EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

A Gene Expression Data Classification and Selection Method using Hybrid Meta-heuristic technique

Rachhpal Singh

PG Department of Computer Science and Applications, Khalsa College, Amritsar

Abstract

The gene expression data selection is an ill-posed problem. The features selection techniques are found to be an efficient way to evaluate the dimensions of huge gene expression data. This feature selection techniques guide the relevant gene selection. In this paper, a hybrid method (MPG) is proposed to get selection of gene expression by using Mutual information way with Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). A simulation environment is developed, which reveals the decrease in gene expression data dimensions and also removes the duplication among the classified gene data sets significantly. The proposed approach suitable for gene data set analysis using different classifier techniques and show the higher efficiency and accuracy of proposed data sets as compared to traditional selection mechanisms.

Keywords: Gene Expression, Genetic Algorithm, Particle Swarm Optimization, Feature Selection Classification.

Received on 20 May 2019, accepted on 13 August 2019, published on 19 August 2019

doi: 10.4108//eai.13-7-2018.159917

1. Introduction

In all the tissue samples the gene expression of all the genes is measured by the DNA microarray and these microarray techniques have many mechanisms, adopted for classification of gene expressions. It also recognizes normal tissues and cancer tissues [1]. This approach has a big datasets for analysis of large samples of gene expressions. This has computing values having genes in thousands cell mixture having range values from 1000 to 30000. Only small samples from 100 to 400 be selected for gene classification [2] [3] and analysis to get the required results. Different parameters were taken using Support Vector Machine (SVM). This will handle statistical oriented calculations in easy way with the help of data mining techniques [4]. This follows the classification rule that was applied to avoid data over fitting from the datasets. This will only applied if sample size is smaller than the number of features and variations of genes. The purpose of this classification method is to set a decision function that deals with the collected data for better classification. Note that this technique fails if either test data was in complete or not available. Also note that execution time increases if data having large number of features and selection criteria. So it is very difficult to execute. To avoid such hindrances in microarray data analysis, a subset of genes having relevant classification is selected to decrease the data dimensionality. Feature selection is best approach for selection of relevant subsets using data mining technique to reduce high-dimensionality of datasets with irrelevant features. Doing so will degrade the clustering performance [5]. Data mining is the artificial intelligent technique with data base management system in the medical field for the evaluation of gene data sets [6]. Feature selection is an optimized technique using number of benchmark datasets. This will increase the performance of data classification or gene classification by reducing selected features [7] [8].

*Corresponding author. Email: rachhpal_kca@yahoo.co.in

Identification of tumours is a complex approach for data clustering having a huge set of gene expressions [9]. Data having different genes expressions is famous for handling big data. It will also decrease the duplication of data and reduction of high-dimensionality data from the assigned data sets [10]. If search space range is high, then a small set of genes be selected for any type of tumour identification.
Feature selection is best classification method from all the classification and selection criteria for removing the duplication of gene data expressions and to control the dimensionality of the data sets. Accuracy and stability of data sets classification in learning method increases the speed of selection and classification process by using feature selection mechanism in gene expressions and gene classifications. This classification approach depends upon the maximization of Mutual Information (MI) with hybrid method using Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). This proposed is named as MGP algorithm or Mutual-information with Genetic and Particle swarm optimization.

When samples of two data sets from gene sets was correlated randomly using maximum Mutual Information (this is dependency level among given data sets and sample size), then it gives better results as compared to traditional or conventional approach. Correlation among all the genes of same data sets or the sample data sets was computed and well interpreted showing the association or relationship between them. Maximization of Mutual Information of genes from the given data sets or sample data sets describe a high informative statement. This is the best way for illustration of all the genes sets as compared to the genes sets having no mutual information mechanism [11]. In past, number of techniques were successfully executed using GA, PSO and SVM for the analysis of microarray data [12]. Also query optimization is an effective technique using GA with PSO to analyse the performance of all evolutionary techniques for gene selection and classification [13].

