International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Enhanced Imputation Method Combining Single and Multiple Methods to Handle Missing Values in Microarray Data

D Saravanakumar [1], S K Mahendran [2]
[1]Research Scholar (Computer Science), Bharathiar University, Coimbatore,
Tamil Nadu, India. https://orcid.org/0009-0005-9239-4362, E-mail: saranji@gmail.com
[2]Asstistant Professor, Department of Computer Science, Government Arts College
Coimbatore, Tamil Nadu. India. E-mail: sk.mahendran@yahoo.co.in

Abstract

Gene Expression Classification (GEC) is a modern healthcare approach for enhancing present medical practices by classifying patient's gene structure to different types of cancer so as to provide effective and personalized treatments especially for all types of cancer. The GEC system aids medical practitioner in providing personalized treatments. The proposed GEC system assess the gene structure of a cancer patient through highly intensive computational intelligence technique named Genetic Algorithm (GA). In GA, the search space is composed of candidate solutions to the problem i.e. the collection of gene expression in the corpus, which is going to be used for training the computation model, which can further be used for testing new cancer patients in order to make accurate prediction about the presence of cancer cells. This will enable doctors to treat different cancer patients differently. In this proposed approach, each gene expression has been represented by a vector termed as chromosomes. In each generation, the chromosomes are selected randomly and fitness is evaluated. The probabilistic similarity function is used to estimate the fitness of the chromosome to predict the patient health condition. Experimental results show that the proposed approach works with relatively better accuracy compared to that of baseline approaches.

Key Words: Gene Expression, Computational Intelligence, Genetic Algorithm, Cancer treatment, Personalized medicine.

1. Introduction

Gene expression is the process by which the instructions in our Deoxyribonucleic acid (DNA) are converted into a functional product, such as a protein. Most of the modern healthcare practices completely ignore the cancer patient's gene expression in planning the treatment. The traditional healthcare practices considers only a viable symptoms of a patients and hence, the computational intelligence enabled gene expression classification method has been proposed so as to incorporate unknown and invisible symptoms at very low level i.e. at gene structure level. This would provide additional biological information about cancer patients, which in turns provide a prediction on type of cancer. Thereby, medical practitioner would have clear track information about patient healthcare stage. This would further enable doctors to provide treatment appropriately. This is the motto behind the proposed approach towards personalized cancer treatment through classifying patient into the different stages of cancer by analyzing gene expression data.. The typical gene expression classification process has been shown in Fig 1.

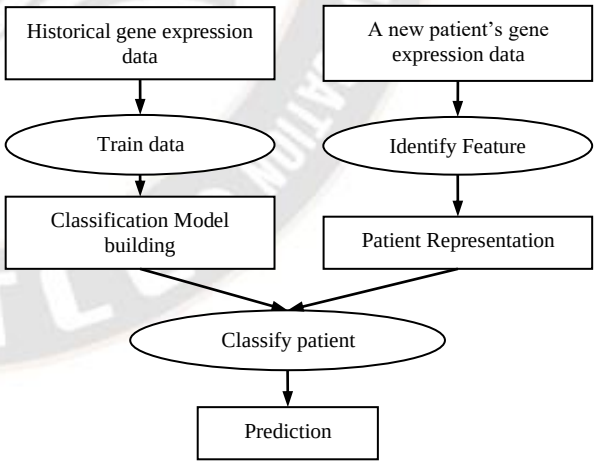

Fig. 1. Gene expression classification process.

In this paper, the Genetic Algorithm (GA) has been adopted to boost classification accuracy in gene expression classification. GA is an algorithm used for simulating the mechanism of natural selection of living organisms and is often used to solve problems having multitudes of solution. In GA, the search space is composed of candidate solutions to the problem each represented by a string termed as a

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

chromosome. Each chromosome has an objective function value, called fitness. A set of chromosomes together with their associated fitness value is called the population. The population at a given iteration of GA is called as a generation.

