Published in final edited form as:

Comput Med Imaging Graph. 2018 July ; 67: 21–29. doi:10.1016/j.compmedimag.2018.04.002.

Exploring diagnosis and imaging biomarkers of Parkinson's disease via iterative canonical correlation analysis based feature selection

Luyan Liu[a], **Qian Wang**[a,*], **Ehsan Adeli**[b], **Lichi Zhang**[a,b], **Han Zhang**[b], and **Dinggang Shen**[b,c,**]

[a]Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, China

[b]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, United States

[c]Department of Brain and Cognitive Engineering, Korea University, Seoul, 02841, Republic of Korea

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder that progressively hampers the brain functions and leads to various movement and non-motor symptoms. However, it is difficult to attain early-stage PD diagnosis based on the subjective judgment of physicians in clinical routines. Therefore, automatic and accurate diagnosis of PD is highly demanded, so that the corresponding treatment can be implemented more appropriately. In this paper, we focus on finding the most discriminative features from different brain regions in PD through T1-weighted MR images, which can help the subsequent PD diagnosis. Specifically, we proposed a novel iterative canonical correlation analysis (ICCA) feature selection method, aiming at exploiting MR images in a more comprehensive manner and fusing features of different types into a common space. To state succinctly, we first extract the feature vectors from the *gray matter* and the *white matter* tissues separately, represented as insights of two different anatomical feature spaces for the subject's brain. The ICCA feature selection method aims at iteratively finding the optimal feature subset from two sets of features that have inherent high correlation with each other. In experiments we have conducted thorough investigations on the optimal feature set extracted by our ICCA method. We also demonstrate that using the proposed feature selection method, the PD diagnosis performance is further improved, and also outperforms many state-of-the-art methods.

Keywords

Iterative canonical correlation analysis; Feature selection; Imaging biomarkers; Diagnosis; Parkinson's disease

[*]Corresponding author at: Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University. [**]Corresponding author at: Department of Radiology and BRIC, University of North Carolina at Chapel Hill, United States.

1. Introduction

Parkinson's disease (PD) is an overwhelmingly neurodegenerative disorders which often starts at 40 years' old and with the age-related deterioration. It also progresses slowly. With PD, patients' movement, balance, and muscle control would be affected. It is still not clear about what causes PD, and there is limited effective treatment. However, the clinical diagnosis of PD is usually prone to subjective biases since it mostly relies on evaluating substantial symptoms of the patients (Calne et al., 1992). As an alternative, machine learning technique can be used as a solution to analyze medical data intelligently and produce the diagnosis automatically. Therefore, there are a lot of studies focus on computer-assisted diagnosis solution for PD, which is able to take advantage of all available data to diagnose and identify biomarkers related to PD (Goebel et al., 2011; Tsanas et al., 2012).

There are several studies in the literature aiming at distinguishing PD from normal control (NC), as well as other similar neural disorders. Goebel et al. (2011) analyzed the single-photon emission computed tomography (SPECT) images, and developed an observer-independent method to classify PD from multiple system atrophy Parkinson (MSA-P), progressive supranuclear palsy (PSP), and normal control automatically. Tsanas et al. (2012) proposed a novel speech signal processing algorithm to compute dysphonia measures, which contribute to better differentiating PD patients from healthy controls. Wenning et al. (2000) evaluated various clinical features (including response to levodopa, motor fluctuation, rigidity, dementia, speech) for distinguishing multiple system atrophy (MSA) from PD. Singh and Samavedham (2015) proposed a novel synergetic paradigm that integrates Kohonen self-organizing map (KSOM) to extract features from magnetic resonance (MR) images for individual-level clinical diagnosis of neurodegenerative diseases. Highly accurate diagnosis is then achieved with the least-square support vector machine (LS-SVM). Chen et al. (2013) proposed an effective and efficient system using fuzzy k-nearest neighbor (FKNN) for PD diagnosis on biomedical voice measurement data. Their experimental results demonstrate that the FKNN system greatly outperforms the SVM-based approaches and other methods in the literature.

Most of the existing works for computer-assisted PD diagnosis study on clinical data or voice measurements, rather than neuroimaging data. Few of them utilize sophisticated machine learning tools to select features and build models for PD diagnosis and biomarker identification. Our goal in this paper is finding a non-invasive and low-cost solution to early PD screening using T1-weighted MR images, which is more available in less-developed areas with limited access to healthcare resources. Our main contribution is to propose a novel iterative canonical correlation analysis (ICCA) based feature selection method to optimize the diagnosis process. Specifically, we commence by extracting features from the T1-weighted MR images, which are associated with individual regions-of-interest (ROIs). The ROIs are divided into two sets, corresponding to white matter (WM) and gray matter (GM), respectively. Therefore, the extracted features can be naturally separated into two groups, corresponding to the descriptions of the patients from the WM and GM perspectives. Next, we adopt canonical correlation analysis (CCA) and propose the iterative CCA (ICCA) method, in order to transform the WM and GM features into an optimal common feature space. The WM and GM features are transformed and then selected in a common space with

ICCA. Finally, the selected optimal features are used for establishing a robust linear discriminant analysis (RLDA) model, which classifies PD patients from normal subjects for disease diagnosis.

Feature selection has proven its importance in tackling problems in translational medical studies including computer aided PD diagnosis. For example, in (Ozcift, 2012), a feature selection scheme based on support vector machine (SVM) was used to help train the rotation forest (RF) ensemble classifiers to improve the diagnosis of PD. The records of voice measurements where each record includes 22 features were treated as input. They firstly selected 10 most important features using linear SVM, and then trained 6 classifiers with the selected features. Subsequently, RF ensemble classification strategy was used to generate the final result. Though the method can get good performance in PD diagnosis, abundant features extracted from medical imaging that prove to be effective and convenient for disease diagnosis are largely ignored (Aël Chetelat and Baron, 2003). In (Adeli et al., 2016), magnetic resonance imaging (MRI) data was used to diagnose PD by the joint feature-sample selection (JFSS) method, which learned the robust and reliable diagnosis model on the selected optimal subset of samples and features only. The JFSS model can discard poor samples and redundant features, where it is difficult to be applied to the small sample problem. In (Cao et al., 2017), multiple heterogeneous features were extracted from medical images, and a multi-kernel framework for feature selection and unevenly distributed data learning was proposed for computer-aided detection of lung nodules. The proposed framework solved the challenging, which were the multiple heterogeneous and high dimensional irrelevant features and the imbalanced distribution between nodule and non-nodule classes, in the field of recognition of nodule. In (Ye et al., 2012), complementary phenotypes for AD including neuroimaging, demographic, genetic and cognitive measures were used for the diagnosis and prognosis of AD. The sparse logistic regression technique (Ye and Liu, 2012) was applied to find the optimal set of features. In (López et al., 2009), principal component analysis (PCA) was used to reduce the dimension of feature space to assist the diagnosis of Alzheimer's disease using single photon emission computed tomography (SPECT) and positron emission tomography (PET). In (Avants et al., 2010), sparse canonical correlation analysis (SCCA) was proved to be a valuable method to distinguish differences between neurodegenerative conditions. With SCCA, authors found that white matter integrity measured by fractional anisotropy (FA) from diffusion tensor imaging (DTI) reduced and cortical thickness measured by high resolution T1-weighted imaging decreased in Alzheimer's disease (AD) and frontotemporal dementia (FTD). Besides, minimal-redundancy-maximal-relevance (mRMR) proves to be an efficient feature selection method in four different datasets including handwritten digits, arrhythmia, NCI cancer cell lines, and lymphoma tissues (Peng et al., 2005). Nie et al. (2010) proposed a robust feature selection method by minimizing the joint l_{21} -norm-based metric during the optimization. The proposed method has shown high performance in five publicly available microarray datasets and also one Mass Spectrometry dataset.

In this paper we further explore the CCA technique, which is capable of establishing the relationship between two high-dimensional vectors of features, and implementing linear mapping to transform them to a common feature space (Zhu et al., 2014). Note that the two feature vectors in our study are extracted from WM and GM regions respectively,

representing two views of different anatomical feature spaces from each patient/normal subject. These two feature vectors thus need to be transformed to a common space, where all features can be evaluated in accordance to their contributions toward PD classification jointly and fairly. Specifically, after linearly transforming the two views of features to a common space estimated by CCA, we learn a regression model to fit the PD/NC labels of the training subjects based on the transformed feature representations. The regression model not only predicts the unknown labels of new test subjects, but also helps identify the importance of relevant features for PD classification. The identified features are often perceived as important biomarkers, which help researchers better understand the mechanism and evolution of the disease potentially. In this paper, we extracted features as the volumes of 90 pre-defined regions of interest (ROIs) within the anatomical automatic labeling (AAL) atlas. Note that our proposed method is not restricted by these extracted features – it is applicable to many other features, including those extracted from different ROIs of different atlases, fiber tracking results from diffusion tensor imaging, etc.