In 1995, K-E (Kennedy and Eberhart) developed a famous population-based stochastic optimization mechanism called PSO. It is a simulation process deals with social behaviour of organisms (like flocking of birds and schooling of fishes). Here, a big search space was done where every single candidate's output considered as a particle. This particle uses its own memory and knowledge obtained by the swarm is to search the best solution. Fitness value evaluated by the fitness function by all the particles and be optimized. So each particle sets its position or by changing its velocity, either according to own experience or neighbouring particle's experiences and that is the best use of best location or position. Particles change their position and take a movement using problem space by finding and follow the current optimum particle positions. Such mechanism works for a fixed no. of times. Its stops its execution till minimum error occurs. This procedure works iteratively. Particle adopts the data characteristics of gene expression having high dimensions and so their classification results are superior as compared to other mechanisms in evolutionary system. Every particle sets its venue or position according to pbest (personal best) and gbest (global best) as fitness values in a search space for gene selection and classification. Also, during searching always avoid from trapping in a local optimum. This can be done by adjusting the fine-tuning and inertia weight. If gbest trapped itself in local optimum, then a search space for every particle be created in same space to avoid trapping and at end gets superior results of classification. This will reduce the number of selected genes. As shown in Figure 1(a), all the particles have converging property near gbest after a defined period and in case gbest not change its value after four iterations, then it is considered that it stuck into local optimum. In these situations, gebest is current fitness value. By resetting its value to zero, this becomes best for gene feature selection and accurate classification as shown in Figure 1(b). Here, local optimum be skipped and search continue for superior gene classification as output. The individual particle be converged towards gbest and reset the value as shown in Figure 1(c). Further to find a new gbest with lesser number of genes is as shown in Figure 1(d). This operation achieves superior gene classification by reducing no. of genes according to the need of selection of genes. Change of particle velocity interpreted as a probability change in searching particle state or location or movement.

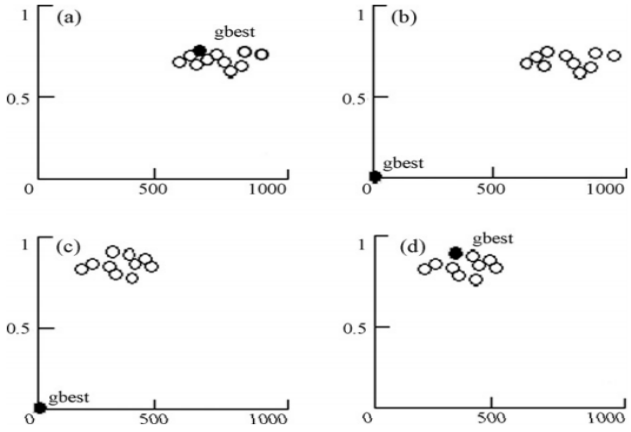

Figure 1(a) Trapping of gbest in local, **(b)** Resetting of gbest to zero, **(c)** Resting gbest and particle movement, **(d)** Convergence of particles towards updated gbest and improvement in individual position.

Furter to improve the gene celection and classification GA a popular heuristic technique is used. It is efficient and useful in case of problem is complex in large size with number of hindrances. GAs are used to optimize queries quickly [14]. It support natural selection process to select best gene selection and abandon worst genes to

get best solution [15]. GA initializes genes as population generation for solving search space problem with the help of genetic operators. Further generated gene population follows iterative procedure as evolution of new gene sets for classification. Selection, Crossover and Mutation are basic operations used to get more genes generations during iteration process. An objective function as fitness function computed during every iteration. If currently generated gene population has better fitness than existing gene population, then old fitness is replaced by new and continues till either n generation occurs or met the stopping criteria [16]. GA operators significantly obtained the optimal solution. To get optimized selection and classification solutions, GA has following components:

Population: Set of gene chromosome considered as population and generated with random variables by using GA's selection operator.

Fitness function: Also called objective function for gene selection and classification to find exact optimal solution.

Selection: It is primarily applied on gene population having ranking metric and is similar to select a hockey team from different countries sports person's as gene population.

Crossover: Swapping of gene population after selection process either selecting one point crossover or two point crossover is the crossover operation of genes for best selection and classification of genes.

Mutation: Rarely used operation to get the best solution. It preserve genetic diversity from one generation of a population of chromosomes to next one i.e. invert the randomly selected bits based on the mutation probability [17].

