MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-731

Fusing Heterogeneous Data for Alzheimer's Disease Classification

Parvathy Sudhir Pillai[a], Tze-Yun Leong[a,b], Alzheimer's Disease Neuroimaging Initiative

[a]*Medical Computing Laboratory, School of Computing, National University of Singapore, Singapore*
[b]*School of Information Systems, Singapore Management University, Singapore*

Abstract

In multi-view learning, multimodal representations of a real world object or situation are integrated to learn its overall picture. Feature sets from distinct data sources carry different, yet complementary, information which, if analysed together, usually yield better insights and more accurate results. Neuro-degenerative disorders such as dementia are characterized by changes in multiple biomarkers. This work combines the features from neuroimaging and cerebrospinal fluid studies to distinguish Alzheimer's disease patients from healthy subjects. We apply statistical data fusion techniques on 101 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We examine whether fusion of biomarkers helps to improve diagnostic accuracy and how the methods compare against each other for this problem. Our results indicate that multimodal data fusion improves classification accuracy.

Keywords:

Multimodal;Data fusion; Heterogeneous; Alzheimer's disease.

Introduction

Multimodal data fusion refers to the fusion of multiple data sources, their associated features, and (or) intermediate decisions to perform an analysis task [1]. This multimodel method has found widespread use in areas such as multimedia and sensor analyses to integrate views obtained from audio and video signals, texts and images, and others. Recent studies in medical informatics have benefitted from combining multiple data sources to better understand disease processes. In this paper, we study the impact of multimodal data fusion on classifying Alzheimers' Disease (AD) patients.

Dementia is a spectrum of neuro-degenerative disorders that lead to memory and cognitive decline, severe enough to disable a person to perform activities of daily living. AD, the most common subtype, affects close to 75% of the demented population. As of 2010, there are around 36 million affected individuals worldwide, and an enormous amount is spent on their care [2]. No definitive prevention methods/cures are available for AD. Hence, we need efficient methods to screen and study the disease early on, so that timely interventions may delay its progression.

Dementia severity is assessed by psychometric tests like Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR), neuroimaging, protein and genomic tests, and others. Biomarkers acquired from these tests provide indicators about a person's state. The sensitivity of biomarkers varies over the stages from normal aging through Mild Cognitive Impairment (MCI) to Dementia, as evident from Figure 1 [3]. Recently, pattern classification methods have been applied to analyze these biomarkers in combinations [10, 11, 12], as the information from different biomarkers is complementary in nature. While structural Magnetic Resonance Imaging (s-MRI) has good spatial resolution to identify atrophied brain regions, functional imaging such as Fluodeoxyglucose Positron Emission Tomography (FDG-PET) reveals hypometabolism in the affected brain areas. Protein studies of the Cerebrospinal Fluid (CSF) indicate the presence of beta amyloid (Aβ42) and tau (τ) proteins which form plaques and tangles in the brain, characteristic of AD. Combining multiple related data sources yields a fused representation of the object under study. Analysing this representation yields a comprehensive picture that benefits from the interplay of statistical dependences of the data sources. Further, the analysis reduces noise in the data by averaging it out over the independent data sources.

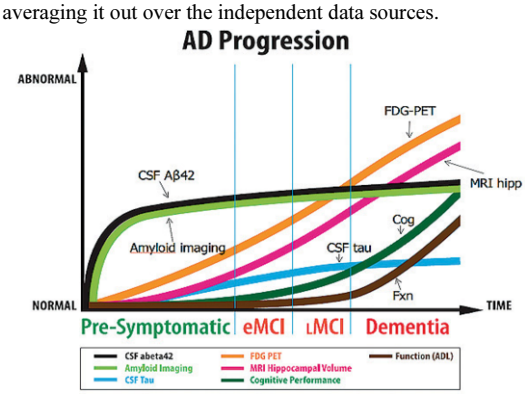

Figure 1. Biomarker sensitivity to Dementia related changes in the human brain across stages. Used with permission from the website of National Institute of Aging [3].

