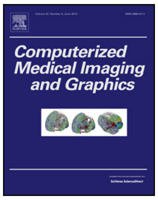
Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging☆

Shen Lu[a], Yong Xia[b,*], Weidong Cai[a,*], Michael Fulham[c,d], David Dagan Feng[a,e], Alzheimer's Disease Neuroimaging Initiative

[a] *Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, NSW 2006, Australia*
[b] *Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[c] *Department of Molecular Imaging, Royal Prince Alfred Hospital, NSW 2050, Australia*
[d] *Sydney Medical School, University of Sydney, NSW 2006, Australia*
[e] *Med-X Research Institute, Shanghai Jiaotong University, Shanghai 200030, China*

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 9 January 2017
Accepted 10 January 2017

Keywords:
FDG-PET
Alzheimer's disease
Mild cognitive impairment
Robust optimization

ABSTRACT

Alzheimer's disease (AD) is the most common type of dementia and will be an increasing health problem in society as the population ages. Mild cognitive impairment (MCI) is considered to be a prodromal stage of AD. The ability to identify subjects with MCI will be increasingly important as disease modifying therapies for AD are developed. We propose a semi-supervised learning method based on robust optimization for the identification of MCI from *[18F]Fluorodeoxyglucose* PET scans. We extracted three groups of spatial features from the cortical and subcortical regions of each FDG-PET image volume. We measured the statistical uncertainty related to these spatial features via transformation using an incomplete random forest and formulated the MCI identification problem under a robust optimization framework. We compared our approach to other state-of-the-art methods in different learning schemas. Our method outperformed the other techniques in the ability to separate MCI from normal controls.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative brain disorder that is characterized by progressive memory loss, cognitive impairment and the inability to perform usual daily activities (Teune et al., 2010). It is the most common type of dementia, accounting for about 65% of all dementia cases globally and the number of patients is increasing every year as people live longer (Devous, 2002). Mild cognitive impairment (MCI) is considered as the prodromal phase of AD (Albert et al., 2011). Individuals with MCI show greater cognitive impairment than expected for their age, but they do not meet the criteria for dementia (McKhann et al., 2011). The conversion rate of MCI to AD is estimated to be between 10%–25% per year (Grand et al., 2011). Although there are no current disease modifying agents to halt the progression of AD there are a number of clinical trials underway in patients with pre-symptomatic disease (Morris et al., 2012). Thus as effective therapies become available the early identification of patients with MCI will be of tremendous benefit to patients and their families.

The pathology of AD includes cortical and subcortical atrophy together with the deposition of β-amyloid. Two widely used AD biomarkers are structural imaging with magnetic resonance (MR) imaging (Fox and Schott, 2004) and functional imaging with *[^{18}F]Fluorodeoxyglucose* positron emission tomography (FDG-PET) (Devous, 2002). The advantage of FDG-PET over MR imaging is that PET can detect reduced cerebral glucose metabolism before structural change is evident on MR imaging. The separation of patients with MCI from normal controls (NCs) by the visual analysis of FDG-PET images, however, is difficult. Visual interpretation of these studies is also operator-dependent and related to the skill and experience of the reader. A reliable and robust computer-aided method could improve this situation.

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Ma-nuscript_Citations.pdf.

* Corresponding authors.
E-mail addresses: yxia@nwpu.edu.cn (Y. Xia), tom.cai@sydney.edu.au (W. Cai).

http://dx.doi.org/10.1016/j.compmedimag.2017.01.001
0895-6111/© 2017 Elsevier Ltd. All rights reserved.

Machine learning theory has been applied to the dementias and Davatzikos et al. used a voxel-based nonlinear multivariate analysis to separate AD from Frontotemporal dementia (FTD) with MR imaging (Davatzikos et al., 2008). In their subsequent study (Davatzikos et al., 2011), they applied a similar method to combinations of features extracted from MR images and cerebrospinal fluid (CSF) to predict progression from MCI to AD. Although there are a number of pathological studies of MCI with PET (Devous, 2002; Sun et al., 2013), the use of computerized classification methods based on PET data is not prominent in the literature. In a previous study (Xia et al., 2014a), we implemented a method that combined multi-kernel learning (MKL) and a genetic algorithm (GA) to differentiate between AD, FTD and NC with FDG-PET images. We used GA to obtain the optimal kernel weights for combining different kernel matrices and then trained a MKL machine to classify the three classes at the same time. In a subsequent study (Xia et al., 2014b), we used an automated classification method for dealing with AD and NC using infinite kernel learning (IKL). We exploited the importance of cerebral features in the AD/NC classification task using this method. We investigated the early identification of different dementia sub-types using FDG-PET and reported superior classification accuracy and efficiency, but we did not address the problem of separating MCI and NCs. Zhang et al. (2011) combined a number of biomarkers (MR, PET and CSF) together and used MKL to classify AD, MCI and NC. They reported good differentiation of AD from NC but they had a lower accuracy (76.4%) for separating MCI from NCs. In addition, in the clinical setting it is difficult to obtain all three biomarkers due to costs and the reluctance for subjects to undertake a lumbar puncture. Recently, Gray et al. (2013) proposed a multi-modality classification process based on the embedding of feature similarities among MR, FDG-PET, CSF, and genetic information via random forest (RF). They reported 75% classification accuracy between MCI and NCs which was poorer than the 89% accuracy in separating AD from NCs.

