International Journal of Machine Intelligence, ISSN: 0975–2927, Volume 1, Issue 2, 2009, pp- 55-61

Statistical classification of magnetic resonance images of brain employing random forest classifier

Joshi S.[1*], Deepa Shenoy P.[2], Venugopal K. R.[2], Patnaik L.M.[3]
[1*]Department of Computer Science and Engineering, MGR University, Chennai, sanjoshi17@yahoo.com
[2]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, K R Circle, Bangalore – 01
[3]Vice Chancellor, Defence Institute of Advanced Technology, Pune, India

Abstract- Data mining in brain imaging is an emerging field of high importance for providing prognosis, treatment, and a deeper understanding of how the brain functions. Dementia due to Alzheimer's disease constitutes the fourth most common disorder among the elderly. Early detection of dementia and correct staging of the severity of dementia is critical to select the optional treatment. The present study was designed to classify and categorize brain images of dementia patients into three distinct classes i.e., Normal, Moderately diseased, and Severe. Decision Forest Classifier was employed to classify the various Magnetic Resonance Images (MRIs) of dementia patients. Results of screening the MRIs are organized by classification and finally grouped into the three categories, i.e., Normal, Moderate and Severe. Experimental results obtained indicated that the proposed method performs relatively well with the classification accuracy reaching nearly 99.32% in comparison with the already existing algorithms.
Key words- Data mining, Machine learning, Dementia, Alzheimer's disease, Random forest classifier

Introduction

Neurodegenerative diseases are now generally considered as a group of disorders that seriously and progressively impair the functions of the nervous system through selective neuronal vulnerability of specific brain regions. Alzheimer's disease is the most common neurodegenerative disease [1], which affects the brain and hence memory. It is a chronic, progressive organic brain disorder characterized by disturbance of multiple cortical functions, including memory, judgment, orientation, comprehension, learning capacity and language [2]. Clinically, the disorder is characterized by a gradual but progressive decline in memory and other cognitive domains and the frequent occurrence of non-cognitive behavioral symptoms. Neuropathologically, the cardinal features of Alzheimer's disease include neuritic plaques, neuro-fibrillary tangles, and the loss of synapses and neurons [3]. Alzheimer's disease has been identified as a protein misfolding disease due to the accumulation of abnormally folded amyloid beta protein in the brains of Alzheimer's disease patients [4]. Amyloid beta (Aβ) is a short peptide that is an abnormal proteolytic byproduct of the transmembrane protein amyloid precursor protein (APP), whose function is unclear but thought to be involved in neuronal development [5]. Neuropathogenesis is proposed to be a result of the accumulation of amyloid beta peptides in the brain together with oxidative stress mechanisms and neuro-inflammation [6]. Alzheimer's disease begins as a deficiency in the production of the neurotransmitter acetylcholine. Patients with Alzheimer's disease show loss of cognitive, intellectual, functional and social abilities, and therefore become fully dependent on their caregiver. It is estimated that in 2010 over five million people will be diagnosed with probable Alzheimer's disease in the United States alone [7-8]. Increasing age is the greatest risk factor for Alzheimer's disease; one-tenth of elderly over 65 years of age develop Alzheimer's disease, whereas nearly half of those over age 85 are diagnosed with probable Alzheimer's disease. Certain people in the population are at greater risk of developing Alzheimer's disease due to various genetic risk factors associated with Alzheimer's disease such as Apolipoprotein (APOE) polymorphism [9]. A person with Alzheimer's disease is expected to live an average of 8 years and up to 20 years after the onset of symptoms. An association between cholesterol and the development of Alzheimer's disease was suggested in the early 1990s and ever since, an increasing amount of research has confirmed that there is a link between cholesterol and the development of AD. A high cholesterol levels in mid-life is a risk for Alzheimer's disease [10]. The National Institute of Health predicts, if the current trend continues, there will be more than 8.5 million Alzheimer's disease patients by the year 2030 in USA alone [11].In this paper we classify and categorize brain images into three distinct classes i.e., Normal, Moderately diseased and Severe. A Normal image is one in which the formation of plaques or neuro-fibrillary tangles are completely absent. A Moderate is a stage where we can find the symptoms of dementia, In Severe we can observe the accumulation of abnormally folded amyloid beta protein in the brains. The concept of decision tree is used for classification which consists of two stages; in the first stage the data extracted from the database is trained to correctly classify the images, by constructing a decision tree. In the second stage for every new tuple that is appended to the database, the decision tree is applied to classify it, termed as categorization. It is then tested on

