Journal of Universal Computer Science, vol. 24, no. 6 (2018), 682-710
submitted: 30/5/17, accepted: 15/1/18, appeared: 28/6/18 © J.UCS

Cancer Classification by Gene Subset Selection from Microarray Dataset

Asit Kumar Das
(Department of Computer Science and Technology
Indian Institute of Engineering Science and Technology, Shibpur, India
akdas@cs.iiests.ac.in)

Soumen Kumar Pati*
(Department of Computer Science and Engineering
St. Thomas' College of Engineering and Technology, Kolkata, India
soumenkrpati@gmail.com, *corresponding author)

Hsien-Hung Huang*
(Department of Emergency, Jen-Ai Hospital, Taichung, Taiwan
hhhuang@gmail.com, *corresponding author)

Chi-Ken Chen
(Department of Laboratory Medicine, Jen-Ai Hospital, Taichung, Taiwan
cck2203@mail.jah.org.tw)

Abstract: Microarray dataset contains huge number of genes, many of which are irrelevant regarding cancer classification and as a result classification accuracy is reduced. Therefore, the dataset should be pre-processed to filter out these redundant genes. In this paper, initially a Pareto optimality based Multi-objective Genetic Algorithm has been proposed where non-linear cellular automata is employed to overcome the demerits of random initialization to generate initial population in high dimensional space. The fitness functions are defined based on both attribute dependency and boundary region exploration of rough set theory and Log-Likelihood ratio to select the informative genes. The chromosomes are hybridized by applying multi-point crossover; whereas proximity mutation builds on Flip-bit mutation with a little modification to produce fittest offspring. Finally, the gene subset with strong biological significance in cancer treatment is obtained from the Pareto dominant solutions. Performances are investigated on publicly available microarray cancer datasets and compared with the state-of-the-art methods to demonstrate the effectiveness of the proposed method.

Keywords: Multi-objective genetic algorithm, Gene selection, Cellular automata, Rough set theory, Log-Likelihood ratio, Proximity mutation
Categories: G.1.6, M.7, F.1.1, F.4.1, H.0, I.2.4

1 Introduction

Make Gene expression microarray data are typically known to possess large sets of observations, represented by hundreds or even thousands of coordinates with seemingly unknown correlations [Kossenkov and Ochs, 10]. This high dimensionality has presented many challenges in analysing the data, especially when correlations among the observations are complex. In DNA microarray data analysis [Kossenkov

and Ochs, 10], biologists generally measure the expression levels of genes (typicallyin the range of 2000–30,000) in the tissue samples (typically in the range of 5–150) from patients, and try to deduce how the genes of patients are related to the type(s) of cancer they had. An investigative system [Mansouri and Khademi, 15]considered using the large set of genes, will have higher computational cost, slower learning process and poor classification accuracy due to the occurrence of high dimensionality. So, from a large number of genes, selection of the most relevant, informative, discriminative, and compact subset is the goal of a gene selection process for accurate diagnosis [Pati and Das, 17], [Pati et al., 13], [Das and Pati, 12]. Inherently, gene selection is a combinatorial optimization problem [Lan and Vucetic, 11],[Salem et al., 17] which searches an optimal gene subset from a pool of 2^N competing candidate subset in a dataset of N genes.

Evolutionary Algorithms (EAs) [Xiao et al., 15], [Farahat et al., 15] are applied for the iterative refinement of a group of candidate solutions to an optimization problem. A standard genetic algorithm (GA) deals with single fitness function but most of our real life problems are inherently multi-objective in nature where simultaneous satisfaction of more than one conflicting objectives are required. Simplicity is the major merit of MOEA as an optimization method, where no fitness modification is required. The purpose of this algorithm is to approximate a set of Pareto optimal solutions [Sikdaret al., 15], [Lazar et al., 12] instead of a single one because the objectives are often conflict with each other and improvement of one objective may lead to deterioration of another. Thus, a single solution, which can optimize all objectives simultaneously, does not exist. Therefore the Pareto optimal solutions are important to a decision maker instead of the best trade-off solutions. However, treating constraint violation as an extra objective increases the computational complexity of the algorithm, and thereby may slow down the algorithm [Das, 01]. Many MOEA methods, like NSGA [Saeys et al., 07], NSGA-II [Mitra et al., 02], MOEA/D [Song et al., 07] etc. efficiently handle the feature selection problem in high dimensional space [Mansouri and Khademi, 15], [Kossenkov and Ochs, 10]. But the number of selected features is not so small by these methods.

The work presented in [Zhao and Liu, 07]uses rough set theory for cancer classification using single biomarker gene and obtained fairly acceptable results.The work described in [Garey and Johnson, 79]is a rough set based soft computing method, where single or double genes are obtained for cancer classification. In[Pawlak, 98], an improved GA based gene selection and SVM classification is done which gives better accuracy but at the expense of more than 15 genes per trial on an average. In [Zhong et al., 01], a multi-objective GA based gene selection method has been proposed which gives satisfactory classification accuracy but more expenses in terms of number of selected genes. In [Jing, 14], the stepwise Fisher's linear discriminant function used for selecting an optimal (or near optimal) subset of genes with satisfactory results. A novel hybrid approach [Devroye et al., 96] that combines gene ranking and clustering analysis selects biomarker genes with fairly acceptable results. In paper [Pal and Mitra, 99], a combinational feature selection method in conjunction with ensemble neural networks is explored to improve the accuracy but the method selects at least 30 genes. A fuzzy rule-based gene selection method is proposed in [Gupta and Kapoor, 94] which achieved better accuracy with more number of genes. In [Gu et al., 15], a multiple-filter-multiple-wrapper (MFMW)

approach is proposed that makes use of multiple filters and multiple wrappers to improve the accuracy to identify potential biomarker genes. An ensemble machine learning based gene selection method is proposed in [Price et al., 05] which are very expensive in terms of both the number of genes and accuracy. In [Souam et al., 13], a signal to noise ratio based method is presented for prostate cancer dataset to classify cancer and non-cancer samples. A Fuzzy-Rough-Neural based f-Information (FRNf-I) method is proposed in [Vatolkin et al., 12] that computes *f*-information measure easily and selects less number of genes with more classification accuracy. The work in [Shelokar et al., 13] proposed an algorithm that combines a simulated annealing schedule specially designed for gene subset selection with the incrementally computed joint entropy to select less number of genes with high classification accuracy.The paper is organized into four sections. Section 2 review related literatures. Section 3 describes the gene subset selection method on Pareto optimality based multi-objective genetic algorithm. Section 4 Experimental results pertaining to the performance evaluation of the proposed method compare to the existing state-of-the-art algorithm is presented. Finally, the paper is concluded in Section 5.

2 Literature Review

In this study, we have proposed a Pareto optimal based multi-objective genetic algorithm (PMOGA) to find some non-dominated solutions where each of the solution contains distinct number of informative genes without sacrificing any knowledge or information in the microarray dataset. Initial population generation is a basic and crucial task in evolutionary algorithms [Xiao et al., 15], [Farahat et al., 15]. The random initialization [Maaranen, 04]is the most frequently used method to generate initial population in case of non-availability of information about the solution, but it takes long computational time, especially when the solution space is difficult to explore. The generation of quasi-random sequences is more difficult [Maaranen, 04] and loses its importance in case of a higher dimensional dataset. To overcome such demerits, the proposed method uses non-linear uniform hybrid Cellular Automata (CA) [Neumann, 96], which is well appreciated for its capability as an excellent random pattern generator for generating initial population of binary strings. In paper [Mitchell et al., 00], a recent review work was completed where the GA was used to evolve cellular automata for two computational tasks, like density classification and synchronization. In paper [Back and Breukelaar, 05], GA is used to evolve behaviour in cellular automata. In [Seredynski and Skaruz, 12], a large space of automata rules is explored efficiently by a GA, which locates through a quality rule. Cellular automata are excellently used as random pattern generator in many fields (like, random number generatorinMathematics,stream cipher in cryptography etc.). This concept is used here for the generation of initial population. The combination of genetic algorithm and cellular automata for initial population generation is an innovative concept for important gene subset selection from the microarray datasets. Generally, normal GA deals with single objective function but most of our real life problems are multi-objective in nature where simultaneously more than one conflicting objective functions are required to be satisfied. However, the conventional MOEAs are faced with unsuccessful convergence in the Pareto Front and increase the computational cost of the system with an increasing number of

objectives [Wang et al., 14]. The goal of the multi-objective optimization is to approximate the Pareto Front in the objective space so that no further improvement on any objective can be achieved without harming the rest of objectives. Thus researchers have designed a number of algorithms to overcome the obstacles. So the proposed PMOGA is used two conflicting objective functions which can easily find the solutions in the Pareto Front and reduce the computational cost of the experiment. Two objective functions for PMOGA are defined using (i) the attribute dependency (obtained using positive region) and explores the boundary region of RST [Pawlak, 98], [Jing, 14] and (ii) Log-Likelihood ratio measurement method [Jerzy and Pearson, 33]to select more precise and informative genes in the microarray dataset. To create new individuals, many genetic algorithm based papers [Jing, 14], [Odibat and Reddy, 14]perform single point crossover but the proposed method uses multi-point crossover with the motivation that, the new individuals generated are more similar to one of their high quality parents than they are in single-point crossover. Thus, convergence is expected to occur earlier. The mutation operator is the exploitation function of the search space in the GA. Generally in binary encoded strings, flip-bit (i.e., single-bit or multi-bit) mutation is applied which is basically inverting or "flipping" a randomly selected bit in the parent with an extremely small mutation rate. But the problem of flip-bit mutation is that, if the most of the flipping positions of the chromosome are '0' (i.e., inactive genes) then these are converted to '1' (i.e., active genes) that increases the active genes in the chromosomes and degrade our objectives to select minimum number of informative genes. To overcome these demerits, a unique proximity mutation methodology is used in the paper for mutating the genes. Thus, the proposed method preserves the diversity of the population applying multipoint crossover and proximity mutation techniques. The replacement strategy for creation of the next generation population is based on the Pareto optimal concept [Olmo et al., 12], [Shelokar et al., 13]with respect to both objective functions and after final generation of the PMOGA we get some non-dominated Pareto optimal gene sets. The PMOGA shows very promising result with less computational complexity as there is no need of global calculation typical of other Pareto based MOEA [Olmo et al., 12], [Shelokar et al., 13]. It uses a steady state selection mechanism, no need for fitness sharing parameter used in NSGA [Srinivas and Deb, 95]or crowding distance used in NSGA-II [Deb et al., 02]or converting the multi-objectives problem into scalar objective problem and use of weighted aggregation concept of the individual objectives in MOEA/D [Zhang and Li, 07].

