A Classifier to Detect Abnormality in CT Brain Images

Hassan Najadat, Yasser Jaffal, Omar Darwish, Niveen Yasser

Abstract— Medical images are among important data sources available, since these images are usually used by physicians to detect different diseases. Extracting features from brain CT images helps in building a machine classifier that able to classify new brain images without human interference. In this paper, we used a data set of 25 CT brain images with different diagnoses, and built a decision tree classifier that is able to predict general abnormality in human brain. The preprocessing uses the three stages described by Peng et al with modifications. The process of feature extraction was mainly to identify the regions of interest and extract analytical data from those regions. The model was evaluated using hold out method and N-fold evaluation. The results showed that the classifier is able to detect abnormality, even with a small training data set.

Index Terms— **Brain images, Decision tree classifier, feature extraction, Image Mining**

I. Introduction

DATA mining is becoming more important in the medical field. Thousands of available medical records represent an interesting data set, from which important roles can be extracted using data mining techniques.

Medical images are among important data sources available, since these images are usually used by physicians to detect different diseases. A huge number of medical images collected over years, along with related medical diagnosis, present a valuable data set that helps in building a model to classify future cases.

Brain CT images [1] are multi-layered images that provide figures for different levels of brain. CT images are useful in diagnosing various diseases like Atrophic, Hemorrhage, Hematoma, Infract, and Craniotomy.

Extracting features from CT images helps us in building a machine classifier that will be able to classify new brain images without human interference. In this project, we used a data set of 25 CT brain images with different diagnoses, and built a classifier that was able to detect abnormality in brain images.

Manuscript received Dec. 29th, 2010; revised Jan. 16th, 2011.

Hassan Najadat is an assistant professor in Computer Information Systems Department at Jordan University of Science and Technology, P.O.BOX 3030, Irbid, 22110, Jordan. E-mail: najadat@just.edu.jo.

(Corresponding author) Yasser Jaffal, Omar Darweesh, and Niveen Yasser are graduate students in Department of Computer Science, Jordan University of Science and Technology.

In this research, we conducted a 3-phase method. In preprocessing phase, CT images were cleaned, corrected, and normalized. In feature extraction phase, final images from the first phase were processed to extract the most important features in tabular representation. Finally, in model building phase, the extracted data from previous phases were utilized to build the classification model, which is able to detect general abnormality from whole CT image rather than detecting a specific disease from part of the image.

Preprocessing steps are described in section 2, which focuses on image-based preprocessing and data cleaning. In section 3, we describe the methodology used for feature extraction from cleaned images. Section 4 outlines model building process and interesting classes we focused on. Related work is outlined in section 5, and we finally conclude in section 6.

II. RELATED WORK

Y. Li [4] introduced in his paper a new idea of image sequence similarity patterns (ISSP) for medical image database. His patterns were focused for medical images. He came up with new algorithms with the assistance of the domain knowledge in order to find ISSP for similarity retrieval. His experiments proved that the results of similarity retrieval were valid. P. Haiwei et al [5] emphasized that for different brain data, expert of information science have to meet two main challenges: how to process exactly the brain data and how to mine hidden information in the data so they propose some statistical strategies, such as principle component analysis (PCA), independent component analysis (ICA), structure equation model (SEM), dynamic causal model (DCM) and Time-Frequency. Ruan et al [6] introduced a fully automatic three-dimensional classification of brain tissues for Magnetic Resonance (MR) images. Hiroshi [7] offered an algorithm in order to discover rules from functional brain images. His algorithm depended on two phases. In phase one he depended on nonparametric regression. While the second phase focused on the rule extraction from the linear formula, which was gained by the nonparametric regression. His emphasized that the algorithm works well for artificial data.

Megalooikonomou et al [8], suggested data mining techniques that could be employed in order to analyze brain images. The suggested methods concentrated on two categories of brain imaging data: structural and functional.

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

They showed statistical techniques that support the detection of remarkable associations and patterns between brain images and other clinical data. They used a number of applications for these methods, such as the analysis of task-activation, lesion-deficit, and structure morphological variability; the development of probabilistic atlases; and tumor analysis.

