Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Computer Aided Detection of SARS Based on Radiographs Data Mining

Xie Xuanyang, Gong Yuchang, Wan Shouhong, Li Xi
Department of Computer Science & Technology
University of Science and Technology of China Hefei, Anhui 230022, CHINA

Abstract - **This paper introduces our work on how to use image mining techniques to detect SARS, the Severe Acute Respiratory Syndrome, automatically as the prototype of computer aided detection/diagnosis (CAD) system. Data used in this paper are digitalized PA(Posterior Anterior) X-ray images stored in the real-life Picture Archiving and Communication System (PACS) of the 2nd Affiliation Hospital of Guangzhou Medical College. Association rule mining was applied first but results showed there was no significant difference between the locations of the lesions or infiltrate. Classification based on image textures was performed. A sample set contains both the pneumonia and SARS X-ray images was built in the first place. After modeling each sample by a feature vector, the sample set was partitioned to match the detection purpose: classification. Three methods were used: C4.5, neural network(NN) and CART. Final result shows that 70.94% SARS cases can be detected by CART. Data preparation, segmentation, feature extraction and data mining steps, with corresponding techniques are included in this paper. ROC charts and confusion matrix by all three methods are given and analyzed.**

I. INTRODUCTION

Computer aided detection/diagnosis (CAD), which is used by physicians to detect and diagnose lesions, has been successfully applied in solving many problems [1,2]. SARS, also named as 'Atypical Pneumonia' in China, was first found in Guangdong, China, 2002. By July 31, 2003, 5327 patients had been infected with it in China, accounts to 65.6% of all the cases reported in the world [4]. There is a pressing necessity of developing a computer aided detection of SARS system, especially for the countries or regions that have no experience in dealing with this disease.

This paper introduces our work on how to use image data mining techniques to detect SARS cases from typical pneumonia cases. As noted in [5], the X-ray radiograph plays important role in diagnosing whether a case belongs to SARS. Typical image mining process is used, which include data preparation, segmentation, feature extraction and mining. All these steps are discussed in following sections. Experiment results of all steps are included too.

II. MATERIALS USED AND DATA PREPARATION

Posterior anterior (PA) X-ray chest radiographs were used. Data set includes 818 images of about 2048x2048 in size, 14 bits in grey level, which include 234 images for SARS patients and 584 for 'typical' pneumonia cases. Default window level and window width are used to scale 14-bit grey level depth to 8-bit grey level.

All data are stored in PACS of the 2nd Affiliation Hospital of Guangzhou Medical College, which was developed by a cooperate team of our laboratory. Various modalities of data were stored in PACS, such as digital X-ray (DX), computed tomography (CT), magnetic resonance (MR),etc. Only DX images of chest radiographs are needed. The selection of DX image can be easily done by query each image's modality because all images are compatible with DICOM, an industrial standard for medical imaging. A good survey in [3] analyzed lots of papers to show that almost all computer aided diagnosis treated only PA images. Then the task of data preparation remains to filter out lateral images. Simple strategy can be applied [9]. Our experiment results show that the accuracy can reach 99.12%.

III. SEGMENTATION THE REGION OF INTEREST

Segmentation is the process of selecting region of interesting (ROI) parts of an image automatically. The purpose of this step is to outline the two lung fields. Simple grey-level based method cannot fulfill this task. Active

shape model [6] and its multi-resolution version: MRASM[8] are adopted because they are robust enough to accommodate the complex image resulted from overlapping organs.

MRASM uses a set of points, called a *shape*, to represent the ROI. A shape can be represented by a vector $\mathbf{x}$ in $\mathbb{R}^{2n}$, as in (1), where n is the number of landmark points.

$$\mathbf{x} = \{x_1, y_1, x_2, y_2, \ldots, x_n, y_n\}^T . \quad (1)$$

MRASM uses a training set $\{\mathbf{x}^i\}$ to capture the mean shape $\bar{\mathbf{x}}$ and variances of each landmark points. After training, the mean shape $\bar{\mathbf{x}}$ is used as a template to match new images by adjusting parameters heuristically. Gaussian image pyramid was built to speedup the search. The search procedure starts at the root of the tree, which is small and coarse, and gets the current shape match the ROI. After that, the search goes down one level to refine the result.

