Journal of Engineering and Applied Sciences 14 (21): 8080-8087, 2019
ISSN: 1816-949X

ECAGS: An Enhanced Cancer-Association based Gene Selection Technique for Cancer Patterns Classification and Prediction

[1]S. Subasree, [2]N.P. Gopalan and [3]N.K. Sakthivel
[1]Bharath University, Chennai, Tamil Nadu, India
[2]National Institute of Technology, Tiruchirappalli, Tamil Nadu, India
[3]Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

Abstract: Microarray based Cancer Pattern Classification and Prediction technique is one of the most efficient mechanisms in Bioinformatics research. This research work studied and analyzed thousands of genes simultaneously to understand the pattern of the gene expression. This research work focuses to identify and prioritize genes that are important for gene patterns classification and prediction. This research work proposed an Enhanced Cancer-Association based Gene Selection technique for Cancer Patterns Classification and Prediction (ECAGS). The proposed classifier is implemented and studied thoroughly in terms of memory utilization, execution time (processing time), classification accuracy, sensitivity, specificity and F score. The experimental results were compared with our previous model called an Enhanced Multi-Objective Particle Swarm (EMOPS). From our experimental results, it was noticed that the proposed model outperforms our previous model in terms of memory utilization, execution time (processing time), classification accuracy, sensitivity, specificity and F score.

Key words: Bioinformatics, cancer pattern classification, classification accuracy, dimensionality reduction, gene prioritization, gene association

INTRODUCTION

The average life time of human beings increases day by day because of lot of improvement in medical science. The Medical Science focuses towards disease management to identify the diseases properly and accurately (Subasree *et al*., 2016; Nikam, 2015). These gene patterns are usually available in microarray data in general are images and these microarray images could be converted into various gene expression. These gene expressions have been usually used for gene pattern classifications. The normal microarray sample data sets and cancer patterns samples can be classified with the help of classifiers. These diseases indicators such as genes and genomes are helping us to find the diseases (Zhang *et al*., 2015; Trivedi *et al*., 2016). Researchers have to identify genes those are needed to consider for analysis and this approach is called gene prioritization. It uses four different strategies like filtering, text mining, similarity profiling and data fusion. Filters are used to identify the ideal genes for classification. Text mining is used to find various diseases relevant keywords. Similarity profiling and data fusion are used to identify the similarity between the candidate genes from various known genes. It uses Gene Selection Expression Heterogeneity algorithm (GSEH) (Kim *et al*., 2018; Pardo *et al*., 2016) for the above mentioned purpose.

Literature review: In this study, this research work is planned to discuss a few recently proposed classifiers namely:

- Hybrid Ant Bee Algorithm (HABA)
- Multi-objective Particle Swarm Optimization (MPSO)
- EMOPS: An Enhanced Multi-Objective Pswarm based classifier

Hybrid Ant Bee Algorithm (HABA): Ant colony optimization, Gu (2016), Kumar *et al*. (2014) and Chakraborty and Maulik (2014) does maintain a colony of ants and make possible Permissible Ranges (PRs) in association with values proposed for a design model. Here, each and every ant is permitted to select a permissible range which will represent the path. When all ants have chosen their paths, then, the discrete value associated with the selected path is taken and for all ants, this is considered as candidate value. Then, the system evaluates the artificial bee colony approach by combining

Corresponding Author: S. Subasree, Bharath University, Chennai, Tamil Nadu, India the candidate values of all the ants and this initializes the food source and the objective function can be evaluated with three phases and those phases named as:

- Employed bee phase where food sources assigned to Bees
- Onlooker bee phase where a decision is taken by Bees
- Scout bee phase where ants making out the random search

MATERIALS AND METHODS

The proposed Ant Bee algorithm combines the strength of Artificial Bee Colony (ABC) and Ant Colony Optimization (ACO). The procedure of the Ant Bee algorithm is described below Algorithm 1.