As GA has two basic operations crossover and mutation having two probabilities. Pc is crossover probability and Pm is mutation probability used for gene selection. If these probabilities are not accurate assumption, then it creates a premature convergence or non-convergent approach during the gene expression search and classification. The proposed mechanism MPG improves the conventional GA with PSO by adjusting the Pc and Pm values to search best global optimized solution. Further Maximum Mutual Information (MI) process combines with two hybrid PSO and GA algorithms to launch a new strategy for solving such complex problem. Here a novel hybrid algorithm proposed having gene feature selection mechanism combining MI with PSO and GA to remove the duplication of gene data samples and decrease the gene expression data dimensionality using SVM. The proposed technique shows the better results by classification accuracy rates comparisons with some of existing gene feature selection approaches. Different classifier methods are applied to selected gene datasets to test robustness of proposed algorithm. All classifiers show classification accuracy rates higher than 80% during simulation.

In section 2, literature review done. In section 3, general scheme or methodology of hybrid approach is discussed. In section 4, simulation results as experiment are discussed and at last in section 5, it was concluded.

2. Related Work

Diaz et. al. studied the random forest classification algorithm for microarray data and showed excellent performance by considering the noise like most predictive variables [18]. Guyon et. al. addressed selection problem of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays and proposed a new gene selection method using Support Vector Machine (SVM) technique [19]. Jäger et. al. derived a mechanism for improving and evaluating the selection procedure of informative genes from microarray data and proposed three pre-filter different approaches based on correlation and clustering to get similar genes groups [20]. Li. et. al. tested on datasets colon and leukemia and examined the reproducibility, sensitivity and gene classification and selection stability by combining GA and the *k*-Nearest Neighbor techniques for identification of genes to discriminate various samples of classes [21]. Yeung et. al. presented a Bayesian model averaging (BMA) approach for classification and selection of genes of microarray data and selected smaller numbers of relevant genes as compared to other approaches and achieved a high accuracy prediction values on microarray datasets [22]. Huerta et. al. proposed combined technique for SVM with GA for high dimensional classification of microarray data based on fuzzy logic having pre-filtering technique and introduced most informative genes identification method [23]. Wang et. al. showed the performance of classification microarray data sets and demonstrated combined approach of different feature selection and classification approaches by selecting relevant genes [24]. Zhu et. al. proposed a new gene selection algorithm known as Markov blanket-embedded genetic algorithm (MBEGA) that insert or remove genes features with the help of GA and improved the output with fine-tuning search technique [25]. Liu et. al. used a parallel GA for filtration and classification of informative genes data sets having various tissues of numerous classes [26]. Peng et. al. discussed GA and SVM with all the pairing for identification of multiclass cancer and eliminated all the post-processing steps by removing the redundancy handling cancer-related predictive gene data sets [27].

Alba et. al. compared PSO and GA with SVM to classify high dimensional microarray data and found low samples of informative genes from huge amount of data sets using a binary representation in Hamming space [28]. Li et. al. used a hybrid approach combining PSO with GA and SVM for selection and classification of genes from a set of large amount of data sets by testing by considering three benchmark gene expression datasets that reduced datasets dimensionality and confirmed about

informative gene subset with improved classification accuracy [29]. Sahu et. al. proposed an optimized and filtering technique having different feature selection mechanism to classify microarray data for high dimensional cancer sets by combining PSO with signal-to-noise ratio and probabilistic neural network for clustering of data sets with the help of *k*-means clustering approach to rank every gene of each cluster [30]. Ghamisi et. al. proposed a new feature selection method based on combined GA with PSO on validation samples for best fitness value using SVM classifier [31]. Shen et. al. presented a paper for tumour classification by combining PSO with SVM to select genes and evaluate or classify by taking microarray data and minimize the high dimension data [32]. Shen et. al. developed a PSO and tabu search hybrid algorithm known as HPSOTS for selection of genes and classification of tumour on three microarray different data sets by minimizing the high dimension data sets [33]. Babaoglu et. al. studied the binary PSO and GA to search the efficiency of feature selection on determination of coronary artery disease existence based upon exercise stress testing data [34]. Chen et. al. developed a novel approach having PSO utilization combining with a decision tree technique work as a classifier and compared the performance of well-known benchmark classification methods with the proposed technique based on statistical analysis for all the test datasets which are compatible with SVM [35]. Arora et. al. proposed real-world complex problems solving technique for feature selection [36]. Sharma et. al. proposed two nature-inspired computing mechanism query optimization in the medical field [37]. Kaur et. al. presented role of big data in healthcare applications technique to handle medical industry related problems [38].