Finally, the target gene subset consists of minimum number of genes providing maximum classification accuracy is identified with the help of an evaluation function defined on both accuracy and number of genes. The accuracy part of the evaluation function depends on Support Vector Machine (SVM), which is very efficient classifier in two-class information system. Some papers [Mohamad and Deris, 05], [Mohamad et al., 09], [Alba et al., 07], [Gonz´alez-Navarro and Belanche-Mu˜noz, 14]have also used classification accuracy as fitness function computed by SVM classifier in every generation before selecting informative gene subset or without filtering unimportant genes. So, these algorithms [Mohamad and Deris, 05], [Mohamad et al., 09], [Alba et al., 07], [Gonz´alez-Navarro and Belanche-Mu˜noz, 14]are more time consuming but the proposed evaluation function is applied only on non-dominated Pareto optimal sets obtained from PMOGA and are able to identify

the target gene subset effectively. Figure 1 shows the basic structure of proposed gene subset selection method (GSSM).

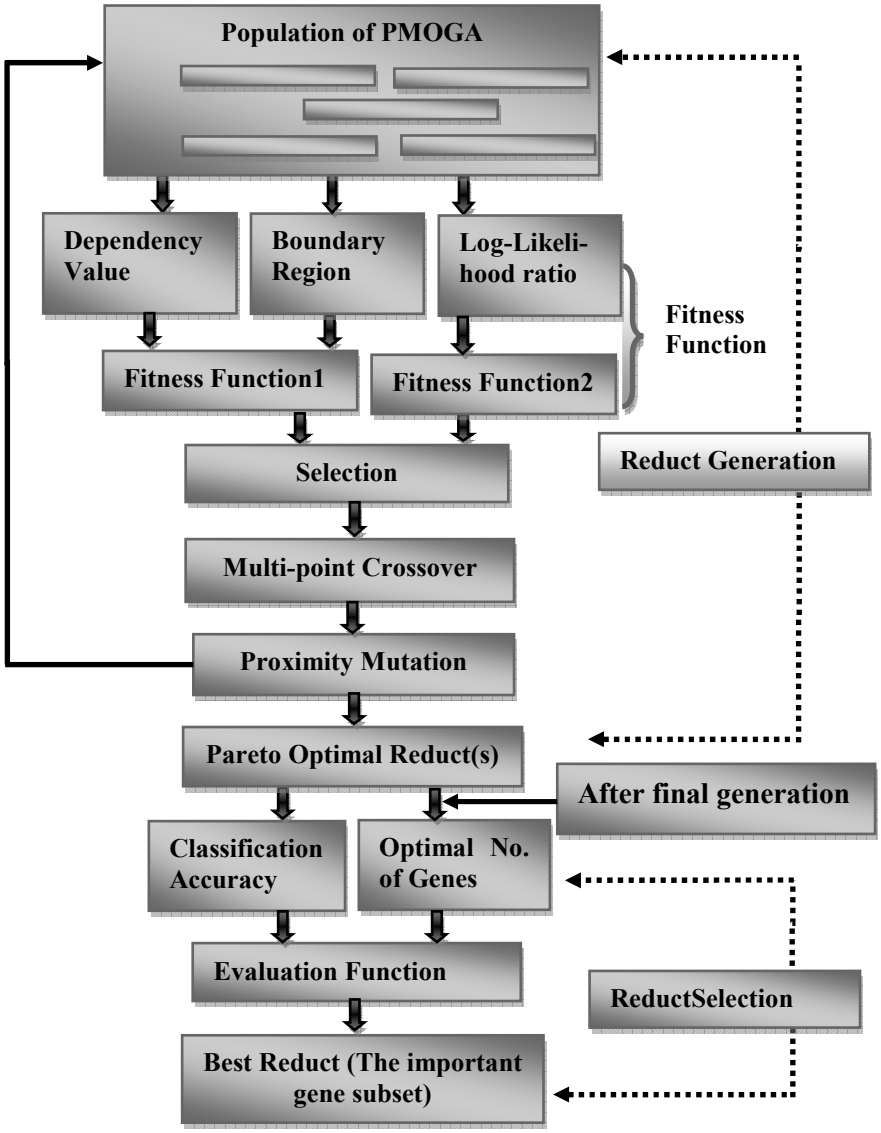

Figure 1: Architecture of the proposed GSSM method

3 Gene Subset Selection

In this section, we describe the gene subset selection method (GSSM) for optimal number of gene selection with maximum classification accuracy of Microarray dataset. Firstly, the proposed GSSM selects some Pareto optimal gene subsets using PMOGA, and then desired subset is selected using an evaluation function computed by the minimum number of genes providing maximum classification accuracy.

Gene selection [Odibat and Reddy, 14] based on single criteria may not always yield better result due to varied characteristics of gene dataset. Multiple criteria if combined for gene selection algorithm generally provides more informative genes compare to a single one, resulting better performance of the algorithm. The paper proposes a novel multi-objective GA using Pareto optimal concept for gene selection (PMOGA), which effectively reduces dimensionality of the gene dataset without sacrificing sample classification. The method uses innovative initial population generation concept with the help of cellular automata, steady state selection strategy, multi-point crossover operation and proximity mutation to maintain diversity in the population.

Gene Initial Population generation is a crucial task in evolutionary algorithms. If no information about the solution is available, then random initialization is the most commonly used method to generate initial population but it takes long computational time, especially when the solution space is difficult to explore [Maaranen, 04]. So, the paper presents a novel technique for initialization of population by applying cellular automata concept to make simple and faster initialization in high dimensional space.

Abstract Cellular Automata (CA) [Neumann, 96], a pseudorandom pattern generator, plays an important role for any population based stochastic search method. In our work, the non-linear hybrid uniform cellular automata have been used for generating the initial population covering majority portion of the search space. As most of the search space can be explored, optimization becomes more efficient using the proposed population generation approach. The model is represented as a large number of cells organized in the form of a lattice where each cell has the capability of self-reproduction and is as powerful as universal Turing machine [Herken, 95]. The proposed method generates next state of a cell using its own state and states of its neighbouring cells based on the rules R_1, R_2, R_3, and R_4 as defined in Equation (1). In this paper, we have considered only 3-neighborhoods namely; left neighbour, self or current and right neighbour one-dimensional cellular automata and each cell only have any one of two states (‘0’ or ‘1’).

$$\begin{aligned} R_1&: \ Next_{state(i)} = \big(L(i) \wedge C(i)\big) \vee \big(\sim L(i) \wedge R(i)\big) \\ R_2&: \ Next_{state(i)} = \big(L(i) \wedge R(i)\big) \vee \big(C(i) \wedge \sim R(i)\big) \\ R_3&: \ Next_{state(i)} = L(i) \oplus C(i) \oplus R(i) \\ R_4&: \ Next_{state(i)} = C(i) \oplus \big(L(i) \vee \sim R(i)\big) \end{aligned} \tag{1}$$

L(i) is the left cell value of current cell *i*, *C(i)* is the *i*-th current cell value and *R(i)* is the right cell value of *C*(*i*). For every cell to generate the next state, a feasible rule is chosen dynamically. Among the rules, R_1, R_2, and R_4 are non-linear while R_3 is

linear and therefore named non-linear hybrid CA. These rules are used for population generation as described with Example 1.

Example 1: The Binary chromosomes are randomly generated having length equal to the number of features in the data set. Say, there are five features in the data set and a randomly generated chromosome (seed) is *11011*. To each cell any rule from R_1 to $R_4$is randomly assigned and the next state value for corresponding cell is obtained, as shown in Figure 2, where *R(i)* and *L(i)* for any cell are obtained from *C(i)* value of right and left cell, respectively (shown in Figure 2 by arrow lines) and generate a binary pattern *10101* as next state after applying rules. The same process is repeated for a certain number of times to obtain all chromosomes in the population.

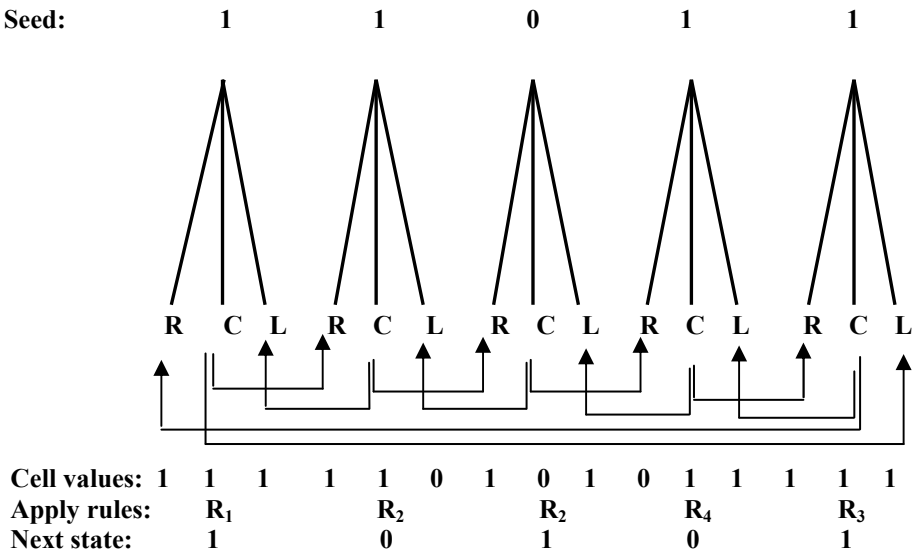

Figure 2: Generation of next-state population using Cellular Automata

In Figure 2, for first bit, the R_1 is applied and $Next_state(1) = (L(1) \wedge C(1)) \vee (\sim L(1) \wedge R(1)) = (1 \wedge 1) \vee (\sim 1 \wedge 1) = 1 \vee 0 = 1$.

