IJCSC 0973-7391 Volume 8 • Issue 2 March 2017 - Sept 2017 pp. 12-17 available online at www.csjournalss.com

Comparative Study of Data Mining Classification Methods in Brain Tumor Disease Detection

[1]Varun Jain, [2]Sunila Godara

[1,2]Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract: In order to detect brain tumor, a large amount of data is provided by Magnetic Resonance Imaging technique. Clustering can also be done in order to cluster the MRI data [16]. As human inspection results to low accuracy, we use data mining classification methods to achieve high accuracy such that it helps in further treatment. In this paper we analyze different data mining classification methods: Naïve Bayesian, Decision tree (LMT), Support Vector Machine (SMO) and Artificial Neural Network (MLP) on Primary Tumor data set. Performance of these techniques is compared in the terms of sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. Further 10-fold cross validation method is applied in order to avoid over-fitting. . As per our results error rates for Naïve Bayesian, Decision Tree, ANN and SVM are 0.114, 0.178, 0.088 & 0.061 respectively and accuracy are 92.03%, 93.51%, 91.44%, 93.80% respectively. Our analysis shows that out of these four classification models, SVM (SMO) is the best classifier to detect brain tumor disease with high accuracy and lowest error rate.

I. Introduction

Human brain represents only 2% mass of total body but uses 20% body's energy [1]. Brain controls all the activities of the human body. So the brain needs to operate with its maximum efficiency. Now-a-days, a lot of people are suffering from brain tumor which causes even death, if not treated at time. Brain Tumor is a cluster of abnormal cells growing rapidly in the brain and clustering is also used for grouping of similar cells[16]. It may occur to any person at any age and appear at any location in the brain. Tumor is further categorized in two: malign and benignant. Benignant tumors have homogeneous structure and don't contain cancer cells while malign have heterogeneous structure and contain cancer cells. Benign tumors are either radio-logically or surgically destroyed and have rare chances of grow back. Malignant are life threatening tumor and can be treated by chemotherapy, radiotherapy or their combination. So, need to diagnose the tumor at an early stage is essential for future treatments.

MRI (Magnetic Resonance Imaging) has proven out as a powerful tool in detection of brain tumor with the help of MR Images. It is a non-invasive technique which produces very detailed 2D and 3D images of the organ inside the brain in every direction. As the large amount of data provided through MRI technique, so it is impractical to develop a method which can classify the images in normal or abnormal through human inspection. [2]

Data Mining has been known for evolving out some important features from large amount of data. Due to this specialization of data mining, this field is used in combination with medical science for the accurate diagnosis of the patient disease. A no. of classification methods has been evolved under data mining. In order to achieve best accuracy model, we will compare the accuracy determined by the different classification models given as: SVM, decision tree classifier, Naïve Bayesian Classifier and KNN algorithm on the specific datasets: primary tumor dataset obtained from the UCI Web Repository.

II. Brain Tumor Detection Models

Under this section we will discuss following data mining classification models to detect brain tumor:

A. Decision Tree

Decision trees are the powerful and greedy classification algorithms. The most popular are Quinlan's ID3, C4.5 and CART algorithm. As the name implies, a tree is constructed in a top-down recursive divide and conquer manner. At start, all the observations are at the root. Then the test attributes are selected on the basis of some heuristic or statistical measure, (e.g. information gain). It splits the input observations into two or more subgroups. This process is repeated recursively until the complete tree is constructed. Our main objective is to find the variable-threshold pair which best splits the observations into subgroups. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used inID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID).

Below Fig. 1 [7,15] shows an example of decision tree on patient diagnosis. Internal nodes represent test on one or more attributes and terminal nodes show decision outcomes. Decision tree summarizes the data as: If a patient has swollen glands, then diagnosis has steep throat. If no swollen glands, then check for fever. If patient has fever, then diagnosis done as patient has cold otherwise patient has allergy.

