Technology and Health Care 24 (2016) S601–S605
DOI 10.3233/THC-161187
IOS Press

A hierarchical two-phase framework for selecting genes in cancer datasets with a neuro-fuzzy system

Jongwoo Lim, Bohyun Wang and Joon S. Lim*
IT College, Gachon University, Seongnam, Korea

Abstract. Finding the minimum number of appropriate biomarkers for specific targets such as a lung cancer has been a challenging issue in bioinformatics. We propose a hierarchical two-phase framework for selecting appropriate biomarkers that extracts candidate biomarkers from the cancer microarray datasets and then selects the minimum number of appropriate biomarkers from the extracted candidate biomarkers datasets with a specific neuro-fuzzy algorithm, which is called a neural network with weighted fuzzy membership function (NEWFM). In this context, as the first phase, the proposed framework is to extract candidate biomarkers by using a Bhattacharyya distance method that measures the similarity of two discrete probability distributions. Finally, the proposed framework is able to reduce the cost of finding biomarkers by not receiving medical supplements and improve the accuracy of the biomarkers in specific cancer target datasets.

Keywords: Microarray data, feature selection, neuro network fuzzy algorithm

1. Introduction

Recently, as one of diagnostic tools, analysis of diseases by using gene expression microarray datasets has been worldwide used in bioinformatics fields. It needs to use computational methods in that gene expression microarray datasets are very large number of various genes in the gene datasets. It has been a challenging problem to identify the genes that are relevant or not to a clinical diagnosis [1,5,15]. As feature selection methods, mutual information [6,7,13], the t-test [14], threshold number of misclassifications (TNoM) score [8], and the Bhattacharyya distance [3,4,21] have been widely used in finding relevant genes. Feature selection methods have been used for pattern recognition and machine learning [10].

In these days, feature selection is used for identifying the relevance and influence of the selected gene in gene expression data and it is able to improve the comprehensibility from the results. As specific classifiers in machine learning, k-nearest neighbor (k-NN) [8], support vector machine (SVM) [8,11,12], and rough set [13] have all been used to verify the efficiency after selecting the genes.

In this context, we propose a hierarchical two-phase framework for selecting effective biomarkers that extracts candidate appropriate biomarkers from the cancer microarray datasets and then selects

*Corresponding author: Joon S. Lim, IT College, Gachon University, Seongnam, Korea. E-mail: jslim@gachon.ac.kr.

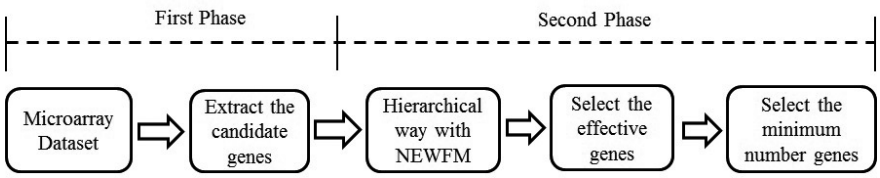

Fig. 1. Overview of the proposed framework.

the minimum number of appropriate biomarkers from the extracted candidate biomarkers with a specific neuro-fuzzy algorithm, which is called a neuro network with weighted fuzzy membership function (NEWFM) [15–17].

The proposed framework is to classify tumor biopsies and normal biopsies from a colon cancer data set, and acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) from a leukemia data set. As the first phase of the proposed process, the Bhattacharyya distance method is used for extracting candidate genes as biomarkers. By using it, we extracted 100 candidate genes in each from the colon dataset that has 2000 genes and the leukemia dataset that has 7129 genes in total. As the second phase of the proposed framework, the minimum number of appropriate genes is selected based on the first phase result.

The minimum of appropriate genes as biomarkers show the highest accuracy by using the NEWFM method. As the minimum, 4 of 100 colon cancer genes and 4 of 100 leukemia genes were selected from the first phase results in 100 colon cancer genes and 100 leukemia genes. The minimum 4 (colon cancer) and 4 (leukemia) genes were used as weighted fuzzy membership functions that preserved the disjunctive fuzzy information and characteristics [15].

