MULTICLASS CLASSIFICATION FOR CHEST X-RAY IMAGES BASED ON LESION LOCATION IN LUNG ZONES

Mohd Nizam Saad[1],Zurina Muda[2a],Noraidah Sahari[2b] and Hamzaini Abd Hamid[3]

[1]*School of Multimedia Technology & Communication, Universiti Utara Malaysia, Sintok, Kedah, Malaysia, nizam@uum.edu.my*
[2ab]*Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, UKM Bangi, Selangor, Malaysia, zurina@ftsm.ukm.my & nsa@ftsm.ukm.my*
[t]*Radiology Department, National University Medical Center Malaysia, Bandar Tun Razak, Kuala Lumpur, Malaysia, drzanid@yahoo.com*

ABSTRACT. Innovation in radiology technology has generated numerous kinds of medical images like the chest X-ray (CXR). This image is used to find common problem in lung like the lesion through scanning process in lung area which is divided into six zones. By classifying the CXR images with common feature like the lesion location, we can ensure efficient image retrieval. Recently, Support Vector Machine (SVM) has turn out to be a well-known method for image classification. While many previous studies have reported the achievement of SVM in classifying images, yet there is still problem with this technique for multiclass classification. Since SVM is a binary classification technique, its ability is limited to classifying features between two classes at one time. Therefore, it is difficult to classify CXR images which contain many image features. Realizing the problem, we proposed an application method for multiclass classification with SVM to the CXR images based on the lesion position in the lung zones. The multiclass classification application is executed on the CXR images taken from Japan Society of Radiology Technology dataset. Lesion coordinates were selected as the classification input while the lung zones becomes the labels. The multiclass classification is performed with RBF kernel and the classification accuracy is tested to attain the classifiers performance. Overall, it can be concluded that the percentage of the classification accuracy is high with the highest accuracy percentage recorded at 98.7% while the lowest was 94.8%. Meanwhile, the average classification accuracy was recorded at 96.9%. The result obtained revealed that the SVM classifiers generated have successfully classified the lesion location correctly according to the lung zones.

Keywords: multiclass image classification; support vector machine, chest x-ray image, JSRT image dataset

INTRODUCTION

The chest X-ray (CXR) has been recorded to be the most medical image produced among others. It comprises almost one third of all radiology images produced in hospital (Tao, Peng, Krishnan, & Zhou, 2011). The reason for it mass production is because it is the easiest and cheapest radiology procedure to produce, yet it is very important to diagnosis any abnormalities at the early stage of the treatment. CXR is used to identify unusual objects found in chest

anatomy such as the lung, mediastinum and ribs. For the lung, the most common radiology procedure done is to detect the lung lesion (or pulmonary lesion) found in the lung area. Normally, radiologist detect the lesion existent by scanning the lung area which are divided into six zones namely left upper zone (LUZ), left middle zone (LMZ), left lower zone (LLZ), right upper zone (RUZ), right middle zone (RMZ) and right lower zone (RLZ). These zones are formed by dividing the lung area into two horizontal divisions and three vertical divisions (Mohd Nizam Saad, Muda, Sahari, & Hamid, 2014).

Ideally, CXR images whose lesion location have already been identified should be grouped together based on this feature for the sake of easiness in image retrieval. This ensures image search and access to be completed in the most effective ways. However, images need to be classified before they can be grouped together. In image processing field, image classification refers to the process of relating image attributes to the known features and the algorithm used to influence the classification is known as image classifiers (Anthony, Gregg, & Tshilidzi, 2007). Recently, Support Vector Machine (SVM) has attracted the attention of researcher community to occupy the technique as a method for image classification.

SVM starts as a part of statistical learning theory and later being modified to become a supreme method for image classification (Vapnik, 1998). SVM is applied to classify image features for various usages like categorizing and pre-filtering image to reduce search space, recognizing image features from multimodal devices and annotating image automatically based on specified image features. Although SVM has become the first choice for image classification, yet it still has loophole when dealing with multiclass classification. This is because; SVM in nature is a binary classification method where it is created to classify features between two classes (-1 , 1) at one time (Hsu, Chang, & Lin, 2010). Meanwhile in reality, images like the CXR contain more than two features and sometime up to hundred depending on how the features are extracted. Even the common low level image features have three feature that are color, texture and shape (Wang, Mohamad, & Ismail, 2010). Therefore, if SVM is used to classify these image features, it would be a repetitive task.