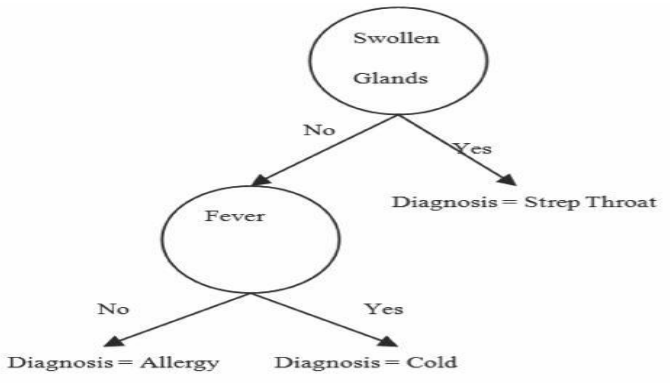


Fig. 1: Decision Tree

B. Artificial Neural Networks

Artificial Neural Networks are the biologically inspired networks which have the tendency to model extremely complex non-linear functions. ANNs are the highly sophisticated analytical techniques having the capability of learning the existing data. Multi-layer perceptron (MLP) with back-propagation is a supervised learning algorithm which is the one of popular ANN architecture. The use of this algorithm is started by psychologists and neurobiologists in order to develop test computational analogues of neurons. Fig. 2 shows MLP feed forward Neural Network. A neural network has a set of connected input/output units where each connection has a weight associated with it. The main usage of neurons in input layer Xi to divide the input signals among neurons in hidden layer. Every neuron j in hidden layer sums its input signals with connections Wji from the input layer and output function given as

$$Yj = f(\Sigma\ Wji\ Xi)$$

The final hidden layer's outputs are input to units of the **output layer**, which evolves the network's prediction value. The network is **feed-forward** because no feedback of output unit to an input unit or to an output unit of a previous layer. The output in the output layer is determined in an identical manner [12].

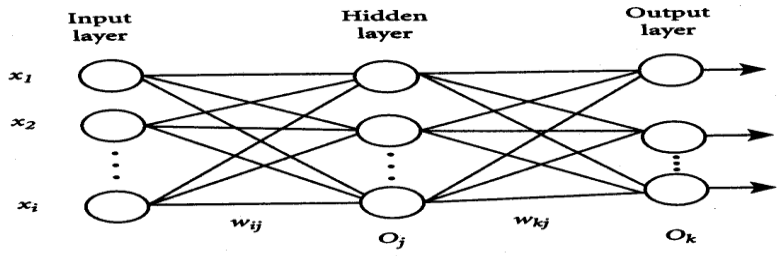


Fig. 2: MLP

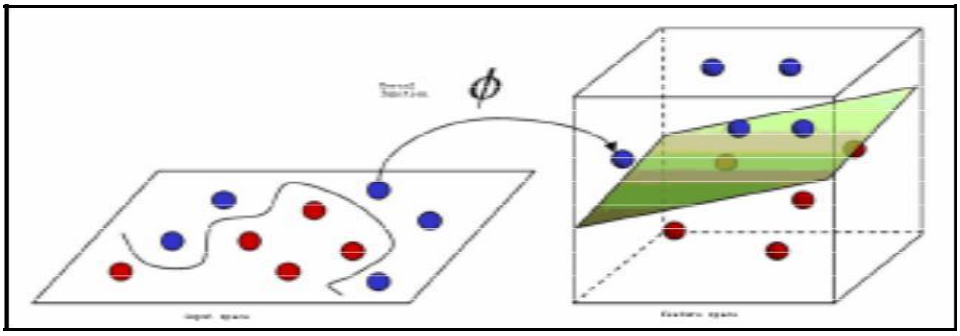
The back-propagation algorithm is used to train neural networks. This algorithm iteratively processes a set of training tuples and then compares the actual target value with the predicted value. The weights are modified for each training tuple in order to minimize the MSE (Mean Squared Error) between actual target value and network's predicted value

$$MSE = \tfrac{1}{2}\ N\ \Sigma\ (X\text{-}X')2$$

Where N is the number of experimental data points utilized for the training.

C. Support Vector Machine

SVM is an up to the minute classification algorithm used for the classification of both linear and non-linear data. This classifier is derived from statistical learning theory given by Vipnik in 1992. SVM classifier approaches the problem by finding out the hyper-plane with largest margin, i.e. maximal marginal hyper-plane. For the data which is not linearly separable, it transforms the original training data into a higher dimension by doing non-linear mapping. By transforming it into high dimensional space, it searches for linear optimal separating hyper-plane. This transformation technique into high dimension always helps in searching for an optimal hyper-plane using support vectors and margins [13]. SVM performs classification by finding optimal MMH and minimizing the classification errors. Fig 3 [7, 15] shows SVM topology in hyperspace:



D. Naïve Bayesian Classifier

Bayesian classifier demonstrated as a statistical classifier which performs probabilistic prediction, i.e. class membership probabilities. Its foundation based on bayes theorem which described as below given training data X, and posterior probability of hypothesis H is P (H|X):

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

A Bayesian classifier has close performance with decision tree and ANN classifier. Each training example can affect the probability that a hypothesis is correct either increase or decrease — some prior knowledge can be combined with observed data. Let G be set of training tuples attached with class labels. Each tuple is represented by attribute vector given as $\mathbf{X} = (x_1, x_2,\ldots,x_{i,\ldots,} x_n)$. Let there are a total of *z* classes $C_1, C_2,\ldots, C_z$. Classification is to get the maximum posteriori probability, i.e., $P(C_i|\mathbf{X})$. This can be obtained with the help of Bayes theorem. As P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized.

III. Data Source

To compare these data mining classification techniques Primary Tumor dataset from UCI repository was used. The Primary Tumor dataset has 17 attributes and 339. Table 1 below lists these attributes:

Table1: Primary Tumor Data Set Description

No.	Name	Description	No.	Name	Description
1	Class	lung, head & neck, esophasus, thyroid, stomach	10	Penitoneum	Yes, no
2	Age	<30,30-59,>=60	11	Lever	Yes, no
3	Sex	Male, female	12	Brain	Yes, no
4	Histologic-type	Epidermoid, Adeno, Anaplastic	13	Skin	Yes, no
5	Degree-of-diffe	Well, fairly, poorly	14	Neck	Yes, no
6	Bone	Yes, no	15	Superclavicular	Yes, no
7	Bone-marrow	Yes, no	16	Axillar	Yes, no
8	Lung	Yes, no	17	Mediastinum	Yes, no
9	Pleura	Yes, no	18	Abdominal	Yes, no

IV. Results

These data mining classification model were developed using data mining classification tool Weka version 3.6. Initially dataset had 17 attributes and 339 records for Primary Tumor data set. Algorithm for attribute selection was applied on dataset to preprocess the dataset. After attribute selection missing values records were identified and were deleted from dataset. After deleting records with missing values we were left with modified records. On these records data mining classification techniques Naïve Bayesian, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) were applied. Sensitivity, specificity and accuracy are obtained from the confusion matrix. Confusion matrix is the representation of classification results in the form of matrix.

Table 2: Confusion Matrix

	Classified As Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

The upper left cell shows the number of sample which are classified as True while they were actually true, i.e., TP and the lower right cell shows the number of samples which are classified as false by classifier while actually they were false, i.e., TN. The lower left cell and upper right cell shows the number of samples misclassified by classifier. The upper right cell is showing the number of samples classified as false by classifier while actually they were true, i.e., FN, and the lower left cell showing the number of samples classified as true by classifier while actually false, i.e., FP. Below formulae were used to calculate sensitivity, specificity and accuracy:

$$\text{Sensitivity} = TP / (TP + FN) \quad , \quad \text{Specificity} = TN / (TN + FP)$$

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

Table 3: Comparison different classification algorithm

	Sensitivity	Specificity	Accuracy
Naïve Bayes	9.52 %	100%	92.03%
Decision Tree LMT	0%	99.68%	93.51%
ANN (MLP)	14.28%	96.54%	91.44%
SVM (SMO)	0%	100%	93.80%

Figure 4: Graphical representation of sensitivity, specificity and accuracy.

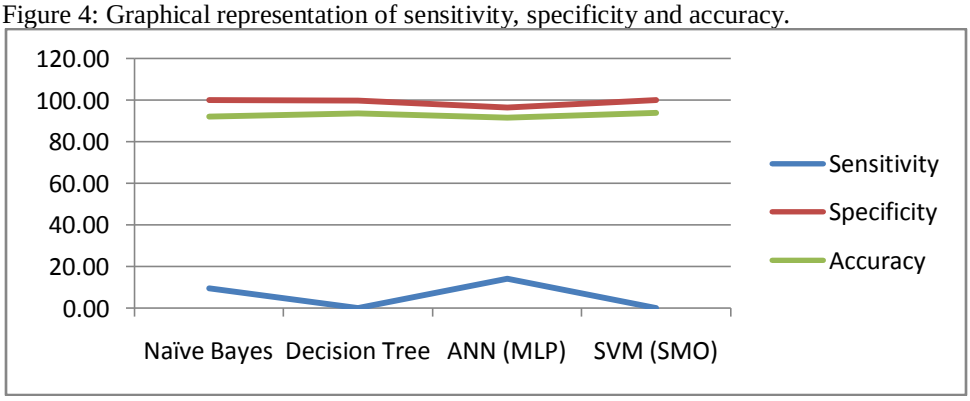


Figure 4 show the graphical representation of accuracy, sensitivity and specificity of Naïve, LMT, ANN, SMO classifier concluded from table 3. It shows that out of four, SMO classifies best.