In the remainder of this paper, we describe the proposed framework for selecting the minimum of appropriate genes as biomarkers in Section 2. The accuracy of selecting appropriate genes and the number of genes is evaluated on experimental results that will be shown in Section 3. We also conclude our proposed framework for selecting the appropriate genes and the minimum number of genes in Section 4.

2. Proposed two-phase process

We propose a hierarchical two-phase framework for selecting the minimum number of appropriate genes with NEWFM. Overall, the proposed framework consists of two phases as following this:

The description of each phase is following this:

– First Phase: The proposed framework extracts the candidate genes as biomarkers for the given cancer datasets by using one of similarity measures, called Bhattacharyya distance [3,4,21] that is used to calculate the correspondence of discrete probability distributions. For this phase, we used the following equation:

$$D_B(x,\ y) = \frac{1}{4}\ln\left\{\frac{1}{4}\left(\frac{\sigma_x^2}{\sigma_y^2} + \frac{\sigma_y^2}{\sigma_x^2} + 2\right)\right\} + \frac{1}{4}\left\{\frac{(\mu_x - \mu_y)^2}{\sigma_x^2 + \sigma_y^2}\right\},$$

where $D_B(x, y)$, σ_x is the variance of the distribution and μ_x is the mean of the distribution. Thus the more distinguish from each other genes, the bigger value of Bhattacharyya it has. In this experiment, we extract 100 genes from the highest value to 100^{th} value in each cancer datasets such as leukemia and colon datasets.

Table 1
Selected genes from Leukemia dataset and Colon cancer dataset

Datasets	Selected genes	Description
Leukemia dataset	D88270_at	GB DEF = (lambda) DNA for immunoglobin light chain.
	M21624_at	TCRD T-cell receptor, delta.
	X03934_at	GB DEF = T-cell antigen receptor gene T3-delta.
	M54995_at	PPBP Connective tissue activation peptide III.
Colon cancer dataset	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
	R36977	26045 P03001 TRANSCRIPTION FACTOR IIIA.
	U10117	Human endothelial-monocyte activating polypeptide II mRNA, complete cds.
	M63391	Human desmin gene, complete cds.

– Second Phase: From the first phase results, the proposed framework selects more effective and less number of genes as biomarkers. In this phase, we separate the extracted genes into 5 clusters based on the Bhattacharyya distance results and those genes in 5 clusters are input into the NEWFM using the bounded sum of the weighted fuzzy membership functions [15] to select the minimum number of genes with the higher accuracy rate simultaneously. The proposed framework selects again the genes with the higher accuracy rate and the result of NEWFM until reaching the highest accuracy with the minimum number of genes.

Overall, the proposed hierarchical two-phase framework means that we first separated the extracted genes into 5 clusters by using the Bhattacharyya distance value and then select the best 4 genes from each 5 cluster based on the result of NEWFM using the bounded sum of the weighted fuzzy membership functions. In hierarchical way, we also select the minimum number of the effective genes from the selected 20 genes based on the result of NEWFM.

3. Experimental results

We used the two well-known datasets that ALL-AML Leukemia dataset and Colon tumor dataset. Those datasets were from the public Kent Ridge Bio-medical Data Repository [2].

The leukemia data set contains 72 samples that are divided into two variants of leukemia: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). Gene expression levels in these 72 samples were measured using high density microarrays reporting the expression levels of 7129 genes. Rewrite, it is copied.

The colon cancer data set consists of 62 samples of colon epithelial cells that are divided into two variants of colon tissue: 40 colon tumor samples and 22 normal colon samples. Gene expression levels in these 62 samples were measured using high density microarrays. 2000 genes were selected based on the confidence in the measured expression levels.

Table 1 shows the selected genes in the given cancer datasets with the proposed framework as follows.

The accuracies for selecting the minimum number of appropriate genes are shown in Table 2. The accuracy is the probability of obtaining correct genes in all genes that are from the given datasets. It is defined as:

$$\text{Accuracy} = \frac{\textit{All true classified genes}}{\textit{All classified genes}},$$

where all classified genes are the number of genes that were classified from the given datasets and all true classified genes are true positive (TP) genes and true negative genes (TN) from our experimental results.

