diffusion tensors. FDA usually assumes single Gaussian distribution for each class, and may fail on high dimensional complex DTI data, especially when only limited samples are available in most clinical studies. In [7], a shaving procedure, which iteratively discards small weights in FDA, has been applied to a single scalar measure of anisotropy to select a certain percentage of voxels for classifying brains induced by schizophrenia. The method provides a cross-validation accuracy only between 65% and 75% when using a single type of scalar features. Nonlinear classifiers, such as Support Vector Machine (SVM), have been used to effectively classify structural MRI data [8], but it has not been applied to a combination of DTI features, and its performance is linked with the choice of features different from DTI data.

In this paper, we present a novel framework for DTI-based brain pattern analysis. Our method can identify brain regions where disease induces abnormality of brain tissue, and can combine information from multiple measures extracted from tensors to classify diseased and healthy brains. The framework includes three major parts: ranking voxels based on their discriminant capability, selecting voxels and regions from which features are extracted for pattern analysis, and classifying brains based on the extracted features. First, rooted in the pattern classification theory, our method estimates the underlying overlap between different groups using a semi-parametric Bayes error estimation. Unlike the linear discriminant analysis or t-test, this method does not assume global Gaussian distribution for each class, but only considers data distribution to be locally smooth, thus being able to handle the complex structure in DTI data. Also multiple measures of anisotropy and diffusivity that are extracted from diffusion tensors are incorporated into the Bayes error estimation. Second, individual voxels are ranked according to their estimated Bayes error, and are then grouped into local regions to identify brain patterns induced by disease. Finally, multiple scalar measures are combined through Kernel Principal Component Analysis (KPCA) in local regions, and are input to nonlinear classifiers for region-based group classification. The presented methods are validated on real data sets, and experimental results demonstrate encouraging classification results.

2 Methods

We present in this section a framework of identifying brain patterns induced by diseases with the use of DTI data. In this paper, scalar values are represented by regular lower case letters, while bold lower case letters, such as $\mathbf{x}$, and $\mathbf{y}$, represent vectors. Capital bold letters, such as $\mathbf{V}$, represent sets of vectors. In feature selection and classification,

all the DT images are from two classes $C_i \in \{-1, 1\}$, which represent two groups: healthy controls (positive class) and patients with disease (negative class).

2.1 Scalar Features of Diffusion Tensors

In diffusion tensor images, a diffusion tensor at a voxel is a 3×3 positive-definite symmetric matrix D, which can be represented by its eigen-decomposition as $D = \lambda_1 \mathbf{g}_1 \mathbf{g}_1^T + \lambda_2 \mathbf{g}_2 \mathbf{g}_2^T + \lambda_3 \mathbf{g}_3 \mathbf{g}_3^T$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ are the eigenvalues and eigenvectors of D respectively. Multiple measures, called "scalar features" in the paper, can be extracted from tensor data. For classification, we focus on two types of features, i.e., anisotropy and geometry based scalar features. We choose Fractional Anisotropy (FA) to characterize the anisotropy of diffusion tensor. The geometry based features characterizes different aspects of tensors in terms of anisotropy and diffusivity: linearity ($C_l = \frac{\lambda_1 - \lambda_2}{\lambda_1}$), planarity ($C_p = \frac{\lambda_2 - \lambda_3}{\lambda_1}$) and sphericity ($C_s = \frac{\lambda_3}{\lambda_1}$). Such scalar features provide different information about brain structures, and their combination is expected to provide comprehensive information of the underlying tissue structure. In this paper, we demonstrate that the combination of multiple scalar features can provide better discriminative accuracy than using single features. Since $C_l + C_p + C_s = 1$, our method uses only two of them, i.e, C_l, C_p, as well as FA.

2.2 Bayes Error Estimation for DTI Feature Selection

A common difficulty in classification of DT images is that they are high dimensional, but there are only a limited number of samples available in any disease-specific classification problem. Therefore for the purposes of analysis, it is important to select a subset of voxels in the brain, which comprehensively represents the underlying abnormality induced by disease. This is equivalent to a typical feature selection task in machine learning [9]. In DTI classification, we re-formalize the feature selection as selecting a subset of voxels, such that the scalar features, including possible combinations of FA, C_l and C_p, in these voxels can well discriminate brains with pathology from healthy brains. Although numerous methods have been proposed to rank and select features [9], identifying a "better choice" is still problem-dependent and hence challenging, in most scenarios. Due to the high dimensionality of medical images, it has been shown that the feature ranking based on the discriminant criteria, followed by classifier-dependent feature selection, is a realistic solution to classification of medical images [8]. However, there is little work in this regard in DT images.