In this work our aim was the early identification of patients with MCI using FDG-PET imaging. We used an incomplete random forest – robust support vector machine (IRF-RSVM) approach to address the problem where subjects with MCI have similar imaging to NCs and the spatial resolution of FDG-PET is poorer than structural imaging. The idea was to build an incomplete random forest using FDG-PET image features and model the outputs of the random forest as a noise corrupted feature dataset, and then minimize a loss function in terms of these noisy data within a robust programming framework.

2. Background

2.1. Random forest (RF)

Random forest is an ensemble learning method, which builds a number of decision trees (Criminisi et al., 2012; Breiman, 2001) with random factors. Basically, RF injects randomness into its learning process in two forms: random sampling and random parameterization. Random sampling arbitrarily selects training examples to train each decision tree. Random parameterization chooses training parameters during the training of each decision tree in an unplanned fashion. Both or either of these two forms of randomness can be used in the training process. The introduced randomness prompts variation and diversity among the decision trees that are built. Each decision tree in the forest is a binary tree on which each non-leaf node, a so-called weak learner, is trained by solving an optimization problem to determine the best data feature to use to split the dataset. For features with a numerical value, we simply threshold the data set at the current node so that examples, where the value of the feature used for splitting is less than the threshold, go to the left branch of this node and other examples go to the right branch. The process continues on subsequent nodes until a stopping criterion is met.

2.2. Support vector machine

In general, the goal of machine learning is to learn distinguishable patterns from training data belonging to different classes, and then use these patterns to classify new (unseen) data (test data) to some extent. Kernel based maximum margin learning methods have been very widely used in machine learning research during the last decade (Schölkopf and Smola, 2002; Xu et al., 2004; Vapnik, 2017). Basically, kernel based method constructs kernels in reproducing kernel Hilbert space (RKHS) based on data, and finds a separating hyperplane that separates data belonging to different classes with maximum margins by minimizing a structural empirical risk functional (Schölkopf and Smola, 2002; Vapnik, 2017). Within this family of methods, support vector machine (SVM) is the most well-known method and has been used in many scientific and industrial applications (Schölkopf and Smola, 2002).

SVM finds the optimal separating hyperplane by solving a linearly constrained quadratic optimization problem (QP), which can be written as:

$$\underset{\mathbf{w},b,\xi}{\text{minimize}}\frac{1}{2}\|\mathbf{w}\|_2^2+C\sum_{k=1}^{p}\xi_k \tag{1}$$

$$s.t.y_k\left(\mathbf{w}^T\mathbf{x}_k+b\right)\geq 1-\xi_k,\xi_k\geq 0,\forall k=1,\ldots,p$$

where $\mathbf{x}$ is the training data vector with label $y\in\left\{-1,1\right\}$, p is the number of training data, ξ is slack variable which allows some data to be misclassified, the weight vector $\mathbf{w}$ and bias b are optimization variables that define the hyperplane. Solving the optimization problem (1) results in a separating hyperplane that separates training data, and at the same time, maximizes the margins between training data on both sides of the hyperplane (Schölkopf and Smola, 2002). After solving (1), the prediction of testing data label is made by evaluating the function below for each testing data $\mathbf{x}'$:

$$f\left(\mathbf{x}';\mathbf{w}^*,b^*\right)=sgn\left(\mathbf{w}^{*T}\mathbf{x}'+b^*\right) \tag{2}$$

where $\mathbf{w}^*$ and b^* are the optimal solutions of (1), sgn(·) gives the sign of the operator and the sign indicates the class membership of testing data $\mathbf{x}'$.

2.3. Inductive learning and transductive learning

Theoretically, there are two types of machine learning schemas, inductive and transductive learning. In the inductive learning setting, a learner is trained using a set of observed data called training data and is then tested on a set of previously unseen data called test data. This setting is extremely common in machine learning research. Transductive learning differs from inductive learning in that, during the training phase a participant has visibility of training data and test data and a participant can potentially make use of the information, exposed by the test data, such as the probability distribution information (Vapnik, 2017; Joachims, 2017). Hence, transductive learning is ideal when the size of the experimental data is small. In this work we tested the proposed method in the inductive and transductive learning settings.

3. Data and materials

The FDG-PET image data we used were from the Alzheimer's Disease Neurodegenerative Initiative (ADNI) cohort (http://adni.loni.usc.edu). ADNI is a multi-center program funded by a public-private

Table 1
T-test results of comparing mean glucose metabolism values within ROIs containing more than 100 voxels between MCI group and CN group.

	MCI Group	NC Group	p-value
Mean ROI glucose metabolism of left angular gyrus	1.20	1.33	p = 3.104e-06
Mean ROI glucose metabolism of right angular gyrus	1.22	1.32	p = 7.071e-05
Mean ROI glucose metabolism of posterior cingulate	1.29	1.42	p = 1.018e-05

Table 2
Summary of demographics of data set with mean ± std.

