Multiobjective Simulated Annealing-Based Clustering of Tissue Samples for Cancer Diagnosis

Sudipta Acharya, *Student Member, IEEE*, Sriparna Saha, *Member, IEEE*, and Yamini Thadisina

***Abstract*—In the field of pattern recognition, the study of the gene expression profiles of different tissue samples over different experimental conditions has become feasible with the arrival of microarray-based technology. In cancer research, classification of tissue samples is necessary for cancer diagnosis, which can be done with the help of microarray technology. In this paper, we have presented a multiobjective optimization (MOO)-based clustering technique utilizing archived multiobjective simulated annealing(AMOSA) as the underlying optimization strategy for classification of tissue samples from cancer datasets. The presented clustering technique is evaluated for three open source benchmark cancer datasets [Brain tumor dataset, Adult Malignancy, and Small Round Blood Cell Tumors (SRBCT)]. In order to evaluate the quality or goodness of produced clusters, two cluster quality measures viz, adjusted rand index and classification accuracy ($\%CoA$) are calculated. Comparative results of the presented clustering algorithm with ten state-of-the-art existing clustering techniques are shown for three benchmark datasets. Also, we have conducted a statistical significance test called *t*-test to prove the superiority of our presented MOO-based clustering technique over other clustering techniques. Moreover, significant gene markers have been identified and demonstrated visually from the clustering solutions obtained. In the field of cancer subtype prediction, this study can have important impact.**

***Index Terms*—Archived multiobjective simulated annealing (AMOSA), adjusted rand index (ARI), clustering, %CoA index, gene marker, multiobjective optimization (MOO).**

I. Introduction

THE study of the gene expression profiles of different tissue samples over different experimental conditions has become possible with the advent of microarray technology. This study has important role in cancer research. Classification of tissue samples is necessary for cancer diagnosis, which can be done with the help of microarray technology. For classification of different types of tissues, it is necessary to organize microarray dataset as samples versus gene fashion. In microarray datasets, classification of tissue samples into different classes like benign (noncancerous) cells or malignant (cancerous) cells can be termed as binary cancer cell classification. But when tissue samples belong to different cancer sub types, it is called multiple class cancer cell classification problem. It is more complex compared to problem of binary cancer cell classification. In this paper, we have solved the multiple-class cancer cell classification problem.

A microarray gene expression dataset can be represented as a 2-D matrix M containing d number of genes and n number of tissue samples. So the size of the matrix is $n \times d$. Each element e_{kj} of the matrix M represents expression level of kth gene corresponding to jth tissue sample. Any clustering [1] method is aimed to partition n number of data points into K number of clusters or groups. Depending on any similarity or dissimilarity metric, the partitioning is done. Number of clusters, K, is not always known in advance.

Gene markers are those genes which are significantly responsible for distinguishing one tumor class from another hence one tissue sample from another. Using the clustering solutions generated by our presented MOO-based clustering algorithm, we identify relevant gene markers using signal-to-noise (SNR)-based ranking methodology. The relevancy of our identified gene markers has been shown with the help of heatmap.

Archived multiobjective simulated annealing (AMOSA) [2], an existing simulated annealing (SA)-based algorithm for multiobjective optimization (MOO) has been successfully applied in the domain of classification and clustering of data. In [2], it has been experimentally shown that AMOSA performs much better than NSGA-II [3] and some other recent MOO techniques for solving many objective optimization problems. Inspired by these observations, in the current paper we have developed a MOO-based clustering technique using the search capability of AMOSA. The proposed algorithm is applied for partitioning samples from cancer datasets. It optimizes three internal cluster validity indices namely Xie-Beni (XB) index [4], fuzzy C-means (FCM) index [4], Pakhira, Bandyopadhyay, and Maulik (PBM) index [4] simultaneously during clustering process.

A. Related Work

Several previous works exist for classification of tissue samples of cancer datasets. But most of them are either supervised or semisupervised classification [5], [6] techniques. These classification methodologies help in cancer diagnosis by classifying tumor samples as benign or malignant or any other subtypes [7]–[9]. But in many cases, it may be possible that labeled tissue samples are not available. For example microRNA datasets used in [10] or real life gene expression datasets used in [11] are some unlabeled datasets. No labeled data information is provided there. In those cases role of unsupervised classification or clustering comes into play. In this paper, we have proposed some unsupervised classification techniques to classify cancer tissue samples.

Manuscript received September 1, 2014; revised January 1, 2015; accepted February 9, 2015. Date of publication February 20, 2015; date of current version March 3, 2016.