Algorithm 1; Generate initial solution space:

```
Evaluate the fitness of objective function
if (Fitness Function Converged)
{ declare best solution stop()}
Spilt the database as clusters
ACO()
//probabilistic based optimization {Set Parameters, Initialize Pheromone Trails Construct path Select and Construct Ant Solution  Update Pheromones}
ABC()
// Optimizes through ABC algorithm
// Cluster based optimization based on intelligent foraging behaviour of bee
{
// No. of parameters D; //Function fn;//No. of Bees NB
// Lower Bound lb
//Upeer Bound ub
Declare par, fn,D,NP,lb,ub,limit
 Initialization of parameter par = 0
If(NP<limit)
{abc_optim(par, fn, D = length(par))}}
Combine the results of ABC() and ACO()
Construct solution
```

Multi-objective Particle Swarm Optimization (MPSO): The particle swarm optimization, Behravan *et al.* (2016), Leskovec *et al.* (2014), Li *et al.* (2015), Qiu *et al.* (2016) and Coello and Lechuga (2002) is one of the popular existing population based optimization techniques. The various candidate solutions are named as particle and the population of these particles is termed as swarm. Let us consider that there were, N particles in swarm to achieve optimal fitness. The particle best position pbest and global best position gbest need to update to attain and compute fitness The MPSO was developed, Gu (2016) by the researchers, Mukhopadhyay and Mandal (2014) as follows:

Input:

- Data matrix
- Cluster center (C)
- Particles (N)
- Samples (S)
- Assign thr = 0.5, Sample Velocity (SV)

Output A:

- Initialize random sample locations and SVs as well
- i): Genes xn, samples gene set GN and fitness Pn
- Initialize random sample locations and SV as well
- ii): Calculate cell boundary(xnd) for all cluster centres till xnd≥threshold
- Calculate cell boundary and average Velocity Vnd
- Select centres by evaluating and combining
- Take average calculation by crowding distance sorting for all derived solutions Select the best sample Gene Gn

EMOPS: An Enhanced Multi-Objective P Swarm based classifier for poorly understood cancer patterns: It was also noted that the Multi-objective Particle Swarm Optimization (MPSO) (Chakraborty and Maulik, 2014; Mukhopadhyay and Mandal, 2014; Yoon *et al.*, 2010) is relative out performing other two classifiers. To improve the performance of the Multi-objective Particle Swarm Optimization (MPSO), this study enhanced Multi-Objective Particle Swarm Optimization (MPSO) (Chen *et al.*, 2014; Ma *et al.*, 2015) and named as an Enhanced Multi-Objective Pswarm Based Classifier (EMOPS) proposed by Subasree *et al.* (2018) and the related procedures are described in the following study.

Procedure of Enhanced Multi-Objective Particle Swarm based classifier (EMOPS): As discussed in the previous section, the Multi-objective Particle Swarm Optimization (MPSO) considers the total number of particles to achieve optimal fitness. The particle best position pbest and global best position gbest will update to attain and compute fitness. This research work noticed that the position and parameter values need to optimize in such a way to achieve a high level of classification accuracy, i.e., need to determine optimized centre values to improve and achieve higher classification accuracy. To achieve higher classification accuracy, this research proposed an efficient model called an Enhanced Multi-Objective Pswarm Based Classifier (EMOPS). The procedure of EMPOS will consider multiple competing solutions to find global best position gbest which will improve classification and prediction accuracy. The procedure for the Enhanced Multi-Objective Particle Swarm based classifier (EMOPS) is given below:

Input:

- Data matrix
- Cluster center (C)
- Particles (N)
- Samples (S)
- Assign thr = 0.5, Sample Velocity (SV)

Output A:

- Initialize random sample locations and SVs as well
- i): Genes xn, samples Gene set Gn and fitness Pn
- Initialize random sample locations and SVs as well
- i) Calculate cell boundary (xnd) for all cluster centres till xnd≥threshold
- Cell boundary and average Velocity Vnd
- Calculate
- Strong-dominance updating strategy
- Compute crowding distance and refresh for next
- Iteration
- Estimates the largest rectangle size
- Calculate the average distance of its two neighbouring solutions
- Select centres by evaluating and combining
- Take average calculation by crowding distance sorting for all derived solutions
- ii): Select the best sample Gene Gn
- Select the global best position gbest

An Enhanced Cancer-Associated Gene Selection technique for Cancer Patterns Classification and Prediction (ECAGS): In this study, the approaches and various steps involved for dimensionality reduction, gene selection, matrix construction and association and ranking prediction in gene selection expression heterogeneity are discussed. Methodology of the proposed enhanced cancer-association based gene selection technique is shown in Fig. 1. As demonstrated in Fig. 1, this research work developed a tool that achieved higher classification accuracy. The various steps involved for this purpose is discussed below.

Step 1: Collection of cancer pattern gene sequence data sets from database
Step 2: Construct positive gene data sets for training and testing as well
Step 3: Perform dimensionality reduction through PCA
Step 4: Gene selection through matrix score
Step 5: Compute rank through gene association and create gene rank list
Step 6: Pattern classification with better prediction

Dimensionality reduction: Dimension reduction is one of the significant approaches in many real world studies because big data always has high dimensions and mapping the data from high dimensions to low dimensions is essential to increase the efficiency of data analysis and handling. To perform this dimensionality reduction, it is needed deep learning techniques that are used to focus the following purposes.

