Military Technical College

Kobry El-Kobbah,

Cairo, Egypt

8th International Conference

on Electrical Engineering

ICEENG 2012

Construction Path analysis Model from DNA Microarray Data

A. Mohmad Badr Al Senousy [*], B. Hesham Mohamed El-Deeb [†], C. Khaled Badran [‡] and D. Ibrahim Ali Al-Khlil[§]

Abstract: A fundamental problem in human health is to predict the effect of genes that cause disease; this is an important step to diagnosis and treatment. Also prediction gene functions are still a challenge for biologists in the post-genomic era. DNA microarray simultaneously monitors expression levels thousands of genes; the massive gene expression data provide us with unique opportunities to analyze the functional and regulatory relationships among genes.

In this paper, new approach to estimate the relationship among genes and the effects of each gene on diseases is proposed. This approach consists of four main steps: **First**, extract the subset of high informative genes. **Second**, constructing genes network; thus, we propose to predict a gene's functions according to its context graph, which is defined as the gene interaction network composed of the genes interacting directly and indirectly based on focal gene. **Third**, path analysis model is used to estimate the effect of genes on the disease and among each other. **Fourth**, obtains the relation among genes and the target (disease) by applying logistic regression analysis for predict new samples. This approach evaluated by lung cancer Microarray dataset.

The proposed path diagram fit the subset of top ranked genes expression dataset, the Goodness of Fit Index (GFI) was > 0.832.

The value of this approach is that it not only tackles the measurement problem by path analysis but also provides a visualization of the relationship among genes. The proposed approach also is useful for feature reduction, due to evaluate all genes simultaneously vs. lung state.

---

* Prof. Dr. in Sadat Academy for Management Science (SAMS), Department of Computer science, *badr_senousy_arcoit@yahoo.com*

† Dr. in Modern University for Technology and Information (M.T.I), Department of Computer science, *hmeldeeb14@yahoo.com*

‡ Military Technical College (M.T.C), Department of Computer science, *khaledBadran@hotmail.com*

§ Military Technical Collage (M.T.C), Department of Computer science, *ibrahim.alkhlil@gmail.com*

Keywords: Microarray gene expression data, Genes selection, Genes ranking, cancer, Genes network, path analysis, Pearson Correlation.

1. Introduction

A DNA microarray technique allows to simultaneously observing the expression levels of thousands of genes during significant biological processes and across collections of related samples [1], microarray data applied to make possible drug and therapeutics improvement, disease diagnosis, and comprehensible basic cell biology. The datasets from microarray analysis enables the measurement of molecular signatures of diverse cells, becomes an important application of data mining, artificial intelligence and machine learning techniques to provide bioinformatics knowledge [2].

A few numbers of genes are highly related with diseases, these genes are called informative genes, which have expression pattern strongly correlated with the diseases. By our the previous study [3], the subset of top ranked genes that gave the high classification accuracy was extracted by two techniques of Attribute selection technique, "Support Vector   Machine-Recursive Feature Elimination (SVM-RFE)[4,5] and Information Gain Attribute Selection IGAS"[6]. 10 top ranked genes gave high classification accuracy by the most popular classifiers, which belong to six main categories, (1) Bayes classifiers (Bayesian Network and Naive Bayes), (2) lazy classifier (K-nearest-neighbor), (3) rule base classifiers (PART and Decision Table, (4) function base classifiers (SVM and Artificial Neural networks ANN), (5) meta classifier (AdaBoost(c4.5) and Baggin(c4.5)), (6) single Decision Tree (C4.5).

Gene nets, is a general class of gene nets where any set of genes may have connections. Theoretically genes that interact directly or indirectly may have the same or similar functions in the biological processes in which they are involved and together contribute to the related cancer diseases. The complicated relations between genes can be clearly represented using network theory [7, 8]. Thus Pearson's correlation or Euclidean distance criterions are used to measure the similarity between genes for constructing genes network [9]. But Pearson's correlation or Euclidean distance are not able to completely capture the relationship among all candidates genes expression profiles and lung cancer simultaneously [10, 11]. Therefore, to identify interactions between genes and to estimate the relationship between genes and lung cancer path analysis model is used.

The rest of this paper organized as follows: Section 2 describes path analysis problem and related materials. Section 3 describes lung cancer microarray dataset, Section 4 presents the proposed model to constructing genes network. Section 5 describes constructing genes network in practical. Section 6 analysis and discussions. Section 7 presents a comparison between single decision tree model and proposed path analysis model. Paper concluded in section 8.

2. Path Analysis Problem

Let $X_1, X_2, \ldots, X_m$ be exogenous variables, correlated and have both direct and indirect effects (through endogenous variables $X_{m+1}, X_{m+2}, \ldots, X_n$) on dependent variable Y.

Let M be the correlation coefficient matrix among these variable.

It is required to construct graph model G (Path diagram) depending on underlying relationship between these variables

Path analysis function calculates the path coefficient between exogenous and endogenous variables, to predict the effect variables on the dependant variable Y.