In addition, we argue that only a few brain ROIs, as well as associated features, are relevant for computer-assisted PD diagnosis. The redundancy within initially extracted features can be high, which make it difficult to estimate the common feature space by a one-shot CCA transformation. The inaccurate estimation of CCA common space is also inaccurate to the selection of a limited number of optimal features and the subsequent learning of regression model. To this end, we develop the ICCA method for feature selection, in which we iteratively optimize the estimation of the common space and gradually discard the irrelevant features (Liu et al., 2016). Specifically, in the first iteration of ICCA, we transform the features of WM/GM views into a tentative common space, and build the PD regression model accordingly. The regression coefficients of the model measure the contributions of individual features, and guide us to eliminate the most irrelevant features for PD classification. The subsequent iterations will update the common space based on the tentatively preserved features, which will be further selected once the common space is updated. In the final, the two feature vectors of WM/GM views are transformed to the common space, while only the limited number of the optimal features is preserved. Though the framework in this study is similar with our previous work in (Liu et al., 2016), we further optimize the feature selection strategy in this study. Specifically, we adopt different strategy to discard the redundant features iteratively in order to ensure that the discarded features are insignificant to subsequent diagnosis. We also provide more details and experiments to analyze the performances of our method and its impact to computer-aided PD diagnosis.

The ICCA-based feature selection reveals the optimal common feature space for multi-view learning of neuroimaging data. We then utilize robust linear discriminant analysis (RLDA), which is an extension of robust regression for classification since classification problems can be cast as a particular case of binary regression (Huang et al., 2016), to complete the PD classification task based on the selected features. The RLDA can solve the small-sample-size problem and also the bias problem caused by outliers at the same time. In the next Section, we will give more details about robust regression and RLDA. Note that, with the linear transformation of the features in CCA, our ICCA provides locally linear feature transformation capabilities that contribute to sophisticated feature selection across two different anatomical views. The remaining parts of the paper are organized as follows. In

Section 2, we provide related works on feature selection as well as robust regression, since the RLDA classifier used in this paper is an extension of robust regression for classification. In Section 3, we present details about the design of our classification framework, especially the ICCA feature selection method. The experimental results of the proposed method and the comparisons with state-of-the-art feature selection approaches are provided in Section 4. The top features selected by the proposed method for PD classification are also discussed. Finally, we conclude the paper in Section 5.

2. Related work

2.1. Feature selection methods

Based on whether using label information in the training data, feature selection can be divided into the supervised and unsupervised methods (Zhao and Liu, 2007). Feature selection techniques which rely on features only and without considering the class labels of the training subjects are *in general* belonging to the unsupervised feature selection category, including PCA, t-test, etc. These methods are simple and hence widely used in many applications. For instance, Song et al. (2010) exploited the eigenvectors of the PCA covariance matrix to evaluate the significance of each individual feature component in the original data. With the evaluated feature components, their method only takes a few eigenvectors into account. Their experimental results on face recognition show that their method could select an appropriate amount of features without jeopardizing the recognition accuracy. Zhou and Wang (2007) modified the t-test ranking measure and used it to discover the significant single Nucleotide polymorphisms (SNPs).

In addition to the above, there are some popular filter-based feature selection methods, which rank features with various criteria and then select them. The criteria include information gain (Azhagusundari and Thanamani, 2013), Fisher score (Gu et al., 2012), ReliefF (Robnik-Šikonja and Kononenko, 2003), Laplacian score (He et al., 2005) and Trace Ratio (Nie et al., 2008), all of which have been extensively studied in different fields. Azhagusundari and Thanamani (Azhagusundari and Thanamani, 2013) proposed an algorithm that combines discernibility matrix and information gain to select features. They demonstrated that better results, in terms of the number of selected features and classification accuracy, can be achieved than applying the two criteria separately. Gu et al. (2012) presented a generalized Fisher score by reformulating the feature selection problem as a quadratic constrained linear programming. They overcame the shortcoming of the conventional Fish score feature selection method, in which each feature is independently considered. The selected suboptimal subset of features is shown to outperform the conventional Fisher score and many state-of-the-art feature selection methods, including Laplacian score, Hilbert Schmidt Independence criterion, and Trace Ratio criterion. In (Robnik-Šikonja and Kononenko, 2003), authors thoroughly investigated Relief and ReliefF, and demonstrated their robustness as well as tolerance to noises.

The class labels of the training subjects can contribute to supervised feature selection methods. In this way, the selection of features is perceived as part of learning upon the training data, which can usually be modeled by optimizing an objective function. Some popular examples are sparse learning (Ye and Liu, 2012) and Least Angle Regression

(LARS) (Efron et al., 2004). In (Tsanas et al., 2010), the Least Absolute Shrinkage and Selection Operator (LASSO) is used to perform feature selection. LASSO minimizes the class label regressed from the features from its ground truth, while the regression coefficients measuring the contributions of individual features are penalized by the l1-norm. The l1-norm penalization encourages many coefficients to vanish, while only the most useful features are selected to contribute to the regression of the class labels. Wang et al. (2011) proposed a novel Sparse Multi-tAsk Regression and feaTure selection (SMART) model, in which they included both l1-norm and l2,1-norm regularizations for selecting features. In (Nie et al., 2010), Nie et al. proposed to measure both the loss function and the regularization of coefficients with the l2,1-norm jointly. The l2,1-norm based loss function is robust to outlier subjects, while the l2,1-norm regularization selects features across all subjects with joint sparsity. In (Armanfard et al., 2016), Armanfard et al. proposed a novel localized feature selection (LFS) method where the optimal feature subset is associated with each region of the sample space. In (Li et al., 2017), Li et al. proposed a granular feature selection method for multi-label learning with a maximal correlation minimal redundancy criterion based on mutual information to select a more relevant and compact feature subset as well as explore the label dependency.

The performance of feature selection can be limited especially when dealing with complex data, in which features are highly correlated and thus redundant. For instance, Peng et al. (2005) presented a two-stage feature selection algorithm based on maximal statistical dependency. By combining minimal-redundancy-maximal-relevance and other sophisticated feature selectors, they demonstrated promising results over several classifiers and datasets. In (Senawi et al., 2017), Senawi et al. proposed a maximum relevance-minimum multicollinearity (MRmMC) method in which relevant features are measured by correlation characteristics based on conditional variance while redundancy elimination is achieved according to multiple correlation assessment using orthogonal projection scheme. In (Naghibi et al., 2015), Naghibi et al. proposed a parallel search strategy on semidefinite programming, which can search through the subset space in polynomial time, with mutual information between features and class labels considered as measure function. In (Zhu et al., 2014), Zhu et al. proposed a novel canonical feature selection method, which efficiently integrates the correlation information between structural and functional neuroimaging data into a sparse multi-task learning framework. They assumed that the structural and functional feature descriptors of a single subject are highly redundant. By projecting the features of these two different modalities (and thus two views) into a common space with CCA, they were able to maximize the correlation between features and also conduct task-related feature selection.

2.2. Robust regression

In this study, we use RLDA for classifying subjects with selected features. RLDA can be considered as a special case of robust regression (Huang et al., 2016). As also confirmed in the literature, regression methods can be extended for the challenging classification tasks (Naghibi et al., 2015; Wang et al., 2010; Huang et al., 2011). Nevertheless, the lack of robustness against noise is one of the major drawbacks of most existing regression approaches, especially when the outliers affect the normal distribution of subjects within

high-dimensional feature space (Huang et al., 2016). Therefore, robust regression methods (Rousseeuw and Leroy, 2005; Van Huffel and Vandewalle, 1991) have developed during the past decades. In (Huber, 2011), Huber introduced M-estimation for regression, which shows robustness for outlier subjects. Rousseeuw and Leroy (2005) developed Least Trimmed Squares, which can find a data subset that minimizes the squared residual sum. However, these methods can only remove outliers among subjects, and therefore they cannot deal with outliers within subjects, or in other words noises in the feature values. Hence, Error-In-Variable (EIV) approaches (Van Huffel and Vandewalle, 1991) are proposed to deal with noises in the features, even though the existing EIV approaches rely on strong assumptions on classification errors.