	Subjects	
	MCI	NC
Number of subjects	120	152
Age (years)	76.06 ± 7.68	76.35 ± 4.85
Weight (kg)	76.31 ± 15.27	75.84 ± 14.28
Gender (M/F)	79/41	80/72
MMSE	26.20 ± 1.82	28.89 ± 1.53

partnership and non-profit organizations to provide standardized longitudinal medical image data to global researchers for neurodegenerative disease research. In ADNI 1 all subjects were followed for 2–3 years and assessed every 6–12 months. In our study, we used 272 FDG-PET MCI and NC studies from ADNI; 120 MCI subjects and 152 NCs. All MCI subjects had Mini-Mental State Examination (MMSE) scores between 24 and 30 (inclusive), Clinical Dementia Rating (CDR) of 0.5 and no sign of significant levels of impairment in other cognitive domains. We also verified the positivity of the FDG-PET MCI scans by a two-group independent *t*-test between the MCI group and NC group. We downloaded the FDG-PET study conducted by the University of California, Berkeley from ADNI. This study contains mean glucose metabolism (normalized to pons) of angular gyrus, temporal lobes and bilateral posterior cingulated for all FDG-PET scans included in our study. Following the approach of Anchisi et al. (2005), we used regions of interest containing more than 100 voxels from MCI group and NC group in our paired *t*-test. The *t*-test results are shown in Table 1. The results show that all MCI subjects in our study are FDG-PET positive ($p < 0.001$). All images used in our study were baseline/initial scans; these data had been through a pre-processing pipeline that included: co-registration, averaging, voxel normalization, and isotropic Gaussian smoothing (ADNI, 2016). This pre-processing work is done by the ADNI participants and it makes any subsequent analysis simpler as the data from different PET scanners are then uniform. The demographic information of all 272 subjects and the Mini-Mental State Examination (MMSE) scores are shown in Table 2.

4. Methods

4.1. Feature extraction

Our aim was to extract spatial features from voxel volumes representing cerebral cortical and subcortical regions on each PET image. To ensure good spatial localization we registered each PET image to a brain atlas. We used the automated anatomical labeling (AAL) cortical parcellation map (Tzourio-Mazoyer et al., 2002) to identify the anatomical volumes of interest (VOIs) where spatial features were to be extracted. Sun et al. (2013) reported on the important role that the AAL map plays in computer-based functional brain image analysis for identifying dementia. The AAL image template contains 116 manually drawn and accurately reconstructed anatomical VOIs, and it has dimension of 91 × 109 × 91 with voxel size of $2 \times 2 \times 2\,\text{mm}^3$. To achieve the best image registration result, we spatially normalized each of the study images to the PET image template provided by statistical parametric mapping (SPM) software. This PET template has the same dimension and voxel size as the AAL template. We then registered the AAL template to each normalized FDG-PET image using SPM with nearest-neighbour interpolation to obtain individual-level AAL template. Fig. 1(a)/(b) shows the normalized FDG-PET image and individual-level AAL template for MCI/NC subject.

We overlaid the individual-level AAL template on top of the normalized FDG-PET image and identified the 116 anatomical VOIs on each FDG-PET image. We normalized the voxel values of each of the 116 VOIs on each FDG-PET image to the corresponding subject's cerebellum vermis, whose locations and voxels were also labeled on the individual-level AAL template. These normalized voxel intensity values were then mapped to the range [0, 1]. After that, we extracted three groups of spatial features from the 116 VOIs defined on the AAL template. They were: mean voxel values from the 116 VOIs, standard deviations of voxel values from the 116 anatomical VOIs, and mean voxel value differences between 54 pairs of the anatomical VOIs on left and right brain hemispheres. We then concatenated these three feature groups together to form a feature vector of dimension 286 for each image. The main reason for extracting these spatial features from the normalized images is that the glucose consumption metabolic rates in cortical regions, which is reflected on PET image voxel intensities, offers a major clue for dementia diagnosis (Teune et al., 2010). In addition, the work of Frisoni et al. (2007) also suggested that the metabolic asymmetry in the left-right brain regions caused by atrophy strongly connects to the factors causing dementia symptoms. Therefore, we employed the mean and standard deviation of voxel values of each cortical volume as features to describe the spatial characteristics of these volumes. Higher order spatial features were not used since it would increase the feature dimensionality significantly.

Let $\mathbf{X} \in \mathbb{R}^{272\times 286}$ denote the column matrix containing all spatial features, and let $\mathbf{x}_i = \left\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,286}\right\}^T, i = 1, \ldots, 272$ be the feature vector for the ith image, where T is matrix transpose. Finally, let $\mathbf{Y} = \left\{-1, 1\right\}, \mathbf{Y} \in \mathbb{R}^{272}$ denote the label vector (MCI: −1, NC: 1) for the images. Note that $\mathbf{X}$ is mean centered and standardized.

4.2. Feature transformation

In our method we do not use the feature matrix $\mathbf{X}$ directly. Instead, we used a transformed version of $\mathbf{X}$, because we wanted to better model the classification problem with noise corrupted images in the robust optimization framework. We attempted to model image noise caused by perturbation to the data $\mathbf{X}$ as a perturbation to the statistical distribution of $\mathbf{X}$. Therefore, we use $\tilde{\mathbf{X}}$ to denote the transformed data matrix, and $\tilde{\mathbf{Y}}$ the transformed label vector.

