Indian Journal of Science and Technology Vol 9(13), DOI: 10.17485/ijst/2016/v9i13/84496, April 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Diagnosis of Alzheimer's Disease using Rule based Approach

S. R. Bhagya Shree[1*], H. S. Sheshadri[2] and Muralikrishna[3]

[1]PESCE, Mandya - 571401, Karnataka, India; srbhagyashree@yahoo.co.in
[2]PET Research Center, Mandya - 571401, Karnataka, India; hssheshadri@gmail.com
[3]CSIH Memorial Hospital, Mysore - 570021, Karnataka, India; muralidoc@gmail.com

Abstract

Objectives: In the world of modern medicine, though there is lot of medical achievements, diseases like dementia still continue to pest human race. Early Diagnosishelps the diseased to have quality life. This study focuses on diagnosing the subject using classification technique by employing neuropsychological test. **Methods:** The dataset of 466 subjects are collected by conducting neuropsychological tests. The readings are recorded and compared with the cut off suggested by 10/66 Research group. The data set is preprocessed; rule based classifier is used for classification. The preprocessing techniques namely Imbalance reduction, Randomization are applied. Features are selected using Wrapper method. The data set is classified using various methods. **Results:** The Jrip classifier has given 100% accuracy. **Conclusion:** In West, some of the researchers have used neuropsychological tests with machine learning for diagnosis of Dementia. But in East, the work done is comparatively less. Most of those researchers have used MMSE and other neuropsychological batteries. The importance of this work is that the authors have used CSID battery; as it is designed by Alzheimer`s Disease International, this is tailor made battery for diagnosis. Jrip classification is applied and it fetched 100% accuracy. Future work is to compare the authors` diagnosis with 10/66 diagnosis by considering 10/66 as gold standard.

Keywords: Alzheimer`s Disease, Classification, JRip, Preprocessing, 10/66

1. Introduction

Dementia is an age related disease associated with loss of cognitive functions. Around 60% of Demented are affected from Alzheimer`s Disease (AD)[1].

Figure 1 shows the prevalence of Dementia in different regions. From Figure 1 it is clear that by 2050 the number of demented in high income countries is almost constant where as it is expected to increase exponentially in Lowand Middle Income Countries (LMIC)[2]. The main risk factors of the disease are Age, Genetics, smoking, consuming alcohol and Cholesterol[3]. Diagnosis can be done by Consulting the General physician, undergoing Neuro Psychological tests and taking Magnetic Resonance Imaging or PET scans[4]. Alzheimer Disease International says that, 66% of the demented are AD patients and only 10% of them will be diagnosed on time[5]. From this it is evident that the diagnosis at early stage is helpful to mankind. The paper's organization is as follows. Section 2 explains the literature survey. Section 3 describes the proposed work. In section 4 the authors have discussed about results. In the last section the authors conclude the paper.

2. Literature Survey

There are various neuro psychological tests like Mini Mental Status Examination (MMSE), BDIMC, Alzheimer's Disease Assessment Scale Cognitive subscale (ADASCOG), Blessed Orientation-Memory-Concentration (BOMC), Montreal Cognitive Assessment (MoCA),

General Practitioner Assessment of cognition etc. All these tests have their own advantages and disadvantages. In addition to that, these tests are meant for a particular group of people. The most popular MMSE is associated

*Author for correspondence

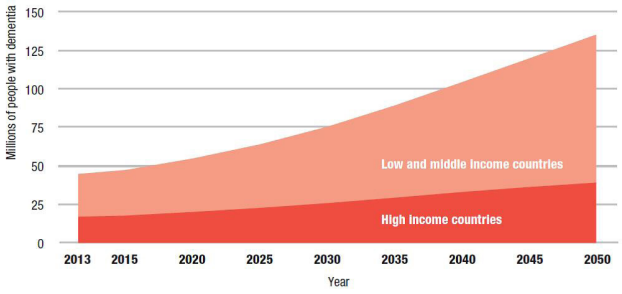

Figure 1. Number of people with dementia in low and middle income countries compared to high income countries.

with a few disadvantages namely insensitiveness to early changes of dementia**,** bias according to age, race, education and socioeconomic status[6].

In their work in[7] have found the relationship among Quality of Life, Depression and Subjective Health Status of the Elderly with Chronic Disease in Korea. The authors have used MMSE for finding cognitive impairment of the subjects. In this work the authors say that health status, depression, education level are the influencing parameters.