- Deep learning uses unsupervised training which focus to eliminates the need of labels for training
- Local optima can be prevented
- Data can be separated more easily
- Meaningful representations can be made

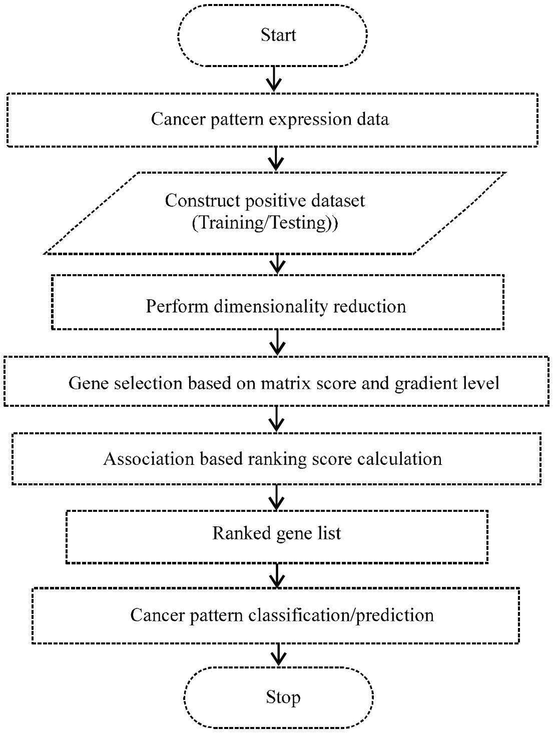
Start
Cancer pattern expression data
Construct positive dataset (Training/Testing))
Perform dimensionality reduction
Gene selection based on matrix score and gradient level
Association based ranking score calculation
Ranked gene list
Cancer pattern classification/prediction
Stop

Fig. 1: Methodology of the proposed enhanced cancer association based gene selection technique

The detailed collaborative filtering for dimensionality reduction is shown in Fig. 2.

Gene prioritization: By using a collaborative filtering, the predicted gene expression matrix is constructed. It has predicted values and it can be compared to the original values of the same region in the original gene expression matrix. The difference between the predicted expression levels of a gene and original expression levels of gene in a class and predicted expression levels of gene and the original expression levels of the gene in other class are dissimilar which has high possibility respect to the disease. The prioritization score R is ith gene is calculated as follows and the procedure is shown in Fig. 3 (Eq. 1):

$$R_i = \left|\left(\frac{\sum_{j-1}^{m}\left|OM_{1ij}-PM_{1ij}\right|}{m} - \frac{\sum_{j-1}^{m}\left|OM_{2ij}-PM_{2ij}\right|}{n}\right)\right| \quad (1)$$

Where:

OM_{1ij} = Expression level of the ith gene and jth sample of class 1 in matrix OM
OM_{2ij} = Expression level of the ith gene and jth sample of class 2 in matrix OM
PM_{1ij} = Expression level of the ith gene and jth sample of class 1 in matrix PM
PM_{2ij} = Expression level of the ith gene and jth sample of class 2 in matrix PM