Motivated by these facts, we examine the effectiveness of statistical methods for fusing biomarker data to distinguish AD patients from healthy subjects (HS). On a subset of data from ADNI, we compare three data fusion methods based on:

1. Canonical Correlation Analysis (CCA)
2. Multiple Kernel Learning (MKL)
3. Collective Matrix Factorization (CMF)

While CCA ensures that the fused representation has maximally correlated features, MKL learns the optimal way to combine the features to yield the best classification accuracy. CMF is a comparatively recent method that jointly factorizes matrices that share a common dimension. We explain, implement, and test these methods on the ADNI data to compare their accuracies of classification over unimodal and prior multimodal studies.

Related Work

Quantitative fusion of medical data is very challenging because of the heterogeneity of the modalities. Two main approaches exist for combining heterogeneous information. The first approach, known as early fusion, aggregates data at the feature level into a single representation before analysis. Kernel space combination proposed by Lanckriet et al. to combine amino acid sequences and gene expressions [4], vector concatenation of Principal Component Analysis (PCA) reduced features used by Lee et al. to fuse mass spectrometry and histology information [5], and Artificial Neural Networks (ANN) used by Baez et al. to integrate various neuropsychological test scores [6] all fall under this category. Though these methods preserve inter-source dependencies, they suffer from the curse of dimensionality and hence require a large amount of training data to learn a relevant model. The second approach, known as late fusion, combines decisions from models learnt on the individual feature spaces. Various rules such as weighted combination [7], majority voting [8], likelihood maximization [9] of the decision variables have been proposed. As the fusion is at the level of decisions, there are no concerns with the dimensionality of the data. However, these methods fail to retain inter-source dependencies.

Multimodal assessments of AD and MCI were found to classify diseased individuals more accurately than unimodal methods. Zhang et al. combined MRI, PET and CSF biomarkers using multiple kernels and a coarse grid search to find the optimal kernel combination on a Support Vector Machine (SVM) classifier [10]. As compared to this discriminative approach which models the conditional distribution of variables for predicting the class labels from features, Young et al. used a variation of kernel combination with a generative Gaussian Process (GP) classifier [11]. This generative approach models the joint distribution of variables and uses likelihood maximization to learn the optimal parameters; it is shown to perform on par with the earlier discriminative approach. Gray et al. applied Random Forest (RF) proximity measures to combine MRI and PET features [12]. Though these methods provide good classification accuracy, they cannot in general support understanding of the data and their interactions. Moreover, these methods do not handle missing data or specific data types such as ordinal data.

There are three general multiview learning approaches: weighted view combination, multiview dimension reduction, and subspace learning. Inspired by the promising results of previous multimodal analyses, we aim to explore the effectiveness of three representative methods from the categories, CCA (multiview dimension reduction), MKL (weighted view combination), and CMF (subspace learning), for combining multimodal biomarker features for AD diagnosis. While Zhang et al. [10] and Young et al. [11] used MKL, only linear combination of kernels was explored. CCA and CMF have not been used in the context of fusing biomarkers for AD diagnosis. The fused representation should generalize well to related problems of supervised learning such as classification and unsupervised learning for understanding the association between biomarkers.

Methods

The goals of data fusion are as follows:

1. Reducing the dimensionality of the participating views, so that the fused representation has the most representative components of the individual views.
2. Explaining the nature of relationships between datasets by measuring the relative contribution of each variable to an analysis task.
3. Learning a joint subspace from the different views that supports interpreting the datasets well enough to handle missing data.

We explore the ability of three popular data fusion techniques in attaining these goals. In the implementations, we consider biomarker data as matrices where the rows correspond to subjects and columns to features.

Canonical Correlation Analysis

CCA seeks to find linear projections of two sets of multidimensional variables, so that the projections are maximally correlated [13]. Correlation as a relationship is heavily dependent on the chosen coordinate system; therefore, even if there is a strong linear relationship between two sets of multidimensional variables, the relationship might not be visible as a correlation.