Robust regression methods are further extended for robust classification applications. A robust extension of Linear Discriminant Analysis (LDA) is proposed in (Croux and Dehon, 2001), in which authors substituted their robust counterparts for the empirical estimation of the class mean vectors and covariance matrices. In (Kim et al., 2005; Zhang and Yeung, 2010), authors proposed a worst-case Fisher Discriminant Analysis/Linear Discriminant Analysis (FDA/LDA) to increase the separating ability between classes in unbalanced sampling by minimizing the upper bounds of LDA cost function. It is worth noting that these classifiers are only robust to outliers among the subjects, yet not robust to the intrinsic noises in extracted features. To deal with this problem, Fidler et al. (2006) modified LDA and made it robust to outliers within subjects through application of PCA on the training data. A robustly estimated basis is computed to replace the minor PCA components, after which they combined these two bases into a new one, and then the data was projected onto the combined basis. Finally, the LDA is calculated in the training phase. In the testing phase, the coefficients of the test data on the recombined basis are estimated, and the class label(s) for the test data are determined by mapping the learned LDA on the estimated coefficients. This method can suppress the outliers outside the PCA subspace, but it cannot deal with the problem of learning LDA with outliers in the PCA subspace of the training data. Zhu and Martinez (Jia and Martinez, 2009) extended the SVM algorithm, denoted as Partial Support Vector Machine (PSVM), to make it applicable for the cases with missing features in the subjects. They have indicated that their method is also robust to some levels of noises. Of note, the robust LDA (Huang et al., 2016) used in this study is an extension of the robust regression study, which is inspired by existing work on robust PCA (De La Torre and Black, 2003). It enjoys the following advantages compared to the aforementioned methods: 1) it is a convex approach; 2) except for sparsity, there is no assumption imposed on the noise in the data; 3) it automatically cleans the noise in the features when learning a classifier. More implementation details about robust LDA classifier will be provided in Section 3.

3. Method

The framework of our proposed method for PD diagnosis in this paper can be illustrated as follows: suppose that we get two feature vectors (feature vector 1 and feature vector 2) already. These features are then fed into the ICCA feature selection framework, where the two feature vectors are transformed onto their common feature space using CCA. The new canonical representation of each feature vector is computed in this common space. Next, we build a linear regression model to fit the PD/NC labels of the training subjects based on the

canonical representations of their features. The regression model then yields the coefficients that describe the contributions of the canonical representations, from which the importance of the original features can be computed. We thus conduct conservative selection upon the features, by discarding just 5 gray matter and 5 white matter features in the early 50 iterations and then only eliminating the least important gray matter and white matter features in the later iterations. The rest of the features are fed into the next iteration of feature selection, as the features are transformed to the updated common space and then selected accordingly. This iterative procedure (or ICCA) ends when only a small set of optimal features are remaining. Finally, we build a robust classifier based on the selected features, and further apply it to diagnose the new test subject with unknown PD/NC label.

3.1. Canonical correlation analysis (CCA)

CCA is a method to exploit the correlated relationship between two multi-dimensional feature sets. Specifically, CCA intends to find a pair of basis vectors, such that the correlation between the two sets of feature vectors is maximized after they are transformed following the basis vectors, respectively. The transformed features are highly correlated and thus perceived to be within the common feature space (Zhu et al., 2014).

Consider two multivariate feature vectors of the form $(\mathbf{x}^1, \mathbf{x}^2)$. Suppose we are given n subjects $\mathbf{X} = \left\{\left(\mathbf{x}_1^1, \mathbf{x}_1^2\right), \ldots, \left(\mathbf{x}_n^1, \mathbf{x}_n^2\right)\right\}$. Let $\mathbf{X}^1$ denote $\left[\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1\right]$ and $\mathbf{X}^2$ denote $\left[\mathbf{x}_1^2, \ldots, \mathbf{x}_n^2\right]$. Define a transform matrix $\mathbf{B}^1$ to project $\mathbf{X}^1$, and $\mathbf{B}^2$ to project $\mathbf{X}^2$. The projected $\mathbf{X}^1$ and $\mathbf{X}^2$ are computed as $\mathbf{Z}^1 = \langle \mathbf{B}^1, \mathbf{X}^1 \rangle = \mathbf{B}^{1'}\mathbf{X}^1$, and $\mathbf{Z}^2 = \langle \mathbf{B}^2, \mathbf{X}^2 \rangle = \mathbf{B}^{2'}\mathbf{X}^2$, where $\langle \bullet, \bullet \rangle$ denotes the Euclidean inner product. The new representations $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are called the canonical representations for $\mathbf{X}^1$ and $\mathbf{X}^2$, respectively. In CCA particularly, we determine a pair of $\mathbf{B}^1$ and $\mathbf{B}^2$ to maximize the correlation between the canonical representations $\mathbf{Z}^1$ and $\mathbf{Z}^2$, which can be formularized as the following:

$$\rho = \max_{\mathbf{Z}^1, \mathbf{Z}^2} corr\left(\mathbf{Z}^1, \mathbf{Z}^2\right) = \max_{\mathbf{Z}^1, \mathbf{Z}^2} \frac{\left\langle \mathbf{Z}^1, \mathbf{Z}^2 \right\rangle}{\left\|\mathbf{Z}^1\right\| \left\|\mathbf{Z}^2\right\|} = \max_{\mathbf{B}^1, \mathbf{B}^2} \frac{\left\langle \mathbf{B}^{1'}\mathbf{X}^1, \mathbf{B}^{2'}\mathbf{X}^2 \right\rangle}{\left\|\mathbf{B}^{1'}\mathbf{X}^1\right\| \left\|\mathbf{B}^{2'}\mathbf{X}^2\right\|}. \quad (1)$$

If we use $\hat{E}[f(\mathbf{X}^1, \mathbf{X}^2)]$ to denote the empirical expectation of $f(\mathbf{X}^1, \mathbf{X}^2)$, where $\hat{E}\left[f\left(\mathbf{X}^1, \mathbf{X}^2\right)\right] = \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{x}_i^1, \mathbf{x}_i^2\right)$, then the correlation expression can be rewritten as:

$$\rho = \max_{\mathbf{B}^1, \mathbf{B}^2} \frac{\hat{E}\left[\left\langle \mathbf{B}^1, \mathbf{X}^1 \right\rangle \left\langle \mathbf{B}^2, \mathbf{X}^2 \right\rangle\right]}{\sqrt{\hat{E}\left[\left\langle \mathbf{B}^1, \mathbf{X}^1 \right\rangle^2\right] \hat{E}\left[\left\langle \mathbf{B}^2, \mathbf{X}^2 \right\rangle^2\right]}} = \max_{\mathbf{B}^1, \mathbf{B}^2} \frac{\hat{E}\left[\mathbf{B}^{1'}\mathbf{X}^1\mathbf{X}^{2'}\mathbf{B}^2\right]}{\sqrt{\hat{E}\left[\mathbf{B}^{1'}\mathbf{X}^1\mathbf{X}^{1'}\mathbf{B}^1\right] \hat{E}\left[\mathbf{B}^{2'}\mathbf{X}^2\mathbf{X}^{2'}\mathbf{B}^2\right]}} \quad (2)$$

$$= \max_{\mathbf{B}^1, \mathbf{B}^2} \frac{\mathbf{B}^{1'}\hat{E}\left[\mathbf{X}^1\mathbf{X}^{2'}\right]\mathbf{B}^2}{\sqrt{\mathbf{B}^{1'}\hat{E}\left[\mathbf{X}^1\mathbf{X}^{1'}\right]\mathbf{B}^1\mathbf{B}^{2'}\hat{E}\left[\mathbf{X}^2\mathbf{X}^{2'}\right]\mathbf{B}^2}}.$$

Now note that the covariance matrix of $(\mathbf{X}^1, \mathbf{X}^2)$ is

$$C\left(\mathrm{X}^1, \mathrm{X}^2\right) = \hat{E}\left[\begin{pmatrix}\mathrm{X}^1\\ \mathrm{X}^2\end{pmatrix}\begin{pmatrix}\mathrm{X}^1\\ \mathrm{X}^2\end{pmatrix}'\right] = \begin{bmatrix}C_{\mathrm{X}^1\mathrm{X}^1} & C_{\mathrm{X}^1\mathrm{X}^2}\\ C_{\mathrm{X}^2\mathrm{X}^1} & C_{\mathrm{X}^2\mathrm{X}^2}\end{bmatrix} = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}. \quad (3)$$

Hence, we can rewrite the function ρ as:

$$\rho = \max_{\mathbf{B}^1, \mathbf{B}^2} \frac{\mathbf{B}^{1'}\Sigma_{12}\mathbf{B}^2}{\sqrt{\mathbf{B}^{1'}\Sigma_{11}\mathbf{B}^1\mathbf{B}^{2'}\Sigma_{22}\mathbf{B}^2}}. \quad (4)$$

The maximal canonical correlation coefficient vector is, equivalently, the maximum of ρ with respect to $\mathbf{B}^1$ and $\mathbf{B}^2$.