This infers the need of a screening test which may be used irrespective of gender, religion, culture and education. 10/66 Research group founded by Alzheimer`s Disease International has designed a battery which overcomes all these disadvantages and is applicable to everyone. After designing the battery, the researchers have studied the subjects of various age groups in different developing countries. These authors have set normative scores for various parameters. These scores vary depending on the country, age, education etc. Comparing the scores of the subjects under consideration with normative values, the neuro psychologist will be able to decide whether the subject is demented or not[8]. In this work the focus is on diagnosis of AD using 10/66 battery. In this battery the neuropsychologists ask a set of predefined questions. The subjects are classified according to the score. To avoid human error and to extract the hidden information knowledge discovery process is used.

Discovering the knowledge is a sequential process comprising of Data Cleaning, integration, selection, Transformation, mining, Pattern evaluation and Knowledge presentation[9]. Various techniques are used for discovering the knowledge, namely Association, Classification, Visualization, Clustering, Collaborative filtering, etc. Of all, mostly used techniques are association, classifiers, visualization and clustering[10]. In the proposed work classification is preferred compared to association because, association involve non numeric attributes. Classification is preferred compared to clustering as clustering is used to group items that will fall naturally together[11]. In this work, authors have used classification.

There are many researchers who have used classification algorithms for the diagnosis of various diseases. In[12] authors have worked on comparative study of classification algorithms on blood transfusion. In this the authors have found out the performance of the various classifiers namely Nave Bayes, J48 and Random forest. In[13] authors have used various data mining techniques in the diagnosis of Diabetes.

The authors in their paper have discussed about applying machine learning techniques for diagnostically differentiating Mild Cognitive Impairment (MCI) and dementia using Clinical Dementia Rating and Clinical Diagnosis[14]. In their work in[15] have focused light upon the correlation between behavior and burden that can be observed in caregivers of elderly demented patients using independent t-test and ANOVA. This work indicates the burden that is added to the caregivers. This infers that early diagnosis, followed by proper medication, postpones the further damage. This in turn reduces the burden on caregivers and the other members of the family.

In their work in[5] have used classification algorithms for diagnosis of AD. In this work the authors had created a data set of 250 and conclude that Naïve Bayes, JRip and Random forest give better results compared to J48. The disadvantage of this work is that the dataset is not preprocessed.

The authors in their work have compared various feature selection and feature extraction methods using wrapper and filter methods. They conclude by saying feature selection using Wrapper over performs filter method in terms of accuracy and efficiency[16]. In the proposed work wrapper method is used for selecting the features.

The various data mining tools like WEKA, See5, Wiz Why are used in many of the research works. WEKA is used by most of the researchers. In[17] have used WEKA in the detection of Breast cancer. The authors in their work have used WEKA for Preprocessing, Classification and Clustering[18]. In their work in[19] have mentioned that WEKA predicts majority of the data.

3. Proposed work

3.1 Collection of Data

Details of subjects are collected. Global Cognitive Function measured by administering the Community Screening Instrument for Dementia (CSI 'D') includes a 32 item

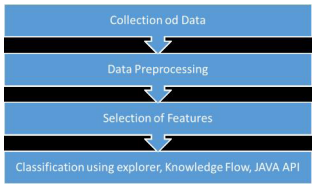

Figure 2. Flow of the work.

cognitive test assessing orientation, comprehension, memory, naming and language expression. All these put together generates a global cognitive score (CSID COGSCORE). The results of the data are tabulated in the form of a spread sheet. The data will be then converted from spread sheet format to WEKA readable ARFF format[9].

3.2 Data Preprocessing

The preprocessing contains a series of processes like Imbalance reduction, Randomization, model evaluation, feature selection etc.

3.2.1 Imbalance Reduction

The data set consists of positive and negative instances. The imbalance in the number of instances may lead to under performance of classification methods and experience over fitting. Synthetic Minority Oversampling Technique (SMOTE) is applied to reduce the imbalance.

3.2.2 Randomization

After the application of SMOTE, the number of negative instance will accumulate at the end of the ARFF file. If 10 fold cross validation is applied, to this data set, the data set will be divided into 10 folds. In that case the last fold will have only negative instances. To overcome this problem, unsupervised filter namely "randomize" is applied. After application of this technique the data set will have same number of records but they will be randomly distributed throughout the ARFF data file.