Mathematically, if x and y are two multidimensional random variables with zero mean and $w_a{}^T x$ and $w_b{}^T y$ are their corresponding linear projections, maximizing their correlation, ρ, corresponds to solving Equation (1). If C_{ab} is the cross-covarianve, C_{aa} and C_{bb} are the auto-covariance matrices,

$$\max_{w_a, w_b} \rho = \frac{w_a^T C_{ab} w_b}{\sqrt{w_a^T C_{aa} w_a}\sqrt{w_b^T C_{bb} w_b}} \quad (1)$$

CCA is often formulated as a generalized eigenvalue problem where the maximum correlation corresponds to the largest eigenvalue.

$$\begin{pmatrix} 0 & C_{ab} \\ C_{ba} & 0 \end{pmatrix} \begin{pmatrix} w_a \\ w_b \end{pmatrix} = \rho \begin{pmatrix} C_{aa} & 0 \\ 0 & C_{bb} \end{pmatrix} \begin{pmatrix} w_a \\ w_b \end{pmatrix} \quad (2)$$

Several extensions to the original CCA have been proposed to include more than two views, and to find non-linear relationships between views. Currently, we restrict ourselves to the linear version because it is faster and involves easily interpretable components. The most commonly used approach to include three or more, say p data sources is to sum up the correlations (mCCA). The generalized eigenvalue problem then accounts for maximizing the sum of the correlations. This formulation is depicted in Figure 2.

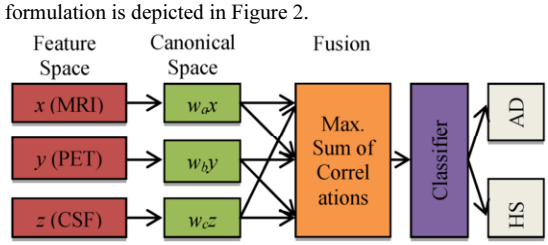

Figure 2. CCA based classification

Tripathi et al. [13] proposed a two step procedure for summing up the correlations. First, the correlations within a data source are removed by a process called whitening. This is done by multiplying the individual data matrices with the square-root of their respective covariance matrices to find components shared between the views. Second, Principal Component Analysis (PCA) is applied to the column-wise concatenation of the whitened data sources. The original data is further projected on to the largest d PCA coefficients. The choice of d is based on the amount of shared variance. The smallest d, after which there is no significant increase in shared variance, is the optimal dimension of of the projection. The projection yields the fused representation which is then used by a classifier to learn the model.

Multiple Kernel Learning

Kernel methods such as SVM, which are based on similarity measures between data points, have been used with great success for dimensionality reduction and classification. Kernelization projects the native space data to a higher dimensional feature space. Non-linear relations between variables in the original space become linear in the transformed space. The projection, ϕ is given by the mapping,

$$\phi: x = (x_1, \dots, x_n) \rightarrow \phi(x) = (\phi_1(x), \dots, \phi_N(x)) \quad (3)$$

To project the data we use the kernel trick, wherein we apply kernel functions, $\kappa_1, \dots, \kappa_p$, to get the corresponding kernel matrices $K_1, \dots, K_p$. Each kernel, $K = <\phi(x), \phi(z)>$ is an inner product of data points. Examples of kernel functions include the linear, radial bias function and others.

Using more than one kernel often produces a better model. In MKL, data is represented as a combination of base kernels [10]. Each base kernel represents a different modality / feature of the entity. MKL seeks to find the optimal combination of the base kernels so that the analysis tasks which follow are benefitted the most. Classification tasks are especially well represented through MKL, as the optimal combination is the one that gives the maximum classification accuracy.

The dual form of MKL optimization, as it is solved by conventional solvers like LIBSVM [14], is

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{p} \beta_m \mathbf{k}^{(m)}(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$$
$$\mathbf{s.t} \quad \sum_{i=1}^{n} \alpha_i y_i = 0; 0 \le \alpha_i \le C; i = 1, \dots, n \quad (4)$$

From a set of n training samples, the features of the i-th sample from the m-th modality are in the vector $x_i^{(m)}$, and its corresponding class label, y_i is either +1 or -1. α's are the Lagrange multipliers which are the variables obtained on converting the primal support vectors to the dual problem. The kernel function applied on each pair of the samples from a modality m, is $\mathbf{k}^{(m)}$. The weights on the m-th modality kernel, represented as β_m are optimized using a grid search or as a separate optimization problem with fixed α. For each new test sample, s, the kernel functions are computed against the training samples. The MKL overview is depicted in Figure 2.

