IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 4, Aug-Sept, 2013
ISSN: 2320 - 8791
www.ijreat.org

IMPROVED IMAGE MINING TECHNIQUE FOR TEXTUAL IMAGES AND MEDICAL IMAGES USING IMAGE FEATURES AND CLASSIFIER

[1]Dr.T.Amitha B.E, M.Tech, Ph.D, [2]Dr.B.Raghu B.E, M.Tech, Ph.D

[1]Professor/ Dept of Computer Science and Engineering
Dhanalakshmi College of Engineering, Chennai-601301, Tamil Nadu, India

[2]Professor/ Dept of Computer Science and Engineering
Sri Ramanujar engineering College, Chennai-601301, Tamil Nadu, India

Abstract

This paper proposes an improved image mining technique for classifying textual images and medical images. Intensity histogram features and GLCM texture features are used to classify the textual images into following types as scene text image, caption text image and document images. In the first stage, the histogram features are extracted from gray scale images to classify document images and in the second stage the GLCM features are extracted from binary images to classify between scene text image and caption text images. In both the stages a decision tree classifier (DTC) is used. For medical image classification Association rule is generated. It classifies the CT scan brain images into three categories namely normal, benign and malign. This method combines low level features from images and high level knowledge from specialist to improve the accuracy and sensitivity of the results.

Key words: DTC, GLCM, Textual image, Weka, Scene text image

1. Introduction

In this paper we propose a method for mining textual images and medical images using combination of features and generating association rules. Textual images can be of many different types: Scene text image, caption text image and document image. Scene text**:** It contains important semantic information such as names of streets, institutes, traffic information, etc. Caption text: It contains news videos annotates information on where and when and who of the happening events. Document text: It contains mix of images and text but the text is not embedded in image. It is important to classify these images so that most relevant images can be used for various applications like "intelligent glasses" for the blind to read sign boards and to annotate news videos based on their content and other areas of machine vision.

In health care centers and hospitals, millions of medical images have been generated daily. Nowadays, physicians are provided with computational techniques for diagnosis process. It has been reported that Brain tumor is one of the major cause of death in human, in which physicians have faced challenging task in feature extraction and decision making. To diagnosis this problem Computerized Tomography (CT) method that uses radiotherapy rays found to be the most reliable for detecting tumors. Due to high volume of CT images, the accuracy of decision making tends to decrease for the physicians. CT scan brain images are among the most difficult medical images to read due to their low contrast and differences in the type of tissues. It increases the demand to improve the automatic digital reading for effective decision making. The proposed method is based on associative classification scheme that has an advantage of selecting only the most relevant features during mining process and obtaining multiple key words when processing a test image.

2. Proposed System Objective

The main objective of the proposed system is to classify the textual images using Decision Tree classifier (DTC) and to improve the accuracy of medical image (CT scan) by generating Association

Textual image: Various low-level image features are used for differentiating the three types of

images: Scene text image (ST), Caption text image (CT) and Document text image (DOC). Various image features such as intensity histogram features, texture features, edge features were investigated for efficiently discriminating the three classes of image. A decision tree classifier is a set of hierarchical rules applied to the input data and are split into groups. Each split (node) is such that the descendant nodes contain more homogeneous data samples. The decision tree classifier is supervised because it relies on training samples to grow.

Medical image: The proposed system of medical image consists of two phases, i) Training phase and ii) Testing phase. Data cleaning and feature extraction are common for both the training set and testing set of brain images. In the training phase, features are extracted from the images and are represented in the form of feature vector and are discretized into intervals and the processed feature vector is merged with the keywords related with the training images. Finally association rule mining produces a pruned set of rules representing the actual classifier. In the test phase the feature vector are submitted to the classifier which makes use of the association rules to generate keywords for diagnosing the test image. These keywords have been used to classify the three categories of CT scan brain images as normal image, benign (tumor without cancerous tissues) image and malignant (tumor with cancerous tissues) image. Advantages: Fast feature extraction

3. Image conversion

There are three types of textual images: scene text image, caption text image and document text image .For textual image there is two types of conversion one is converting the image into gray scale and other is converting the image into binary. When the image is converted into gray scale histogram features are extracted, from the histogram features it is confirmed that whether the image is a document text image or not. If the image is not document text image then the image is converted into binary image and from that image GLCM features are extracted.

