International Journal of Pure and Applied Mathematics
Volume 116 No. 10 2017, 275-279
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

MICROARRAY GENE EXPRESSION ANALYSIS USING GENETIC ALGORITHM

Dr.C.Nalini[1]
ProfessorBIST, BIHER, Bharath University, India
[1]Nalini.cse@bharathuniv.ac.in

Abstract: Microarray studies and gene expression analysis have received tremendous attention over the last few years. In this paper we present a Genetic Algorithm (GA) based analysis of gene expression dataused to select genes that are highly related to tumorclassthat can be used to predict treatment.The genetic algorithm is used to select the best features from original data space, and the corresponding gene information is detected from the selected features. In the algorithm, gene expression data is the initial population. The population is subjected to fitness computation. The population that does not meet the fitness criteria it is updated by the genetic operations like selection, crossover and mutation and provide sequence of new populations is formed. This approach provides prediction accuracy in classification of gene data. For the experimental setup, leukemia gene expression data was used and is taken from Broad Institute, University of Cambridge.
KEYWORDS: gene expression, microarray, data mining, prediction, genetic algorithm.

1. Introduction

The improvement of microarray advances has empowered observing of the expression levels of thousands of qualities at the same time. Propels in microarray innovations have brought about a huge increment in the measure of natural information accessible. Given thousands of genes and Hundreds of samples, a very large number of gene expression profiles areanalysed [1].Microarray expressions are measured through hybridization process, the levels of genes expressed in biological samples. Microarray investigations can either screen every quality a few circumstances under fluctuating conditions or dissect the qualities in a solitary situation yet in various sorts of tissue [2]. In this paper, we focus on the latter where one important aspect is the selection of the recorded samples. This can be used to categorize different types of cancerous tissues with different class types are identified [3].The cancer gene expression datasets generated by microarrays generally have a high dimensionality as they are derived from a small number of patient samples. This can be problematic as when the number of samples is much smaller than the number of features, such data must be treated to reduce gene dimensionality. Also, it is not clear which genes are important and which can be omitted without reducing the classification performance [4-9].Genetic Algorithm(GA) is an evolutionary search and optimization algorithm based on the technique of natural genetics and selection[10-15].GA is used to select relevant subset of genes thereby reducing gene dimensionalityand the selected relevant subset of genes are highly related to tumour classes [16-18].

2. Related works

E.Bonilla Huerta (2010) proposed an inserted approach for determination of subset of qualities that groups malignancy class of small scale cluster information. The proposed approach starts with a channel procedure that pre-chooses an arrangement of qualities. A half and half hereditary calculation joined with fisher straight discriminant investigation. In this approach, Linear Discriminate Analysis (LDA) is utilized to get to the wellness of an applicant quality subset, and to illuminate the hybrid and change administrators of GA. This GA and LDA hybridization makes the hereditary look profoundly compelling for distinguishing little and instructive quality subsets.

Haibo Yao (2003) proposed directed measurement decrease and highlight extraction technique, hereditary calculation based specific vital part examination (GA-SPCA), is a commonsense approach to remove hyper unearthly picture highlights for various remote detecting applications. The first picture band number is initially diminished in the band choice stage. The picture measurement is therefore lessened and elements are extricated through a primary part examination stage. A hereditary calculation is utilized to streamline the band choice process. By utilizing this strategy, the first picture groups can be lessened to one or a few essential part picture groups, while as yet holding the vast majority of the application information data in the first picture. Exploratory outcomes from three ground reference datasets were empowering. By evacuating picture groups that don't add to data extraction for a particular application, the outcome connection progressed. The GA-SPCA technique accordingly can give a standard way to deal with hyper ghostly picture dimensionreduction and highlight extraction, and in

addition give valuable data to imaging sensor advancement.

Danh Cong Nguyen (2014) displayed a technique for assessing and utilizing Gene Regulatory Network (GRN)to foresee the potential for disease and finding the ideal approach to influence these administrative decides so that in the long run,the growth qualities wind up in a non-expressive mode. A PBN model was produced for a specimen of tumor and non-tumor cells from microarray information. After estimation of the comparing GRNs through relapse investigation and estimation of their long haul conduct by reproduction, it was shown that an improvement approach will help with recognizable proof of GRNs whose modifications would avert development of the disease. A hereditary calculation based recreation enhancement process was utilized to decide the specific principles inside the GRN to be changed and the substitution guidelines to be actualized. The recreation consequences of the proposed procedure demonstrated that the move of the cell with tumor making qualities a growth cell could be anticipated with a high likelihood. [18]

