MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190243

Modeling Multi-View Dependence in Bayesian Networks for Alzheimer's Disease Detection

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

[a] *School of Computing, National University of Singapore, Singapore*

Abstract

Early detection of Alzheimer's disease is important for deploying interventions to prevent or slow disease progression. We propose a multi-view dependence modeling framework that integrates multiple data sources to distinguish patients at different stages of the disease. We design interpretable models that can handle heterogeneous data types including neuro-images, bio- and clinical markers, and historical and genotypical characteristics of the subjects. We learn the dependence structure from data with guidance from domain knowledge in Bayesian Networks, visualizing and quantifying the conditional probabilistic dependence among the variables. Our results indicate that the hybrid dependence models also improve prediction performance.

Keywords:

Alzheimer Disease, Bayesian networks, Classification

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder that leads to cognitive decline of the elderly and renders them incapable of performing routine activities of daily living. The neuronal degeneration is often irreversible [1]; common clinical symptoms such as memory loss or speech impairment may only appear at a later stage. It is important to identify the relevant factors that may individually or collectively impact the cognitive condition or state to assist in accurate and early diagnosis. Such factors may be categorized based on demographics, physical examinations, clinical tests, cognitive assessments, etc. Each category presenting information about an individual froms a distinct *view* or perspective. We propose a framework for detecting AD at different stages by fusing multiple views of patient-related data and explicating the dependence among the factors or variables.

In practice, AD severity is usually assessed by psychometric tests such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog). Disease markers acquired from neuroimaging and protein studies are considered better indicators of AD in the early stages of Mild Cognitive Impairment (MCI) [1]. Genotype and non-behavioral background factors from demographics, family and medical history have been associated with cognitive decline. Linking potential diagnoses to background factors helps to identify the individuals at risk of developing AD and indicates ways to plan for its prevention and intervention.

We aim to build a disease model that accommodates the *correlational*, *causal* and *complementary* semantics of dependence and uncertainty from heterogeneous, multi-view data, possibly from different modalities. Clinical data is heterogeneous and variables relevant to AD coud be discrete (e.g., medical history of diabetes is binary) or continuous (e.g., individual's age). Knowledge discovery from medical data can be broadly categorized into correlation-based and causality-based. Correlation identifies how close two variables are to having a linear relationship with each other and indicates a predictive relationship. Most multi-view disease models identify correlations of bio- and clinical markers for predicting the clinical status. On the other hand, a variable A causally influences variable B if we manipulate A to different values, measure the effects on B, and observe changes in the probability distribution of B under different values of A. Incorporating the causal contributions from background factors in the model helps us understand their interactions that manifest as the individual's cognitive state. The cognitive state cannot be causally established from the background factors in a fully data-driven approach using only observational data [4]. Knowledge about disease epidemiology can help identify beneficial or risky causal factors. Complementarity separates the unique knowledge in a view, leading to better predictions through combining different views [2].

We propose a hierarchical probabilistic graphical model that incorporates two types of views: *markers* including MMSE, neuro-images that measure the cognitive state, etc., and *background* including genotypic, demographic variables that could possibly impact the cognitive state. The variables in each high-dimensional view are linear transformations of the underlying continuous, low-dimensional latent traits. The dependence between the predictor variables and the cognitive state is depicted in Bayesian Networks (BNs). We learn relevant evidence-based relations among the variables using prior knowledge of domain rules elicited from experts. We examine the key classification metrics: accuracy, precision and recall to determine the disease stage of a set of individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [5].

Related Work

Multiple linear/logistic regressions are the most popular correlation-based approaches used by clinicians and epidemiologists. Classifiers such as naive Bayes (NB), decision tree, back-propagation neural network (NN), and support vector machine (SVM) [6] have also been used for disease prediction from clinical data. With the recent progress in multi-view machine learning, combinations of markers distinguish AD patients from cognitively normal (CN) controls with high accuracy. Multiple kernel learning [2], canonical correlation [15] and shared subspace learning [3] are applications of multi-view learning that target the correlations between views of bio-

or clinical markers. These methods treat all views uniformly as results of the clinical state.

