A Data-Mining Framework for Classification of High Resolution Magnetic Resonance Images

Christian Böhm
University of Munich
boehm@dbs.ifi.lmu.de

Annahita Oswald
University of Munich
oswald@dbs.ifi.lmu.de

Claudia Plant
Technical University of Munich
plant@lrz.tum.de

Bianca Wackersreuther
University of Munich
wackersb@dbs.ifi.lmu.de

ABSTRACT

Magnetic resonance imaging (MRI) allows to display brain structures with highest resolution. To fully exploit the potential of this imagining modality, data mining methods are required to reveal subtle differences in brain structure caused by disorders such as Mild Cognitive Impairment (MCI) and early stage Alzheimers disease (AD). In this paper, we propose a data mining framework which combines elements from feature selection, clustering, classification and provides a concise visualization of affected areas in the brain.

1. INTRODUCTION

Recently, magnetic resonance imaging (MRI) has become very popular since this non-invasive imaging method allows to visualize function and structure of body parts in high resolution without exposing subjects to radiation. MRI works by a powerful magnetic field which aligns the nuclear magnetization of hydrogen atoms in the body. The magnetization is systematically changed by radio frequency fields to produce a signal which is detectable by the scanner. MRI allows to distinguish different soft tissues much better than Computed Tomography (CT) and therefore is especially suitable for brain imaging. In clinical practice MRI is e.g. commonly applied for diagnosis and monitoring of different types brain tumors. Most tumors can reliably be detected by an expert by inspection. However there is a wide range of diseases which cause very subtle alterations of the brain and thus can not be diagnosed by visual inspection of a single MRI image, e.g. Alzheimers Disease (AD) and Mild Cognitive Impairment (MCI).

Demographic changes lead to an increasing prevalence of AD, the most frequent form of age-related dementia. MCI is often regarded as an early stage of AD. Persons with MCI have cognitive deficits but still manage to organize their everyday life. Many persons with MCI develop AD within five years, however some remain stable, and some even recover to normal. The diagnosis of AD and MCI mainly relies on clinical criteria so far. Widespread questionnaires include e.g. the evaluation of the cognitive performance by relatives, which has turned out to be highly subjective. Sensitive and specific early stage diagnosis of AD is of prime importance to therapeutic interventions. First studies demonstrate that MRI provides the potential to contribute to the diagnosis of AD, and moreover, to evaluate the individual risk of subjects with MCI to develop AD.

Data mining and pattern recognition methods are required to extract from millions of voxels within an MRI image the minimal set of voxels that shows systematic abnormalities in those subjects that convert to AD. Some studies demonstrated that specific spatial patterns of brain atrophy are correlated with the progression to AD in subjects with MCI. ([1, 10]). In this paper we propose a two-step data mining framework combining a distribution free feature selection algorithm at the first stage and, at the second stage, different multivariate classifiers. Feature selection circumvents potential problems of previous approaches including lack of statistical power due to multiple testing [4] or the lack of supervision in dimensionality reduction (e.g. PCA) [10]. We apply cross-validated classifiers including support vectors machine (SVM), Bayesian classification, and voting feature intervals (VFI) to derive the minimal set of voxels for optimized prediction of diagnosis (AD vs HC) or prediction of conversion. To obtain a spatial contiguous result and to improve the predictive validity we apply a modified clustering algorithm on the selected feature subset. Founded on density-based clustering, this algorithm determines a grouping of the features based on their location and discriminatory power. Our technique allows a highly accurate identification of subjects with AD and MCI and can also be applied to support the prognosis whether subjects with MCI will convert to AD. In addition, our method provides a concise visualization of the brain regions affected by MCI and AD and thus may contribute to a better understanding of the mechanisms and progress of AD.

2. METHOLOGY

Data. MRI examinations of the brain were performed on a 1.5 Tesla MRI scanner. The preprocessing of the sans was conducted with the statistical software package SPM2 (http://www.fil.ion.ucl.ac.uk/spm/). In addition, high dimensional normalization and segmentation into brain and cerebrospinal fluid was performed as described in [10].