Jun Li (2013) proposed SePCA, a group of generative inert component models. The proposed show handles information of general sorts utilizing exponential family conveyances, which are parameterized by the dormant variables and coefficients. By applying programmed importance assurance (ARD) to the coefficients, SePCA consequently decides the suitable number of inert components for speaking to the information. Exponential family circulations assume the basic part of connecting realvalued elements and coefficients to information of general sorts. This empowers ARD to work on general sort information populace. The calculation of ARD quantitatively satisfies the instinct that a component ought to be valuable for speaking to the information. The discourse prompts to an example weight parameter ν. SePCA decides a ν in the preparation arrange. Specifically, the model is influenced by the specimen size and clamor. To illuminate, another conceivable expansion is to study display determination in administered and semi-managed settings. Semi-regulated learning issues are regularly talked about inside the system of locally straight subspace models, where expansion of SePCA might be considered to decide the model many-sided quality.[14]

3. Genetic algorithm

Hereditary calculation is a transformative hunt and enhancement calculation in view of regular hereditary qualities and determination. Quality expression dataset is the underlying population.The calculation starts by selecting an irregular beginning populace. At that point it makes a grouping of new populaces. At every progression, the calculation utilizes the people in the present era to make the following population.The shaping of future era of arrangements is accomplished by numerical operationsthat take after hybrid and transformation. The procedure begins bychoosing n possible arrangements as n strings as theoriginal populace. Every arrangement on this populace is evaluatedvia the target work for the issue and doled out a wellness esteem. The hereditary administrators of the procedure is compressed in thefollowing steps.

Hybrid:

Each match of arrangements is consolidated to generatea new combine of arrangements. This blend, frequently alluded to ascrossover, is proficient by breaking every chain at certainlocation and trading the half-chains between two stringsto accomplish two new strings. These strings are added to the populace, and their wellness qualities are assessed.

Determination:

Another pool of size n is then shaped by a processreferred to as choice. In this procedure, the new pool is formedby giving a higher likelihood of determination to strings with higherfitness values. This populace then turns into the objective for once more round of hybrid process.

Transformation:

Mutation causes singular strings to be changedaccording to some probabilistic run the show. Generally, just a little partof strings is changed by transformation, bringing about the posterity toinherit a large portion of the elements of the parent.

Termination:

The procedure ends when applicationsof the calculation don't bring about a noteworthy change inthe general wellness, or the calculation emphasis breaking point is come to. The hunt stops, and one of the bestsolutions is picked as the ideal for the issue.The Genetic algorithm performs the following steps:

Input: ,Microarray gene expression dataset

Step 1: Initialize population set ( ) from the gene expression dataset ,where $\{1 < i < n , 1 < j < k\}$ n is number of gene patterns and k is number of samples.

Step 2: The fitness of each chromosome is computed by determining

$f(x) = X_N$ –m, --→ Eq.(1)

Where, m=mean in XN

F(x) = 1/ (1 + f(x))--→ Eq.(2)

Step 3: The chromosomes (X) with maximum fitness is selected and is placed in theselection pool for crossover and mutation.

Selection: The probability of selecting i-th strings is

$$p_i = \frac{F_i}{\sum_{j=1}^{n} F_j} \text{ Eq. (3)} \longrightarrow$$

Step 4: Crossover operation is performed over the chromosomes at crossoverrate (Pc) at the selection pool.

Step 5: Mutation is performed at the child chromosomes described as $Xchild$.

Step 6: The new child is subjected to fitness computation. When the fitness criteria is satisfied, the new chromosomes are placed in selection pool. Otherwise, repeat the step(1) until it satisfies the same.

Step 7: Repeat the process from step(3) for number of iterations, until the best chromosomes with maximum fitness is selected ($Gbest$).

Output: A subset of informative genes.

4. Results and discussion

Datasets:

Our proposed approach for gene expression analysis is applied to the leukemia dataset in order to investigate their performances.

The leukemia gene expression data was used and is taken from Broad Institute, University of Cambridge.

Dataset consists of 7129 number of genes and 72 number of different samples.

Experimental Results:

A few arrangements of parameters of the hereditary calculation are utilized for this issue. The parameter set recorded underneath ended up being the most effective among the ones chose, i.e. Max era: 100 Crossover likelihood Pc = 0.5 Mutation likelihood Pm = 0.05.

The implementation of the experiment is by usingMatrixLaboratory (MATLAB) tool

The final results of selected genes based on the crossover and mutation probability of GAis obtained after the 59th number of iterations and the best individuals of the final population after optimization obtainedis plotted in the graph shown in Fig(1)

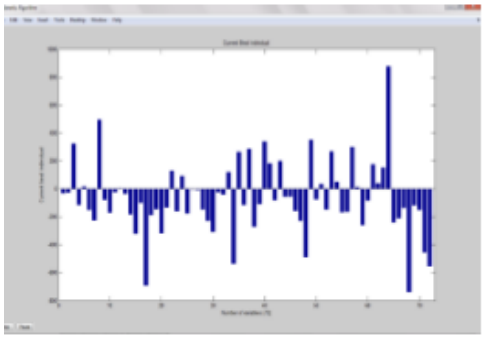

Figure 1

5. Conclusions

The paper presents a genetic algorithm for selecting subset of informative genes there by reducing the dimension of gene expression data. The selected genes can be used for diagnosis of leukemia gene expression data, thus providing treatment. There are many interesting aspects for future work. Further, this work can be extended to classification by neural networks and analyze the performance evaluation based on diagnosis result. Also, this can be further analyzed on multi-class problems and with different gene expression datasets

