Detecting Brain Disease Based on Pixel's Clustering [1]

Pan Haiwei, Jianzhong Li, Zhang Wei

Department of Computer Science, Harbin Institute of Technology, Harbin, P. R. China
heaven_008@sina.com, heaven_008@sohu.com

Abstract

The high incidence of brain disease, especially brain tumor, has increased significantly in recent years. It is becoming more and more concernful to discover knowledge through mining medical brain image to aid doctors' diagnosis. Image mining is the important branch of data mining. It is more than just an extension of data mining to image domain but an interdisciplinary endeavor. Image Clustering is a basilic parts of image mining. In this paper, we will use clustering method on the most fundamental elements of image – pixels to detect the possible space occupying (SO) and utilize divide and conquer method to generate clusters included in each image of all objects. Then we determine the real SO according to the property of image sequence that comes from the Merge process. We also analyze the guidance effect of the domain knowledge in data mining.

Keywords:

Data mining; image mining; clustering

1 Introduction

Advances in image acquisition and storage technology have led to tremendous growth in very large and detailed image databases [1]. A vast amount of image data is generated in our daily life and each field, such as medical image (CT images, ECT images and MR images etc), satellite images and all kinds of digital photographs. These images involve a great number of useful and implicit information that is difficult for users to discover.

Image mining can automatically discover these implicit information and patterns from the high volume of images and is rapidly gaining attention in the field of data mining. Image mining is more than just an extension of data mining to image domain. It is an interdisciplinary endeavor that draws upon computer vision, image processing, image retrieval, machine learning, artificial intelligence, database and data mining, etc. While some of individual fields in themselves may be quite matured, image mining, to date, is just a growing research focus and is still at an experimental stage. Research in image mining can be broadly classified to two main directions: (1) domain-specific applications; (2) general applications [2]. The focus in the first direction is to extract the most relevant image features into a form suitable for data mining [4,8,9] and the latter is to generate image patterns that may be helpful in understanding of the interaction between high level human perceptions of image and low level image features [1,10]. Data mining in medical images belongs to the first direction.

Brain tissue is human's advanced nerve center, so its function is particularly important. The disease affecting the brain has received much attention in the domain of medicine. Especially during these years, the incidence of brain disease (especially brain tumor) has increased significantly and the quality of human's living even their lives has been endangered greatly. Therefore, the early diagnosis of brain diseases is becoming more and more crucial and is directly working on patients' treatment. That is why data mining in medical image for assisting medical staff is so significant. Furthermore, it is a greater challenge because of referring to the special domain.

Computerized Tomography (CT) is one of the most important techniques that are used to diagnose by medical doctors. Brain CT scan of each patient (as an object) is an image sequence in which each one is an image of a layer every a few millimeters from calvaria. There exists a certain spatial relationship between images in a sequence. We will try to discover knowledge from this kind of image dataset by means of data mining technique.

At present, the main work of data mining on medical images [3,4,5,6,7] has two characteristics: (1) research content is the images in the medical image database, not the objects with medical images. For example, it is possible to classify images of the same object into the different class because they always have different morphological SO. This determines the type of knowledge that will be mined; (2) research method is to extract features from images to form feature attributes and use data mining on these attributes, not to consider the fundamental element – pixel's significance. In fact, medical doctors make a diagnosis mainly according to medical knowledge and the tone of pixels. Also, these work paid little attention to the guidance effect of domain knowledge to data mining.

This paper presents a new method for using image mining to detect the brain diseases from brain CT images of one object. The novelty includes two directions. The first is to make use of medical domain knowledge efficiently to guide data mining. The second is that we utilize two different clustering algorithms on pixels to judge the possible brain diseases that are called space occupying pathology (SO) by doctors, as shown in figure 1(c). In this work, pixels clusters of each image in objects are generated by using divide and conquer

[1] This work is supported partially by National Science Funds of China, national 973 plan of China (G1999032704) and national 863 plan of China (2001AA415410).

methods. Detection is completed according to the properties of each object's image sequence (continuity, discontinuity and length) that come from the Merge process. This paper also illustrates the influence of the medical domain knowledge.

The rest of the paper is organized as follows: section 2 presents the data collection. Pre-processing is presented in section 3 and Detecting SO Based on Pixel's Clustering is introduced in section 4. Conclusions and future research are presented in section 5.

