November-December 2019
ISSN: 0193-4120 Page No. 2307 - 2314

Feature Selection and Machine Learning Method for Classification of Lung Cancer Types

Byungju Shin[1], Bohyun Wang[2], Joon S. Lim[3]*
[1]Department of Computer Science, Gachon University, Korea, phillogic@naver.com
[2] Department of Computer Science, Gachon University, Korea, bhwang99@hanmail.net
[3] Department of Computer Science, Gachon University, Korea, jslim@gachon.ac.kr
* Corresponding Author

Article Info
Volume 81
Page Number: 2307 - 2314
Publication Issue:
November-December 2019

Article History
Article Received: 5 March 2019
Revised: 18 May 2019
Accepted: 24 September 2019
Publication: 12 December 2019

Abstract
Microarray technology and computational methods have enabled researchers to obtain significant amount of gene expression data for lung cancer, which have allowed them to select genes that are specific to particular types of lung cancer. In this paper, a relational matrix is proposed, which is used to find genes with high correlations, depending on the types of lung cancer by using microarray expression data. We perform machine learning on the genes discovered from the relational matrix by using the weighted neuro-fuzzy algorithm to accurately classify the types of lung cancer. In addition, some genes among the discovered genes were investigated in the relative pathways, and p-values were obtained to analyze the validity of those genes in the given pathways. The relational matrix is constructed by enumerating the number of meaningful relationships identified through observations of the changes in gene expression values between different types of lung cancer. The weighted neuro fuzzy algorithm uses a bounded sum function into which the three functions are combined during learning and classification. We obtained 405 type-dependent genes using the proposed relational matrix and classified 203 samples into five types of lung cancer by using those genes. We obtained a classification accuracy of 99.5% for all samples; the results of Leave One Out Cross Validation test showed an accuracy of 87.19%. Moreover, we obtained valid p-values from 12 pathways in KEGG.

Keywords: *Feature Selection, Lung Cancer Types, Relational Matrix, Weighted Neuro Fuzzy Algorithm*

1. INTRODUCTION

Cancer diagnosis through morphological observation often results in misdiagnosis [1]. Cancer is a complex genetic disease involving many genes, proteins, and pathways. Therefore, using genetic information when diagnosing cancer or sub-classifying several types of cancer is an effective approach. Before the advent of microarray technology, research was focused on one protein or a small number of genes. However, because of advances in microarray technology, researchers were able to obtain vast amounts of gene expression data. Moreover, computational methods have enabled researchers to process large amounts of these genetic data.

It is important to distinguish between cancerous and non-cancerous tumors using genetic information, but it is also important to judge the subtypes of a cancer upon occurrence, to facilitate personalized medicine. If genes that can classify cancer subtypes are discovered from tens of thousands of microarray gene data without noise, it will be very helpful for cancer therapy; however, finding the subsets corresponding to

November-December 2019
ISSN: 0193-4120 Page No. 2307 - 2314

the cancer types from a vast amount of gene information remains a challenging research topic.

Globally, lung cancer is the leading cause of cancer-related mortalities. Lung cancer can be categorized into several sub-types. Different types of tumors are associated with different types of lung cancers; these different types of tumors may have different histological characteristics and clinical results, such as drug response [2]. The lung cancer subtypes are not easily distinguishable, and proper classification is a critical issue during pre-therapy [2].

Over the past few decades, researchers have proposed many computational methods for processing, classifying, and analyzing very large amounts of microarray expression data. SVM [3], clustering [4], k-nn [5], and so on have been used for classification of the expression data, which are similar to statistical methods [1].

In this paper, we use a combination of a statistical method and machine learning to improve classification accuracy. We discovered gene groups corresponding to lung cancer subtypes, and investigated those genes in several relative pathways. Then, we obtained p-values for those pathways. In addition, we classified lung cancer types using the obtained genes and evaluated them using the Leave One Out Cross Validation(LOOCV) method [6] using a weighted neuro fuzzy algorithm. In this study, we use the relational matrix as the statistical method for extracting genes that can be used to classify lung cancer subtypes. The relational matrix is constructed by enumerating the number of meaningful relationships identified through the observation of changes in the expression values of genes between lung cancer subtypes.