The authors are with the Department of Computer Science and Engineering, Indian Institute of Technology, Patna 800013, India (e-mail: sudiptaacharya.2012@gmail.com; sriparna.saha@gmail.com; yaminireddy6@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2404971

Evolutionary algorithms or genetic algorithms (GAs) [11] are some widely used optimization techniques utilized for unsupervised clustering [4]. A single fitness function or cluster quality measure is used in majority of the already existing GA-based clustering techniques [12] in order to measure the goodness of the encoded partitions. In this paper, we represent the problem of cancer tissue sample clustering as a MOO [3], [4] problem.

In [13], a MOO-based clustering technique is developed using the search capability of NSGA-II (nondominated sorting GA-II) [3] for gene marker identification from cancer tissue samples. Thereafter, a novel method is proposed to combine the solutions of the final Pareto optimal front using the principles of support vector machine (SVM) [14]. Note that in general time complexities of MOO-based clustering techniques are much higher compared to the traditional clustering techniques like *K*-means etc. Thus, the postprocessing technique proposed in [13] further increases the time complexity. It involves the training and testing time of SVM which would increase with the increase in sample size.

B. Scope of This Work

In this paper, we have proposed a new MOO-based clustering technique for cancer cell classification and for identification of relevant gene markers. Our proposed clustering mechanism is applied on three open source cancer datasets, namely Brain tumor, Adult malignancy, and Small Round Blood Cell Tumor (SRBCT). Our proposed technique is also compared with some state-of-the-art clustering algorithms with respect to two clustering performance metrics namely, adjusted rand index (ARI) [13] and percentage classification accuracy (CoA%) [13]. Experimental results support our assumptions that the proposed AMOSA-based clustering technique will perform much better than the existing clustering algorithms. Obtained results prove that the proposed AMOSA-based clustering technique without using any postprocessing mechanism (without using the advantages of SVM) performs much better than MOGASVM approach, which utilizes the advantages of both NSGA-II [3] and SVM, as well as other chosen clustering algorithms. Thus, the proposed approach is highly relevant in the field of cancer subtype classification and gene marker identification where with a less time-complex system we can achieve better results.

The main contributions of the current paper are as follows.

1) The cancer tissue classification problem is treated as a MOO problem. Thereafter, a modern MOO technique based on the concepts of SA, namely AMOSA, a newly developed multiobjective simulated annealing-based optimization technique [2] is utilized to develop a clustering technique to solve the particular problem of cancer tissue classification.
2) Experimental results on three open access datasets show that AMOSA-based clustering technique outperforms all the state-of-the-art clustering techniques including a recently introduced MOO-based clustering technique, MOGASVM utilizing the search capability of NSGA-II [3], a GA-based MOO technique. MOGASVM is a combination of MOO-based clustering along with a postprocessing technique based on the principles of SVM. Here, a SVM-based methodology is developed to combine the solutions of the final Pareto optimal front. But without taking help of any postprocessing technique, the proposed AMOSA-based clustering technique outperforms MOGASVM in terms of existing quality measurements.
3) Thus, the proposed approach provides a way to obtain relevant partitioning of cancer tissues with less complexity.
4) Gene markers identified by the proposed technique are also highly relevant.
5) Note that the proposed technique is based on the search capability of AMOSA. In [2], it has been experimentally proved that AMOSA performs much better than the existing techniques for solving MOO problems. AMOSA often reaches the final Pareto optimal front but the existing MOO-based techniques are not capable of reaching the global Pareto optimal front. This is because in AMOSA there is a positive probability of accepting some bad solutions. This feature helps AMOSA to reach the global Pareto optimal front like the traditional optimization techniques (e.g., GAs and SA). But the existing multiobjective evolutionary algorithms are designed in such a way so that they often get stuck at local Pareto optimal front. This is the reason behind getting improved results by AMOSA-based technique.

The superiority of our proposed clustering algorithm based on AMOSA over other chosen algorithms is illustrated both visually and quantitatively. A statistical significance test is conducted in order to prove the superiority of our presented clustering algorithm over other algorithms. Finally, it has been shown that clustering solutions generated by our proposed MOO-based algorithm can be used to identify relevant gene markers for Brain tumor dataset.

II. AMOSA-Based Multiobjective Clustering Technique

This section describes the proposed MOO-based clustering technique in detail. The proposed technique is a generalized unsupervised clustering algorithm that is capable of partitioning the given dataset based on the available unlabeled data. During clustering, it does not utilize any labeled information. After execution of the proposed technique, a set of optimal solutions is obtained. In order to select a single solution from the final Pareto front, some labeled data (10%) are utilized. Thus, labeled data are not required for execution of the MOO-based clustering technique. Some partial labeled information is required for selecting a single solution, but any other method could have been used here.