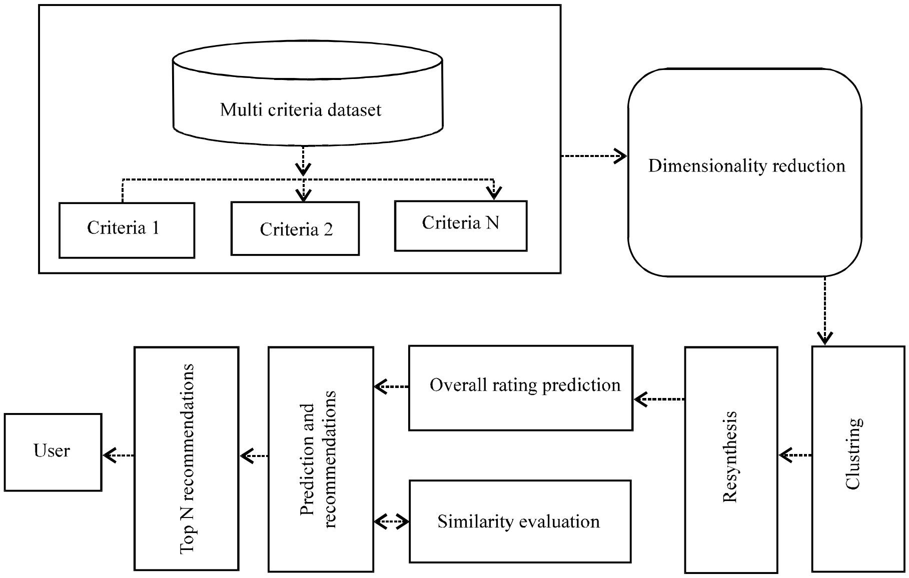

Fig. 2: Collaborative filtering using dimensionality reduction

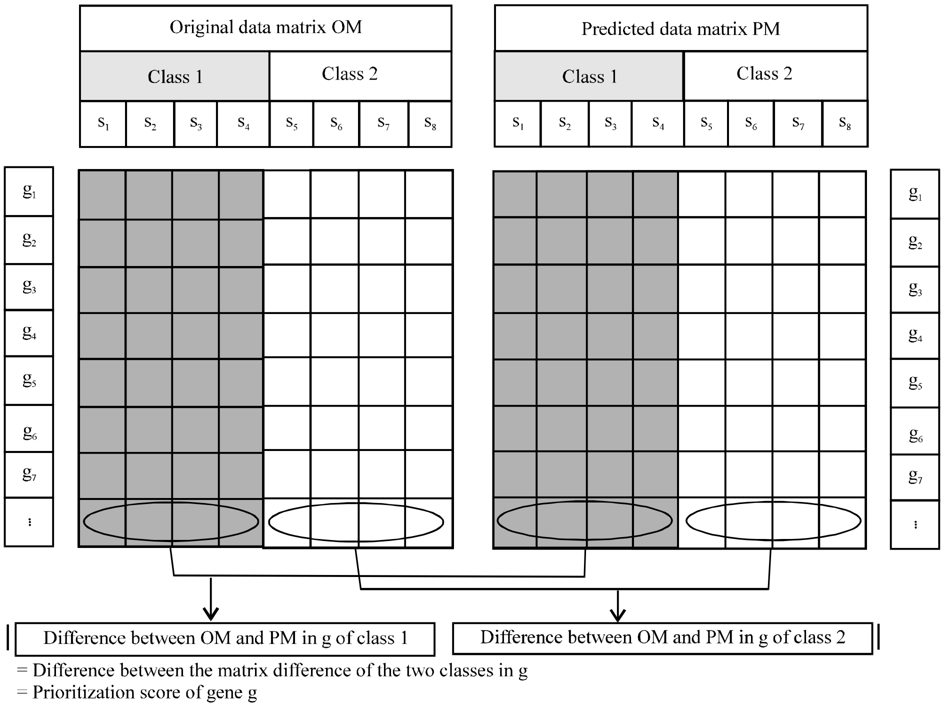
Original data matrix OM
Predicted data matrix PM
Class 1
Class 2
s1 s2 s3 s4 s5 s6 s7 s8
g1 g2 g3 g4 g5 g6 g7 ⋮
Difference between OM and PM in g of class 1
Difference between OM and PM in g of class 2
= Difference between the matrix difference of the two classes in g
= Prioritization score of gene g

Fig. 3: Calculation of gene prioritization

Principal Component Analysis (PCA): The principal component analysis is to employed in medical diagnosis, especially, in gene prediction. This method is commonly known as PCA. Most of the popular molecular dynamics packages inevitably provide PCA tool to analyze for identifying/predicting important genes. This algorithm uses the following steps to identify the important genes for analysis.

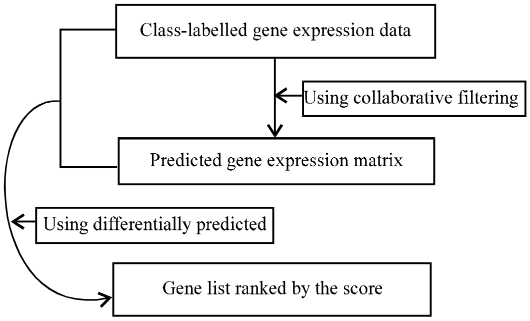
Class-labelled gene expression data
Using collaborative filtering
Predicted gene expression matrix
Using differentially predicted
Gene list ranked by the score

Fig. 4: GSEH process levels

Step 1: Make the input data zero-mean, i.e., substract the mean from each observation of each feature
Step 2: Calculate the covariance matrix for the zero-mean data
Step 3: Calculate
- The eigenvectors and eigenvalues for the covariance matrix
- The eigenvectors are the principal components
- Arrange the first principal component has the highest eigenvalue, the second principal component the second highest, etc

Step 4: Find
- The new feature vector and perform feature selection
- If the eigenvalues of some principal components is zero, the feature selection is done automatically

Step 5: Transform the data to the new feature system

Gene Selection Expression Heterogeneity technique (GSEH): The Gene Selection Expression Heterogeneity (GSEH) employs collaborative filtering to identify and select biologically meaningful candidate genes. The GSEH methodology is categorized into two levels.

