NeuroImage 58 (2011) 785–792

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier

Ahmed Abdulkadir [a,b,*,1], Bénédicte Mortamet [b,1], Prashanthi Vemuri [c], Clifford R. Jack Jr. [c], Gunnar Krueger [b], Stefan Klöppel [a] and The Alzheimer's Disease Neuroimaging Initiative [2]

[a] Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Freiburg Brain Imaging, University Medical Center Freiburg, Freiburg, Germany
[b] Advanced Clinical Imaging Technology, Siemens Suisse SA, Healthcare Sector IM&WS-Centre d'Imagerie Biomédicale (CIBM), Lausanne, Switzerland
[c] Department of Radiology, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 19 January 2011
Revised 9 June 2011
Accepted 10 June 2011
Available online 25 June 2011

Keywords:
Magnetic resonance imaging
MRI
Support vector machines (SVM)
Alzheimer's disease
Multi-site study

ABSTRACT

Fully automated machine learning methods based on structural magnetic resonance imaging (MRI) data can assist radiologists in the diagnosis of Alzheimer's disease (AD). These algorithms require large data sets to learn the separation of subjects with and without AD. Training and test data may come from heterogeneous hardware settings, which can potentially affect the performance of disease classification.

A total of 518 MRI sessions from 226 healthy controls and 191 individuals with probable AD from the multicenter Alzheimer's Disease Neuroimaging Initiative (ADNI) were used to investigate whether grouping data by acquisition hardware (i.e. vendor, field strength, coil system) is beneficial for the performance of a support vector machine (SVM) classifier, compared to the case where data from different hardware is mixed. We compared the change of the SVM decision value resulting from (a) changes in hardware against the effect of disease and (b) changes resulting simply from rescanning the same subject on the same machine.

Maximum accuracy of 87% was obtained with a training set of all 417 subjects. Classifiers trained with 95 subjects in each diagnostic group and acquired with heterogeneous scanner settings had an empirical detection accuracy of $84.2 \pm 2.4\%$ when tested on an independent set of the same size. These results mirror the accuracy reported in recent studies. Encouragingly, classifiers trained on images acquired with homogenous and heterogeneous hardware settings had equivalent cross-validation performances. Two scans of the same subject acquired on the same machine had very similar decision values and were generally classified into the same group. Higher variation was introduced when two acquisitions of the same subject were performed on two scanners with different field strengths. The variation was unbiased and similar for both diagnostic groups. The findings of the study encourage the pooling of data from different sites to increase the number of training samples and thereby improving performance of disease classifiers. Although small, a change in hardware could lead to a change of the decision value and thus diagnostic grouping. The findings of this study provide estimators for diagnostic accuracy of an automated disease diagnosis method involving scans acquired with different sets of hardware. Furthermore, we show that the level of confidence in the performance estimation significantly depends on the size of the training sample, and hence should be taken into account in a clinical setting.

Introduction

Fully automated methods detecting presence or absence of Alzheimer's disease (AD) based on structural magnetic resonance imaging (MRI) data can help radiologists (Klöppel et al., 2008; Magnin et al., 2009; Plant et al., 2010; Vemuri et al., 2008). AD is associated with formation of extracellular amyloid immunoreactive senile plaques and tau immunoreactive neurofibrillary tangles (Braak and Braak, 1991). It is also associated with progressive atrophic changes that can be detected by structural MRI. Subjects with AD typically show patterns of gray matter (GM) atrophy involving the medial temporal lobe, particularly the hippocampus and entorhinal cortex,

* Corresponding author at: Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Freiburg Brain Imaging, University Medical Center Freiburg, Freiburg, Germany. Fax: +49 761 270 54160.
E-mail address: ahmed.abdulkadir@epfl.ch (A. Abdulkadir).

[1] These authors contributed equally to this work.

[2] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.06.029

among other brain regions, with simultaneous expansion of the ventricles (Baron et al., 2001; Fox et al., 1996; Jack et al., 1992; Whitwell et al., 2007). Due to the characteristic atrophy pattern, the GM is an informative biomarker to detect AD with structural MRI (Klöppel et al., 2008; Magnin et al., 2009; Vemuri et al., 2008).

An increasing number of multi-center studies aim to combine data from different scanners to increase statistical power and fields of applications. Studies suggest that data from different sites can be pooled, but at the same time that systematic inter-scanner differences can occur. Stonnington et al. (2008) compared the variation of data acquired on six distinct scanners of same vendor/type on a voxel-by-voxel level with a mass univariate test on GM probability maps and concluded that the effect of AD is significantly larger than the inter-scanner effects. On the other hand, several studies indicate that the effects of inter-scanner variability are far greater than intra-scanner variability (Huppertz et al., 2010; Moorhead et al., 2009). Similarly, bias field correction and variation in image quality such as signal to noise ratio (SNR) have an impact on the segmentation (Acosta-Cabronero et al., 2008; Klauschen et al., 2009; Shuter et al., 2008). Previous classification methods detecting presence of AD from structural MRI data indicate that performance improved when a high number of samples were used for training (Franke et al., 2010; Klöppel et al., 2009). This may entail the need to pool data from different manufacturers and hardware settings.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) is a large, multi-center, multi-vendor study that acquires structural MRI of cognitively normal healthy controls (CN), mild cognitive impaired (MCI) and AD-probable (AD-p) elders. The ADNI protocols on each scanner type are adjusted such that all sites report comparable results at all times (Jack et al., 2008). Intensive quality control and the use of a phantom, assure low inter-scanner variation and high stability of the image quality (Gunter et al., 2009).