The data transformation process took the form of incomplete random forest, whose main difference compared to the classic random forest is that the decision trees in the incomplete random forest are never fully grown. That is to say, the training cycle of each decision tree was terminated before it reached the state when each leaf tree node only contains data points from single class or single data point. The main reasons for using this variation of random forest are two folds. Firstly, the computational cost of training a random forest containing hundreds of fully grown tree in our study might be prohibitive. Secondly and more importantly, using incomplete random forest allows us to obtain 'clusters' of data points at the leaf nodes of each tree. These clusters contain sufficient amount

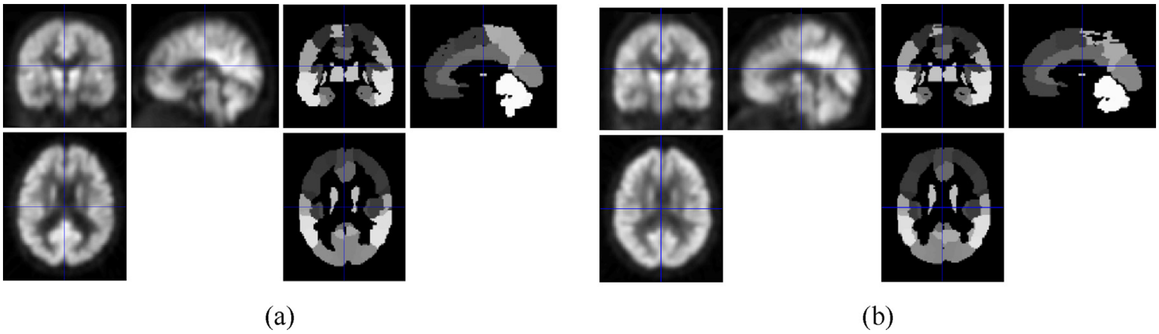


Fig. 1. Normalized FDG-PET image (left) and individual-level AAL template (right) for (a) MCI and (b) NC subject.

of data belonging to the two different classes (MCI and NC), where the modes of their probability distributions can be properly estimated. In contrary, for a fully grown decision tree, the statistics of the probability distributions at leaf nodes cannot be estimated properly due to large portions of leaf nodes containing only single or very few data points. Our method treated all data points at the tree leaf nodes as noise corrupted data and replaced them with statistics estimated based on data in their local neighbourhood.

Let F be the incomplete random forest we build and denote the decision tree as T_m, $m = 1, 2, ..., N_T$, where N_T is the total number of trees in F. T_m is only allowed to grow up to d level where d is a predefined parameter.

To construct F, we iteratively built each decision tree T_m using $\mathbf{X}$ by branching $\mathbf{X}$ at each non-leaf tree node following top-down order. Starting from the root node of T_m, at each non-leaf node of the tree we randomly selected a number of different features $k = 1, ..., n_k$ (where n_k is a predefined parameter) from the 286 spatial features and calculate the branching threshold θ_k using

$$\theta_k = \left(\max\left(\mathrm{x}_{:,k}\right) - \min\left(\mathrm{x}_{:,k}\right)\right)/2 \tag{3}$$

where max and min indicate the maximum and minimum values in a vector, respectively. $\boldsymbol{x}_{:,k} \in \mathbb{R}^{272}$ is the column vector of kth feature values in x_i, i = 1, ..., 272. We then selected the best branching threshold θ^* from the candidates $\theta_1, ..., \theta_{n_k}$ by solving the optimization problem below

$$\theta^* = \underset{\theta_k}{\mathrm{argmin}}\, I,\ k = 1, ..., n_k \tag{4}$$

where I is the unsupervised information gain (Criminisi et al., 2012) defined by

$$I = \log\left(|\Lambda\left(S_h\right)|\right) - \sum\nolimits_{b=\{L,R\}} |S_h^b| \log\left(|\Lambda\left(S_h^b\right)|\right)/|S_h| \tag{5}$$

where h is the current non-leaf node being branched, $S_h \subset \mathbf{X}$ is the dataset at node h before branching, b is the branching direction which can only be either L – left branch of node h or R – right branch of node h, $S_h^b \subset S_h$ is thus the dataset assigned to the respective branch (left or right) of node h, Λ is covariance operator and $|\cdot|$ is set cardinality. Note that the calculated unsupervised information gain I may not be a real number due to the presence of the covariance operator, in which case the candidate θ_k is discarded and a new θ is randomly selected to replace it. This branching/optimization process is carried on until the predefined tree depth d is reached.

Once every T_m in F is built following the procedure outlined above, the transformed feature data and labels are collected from leaf nodes of each tree T_m. Each leaf node of T_m is treated as a subspace containing two clusters, one for each of the two data classes (MCI, NC). It is straightforward to calculate the mean μ and covariance $\sum$ from each cluster. For T_m we obtain $\boldsymbol{\mu}_m = \left\{\mu_+^1, ..., \mu_+^l, \mu_-^1, ..., \mu_-^l\right\}^T$, $\sum_m = \left\{\sum_+^1, ..., \sum_+^l, \sum_-^1, ..., \sum_-^l\right\}^T$, and $\mathbf{y}_m = \left\{1, ..., 1, -1, ..., -1\right\}^T$, $|\mathbf{y}_m| = |\boldsymbol{\mu}_m| = |\sum_m|$ where l is the number of leaf nodes, + is MCI class, – is NC class, $\mathbf{y}$ is the corresponding label vector. To this end, the transformed data is $\tilde{\boldsymbol{X}} = \left\{\mathbf{S}_\mu, \mathbf{S}_{\sum}\right\} = \left\{\left\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m\right\}, \left\{\sum_1, ..., \sum_m\right\}\right\}$ and the transformed label vector is $\tilde{\boldsymbol{Y}} = \left\{\mathbf{y}_1, ..., \mathbf{y}_m\right\}$.