3.2. CCA-based feature selection

Given n subjects, their d-dimensional feature vectors are considered as individual columns in $\mathbf{X}^1 \in \mathrm{R}^{d\times \mathrm{n}}$ and $\mathbf{X}^2 \in \mathrm{R}^{d\times \mathrm{n}}$, which correspond to the views of WM and GM, respectively, in our application. The class labels for the subjects are stored in $\mathbf{y} \in \mathrm{R}^{n\times 1}$, where each entry is either "1" (PD) or "0" (NC) to indicate the class label that a subject is associated with. Let $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2] \in \mathrm{R}^{d\times 2\mathrm{n}}$, and $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ be its covariance matrix. CCA can find the basis matrices $\mathbf{B}^1 \in \mathrm{R}^{d\times d}$ and $\mathbf{B}^2 \in \mathrm{R}^{d\times d}$ to maximize the correlation between the transformed $\mathbf{X}^1$ and $\mathbf{X}^2$. The two basis vectors $\mathbf{B}^1$ and $\mathbf{B}^2$ can be estimated by solving the following optimization problem:

$$\left(\mathbf{B}^1, \mathbf{B}^2\right) = \operatorname*{argmax}_{\mathbf{B}^1, \mathbf{B}^2} \frac{\mathbf{B}^{1'}\Sigma_{12}\mathbf{B}^2}{\sqrt{\mathbf{B}^{1'}\Sigma_{11}\mathbf{B}^1}\sqrt{\mathbf{B}^{2'}\Sigma_{22}\mathbf{B}^2}}. \quad (5)$$

Note that the generalized Eigen-decomposition in (Zhu et al., 2012) gives the optimal solution of $(\mathbf{B}^1, \mathbf{B}^2)$. Therefore, the canonical representations of all features in the common space are obtained by $\mathbf{Z}^m = \mathbf{B}^{m'}\mathbf{X}^m$, $m = \{1,2\}$.

Next, we build a sparse linear regression model based on the canonical representations, aiming to fit the class labels with the latest canonical representations. Specifically, the non-zero weights are assigned to a limited number of the canonical representations following:

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{W}} \frac{1}{2} \| \mathbf{y} - \mathbf{Z}^{\mathrm{T}}\mathbf{w} \|_{\mathrm{F}}^2 + \beta\|\mathbf{w}\|_1 + \gamma\| \mathbf{w} \|_{\mathrm{CCA}}^2. \quad (6)$$

Here, $\mathbf{Z} = [\mathbf{Z}^1, \mathbf{Z}^2] \in \mathrm{R}^{2d\times \mathrm{n}}$ is the canonical representation matrix, $\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2] \in \mathrm{R}^{2\mathrm{d}\times 1}$ is the regression coefficient matrix, and β and γ are the trade-off scalar parameters. The l_1 -

norm penalty $\|\mathbf{w}\|_1$ tends to yield sparse coefficients for the canonical representations. And $\|\mathbf{w}\|_{\mathrm{CCA}}$ denotes the following canonical regularizer (Nie et al., 2010):

$$\|\mathbf{w}\|_{\mathrm{CCA}}^2 = \sum_{i=1}^{d} \frac{1-\lambda_i}{\lambda_i}\left(w_i^2 + w_{i+d}^2\right). \quad (7)$$

Note that $\{\lambda_1, \ldots, \lambda_d\}$ denotes the set of the canonical correlation coefficients, and w_i and w_{i+d} are the weights corresponding to the same feature index (and thus ROI) of the two views (i.e., GM and WM). In the feature selection process, more correlated canonical representations across the two views, which means with large canonical correlation coefficients, tend to be selected and vice versa.

It is worth noting that the canonical regularizer finds canonical representations that are highly correlated across the GM/WM feature views simultaneously. Besides, it can also be considered that the higher λ_i comes with the increases of w_i and w_{i+d} after optimization. In the conventional methods, the feature matrix in the original space is generally used to build the regression model, find the relationship between feature matrix and output, and use this kind of relationship to select features. These methods are able to select significant features and remove redundant features. However, when there are complex relationships among features, it is really hard to discard redundant features. The CCA-based feature selection uses canonical representations for regressors, which are highly correlated but mutually independent within each representation. Therefore, it is more convenient to identify redundant information in the canonical space, and thus the CCA-based feature selection would be more predictive than the traditional sparse feature selection methods (Zhu et al., 2014).

3.3. ICCA-based feature selection

As stated in Section 3.2, the feature selection method by CCA is more powerful in predicting the response variables (i.e., the PD/NC labels) than the conventional sparse feature selection methods, since CCA considers the correlation between the two sets of features from two different views, as well as the correlation between features and response variables. Even though there are many advantages of the CCA based feature selection, its main limit lies in the inaccurate one-shot estimation of the common feature space. As all features are linearly transformed to the common space and selected accordingly, it might be not precise enough to identify and preserve the optimal features. Therefore, we propose a novel ICCA feature selection method, in which we iteratively update the estimation of the common feature space and discard the most irrelevant features for iterative feature selection. In this way, we are able to acquire the best features by approximating locally linear operation upon the input feature vectors.

In the Eq. (6), we can get the regression coefficient matrix $\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2]$ in the tentatively estimated common/canonical space, which present the weights for canonical representations $\mathbf{Z} = [\mathbf{Z}^1, \mathbf{Z}^2]$ in canonical space. From the equation, $\mathbf{Z}^m = \mathbf{B}^{m'}\mathbf{X}^m$, $m = \{1,2\}$, we can see that canonical representations $\mathbf{Z}$ ($\mathbf{Z} = [\boldsymbol{Z}^1, \boldsymbol{Z}^2]$) are linear combinations of $\mathbf{X}$ ($\mathbf{X} = [\boldsymbol{X}^1, \boldsymbol{X}^2]$)

(WM/GM features in the original feature space) warped by $\mathbf{B}$ ($\mathbf{B} = [\mathbf{B}^1, \mathbf{B}^2]$). Therefore, the weights in $\mathbf{w}$ ($\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2]$), which learned from Eq. (6), are also linearly correlated with the importance of $\mathbf{X}$ ($\mathbf{X} = [\boldsymbol{X}^1, \boldsymbol{X}^2]$) (WM/GM features). Then, the importance or weights of the original WM/GM features $\mathbf{X}$ ($\mathbf{X} = [\boldsymbol{X}^1, \boldsymbol{X}^2]$) can be computed by $\tilde{\mathbf{w}}^m = \left(\mathbf{B}^m\right)^{-1}\mathbf{w}^m$, $m = \{1,2\}$, where $\tilde{\mathbf{w}}^m$ records the weights of the m-th view of the original features for discriminant PD from NC. Given the estimated $\tilde{\mathbf{w}}^1$ and $\tilde{\mathbf{w}}^2$, the least important WM or GM features for discrimination in the original feature space can be eliminated. Note that we optimize feature selection here in a more effective way compared to our previous work in (Liu et al., 2016). In (Liu et al., 2016), we discarded one most insignificant feature from WM and GM feature vectors, respectively, which is time consuming and ineffective. In order to solve this problem, we improve feature selection strategy in this study. In the first a ($a = 50$) iterations, we discard b ($b = 5$) WM and GM features, respectively. The values of a and b are determined by experiments. Then, in the following iterations, we discard a pair of WM and GM features. With the updated WM/GM features in the original space in each iteration, $\mathbf{X}$ ($\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2]$), CCA is applied to update the common feature space and transforms the updated WM/GM features to refined canonical representations in the common space accordingly. This transforming-eliminating scheme is iteratively executed until the number of the iterations exceeds a predefined threshold, or our desired number of features is obtained.

In conclusion, our proposed ICCA-based feature selection method consists of the following five steps:

1. Compute basis vectors $\mathbf{B}^1$, $\mathbf{B}^2$ using CCA via Eq. (5);
2. Transform the original feature matrices $\mathbf{X}^1$, $\mathbf{X}^2$ into the canonical space ($\mathbf{Z}^1$, $\mathbf{Z}^2$) by $\mathbf{Z}^m = \mathbf{B}^{m'}\mathbf{X}^m$, $m = \{1,2\}$;
3. Compute the coefficient vectors $\mathbf{w}^1$, $\mathbf{w}^2$ in the canonical space, which are stored in $\mathbf{w}$ ($\mathbf{w} = [\mathbf{w}^1; \mathbf{w}^2]$), using Eq. (6);
4. Compute the feature importance vectors in the original feature space (i.e., $\tilde{\mathbf{w}}^1$ and $\tilde{\mathbf{w}}^2$), by $\tilde{\mathbf{w}}^m = \left(\mathbf{B}^m\right)^{-1}\mathbf{w}^m$;
5. Discard features from $\tilde{\mathbf{w}}^1$ and $\tilde{\mathbf{w}}^2$ based on the optimized strategy mentioned above, and update the feature matrices $\mathbf{X}^1$, $\mathbf{X}^2$;
6. Repeat Steps 1–5 until the stopping criterion is attained.