As fitness function determines quality of a solution in the population, so a strong fitness function is imperative for obtaining good result. Contrary to single objective GA, multi-objective GA deals with simultaneous optimization of several incommensurable and often competing objectives. The objectives often conflict with each other. Improvement of one objective may lead to deterioration of another. Our method uses a bi-objective fitness function with two parameters based on attribute dependency value with exploring boundary region in RST [Pawlak, 98]and Log-Likelihood ratio [Jerzy and Pearson, 33]in information theory. These two objectives are conflicting in nature and are used to approximate a set of Pareto optimal solutions.

Rough set theory [Pawlak, 98]is a mathematical tool to deal with incomplete, imprecise or uncertain information from granularity in the domain of discourse. The

granularity is based on the indiscernibility relation generated by information about objects of interest that are indistinguishable from each other.Let, *I=(U, A)* be an information system where *U* is the finite, non-empty set of objects (called the *universe*) and *A*=(*C* ∪ *D)*is a finite, non-empty set of *attributes* with *C* and *D* as the condition and decision attributes. Each attribute $a \in A$ can be defined mathematically, as a function described in Equation (2).

$$f_a: U \rightarrow V_a, \forall a \in A \tag{2}$$

Where, V_a is the set of values of attribute *a*, called the *domain* of *a* and f_a is the function representing on attribute *a*. For calculating dependency value [23] of a target set *X* with respect to an gene subset *P*, universe of discourse *U* is partitioned into equivalence classes $[x]_P$ using an indiscernible relation *IND(P)*, in Equation (3).

$$IND(P) = \{(x, y) \in (U \times U) | \forall a \in P, f_a(x) = f_a(y)\} \tag{3}$$

Where, $f_a(x)$ is the function representing the value of object *x* on attribute *a*. Similarly, equivalence classes $[x]_D$ are formed using Equation (3) for the subset *D* consisting of decision attributes. Thus, two different partitions *U*/*P* and *U*/*D* of equivalence classes $[x]_P$ and $[x]_D$ are obtained. Now each class $[x]_D$ in *U*/*D* is considered to be the target set *X*, (i.e., *X*∈*U*/*D*). The lower approximation set $\underline{P}X$ under *P* is computed using Equation (4), whose elements are certainly member of *U*/*P*. The positive region $POS_P(D)$ is obtained by taking union of lower approximations $\underline{P}X$under *P* for all *X* in *U*/*D*, using Equation (5). Dependency value of decision attribute *D* on *P* (i.e.,$\gamma_P(D)$) is calculated using Equation (6), the value ranges from 0 to 1. Lower and upper approximation of a set *X* is shown in Figure 3, which clearly shows that more the objects in positive region implies less number of objects in boundary region and so dependency value increases.

We use $\gamma_P(D)$ as the first fitness function, which is to be maximized for utilizing dependency of attributes to infer a decision (*D*). More dependency of decision attribute with respect to an attribute subset implies that the attributes are more significant.

$$\underline{P}X = \{x | [x]_P \subseteq X\} \tag{4}$$

$$POS_P(D) = \cup_{X \in U/D} \underline{P}X \tag{5}$$

$$\gamma_P(D) = \frac{|\, POS_P(D)|}{|\, U\,|} \tag{6}$$

The upper approximation $\bar{P}X$ of target set *X*, for all $X \in U/D$ under attribute subset *P* is computed using Equation (7) which contains the set of attributes which possibly belong to the target set *X* and the boundary region $BND_P(D)$, as shown in Figure 3, for the decision system is obtained using Equation (8) which possesses the degree of uncertainty as the objects in this region may or may not belong to the target set.

$$\bar{P}X = \{x\, | [x]_P \cap X \neq \emptyset\} \tag{7}$$

$$BND_P(D) = \cup_{X \in \frac{U}{D}} (\bar{P}X) - \cup_{X \in \frac{U}{D}} \left(\underline{P}X\right) \tag{8}$$

Obviously from the definition of positive region, equivalence class $[x]_P$ in *U/P* that is not a subset of *X* in *U/D*, falls in the boundary region $BND_P(D)$. If more $[x]_P$ falls in the boundary region, then the dependency value $\gamma_P(D)$ will decrease. An equivalence class $[x]_P$ falls in $BND_P(D)$ because of some objects in $[x]_P$ thatdo not belong to *X*. If very few objects of $[x]_P$ are responsible for placing it in the boundary region, then the class $[x]_P$ almost agrees to the target set *X*, i.e., a class $[x]_D$ in *U/D*. So dependency should not be the only criterion for reduct generation. To overcome this shortcoming, boundary region is explored by computing similarity factor $\delta_P(D)$ of set $CB_P(D)$ (classes $[x]_P$ of *U/P* whose objects lie in the boundary region, formed using Equation (9)) to *U/D*, formulated using Equation (10).

$$CB_P(D) = [x]_P \mid \left([x]_P \in \frac{U}{P}\right) \wedge \left(X \in POS_P(D)\right) \tag{9}$$

$$\delta_P(D) = \frac{1}{BND_P(D)} \sum_{[x]_P \in CB_P(D)} \max_{[x]_D \in U/D} (|[x]_P \cap [x]_D|) \tag{10}$$

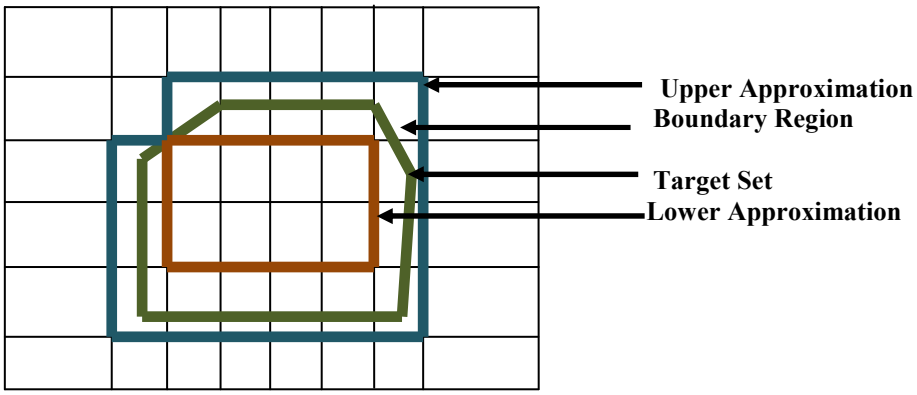

Figure 3: Illustrate the boundary region of a target set

In Equation (10), summation of maximum number of common objects between an element $[x]_P \in CB_P(D)$ and all elements $[x]_P$ in *U/D* is calculated, for all $[x]_P$ in $CB_P(D)$ and then it is divided by the total number of objects in $BND_P(D)$. So, if veryfew objects of $[x]_P$ are responsible for placing it into the boundary region, then the class $[x]_P$ almost agree the target class, i.e., a class $[x]_D$ in *U/D* and similarity factor $\delta_P(D)$ will increase, where in the same situation dependency value $\gamma_P(D)$ decreases. Since, for a decision system, these two factors namely, dependency value $\gamma_P(D)$ and similarity factor $\delta_P(D)$ needs a maximization, so the fitness function *F(ch)* for chromosome *ch* of associated GA-based optimization problem is considered as the weighted average of these two factors, computed using Equation (11).

$$F(ch) = \mathcal{W}.\gamma_P(D) + (1 - \mathcal{W}).\delta_P(D) \tag{11}$$

Where, $\mathcal{W}$ is the weight factor of$\gamma_P(D)$, which is taken as 0.5 in our experiment by examining several test conditions. Obviously, higher the fitness value *F(ch)*, better the quality of the chromosome (or encoded string) *ch*.

Example 2: Illustration for finding Fitness value *F*(*ch*): For a decision system with21 objects, let *P* is the subset of the conditional attribute set *C* and *D* is the decision attribute, so that the equivalence classes of objects induced by the indiscernibility relations *IND*(*P*) and *IND*(*D*) on *P* and *D* are as follows: *U*/*P* ={{1, 2, 3, 4}, {5, 6, 7, 14}, {8, 10, 11, 18, 21}, {9, 12, 13, 15, 16, 17, 19, 20}} and *U*/*D* = {{1, 2, 3, 4}, {5, 6, 7, 8, 9, 10, 11}, {12, 13, 14, 15, 16, 17, 18, 19,20, 21}}.

So the positive region for the target sets in *U*/*D* is obtained by Equation (6) as $POS_P(D)$ = {1, 2, 3, 4} and the boundary region is obtained by Equation (8) as $BND_P(D)$= {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21} which can be partitioned into classes using Equation (9) as $CB_P(D)$= {{5, 6, 7, 14}, {8, 10, 11, 18, 21}, {9, 12, 13, 15, 16, 17, 19, 20}}. Thus, for a decision system, positive region and boundary region under condition attribute subset *P* are obtained, as shown inFigure4. The positive region helps to compute $\gamma_P(D)$ and boundary region tocompute$\delta_P(D)$. Using Equation (6), the dependency is computed as $\gamma_P(D)$= 4/21 and similarity factor is computed using Equation (10) as $\delta_P(D)$ = 13/17, since sum of the maximum number of overlapping objects for classes in $CB_P(D)$ with the classesin*U*/*D* is 13 out of total 17 objects in the boundary region. Therefore, fitness value for a chromosome *ch*, encoded as '1' for genes in *P* and '0' for other genes is given by Equation (11) as $F(ch) = \mathcal{W}$. (4/21) + (1 −$\mathcal{W}$).(13/17) = 0.48, considering $\mathcal{W}$= 0.5.

