J Biochem Tech (2018) Special Issue (2): 170-174
ISSN: 0974-2328

Application of Neural Network for Classification of Breast and Lung Cancer Patients using DNA Microarray Data

Leila Nezamabadi Farahani, Hossein Mahjub*, Javad Faradmal, Jalal Poorolajal and Massoud Saidijam

Received: 25 December 2017 / Received in revised form: 14 May 2018, Accepted: 18 May 2018, Published online: 05 September 2018

Abstract

Objectives: One of the major challenges facing the cancer biology is finding the best treatment, (the most efficient and least side effects). Using microarray data led to fundamental changes in prediction clinical outcomes. Analyzing microarray data due to large number of variables in comparison with the number of samples needs for appropriate methods. The aim of the present study was to evaluate and to compare two data mining methods for classification of cancer patients. Methods: This study used two public dataset (lung and breast cancer) for classification of tumor types and other outcomes. We applied wavelet transform for feature extraction and neural network as a classifier. The accuracy criterion was used to evaluate artificial neural network performance. Results: Accuracy of artificial neural network in lung and breast cancer data was 100%. Dimension reduction did not change the accuracy for lung cancer dataset but it slightly declined for the breast cancer dataset. Conclusion: Artificial neural network was highly efficient in determining tumor type using microarray data compared with other classification methods. The results indicated that feature extraction and dimension reduction with wavelet transform did not change the accuracy of artificial neural network for data with large enough sample size.

Keywords: Neural Networks, Gene Expression, Wavelet Transform, Feature transformation, Neoplasm.

Leila Nezamabadi Farahani
Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.

Hossein Mahjub*, Jalal Poorolajal
Research Center for Health Sciences and Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.

Javad Faradmal (PhD)
Modeling of Noncommunicable Disease Research Center and Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.

Massoud Saidijam (PhD)
Department of Molecular Medicine and Genetics, Hamadan University of Medical Sciences, Hamadan, Iran.

***Email:** mahjub@umsha.ac.ir.

Introduction

Based on WHO reports, more than 7.6million deaths (13% of all deaths) caused by cancer during 2008 (Globocan. 2008 January 2013). It is also expected to influence the life of more than 13 million people by the year of 2030. Because of its immense impact on human life it has become one of the biggest threat to the public health (Iyer et al., 2006). Cancer can occurs by uncontrolled growth and spread of abnormal cells and can affect almost all tissues of the body (Hejmadi, 2010).

Breast and lung cancers are among the major causes of deaths by cancers Globocan. 2008 January 2013). Distinguishing malignant pleural mesothelioma (MPM) and adenocarcinoma (**AC**) in patients with lung cancer and positive estrogen receptor (ER+) from negative estrogen receptor (ER-) in breast cancer patients by traditional methods, based on morphological appearance, is subjective. Also these methods, have serious limitations and may lead to significant variation (Cummings et al., 2011; Demir, & Yener, 2005; Gordon et al., 2002). These subtypes differ in their biology and patients with different subtypes demonstrate dissimilar results with the same modalities (de Ronde et al., 2010). Therefore, an accurate diagnostic method can help finding appropriate treatments.

The role of human DNA and genes on cancers is studied in the literatures. In the past, to study the properties, activity of genes and proteins, discovery of the molecular processes within cells and tissues, often a gene or a few genes were specifically and individually studied.

Microarray technology, which was introduced the first time by Schena, probes expression of thousand of genes simultaneously for studying expression of the cells(Schena et al., 1995; Lipshutz et al., 1995; Draghici, 2010). It helps scientists to work on human DNA in order to find the relationship between gene expressions and diseases (Mitra et al., 2012). Hence, the use of DNA microarrays has increased in recent years to find cancer subtypes or to identify the characteristics of the tumor cells (Arpino et al., 2012).

High dimensionality of the DNA microarray data leads the researchers to use advanced and complex methods such as artificial neural networks (ANN) to analysis such data. Modeling nonlinear relations and capturing high degree of interaction among variables are the features of ANN which makes it more flexible than classical statistical methods. The procedure of ANN consists of the training, testing, and validation phases which helps to achieve the best prediction for the outcomes.

