Informatics in Medicine Unlocked 6 (2017) 28–35

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Devising an interpretable calibrated scale to quantitatively assess the dementia stage of subjects with alzheimer's disease: A machine learning approach

C.R. Aditya[a,c,*], M.B. Sanjay Pande[b,c]

[a] Department of CSE, VVIET, Mysuru, Karnataka, India
[b] Department of CSE, SITAR, Channapatna, Karnataka, India
[c] Visvesvaraya Technological University, Belgaum, Karnataka, India

ARTICLE INFO

Keywords:
Alzheimer's disease
Dementia
Multifactor dimensionality reduction
Knowledge discovery
Similarity measure
Affiliation analysis

ABSTRACT

Background: Machine learning and data mining techniques have been successfully applied on MRI images for detecting Alzheimer's disease (AD). But only a few studies have explored the possibility of AD detection from non-image data. These studies have applied traditional data visualization and classification algorithms. There is a need for new sophisticated learning algorithms, for detecting and quantifying the severity of AD by exploring the complex interactions between the features in AD subjects.

Method: In this work, a supervised learning model to effectively capture the complex feature interactions, in the sample space of AD data, is presented for knowledge discovery. The discovered knowledge is further used to quantify the similarity of a test subject to the demented class.

Results: Evaluation of the proposed model, on OASIS database of Alzheimer's subjects, validates the well established risk factors and identifiers for AD: Age, Socio-Economic Status, MMSE Score and Whole Brain Volume. The Test subjects are affiliated to either non-demented (ND) or AD class, with non-overlapping and measurable similarity indices: Female ND (CDR=0) [0.48–2.90], Female AD (CDR=0.5) [90.16–774.51], Female AD (CDR=1) [1633.90–7182.23], Female AD (CDR=2) [55258.51–66382.44], Male ND (CDR=0) [0.69–3.66], Male AD (CDR=0.5) [99.18–647.51] and Male AD (CDR=1) [3880.16–6519.40].

Conclusion: The outcome of the work clearly demonstrates that, supervised learning model can be used effectively to quantify the severity of AD on a standard measurable scale. This scale of distance can be used as a supplement for clinical dementia rating.

1. Introduction

Alzheimer's disease (AD) is the most common type of dementia disease involving degeneration of the brain which is irreversible and gradually ends up with the complete brain failure. According to the statistics of Alzheimer's Association, AD accounts for 60–80% of the dementia cases [1]. In 2006, there were 26.6 million sufferers world-wide, and is expected to double by 2030 and triple by 2050 as projected by world health organisation. AD is predicted to affect 1 in 85 people globally by 2050, and at least 43% of prevalent cases need a high level of care [2,3]. Aging and other factors increase the possibility of neuron degeneration and can lead to AD. As the world is evolving into an aging society, the burdens and impacts caused by AD on families and the society will be increasingly pronounced. Studies have shown that AD is influenced by several factors such as age, education and socio-economic status. In normal aging, whole-brain volume decline begins in early adult hood and accelerates in advanced aging [4–8]. Preferential volume loss of gray matter [9] and regionally specific thinning of the cortex are also noted [10]. Level of education, sex, socioeconomic status, and cardiovascular health have been identified as contributing factors in volume decline in advanced aging, suggesting that subclinical health conditions contribute to age related changes in brain structure [11–15]. Individuals with clinically diagnosed AD show substantially reduced over all brain volumes relative to age matched peers as well as regional volume loss that has been well documented in the hippocampal formation, among other regions [16–20].

Many models have been created to analyse and detect AD from MRI images [21–23]. Techniques such as neural networks, support vector machines, decision tree classifiers have been successfully applied on MRI images to find the region of interest responsible for causing AD

* Corresponding author at: Department of CSE, VVIET, Mysuru, Karnataka, India.
E-mail address: aditya.ise@gmail.com (C.R. Aditya).

http://dx.doi.org/10.1016/j.imu.2016.12.004
Received 9 September 2016; Received in revised form 29 December 2016; Accepted 30 December 2016
Available online 03 January 2017
2352-9148/

[24–26]. But only few studies have explored the possibility of AD detection from non-image data using machine learning approaches [27,28]. The challenges faced by these machine learning approaches are, capturing complex interactions among features in the data set and the ability to handle high-dimensional data(if required). Even though the existing approaches (both image based and non-image based methods) classify the test subjects to either normal or abnormal class, they fail to quantify the degree of abnormality on a standard scale. In medical data analysis such as detecting AD, it is important that a standard measurable scale is formulated to compare the severity of the disease. It is required to affiliate whether the sample is normal or abnormal on a measurable scale. Computational models provide a better means of modelling complex systems (such as the nervous system, neuro-disorders, etc.). To arrive at a good model for analysis, we need to design the knowledge base (KB) and ensure accuracy in the classification of data samples. Creation of a reference KB implies consolidating a huge database of control/training set into knowledge parameters, which provide meaningful and useful comprehension for later analysis. In this study we explore the possibility of analysing AD subjects from non-image data using supervised learning approach, and quantify their abnormality using measurable similarity indices. The improvements needed over existing models, which are addressed by the proposed model, are as follows:

1. Inter-feature relationship has to explored and captured in a presentable knowledge base.
2. Affiliation has to be carried out and represented through a suitable similarity index.
3. High Dimensional data (If any) must be handled.