A model of gene expression classification model predicts whether the patient biological structure shows any symptom of cancer or not. In this proposed scheme, the more importance has been extended to feature extraction process. In this regard, a few well known models of representing patient health records have been demonstrated.

2. Problem Statement

The most of the classification strategies proposed in the literature under perform in case of high dimensional data. therefore, the proposed method adopt the computational intelligence strategy named Genetic Algorithm, which explore high dimensional data through varying number of generation or iteration in order to predict cancer disease by analyzing gene expression data. Hence, the aim of the proposed work is to develop a novel classification strategy for gene expression data in order to enhance the accuracy of prediction especially improve performance in high dimensional data problems.

3. Solution outline

The fitness function to be used in GA for making accurate prediction of cancer disease by analyzing gene expression data uses probabilistic weighted model. In probabilistic model, each feature T in a gene expression data D assigned a probability P(T|D) instead of just marking presence/absence of the feature. Thus, every feature in a gene expression will be weighted by P(T|D), where P(T|D) is the probability i.e. associated with each feature in a gene expression data. For example, a patient health records, which contains a set of features T in a sequence of gene expression D, possesses the highest probability P(T|D) will be given higher rank and appropriately prediction will be made by using probabilistic Bayes' rule as given below equation 1.

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)} \qquad (1)$$

Where, P(T) is constant for any given feature T, P(T|D)P(D) is a quantity proportional to the value of P(D|T), and P(D) is the prior-probability of being predicted as malignant or benignant. This P(D) will be estimated based on the historical patient gene expression data i.e. P(D) can be estimated by statistics based on corpus. P(T|D) would be estimated by storing the conditional probability table in order to estimate the maximum likelihood in order to classifiy the patient health record consisting of gene expression as malignant or benignant.

The proposed work concentrates on applying GA with an adaptation of the above said probabilistic model. The probabilistic classification method has been used for fitness evaluation which leads to a better classification performance than that obtained by using a traditional approaches presented in the literature.

4. Genetic Algorithm based Information Retrieval

GA is a generative procedure which maintains a set of population of feasible solutions. The fitness of each candidate solution is evaluated at each iteration i.e. generation. The chromosomes i.e. candidate solution that possess higher fitness values will be chosen for next generation process. These higher fitness candidate solutions will be selected for reproduction operation i.e. crossover and mutation to form new and better candidate solutions. The new set of population will be processed as above until we get an optimal solution or reach maximum generation. The approach presented in this paper investigates the use of GA in GEC.

The probabilistic classification model based on GA works as follows. The features in each gene expression data in a collection will be assigned a probability P(T|D). The data of patient susceptible of cancer can be formulated a query using a set of features T, then the gene expression D, which possesses the highest probability P(T|D) will be given higher rank and classified as malignant or benignant for the patient. Here, GA approach has been adopted to find out the gene expressions which possesses higher probability for the patient health record. The detailed probabilistic model explanation has been given in [2].

Representation of Chromosomes: The gene express corpus uses binary representation. For example, if there are 801 instances in the corpus, then the chromosomes to represent each of the documents will have ceil($\log_2(801)$) binary strings i.e. 10. We take 801 instances of gene expression as our data collection. Out of which 10 or 15 can be the initial population size.

A collection of n instances can be represented in the model by a term-document matrix. An entry in the matrix corresponds to the weight of a feature in the gene expression data is computed using the probability P(T|D); zero means the feature has no significance in the gene expression or it simply does not exist in the gene expression.

$$\begin{array}{r} 0000000000 \\ 0000000001 \\ \vdots \\ \vdots \\ 1111111111 \end{array} \begin{bmatrix} & T_1 & T_2 & T_3 & \cdots & T_t \\ D_1 & w_{11} & w_{12} & w_{13} & \cdots & w_{1t} \\ D_2 & w_{21} & w_{22} & w_{23} & \cdots & w_{2t} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nt} \end{bmatrix}$$

Fig. 2. Representation of documents as chromosomes.

