Prediction of Alzheimer's Disease using Oasis Dataset

Chandni Naidu, Dhanush Kumar, N Maheswari, M Sivagami, Gang Li

Abstract: ***Alzheimer's Disease (AD) is hard to predict in the early stage. But giving treatment at an early stage of AD is more effective and causes less damage to people. Various approaches like Random Forest, Support Vector Machine, Gradient Boosting and Lasso Regression have been applied to identify the best parameters for the Alzheimer's Disease prediction. Accuracy results are tabulated. Alzheimer's Disease has been predicted using Open Access Series of Imaging Studies (OASIS) dataset. Random Forest has the best accuracy rate of 97.94% and SVM has the least accuracy rate of 93.6%.***

Keywords: OASIS; Alzheimer's Diease; MRI; Random Forest; Gradient Boosting; Lasso; SVM

I. INTRODUCTION

Alzheimer's Disease is a progressive chronic neuro disease that destroys memory and other important memory functions and leads to short term memory loss and paranoid suspicion which are mistaken to be the effects of stress or ageing [1]. About 5.1 million people of the USA's total population suffer from this disease [2]. There is no proper medical treatment for AD. The only way is to control AD by using continuous medication. AD is chronic so it could stay for few years or stay for a lifetime. To have a better effect of medication we must prescribe it at an early stage to avoid a large amount of damage to the person's brain. So, detection of this at an early stage is a tedious task and very costly as we need a large amount of data, tools for prediction and a highly-experienced doctor [1]. When compared to visual assessment by a medical expert, automated systems are unbiased towards human mistakes and can be incorporated in medical decision support systems. [3] When we look back at the previous works related to AD, researchers have worked on this disease used image (MRI scans), biomarkers (chemicals and blood flows) and numerical values extracted from the MRI scans. By doing so they could predict whether the person is demented or non-demented. If the prediction of Alzheimer's is automated then there will be less human interaction and the time for prediction is also faster. The overall cost for automation can be reduced and the results are more accurate. The prediction techniques are applied on the data extracted from MRI scans to predict whether the patient is demented or not.
A person is classified as demented if he/she is suffering from early stage Alzheimer's Disease. This helps us achieve better accuracy.

Revised Manuscript Received on March 25, 2019.
Chandni Naidu, SCSE, Vellore Institue of Technology, Chennai
Dhanush Kumar, SCSE, Vellore Institue of Technology, Chennai
N Maheswari, SCSE, Vellore Institue of Technology, Chennai
M Sivagami, SCSE, Vellore Institue of Technology, Chennai
Gang Li, School of Information Technology, Deakin University, Australia

II. LITERATURE SURVEY

Recent studies have worked with system that predict using MRI scans (image) [4], image along with biomarkers [5] and [3] and biomarkers [6], [7] and [8]. Biological Markers (biomarkers) is defined as "cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids" by Hulka and colleagues [9]. Using biomarkers is costly unless the laboratory procedure is relatively simple and automated. Selection of correct set of biomarkers is very important to get accurate prediction value. The choice of biomarkers is not trivial because they are derived from body fluids or human tissues. A lot of factors such as changes in storage environment or improper storage of samples affect measurement of biomarkers. [9] Using MRI images is a better option because image does not vary in values and gives a better prediction accuracy [9]. The existing deep learning methods have few limitations because they train deep architectures from scratch. Training from scratch requires huge amount of training data and the data can be expensive. It also requires huge computational resources [10]. Few have used machine learning methods like SVM and feedforward neural networks on MRI images. Using few values like MMSE along with image data helps us improve the accuracy of the prediction model.

III. METHODOLOGY

We have used data mining techniques for prediction of Alzheimer's Disease using OASIS dataset [11].

A. Random Forest

Bagging and Boosting are two well-known ensembles learning methods that generates classifiers and aggregates them. Successive trees are generated in classification using Boosting and then weighted votes are taken at the end for better accuracy of prediction. Whereas in the case of bagging the tress are indecently contracted and at the end just a simple vote is taken for prediction. [12]

Random Forest, in addition to construction of tress with different data it also changes how the tress is constructed. Random Forest performs well because each node is split using the best sub set of predictors randomly chosen at the node and this is also robust with outfitting. [12]

Random Forest works by firstly drawing a nTree using the training dataset. Then for each data from the training dataset it constructs an unpruned tree. Each node chooses a data m among all the predictors and then uses the best split on these data. Then finally by taking the votes (maximum votes for classification and average votes for regression) of the aggregated nTress it predicts the new data.

