An approach to preprocess data in the diagnosis of Alzheimer`s Disease

Bhagya Shree S R
Research Scholar
PET Research Center
Mandya, India
srbhagyashree@yahoo.co.in

Dr.H.S.Sheshadri
Prof. Department of E & C
PES College of Engineering
Mandya, India
hssheshadri@gmail.com

***Abstract*- The number of people surviving in older age is more. This is mainly due to the developments that have taken place in the field of medicine. Theseold people are prone to many age related diseases. There are numerous neuro degenerative brain related diseases. Dementia is one among them. The people affected by Dementia will have lapse of memory. Alzheimer's disease is one of the types of dementia. Diagnosis of the disease is a time consuming task. To reduce the time needed for diagnosis the medical practitioners use system based approach. To help the practitioners researchers have developed various tools and techniques.**

In this paper the authors focus on classifications of subjects as diseased or not. Before doing classification the data has to be preprocessed. Preprocessing of data is done by applying techniques such as preparation of data, selection of attributes, balancing data, model evaluation and feature selection etc.The authors have collected the data of 466 subjects. The preprocessing techniques are applied on the data set. The subjects are classified using Naïve bayes and J48. The accuracy of the classifications are compared and Naïve bayes is found better.

Keywords-Neuro psychological tests, SMOTE, Wrapping method,Naïve bayes, J48.

I. INTRODUCTION

All over the world there are 44million people suffering from dementia [1].Dementia means loss of memory. This Disease is classified into various types and some of them are Alzheimer's disease, Parkinson's disease, Front temporal lobar degeneration, vascular dementia etc., [2]. Most of the demented patients are grouped under Alzheimer's disease. There are around 38million people suffering from Alzheimer's disease.According to the special report done by Alzheimer`s association done in2013, in 2013, an estimated 5.2 million Americans of all ageshave Alzheimer's disease. This includes anestimated 5 million people of age 65 and olderand approximately 200,000 individuals under age65 who have younger-onset Alzheimer's. One in nine people age 65 and older (11 percent)has Alzheimer's disease. About one-third of people age 85 and older(32 percent) have Alzheimer's disease. Of those with Alzheimer's disease, an estimated4 percent are under age 65, 13 percent are 65 to 74,44 percent are 75 to 84, and 38 percent are 85or older [3]. These facts indicate the need of early diagnosis. There are various risk factors that contribute to the development of the disease. They are Age, Genetics, Smoking, alcohol Intake, Cholesterol, Down Syndrome etc. [4]. The symptoms of the Alzheimer`s diseases are poor decision making, poor judgment, misplacing things, impairment of movements, problem with verbal communication, abnormal moods, complete loss of memory. The diagnosis of AD is done at three different stages namely consulting the General Physician, Undergoing neuro psychological tests and taking MRI scans.Alzheimer's disease and otherdementias are caused by damage to neurons that cannotbe reversed with current treatments [4]. Diagnosis of the disease at the early stage will help the patients to have quality life for the rest of their life. The authors have focused on diagnosis of the disease for neuro psychological test. For diagnosis of the disease machine learning approach is used. In this paper authors focus on various preprocessing techniques that have to be applied to the dataset.The classification techniques Naïve bayes and J48 are applied on the dataset and the results are compared.

II. LITERATURE SURVEY

There are various neuro psychological tests like MMSE, BDIMC, COG, BOMC, MOCA, AD8 and GP CoG etc. Each of these tests has its own advantage and disadvantage and moreover, the tests are meant for a community of people. Of the all MMSE is very popular. But even that has a disadvantage. The disadvantage of MMSE is it is insensitive to early changes of dementia. This indicates the need of a screening test which may be used to the subjects irrespective of gender, religion, culture and education. To overcome this problem 10/66 research group founded by Alzheimer` Association has studied the subjects of various age groups in different developing countries. Theses researchers have designed a battery and they have set normative scores. The paper focuses on diagnosis of AD using 10/66 battery by knowledge discovery from data [5]. This 10/66 battery is preferred compared to the most popular MMSE battery as it is applicable to anyone irrespective of gender, religion, culture and education [6].