3. Hybrid Mutual Information Meta-heuristic Feature Selection Algorithm (MPG)

A. Maximization of Mutual Information

Mutual information is maximized with the help of random variables [41]. It means one random sample (p) has the dependent information on the second random sample (q). So overall maximization of mutual information having a defined gene expression dataset is defined in the equation (1) below:

$$MI(p,q) = \sum_{P\epsilon R} \sum_{Q\epsilon S} a\,(p,q) \log_2 \frac{a(p,q)}{a(p)a(q)} \qquad (1)$$

where a(p) denotes probability density of variable p, a(q) represent probability density of variable q and a(p, q) shows joint probability density.

MI denotes the maximum mutual information of x in y and is illustrated in equation (2) as below. Here consider M genes in a data set quantity and D shows gene expression with outline x data range having a dataset of genes in y class. E represents gene expression with outline x data range having a dataset of genes not in y class. Similarly, F shows gene expression without outline x data range having a dataset of genes in y class:

$$MI(x,y) = log\frac{a(x/y)}{a(x)} = log\frac{a(x,y)}{a(x)\times a(y)} = log\frac{DXM}{(D+F)\times(D\times E)} \qquad (2)$$

As shown in equation (2), irrelevance of gene expression outline x to class y shows a result zero as MI(x, y)=0. Further the final formula of maximum mutual information is shown in the equation (3).

$$MI(x) = \sum_{l=1}^{m} a\left(\frac{X_l}{y}\right) log\frac{a\left(\frac{X_l}{y}\right)}{a(X_l)} \qquad (3)$$

Here m is the class quantity number in a dataset and the main objective of this operation is to get genes with strong dependency as compared to other groups of genes in the same class. In particular times computing the MI will describe genetic filtering.

B. Particle Swarm Optimization

PSO proposed by Kennedy and Eberhart in 1995 [13] at first for optimization. In PSO, swarm are executed for counted number of particles and it is similar to individual population as Evolutionary Algorithms (EA). At every iteration or loop, all the particles take a movement in problem space to search global optima. Every particle has a velocity vector and a current position vector for giving a direction to movement of particles.

$$vel^{K+1} = w.vel\frac{K}{i} + \emptyset_1 rand_1.\left(pbest_i - x\frac{K}{i}\right) + \emptyset_2 rand_1.\left(gbest_i - x\frac{K}{i}\right) \qquad (4)$$

$$x_i^{K+1} = x\frac{K}{i} + vel_i^{K+1} \qquad (5)$$

Equations (4) represent velocity computation and equation (5) shows updated position of given particle i at a some defined iteration k. Equation (4) compute a new velocity vel_i for every particle which is also known as potential output based on previous velocity value and location of the particle to find best fitness as $pbest_i$. Further searching continues in global space to find the global best fitness. By computing the global population or from a set of local neighbourhood so that local neighbourhood gives a direction for global best which is also called $gbest_i$. Social weight or individual are denoted by ϕ_1 and ϕ_2 respectively. Finally, $rand_1$ and $rand_2$ are two random number variables having a range from 0 to 1 and w denotes inertia weight factor. Equation (5) is used for updating of every particle's location x_i in the defined solution space.

C. Genetic Algorithm (GA)

GA has two critical operations as crossover and mutation. New individuals are generated by the crossover operation globally. Similarly new individuals are generated by

mutation operation locally. These operations are two methodology of GA for global and local search [40]. Two probabilities Pc for crossover, Pm for mutation are applied to find the convergence of GA for searching the optimal output. Pc and Pm are two pre-defined variables in the standard GA that have fixed value GA search space and helpful in gene selections and classifications. If Pc has the too large value, then the global search is also become too complex or coarse and so optimal output can be missed. Further if Pc has too small value, then it stuck or put into local minima. Similarly if Pm value is too large, then genetic procedure works same as a random search. Also if Pm has smallest value, then will supress the exploratory capability of search. During the searching operation, particle search with a defined velocity and position factor is a more effective to permit GA to adjust the values of Pc and Pm that resembles with mutual information. This whole procedure generates a MPG algorithm for gene selection and classification. In MPG, the Pc and Pm values are computed as described in equations (6) and (7) below:

$$Pc = \begin{cases} K_1 \dfrac{(FN_{max} - FN^1)}{(FN_{max} - FN_{avg}}, FN^1 \geq FN_{avg} \\ K_2, \qquad\qquad FN^1 < FN_{avg} \end{cases} \quad (6)$$

$$Pm = \begin{cases} K_3 \dfrac{(FN_{max} - FN)}{(FN_{max} - FN_{avg}}, FN \geq FN_{avg} \\ K_4, \qquad\qquad FN < FN_{avg} \end{cases} \quad (7)$$

In equation (6), FN_{max} shows maximum value of all individual's fitness during MPG search operation, FN_{avg} denotes average fitness value, FN^1 tells us the high fitness value [38] and K_1, K_2, K_3 and K_4 denotes four variables having range from 0 to 1.

D. Maximization of Mutual Information with PSO and GA (MPG)

Combining maximum mutual information with particle swarm optimization and genetic algorithm, a gene selection mechanism was proposed known as MPG selection and classification algorithm. SVM work as a classifier in selected operation for best selection. So SVM in MPG has the best fitness value that is more efficient and useful for classification accuracy. In equations (6) and (7), set the values for k1 = 0.8, k2 = 0.9, k3 = 0.2 and k4 = 0.002 and also have maximum number of iteration to 800 that is used for selection of best fitness. Let take gene expression dataset X has x1 and x2 gene samples. MPG procedure for selection and classification has the following steps:

(1) Compute mutual information for all the genes in dataset X by implementing MI with a given number of times. It produces a subset Y of X from the whole selected MI genes. Number of genes in Y subset is 200.

(2) Compute the particle velocity and position during the MPG iterations by using equation (4) and (5) with the updated value of velocity and position in a defined number of iterations or upto a maximum value.

(3) Initialize population for MPG and compute fitness value for every individual. Note that population size is considered according to the space of the problem. It means if size is larger, then it is very easy to find the fitness value from MPG for best optimal results. It will elapse for longer time and for the large and complex iterations. Here 300 is taken the population size and is denoted as M. Every individual has number of genes from the subset Y and take sample size as a1 for every gene.

(4) Apply coding process to encode individuals in a population and after coding, every individual value corresponds to a chromosome having some defined length.

(5) Compute all the fitness values for FN_{max}, FN_{avg} and F^1.

(6) Individuals having high fitness value can be selected by setting a threshold value.

(7) Compute the Pc and Pm for best fitness value after finding particle velocity and position with updating operation. It will generate a new population.

(8) Finally a test was done to find current optimal fitness function or value that satisfy the target or meet the termination criteria. If it meets then, move to the step (9) else move to step (4).

(9) Finally get the optimal subset values of genes for selection and classification according to the decoding rules.

4. Experimental Results and Comparisons

Four type gene expression datasets of Appendix cancer, Cervical Cancer, Gallblader cancer and Kidney cancer are taken for experimental purpose. Sample size, quantity of genes taken and distribution of every class of data sets taken for simulation is as shown in Table 1.

Table 1. Gene expression datasets

Datasets	Sample size	Number of genes	Class distribution	Sample distribution
Appendix	230	8700	Negative Positive	130 100
Cervical	130	3000	Negative Positive	80 50
Gallbladder	150	12400	Negative Positive	90 60
Kidney	80	2200	Negative Positive	50 30

Parameters Settings:

Proposed approach of maximization of mutual information mechanism with hybrid meta-heuristic

particle swarm optimization and genetic algorithm using SVM is implemented in Matlab software having version 6.1 for MS-Windows 10 on i5 7th generation processor with 8GB RAM. Gavin Cawley's toolbox of SVM software was taken as test data for classification of data gene sets [42]. Some of the parameters are used for simulation purposes. The parameters considered for the proposed technique for the selection and classification of genes is as shown in Table 2.