The first level involves constructing a predicted gene expression matrix using collaborative filtering and the second level uses calculation of the rank scores of the genes using a comparison between the predicted and original gene expression matrix. After the calculation of the scores, the genes can be ranked in order and k top-ranked genes can be identified and selected. The GSEH process levels are explained in the following Fig. 4 and 5.

As shown in Fig. 4, all the gene expressions are labelled and classified as various clusters. The clustered data is filtered to eliminate dissimilarity data. Then clustered data were employed for predicting various gene patterns. All these gene patterns are ranked and scored. The sorted list will be recorded to identify/predict the relatively better gene expressions related to diseases patterns.

The detailed dimensionality reduction is performed through deep learning approach that eliminates dissimilarity data/labels to optimize the size of data sets that will facilitate to improve classification accuracy. This method that involved for dimensionality reduction is discussed in the study.

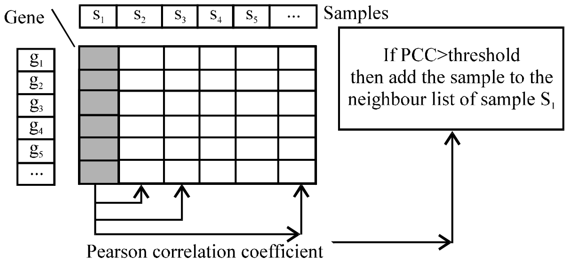
Gene
S1
S2
S3
S4
S5
...
Samples
g1
g2
g3
g4
g5
...
If PCC>threshold then add the sample to the neighbour list of sample S1
Pearson correlation coefficient

Fig. 5: Selection of neighbour samples using Pearson correlation coefficient

Predicted gene expression matrix construction: The gene expression data can be reconstructed by using commonly accepted mechanisms of collaborative filtering method. It has various form of filtering mechanisms it uses user-based collaborative filtering. The user based collaborative filtering is consists of two processes. The first step is focused to select the neighbour samples for a given sample. A neighbour sample is nothing but the sample it has the characteristics of similarity to the given sample and it uses the pearson correlation coefficient to find similarity between the selected samples. Pearson correlation coefficient P_{xy} can be calculated as follows:

$$P_{xy} = \frac{\sum\left[\left(X_i - \bar{X}\right).\left(Y_i - \bar{Y}\right)\right]}{\sqrt{\sum\left(X_i - \bar{X}\right)^2 . \sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}} \qquad (2)$$

Where:

X and $\bar{X}$ = The given sample and average gene expression for X
Xσ = The standard deviation of the average gene expression for X
X_i = Indicates the ith gene expression of sample X
Y = The remaining samples excluding sample X

The Pearson correlation coefficient lies between -1 to 1. The more association ie similar sample is identified, if the value is closest to 1 and non-association by means of value nearer to -1. To identify the gene expression levels it group similar expression patterns together in Pearson correlation coefficient. With the help of other samples the neighbors of the target will be identified. Compare the correlation between target and other sample is calculated. Select the Pearson correlation equal to or greater than threshold C are chosen as neighbors. If, we want to identify the neighbors of all cells by repeating this process for all cell in the dataset.

The next step is to find the gene expression level of a given cell based on neighbour's gene expression levels. The larger Pearson correlation coefficient than the threshold of c is responsible for predicting the gene expression levels of the samples. This collaborative filtering method focus to produce a prediction with average weighted neighbour of the samples. The predicted gene expression level is calculated as follows:

$$V_{ij} = \bar{S}_J + \frac{\sum S_n \in \text{Neighbour}\left((E_{in} - \bar{S}_n).P_{sj}P_{S_n}\right)}{\sum S_n \in \text{Nieghbour}\left|P_{s_j}P_{S_n}\right|} \quad (3)$$