Random Forest produces two additional pieces of information importance of predictor variables and measure of internal structure of data. Importance of each variable is due to the complex interaction with other variables. The importance is determined by predicting the error rate when the value of the variable is changed while keeping the other variables unchanged. Any two elements of the proximity matrix of random Forest is a part of the tree contracted and they terminate at the same node. Such cases should be repetitive.

B. Gradient boosting

Gradient Boosting (GBM) is based on a different constructive strategy of ensemble formation. In this method, new models are added sequentially. Constructing new base-learners such that they are maximally correlated with the negative Gradient of the loss function which is associated with the ensemble is the principle idea behind this algorithm.[13] Choice of loss function depends on the researcher. GBM is flexible and customisable for any data driven task. In supervised learning the major issue is to provide proper combinations of class values for training dataset. The unknown functional dependence is to be reconstructed with estimates such that loss function is minimized. [13] Gradient Boosting involves loss function, weak learner and an additive model. Weak learners here are usually decision trees. In additive model, trees are added sequentially and a Gradient descent is used to minimize the loss while adding.[13] After calculating loss, a new tree that reduces loss is added. Output of new tree is added to the existing output sequence to manipulate the final output. [13]

C. Lasso

Least Absolute Shrinkage and Selection Operator aka Lasso's main goal is to reduce the prediction error. Lasso performs two main tasks, feature selection and regularization.[14] This method puts a constraint on the sum which is, the sum must be less than the upper bound value. Lasso sets some of the variable values to zero. This technique is called as shrinkage. While performing the feature selection the variables which are still non-zero even after shrinkage is selected [14]. The strength of the penalty is controlled by the tuning parameter(λ). The coefficients are forced to be zero when λ is large, this helps in reducing the dimensionality. The larger the λ more coefficients are forced to be zero.

Lasso has many advantages. It provides a very good accuracy rate for prediction. In case of the tuning parameter λ, when λ increases, bias increases and variance decreases. Overfitting is reduced by eliminating irrelevant variables that are not associated with the response variable. [14]

D. SVM

Support Vector Machine or simply SVM is a powerful machine method developed for statistical learning [15]. SVM is a supervised learning method that analyses data and recognises patterns. The empirical classification error and geometric margin is reduced in SVM. An attribute is a predictor variable and a feature is a transformed attribute which defines a hyperplane [15]. A vector is a set of features that represents one instance. The main goal of SVM is to find an optimal hyperplane which separates cases of clusters of vector with one category of variables on one side and the other category variables of on the other side. These vectors that are closer to the hyperplane are the support vector. SVM is useful classification technique that uses training and test data. Each training data's instance has one target value and several attributes. Finally, SVM produces a model that predicts target values of test data.

IV. RESULTS AND DISCUSSIONS

A. Dataset

The dataset for cross-sectional data was obtained from OASIS [11]. Cross-sectional MRI is the imaging based on axial slices. The cross-sectional data has multiple attributes categorized as follows: Demographic data: Gender (M/F), Handedness (Hand), Age, Education (Educ), socioeconomic status(SES). Education codes correspond to the following levels of education: 1: less than high school grad., 2: high school grad., 3: some college, 4: college grad., 5: beyond college. Clinical data: Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR; 0= nondemented; 0.5= very mild dementia; 1= mild dementia; 2= moderate dementia).Derived anatomic volumes: Estimated total intracranial volume (eTIV), Atlas scaling factor(ASF), Normalized whole brain volume (nWBV).

B. Preprocessing

It was found that there are missing values in SES, MMSE and CDR attributes. Missing values can be handled in different ways: Ignore the data row, use a constant to fill in for missing values, use attribute mean, use attribute mean for sample belonging to the same class and use data mining algorithm to predict the most probable value. Ignore the data row method is usually used when the class value is missing. This is done because such rows reduce the performance of the model. In using a constant to fill in for missing values, the missing value is filled with values like "unknown", "N/A", 0 or minus infinity.