An example of the different models of trait effects on output y is illustrated in fig. 1. Fig. 1 (A). shows a multiple regression model where each trait operates simultaneously on fitness y. Fig. 1 (B). is the path analysis model showing four traits at four time periods.
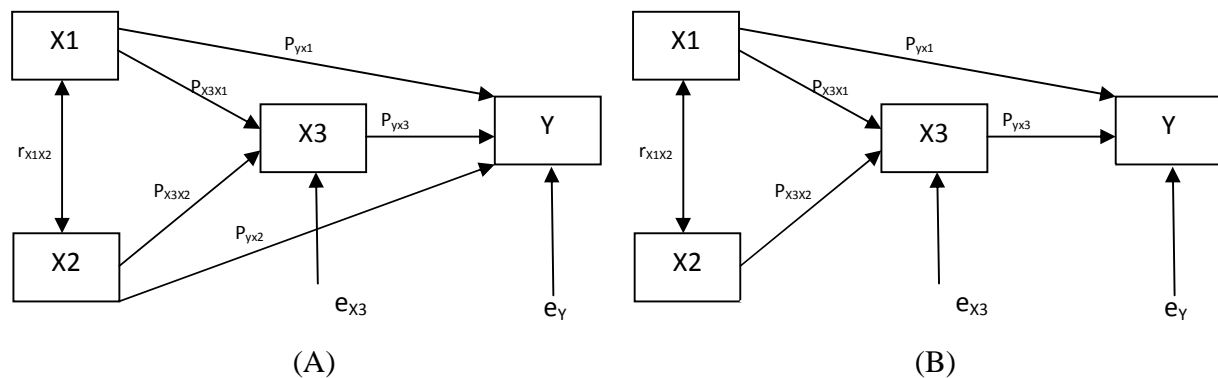

(A)   (B)

Fig. 1. An example of a general recursive causal system, (A) Illustration of possible relations among variables and output. (B) The result of significant variable.

In the model fig. 1A, P_{YX1}, P_{X3X1}, P_{X3X2}, … are path coefficients and r_{X1X2} the correlation between the exogenous variable X1 and X2, e_{X3} and e_Y is uncorrelated error related to the endogenous variable X3, Y respectively.

	X1	X2	X3	Y
X1	1			
X2	r_{12}	1		
X3	r_{13}	r_{23}	1	
Y	r_{1y}	r_{2Y}	r_{3Y}	1

Let $M =$

Where *M* the correlation matrix among variables that described as:

$r_{12} = p_{21}$ *for one-headed arrow, X1 to X2* $r_{1Y} = p_{Y1} + p_{YX2}r_{12} + p_{Y3}r_{13}$
$r_{13} = p_{X3X1} + p_{X3X2}r_{12}$ $r_{2Y} = p_{Y1}r_{12} + p_{YX2} + p_{YX3}r_{23}$
$r_{23} = p_{X3X1}r_{12} + p_{X3X2}$ $r_{3Y} = p_{YX1}r_{13} + p_{YX2}r_{23} + p_{YX3}$

From given or calculate correlation matrix the path coefficients can be compute for each path in the graph.

$$X_i = \sum_{h=1}^{i-1} P_{hi} X_h + e_i \quad , i=1, 2, \ldots, n; \; n:\text{number of variable,}$$

Hence the target Y for one-headed arrow, "between" X_1 and X_2 in fig. 1A calculated as:

$$Y = P_{YX1}X_1 + P_{YX2}X_2 + P_{YX3}X_3 \qquad (1)$$

2.1 Relations among variables (Genes) and Output

The results of path analysis model are path coefficients which are continuous values, but in case of target variable is a binary variable such as (lung cells stat). It's necessary to convert the path analyses output into binary value to obtains the probability of target value belong to one category. Thus Logistic regression is suggested to convert the continuous output into categorical.

Logistic regression (general Model)

If $x_1 \ldots x_k$ are a collection of independent variables and y is a binomial outcome variable with probability of success $= p$, then the multiple logistic regression model is given by:

Logistic regression: is a regression model for Bernoulli-distributed dependent variables. It is a linear model that utilizes the logit as its link function. Logistic regression has been used extensively in medical and social sciences [12, 13].