4. Extracting features

After converting the textual image into gray scale and binary the histogram features and GLCM features are extracted. From the histogram features of gray scale image, the skewness and mean are deduced. If the skewness is less than 0 and the mean is too high then the image is Document text image (DOC) .If the image is not a DOC then the image is converted into binary image, from the binary image GLCM features (entropy, energy, contrast) are extracted. When the contrast and energy are more for the image then the image is termed as scene text image (ST) otherwise it is a caption text image (CT).

Histogram features are extracted from medical image .From the mean calculated for the histogram features the medical image is classified into three types: Normal, Benign and Malignant.

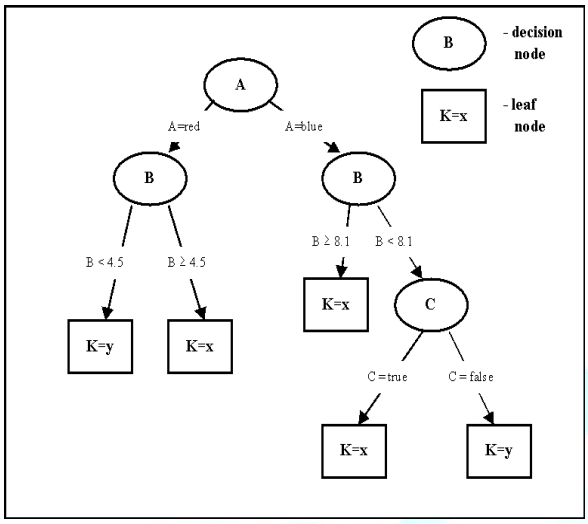
5 .Image classification

A decision Tree is visualized from the values obtained from Scene text image, Caption text image and Document text image and medical images by using a J48 classifier. A threshold curve is also obtained of the scene text image (ST), caption text image (CT) and document text image (DOC).

5.1 Decision Tree Classifier

Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved.

In some applications, the accuracy of a classification or prediction is the only thing that matters. In such situations we do not necessarily care how or why the model works. In other situations, the ability to explain the reason for a decision, is crucial. In marketing one has describe the customer segments to marketing professionals, so that they can utilize this knowledge in launching a successful marketing campaign. This domain expert must recognize and approve this discovered knowledge, and for this we need good descriptions. There are a variety of algorithms for building decision trees that share the desirable quality of interpretability. A well known and frequently used over the years is C4.5 (or improved, but commercial version See5/C5.0).



6. Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite machine learning software written in java, developed at the university of Waikato, New Zealand**.** Weka is free software available under the GNU General public License .WEKA was first implemented in its modern form in 1997. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API, so you can embed WEKA, like any other library, in your own applications to such things as automated server-side data-mining tasks

Three graphical user interfaces in weka are 1. The Explorer (exploratory data analysis) 2. The Experimenter (experimental environment) 3. The Knowledge Flow (new process model interface Weka Functions and Tools: Preprocessing Filters**,** Attribute selection, Visualization.

7. Feature Extraction

7.1. Intensity histogram features:

Document text image (DOC) has two special features which distinguish themselves from ST and CT images. The background and the text color are mostly uniform when compared to other types of images. There are more text characters in a document image compared to other types of images.

The intensity of the pixels helps to differentiate the text pixels from the background pixels especially in the case of a DOC image. The intensity histogram was created by quantizing the grayscale intensity values in the range 0-255 and then making a bin histogram for these values. The mean variance, skewness, etc. are derived from the intensity histogram features. skewness property represents the frequency of the pixels of a certain intensity, when one particular pixel intensity dominates the others intensity, when one particular pixel intensity dominates the others, the intensity histogram of such an image must be positively or negatively skewed .from our experiment, we have found that the histogram of DOC images was almost always negatively higher for DOC image than either ST or CT images