The knowledge Discovery process is a procedure that comprises of Data Cleaning, Data integration, Data selection, Data Transformation, Data mining, Pattern evaluation, Knowledge presentations [7].Data mining finds its application in the field of biomedical engineering. Researchers have used data mining for the diagnosis of various diseases.

Abhishek Taneja in his paper has discussed about using data mining for the prediction of heart disease

[8].Tarigoppula V.S sriram et.al has used classification algorithms to detect Parkinson's disease [9].

Breetha S and Kavinila R have discussed about using hierarchical clustering in the diagnosis of cancer and classification of cancer [10].Rashedur M. Rahman and FarhanaAfroz have tested the various classification techniques using various tools likeWEKA, Mat lab, Tanagra for the data sets of diabetes patients [11].

Various techniques are used for discovering the knowledge namely, Association, Sequential pattern, Classification, Decision trees, Neural networks, Visualization, Clustering, Collaborative filtering, Data transformation and cleaning, Deviation and fraud detection, Estimation and forecasting, Bayesian and dependency networks, OLAP anddimensional analysis, Statistical analysis, Text analysis, Web mining etc.

Jyothi Sony has used supervised machine learningnamely Naïve Bayes, K-NN, Decision List algorithm to analyze the datasets of heart disease patients [12].

Tina R. Patil and Mrs. S. S. Sherekar in their paper have done the performance analysis of Naive Bayes and J48 Classification Algorithm for Data Classification [13].

Jehad Ali et.al in their paper compared the classification results of Random Forest and the J48 for classifying twenty versatile datasets. They concluded that random forest gave better results for same number of attributes with large datasets while J48 is handy and it suits only for those with small datasets [14].

Plamena Andreeva and group have tested theparameters of data sets of three different diseases namelybreast cancer, Diabetes Pima and IRIS and published ascholarly article in Google scholar. They have analyzed thedata using various types of classification by using differenttools namely See5, Wiz Why and WEKA. From the results theauthors suggest that WEKA is better in terms of usage, consistency etc. The authors also say that, of the all, WEKApredicts the majority of the data [15].

In this paper the authors focus on application of various pre-processing techniques on the data set. The authors have applied classification techniques and compared the result. WEKA tool is used for implementation.

III. Problem Definition

Data set consist of 466records of subjects aging from 50 to 80 years. Real world data can be incomplete, noisy or it may be lacking in terms of attributes of interest.

The main objectives are:

- To apply the various preprocessing techniques.
- Choosing the appropriate technique.
- Applying classification techniques naïve bayes and J48
- Comparing the results of naïve bayes and J48

IV. Architecture

Data preprocessing is a very important step in knowledge discovery process, as decisions are based on the quality of data. Detection of data anomalies, rectifying the errors and reducing the data to be analyzed will lead to huge pay offs for decision making [16].

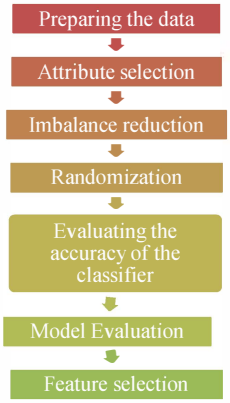

Fig. 1. Flow diagram of various preprocessing techniques.

a) Preparing the data:

The data is often present in the form of spread sheet.HoweverWEKA native data storage format isARFF. The data will be converted from spread sheet to CSV format. Having done this the CSV file is converted to ARFF file. Thus the data has to be converted from spread sheet format to ARFF format[17].

b) Attribute selection:

All the attributes that are present in the file may not be useful. Hence the attributes which are not required are removed using "remove" command [18].

c) Imbalance reduction

The data set consists of positive and negative instances. One of these may be less in number compared to other.This imbalance may lead to under performance of classification methods and experience over fitting.This problem can be overcome by resampling the dataset by applying the Synthetic Minority Oversampling Technique (SMOTE).

c) 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.

d) Evaluating the accuracy of the classifier

To obtain a reliable estimate of classifieraccuracy, hold out, random sub sampling, cross validation and boot strap are commonly used techniques. In hold out method the given data are randomly partitioned into two independent sets,Test set and training set. Typically training set will have more instances than test set.

e) Model Evaluation

The data set can be evaluated from,

1. Training set: In this case, the result of each model can be saved and can be visualized.
2. Cross validation: In case of 10 fold cross validation, WEKA develops 10 models, when it displays the result, it uses the average performance of those 10 models. It deletes theremaining models.

