Rough K-means and support vector machine based brain tumor detection

Conference Paper · September 2017

DOI: 10.1109/ICACCI.2017.8125826

CITATIONS
6

READS
12

2 authors:

Amiya Halder
St. Thomas College of Engineering and Technology
50 PUBLICATIONS **270** CITATIONS
SEE PROFILE

Oyendrila Dobe
Michigan State University
7 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Rough K-means and Support Vector Machine based Brain Tumor Detection

Amiya Halder
Department of CSE
STCET, Kolkata-23, India
Email: amiya.halder77@gmail.com

Oyendrila Dobe
Department of CSE
STCET, Kolkata-23, India
Email: oyendrila.dobe@gmail.com

***Abstract*—In this paper, we present a proposed algorithm to classify brain MRI as tumor-free or tumor present. For computing difference between normal and abnormal MR images, a set of features is calculated. The number of features of the original feature set is reduced by rough set based K-means algorithm and classification of the dataset into tumor-free or tumor-present category has been done by support vector machine (SVM). The proposed algorithm gives better classification result for a smaller set of training dataset.**

Keywords:-Brain tumor; Feature selection; SVM; Rough K-means.

I. Introduction

Tumor is defined as a growing mass of unwanted and uncontrolled cells. The skull being rigid, unwanted growth of cells inside such restricted perimeter can cause problems. Brain tumors are either malignant or cancerous. When tumors grow, they increase the internal pressure and can lead to fatal damage. They can be classified into two types- primary tumors originate inside the brain; while secondary tumors occur when cancerous cells spread to the brain from another organ. A brain tumor can be genetically inherited, although the percent of such cases hardly exceed 5-10%. Continuous exposure to radiation and chemicals can also cause brain tumors. It usually is a slow growing abnormality that reaches a fatal level before it is detected. The approach of treatment will depend on the type of tumor, the size of tumor, the location of tumor and general health of the patient. Early treatment can prevent any future complications that might occur due to excess pressure that arises in the brain due to the tumor. Since it is treatable if detected in initial stages, several researches have been directed towards the detection process. The detection algorithm suggested here is found to give better results in several stages including feature set dimension reduction and support vector machine training, than previously suggested algorithms.

II. Related works

In order to detect the tumor, several scientific technologies have been developed. Usually a dye is injected in order to trace the blood flow in the tumor region. An X-ray scan can be used to detect calcium deposits, if the tumor has moved a bone, due to pressure; a computed tomography (CT) scan uses a specified dye to make it easier for doctors to detect blood vessels; an angiography also involves injecting a dye, to trace the origin and nature of blood flow in the tumor region; a biopsy involves collection of tissue sample to detect if the tumor is benign or malignant [1]. Among these Magnetic Resonance provides the best detection result. It also uses a dye to detect tumor, however it eliminates the use of radiation and provides deeper coverage than other scan results. Hence MRI scans have been used rigorously in the field of medical image analysis. An ANN approach has been used by Abhijeet Zamre et al. [2]; a pattern recognition algorithm using statistical measures have been proposed by Nathali Richard et al. and Zhang et al. [3] [4]; an effective fast-marching method with mathematical morphology is proposed by Zang [5]; a rule-based algorithm has been proposed by Matesin [6]; a local thresholding technique on CT scan images, utilizing maximum entropy principle has been proposed by Ruthmann et al. [7]; a combination of K-means clusters and neural network has been suggested by Loncaric et al. [8]; a combination of fuzzy c-means algorithm along with SVM has been proposed by A. Halder et al. [9]. Other suggested techniques include use of knowledge based techniques by Clark et al. [10], Bayesian classifier by Lauric et al. [11], genetic algorithm by Ganesan et al. [12], genetic algorithm along with morphological operations have been used by Halder et al. [18], LDA by Majos. C et al. [13], a combination of SVM and artificial neural network has been suggested by Ahmmed [19]. In this paper we are suggesting the utilization of rough set theory in classifying MRI scans. Rough set theory provides an efficient new method of clustering and is shown to give better results using a smaller set of features than fuzzy c-means [9].

III. Proposed Method

The proposed method initially extracts features from brain MRIs [9], [14]. The feature set dimensionality is then reduced using rough set based K-means algorithm [15]. Further, a support vector machine has been trained with a radial basis function [16] to classify MRI scan images as abnormal or normal. Each of the components mentioned above has been defined in details below.