The different steps of the proposed algorithm are described below.

A. Input Preprocessing

Before the application of any clustering algorithm, some preprocessing steps are required to be performed on input data to

make it compatible to the proposed algorithm. In our study, we have performed some preprocessing on each dataset. Across all samples, those genes are selected which are having maximum variability. Initially variances of all genes over all samples have been calculated. Next, a sorted list of genes according to their variances is created. Top 200 genes having largest variances are selected from that list. It is desirable that genes having larger variances are more capable of distinguishing tumor samples having different classes than genes having lower variances. In the next preprocessing step, log-transformation is done on the expression values of genes. At the end, each tissue sample is normalized to variance 1 and mean 0.

B. String Representation and Archive Initialization

AMOSA starts its execution after initializing the archive with some alternative random solutions. It utilizes the concept of string to represent each individual solution. To encode the clustering problem in the form of a string, center-based representation is used. Each archive member represents one clustering solution by itself, i.e., one way of partitioning different tissue samples into different clusters. Different archive members have different lengths. Let us assume that our chosen dataset contains n number of samples and each sample has d number of gene expression values. n and d are specific to a dataset. Let us assume that archive member i represents the centroids of K_i clusters and then the array or archive member has length l_i where $l_i = d * K_i$.

Each data point represents a sample of d number of gene expression values, and each cluster centroid c_k is defined by a vector of d expression values.

Each centroid used in string encoding is atomic in nature, i.e., during mutation if we insert one centroid, then all the contained expression values will be inserted. Similarly, if we perform deletion during mutation, all expression values of the chosen centroid will be deleted.

The number of centroids, K_i, encoded in a string i is chosen randomly between two limits $K_{\min}$ and $K_{\max}$. The following equation is chosen to determine this value:

$$K_i = (\text{rand}() \text{mod}(K_{\max} - 1)) + 2. \tag{1}$$

Here, $\text{rand}()$ is a function returning a random integer number and $K_{\max}$ is the upper-limit of the number of clusters. The minimum number of clusters is assumed to be 2. The number of whole clusters present in a particular string/member of archive can therefore vary in the range of 2 to $K_{\max}$. For the initialization step, these K_i cluster centroids represented in a string are some randomly generated samples from the cancer dataset.

C. Assignment of Points and Computation of Objective Functions

After the initialization of archive members with some randomly generated cluster centroids, assignment of n samples or data points (where $n =$ total number of tissue samples in a particular dataset) to different clusters is performed. Next, we compute three cluster quality measures, XB index [4], FCM index [4], PBM index [4], which are used as three objective functions for each solution or string. Thereafter using the search methodology of AMOSA, we simultaneously optimize these three objective functions.

1) *Membership of Tissue Samples to Different Clusters:* In this part, the membership values of tissue samples or data points to different clusters are calculated using the well-known FCM algorithm [4]. To achieve this, for each data point its distance is measured with respect to each cluster center separately. According to FCM algorithm, data points which are more similar to a particular center have more membership values. Thereafter the degree of membership of x_i with respect to cluster center c_j, represented as μ_{ij} is computed as follows:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{K} \left[\frac{d_e(x_i, c_j)}{d_e(x_i, c_l)}\right]^{\frac{1}{m-1}}} \tag{2}$$

where m represents a real number having value greater than 1. Here, we have used Euclidean distance for measuring distances between data point and cluster center.

2) *Objective Functions:* To measure the quality of each solution, three objective functions are calculated. To get some optimized solutions, values of XB [4] and FCM [4] indices corresponding to a particular solution should be minimized and the value of PBM index [4] should be maximized. These three objective functions are optimized simultaneously using the search capability of AMOSA. The mathematical explanations of different objective functions are described in the supplementary file.

D. Search Operators

In order to explore the search space, perturbation operations are used in SA to generate new solution from the current solution. In case of AMOSA based clustering, we have used three different mutation operators. These are defined as follows:

1) *Mutation 1:* This is used to change each cluster center by some small amount. Each cluster center encoded in a string is modified with a random variable which is drawn using a Laplacian distribution, $p(\epsilon) \propto e^{-\frac{|\epsilon - \mu|}{\delta}}$. Scaling factors μ and δ are used to measure magnitudes of mutation. The value of scaling factor δ is generally 1.0. The Laplacian distribution is used to generate a value near the old value, and the old value is replaced with the newly generated value. If a center is selected for mutation, then for all of its dimensions mutation is applied.
2) *Mutation 2:* In this type of mutation, the size of the string is decreased by one. From the string, a cluster center is chosen randomly and then deleted. As each cluster center is considered to be indivisible, so by deleting a cluster center all of its dimensional values are removed.
3) *Mutation 3:* This mutation is used to increase the size of the string by one. This is performed by inserting a new center in the string. Similar to second-type mutation here also each center is considered to be indivisible.