datasets by the method of cross-validation techniques which is cost effective. A detailed study on Diagnosis of Dementia has been proposed by many researchers. This section presents a brief survey of related work. WR Shankle et. al [12] have applied Knowledge Discovery and Data mining methods in conjunction with Electronic Medical Records of normal aging and demented subjects to automate the screening and diagnosis of Alzheimer's Disease and Vascular Disease. Classification and mining of brain image data using Adaptive Recursive Partitioning method and use of statistical methods for the diagnosis of Alzheimer's disease presented in [13]. Detecting discriminative Regions of Interest (ROIs) and mining associations between their spatial distribution and other clinical assessment is proposed in [14]. In this they used Naïve static partitioning approach for the diagnosis of Alzheimer's disease. Vasileios Megalooikonomou et al [15] proposed a framework for detecting, characterizing and classifying spatial Region of Interest (ROIs) in medical images such as fMRI for classifying Alzheimer's disease by applying characterization technique. Statistical as well as non-statistical methods for classifying three dimensional probability distributions of regions of interest (ROIs) in brain images for Alzheimer's disease is presented in [16]. Enhancement in the life-span of human beings in developed and developing countries has resulted in proportionate increase in the number of patients suffering from senile dementia. Alzheimer's disease is said to be the leading cause of dementia in elderly individuals. Alzheimer's disease individuals exhibit deterioration in mental functions rendering them incapacitated to perform normal daily activities. However, evidence shows that Alzheimer's disease can also afflict young individuals as early as 40 years of age. It is estimated that in 2010 over five million people will be diagnosed with probable Alzheimer's disease in the United States alone. Increasing age is the greatest risk factor for Alzheimer's disease; one-tenth of elderly over 65 years of age develop Alzheimer's disease, whereas nearly half of those over age 85 are diagnosed with probable Alzheimer's disease. The National Institute of Health predicts, if the current trend continues, there will be more than 8.5 million Alzheimer's disease patients by the year 2030 in USA alone. Hence there is an urgent need to understand the disease, to develop prophylactic strategies and minimize the complications associated with this dreaded disease by effective and timely management. It is very essential to diagnose and classify the disease in the beginning stage.

Model

The Architecture and modeling of the current paper is depicted in Fig. (1). This model begins with the collection of brain images from various sources. Since the images found in various formats preprocessing becomes a necessity. In the feature selection process various statistical features are extracted by applying statistical functions on theses images which is then organized to form transactional data base. Classification is performed on these data, and then the new images are categorized accordingly.

Fig. 1- Architecture of Classification of Brain Images

1) Data Collection*:* Images for our experimental study were collected from various sources such as Alzheimer's disease Research Center (ADRC), National Institute on Aging, USA, and National Institute of health, and from various neuroimaging centers across the country. These images were used in our experiments for pre-processing and feature extraction for automatic classification of images.

2) Data Preprocessing: Data from real-world sources are often erroneous, incomplete, and inconsistent. Most of the collected brain images are noisy and inconsistent. Hence, it is very difficult to interpret these images in their original form. Hence a pre processing technique which involves data cleaning and data transformation is applied which assists in removing outliers, noise and inconsistencies. In our study most of the collected images contains some labels and noise that need to be eliminated, this is done by cropping the image. Cropping basically removes the unwanted part of the image i.e., the extra peripheral region which is not of interest. This Cropping operation was done automatically by sweeping through the image, and finding those areas in the image that had a mean intensity less than a certain threshold, these parts of an image were cut horizontally and vertically. Most of the collected images contain some noise which is to be removed. For noise smoothing the filtering techniques were used. The three different classifications of the filters such as such as spatial low-pass, high-pass and band-pass filters. In our study we used low pass filters which is used for noise smoothening and interpolation. Fig. (2) shows the original image and the corresponding filtered image is shown in Fig. (3).