4.3. Classification

In the classification stage, we train a classifier within the robust optimization (RO) framework (Ben-Tal and Nemirovski, 1998) using transformed data $\left\{\mathbf{S}_\mu, \mathbf{S}_{\sum}, \tilde{\boldsymbol{Y}}\right\}$. Assuming that $\left\{\mathbf{S}_\mu, \mathbf{S}_{\sum}, \tilde{\boldsymbol{Y}}\right\}$ are noise corrupted, which is appropriate as it is accepted that FDG-PET images usually have a low signal-to-noise ratio due to the limited resolution of PET scanners (Feng, 2011), to model the uncertainty associated with these noisy data, we consider the modified version of the inequality constraint in the original SVM problem

$$Pr\left(y_k\left(\mathbf{w}^T\mathbf{x}_k + b\right) \geq 1 - \xi_k\right), \xi_k \geq 0, \forall k = 1, ..., p \tag{6}$$

where $\delta \in [0, 1)$ is a user defined parameter. The probabilistic constraint simply requires each feature vector $\mathbf{x}_k$ to lie on the correct side of the optimal hyperplane with a certain confidence value δ. Solving SVM problem with this constraint is extremely difficult. Therefore, we transformed it into a deterministic constraint with the assumption that the feature data is drawn from a multimodal Gaussian distribution characterized by mean and covariance (Shivaswamy et al., 2006). Our transformed datasets $\left\{\mathbf{S}_\mu, \mathbf{S}_{\sum}, \tilde{\boldsymbol{Y}}\right\}$ naturally fit into this new deterministic constraint, which is written as

$$\tilde{\boldsymbol{Y}}\left(\mathbf{S}_\mu\mathbf{w} + b\right) \geq 1 - \boldsymbol{\xi} + \boldsymbol{\gamma}\|\mathbf{S}_{\sum}^{1/2}\mathbf{w}\|_2 \tag{7}$$

where we introduce a new parameter vector $\boldsymbol{\gamma}$, $|\boldsymbol{\gamma}| = |\tilde{\boldsymbol{Y}}|$. $\boldsymbol{\gamma}$ is computed from the leaf nodes of decision trees in a way similar to (Huang et al., 2013). For *i = 1, ..., lm*, where *l* is the number of leaf nodes of each tree, *m* is the number of trees in forest *F*

$$\gamma_i = \begin{cases} n_i/n_{MCI}, y_i = 1 \\ n_i/n_{NC}, y_i = -1 \end{cases} \tag{8}$$

where n_i is the number of feature vectors dwelled at leaf node i, n_{MCI} and n_{NC} are the total number of MCI cases and NC cases in the whole dataset, respectively. Finally, the robust SVM problem is formulated as

$$\underset{\mathbf{w},b,\xi}{\text{minimize}}\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{k=1}^{p}\xi_k \tag{9}$$

$$s.t.\tilde{\boldsymbol{Y}}\left(\mathbf{S}_{\mu}\mathbf{w}+b\right) \geq 1 - \boldsymbol{\xi} + \boldsymbol{\gamma}\|\mathbf{S}_{\sum}^{1/2}\mathbf{w}\|_2$$

$$\boldsymbol{\xi} = \left\{\xi_1, \ldots, \xi_k\right\}, \xi_i \geq 0, i = 1, \ldots, k$$

Evidently, (9) is a convex problem. In our study, we solve this problem using CVX Matlab toolbox (Grant et al., 2008). In order to efficiently solve this problem with CVX, we reformulate (9) into an equivalent second-order cone programming (SOCP) problem

$$\underset{\mathbf{w},b,\xi}{\text{minimize}} t + C\sum_{k=1}^{p}\xi_k \tag{10}$$

$$s.t.\|\mathbf{w}\|_2 \leq t$$

$$s.t.\tilde{\boldsymbol{Y}}\left(\mathbf{S}_{\mu}\mathbf{w}+b\right) \geq 1 - \boldsymbol{\xi} + \boldsymbol{\gamma}\|\mathbf{S}_{\sum}^{1/2}\mathbf{w}\|_2$$

$$\boldsymbol{\xi} = \left\{\xi_1, \ldots, \xi_k\right\}, \xi_i \geq 0, i = 1, \ldots, k$$

Once the optimal solution $\left\{\mathbf{w}^*, b\right\}$ is found by solving (10), predictions of feature vectors extracted from new PET image are made by evaluating Function (2).

5. Experiments

5.1. Benchmark methods

We compare the proposed RF-RSVM method to three baseline methods:

1. Supervised SVM (Schölkopf and Smola, 2002): We applied the soft margin SVM as described in the section on background.
2. Laplacian SVM (LapSVM) (Belkin et al., 2006; Melacci and Belkin, 2011): LapSVM regularizes the standard SVM cost function with a data dependent penalty term with the assumption that the intrinsic structure of the data is embedded within a low dimensional manifold. It approximates this new penalty term by modeling the structure of the data using graph Laplacian.
3. Method proposed by Huang et al. (2013): Huang et al. conducted clustering based on a dataset using the k-nearest neighbour algorithm, and then merged similar clusters, followed by solving the SOCP problem (10). The method showed good performance on non-medical imaging datasets.

Only the soft margin SVM is supervised learning method while the other two methods are both semi-supervised.

5.2. Experimental settings

To ensure that our method had good generalizability, we applied 3-fold cross validation for our method and the three benchmark methods. We first divided the whole dataset evenly into 3 subsets (the residual is randomly assigned to one of the subsets), each contained 20% labeled data examples and the rest of data were treated as unlabeled. This is corresponding to the scenario in clinical settings, the number of images without definite diagnosis are usually much larger than the number of accurately labeled images. The aim of our study is to design and implement a method jointly using unlabeled and labeled images to accurately separate MCI subjects and NCs. Within the labeled and unlabeled groups of data in each subset, we further restricted that 50% of data in this group were MCI subjects and 50% were NCs. We used the inductive learning schema in our first experiment, and trained the target classifier using any 2 out of the 3 subsets, then tested the target classifier on the leftover subset. Initially, $d = 3$, $N_T = 50$ were used to construct the unsupervised random forest in RF-RSVM. Hyperparameters required in the benchmark methods were set empirically or selected by an inner 3-fold cross-validation using the training data.

