International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 3392~3398
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp3392-3398

Feature Extraction of Chest X-ray Images and Analysis Using PCA and kPCA

Roopa H[1], Asha T[2]
[1]Department of I.S.E, Bangalore Institute of Technology, K. R. Road, Bangalore, India
[2]Department of C.S.E, Bangalore Institute of Technology, K. R. Road, Bangalore, India

Article Info

Article history:

Received Oct 9, 2017
Revised Feb 22, 2018
Accepted Mar 8, 2018

Keyword:

Classification
Feature extraction
Image mining
Tuberculosis

ABSTRACT

Tuberculosis (TB) is an infectious disease caused by mycobacterium which can be diagnosed by its various symptoms like fever, cough, etc. Tuberculosis can also be analyzed by understanding the chest x-ray of the patient which is revealed by an expert physician .The chest x-ray image contains many features which cannot be directly used by any computer system for analyzing the disease. Features of chest x-ray images must be understood and extracted, so that it can be processed to a form to be fed to any computer system for disease analysis. This paper presents feature extraction of chest x-ray image which can be used as an input for any data mining algorithm for TB disease analysis. So texture and shape based features are extracted from x-ray image using image processing concepts. The features extracted are analyzed using principal component analysis (PCA) and kernel principal component analysis (kPCA) techniques. Filter and wrapper feature selection method using linear regression model were applied on these techniques. The performance of PCA and kPCA are analyzed and found that the accuracy of PCA using wrapper approach is 96.07% when compared to the accuracy of kPCA which is 62.50%. PCA performs well than kPCA with a good accuracy.

Corresponding Author:

Roopa H,
Department of Information Science and Engineering,
Bangalore Institute Of Technology,
K R Road, Bangalore-560004, Karnataka, India.
Email: roopatejas@gmail.com

1. INTRODUCTION

Medical images are complex in nature. The information present in a medical image must be analyzed and investigated for any specific application. Raw image contains many properties associated with it called as features. Features can be of low level feature and high level feature. From the original image, low level features are extracted but high level features are extracted based on low level features. Features contain information like color, texture, shape or context.

Tuberculosis (TB) [9] is an infectious disease caused by mycobacterium which usually affects lungs.TB is transmitted from one person to another through air when an infected person sneezes or coughs. TB mostly occurs in lungs but can also occur in other parts of the body like brain, spine, kidney and bones. The main symptoms of TB are loss of weight, fever, chills, weakness, chest x-ray image findings and cough.Every year World TB Day recognized on March 24 is an opportunity to raise awareness about TB and support worldwide TB prevention and control efforts. The World Health Organization (WHO) [12] monitors the level of TB in every country in the world. Worldwide more than 9 million people develop TB annually and 1.5 million die from the disease.

Features represent information in an image. To understand the information present in an image, it must be analyzed and measured in various angles to get relevant information in a particular domain like

medical image, satellite images, etc., so, feature extraction [13], [14] plays an important role in analyzing medical image like TB image. PCA is a linear method which uses an orthogonal transformation to convert set of values of possibly correlated features into a set of uncorrelated features named as principal components. It puts all data along the axes where the variance of first principal component is highest when compared to other principal components. Here the axes correspond to the largest Eigen values of the data. kPCA is a nonlinear technique which is an extension of PCA that uses kernel methods. Principal components are computed using kernel functions by nonlinear mapping of input feature space. These features are finally placed in a nonlinearly transformed space. Features extracted [10] are analyzed and compared on linear and nonlinear feature space, then their performance are measured using datamining models like linear regression model. Linear regression model is a method to find a relationship between one dependent variable and series of changing independent variables by fitting a linear equation to observed data.

Perner *et al* [2] have discussed about framework of image mining, developed data mining and image processing tool which is helpful for medical image analysis. Descriptions of list of attributes as given by experts are stored in a database then a classification technique decision tree induction tree is applied to this to extract expert knowledge. This tool was used for various applications like breast MRI data, etc. Asha *et al* [8] used data mining techniques like Association Rule Mining (ARM) on TB data sets to improve TB disease prediction. The symptoms of TB are considered and many descriptive rules were written and these were combined with an association classification technique used for predicting TB.

Pedro *et al* [3] extracted texture and shape based features from MRI data, applied statistical association rules, and used continuous feature selection concepts to find patterns from these data. M.Suganthi *et al* [6] used Multi objective Genetic Algorithm (MOGA) to extract texture and shape based features from breast tumor data. Li –Yeh Chung *et al* [7] used feature selection and Taguchi genetic algorithm together on DNA microarray data then KNN with Leave One Out Cross Validation (LOOCV) was used to evaluate the performance.

Mahmoodabadi *et al* [4] used PCA to extract features of brain MRS data then Simple Genetic Algorithm (SGA) is used to discriminate these features. Liu Yihui *et al* [5] extracted wavelet features from microarray data then Support Vector Machine (SVM) was used for classification. Zyout *et al* [11] extracted textual pattern from mammogram images and Particle Swarm Optimization (PSO) was applied to select the most discriminative features then SVM was used for classification. Roopa *et al* [15] used city block distance measure for segmenting chest x-ray image which helps in diagnosis of TB.

Research on feature extraction of various image data on different domain applications have been carried on but not on the chest x-ray image related to TB disease.The aim of this paper is to analyze chest x-ray image to extract relevant features that represents symptoms of TB by applying image processing concepts.The extracted features are examined using PCA and kPCA, then the transformed feature space is subjected to linear regression model for classifying the TB disease.

The paper is organized as follows. Section 2 discusses about proposed method for extracting and selecting features from x-ray images, analyze using PCA and kPCA by applying linear regression model.Experimental illustration and results are described in detail in Sections 3. This paper is concluded in Section 4.

2. RESEARCH METHOD

Chest x-ray image contains relevant, irrelevant and redundant information. Features which are relevant and informative with respect to TB disease should be considered. The x-ray image must first be preprocessed and then important features are extracted from the affected region of x-ray using image processing methods. The following figure Figure 1 shows the proposed steps involved in extracting and selecting features from a x-ray image, then analyse it using data mining classification technique.

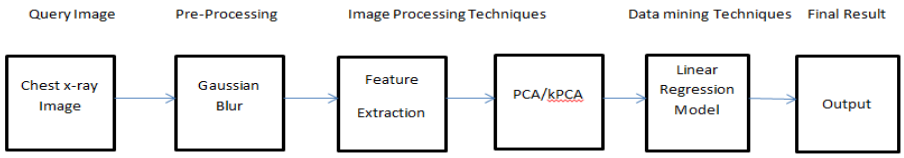

Figure 1. X-ray image feature extraction and selection

The main steps involved in analyzing and extracting features from x-ray image by applying image processing and datamining techniques are:

First, an x-ray image is taken as input.

Preprocessing:

Preprocessing of the x-ray image is done to remove noise and redundant data if present any. This is done using Gaussian blur filter method.

Feature Extraction:

Geometric features and Texture based features can be used to measure the characteristics of TB. Both the shape descriptors and texture descriptors were extracted from x-ray image. Extracting as many features from the region of interest of TB is one the concern in this work which is done by applying roipoly () function using matlab software.

Shape based features were used to measure the region of interest in a TB image. The statistics like Area, Perimeter, Coordinates of region centroid, Major Axis Length, Minor Axis Length, Eccentricity, Orientation, etc., are some of the characteristics of shape based feature extraction which is obtained by analyzing external boundary of the x-ray image. These features are obtained by using function called as regionprops ().

Texture descriptors proposed by Haralick [1] defines fourteen statistics that can be calculated from the co-occurrence matrix of the image. Texture feature extraction refers to surface characteristics and appearance of an object in an image. Entropy can be found by using entropy(), graycrops()function can be used to extract Homogeneity, Contrast, Energy, Correlation. numel(UL)where UL represents uniformity, Mean, Standard Deviation can be calculated using SD=sqrt(VR) Where VR represents the variance and Skewness by using function skewness() .All these features values were extracted from chest x-ray images. Principal Component Analysis(PCA) and Kernel Principal Component Analysis(kPCA):

PCA is a linear method when applied to the extracted features of step 3 results in axes that contains corresponding principal Eigen vectors which represents the data that is spread out along this axes. It puts all features along the axes where the variance of first principal component is highest when compared to other principal components. Here the axes correspond to the largest Eigen values of the data. PCA maximizes variance of extracted features which are uncorrelated.

kPCA is a nonlinear technique which is an extension of PCA that uses kernel methods. kPCA is applied to extracted features where principal components are computed using Radial Basis kernel functions with gamma=1.00.These features are finally placed in a nonlinearly transformed space. Linear Regression classification model:

The performances of transformed feature space are measured using linear regression model. Linear regression model is applied on the result of PCA and kPCA. Here features are selected using M5 prime method by eliminating collinear features with a minimum tolerance of 0.05.

Final Result:

Features obtained from x-ray images are classified as affected or normal image by using linear regression model and the performance of PCA and kPCA with respect to filter and wrapper feature selection methods were evaluated using 10-fold cross validation technique.

3. RESULTS AND DISCUSSION

Consider TB affected image, initially the image is preprocessed to remove noise using Gaussian function. Then texture and shape based features are extracted using matlab based on image processing concepts. The results of feature extraction method of TB x-ray image are illustrated in Figure 2, Figure 3 and Figure 4 respectively.

Figure 2. TB image

Figure 3. Marked image

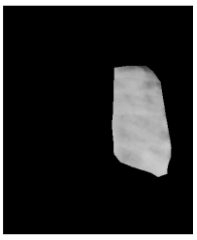

Figure 4. Masked image

The affected region of TB image is marked and then the marked region is masked to extract shape based and texture based feature which are shown in Figure 3 and Figure 4 respectively.By using the image processing methods, texture based and shape based features values extracted from the masked image of TB are stored in a file. The extracted values of a single TB image are shown in the Table 1.

Table 1. Extracted Feature Values of a single TB Image

Entropy	2.0954
Skewness	0.2724
SNR	6.1205
Homogeneity	0.9964
Contrast	0.16
Energy	0.4991
Correlation	0.9933
Total Mean	110.4313
Variance	110.0112
Standard Deviation	10.4886
Uniformity	111
Area	1744
Perimeter	172.625
Major Axis	65.7998
Minor Axis	36.1286
Eccentricity	0.8358

For the implementation we have considered 47 TB affected images and 30 normal images. Features from x-ray images were extracted using matlab software and feature selection and evaluation using classification model was performed by using Rapidminer software.

All extracted features values from these 77 images are stored in a file. This file was used as an input for PCA and kPCA. First filter based feature selection method was applied on PCA and kPCA separately, and then this was fed to linear regression classification model to classify the x-ray image as affected or normal. Later wrapper based feature selection method was applied on PCA and kPCA separately, then again this was fed to linear regression classification model to classify the x-ray image as affected or normal The performance of the classification model was analyzed using 10-fold cross validation technique.

Features extracted from TB and normal images are represented in a linear and nonlinear feature space. Filter and Wrapper feature selection method are applied on these extracted features. PCA is applied on the obtained feature space which maximizes the variance of extracted features that are uncorrelated. kPCA is also applied on this feature space so that the features are placed in a nonlinearly transformed space. Features from feature space are subjected to linear regression model. The performance of linear regression classification model was evaluated using 10 fold cross validation and the results are shown in Table 2 for filter approach and Table 3 for wrapper approach. The overall results of linear regression classification model are given in Table 4.

Table 2. Result of Linear Regression Classification Model where Feature Selection was done using Filter Method

PCA	kPCA
PerformanceVector PerformanceVector: accuracy: 95.65% ConfusionMatrix: True: Affected Normal Affected: 13 0 Normal: 1 9 precision: 90.00% (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 13 0 Normal: 1 9 recall: 100.00% (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 13 0 Normal: 1 9 AUC (optimistic): 0.929 (positive class: Normal) AUC: 0.929 (positive class: Normal) AUC (pessimistic): 0.929 (positive class: Normal)	**PerformanceVector** PerformanceVector: accuracy: 60.87% ConfusionMatrix: True: Affected Normal Affected: 14 9 Normal: 0 0 precision: unknown (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 14 9 Normal: 0 0 recall: 0.00% (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 14 9 Normal: 0 0 AUC (optimistic): 1.000 (positive class: Normal) AUC: 0.500 (positive class: Normal) AUC (pessimistic): 0.000 (positive class: Normal)

Table 3. Result of Linear Regression Classification Model Where Feature Selection was done using Wrapper Method

PCA	kPCA
PerformanceVector PerformanceVector: accuracy: 96.07% +/- 8.21% (mikro: 96.10%) ConfusionMatrix: True: Affected Normal Affected: 47 2 Normal: 1 27 AUC (optimistic): 0.980 +/- 0.060 (mikro: 0.980) (positive class: Normal) AUC: 0.980 +/- 0.060 (mikro: 0.980) (positive class: Normal) AUC (pessimistic): 0.980 +/- 0.060 (mikro: 0.980) (positive class: Normal) precision: 96.67% +/- 10.00% (mikro: 96.43%) (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 47 2 Normal: 1 27 recall: 95.00% +/- 15.00% (mikro: 93.10%) (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 47 2 Normal: 1 27	**PerformanceVector** PerformanceVector: accuracy: 62.50% +/- 5.87% (mikro: 62.34%) ConfusionMatrix: True: Affected Normal Affected: 48 29 Normal: 0 0 AUC (optimistic): 1.000 +/- 0.000 (mikro: 1.000) (positive class: Normal) AUC: 0.500 +/- 0.000 (mikro: 0.500) (positive class: Normal) AUC (pessimistic): 0.000 +/- 0.000 (mikro: 0.000) (positive class: Normal) precision: unknown (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 48 29 Normal: 0 0 recall: 0.00% +/- 0.00% (mikro: 0.00%) (positive class: Normal) ConfusionMatrix: True: Affected Normal Affected: 48 29 Normal: 0 0

Table 4. Comparitive Results of PCA and kPCA using Filter and Wrapper Feature Selection Method

Feature Selection Method	Filter Approach		Wrapper Approach	
	PCA	kPCA	PCA	kPCA
Accuracy	95.65%	60.87%	96.07%	62.50%
Precision	90%	unknown	96.07%	unknown
Recall	100%	0.00%	95%	0.00%
AUC(optimistic).	0.929	1	0.98	1
AUC	0.929	0.5	0.98	0.5
AUC(pessimistic)	0.929	0	0.98	0
RMSE	0.418±0.00	0.463±0.00	0.422±0.002	0.493±0.004
Squared Error	0.174±0.079	0.214±0.081	0.179±0.017	0.243±0.004

4. CONCLUSION

The current work proposed uses image processing concepts to extract relevant and important features from a chest x-ray image to diagnoise whether a person is TB infected or not , which broadly encompass data preprocessing and feature extraction process. Filter and wrapper feature selection methods are implemented on these extracted features. The features obtained are highly projected to a feature space where PCA or kPCA are applied on this space. Then linear regression classification model is applied to analyze the TB disease. The performance of PCA and kPCA are examined and found that the accuracy of PCA using wrapper approach 96.07% is better when compared to the accuracy of kPCA which is 62.50%. Future work will be considered on more images and carry out the implementation on this huge data set.

REFERENCES

[1] Haralick, Robert M, "Statistical and Structural Approaches to Texture", *Proceedings of the IEEE,* 67.5, pp. 786-804, 1979.

[2] Perner, Petra, "Image Mining: Issues, Framework, a Generic Tool and its Application to medical-image diagnosis", *Engineering Applications of Artificial Intelligence*, 15.2, pp. 205-216, 2002.

[3] Bugatti, Pedro Henrique, *et al*., "Content-based retrieval of medical images by continuous feature selection", *Computer-based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on. IEEE,* 2008.

[4] Mahmoodabadi, S. Zarei, *et al*., "PCA-SGA Implementation in Classification and Disease Specific Feature Extraction of the brain MRS signals", *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008.

[5] Liu, Yihui, "Wavelet Feature Extraction for High-dimensional Microarray Data", *Neurocomputing*, 72.4, pp. 985-990, 2009.

[6] Suganthi, M., and M. Madheswaran, "Mammogram Tumor Classification using Multimodal Features and Genetic Algorithm", *Control, Automation, Communication and Energy Conservation, 2009. INCACEC 2009*. *International Conference on. IEEE*, 2009.

[7] Chuang, Li-Yeh, *et al*., "A Hybrid Feature Selection Method for DNA Microarray Data", *Computers in Biology and Medicine*, 41.4, pp, 228-237, 2011.

[8] Asha. T, Dr. S. Natarajan, Dr. K. N. B. Murthy, "A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction", *IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications*, AIT, 2011.

[9] Asha, T., *et al*., "Data Mining Techniques in the Diagnosis of Tuberculosis", INTECH Open Access Publisher, 2012.

[10] Rad, Abdolvahab Ehsani, Mohd Shafry Mohd Rahim, and Alireza Norouzi, "Digital Dental x-ray Image Segmentation and Feature Extraction", *Indonesian Journal of Electrical Engineering and Computer Science*, 11.6, pp. 3109-3114, 2013.

[11] Zyout, Imad, Joanna Czajkowska, and Marcin Grzegorzek, "Multi-scale Textural Feature Extraction and Particle Swarm Optimization Based Model Selection for False Positive Reduction In Mammography", *Computerized Medical Imaging and Graphics*, 46, pp. 95-107, 2015.

[12] GlobalTuberculosisReport2015-http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf

[13] Ali, Rozniza, Amir Hussain, and Mustafa Man, "Feature Extraction and Classification for Multiple Species of Gyrodactylus ectoparasite", *Indonesian Journal of Electrical Engineering and Computer Science*, 13.3, pp. 503-511, 2015.

[14] Faridah and Balza Achmad, "Lip Image Feature Extraction Utilizing Snake's Control Points for Lip Reading Applications", *International Journal of Electrical and Computer Engineering*, 5.4, p. 720, 2015.

[15] Roopa H and Asha T, "Segmentation of X-Ray Image using City Block Distance Measure", *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kumaracoil, pp. 186-189, 2016.

BIOGRAPHIES OF AUTHORS

Roopa. H received M.Sc degree in Mathematics from Bangalore University in 2002 and M.Tech in Computer Science and Technology from University of Mysore in 2006.She is currently working as an Assistant Professor in the Department of Information Science & Engineering, Bangalore Institute of Technology, and Bangalore. She is currently pursuing her PhD in Computer Science at VTU in the area of datamining.

Dr. Asha. T is a Professor & PG Coordinator in the Department of Computer Science & Engineering, Bangalore Institute of Technology, Bangalore. She obtained her Ph.D in Computer and Information Science from Visvesvaraya Technological University, Karnataka. She has published around 24 papers in International/National journals and Conferences. Her research interests include Data Mining, Medical Informatics, Pattern Recognition, Big data management etc., email: asha.masthi@gmail.com