3) Feature Selection: In the current system the brain image features are extracted using statistical approach. These features are collected and then organized to form a transactional database, and each attributes in the database represents a particular feature of a brain. These features are represented in the form

{Image _id, F0, F1, F2……Fn, class label},

Where F1……Fn represents the various features of the image. Some of the important features are (i) Mean, (ii) Variance,
(iii)Skewness, (iv)Kurtosis , (v)Standard Deviation (vi)Discrete Fourier transformation (vii)Discrete cosine transformation
In general, the nth moment about the mean is given by
$\mu_{n\,(r)} = \Sigma\,(r_i\text{-}m)^n * p\,(r_i)$ where i = 0 to L -1
Where i= 0 to L-1; r_i is a random variable indicating intensity,
P (ri) is the Normalized histogram component corresponding to the ith value of r, L is the intensity levels, and m represents the mean. The mean represents the average intensity which is given by,
$\text{Mean} = \Sigma\, r_i * p\,(r_i)$ (i) where i = 0 to L -1
The variance is the second moment defined as
$\text{Var} = \Sigma\,(r_i\text{-}m)^2 * p\,(r_i)$ (ii) where i = 0 to L -1
Similarly Skewness, which is the third moment, can be defined as
$\text{Sk} = \Sigma\,(r_i\text{-}m)^3 * p\,(r_i)$ (iii) where i = 0 to L -1
And Kurtosis is defined as:
$\text{Kurt} = \Sigma\,(r_i\text{-}m)^4 * p\,(r_i)$ (iv) where i = 0 to L -1

4) Classification: Classification is the most commonly used data mining technique, which involves in separating the data into segments which are non-overlapping. Classification can be viewed as forecasting a discrete value. Any approach to classification assumes some knowledge about the data [17] .Hence a training set is used to identify specific parameters. Training data requires sample input data, domain expertise, and a classification assignment to the data. The confusion matrix lists the correct classification against the predicted classification for each class. The number of correct predictions for each class falls along the diagonal of the matrix. All other numbers are the number of errors for a particular type of misclassification error. The outcome of classification can be described as

- True positive (TP): A tuple t_i predicted to be in class C_j, and is actually in it.
- False positive (FP): A tuple t_i predicted to be in class C_j, but is actually not in it.
- True negative (TN): A tuple t_i not predicted to be in class C_j, and is actually not in it.
- False negative (FN): A tuple t_i not predicted to be in class C_j, but is actually in it.
- The Selectivity and Sensitivity are used to determine the accuracy of the classifier.

$$\text{Selectivity} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

A confusion matrix is used to indicate an accuracy while classifying a class with m classes which is an m*m matrix. A general confusion matrix for two classes is shown in the Table 1.

5) Categorization: The knowledge extracted from decision tree is represented as classification using IF-THEN rules. One rule is being uniquely created for every path from root to leaf node. Each attribute value pair along a path forms the IF path i.e., the rule antecedent. Prediction is held in the leaf node which holds the rule consequent (THEN part). This concept is very easy for understanding especially when tree is large.

Problem definition
Given a sample database consists of 153 instances. Electronic medical records such as brain images were collected from various sources such as Alzheimer's disease Research center (ADRC), National Institute on Aging and National Institute of health, and from neuroimaging centers. The objective is to develop an efficient method to correctly classify the images into distinct classes such as Normal, Moderate and Severe, secondly, to increase the accuracy of classification when compared it with the existing results. Random Forest [18], a meta learner is made up of many individual trees, and is designed to work very fast especially when large data sets are used. Every tree in the forest is unique and diverse since it is built using random samples. Since every tree in the forest is trained independently from all other trees it is a strong candidate for parallelization. Random forest classification method is unexcelled in accuracy among current algorithms, it can handle thousands of input variables without variable deletion and estimates of what variables are important in the classification. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.The most important features which we used in Random Forest Classifier are:
1.Algorithm can handle missing value replacement for the training set.
The most important feature is replacing missing values. Suppose ,if the mth variable is not categorical, the method computes the median of all values of this variable in class j, then it uses this value to replace all missing values of the mth variable in class j. If the mth variable is categorical, the replacement is the most frequent non-missing value in class j. These replacement values are called fills.

