# Merged Ensemble Clustering in fMRI Analysis

E. Dimitriadou[1], M. Barth[2,3], K. Hornik[4], E. Moser[2,3]

[1]Dept. of Statistics, University of Technology, Vienna, Austria, [2]Dept. of Radiodiagnostics, Medical University, Vienna, Austria, [3]MR Centre of Excellence, Medical University, Vienna, Austria, [4]Dept. of Statistics, University of Economics, Vienna, Austria

## Introduction:

Exploratory data analysis (EDA) methods may be used to identify important aspects within fMRI data sets such as artifacts and unexpected functional activation, because they do not require exact knowledge of the stimulation paradigm, the individual hemodynamic response, or the noise distribution they rather search for various "interesting" time courses [1]. However, EDA results critically depend on the clustering method [2] and corresponding parameters (initialization, metric, fuzziness) used. To overcome this problem we suggest to use a combination of ensemble methods and merging. Ensemble methods use several clustering runs which first result in a set of individual (different) solutions and which are then suitably aggregated to finally obtain a single optimized solution by integrating pre-existing knowledge from all available runs. The application of pluralistic strategies was already proposed for functional MRI analysis [3,4], however, no thorough studies have been performed up to now to find the appropriate approach.

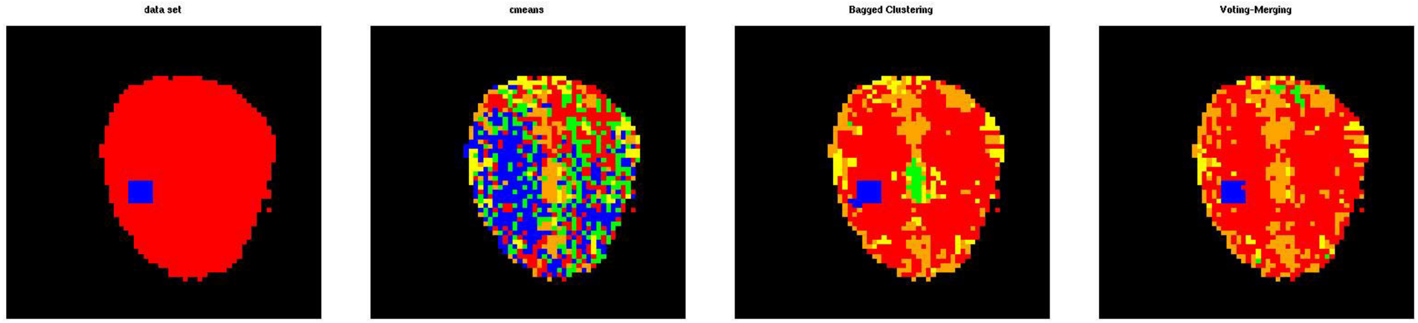
## Materials and Methods:

Quantitative performance assessment was done using a hybrid data set constructed from a baseline in vivo MRI data set. The activation was added artificially with 3 different contrast-to-noise (CNR) ratios of 1.33, 1.66 and 2, respectively, so that the activated pixels (in total 25 true positives -TPs) are known. The data sets and a detailed description are available at http://www.ci.tuwien.ac.at/research/oenb/oenb_data.html.