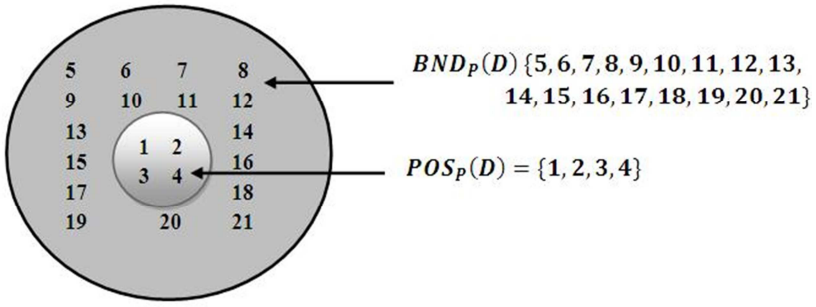

Figure 4: Positive and boundary regions of a given gene set

Likely-hood ratio (*LLR*) is a general and powerful method of testing model assumptions. The *LLR* [Jerzy and Pearson, 33] quantifies the proximity of two probability distributions in information theory. It is a measure in statistics that quantifies how close a probability distribution *p(x)* is to a model distribution *q(x)*. The *LLR*, which is non-negative and non-symmetric in two probability distributions *p (x)* and *q (x),* is defined in Equation (12).

$$LLR = \frac{1}{n}\sum_{i=1}^{n} log_2\left(\frac{p(x_i)}{q(x_i)}\right) \qquad (12)$$

The *LLR* is used as another fitness function and to be minimized that governs maximum similarity between *p(x)* and *q(x)*.

Reproduction directs the search towards the best existing individuals but is unable to create new individuals. To create new individuals, crossover operation is required. Two new offspring are generated from a selected pair of parents applying crossover operation with probability c_p. In our method 2-point crossover has been used generating two random numbers, indicating positions of bits in chromosome. Then, the substrings of the parent strings, lying between the two randomly generated positions, are interchanged. Thus, two new individuals are created. The motivation for using 2-point crossover is that, the new individuals generated are more similar to one of their high quality parents than they are in 1-point crossover so that convergence is expected to occur earlier.

In single-bit mutation, a gene is randomly selected to be mutated and its value is changed depending on the encoding type used but it lacks diversity in population as the first bit of the binary string generally does not change. In multi-bit mutation, multiple genes are randomly selected for mutation and there values are changed depending on the encoding type used. So, both of the mutation is depended on the flip-bit mutation and random bit number generation with respect to mutation probability m_p, which is inefficient in high dimensional space. Finding the minimum number of active genes in the chromosome is one of the objectives of our proposed method, so flip-bit mutation methods may diversify the population. To overcome these demerits, proximity mutation is used in the paper for mutating the chromosomes, which builds on flip-bit mutation but modifies it greatly to produce fittest offspring. The proposed mutation method works in the following manner.

Let, two random positions are generated and count the number of '0's and '1's between these positions, say c_0 and c_1, respectively. Then we have two cases to analyse:

> **Case 1:** If $c_0 \leq c_1$, then we swap '0'and '1' and generate the offspring.
> **Case 2:** If $c_0 > c_1$, then if $(c_0 - c_1) \leq C$, then they are flipped else, the chromosome remain unchanged. Where, C is a positive constant which depends on the optimization goals.

These two cases ensure the reduction in number of active genes(i.e., '1's) in the chromosome, which would generate a greater diversity in terms of lower number of '1's than flip-bit mutation that is our objective with regard to the least number of ones in the chromosome.

In our optimization problem, the two objectives are conflicting in nature and cannot be optimized simultaneously. Therefore, it is necessary to have a decision making process in which preference information is used in selecting an appropriate trade-off. The replacement strategy of PMOGA is based on Pareto optimality concept. Figure 5 demonstrates the measurement of dominance based on the concept of Pareto optimality, defined below.

> **Definition 1 (Strongly dominated solution):** A solution X_2 is said to be strongly dominated by another solution X_1, if the solution X_1 is strictly better than solution X_2 with respect to all objectives.

Definition 2 (Non-dominated solution): The solutions $X_1$and X_2 are said to be non-dominated to each other, if some objective values of each solution are higher than that of the other.

Definition 3 (Dominated solution): A solution $X_1$is said to be dominated by another solution X_2, if the solution X_1 is strictly worse than solution $X_2$with respect to all objectives.

In Figure 5, let F_1 and F_2 values are considered as two objective functions. Thus, a solution defined by corresponding decision vector can be better than, worse, or equal to, but also indifferent from another solution with respect to the objective values. Here, better means a solution is not worse in any objective and better in at least one objective function. The solution represented by point *P* is worse than the solution represented by point *Q*, and the solution with *R* is better than that of *Q*. But, it cannot be said that *R* is better than *S* or vice versa because one objective value of each point is higher than the other one. These are called non-dominated or Pareto optimal solution represent by dotted line in Figure 5. The solution *T* is strongly dominated compared to all other solutions with respect to both objective values. So after mutation, both fitness values are evaluated for offspring and the elitism property is maintained replacing parent with its offspring based on strong dominated or non-dominated properties.

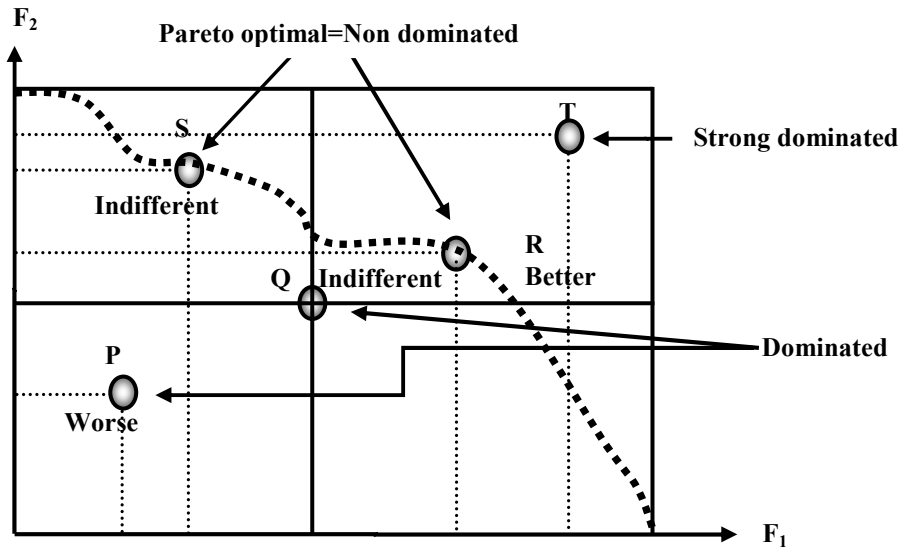

Figure 5: The concept of Pareto optimality

We describe a pseudo-code of the proposed reduct generation method (PMOGA) below.

Procedure: *PMOGA (DS)*
Input: Population size: *M*
Maximum number of generations: *G*
Crossover probability: c_p
Mutation rate: m_p

Output: Strongly dominated and non-dominated solutions (Reducts)
BEGIN
Generate initial Population P of size M using nonlinear hybrid uniform cellular automata;
Evaluate fitness values of all chromosomes;
Set, $c = 0$;
Repeat
FOR $i=1$ to M **DO**
First_ parent = P_i;
Select another parent randomly from the remaining population;
Apply multipoint crossover with probability c_p and produce two offspring;
Use Proximity mutation to the offspring with mutation rate m_p;
Evaluate fitness values of the offspring;
IF (both the offspring either strongly dominate or non-dominate with the parents) **THEN**
Both the parents are replaced by these offspring;
IF (only one offspring either strongly dominates or non-dominated to parents) **THEN**
This offspring replaces the dominated parent;
IF (both the offspring are dominated by the parents)
THEN
The offspring are discarded;
END FOR
$c = c+1$**;**
Until ($c<=G$);
Return strongly dominated and non-dominated chromosomes with fitness values;
END

Example 3: Suppose [F_1, F_2] be an objective vector, where maximum of $F_1$and minimum of $F_2$are desired. Let, parents $P_1$and P_2 are chosen randomly with their objective values [0.77, 0.35] and [0.59, 0.15] respectively. Say, after crossover and mutation phase two new offspring $C_1$and C_2 are produced with their objective values. Then the following situations may occur.

(a) If the objective values of C_1 and C_2 are [0.92, 0.10] and [0.85, 0.05] respectively, then P_1 and P_2 are replaced with C_1 and C_2 according to strong dominance property.

(b) If the objective values of C_1 and C_2 are [0.90, 0.12] and [0.75, 0.18] respectively, then P_1 and P_2 are replaced with C_1 and C_2 according to strong and non-dominance property.

(c) If the objective values of C_1 and C_2 are [0.81, 0.37] and [0.51, 0.09] respectively, then P_1 and P_2 are replaced with C_1 and C_2 according to non-dominance property.

(d) If the objective values of C_1 and C_2 are [0.83, 0.43] and [0.47, 0.38] respectively, then P_1 or P_2 is replaced by C_1 and C_2 is removed from population according to non-dominance and dominance property.

If the objective values of C_1 and C_2 are [0.55, 0.41] and [0.42, 0.37] respectively, then both the C_1 and C_2 are removed from population according to dominance property.

The PMOGA produces some Pareto optimal non-dominated solutions (i.e., reducts) representing various gene subsets. Now our objective is to select the best gene subset among these solutions that gives maximum classification accuracy and minimum numbers of genes based on defined evaluation function. The evaluation function uses two different measures (a) Classification Accuracy and (b) Optimal Number of Genes. The microarray dataset has two classes of samples, one is normal and other is cancerous. The evaluation function is defined on the classification accuracy of SVM classifier applied on data subset corresponding to the selected reduct. The SVM classifier is a function based classifier and more effective in two class system. Also our target is to find informative gene subset that contains minimum number of genes with maximum accuracy. Pareto front contains many non-dominated chromosomes, some of which may contain large number of genes and some other may contain less number of genes. The evaluation function is so defined that we are sacrificing some accuracy for selecting less number of genes, as less number of important genes fasten the subsequent data analysis task like cancer detection and classification, precautions etc. At the same time, degradation of classification accuracy for involvement of very few genes is not desirable as it may wrongly classify the cancer diseases. Hence, a trade-off between these two measures is important for selecting appropriate chromosomes from the Pareto front for gene subset selection.