To identify risky/beneficial factors from background views, it is necessary to understand how changes in the variables affect the clinical state. Previous studies that explore associations between background factors and diseases are mostly hypothesis-driven; the validity of a hypothesis is tested with the available data. Structural equation models and dynamic causal networks [8] operate on causal assumptions given by domain experts about the disease. However, since the data may be from different sources, it is hard to include all the assumptions that may be valid in one source but not in another. Also, those relationships that are not included a priori are possibly missed.

Jin et al. [9] analyzed 16 multimodal features from ADNI, including age, sex, education, hippocampal volume, 2 average Positron Emission Tomography intensity measures, 7 Single Nucleotide Polymorphisms (SNP), MMSE and ADAS-Cog for cognitive score prediction in a BN. They included the notion of causality in determining the conditional probabilsitic dependence. However, feature or variable selection was done manually and solely based on domain knowledge, without considering other potentially important markers from the corresponding modalities. Hence, they do not address correlations among the variables within a modality.

Methods

A Bayesian Network (BN) is a probabilistic model that consists of two parts: a graphical model visualizing the dependencies among the variables and a probability model quantifying the dependencies. Learning the structure of a BN is done either purely from domain knowledge or the dependencies learnt from the data or a combination of both.

We learn probabilistic dependence models from multi-view clinical data using a hybrid approach i.e., score-based BN structure learning [9] guided by inputs from experts about the direction of dependencies. We adopt the prior knowledge of a causal approach in which genetic variables like SNPs and demographic variables such as age and sex are fixed before other variables and are not influenced by them. We assume that i) the markers assume certain values as manifestations of the clinical status and ii) non-behavioral background variables may influence the occurrence of AD and not vice-versa. We use a multi-level BN [10] to simulate this hypothetical data generation hierarchy: background traits ⇒ cognitive state ⇒ marker measures. We summarize the procedure to learn the dependency model, i.e., the BN of multi-view data in Figure 1.

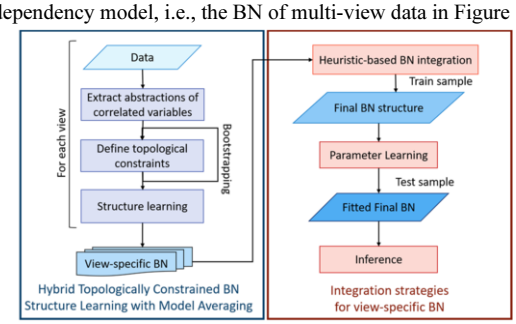


Figure 1. Modeling multi-view dependence via BN integration

We consider a supervised setting, where $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M_{mark})}$ represent the multi-view markers, $\mathbf{B}^{(1)} \ldots \mathbf{B}^{(M_{bg})}$, the background views and $\mathbf{Y}$ the response variable, i.e., clinical status of AD, the intermediate stage of Mild Cognitive Impairment (MCI) or normal controls (NC). $\boldsymbol{M_{mark}}$ is the number of marker views, while $\boldsymbol{M_{bg}}$ is the number of background views and $\boldsymbol{N}$ the number of subjects. We divide the variables in the dataset into mutually exclusive views based on their modality and the features they describe. The target variable, i.e., the clinical status of the subject is included in each view; it connects the background variables and the disease markers during view integration.

Extracting abstractions of correlated variables

In using multi-view patient data, we face the 'curse of dimensionality', which results in complex models that overfit and are not very "interpretable". There is also the issue of multi-collinearity among the variables within a view. We overcome these by abstracting out latent factors/traits from the variables which simultaneously explain their correlations and achieve dimensionality reduction. The latent factors are continuous-valued, follow a normal distribution with zero mean and unit covariance and, emulate a continuum: low to high, sick to healthy, etc. We use Bayesian matrix factorization (BMF) [14] to extract latent factors from continuous-valued views. We represent markers, $\mathbf{X}^{(j)}$, and background variables, $\mathbf{B}^{(k)}$ as linear transformations of uncorrelated low-dimensional latent factors, $\mathbf{Z}^{(j)}$ and $\mathbf{U}^{(k)}$ with $\boldsymbol{l}^{(j)}$ and $\boldsymbol{l}^{(k)}$ numbers of latent factors respectively, as depicted in Equation 1.