3.4. Robust linear discriminant analysis classification

As previously stated in Section 2.2, in this paper we incorporate the RLDA technique for the PD/NC classification work, using the optimal feature subset selected by the proposed ICCA method. RLDA can be seen as a special case of robust regression (RR) method (Huang et al., 2016), since binary classification can be treated as a special case of regression.

Given $\tilde{\mathbf{X}} \in \mathbf{R}^{n \times \tilde{d}}$ as the input matrix, which consists of $\tilde{d}$-dimensional samples with noises. LDA learns a linear transformation, which maximizes the inter-class variance while

minimizing the intra-class variance. However, when learning upon the high-dimensional data, LDA cannot overcome the small sample size problem. In order to solve this problem, LDA can be formularized as a least-squares (LS) problem (Huang et al., 2016), in which $\widetilde{\mathbf{X}}$ is directly mapped to the class labels represented by an indicator vector. Thus, LS-LDA reduces to:

$$\underset{\mathbf{t}}{\operatorname{argmin}} \frac{\eta}{2} \left\| \left(\mathbf{y}^{\mathrm{T}}\mathbf{y}\right)^{-\frac{1}{2}} (\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{t}) \right\|_2^2. \quad (8)$$

The normalization factor $(\mathbf{y}^T\mathbf{y})^{-1/2}$ compensates for different sample size per class and $\mathbf{t} \in \mathscr{R}^{d \times 1}$ represents a regression vector. However, when $\widetilde{\mathbf{X}}$ is corrupted by the noise, the LS-LDA suffers from a biased estimation of **t**. The robust LDA (RLDA), by explicitly decomposing the data matrix into a noise-free data matrix and the error matrix, partially resolves the problems above.

In RLDA, by modeled the data as $\widetilde{\mathbf{X}} = \mathbf{D} + \mathbf{E}$, where **D** is the underlying noise-free or clean component and **E** contains the noise or error in the data, the noise in data can be partially considered and removed. The formula of RLDA is shown in Eq. (9), where $\mathbf{H} = (\mathbf{I}_n - 11^{\mathrm{T}}/n)$ is a centering matrix and $\mathbf{t} \in \mathscr{R}^{d \times 1}$ is the mapping matrix that learned from the centered clean or noise-free data **HD** only. From the Eq. (9), we can see that, in RLDA, **t** is mapped from $\widetilde{\mathbf{X}}$ to fit the class labels in $\mathbf{y} \in \mathbf{R}^{n \times 1}$, and $\widetilde{\mathbf{X}}$ is decomposed into **D** and **E**, thus the mapping **t** is computed using the noise-free data **D**, which yields the desired effect that we can partially decrease the influence of noise on data $\widetilde{\mathbf{X}}$.

$$\underset{\mathbf{t},\mathbf{D},\mathbf{E}}{\operatorname{argmin}} \frac{\eta}{2} \left\| \left(\mathbf{y}^{\mathrm{T}}\mathbf{y}\right)^{-\frac{1}{2}} (\mathbf{y} - \mathbf{HDt}) \right\|_2^2 + \|\mathbf{D}\|_* + \lambda \|\mathbf{E}\|_1 \text{ s. t. } \widetilde{\mathbf{X}} = \mathbf{D} + \mathbf{E}, \quad (9)$$

Since $\widetilde{\mathbf{X}}$ is decomposed into clean (noise-free) data **D** and noise, and mapping matrix **t** is only learned from the clean data. Therefore, unlike traditional method (e.g., LDA, LS-LDA), it avoids projecting the noise **E** in the estimation, and thus conduce to an relatively unbiased noise-free estimation. After **t** is learned in the training phase, a testing data, $\widetilde{\mathbf{X}}_{\text{test}}$, is projected by **t** and a k-Nearest Neighbor (k-NN) algorithm is used to determine the class label of the test data. The above RLDA formulation can be easily solved by augmented Lagrangian multipliers method. For more details, please refer to (Huang et al., 2016).

4. Experiments

In this section, we apply the proposed ICCA method to the PD/NC diagnosis to demonstrate its validity. Here we employ the Parkinson's Progression Makers Initiative (PPMI) dataset (Zhu et al., 2012) for the evaluations. The T1-weighted MR images were acquired using 3 T SIEMENS MAGNETOM TrioTim Syngo. There are 112 subjects (56 PD, and 56 NC)

randomly selected from the PPMI dataset for this study. The parameter settings are given as follows: acquisition matrix = 256 × 256 mm^2, 176 slices, voxel size = 1 × 1 × 1 mm^3, echo time (TE) = 2.98 ms, repetition time (TR) = 2300 ms, inverse time = 900 ms, and flip angle = 9°.

All T1-weighted MR images are pre-processed by skull stripping using a learning-based algorithm for brain extraction and labeling, which performs multiple complementary brain extractions on a given testing imaging by using a meta-algorithm (Shi et al., 2012), cerebellum removal, and tissue segmentation into WM and GM using a hidden Markov random field model proposed in (Zhang et al., 2001). Then, the anatomical automatic labeling (AAL) atlas with 90 pre-defined ROIs is registered to the native space of each subject, using multi-modal affine registration (FLIRT) in FSL (Smith et al., 2004), followed by HAMMER registration algorithm (Shen and Davatzikos, 2002). For each ROI, we estimate the WM/GM tissue volumes as the WM/GM feature, respectively. In this way, we extract 90 WM and 90 GM features for each subject. The features are naturally grouped into two vectors, to be handled by the ICCA based feature selection and then the RLDA based classification. Fig. 1 shows the pipeline of our proposed ICCA based feature selection algorithm for PD diagnosis.

4.1. Evaluation protocol

In order to evaluate our proposed method, we compare our method with state-of-the-art feature selection methods, including PCA (López et al., 2009), FSASL (unsupervised feature selection with adaptive structure learning) (Du and Shen, 2015), CCA (Zhu et al., 2014), sparse feature selection (Ye et al., 2012), Elastic Net (Zou and Hastie, 2005), robust feature selection method (Nie et al., 2010) and minimal-redundancy-maximal-relevance (mRMR) (Peng et al., 2005). To this end, we use the same nested cross-validation to evaluate the performances of individual methods on classifying PD patients from NC. Specifically, in the outer layer of the nested cross-validation, an 8-fold scheme is adopted to partition the dataset into 7 training folds and 1 testing fold. In the inner layer of the nested cross-validation, we further conduct 10-fold analysis toward the 7 folds of training data. The inner 10-fold cross-validation is used to tune the parameters automatically. With a fixed subset of features that are selected in the inner 10-fold cross-validation, the diagnostic ability is tested in the outer 8-fold cross-validation. Note that with 12 CPU cores (2.76 GHz) and 48 GB memory size, it takes about 2 s for each inner loop, 1 h in the training stage with parallel computing, and about 25 s to test a new subject. The implementations are all done in MATLAB.

4.2. Comparisons with the state-of-the-art feature selection methods

In this study, we compare the PD/NC classification performance of all the configurations along with the alternative methods available. Specifically, we compare our ICCA-based feature selection with several popular feature selection or feature reduction methods, including PCA (López et al., 2009), FSASL (unsupervised feature selection with adaptive structure learning) (Du and Shen, 2015), CCA (Zhu et al., 2014), sparse feature selection (Ye et al., 2012), Elastic Net (Zou and Hastie, 2005), robust feature selection method (Nie et al., 2010) and mRMR (Peng et al., 2005). No feature selection (NoFS) is also compared here, where all features are input to the RLDA classifier without any feature selection. For Elastic

Net, a function namely “lasso” in MATLAB is used for implementation. For the sparse feature selection method, we use the $l_{2,1}$ -norm regularization, instead of the l_1 -norm regularization, as in the traditional lasso regression model. In PCA-based feature selection, only the top 1% principal components are preserved, which result in much fewer yet more powerful features. The details about the one-shot CCA feature selection are described in (Zhu et al., 2014). The details of the FSASL method can be found in (Du and Shen, 2015), where the code is also available online.

Alternatively, we also measure the performance when selecting upon the canonical representations rather than the original WM/GM features. State succinctly, we directly eliminate the least important canonical representations in every iteration of ICCA, instead of using the inverse-transform and elimination steps in the original WM/GM features. The remaining canonical representations are used for re-estimating the new common space by CCA and further selected. We consider this configuration as “cascaded-CCA”, that several CCA based feature selection steps are cascaded rather than iteratively executed.