In this study we used data from 56 different sites that participated in the ADNI study to assess the change in detection performance of an AD classifier trained with images acquired either with homogenous or heterogeneous hardware. As in previous work (Klöppel et al., 2008), we used a fully automated processing pipeline and a support vector machine (SVM) classifier (Vapnik, 1998). The process that computes spatially normalized GM probability maps in a common template space from structural T1MRI images was found to outperform other approaches in a recent comparison using multi-site data from ADNI (Cuingnet et al., 2011). We set out to investigate the impact of heterogeneity of the acquisition hardware on the classifier outcome. First, as coarse measure of the performance, we computed the accuracy of classifiers trained on homogenous hardware (pure set). Then we computed the ranges of accuracies that can be expected from classifiers trained on randomly selected images from heterogeneous hardware (mixed sets) with the same sample sizes as the pure sets. These distributions were then compared to the previously observed accuracies of each pure set. Second, in order to quantify hardware-related effects we introduced the analysis of the SVM decision value. Positive values indicated AD-p and negative values indicated CN. Ideally, the decision value should depend only on the subject, not on the hardware. The further away from zero, the higher is the confidence of the classifier in its decision. With the intention to determine the minimal uncertainty of this value due to acquisition noise and pre-processing, we quantified the variation of the decision value between back-to-back scans of subjects. Then we quantified the variation of the decision value between scans of same subjects on both field strengths.

Materials

Participants and image acquisition

Our data included T1-weighted MR images from 417 individuals of which 226 were cognitively normal healthy controls (Mini-mental state examination (MMSE): 29.1 ± 1.0, age: 76.1 ± 5.0) and 191 had probable AD (MMSE: 23.3 ± 2.1, age: 75.5 ± 7.5). All images were obtained from ADNI. Inclusion criteria for participants were according to the protocol described in http://www.adni-info.org/scientists/AboutAdni.aspx#. Individuals assigned to the AD-p group met NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984). We first selected all ADNI CN and AD-p subjects with a baseline MRI scan (all were scanned on 1.5 T, a subset also on 3 T). We excluded 2 AD subjects that progressed to some other dementia during follow-up. The median follow-up time for all patients was 24 months. The interquartile ranges (IQR) by field strength are listed here: 1.5 T–IQR: 24–36 months and 3 T–IQR: 24–31 months. Three subjects were further excluded because the required baseline images were not available. A total number of 417 subjects were included. The list of all images is attached in the supplementary material. T1-weighted sagittal volumes were obtained using the magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence with imaging parameters TR = 2300 ms, TI = 900 ms, flip-angle = 9°at 3 T (and TR = 2400 ms, TI = 1000 ms, flip angle = 8°at 1.5 T) minimum full TE, sagittal slices = 160. All 1.5 T subject acquisitions used 1.25×1.25 mm^2 in-plane spatial resolution and 1.2-mm thick sagittal slices. The 3 T subject acquisitions also used 1.2-mm thick sagittal slices, but were acquired with 1.0×1.0 mm^2 in-plane spatial resolution. Back-to-back scans were acquired from each subject within each scanning session and an image analyst at Mayo clinic rated the image quality of each scan. Quality criteria included blurring/ghosting, flow and susceptibility artifacts. For the analysis based on accuracy we included the ADNI baseline scan (Timepoint 1) with the best quality rating to avoid misclassifications due to low quality, e.g. caused by motion artifacts. For the analysis of the impact when changing field strength, we included further 192 back-to-back scans with a lower or equal quality compared to the other image acquired at the same session. The ADNI structural brain imaging data can be downloaded with or without certain processing steps applied (see http://www.loni.ucla.edu/ADNI/Data/ADNI_Data.shtml). Availability of pre-processing steps depends on manufacturer and coil system (Jack et al., 2008). We included images that were corrected for system-specific image geometry distortion due to gradient non-linearity (GradWarp) and, if available, additional image intensity non-uniformity (B1 correction).We excluded subjects with diagnosed MCI to reduce biological variability, as this diagnostic group is arguably the most heterogeneous. The scanner configurations considered were (a) manufacturer, namely Siemens Healthcare, GE Healthcare and Philips Medical Systems, (b) magnetic field strength, namely 1.5 T and 3 T, and (c) coil system, namely single-channel birdcage coils (BC) and multi-channel phased-array head coils (PA). We focused on these parameters as they were explicitly taken into account during the establishment of the MRI protocols for the ADNI study (Jack et al., 2008). Other configurations like scanner software version, detailed coil configuration or coil type were not considered. Platform-specific lists of sequence parameters are available at http://www.loni.ucla.edu/ADNI/Research/Cores/.

Each of the 417 individuals had a baseline scan at 1.5 T. Among these, 101 participants had a second scan within 2 to 102 days (24 ± 15 days) in a scanner with 3 T. For the rest of the article, we will refer to the 316 images of higher quality of individuals that did not have a scan at 3 T as SOLO_1.5 T and we will refer to the two sets of 101 images from individuals that had an image at both magnetic field strengths as PAIR_1.5 T and PAIR_3.0 T respectively. All resulting 26 subgroups are listed in Supplementary Table 1. There was a trend towards age difference in two of these groups. The subgroup with lowest MMSE of the AD-p group had 22.6 ± 2.0 [18–26], and the highest MMSE score of AD-p group was 24.1 ± 2.2 [20–28] ($p = 0.03$). No significant differences in the MMSE between control groups were observed.