2. Gini importance.
Every time a split of a node is made on variable m the gini impurity criterion for the two descendent nodes is less than the parent node. Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

3. The out-of-bag (oob) error estimate.
In this random forests algorithm, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run. In the present classifier algorithm, the concept of ten fold cross validation is used. For cases when the amount of data for training and testing is limited, cross validation technique is used. The data is divided in to 10 parts in which the class is approximately the same proportions as in the full data sets. Each part is held out in turn and the learning scheme on the remaining nine-tenth; then its error rate is calculated on the hold out. Thus the learning procedure is executed a total of ten times on different training sets to yield an overall error estimate.

Algorithm
Decision forest Classifier (DFC):
A Decision Forest Classifier consists of a collection of individual tree classifier. The concept of a general random forest is as follows:
Step 1: Let the number of training cases be N, and the number of variables in the classifier be M.
Step 2: We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
Step 3: For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set
Step 4: for I = 1 to *N*
Call construct forest ( ):
Step 5: Draw a bootstrap sample from the data, term those not in the bootstrap sample as *out_of_bag* data.
Step 6: Grow a random tree, where at each node the best set is chosen among m randomly selected variables. The tree is grown to a maximum size and it should not be pruned.
Step 7: use the tree to predict *out_of_bag* data.
Step 8: in the end use the predictions on *out_of_bag* data to form majority votes.
Let the feature space be represented as F = {f_1, ….f_M }, where M is the dimension of F. A database is created as *db* = {t_1,.....t_I} where I is the size of db. Each record $t_i \in$ *db* is represented as a real valued vector t_i = { $t_{i,1}$, ... $t_{i,M}$ }. To train the random forest, we require a training set S = {(s1, v_1)… (s_N, v_N)}, which intern is extracted from brain images after it is pre-processed and certain features are then extracted from it. We apply Random forest with relevance feedback and train a 3-class classifier *h* to classify the database objects as normal, moderate, severe.

Pseudocode for construct_forest ( )
Inputs:
a) training set $S = \{(s_1, v_1), \ldots (s_N v_N)\}$; where v_n = 0,1, or 2
b) Feature space $F = \{f_1 \ldots f_M\}$;
c) J: the number of tree classifiers to grow.

Step 1: *Initialization* set the random forest h ← { }
Step 2: *for* $j = 1$ to J do
Step 3: *Generate* the bootstrap sample set $S_j \subset S$;
Step 4: *Train* a tree classifier (CART) h_j with S_j;
Step 5: Set h ← h ∪ {hj};
Step 6: *end for*.

Results and discussion
WEKA software (Waikato Environment for Knowledge Analysis) is used for the simulation purpose. The decision forest implementation in WEKA allows choosing the number of trees and also controlling the random attributes required at each node. It contains tools for classification, regression, clustering, association rules, and visualization. Random forest classifier was applied to the data set which consists of 153 instances, with seven attributes of which 39 are Normal, 38 are Moderate and 74 are Severe. In our study, the features of brain image are extracted from statistical approach. These features formed the input parameters for the classification stage. The classifier accuracy was estimated by using the test option 10 - fold cross validation. Here 90% of the data is used for training and remaining 10% is used for testing. The experiment was conducted on the data set, and the average is computed. It divides the available samples into s disjoint subsets where 1≤ s ≤10. (s-1) subsets are used for training and remaining subset for testing. As we know the performance accuracy is measured in terms of Classification accuracy. The classification accuracy is computed using the confusion matrix, which helps in understanding the correctness of a test set model. In the present experiment we have taken three classes for classification Normal, moderate and Severe. Our test set consists of 153 instances of which 39 are Normal, 38 Moderate and 74 Severe. The Classification of Decision Forest Classifier over the given set of data with various algorithms is presented in Table 2. The accuracy of the Random forest Algorithm is found to be 99.34%. The confusion matrix obtained for Test data is shown in Table 3. It is observed that the numbers of correctly classified instances are good. It is important to note that the Random forest classification gives very good accuracy compared with the existing results which is shown in the Table 4. The comparison of various classification

