International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12319-12321
© Research India Publications. http://www.ripublication.com

Application of Machine Learning to Determine the Characteristics of Adjacent Normal Tissues in Liver Cancer

Wafaa Khazaal Shams[1], **Zaw Z. Htike**[2]

[1]*Faculty of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*
[2]*Faculty of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*

Abstract

This study applies machine learning methods to gene expression data from normal tissue of patients with liver cancer to predict whether this tissue is 'healthy', 'cirrhotic' (liver damage), 'non tumor', or 'tumor'. The method is based on using Principle Component Analysis (PCA) combined with the Regularized Least Squares (RLS) classifier. The results show a high accuracy with 10-fold cross validation for discrimination among tissue types. Results indicate the capability of gene expression profiling to successfully discriminate between tumor tissue and normal tissue, however there is a clear and strong overlap between non-tumor tissue and cirrhotic tissue. Further, we used the same classification model to predicate the probability of detecting each class separately. Tumor gene expression can be predicated successfully.

Keywords: Adjacent normal tissue; cancer classification; PCA; RLS.

INTRODUCTION

Machine learning techniques and statistical methods are widely used in gene expression array analysis [1-4]. Gene expression arrays contain thousands of genes extracted from tissues under different conditions. Analysis of these arrays is not trivial task, even with simple processes. Machine learning makes such processes easier and more efficient, including feature selection methods to identify the relevant genes, feature reduction to reduce array size, and classification methods to assign class labels to cell samples with unknown biological conditions.

One of the most prevalent cancers in Asia is hepatocellular carcinoma (HCC) which is the leading cause of death in this region. Until now the only effective treatment has been surgery [5]. Recently studies done on HCC report that tumor expression profiles can be used for cancer classification [6-7]. Generally, cancer cells have genomic instability, and therefore identifying gene instability is important for early detection of tumor tissue. However, these changes may be simply genomic changes that not develop into tumor tissue. Lamb[7] studied large scale genomic changes that are found between adjacent tissue (AN) and tumor tissue (TU) samples and found that 75% of the gene expression array had expression differences between these types of tissues. Lamb [7] tested the connectivity of the gene network of AN and TU to find the changes that are relevant to the progression of the tumor and separate them from the changes that are simply genomic changes that are not relevant to tumor formation.

In this study, we used Regularized Least Squares (RLS) to distinguish tumor tissue from other type of tissue in the gene expression array. Because of the huge amount of data in the gene array, PCA was applied to reduce the gene array size. PCA is a known method for data reduction which reserves the same information with fewer components. Next we tried to assess how genes can be closed from each other and the type of gene, based on a blind test.

PRINCIPAL COMPONENT ANALYSIS

PCA is a familiar statistical method for data reduction that transforms a number of correlated variables into a number of uncorrelated variables, called principal components [8], that allows reduction of the data dimensions while preserving information on the variable interactions [9]. The data is transformed into a new coordinate space. Hence, the first principal component has the most information and the highest variability. In the new space, gene expression appears to be classified into groups, as show in Figure 1.

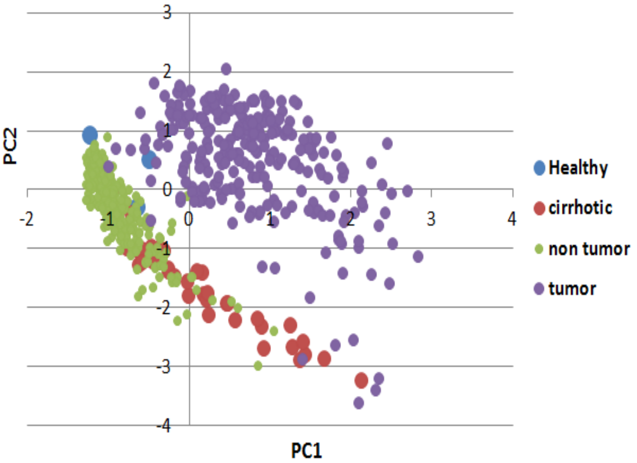

Figure. 1: A scatter plot of the first and second principle components for the data set.