$$\boldsymbol{X}_{pq}{}^{(j)} = \sum_{r=1}^{l^{(j)}} \boldsymbol{Z}_{pr}^{(j)} \boldsymbol{V}_{rq}{}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \qquad (1)$$

$$\mathbf{Z}_{\boldsymbol{p}}^{(\mathrm{j})} \sim \boldsymbol{\mathcal{N}}_{l^{(j)}}\,(\mathbf{0}, \mathbf{I}_{\mathrm{l}^{(\mathrm{j})}}), \mathbf{V}_{\boldsymbol{q}}^{(j)} \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\, \boldsymbol{\lambda}_{\mathbf{q}}^{(\mathrm{j})^2}), \boldsymbol{\varepsilon}^{(j)} \sim \boldsymbol{\mathcal{N}}(\mathbf{0},\, \boldsymbol{\sigma}_{\mathbf{q}}^{(\mathrm{j})^2})$$

where $\mathbf{V}^{(\mathrm{j})}$ is the weight matrix and $\boldsymbol{\varepsilon}^{(j)}$ is the noise. Further, we specify a Gamma prior for the inverse variances, $\boldsymbol{\sigma}_{\mathbf{q}}^{(\mathrm{j})^{-2}}$ of the noise term, $\boldsymbol{\varepsilon}^{(j)}$ i.e., $\boldsymbol{\sigma}_{\mathbf{q}}^{(\mathrm{j})^{-2}} \sim$ Gamma $(\boldsymbol{a_\sigma}, \boldsymbol{b_\sigma})$, where $\boldsymbol{a_\sigma}$, $\boldsymbol{b_\sigma}$ are hyperparameters. The impact of each factor in $\mathbf{Z}^{(j)}$ on the actual variables in $\mathbf{X}^{(j)}$ can be derived by inspecting $\mathbf{V}^{(j)}$. The higher a variable weighs, higher is the corresponding factor's ability in capturing its correlations with other variables. Specifically, the variables with higher weights in a factor are interpreted as a cluster because they are similar due to their high correlations. To make the clusters interpretable, we need to reduce the number of variables with higher weights on many factors. Thus we apply sparsity constraints on the weights to have fewer non-zero weights and seggregate the strongest signals. We apply the constraint as a prior probability to the weight matrix. However, with the multiple heterogenenous views of data, our model needs to handle outliers and unknown sparsity structures of variables. For this we apply the horseshoe prior on the weight matrix, $\mathbf{V}^{(\mathrm{j})}$, with a shrinkage parameter, $\boldsymbol{\lambda}_{\mathbf{q}}^{(\mathrm{j})} \sim$ Half- Cauchy (0,1), that is robust at handling unknown sparsity and leaves out large outliers [12].

For categorical-valued views, we use multi-dimensional Item-Response Theory (IRT) models [16] to express the observed variables as resulting from continuous latent traits. For example, the probability of a person achieving a certain score in a test is a consequence of their ability, the test's difficulty and the discrimination power of a question. It is then possible to assume that individuals with the same ability answer a question correctly with a certain probability. Cognitive capability is an interpretation of this ability with respect to MMSE. IRT relates the latent ability, $\boldsymbol{\theta_i}$, to the probability of getting the correct answer to a question through the Item Characteristic Curve (ICC). Difficulty, $\boldsymbol{\delta_j}$, is practically the point on the ICC where the probability for answering correct is 50%, whereas, discrimination, $\boldsymbol{\alpha_j}$, is the slope of the ICC. An

ICC with a steep slope imparts sudden value jumps even for a small change in ability. This in turn implies that the particular question's discrimination power is very high. Meanwhile, if the item achieves a median (50%) probability at higher values of cognitive ability, the item is difficult. Equation 2 shows the probability of choosing the right option, $\boldsymbol{j}$, for a question, $\boldsymbol{i}$.

$$\boldsymbol{P(X_{ij} = 1|\theta_i,\ \alpha_j,\ \delta_j) = \frac{exp(\alpha_j(\theta_i - \delta_j))}{1+exp(\alpha_j(\theta_i - \delta_j))}} \quad (2)$$

where $\boldsymbol{\theta_i \sim \mathcal{N}(0, 1)}$. For ordinal-valued views, where there are more than two possible values of responses, we use the Partial Credit Model (PCM). The above equations apply for both marker and background views.