A meaningful relationship indicates an identical trend or a reverse trend among genes. The meaningful relationship is calculated as a score and is stored in each element of the generated relational matrix. The relational matrix enables the discovery of groups of genes associated with lung cancer types on the basis of the relationship extraction method; such groups are used to classify lung cancer gene samples into subtypes through machine learning, by using the neuro-fuzzy algorithm.

We discovered more than 400 genes that have strong relationships with their corresponding subtypes by investigating the matrix after constructing the proposed relational matrix. When the types of lung cancer were classified using the weighted neuro fuzzy algorithm based on those genes, accuracy was high. The classification results showed an accuracy of 99.5% for five types of genes, and the accuracy of the LOOCV test was 87.19%.

The remainder of this paper is organized as follows. We propose the relational matrix and the weighted neuro fuzzy algorithm in Section 2. The experimental results and our conclusions are presented in Section 3 and Section 4, respectively.

2. MATERIALS AND METHODS

2.1 Samples

In experiment, we used the 12,600-gene expression profile of 203 lung cancer samples. Among the 203 samples, 17 samples were non-cancerous, 139 were adenocarcinomas, 6 were small cell lung cancer, 21 were squamous cell lung carcinomas, and 20 were pulmonary carcinoids. Only 3312 genes of 12,600 genes were used for learning [8], [9]; these genes are considered as the most valuable genes for this analysis [7]. Therefore, the size of the dataset was 3312 × 203.

2.2 Relational Matrix

The relational matrix is an R × R table, in which R indicates the number of genes, with relational information among genes. The types of lung cancer are added to the relational matrix as genes for the purpose of identifying the relationship

November-December 2019
ISSN: 0193-4120 Page No. 2307 - 2314

between the cancer type and the gene. Each element of the matrix is set to five digits, from 0 to 4. The relational matrix (referred to as RM) is constructed as follows:

Step 1. Preprocessing of gene expression values:

All gene expression values are transformed from 0 to 4. Table 1 shows the pseudo code for the transformation processing of gene expression values. The minimum expression value is subtracted from the maximum expression value in all samples of one gene, and then the result is divided by 5 in line 02. This indicates that the interval between values that a gene can have in all samples is separated by 5 in order. Here, 5 can be regarded as a heuristic. When a number smaller than 5 was used, a loss of information was observed. For a number greater than 5, the result was identical; however, the computational complexity increased.

Table 1. Pseudo Code for Transformation Processing of Gene Expression Values

```
// i is the ith gene of 3312 genes
// j is the jth sample of 203 samples
// SN is the number of samples
// g_i is ith gene , g_ij is the jth sample of ith gene, tg_ij is the transformed value of g_ij
// Max(g_i) is maximum expression value of all samples of gene i
// Min(g_i) is minimum expression value of all samples of gene i
01  For i = 1; i <= 3312 ; i++
02      add = (Max(g_i) - Min(g_i)) / 5
03      f = Min(g_i), t = f + add
04      For j = 1 ; j <= SN   ; j++
05          For k = 0 ; k < 5   ; k++
06              if f ≤ g_ij < t then set tg_ij to k
07              f = t, t = t + add
08          End for
09      End for
10 End for
```

From line 03 to 09, the interval that includes each sample value of one gene is searched. In accordance with the interval in which the expression value is included, an expression value is set from between 0 and 4. Table 2 shows an example of the transformed values. In table 2, the lung cancer type was added in the last column; therein, type is considered as a gene for identifying the relationship between types and genes in step 2 (the values in the type column use values between 0 to 4 for the five types of lung cancer).

