RESEARCH ARTICLE

Editorial Process: Submission:11/03/2017 Acceptance:01/09/2018

Informative Gene Selection for Cancer Classification with Microarray Data Using a Metaheuristic Framework

Pyingkodi M[1]*, Thangarajan R[2]

Abstract

Objective: Cancer diagnosis is one of the most vital emerging clinical applications of microarray data. Due to the high dimensionality, gene selection is an important step for improving expression data classification performance. There is therefore a need for effective methods to select informative genes for prediction and diagnosis of cancer. The main objective of this research was to derive a heuristic approach to select highly informative genes. **Methods:** A metaheuristic approach with a Genetic Algorithm with Levy Flight (GA-LV) was applied for classification of cancer genes in microarrays. The experimental results were analyzed with five major cancer gene expression benchmark datasets. **Result:** GA-LV proved superior to GA and statistical approaches, with 100% accuracy for the dataset for Leukemia, Lung and Lymphoma. For Prostate and Colon datasets the GA-LV was 99.5% and 99.2% accurate, respectively. **Conclusion:** The experimental results show that the proposed approach is suitable for effective gene selection with all benchmark datasets, removing irrelevant and redundant genes to improve classification accuracy.

Keywords: Cancer diagnosis- gene treatments- genomics- gene expression data-genetic algorithm-classification

Asian Pac J Cancer Prev, **19 (2)**, 561-564

Introduction

Recent and rapid growth in DNA microarray technology help the researchers to analyze the expression of thousand of genes in a single experiment quickly and in an efficient manner (Shalon et al., 1996; Balamurugan et al., 2017). Gene expression profiling by microarray method has appeared as a capable technique for classification and diagnostic prediction of cancer. A typical Deoxyribonucleic acid (DNA) microarray analysis involves a multi-step procedure: Specific genes are represented by fabrication of microarrays which has properly designed oligonucleotides; hybridization of cDNA populations onto the microarray; scanning hybridization signals and image analysis; transformation and normalization of data; and analyzing data to detect differentially expressed genes as well as the sets of genes that are co regulated (BilBan et al., 2009). However, microarray dataset suffers from the curse of dimensionality, the limited number of samples, and the irrelevant and noise genes, all of which make the classification task for a given sample more challenging (Balamurugan et al., 2016).

Classification is an important task in machine learning (Cho and Won, 2003). Without prior knowledge, it's hard to determine which genes are useful. Therefore, a large number of genes are usually introduced into the dataset, including relevant, irrelevant and redundant genes. However, irrelevant and redundant genes are not useful for classification. As such, effective methods of selecting genes for cancer are critically necessary. Gene selection process aims to select the minimum number of relative and informative genes that are more predictive in classification process. The gene selection uses an optimization algorithm to select a subset of the genes, which has the most classification information, from the original gene microarray data (Pyingkodi and Thangarajan, 2017). The most commonly used gene selection methods can be divided in to filter, wrapper, and embedded ones. The optimal gene selection problem is considered as NP-hard problem. Therefore, it is better to use heuristic approaches such as bio-inspired evolutionary algorithms in order to solve this problem. In this work, we propose a new method, which couples Genetic Algorithm (GA) with Levy Flight approach, to select gene subset from cancer microarray data.

Materials and methods

Problem statement

The most important task in computation biology is selecting informative genes or combinations of genes with a high prognostic potency from the microarray data for cancer classification (Yuanyuan et al., 2017). The biggest issue in gene expression data is its high

*[1]Department of Computer Applications, [2]Department of Computer Science and Engineering, Kongu Engineering College Erode, TamilNadu, India. *For Correspondence: pyingkodi@yahoo.co.in*

curse dimensionality. It contains huge number of genes (rows) and a small number of samples (columns). There is a need for selection methods to select significant genes for disease prediction and diagnosis. From the large dataset, feature selection is a Non-deterministic Polynomial-time hard (NP-hard) problem (Momiao et al., 2001; Premalatha et al., 2017). It involves finding a subset of the original features so that a classifier built with this subset would perform better than a classifier built from the entire set of features. Feature or gene selection methods remove irrelevant and redundant features to improve the classification accuracy.

Though, most of the existing methods of microarray-based cancer classification works too many genes to achieve accurate classification. Since many classification methods are not scalable to the high dimensions, they are inapplicable to analyze raw gene expression microarray data. Classification accuracy is the overall correctness of the classifier and is calculated as the sum of correct cancer classifications divided by the total number of classifications. It is computed by the expression shown below:

$$Classification\ accuracy = \frac{CC}{N} \times 100 \quad (1)$$

where N is the total number of the instances in the initial microarray dataset. And, CC refers to correctly classified instances. The primary objective of the research work is to identify the significant genes that improve the classification accuracy.