From the observations the authors conclude that the model saved with cross validation and with the training set are same.

f) Feature selection

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

i. Wrapper method: wrapper method will create all possible subsets from the data set. Then the classification algorithm is used to induce classifiers from the feature in each subset. To find a subset, evaluator will use one of search techniques such as random search, first search, depth search etc.,

ii. Filter method: Filter method uses an evaluator and a ranker to rank all features in a particular dataset. It arranges the attributes according to the rank. By omitting the feature with lowest ranking one at a time, the dominant features can be identified. In this paper authors use wrapper method to select the features of interest.

V. RESULTS AND DISCUSSIONS

The figures below show the results after applying various preprocessing techniques.

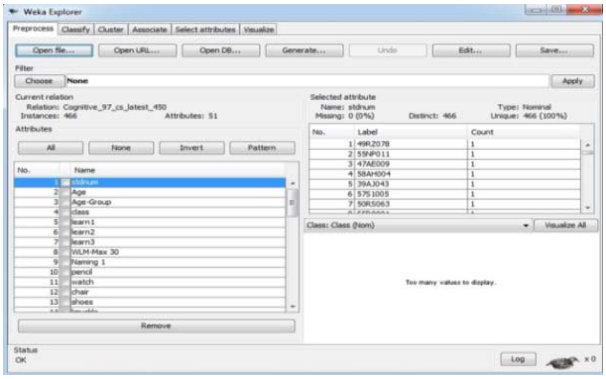

Fig. 2. The CSV file loaded to WEKA

The CSV file is loaded into WEKA. As it can be seen there are 466 instances and 51 attributes.

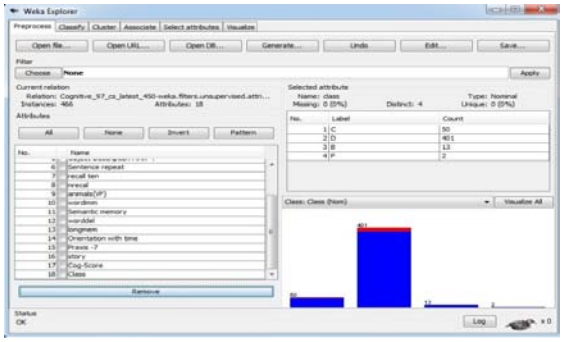

Fig. 3. Attribute selection

The unwanted attributes are selected and removed and the numbers of attributes are reduced to 18.

To reduce the imbalance SMOTE filter is applied. Fig 4 depicts the result after the application of SMOTE.

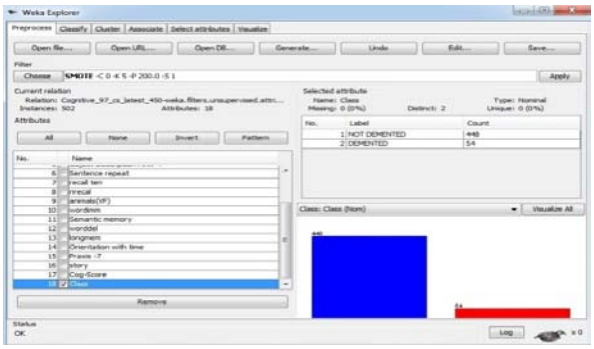

Fig. 4. Imbalance reduction

After applying SMOTE the numbers of positive instances areincreased from 18 to 54, which are accumulated at the end of ARFF file.Fig 5 shows the ARFF file having large number of positive instances accumulated at the end of the file.