III. CT IMAGE PREPROCESSING

CT images are multidimensional in terms of imaging angle. Images are taken from top-down, left-right, and front back perspectives. The data set used in this work had its best quality and most complete images from the front-back perspective, so it was the main data which were used in the steps described in this section. All images were collected from King Abdullah I Educational Hospital in Jordan.

For preprocessing phase, we used a process of three stages described by Peng et al in [2], but with little modifications. Instead of delaying average image generation to the last step [3], we performed it first. This method had the following advantages on the preprocessing: (1) the quality of the final image higher is preserved, because original images with best quality were used to generate the average image, rather than modified images. (2) The time needed for data set preprocessing is reduced, since each instance was represented by one image rather than set of images. This reduced data set size to 1/8 of its original size. Figure 1 shows an example of multiple layers of single CT image (a), and the average image of all layers (b).

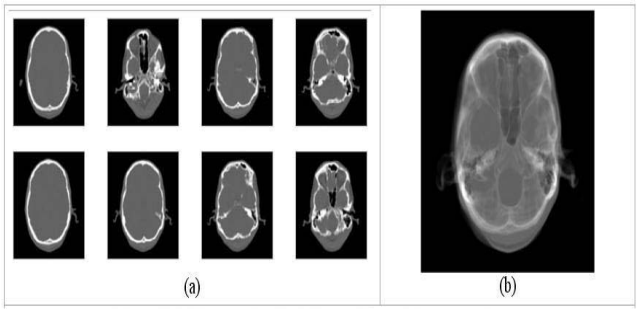

Fig.1. (a) Images of multiple layers. (b) Resulting average image

In our method, we applied preprocessing steps identified in [2] on average images like the one shown in Figure 1-b. After getting average images, correcting lean angle was applied. It was suggested in [2] that lean image correction should be performed after background removal. However, it was practically better for image quality to change the order, because it guaranteed no loss in any part of image when rotated. Figure 2 shows an example of a CT image before and after lean correction.

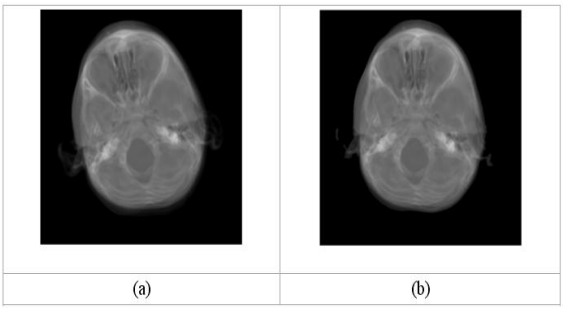

Figure .2. Image lean correction effect. (a) Before correction. (b) After correction

In the following step, background removal was applied. This step was important to standardize the view before normalization. It was also useful for noise removal and size reduction. Figure 3 shows background removal result. Finally, all images are normalized to 256*256 to be ready for feature extraction phase. Normalization importance lies in avoiding any dissimilarity between images that might occur due to differences in skull or brain size. Additionally, normalization make feature extraction process faster.

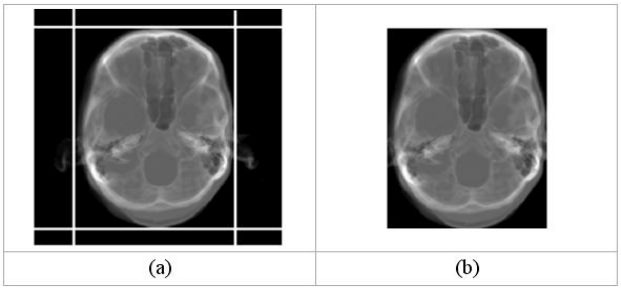

Figure .3. Background region removal[1]

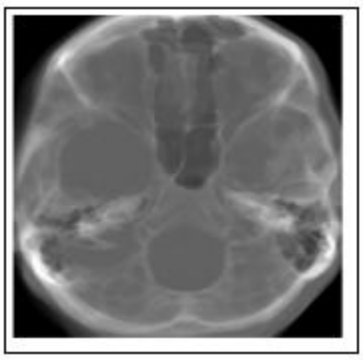

Figure.4. Normalized image

IV. FEATURE EXTRACTION

Feature extraction is the process of transforming the input data into the set of features to be analyzed. The feature extracted should be carefully chosen to be representative for all the relevant information from the input images in order to achieve good result from the next stages of the analysis process.

