

Interactive Naive Bayesian network: A new approach of constructing gene-gene interaction network for cancer classification

Xue W. Tian^a and Joon S. Lim^{b,*}

^a*Department of Educational Technology, Teachers College, Qingdao University, QingDao, China*

^b*IT College, Gachon University, Seongnam, South Korea*

Abstract. Naive Bayesian (NB) network classifier is a simple and well-known type of classifier, which can be easily induced from a DNA microarray data set. However, a strong conditional independence assumption of NB network sometimes can lead to weak classification performance. In this paper, we propose a new approach of interactive naive Bayesian (INB) network to weaken the conditional independence of NB network and classify cancers using DNA microarray data set. We selected the differently expressed genes (DEGs) to reduce the dimension of the microarray data set. Then, an interactive parent which has the biggest influence among all DEGs is searched for each DEG. And then we calculate a weight to represent the interactive relationship between a DEG and its parent. Finally, the gene-gene interaction network is constructed. We experimentally test the INB network in terms of classification accuracy using leukemia and colon DNA microarray data sets, then we compare it with the NB network. The INB network can get higher classification accuracies than NB network. And INB network can show the gene-gene interactions visually.

Keywords: Gene interaction, naive Bayesian, differently expressed genes (DEGs), leukemia, colon, DNA microarray

1. Introduction

Genetic interactions have long been studied in model organisms as a means of identifying functional relationships among genes or their corresponding gene products, with the nature of the relationships depending on the types of interactions [1-6]. The availability of complete genome sequences facilitates the development of high-throughput assays that can probe cells at a genome-wide scale [7-9]. Such assays measure molecular networks and their components at multiple levels. These include mRNA transcript quantities; protein-protein and protein-DNA interactions; chromatin structure; and protein quantities, localization, and modifications [10-12]. These rich data illuminate the working of cellular processes from different perspectives and offer much promise for novel insights into these processes [13-16]. However, there is still a tendency to look for the smallest and most accurate set of genes that are able to distinguish between two or more phenotypes [17-20]. We propose an interactive naive Bayesian (INB) network to classify gene phenotypes with the smallest gene set and map the relation-

*Address for correspondence: Joon S. Lim, IT College, Gachon University, Seongnam, South Korea. Tel.: 82-031-750-5750; Fax: 82-031-750-5662; E-mail: jslim@gachon.ac.kr.

ships of the genes with the set. We experimentally test the INB network in terms of its classification accuracy.

There are many approaches to classification, including decision trees, neural networks, support vector machines, and the Bayesian network. The Bayesian approach is the most commonly used when dealing with uncertainty because it is based on the probability theory [21, 22]. A Bayesian network is a graphical model that encodes probabilistic relationships of variables. Its main distinguishing feature from classical statistical inference approaches is the use of subjective or personal beliefs (prior probabilities) in the analysis [23, 24]. These probabilistic approaches make strong assumptions about how the data is generated and posit a probabilistic model that embodies these assumptions. Naive Bayesian (NB) network classifiers are very robust in terms of irrelevant attributes, and the classification takes into account evidence from many attributes to make the final prediction [24, 25]. NB classifiers have been adapted to handle continuous attributes primarily using Gaussian distributions or discretizing the domain, both of which present certain disadvantages. In the former approach, the probability density of the attributes is not always well fitted by a Gaussian distribution. In the latter approach, there can be loss of information.

We determine a weight for the INB classifier by assigning a weight between each node and its interactive parent on the NB network. This decreases the independence and increases the classification performance. The weights are constructed by a neuro-fuzzy network called a neural network with weighted membership functions (NEWFM) [26, 27]. With NEWFM we can determine the relationships between each attribute. We use the relationships to calculate the weight for each node of the Gaussian NB network. The INB network can also provide a gene interaction network using the weight between the node and its interactive parents.

We apply the INB network to classify leukemia and colon DNA microarray data sets [28, 29]. With the successful completion of the Human Genome Project (HGP), we are entering the post-genomic era. Facing massive amounts of data, traditional biological experiments and data analysis techniques encounter significant challenges. In this situation, cDNA microarrays and high-density oligonucleotide chips, which are novel biotechnologies, are global (genome-wide or system-wide) experimental approaches that are used effectively in the systematic analysis of large-scale genome data [30]. In this paper, we apply INB classifiers for the analysis of DNA microarray data. The experimental results demonstrate that INB classifiers are more reliable than NB classifiers.