In our framework, we adopt a fundamental method of ranking voxels by estimating their corresponding Bayes error. By its definition, the Bayes error is the minimal error rate that a probabilistic classifier can achieve when applying the Bayes rule [10]. Our method directly estimates Bayes error at each voxel, based on the k-NN and Parzen window method [11,10]. Unlike linear statistical methods that usually assume global Gaussian distributions, this method only assumes the local smoothness in kernel density estimation, therefor it can naturally handle non-Gaussian data. Also this method is able to combine multiple scalar features for voxel selection.

In the Parzen window kernel density estimation, the likelihood probability of **x** for class C_i is given as Eqn. (1):

$$P(\mathbf{x}|C_i) = \frac{1}{nh^d} \sum_{\mathbf{y} \in C_i} K(\frac{\mathbf{y} - \mathbf{x}}{h}) \tag{1}$$

In Eqn. (1), **x** and **y** are feature vectors, which are combinations of scalar features extracted at a voxel from different brains. **y** corresponds to a DT image of the class C_i. n is the total number of samples, d is the feature dimensionality, K is the kernel function, and h is the kernel window size. In our framework, we adopt the widely used Gaussian kernel. Since the $K(\mathbf{y} - \mathbf{x})$ is small when **y** is far from **x**, we only use the data in a k-nearest neighborhood in the kernel density estimation to reduce computational complexity. From the Bayes rule, the posterior probability of class C_i given **x** is

$$P(C_i|\mathbf{x}) \propto P(C_i) \sum_{\mathbf{y} \in C_i, \mathbf{y} \in NN_\mathbf{x}} K(\frac{\mathbf{y} - \mathbf{x}}{h}) \tag{2}$$

where NN_x is a k-nearest neighborhood of **x**. Two parameters in the Gaussian kernel function, i.e., the covariance matrix and kernel size h, can be optimally estimated from training samples assuming local smoothness [12]. According to its definition, the Bayes error rate P_e can be estimated as Eqn. (3):

$$P_e = \sum (1 - \max_i \{(C_i|\mathbf{x})\}) P(\mathbf{x}) \tag{3}$$

Our method estimates error rate P_e at each individual voxel in the brain. The range of P_e is between 0 to 0.5, where 0 means perfect separability and 0.5 means total overlap between classes at a voxel level. The voxels with small error rates are more discriminative of different patterns.

2.3 Region Selection, Feature Extraction, and Classification

For each voxel **v** in the 3D brain image, the Bayes error rate $P_e(\mathbf{v})$ is estimated based on Eqns. (2) and (3), to represent its discriminative capability for pattern classification. To select a subset of voxels for pattern classification, only those voxels with error rate less than a threshold P_{th} are preserved. The selected voxels sets are denoted as $\mathbf{V}^s = \{\mathbf{v}|P_e(\mathbf{v}) < P_{th}\}$. Usually brain tissue changes occur in regions instead of at single voxels; therefore, the selected voxels are further grouped into local regions by connected component analysis. Small local regions are removed to eliminate noisy regions caused by registration error. As a result of region selection, we obtain a set of disjointed local regions, with each local region containing a subset of connected voxels in the brain.

Features are then extracted from the selected regions to characterize brain patterns. Our method first extracts Kernel Principal Component Analysis (KPCA) features from each of the local regions, to remove the nonlinear correlation among scalar features and reduce feature dimensionality. All the local features are then combined into a pattern feature vector $\mathbf{f} = \{\mathbf{f}_1, ..., \mathbf{f}_i, ...\}$ where $\mathbf{f}_i$ is the KPCA feature extracted at i-th local

region. The feature vector **f** represents the global brain pattern of a DT image, and is ready to be used for pattern recognition.

For each testing DT image, extracted pattern features **f** are used to predict its class label (healthy/patient). The classification can validate whether selected regions reflect the underlying differences between the healthy and patient groups, and can also provide numerical outputs indicating the degree of class membership, and thereby can assist in disease diagnoses. We adopt a probabilistic k-NN classifier, which follows the same principle of k-NN/Parzen window based kernel density estimation as described in Section 2.2. For a DT image **x** for testing, its posterior probability of belonging to class C_i is estimated as $P(C_i|\mathbf{x}) \propto P(C_i) \sum_{\mathbf{y} \in C_i, \mathbf{y} \in NN_{\mathbf{x}}} K(\frac{|\mathbf{f_y} - \mathbf{f_x}|}{\sigma'})$, where $\mathbf{f_x}$ and $\mathbf{f_y}$ are pattern features extracted from the testing image **x** and training images **y**, and σ' is the kernel parameter. The predicted class is the one with the maximal posterior probability, i.e., $\hat{C} = \arg_{C_i} \max P(C_i|\mathbf{x})$. In our experiments shown in Section 3, the k-NN classifier has outperformed the commonly used Support Vector Machine (SVM) [13].