Two methods were used to perform merged ensemble clustering (MEC). The first, Voting-Merging (VM) [5, 6] is a scheme which consists of 3 procedures, namely the repeated runs of a clustering algorithm (with a large number of initial clusters), the voting procedure receiving these results, and finally the merging procedure by which the number of clusters in a data set are obtained. The 3 levels of the algorithm are applied sequentially so that they do not interfere with each other. As a result of the voting procedure the data points are typically not uniquely assigned to one cluster, but there is a "fuzzy" partition. That is, after voting of N runs we get for every data point x and every cluster j a value which gives the fraction of times this data point has been assigned to this cluster with a certain degree of membership the so-called "sureness". The "AveSure" of a cluster is defined as the average "sureness" of all the points in it. This way, we can set a measure of how strong the points of cluster k belong to cluster j. This measure defines a (non-symmetric) neighborhood relation between two clusters. Thus we define that a cluster k is the closest cluster to cluster j, if the points belonging to k have a higher „AveSure“ for cluster j than for any other cluster. We use this neighborhood relation to develop a merging procedure that starts with the number of clusters obtained from the clustering procedure and merges clusters which are closest to each other. Two clusters (or a whole set of clusters, even chains) are merged, if the clusters are mutually closest to each other. After merging the memberships of the points are recomputed and merging is repeated until some final criterion is met (for example a minimum threshold for the "AveSure" of the clusters, or until a desired number of clusters is reached). The second method, Bagged Clustering [7] consists of 2 steps. First, bootstrap samples of the original data are created by drawing with replacement. The base clustering method is run on each of these samples also with a large number of initial clusters. The centers of all these clustering runs are merged by being clustered by an hierarchical method. If no bootstrapping takes place, then the whole data set is being repeatedly clustered. In our simulations no bootstrapping took place. Fuzzy c-means (FCM) is used as the basic clustering method [8]. Preprocessing included a normalization by subtracting the median value. For processing, a fuzzy index (FI) of 1.1, initial number of clusters (INC) of 40, random initialization, and the Euclidean distance measure were used. Both methods have been performed with the results of N=30 clustering runs. For Bagged Clustering the ward hierachical method was used to merge the centers. All algorithms used for evaluation are implemented in the software package R (http://www.R-project.org), an open-source system for data analysis and graphics implementing the award-winning S language ("GNU S").

## Results and Discussion:

The maps in Figure 1 show the hybrid data set with the true activation (blue) overlaid [far left], a typical result of a c-means run with 5 clusters initialization [center left], a combination of all resulting 5 clusters from the merged clusters using the Bagged Clustering method [center right] and using the VM algorithm [far right] for the data set with CNR 2.00. For the other CNR levels (1.33 and 1.66, 2.00) the VM method results in a partition with 23 and 24 TPs - from the total of 25 - in the activation cluster, respectively, having no false positives at all (FPs). Bagged clustering gets all 25 TPs for data sets of CNR 2.00 and 1.66 with 4 FPs and 50 FPs, respectively. It fails for the 1.33 CNR data set as more clusters need to be initialized in the FCM runs. A comparison with the standard FCM - without any ensemble or merging - using 30 repetitions of the algorithm (with 40 initialized clusters) give a average number of 14-15 TP pixels and a mean of 21-22 FPs (depending on the data set CNR).



Merged ensemble clustering as an unsupervised EDA method provides a single final cluster result which represents the true activation better than using individual results. It thereby reduces the instability induced by the clustering method and corresponding parameter choice.

## References:

[1] C. Windischberger et al, Fuzzy cluster analysis of high-field functional MRI data, Artificial Intell. Medicine, 2003. [2] E. Dimitriadou et al, Artificial Intell. Medicine, A Quantitative Comparison of functional MRI Cluster Analysis, in press. [3] N. Lange et al, Plurality and Resemblance in fMRI Data Analysis, NeuroImage, 1999. [4] M. Jarmasz and R. Somorjai, EROICA: Exploring Regions of Interest with Cluster Analysis in Large Functional Magnetic Resonance Imaging Data Sets, Concepts in Magnetic Resonance Part A, 2003.[5] E. Dimitriadou et al., A Combination Scheme for Fuzzy Clustering, Intern. Journal of Pattern Recognition and Artificial Intelligence, 2002. [6] E. Dimitriadou et al, Voting-Merging: An Ensemble Method for Clustering, Proc. Artificial Neural Networks (ICANN 01), 2001. [7] F. Leisch and K. Hornik, Stabilization of k-Means with Bagged Clustering, Proc. Joint Statistical Meetings, 1999. [8] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.