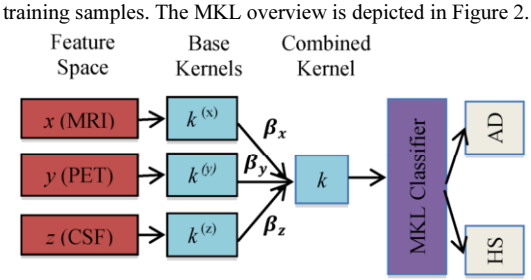

Figure 3. MKL based classification

Zhang et al. [10] and Young et al. [11] used coarse grid search and likelihood maximization approaches respectively, to find the optimal kernel weights, β. They used only one linear base kernel for each of the feature sets and constrain the β's to sum to 1 ($\beta: \|\beta\|_1 = 1; \beta \ge 0$). This however may yield sparse solutions with certain kernels not being well-represented. Recent research has shown that including the base data sets in more than one kernel each differing in their selection of kernel parameters, improves performance [15]. Regularized MKL based on $l2$ norm ($\beta: \|\beta\|_2 = 1; \beta \ge 0$) and $l12$ mixed norm ($\beta: \|\beta\|_2 \le 1; \beta \ge 0$) for constraining β have been proposed. Though $l2$-regularized MKL yields non-sparse solutions, it no longer remains a convex optimization problem and hence is difficult to solve as the sample size increases. $l12$-regularized MKL involves more than one base kernel from a single modality. It enforces sparsity across modalities, while allowing more than one discriminative kernel to be chosen from the same modaility. In other words, there is sparsity across modalities and non-sparsity within modalities, thereby making it a convex optimization problem.

Collective Matrix Factorization

CMF is a technique in relational learning for predicting the unknown values of a relation, given a database of entities and their relations. It learns the low-rank approximations of the matrices which share entities [16].

Given a set of M matrices which describe the relations among E entities, CMF approximates them to low-rank factorizations. The matrices are approximated as a rank-L product and additional row and column bias terms. If r_m and c_m are the entity sets corresponding to the row and column respectively of the m-th matrix, on factorization, its element in the i-th row and j-th column is represented as:

$$x_{ij}^{(m)} = \sum_{l=1}^{L} u_{il}^{r_m} u_{lj}^{c_m} + b_i^{(m,r)} + b_j^{(m,c)} + \varepsilon_{ij}^{(m)} \quad (5)$$

Where, $[u_{ik}^{(e)}]$ is the rank-L approximation of entity set e, $b_i^{(m,r)}$ and $b_j^{(m,c)}$ are the row and column biases respectively and $\varepsilon_{ij}^{(m)}$ is the element-wise noise. The matrices which share the same entity set share the same low-rank matrix approximation. Recent works arrange all the M matrices into a large square grid, whose dimension is the sum of cardinalities of all the entity matrices. However, in the resulting symmetric matrix, Y, only blocks corresponding to the M matrices are observed and the rest of the elements are left unobserved. The CMF model is then formulated as a symmetric matrix factorization,

$$Y = UU^T + \varepsilon \quad (6)$$

where, U is the column-wise concatenation of different $[u_{ik}^{(e)}]$ matrices and bias terms are dropped for simplicity [16].