Table 2. The Example of Transformed Values

	tg1	***tg2***	**…**	***tg3312***	**type**
s1	0	3	…	4	0
s2	4	0	…	3	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
s203	2	1	…	2	4

Step 2. Computing the relationship score between two genes:

In the proposed matrix, the score of the relationship between two genes *tgi* and *tgj* is computed. The relationship score *ri,j* is calculated using Eqn. (1).

$$r_{i,j} = \sum_{n=1}^{202}\left(\sum_{m=n+1}^{203} f(tg_{i,n\to m}, tg_{j,n\to m})\right) \quad (i,j = 1,2,\ldots,3312) \qquad (1)$$

where *tgi* and *tgj* are the transformed values of gene *i* and gene *j*, respectively. The *tgi,n->m* represents *tgi,n* – *tgi,m* and is indicative of the change between the *n*th and the *m*th sample values of *tgi*. The function *f*(*tgi,n->m* , *tgj,n->m*) returns the value, which indicates the relationship between *tgi* and *tgj*. The function can be represented as Eqn. (2). If both *tgi,n->m* and *tgj,n->m* are positives or negatives, it is considered that *tgi* and *tgj* have an activator relationship because *tgi* and *tgj* are changed in a manner that is similar to the trend between the *n*th sample and the *m*th sample, and the function returns 1. If only one of *tgi,n->m* and *tgj,n->m* are positive or negative, it is considered that *tgi* and *tgj* have a repressor relationship because as reverse change trend, and the function returns -1. Empirically, it is considered that the change is

November-December 2019
ISSN: 0193-4120 Page No. 2307 - 2314

meaningful only when $tgn->m$ is greater than 2 or is equal to 2.

$$f(tg_{i,n\to m}, tg_{j,n\to m}) = \begin{cases} 1, \ if \ (tg_{i,n} - tg_{i,m})(tg_{j,n} - tg_{j,m}) \geq 4 \\ -1, \ if \ (tg_{i,n} - tg_{i,m})(tg_{j,n} - tg_{j,m}) \leq -4 \\ 0, \ else \end{cases} \quad (2)$$

For finding genes that change together in type change, Eqn. (3) and (4) are used.

$$r_{type,i} = \sum_{n=1}^{typeA}(\sum_{m=1}^{typeS} f(type_{n\to m}, tg_{i,n\to m})) \quad (i = 1,2,\ldots,3312) \quad (3)$$

where typeA is the number of samples contained in any lung cancer type, and typeS is the number of all samples except typeA.

In formula (2), 1 indicates an activator relationship, and -1 indicates a repressor relationship; however, 1 and -1 in Eqn. (4) are not indicative of such relationships.

$$f(type_{n\to m}, tg_{j,n\to m}) = \begin{cases} 1, \ if \ (type_n - type_m)(tg_{i,n} - tg_{i,m}) \geq 2 \\ -1, \ if \ (type_n - type_m)(tg_{i,n} - tg_{i,m}) \leq -2 \\ 0, \quad else \end{cases} \quad (4)$$

Table 3. The constructed relational matrix

	tg1	***tg2***	***…***	***tg3312***	***t0***	***t1***	***t2***	***t3***	***t4***
tg1	*r1,1*	1500	…	400					*rtype,i*
tg2	1500		…	20000				50000	
. . .	. . .	. . .	. . .	. . .					
tg3312	400	1	…	30000					
t0									
t1									
t2									
t3		50000							
t4	*rtype,i*								

Step 3. Applying the threshold:

After generating table 3 in step 2, a threshold is applied to each score of the matrix. If the value obtained by dividing the score by the maximum number that the score can have is smaller than the threshold, it is considered that the value is not valid and the value is replaced by 0. Otherwise, the value is replaced by 1. We empirically determined the threshold as 0.7 [23-26].

Step 4. Finding a strong relationship group:

If the element of matrix, ri,j was set to 1, two genes *gi* and *gj* interact with each other. For $rtypes,j$, it is same. We were able to identify the group of genes related to one type of lung cancer by scanning the relational matrix.

3. EXPERIMENTAL RESULTS

We obtained genes with strong relationships corresponding to the lung cancer types from the relational matrix and obtained p-values in several pathways for those genes. In addition, we performed machine learning using the neuro fuzzy algorithm in order to classify the lung cancer types. We evaluated the classification results by using the LOOCV method.