accuracies of existing methods and the proposed methods are shown in the Fig (4).Over the past decade functional magnetic resonance imaging (fMRI) has emerged as a powerful new instrument to collect vast quantities of data about activity in the human brain. The study of human brain function has received a tremendous boost in the recent years from the advent of fMRI, a brain imaging method that dramatically improves our ability to observe correlates of neural activity in human subjects at high spatial resolution across the entire brain. Electronic Medical Records such as MRI, and fMRI were used to examine the Neurodegenerative disorders. It is a great challenge to neurologists to get the information from electronic records through experience. In this paper classification and categorization of Electronic Medical Records are performed by random Forest classifier for three cases such as Normal, Moderate and Severe. Experimental results on the image data set have proved to be efficient, resulting in an accuracy of 99.32% and are better than the accuracies obtained from the existing methods. This work can be of enormous use as it can be used to extend the model to other neurodegerative disorders such as Huntington's disease and Parkinson's disease for the classification. This can be used by the Neurologists and radiologists to get the information from Electronic Medical Records to classify disease more accurately rather than deciding through experience.

Acknowledgment

The authors would like to thank Dr. Dallas Anderson, National Institute of Aging, USA for providing the MRIs data, Nancy Lombardo, Alzheimer's Association, Chicago and Dr. Hanumanthachar Joshi, International Society to Advance Alzheimer Research and Treatment (ISTAART), USA, for comments and information on various MRIs.

References

[1] Scatena R., Martorona Bottani P., Botta G., Pastove P. and Giardina. (2007) *Expert Opin. Investig. Drugs*, 16, 59-72.

[2] Jay M. E. (2005) *JAOA,* 3, 145-158.

[3] Caselli R .J., Beach T. J., Yaari R. and Reiman E. (2006) *J. Clini. Psychiatry*, 67, 1784 -1800.

[4] Hashimoto M., Rockenstein E., Crews and Masliah E. (2003) *Neuromolecular. Med.,* .4, 21-36.

[5] Kerr M. and Small D. (2007) *J. Neurosis.* 80, 151-159.

[6] Hanumanthachar Joshi, Parle M. and Disale C. (2008) *Alzheimer's and Dementia*, 4, 760-763.

[7] Hanumanthachar Joshi and Parle M. (2007) *Colombia medica*, 38(2), 132-139.

[8] Hanumanthachar Joshi and Disale C. (2008) *Frontiers in Neuroscience*, 8, 12-18.

[9] Francesco C., Audrey M. N., David R. M., Ercolano M. and Paolo. (2006) *Evid Based Complement Alternat Med,* 3, 411-424.

[10] Jogren M. S., Mielxe M., Gustafson D., Zandi P. and Skoog I. (2006) *Mech. ageing dev*, 127, 138-147.

[11] [Geldmacher D. S. (2003) *J. Am. Geriatr.soc,* 51, 89-95.

[12] Subramani M., William Rodman Shankle, Michael J., Pazzani, Padhraic Smyth and Malcolm B. (1999) *J. Inter Neuropsycholog Soc.,* 12, 377-392.

[13] Vasileios Megalooikonomou, Despina Kontos and Dragoljub Pokrajac. (2004) *Artificial Intelligence*, 3 -6

[14] Despina Kontos and Vasileios Megalooikonomou. (2004) *SPIE*, 53-70.

[15] Vasileios Megalooikonomou (2004) *International Conference on acoustics, speech, and signal processing, Montreal, Canada*. *Published by IEEE Computer Society in IEEE Xplore USA*, 614-615.