A Modified Genetic Algorithm with Levy Flight for Cancer gene selection

Genetic Algorithm (GA) is an evolutionary approach which is inspired on the principle of survival of the fittest (Goldberg, 2009). It is derived from the theory of evolution described by Charles Darwin in The Origin of Species. Genetic algorithms are search methods that employ processes found in natural biological evolution (Jong, 2005). GAs comprise a subset of these evolution-based optimization techniques focusing on the application of selection, mutation, and recombination to a population of competing problem solutions (Mitchell, 2002). It has received much consideration regarding its application potential as continuous and discrete optimal problem. Therefore, in recent years various optimization problems have been resolved by GA method. However, for a complex optimization problem, classical GA algorithm can be easy to cause stagnation premature convergence behavior.

The overall performances of the metaheuristic algorithms are mainly dependent on two properties of the algorithm: diversification and intensification, also mentioned as exploration and exploitation. Generally, in GA method cross over and mutation are the two important parameters that largely affect the performance of GA algorithm. However, Mutation is an in effective way of leading the population to escape from local minima. We present a modified GA-LV method for selecting significant genes. In GA-LV a fraction of the worst chromosomes are abandoned and new ones are built via levy flight principles (Yang and Deb, 2009). After few (every 20) generation, the chromosome will replace by levy flight. Through a repeated process of reproduction, that is genetic principle with levy flight, the population is led towards the global optimum. Levy flights is preferred over other simple random walks because it leads to better overall performance of the CS. The general equation for the Levy flight is given by:

$$x_i(t+1) = x_i(t) + \alpha \oplus \text{Levy}(\lambda) \quad (2)$$

Where t indicates the number of the current generation and $\alpha > 0$ represents the step size, which should be related to the scale of the particular problem under study. The symbol θ is used to indicate the entry wise multiplication. Note that is essentially a Markov chain, since next location at generation t+1 only depends on the current location t at generation and a transition probability, given by the first and second terms respectively. This transition probability is modulated by the Levy distribution as:

$$\text{Levy} \sim u = t^{-\lambda} \quad (3)$$

which has an infinite variance with an infinite mean. Here the steps essentially form a random walk process with a power-law step-length distribution with a heavy tail.

Pseduocode of GA-LV

choose initial population

evaluate each individual's fitness

determine population's average fitness

repeat

select best-ranking individuals to reproduce

mate pairs at random

apply crossover operator

apply mutation operator

evaluate each individual's fitness

determine population's average fitness·

If population stagnates for designated number of iterations then apply levy flight to generate new chromosomes·

until terminating condition (e·g· until at least one individual has the desired fitness or enough generations have passed)

Results

Datasets

In this work, the cancer gene expression datasets from Kent ridge biomedical data repository (http://leo.ugr.es/elvira/DBCRepository/) are used for experimental purpose. In order to evaluate the performance of the proposed method, five well known cancer gene

Table 1. Description for the Test Databases

Number	Name of data set	Number of examples	Number of genes	Classes
1	Leukemia	72	7129	2
2	Prostate	102	12600	2
3	Colon	62	2000	2
4	Lung	181	12533	2
5	Lymphoma	77	7129	2

expression datasets are analyzed. The details of the datasets are given in Table 1. In the columns Class1 and Class2, the numeric value within the bracket denotes the number of samples. In this work, the essential parameters of GA-LV are adopted from benchmark algorithms (Yang and Deb, 2009; Goldberg, 2009) .

Figure 1 shows the accuracy obtained for selected top 10 genes from KNN, SVM, NBC, GA and GA-LV for leukemia, prostate, colon, lung, lymphoma datasets (Duan et al., 2005). The results show that the proposed GA-LV algorithm outperforms existing statistical methods and GA in all five Cancer gene expression data sets.

Comparative Analysis

We compare GA-LV with KNN, SVM and NBC. Since these three algorithms are running in the same environment, parameters, and data sets, the results are absolutely comparable. Table 2 lists the highest accuracy in 100 independent executions of each method for each data set. The performance comparison shows that, compared to KNN, SVM and NBC has an obvious advantage. The proposed feature selection method gives 100% classification accuracy for Leukemia, Lung Cancer Michigan, Lymphoma with all the classifiers. For colon outcome dataset and Prostate outcome dataset the classification accuracy obtained by all the classifiers is

Table 2 Comparison of Accuracy with the Proposed Algorithm

Dataset	KNN	SVM	NBC	GA	GA-LV
Leukemia	100	100	100	100	100
Prostate	86.06	91.58	93.02	97.32	99.51
Colon	93.55	92.52	91.94	95.94	99.22
Lung	90.77	99.96	100	100	100
Lymphoma	97.41	99.79	100	100	100

Table 3 Top 5 Genes with the Highest Selection Frequency of Leukemia Data Set

Gene Name	Probe id	Gene Description
NCBP1	D32002_s_at	Nuclear cap binding protein subunit 1, 80kDa
LCK	U23852_s_at	Lymphocyte-specific protein tyrosine kinase
CLDN10	U89916_at	Claudin 10
MYH7	Z20656_rna1_s_at	Myosin, heavy chain 7, cardiac muscle, beta
PTGDR	U31099_at	Prostaglandin D2 receptor (DP)