Our GA approach takes an initial population chromosomes corresponding to the top 15 documents retrieved from the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

current web search engine with respect to the query.

Fitness Function: Fitness function is an evaluation measure which evaluates the candidate solution. The solution to the gene expression classification problem is achieved as long as the fitness function is good and appropriate.

Typically, the fitness function results in a probability value between 0 and 1. Fitness value close to 1 is assumed to be patient is malignant whereas the value near to 0 is assumed to be patient is benignant. The fitness function is defined as shown in Eq. (2).

$$P(q \mid d) = \prod_{q \in q} \sum_{w \in d} (P(q \mid w) P(w \mid d)) \quad (2)$$

Where, q is the patient identifier, d is the gene expression sequence, P(w|d) is the unigram probability of feature w in d, and P(q|w) is the probability of translating w into a patient identifier q. The unigram probability model is defined as shown in Eq. (3 to 6). In order to bridge the lexical gap between patient and gene expression data, the translation based approach allows a gene expression data to be translated semantically as shown below.

$$P(w \mid d) = \frac{Count(w)}{Count(words_d)} \quad (3)$$

$$P(q \mid d) = \alpha P(q \mid C) + (1-\alpha) \sum_{w \in d} (P(q \mid w) P(w \mid d)) \quad (4)$$

$$P(q \mid C) = \frac{Count(q\_in\_C)}{|C|} \quad (5)$$

$$P(w \mid d) = \frac{Count(w\_in\_d)}{|d|} \quad (6)$$

Where, α is a parameter for balancing the importance of the presence of a feature across entire gene expression collection C and the gene sequence d. α is experimentally tuned and set as 0.5. |C| and |d| are the sizes of collection and the individual gene expression.

Selection: The proposed approach selects the initial candidate solution i.e. initial population by random selection. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated; multiple individuals are selected from the current population based on their fitness, and modified to form a new population. The new population is used in the next generation i.e. iteration of the algorithm. There are many different selection techniques are available to select the individuals to be used over the next generation. In this work, we have used two selection mechanisms namely roulette-wheel selection and elitism.

1) Roulette-wheel selection: the fittest is the solution with the most chances to be chosen. Conceptually, this can be represented as a game of roulette wherein each individual gets a slice of the wheel. The fittest one gets larger slice than the less fit ones.

Table 1. Chromosomes and its fitness value

Chromosomes	Fitness value	% of chance to be chosen
100010110111	0.021	2%
110000101110	0.059	6%
100011001101	0.127	14%
110100010111	0.415	32%
111100110100	0.29	46%

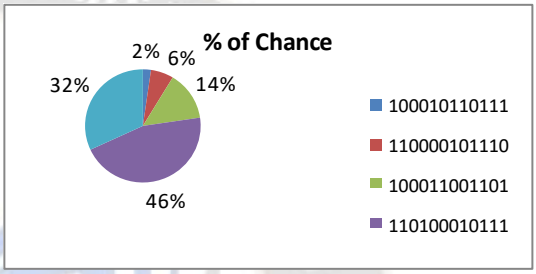

Fig. 3. Roulette-wheel selection of chromosome

2) Elitist selection: The fit members of each generation are guaranteed to be selected. Fitness value of a chromosome which is above 0.7 is retained for further generations.

Reproduction Operators: Once the selection process has chosen fit individuals, they need to be randomly altered with the hope of improving their fitness for the next generation. There are two basic strategies to perform reproduction.

1) Crossover: The chromosomes of the two parents are randomly recombined to form offspring. Uniform crossover operation is adopted for recombination. Here, a random mask is generated which determines which bits are to be copied from one parent and which from the other parent. The density of bits '1' and '0' in mask determines how many bits from first and second parent are to be taken to form new offspring. The probability of crossover (P_c) is the probability that the crossover will occur at a particular mating i.e. not all mating must be reproduced by crossover. The probability of crossover is assumed to be 0.5.