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

A. Data Description

Three sets of data have been created, consisting of normal as well as tumor affected brain scans. The scans have been used for feature set extraction, feature set size reduction, training and testing the SVM. The MR scan images utilized here has been extracted from the IXI-dataset [17]. Both T1 and T2 types of MR scan images have been utilized to check the effectiveness of the proposed method.

B. Normalization

All the scan images have initially been normalized to grayscale and their intensity values have been limited to the range 0-255. A gray level covariance matrix is generated for the purpose of feature extraction from each scan image.

C. Feature Sets

An image is represented by pixel values at all coordinates within the dimensions of the image. Instead of using the whole image in the process, we have extracted certain features from each scan image. The values of the extracted features provide us with the necessary information about an image that can be used to capture the minute differences in the image. Since dyes have been used in MRI scan, each tumor affected region is described in the image using a different color than the rest of the image. These features help to alternatively represent an image accurately. Hence they can be used as input to the classifier for image classification. We consider the grayscale image $g(x, y)$, $I(i)$ as the intensity level of an image, L_n as the gray levels in the image and $pd(i)$ as the probability density.The following features have been utilized to represent the scan images accurately:

1) Entropy of an image: It is defined as the amount of information which must be coded for by a compression algorithm. Low entropy images have very low contrast. Similarly, a perfectly flat histogram based image will have a zero entropy.

$$En = -\sum_{c=0}^{L_n-1}\sum_{d=0}^{L_n-1}(pd(c,d) * log_2(pd(c,d)+1)). \quad (1)$$

2) Contrast of an image: It refers to the difference in luminance that makes an object distinguishable from its background

$$con = \sum_{c=0}^{L_n-1}\sum_{d=0}^{L_n-1}(c-d)^2 * pd(c). \quad (2)$$

3) Energy of an image: It refers to how the variance of a signal is spread with respect to frequency.

$$ener = \sum_{c=0}^{L_n-1}\sum_{d=0}^{L_n-1} pd(c-d)^2. \quad (3)$$

4) Homogeneity: It measures the similarity among the pixel intensity in the image.

$$homo = \sum_{c=0}^{L_n-1}\sum_{d=0}^{L_n-1}\frac{pd(c,d)}{(1+(c-d)^2)}. \quad (4)$$

5) Correlation among pixels: It measures the extent of relation between the reference pixel and its neighboring pixel.

$$cor = \frac{1}{sd_w sd_v}\sum_{c=0}^{L_n-1}\sum_{d=0}^{L_n-1} c * d * pd(c,d)^2 - mn_w mn_v. \quad (5)$$

Where sd_w, sd_v are standard deviations and mn_w, mn_v are means of $pd(c)$, $pd(d)$.

6) Variance of the image: It measures the extent of variation among the pixel of the image.

$$vary = \sum_{c=0}^{L_n-1}(c-m_w) * pd(c). \quad (6)$$

where m_w is the mean of pixel values.

7) Standard deviation: It is the extent of how much the pixel values has deviated from the mean value.

$$stndv = \sqrt{\sum_{c=0}^{L_n-1}(c-m_w) * pd(c)}. \quad (7)$$

8) Sumvariance: It is used to measure the variance of the sum of pixels in the respective columns and rows.

$$sumvar(svar) = \sum_{c=0}^{2(L_n-1)}(c-sen)^2 * pd_{w+v}(c). \quad (8)$$

9) Sum average: It measures the average of the sum of the pixels in the respective columns and rows.

$$sumaverage(su) = \sum_{c=0}^{2(L_n-1)} c * pd_{w+v}(c). \quad (9)$$

10) Sum entropy: It measures the entropy among the sum of rows and columns.

$$sumen(sen) = \sum_{c=0}^{L_n-1} pd_{w+v}(c) * log(pd_{w+v}(c)). \quad (10)$$

11) Inertia: It refers to the ability of the image pixels to resist changes.

$$inertia(int) = \sum_{c=0}^{L_n-1}\sum_{d=0}^{G_n-1}(c-d)^2 * pd(c,d). \quad (11)$$

12) Kurtosis:It refers to the measurement of flatness of the histogram.