The logit model takes the form:

$$\mathbf{logit}(p) = \log\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + e_i \qquad (2)$$

Where i=1,2,..n; n number of the sample, p_i=Pr (y_i=1), y_i the target variable, fig. 2. shows the logistic function, with y on the horizontal axis and $f(y)$ on the vertical axis
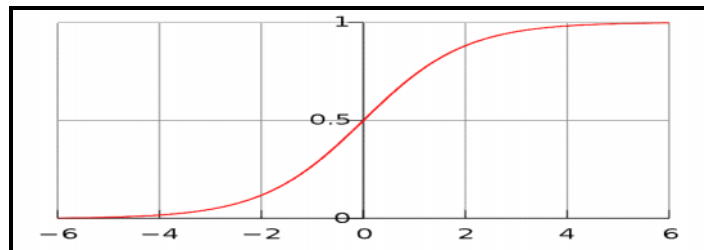

Fig. 2. The logistic function

2.2 Pearson Correlation Coefficient for Measure Genes Similarity

Pearson's correlation coefficient measures linear relationship between two variables by comparing their strength and direction. Relationship between two variables is expressed by -1 to +1. If the variables are perfectly linear related by an increasing relationship, the Correlation Coefficient gains the maximum value i.e. +1. On the other hand, if the variables are perfectly linear related by a decreasing relationship, the correlation value gains -1. And a value of 0

expresses that the variables are not linear related by each other. In general, if the correlation coefficient is greater than 0.8, it expresses strong correlation between variables.

Let X and Y are interval or ratio variables. They are normal distribution and their joint distribution is bivariate normal. So the formula of Pearson's Correlation Coefficient is:

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X)(SS_Y)}} \qquad (3) \quad \text{OR}$$

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\left(X^2 - \frac{(\sum X)^2}{n_x}\right)\left(Y^2 - \frac{(Y)^2}{n_y}\right)\right]}} \qquad (4)$$

Where $\sum X$ sum of all the X scores is is, $\sum Y$ is sum of all the *Y* scores, $\sum X^2$ is square of each *X* score and then sum of them. $\sum Y^2$ is square of each *Y* score and then sum of them. $\sum XY$ is multiply of each *X* score by its associated *Y* score and then add of the resulting products together. This is also called cross product. *n* refers to the number of "pairs" of data [14].

3. Lung Cancer Microarray Dataset

The dataset where used in this study primary lung cancer specimens, was collected from European Bioinformatics Institute (EBI) available online. http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-31908 , the dataset published on 2011-09-05, contains from 40 normal samples vs. 60 primary lung cancer. Each sample in the dataset contains from 22646 genes.

4. Proposed Model to Constructing Genes Network

As we mention above Microarray dataset contains thousand of genes unrelated with diseases, thus we can't deal with a huge numbers of genes in single network, in addition to the redundant genes don't provide any additional information, thus, for the simplicity we will dealing with microarray signature (most important genes).
The proposed genes network constructed as the follows:
1. Extract the microarray signature (subset of 10 top ranked genes) which ranked by information Gain Ratio (IGR) criterion.
2. Assign the gene that take higher IGR value as focal gene.
3. Calculate Pearson's correlation coefficient (PCC) between focal gene and each other genes, Only those genes with the absolute PCC value higher than (0.7) thresholds are linked to the focal gene.
4. Assign the gene that gave next IGR as focal gene. Return form 3.
Finally, a Directed networks are then created from these gene sets. In the network, a vertex set $V = \{g_1, g_2, …, g_i\}$ represents the genes.

$E = \{\{g_i, g_j\} \mid g_i, g_j \in V \quad |P_{ij}| \geq T\}$ is the edge set that represents the relationship between the genes, where T is the given threshold of PCC, and P_{ij} is the PCC value of g_i and g_j. Fig.3. shows the gene network.

5. Practical Constructing Genes Network

To extracting the subset of 10 top ranked genes according to information gain ratio criterion [3], ELBRAV algorithm [15] was used. This subset of genes gave the higher classification accuracy on majority of cancerous microarray datasets.

Software package SPSS 16.0 [16] was used to calculate the Pearson's correlation to measure the similarity among genes and lung cancer "lung cells state". Genes that have strong correlation coefficient have the same function and linked with same biological pathway [17]. Thus the genes those have Pearson's correlation value greater than |0.7| are linked together. The 10-top ranked genes that extracted from the dataset labeled as being in the microarray. Fig.3. shows the path diagram (genes network) that constructed based on Pearson's correlation value. In the path diagram Fig.3. The rectangles represent genes (variables) and the directed arrows represent the linkage between genes (interaction between two genes or effect of genes on lung cell state).

Fig.3. Genes network generated from subset of top ranked genes (assigned as null path analysis model)

6. Analysis and Discussion

Model Hypothesis

The genes network that generated from subset of top ranked genes assigned as null path analysis model. (cf. Fig.3.)

The main objective here the question; is the genes network fit the subset of lung cancer microarray dataset?

- Ho: The genes network is assigned as null path model of hypothesized causal pathways among measured genes expression profiles and between lung cancers.

- R: is the gene network fit the subset of lung cancer microarray dataset?