In the equation predicted expression value for i-th row and jth column can be found in $V_{ij}\bar{S}_J$ is the average expression in all the genes of the sample S_j. S_n is one of the neighbor samples and E_{in} is the ith gene expression level of the neighbor sample S_n the Pearson correlation coefficient between the sample S_j and neighbor sample S_n is represented by $P_{Sj}P_{Sn}$:

$$V_{11} = \bar{S}_1 + \frac{\sum S_n \in \text{Neighbour}\left((E_{in} - \bar{S}_n).P_{s_j}P_{s_n}\right)}{\sum S_n \in \text{Nieghbour}\left|P_{s_j}P_{s_n}\right|} = \bar{S}_1 + \frac{\sum\left(\in_{13} - \bar{S}_3\right)P_{s_1}P_{s_3}S_n \in \text{Neighbour}\left((E_{15} - \bar{S}_5).P_{s_1}P_{s_5}\right)}{\left|P_{s_1}P_{s_5}\right| + \left|P_{s_1}P_{s_5}\right|} \quad (4)$$

If the sample S_j has no neighbors, if $V_{ij} = E_{ij}$ which indicates that the predicted value has the same gene expression as the original value, since, tit has no neighbours of S_j. By using this we can predict the level of expression for certain gene of a given sample using this equation. For example, we can predict the level of gene expression level by using gene g_1 and sample S_1.

Association based ranking: The rank calculation of the Gene Selection Expression Heterogeneity (GSEH) is discussed here. The following procedure shows the calculation of rank for selecting gene g_i for each class i:

```
Step 1: Input the gene expression data OM(i, X_j), Pearson
        correlation coefficient threshold t
Step 2: Select sample S from OM(i, X_j)
Step 3: Select Sample S' from OM(i, X_j)
Step 4: Calculate Pearson correlation coefficient p between S and S' in the
        same class
Step 5: If p≥t, then add sample S' neighbour list of sample S
Step 6: For each gene g_i from S
Step 7: Calculate predicted gene expression of g_i
Step 8: Construct predicted gene expression matrix PM(i, X_j) with
        predicted gene expression
Step 9: For each class of i for each gene gi for each S_j in class i
Step 10: Compute matrix difference d of gene g_i for each class i
Step 11: Calculate rank score r_i of gene g_i by using d of each class i
```

RESULTS AND DISCUSSION

This research work is implemented the proposed model Enhanced Cancer-Association based Gene Selection Technique (ECAGS) and studied thoroughly (Fig. 6-11).