The evaluation function is defined as the linear combination of the classification accuracy (*CA*) and number of genes in a reduct (*GR*). As *CA* value may dominates the *GR* value, so *CA* and *GR* values are normalized into (0, 1) and finally, as accuracy is our main concerned so a weight factor α is assigned to normalized *CA* value and β is assigned to *GR* value, where, $\alpha > \beta$ and $\alpha + \beta = 1$. Thus, the evaluation function is defined in Equation (13).

$$EF = \left(\alpha \times \frac{CA}{100}\right) + \left(\beta \times \left(\frac{S - GR}{S}\right)\right) \tag{13}$$

Where, *S* is the total number of genes of dataset. If *GR* is small and *CA* is high for any reduct, then *EF* value is high and gives better reduct with respect to objective functions of the proposed method. So, maximum of *EF* value allows us to select a better gene subset with respect to maximum accuracy and minimum number of genes.

4 Experimental Results and Performance Evaluation

Performance evaluation of the proposed GSSM method and comparative study with some state of the art methods are discussed in this section.

4.1 Dataset description and Parameter setup

Experiments are carried out on benchmark microarray dataset collected from the 'Kent Ridge Bio-medical Data Set Repository' publicly available in[Kent Ridge]that contain high volume of unwanted genes with random noise and the samples are linearly inseparable. The microarray datasets are summarized in Table 1.

The parameters used in PMOGA are listed in Table 2. These parameters are selected after several evaluation of the proposed algorithm on test dataset until reach to the best configuration in terms of the quality (accuracy and other statistical measures) of solutions and the computational cost. The final settings of the parameters are listed in Table 2.

Dataset	#Genes	Class Name	#Samples (class1/class2)
Leukemia	7129	ALL/AML	72(47/25)
Lung cancer	12533	MPM/ADCA	181(31/150)
Prostate cancer	12600	Tumor/Normal	102(52/50)
Breast cancer	24481	Relapse/non- Relapse	78(34/44)
Colon cancer	2000	Negative/Positive	62(40/22)
DLBCL data	6817	DLBCL/FL	58(32/26)

Table 1: Summary of Microarray dataset

Parameter	Value
Population size (M)	130
Number of generations (G)	700
Probability of crossover (c_p)	0.76
Probability of mutation (m_p)	0.09
Mutation Constant (C)	10

Table 2: Parameters of PMOGA

4.2 Performance Evaluation of Pareto front Solutions

Two objective functions are used in the proposed PMOGA and a non-dominated set of chromosomes, well distributed on the Pareto fronts are obtained. Figure 6 to Figure 11 show some Pareto fronts solutions obtained by our algorithm for experimental datasets. Also, it is observed that the non-dominated members are well distributed along the front, indicating the correct selection of two objective functions for optimal gene subset selection from the dataset.

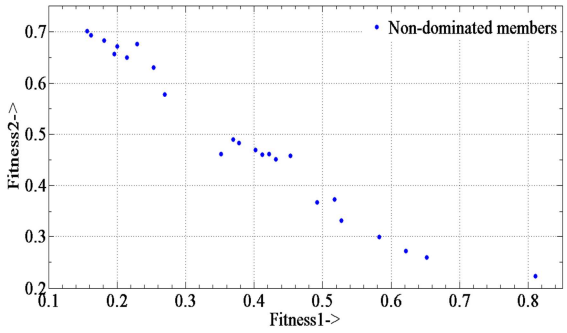

Figure 6: Non-dominated (Pareto fronts) solutions for Leukemia data

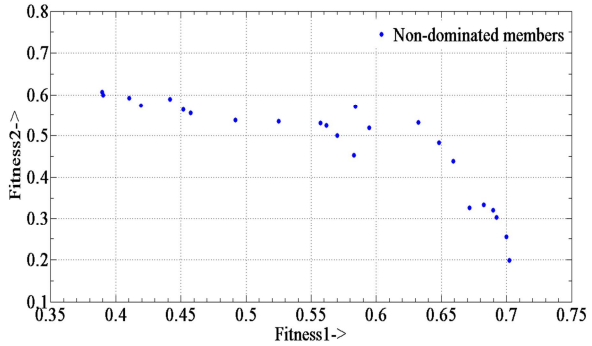

Figure 7: Non-dominated (Pareto fronts) solutions for Lung cancer data

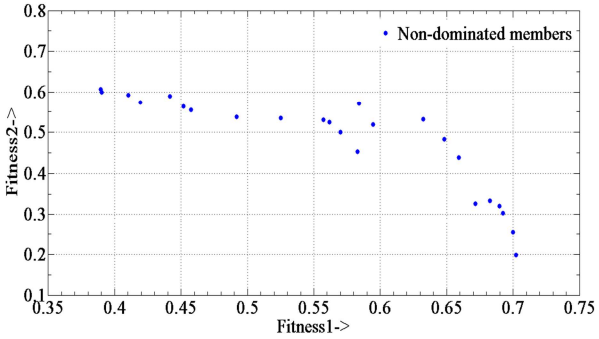

Figure 8: Non-dominated (Pareto fronts) solutions for Prostate cancer data

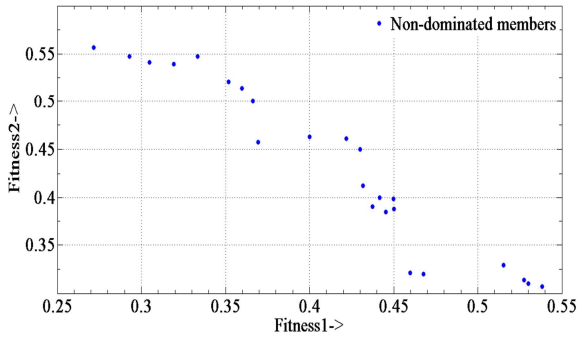

Figure 9: Non-dominated (Pareto fronts) solutions for Breast cancer data

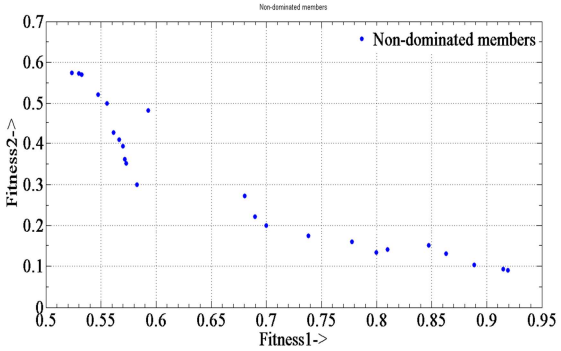

Figure 10: Non-dominated (Pareto fronts) solutions for Colon cancer data

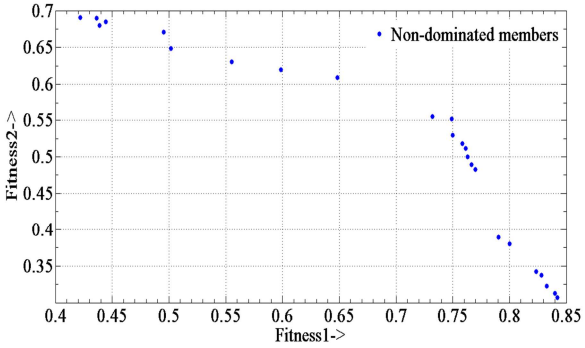

Figure 11: Non-dominated (Pareto fronts) solutions for DLBCL data

We run the algorithm several times and after final generation the Pareto fronts with respect to both fitness functions for the datasets is considered for measuring minimum (i.e., *Min.*), maximum (i.e., *Max.*), mean (i.e., *Avg.*) and standard deviation

(i.e., *Std.*) among all chromosomes in the fronts, listed in Table 3. Also the average value of these measures for consecutive 50 such runs on the datasets are presented in Table 4.This signifies that, the chromosomes in the front are very close to each other and remain close whatever may be the number of runs. This implies that the method give us the stable Pareto fronts which is desired for optimal gene selection problem.

Dataset	Fitness1				Fitness2			
	Min.	Max.	Avg.	Std.	Min.	Max.	Avg.	Std.
Leukemia	0.2776	0.8325	0.6071	0.4406	0.2027	0.6541	0.4144	0.1752
Lung	0.4305	0.8707	0.7219	0.1738	0.0933	0.7638	0.3006	0.3505
Prostrate	0.4083	0.7522	0.5807	0.3002	0.1467	0.5729	0.3482	0.1361
Breast	0.3681	0.6080	0.4890	0.1724	0.2710	0.5206	0.4092	0.7419
Colon	0.5943	0.9271	0.8233	0.2073	0.1785	0.5464	0.1923	0.1310
DLBCL	0.4872	0.8636	0.7351	0.1540	0.2854	0.6260	0.4307	0.1864

Table 3: Statistical measures of the population after final generation for a run

Dataset	Fitness1				Fitness2			
	Min.	Max.	Avg.	Std.	Min.	Max.	Avg.	Std.
Leukemia	0.3711	0.5375	0.4839	0.0926	0.5261	0.6920	0.6337	0.2044
Lung	0.6055	0.7047	0.6710	0.1472	0.2205	0.4033	0.3064	0.1095
Prostate	0.3520	0.5200	0.4828	0.3204	0.2094	0.4225	0.3500	0.1704
Breast	0.2996	0.5372	0.4170	0.2374	0.6011	0.7408	0.7170	0.4586
Colon	0.5264	0.6066	0.5527	0.1506	0.4309	0.5674	0.5281	0.1407
DLBCL	0.7340	0.8429	0.7903	0.2003	0.3875	0.6128	0.4327	0.1816

Table 4: Avg. statistical measures of the population after final generation for 50 runs

4.3 Evaluation of Reducts

A chromosome with minimal number of genes providing the highest classification accuracy (measured by SVM classifier) is considered as the best gene subset for cancer classification. The SVM used RBF Kernel in our experiments.The experiment is independently conducted several times on each dataset to evaluate the reducts using Equation (12) with $\alpha = 0.7$ and $\beta = 0.3$, set experimentally. In our experiments, '*10-fold cross validation*' is used to evaluate classification performance where in each iteration 90% samples (9-fold) are used for training and 10% (1-fold) other samples are used for test purpose.Table 5 shows the value of the evaluation function for the chromosome with number of genes and corresponding accuracy for five such runs.