7.2 GLCM features

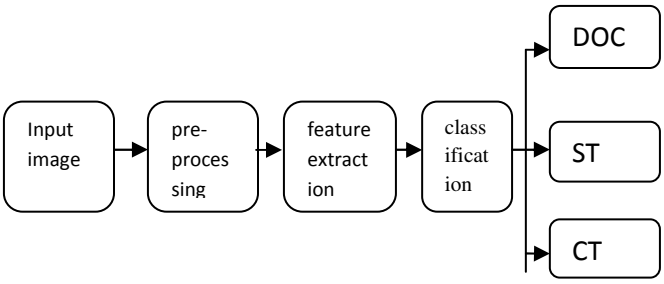
The Grey Level Co-occurrence Matrix, GLCM (also called the Grey Tone Spatial Dependency Matrix). The GLCM described here is used for a series of "second order" texture calculations.

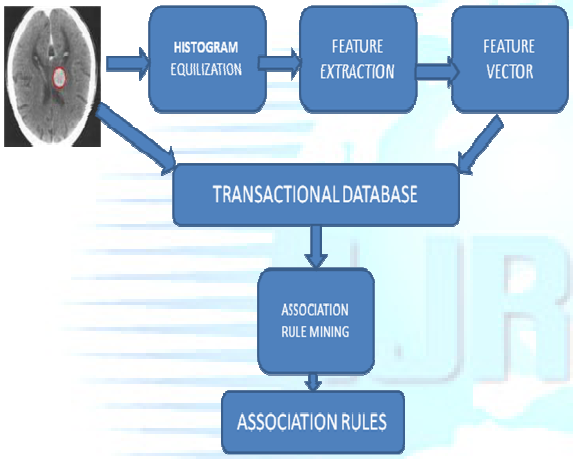
First order texture measures are statistics calculated from the original image values, like variance, and do not consider pixel neighbor relationships.

Second order measures consider the relationship between groups of two (usually neighbouring) pixels in the original image.

GLCM texture considers the relation between two pixels at a time, called the reference and the neighbor pixel. In the illustration below, the neighbor pixel is chosen to be the one to the east (right) of each reference pixel. This can also be expressed as a (1,0) relation: 1 pixel in the x direction, 0 pixels in the y direction.

8. System Architecture





9. Conclusion

We have provided a method of mining images using image classification which uses a combination of low level features to classify images into one of three types: Scene text image, Caption text image, Document text image. This is the first work and does not consider such asset of features and does not depend on the language of the text in the image.

An improved image mining technique for brain tumor classification using pruned association rule has been developed and the performance is evaluated. The developed brain tumor classification system is expected to provide valuable diagnosis techniques for the physicians.

10. Future Enhancement

In future, we plan to consider those images in which there is a combination of scene and caption text within a single image. We also plan to investigate many other low-level features such as those based on the GLRM (Gray Level Run Length Matrix), edge strength and edge intensity features for robust classification. Though at this juncture we have stopped with classification of images, we plan to extend it into the domain of Content-Based image Retrieval (CBIR).

We also plan to improve the accuracy and sensitivity of the brain images to a much greater extent. In future we have considered merging the preprocessed CT scan brain images with the angle representation.

11. References

[1]. S.Chitrakala, P.Shamini, Dr.D.manjula "Multi-class Enhanced Image Mining of Heterogeneous Textual Images Using Multiple Image Features"2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India 6-7 March 2009

[2]. P.Rajendran, M.Madheswaran "An Improved Image Mining Technique For Brain Tumour Classification Using Efficient classifier" (IJCSIS) International Journal of Computer science and Information Security,Vol 6,No. 3,2009

[3]. Wei-HaoLin,Rong Jin, Alexander Hauptmann,"Meta-classification of Multimedia Classifiers", International Workshop on knowledge discovery in multimedia and complex data ,2002

[4]. L.Breiman, J. H. Friedman, R. A.Olshen, and C.J. Stone, Classification and Regression Trees. New York: Chapman & Hall, 1984.

[5]. S.RasoulSafavian and David Landgrebe, "A survey of decision tree classifier methodology", IEEE Transactions on system, Man, and Cybernetics, Vol. 21,No. 3,MayJune 1991.