Realizing the repetitive classification tasks may hinder the effectiveness to cluster CXR images, in this paper, we proposed an application method for multiclass classification with SVM in order to classify the CXR images based on the lesion position in the lung zones. By occupying the method, an improved image classifier can be produced which powerful enough to classify the image. As a mean for sharing our experience in working with the method, the rest of the paper is layout as follow. Section 2 explain multiclass classification method for classifying CXR images based on the lesion position in lung zones. Section 3 discusses about the image dataset for the study while Section 4 elaborate the experiment that we have conducted. Section 5 presents the result based on the experiment and finally, Section 6 conclude all the works that we have done in the study.

SVM MULTICLASS CLASSIFICATION FOR THE CXR IMAGES BASED ON THE LUNG LESION POSITION

The multiclass classification problem refers to assigning each of the observations (CXR images) into one of the predefined *K* classes (Rahman, Bhattacharya, & Desai, 2007). Among researchers, this technique is also referred as a one-against-all classification technique (Mueen, Zainuddin, & Baba, 2008). In this paper, we examine a multiclass probability estimate technique by combining all pairwise comparisons of binary SVM classifiers. Therefore, given N training data (or the feature vector) $(x_i,y_{i,})$ i=1,2,3….*l* where $x_i \in R^n$, and their labels $(y_1,y_2,y_3, \ldots, y_n)$, where $y \in (+1,-1)_n$, then the general form of the binary linear classification function can be written as follow:

$$F(x) = w.z + b \quad (1)$$

which corresponds to a separating hyperplane:

$$w.z + b = 0 \quad (2)$$

where z is an input vector, w is a weight vector, and b is the bias. The goal of SVM is to find the parameters w and b for the optimal hyperplane to maximize the geometric margin $\frac{2}{\|w\|}$ hyperplanes, subject to the solution of the following optimization problem (Hsu et al., 2010):

$$\min_{w,b,\xi} \quad = \frac{1}{2} w^T w \; T + C \sum_{i=0}^{1} \xi_i \quad (3)$$

$$\text{subject to} \quad y_i = (w^T \phi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

where $\xi_i > 0$ and $C > 0$ are the penalty parameters of the error term and regularization parameter respectively. In order to ease reader understanding regarding the multiclass classification method applied in this study, Figure 1 illustrate the classification process as recommended by (Zhang, Islam, & Lu, 2012).

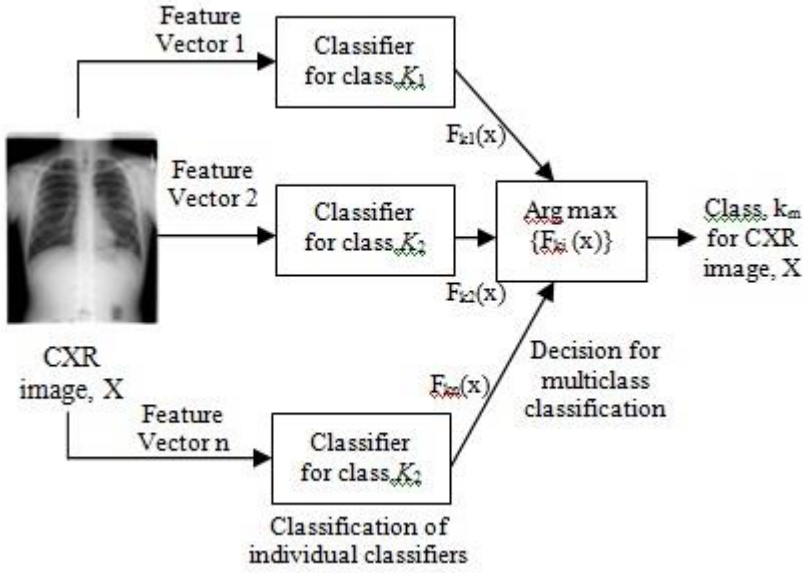

Figure 1. CXR Classification Process

As presented in Figure 1, the CXR image feature vector; (x_i, y_i), will be extracted first before the SVM classifier can be generated. In this study, the feature vector will be the coordinate of the lesion. Once extracted, these feature vectors will be used as an input to train the SVM classier. Before the training process, the classifier need to be label; y, so that each SVM classifier class; K , can be generated. Therefore, for this study, we have decided to use the lung zones as the label for the training process. By having the labels, we can calculate the number of classifier needed using the following equation:

$$\text{Number of classifier} = K(K\text{-}1)/2 \quad (4)$$

For example, since we have six labels which are derived from six lung zones, therefore, there will be 15 SVM classifiers needed to run the multiclass classification for the CXR images.

IMAGE DATASET

The CXR images used in this study are from a public chest radiograph dataset of Japan Society of Radiological Technology (JSRT) at http://www.jsrt.or.jp/jsrt-db/eng.php (Shiraishi et al., 2000). There are 247 CXR images available in the dataset which are clustered into two parts that are images with lung nodule (154 images) and images without lung nodule (93 images). The images were scanned from films and have standard resolution of 2048 x 2048 pixels with 12 bit gray levels at 4096 grayscale. Apart from the image files, JSRT also provides a text file (.txt) for additional information regarding the images which contain patient age, gender, diagnosis and detected location of the nodule. Figure 2 shows an example of CXR image with lung nodule taken from JSRT dataset.

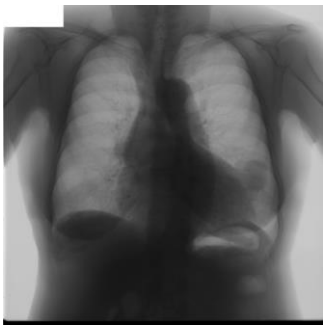

Figure 2. CXR Image With Nodule Taken from JSRT Dataset

EXPERIMENT

The aim of the experiment is to do multiclass classification for the CXR images based on the lesion location in the lung zones. The lesion coordinates (154 coordinate pairs from CXR image with nodule) are taken from the text file downloaded together with the image dataset from the JSRT website. These coordinates are scaled down to quarter because the original numbers is big since they are referring to the actual CRX image in the dataset. Thus, reducing the coordinate and the image would be the best way to manipulate the images. We have used Matlab as the tool to execute the multiclass classification with additional SVM library named LIBSVM (Chang & Lin, 2011). According to Hsu et al.(2010), there are six tasks should be taken when applying multiclass classification. Firstly, the classification data must be transformed in SVM package which means that the data must be transformed into real numbers. Since our classification data are in the form of lesion coordinate (x-axis and y-axis), therefore no data transformation is required. Secondly, data scaling task must be performed in order to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. In their paper, Hsu et al.(2010), recommended that the linear scaling for each attributes is set to the range of [-1,+1] or [0,1]. Using this recommendation, we have divided our lesion coordinate by 1000 so the numerical range falls between the range of zeros to one [0, 1]. For example, the original coordinate for the first CXR image is (1634,692), reducing the coordinate to quarter makes it (409,173) then applying the data scaling resulting the coordinate to become (0.409,0.173).

The third task in multiclass classification is to select the kernel type. There are four famous kernel in SVM classification that include linear, polynomial, radial based function

(RBF) and sigmoid (Hsu et al., 2010). In this experiment, we have applied RBF kernel for the multiclass classification.

$$\text{RBF:} \qquad K(x_i,x_j)= \exp(-\gamma\|x_i - x_j\|^2), \gamma > 0) \tag{5}$$

where γ is the kernel parameters.

The RBF is chosen because it is a nonlinear (Gaussian) kernel and able to map the lesion coordinate into higher dimensional space. It can handle cases when relation between class labels and the coordinate is nonlinear. The fourth task is to separate the training and test data from the JSRT dataset via cross-validation technique. By separating the data in the duo, the accuracy prediction on the training data can be obtained and this can reflects the performance on classifying the test data. In this study, the separation is done based on random sub-sampling where the JSRT dataset is divided into equal halves between the training and test data. Therefore, there are 77 lesion coordinates for both training and test data. After the separation, the fifth task is to use the best parameter to train the training data. With the help of LIBSVM library, we have set C = 4 and γ=0.2 for the RBF kernel parameter (refer equation (3) and (5)). Readers are advised to read details specification of these parameters in (Chang & Lin, 2011). Finally, the last task is to run classification test for the test data. The final task is straightforward where the SVM classifiers are used to classify the lesion coordinates in the test dataset. Later the classification result occurs with the lesion coordinates are clustered into six lung zones based on their location in the image.

RESULTS AND DISCUSSION

For the experiment results, we have chosen accuracy test to evaluate the performance of the SVM classifiers classifying the lesion coordinated in the lung zones. According to Chang & Lin (2011), the classification accuracy can be derived using this equation:

$$\text{Accuracy} = \frac{\text{\# correctly predicted data}}{\text{\# testing data}} \text{ x } 100\% \tag{6}$$

Table 1 summarized experiment results where the accuracy result obtained in the multiclass classification experiment for the lesion location in the lung zones.

Table 1. Classification Accuracy

Lung zones	Classification accuracy (%)	# correctly predicted data / # testing data
LUZ	98.7%	(76/77)
LMZ	97.4%	(75/77)
LLZ	97.4%	(75/77)
RUZ	96.1%	(74/77)
RMZ	94.8%	(73/77)
RLZ	97.4%	(75/77)

Overall, it can be seen in Table 1 that the percentage of the classification accuracy for all zones is high (more than 90%). The LUZ has the highest classification accuracy at 98.7% where the SVM classifiers had correctly predicted 76 out of 77 tested data. Meanwhile the classification accuracy for three lung zones namely LMZ, LLZ and RLZ are similar at 97.4%. On the other hand, the classification accuracy for the RUZ and RMZ is slightly lower than the others at 96.1% and 94.8% respectively. Based on the percentages listed in Table 1, the aver-

age classification accuracy for all lung zones was at 96.9%. This percentage in this experiment indicates that the performance of the SVM classifiers in the multiclass classification is excellent.

DISCUSSION AND CONCLUSION

In this paper, we have shared our experience in applying multiclass classification for the lesion location in lung zones for the CXR images. The classification is successfully done using SVM technique where the lesion coordinates are used as the classification input while the lung zones become the classification labels. Overall, the classification accuracy shows high achievement with the average accuracy has been recorded at 96.9%. The figure has indicated that the performance of our SVM classifiers is outstanding. In the future, we hope to perform further test for our classifiers like integrated them into CBIR system. By this integration, we can test whether classifying images into predefined group help the image retrieval.

ACKNOWLEDGEMENT

This study was financially supported by the FRGS/1/2014/ICT05/UKM/02/2 grant from the Ministry of Education Malaysia.

REFERENCES

Anthony, G., Gregg, H., & Tshilidzi, M. (2007). Image Classification Using SVMs: One-against-One Vs One-against-All. In *Proceedings of the 28th Asian Conference on Remote Sensing. 2007*.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27:1–27:27.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). *A Practical Guide to Support Vector Classification* (Vol. 101). doi:10.1177/02632760022050997

Mohd Nizam Saad, Muda, Z., Sahari, N., & Hamid, H. A. (2014). Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology. In *4th IEEE International Conference on Control Systems, Computing and Engineering*.

Mueen, A., Zainuddin, R., & Baba, M. S. (2008). Automatic multilevel medical image annotation and retrieval. *Journal of Digital Imaging*, *21*(3), 290–5. doi:10.1007/s10278-007-9070-3

Rahman, M. M., Bhattacharya, P., & Desai, B. C. (2007). A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *IEEE Transactions on Information Technology in Biomedicine*, *11*(1), 58–69.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., … Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, *174*(1), 71–74.

Tao, Y., Peng, Z., Krishnan, A., & Zhou, X. S. (2011). Robust learning-based parsing and annotation of medical radiographs. *IEEE Transactions on Medical Imaging*, *30*(2), 338–50.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Wang, H. H., Mohamad, D., & Ismail, N. A. (2010). Semantic Gap in CBIR : Automatic Objects Spatial Relationships Semantic Extraction and Representation. *International Journal Of Image Processing*, *4*(3), 192–204.

Zhang, D., Islam, M. M., & Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition*, *45*(1), 346–362.