REGULARIZED LEAST SQUARES (RLS)

Regularized Least Squares (RLS) is one of the machines learning systems used in this study. RLS is also known as the Tikhonov regularization problem, with a square loss function that minimizes the given equation [10].

$$\min_{f\in\boldsymbol{H}} \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda\|f\|_{\mathrm{K}}^2\right)$$
(1)

where n is the number of samples, X_i is the data samples of I of the training set, Y is the binary outcome, $\|f\|$ is the norm of f (the expected values) in Hilbert space defined by kernel K, λ is the regularization parameter that is computed from the kernel of the training data. We use Gaussian kernel K [11].

$$K(X,X) = e^{\frac{\|x_i - x_j\|^2}{\sigma^2}} \qquad (2)$$

The solution of equation (13) is given by:

$$f(X) = \sum_{i=1}^{\ell} C_i \mathrm{K}(X, X_i) \qquad (3)$$

where $c = (\mathrm{K} + \lambda\ell\mathrm{I})^{-1}Y$, (4)

I is the identity matrix. C values depend on λ. In this study we utilized the toolbox from MIT [12]. λ and sigma values were determined omitting one cross validation for the training sets. After computing C, the outcome predication values of the test data will be:

$$f(X_t) = \sum_{i=1}^{\ell} C_i \mathrm{K}(X_t, X_i). \qquad (5)$$

RESULTS AND DISCUSSION

To investigate the ability of gene expression arrays to identify human liver tumor and adjacent normal tissue, we combined tissues from liver tumor, adjacent normal tissue (AN), healthy tissue, and cirrhotic (liver damage) tissue and then analysed this sample with the classification process. As mention in Section 2, we used PCA to reduce the gene array. PCA has been used from whole gene expression arrays. A 10-fold cross validation was used to evaluate the classification process, hence the data was divided into 10 sets, with nine sets used for training and the remaining set used for testing. Next, the average accuracy for the 10 fold analysis was computed for each tissue type, as well as for the sample with misclassification. Table 1 shows the confusion matrix. Obviously, the accuracy for detecting healthy genes and tumor tissue are the highest at 100% and 99.6%, respectively. The accuracy for detection of AN tissue and damaged tissue are 97 % and 90%, respectively. Similarly, Table 2 show the confusion matrix for detection pf each tissue types using the whole gene expression array without reduction of the data set. The results are extremely similar to Table 1. This indicates that PCA is suitable for conservation of information in this type of data set. PCA was not compatible with all type of data sets, especially when they are not linear.

Table 1: Average accuracy for detection of each class using PCA with RLS.

	healthy	'cirrhotic	non-tumor	tumor
healthy	100	0	0	0
'cirrhotic (liver damage 2	2.5	90	7.5	0
Non-tumor	1.2	1.2	97	0.4
tumor	0	0	0.3	99.6

Table 2: Average accuracy of detection for each class using whole gene array with RLS.

	healthy	'cirrhotic	non-tumor	tumor
healthy	100	0	0	0
'cirrhotic (liver damage 2	2.5	85	10	2.5
Non-tumor	0.8	1.2	97.2	0.8
tumor	0	0	0.3	99.6

From both tables, one can see that the TU gene has significant characteristics that can be recognized from others tissue types by HCC. There is also a strong overlap between AN tissue and damaged tissue. This was shown in a scatter plot (Figure 1) using PCA. To achieve our objective in this study, we used another classification model that is based on how the data are closed to each other. For this we used three tissue types as training and the left one type for use in testing. In this case we label the classes by integer number as (1, 2, 3, 4) and used 10 fold cross validation to validate the used labels. The classification data resulted in the same analysis as in the binary labels. Figure 2 shows the predicted values of non-tumor tissue. Clearly, most the samples are in Classes 1 and 2, which are healthy and damage tissue, respectively. Figure 3 shows the predicted value of tested tumor tissue. Clearly, it doesn’t belong to any class, hence the value occurring near zero. Figure 4 shows the predicted value for damage tissue. The values show a clear distribution among the classes, non-tumor, tumor, and healthy tissue.