The classification results are presented in Table 1., which shows that the proposed ICCA based feature selection method achieves better performance than the competing methods. In particular, our method improves by 11.5% on ACC and 16% on AUC, respectively, compared to the baseline (NoFS). Moreover, compared to feature selection through Elastic Net, PCA, SFS, CCA, FSASL, RFS and mRMR, our method achieves 6.2%, 5.3%, 5.4%, 4.4%, 5.2%, 4.4% and 3.8% accuracy improvements, respectively. Also, our methods achieve 6.5%, 11.3%, 6.6%, 6.7%, 4.7%, 7.1%, 5.5% AUC improvements, compared to Elastic Net, PCA, SFS, CCA, FSASL, RFS, mRMR respectively. It can also be observed from the last two lines in Table 1 that the proposed ICCA based feature selection outperforms the cascaded-CCA scheme. We attribute this to the observations that the CCA mapping to the common space is unsupervised and, after CCA mapping, the two feature vectors become highly correlated in the canonical space. This leads to the redundancy occurred in the canonical representations, which further mislead the feature selection and thus influence the subsequent classification. It is also concluded that such limitation can be solved by inversely transforming the representations back to the original feature space before conducting feature selection in the ICCA strategy.

4.3. The Most discriminant regions

Recently, there are a lot of works to study the WM/GM changes on PD patients. They also identify several significant brain regions that are affected by PD mostly. For example, the WM lesions are associated with some motor and cognitive deficits in PD (Bohnen and Albin, 2011). Significantly increased WM density in occipital lobes, posterior cingulate gyrus, and paracentral lobule are found in PD with olfactory dysfunction compared with the normal controls (NC) (Ding et al., 2011). GM volumes are found to be significantly decreased in the patients with PD compared with NC (Xia et al., 2013). And reduced gray matter volumes in PD patients are observed in the temporal lobe, occipital lobe, right frontal lobe, left parietal lobe, and some subcortical regions (Burton et al., 2004).

In this section, we intend to evaluate these optimal feature subsets by ICCA, which are expected to be the biomarkers of PD that correspond to their specific WM/GM regions. Note

that we further select the most discriminative regional features, which are available in at least 60% of feature subset in all the cross-validation folds. Fig. 2 presents these selected regional features, which have ‘precuneus’, ‘thalamus’, ‘hippocampus’, ‘superior temporal pole’, ‘postcentral gyrus’, ‘middle frontal gyrus’, and ‘medial frontal gyrus’.

It is worth noting that the most discriminative regional features shown in Fig. 2 fit with the PD pathology, which is also proven by the previous clinical researches. As reported in (Hanakawa et al., 1999; Halliday, 2009), the PD patients reveal under-activation in precuneus and temporal cortex, and the pathology in thalamus also contribute to the abnormal neural activity characteristic of PD. In additional, temporal pole and hippocampal atrophy are reported as an early sign of PD in (Burton et al., 2004; Potgieser et al., 2014; Camicioli et al., 2003). In (González-Redondo et al., 2014) bilateral areas of atrophy in middle frontal gyrus and bilateral GM loss in medial-superior frontal are reported in PD patients. It is found in (Beyer and Aarsland, 2008) that the areas of reduction in the absolute amount of GM are bilateral in the medial frontal gyrus, right precuneus, inferior parietal lobule, and so on. In (Song et al., 2011), gray matter densities are found significantly decreased in bilateral temporal, right postcentral, and some other brain regions in PD of dementia patients.

We also conduct t-tests for the statistical analysis on GM/WM volumes computed from the 7 significant ROIs between PD and NC, with a p-value of 0.05 as the threshold for the statistical significance. One interesting finding is that only three regions show significant differences between PD and NC, namely ‘precuneus’ ($p = 0.0065$), ‘hippocampus’ ($p = 0.0164$), and ‘middle frontal’ ($p = 0.0429$). This implies that the GM/WM volumes extracted from brain regions may not be simply used for prognosis. Instead, when combining them together and using the highly efficient machine learning tools for the analysis, we can achieve high prognosis capability.

Except for our proposed ICCA based feature selection method, we also explore the most discriminant features selected by the traditional feature selection method (mRMR) and state-of-the-art (CCA-based feature selection), with their performances shown in Section 4.2. The most significant features selected by mRMR include “middle occipital gyrus”, “putamen”, “medial orbitofrontal cortex”, “caudate nucleus”, “superior frontal gyrus (orbital part)”, and “anterior cingulate gyrus”. Four out of the six significant regional features selected by mRMR are in line with the existing studies such as “caudate nucleus” (Corrigan et al., 1998), “superior frontal gyrus (orbital part)” (Zhang et al., 2017a), “putamen” (Kish et al., 2007), and “anterior cingulate gyrus” (Allman et al., 2001). The most discriminant regional features selected by CCA includes “precentral gyrus”, “anterior cingulate and paracingulate gyri”, “hippocampus”, “precuneus”, “caudate nucleus”, and “postcentral gyrus”. Similarly, most of the extracted regional features are in line with the existing studies such as “hippocampus” (Burton et al., 2004; Potgieser et al., 2014; Camicioli et al., 2003), “precuneus” (Beyer and Aarsland, 2008), “caudate nucleus” (Corrigan et al., 1998), and “postcentral gyrus” (Song et al., 2011).

We have also conducted *t*-test to explore the statistical significance of the selected features by mRMR and CCA, and obtain a *p*-value of 0.05. For the most significant regional features

selected by mRMR, there is no significant difference found between GM/WM volumes computed from the 6 significant ROIs from PD and NC, respectively, which is consistent with the finding metioned above. With regard to the most discriminant features selected by the CCA, only two regions show significant differences between PD and NC, which are "precuneus" ($p = 0.0065$) and "hippocampus" ($p = 0.0164$). All these experiments shown above prove an interesting phenomenon – although it seems like that these features don't have diagnosis abilities when used separately, when used together with modern machine learning methods, their joint diagnosis ability can be carved out.

5. Conclusion

In this paper, we present the novel ICCA based feature selection method for the PD diagnosis. Extended from the CCA based feature selection approach, our proposed ICCA based feature selection method is not only capable of locally linear mapping ability, but also able to explore the relationship among features and underlying structure that deeper into the feature space. It is also worth noting that the two view of feature vectors in the canonical space (i.e., canonical representations) would become closer and closer, when iterations of learning in ICCA feature selection framework is increased. This property also results in the decreased number of the selected features. We also conduct extensive PD/NC classification experiments for performance comparison between the proposed method using our ICCA strategy and the state-of-art approaches. The results show that our method achieves the highest accuracy output, which demonstrate its validity in improving the PD diagnosis process using the T1-weighted MR images.

As previously stated in Section 3, in each iteration of ICCA method, we discard the least important features from GM and WM, respectively. Note that here the number of discarded features are set up conservatively in order to avoid missing the potential features that maybe important for PD/NC classification. A smarter dropping out way could be used in the feature selection strategy to optimize the whole feature selection framework in our future work. Besides, another limitation of our proposed ICCA based feature selection method is that it can only explore relationships between two views of features one time. Thus, we will also find the solution to extend the proposed method so that it can handle more than two views of features simultaneously, which can make it more feasible in the feature selection domain.

Except for the further optimization on the efficiency and performance of proposed ICCA based feature selection method, another future work is improving the classification method for PD diagnosis, which is currently implemented using RLDA classifier in this paper. Although RLDA is able to achieve favorable classification performance which has been validated in Section 4.1, it is not able to model the nonlinear relationship between features and labels since RLDA is a linear classifier. Therefore, using a nonlinear classifier in the classification would make the proposed method be able to utilize feature information in the different way, which may conduce to at least equal or better than linear classifier. Finally, there is only structure information in T1 MR images used in this study. Thus, we will also turn to other modalities for the PD diagnosis apart from the T1-weighted MR images, such as diffusion tensor imaging (DTI) and function MRI (fMRI). The integrated feature information from multi-modality images can provide more comprehensive details, which can

be utilized by the proposed framework for better classification performance. We will also incorporate alternative machine learning techniques in the fields of image segmentation (Zhang et al., 2016, 2017b), image super-resolution (Zhang et al., 2017c) and classification (Zhang et al., 2017d; Chen et al., 2017), and seek further performance improvements in PD diagnosis.

Acknowledgements

This work is supported by National Key Research and Development Program of China (2017YFC0107600), National Natural Science Foundation of China (NSFC) Grants (Nos. 61473190, 61401271, 81471733), Science and Technology Commission of Shanghai Municipality (16511101100, 16410722400). This work was also supported in part by NIH grants (EB006733, EB008374, AG041721, AG049371, AG042599, and EB022880).