Learning the structure of dependence in a view

Directed edges in a BN represent the probabilistic dependence between variables, with the child node being conditionally dependent on the parent. Therefore, learning the structure of dependence among the latent factors from a view translates to learning the edges of the BN. However, as the number of variables in a view increases, an exhaustive search in the space of possible graph structures becomes infeasible. In clinical data, there my be many variables of interest in each view of different data types. We use the Conditional Gaussian BN (CGBN) framework [9] with multinomial and Gaussian distributions to model discrete and continuous nodes respectively. The discrete nodes are parameterized by conditional probability tables, with their values depending on their discrete parents. The mean of a continuous node is deduced as a linear regression of the continuous parents on each configuration of its discrete parents. Variance, however, depends only on the discrete parents. CGBN disallows discrete nodes with continuous parents. We incorporate the following techniques in the CGBN structure learning procedure to reduce the search space.

Incorporating topological constraints

Certain constraints which represent either scientific laws, common sense, expert opinions, accumulated personal experiences, etc., help to build better structures. These are usually qualitative constraints and do not signify the strength of dependence between variables. Following the convention in Li and Leong [3], we define the domain constraints in Table 1. The values of demographic (e.g., "Age", "Sex") and genetic variables (e.g., SNPs) are fixed before other variables and are not influenced by them; these should be roots in the BN. On the contrary, the values of bio- or clinical markers are dependent on other variables and do not affect other variables; these should be leaves in the network. Older age is known to be linked to AD and hence there is a directed edge from "Age" to the clinical status. We also restrict edges from clinical status and latent markers to latent background and, latent markers to status.

Table 1. Types of topological constraints in a BN

Constraints	Description
Roots	Cannot be children of other nodes
Leaves	Cannot be parents of other nodes
Known links	Links that domain experts know exist
Forbidden links	Links that domain experts know don't exist
Ordering	Chronological or logical ordering of variables

BN structure learning with search space reduction

We achieve further search space reduction by trying to build edges only between variables that are correlated. For each node the candidate neighbors with a definitive correlation, we create the dependency structure of a view using a search and score algorithm. We use a greedy hill climbing strategy that searches for a possible distribution of edges, while trying to maximize the Bayesian information criterion (BIC) score [10]. BIC measures the goodness of fit of the structure given the data, but also penalizes complicated structures with many parameters to learn. The search begins from an empty graph. At each step, the greedy algorithm performs an edge operation (adds, deletes or reverses a directed edge) that increases BIC maximally.

Model averaging

We use the bootstrap approach to assess the confidence of network structures learnt from a few hundreds of instances. Accuracy is a measure of confidence on the presence of an edge and its direction in the BN structure [9]. For a number of iterations, the algorithm re-samples the same number of instances, N as the dataset, D, with replacement. It further learns the network structure from this re-sampled dataset. The structures resulting from each iteration are averaged to help identify the nodes and edges which appear in at least half of them. We use only those edges which have edge strengths more than 0.5 and probability of edge direction more than 0.5. The edge strength signifies the confidence on the direct dependence relationship between two nodes and is estimated as its empirical frequency over the set of networks learned from bootstrap samples. The probability of the edge direction is computed conditional on the edge being present in the network.