To date, support vector machine, random forest, AdaBoost C4.5, BaggingC4.5, LibSVMs (Nanni et al., 2012; Saini et al., 2013; Hu et al., 2006; Nikumbh et al., 2012) methods are used to distinguish MPM and AC in lung cancer patients and ER+ and ER- in breast cancer ones. Using artificial neural network can model complex nonlinear relationships and high degree of interactions, and there is no limitation in the number of input variables and samples. Increasing the sample size will improve the network learning (vittingoff et al., 2012). In the current study, we intended to use ANN to distinguish the lung cancer patients with MDM and AC sub-types and breast cancer patients with ER+ and ER- using DNA microarray analysis. Furthermore, we calculate the effect of wavelet as a pre-processing method in the prediction accuracy of the ANN models.

Material and Methods

Datasets

In the current study, we use information of two different groups of patients, lung and breast cancer patients. The lung cancer datasets contains 181 tissues (31 MPM and 150 AC) with 1626 gene expressions for each sample. These specimens were gathered from patients who had gone surgery at Brigham and Women's Hospital from 1993 to 2001. The full description of these data has been published elsewhere (Gordon et.al 2002). Also we used (West el al., 2001) dataset for modeling the ER- and ER+ in breast cancer patients. In this dataset, tissues from 49 breast cancer patients were collected and expression of 1198 genes were recorded for each patient. This dataset was gathered by the Duke Breast Cancer SPORE tissue bank of frozen tumors. Tumors were diagnosed as invasive ductal carcinoma and their largest dimension was between 1.5 and 5 cm.25samples of these were ER+, and 24 samples were ER-.

Both data sets that mentioned above are available at:

http://algorithmics.molgen.mpg.de/Static/Supplements/CompCancer/datasets.htm.

Statistical analysis

Pre-processing phase

First, we adapted wavelet shrinkage for denoising the data.

In a way to we performed a multiple 1-D discrete wavelet transform of the data (Aminghafari et al., 2006).

Number of wavelet tried to choose the best basis and Haar wavelet selected for this section. With wavelet transform we considered noisy data as a regression model.

$$y(t)^i = f(t)^i + \varepsilon(t)^i , \quad i = 1, \dots . , n, \qquad t = 1, \dots . p$$

Where $y(t)^i$ is observed data, $\varepsilon(t)^i$is a Gaussian white noise of unknown variance σ^2, and $f(t)^i$is an unknown function and can be reconstructed through the observations.With wavelet transform as Figure1 shows, in each level, data decomposes into detail and approximation coefficient. Noise reduction was performed with thresholding detail coefficients and reconstruction the data.

Second, different wavelets such as Haar, Biorthogonal3.3, Symlet4 and Daubchies5 were applied to the data to reduce the dimension. We applied multiple 1-D wavelet decomposition to 6^{th} level. Approximation and detail coefficient in each level were used as neural network input, separately.

Artificial Neural Networks

We applied artificial neural network in order to classifying types of tumor using gene expressions.

Artificial neuarl network (ANN) is a computational model, simulated from the structure of the neuron. ANN learn the pattern between input and responses, then predic responses for new samples. There are different neural network in structure, learning and type of weighted connections. The proccesing elements are input, output and hidden layers and weighted connections (Hastie et al., 2008; Dreyfus, 2005).

A one-hidden-layer feed-forward neural network was used with 20 neurons in the hidden layer. Number of neurons in the hidden layer was obtained by trial and error to get the best performance. Sigmoid activation function was used for hidden and output neurons.

In our analysis, three phases of training, validation and test were carried out.

In the training phase, we randomly choose 60% and 70% of the data as training in lung and breast cancers, respectively. In the next step, 15% and 5% were used as validation for checking the generalizability of the network and stop training when network performance in validation set stop improving. Since our data did not have a lot of subjects, these percentages were set to have enough data for each phase. In training phase, we trained the model with training dataset and "gold standard" by comparing the gold standard with expected output and minimized the error during training.

In the test phase, we used 25% of the data for each dataset in order to check the validation of the proposed models.

In this study we used accuracy of the test, train and validation sets separately to evaluate the performance of neural network for

classification type of tumors. TN, TP, FN and FP denote the number of true negative, true positive, false negative and false positive respectively. Accuracy is defined as (TN+TP)/ (TN+FN+TP+FP)*100. Firstly we did the preprocessing with denoising and feature extraction and then feed data into the neural network and compared their accuracy.

software

All analyses of this study were performed with MATLAB ver12.

Result

Breast cancer dataset

Table 1 shows the performance of neural network for breast cancer dataset. For denoised and original datasets the accuracy was 100. Table 2 shows the classification's accuracy for dimension reduced data with Haar wavelet. We feed detail and approximation coefficient separately as input to the neural network. Different wavelet basis such as Haar, Biorthogonal3.3, Symlet4 and Daubchies5 tried for dimension reduction. They have similar result and here we presented the result of Haar wavelet. There wasn't significant difference between AUC in different levels and coefficients (P=0.399). After four and six level decomposition, number of features reached 75 and 19 respectively.