$$kurtosis(kur) = a^{-4}\sum_{c=0}^{L_n-1}((c-m_w)^4 * pd(c)) - 3. \quad (12)$$

D. Rough K-means based Reduction of Feature Set

The above discussed features together, approximately describe an image and it can be used to distinguish among them. However, few features act as redundant, and they unnecessarily increased the size of dataset, increasing the execution time of the algorithm. Hence, in this step we propose to remove them using an innovative clustering. Rough based K-means [15] algorithm utilizes rough set theory to improve K-means algorithm used for clustering. It is better than K-means clustering since it considers the fact that boundary pixels may be associated with more than one cluster. In traditional K-means, the clusters formed are crisp and the data points belonging to each cluster do not overlap. However, a boundary pixel may have an equal probability of belonging to more than one cluster. Such cases are efficiently resolved by introducing rough set concepts to the traditional K-means algorithm. Let us consider a given set of objects, $V = \{v_1, v_2, , v_n\}$, and we are supposed to divide them in to k clusters, $C = \{c_1, c_2, .., c_k\}$. Each cluster center will have its lower approximation set $\underline{L}(c_i)$ and upper approximation set $(\overline{U}(c_i))$. They will be occupied by objects using the following:

1) Any object v_i shall be a part of at most any one lower approximation of a cluster center. This ensures that no two lower approximations overlap.
2) Any object v_i which is a member of a lower approximation of a cluster center is also part of its upper approximation by default $(v_i \in \underline{L}(c_i) \rightarrow v_i \in \overline{U}(c_i)$. This suggests that lower approximation of a cluster center is a subset of its corresponding upper approximation $\underline{L}(c_i) \subseteq \overline{U}(c_i))$.
3) If any object v_i is not part of the lower approximation set of any cluster center, it belongs to upper approximation sets of two or more cluster centers. This introduces the fact that an object cannot belong to a single boundary cluster. Hence the traditional K-means encounters a change in cluster center calculation. Alongside the original concept, an extra overhead is added, where we decide to which cluster a boundary object should be assigned.

For each of the above feature, a threshold value has been generated by calculating the feature value for a number of normal brain MRI scan image. An average of the set of values has been considered as the threshold value for the corresponding feature. The algorithm is given below that elaborates the process.

1) Calculate the set of features from each image.
2) In every feature dataset choose K no. of cluster centers randomly.
3) Using Rough based K-mean algorithm, segment the dataset into K clusters.
4) Choose maximum cluster center value in each feature dataset.
5) If the threshold value of a feature is less than maximum cluster value then
   a) The feature is selected otherwise
   b) The feature is not selected

E. Classification of Images Using Support Vector Machine

Support vector machine (SVM) is a method which is widely used in machine learning for data classification. SVM is a supervised learning method. It creates decision planes that define boundaries between classes. A decision plane or hyperplane acts as a boundary to separate different class memberships [9]. A random hyperplane is represented as,

$$\overrightarrow{\alpha}.\overrightarrow{\beta} - \gamma = 0 \tag{13}$$

Where $\overrightarrow{\alpha}$ is the normal vector to the decision plane and $\frac{\gamma}{|\alpha|}$ is the offset, along the direction of the normal to the hyperplane, from the origin. If the training data set is separable linearly, two parallel hyper planes are created which separate the data into two classes, so that the intra-class distance is as small as possible and the inter-class data is as large as possible. The region between these two hyper planes is called the "margin", and the maximum-margin hyperplane is the hyperplane that lies midway between them. These hyper planes can be represented as,

$$\overrightarrow{\alpha}.\overrightarrow{\beta} - \gamma = 1 \tag{14}$$

and

$$\overrightarrow{\alpha}.\overrightarrow{\beta} - \gamma = -1 \tag{15}$$

Here a radial basis kernel has been used in the SVM. Radial basis function(RBF) is one of the most popular kernel function. It adds a little 'bump' around each data set. The most widely used RBF is Gaussian functions given below,

$$g(r) = \sum_{p=1}^{m} \alpha_p exp(-\gamma||x_p - x||^2) + b \tag{16}$$

In SVM, the training phase is used to input values which fix support vectors using which the algorithm learns to generate the hyperplane to divide the data into a specified number of classes. During the testing phase, the input data are plotted on the n-dimensional; utilizing the previously set support vectors, the class to which the data belongs is decided.