KDD '08 Las Vegas, Nevada USA

Notations. Given a data set DS consisting of MRI scans of n subjects $s_1, ..., s_n$ labeled to a set of k discrete classes $C = c_1, ..., c_k$ (in our study e.g. HC and AD), we denote the class label of subject s_i by $si.c$. For each subject we have a MRI image which we regard as a feature vector V composed of d voxels $v_1, ..., v_d$. In our study each scan of each subject consists of $d = 4,035,528$ voxels.

2.1 Feature Selection

In the first step, we select the most discriminating features from the image using a feature selection criterion. We use the Information Gain ([6]) to rate the interestingness of a voxel for class separation, because it is highly efficient to compute and is not restricted to linear correlations but captures arbitrary dependencies between features and class labels.

Entropy of the class distribution. The entropy of the class distribution $H(C)$ is defined as

$$H(C) = \sum_{c_i \in C} p(c_i) \cdot log_2(p(c_i)) \quad (1)$$

whereas $p(c_i)$ denotes the probability of class c_i. $H(C)$ corresponds to the required amount of bits to tell the class of an unknown subject and scales between 0 and 1.

Information Gain of a voxel. Now we can define the Information Gain of a voxel v_i as the amount by which $H(C)$ decreases by the additional information provided by v_i on the class, which is described by the conditional entropy $H(C|v_i)$.

$$IG(v_i) = H(C) - H(C|v_i) \quad (2)$$

The Information Gain scales between 0 and 1, whereas 0 means that the corresponding voxel provides no information on class label of the subjects. An Information Gain of 1 means that the class labels of all subjects can be derived from the corresponding voxel without any errors. To compute the conditional entropy, features with continuous values, as in our case, need to be discretized using the algorithm of Fayyad and Irani [5]. This method aims at dividing the attribute range into class pure intervals. The cut points are determined by the Information Gain of the split. Since a higher number of cut points always implies higher class purity but may lead to overfitting, an information-theoretic criterion based on the Minimum Description Length principle is used to determine the optimal number and location of the cut points.

2.2 Clustering

DBSCAN. After feature selection, we apply the density-based clustering algorithm DBSCAN [3] to identify groups of adjacent voxels with a high discriminatory power and to remove noise. As a density-based algorithm, DBSCAN detects clusters as areas of high object density in the feature space which are separated by areas of lower object density. This idea is formalized by two two parameters: ϵ specifying a volume and *MinPts* specifying a number of objects. An object O is called *core object* if it has at least *MinPts* objects in its ϵ range, i.e. $|N_\epsilon(O)| >= MinPts$, whereas N_ϵ (O) = {O'| dist(O, O') <= ϵ}. An object P is directly density reachable from another object O w.r.t ϵ and *MinPts* if O is a *core object* and P $\in N_\epsilon$(O). An object P is density reachable from an object O w.r.t. ϵ and *MinPts* if there exists a sequence of objects $P_1, \cdots, P_n$ such that P_1 = O and P_n = P and P_{i+1} is directly density reachable w.r.t. ϵ and *MinPts* from P_i for $1 \leq i \leq n$. Two objects O and P are density connected w.r.t. ϵ and *MinPts* if there exists an object Q such that both O and P are density reachable from Q. A density based cluster is the maximum set of density connected objects.

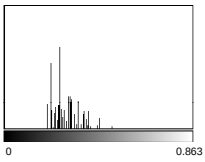


Figure 1: Information Gain Histogram AD vs HC.