Lung cancer dataset

Table 3 demonstrates the performance of neural network for lung cancer dataset. For denoised and original datasets, the results were exactly analogous to breast cancer results. Table 4 shows the classification's accuracy for dimension reduction data using Haar wavelet. The Tables shows neural network's result for lung cancer data set in original data and data with dimension reduction has same accuracy.

Discussion

Breast cancer is the most prevalent cancer and most common cause of cancer death among females in all races. According to WHO report in 2008, 23% of the total cancer cases and 14% of the all cancer deaths among females were due to breast cancer. Lung cancer had the highest number of death from cancer, account for about 18% of all deaths in both males and females.

Timely and accurate diagnosis of the tumor type or stage in cancer will increase the chance of cure and preventing disease progression with finding appropriate treatment.

Pathological discrimination between MPM and AD in lung cancer and detection ER+ from ER- in breast cancer with current methods is hard and misleading and may lead to misclassification and tumors with same classes respond differently to the same treatment.

Estrogen receptor-positive cancers suggested that the growth of tumor cells, like other breast cells, is influenced by estrogen hormones. There are different receptors on the cell's surface and in their nucleus and cytoplasm. Hormones and other chemical messengers bind to receptors and lead to cell growth. It is important to distinguish the hormone receptor-positive and negative to find the best therapeutic measure. Hormone therapy in receptor-positive cancers prevents cancer cell growth by inhibition reaching estrogen to cancer cells. Hormone therapy could not used for receptor-negative breast cancer because they do not have hormone receptors.

In this paper we used multilayer perceptron neural network for classification of tumor types and subtypes. Two datasets with different sample sizes were used. Results showed that artificial neural networks are well able to classify both datasets. Due to limitations of working with high-dimensional data, the discrete wavelet transform is used for feature extraction. Detail and approximation coefficients obtained from the wavelet transform, applied as neural network input.

In lung cancer data set, the accuracy of neural network in both original dataset and dimension reduced dataset was complete. In original data set, 1626 gene expressions (variable) were used. The accuracy of the neural network did not change with reduction variables to 26 features. While in breast cancer dataset the accuracy of the neural network was reduced with reduction the dimension of data. The difference between these two datasets was in their sample size. Comparison of the results suggested that in datasets with large sample size, reduction in dimension to very low levels, don not reduce the neural network's accuracy. Instead of using a large number of gene expressions, we could apply a few numbers of features obtained wavelet transform without accuracy reduction.

In 2012, (Nikumbh et al., 2012) applied Biogeography–based Optimization for dimension reduction, Support Vector Machine (SVM) and Random Forests (RF) for classification of west breast cancer data set. They suggested that accuracy of classification with 10-fold cross validation with SVM and reduce gene expressions to 15 is 99.56%, and accuracy for RF with 20 gene expressions is 94.38%.

In 2006, (Hu et al., 2006) studied on several classification methods for discrimination MPM and AC in lung cancer and compared the accuracy of these methods including C4.5, RF, AdaBoost C4.5, Bagging C4.5 and LibSVMs. The average of accuracy with 10-fold cross validation in RF is 99.5%, in C4.5 is 98.3%, AdaBoost is C4.5, Bagging C4.5 is 97.8% and with LibSVMs is 100%.

Based on the results of this study, we indicated that artificial neural network in comparison with other classification methods is a more accurate method and wavelet transform is an efficient way for reducing data dimension without lossing important information.

Conclusion

This study indicated that adapting of classification methods for tumor types detection with gene expressions is an objective and exact procedure compared with traditional methods based on morphological appearance. Artificial neural network could correctly classify the ER+ , ER- in breast and AD , MPM in lung cancer patients. Also this paper indicated that insist of high number of genes, fewer features can be used without any reducing in artificial neural network performance.

Acknowledgments

This is a part of MSc thesis in Biostatistics. The authors would like to thank to all the department staff for their valuable advice and guidance during the project.

Conflict of interest statement

The authors have no conflict of interests to declare.

Funding

The authors thank the Deputy of Research and Technology of Hamadan University of Medical Science for approving the project and providing financial support.

References

Aminghafari, M., N. Cheze, and J.-M. Poggi, Multivariate denoising using wavelets and principal component analysis. Comput. Stat. Data Anal, 2006. 50(9): p. 2381-2398.