IV. Experimental Results

A dataset is created, of 150 MRI images, containing both normal and abnormal scan images. In the feature set dimensionality reduction step, from the set of 12 features, 6 are found to be sufficient in properly defining the scan images (eliminate the redundant features). In contrast to this, a similar algorithm using FCM [9], reduced the feature set from 12 to 9, to generate similar result. Hence rough based K-means is found to be more efficient in segmenting the dataset. For the training phase varying size of dataset has been used to compare a similar algorithm using FCM. As predicted, the proposed algorithm is seen to provide better results. The test dataset has 150 scan images consisting of both tumor-free and tumor-present scan images. Out of them 80 are normal images, i.e, tumor-free scans, and 70 are abnormal images, i.e, scans where tumor are known to be present.

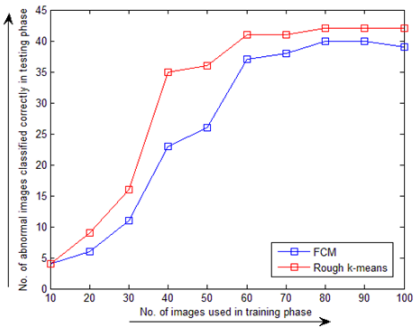

Fig. 1. Plot to compare test phase results using FCM and rough based segmentation of dataset.

TABLE I
RESULTS OF THE PROPOSED METHOD

Method	Normal image	Identified correctly	Abnormal image	Identified correctly
Fuzzy c-means	80	80	70	59
Proposed method	80	80	70	68

TABLE II
ACCURACY MEASURE OF THE DIFFERENT METHODS.

Different method	**Accuracy**
Fuzzy C-means algorithm	85.45%
K-means method	87.87%
Bayes classifier	89.23%
Genetic Algorithm	93.23%
Support Vector Machine approach	92.33%
Neural Network based approach	96.21%
FCM with SVM	97.89%
Proposed Method	99.05%

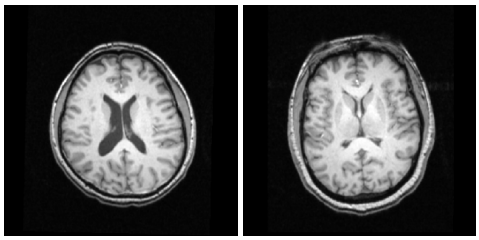

Fig. 2. Normal brain images.

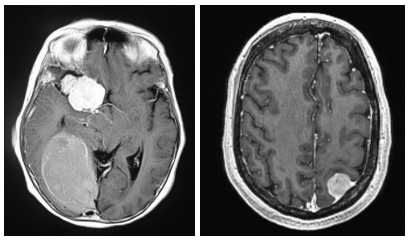

Fig. 3. Brain MR images affected by tumor.

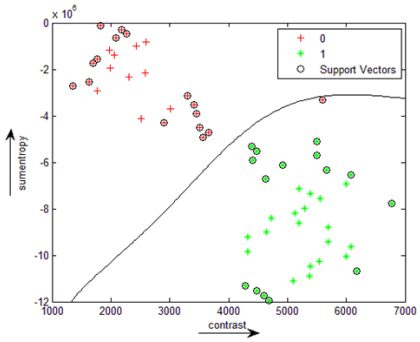

Fig. 4. Plot to show support vectors fixed in training dataset by SVM from two different classes

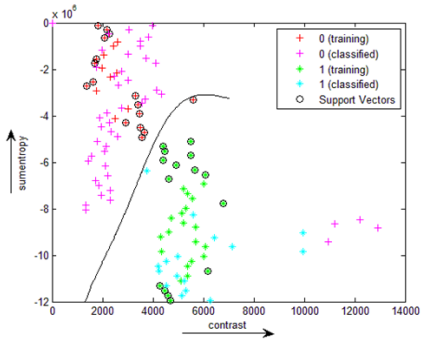

Fig. 5. Plot to show training and test sessions, along with fixed support vectors in SVM being partitioned two classes.