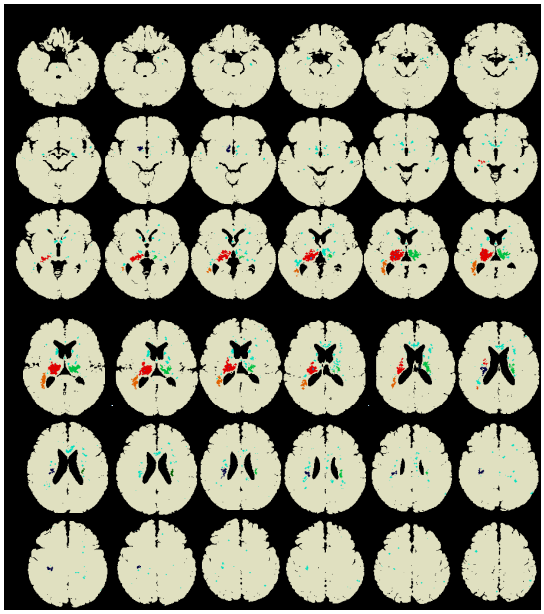
Modifications. For clustering voxels in our specific set of images we adapt the algorithm by redefining *core object* and *direct density reachability* as follows: We call a voxel v_i a *core object* if the IG of v_i is larger than a given threshold t_{core} and v_i is surrounded by at least *MinVox* voxels having an IG of at least t_{border}. Since the voxels in our specific set of images have either a significant IG value or an IG of zero (see Figure 1), we don't have to distinguish between t_{core} and t_{border}. So we set t_{core} and t_{border} to the minimum IG in the data set and use *MinVox*=6, which means that we require a core object to be situated in a neighborhood of highly discriminative voxels.

2.3 Classification

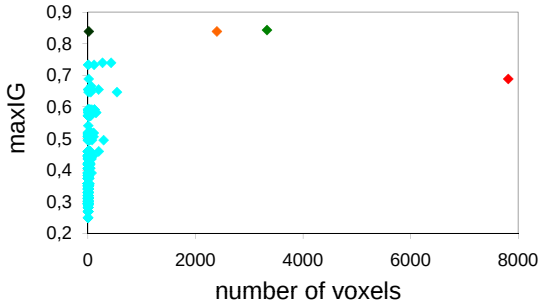
After clustering, the selected features represent spatially coherent regions which exhibit significant differences among the groups. At this stage, classification algorithms can be applied to validate the discriminatory power of these selected clusters. We applied the Linear Support Vektor Machine (SVM) [9], the Bayesian Classifier [7] and the classification approach by voting feature intervals (VFI) [2] as these are three approaches with very different algorithmic paradigms.

2.4 Validation

Cross-validation is an appropriate validation scheme in the case of few training examples w.r.t. the dimensionality of the data [8]. The data set is divided into n folds of size n-1 subjects each. In each fold, n-1 subjects are used for training, i.e. we perform the steps feature selection and clustering on these n-1 subjects and obtain a pattern of highly selective clusters. The remaining subject is used as test object, i.e. we predict the class label of this subject by applying a classifier in the feature space defined by the clusters derived in the training phase. To avoid to apply the whole framework (feature selection, clustering, classification) n-times, we pursue a "worst case" strategy in the feature selection step. As overall Information Gain of a voxel we retain the minimum value among all n folds. Therefore only voxels which are relevant in all folds are selected. Thus, feature selection has to be performed only once. Since clustering refines the result of feature selection by reducing the number of voxels, also the clustering step needs to be performed only once. After clustering, classification is performed with leave-one-out validation. In contrast to cross-validation, the train-and-test methodology employs two fully different data



(a) Anatomical location



(b) Cluster Size and maximum Information Gain.

Figure 2: Selected Features AD vs. HC after Clustering. Best viewed in color.

sets as training and test data. To evaluate the quality of the classification result, we report three established measures: accuracy, sensitivity and specificity.