Integration of multiple views

We learn the dependence model of the multi-view data by selectively combining nodes and edges from the view-specific BNs. In the second phase of structure learning, edges between latent variables from different views are established. From each view-specific BN, we segregate the nodes that form the Markov blanket (MB) of the clinical status node. The set of children, parents, and spouses of the node X is its MB [9]. In a BN, a variable is conditionally independent of all the other variables given its MB. Since the bootstrap approach iterates a number of times while resampling of the training data with replacement, we have as many candidate structures as there are iterations. If the edge between nodes X and Y is not present in any of the candidates, the edge strength of $E_{X,Y}$ is 0. If it appears in all the candidates, the strength is 1. Since the edge strength, Γ is a positive value between 0 and 1, it can be regarded as a probability of the validity of the edge in the BN, i.e., $\Gamma_{E_{X,Y}} = \frac{|E_{X,Y}|}{n_{iterations}}$. Thus, we follow a Bayesian approach with only the MB nodes from view-specific BNs and edges across views to be learned afresh with a default uniform prior and edges within views learned with a prior probability.

Parameter learning and inference

We learn the parameters of the integrated BN structure through maximum likelihood estimation with Laplace smoothing over the n_{train} training samples. We compute the predictions of the clinical status by averaging likelihood weighting simulations. The predicted value is the one with the highest conditional probability for a discrete target, and the highest expected value of the conditional distribution for a continuous target.

Experiments

Data: We work with a dataset of 589 subjects from ADNI which includes their background variables: demographic (4), genotypic (SNP~900) and medical history (25), and markers: grey matter volumes from baseline MRI (90), and cognitive measures: Mini Mental State Examination (MMSE) and

Clinical Status (AD/MCI/CN). Table 2 shows the demographics of the subjects included in our study.

Setup: We use the Stan probabilistic programming language to build view-specific abstractions and decide on the dimensions through multiple cross-validations [14]. We build and represent the BNs programmatically using the bnlearn R-package [9]. The integrated BN is used to predict the clinical status (2- AD, 1-MCI and 0-CN). For all the classification problems, we learn one BN structure per view with disease status as the target variable. This is followed by structure and parameter learning for the integrated BN for different view combinations. We report the performance of all models averaged over 5-fold cross validations. We evaluate the classification of a concatenated data vector comprising data from the three modalities, viewing it as a baseline study.

Table 2. Subject Demography

	AD (n = 128) 66 M, 62 F		**MCI (n = 287)** 179 M, 108 F		**CN (n = 174)** 98 M, 76 F	
	Mean	Range	Mean	Range	Mean	Range
Age	75.4 (±7.3)	58.4-87.7	75.3 (±7.2)	55.1-88.8	75.2 (±5.2)	62-84.7
MMSE	23.77 (±1.9)	20-26	27.12 (±1.6)	24-30	29.04 (±1.2)	25-30

Assumptions: We learn view-specific BNs using only the constraints that apply to the variables within. As the disease state is a categoric variable in classification, we discretize its possible parent variables such as "Age", "Years of education" according to the rules of CGBN. Only views with more than 5 variables are subject to the extraction of low-dimensional abstractions. We use a subset of SNPs (~900) as previously reported in [3]. Each SNP is matched to the gene/s located within a distance of 20kb. We learn a latent score for each gene per person using PCMs and use these for further processing.

Results

BMF captures the dense correlations among 90 regions of interest (RoI) captured into 25 latent factors. Figure 2 depicts the weights of the top 10 features (y axis) on the 25 factors, with longer bars indicating higher absolute weight in the corresponding color-coded factor. Left Amygdala has the longest bar, and hence the highest absolute weight. The latent factors are interpreted based on the hemisphere, lobe and area of the brain surface where the constituent RoI belongs. In Figure 3, the discrimination and difficulty parameters of the ICC for MMSE questions are elaborated in the left and right subplots. The five most discriminating questions are related to attention in computing alphabets in a series (MMO, MMR, MMW), language (MMWATCH), reading (MMREAD) and orientation (MMSTATE) while the five most difficult questions are alphabets in a series (MMD, MMO, MMR), orientation (MMONFLR) and immediate recall (MMTREE).