The main theme of this research paper is to devise a computational model for critical analysis of an input (a test subject). Two phases are involved in the suggested computational model: (i) Learning Phase (ii) Recognition Phase. Learning phase involves building up a reference KB using known demented and non-demented (ND) subjects. In the recognition phase a test subject is contrasted with the reference KB to decide the label (ND/AD), and further the degree of belongingness (affinity) of the subject with respect to AD class.

This research communication is organized in the following way: Section 2 provides a detailed picture of computational model and the dataset used for evaluating the proposed model. Section 2.1 describes the data-set used for validating the model. *Multi-factor Affiliation Analysis* [29], a powerful computational model for KB creation and affiliation analysis of target data is explained in Section 2.2. Section 3 includes experimentation results and Section 4 includes useful discussions. Section 5 provides the conclusion.

2. Materials and methods

2.1. Data-set

The proposed model is evaluated on data obtained from the Open Access Series of Imaging Studies [30]. The dataset consists of a collection of 354 observations for 142 subjects aged 60–96. For some patients the observations are recorded more than once. The dataset contains both men and women subjects who are all right-handed. The data also includes the *education level* (EDUC) and *socio-economic status* (SES) of the subjects. Moreover, some other medical statistics exist in the dataset, including *intracranial volumes* (eTIV) and *brain volumes* (nWBV) of the subjects. The two groups of subjects are demented and non-demented in which the patient has the AD or not, respectively. Patients that develop the AD during the tests are grouped as *converted*.

The features in the dataset are explained in Table 1. *Clinical Dementia Rating* (CDR) can only take values 0, 0.5, 1 and 2. CDR being equal to 0 corresponds to non-demented subject. CDR being equal to 0.5, 1 and 2 corresponds to very mild dementia, mild dementia and moderate dementia respectively. This medical test carries significance to entitle a subject as Alzheimer patient. The *minimental state examination* (MMSE) is a questionnaire test that has 30 questions covering cover arithmetic, memory, and orientation to examine the cognitive situations of individuals.

2.2. Multifactor affiliation analysis (MAA)

As pointed out in introduction section, the improvements needed over existing models are the ability to effectively capture inter-feature relationships in data and to handle high dimensional data(if any). Also the knowledge base has to be further used in quantifying the similarity of a test subject with a target class. Non-parametric methods such as multifactor dimensionality reduction (MDR) and combinatorial partitioning method [31–33] were developed to handle high dimensional data and to uncover complex relationships between the data features. MAA, an MDR based model, effectively captures the inter-feature relationships in the form of a knowledge-base containing affiliation weights, which can be further used for quantifying the test subjects on a measurable scale. MAA uses the Case-Control ratio comparison concept of MDR [34] to reduce the $m*n$ feature space into $m*1$ dimensional space. As the feature dimension is reduced, a knowledge base, which captures the complex inter-feature relationships, is generated in the form of *multi-factor affiliation table* (MAT). The affiliation of a test subject with the AD class is carried out by comparing the feature values of the test subject with the knowledge in MAT.

2.2.1. MAT generation

Initially the feature values are converted to a value 'n' or 'a', depending on whether the values are close to the mean of ND or AD subjects respectively. In the changed representation space of data, the possible combinations of values for any two selected features are: $n-n$, $n-a$, $a-n$ and $a-a$. The occurrence count of each of these combinations in ND and AD subjects is compared, and selected features are merged to form a new feature. The value of new merged feature for each combination of values in old features and multifactor affiliation weight, calculated using equation (1), is updated in the MAT. This process is repeated till $m*n$ dimensional dataset gets reduced to $m*1$ dimensional dataset. MAT thus created will contain one row for each combiation of feature values and one column for every new feature created. So a MAT generated for an $m*n$ dataset will have 4 rows and $n-1$ columns. The interpretation of MAT is summarized in Table 2.

Multifactor affiliation weight (MAW), updated for each combination of values in selected features, quantifies the degree of occurrence of a value combination in AD group. MAW is designed such that it varies between zero to one, where one is the highest quantifying factor for a value combination. Let $ND_{occurance}$ represent the occurrence count of a value combination in ND subjects and $AD_{occurance}$ represent the same in AD subjects. Then MAW exhibits the following properties:

Property 1: If $AD_{occurance} = 0$ and $ND_{occurance} = m$, then $MAW=0$. A value zero for MAW indicates that the occurance of a value combination is unique to ND subjects, and AD subjects do not exhibit that pattern of combination.

Property 2: If $AD_{occurance} = m$ and $ND_{occurance} = 0$, then $MAW=1$. If the degree of occurance of a value combination is unique and at its maximum value for AD subjects, then MAW must be equal to one.

Property 3: If $AD_{occurance} >= ND_{occurance}$ then $0.5 < MAW < 1$, else $0 < MAW < 0.5$. As the occurance count of a value combination increases in AD subjects then the value of MAW increases towards 1, else it decreases towards 0.

$$MAW = 0.5 + \left[\frac{AD_{occurance}}{2*m} - 0.5*\frac{ND_{occurance}}{2*m}\right] \quad (1)$$