Arpino, G., et al., Gene expression profiling in breast cancer: A clinical perspective. Breast, 2013.

Cummings, M.C., et al., Molecular classification of breast cancer: is it time to pack up our microscopes? Pathology, 2011. 43(1): p. 1-8.

de Ronde, J.J., et al., Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response. Breast Cancer Res Treat, 2010. 119(1): p. 119-26.

Demir, C. and B. Yener, Automated cancer diagnosis based on histopathological images: a systematic survey. 2005, Citeseer: Rensselaer Polytechnic Institute.

Draghici, S., Data analysis tools for DNA microarrays. Vol. 4. 2010, London: Chapman & Hall /CRC Press.

Dreyfus, G., Neural Networks Methodology and Applications. 2005, New York: Springer.

Globocan. 2008 January 2013 [cited 2013; Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.

Gordon, G.J., et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res, 2002. 62(17): p. 4963-7.

Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. 2008, California.

Hejmadi, M., Introduction to cancer biology. 2010, ventus publishing.

Hu, H., et al. A Comparative Study of Classification Methods for Microarray Data Analysis. in Proc. Fifth Australasian Data Mining Conference 2006. Australia.

Iyer, A.K., et al., Exploiting the enhanced permeability and retention effect for tumor targeting. Drug Discovery Today, 2006. 11(17): p. 812-8.

Lipshutz, R., D. Morris, and M. Chee, Using oligonucleotide probe arrays to access genetic diversity. Biotechniques, 1995. 19(3): p. 442-7.

Mitra, P.S., et al., Analysis of the toxicogenomic effects of exposure to persistent organic pollutants (POPs) in Slovakian girls: Correlations between gene expression and disease risk. Environ Int, 2012. 39(1): p. 188-99.

Nanni, L., S. Brahnam, and A. Lumini, Combining Multiple Approaches for Gene Microarray Classification. Bioinformatics, 2012. 28(8): p. 1151-7.

Nikumbh, S., S. Ghosh, and V.K. Jayaraman, Biogeography-Based Informative Gene Selection and Cancer Classification Using SVM and Random Forests, in IEEE World Congress on Computational Intelligence. 2012: Brisbane, Australia

Saini, A., J. Hou, and W. Zhou, Hub-Based Reliable Gene Expression Algorithm to Classify ER+ and ER-Breast Cancer Subtypes. Int J Biosci Biochem Bioinforma, 2013. 3(1).

Schena, M., et al., Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science, 1995. 270(5235): p. 467-70.

vittingoff, E., et al., Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models 2012, USA: Springer.

West, M., et al., Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A, 2001. 98(20): p. 11462--7.

Table 1: Performance of the neural network for breast cancer dataset.

Datasets	**N**umber of gene expressions	Accuracy of train **(%)**	Accuracy of test **(%)**	Accuracy of validation **(%)**
Original Data	1198	100	100	100
Denoised Data	1198	100	100	100

Table 2: Classification's accuracy for dimension reduced data in breast cancer with Haar wavelet

Decomposition Level	**Coefficient**	**Number of coefficient**	**Classification Accuracy (%)**	**AUC**
4	Approximation	75	98	0.99
4	Detail	75	98	0.99
5	Approximation	38	96	0.96
5	Detail	38	98	0.99
6	Approximation	19	98	0.99
6	Detail	19	94	0.93

Table 3: Performance of the neural network for lung cancer dataset.

Datasets	N	Atr (%)	Ate (%)	AV (%)
Original Data	1626	100	100	100
Denoised Data	1626	100	100	100

Table 4: Classification's accuracy for dimension reduced data in lung cancer with Haar wavelet

Decomposition Level	Coefficient	Number of coefficient	Classification Accuracy (%)	AUC
4	Approximation	102	100	1
4	Detail	102	100	1
5	Approximation	51	100	1
5	Detail	51	100	1
6	Approximation	26	100	1
6	Detail	26	100	1

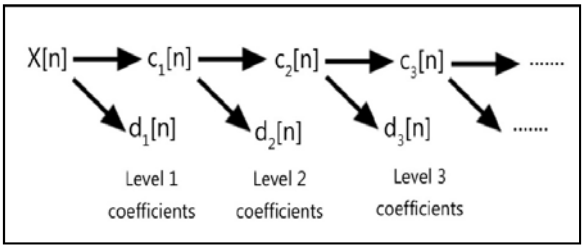

Figure 1: data decomposition into detail (d_i) and approximation coefficient (c_i)