2. Materials and method

2.1. Preprocessing of the data set

2.1.1. Materials

In [28], the authors present methods for analyzing gene expression data obtained from DNA microarrays in order to classify types of cancer. Their leukemia data is available on-line. The data is split into two subsets: A training set and a test set. Their training set consists of 38 samples (27 ALL and 11 AML) from bone marrow specimens. Their test set has 34 samples (20 ALL and 14 AML), prepared under different experimental conditions and including 24 bone marrow and 10 blood sample specimens. All samples have 7129 features, corresponding to some normalized gene expression value extracted from the micro-array image. We retained the exact same experimental conditions for ease of comparison with their method.

Table 1
The selected sets of DEGs

Data Set	DEG
Leukemia	L33930_s_at, X82240_rna1_at, J04990_at, M30703_s_at, D88270_at, M21624_at
Colon	H04282, M64673, R73606, M97676, M55268, T55558, L10717

In [29], the authors describe and study a data set that is available on-line. Gene expression information was extracted from DNA microarray data resulting, after pre-processing, in a table of 62 tissues \times 2000 gene expression values. The 62 tissues include 22 normal and 40 colon cancer tissues. The matrix contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Some genes are non-human genes. Since there was no defined training and test set, we used all 62 samples for both training and test.

2.1.2. General remarks on figures

We use the Bhattacharyya distance (BD) [31] and a neuro-fuzzy network [27] as our differently expressed gene (DEG) selection method for microarray data analysis. The detail of this method is represented on the previous paper [32, 33]. Finally, we select six DEGs from the leukemia data set and seven DEGs from the colon data set [32, 33]. The selected sets of DEGs are shown in Table 1.

2.1.3. Data normalization

After DEGs selection, we normalize the values of DEGs on the scale of [0, 1] and transform them with a concentrated distribution [34]. The set of normalized expression values for the i th DEG with s samples, N_i , is given by

$$N_i = \left\{ n_{i,j} \mid n_{i,j} = \frac{1}{1 + e^{-(d_{i,j} - \min(D_i)) / (\max(D_i) - \min(D_i))}}, \forall j = 1, 2, \dots, s \right\} \tag{1}$$

Table 2
The representations of DEGs and their corresponding original and normalized expression values

Genes	D_1		...	D_i		...	D_n		Class
	Original	Normalized		Original	Normalized		Original	Normalized	
Sample Values	$d_{1,1}$	$n_{1,1}$...	$d_{i,1}$	$n_{i,1}$...	$d_{n,1}$	$n_{n,1}$	c_1^*
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	$d_{1,j}$	$n_{1,j}$...	$d_{i,j}$	$n_{i,j}$...	$d_{n,j}$	$n_{n,j}$	c_j^*
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	$d_{1,s}$	$n_{1,s}$...	$d_{i,s}$	$n_{i,s}$...	$d_{n,s}$	$n_{n,s}$	c_s^*
Sample Set	D_1	N_1	...	D_i	N_i	...	D_n	N_n	$C \{1,2\}$

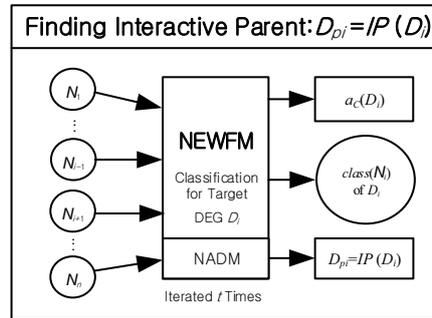


Fig. 1. Finding interactive parent D_{pi} of D_i with the highest classification accuracy $a_C(D_i)$.

where D_i is the set of expression values of the i th DEG D_i in microarray samples, $\min(D_i)$ and $\max(D_i)$ are the minimum and maximum values in D_i , and $d_{i,j}$ and $n_{i,j}$ are the j th expression value and normalized value in D_i and N_i , respectively. Table 2 shows the representations of DEGs and their corresponding original and normalized expression values. C is the set of class variables, c_j^* is the class variable of the j th samples set $\{d_{1,j}, \dots, d_{i,j}, \dots, d_{n,j}\}$ and $\{n_{1,j}, \dots, n_{i,j}, \dots, n_{n,j}\}$, it can be represented by class 1 or class 2.

2.2. Interactive Naive Bayesian network

In order to effectively relax the conditional independence assumptions of the NB network, we suppose every attribute (or gene) has an interactive relationship with other attributes (or genes). This proposed model, called interactive naive Bayesian network (INB), contains attributes each of which has an interactive attribute as a parent and one classification node connected to all attributes as another parent. We use a non-overlap area distribution measurement method (NADM) [27] to measure the interactive relationship between attributes.