2 Data Collection

The dataset utilized in our experiments was real data from hospital. The main reason that we study on real brain CT images instead of any simulative data is to avoid reducing the accuracy of the detection and the reliability of the discovered knowledge. To have access to real medical images is a very difficult undertaking due to legally privacy issues and management of hospital. But with some specialists' help and support, we got 103 pieces of precious data, which included 11 normal objects' CT scans and 92 abnormal objects data including CT scan and clinical data. As brain CT scan of each object comprises several or more than ten images and there is a certain spatial relationship between these images. This greatly increases the quantity and complexity of data that is to be processed. These collected data is gained randomly. That is, the implicit knowledge hidden in these images is unknown to specialists and us. Since the images we got were original CT scans, we should digitize them to no loss, no compression and 256 gray scale images through special medical scanner.

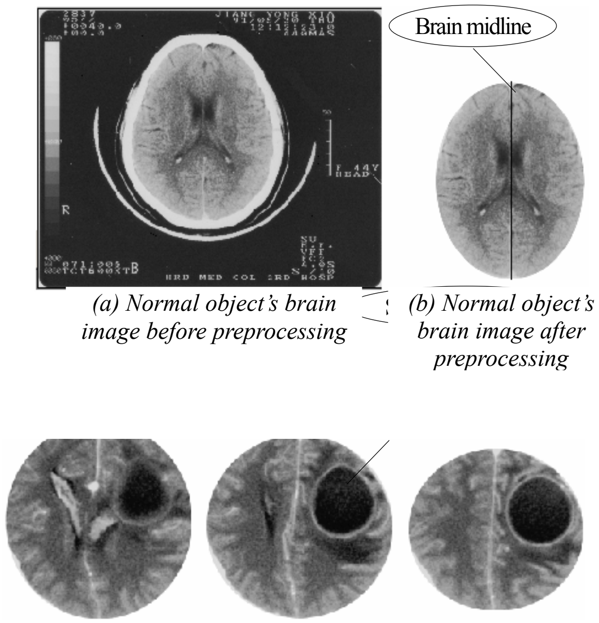
3 Pre-processing

In preprocessing, our work uses domain knowledge effectively to remove the noisy data. A brain CT image mainly consists of three parts: noisy data, skull and cerebrum. The noisy data includes the black background and some additive information on it, such as CT identification, date and patient's name etc. These information is not only helpless but revealing patient's privacy. We are only interested in cerebrum. So we use domain knowledge (for short DK1) and cropping technique of image processing to gain it.

DK1 --- human's brain skull has the highest density and surrounds the cerebrum.

That is, the skull is a cricoid area in the image with the whitest pixels that separate cerebrum from the noisy data, see figure 1(a). It becomes easy to remove the noisy data by using cropping technique and keep the interesting region with the guidance of DK1. Here, we define the effect factor of DK1 to the preprocessing as:

Factor = (the number of pixels in the images cropped with guidance of DK1) / (the number of pixels in the images without cropping)



(a) Normal object's brain image before preprocessing

(b) Normal object's brain image after preprocessing

(c) Abnormal object's brain image sequence after preprocessing, further from calvaria from the first to the last

Figure 1

Obviously, the less the factor is, the better the guidance of DK1 will be.

For the brain images, doctors mostly concern the information about gray level and location of the pixels. Therefore, we keep the original pixels of the interesting regions of the images in the preprocessing, avoiding the binarization of the images to lose the pixels' information and the extraction of the features to ignore the pixels' significant.

After image preprocessing, all the objects are formed as follows, see table 1. Each object has a unique identification (ID) and its image part (IM) is a preprocessed image sequence where every code is composed of id, a dot and the image sequence number, and the arrow represents the spatial relationship.

Table 1

ID	IM
001	001.01→001.02→001.03→…→001.07
002	002.01→002.02→002.03→…→002.13
…	…
n	n.01→n.02→n.03→…→n.10

4 Detecting SO Based on Pixel's Clustering

The following domain knowledge (for short DK2) is used to direct clustering algorithm.

(1) The normal persons have nearly the same brain

structure that is evident to be lateral symmetry. That is, the distribution of density in the left hemisphere of the brain is almost identical with the right, see figure 1 (b). If there is SO in either side, its density will change and destroy the symmetry;

(2) If one object has SO, it is more possible that this SO will be shown in some continuous images, see figure 1 (c).