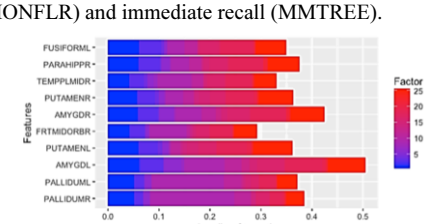


Figure 2. RoIs with highest absolute weights on factors

Table 3 shows the genes abstracted from their corresponding SNPs and factored brain RoIs that report the 5 highest associations with AD (p-values <= 0.05). Table 4 compares the performance of our approach for the multiclass classification of AD vs. MCI vs. CN with the baseline, single views, and the state-of-the-art classification for AD [7]. Figure 4 presents the integrated BN structure of demography and the abstractions from marker views of MRI, MMSE and background views of genes and medical history. The MRI factors are color-coded in navy, genes in yellow, clinical status (DX) in red, demographic variables in pink, medical history in ivory and MMSE in green.

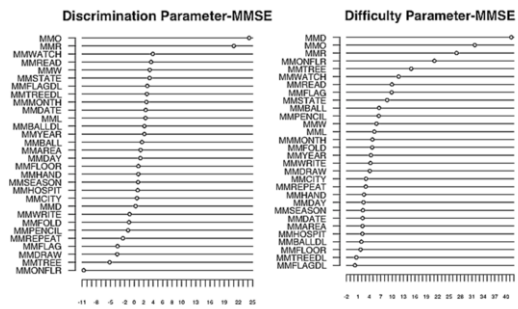


Figure 3. Discrimination and difficulty of MMSE (low-high)

Table 3. Genes, Brain areas identified by our model as associated with AD

Gene	p-value	Brain Area	p-value
APOE	<.001	Left Hippocampal	<.001
TOMM40	<.001	Right Hippocampal	<.001
SLC6A11	<.001	Left Amygdala	<.001
IL16	<.001	Right Amygdala	<.001
FGD6	.002	Left Insula Medial	<.001

Table 4. Key performance metrics for classifying AD/MCI/CN

Study/Data	**Accuracy%**	**Precision**	**Recall**
MMSE	71.7	0.58	0.91
Baseline	59.44	0.64	0.58
(Our work)	**80.18**(±8.89)	**0.73**(±.06)	0.68(±**.09**)
[7] Used MRI, PET	72.9	NA	NA

Discussion

Most clinicians use only total MMSE scores to assess AD staging. This reports an accuracy of 71.7%. Our method achieves an accuracy of 80.18% along with a precision of 0.73 and a recall of 0.68. The reported measures are averaged over predictions of AD, MCI and HC subjects by the model. The performance of the baseline (vertical concatenation of views) is not very impressive due to the inclusion of more features compared to training samples, i.e., *'large p small n'* that leads to model overfit. It underperforms grossly while working with 589 individuals and ~1030 features. The results in Table 4 show that our method fares better than the state-of-the-art approach to identify the stage of AD [7]. By categorizing the data views and defining edge directions, we are also able to visualize and interpret the model structure (refer to Figure 4). We notice that MRI factors are correctly modeled as interacting mostly through the disease variable and form a connected network. As we include only the variables from the Markov blanket (MB) of the disease variable, the less relevant ones are filtered out from each view due to the conditional independence property. The inclusion of the background features from genes, medical history and demography in the final BN in the MB of the clinical status variable (DX) implies that they are not

overshadowed by the markers. We also compare our results with a recent work [8] which used BN structure learning to predict the MMSE scores using a highly selective list of 16 biomarkers from various modalities. They report a mean square error of 2.81, while we achieve 2.44 ($\pm$0.29). Feature selection from each modality is done manually in their prediction model. Our approach also facilitates relevant biomarker discovery. The latent factors extracted from the marker and background views and visualization of correlated features lead to hitherto unexplored associations with the clinical status. Finally, in Figure 5, we report the impact of significant clinical features in the MB of the clinical status variable in the network we learnt. We examine the probabilities of a clinical status of AD or MCI conditioned on different values of the variables (scaled/standardized for continuous variables) in the MB of the clinical status. The results show that lower grey matter volume of the left hippocampus and insula region are indicative of a higher probability for cognitive impairment. The presence of APoE and TOMM40 SNPs strongly increase the chance of cognitive impairment and so does a lower education, MMSE score and higher age. These results are reasonable based on published evidence.