Hypothesis Testing

To estimates the path coefficient (regression weights) for model (Fig.3.) Maximum likelihood estimation function (MLF) [18] was used. The MLF exists in the AMOS (v16) software package [19], fFig. shows standardized path coefficient among genes and lung cells state that calculated based on MLF.
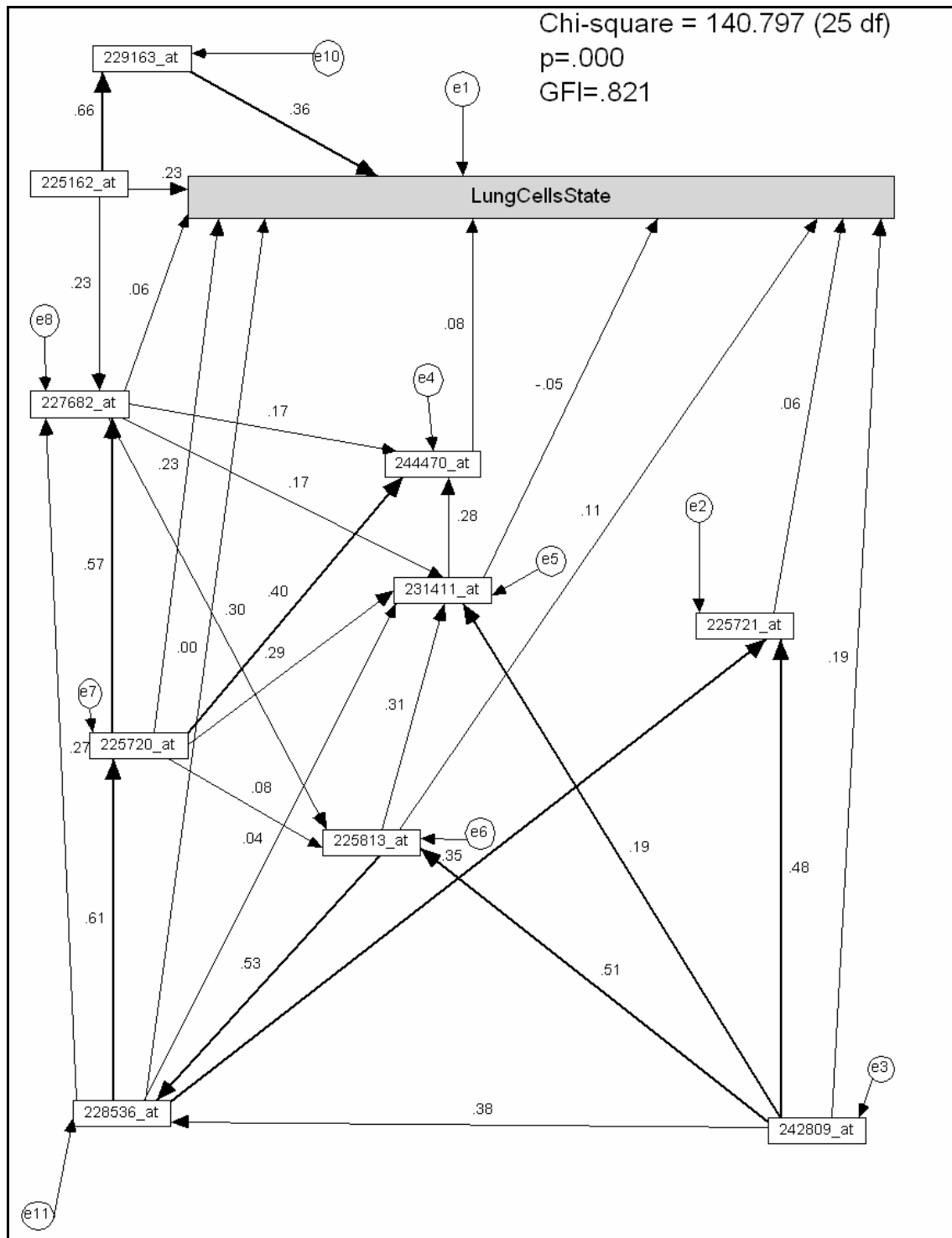
Fig. 4. Standardized path coefficient among genes and lung cells state.

6.1 Model Properties and Evaluation

The model (Fig.) properties as the follows:

- Number of variables (genes) (21), number of observed variables (11), number of unobserved variables (10), number of exogenous variables (11), number of endogenous variables (10).
- Degrees of freedom: Number of distinct sample moments (66), number of distinct parameters to be estimated (41), thus degrees of freedom (25)
- ^{2} = 140.797, probability level = .000.
- Goodness of Fit Index (GFI) [20] of the model GFI =0.821.
- Root Mean Square Residual (RMR), RMR= 0.282.
- Root Mean Square Error of Approximation (RMSEA) = 217.
- Comparative Fit Index (CFI) [21], CFI > 0.894.

The model in fFig. Contains several path coefficient values is not significantly different from zero at the p- value 0.05, these paths are organized in Table 1.

Due to the path diagram in (Fig.) contains several paths that are not significantly different from zero at the 0.05 p-value; the path diagram simplified by eliminating these paths and variables (genes) that insignificantly affected on lung cells state, the modified path diagram shown in fFig.