The process of feature extraction in our work is mainly divided into two steps: The first step is to identify the regions of interest, and the second to extract analytical data from those regions.

The analysis of images to find the regions that is affected by certain diseases is a very important process, since the effect of a certain diseases is concentrated at a specific area, so the identification of this area is critical to get accurate result from the whole classification process. For this purpose we use a MATLAB [11] toolbox that is widely used for this process. The tool box implements multiple algorithms for component analysis that is contributed by brain image specialists.

The process of region identification includes three stages: first Pre-Analysis where multiple Functions used to get the parameters required for the analysis. Then the Analysis stage, here specific function used for running the analysis using the parameter file information, and finally Display

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

functions that is used to display the result file, there is multiple algorithm implemented for each stage that is previously tested and implemented in the tool box [12, 13].

In our project we handle preprocessing as an individual phase as descried in the previous section, and because of the nature of the available data, and the class label that is available for our data set beside the size of the data set we choose to extract the analytical data form the whole brain image area after preprocessing stage.

The analytical data extracted from the images using 18 algorithms implemented in java programming language. Each algorithm extracts one feature or more from the selected area of the image which does not include the whole background.

The following algorithms include: (1) the area algorithm which calculates the number square pixels in the selected area, (2) Mean Gray Value algorithm which calculates the average gray value within the selection by taking the sum of the gray values of all the pixels in the selection divided by the number of pixels, (3) the standard deviation which calculates the standard deviation of the gray values used to generate the mean gray value, (4) the modal gray value algorithm which calculates the most frequently occurring gray value within the selection, (5) min and max gray level algorithm which generates two values: the minimum and the maximum gray values within the selection, (6) centroid algorithm which calculates the center point of the selection by computing the average of x and y coordinates of all of the pixels in the selection which results the brightness-weighted average of the x and y coordinates, and (7) perimeter algorithm which calculates the length of the outside boundary of the selection which produces the smallest rectangle enclosing the selection. The main feature of the rectangle are the coordinates of the upper left corner of the rectangle i.e. BX, BY, width and height, coordinates of the center of the ellipse which are displayed as X and Y if Centroid is checked. The Shape Descriptors (previously Circularity) calculate and display the following shape descriptors:

1) Circ. (circularity): 4π*area/sqr(perimeter). A value of 1.0 indicates a perfect circle. As the value approaches 0.0, it indicates an increasingly elongated shape. Values may not be valid for very small particles.
2) AR (aspect ratio): major axis/minor axis.
3) Round (roundness): 4*area/(π*sqr(major axis)), or the inverse of the aspect ratio.
4) Solidity: area/convex area. Feret's Diameter - The longest distance between any two points along the selection boundary, also known as maximum caliper.
5) Integrated Density algorithm calculates the sum of the values of the pixels in the image or selection. This is equivalent to the product of Area and Mean Gray Value. The Median algorithm output is the median value of the pixels in the image or selection. The Skewness- algorithm provides the third order moment about the mean. The Kurtosis algorithm result is the fourth order moment about the mean.
6) Area Fraction calculates the percentage of pixels in the selection

All results from the previous algorithms are collected as data sheet file, which will be used as an input to the next phase to build the classifier.

V. Building and Evaluating Classifier

Using data extracted from images, we built a classification model based on abnormality. In our context, abnormality is the existence of one of the following diseases in the medical record of each image:

1) Atrophic
2) Hemorrhage
3) Hematoma
4) Infract
5) Craniotomy

We first lined up medical reports (which were available in hard copy only) in a CSV file, which included case ID and positive/negative values for mentioned diseases. Figure 5 shows a sample of the CSV file viewed in Microsoft Excel.

The class attribute was generated by applying logical NOT on the normal attribute, so we set interesting value of 1 in the abnormal instances. The final step in data preprocessing was to combine digitalized medical reports and images attributes file from feature extraction process together based on case ID.

Weka software [9] was applied to define the classification systems. Unpruned C4.5 decision tree [10] classifier was utilized on a final set of 17 instances, which included 4 abnormal instances. Except images width and height, all attributes extracted from feature extraction phase were provided to classifier building algorithm. We evaluated our classifier by two methods: hold out method and using N-fold validation with 2 folds. Confusion matrices for both evaluations are listed in Table 1 and Table 2.