Table 4 shows the number of genes that can represent each lung cancer type. 'Normal' indicates that it is not lung cancer. For example, when the normal samples were compared with the other samples, 74 genes showed significant changes. We did not find any genes for adenocarcinomas. This may be attributed to the very large number of samples, relative to the number of other types of samples. Nevertheless, a total of 405 genes were able to accurately distinguish the five types. We obtained an accuracy of 99.5% when both training and learning were performed for classifying the samples into the five types. Moreover, the accuracy of the LOOCV test was 87.19%.

Table 4: The number of genes with high interactions according to the types of lung cancer

Types of lung cancer	The number of genes depending on types
Adenocarcinomas	0
Normal	74
Small cell lung cancer	79
Squamous cell lung carcinomas	5
Carcinoid	247

We obtained valid p-values for three lung cancer types in the 12 pathways found in KEGG [10]. Table 5 shows the pathways and p-values. The genes included in Normal, Small Cell Lung Cancer, and Carcinoid types were discovered in pathways and the p-values were shown, as listed in table 5, for each pathway.

Table 5: The number of genes with high interactions according to the types of lung cancer

Lung cancer types	Pathway	Count	%	P-value
normal	Malaria	5	7	0.00017
	Cell adhesion molecules (CAMs)	4	5.6	0.048
small cell lung cancer	Pathogenic Escherichia coli infection	8	11.3	0.000000014
	Gap junction	7	9.9	0.000011
	Cell cycle	7	9.9	0.00008
	Phagosome	7	9.9	0.00025
	DNA replication	4	5.6	0.0012
carcinoid	Synaptic vesicle cycle	10	4.3	0.00000014

	Vibrio cholerae Infection	8	3.4	0.0000064
	Oxidative phosphorylation	11	4.7	0.000011
	Insulin secretion	8	3.4	0.00014
	Vasopressin-regulated water reabsorption	6	2.6	0.0003

4. CONCLUSION

In this study, we classified the types of lung cancer with high accuracy through the proposed relational matrix. The relationships between lung cancer type and genes as well as the interaction among genes were identified using the relational matrix. In addition, p-values were obtained in some pathways. If the threshold and the number of intervals are changed, we conjecture that diverse relationships between lung cancer subtype and genes will be discovered.

A wide variety of pathways related to the function of lung cancer are known, but a new and important pathway is emerging recently. Our study focused on finding genes or pathways that differentiate between lung cancer subtypes, rather than identifying specific pathways or genes for individual subtypes of lung cancer using the weighted neuro fuzzy algorithm. Thus, 405 genes or related pathways should be interpreted as the distinguishability of the subtypes in lung cancer.

In the future, we will attempt to reduce the number of genes required for the classification of lung cancer types and to increase the p-values in the pathways.

ACKNOWLEDGEMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2019-2017-0-01630) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

1. Xiaosheng Wang, and Osamu Gotoh, **Microarray-Based Cancer Prediction Using Soft Computing Approach**, Cancer Inform, 7, 2009, pp. 123–139.
2. Hongye Liu, Alvin T. Kho, Issac S. Kohane, and Yao Sun, "**Predicting Survival within the Lung Cancer Histopathological Hierarchy Using a Multi-Scale Genomic Model of Development**," PLOS MEDICINE, 3(7), 2006, pp.1090–1102.
3. Komura D1, Nakamura H, Tsutsumi S, Aburatani H, and Ihara S, "**Multidimensional support vector machines for visualization of gene expression data**," Bioinformatics. Feb 2005, 15;21(4):439-44. SVM.
4. Sultan M1, Wigle DA, Cumbaa CA, Maziarz M, Glasgow J, and Tsao MS abd Jurisica I, "**Binary tree-structured vector quantization approach to clustering and visualizing microarray data**," Bioinformatics. 2002;18 Suppl 1:S111-9.
5. Paul TK1, and Iba H, "**Gene selection for classification of cancers using probabilistic model building genetic algorithm**," Biosystems. Dec 2005 ,82(3), pp.208-25.
6. Ben-Dor A1, Bruhn L, Friedman N, Nachman I, Schummer M, and Yakhini Z, "**Tissue classification with gene expression profiles**," J Comput Biol. 2000,7(3-4), pp.559-83.
7. Arindam Bhattacharjee, William G. Richards, Jane Staunton, el al, "**Classification of human**

November-December 2019
ISSN: 0193-4120 Page No. 2307 - 2314

lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," Proc Natl Acad Sci U S A., 98(24), 2001, pp.13790–13795.