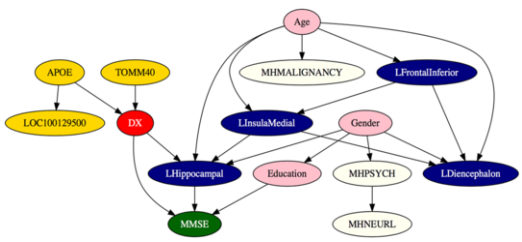
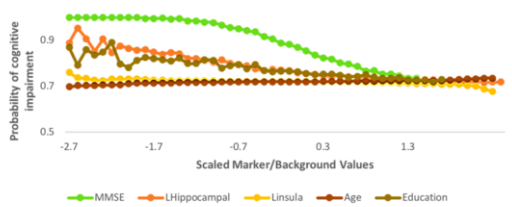


Figure 4. Multi-view integrated BN



Categories	APOE	TOMM40
0	0.62	0.42
1	0.76	0.69
2	0.92	0.72

Figure 5. Conditional probabilities for cognitive impairment

The CGBN framework, however, does have several restrictions on the nodes and their distributions. In particular, only direct conditional dependencies can be viewed. There could be hidden or non-linear associations that are not explicitly captured.

Conclusion

We have introduced a multi-view disease staging framework that takes into account the dependence semantics among the variables from disparate data sources. Our model achieves comparable and sometimes better prediction performance in identifying individuals at different stages of AD as compared with the state-of-the-art approaches. Moreover, our framework is scalable, takes into account the heterogeneity and the multitude of the data types, modalities, and quantities. It also serves as the basis of a general approach to clinical decision support with interpretable recommendations from multi-view data.

Acknowledgements

We would like to thank Dr. Feng Lei and his team at the Department of Psychological Medicine, NUS Yong Loo Lin School of Medicine, for providing the medical background and guidance. This project was supported by an Academic Research Grant No. MOE2016-T2-2-068. Data used in the preparation of this article was obtained from the ADNI database [5].

References

[1] Jack Jr CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The Lancet Neurology. 2013 Feb 1;12(2):207-16.

[2] Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage. 2011 Apr 1;55(3):856-67.

[3] Xu Z, Zhe S, Qi Y, Yu P. Association discovery and diagnosis of Alzheimer's Disease with Bayesian multiview learning. Journal of Artificial Intelligence Research. 2016 Jun 23;56:247-68.

[4] Li G, Leong TY. Biomedical knowledge discovery with topological constraints modeling in Bayesian networks: A preliminary report. In: Proceedings of the 12th World Congress on Health (Medical) Informatics. 2007 (p. 560). IOS Press.

[5] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics. 2005 Nov 1;15(4):869-77.

[6] Williams JA, Weakley A, Cook DJ, Schmitter-Edgecombe M. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In: Workshops at the twenty-seventh AAAI conference on artificial intelligence 2013 Jun 29.

[7] Zhu X, Suk HI, Lee SW, Shen D. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Transactions on Biomedical Engineering. 2016 Mar;63(3):607-18.

[8] Jin Y, Su Y, Zhou XH, Huang S. Heterogeneous multimodal biomarkers analysis for Alzheimer's disease via Bayesian network. EURASIP Journal on Bioinformatics and Systems Biology. 2016 Dec 1;2016(1):12.

[9] Scutari M. Learning Bayesian Networks with the bnlearn *R* Package, Journal of Statistical Software. (2010). 35(3):1-22.

[10] Pradhan M, Provan G, Middleton B, Henrion M. Knowledge engineering for large belief networks. In: Proceedings of the Tenth international conference on Uncertainty in artificial intelligence 1994 Jul 29 (pp. 484-490).

[11] Nakajima S, Sugiyama M. Theoretical analysis of Bayesian matrix factorization. Journal of Machine Learning Research. 2011;12(Sep):2583-648.

[12] Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. In: Artificial Intelligence and Statistics 2009 Apr 15 (pp. 73-80).

[13] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Sage; 1991.

[14] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017 Jan 1;76(1):1-37.

[15] Zhu X, Suk HI, Lee SW, Shen D. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain imaging and behavior. 2016 Sep 1;10(3):818-28.

Address for correspondence

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

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