Table 6shows the results of the proposed method in terms of statistical measures, reporting the b*est* solution found, average *(Avg.)* and Standard Deviation *(Std.)* of 50 independent runs and Table 7 shows the gene names and classification accuracies.

Dataset	Run#	#Genes	Acc. (%)	EF Value
Leukemia	1.	3	100	0.9999
	2.	3	100	0.9999
	3.	3	100	0.9999
	4.	3	100	0.9999
	5.	3	100	0.9999
Lung	1.	4	100	0.9999
	2.	4	100	0.9999
	3.	4	100	0.9999
	4.	4	100	0.9999
	5.	5	100	0.9998
Prostate	1.	2	100	0.9999
	2.	2	100	0.9999
	3.	2	100	0.9999
	4.	2	100	0.9999
	5.	2	100	0.9999
Breast	1.	4	93.58	0.9550
	2.	4	93.58	0.9550
	3.	5	92.31	0.9461
	4.	5	92.31	0.9461
	5.	5	92.31	0.9461
Colon	1.	2	100	0.9997
	2.	2	100	0.9997
	3.	2	100	0.9997
	4.	2	100	0.9997
	5.	3	98.39	0.9882
DLBCL	1.	1	98.68	0.9907
	2.	1	98.68	0.9907
	3.	1	97.40	0.9817
	4.	1	97.40	0.9817
	5.	1	97.40	0.9817

Table 5: Results for five different runs using proposed GSSM

Dataset	#Gene (Best)	#Gene (Avg.)	#Gene (Std.)	Accuracy (Best)	Accuracy (Avg.)	Accuracy (Std.)
Leukemia	3	2.38	0.5963	100	99.4997	0.8683
Lung	4	4.08	0.2713	100	100	0.0000
Prostate	2	2	0.0000	100	100	0.0000
Breast	4	4.20	0.4899	93.58	92.8942	0.6095
Colon	2	2	0.0000	100	99.16	0.8319
DLBCL	1	1	0.0000	98.68	97.62	0.7085

Table 6: Statistical results of GSSM method by 50 independent executions

Dataset	Accuracy(#gene)	Gene name
Leukemia	100(3)	L12052_at,M23197_at, U50136_rnal_at
Lung	100(4)	36245_at, 37205_at, 32046_at,37957_at
Prostate	100(2)	37639_at,39939_at
Breast	93.58(4)	AB022847, NM_012109, NM_007321, NM_006191
Colon	100(2)	U29092, M55543
DLBCL	98.68(1)	M35878_at

Table 7: Selected gene subset with SVM classifier accuracy

4.4 Statistical performance analysis

Receiver Operating Characteristics (ROC) curve is a graphical representation of the relationship between both sensitivity and specificity and it helps to visualise the performance of the classifier. The sensitivity and specificity is defined by Equation (14) and Equation (15) respectively. ROC curve is defined between True Positive Rate (*TPR*) and False Positive Rate (*FPR*) within an area, starting from coordinate (0, 0) and ending at coordinate (1, 1). Figure 12 shows the two dimensional ROC curve for the used datasets. Where, *FPR* (i.e., 1 – specificity) is represented by *x*-axis and *TPR* (i.e., sensitivity) is represented by *y*-axis. The graphical interpretation of ROC curve is that if the points on the ROC curve are closer to the ideal coordinate (i.e. provides more area in the ROC space), then the test is more accurate but the points on the ROC curve closer to the diagonal (i.e. provides less area in the ROC space) implies the test is less accurate.

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{15}$$

Where, *TP* is the positive object classified as positive, *FP* is the negative object classified as positive, *TN* is the negative object classified as negative and *FN* is the positive object classified as negative.

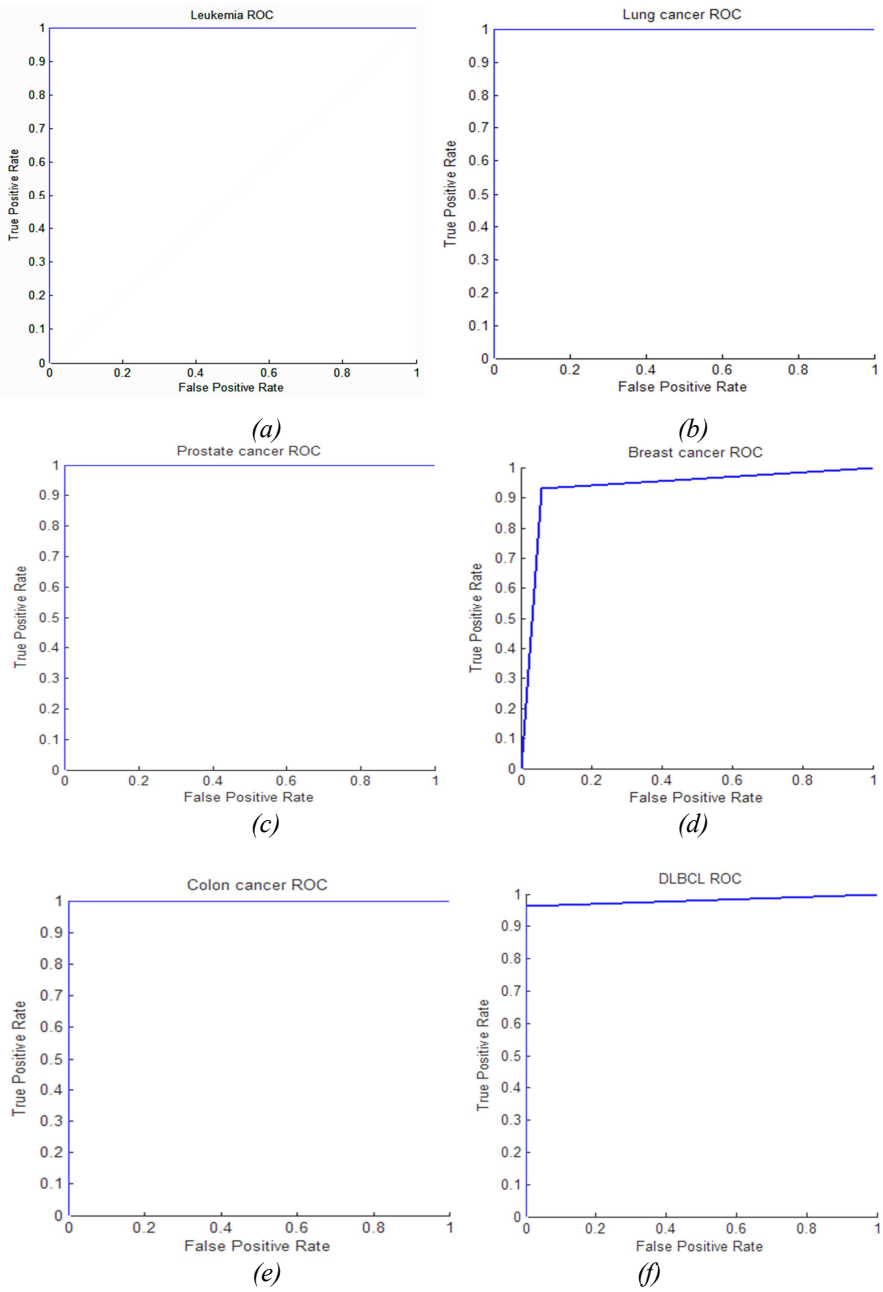

Figure 12: ROC curve for (a) Leukemia (b) Lung (c) Prostate (d) Breast (e) Colon and (f) DLBCL cancer data

From Figure 12 (a) to (f), it is observed that the ROC curves for all the datasets are reasonably closer to the upper left corner (area of ROC curve is high) that confirms the higher sensitivity/specificity rate and overall accuracy of the proposed method is up to the mark.

4.5 Statistical performance analysis

To show the effectiveness of the proposed method, a performance comparison is made between the proposed GSSM and the methods described in the literature and the results are shown in Table 8. This result shows that the proposed gene selection method GSSM has the ability to obtain highly informative genes and achieves comparatively better classification performance than other methods. The results for existing methods are collected from corresponding papers where classifier with the maximum accuracy is listed in Table 8. Our method takes the minimum number of selected genes and measures the accuracy of SVM classifier on reduced dataset. The Table also shows the average accuracy of the proposed and other compared methods for each of the dataset.

The parameters of the proposed and compared genetic algorithm based gene selection methods (listed in Table 8) are listed in Table 9.

4.6 Biological Significance of Selected Genes

The optimal gene subset and corresponding classification accuracy is listed in Table 7. We can provide a brief biological significance of some of the most commonly obtained genes since they are recently used in relevant medical literatures. Some of important genes of different data set are listed below.

M23197_at and L12052_at: In Leukemia dataset, gene M23197_atis up regulated and important biomarker of AML [Golub et al., 99], [Lamba et al., 09]. It is reported to be correlated with the prognosis and diagnosis of cancer and listed of the 50 informative genes distinguishing ALL from AML [Golub et al., 99] and marked as highly expressed gene in AML.Gene L12052_at is used in drugs like Anagrelide or Milrinone and effective in treating patients with certain kinds of leukemia[Silver, 05].These genes are belonging to a set of three genes (reported in Table 7 for Leukemia dataset) with 100% accuracy selected by GSSM.

36245_at and 37957_at: Gene 36245_at is found in human plasma membrane and used in several drugs, like Risperidone, Blonanserin, and Mirtazapine, 37957_at is activated during amino acid deprivation and associated with different diseases, like neurodegenerative diseases, lung cancer, pathogen infections, myopathies etc. [Scherz-Shouval, 07]. These genes are belonging to a set of four genes (reported in Table 7 for Lung cancer dataset) with 100% accuracy selected by GSSM.