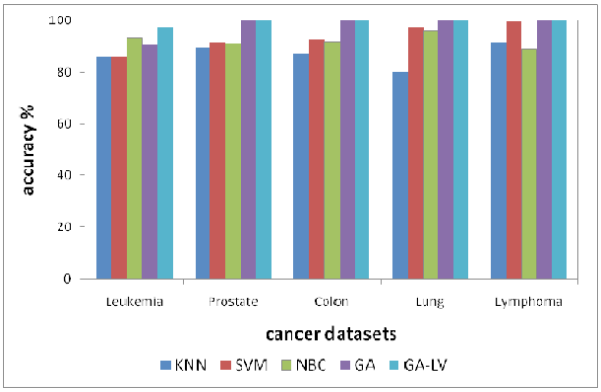

Figure 1. Datasets Versus Accuracy on Various Approaches

99.22% and 99.51% respectively. In terms of the correct rate, the search capability of GA-LV is stronger than the other three competitors. GA-LV is much better than existing statistical methods on all datasets. Moreover proposed method returns better performance than network motif-based method for selecting high-stability significant expression-correlation differential motifs (HSCDMs) (Lina et al., 2013).

Biological Analysis of Selected Genes for Leukemia Cancer data

Finally, the best subsets of genes were found for each data set. We add up all subsets having the highest accuracy and list the selected genes. For leukemia data set, the top 10 genes with the highest selection frequency of each microarray data are presented in Table 3. The roles or activities of the selected genes have been confirmed by both the biological experiment, as well as by the GA based method. It means that they have a higher chance of acting as biomarkers for the disease. These genes are proved to be the potential reason for leukemia cancer.

Discussion

Tumors are generally labeled on the basis of histological features. To achieve a thorough and comprehensive insight of the biology in tumors, microarrays provide a powerful strategy. Microarray establishes the gene expression signatures associated with different phenotypes. Gene expression profiles are useful to separate tumors into new and well established tumor types. In this paper, a modified Genetic Algorithm with Levy flight (GA-LV) classifier has been proposed and implemented to improve the classification accuracy in cancer prediction. Classification accuracy using various data mining algorithms like KNN, SVM and NBC induction was investigated. GA-LV shows promising results compared with KNN, SVM and NBC methods. The classification accuracy obtained by GA-LV based feature selection method outperforms statistical measures based method. It is observed that the genes selected by the GA-LV for different cancer datasets, have significant biological relevance. The results show that, more number of gene involvements does not improve the accuracy of the classifier. Only the informative gene selection leads to improve the classification accuracy. The

classification accuracy of our method was almost higher than ones of all individual gene approaches and HSCDMs.

References

Alon U, Barkai N, Notterman DA, et al (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. *Proc Natl Acad Sci U S A*, **96**, 6745-50.

Balamurugan R, Natarajan AM, Premalatha K (2017). Cuckoo search with mutation for biclustering of microarray gene expression data. *Int Arab J Info Tech*, **14**, 300-6.

Balamurugan R, Natarajan AM, Premalatha K (2016). A modified harmony search method for biclustering microarray gene expression data. *Int J Data Min Bioinf*, **16**, 269-9.

BilBan M, Buehler LK, Head S, Desoye G, Quaranta V (2002). Normalizing DNA microarray data. *Mol Biol*, **4**, 57-4.

Chen L, Qu X, Cao M, et al (2013). Identification of breast cancer patients based on human signaling network motifs. *Sci Rep*, **3**, 1-7.

Cho SB, Won HH (2003). Machine learning in DNA microarray analysis for cancer classification. *Proc Asia-Pac Bioinf Conf Bioinf*, **19**, 189-8.

Duan KB, Rajapakse JC, Wang H, Azuaje F (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans Nanobioscience*, **4**, 228-4.

Goldberg DE (2009). Genetic algorithms in search, optimization and machine learning, Addison-Wesley, Boston, pp 28-41.

Jong KD (2005). Genetic algorithms: A 30 year perspective, in perspectives on adaptation in natural and artificial systems, Oxford University Press, London, pp 11-8.

Li Y, Kang K, Krahn J M, et al (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics*, **18**, 508.

Mitchell M (2002), An introduction to genetic algorithms, Prentice Hall, New Delhi, India, pp 87-93.

Momiao X, Li W, Zhao J, Li J, Eric, B (2001). Feature (Gene) selection in gene expression-based tumor classification. *J Mol Gene Meta*, **73**, 239-7.

Premalatha K, Balamurugan R, Kannimuthu K (2017). Stellar mass black hole for engineering optimization. Handbook of recent developments in intelligent nature-inspired computing, IGI global book series advances in computational intelligence and robotics, pp 62-0.

Pyingkodi M, Thangarajan R (2017). Meta-analysis in autism gene expression dataset with biclustering methods using random cuckoo search algorithm. *Asian J Res Soc Sci Humanit*, **7**, 186-4.

Yang X S, Deb S (2009). Cuckoo search via Lévy flights. *NaBIC*, **2009**, 210-4.