During our experiment, we found that a good initial position and scale parameter for $\bar{\mathbf{x}}$ can make the iteration converge more rapidly. This is easily to be understood: less time is needed if the target is nearby. We introduce the concept *window of lung fields* to initialize these parameters [9], which is a four-element vector defined as the bounding box of lung fields:

$$\mathbf{w} = \{row_1, row_2, col_1, col_2\}^T \quad (2)$$

, where the four elements represent upper/lower and left/right bounds of $\mathbf{w}$. The initial position $(dX,dY)^T$ and scale parameter s can be derived easily and other applications of *window of lung fields* can be found in [8]. Fig.1 shows that only two or three steps are needed for the convergence if the window of lung fields is applied. 561 X-ray chest radiographs were used in our experiment. Results showed only a few iterations are needed in more than 90% cases.

(a) First 5 Iteration of MRASM

(b) First 5 Iteration of ASM with window of lung fields used

Fig.1. Comparison of iteration between whether the window is used or not

IV. Feature Extraction and Feature Vectors

According to the common stand used by radiologists, three zones are divided vertically (inner, middle, outer) and another three are divided horizontally (upper, middle, lower), as demonstrated in Fig. 2. These blocks were also used to build association rules, but results (in the next section) showed that there are no obvious differences between locations of lesions.

Features were extracted from each one of the 18 blocks. These features include: mean, standard deviation, skewness and kurtosis grey level to describe the average data distribution of each block; energy, entropy, correlation, inertia and local contrast generated from the co-occurrence matrix of five directions, say $0, \pi/4, \pi/2, 3\pi/4$, to describe the texture of each block. Thus for each block, 24 features are extracted. So for each image, there are 24x18=432 features generated. Finally, another three features are added to represent the overall condition of lung fields, see Fig. 3.

1. Cardiothoracic ratio (CTR): A cardiothoracic ratio of more than 50% is considered abnormal. For clarity, we draw the two diameters in different horizontal levels in Fig. 4, then CTR can be simply calculated as $CTR = L1/L2$;
2. Costophrenic angles(CPA): a obliterated angle reminds abnormality. Both left and right CPA , $\alpha 1$ and $\alpha 2$, are calculated.

In summary, each image is represented by a feature vector in $\mathbb{R}^{435}$: $\mathbf{F} = \{\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^{18}, CTR, \alpha_1, \alpha_2\}$, where $\mathbf{f}^i$ denotes the feature vector of the ith block.

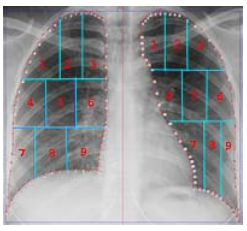

Fig.2. Dividing lung

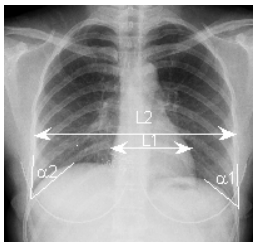

Fig.3 .CTR and CPA

V. Data Mining and Experiments Results

Because our 'simplicity first principle', association rule mining was chosen at first. We want to find whether there exist differences of lesions' locations between SARS and

pneumonia cases. One radiologist helped us to delineate which of the 18 blocks are considered abnormal. An image set contains 75 SARS and 125 pneumonia cases, was presented to the expert. Let R1~R9 denote the 9 blocks of right lung fields, L1~L9 denote the left ones. Then the expert just tick off each block to indicate it is abnormal. For direct comparison of the two classes, count of each position is normalized in percentage. Fig.4. shows that there seems no significant difference of the locations of abnormal between SARS and pneumonia. *t*-test ($p_{value} \to 1$)confirms this observation. Association rule mining was performed on the tick matrix resulted from the expert. TABLE I gives the results. The rules are sorted ascending by confidence (CON), support (SUPT) and lift. Then the top 10 rules are given in TABLE I. Many rules are identical. So this again confirms that association rule cannot be used to classify SARS and pneumonia. The classifier based on the feature vector was built and results are given below.