4.1 Basic Definitions

Definition 1. Let $S=\{S_i \mid i=1\ldots m\}$ be object set; Let $S_i=\{IM_{i1}, IM_{i2}, \ldots, IM_{in}\}$ be object or ordered image set, where

(1) $IM_{i1}$and IM_{in} are the nearest and farthest image from calvaria;

(2) For any $IM_{i1}, IM_{i2}, \ldots, IM_{ip}$, IM_{ip} must be the farthest image from calvaria;

For any preprocessed image IM_p, it is halved by brain midline (shown as figure 1(b)) and is composed of two parts: $IM_p(L)$ and $IM_p(R)$ present the left and right hemisphere image respectively.

Definition 2. For any IM_p and IM_j, they are adjacent if

(1) $IM_p \in S_i$ and $IM_j \in S_i$;

(2) p=j+1 or p=j-1；

Definition 3. Pixel set of IM_p is defined as $P=\{p_i|p_i$ is the pixel with coordinate (x_i, y_i) in the image $IM_p\}$, P(L) and P(R) are pixel set of $IM_p(L)$ and $IM_p(R)$ respectively.

According to the symmetry of the brain structure in DK2, we assume that the number of pixels in $IM_p(L)$ and $IM_p(R)$ are equal. That is, $|P(L)|=|P(R)| = |P|/2$.

Definition 4. For any $p_{li} \in P(L)$, $p_{rj} \in P(R)$, they are symmetric pixel if the line between p_{li} and p_{rj} is halved vertically by brain midline. They are denoted as p_{li} and p_{ri} below.

Definition 5. We partition all pixels in pixel set P into m blocks. Pixels in the same block have the same grey level and pixels in the different blocks have the different grey level. Let $G(P)=\{g_1, g_2, \ldots, g_m\}$ be P's grey-scale (GS) set if G(P) is an ascending sort set of g_1', g_2', …, g_m' and g_i' is grey level of pixels in the i^{th} block, where g_i (i=1,…,m) is the i^{th} GS, g_1' and g_m' are P's minimum and maximum GS respectively. The GS of pixel p_i is denoted as $g(p_i)$.

Definition 6. we call $g_{mean}(P)$ the mean GS if

$$g_{mean}(P)= \sum_{i=1}^{|P|} g(p_i)/|P|.$$

Definition 7. For any P and distance function $DisA=|g_k - g_{mean}(P)|$, mid-value GS is a middle value in the GS set that minimizes DisA. Mid-value GS set is a set of mid-value GS.

Theorem Mid-value GS set includes not more than two values.

Proof The value of DisA has two possibilities:

(1) If $g_k - g_{mean}(P) <0$, then $DisA= g_{mean}(P) - g_k$;

(2) If $g_k - g_{mean}(P) \geqslant 0$, then $DisA= g_k - g_{mean}(P)$;

In the first case, if there exist more than two GS that make the value of DisA be a certain minimum γ : g_1', g_2', … , g_k'(k>2), then they must satisfy the equation g_1'=g_2'= … =g_k'. This doesn't agree with the definition of GS set. Therefore, it is only one GS that minimizes DisA in the first case, denoted as g_{mida}.

Case 2 is proved similarly and thus it is also only one GS that minimizes DisA, denoted as g_{midb}.

If $g_{mean}(P) – g_{mida} = g_{midb} - g_{mean}(P)$, then mid-value set includes two elements. Otherwise, it only includes one of g_{mida} and g_{midb} that minimizes DisA. ■

Definition 8. For any P, if

(1) Mid-value GS set includes one element g_{mid}, and g_s is the minimum value between g_{mean} and g_{mid};

(2) Mid GS set includes two elements, g_s is the minimum value between these two values;

g_s is called Benchmark GS and another one is denoted as g_s'.

Definition 9. For pixel set P, let

$g^{(l)}=\{g_i \mid g_1 \leqslant g_i \leqslant g_1+|g_1-g_s|/2\}$ be low bounded GS;

$g^{(h)}=\{g_i \mid g_m-|g_m-g_s'|/2 \leqslant g_i \leqslant g_m\}$ be high bounded GS;

$g^{(b)}= g^{(l)} \cup g^{(h)}$ be bounded GS;

Definition 10. For pixel set P, let

$P^{(l)}=\{p_i \mid g(p_i) \in g^{(l)}\}$ be low bounded pixel set;

$P^{(h)}=\{p_i \mid g(p_i) \in g^{(h)}\}$ be high bounded pixel set;

$P^{(b)}= P^{(l)} \cup P^{(h)}$ be bounded pixel set and pixels in $P^{(b)}$ are bounded pixels.