	Normal	Atrophic	Hemorrhage	Hematoma	Infarct	Craniotomy
232051	1	0	0	0	0	0
189494	1	0	0	0	0	0
108829	1	0	0	0	0	0
15046	0	1	0	0	0	0
271374	0	0	1	1	0	0
271693	1	0	0	0	0	0
271624	1	0	0	0	0	0
265109	1	0	0	0	0	0
271760	1	0	0	0	0	0
125409	1	0	0	0	0	0
271124	0	0	0	0	1	0
10191	1	0	0	0	0	0
176762	0	1	0	0	0	0
270813	0	0	0	0	0	1
169083	1	0	0	0	0	0
213854	0	0	0	0	1	0

Figure.5. Image lean correction effect. (a) Before correction. (b) After correction

Table 1: CM for training set evaluation

	Abnormal	Normal	Total
Abnormal	2	2	4
Normal	0	13	13

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 2: CM for 2-fold evaluation

	Abnormal	Normal	Total
Abnormal	1	3	4
Normal	4	9	13

Using data from Table 1 and Table 2, we computed sensitivity, specificity, and accuracy for both evaluations. These measures are listed in Table 3.

Table 3: evaluation measurements

Evaluation Test	Sensitivity	Specifity	Accuracy
Training set	0.5	1	0.88
2-Fold	0.25	0.69	0.59

Test results have shown that our model was able to correctly classify one of the abnormal instances, even the number of instances used to build the model were few. However, this reflects the validity of feature extraction from the entire brain image to detect multiple diseases at once, which was the aim of this research.

VI. Conclusions and Future Work

As shown by classifier evaluation measurements, our method has shown ability to detect abnormality in human brain using average front-back CT image. However, these results could have been better if a larger brain images data set was available. Another problem was the unavailability of header files for brain images in the provided data set. Our focus in the future will be on applying our method on larger data set. It is important that feature extraction phase applied on CT images with header.

However, our modifications on image preprocessing steps have shown enhancements in feature extraction with the absence of header files. This indicates the applicability of our approach on CT images for which no headers are available.

References

[1] G. Simpson, Thoracic Computed Tomography: Principles and Practice. Australian Prescriber, Vol32-4, pp. 105-107, 2009

[2] F. Peng, K. Yuan, S. Feng, and W. Chen. Pre-Processing of CT Brain Images for Content-Based Image Retrieval. International Conference on BioMedical Engineering and Informatics, pp. 208-212, 2008

[3] W. Liu1, F. Peng, S. Feng, J. You, Z. Chen, J. Wu, K. Yuan, and D. Ye. Semantic Feature Extraction for Brain CT Image Clustering Using Nonnegative Matrix Factorization. LNCS 07, pp. 41-48, 2007

[4] Y. Li. Information Mining in Brain Data. Neural Networks and Brain, 2005. ICNN&B '05. International Conference, Volume 2, pp. 1274 – 1278, 2005

[5] P. Haiwei, X. Xie, Z. Wei, and J. Li. Mining Image Sequence Similarity Patterns in Brain Images. Springer Berlin / Heidelberg, 2006

[6] S. Ruan, C. Jaggi, J. H. Xue, M. J. Fadili, and D. Bloyet. Brain Tissue Classification of Magnetic Resonance Images using Partial Volume Modeling. IEEE Trans. Med. Imaging, pp. 1179-1187 , 2000

[7] T. Hiroshi. On Data Mining from Functional Brain Images. SIG-FAI99, pp. 95-98, 1999

[8] V. Megalooikonomou, J. Ford, L. Shen, F. Makedon, and A. Saykin. Data mining in brain imaging. Stat. Methods Med. Res. 9, pp. 359–394, 2000

[9] S. R. Garner. WEKA: The Waikato Environment for Knowledge Analysis. University of Waikato, Hamilton, 1995

[10] R. Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993

[11] MATLAB – The Language of Technical Computing. <http://www.mathworks.com/products/matlab/>, 2009

[12] M. S. Nixon, and A. S. Aguado. Feature Extraction and Image Processing, First edition, 2002

[13] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar. ICA OF FUNCTIONAL MRI DATA AN OVERVIEW, April 2003, Nara, Japan

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)