37639_at: Gene 37639_atis a cell surface serine protease and takes an essential part in cell growth and maintenance of cell morphology. It is well associated with prostate cancer, benign prostatic hyperplasia [GenCards, 09]. This gene is belonging to a set of two genes (reported in Table 7 for Prostate cancer dataset) with 100% accuracy selected by GSSM.

AB022847 andNM_006191: Gene AB022847is located in plasma membrane and used in different drugs like, Radaxafine, Amphetamine or Venlafaxine [Loprinzi, 00]. The gene NM_006191 is shown to be a transcriptional corepressor that slows down

the growth of human breast cancer cell lines [Akinmade et al., 08]. These genes are belonging to a set of four genes (reported in Table 7 for Breast cancer dataset) with 93.58% accuracy selected by GSSM.

Dataset	Method	#Genes	Classification Method	Accuracy (%)
Leukemia	Monte Carlo, Step wise[Xiong et al., 01]	2	FLDA	95.80
	Hyk Gene [Wang et al., 05]	4	KNN	98.61
	GA [Schaefer, 10]	100	Fuzzy	98.61
	FRNf-I [Kumar et al., 15]	3	ANN	99.01
	New-GASVM [Mohamad and Deris, 05]	40	SVM	**100**
	MOGASVM [Mohamad et al., 09]	2252	SVM	97.37
	α-value [Wang and Gotoh, 10]	1-100	NB	**100**
	Simulated Annealing [Gonz´alez-Navarro and Belanche-Mu˜noz, 14]	3	SVM	99.62
	GSSM	2	SVM	**100**
	Average Accuracy			98.78
Lung	α-value [Wang and Gotoh, 10]	1-100	NB	**100**
	MFMW [Leung and Hung, 10]	6	C4.5	98.34
	FRNf-I [Kumar et al., 15]	4	ANN	99.40
	GSSM	4	SVM	**100**
	Average Accuracy			99.44
Prostate	Discretization [Tan and Gilbert, 03]	3071	DT	73.53
	Signal to noise ratios [Singh et al., 02]	16	KNN	85.70
	α-depended degree + decision rules[Wang and Gotoh, 09]	1	Classification Rule	91.00
	α-value [Wang and Gotoh, 10]	1-100	SVM	98.04
	GSSM	2	SVM	**100**
	Average Accuracy			89.65
Breast	α-value [Wang and Gotoh, 10]	1-100	DT	88.46
	Simulated Annealing [Gonz´alez-Navarro and Belanche-Mu˜noz, 14]	6	SVM	86.90
	GSSM	4	SVM	**93.58**
	Average Accuracy			89.65
Colon	Monte Carlo, Step wise [Xiong et al., 01]	3	FLDA	93.50
	Rank sum, PCA, Clustering [Liu, et al., 04]	30	Ensemble ANN	91.94
	GA [Schaefer, 10]	50	Fuzzy	85.48
	MFMW [Leung and Hung, 10]	6	C4.5	95.16
	FRNf-I [Kumar et al., 15]	5	ANN	98.40
	New-GASVM [Mohamad and Deris, 05]	30	SVM	98.39
	MOGASVM [Mohamad et al., 09]	446	SVM	96.16
	α-value [Wang and Gotoh, 10]	1-100	DT	91.93
	Simulated Annealing [Gonz´alez-Navarro and Belanche-Mu˜noz, 14]	5	SVM	89.19
	GSSM	2	SVM	**100**
	Average Accuracy			94.02
DLBCL	α-value [Wang and Gotoh, 10]	1-100	SVM/KNN/DT/NB	84.48
	GSSM	1	SVM	**98.68**
	Average Accuracy			91.58

Table 8: Comparative study between GSSM and other methods described in literature

Method	Population size	Crossover rate	Mutation rate
New-GASVM [35]	100	0.70	0.01
MOGASVM [36]	100	0.70	0.01
GA [41]	NA	0.90	0.10
GSSM	130	0.76	0.09

Table 9: Comparison of GA environment between GSSM and GA based methods

5 Conclusions

Systematic and unbiased approach to cancer classification is an important treatment of the disease and drug discovery. Biologists focus on a small subset of genes that dominate the outcomes before conducting in depth analysis and expensive experiments with a larger set of genes. Therefore, automated discovery of this small and informative gene subset is highly desirable. In the paper, a novel multi-objective genetic algorithm has been proposed to select non-dominated solution set providing minimum number of relevant genes for cancer classification. The method uses two fitness functions separately based on the concepts of both rough set theory and information theory. Here, rough set theory is used to remove imprecise and vague data and collect only the precise one computing positive region and exploring boundary region of the target sets and on the other hand, Log-Likelihood ratio is considered as another fitness function to select only the informative genes. Nonlinear uniform cellular automata concept is used to generate initial population in high dimensional space and 2-point crossover and proximity mutation operation are used to maintain the diversity in the population. At last, an evaluation function is defined to select the minimum number of genes with the maximum classification accuracy from a set of non-dominated solution set.In Future enhancements to this work may include the use of neural network for not computing experimentally rather fixing theoretically the weight $\mathcal{W}$used to compute fitness function. Other optimization techniques like Ant-colony optimization, particle swarm optimization (PSO), Differential Evaluator (DE), and so on may be applied for the same purpose and a through comparative study of the results is very useful for gene subset selection for cancer classification.

Compliance with ethical standards

Conflict of interest: The authors declare that there are no conflicts of interest in this paper.

Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

Acknowledgments

This work is supported by the Chung Shan Medical University and Jen-Ai Hospital: CSMU-JAH-106-01.

References

[Akinmade et al., 08] Akinmade, D.,Talukder, A., Zhang, Y., Luo, W., Kumar, R., Hamburger, A.: Phosphorylation of the erbb3 binding protein ebp1 by p21-activated kinase 1 in breast cancer cells, British Journal of Cancer, 98:1132–1140, 2008

[Alba et al., 07] Alba, E., Garc´ıa-Nieto, J., Jourdan, L.,Talbi,E.G., Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms, IEEE Congress on Evolutionary Computation, 284 – 290, 2007, DOI:10.1109/CEC.2007.4424483

[Back and Breukelaar, 05] Back, T.,Breukelaar, R.: Using Genetic Algorithms to Evolve Behavior in Cellular Automata, C.S. Calude et al. (Eds.): UC 2005, LNCS 3699:1–10, 2005

[Das, 01] Das,S.: Filters, wrappers and a boosting-based hybrid for feature selection, In ICML, 1:74-81, 2001

[Das and Pati, 2012] Das, A.K., Pati, S.K.: Gene Subset Selection for Cancer Classification Using Statistical and Rough Set Approach, Swarm, Evolutionary and Memetic Computing (SEMCCO), LNCS 7677, Springer, 294-302, 2012

[Deb et al., 02] Deb, K., Pratap, A., Agarwal, S.,Meyarivan, T.A.: A fast and elitist multi objective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, 6(2):182-197, 2002

[Devroye et al., 96] Devroye, L.,Gyorfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, ISBN 978-1-4612-0711-5, 1996

[Farahat et al., 15] Farahat, A.K.,Elgohary, A.,Ghodsi, A.,Kamel, M.S.: Greedy column subset selection for large-scale data sets, Knowledge and Information Systems, 45(1):1-34, 2015

[Garey and Johnson, 79] Garey, M.R., Johnson, D.S.: Computers and intractability: A guide to the theory of NP-completeness, W.H. Freeman & Co., USA, ISBN: 0716710447, 1979

[GenCards, 09] GenCards,Weizmann Institute of Science: http://www.genecards.org, 2009

[Golub et al., 99] Golub, T.R.,Slonim, D.K., Tamayo, P., Huard, C.,Gaasenbeek, M.,Mesirov, J.P.,Coller, H., Loh, M.L., Downing, J.R.,Caligiuri, M.A.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,Science, 286 (5439):531-537, 1999

[González-Navarro and Belanche-Muñoz, 14] González-Navarro, F.F., Belanche-Muñoz, L.A.: Feature Selection for Microarray Gene Expression Data Using Simulated Annealing Guided by the Multivariate Joint Entropy, Computacion y S ystemas, 18(2):275-293,2014

[Gu et al., 15] Gu, F., Liu, H.L., Tan, K.C.: A hybrid evolutionary multi-objective optimization algorithm with adaptive multi-fitness assignment, Soft Computing, 19(11):3249-3259, 2015

[Gupta and Kapoor, 94] Gupta, S.C., Kapoor, V.K.: Fundamental of Mathematical Statistics, Sultan Chand Sons, A.S. Printing Press, India, 1994

[Herken, 95] Herken, R.: The Universal Turing Machine – A Half-Century Survey, Springer Verlag, 1995

[Horng et al., 08] Horng, J., Wu, L-C ,Liu, B-J,Kuo, J-L, Zhang, J-J: An expert system to classify microarray gene expression data using gene selection by decision tree, Expert Systems withApplications.36:9072–9081, 2008

[Jerzy and Pearson, 33] Jerzy, N., Pearson, E.S.: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 23:289–337, 1933

[Jing, 14] Jing, S.Y.: A hybrid genetic algorithm for feature subset selection in rough set theory, Soft Computing, 18(7):1373-1382, 2014

[Kerber, 92] Kerber, R.: ChiMerge: Discretization of numeric attributes, Tenth National Conference on Artificial Intelligence, 123–128, 1992

[Kent Ridge] Kent Ridge Bio-medical Data Set Repository: URL http://datam.i2r.a-star.edu.sg/datasets/krbd

[Knowles and Corne, 00] Knowles, J.D.,Corne, D.W.: M-PAES: A memetic algorithm for multi-objective optimization, IEEE Congress on Evolutionary Computation, 1:325–332, 2000

[Kossenkov and Ochs, 10] Kossenkov, A.V., Ochs, M.F.,:Matrix factorization methods applied in microarray data analysis, International Journal of Data Mining and Bioinformatics, 4(1):72–90, 2010

[Kumar et al., 15] Kumar, P.G., Rani, C.,Mahibha, D., Victoire, T.A.A.: Fuzzy-rough-neural-based *f*-information for gene selection and sample classification, Int. J. Data Mining and Bioinformatics, 11(1):31-52, 2015