Definition 11. $\triangle g(P)$ is IM_p's difference set if for any symmetrical pixel p_{li} and p_{ri} in P, $\triangle g(P)=\{ \triangle g_i| \triangle g_i=g(p_{li})-g(p_{ri}), i=1,2,\ldots,|p|/2 \}$.

Definition 12. For any image sequence $<IM_{ij}, \ldots, IM_{ik}>$, if

(1) only the first and last image IM_{ij} and IM_{ik} have one adjacent IM;

(2) other IM_{ip} (if existed) has two adjacent IM;

we called that it has the property of continuity.

Definition 13. Image sequence $<IM_{ij}, \ldots, IM_{ik}>$ has the property of discontinuity if it can't satisfy the property of continuity.

Definition 14. For any pixel p_i and a certain integer ε , the assemble of pixels is called ε -adjacent area (ε -AA) if distance between p_i and pixels in the assemble is not more than ε .

Definition 15. p_i is core pixel (c-pixel) if its ε -AA involves no less than MP pixels that satisfy some conditions.

Definition 16. p_i is immediate density reachable from p_j if p_i is in the ε -AA of p_j and p_j is c-pixel.

Definition 17. p_1 and p_k are density reachable if for some pixels p_1, p_2, …, p_k, any pixel p_{i+1} is immediate density reachable from p_i.

Definition 18. p_i and p_j are density connective if there exists pixel p_k that is density reachable from not only p_i but also p_j.

4.2 Clustering Algorithm with the Guidance of DK2

It is very crucial step for medical doctors to determine whether there is a space occupying or not in the brain images. In this paper, we use clustering method on the pixels of images to detect the possible SO. Firstly, for any IM_p, we compute to get the IM_p's difference set $\triangle g(P)$, then sort the absolute value of each element in $\triangle g(P)$ to yield the set $\triangle g'(P)=\{|\triangle g_i| \; |\triangle g_i \in \triangle g(P)$ and for any $|\triangle g_i|$, it must be maximal in the former i elements$\}$. Each element of $\triangle g'(P)$ is regarded as an atomic cluster and hierarchical clustering from bottom to top will not stop in the light of the following similarity function of difference between pixels' GS until the number of clusters is equal to a specified value k.

similarity(C_i,C_j) = min$|T_i-T_j|$, where T_i and T_j are mean value of all $|\triangle g_i|$ in cluster C_i and C_j respectively.

The algorithm is as follows:

Clustering algorithm I:

Input: the set $\triangle g'(P)$ and the number of clusters k

Output: k clusters that satisfy the similarity function similarity(C_i,C_j)

```
1.  Each element of $\triangle g'(P)$ is regarded as an atomic cluster and compute $|T_i-T_j|$ of the adjacent clusters;
2.  Clustering in terms of similarity$(C_i,C_j)$;
3.  While (the number of clusters is not equal to k) {
4.      Compute $|T_i-T_j|$ of the adjacent clusters;
5.      Clustering in terms of similarity$(C_i,C_j)$;}
```

According to "If there is SO in either side, its density will change and destroy the symmetry" in DK2, we can deduce that if there exists SO in the image, then pixels' GS of SO should change and the value of the elements corresponding to these pixels in $\triangle g'(P)$ will be much greater than zero. Otherwise, the value of these corresponding elements in $\triangle g'(P)$ will approximate zero. Therefore, we set the number of clusters to 2 as the termination condition. The first step of Clustering algorithm I scans $|P|/2$ elements in the set $\triangle g'(P)$ one time and time complexity is $O(|P|/2)$. The second step is to select the minimum and time complexity is $O(|P|/2)$, too. In the third step, the loop times is related to the speed of clustering. To the worst, only two clusters is clustered to a bigger cluster in each loop. The number of time is $|P|/2-3$ and time complexity of the 4th and 5th step is $O(|P|/2-i)$, where i is the ith loop. Accordingly, the 3rd and 5th step for the worst condition need $(n-3)(n+2)/2$ operations and time complexity is $O(n^2)$.

According to the symmetry of the brain structure, we single out the cluster with the greater $|\triangle g_i|$ from clustering algorithm I (denoted as high difference cluster) to be the main study objects of the next step. It means that data size to be processed may be reduced. For Each $\triangle g_i$, there are two corresponding symmetric pixels p_{li} and p_{ri}. All symmetric pixels $g(p_{li})$ and $g(p_{ri})$ in the high difference cluster are judged to be whether bounded GS or not, then bounded pixel set is generated.