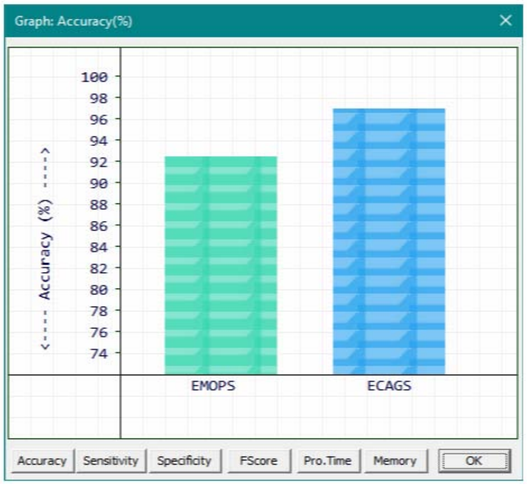

Fig. 6: Accuracy vs. classifiers

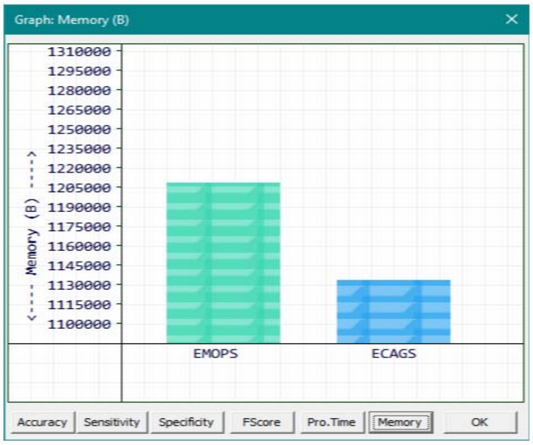

Fig. 7: Memory usage vs. classifiers

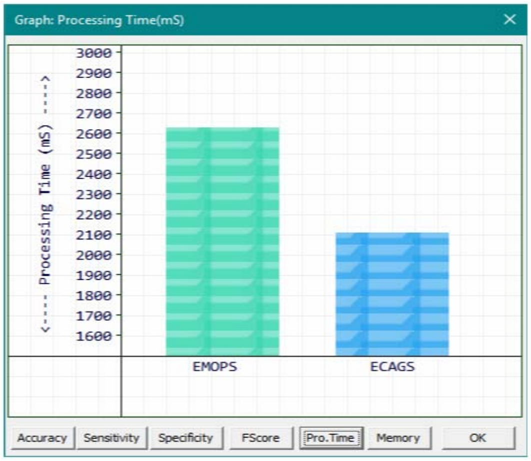

Fig. 8: Processing time vs. classifiers

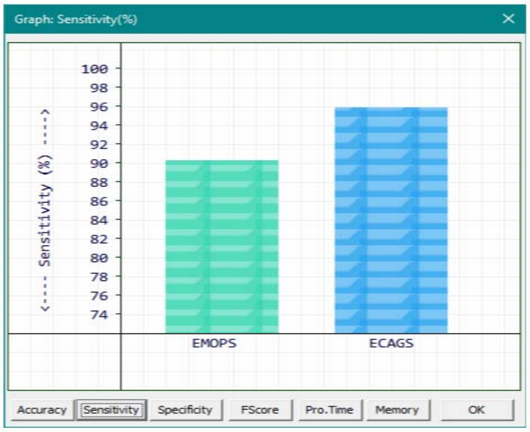

Fig. 9: Sensitivity vs. classifiers

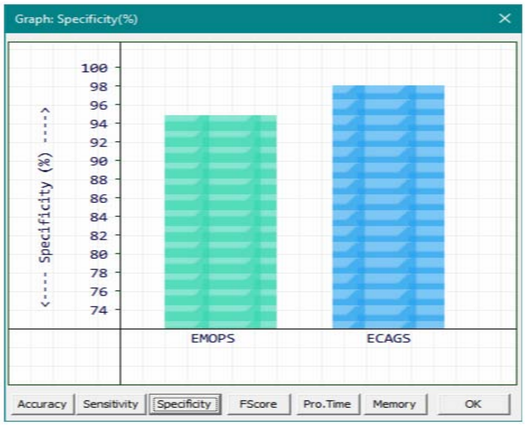

Fig. 10: Specificity vs. classifiers

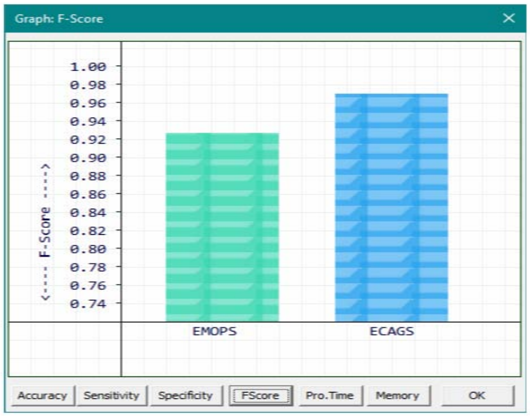

Fig. 11: F score vs. classifiers

The Cancer Genome Sequence datasets namely NCBI.CGS.MER and NCBI.CS.MER are used to analysis the proposed model.

Table 1: Comparison between EMOPS vs ECAGS

Parameters	EMOPS	ECAGS
Memory (B)	1225700	1150700
Processing time (msec)	2640	2121
Accuracy (%)	92.72	97.22
F-score	0.93	0.97
Sensitivity (%)	90.52	96.10
Specificity (%)	95.17	98.39

The performance of the proposed classifier is analyzed in terms of execution time (processing time), classification accuracy, sensitivity, specificity, F score and memory utilization. This research is developed an interfacing tool with the VC++ programming language to extract and validate the gene expressions which are downloaded from NCBI. The validated data is fed into BioWeka simulation tool for analyzing the performances of the proposed classifier in terms of execution time (processing time), classification accuracy, sensitivity, specificity, F score and memory utilization.

The experimental results were shown in Fig. 6-11 and consolidated report was given in Table 1. It was compared with our previous classifier namely Enhanced Multi-Objective Pswarm EMOPS. From the experimental results, it was noticed that the proposed model outperforms our existing classifier in terms of execution time (processing time), classification accuracy, sensitivity, specificity, F score and memory utilization.

CONCLUSION

This research work proposed an efficient cancer pattern classifier called an Enhanced Cancer-Association based Gene Selection technique for Cancer Patterns Classification and Prediction (ECAGS) and studied thoroughly. The proposed classifier is implemented and studied thoroughly in terms of memory utilization, execution time (processing time), classification accuracy, sensitivity, specificity and F score. The experimental results were compared with our previous model and from results, it was established that the proposed model outperforms our previous model in terms of memory utilization, execution time (processing time), classification accuracy, sensitivity, specificity and F score.