3.2.3 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. Basically there are two methods,

- Wrapper method: In this method all the possible subsets of the dataset are created. The search techniques like random search, first search, depth search are used to find the evaluators.
- Filter method: This method gives a rank to all the features of the dataset. The features are arranges according to the rank. The feature with highest rank is considered. In the proposed work, authors have used wrapper method for feature selection.

3.2.4 Model Evaluation

Two learning performance evaluators are included with WEKA.

- Training Set: In this case, the classifier simply splits the dataset into training and test data.
- Cross Validation: In case of n fold cross validation, WEKA develops n models, it finds the average performance of those n models and displays the results. The remaining models are deleted.

The next step is classification. Classification is done to know how exactly the data is being classified.

Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a classifier works.

The data set is preprocessed and the results are compared with different evaluators. The classification is applied to the preprocessed data before and after selecting the features and the results are compared.

The preprocessed dataset is evaluated using explorer, Knowledge flow and JAVA API.

The model is evaluated using rule based classifier that is JRip. A rule- based classifier uses IF – THEN rules for classification. An IF – THEN rule is an expression of the form,

IF condition THEN conclusion

A rule R can be assessed by its coverage and accuracy.

Accuracy = …. (1)

Coverage = …. (2)

Where N_{covers} is the number of tuples covered by R

|D| be the number of tuples in D

That is, a rule's coverage is the percentage of tuples that are covered by rule[8].

4. Discussion

Model evaluation using explorer

The CSV file is loaded to the WEKA which is as shown in Figure 3. The data set contains 466 instances and 51 attributes. And there are 448 negative instances and 18 positive instances.

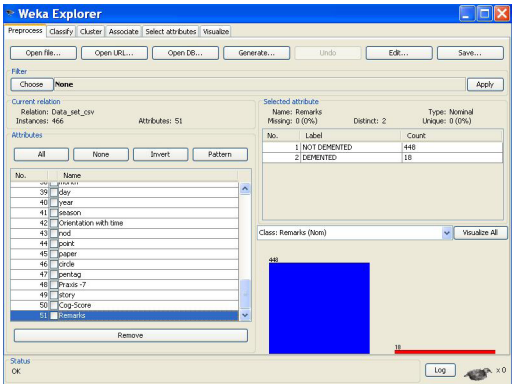

Figure 3. CSV file loaded to explorer.

Table 1. Summary of results: before and after application of SMOTE filter

Parameter	Accuracy	Precision	Recall	ROC Area
With SMOTE	100%	1	1	1
Without SMOTE	100%	1	1	1

The performance parameters with and without SMOTE is shown in Table 1.

After applying SMOTE the numbers of positive instances are increased from 18 to 36, which are accumulated at the end of ARFF file. Application of n fold cross validation will divide the data set into n folds. Under such circumstances the last fold will have only positive instances. Randomization is done to distribute the dataset throughout the ARFF file.

In this paper the authors are working on primary data. The inclusion of imbalance reduction will not have any positive impact on the results, instead the authors find it difficult to judge the validity of the results in comparison with the neuropsychologist`s decision. Hence the data set without SMOTE is considered for further processing.

Both 10 fold Cross validation and training is applied on the data set. Table 2 shows the summary of various parameters.

Feature selection improves the accuracy, efficiency and performance of the classification.

The best first technique of wrapper method used in this work, searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility, setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Wrapper feature selection method selects the Cog score as the term.

The features which are responsible for results are kept and all other features are removed.

Classification is applied on the data set. The below confusion matrix is the result of classification applied on the data set with selected attribute.

Confusion matrix

a	b
448	0
0	18

a: Not demented b: demented

Ensemble is the method to increase the accuracy of a classifier. The method includes Bagging and Boosting. As the accuracy of the classifier is 100% ensemble is not required. The model is evaluated. The results are compared.

Model evaluation using Knowledge flow

Figure 4 shows the classification implemented using knowledge flow. ARFF file is loaded. Model is evaluated using

Table 2. Summary of training set and cross validation results

Method	Accuracy	Precision	Recall	F - Measure	ROC Area
Training Set	100	1	1	1	1
Cross Validation	100	1	1	1	1

Table 3. Results of model evaluation

Method	Accuracy	Precision	Recall	F Measure
Explorer	100%	1	1	1
Knowledge Flow	100%	1	1	1
Java API	100%	1	1	1

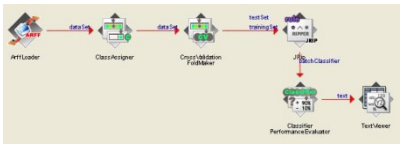

Figure 4. Knowledge flow diagram – classification.

cross validation. Data set is classified. Using Text Viewer classification performance statistics can be visaulized.