Next, the based-on density clustering method is utilized to re-cluster these bounded pixel set and determine the location and size of SO in each brain image. The algorithm is as follows:

Clustering algorithm II:

Input: all bounded pixel set, ε and MP

Output: k clusters

```
1.  Assume that the pixel count of any bounded pixel set is bn and examine ε -AA of these bn pixels;
2.  If ( ε -AA of $p_i$ involves more than MP bounded pixels)
3.  Then mark $p_i$ as the c-pixel;
4.  While (all c-pixels) {
5.  Clustering all density reachable pixels; }
```

For ε and MP in clustering algorithm II, we specify their value through learning on the normal object's brain images. The learning process is as follows: (1) run the first clustering on the normal object's IM_p and achieve the bounded pixel set; (2) count (not clustering) on the bounded pixel sets which, in fact, are noisy data but not SO, and compute the maximum of the radius and the count of all bounded pixel set which are the greatest lower bound of ε and MP. Time complexity of this algorithm is O(nlogn). The k clusters generated from this algorithm are k possible SO.

4.3 Detecting Space Occupying

For any S_i in S, we use divide and conquer algorithm to generate the entire possible SO in each IM_{ip} and then merge the results of all the IM in one object. According to DK2, we analyze the continuity and discontinuity of the possible SO of $IM_{ip}(R)$ and $IM_{ip}(L)$ in S_i and thus those satisfying the property of continuity are real SO.

5 Conclusion and Future Research

The high incidence of brain disease, especially brain tumor, has increased significantly in recent years. It is becoming more and more concernful to discover knowledge through mining medical brain image to aid doctors' diagnosis. It is very crucial step for medical doctors to determine whether there is a space occupying or not in the brain images. This paper uses two clustering algorithm on the most fundamental elements of image – pixels to detect the possible space occupying (SO), then utilizes divide and conquer method to determine the real SO. We also analyze the guidance effect of the domain knowledge in data mining.

It is the first step to determine whether there is SO or not. Further contact with medical knowledge will make us engage in studying the classification and retrieval methods in medical images. Also, we will combine the clinical data with images to study new methods to enhance the accuracy of classifying and retrieving.

References

[1] Zaiane, O.R. et al. (1998). Mining MultiMedia Data. CASCON: Meeting of Minds.

[2] WYNNE HSU, MONG LI LEE, JI ZHANG. Image Mining: Trends and Developments. Journal of Intelligent Information Systems, 19:1, 7–23, 2002.

[3] Vasileios Megalooikonomou, Christos Davatzikos, Edward H. Herskovits. Mining Lesion-Deficit Associations in a Brain Image Database. KDD-99 San Diego CA USA.

[4] Wynne Hsu, Mong Li Lee, Kheng Guan Goh. Image Mining in IRIS: Integrated Retinal Information System. Proceedings of the ACM SIGMOD, May 2000, Dellas, Texas, U.S.A., pp. 593.

[5] Y. Liu, F. Dellaert, W.E. Rothfus, A. Moore, J. Schneider, and T. Kanade. Classification-Driven Pathological Neuroimage Retrieval Using Statistical Asymmetry Measures. Proceedings of the Medical Imaging Computing and Computer Assisted Intervention Conference (MICCAI 2001), Utrecht, The Netherlands, October, 2001.

[6] Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman. Application of Data Mining Techniques for Medical Image Classification. Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001).

[7] Osmar R. Zaiane, Maria-Luiza Antonie, Alexandru Coman. Mammography Classification by an Association Rule-based Classifier. Proceedings of the Third International Workshop on Multimedia Data Mining (MDM/KDD'2002).

[8] Fayyad, U.M., Djorgovski, S.G., and Weir, N. (1996). Automating the Analysis and Cataloging of Sky Surveys. Advances in Knowledge Discovery and Data Mining, 471–493.

[9] Kitamoto, A. (2001). Data Mining forTyphoon Image Collection. In Second International Workshop on Multimedia Data Mining (MDM/KDD’2001).

[10] Ordonez, C. and Omiecinski, E. (1999). Discovering Association Rules Based on Image Content. In IEEE Advances in Digital Libraries Conference.

[11] Burl, MC et al. Mining For Image Content. In systems, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis (1999).

+