Table 2
The results of comparison with other algorithms in terms of accuracy

Datasets	Leukemia	Colon
Cho et al. [18]	94.12% (17)	82.08% (10)
Guyon et al. [12]	100% (4)	90.32% (8)
Wang et al. [20]	100% (5)	91.9% (3)
Proposed framework	100% (4)	95.16% (4)

The number in brackets in Table 2 represents the number of selected genes. In this study, AML and ALL in leukemia datasets and tumor biopsies and normal biopsies in colon datasets were classified with the proposed two-phase framework in terms of accuracy and the minimum number of genes. The accuracy of NEWFM was compared with that determined by Cho [18], Guyon [12], and Wang [19]. As shown above, our proposed framework selects the minimum genes with highest accuracy of classification.

4. Conclusion

In this paper, we proposed a hierarchical two-phase framework for selecting the minimum number of appropriate genes as biomarkers from the given cancer datasets in terms of accuracy. The proposed framework consists of two-phases that the first phase is to extract the candidate genes as biomarkers by using the Bhattacharyya distance measure and the second phase, a hierarchical way for selecting the minimum number of appropriate genes with NEWFM from the extracted genes as biomarkers. As shown in Section 3, the proposed framework showed higher accuracy compared to other methods [12,18,19] with the minimum genes.

Finally, as our future work, to robust our proposed framework of selecting the effective and the minimum number of biomarkers, we will experiment with another bigger size of various cancer type datasets.

Acknowledgement

This work was supported by the Gachon University Research Fund of 2015 (GCU-2015-0102).

References

[1] Alba E, Garcia-Nieto J, Jourdan L, Talbi E (2007). Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. Evolutionary Computation IEEE.
[2] Kent Ridge Bio-medical Data Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/).
[3] Bhattacharyya A (1943). On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35: 99-109.
[4] Reyes-Aldasoro CC, Bhalerao A (2006). The Bhattacharyya space for feature selection and its application to texture segmentation. Pattern Recognition 39(5): 812-826.
[5] Golub T, Slonim D, Tamayo P, Huard C, Caasenbeek JM, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.
[6] Liu X, Krishnan A, Mondry A (2005). An entropy based gene selection method for cancer classification using microarray data. BMC Bioinformatics 6: 1-14.
[7] Peng H, Long F, Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 1226-1238.

[8] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000). Tissue Classification with Gene Expression Profiles. J Computational Biology 7: 559-584.
[9] http://www.genome.jp/kegg/.
[10] Hong Y, Kwong S, Chang Y, Ren Q (2008). Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. Pattern Recognition 41: 2742-2756.
[11] Frank O, Brors B, Fabarius A, Li L, Haak M, Merk S, Schwindel U, Zheng C, Müller MC, Gretz N, Hehlmann R, Hochhaus A, Seifarth W (2006). Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. Leukemia 20: 1400-1407.
[12] Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using support vector machines. Machine Learning 46: 389-422.
[13] Maji P, Paul S (2011). Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning 52: 408-426.
[14] Li J, Su H, Chen H, Futscher BW (2007). Optimal search-based gene subset selection for gene array cancer classification. IEEE Transactions on Information Technology in Biomedicine 11: 398-405.
[15] Lim JS (2009). Finding Features for Real-Time Premature Ventricular Contraction Detection Using a Fuzzy Neural Network System. IEEE Transactions on Neural Networks 20: 522-527.
[16] Lee SH, Lim JS (2011). Forecasting KOSPI based on a neural network with weighted fuzzy membership functions. Expert Systems with Applications 38: 4259-4263.
[17] Lee SH, Lim JS (2012). Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. Expert Systems with Application 39: 7338-7344.
[18] Cho JH, Lee D, Park JH, Lee IB (2004). Gene selection and classification from microarray data using kernel machine. FEBS Letters 571: 93-98.
[19] Wang Y, Makedon FS, Ford JC, Pearlman J (2005). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21: 1530-1537.
[20] Wang L, Khan L (2006). Automatic image annotation and retrieval using weighted feature selection. Multimed Tools Appl 29: 55-71.
[21] Coleman GB, Andrews HC (1979). Image segmentation by clustering. Proc IEEE 67(5): 773-785.