References

[1] Ding Jun Chen, Keith C.C.Chan and Xindong Wu., "Gene Expression Analyses using Genetic Algorithm based Hybrid Approaches", IEEE Congresson Evolutionary Computation(CEC 2008)

[2] Danh Cong Nguyen and FarhadAzadivar.,"Application of Computer Simulation and Genetic Algorithms to Gene Interactive Rulesfor Early Detection and Prevention of Cancer", IEEE Systems Journal, vol. 8, no. 3, September 2014

[3] SantanuGhorai, Anirban Mukherjee, SanghamitraSengupta, and Pranav K. Dutta.,"Cancer Classification from Gene Expression Data by NPPC EnsembleAlgorithm", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 8, No. 3, May/June 2011

[4] Udayakumar R., Kaliyamurthie K.P., Khanaa, Thooyamani K.P., Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, v-29, i-14, pp-86-90, 2014.

[5] Kaliyamurthie K.P., Parameswari D., Udayakumar R., QOS aware privacy preserving location monitoring in wireless sensor network, Indian Journal of Science and Technology, v-6, i-SUPPL5, pp-4648-4652, 2013.

[6] BrinthaRajakumari S., Nalini C., An efficient cost model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, v-7, i-, pp-45-46, 2014.

[7] BrinthaRajakumari S., Nalini C., An efficient data mining dataset preparation using aggregation in relational database, Indian Journal of Science and Technology, v-7, i-, pp-44-46, 2014.
[8] Khanna V., Mohanta K., Saravanan T., Recovery of link quality degradation in wireless mesh networks, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4837-4843, 2013.
[9] Khanaa V., Thooyamani K.P., Udayakumar R., A secure and efficient authentication system for distributed wireless sensor network, World Applied Sciences Journal, v-29, i-14, pp-304-308, 2014.
[10] Udayakumar R., Khanaa V., Saravanan T., Saritha G., Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction, Middle - East Journal of Scientific Research, v-16, i-12, pp-1781-1785, 2013.
[11] Khanaa V., Mohanta K., Saravanan. T., Performance analysis of FTTH using GEPON in direct and external modulation, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4848-4852, 2013.
[12] Kaliyamurthie K.P., Udayakumar R., Parameswari D., Mugunthan S.N., Highly secured online voting system over network, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4831-4836, 2013.
[13] Thooyamani K.P., Khanaa V., Udayakumar R., Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, v-20, i-12, pp-2464-2470, 2014.
[14] R.Kalaiprasath, R.Elankavi, Dr.R.Udayakumar, Cloud Information Accountability (Cia) Framework Ensuring Accountability Of Data In Cloud And Security In End To End Process In Cloud Terminology, International Journal Of Civil Engineering And Technology (Ijciet) Volume 8, Issue 4, Pp. 376–385, April 2017.
[15] R.Elankavi, R.Kalaiprasath, Dr.R.Udayakumar, A fast clustering algorithm for high-dimensional data, International Journal Of Civil Engineering And Technology (Ijciet), Volume 8, Issue 5, Pp. 1220–1227, May 2017.
[16] R. Kalaiprasath, R. Elankavi and Dr. R. Udayakumar. Cloud. Security and Compliance - A Semantic Approach in End to End Security, International Journal Of Mechanical Engineering And Technology (Ijmet), Volume 8, Issue 5, pp-987-994, May 2017.
[17] Thooyamani K.P., Khanaa V., Udayakumar R., Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, v-20, i-12, pp-2604-2612, 2014.
[18] Udayakumar R., Thooyamani K.P., Khanaa, Random projection based data perturbation using geometric transformation, World Applied Sciences Journal, v-29, i-14, pp-19-24, 2014.
[19] Udayakumar R., Thooyamani K.P., Khanaa, Deploying site-to-site VPN connectivity: MPLS VsIPSec, World Applied Sciences Journal, v-29, i-14, pp-6-10, 2014.
[20] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous wireless ad hocnetwork using FRCC." Australian Journal of Basic and Applied Sciences 9.7 (2015): 698-702.
[21] S.V.Manikanthan and K.Baskaran “Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network”, CiiT International Journal of Programmable Device Circuits and Systems,Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue : November 2011, PDCS112011008.
[22] T. Padmapriya and V. Saminadan, “Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks”, International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.
[23] C.Ashwini , J.Muthu , B.Karthikeyan , V.Vinith Raj, “An Efficient Auditing Technique for Secure Cloud Computing Using Asymmetric Cryptographic Algorithm” , International Innovative Research Journal of Engineering and Technology, Vol. 1, no. 1, pp. 23-26, 2015.