The error rate for naïve, LMT, ANN and SVM are 0.114, 0.178, 0.088 & 0.061 respectively.

True positive rate and false positive rate is calculated from confusion matrix given as:

True Positive Rate = TP / (TP + FN)
False Positive Rate = FP / (FP + TN)
Table 4 shows True Positive Rate and False Positive Rate for naïve Bayesian, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM).

Table 4: True Positive Rate and False Positive Rate

	True Positive Rate	False Positive Rate
Naïve Bayesian	0.920	0.850
Decision Tree LMT	0.935	0.938
ANN (MLP)	0.914	0.806
SVM (SMO)	0.938	0.938

This result shows that out of all classification algorithms, Support Vector Machine performs better than all other in aspect of all parameters like sensitivity, specificity, accuracy and error rate. Accuracy of SMO algorithm is near to the perfect point, a little margin close to decision tree which shows SVM (SMO) to be the best detector of brain tumor.

V. Conclusion:

There are different data mining techniques that can be used for the detection and prevention of brain tumor disease among patients. In this paper four classification techniques in data mining to predict brain tumor disease in patients are compared: Naïve Bayesian, decision tree LMT, Artificial Neural Networks and Support Vector Machine. These techniques are compared on behalf of True Positive Rate, False Positive Rate, Sensitivity, Specificity, Accuracy and Error Rate. Our studies showed that Support Vector Machine model turned out to be best classifier for brain tumor detection. In future we intend to improve performance of these basic classification techniques by using some hybrid approach in terms of accuracy and other measuring criteria.

References

1. Fox, Michael D. and Raichle, Marcus E (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature Publishing Group, vol. 8.
2. Chaplot, S.; Patnaik, L.M.; Jagannathan, N.R. (2006). Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network, Biomed. Signal Process Control, 1, 86–92.
3. Maitra, M.: Chatterjee, A. (2011). A Slantlet transform based intelligent system for magnetic resonance brain image classification. Biomed. Signal Process Control, 1, 299–306.
4. Zhang, Y.; Wu, L.; Wang, S. (2011). Magnetic resonance brain image classification by an improve artificial bee colony algorithm. Progress Electromagnetic Resolution, 116, 65–79.
5. Zhang, Y.; Wu, L. (2012). An MR brain images classifier via principal component analysis and Kernel support vector machine. Progress Electromagnetic Res., 130, 369–388.
6. Naik, J.; Prof. Patel, Sagar (2013). Tumor Detection and Classification using Decision Tree in Brain MRI. IJEDR , ISSN:2321-9939.
7. Kumari, M.; Godara, S.(2011). Comparitive Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. IJCST, Vol. 2, ISSN: 2229-4333.
8. Zhang, Y.; Lu, S.; Zhou, X. et al. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors and support vector machine. Simulation, Vol. 92(9), 861–871.
9. Gonzalez, R.C. and Woods, R.E. (2009). Digital Image Processing, Third edition, Prentice-Hall.
10. Navneet, L. and et al. (2012). A Novel Machine Approach for Detecting the Brain Abnormalities From MRI Structural Images. PRIB, pp. 94-105.
11. Chau, M.; Shin, D. (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 183-187.
12. Patil, S.; Kumaraswamy, Y. (2009). Intelligent and effective Heart Attack prediction system using data mining and artificial neural networks. European Journal of Scientific Research, Vol. 31, pp. 642- 656.

13. Han, J.; Kamber, M. Data Mining Concepts and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco.
14. Palaniappan, S.; Awang, R. Intelligent Heart Disease Prediction System Using Data Mining Techniques. Proceedings of IEEE/ACS International Conference on Computer Systems and Applications 2008, pp. 108-115.
15. Godara, Sunila and Singh, Rishipal (2016). Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis. Indian Journal of Science and Technology, vol. 910.
16. Godara, Sunila and Verma, Amita (June 2013). Analysis of Various Clustering Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-3, Issue-1.