After data normalization, one target gene D_i in DEGs is classified in turn by the other genes using a neural network with a weighted fuzzy membership function (NEWFM) [27]. A class is given to all gene samples in N_i for classifying a target gene D_i , denoted by $class(n_{i,j})$, is defined such that:

$$class(n_{i,j}) = \begin{cases} 1, & n_{i,j} < mean(N_i) \\ 2, & n_{i,j} \geq mean(N_i) \end{cases} \quad (2)$$

Then, the every target gene D_i in DEGs is classified in turn by the NEWFM obtaining classification accuracy of D_i , $a_C(D_i)$, with interactive parent gene D_p of D_i , denoted by $D_{pi} = IP(D_i)$, where $i \neq p$ as in Figure 1. The input nodes of the NEWFM are the normalized values of DEGs except D_i for classifying D_i . To find the D_{pi} , the training and test processes are iterated t times. During the iterations, D_p is found by the NADM method in the NEWFM as an interactive parent gene D_{pi} of D_i .

In each iteration, the best gene for the classification of D_i is chosen by the NADM. Every input gene node keeps a variable storing the number of choices. The variable is added by one whenever the gene is chosen. After t time iterations, the gene D_p that has the highest number of variable will be assigned as $IP(D_i)$. The variable is called $h(D_{pi})$ to be used for calculating the weight between the gene D_i and its interactive parent D_{pi} .

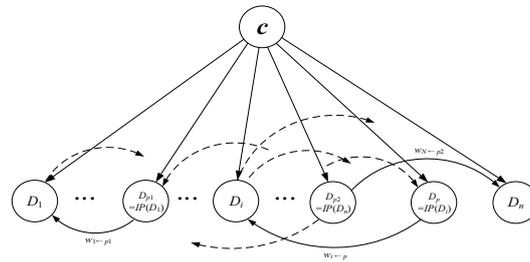


Fig. 2. Structure of the INB network.

After finding D_{pi} with $h(D_{pi})$, the interaction weight calculations between D_i and its interactive parent D_{pi} are continued for enhancing classification power of INB. The weight between D_i and D_{pi} , denoted by $w_{i←p}$ is shown in Eq. (3), where t is the number of classification iterations for D_i , $a_C(D_i)$ is the classification accuracy of D_i , and n is the number of DEGs.

$$w_{i←p} = \frac{h(D_{pi})}{t} * a_C(D_i) \text{ for } C=1, 2, i = 1, 2, \dots, n, \text{ where } i \neq p. \quad (3)$$

The $w_{i←p}$ represents the degree of interactive relationship from D_{pi} to D_i using gene expression data, which is the goal of the genomic revolution for understanding the genetic interacts in a complex living system.

The proposed INB network is constructed by adding the interactive parents with their weights to naive Bayesian network. The higher weight causes the stronger interactive influence from the interactive parent to the target gene. Figure 2 shows the structure of an INB network, where C is the diagnostic class node that is the parent of all other nodes (DEGs) D_i , where $i = 1, 2, \dots, n$. Each node D_i has one more interactive parent D_{pi} represented by arc with the weight on the arc, and also can be an interactive parent of other DEGs or not.

2.3. Classification of interactive Naive Bayesian network

The naive Bayes classifier uses the Bayes theorem in conjunction with the conditional independence hypothesis. The naïve Bayes paradigm is thus based on two conditions over the features of predictive variables and the class variable to predict. Based on the NB network, the joint distribution represented by an INB network is defined as:

$$P(D_1, \dots, D_n, C) = P(C) \prod_{i=1}^n P(D_i | D_p, C) \quad (4)$$

where C is the class node.

The classification corresponding to an INB network on samples set $S_j \{n_{1,j}, \dots, n_{n,j}\}$, c_j^* , is defined as

$$c_j^* = \arg \max_{c_j^* \in C} P(C) \prod_{i=1}^n P(n_{i,j} | n_{p,j}, C) \quad (5)$$

Where

$$P(n_{i,j} | n_{p,j}, C) = P(n_{i,j} | C) + (P(n_{p,j} | C) - P(n_{i,j} | C)) * w_{i←p} \quad (6)$$

The expression values are continuous, in Eq. (6), $P(n_{i,j}|C)$ and $P(n_{p,j}|C)$ are got by Gaussian distribution. In an INB network, attribute dependencies are represented by interactive parents of attributes. The method of defining interactive parents determines the capability of representing attribute dependencies.