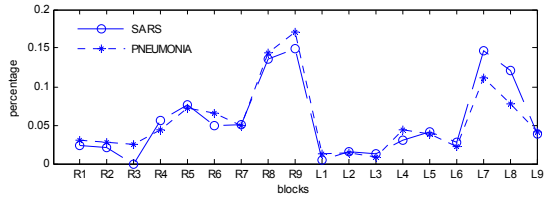

Fig.4. Occurrence frequency of each block

TABLE I ASSOCIATION RULES MINING RESULTS

SARS				PNEUMONIA			
CON	SUPT	LIFT	RULE	CON	SUPT	LIFT	RULE
100	64.38	1.28	L8 > L7	100	40.00	1.74	L8 > L7
100	53.42	1.28	L8 & R9 > L7	100	36.00	1.74	L8 & R9 > L7
100	49.32	1.28	L8 & R8 > L7	100	33.60	1.74	L8 & R8 > L7
100	45.21	1.28	L8 & R9 & R8 > L7	100	32.00	1.74	L8 & R9 & R8 > L7
100	27.40	1.38	R7 > R8	100	22.40	2.66	R4 > R5
100	24.66	1.38	R9 & R7 > R8	100	20.80	1.74	L8 & R5 > L7
100	24.66	1.28	L8 & R5 > L7	100	20.00	2.50	L9 & L7 > L8
100	23.29	1.28	L8 & R8 & R5 > L7	100	20.00	1.74	L9 & L8 > L7
100	21.92	1.28	L8 & R9 & R5 > L7	100	20.00	1.74	L8 & R9 & R5 > L7
100	20.55	2.03	L7 & R7 > L8 & R8	100	20.00	1.74	L8 & R8 & R5 > L7

To make the classification results more credible, the sample set was partitioned into training, validation and testing sets in the ratio of 6:1:3 by simple stratified partition. Three mining techniques were used to build the classifier:

- Decision tree: C4.5 with Gini reduction as the splitting criteria is used. Fig.5 gives the success rate versus the number of leaves. It can be noticed that a 6-leaf tree can reach about 75% success rate. As the tree grows larger, say with 18 leaves, the success rate of validation will decrease.

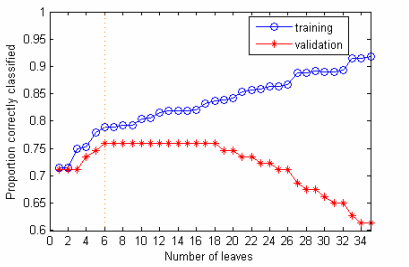

Fig.5. Decision tree building

- Standard back propagation multilayer perceptron neural network(Fig.6): After 42 iterations. The success rate is about 73%, a good lift compared to decision tree.

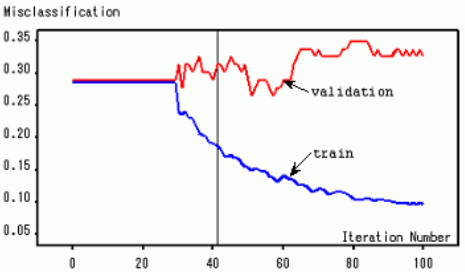

Fig.6. Neural network

- CART: classification and regression tree, a robust and advanced data analysis algorithm. Ten fold cross validation is used to ensure more reliable results. Fig.7 shows how the cost related to number of leaves during training. Though compared to Fig.6, CART uses more leaves (21 leaves is chosen) to reach about 70% overall success rate. However CART performs the best as analyzed hereunder.

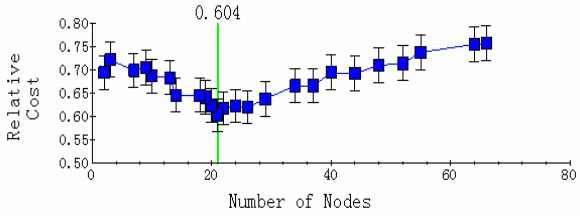

Fig.8. CART training

Total success rate for these three methods are **75.9%**, **73.0%** and **69.32%**, respectively. It seems that C4.5 decision tree performs best. But further analysis showed that contrary to this, CART is the best one. Fig. 8 gives the receiver operating characteristic (ROC) curves of C4.5 and NN. Roughly speaking, the NN performs better than C4.5.

But the CART performs the best as can be observed from Fig. 9, which is drawn separately for clarity. Confusion matrix confirmed these observations numerically (Table II). All rows in the table are presented in percentage to make them comparable. As pointed previously, although the C4.5 makes the highest total success rate, it detects 41.67% of SARS cases. This is not applicable to use. NN performs better but not well enough to match the 70.94% success rate given by CART. The bold faced data are results provided by CART.