REFERENCES

Behravan, I., O. Dehghantanha and S.H. Zahiri, 2016. An optimal SVM with feature selection using multi-objective PSO. Proceedings of the 1st IEEE Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), March 9-11, 2016, IEEE, Bam, Iran, ISBN:978-1-4673-8737-8, pp: 76-81.

Chakraborty, D. and U. Maulik, 2014. Identifying cancer biomarkers from microarray data using feature selection and semi supervised learning. IEEE. J. Transl. Eng. Health Med., 2: 1-11.

Chen, B., W. Zeng, Y. Lin and D. Zhang, 2014. A new local search-based multiobjective optimization algorithm. IEEE. Trans. Evol. Comput., 19: 50-73.

Coello, C.A.C. and M.S. Lechuga, 2002. MOPSO: A proposal for multiple objective particle swarm optimization. Proceedings of the 2002 Congress on Evolutionary Computation part of the 2002 IEEE World Congress on Computational Intelligence, May, 12-17, 2002, Hawaii, pp: 1051-1056.

Gu, X., 2016. A multi-state optimization framework for parameter estimation in biological systems. IEEE. ACM. Trans. Comput. Biol. Bioinf., 13: 472-482.

Kim, H., S.M. Choi and S. Park, 2018. GSEH: A novel approach to select prostate cancer-associated genes using gene expression heterogeneity. IEEE. ACM. Trans. Comput. Biol. Bioinf., 15: 129-146.

Kumar, P.G., C. Rani, D. Devaraj and T.A.A. Victoire, 2014. Hybrid ant bee algorithm for fuzzy expert system based sample classification. IEEE. ACM. Trans. Comput. Biol. Bioinf., 11: 347-360.

Leskovec, J., A. Rajaraman and J.D. Ullman, 2014. Mining of Massive Datasets. 2nd Edn., Cambridge University Press, Cambridge, UK., ISBN: 978-1-107-07723-2, Pages: 467.

Li, B., J. Li, K. Tang and X. Yao, 2015. Many-objective evolutionary algorithms: A survey. ACM. Comput. Surv., 48: 1-35.

Ma, X., F. Liu, Y. Qi, X. Wang and L. Li *et al*., 2015. A multiobjective evolutionary algorithm based on decision variable analyses for multiobjective optimization problems with large-scale variables. IEEE. Trans. Evol. Comput., 20: 275-298.

Mukhopadhyay, A. and M. Mandal, 2014. Identifying non-redundant gene markers from microarray data: A multiobjective variable length PSO-based approach. IEEE. ACM. Trans. Comput. Biol. Bioinf., 11: 1170-1183.

Nikam, S.S., 2015. A comparative study of classification techniques in data mining algorithms. Orient. J. Comput. Sci. Technol., 8: 13-19.

Pardo, A., E. Real, V. Krishnaswamy, J.M. Lopez-Higuera and B.W. Pogue *et al*., 2016. Directional kernel density estimation for classification of breast tissue spectra. IEEE. Trans. Med. Imaging, 36: 64-73.

Qiu, F.Y., L.P. Mo, B. Jiang and L.P. Wang, 2016. Multi-objective particle swarm optimization algorithm using large scale variable decomposition. Chinese J. Comput., 39: 2598-2613.

Subasree, S., N.P. Gopalan and N.K. Sakthivel, 2016. A comparative study and analysis of data mining classifiers for microarray based cancer pattern diagnostics. Proceedings of the International Conference on Informatics and Analytics (ICIA-16), August 25-26, 2016, ACM, Pondicherry, India, ISBN:978-1-4503-4756-3, pp: 1-5.

Subasree, S., N.P. Gopalan and N.K. Sakthivel, 2018. EMOPS: An enhanced multi-objective pswarm based classifier for poorly understood cancer patterns. Intl. J. Eng. Technol., 7: 7-11.

Trivedi, A., D. Srinivasan, K. Sanyal and A. Ghosh, 2016. A survey of multiobjective evolutionary algorithms based on decomposition. IEEE. Trans. Evol. Comput., 21: 440-462.

Yoon, Y., S. Bien and S. Park, 2010. Microarray data classifier consisting of K-top-scoring rank-comparison decision rules with a variable number of genes. IEEE. Trans. Syst. Man Cybern. Part C. Appl. Rev., 40: 216-226.

Zhang, Y., S. Wang and G. Ji, 2015. A comprehensive survey on particle swarm optimization algorithm and its applications. Math. Prob. Eng., 2015: 1-38.