Table 1. Insignificant path coefficients for (model in Fig.), where SE (Standard Error of Regression Weight), CR Critical ratio for regression weight) and P (Probability Value)

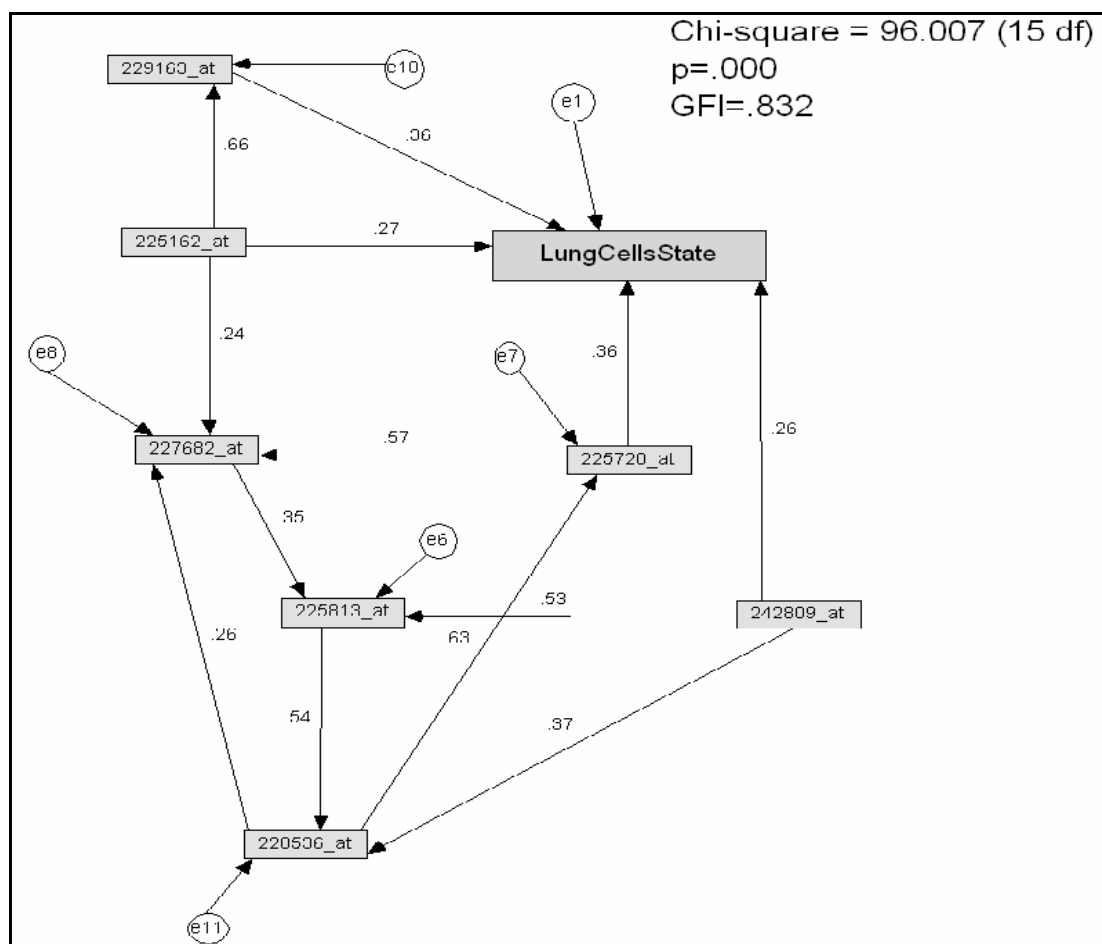| Direct path	Estimate	SE	CR	P
231411_at<---228536_at	0.043	0.112	0.381	0.703
231411_at<---227682_at	0.175	0.099	1.769	0.077
244470_at<---227682_at	0.182	0.12	1.521	0.128
LungCellsState<---227682_at	0.053	0.12	0.442	0.658
LungCellsState<---225813_at	0.095	0.118	0.807	0.419
LungCellsState<---225721_at	0.054	0.082	0.66	0.509
LungCellsState<---225720_at	0.191	0.103	1.864	0.062
LungCellsState<---242809_at	0.155	0.091	1.712	0.087
LungCellsState<---231411_at	-0.047	0.113	-0.418	0.676
LungCellsState<---228536_at	-0.003	0.126	-0.022	0.983
LungCellsState<---244470_at	0.069	0.086	0.804	0.421
225813_at<---225720_at	0.078	0.098	0.797	0.426

Fig. 5 Path diagram resulted after simplified model (Fig.3.)

      The path coefficients of the simplified path diagram were re-estimated as it's shown in Error! Reference source not found. Its notice from Error! Reference source not found. all variables (genes) are significantly affected among each other, as well as on lung cells state. Also the goodness of the simplified model (Error! Reference source not found.) is GFI=0.832 (better than the original model). The path coefficients (regression weights) for path diagram (Error! Reference source not found.) Organized in Table 2, Table 3 presents the total effects in the model Error! Reference source not found.