Table 2. Parameters for proposed technique for selection of genes data sets

Parameter	Appendix	Cervical	Gallbladder	Kidney
Population size	520	550	600	500
Chromosome length	1220	1200	1250	800
No. of generations	2400	2200	2250	2500
Crossover rate	0.99	0.92	0.98	0.95
Mutation rate	0.02	0.02	0.01	0.01

Various parameters taken for proposed technique using GA for the classification of genes data sets in the simulation process is as shown in Table 3.

Table 3. Parameters for proposed technique for classification of genes data sets

Parameter	Appendix	Cervical	Gallbladder	Kidney
Population size	62	68	60	50
Chromosome length	120	100	80	50
No. of generations	700	800	600	500
Crossover rate	0.92	0.93	0.94	0.95
Mutation rate	0.02	0.02	0.01	0.01

After implementation of these parameters in simulation, it was observed that accuracy of genes classification not incremented monotonically with the number of gene incrimination. So the datasets having small numbers samples relatively have the better results. Data sets of gene expression for the selection of genes have minimum number of genes was simulated using a mapping procedure. This mapping procedure of genes for the labelling and classification of genes gave better solution in optimized form. Also note that by increasing number of genes, rate of classification may be decreased or increased due to complex relationship among genes and genes data sets. So accuracy of the classification depends upon the relationship agreement and identifies the co-relationship between classifier and selection algorithm. Nutshell all the classifiers and selection mechanism has high classification accuracy rates than all the datasets taken. This creates robustness in the proposed MPG algorithm. Robustness can be decreased or removed during simulation process either by increasing the number of iterations or decreasing the sample size. By classification and selection criteria information regarding cancer, identification of cancer and clustering of genes was improved. It also increase the efficiency of all the applications related with medical field for genes selection and classification..

MPG algorithm was executed n times (where n= 50, 36 , 40 and 20) with the help of SVM on every data set of the Appendix, Cervical, Gallbladder and Kidney cancer. Rate of average classification computed from the given gene data subset, a popular LOOCV method [39] was implemented. Output in a summarized form for the above considered datasets using simulation process using various iterations is shown in Table 4. It can also be used for comparison purpose for different data sets considered in the simulation.

Table 4. Summarized output of proposed technique using n iterations (simulator MPG with SVM using four state of the art procedures)

Dataset	MPG	[50]	[36]	[40]	[20]
Appendix	100(50)	100(12)	100(18)	94(10)	98(2)
Cervical	95(30)-	92(10)	90(8)-	99(12)	99(20)
Gallbladder	99(40)	90(15)-	96(16)	98(20)	96(10)
Kidney	100(50)	100(10)	98(20)	99(18)	99(12)-

Results were compared using conventional criteria to find the accuracy of the genes selections and classifications. Comparisons show the correct rate with the size of the gene sets (number of genes used) in above table for optimized solution of datasets. In the table 4, data having the symbol "-" means some of the genes were not participating or non-available. Table 5 describe the detailed output of the experiment done with the help of simulator and it was observed that Appendix dataset has 100% classification rate by using maximum 50 genes data sets and has better performance than the previous mechanism used. Here Ex is the executions like Ex1, Ex2, Ex3 and Ex4.

Table 5. MPG with SVM performances on 4 executions

Ex	Ex1	Ex2	Ex3	Ex4	Average
Appendix	100(50)	100(36)	100(40)	100(20)	100
Cervical	99(20)	98(12)	96(30)	98(10)	97.75
Gallbladder	92(10)	94(20)	96(30)	98(30)	95
Kidney	99(40)	96(20)	92(10)	98(30)	96.25

Interesting output obtained with the help of simulator for Appendix, Cervical, Gallbladder and Kidney cancer datasets. Proposed method has highest or averaged accurate classification rate analysis. Experimental results show the significant change in biological genes and much better than previous approaches. Table 5 represent detailed output of 4 independent execution by simulator MPG algorithm by using the SVM and found that performance and output is quite stable.