Mask:	0110011000110	
Parents:	101000111010 0	0011010010001
Offspring:	0011001010101	1010010110000

2) Mutation: It generates new offspring from single parent. Mutation operator is performed to introduce diversity in the population. Mutation is required

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

because, even though crossover effectively search and recombine solutions at some extent, rarely they may lose some potentially useful genetic combination. The mutation operator protects against such an irrecoverable loss. Mutation is assumed to be a random walk through the solution space e.g. randomly mutating chromosome at position 7. The probability of mutation (P_m) of a bit is assumed as 1/l where l is the length of the chromosome i.e. 1 out of 12 bits is chosen at random and modified.

Algorithm: Consider d as gene expression, q as patient identifier, w as features in d.

```
for i=1 step 1 to q
    for j=1 step 2 to w in d
        Compute P(q|d)
    end step 2
end step 1
for k=1 step 3 to |d|
        Order the gene expression sequence
according to P(q|d)
end step 3
```

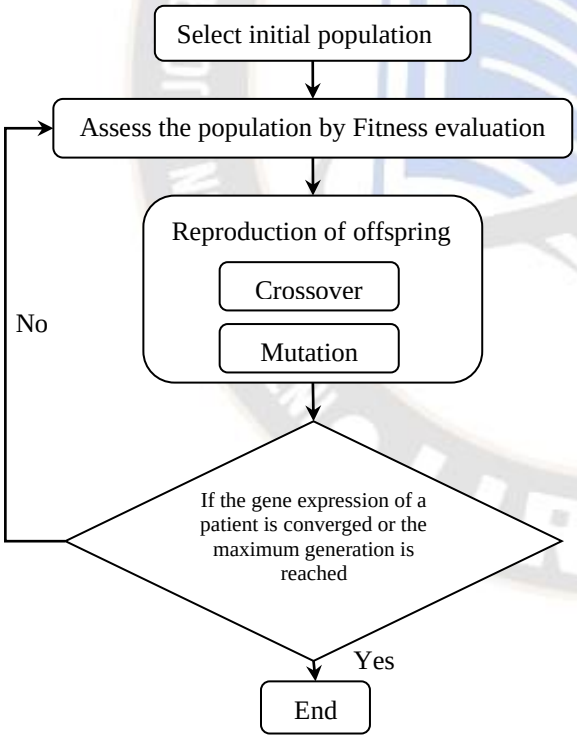

Fig. 4. Steps in Genetic Algorithm for Gene Expression Classification

5. Experimental Evaluation

5.1 The Dataset

The dataset considered for the experiment is a collection of data is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having different types of cancer tumor such as BRCA, KIRC, COAD, LUAD and PRAD. The data is collected from UCI machine learning repository. The dataset contains 801 instances with 20531 attributes. It has a multivariate characteristics. This data has been widely used as bench mark data for many applications of classification and clustering. In the proposed method, the dataset has been used for classification, as it is the ideal aim of the paper. The entire dataset has been split into 70% of training set and 30% of test set.

5.2 Evaluation metrics

In order to evaluate the results of the proposed approach, the following classification metrics have been measured.

Table 2. Accuracy Metrics for two class classifier

	Observed label	
Actual label	True positive	False positive
	False negative	True negative

Precision: This measures the accuracy of the classified results. Precision defines the fraction of gene expressions assigned class i that are actually about class i, where i could take possibly two values such as malignant or benignant.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{7}$$

Recall: This measures the coverage of the predictions by the classification algorithm. Recall defines the fraction of gene expression that are correctly classified.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{8}$$

F-Measure: This measure combines precision and recall i.e. the harmonic mean of precision and recall. Recall and precision are evenly weighted by F-Score.

$$F = 2 * \frac{\Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \tag{9}$$

5.3 Experimental Results

The following experiments have been carried out in order to show the effectiveness of the proposed GA based approach on the PANCAN data set. It is observed that the proposed classification approach works well compared with other baseline works.