Each string goes through any one of the aforementioned types of mutation operation and generates a new string.

E. Selecting Best Clustering Solution From the Pareto Optimal Front

A set of nondominated solutions is produced by any MOO technique [4] on its final Pareto front. We have plotted final Pareto front obtained by our proposed approach for three datasets. These are shown in the supplementary file. Each of the point on the final Pareto front represents one complete clustering solution. Each of these nondominated solutions corresponds to a complete assignment of all data points of chosen dataset to different clusters. In the absence of additional information, any of those solutions can be selected as the optimal solution. In this approach, we have selected the best solution using external cluster validity index, ARI measure [13]. In case of Cancer tissue classification, we have some limited amount of labeled/supervised information available. We have utilized a semisupervised approach to select a single solution. Our assumption is that for 10% data points, class label information is known. Based on these information, ARI value is computed over these 10% data points for each solution on the Pareto optimal front. The solution having highest ARI value is selected as the best solution. The ARI is calculated according to (3).

III. Experimental Setup

A. Datasets

In this paper, we have chosen three publicly available datasets, Brain Tumor (http://algorithmics.molgen.mpg.de/Static/Supplements/), Adult Malignancy (http://algorithmics.molgen.mpg.de/Static/Supplements/), and SRBCT (http://www.ailab.si/supp/bi-cancer/projections/info/SRBCT.htm.). Brain tumor dataset contains total 42 number of tumor samples and five different classes. Adult Malignancy dataset contains 190 number of tumor samples and total 14 classes. For Small Round Blood Cell Tumor dataset, there are total 63 number of samples and four classes.

B. Evaluation Metrics

We have chosen two metrics for evaluating clustering solutions with respect to actual or true clustering solutions. Those are adjusted rand index or ARI [13] and percentage classification accuracy or $\%CoA$ [13]. Higher values of $\%CoA$ and ARI signify good compatibility between the clustering solution and true clustering solution.

C. Gene Marker Identification

In this section, we have shown how relevant gene markers (i.e., genes which are mostly responsible to distinguish different classes of tumor samples) can be identified from the clustering outcome of AMOSA. In this paper, we have identified gene markers for Brain tumor dataset from its clustering output. For this, at first clustering solutions are collected by executing AMOSA on preprocessed Brain tumor dataset. The dataset is clustered into five tumor classes viz. are MGLIO, RHAB, NCER, PNET, and MD class. In order to identify gene markers from MGLIO class, this problem is treated as a two-class classification problem, where one class is MGLIO itself and other one corresponds to remaining tumor class. Now after taking into consideration both of these classes, a statistic called SNR [15] is calculated for each of the genes. It is defined as

$$\mathrm{SNR} = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \times 100 \tag{3}$$

where μ_i and σ_i , $i \in [1, 2]$, respectively, denote the mean and standard deviation of class i for the corresponding gene. Having large absolute SNR value for a gene indicates that expression value of that gene is high in one class and low in another class. For each gene, its SNR value is calculated and sorted in descending order of their values. From that list, top ten genes are selected (among ten, five are up regulated and other five are down regulated) for each sub type of a particular dataset for example MGLIO subtype. For other subtypes, ten gene markers for each type are selected similarly. It has been observed that the final set of ten selected gene markers changes slightly after each execution of the proposed AMOSA-based clustering. So we have reported those gene markers which have highest frequencies over 20 runs. Frequencies of different genes are also reported.

IV. Results and Discussion

The performance of the proposed MOO-based clustering technique using the concepts of AMOSA is compared with other state-of-the-art clustering algorithms like MOGASVM [13], Expectation Maximization Clustering (EM) [16], K-means clustering [1], hierarchical average linkage clustering [1], SiMM-TS clustering [17], Self Organizing Map (SOM) clustering [18], and consensus clustering [19]. Consensus clustering contains three approaches to ensemble cluster, which are Cluster-based Similarity Partitioning Algorithm (CSPA), Meta-Clustering Algorithm (MCLA), and HyperGraph Partitioning Algorithm (HGPA). These three cluster ensemble techniques combine the clustering solutions which are found by EM, SOM, K-means, and average linkage clustering techniques.