5. Conclusion

In this paper, the data set of 466 subjects is collected from **MY**sore studies of **N**atal Effects on **A**geing and **H**ealth cohort (MYNAH). The study was approved by the Ethics and Research Committee at Holdsworth Memorial Hospital.

The subjects are classified based on the neuropsychological battery designed by 10/66 research group. Machine learning approach is used, as researchers have applied machine learning for diagnosis of various diseases. In this work, authors have used rule based JRip approach for classifying the subjects. The dataset is Pre-Processed. The Authors conclude that SMOTE is not requiredas the data collected is primary and inclusion of SMOTE will affect the quality of data. In addition to that SMOTE has no impact on the results. So the CSV file without SMOTE is considered for further processing of data. The results of training and cross validation are one and the same. Hence the training dataset is considered. The features are selected using Wrapper method as Wrapper is proved better compared to filter method. The result of feature selection is Cog score. To find Cog Score all the other attributes need to be considered. Hence feature selection will not have any impact on the process.

The model is evaluated. The results are tabulated Table 3. The novelty of the work is the data set is of primary type. The datasets were obtained from a community dwelling older adults. The data set is realistic in nature. The future work is to compare the researcher`s result with 10/66 and taking suitable measures to reduce the disagreement if there is any.

6. Acknowledgement

We are thankful to the staff members of CSIH Memorial hospital, Mysore for their support. We are grateful to Participants and their family members.

7. References

1. Salmon DP, Bondi MW. Neuropsychological assessment of dementia. Access NIH Public, PubMed Central, US National Library of Medicine National Institutes of Health; 2010 May.
2. Alzheimer`s Disease international. The Global Impact of Dementia 2013–2050; 2013.
3. Thies W, Bleiler. Alzheimer's disease facts and figures. Alzheimer's and Dementia. 2013 Mar; 9(2):208–45.
4. Saling M, Brodaty H, Yates M, Scherer S, Anstey K. Early diagnosis of dementia; 2007.
5. Shree SRB, Sheshadri HS. An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques. Proceedings of IEEE Conference. ICCIC; 2014.
6. Galvin JE, Sadowsky CH. Practical guidelines for the recognition and diagnosis of Dementia. JABFM. 2012 Jun; 25(3):367–82.
7. Ju S, Kim K-S. The relationship among quality of life, depression and subjective health status of the elderly with chronic disease in Korea. Indian Journal of Science and Technology. 2015 Jul; 8(16). DOI: 10.17485/ijst/2015/v8i16/75174.
8. Sosa1 AL, et al. Population normative data for the 10/66 Dementia Research Group cognitive test battery from Latin America, India and China: Across-sectional survey. Access NIH Public, PubMed Central, BMC Neurology. 2009 Aug; 9(48):1–11.
9. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Elsevier; 2012.
10. Soman KP, et al. Insight into data mining theory and concepts. 6th ed. PHI Learning Private Limited; 2012.
11. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques. 2nd ed. Elsevier; 2008.
12. Rani SA, Ganesh H. A comparative study of classification algorithm on blood transfusion. International Journal of Advancements in Research and Technology. 2014 Jun; 3(6):57–60.
13. Rahman RM, Afroz F. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. Journal of Software Engineering and Applications. 2013; 6(3):85–97.
14. Williams JA, Weakley A, Cook DJ, Edgecombe MS. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. Proceedings of the 27th AAAI Conference on Artificial Intelligence; 2013. p. 71–6.
15. Cho YH, Ra JS. Correlation between preventive health behavior and family burden in family caregivers for the elderly with dementia. Indian Journal of Science and Technology. 2015 Oct; 8(26). DOI: 10.17485/ijst/2015/v8i26/81889.

16. Li Y, Liu Y. A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '08; 2008 Sep 15-17. p. 195–200.
17. Fallahi A, Jafari S. An expert system for detection of breast cancer using data preprocessing and bayesian network. International Journal of Advanced Science and Technology. 2011 Sep; 34:65–70.
18. Singhal S, Jena M. A study on WEKA tool for data preprocessing, classification and clustering. IJITEE. 2013 May; 2(6):250–3.
19. Andreeva P, et al. Data mining learning models and algorithms for medical applications. Proceedings of the 18th Conference Systems for Automation of Engineering and Research (SEAR 2004); Varna, BG. 2004. p. 148–52.