References

Adeli E, Shi F, An L, Wee CY, Wu G, Wang T, Shen D, 2016 Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. NeuroImage 141, 206–219. [PubMed: 27296013]

Aël Chetelat G, Baron Jean-Claude, 2003 Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage 18 (2), 525–541. [PubMed: 12595205]

Allman John M., et al., 2001 The anterior cingulate cortex. Ann. N.Y. Acad. Sci 935 (1), 107–117. [PubMed: 11411161]

Armanfard N, Reilly JP, Komeili M, 2016 Local feature selection for data classification. IEEE Trans. Pattern Anal. Mach. Intell 38 (6), 1217–1227. [PubMed: 26390448]

Avants BB, Cook PA, Ungar L, Gee JC, Grossman M, 2010 Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. Neuroimage 50 (3), 1004–1016. [PubMed: 20083207]

Azhagusundari B, Thanamani AS, 2013 Feature selection based on information gain. Int. J. Innovative Technol. Exploring Eng. (IJITEE) 2 (2), 18–21.

Beyer M, Aarsland D, 2008 Grey matter atrophy in early versus late dementia in Parkinson's disease. Parkinsonism Relat. Disord 14 (8), 620–625. [PubMed: 18394949]

Bohnen NI, Albin RL, 2011 White matter lesions in Parkinson disease. Nat. Rev. Neurol 7 (4), 229–236. [PubMed: 21343896]

Burton EJ, et al., 2004 Cerebral atrophy in Parkinson's disease with and without dementia: a comparison with Alzheimer's disease, dementia with Lewy bodies and controls. Brain 127 (4), 791–800. [PubMed: 14749292]

Calne D, Snow B, Lee C, 1992 Criteria for diagnosing Parkinson's disease. Ann. Neurol 32 (S1), S125–S127. [PubMed: 1510370]

Camicioli R, et al., 2003 Parkinson's disease is associated with hippocampal atrophy. Mov. Disord 18 (7), 784–790. [PubMed: 12815657]

Cao P, et al., 2017 A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. Pattern Recognit. 64, 327–346.

Chen H-L, et al., 2013 An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst. Appl 40 (1), 263–271.

Chen Xiaobo, Zhang Han, Zhang Lichi, Shen Celina, Lee Seong-whan, Shen Dinggang, 2017 Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification. Hum. Brain. Mapp 38 (10), 5019–5034. [PubMed: 28665045]

Corrigan FM, et al., 1998 Diorthosubstituted polychlorinated biphenyls in caudate nucleus in Parkinson's disease. Exp. Neurol 150 (2), 339–342. [PubMed: 9527905]

Croux C, Dehon C, 2001 Robust linear discriminant analysis using S-estimators. Can. J. Stat 29 (3), 473–493.

De La Torre F, Black MJ, 2003 A framework for robust subspace learning. Int. J. Comput. Vis 54 (1), 117–142.

Ding H, et al., 2011 Change of olfactory function associated structures in Parkinson's disease: a voxel-based morphometry study. Chin. J. Contemp. Neurol. Neurosurg 11 (1), 54–59.

Du L, Shen Y-D, 2015 Unsupervised feature selection with adaptive structure learning ACM. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Efron B, et al., 2004 Least angle regression. Ann. Stat 32 (2), 407–499.

Fidler S, Skocaj D, Leonardis A, 2006 Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. IEEE Trans. Pattern Anal. Mach. Intell 28 (3), 337–350. [PubMed: 16526421]

Goebel G, et al., 2011 A novel computer-assisted image analysis of [123I] β-CIT SPECT images improves the diagnostic accuracy of parkinsonian disorders. Eur. J. Nucl. Med. Mol. Imag 38 (4), 702–710.

González-Redondo R, et al., 2014 Grey matter hypometabolism and atrophy in Parkinson's disease with cognitive impairment: a two-step process. Brain 137 (8), 2356–2367. [PubMed: 24951642]

Gu Q, Li Z, Han J, 2012 Generalized Fisher Score for Feature Selection. arXiv preprint arXiv: 1202.3725. .

Halliday GM, 2009 Thalamic changes in Parkinson's disease. Parkinsonism Relat. Disord 15, S152–S155. [PubMed: 20082979]

Hanakawa T, et al., 1999 Mechanisms underlying gait disturbance in Parkinson's disease. Brain 122 (7), 1271–1282. [PubMed: 10388793]

He X, Cai D, Niyogi P, 2005 Laplacian Score for Feature Selection. NIPS.

Huang D, et al., 2011 Supervised local subspace learning for continuous head pose estimation IEEE Conference on. 2011. IEEE. Computer Vision and Pattern Recognition (CVPR).

Huang D, Cabral R, De la Torre F, 2016 Robust regression. IEEE Trans. Pattern Anal. Mach. Intell 38 (2), 363–375. [PubMed: 26761740]

Huber PJ, 2011 Robust Statistics. Springer.

Jia H, Martinez AM, 2009 Support vector machines in face recognition with occlusions CVPR 2009. IEEE Conference on. 2009. IEEE. Computer Vision and Pattern Recognition.

Kim S-J, Magnani A, Boyd S, 2005 Robust Fisher Discriminant Analysis. NIPS.

Kish Stephen J., et al., 2007 Preferential loss of serotonin markers in caudate versus putamen in Parkinson's disease. Brain 131 (1), 120–131. [PubMed: 17956909]

Li F, Miao D, Pedrycz W, 2017 Granular multi-label feature selection based on mutual information. Pattern Recognit. 67, 410–423.

Liu L, et al., 2016 Feature selection based on iterative canonical correlation analysis for automatic diagnosis of Parkinson's disease Springer. International Conference on Medical Image Computing and Computer-Assisted Intervention.

López M, Ramírez J, Górriz JM, Salas-Gonzalez D, Alvarez I, Segovia F, Puntonet CG, 2009 Automatic tool for Alzheimer's disease diagnosis using PCA and Bayesian classification rules. Electron. Lett 45 (8), 389–391.

Naghibi T, Hoffmann S, Pfister B, 2015 A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE Trans. Pattern Anal. Mach. Intell 37 (8), 1529–1541. [PubMed: 26352993]

Nie F, et al., 2008 Trace Ratio Criterion for Feature Selection. AAAI.

Nie F, et al., 2010 Efficient and robust feature selection via joint ℓ2, 1-norms minimization. Advances in Neural Information Processing Systems.

Ozcift A, 2012 SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. J. Med. Syst 36 (4), 2141–2147. [PubMed: 21547504]

Peng H, Long F, Ding C, 2005 Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell 27 (8), 1226–1238. [PubMed: 16119262]

Potgieser AR, et al., 2014 Anterior temporal atrophy and posterior progression in patients with Parkinson's disease. Neurodegenerative Dis. 14 (3), 125–132.

Robnik-Šikonja M, Kononenko I, 2003 Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn 53 (1–2), 23–69.

Rousseeuw PJ, Leroy AM, 2005 Robust Regression and Outlier Detection, vol. 589 John wiley & sons.

Senawi A, Wei H-L, Billings SA, 2017 A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pattern Recognit. 67, 47–61.

Shen D, Davatzikos C, 2002 HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439. [PubMed: 12575879]

Shi F, et al., 2012 LABEL: pediatric brain extraction using learning-based meta-algorithm. NeuroImage 62 (3), 1975–1986. [PubMed: 22634859]

Singh G, Samavedham L, 2015 Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease. J. Neurosci. Methods 256, 30–40. [PubMed: 26304693]

Smith SM, et al., 2004 Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219. [PubMed: 15501092]

Song F, Guo Z, Mei D, 2010 Feature selection using principal component analysis International Conference on. 2010. IEEE. System Science, Engineering Design and Manufacturing Informatization (ICSEM).

Song SK, et al., 2011 The pattern of cortical atrophy in patients with Parkinson's disease according to cognitive status. Mov. Disord 26 (2), 289–296. [PubMed: 21370255]

Tsanas A, et al., 2010 Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression IEEE International Conference on. 2010. IEEE. Acoustics Speech and Signal Processing (ICASSP).

Tsanas A, et al., 2012 Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans. Biomed. Eng 59 (5), 1264–1271. [PubMed: 22249592]

Van Huffel S, Vandewalle J, 1991 The Total Least Squares Problem: Computational Aspects and Analysis. SIAM.

Wang H, Ding C, Huang H, 2010 Multi-label linear discriminant analysis. Comput. Vis.–ECCV 2010, 126–139.

Wang H, et al., 2011 Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance IEEE International Conference on. 2011. IEEE. Computer Vision (ICCV).