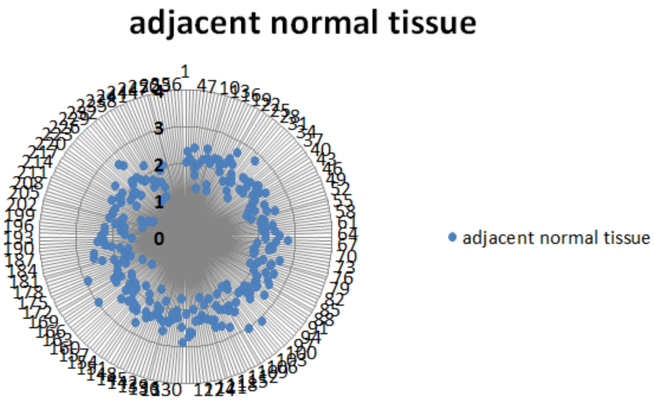

Figure 2: The scatter plot of the predict values for the adjacent normal tissue (non-tumor).

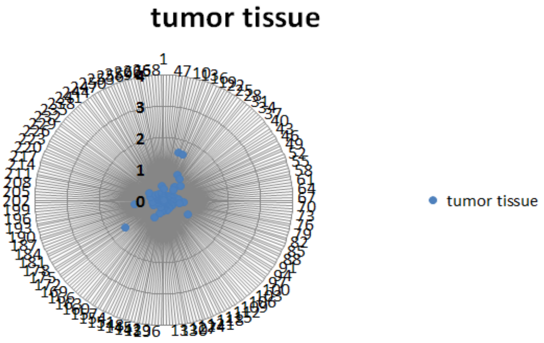

Figure 3: The scatter plot of the predicted values for the tumor tissue.

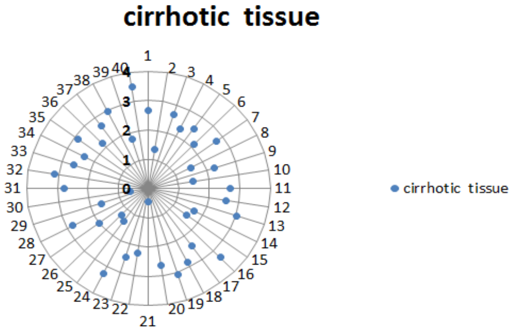

Figure 4: The scatter plot of the predicted value for the cirrhotic tissue.

CONCLUSION

In this study we applied a RLS classifier with PCA to a binary classification problem with gene expression array of tumor liver tissue. The model showed significant accuracy, around 99%, in the detection of tumor tissue from adjacent tissue, damaged tissue, and healthy tissue. However, there was strong overlap between adjacent and damage tissue. Another test has already been done to predict how these tissue relate to each other, based on their gene array data. Results indicate that gene arrays for tumors have significant characteristics that identify tumor tissue. However the non –tumor tissue was indiscernable from damage tissue. Our results did not show correlation between genes of adjacent tissue and tumor tissue.

REFERENCES

[1] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906-914.

[2] Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS one, 7*(7), e39932.

[3] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1-3), 389-422.

[4] 4. Chang, S.-W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics, 14*(1), 170.

[5] Trevisani F, Cantarini MC, Wands JR, Bernardi M (2008) Recent advances in the natural history of hepatocellular carcinoma. Carcinogenesis 29: 1299–1305.

[6] Kim,J.W., Sime,J., Forgues,M., He,P., Ye,Q., Kaul,R. and Wang,X.W. (2002) Molecular characterization of preneoplastic liver diseases by cDNA microarray. Proc. Am. Assoc. Cancer Res., 43, 461–462.

[7] Lamb, J. R., Zhang, C., Xie, T., Wang, K., Zhang, B., Hao, K., . . . Ferguson, M. (2011). Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PloS one, 6*(7), e20090-e20090.

[8] Yeung KY, Ruzzo WL 2001 Principal component analysis for clustering gene expression data. Bioinformatics 17:763–774.

[9] Joliffe,I.T. (1986) *Principal Component Analysis*. Springer, New York

[10] 10.Evgeniou, T., M. Pontil, and T. Poggio, *Regularization networks and support vector machines.* Advances in Computational Mathematics, 2000. **13**(1): p. 1-50.

[11] Aronszajn, N., *Theory of reproducing kernels.* Trans. Amer. Math. Soc, 1950. **68**(3): p. 337-404.

[12] Tacchetti, A., et al., GURLS: a Toolbox for Regularized Least Squares Learning. 2012.