A. Input Parameters

We have executed AMOSA-based clustering technique on three gold standard datasets: Adult Malignancy, Brain tumor, and SRBCT. The proposed algorithm is executed with the following parameter combinations:

$T_{\min} = 0.0001$, $T_{\max} = 100$, $\alpha = 0.9$, HL $= 100$, SL $= 200$, and $iter = 100$.

The parameter values are determined after conducting a thorough sensitivity study. The three main parameters that we have selected by sensitivity study are

1) initial value of temperature ($T_{\max}$),
2) cooling schedule,
3) number of iterations to be performed at each temperature.

According to [2], initial value of the temperature should be so chosen that it allows the SA to perform a random walk over the landscape. As in [2] we have set the initial temperature to achieve an initial acceptance rate of approximately 50% on derogatory proposals. The geometrical cooling schedule α is chosen in the

TABLE I
AVERAGE ARI AND $\%CoA$ SCORES FOR THE ADULT MALIGNANCY, BRAIN TUMOR, AND SRBCT DATA GENERATED BY 20 CONSECUTIVE RUNS OF DIFFERENT ALGORITHMS

Algorithms	Adult-malignancy		Brain Tumor		SRBCT	
	ARI	% CoA	ARI	% CoA	ARI	% CoA
AMOSA(with XB, PBM, FCM indices)	**0.84830**	**97.673120**	**0.755430**	**91.742230**	**0.700913**	**87.7112**
AMOSA (with Specificity, Sensitivity, Classification accuracy)	**0.816124**	**93.4878**	**0.763530**	**92.5962**	**0.498779**	**70.5223**
AMOSA(Using tenfold cross validation)	**0.839784**	**95.24347**	**0.72867**	**90.21434**	**0.5186757**	**78.5354657**
K-means	0.69240	92.54410	0.57640	84.51440	0.3135	70.1903
MOGASVM	0.81720	96.47180	0.71720	88.5150	0.5126	76.6412
SGA	0.74910	95.78580	0.63250	87.14330	0.3198	70.8193
EM	0.72510	94.72940	0.55810	83.14570	0.3376	71.1295
SOM	0.59170	92.8100	0.62140	87.03760	0.3872	71.7845
Avg. Linkage	0.619	93.04370	0.46030	78.28110	0.1021	49.0527
CSPA	0.73310	95.08010	0.60280	85.99840	0.3922	72.0297
SiMM-TS	0.78230	96.01390	0.68920	87.9110	0.4628	74.4853
MCLA	0.73980	95.28130	0.59740	86.45430	0.3902	71.9764
HGPA	0.71920	94.05490	0.52950	83.94160	0.2839	67.4533

TABLE II
p-VALUES PRODUCED BY *t*-TEST COMPARING AMOSA WITH OTHER ALGORITHMS

	p values									
Datasets	K-means	MOGASVM	SGA	EM	SOM	Avg. Linkage	CSPA	SiMM-TS	MCLA	HGPA
Adult Malignancy	3.11E-023	8.3E-011	1.09E-014	1.65E-018	8.79E-023	2.3E-022	2.02E-017	1.41E-013	9.9E-017	2.9E-020
Brain tumor	1.14E-022	9.68E-016	8.42E-019	3.8E-024	5.36E-019	6.45E-028	1.04E-020	3.16E-017	5.33E-020	2.58E-023
SRBCT	2.19E-27	4.36E-264	4.23E-270	1.05E-005	6.92E-005	1.12E-008	1.28E-007	3.05E-006	1.69E-006	1.22E-007

range between 0.5 and 0.99 according to [2]. We have varied the value of α between this range by keeping other parameters constant. Finally, the value of α for which we got the best ARI value of the produced solution is chosen as the value of the cooling rate α. The third factor, i.e., the number of iterations per temperature should be so chosen that the system is sufficiently close to the stationary distribution at that temperature. We have chosen value of $iter = 100$. By further increasing the value of $iter$, the ARI value of resulting solution did not improve. So we kept $iter = 100$.

To get consistent and standard solutions for all the chosen datasets, we have considered the upper mentioned setting of parameters. We have taken results of all ten selected state-of-the-art clustering algorithms for all three datasets. The results are shown in Table I.