3. RESULTS

To assess highly selective brain areas, 19 patients with clinically probable AD, 19 patients with amnestic MCI and 18 healthy controls (HC) underwent MRI and clinical examinations. MCI patients were followed-up in longitudinal clinical and neuropsychological examinations to determine which subjects converted to AD and which remained stable. We built three data sets in order to first apply our framework with leave-one-out cross validation on these data sets. Data set A consists of 19 subjects affected with AD and 18 healthy controls. Data set B consists of 24 subjects with MCI, among which 9 converted to AD (MCI-AD) and 15 remained stable (MCI-MCI) within the follow-up interval. Data set C consists of 24 subjects with MCI (the same subjects as in data set B) and 19 controls (the same subjects as in data set A). Second we predict conversion of patients with MCI to AD by using train and test sets in our framework.

Classification of AD vs HC. 3,969,492 of 4,035,528 voxels (98.36%) have an Information Gain of 0, i.e. they contain no information separating the groups and are therefore excluded from further analysis. Theoretically, combinations of these features may provide valuable information. However, due to the high dimensionality of the data, an exhaustive search for feature combinations is not applicable. The range of IG value among the remaining 66,036 voxels was between 0.23 and 0.86. The minimum IG of 0.23 was relatively high, indicating that the voxels either contain valuable information to separate the classes or are completely irrelevant. Clustering reduces the 66,036 selected features to 18,797. In total, 981 clusters containing as a minimum one core object exhibiting the maximum number of 6 neighbors are obtained. The largest cluster comprises 7,817 voxels. Figure 2(b) summarizes the cluster statistics with respect to the two most important criteria: the size of the clusters and value of IG. For each cluster, the size (number of voxels) is plotted on the x-axis and the maximum IG is displayed on the y-axis. The most interesting clusters with regard to both criteria are highlighted in different colors. The anatomical locations of these clusters have been marked with the same colors in Figure 2(a). The clusters were centered within the medial temporal lobe including the hippocampus, parahippocampus and amygdala and adjacent basal ganglia, the right anterior cingulate gyrus extending towards the prefrontal cortex, and left insula and claustrum. These results are consistent with a previous PCA-based analysis [10]. On the basis of these selected clusters of interest, a classification accuracy of 91.89% with both SVM and Bayes (SVM: sensitivity 94.7%, specificity 91.4%, Bayes: sensitivity 89.5%, specificity 94.4), and 89.19% with VFI (sensitivity: 84.2%, specificity 94.4%) was obtained.

Classification of MCI-AD vs MCI-MCI. When applying feature selection on the brain images of the group of MCI with respect to conversion we obtain 25,284 features with IG greater zero. The minimum occurring IG is 0.36. Clustering drastically reduces the number of features to 2,003. We obtained excellent classification results to separate converters and non-converters: 100% accuracy is obtained with SVM, 91.30% with Bayes (sensitivity 75%, specificity 100%) and 86.96% with VFI (sensitivity 100%, specificity 80%). The most interesting clusters with respect to size and IG for the separation of converters and non-converters are displayed in Figure 3. They are roughly a subset of the most selective regions for AD vs. HC, extending towards the temporal lobe including the superior temporal gyrus.

Classification of HC vs MCI. We obtain 24,955 characteristic features. Clustering reduces the number of features to 1,069. In total, 134 clusters are obtained. On the clustered data, linear SVM performs best with 97.29% in accuracy (sensitivity and specificity 97%). With VFI we obtain an accuracy of 89.12% (sensitivity: 89.5%, specificity: 88.9%) and with Bayes an accuracy of 83.7% (sensitivity: 84.2%, specificity 83.3%). The spatial pattern of the clusters is similar to that described for MCI-AD vs MCI-MCI and therefore not depicted.

Prediction of Conversion. To demonstrate the discriminatory power of the identified patterns and to confirm that

Data sets	Validation	Training data	Test data	Best accuracy
A	Leave-one-out	n.a.	n.a.	91.89% (SVM,Bayes)
B	Leave-one-out	n.a.	n.a.	100% (SVM)
C	Leave-one-out	n.a.	n.a.	97.29% (SVM)
A,B	Train-and-test	A	B	78.26% (VFI)
A,B	Combined	A,B	A,B	86.96% (Bayes)

Table 1: Classification Results.