3. Experiment results

We experimentally test the INB classifier in terms of classification accuracy, using leukemia and colon DNA microarray data sets, and compare it to the NB classifier. The comparison results shown in Table 3 demonstrate that the INB classifier is more reliable than the NB classifier. In Table 3, Sensitivity (Se) is the ability (probability) to classify. Specificity (Sp) is the probability of identifying class 2 correctly. Positive probability (Pp) is the probability that classified class 1 is truly class 1. Negative probability (Np) is the probability that classified class 2 is truly class 2. Accuracy (Acc) is the probability of obtaining a correct classification. The performance results for leukemia show that 95% accuracy can be achieved with the INB classifier. The NB classifier can only attain 94%. For the colon data set classification, the accuracy of the INB classifier is 3% higher than the NB classifier.

After the experiment, we construct a gene interaction network to represent the relationship between DEGs shown in Figure 3. Figure 3a is the gene interaction network of leukemia and Figure 3b shows the gene interaction network of a colon. For leukemia, Figure 3a reflects the interaction relationship by the arcs. The number on each arc is the weight between the two connected genes. As arc $D88270_at \rightarrow L33930_s_at$, weight 407 shows the degree of effect between gene $D88270_at$ and $L33930_s_at$. The higher the degree, the stronger the effect and the darker the arc is.

Table 3
Experimental Results Comparison

Categories	Method	TP	FP	TN	FN	Se	Sp	Pp	Np	Acc
Colon	NB	33	7	17	5	86.84%	70.83%	82.5%	77.27%	80.64%
	INB	35	5	17	5	87.5%	77.27%	87.5%	77.27%	83.87%
Leukemia	NB	44	1	23	3	93.62%	95.83%	97.78%	88.46%	94.44%
	INB	44	0	24	3	93.62%	100%	100%	88.89%	95.83%

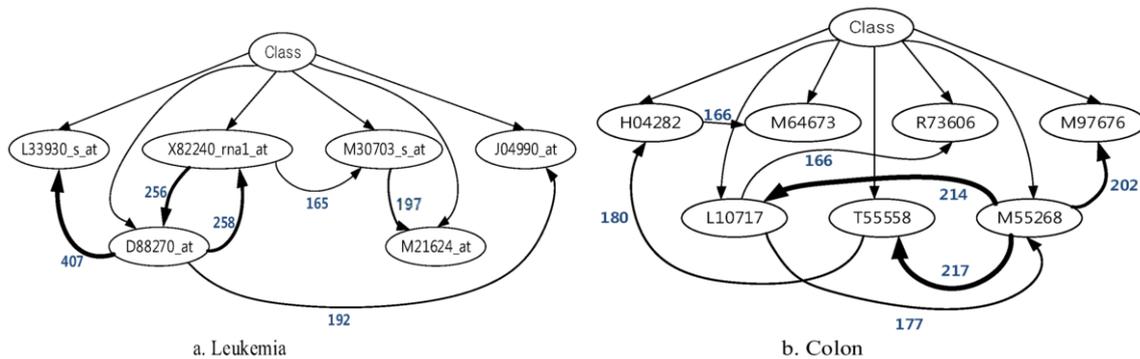


Fig. 3. Gene INB network of leukemia and colon.

4. Conclusions

In this research, a new approach to identify gene interactions has been proposed based on the INB. We selected the six DEGs from the leukemia data set and the seven DEGs from the colon data set to construct and test the INB network. The advantage of the INB network is that it is not only a naive Bayesian network, but also give the gene-gene interactive from the selected DEGs. We combined the NB network and gene interactive network to obtain a novel Bayesian model, INB, for gene classification, and represented the gene interaction network. We experimentally tested the INB classifier in terms of classification accuracy using leukemia and colon DNA microarray data sets, and compared it to the NB classifier. The experimental results show that the INB classifier outperforms the NB classifier. We constructed a gene interaction network to represent the relationship between genes. It represents more informative way than a single variable list of genes. This network approach could be useful also to other research field, such as finding interactions of features for classification approach.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. (2012R-1A1A2044134).