B. Clustering Performance

In Table I, we have reported the average $\%CoA$ and average ARI values obtained by all the chosen clustering algorithms for 20 consecutive runs for Adult malignancy, Brain tumor, and SRBCT datasets. For a given partitioning, in order to determine which cluster corresponds to which cancer type, majority vote-based approach [20] is utilized. We need to determine the cancer class of the majority data points of a given cluster. If the actual class label information of all the data points is given, then it is easy to determine this value. For an obtained cluster, we need to find the actual class labels of all the samples. Now calculate the frequency of occurring of each cancer type within this particular cluster. The cancer type with the highest frequency would be the type of that obtained cluster. In case labeled information is not available, we need to take help of some human annotators to assign some class label to each obtained cluster. From Table I, we can conclude that AMOSA performs much better than all of SOO-based clustering algorithms provided in table (all algorithms in Table I except MOGASVM) for clustering tissue samples of Adult Malignancy, Brain tumor, and SRBCT datasets in terms of ARI and $\%CoA$. These results are as desired because MOO is expected to perform better than SOO. But the interesting part of our obtained results is that AMOSA-based MOO clustering algorithm performs better than another MOO-based clustering algorithm, MOGASVM. MOGASVM clustering algorithm is a combination of NSGA-II and SVM [13] (after getting clustering solutions using NSGA-II [3], those are combined using majority voting concept following the principles of SVM [14]). But without taking advantage of SVM, AMOSA solely performs better than MOGASVM.

We have also conducted experiments on the three datasets using other objective functions like Sensitivity [21], Specificity [21], and Classification accuracy. Note that all these three objective functions require true clustering information of the given datasets. Thus, they are not applicable for unlabeled datasets. In general, the proposed approach is an unsupervised clustering technique. It is applied on a given data set to partition it without using any labeled information. Thus, some internal cluster validity indices are used in the current paper as the objective functions.

The ARI and $\%CoA$ index values obtained using the proposed approach where three objective functions, Sensitivity, Specificity, and Classification accuracy are simultaneously

optimized are also shown in Table I. From this table, we can see for Brain tumor dataset ARI and $\%CoA$ values obtained by this approach are better than our original approach (AMOSA with objective functions XB, PBM, and FCM index). But for Adult Malignancy and SRBCT datasets, our original approach provides better results compared to the approach which considers sensitivity, specificity, and classification accuracy as objective functions.

We have also obtained results after conducting tenfold cross-validation test for computing the objective function values on three datasets. Here, each of the datasets is divided into ten subsets where all the classes are present in equal proportions. At a time, nine of the subsets are merged together, and the original version of the proposed algorithm is applied on it. For each of the obtained solutions on the final Pareto optimal front, class labels are assigned to the data points of the rest subset using the minimum center distance-based criterion. The ARI value is calculated for the points on this subset. The solution having the maximum value of ARI is selected as the best solution for this subset. The aforementioned approach is repeated ten times varying the subset as the test set. For each of the run, cluster quality measures are calculated for the whole dataset. The average values over ten runs for all the datasets are also reported in Table I.

For adult malignancy dataset, AMOSA clustering technique (Considering XB, PBM, and FCM index as objective functions) attains improvements of 3.81%, 1.25% for ARI and $\%CoA$ metrics, respectively, over MOGASVM clustering technique. For brain tumor dataset, these improvements are 5.33% and 3.65% for ARI and $\%CoA$ metrics, respectively, over MOGASVM. For SRBCT dataset, these improvements are 36.7% and 14.44%, respectively. It is very clear from the Table I, that for all three datasets AMOSA outperforms all other state-of-the-art clustering algorithms in terms of ARI and $\%CoA$ values. Results on all the datasets show that the proposed AMOSA-based clustering technique performs much better than MOGASVM in terms of both the performance measurements. Note that MOGASVM utilizes the advantages of both multiobjective GA-based optimization technique (MOGA) and SVM [14]. It uses MOGA to solve the clustering problem and then utilizes the principles of SVM to combine the solutions on the final Pareto optimal front. Thus, the time complexity of MOGASVM involves the time requirement of MOGA and training and testing time of SVM [14]. But experimental results show that without using any extra postprocessing procedure, the proposed AMOSA-based clustering technique produces better results than MOGASVM. Thus, the proposed clustering technique is less time-complex than MOGASVM. It avoids the training and testing time of SVM but still achieves better results than MOGASVM. This proves the effectiveness of the proposed algorithm. Thus, the proposed AMOSA-based clustering technique is more relevant for clustering cancer tissue samples as compared to existing techniques. Experimental results on three bench mark datasets show that it provides better results with reasonable time.