In using attribute mean, mean value for that attribute is used to replace the missing value. In using attribute mean for all samples belonging to the same class, the mean for the attributes of relevant class is calculated and substituted instead of missing value. In using a data mining algorithm to predict the most probable value, value can be determined using regression, decision trees, clustering, etc. We have used constant value 0 to fill the missing values of SES attribute. We have ignored the records where MMSE and CDR values are missing. The values of the gender attribute are converted to 1 and 0, where 0 represents female and 1 represents male.
The cross-sectional dataset consists of 417 records. After ignoring the records with missing MMSE and CDR values we obtained 236 records.

C. Attribute selection

Using the 8 attributes MMSE, eTIV, nWBV, ASF, Age, Education (educ), SES and M/F, we derived 15 combinations keeping attributes MMSE, eTIV, nWBV, ASF common since they are derived from MRI scans for the prediction. We used Random Forest, Gradient Boosting, Lasso regression and SVM classifier to predict the CDR value. The CDR value helps to predict the Alzheimer's Disease in the patients. The dataset was divided into 2 parts, 120 records for training data and 116 records for test data.

```
Procedure Att_Selection
Q = { Set of all attributes }
Q1 = Q
A = { MMSE, eTIV, nWBV, ASF }
Apply prediction techniques using A
Q = Q – A
for j = 1 to Q
  { A= A ∪ Aj
    Apply prediction techniques using A
    While Q != Ø
        for i = j + 1 to Q
            { A = A ∪ Ai
              Apply prediction techniques using A
              A = A – Ai  } }
     A = A – Aj    }
Apply prediction techniques using Q1
```

D. Discussions

Using attribute selection procedure the following combinations are derived:
Combination 0: MMSE, eTIV, nWBV, ASF
Combination 1: MMSE, eTIV, nWBV, ASF, age
Combination 2: MMSE, eTIV, nWBV, ASF, Educ
Combination 3: MMSE, eTIV, nWBV, ASF, SES
Combination 4: MMSE, eTIV, nWBV, ASF, M/F
Combination 5: MMSE, eTIV, nWBV, ASF, age, Educ
Combination 6: MMSE, eTIV, nWBV, ASF, age, SES
Combination 7: MMSE, eTIV, nWBV, ASF, age, M/F
Combination 8: MMSE, eTIV, nWBV, ASF, Educ, SES
Combination 9: MMSE, eTIV, nWBV, ASF, Educ, M/F
Combination 10: MMSE, eTIV, nWBV, ASF, SES, M/F
Combination 11: MMSE, eTIV, nWBV, ASF, age, Educ, SES
Combination 12: MMSE, eTIV, nWBV, ASF, age, Educ, M/F
Combination 13: MMSE, eTIV, nWBV, ASF, Educ, SES, M/F
Combination 14: MMSE, eTIV, nWBV, ASF, age, SES, M/F
Combination 15: MMSE, eTIV, nWBV, ASF, age, SES, M/F, Educ
All the combinations of attributes are considered for prediction. The combination 0: MMSE, eTIV, nWBV, ASF has given better accuracy in all the four prediction techniques Random Forest, Gradient Boosting, Lasso regression and SVM classifier.
To identify the best combination of attributes we used MSE (mean squared error) to identify the error rate and chose the combination which has the least error for prediction.
MSE or Mean Squared error calculates the average of the squares of the deviations(error) which is the difference between the estimated value and the actual value. MSE is a non-negative as it is a measure of quality. If the MSE values are closer to zero the more better they are.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(actual\ value - predicted\ value)^2$$

Figure 1 represents the accuracy rate of the prediction techniques. Random Forest and Gradient Boosting are equally good in accuracy.

Table 1: Accuracy calculation

Technique	Accuracy (%)	MSE
Random Forest	97.94	0.020588235
Gradient Boosting	97.94	0.02058824
Lasso	94.06	0.059404255
SVM	93.6	0.074117647

Random Forest shows significant difference in accuracy for all the combinations of attributes. Gradient Boosting has not shown significant difference. Random Forest has best accuracy of 97.94% for the prediction of Alzheimer's Disease.