Table 2 Standardization path coefficients for path diagram (Error! Reference source not found.)

Direct path	Estimate	SE	CR	P
228536_at<---242809_at	0.355	0.068	5.251	***
225813_at<---242809_at	0.495	0.064	7.789	***
227682_at<---225162_at	0.213	0.046	4.647	***
229163_at<---225162_at	0.656	0.076	8.601	***
LungCellsState<---225162_at	0.221	0.069	3.22	0.001
LungCellsState<---229163_at	0.297	0.069	4.318	***
LungCellsState<---225720_at	0.304	0.06	5.093	***

LungCellsState<---242809_at	0.214	0.059	3.653	***
227682_at<---228536_at	0.244	0.068	3.593	***
225813_at<---227682_at	0.365	0.08	4.582	***
225720_at<---228536_at	0.642	0.078	8.231	***
228536_at<---225813_at	0.555	0.076	7.31	***
227682_at<---225720_at	0.509	0.065	7.877	***

Table 3 Standardization total effects (model Error! Reference source not found.)

	242809_at	225162_at	225720_at	225813_at	228536_at	227682_at	229163_at
225720_at	0.457	0.031	0.075	0.403	0.726	0.147	0
225813_at	0.644	0.088	0.211	0.131	0.236	0.413	0
228536_at	0.712	0.049	0.117	0.628	0.131	0.229	0
227682_at	0.407	0.241	0.576	0.359	0.646	0.131	0
229163_at	0	0.656	0	0	0	0	0
LungCellsState	0.353	0.425	0.326	0.122	0.22	0.045	0.297

In summary: the genes network that generated from subset of top ranked genes is fit the data with GFI =0.821, also we simplified the genes network by eliminating insignificant path coefficient, the resulted simplified model fit the subset of the dataset GFI =0.832 and better than the original model.

From two models the interaction between candidate genes were estimated and the standardization path coefficient reflect the quantitative of these interaction.

7. Single Decision Tree Model vs. Proposed Path Analysis Model

In addition to path analysis model measure the effect of variables (genes) on the target variable (Lung cell stat) also it can used as classification model as shown in the comparative below.

The output of traditional classification model such as single decision tree (c4.5) is a set of roles that partition the dataset according to higher informative attribute. Tested samples are classified according to these roles. As shown in the real practical example fFig.

Fig. the microarray dataset is the input of the C4.5 classifier, the output is set of rules, in testing stage tested samples are input of rules, and the output is category of the new samples.

Pearson correlation between genes that shared in the model (Fig.) and lung cells were calculated as shown in Table 1, its notice the higher informative gene (225721_at) is also obtains higher correlated with lung cancer, but others genes (225035_x_at and 200077_s_at) in the model are uncorrelated with Lung Cancer, Also these genes are uncorrelated with the higher informative gene (225721_at) that mean no interaction with it. We conclude the C4.5 built model from uncorrelated and no interacted genes.
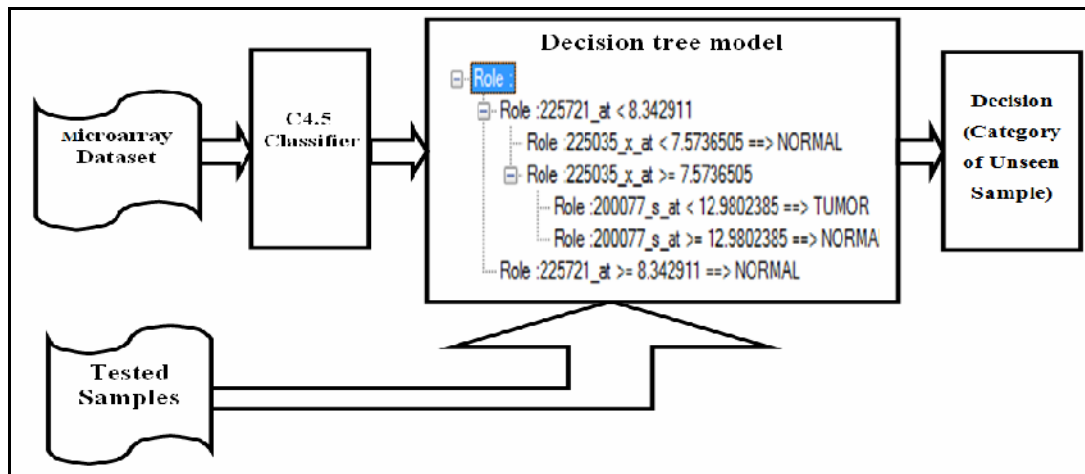
Fig. 6. Real example of decision tree model

Table 1 Pearson correlation between genes and lung cancer, at significant 0.01 levels

	Lung Cancer	225721_at	225035_x_at	200077_s_at
Lung Cancer	1	0.79835	-0.60447	0.367763
225721_at	0.79835	1	**-0.30628**	**0.29778**
225035_x_at	-0.60447	**-0.30628**	1	**0.144496**
200077_s_at	**0.367763**	**0.29778**	**0.144496**	1

In Path analysis model the gene network firstly constructed from correlated and interacted genes (cf. Fig.3.). Path analysis model allows estimating the path coefficient of all paths in the model and visualizing these effects.