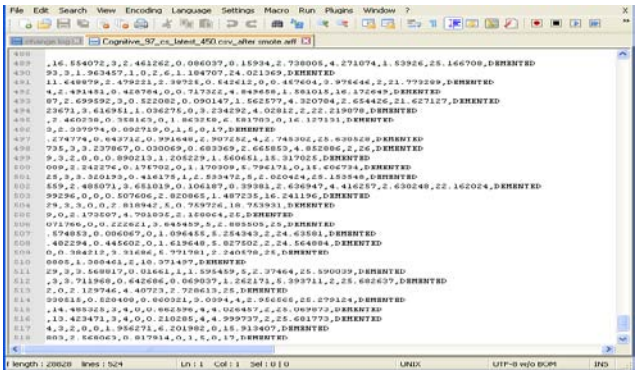

Fig. 5. ARFF file having large number of positive instances.

Fig 6 is the visualization of ARFF data file after randomization.

Classification	*Classification Accuracy*	*Precession*	*Recall*	*Time taken to build the model*
Naïve Bayes	100%	1	1	0.02s
J48	96.5665%	0.998	0.967	0.05s

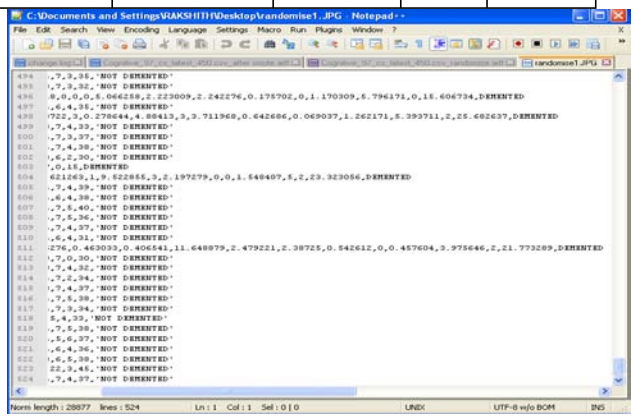

Fig. 6. Randomization result

Wrapper feature selection method, in this method a filter called “best find” is used to select the attribute of interest.

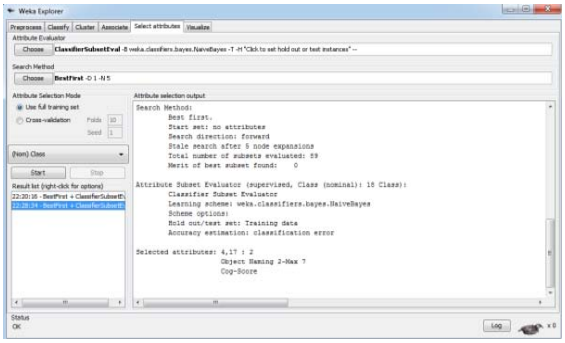

Fig. 7. Selection of attributes which are of less interest.

Fig 7 shows the classifier output which is showing the feature of interest. After selecting the feature, the attributes or features which are of less interest are removed.

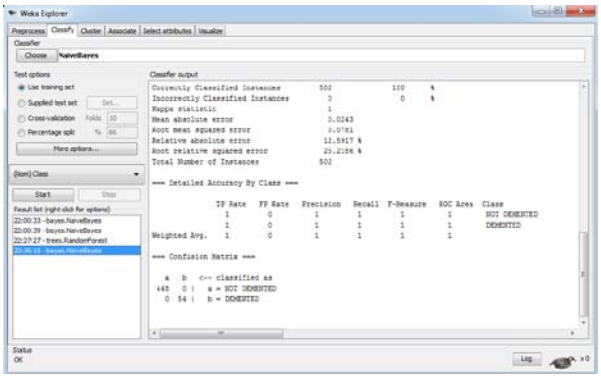

Fig 8: Accuracy is monitored after removing the attributes which are of less interest.

Fig 8 depicts the classifier output with greater accuracy after removing the feature of less interest.

The classification techniques, Naïve bayes and J48 are applied on the data set. The parameters like Classification Accuracy, Precession, recall and time taken to build the model are considered. The summary of the data sets are shown in Table 1.