In order to measure the performance of our chosen algorithm, we have also chosen one internal cluster validity index known as Silhouette index [22]. It can vary from -1 to 1 and a good clustering solution is having higher Silhouette index value. Our proposed approach has obtained Silhouette index values of 0.5679, 0.5832, and 0.72380, respectively, for three different datasets, Adult Malignancy, Brain tumor, and SRBCT.

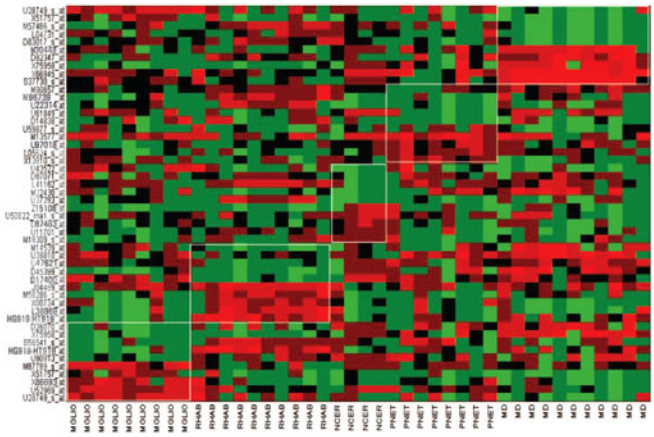

Fig. 1. Heatmap of the expression levels of the most frequently selected top ten gene markers for each tumor subtype in the Brain tumor data.

C. Statistical Significance Test

In Table I, it has been shown that the average $\%CoA$ values obtained by AMOSA are better than those obtained by all the chosen state-of-the-art algorithms (1) MOGASVM, 2) K-means, 3) EM, 4) SGA, 5) average linkage, 6) SOM, 7) SiMM-TS, 8) CSPA, 9) HGPA, 10) MCLA) for all of three chosen datasets. To prove superiority of AMOSA statistically, a statistical significance test is conducted (also known as *t*-test) at 5% significance level. Eleven groups, corresponding to the 11 algorithms (1) AMOSA, 2) MOGASVM, 3) K-means, 4) EM, 5) SGA, 6) average linkage, 7) SOM, 8) SiMM-TS, 9) CSPA, 10) HGPA, 11) MCLA) are created for all datasets.

Now between each two groups (a group corresponding to AMOSA and another group corresponding to any algorithm among ten selected algorithms), the p-values produced by *t*-test are reported in Table II. As null hypothesis, we assume that there is insignificant difference between mean values of two groups. According to alternative hypothesis, there are significant differences in the mean values of two groups. It can be seen that all of the p-values in Table II are less than 0.05 (5% significance level). It strongly indicates that the null hypothesis is wrong, and the better mean values of the $\%CoA$ index produced by AMOSA are statistically significant and have not occurred by chance.

D. Gene Markers for Brain Tumor Dataset

In Fig. 1, the identified relevant gene markers are shown by heatmap of sample versus gene matrix of Brain tumor dataset. In that figure, each row represents each one of the identified gene markers, and each column represents class name of the sample. So there are total 50 rows corresponding to 50 identified gene markers. Each cell in heatmap represents expression level of the

corresponding gene marker in terms of color. High expression level is represented as red color, while green represents low expression level and absence of differential expression values are represented by black. From the figure, it is clear that our selected gene markers have either high expression levels (up regulated) or low expression levels (low regulated) over all samples of respective tumor class.

In the supplementary file, we have reported the top ten gene markers along with their descriptions, frequencies of selection of each gene over 20 runs of AMOSA, and up/down regulation states for the MGLIO, RHAB, NCER, PNET, and MD tumor classes, respectively.

We have also performed some literature search in order to prove the significance of gene markers. We have validated many of our obtained gene markers for each class of brain tumor dataset with different existing literatures. For example, in case of Brain tumor dataset, genes M30448_s and M96739_at are reported to belong to MD tumor class in [23] as similar to our obtained results. Similarly for $PNET$ class, gene markers U97018_at and U22314_s_at are reported in [23]. In papers [24] and [25], gene markers X86693_at, HG919-HT919_at are reported for $MGLIO$ class. For $RHAB$ class gene markers U47621_at, D17400_at and L38969_at are reported in paper [24]. In [26], gene markers D87463_at and Z15108_at are reported for $NCER$ class. These findings are similar to the observations derived from our present study.