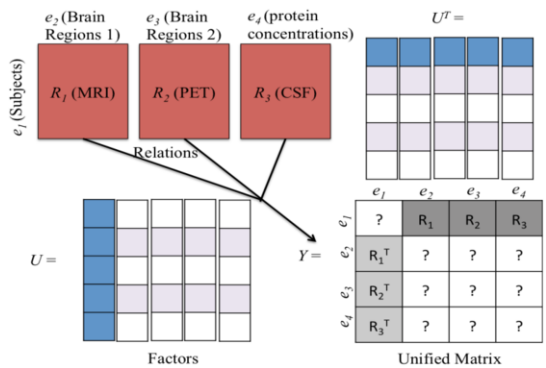

Figure 4. CMF based modeling

Experiments

We implemented the multimodal fusion approaches described above, for integrating the MRI, PET and CSF biomarkers. We used the fused representation to classify a selected study group into patients with AD from healthy subjects (HS). The fused approach is considered successful if the classification task is performed with greater accuracy along with better precision and recall against unimodal classifications. Along with the unimodal approaches, we evaluated the classification of a concatenated data vector comprising data from the three modalities and used it as a baseline study.

The three modalities: MRI, PET and CSF, complement each other in the information they hold [10]; this enables us to draw better insights in a classification task. We used these three biomarkers specifically because, as shown in Figure 1, they compare better than the others in identifying AD early on.

Data

The data for evaluation was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [17]. We worked on baseline MRI images and FDG-PET images that were acquired within 30-60 min post injection. The image details are available at the ADNI website: http://adni.loni.usc.edu/. MMSE (0-30) score of >=27 and CDR (0-3) of 0 are considered normal. The demography of the subjects that we considered are shown in Table 1.

Table 1– Subject Demography

	AD (n = 51; 18F/33M)			HS (n = 52; 18F/34M)		
	Mean	SD	Range	Mean	SD	Range
Age	75.2	7.4	59-88	75.3	5.2	62-85
MMSE	23.8	2.0	20-26	29	1.2	25-30
CDR	0.7	0.3	0.5-1	0	0	0

Preprocessing

The sequence of steps for processing the MRI images included setting the origin to the Anterior Commissure (AC), correcting intensity inhomogeneities, and performing skull stripping. As grey matter atrophy is a prominent feature in AD patients, we segmented the images into grey, white matter and the CSF. This segmentation and the subsequent steps were done using the Statistical Parametric Mapping (SPM) 8 toolbox [18]. To standardize the images of all the subjects, they were normalized to a study specific template created by the SPM DARTEL toolbox [19]. The PET images were co-aligned to the corresponding MRI image using SPM8. Masks of 83 brain regions enlisted in the atlas prepared by Kabani et al. [20] were created using a tool called WFU-PickAtlas [21]. These masks were imposed on the segmented gray matter and PET images to obtain the regional grey matter volume and the average intensity measurements respectively. Thus, we obtained a 1×83 sized feature vector per subject for each of the imaging modalities. The CSF values obtained from ADNI were represented as a 1×3 sized vector per subject representing the total tau, Aβ42 and p-tau values respectively.

We implemented CCA and MKL fusion methods in MATLAB and used the R library 'CMF', for CMF. We tested the individual modalities and the concatenated feature vector (baseline) on the following classifiers:

SVM –This discriminative classifier is accepted to be standard for binary classification. We used the popular LIBSVM [14] tool for our experiments. With unimodal data we used an RBF kernel with default parameters.

GP –We used the GPML toolbox [22] and followed Young et al. [11] for the choice of covariance, mean, likelihood and inference functions.

RF – As an ensemble classifier, we used a MATLAB version of R language's RF library. The number of trees in the classifier were varied according to the dimensionality of the dataset under consideration.

Each method was tested using 10-fold cross validation, categorizing subjects into ten groups based on a random permutation. Nine groups were used for the learning phase and the remaining group formed the test set. The accuracy, precision and recall of the classification tasks were studied. Three prior works in multimodal AD classification were reimplemented and tested with our dataset.

Results

The results of our experiments are tabulated in Table 2. It is evident that a simple concatenation of the feature vectors (SVM (c), GP (c) and RF (c)) provides better classification results than unimodal tests. Prior multimodal biomarker based methods [10, 11, 12] have better classification accuracy than the baseline study (feature concatenation) as expected. However, the results are even better for classification on the fused representation obtained from the statistical methods like CCA and CMF.