Experiment 1: Accuracy of the classification performance has been tabulated by measuring precision at varying size of testset i.e. k = 5, 10, 15, 20, 25, 50, 100, and 500. It is observed that proposed GA results are found to be relatively better than baseline system with the minimum improvement of 15% on precision i.e. accuracy of classification. Especially, the proposed approach works well when the size of the test set is increasing.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Table 3. Precision on PANCAN data set

Classification of 1st k gene expression instances	Precision on PANCAN data set					
	SVM	BEL Network	KNN	Fuzzy SVM	Proposed GA based Approach	Average improvement in %
5	0.65	0.78	0.82	0.88	0.93	0.15
10	0.643	0.71	0.77	0.86	0.91	0.16
15	0.61	0.76	0.78	0.863	0.92	0.17
20	0.56	0.67	0.731	0.84	0.927	0.23
25	0.58	0.63	0.7	0.82	0.925	0.24
50	0.523	0.601	0.71	0.8	0.93	0.27
100	0.512	0.55	0.69	0.78	0.95	0.32
500	0.52	0.57	0.69	0.74	0.945	0.32

Experiment 2: The coverage of the predictions by the classification algorithm has been tabulated by measuring recall at varying size of test set i.e. k = 5, 10, 15, 20, 25, 50, 100, and 500. It is observed that proposed GA results are found to be relatively better than baseline system with the minimum improvement of 13.5% on recall.

Table 4. Recall on PANCAN data set

Classification of 1st k gene expression instances	Recall on PANCAN data set					
	SVM	BEL Network	KNN	Fuzzy SVM	Proposed GA based Approach	Average improvement in %
5	0.64	0.69	0.73	0.86	0.95	0.135
10	0.593	0.65	0.73	0.86	0.94	0.13
15	0.63	0.67	0.72	0.874	0.945	0.135
20	0.62	0.68	0.721	0.86	0.96	0.16
25	0.61	0.64	0.66	0.86	0.956	0.18
50	0.58	0.61	0.64	0.832	0.95	0.21
100	0.57	0.6	0.61	0.81	0.96	0.23
500	0.556	0.595	0.54	0.82	0.956	0.234

Experiment 3: In this experiment, classification computational time combined with feature selection for various classifiers has been empirically tested and measured on PANCAN dataset. The results are tabulated in Table 5.

Table 5. Comparison of computational time on PANCAN data set

Classification of 1st k gene expression instances	Computational time consumed (in sec.) on PANCAN data set				
	SVM	BEL Network	KNN	Fuzzy SVM	Proposed GA based Approach
5	0.5	0.49	0.37	0.7	0.75
10	0.5	0.5	0.4	0.75	0.76
15	0.52	0.56	0.41	0.76	0.8
20	0.53	0.56	0.42	0.78	0.83
25	0.53	0.565	0.45	0.78	0.8
50	0.55	0.565	0.67	1.23	1.14
100	1.06	1.3	1.59	2.03	1.78
500	1.65	1.39	1.63	3.39	2.27

5.4 Analysis

Experiments were carried on varying number of instances from test data for various baseline methods and it is observed that the approach performs well on the test dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

with the crossover probability (P_c) of 0.5 and mutation probability (P_m) of 0.08. It is verified that this setting yields reasonable improvement in precision, recall and computation time as well.

Fig. 7 depicts the fitness value obtained in each generation against the number of generations. It is observed that there is a significant improvement on average fitness as the numbers of generations are increased. Thus, it is suggested that the system would improve classification accuracy as the number of generation gets increased. It is shown that the fitness value obtained over different runs of generation when P_c is set to 0.3, 0.4 and 0.5 respectively while P_m is set to 0.08. Thus, it is implied that P_c =0.5 will yield better average fitness.