The value of path analysis model in addition to estimate the regression weights of all paths simultaneously, its work as a classifier by convert the output (target) value into domain between [0 and1] that represent probability of the target value belong to one of category. Fig. shows path analysis model as classifier.
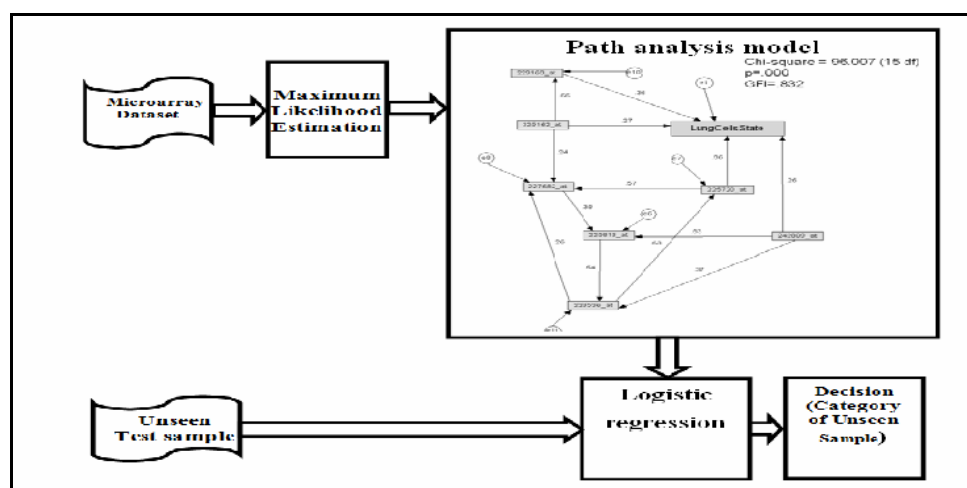

Fig. 7. Path analysis model classifier

Lung cell state calculated as equation (5) for the path analysis model in Fig. and as equation (6) for path analysis model in Error! Reference source not found.

$$\text{Lung cell state} = 0.054 * 225721_{at} + 0.155 * 242809_{at} + 0.193 * 225162_{at} + 0.191 * 22$$

$$\text{Lung cell state} = 0.26 * 242809_{at} + 0.27 * 225162_{at} + 0.36 * 225720_{at} + 0.36 * 229163_{at}$$

By Applying the Logistic regression to obtain the probability if new samples are tumor or normal

The path analysis modeling gave 95% by the model (Fig.) and model (Error! Reference source not found.) vs. 90% accuracy obtained by single decision tree model (Fig.). And it's useful to mention, samples that incorrect classified were the same in two path analysis models (Fig. and (Error! Reference source not found.) So, in this case path analysis modeling can be used as attribute reduction.

8. Conclusions

We presented new approach to predict genes inference based on gene-gene interaction network that constructed from gene expression data. To elucidate this approach, a real lung cancer microarray dataset was used.

In this approach, information gain ratio attribute evaluation is used to extract the signature of cancerous microarray dataset, and Pearson's correlation criterion was used to measure the similarity between genes to construct genes network according to the strength correlation between focal gene and other candidate genes, where the high correlated genes linked together and theoretically do the same genes function. In addition to, estimate the effect of each gene on lung cancer (path coefficient) using path analysis model.

The value of this approach is that it not only tackles the measurement problem by path analysis but also provides a visualization of the relation among genes. In addition to its ease of use, this approach effectively addresses the genes reduction problem. For instance; the genes that gave approximately equal information gain ratio value or Pearson's correlation value, not gave the same path coefficient on lung cancer using path analysis model, due to in the genes reduction methods or correlation methods dealing with each gene and the disease label individually but path analysis modeling analysis all genes simultaneously.

Path analysis modeling can be used as classifier but it is need additional steps, such as data standardization and logistic regression function. In this case path analysis modeling gave the probability of the target (class label) belong to one class, this is additional advantage of path analysis modeling vs. other classification modeling, where path analysis in this study gave 95% classification accuracy vs. 90% that given by C4.5.

This approach provides path diagram that explains the direct and indirect effect of each gene on lung cancer and shows the genes that provide significantly and insignificantly effect on lung cancer, maximum likelihood estimation function is used to calculate the path coefficients, and the proposed path diagram fit the subset of the microarray dataset, hypothesis testing were used to measure the goodness of the proposed path diagram, where the Goodness of Fit Index (GFI) was $\geq$ 0.832.