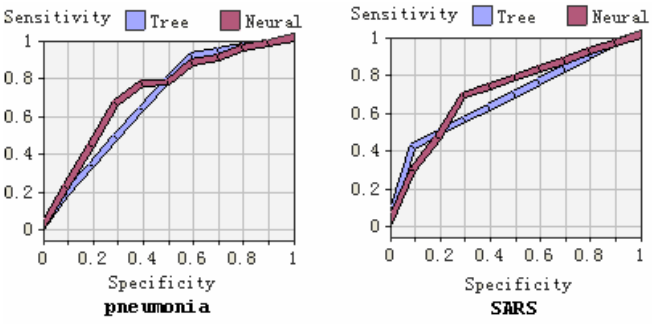

Fig.8. ROC curves of C4.5 and NN

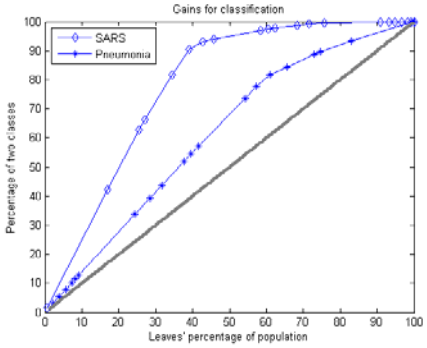

Fig.9. ROC curve of CART

TABLE II: CONFUSION MATRIX for C4.5 / NN /CART

		Predicated class					
Actual class		*SARS*			*Pneumonia*		
		C4.5	*NN*	*CART*	*C4.5*	*NN*	*CART*
	SARS	41.67	45.83	**70.94**	58.83	54.17	**29.06**
	Pneumonia	10.17	15.25	**31.34**	89.83	84.75	**68.66**

VI. CONCLUSION

Image mining for SARS detection, including detailed techniques for each key procedure is presented in this paper. Because simple location association rules cannot distinguish between SARS and normal pneumonia, we built classifier based on image texture. Experiment results from three classification methods of data mining are presented and analyzed. Final conclusion is that CART performs well: 70.94% of SARS cases can be detected.

SARS detection plays central role in computer aided diagnosis (CAD) system. It is believed that the letter 'D' in CAD involves two implications: detect first than diagnosis. Though 70% is inapplicable for real life usage, we believe this prototype is the stepping stone for further research. Some works are ongoing by our group.

REFERENCES

[1] Osmar R.Zaiane, Maria-Luiza Antonie, and Alexandru Coman, Mammography Classification by an Association Rule-based Classifier. *Proc. of the MDM/KDD 2002*, 2002.7, pp.62-69

[2] M. Kakimoto, C. Morita and H. Tsukimoto, Data Mining from Functional Brain Image, *Proc. of MDM/KDD 2000*, 2000.8, pp. 91-97

[3] Bram van Ginneken, Bart M. ter Haar Romeny, and Max A. Viergever, Computer Aided Diagnosis in Chest Radiograph: A Survey. *IEEE Trans. on Medical Imaging*, Vol.20,No.12,2001.12, pp.1228-1241

[4] Summary of SARS case by country, World Health Organization, http://www.who.int/csr/sars, 2003.9

[5] Anil T. Ahuja et.al ,*Radiological Appearance of Recent Cases of Atypical Pneumonia in Hong Kong*, The Dept. of Diagnostic Radiology and Organ Imaging,The Chinese Univ. of Hong Kong. http://www.droid.cuhk.edu.hk/web/atypical_pneumonia,2003.10

[6] T.F.Cootes, C.J.Taylor, D.H.Cooper,etc.al., Active Shape Models-Their Training and Application.*Computer Vision and Image Understanding.* Vol.61,No.1,1995,pp.38-59

[7] T.F.Cootes, C.J.Taylor, A. Lanitis, Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search. *Proc. of the British Machine Vision Conference*, 1994, pp.327-336.

[8] Xie Xuanyang,Li Xi,Zhang Jin,Gong Yuchang,The Window of Lung Fields: Automatic Determination and its Applications in Chest Radiograph Processing, *The 3rd IASTED International Conference on Biomedical Engineering,* 2005-2,p458-086,ISBN 0-88986-476-4

[9] Xie Xuanyang, Li Xi, Xu Yufeng, Wan Shouhong and Gong Yuchang, Mining X-ray images of SARS patients, *Proc. of the 3rd Australasian Data Mining Conference*, 2004.12, Cairns, Australia, Simeon J. Simoff and Graham J. Williams, eds. ISBN 0-646-44379-8