Discussion

The concatenated feature vector consistently performs better across the three types of classifiers than individual biomarkers because of their complementary information. The poor performance of the baseline study in which there is no kernel combination, against the prior multimodal analyses of Zhang et al. and Gray et al. [10, 12], is due to the inclusion of all features and not just those which contribute to classification.

The MKL formulation based on *l12* mixed norm performs worse than Zhang et al.'s [10] but better than Young et al.'s [11] both of which are *l1* norm based. As the mixed norm enforces group sparsity, it chooses features common across all participating modalities. In comparison, *l1* norm and Gray et al.'s [12] RF based method choose features individually across modalities and ignores intermodal relationships. From this we understand that the common feature constraint may overlook certain modality specific features aiding classification.

mCCA and CMF both perform better than the rest of the methods. These methods learn a generic model of the biomarkers, not specific for classification. But they perform the best on classification task as well. This is because the generic model learnt from these techniques is built only on those relevant features or components that are statistically dependent across modalities. Though mCCA in its current form is incapable of handling missing entries it may be extended to handle them. CMF performs slightly poorer than mCCA in the classification task but is the most generic model.

The three methods compare as follows, with respect to achieving the goals of data fusion:

1. mCCA is effective in data exploration to find if there are any associations between the data sources. It saves what is shared between the views and ignores variations within, thereby achieving goals 1 and 2.
2. If the goal is only supervised learning, MKL methods, *l12* and *l1* based optimization [10, 11], can be applied directly as they learn the most distinguishing multimodal features, satisfying goal 1. However, such methods fail when there is missing data. Moreover, these methods lack a proper generative model for each view, and hence cannot be used for the task of understanding the data.
3. CMF handles missing entries by treating them as test data and allows multiple likelihood functions for modeling the data. The benefit of using CMF is that it identifies common factors shared between matrices and factors specific to individual matrices. Matrix factorization results in dimensionality reduction and thus satisfies the three goals of data fusion.

Conclusion

We examined multimodal data fusion on a dataset consisting of heterogeneous biomarker data. We used three categories of

fusion methods based on CCA, MKL and CMF. Further, we used the resultant fused representation for classifying AD patients. We found that classifying based on the fused representation that preserves intermodal relationships yields better results than unimodal classification. Amongst the three methods, mCCA gives the best accuracy on our dataset closely followed by the CMF based method.

Table 2– Region of Interest Based Classification

			Precision		Recall	
Data	**Method**	**Acc.**	**AD**	**HS**	**AD**	**HS**
MRI	SVM	82.7	86.7	79.3	82.8	78.4
	GP	81.5	86.8	78.4	76.8	84.6
	RF	82.7	86.5	85.4	81.7	81.6
PET	SVM	85.5	88.4	86.4	85.3	84.3
	GP	82.6	84.2	81.3	82.1	83.1
	RF	81.5	81.4	87	94	73.6
CSF	SVM	80.6	81.2	81.3	83.1	81.6
	GP	81.5	83.2	83.9	85.9	78.9
	RF	81.6	83.7	81.9	82.6	82.9
MRI + PET + CSF	[10]	92.4	87.9	86.4	88.1	84.7
	[11]	87.5	87.9	89.6	87.7	84.6
	[12]	91.5	91.7	91.7	93.2	90.6
	SVM (c)	86.5	88.6	90	88.2	80.4
	GP (c)	89.3	89.6	93.7	91.5	82.9
	RF (c)	90.5	88.3	96.6	95.5	83.6
	mCCA	**95.1**	**94.8**	**97.1**	**96**	**94.2**
	l12-MKL	88.4	86.6	92.2	92.9	83.9
	CMF	94.4	84.5	96.3	87.3	87.3

Acknowledgments

We would like to thank Dr. Kandiah Nagaendran, Dr. Ming-Ching Wen, and Dr. Tchoyoson Lim, and their teams at the National Neuroscience Institute, Singapore, for providing the medical background and guidance. This project was supported by an Academic Research Grant No. T1 251RES1211 and a research scholarship from the Ministry of Education, Singapore. Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu/).