The performance of the proposed method is shown in TABLE I and TABLE II, where accuracy is compared with K-means, Fuzzy C-means, Genetic Algorithm, SVM, Bayesian classifier and Neural Network. From the accuracy measure, our proposed method gives satisfactory results. Normal and abnormal MR scan images are shown in Fig.2 and Fig.3 respectively. Plot for SVM in training data and for both training as well as classified data are shown in Fig.4 and Fig.5 respectively.

V. CONCLUSION

From the above results it can be concluded that rough based K-means provides a better segmentation result than fuzzy C-means. Hence, even for SVM, in the test dataset images are recognized correctly for less number of images in the training dataset.

REFERENCES

[1] www.healthline.com/health/brain-tumor.
[2] A. Zamre, T. Shah, H. Thadhani, D. Bangar,Detection and classification of brain tumors in MRI images, *IJSET*, vol 2, issue 4, pp. 1-7, 2014.
[3] N. Richard, M. Dojata, C. Garbayvol., Distributed Markovian Segmentation Application to MR brain Scans, *Journal of Pattern Recognition*, vol 40, pp. 3467-3480, 2007.
[4] Y. Zhang, M. Brady, S. Smith, Segmentation of Brain MR Images through hidden Markov random field model and the expectation maximization algorithm, *IEEE Transactions on Medical Imaging*, vol 20, pp. 45-57, 2001.

[5] X. Zang, Y. Wang, J. Yang, Y. Liu, A Novel Method for segmentation of CT Head Images, *International Conference on Medical Image Analysis and Clinical Applications*, vol 4, pp. 717-720, 2010.
[6] M. Milan, L. Sven, P. Damir, A rule based approach to stroke lesion analysis from CT brain Images, *2nd International symposium on Image and Signal Processing and Analysis*, pp. 219-223, 2001.
[7] V.E. Ruthmann, E.M. Jayce, D.E Reo, M.J. Eckaidt, Fully automated segmentation of cerebro spinal fluid in computed tomography,*Psychiatry research: Neuro imaging*, vol 50 , pp. 101-119, 1993.
[8] S. Loncaric and D. Kova Cevic, A method for segmentation of CT head images, *Lecture Notes on Computer Science*, vol 1311, pp. 1388-305, 1993.
[9] A. Halder and O. Dobe, Detection of tumor in brain Mri using fuzzy feature selection and support vector machine, *5th International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1919-1923, 2016.
[10] M. C. Clark, L. O. Hall, D. B. Goldgof, R. Velthuzien, F. R. Murtagh, and M. S. Silbiger, Automatic tumor segmentation using knowledge based techniques, *IEEE Transactions on Medical Imaging*, vol. 17, pp. 187-192, 1998.
[11] A. Lauric, S. Frisken, Soft segmentation of CT brain data, *Technical Report TR-2007-3 Tuffs University*, M A.
[12] R. Ganesan, R. Radhakrishnan, Segmentation of Computed Tomography Brain Images using genetic algorithm, *International Journal of Soft computing*, vol 4, pp. 157-161, 2009.
[13] C. Majos, M. Julia-Sape, J. Alonso, M. Serrallonga, C. Aguilera, J. Juan, J. Gilli, Brain tumor classification by proton MR spectroscopy: Comparison of diagnostic accuracy at short and long TE, *AJNR*, 2004.
[14] D. Selvaraj, R. Dhanasekaran, A Review on Tissue Segmentation and Feature Extraction of MRI Brain images, *International Journal of Computer Science and Engineering Technology*, vol 4, pp. 1313-1332, 2013.
[15] P. Lingras, G. Peters, Rough sets: selected methods and applications in Engineering and management, 2012.
[16] M. Parviainen, Radial Basis Function (RBF) and Support Vector Machines (SVM) networks, unpublished.
[17] http://brain-development.org/ixi-dataset/
[18] A. Halder, A. Pradhan, S. K. Dutta and P. Bhattacharya, Extraction from MRI images using Dynamic Genetic Algorithm based Image Segmentation and Morphological Operation, *International Conference on Communication and Signal Processing*, pp. 1541-1545, 2016.
[19] R. Ahmmed, AS Swakshar, MF Hossain, MA Rafiq, Classification of tumors and it stages in brain MRI using support vector machine and artificial neural network. InElectrical, *International Conference on Computer and Communication Engineering*, pp. 229-234, 2017.