References

- [1] S.F. Elena and R.E. Lenski, Test of synergistic interactions among deleterious mutations in bacteria, *Nature* **390** (1997), 395–398.
- [2] J.A.G.M. De Visser, R.F. Hoekstra and H. van den Ende, Test of interaction between genetic markers that affect fitness in *Aspergillus niger*, *Evolution* **51** (1997), 1499–1505.
- [3] R.E. Lenski and M. Travisano, Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations, *Proceedings of the National Academy of Sciences of the United States of America* **91** (1994), 6808–6814.
- [4] M. Travisano, J.A. Mongold, A.F. Bennett and R.E. Lenski, Experimental tests of the roles of adaptation, chance, and history in evolution, *Science* **267** (1995), 87–90.
- [5] C.L. Burch and L. Chao, Evolution by small steps and rugged landscapes in the RNA virus 6, *Genetics* **151** (1999), 921–927.
- [6] P.E. Turner and L. Chao, Prisoner's dilemma in an RNA virus, *Nature* **398** (1999), 441–443.
- [7] S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genetics* **31** (2002), 64–68.
- [8] N. Guelzim, S. Bottani, P. Bourguin and F. Kepes, Topological and causal structure of the yeast transcriptional regulatory network, *Nature Genetics* **31** (2002), 60–63.
- [9] J. Gough, K. Karplus, R. Hughey and C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, *Journal of Molecular Biology* **313** (2001), 903–919.
- [10] E. Perez-Rueda and J. Collado-Vides, The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12, *Nucleic Acids Research* **28** (2000), 1838–1847.
- [11] E. Perez-Rueda and J. Collado-Vides, Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria, *Journal of Molecular Evolution* **53** (2001), 172–179.
- [12] S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genet* **31** (2002), 64–68.
- [13] S.F. Elena and R.E. Lenski, Test of synergistic interactions among deleterious mutations in bacteria, *Nature* **390** (1997), 395–398.

- [14] A.C. Peters, Lively CM in Epistasis and the Evolutionary Process, J.B. Wolf, E.D. Brodie and M.J. Wade, eds., Oxford University Press, New York, 2000, pp. 99–112.
- [15] L. Jasnos and R. Korona, Epistatic buffering of fitness loss in yeast double deletion strains, *Nat Genet* **39** (2007), 550–554.
- [16] Clifford W. Zeyl, How missing genes interact, *Nat Genet* **39** (2007), 440–442.
- [17] G. Bontempi, A blocking strategy to improve gene selection for classification of gene expression data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4** (2007), 293–300.
- [18] T. Lin, R. Liu, C. Chen, Y. Chao and S. Chen, Pattern classification in DNA microarray data of multiple tumor types, *Pattern Recognition* **39** (2006), 2426–2438.
- [19] L. Wang, F. Chu and W. Xie, Accurate cancer classification using expressions of very few genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4** (2007), 40–53.
- [20] K. Yang, Z. Cai, J. Li and G. Lin, A stable gene selection in microarray data analysis, *BMC Bioinformatics* **7** (2006), 228.
- [21] N. Friedman, D. Geiger and M. Goldszmidt, Bayesian network classifiers, *Machine Learning* **29** (1997), 131–163.
- [22] A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America* **96** (1999), 6745–6750.
- [23] B.W. Morgan, *An Introduction to Bayesian Statistical Decision Processes*, Prentice-hall, New Jersey, 1968.
- [24] D. Heckerman, *A tutorial on learning with Bayesian Network*, Microsoft Research, Redmond, 1996.
- [25] P. Langley, An analysis of Bayesian classifiers, *Tenth National Conference on Artificial Intelligence*, 1992, pp. 223–228.
- [26] J.S. Lim, D. Wang, Y.-S. Kim and S. Gupta, A neuro-fuzzy approach for diagnosis of antibody deficiency syndrome, *Neurocomputing* **69** (2006), 969–974.
- [27] J.S. Lim, Finding features for real-time premature ventricular contraction detection using a fuzzy neural network system, *IEEE Transactions on Neural Networks* **20** (2009), 522–527.
- [28] T.R. Golub, D.K. Slonim and P. Tamayo, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286** (1999), 531–537.
- [29] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America* **96** (1999), 6745–6750.
- [30] E. Themaat, On the use of learning Bayesian networks to analyze gene expression data: Classification and gene network reconstruction, Master Thesis, University of Amsterdam, Artificial Intelligence, June 2005.
- [31] G. Xuan, Bhattacharyya distance feature selection, *Proceedings of the 13th International Conference on Pattern Recognition*, 1996, pp.195–199.
- [32] X.W. Tian, S.H. Lee and J.S. Lim, Gene selection for leukemia classification based on bhattacharyya distance, *Proceedings of KIIS Spring Conference* **23** (2013), 17–18.
- [33] X.W. Tian and J.S. Lim, Bhattacharyya distance for identifying differentially expressed genes in colon gene experiments, *International Conference on Information Science and Applications*, 2013, pp. 1–2.
- [34] H. Jun and M. Claudio, The influence of the sigmoid function parameters on the speed of back propagation learning, *From Natural to Artificial Neural Computation*, 1995, pp. 195–201.