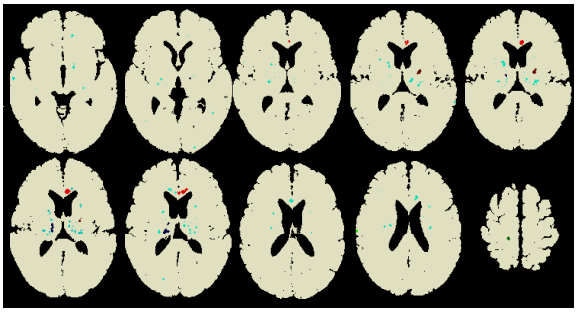


Figure 3: Most Interesting Clusters of MCI-AD vs. MCI-MCI. Best viewed in color.

they are characteristic for AD, we used data set A as training data and data set B as test data. More precisely, we applied feature selection, clustering and training of the classifier on data set A and predicted the conversion of the subjects of data set B based on the learned information. An accuracy of 56.21% was obtained with SVM (sensitivity 62.2%, specificity 53.3%), 73.91% with Bayes (sensitivity 75%, specificity 73.3%). The best result was obtained with VFI with a prediction accuracy of 78.26% (sensitivity 87.5%, specificity 73.3%). The general decrease in performance can be explained by the fact, that the train-and-test setting uses much less training data than cross-validation, and the test data originates from a completely different data set. Considering these aspects, the obtained accuracies are remarkable high. Nevertheless, the accuracies by both Bayes (74%) and VFI (78%) are superior than reported previously (73%) by [10].

Furthermore the performance breakdown may be related to superfluous features in the training data as these can be misleading and may lead to over fitting, especially in the case of high dimensional space and few training examples. Therefore we restricted the feature set to those features characteristic for AD which are also highly selective to distinguish MCI converters and MCI non-converters. On the brain images, 3,341 of the 25,284 features selected to separate MCI-AD from MCI-MCI are also significant for the identification of AD compared to HC. The presumption for the performance breakdown is supported by the improved accuracies of the classifiers. For SVM, the accuracy increases from 56.21% to 82.61% (sensitivity 86.7%, specificity 75%),and for Bayes from 73.91% to 86.96% (sensitivity 87.5%, specificity 93.3%), only VFI shows no change in accuracy. Table 1 provides a summary on the results.

4. CONCLUSION

In this paper we proposed a novel approach to identify regions of high discriminatory power in high resolution brain MRI. Our method combines data mining techniques from feature selection, clustering and classification. The results of our technique show excellent accuracies to identify AD and MCI on high resolution MR images and indicate that it is a valuable complement to existing methods.

To further validate the regions identified for AD and MCI we intend to apply or technique on a larger data set. Going beyond imaging, it is also very interesting to combine these findings with the clinical scores which are currently applied for diagnosis of AD. In addition, this or similar methods could successfully be applied to other pathologies, such es schizophrenia and other imaging modalities such es positron emission tomography.

5. REFERENCES

[1] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick. Detection of prodromal alzheimer's disease via pattern classification of mri. *Neurobiol Aging*, December 2006.

[2] G. Demiröz and H. A. Güvenir. Classification by voting feature intervals. In *ECML*, pages 85–92, 1997.

[3] M. Ester, H.-P. Kriegel, J. Sander, and et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[4] Y. Fan, N. Batmanghelich, C. Clark, and C. Davatzikos. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39:11731–1743, 2008.

[5] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1029, 1993.

[6] M. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.*, 15(6):1437–1447, 2003.

[7] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *UAI*, pages 338–345, 1995.

[8] M. J. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *COLT*, pages 152–162, 1997.

[9] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines, 1998.

[10] S. Teipel, C. Born, M. Ewers, A. Bokde, M. Reiser, H.-J. Möller, and H. Hampel. Multivariate deformation-based analysis of brain atrophy to predict alzheimer's disease in mild cognitive impairment. *NeuroImage*, 38:13–24, 2007.