Table 3
Classification performance of the proposed method and the baseline methods.

Method	Accuracy	Sensitivity	Specificity
RF-RSVM	90.53%	90.63%	93.33%
LapSVM	71.71%	74.04%	71.11%
kNN-SVM	84.27%	77.43%	89.23%
Supervised SVM	87.30%	78.13%	96.77%

5.3. Results and discussion

We applied the proposed method (RF-RSVM) and the baseline methods to classify MCI and NC. The performance of these methods measured by classification accuracy, sensitivity and specificity averaged over the 3 fold cross-validation steps are shown in Table 3. The plain supervised SVM generated the worst results possibly due to the high non-linearity and high similarity of the feature patterns in our dataset. The RF-RSVM outperformed the other two semi-supervised learning methods in terms of accuracy, sensitivity and specificity. The less impressive performance of LapSVM may be related to that the intrinsic structure of our data, which to some extent, violates the smooth manifold assumption that is crucial for LapSVM to perform well. We also depicted the receiver operating characteristic (ROC) curves of the performances of RF-RSVM, LapSVM and kNN-SVM. ROC curve is known as an effective measure of accuracy of diagnostic tests (Hajian-Tilaki, 2013). On each ROC curve, true positive rate (sensitivity) is plotted against false positive rate (1–specificity). True positive rate (TPR) is the conditional probability of correctly identifying the MCI subjects and false positive rate (FPR) is the conditional probability of incorrectly identifying the MCI subjects. Let $\mathbf{T}_+$ denote subjects who were predicted as MCI. Let $\mathbf{D}_+$ and $\mathbf{D}_-$ denote MCI subjects and NCs, respectively. TPR and FPR can be written as

$$TPR = Pr\left(\mathbf{T}_+|\mathbf{D}_+\right) \tag{11}$$

$$FPR = Pr\left(\mathbf{T}_+|\mathbf{D}_-\right) \tag{12}$$

Define the prior probability of identifying a subject as MCI to be i, $i = \left\{0.1,\ 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\right\}$. The ROC curve was generated by computing the probability of MCI ($\mathbf{D}_+$) conditioned on the predicted MCI subjects ($\mathbf{T}_+$) with varying i using Bayes' theorem

$$Pr\left(\mathbf{D}_+|\mathbf{T}_+\right) = \frac{Pr\left(\mathbf{T}_+|\mathbf{D}_+\right)Pr\left(\mathbf{D}_+\right)}{Pr\left(\mathbf{T}_+\right)} \tag{13}$$

where $Pr\left(\mathbf{D}_+\right) = i$, $Pr\left(\mathbf{T}_+\right)$ is the marginal probability defined as

$$Pr\left(\mathbf{T}_+\right) = Pr\left(\mathbf{T}_+|\mathbf{D}_+\right)Pr\left(\mathbf{D}_+\right) + Pr\left(\mathbf{T}_+|\mathbf{D}_-\right)Pr\left(\mathbf{D}_-\right) \tag{14}$$

where $Pr\left(\mathbf{D}_-\right) = 1 - i$. All these probabilities can be computed from the 2-by-2 confusion table obtained from the results of the classification method. ROC curves for these RF-RSVM, LapSVM and kNN-SVMare shown in Fig. 2 complement the findings in Table 3.

The inductive learning scheme used in the first experiment validates the generalizability of classifier built using training data. Transductive learning, on the other hand, carries out training and testing on the same dataset. It is very useful when the size of

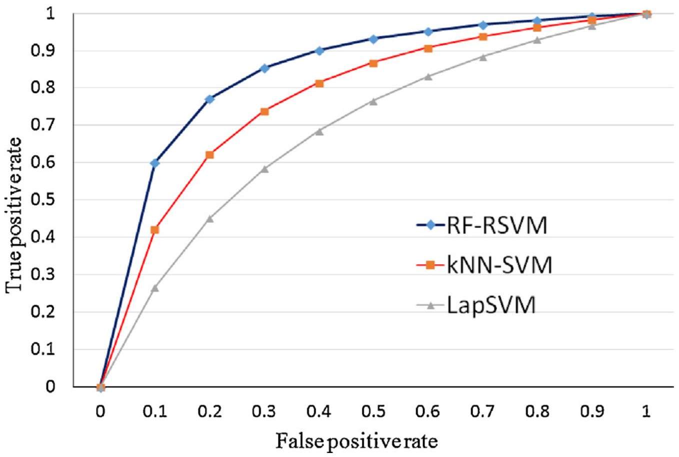


Fig. 2. ROC curves for methods (excl. supervised SVM) compared in Table 3.

Table 4
Classification performance of the proposed method and the baseline methods under transductive learning setting.

Method	Accuracy	Sensitivity	Specificity
RF-RSVM	92.16%	88.46%	96%
LapSVM	70%	80%	60%
kNN-SVM	76.47%	65.38%	88%

the dataset (training + testing) is small. Since the dataset used in dementia related studies usually does not contain tens of thousands images, the proposed method could be tested under the transductive learning setting. For each of the 3 subsets created before, we trained our method, LapSVM, and kNN-SVM in a way that both labeled and unlabeled data within each subset were used for training while only unlabeled data within each subset was used for testing. The same performance metrics as those shown in Table 3 were used and they were averaged over the 3-fold cross validation. Table 4 shows the final performance of the three methods.