3 Experiments

In this study, we apply our method to real datasets. The data set consists of 36 healthy controls (17 male and 19 female) and 34 patients (21 male and 13 female) diagnosed with schizophrenia. The males and females are analyzed separately as the progression of the schizophrenia is known to be different in the genders [14]. All the DTI data has been acquired on a Siemens 3T scanner, at the dimensions of $128 \times 128 \times 40$, with voxel resolution of $1.72 \times 1.72 \times 3.0$ mm. For each subject, DTI images were registered to a common template space using a deformable registration method [6]. Scalar features of FA, C_l and C_p are computed for these normalized images. In the following experiments, abnormal brain regions that can distinguish patients from controls are first identified, and then validated through pattern classification experiments.

3.1 Identifying Abnormal Brain Regions

In the first experiment, we identify brain regions where abnormal tissue changes occur. For each voxel in the brain, the Bayes error is estimated using combinations of scalar features, and then local regions are selected by ranking and filtering voxels, as described in Section 2.3. Since DTI mainly characterizes the white matter, the method is applied to voxels where FA is greater than 0.10 (a loose threshold to mainly discard the CSF). Due to limited space, only the results of female participants with the use of combinations of FA, C_l and C_p are displayed in Figure 1. Figure 1 (a) shows color maps of estimated Bayes error, overlayed on FA maps of a heathy control. For visual effect, the red color shows regions with smaller estimated error rate, meaning higher group separation between patients and healthy people, and blue color represents regions with lower group separation. Figure 1 (b), (c), (d) show identified abnormal regions. Such regions are obtained by setting threshold $P_{th} = 0.40$, which represents a significant group difference at the voxel level. The affected regions are mainly located in the corpus callosum, corona radiata, corticospinal tract, and cingulum bundle. These regions have been confirmed to be affected in schizophrenia by other researchers [2] using completely different methods.

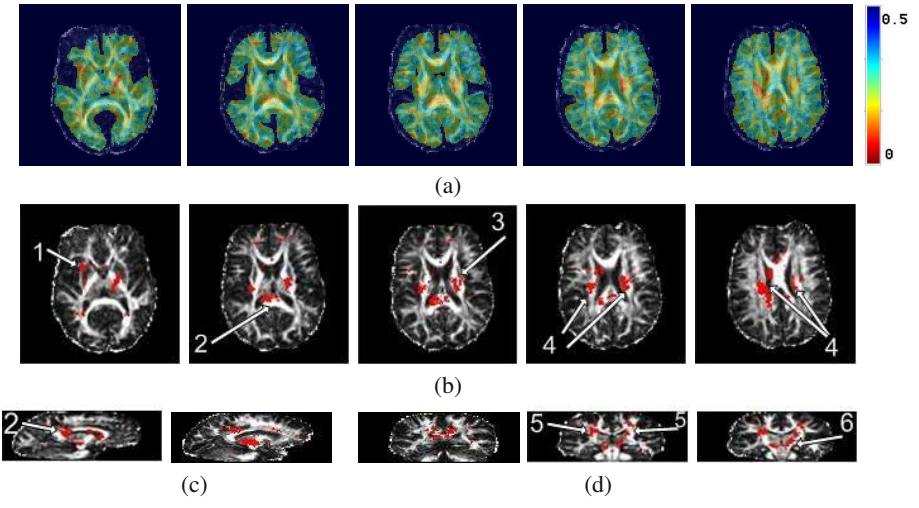


Fig. 1. Bayes error estimation and identified abnormal brain regions in female participants by using combinations of FA, C_l and C_p. (a): map of estimated Bayes error overlayed on FA maps of a healthy control at the axial view. The red color shows smaller estimated error rate, i.e., more discriminative regions; (b) (c) (d): some identified regions (P_{th} = 0.40) at the axial, sagittal and coronal views respectively. Some of those regions lie in: (1) external capsule; (2) corpus callosum; (3) internal capsule; (4) cingulum bundle; (5) corona radiata; (6) corticospinal tract.