8. SY Son, SH Lee, KY Chung, and JS Lim, "**Feature selection for daily peak load forcasting using a neuro-fuzzy system**," Multimed Tools Application, 74(7), 2015, pp.2321-2336.
9. Sang-Hong Lee, and Joon S. Lim, "**Forecasting KOSPI based on a neural network with weighted fuzzy membership functions**," Expert System with Applications, 38(4), 2011, pp.4259-4263.
10. http://www.genome.jp/kegg/.
11. E. Brambilla, and A. Gazdar, "**Pathogenesis of lung cancer signaling pathways: roadmap for therapies**," Eur Respir J. Jun 2009, 33(6), pp.1485–1497.
12. Roozbeh Manshaei, and Pooya Sobhe Bidari, "**Hybrid-Controlled Neurofuzzy Networks Analysis Resulting in Genetic Regulatory Networks Reconstruction**," ISRN Bioinformatics, vol. 2012, 2012, pp.1–16.
13. K. Cho, S. Choo, S. Jung, J. Kim, H. Choi, and J. Kim, "**Reverse engineering of gene regulatory networks**," IET Syst. Biol., 1(3), 2007, pp.149–163.
14. Geeta, R. B., Shobha, R. B., Totad, S. G., & PVGD, P. R. (2014). **Web Pages Categorization Based on Classification & Outlier Analysis through FSVM**. Review of Computer Engineering Research, 1(1), 19-30.
15. I. A. Maraziotis, A. Dragomir, and A. Bezerianos, "**Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks**," IET Syst. Biol., 1(1), 2007, pp.41-50.
16. Takagi T, and Sugeno M, "**Fuzzy identification of systems and its applications to modeling and control**," System Man and Cybernetics IEEE Trans., SMC-15(1), 1985, pp.116-132.
17. L. A. Soinov, M. A. Krestyaninova, and A. Brazma, "**Towards reconstruction of gene networks from expression data by supervised learning**," Genome Biology, 4(1), 2003, article R6.
18. H. Jun, and M. Claudio, "**The influence of the sigmoid function parameters on the speed of backpropagation learning**," Computational Models Of Neurons And Neural Nets, 930, 1995, pp.195–201.
19. S. Y. Kim, S. Imoto, and S. Miyano, "**Inferring gene networks from time series microarray data using dynamic Bayesian networks**," Briefings in Bioinformatics, 4(3), 2003, pp.228–235.
20. M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "**KEGG for representation and analysis of molecular networks involving diseases and drugs**," Nucleic Acids Research, 38(1), 2009, pp.D355–D360.
21. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "**KEGG for integration and interpretation of largescale molecular datasets**," Nucleic Acids Research, 40(1), 2012, pp.D109–D114.
22. Ponzoni, Ignacio, et al., "**Pathway network inference from gene expression data**," BMC systems biology , 8.Suppl 2, 2014, S7
23. Sarma, U.; Karnitis, G.; Zuters, J.; Karnitis, E. 2019. District heating networks: enhancement of the efficiency, Insights into Regional Development 1(3): 200-213. https://doi.org/10.9770/ird.2019.1.3(2)
24. Wichitsathian, S., Nakruang, D. 2019. Knowledge integration capability and entrepreneurial orientation: case of Pakthongchai Silk Groups Residing. Entrepreneurship and Sustainability Issues, 7(2), 977-989. http://doi.org/10.9770/jesi.2019.7.2(13)
25. Jabarullah, N.H. & Othman, R. (2019) Steam reforming of shale gas over Al2O3 supported Ni-Cu nano-catalysts, Petroleum Science and Technology, 37 (4), 386 – 389.

26. Hussain, H.I., Kamarudin, F., Thaker, H.M.T. & Salem, M.A. (2019) Artificial Neural Network to Model Managerial Timing Decision: Non-Linear Evidence of Deviation from Target Leverage, International Journal of Computational Intelligence Systems, 12 (2), 1282-1294.