When dealing with random forest an obvious question is what is the impact that the hyperparameters such as tree depth d and the number trees N_T have on the performance of the proposed method. We carried out experiments to apply RF-RSVM with varying d and N_T to measure the impact. We fixed $d = \{2, 3, 4, 5\}$, respectively, and then increased the number of trees in the random forest from 10 to 50 with increments of 5. The metrics used were identical to the first experiment and the performance charts for these two scenarios are shown in Fig. 3. Fig. 3 shows that increasing d did not improve the overall performance of the proposed method.

In machine learning, when the whole training dataset is labeled (e.g. each PET image contained in the dataset is given a class: MCI or NC), the learning process is called supervised learning, whereas it is called semi-supervised learning if large part of the training dataset are unlabeled. So if we denote the total number of data examples contained in a dataset as N. Let N_L and N_U be the number of labeled data and unlabeled data. $N = N_L + N_U$, and usually, $N_L \leq N_U$. Therefore, semi-supervised learning can play an important role in solving practical problems when most of data labels are unavailable due to the high cost of manual data labeling or when full data labeling is not possible. Our method is essentially a semi-supervised learning method, which is appropriate, since in the clinical setting brain images labeled as dementia are usually not available given the difficulty in making an accurate diagnosis without a post-mortem. The number of unlabeled brain images, or brain images which are suspected to reflect dementia, are abundant.

One of the most important components/processes in our method is feature transformation via incomplete random forest. This transformation is the key to modeling the MCI/NC classification under RO framework. This transformation also introduces some problems. For example, after a decision tree is constructed it is not guaranteed that each leaf node will always contain some feature vectors belonging to MCI and some belonging to NC. Some leaf nodes may only contain feature vectors belonging to a single class – we call those leaf nodes degenerative leaf nodes. We discard all degenerative leaf nodes to avoid numerical difficulties. The main problem with the feature transformation process is a long training time. This issue can be seen from the first constraint in the optimization problem (10). Recall that d is the depth of each decision tree in forest F. Since the decision tree is a binary tree, the number of leaf nodes each decision tree can have is $2^d - 2^{d-1}$, thus the total number of leaf nodes in forest F is $N_T\left(2^d - 2^{d-1}\right)$. Each leaf node contains two clusters (one for MCI, one for NC), as a result, the upper bound of the number of constraints is $2N_T\left(2^d - 2^{d-1}\right)$ (this is an upper bound since some leaf nodes may be degenerative). It is easy to have tens of thousands of constraints even with a moderate number of decision trees and tree depth. This greatly decreases the efficiency of our method. A simple strategy to alleviate this effect would be to combine μ and $\sum$ for each data category (MCI and NC) within each tree by calculating their arithmetic means and we applied this strategy to all our experiments.

6. Conclusion

We implemented a novel computer-aided method for the early identification of baseline MCI subjects among NCs using FDG-PET image data obtained from the ADNI cohort. We formulated the problem within a robust optimization framework with feature data transformed via incomplete random forest to enable semi-supervised learning. Our results show that our method outperforms two other semi-supervised learning methods. In future work we will test the performance of our method on a much larger dataset to determine if the current results are sustained over a larger dataset.