[Lamba et al., 09] Lamba, J.K., Pounds, S., Cao, X., Downing, J.R.,Campana, D., Ribeiro, R.C.,Pui, C.H., Rubnitz, J.E.: Coding polymorphisms in CD33 and response to gemtuzumabozogamicin in pediatric patients with AML: a pilot study, Leukemia, 23(2):402-404, 2009

[Lan and Vucetic, 11] Lan, L., Vucetic, S.: Improving accuracy of microarray classification by a simple multitask feature selection filter, International Journal of Data Mining and Bioinformatics, 5(2):189–208, 2011

[Lazar et al., 12] Lazar, C.,Taminau, J.,Meganck, S.,Steenhoff, D.,Coletta, A.,Molter, C.,Schaetzen, V., Duque, R.,Bersini, H.,Nowe, A.: A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, IEEE/ACM Transaction Computational Biology, Bioinformatics, 9(4):1106-1119, 2012

[Leung and Hung, 10] Leung, Y., Hung, Y.: A multiple-filter-multipe-wrapper approach to gene selection and microarray data classification,IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(1):108–117, 2010

[Liu, et al., 04] Liu, B., Cui, Q., Jiang, T., Ma, S.: A combinational feature selection and ensemble neural network method for classification of gene expression data, BMC Bioinformatics, 5:136-136, 2004

[Loprinzi, 00] Loprinzi, C.L.: Venlafaxine in management of hot flashes in survivors of breast cancer: a randomized controlled trial,The Lancet, 356(9247):2059–2063, 2000

[Maaranen, 04] Maaranen, H., Miettinen, K.,Makela,M.M.: A Quasi-Random Initial Population for Genetic Algorithms, in: Computers and Mathematics with Applications, Computers & Mathematics with Applications. Elsevier, 47(12):1885–1895, 2004

[Mansouri and Khademi, 15] Mansouri, J.,Khademi, M.: Multiplicative distance: a method to alleviate distance instability for high-dimensional data, Knowledge and Information Systems, 45(3):783-805, 2015

[Mitchell et al., 00] M. Mitchell, J.P Crutchfield, R. Das, Evolving Cellular Automata with Genetic Algorithms: A Review of Recent Work, Proceedings of the First International Conference on Evolutionary Computation and Its Applications (EvCA'96), 1:2000

[Mitra et al., 02] Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):301-312, 2002

[Mohamad and Deris, 05] Mohamad, M.S.,Deris, S.: A Hybrid of Genetic Algorithm and Support Vector Machine forFeatures Selection and Classification of Gene Expression Microarray, International Journal of Computational Intelligence and Applications, 5(1):91–107, 2005

[Mohamad et al., 09] Mohamad, M.S., Omatu, S., Deris, S.,Misman, M.F., Yoshioka,M.: A multi-objective strategy in genetic algorithms for gene selection of gene expression data, Artif Life Robotics, 13:410–413, 2009, DOI 10.1007/s10015-008-0533-5

[Neumann, 96] Neumann, J.V.: Theory of Self-reproducing Automata, Arthur W. Burks (Ed.), University of Illinois Press, Champaign, USA, 1996

[Odibat and Reddy, 14] Odibat, O., Reddy, C.K.: Efficient mining of discriminative co-clusters from gene expression data, Knowledge and Information Systems, 41(3):667-696, 2014

[Olmo et al., 12] Olmo, J.L., Romero, J.R., Ventura, S.: Classification rule mining using ant programming guided by grammar with multiple Pareto fronts, Soft Computing, 16(12): 2143-2163, 2012

[Pal and Mitra, 99] Pal, S.K.,Mitra, S.: Neuro-Fuzzy pattern Recognition: Methods in Soft Computing, New York: Willey, 1999

[Pati and Das, 17] Pati, S.K.,Das, A.K.: Ensemble Classifier Design Selecting Important Genes based on Extracted Features, International Journal of Data mining and Bioinformatics, Inder Science, 19(2):117-149, 2017

[Pati et al., 13] Pati, S.K.,Das, A.K.: Gene Selection using Multi-Objective Genetic Algorithm integrating Cellular Automata and Rough Set Theory, Swarm, Evolutionary and Memetic Computing (SEMCCO), LNCS 8298, Springer, 144-155, 2013

[Park et al., 14] Park, B.J., Kim, W.D., Oh, S.K.,Pedrycz, W.: Fuzzy set-oriented neural networks based on fuzzy polynomial inference and dynamic genetic optimization, Knowledge and Information Systems, 39(1):207-240, 2014

[Pawlak, 98] Pawlak, Z.: Rough set theory and its applications to data analysis, Cybernetics and systems, 29:661-688, 1998

[Pensa et al., 04] Pensa, R.G., Leschi, C., Besson, J.,Boulicaut, J.: Assessment of discretization techniques for relevant pattern discovery from gene expression data, In 4th Workshop on Data Mining in Bioinformatics, 2004

[Price et al., 05] Price, K., Storn, R.M.,Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization, Natural Computing Series, Springer, ISBN: 3540209506, 2005

[Saeys et al., 07] Saeys, Y.,Inza, I.,Larrañaga, P.: A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19):2507-2517, 2007

[Salem et al., 17] Salem, H.,Attiya, G., El-Fishawy, N.: Classification of human cancer diseases by gene expression profiles, Applied Soft Computing, 50:124-134, 2017

[Schaefer, 10] Schaefer, G.: Data mining of gene expression data by fuzzy and hybrid fuzzy methods, IEEE Transaction on Information Technology in Biomedicine, 14(1):23–29, 2010

[Scherz-Shouval, 07] Scherz-Shouval, R., Shvets, E.,Fass, E.,Shorer, H., Gil, L., Elazar, Z.: Reactive oxygenspecies are essential for autophagy and specificallyregulate the activity of atg4, The EMBO Journal, 26:1749–1760, 2007

[Seredynski and Skaruz, 12] Seredynski, F.,Skaruz, J.: Evolutionary Algorithms and Cellular Automata towards Image Reconstruction, Third IEEE International Conference on Emerging Applications of Information Technology (EAIT), 283-286, 2012

[Shelokar et al., 13] Shelokar, P.,Quirin, A.,Cordón, O.: MOSubdue: a Pareto dominance-based multi objective Subdue algorithm for frequent sub graph mining, Knowledge and Information Systems, 34(1):75-108, 2013

[Sikdaret al., 15] Sikdar, U.K.,Ekbal, A.,Saha, S.,Uryupina, O.,Poesio, M.: Differential evolution-based feature selection technique for anaphora resolution, Soft Computing, 19(8): 2149-2161, 2015

[Silver, 05] Silver,R.T.: Anagrelide is effective in treating patients with hydroxyurea-resistant thrombocytosis in patients with chronic myeloid leukemia, Leukemia, 19(3):3943, 2005

[Singh et al., 02] Singh, D.,Febbo, P.G., Ross, K., Jackson, D.G.,Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P.: Gene expression correlates of clinical prostate cancer behavior, Cancer Cell. 1(2):203-209, 2002

[Song et al., 07] Song, L.,Smola, A.,Gretton, A.,Borgwardt, K.M.,Bedo, J.: Supervised feature selection via dependence estimation, In Proceedings of the 24th International Conference on Machine Learning, 823-830, 2007, DOI:10.1145/1273496.1273600

[Souam et al., 13] Souam, F.,Aïtelhadj, A., Baba-Ali, R.: Dual modularity optimization for detecting overlapping communities in bipartite networks, Knowledge and Information Systems. 40(2):455-488, 2013

[Srinivas and Deb, 95] Srinivas, N., Deb, K.: Multi-objective function optimization using non-dominated sorting genetic algorithms, IEEE Transactions on Evolutionary Computation, 2(3):221–248, 1995

[Tan and Gilbert, 03] Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification, Appl. Bioinformatics, 2(3 Suppl):S75-83, 2003

[Wang et al., 05] Wang, Y., Makedon, F.S., Ford, J.C., Pearlman, J.: HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data, Bioinformatics, 21(8):1530–1537, 2005

[Wang et al., 14] Wang, H., Jiao, L., Yao, X.: An Improved Two-Archive Algorithm for Many-Objective Optimization, IEEE Transactions on Evolutionary Computation,doi: http://dx.doi.org/10.1109/TEVC.2014.2350987, 2014

[Wang and Gotoh, 09] Wang, X.,Gotoh, O.: Cancer classification using single genes,Genome Information, 23(1):179-88, 2009

[Wang and Gotoh, 10] Wang, X.,Gotoh, O.: A Robust Gene Selection Method for Microarray-based Cancer Classification, Cancer Informatics, 9:15–30, 2010

[Vatolkin et al., 12] Vatolkin, I., Preuß, M., Rudolph, G.,Eichhoff, M.,Weihs, C.: Multi-objective evolutionary feature selection for instrument recognition in polyphonic audio mixtures, Soft Computing, 16(12):2027-2047, 2012

[Xiao et al., 15] Xiao, J., Xiao, Y., Huang, A., Liu, D., Wang, S.: Feature selection based dynamic transfer ensemble model for customer churn prediction, Knowledge and Information Systems, 43(1):29-51, 2015

[Xiong et al., 01] Xiong, M., Li, W., Zhao, J.,Jin, L.,Boerwinkle, E.: Feature (gene) selection in gene expression-based tumor classification, Journal of molecular Genetics and Metabolism, 73(3):239–247, 2001

[Zhao and Liu, 07] Zhao, Z., Liu, H.: Semi-supervised Feature Selection via Spectral Analysis, Proceedings of SIAM International conference on Data Mining (SDM), 641-646, 2007

[Zhang and Li, 07] Zhang, Q., Li, H.: MOEA/D: a multi-objective evolutionary algorithm based on decomposition, IEEE Transaction on Evolutionary Computation, 11(6):712-731, 2007

[Zhong et al., 01] Zhong, N., Dong, J., Ohsuga, S.: Using rough sets with heuristics for feature selection, Journal of Intelligent Information Systems, 16:199–214,2001