V. Conclusion

In this paper, we have formulated the problem of clustering of cancer tissue samples of cancer dataset as a MOO problem and solved it with the help of a MOO-based clustering approach. Three cluster quality measures, XB, PBM, FCM indices are used as three objective functions. The performance of the MOO-based clustering technique is evaluated on three gold standard datasets, Brain tumor, Adult Malignancy, and Small Round Blood Cell Tumors. Results show that the proposed algorithm not only outperforms the existing single objective clustering techniques but also achieves better results than a recently developed MOO-based clustering technique namely MOGASVM which combines the fruitfulness of both MOO and SVM. It uses a modern multiobjective evolutionary algorithm, NSGA-II as the underlying optimization strategy and finally nondominated solutions on the final Pareto optimal front are combined using the advantages of SVM. But the use of AMOSA as the background optimization strategy of the proposed clustering technique helps it to attain better results compared to MOGASVM without using any further postprocessing mechanism. The experimental results conclude the effectiveness of the proposed clustering technique which finds better solutions within reasonable time frame. We have also identified the relevant gene markers from the clustering output and relevancy of them is shown visually with the help of Heatmap.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[2] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," *IEEE Trans. Evol. Comput.*, vol. 12, no. 3, pp. 269–283, Jun. 2008.

[3] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[4] S. Bandyopadhyay and S. Saha, "Unsupervised classification: Similarity measures," in *Classical Metaheuristic Approaches, Appl.* New York, NY, USA: Springer, 2012.

[5] L. An and R. W. Doerge, "*Dynamic Clustering of Gene Expression*, *ISRN Bioinformatics*, vol. 2012, art. no. 537217, pp. 1–12, 2012, doi: 10.5402/2012/537217.

[6] Y. Wang and Y. Pan, "Semi-supervised consensus clustering for gene expression data analysis, *BioData Mining* vol. 7.1, pp. 1–13, 2014.

[7] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt,, "Distinct types of diffuse large b-cell lymphomas identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[8] K. Y. Yeung and R. E. Bumgarner, "Multiclass classification of microarray data with repeated measurements: Application to cancer," *Genome Biol.*, vol. 4, no. 12, R83-R83, 2003.

[9] M. C. P. de Souto, I. G. Costa, D. SA de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: A comparative study," *BMC Bioinformatics*, vol. 9, no. 1, pp. 497, 2008.

[10] S. Paul and P. Maji, "City block distance for identification of co-expressed MicroRNAs," *SEMCCO*, no. 2, pp. 387–396, 2013.

[11] S. Saha, A. Ekbal, K. Gupta, and S. Bandyopadhyay, "Gene expression data clustering using a multiobjective symmetry based clustering technique," *Comput. Biol. Med.* vol. 43.11, pp. 1965–1977, 2013.

[12] S. C. Dinger, M. A. Van Wyk, S. Carmona, and D. M. Rubin, "Clustering gene expression data using a diffraction inspired framework," *Biomed. Eng. Online*, vol. 11, no. 1, p. 85, 2012.

[13] A. Mukhopadhyay, S. Bandyopadhyay, and U. Maulik, "Multi-Class clustering of cancer subtypes through SVM based ensemble of pareto-optimal solutions for gene marker identification," *PLoS ONE* vol. 5, no. 11, p. e13803 2010.

[14] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, pp. 264–323, 1999.

[17] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *Bioinformatics*, vol. 23, pp. 2859–2865, 2007.

[18] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Nat. Acad. Sci.*, USA, vol. 96, pp. 2907–2912, 1999.

[19] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[20] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Trans. Syst., Man, Cybern., A, Syst. Humans*, vol. 27, no. 5, pp. 553–568, Sep. 1997.

[21] R. Parikh, A. Mathai, S. Parikh, and G. C. Sekhar, "Understanding and using sensitivity, specificity and predictive values," *Indian J. Ophthalmol.*, vol. 56.1, pp. 45–50, 2008.

[22] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[23] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[24] Y. S. Tsai, C. T. Lin, G. C. Tseng, I. F. Chung, and N. R. Pal, "Discovery of dominant and dormant genes from expression data using a novel generalization of SNR for multi-class problems, *BMC Bioinformatics*, vol. 9, no. 1, p. 425, 2008.

[25] Y. Wang, "Integrative methods for gene data analysis and knowledge discovery on the case study of KEDRIs brain gene ontology," Ph. D. dissertation, School Comput. Mathemat. Sci., Auckland Univ. Technol., Auckland, New Zealand, 2008.

[26] T. Golub, E. S. Lander, S. Pomeroy, and P. Tamayo, "Brain tumor diagnosis and outcome prediction," Google Patents number: WO 2002061144 A2, 2002.

[27] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

Authors' photographs and biographies not available at the time of publication.