5. Conclusion

Proposed hybrid meta-heuristic based feature selection algorithm combining the maximizing mutual information algorithm with PSO and GA known as MPG algorithm. This selection algorithm shows effectively the dimension of genes expression data sets in their original format and reduces redundancy problem in data set. MPG selection mechanism decreases the number of genes with high classification accuracies. Also simulation was done for all classification accuracy datasets with existing feature selection mechanisms. It demonstrates the effectiveness of proposed algorithm. Different classifiers are applied to decrease dataset. Further this algorithm is used for efficiency improvement. Simulator can be developed for cloud environment in future using this MPG technique to improve the time complexity.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J Levine. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *In Proc. Natnl. Acad. Sci. USA*, **96:** 6745-6750.

[2] Sharma, Samriti, Gurvinder Singh, and Dhanpreet Singh, (2019) Role and Performance of Different Traditional Classification and Nature-Inspired Computing Techniques in Major Research Areas, *EAI Transactions on Scalable Information System*, **6(21)**:1-17.

[3] Suguna, S. Kanimozhi, et al., (2019), Application of Nature—Inspired Algorithms in Medical Image Processing, *Advances in Nature-Inspired Computing and Applications. Springer, Cham*, 61-100, ch. 1.

[4] Sharma, Manik, Samriti Sharma, and Gurvinder Singh, (2018), Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining., *Data*, **3(4)** : 54 (1-16).

[5] Kumar, Lalit, and Kusum Kumari Bharti, (2019), An Improved BPSO Algorithm for Feature Selection, *Recent Trends in Communication, Computing, and Electronics. Springer, Singapore*, **524**: 505-513.

[6] Sharma, M., G. Singh, and R. Singh, (2017), Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques, *IRBM* , **38(6)**: 305-324.

[7] Sayed, Gehad Ismail, Aboul Ella Hassanien, and Ahmad Taher Azar, (2019), Feature selection via a novel chaotic crow search algorithm, *Neural Computing and Applications*, **31(1)**: 171-188.

[8] Sharma, Manik, Gurvinder Singh, and Rajinder Singh, (2016), Design and analysis of stochastic DSS query optimizers in a distributed database system, *Egyptian informatics journal,* **17(2)**: 161-173.

[9] Yu. Z. Chen, H., You, J., Han, G., & Li, L., (2013), Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **10(3)**: 657-670.

[10] Sherlock, G., (2000). Analysis of large-scale gene expression data. *Current opinion in immunology*, **12(2):** 201-205

[11] Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. *The Journal of Machine Learning Research*, **3:**1415-1438.

[12] T. S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. (2000), Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics,* **16(10)**:906–914.

[13] Sharma, Manik, Gurvinder Singh, and Rajinder Singh, (2019), A review of different cost-based distributed query optimizers, *Progress in Artificial Intelligence,* **8(1)**: 45-62.

[14] Sharma, Manik, et al., (2013), Design and comparative analysis of DSS queries in distributed environment, *International Computer Science and Engineering Conference (ICSEC). IEEE.*

[15] Mirjalili, Seyedali, et al., (2019), Grey Wolf Optimizer: Theory, Literature Review, and Application in Computational Fluid Dynamics Problems." *Nature-Inspired Optimizers. Springer, Cham*, **811**:87-105.

[16] Chandrashekhar Azad1, Vijay Kumar Jha, (2015), Genetic Algorithm to Solve the Problem of Small Disjunct In the Decision Tree Based Intrusion Detection System, *J. Computer Network and Information Security,* **8**: 56-71

[17] Singh Rachhpal, (2017), Cuckoo Genetic Optimization Algorithm for Efficient Job Scheduling with Load Balance in Grid Computing, *I. J. Computer Network and Information Security, MECS Publication*, **8**:59-66

[18]] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres, (2006), Gene selection and classification of microarray data using random forest." *BMC bioinformatics,* **7(3)**:1-13.

[19] Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, (2002), Gene selection for cancer classification using support vector machines." *Machine learning.***46(1-3)**: 389-422.

[20] Jäger, Jochen, Rimli Sengupta, and Walter L. Ruzzo, (2003), Improved gene selection for classification of microarrays, *In Biocomputing*, **2002**:53-64.

[21] Li, Leping, Clarice R. Weinberg, Thomas A. Darden, and Lee G. Pedersen, (2001), Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics,* **17(12)**: 1131-1142.

[22] Yeung, Ka Yee, Roger E. Bumgarner, and Adrian E. Raftery, (2005), Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data, *Bioinformatics,* **21(10)**: 2394-2402.