[16] Lazarevic A., Pokrajac D., Megalooikonomou V., and Obradovic Z. (2001) *4th International Conference on Neural Networks and Expert Systems in Medicine and Healthcare,* Milos Island, Greece 20-22.

[17] Han M. and Kamber (2001) *Data Mining: Concepts and Techniques, Published by Morgan Kaufmann, San Francisco,* 279-284.

[18] Leo Breiman (1999) *Random Forests-Random Features Technical report published by university of California, Berkeley,* 567.

Table 1- General Confusion Matrix for Two Classes

	Class Positive (C+)	Class Negative (C-)
Prediction Positive (R+)	True Positive (TP)	False Positives (FP)
Prediction Negative (R-)	False Negatives (FN)	True Negatives (TN)

Table 2- Classification Accuracy of Machine Learning Algorithms

Algorithms	Number of Runs	Classification Accuracy	Sensitivity	Selectivity
C4.5	10	99.2329	0.921	0.957
C4.5 rules	10	98.0262	0.951	0.984
Random Forest	10	99.3421	0.988	0.9765
Naïve Bayes	10	99.0878	0.986	0.949
PART Rule	10	99.0576	0.976	0.986

Table 3- Confusion Matrix Over a Test Set

	Normal	Moderate	Severe
Normal	22 (C_{11})	00(C_{12})	00(C_{13})
Moderate	00(C_{21})	38(C_{22})	01(C_{23})
Severe	00(C_{31})	00(C_{32})	14(C_{33})

Table 4-Classification Accuracies of Various Existing Methods

Existing Methods	Classification Accuracy	Sensitivity	Selectivity
Characterization technique	82.76	0.876	0.897
Adaptive Recursive Partitioning Method	90.97	0.912	0.890
Quantitative Characterization Technique	94.078	0.937	0.956

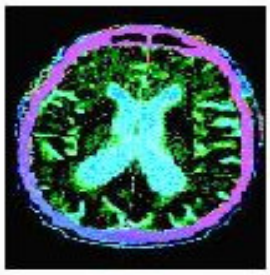

Fig. 2-Brain image with noise

Fig. 3- Filtered image without noise (By Salt and Pepper Method)

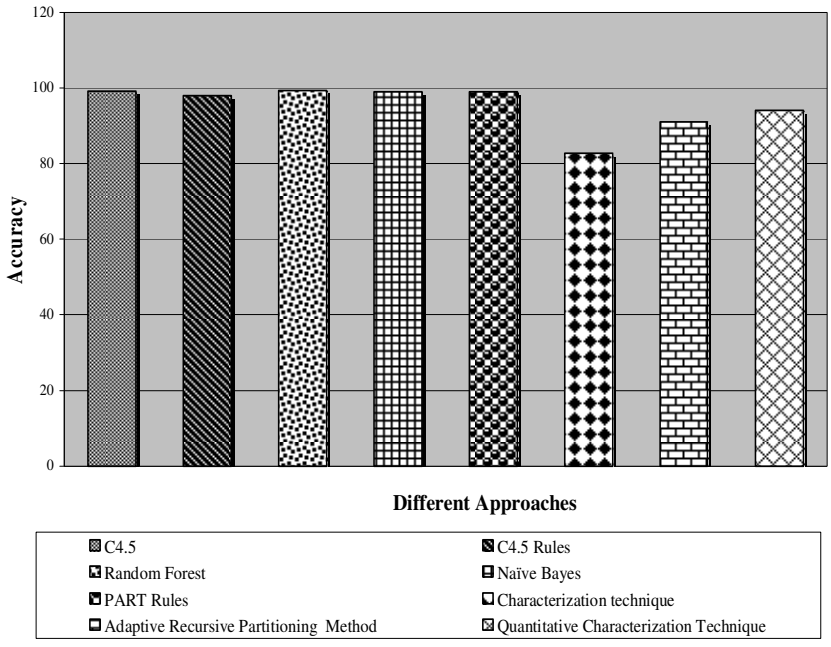

Fig. 4- Comparison of classification accuracies obtained from various methods