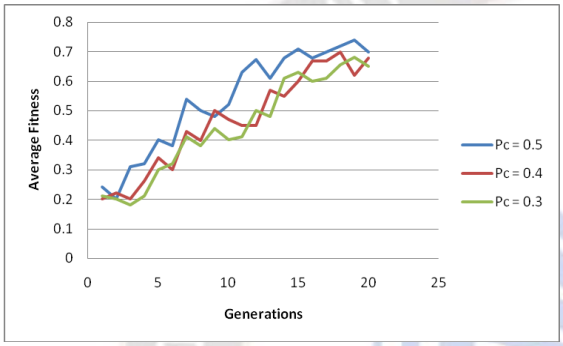

Fig. 5. Average Fitness vs. Number of Generations of GA based classification method

The proposed GA based gene expression classification approach has been implemented in python. The proposed algorithm along with baseline methods were run on standard system configuration with Core i7 machine with 8GB RAM.

6. Conclusion

Gene Expression Classification approach has been presented in this paper with appropriate empirical evidences about the results achieved. In general, the genetic algorithm terminates when either a maximum number of generations are reached or an acceptable fitness level has been achieved for the population of candidate solutions. If the algorithm is terminated due to a maximum number of generations specified, a satisfactory solution may or may not have been achieved. As a future work, this work could be extended to incorporate domain knowledge from expert medical practitioner along with system preferences on weighting the attributes associated with cancer tumor. The empirical results are obtained for a varying number of instances from 30% of test set. It is observed that the proposed classification approach work well in both the cases, when the number of instances from test set is small [5 to 25] and high [50 to 500] as well.

References

[1] Hechenbichler, K., Schliep, K, Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Collaborative Research Center, LMU University, Munich,Germany Tech. 2006

[2] Veningston. K, Shanmugalakshmi. R. Computational Intelligence for Information Retrieval using Genetic Algorithm. INFORMATION-An International Interdisciplinary Journal, Published by International Information Institute, Japan, Vol.17, No.8, pp. 3825-3832, 2014.

[3] Lotfi, E., Keshavarz, A, Gene expression microarray classification using PCA–BEL. Comput. Biol. Med. 54, 180–187, 2014.

[4] Veningston. K, Shanmugalakshmi. R. Statistical language modeling for personalizing Information Retrieval. In Proc. IEEE International Conference on Advanced Computing and Communication Systems, pp. 1- 6, 2013.

[5] Moteghaed, N.Y., Maghooli, K., Garshasbi, M., Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine. J. Med. Signals Sens. 8 (1), 2018.

[6] Mei, Z., Shen, Q., Ye, B. Hybridized KNN and SVM for gene expression data classification. Life Sci. J. 6 (1), 61–66, 2009.

[7] Nguyen, T., Khosravi, A., Creighton, D., Nahavandi, S., Hidden Markov models for cancer classification using gene expression profiles. Inf. Sci. (Ny) 316, 293–307, 2015.

[8] Parvin, H., Alizadeh, H., Minati, B., A modification on K-Nearest Neighbor classifier. Glob. J. Comput. Sci. Technol. 10 (14), 37–41, 2010.

[9] Veningston K, Seifedine Kadry, Haydar Sabeeh Kalash, B. Balamurugan, R. Sathiyaraj. Intelligent Social Network based Data Modeling for Improving Health Care. Springer Health and Technology, Vol.9, Issue No.31, 2019.

[10] Veningston, K, Shanmugalakshmi, R, Personalized information retrieval system using computational, Ph.D thesis, Anna University, Chennai, INDIA, 2015.

[11] [Dataset] Samuele Fiorini, samuele.fiorini@dibris.unige.it, University of Genoa, redistributed under Creative Commons license (http://creativecommons.org/licenses/by/3.0/legalcode) from https://www.synapse.org/#!Synapse:syn4301332.