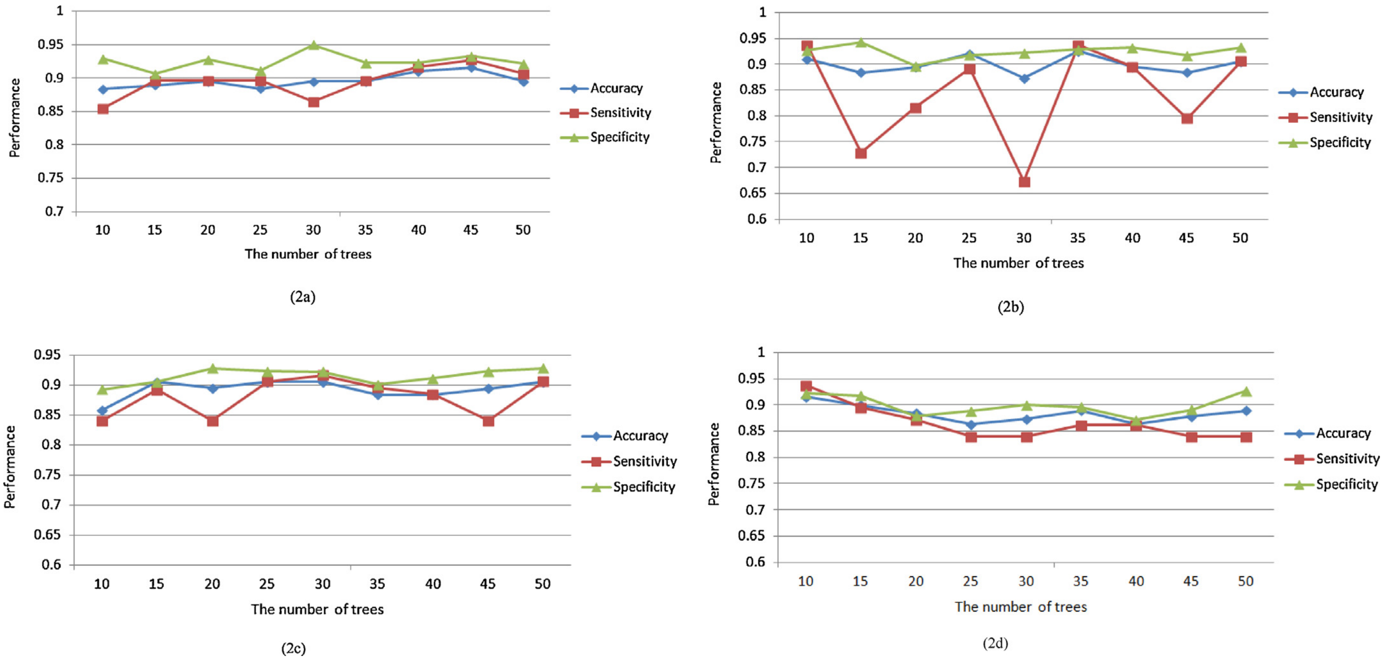
Conflict of interest statement

The authors do not hold any conflicts of interest that could inappropriately influence this manuscript.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61471297, in part by the Natural Science Foundation of Shaanxi Province, China, under Grant 2015JM6287, in part by the Fundamental Research Funds for the Central Universities under Grants 3102014JSJ0006, and in part by the Returned Overseas Scholar Project of Shaanxi Province, China.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis



(2a)

(2b)

(2c)

(2d)

Fig. 3. Accuracy, sensitivity, and specificity changes as the number of trees increases with the tree depth fixed as 2 (2a), 3 (2b), 4 (2c) and 5 (2d).

Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

ADNI, 2016. http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/.

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., et al., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. 7, 270–279.

Anchisi, D., Borroni, B., Franceschi, M., et al., 2005. Heterogeneity of brain glucose metabolism in mild cognitive impairment and clinical progression to alzheimer disease. Arch. Neurol. 62, 1728–1733.

Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7, 2399–2434.

Ben-Tal, A., Nemirovski, A., 1998. Robust convex optimization. Math. Oper. Res. 23, 769–805.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends® Comput. Graph. Vis. 7, 81–227.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41, 1220–1227.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32, 2322.e19–2322.e27.

Devous, M.D., 2002. Sr Functional brain imaging in the dementias: role in early detection, differential diagnosis, and longitudinal studies. Eur. J. Nucl. Med. Mol. Imaging 29, 1685–1696.

Feng, D.D., 2011. Biomedical Information Technology. Academic Press.

Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. Lancet 363, 392–394.

Frisoni, G.B., Pievani, M., Testa, C., Sabattoli, F., Bresciani, L., Bonetti, M., et al., 2007. The topography of grey matter involvement in early and late onset Alzheimer's disease. Brain 130, 720–730.

Grand, J., Caspar, S., Macdonald, S., 2011. Clinical features and multidisciplinary approaches to dementia care. J. Multidiscip. Healthc. 4, 125–147.

Grant, M., Boyd, S., Ye, Y., 2008. CVX: Matlab Software for Disciplined Convex Programming.

Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 65, 167–175.

Hajian-Tilaki, K., 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J. Intern. Med. 4, 627–635.

Huang, G., Song, S., Gupta, J.N., Wu, C., 2013. A second order cone programming approach for semi-supervised learning. Pattern Recogn. 46, 3548–3558.

Joachims, T., 2017. Transductive Learning via Spectral Graph Partitioning. ICML, 2003, pp. 290–297.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., et al., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. 7, 263–269.

Melacci, S., Belkin, M., 2011. Laplacian support vector machines trained in the primal. J. Mach. Learn. Res. 12, 1149–1184.

Morris, J.C., Aisen, P.S., Bateman, R.J., Benzinger, T.L.S., Cairns, N.J., Fagan, A.M., et al., 2012. Developing an international network for Alzheimer's research: the dominantly inherited Alzheimer network. Clin. Investig. 2, 975–984.

Schölkopf, B., Smola, A.J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond Cambridge, Mass. MIT Press.

Shivaswamy, P.K., Bhattacharyya, C., Smola, A.J., 2006. Second order cone programming approaches for handling missing and uncertain data. J. Mach. Learn. Res. 7, 1283–1314.

Sun, H., Hu, B., Yao, Z., Jackson, M., 2013. A PET study of discrimination of cerebral glucose metabolism in Alzheimer's disease and mild cognitive impairment. In: Biomedical Engineering and Informatics (BMEI), 2013, 6th International Conference on: IEEE, pp. 6–11.

Teune, L.K., Bartels, A.L., de Jong, B.M., Willemsen, A.T., Eshuis, S.A., de Vries, J.J., et al., 2010. Typical cerebral metabolic patterns in neurodegenerative brain diseases. Mov. Disord. 25, 2395–2404.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289.

Vapnik, V., 2017. The Nature of Statistical Learning Theory. Springer Science & Business Media, 2013.

Xia, Y., Lu, S., Wen, L., Eberl, S., Fulham, M., Feng, D.D., 2014a. Automated identification of dementia using FDG-PET imaging. BioMed Res. Int. 2014.

Xia, Y., Lu, S., Wei, W., Feng, D.D., Zhang, Y., 2014b. Non-sparse infinite-kernel learning for automated identification of Alzheimer's disease using PET imaging. In: Control Automation Robotics & Vision (ICARCV), 2014, 13th International Conference on: IEEE, pp. 855–860.

Xu, L., Neufeld, J., Larson, B., Schuurmans, D., 2004. Maximum margin clustering. Adv. Neural Inf. Process. Syst., 1537–1544.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867.