3.2 Validation of Classification

The identified regions indicate possible existence of disease-induced abnormal tissue changes in the brain. To evaluate how such regions can be used to distinguish between brain patterns of patients and healthy controls, cross-validation experiments [10] are performed. *We first validate that our framework provides better accuracy than a commonly used correlation based feature selection method, through experiments on single features, and then we demonstrate that combinations of multiple features can further improve classification accuracy.* In the cross-validation, only the training set will be used for the feature selection and classifier training. The Bayes error estimation based feature selection is compared with a commonly used feature selection method based on linear correlation coefficients, applied to both SVM and k-NN classifiers. In the correlation based feature selection, the linear correlation coefficient γ between a scalar feature variable x and class label c is calculated as $\gamma = \frac{\sum_j (x_j - \bar{x})(c_j - \bar{c})}{\sqrt{\sum_j (x_j - \bar{x})} \sqrt{\sum_j (c_j - \bar{c})}}$ [9]. Such correlation indicates the dependence between features and class label, and is optimal when each class is a single Gaussian distribution. Following the procedure of feature extraction described in Section 2.3, extracted features are then input to two types of classifiers, i.e., SVM and k-NN, for the prediction of class label. In all of the experiments, classifier parameters are optimized automatically through cross-validation within training samples only. Such parameters include the kernel type and kernel size in SVM, and the size of nearest neighborhood and the kernel size in k-NN. The threshold P_{th} is empirically set to select a certain number of voxels, e.g. between 600 to 1000 voxels. When more

Table 1. Recognition error of leave-one-out (LOO) cross-validations using single scalar features. Lower error indicates better performance.

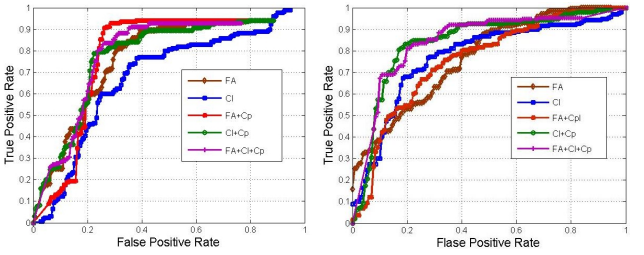
Features	Male				Female			
	Correlation		Bayes Error		Correlation		Bayes Error	
	SVM	k-NN	SVM	k-NN	SVM	k-NN	SVM	k-NN
FA	28.95%	26.32%	26.32%	**23.68**%	28.13%	25.00%	25.00 %	**18.75**%
C_l	36.84%	31.58%	31.21%	28.95%	37.50%	33.75%	25.00 %	21.88%
C_p	39.47%	34.21%	39.47%	36.84%	31.25%	31.25%	28.13 %	28.13%

Table 2. Recognition error of 10-fold cross-validations using single and combined scalar features. Lower error indicates better performance.

Single features	Male		Female		Combined features	Male		Female	
	SVM	k-NN	SVM	k-NN		SVM	k-NN	SVM	k-NN
FA	32.11%	25.79%	28.75 %	26.87%	(FA, C_p)	23.68%	**19.68**%	33.75%	28.13%
C_l	32.61%	31.58%	29.69 %	25.00%	(C_l, C_p)	27.11%	21.84%	24.37%	**17.81**%
C_p	40.79%	37.87%	34.06%	30.63%	(FA, C_l, C_p)	27.89%	22.37%	25.62 %	19.06%

voxels are added, the classification accuracy starts decreasing due to more noise and limited number of samples. The comparison of leave-one-out (LOO) cross-validation accuracy using single scalar features are summarized in Table 1. Results show that our method outperforms the correlation based method, for both SVM and k-NN classifiers. Among the three scalar features, FA and C_l provide better accuracy than C_p. Furthermore, for both feature selection methods, the k-NN classifier provides a slightly better accuracy than commonly used SVM.