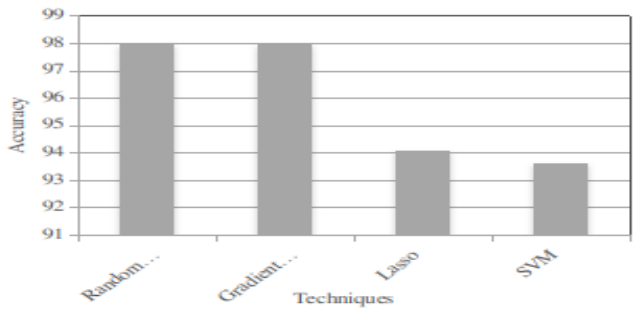

Figure 1: Accuracy of prediction techniques

.

V. CONCLUSION

We have shown that various prediction techniques can be applied to the OASIS dataset for Alzheimer's

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Disease prediction. Missing values have been taken care by preprocessing. Best combinations are identified and the prediction has been performed. Random Foresting achieves more accuracy than other techniques. In future, the data and the images of the OASIS dataset can be combined for Alzheimer's Disease prediction.

REFERENCES

1. Xiaojing Long, Lifang Chen, Chunxiang Jiang, Lijuan Zhang (2017), Alzheimer's Disease Neuroimaging Initiative, Prediction and classification of Alzheimer disease based on quantification of MRI deformation, PLoS ONE 12(3): e0173372. doi:10.1371/journal. pone.0173372
2. Alzheimer's Association. (2015) Alzheimer's Disease facts and _gures. Alzheimer's and Dementia: The Journal of the Alzheimer's Association, 11(3):332
3. A.V. Lebedev, E. Westman, G.J.P. Van Westen, M.G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki B. Vellas, S. Lovestone, A. Simmons (2014), and for the Alzheimer's Disease Neuroimaging Initiative and the AddNeuroMed consortium, Random Forest ensembles for detection and prediction of Alzheimer's Disease with a good between-cohort robustness, ELSEVIER Neuroimage Clin. ; 6: 115–125
4. Marcia Hon, Naimul Mefraz Khan (2017), Towards Alzheimer's Disease Classification through Transfer Learning, arXiv:1711.11117v1 [cs.CV]
5. Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, and the Alzheimer's Disease Neuroimaging Initiative1 (2013), Accurate multimodal probabilistic prediction of conversion to Alzheimer's Disease in patients with mild cognitive impairment, ELSEVIER Neuroimage Clin.; 2: 735–745
6. Hinrichs C, Singh V, Xu G, Johnson SC, ADNI (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population, Neuroimage ;55:574–89.
7. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S, et al (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's Disease in patients with mild cognitive impairment. Neu-roimage Clin. ; 2:735–45.
8. Liu Y, Mattila J, Ruiz MAM, Paajanen T, Koikkalainen J, van Gils M, et al. (2013) Predicting AD conversion: comparison between prodromal AD guidelines and computer assisted predictAD tool. PLoS One; 8:e55246.
9. Gertrude H. Sergievsky Center and the Taub Institute for Research (2004), Biomarkers: Potential Uses and Limitations, NeuroRx, v.1(2)
10. Marcia Hon, Naimul Mefraz Khan (2017), Towards Alzheimer's Disease Classification through Transfer Learning, arXiv:1711.11117v1 [cs.CV]
11. Open Access Series of Imaging Studies (OASIS), online: http://www.oasis- brains.org
12. Andy Liaw and Matthew Wiener (2002), Classification and Regression by Random Forest, Rnews, Vol. 2/3
13. Alexey Natekin and Alois Knoll (2013), Gradient boosting machines, a tutorial, Front Neurorobot. ; 7: 21.
14. Valeria Fonti, Dr. Eduard Belitser (2017), Feature Selection using LASSO, VU Amsterdam
15. Durgesh K. Srivastava, Lekha Bhambhu (2012), Data Classification Using Support Vector Machine, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 National Conference on Emerging Trends in Engineering & Technology (VNCET-30 Mar'12)
16. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
17. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
18. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
19. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
20. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.
21. Rajesh, M., K. Balasubramaniaswamy, and S. Aravindh. "MEBCK from Web using NLP Techniques." Computer Engineering and Intelligent Systems 6.8: 24-26.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication