Lung cancer detection using the SOM-GRR based radial basis function neural network

D U Wutsqa and A Farhan

1Department of Statistics, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
2Department of Mathematics, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

E-mail: dhoriva_uny@uny.ac.id, alvinfarhan95@gmail.com

Abstract. This study intends to detect lung cancer using chest X-ray. We propose the SOM-GRR based radial basis function neural network (RBFNN) model. The self-organizing maps (SOM) based deals with unsupervised learning and the global ridge regression (GRR) based deals with supervised learnings that are carried out in developing RBFNN model. The gray level co-occurrence matrix (GLCM) extraction is performed to obtain the features of chest X-ray which are used as RBFNN input variables. We consider thirteen features, namely contrast, correlation, energy, homogeneity, sum entropy, variance, inverse difference moment, sum average, sum variance, entropy, difference entropy, maximum probability, and dissimilarity. The best model is obtained by evaluating its performance in training and testing data sets. The RBFNN model yields 93 % and 88 % accuracy in training and testing data sets, respectively.

1. Introduction

Lung cancer is a malignant lung tumor that starts from the bronchial epithelium that is characterized by abnormal, unlimited cell growth and it damages to normal tissue cells. Lung cancer causes about one third of all cancer deaths in men [1]. The main cause of lung cancer is cigarette smoke because it contains more than 4,000 chemicals, 63 types of which are carcinogens and toxic [2]. Other causes of lung cancer are radiation, air pollution, family history of lung cancer [3].

As a first step, the detection of lung cancer can be done by lung radiological examination. Lung image or often called chest X-ray is a radiographic projection of the lung. The chest X-ray will show a different image between the normal and the abnormal lungs. Abnormal lung image will show the presence of nodules (round or oval), but the normal lung does not indicate the presence of nodules. Recently, soft computing is an interesting approach to detect cancer, specifically lung cancer. Some works on lung cancer detection have been reported, including kernel nearest neighborhood [4], artificial neural networks [5], and radial basis neural networks [6-7]. In this work, we are interested in exploring more about the application of the RBFNN model in lung cancer detection with different learning processes.

Parameter estimation in the RBFNN model consists of unsupervised learning to estimate the parameters of the Gaussian function and supervised learning to estimate the weights between the hidden layer and the output layer. The usual unsupervised used method is the K-means clustering [6-7]. Here, the Self-Organizing Maps (SOM) method is proposed for unsupervised learning. SOM is one of the artificial intelligent approaches that are configured for grouping data [8]. The SOM method has a function like that in K-means clustering method. The ability of SOM is slightly better than K-means in recognizing clusters that statistically could not be recognized due to some assumptions, such as the
absence of correlation and not normally distributed [9]. The used of SOM as unsupervised learning in the RBFN model has been reported as an effective method to estimate riverine fish diversity [10], Tamil handwritten character recognition [11], induction machine fault detection [12], and breast cancer detection [13].

Meanwhile, the standard supervised learning method used in the RBFN model is the ordinary least square (OLS). The OLS is not appropriate for the data with a high correlation between the input variables since it leads to no unique solution problem. Deals with these kinds of data, Orr [14] recommends the global ridge regression (GRR) method for supervised learning. Considering the advantages of SOM and GRR, this research proposes a SOM-GRR based Radial Basis Function Neural Network method to detect lung cancer using chest X-ray data.

2. Method

2.1. Feature extraction

The texture is an important characteristic that can be used to identify the object or area of an image. Textural features based on spatial gray are commonly used in classifying images. The gray-level co-occurrence matrix (GLCM) has been used as a method to extract the features from the image, such as to extract the features from the colposcopy images [15-16] and mammographic images [17], and magnetic resonance imaging [18]. The features obtained from GLCM can help in understanding the overall image details in terms of texture [19]. In this work, we consider 13 features extracted from the chest X-ray using GLCM method [20-22]

1. Contrast = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)(e-g)^2\]
2. Correlation = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} \frac{(e-g)h(e,g))}{\sigma_a\sigma_b}\]
3. Energy = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)^2\]
4. Homogeneity = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} \frac{h(e,g)}{1+|e-g|}\]
5. Sum entropy = \[-\sum_{e=2}^{2Ng} h_{a+b}(s)\log(h_{a+b}(s))\]
6. Variance = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)(e-\mu)^2\]
7. Inverse difference moment = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} \frac{h(e,g)}{1+|e-g|^2}\]
8. Sum average (sa) = \[-\sum_{s=2}^{2Ng} (s)(h_{a+b}(s))\]
9. Sum variance = \[\sum_{s=2}^{2Ng} (e-sa)^2 h_{a+b}(s)\]
10. Entropy = \[-\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)\log_2\{h(e,g)\}\]
11. Difference entropy = \[-\sum_{q=0}^{Ng-1} (h_{a-b}(q))\log(h_{a-b}(q))\]
12. Maximum probability = \[\max_{e,g}\{h(e,g)\}\]
13. Dissimilarity = \[\sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)|e-g|\]

The following notations are used in the above features.

\[h(e,g)\] = pixel on eth row and gth column, \(e = 1,2,\ldots,Ng;\ g = 1,2,\ldots,Ng\)

\(Ng\) = the number of grayscales of the image.

\(\mu_a = \sum_{e=1}^{Ng} \sum_{g=1}^{Ng} h(e,g)\) (mean of row element on image histogram)
\[ \mu_b = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} e h(e, g) \] (mean of column element on image histogram)

\[ \sigma_a = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} ((g - \mu_a)^2 h(e, g)) \] (variance of row element on image histogram)

\[ \sigma_b = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} ((e - \mu_b)^2 h(e, g)) \] (variance of column element on image histogram)

\[ h_{a+b}(s) = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} h(e, g); e + g = s; s = 2, 3, ..., 2N_g \]

\[ \mu = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} egh(e, g) \]

\[ h_{a-b}(q) = \sum_{e=1}^{N_g} \sum_{g=1}^{N_g} h(e, g); |e - g| = q; q = 0, 1, ..., (N_g - 1) \]

### 2.2. SOM-GRR based RBFNN Model

The specific characteristic of the RBFNN is its activation function which is included in the radial basis function class. The most popular activation function is Gaussian. It has two parameters, those are the cluster center and maximum distance. The architecture of the RBFNN consists of three layers, i.e. the input layer, one hidden layer, and the output layer as displayed in Figure 1.

![RBFNN architecture](image)

**Figure 1.** RBFNN architecture

The RBFNN model with Gaussian activation function can be expressed as [14]

\[ y = \sum_{k=1}^{q} w_k \varphi_k(x) + w_0 \] (1)

\[ \varphi_k(x) = \exp \left[ -\frac{(x - c_k)^T (x - c_k)}{r_k^2} \right] \]

where \([x]^T = [x_1, x_2, ..., x_j, ..., x_p]\) is the input vector, \(w_k\) is the weight between the hidden layer and output layer, \(w_0\) is the bias, \([C_k]^T = [c_1, c_2, ..., c_j, ..., c_p]\) is \(k\)th cluster center, \(r_k\) is \(k\)th cluster maximum distance, \(k = 1, 2, ..., q\), \(p\) is the number of input variables, and \(q\) is the number of hidden neurons.
The Gaussian parameters are estimated using the Self Organizing Map (SOM). It was introduced by a Finnish scientist Teuvo Kohonen in 1982. So, it is also referred to as Kohonen SOM. The SOM is a type of artificial neural network employing unsupervised learning in which the learning does not require supervision (target output). The SOM network consists of the input layer and the output layer (see Figure 2).

\[
D(k) = \sum_{j=1}^{p} (\rho_{kj} - x_j)^2
\]

where \(\rho_{kj}\) is the weight of the SOM network. The algorithm in SOM is done by finding the winning unit, which is the unit that has a \(D(k)\) minimum. The new weight is updated on the winning unit as

\[
\rho_{kj}(new) = \rho_{kj}(old) + \alpha [x_j - \rho_{kj}(old)]
\]

The iteration process is performed by decreasing the learning rate \(\alpha\) and reducing the radius of the topological neighborhood at specified times until it reaches the stopping condition. The Euclidian distance of each object in each cluster is calculated using the final weights. The object is placed in the cluster with the minimum Euclidian distance. Then, the Gaussian parameters are estimated by calculating the cluster center and maximum distance.

The weights between the hidden layer and output layer are estimated utilizing global ridge regression. It obtains the optimum weights by minimizing the cost function [14]

\[
CF = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{q} w_k^2
\]

where \(y_i\) is the target or the observed object, \(\hat{y}_i\) is the output of the model, \(n\) is the number of images, \(\lambda\) is the regulation parameter, \(w_k\) is the weight between the hidden layer and the output layer. The optimum weights of the RBFNN model are computed by using the formula

\[
\hat{w} = (\varphi^T \varphi + \lambda I_{q+1})^{-1} \varphi^T y
\]
where $\hat{w}^T = (w_1, w_2, ..., w_q, w_0)$, $\phi^T$ is the $(q + 1) \times n$ dimensional with element $q_k(x_i)$ and the $(k + 1)th$ column is the unit column vector, $I_{q+1}$ is the $(q + 1) \times (q + 1)$ dimensional identity matrix, and $y^T = (y_1, y_2, ..., y_n)$.

2.3. The procedure of SOM-GRR based RBFNN for lung cancer detection

Lung detection using SOM-GRR based RBFNN involves many steps. The procedure is arranged in the following steps.

**Step 1.** Image preprocessing.

In this work, it is conducted by changing the file format from Disk Image File becomes Joint Photographic Experts Group, removing the image background, changing the images from red blue green to grayscale, and changing the pixel size.

**Step 2.** Feature extraction using GLCM.

**Step 3.** Data normalization.

Since the features have different sizes, they need to be normalized.

**Step 4.** Data splitting.

The data are split into 80 % training data and 20 % testing data. This step deals with the cross-validation process. The model is built using the training data, then it validates using testing data to evaluate the generalization ability of the model to out sample data.

**Step 5.** Parameter estimation.

The parameters of the model are estimated using SOM unsupervised learning and GRR supervised learning. The best RBFNN model is obtained by examining its performance on the various number of hidden neurons or clusters.

**Step 6.** Model performance evaluation

The performance of RBFNN model is evaluated relied on sensitivity, specificity, and accuracy criteria both in training and testing data. Those criteria are expressed as [24]

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \tag{6}
\]
\[
\text{Specificity} = \frac{TN}{TN+FP} \times 100\% \tag{7}
\]
\[
\text{Accuracy} = \frac{TN+TP}{TP+TN+FP+FN} \times 100\% \tag{8}
\]

where

- True Positive (TP) : the original classification of the image and learning output represent cancer.
- True Negative (TN) : the original classification of the image and learning output represent normal.
- False Positive (FP) : the original classification of the image represents normal and learning output represent cancer.
- False Negative (FN): the original classification of the image represents cancer and learning output represent normal.

3. Result and Discussion

This study uses a Chest X-ray obtained from a digital image database available at the Japanese Society of Radiography Technology [25]. The image has 2048x2048 pixels in the Disk Image File format. The data consist of 68 images of cancer lungs and 57 images of normal lungs. The first step for classifying lung cancer is image preprocessing. The examples of the Chest X-Ray before and after preprocessing are shown in Figure 3.
The image in figure (b) focuses on the lung image without a background, so it is hoped that the extraction results will not be interrupted by the background. The extraction process using GLCM to all images produces 13 features. They are contrast, correlation, energy, homogeneity, sum entropy, sum of square variance, inverse difference moment, sum average, sum variance, entropy, difference entropy, maximum probability, and dissimilarity. These features are set as inputs of SOM-GRR based RBFNN model. The model only has a single output neuron whose values are 1 for normal and 2 for cancer. The data are split into 100 images (80%) for training data and 25 images (20%) for testing data. The distribution of the lung condition is 56 cancer and 44 normal, 12 cancer and 13 normal for training and testing data, respectively. The best SOM-GRR based RBFNN model is attained by attempting the number of clusters from 2 to 10. The results are presented in Table 1.

Table 1. The accuracy (%) of SOM-GRR based RBFNN

| Number Clusters | Training | Testing |
|-----------------|----------|---------|
| 2               | 85       | 72      |
| 3               | 90       | 72      |
| 4               | 88       | 88      |
| 5*              | 93       | 88      |
| 6               | 91       | 76      |
| 7               | 92       | 80      |
| 8               | 93       | 72      |
| 9               | 92       | 80      |
| 10              | 92       | 76      |

Table 1 shows that the highest accuracy of the SOM-GRR based RBFNN model is on the number of clusters 5 with a percentage of training data at 93% and testing data at 88%. Increasing the number of clusters more than 5 does not improve the performance of the model in testing data. The SOM-GRR based RBFN model delivers the network architecture with 13 inputs, 5 neurons and bias in the hidden layer, and a single neuron in the output layer. The complete SOM-GRR based RBFNN performance for detecting lung cancer is depicted in Table 2.
The SOM-GRR based RBFNN employed on the images without quality enhancement. It outperforms our previous results using K-means-GRR based RBFNN [6-7]. The high frequency emphasis filtering (HFEF) and histogram equalizer (HE) [6] and the point operation (PO) [7] are implemented to enhance the image quality. Our result has higher accuracy than the results of [6-7], even on the image with quality enhancement. The accuracy of the previous works is listed in Table 3.

Table 3. The accuracy (%) of K-means-GRR based RBFNN

| Method       | Training | Testing |
|--------------|----------|---------|
| HFEF & HE[6] | 92.5     | 80      |
| PO [7]       | 88.75    | 80      |

4. Conclusion

A SOM-GRR based RBFNN is developed as a tool for lung cancer detection. The SOM is used as unsupervised learning to estimate the Gaussian parameters, and the GRR is used to estimate the weights between the hidden layer and output layer. The inputs of the model are the features extracted from the chest radiograph by using the GLCM method. The results show the high performance in training data and a slight decrease in testing data. SOM learning is more effective than K-means learning to obtain high accuracy for lung cancer detection using RBFNN model.

References

[1] Perhimpunan Dokter Paru Indonesia 2003 Kanker Paru : Pedoman Diagnosis dan Penatalaksanaan di Indonesia (Lung Cancer: Guidelines for Diagnosis and Management in Indonesia) (Jakarta: Indonesian Lung Doctors Association)
[2] Kementerian Kesehatan Republik Indonesia 2016 Pedoman Nasional Pelayanan Kedokteran Kanker Paru (Guidelines for Lung Cancer Medical Services) (Jakarta: Ministry of Health Republic of Indonesia)
[3] American Cancer Society 2017 Lung Cancer Risk Factors. Accessed from https://www.cancer.org/cancer/lung-cancer/prevention-and-early-detection/risk-factors.html on 2 March 2017 at 07.00 am
[4] Bhuvaneswari P, Brintha A and Therese 2015 Detection of cancer in lung with K-NN classification using genetic algorithm Procedia Materials Science 10 pp 433–40 (Science Direct Elsevier)
[5] Zhou, Zhou Z H, Jiang Y, Yang Y B and Chen S F 2002 Lung cancer cell classification based on artificial neural network ensembles Artif Intell Med Vol 24 pp 25–36
[6] Abadi A M, Wutsqa D U and Pamungkas L R 2017 Detection of lung cancer using radiograph images enhancement and radial basis function classifier Proc. 10th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (Shanghai) 14–16 October (IEEE)
[7] Wutsqa D U and Mandadara H L R 2017 Lung cancer classification using radial basis function neural network model with point operation Proc. 10th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (Shanghai)14–16 October (IEEE)
[8] Siang J J 2009 *Jaringan Syaraf Tiruan dan Pemrogramannya Menggunakan MATLAB (Artificial Neural Networks and Programming using MATLAB)* (Yogyakarta: Penerbit Andi)

[9] Huang D S and Zao W B 2005 Determining the center of radial basis probabilistic neural network by recursive orthogonal least square algorithms *Appl Math Comput* Vol 162 pp 461–73

[10] Chang F J, Tsai W P, Chen H K, Yam R S W and Herricks E E 2013 A Self-organizing radial basis network for estimating riverine fish diversity *J. Hydrol* Vol. 476 pp 280–89

[11] Baboo S S, Subashini P and Krishnaveni M 2009 Combining self-organizing maps and radial basis function networks for Tamil handwritten character recognition *ICGST-GVIP J.* Vol. 9 pp 1–7

[12] Wu S and Chow T W S 2004 Induction machine fault detection using SOM-based RBFNN *IEEE Transactions on Industrial Electronics* Vol 51 pp 183–94

[13] Mu T and Nandi A K 2007 Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier *J. Franklin Inst.* Vol. 344 pp 285–311

[14] Orr M J 1996 *Introduction to Radial Basis Function Networks Centre for Cognitive Science* (Edinburgh : University of Edinburgh)

[15] Wutsqa D U and Marwah M 2017 Median filter noise reduction of image and backpropagation neural network model for cervical cancer classification *J. Phys. Conf. Ser.* Vol. 855

[16] Wutsqa D U, Subekti R, and Kusumawati R 2019 Fuzzy neural network with intensity adjustment and median filter for classifying cervical cancer *Int. J. Advance Soft Compu. Appl.* Vol. 11, No. 3 pp 173–87

[17] Wutsqa D U and Rahmada P S 2017 Point operation to enhance the performance of fuzzy neural network model for breast cancer classification *J. Eng. Appl. Sci.* Vol. 12 No. 17 pp. 4405–10

[18] Abadi A M, Wustqa D U and Nurhayadi 2019 Diagnosis of brain cancer using radial basis function neural network with singular value decomposition Method *Int. J. Mach. Learn. Comput.* Vol. 9 No.4 pp 527–32

[19] Gadkari D 2004 *Image quality analysis using GLCM* (Orlando: University of Central Florida)

[20] Haralick R M, Shanmugam K, and Dinstein I 1973 Textural features for image classification *IEEE Trans Syst Man Cybern* Vol. SMC-3 No. 6 pp 610–21

[21] Anami B S and Burkpalli V C 2009 Texture based identification and classification of bulk sugary food objects Vol. 9 Issue 4 pp 9–14

[22] Sharma M and Mukharjee S 2013 Brain Tumor Segmentation Using Genetic Algorithm and Artificial Neural Network Fuzzy Inference System (ANFIS). In: Meghanathan N, Nagamalai D, and Chaki N (eds) *Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing*, vol. 177 pp 329–39 (Springer, Berlin, Heidelberg)

[23] Faussett L 1994 *Fundamentals of Neural Network: Architecture, Algorithms, and Applications* (Upper Saddle River, New-Jersey: Prentice-Hall)

[24] Wong H B and Lim G H 2011 measures of diagnostic accuracy: sensitivity, specificity, PPV, and NPV. *Proc. of Singapore Healthcare* Vol 20 pp 316–18

[25] Japanese Society of Radiology Technology 1997 *Digital Image Database* Accessed from http://www.jsrt.or.jp on 4 March 2017 at 07.00 pm