We further validate that the DTI pattern classification accuracy can be significantly improved with the use of combined features. Since only limited number of samples are available, for a robust evaluation of our method, we adopt multiple times of 10-fold cross-validations. The final ROC is obtained from the outputs from multiple times cross-validation. Compared to a single cross-validation, such multiple cross-validations provide a more robust evaluation that is less sensitive to parameter tuning, especially when the data number is limited. Comparisons between pattern recognition using single and combined scalar features based on 10 times 10-fold cross-validation are summarized in Table 2, where the 10-fold cross-validation accuracy is slightly worse than the leave-one-out cross-validation. From the table, it can be observed that combining different scalar features significantly improves the pattern classification accuracy. For example, when combining FA and C_p features, the cross-validation recognition error of male participants is decreased to 19.68% from 25.79% of using FA features. Other combinations, such FA+ C_p, also improves the recognition accuracy. For female participants, the combination of FA and C_p greatly improves the recognition accuracy, i.e, recognition error decreased from around 25.00% to 17.81%. Furthermore, from all the experiments, the k-NN classifier consistently outperforms the SVM classifier. ROC curves corresponding to different combinations of scalar features, displayed in Figure 2, further confirm the advantages of combining scalar features.



(a) (b)

Fig. 2. Comparison of ROCs from 10-fold cross-validations using different combination of scalar features. (a) ROCs of male participants; (b) ROCs of female participants.

We notice that, although in general, combinations of different scalar features improve the recognition accuracy, the optimal combinations that achieve the best accuracy are different for female and male sets. Hence the regions that are involved in the classification are also different for the two genders. This supports the clinical observation that the progression of schizophrenia is different in men and women [14]. Also, our experiments show that the combinations of the planar index C_p, which can be seen as a measure of diffusivity, with other features of anisotropy can significantly improve the recognition accuracy, although C_p on its own cannot effectively distinguish patients from healthy people. This demonstrates that C_p carries complimentary information to the other two features (FA and C_l) which characterize the anisotropy. The results confirm the benefits of combining multiple tensor features by our method. Furthermore, the combination of all the features does not necessarily provide the best accuracy. This is probably caused by insufficient samples in estimating pattern overlap when the feature dimensionality increases.

4 Conclusion

We have presented a novel method of identifying regions in the brain affected by disease, based on the pattern classification of DTI data. The method is able to detect nonlinear group differences between patients and controls, and can be used to distinguish disease-affected brains from healthy controls, and potentially aid in disease diagnosis. We show that the combination of multiple scalar features from tensors characterizing the anisotropy and the shape can improve the classification accuracy due to the complementary nature of physical properties that the features characterize. Our future work will focus on generalizing the framework to other medical image group classifications such as in aging and to multi-modality data.

Acknowledgement

This work is supported by NIH grants R01-MH-079938 and 1R01MH73174-01. The authors would like to thank the Brain Behavior Laboratory, University of Pennsylvania, for providing us the data.

References

1. Basser, P.J., Mattiello, J., LeBihan, D.: MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66, 259–267 (1994)
2. Kubicki, M., McCarley, R.W., Westin, C.F., Park, H.J., Maier, S., Kikinis, R., Shenton, M.E., Jolesz, F.A.: A review of diffusion tensor imaging studies in schizophrenia. Journal of Psychiatric Research 41, 15–30 (2007)
3. Ashburner, J., Fristonn, K.J.: Voxel-based morphometry: The methods. NeuroImage 11, 805–821 (2000)
4. Westin, C.F., Maier, S.E., Mamata, H., Nabavi, A., Jolesz, F.A., Kikinis, R.: Processing and visualization of diffusion tensor MRI. Med. Image Anal. 6(2), 93–108 (2002)
5. Lenglet, C., Rousson, M., Deriche, R.: DTI segmentation by statistical surface evolution. IEEE Trans. Med. Img. 25(6), 685–700 (2006)
6. Xu, D., Mori, S., Shen, D., van Zijl, P.C.M., Davatzikos, C.: Spatial normalization of diffusion tensor fields. Magnetic Resonance in Medicine 50, 175–182 (2003)
7. Caan, M., Vermeer, K., van Vliet, L., Majoie, C., Peters, B., den Heeten, G., Vos, F.: Shaving diffusion tensor images in discriminant analysis: a study into schizophrenia. Med. Image Anal. 10, 841–850 (2006)
8. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: COMPARE: Classification of morphological patterns using adaptive regional elements. TMI 26, 93–105 (2007)
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
10. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. John Wiley Sons, Chichester (2000)
11. Fukunaga, K., Hummels, D.M.: Bayes error estimation using Parzen and k-NN procedures. IEEE Trans. Pattern Anal. Mach. Intell. 9(5), 634–643 (1987)
12. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman & Hall / CRC (1986)
13. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 4
14. Nopoulos, P., Flaum, M., Andreasen, N.C.: Sex differences in brain morphology in schizophrenia. Am. J. Psychiatry 154, 1648–1654 (1997)