Wenning G, et al., 2000 What clinical features are most useful to distinguish definite multiple system atrophy from Parkinson's disease? J. Neurol. Neurosurg. Psychiatry 68 (4), 434–440. [PubMed: 10727478]

Xia J, et al., 2013 Magnetic resonance morphometry of the loss of gray matter volume in Parkinson's disease patients. Neural Regener. Res 8 (27), 2557.

Ye J, Liu J, 2012 Sparse methods for biomedical data. ACM SIGKDD Explorations Newsl. 14 (1), 4–15.

Ye J, et al., 2012 Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 12 (1), 46. [PubMed: 22731740]

Zhang Y, Yeung D-Y, 2010 Worst-case linear discriminant analysis. Advances in Neural Information Processing Systems.

Zhang Y, Brady M, Smith S, 2001 Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57. [PubMed: 11293691]

Zhang Lichi, Wang Qian, Gao Yaozong, Wu Guorong, Shen Dinggang, 2016 Automatic labeling of MR brain images by hierarchical learning of atlas forests. Med. Phys 43 (3), 1175–1186. [PubMed: 26936703]

Zhang Li, et al., 2017a Brain metabolic correlates of fatigue in Parkinson's disease: a PET study. Int. J. Neurosci 1–16 just-accepted:.

Zhang Lichi, Wang Qian, Gao Yaozong, Li Hongxin, Wu Guorong, Shen Dinggang, 2017b Concatenated spatially-localized random forests for hippocampus labeling in adult and infant MR brain images. Neurocomputing 229, 3–12. [PubMed: 28133417]

Zhang Jinpeng, Zhang Lichi, Xiang Lei, Shao Yeqin, Wu Guorong, Zhou Xiaodong, Shen Dinggang, Wang Qian, 2017c Brain atlas fusion from high-thickness diagnostic magnetic resonance images by learning-based super-resolution. Pattern Recognit. 63, 531–541. [PubMed: 29062159]

Zhang Lichi, Zhang Han, Chen Xiaobo, Wang Qian, Yap Pew-Thian, Shen Dinggang, 2017d Learning-based structurally-guided construction of resting-state functional correlation tensors. Magn. Reson. Imaging 43, 110–121. [PubMed: 28729016]

Zhao Zheng, Liu Huan, 2007 Spectral feature selection for supervised and unsupervised learning ACM. Proceedings of the 24th International Conference on Machine Learning 1151–1157.

Zhou N, Wang L, 2007 A modified T-test feature selection method and its application on the HapMap genotype data. Genomics Proteom. Bioinform 5 (3–4), 242–249.

Zhu Xiaofeng, et al., 2012 Dimensionality reduction by mixed kernel caanonical correlation analysis. Pattern Recognit. 45 (8), 3003–3016.

Zhu X, Suk H-I, Shen D, 2014 Multi-modality canonical feature selection for Alzheimer's disease diagnosis Springer. International Conference on Medical Image Computing and Computer-Assisted Intervention.

Zou H, Hastie T, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2), 301–320.

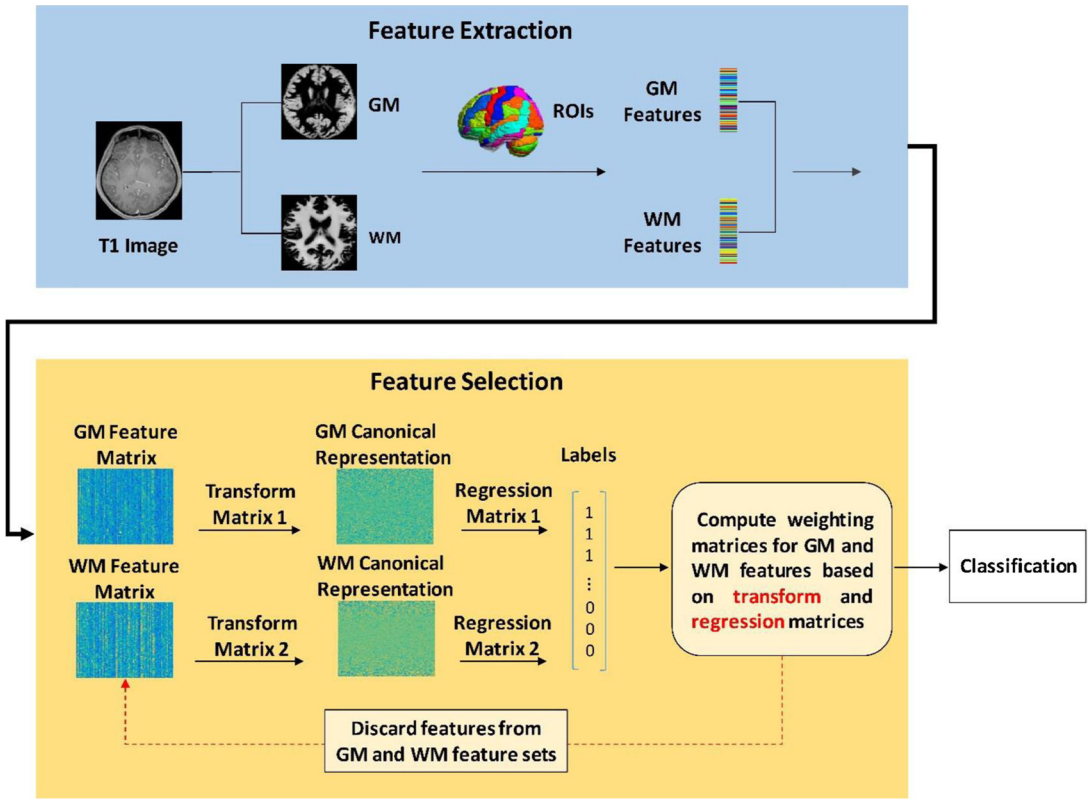

Fig. 1.

Pipeline of the automatic diagnosis system with the proposed ICCA based feature selection. We extract ROI based gray matter (GM) and white matter (WM) tissue volumes as the features. The GM and WM feature vectors per subject (or matrices for the dataset) are transformed into the canonical space by CCA. Then, we use the regression model to map the canonical representations, corresponding to the features transformed into the canonical space, toward the PD/NC labels. Finally, the regression coefficients are transformed back into the original GM and WM feature spaces, where feature selection is conducted according to the importance of the original features. The above procedure is iterated, and the finally selected features are used for classification-based PD diagnosis.

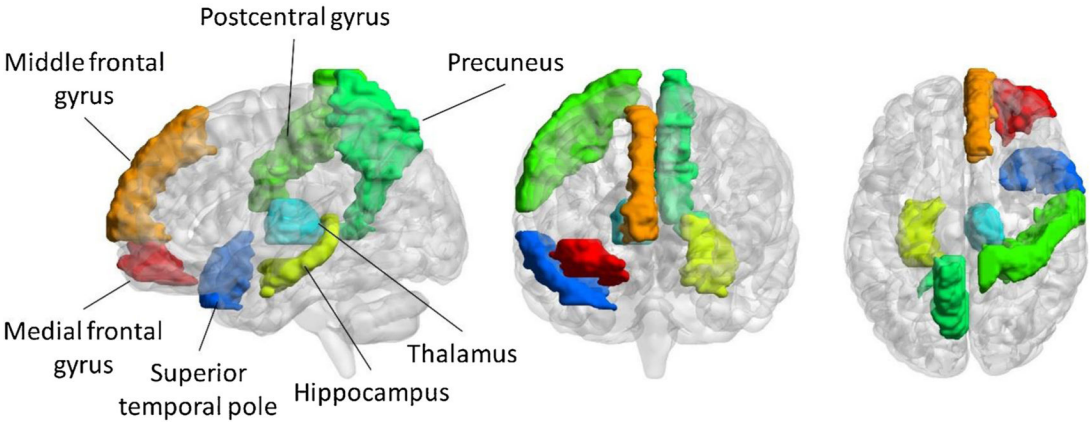

Fig. 2.
Visualization of the most discriminative ROIs used for automatic diagnosis of PD.

Table 1

PD/NC classification comparison (ACC: Accuracy; AUC: Area Under ROC Curve).

Method	ACC(%)	AUC(%)	SEN(%)	SPE(%)
NoFS	58.0	55.1	32.5	82.5
Elastic Net	64.3	64.6	54.1	87.8
SFS (with $l_{2,1}$ norm)	65.1	64.5	57.1	73.2
PCA	65.2	59.8	48.2	82.1
FSASL	65.3	66.4	56.2	77.6
CCA	66.1	64.4	53.6	78.6
RFS	66.1	63.9	53.6	**89.3**
MRMR	66.7	65.6	52.9	61.7
Cascaded-CCA (Proposed)	68.8	69.3	62.5	71.6
ICCA (Proposed)	**70.5**	**71.1**	**62.5**	78.6