[23] Huerta, Edmundo Bonilla, Béatrice Duval, and Jin-Kao Hao, (2006), A hybrid GA/SVM approach for gene selection and classification of microarray data, In *Workshops on Applications of Evolutionary Computation*, *Springer, Berlin, Heidelberg,* **3907**:34-44.

[24] Wang, Yu, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W. Mewes, (2005), Gene selection from microarray data for cancer classification—a machine learning approach, *Computational biology and chemistry,* **29(1)**: 37-46.

[25] Zhu, Zexuan, Yew-Soon Ong, and Manoranjan Dash, (2007), Markov blanket-embedded genetic algorithm for gene selection, *Pattern Recognition* , **40(11)**: 3236-3248.

[26] Liu, Juan, Hitoshi Iba, and Mitsuru Ishizuka, (2001), Selecting informative genes with parallel genetic algorithms in tissue classification, *Genome Informatics,* **12**: 14-23.

[27] Peng, Sihua, Qianghua Xu, Xuefeng Bruce Ling, Xiaoning Peng, Wei Du, and Liangbiao Chen, (2003), Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, *FEBS letters,* **555(2)**: 358-362.

[28] Alba, Enrique, Jose Garcia-Nieto, Laetitia Jourdan, and El-Ghazali Talbi, (2007), Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, *IEEE Congress on Evolutionary Computation*, **12**: 284-290.

[29] Li, Shutao, Xixian Wu, and Mingkui Tan, (2008), Gene selection using hybrid particle swarm optimization and genetic algorithm, *Soft Computing,* **12(11)**: 1039-1048.

[30] Sahu, Barnali, and Debahuti Mishra, (2012), A novel feature selection algorithm using particle swarm optimization for cancer microarray data, *Procedia Engineering,* **38**: 27-31.

[31] Ghamisi, Pedram, and Jon Atli Benediktsson, (2014), Feature selection based on hybridization of genetic algorithm and particle swarm optimization." *IEEE Geoscience and remote sensing letters,* **12(2)**: 309-313.

[32] Shen, Qi, Wei-Min Shi, Wei Kong, and Bao-Xian Ye, (2007), A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification, *Talanta.* **71(4)**: 1679-1683.

[33] Shen, Qi, Wei-Min Shi, and Wei Kong, (2008), Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, *Computational Biology and Chemistry,* **32(1)**: 53-60.

[34] Babaoglu, İsmail, Oğuz Findik, and Erkan Ülker, (2010), A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine, *Expert Systems with Applications,* **37(4)**: 3177-3183.

[35] Chen, Kun-Huang, Kung-Jeng Wang, Min-Lung Tsai, Kung-Min Wang, Angelia Melani Adrian, Wei-Chung Cheng, Tzu-Sen Yang, Nai-Chia Teng, Kuo-Pin Tan, and Ku-Shang Chang, (2014), Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm, *BMC bioinformatics,* **15(49)**:1-10.

[36] Arora, Sankalap, et al., (2019), A New Hybrid Algorithm Based on Grey Wolf Optimization and Crow Search Algorithm for Unconstrained Function Optimization and Feature Selection, *IEEE Access*, 7: 26343-26361.

[37] Sharma, M., G. Singh, and R. Singh, (2018), Clinical decision support system query optimizer using hybrid Firefly and controlled Genetic Algorithm, *Journal of King Saud University-Computer and Information Sciences. **xxx(2018)**:1-12 (In Press).*

[38] Kaur, Prableen, Manik Sharma, and Mamta Mittal, (2018), Big Data and Machine Learning Based Secure Healthcare Framework, *Procedia computer science,* **132**: 1049-1059.

[39] T. Joachims, (2000), Estimating the Generalization Performance of a SVM Efficiently, *Proceedings of the International Conference on Machine Learning (ICML), Morgan Kaufman.* **12***:1-20.*

[40] Jakobović, D., & Golub, M. (1999), Adaptive genetic algorithm, *CIT. Journal of computing and information technology*, **7(3)**: 229-235

[41] Lu, Huijuan, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. (2017), A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing,* **256**: 56-62.

[42] Cawley, Gavin C. and Nicola LC Talbot. (2006), Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, **22(19)**: 2348-2355.