Table 1: Summary of data sets

The confusion matrices of Naïve bayes and J48 are shown below

=== Confusion Matrix J48===
a b <-- classified as

a	b
433	15
1	17

a = NOT DEMENTED
b = DEMENTED

=== Confusion Matrix Naïve bayes===
a b <-- classified as

a	b
488	0
0	18

a = NOT DEMENTED
b = DEMENTED

VI. Conclusion and Feature Work

As the real word data tends to be incomplete, noisy and inconsistent, the data has to be preprocessed. Various preprocessing methods have been discussed. In preparation of data the missing values can be filled and noisy data can be smoothened. As the dataset is of primarytype there is no

missing of data. The number of attributes is reduced from 51 to 18. The number of instances isbalanced using SMOTE filter.The data which is stored in ARFF format is randomized by using randomize filter. In order to obtain a reliable estimate of classifier accuracy, hold out technique is used. Wrapper feature selection technique is used to select the appropriate feature. By removing the features of less interest accuracy can be improved. The classification techniques are applied and the results are compared. Naïve Bayes performed better than J48.

Future work includes designing an embedded system to facilitate the diagnosis.

ACKNOWLEDGMENT

The authors are thankful to Dr. Murali Krishna, Earlier Scientist Research Fellow, Wellcome DBT Allianz, CSI Holdsworth Memorial Mission Hospital, Mysore, Dr. L Basavaraj, Principal, ATME, Mysore and to the research colleagues who supported with the data in respect of the Alzheimer's disease.

REFERENCES

[1] http://www.capitalfm.co.ke/lifestyle/2013/12/06/44-million- now-suffer-from-dementia-worldwide/

[2] [Viswanathan A, Rocca WA, Tzourio C. Vascular risk factors and dementia: How to move forward? Neurology 72:Pp368–74,2009;.

[3] Thies w, bleiler l, 2013Alzheimer`s facts and figures,"Alzheimer`s dement (journal Of Alzheimer's association), Elsevier Inc. Mar-2013.

[4] Michael saling, Henry Brodaty, Dr. Mark Yates, Dr. Sam Scherer, Professor Kaarin Anstey, "Early Diagnosis of Dementia",2007.

[5] Bhagya shree S. R, Dr. H. S. Sheshadri "An Approach in the Diagnosis of Alzheimer Disease - A Survey"International Journal of Engineering Trends and Technology (IJETT) – Volume 7 Number 1- Jan 2014 ISSN: 2231

[6] Ana Luisa Sosa1 et.al ,\Population normative data for the 10/66 Dementia Research Group cognitive test battery from LatinAmerica, India and China: across-sectional survey," Access NIH public, PubMed central, BMC Neurology,Vol.9,pp 1-11, Aug2009.

[7] Jiawei Han, Micheline Kamber, JianPei,\Data Mining: Concepts and Techniques ," Elsevier , Third edition, 2012.

[8] Abhishek Taneja ,\Heart Disease Prediction System Using Data Mining Techniques," Oriental Journal of Computer Science & technology, Vol. 6,Issue 4, pp 457-466, December 2013 .

[9] Tarigoppula V.S Sriram et.al "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 1,PP 44-52, September 2010.

[10] Breetha S, Kavinila " Hierarchical clustering for cancer discovery using Range check and delta check" International Journal of Scientific and Research Publications, Volume 3, Issue 4, April 2013.

[11] Rashedur M. Rahman et.al "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis" Journal of Software Engineering and Applications, 6, PP 85-97,2013.

[12] Jyoti Soni, UjmaAnsari,Dipesh Sharma and Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[13] Tina R. Patil, Mrs. S. S. Sherekar "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013.

[14] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood "Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[15] Plamena Andreeva1, Maya Dimitrova1, Petia Radeva2 "Data mining learning models and algorithms for medical applications" http://scholar.google.co.in/scholar /data mining.

[16] Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber, JianPei published by Elsevier, Third edition, 2012

[17] Data mining: Practical machine learning tools and techniques by Ian H Witten and Eibe Frank published by Elsevier, second addition 2008.

[18] Insight into data mining theory and practice by K P Soman, ShyamDiwakar and V.Ajay published by PHI learning private limited 2012.