9. References

[1] R. Arma, I , L. Marcos, V. Taboada, E. Ucar, B. Irantzu, A. Fullaondo, L. Pedro, and A. Zubiaga, "Microarray Analysis of Autoimmune Diseases by Machine Learning Procedures", IEEE Transaction on Information in biomedicine , vol. 13, no. 3, pp 341-350, 2009.

[2] M. B. Senousy, H. M. El-Deeb, K. Badran and I. A. Al-Khlil, " Suite of Decision Tree-Based Classification Algorithms on Cancer Gene Expression Data", Egyptian Informatics Journal, vol. 12, Issue 2, 2011

[3] M. B. Senousy, H. M. El-Deeb, K. Badran and I. A. Al-Khlil, "Gene Ranking Techniques via Attribute Evaluation Algorithms for DNA Microarray Analysis", ASAT international conference on Aerospace Sciences & aviation technology, pp. 153-167, 2011.

[4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, 46(1-3):389–422, 2002.

[5] Mahesh Pal and Giles M, "Feature Selection for Classification of Hyperspectral Data by SVM", IEEE Transactions on geosciences and remote segnaling vol. 48, no. 5, 2010.

[6] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data", BMC Bioinformatics, vol. 6, no.76, 2005.

[7] Hui Liu, Jianzhong Su, Junhua Li, Hongbo Liu, Jie Lv, Boyan Li, Hong Qiao and Yan Zhang, "Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network", BMC Systems Biology, vol. 5, no.158, 2011.

[8] Xuebing Wu and Shao Li, "Cancer Gene Prediction Using a Network Approach", Chapter 11, in: Cancer Systems Biology (Ed. Edwin Wang). Series: Chapman & Hall/CRC Mathematical & Computational Biology, USA: CRC Press, pp. 191-212, 2010.

[9] Xiao-Gang Ruan, Jin-Lian Wang and Jian-Geng Li, "A Network Partition Algorithm for Mining Gene Functional Modules of Colon Cancer from DNA Microarray Data", Geno. Prot. Bioinfo. vol. 4 no. 4, pp. 245- 252, 2006

[10] Xing-Ming Zhao, Luonan Chen and Kazuyuki Aihara, "Gene Function Prediction with the Shortest Path in Functional Linkage Graph", The First International Symposium on Optimization and Systems Biology (OSB'07), pp. 68–74, 2007

[11] Xianghong Zhou, Ming-Chih J. Kao, and Wing Hung Wong, "Transitive functional annotation by shortest-path analysis of gene expression data", PNAS, vol. 99, no. 20, pp. 12783–12788, 2002

[12] Dodd, L.E., Wagner, R.F., Armato, S.G., McNitt-Gray, M.F., Beiden, S., Chan, H.P., Gur, D., McLennan, G., Metz, C.E., Petrick N., Sahiner, B., Sayre J., "Assessment Methodologies and Statistical Issues for Computer-Aided Diagnosis of Lung Nodules in Computed Tomography", Acad. Radiol., vol. 11, pp. 462-474, 2004.

[13] Song, J.H., Venkatesh, S.S., Conant, E.A., Arger, P.H., Sehgal, C.M. "Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses", Acad. Radiol., vol. 12, pp. 487-495, 2005.

[14] MOHAMMED M MAZID, A. B. M. SHAWKAT ALI and KEVIN S TICKLE, "Input

space reduction for Rule Based Classification", INFORMATION SCIENCE and APPLICATIONS, Issue 6, vol.7, 2010.

[15] M. B. Senousy, H. M. El-Deeb, K. Badran and Ibrahim Ali Al-Khlil, "Ensample Learning Based on Ranking Attribute Value (ELBRAV) for imbalanced Biomedical Data Classification", Egyptian Computer Science Journal (ECS Journal), vol. 36, no. 1, 2012.

[16] "Command Syntax Reference", Copyright © 2007 by SPSS Inc. http://www.spss.com

[17]   Xin Li, Hsinchun Chen, Jiexun Li, and Zhu Zhang, "Gene Function Prediction with Gene Interaction Networks: A Context Graph Kernel Approach", IEEE Information Technology in Biomedicine, vol. 14 Issue1, pp.119-128, 2010.

[18] In Jae Myung, "Tutorial on maximum likelihood estimation", Journal of Mathematical Psychology, vol. 47, pp. 90-100, 2003.

[19] Jeremy J. Albright and Hun Myoung Park , "Confirmatory Factor Analysis using Amos, LISREL, Mplus, SAS/STAT CALIS", University Information Technology Services Center for Statistical and Mathematical Computing Indiana University, © 2006-2009, Last modified on April 2009

[20] Tanaka J.S. and Huba G.J., "A fit index for covariance structure models under arbitrary GLS estimation", British Journal of Mathematical and Statistical Psychology, vol. 38, pp. 197–201, 1985.

[21] McDonald, R.P. and Marsh, H.W., "Choosing a multivariate model: Noncentrality and goodness of fit", Psychological Bulletin, vol. 107, pp. 247-255, 1990.