References

[1] Atrey P, Hossain M, El Saddik A, and Kankanhalli M. Multimodal fusion for multimedia analysis: a survey. Multimedia Systems, 2010; 16(6): 345–379.

[2] Alzheimer's Disease International. The Global Impact of Dementia 2013–2050, 2013. Available from: https://www.alz.co.uk/research/GlobalImpactDementia2013.pdf (accessed 11 December 2014).

[3] National Institute of Aging. Advances in Detecting Alzheimer's Disease. 2011-2012 Alzheimer's Disease Progress Report, 2012. Available from: http://www.nia.nih.gov/alzheimers/publication/2011-2012-alzheimers-disease-progress-report/advances-detecting-alzheimers (accessed 11 December 2014).

[4] Lanckriet GRG, M Deng, N Cristianini, MI Jordan, and WS Noble. Kernel-Based Data Fusion And Its Application To Protein Function Prediction In Yeast. Pacific Symposium on Biocomputing, 2004.

[5] Lee G, Doyle S, Monaco J, Madabhushi A, Fledman M, Master S, and Tomaszewski J. A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: Preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology. IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009.

[6] Baez PB, Viadero C, Perez del Pino M, Prochazka A, and Suarez-Araujo C. Humann-based systems for differential diagnosis of dementia using neuropsychological tests. 14th Intl. Conference on Intelligent Engineering Systems, 2010.

[7] Hua XS, and Zhang HJ. An attention-based decisionfusion scheme for multimedia information retrieval. 5th Pacific-Rim Conference on Multimedia, 2004.

[8] Radova V, and Psutka J. An approach to speaker identification using multiple classifiers, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997.

[9] Wu Y, Chang EY, and Tseng BL. Multimodal metadata fusion using causal strength. 13th annual ACM international conference on Multimedia, 2005.

[10] Zhang D, Wang Y, Zhou L, Yuan H, and Shen D. Multimodal classification of alzheimer's disease and mild cognitive impairment. NeuroImage, 2011; 55(3): 856 – 867.

[11] Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, and Ourselin S. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. NeuroImage, 2013; 2(0):735 – 745.

[12] Gray K, Aljabar P, Heckemann R, Hammers A, and Rueckert D. Random forest-based manifold learning for classification of imaging data in dementia. In: Machine Learning in Medical Imaging. Lecture Notes in Computer Science, 2011; 7009: 159–166.

[13]Tripathi A, Klami A, Kaski S. Simple integrative preprocessing preserves what is shared in data sources. BMC Bioinformatics; 2008; 9(1):111.

[14] Chang CC and Lin CJ, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011; 2(27): 1-27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15]Xu Z, Jin R, Yang H, King I, Lyu M. Simple and Efficient Multiple Kernel Learning by Group Lasso. International Conference on Machine Learning, 2010.

[16]Klami A, Bouchard G, and Tripathi A. Group Sparse Embeddings in Collective Matrix Factorization. International Conference on Learning Representations, 2014.

[17] ADNI | Alzheimer's Disease Neuroimaging Initiative. http://adni.loni.usc.edu/ (accessed on 11 December 2014].

[18] Friston K, Ashburner J , Kiebel S, Nichols T, and Penny W. Statistical parametric mapping: The analysis of functional brain images. Elsevier, London, 2006.

[19] Ashburner J. A fast diffeomorphic image registration algorithm. NeuroImage, 2007; 38(1): 95 – 113.

[20] Kabani N, MacDonald D, Holmes CJ, and Evans A. A 3D atlas of the human brain. Neuroimage 7, 1998; S717.

[21] Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fmri data sets. NeuroImage, 2003; 19: 1233– 1239.

[22] Rasmussen CE and Nickisch H. Gaussian Processes for Machine Learning (GPML) Toolbox. J. Mach. Learn. Res. 11, 2010; 3011-3015.

Address for correspondence

Parvathy Sudhir Pillai (parvathy@comp.nus.edu.sg)

Medical Computing Laboratory, School of Computing, National University of Singapore, Computing 1, 13 Computing Drive, Singapore-117417