Cancer tissue sample classification using point symmetry-based clustering algorithm

Sudipta Acharya* and Sriparna Saha

Department of Computer Science and Engineering,
IIT Patna, India
Email: sudiptaacharya.2012@gmail.com
Email: sriparna.saha@gmail.com
*Corresponding author

Abstract: Clustering or unsupervised classification techniques can be used to solve different types of classification problems of different domains. Symmetry is an important property for any real life object. Therefore, symmetry-based distance measurements play some important roles in identifying some patterns or clusters of real life datasets. In this paper, inspired by the symmetric property, we have proposed a point symmetry-based clustering algorithm which has been used to identify clusters of tissue samples from some real life cancer datasets. Our proposed algorithm is also multi-objective-optimisation (MOO) based, i.e., optimises more than one objectives simultaneously. We have also shown the superiority of our proposed algorithm with respect to some state-of-the-art clustering algorithms.

Keywords: multi-objective-optimisation; MOO; clustering; AMOSA; gene marker; point symmetry-based distance; ARI index; %CoA index.

Reference to this paper should be made as follows: Acharya, S. and Saha, S. (2018) 'Cancer tissue sample classification using point symmetry-based clustering algorithm', *Int. J. Humanitarian Technology*, Vol. 1, No. 1, pp.19–39.

Biographical notes: Sudipta Acharya received her MTech in Information Technology from the National Institute of Technology Dugrapur, in 2013. She is currently working towards her PhD degree at the Indian Institute of Technology, Patna. She is author and co-author of 14 papers. Her research interests include multiobjective optimisation, pattern recognition, evolutionary algorithms and data mining.

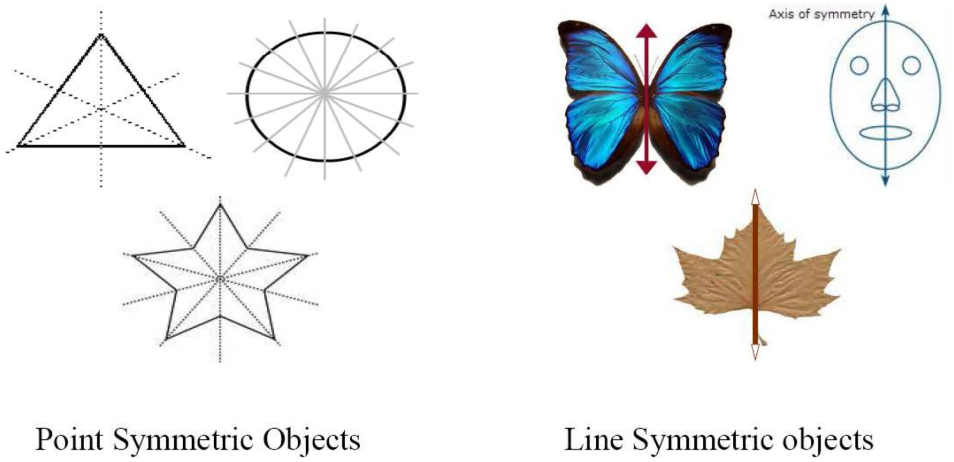
Sriparna Saha received her MTech and PhD degrees in Computer Science from the Indian Statistical Institute Kolkata, Kolkata, India, in 2005 and 2011, respectively. She is currently a faculty member in the Department of CSE, IIT Patna, Patna, India. She is the author of a book published by Springer-Verlag. She has authored or co-authored more than 100 papers. Her current research interests include pattern recognition, multiobjective optimisation and biomedical information extraction. She is the recipient of the Lt Rashi Roy Memorial Gold Medal from the Indian Statistical Institute for outstanding performance in MTech (Computer Science). She is the recipient of the Google India Women in Engineering Award, 2008.

This paper is a revised and expanded version of a paper entitled 'Multi-objective clustering of tissue samples for cancer diagnosis' presented at ICACCI 2014, Greater Noida, Delhi, 24–27 September 2014.

1 Introduction

In the field of pattern recognition, clustering (Everitt et al., 2001) has several applications for solving different real life problems. One of the real life problems where clustering methodologies can play some important role is the classification of cancer tissue samples from different cancer datasets. The study of different gene expression profiles by performing tissue sample classification has become possible with the help of microarray technology.

Figure 1 Point symmetric and line symmetric objects (see online version for colours)



According to this technology, cancer datasets are represented in the form of matrix (in gene versus sample fashion) in order to classify tissue samples. In this paper, we have solved a multi-class cancer tissue sample classification problem with the help of a clustering algorithm.

In any clustering algorithm, for finding similarity between two data points or assigning different data points in different clusters, distance measurements play very important role. In most of the existing literatures (Acharya et al., 2014; Mukhopadhyay et al., 2010; Paul and Maji, 2013) for measuring similarity between two given data points, Euclidean distance or city block distance (Paul and Maji, 2013) is in general used. But many of the real life objects like human face, leaves of trees, etc. are symmetric in nature and this is also true for many real life datasets. So, to determine some patterns or clusters from real life datasets, symmetry-based similarity measurement is the most suitable in order to find similarity between two data points.

Symmetry measurement can be of two types: point symmetry (PS) and line symmetry (LS). The different real life objects which are point or line symmetric are shown in Figure 1. Inspired by these two types of symmetry, PS-based distance (Bandyopadhyay and Saha, 2007) and LS-based distance (Saha and Bandyopadhyay, 2011) were developed. It has been shown in the literature that symmetry-based distance measurements perform better than the traditional distances like Euclidean distance or city-block distance, etc. Inspired by these observations, in the current work, we have used symmetry-based similarity measurement to develop a multi-objective clustering technique to solve the cancer tissue sample classification problem from three real life benchmark cancer datasets, brain tumour, adult malignancy and SRBCT.

Our proposed technique is also compared with 11 state-of-the-art clustering algorithms viz. K-means clustering (Jain and Dubes, 1988), GA-based single objective clustering that optimises an objective function which is a combination of cluster separation and compactness (SGA), expectation maximisation (EM) clustering (Jain et al., 1999), hierarchical average linkage clustering (Jain andDubes, 1988), SiMM-TS (Bandyopadhyay et. al., 2007), consensus clustering (Strehl and Ghosh, 2002), self-organising map (SOM) clustering (Tamayo et al., 1999) and a recently developed multi-objective-optimisation (MOO)-based clustering technique called MOGASVM (Mukhopadhyay et al., 2010).

We have also compared our work with (Acharya et al., 2014) where a Euclidean distance-based multi-objective clustering technique utilising the optimisation strategy, archived multi-objective simulated annealing (AMOSA) (Bandyopadhyay et al., 2008) is proposed. We have shown in the current work that the incorporation of PS-based distance in the clustering algorithm in place of Euclidean distance helps to improve the performance. We have performed clustering on three micro array datasets, brain tumour, adult malignancy and SRBCT datasets and shown that our newly proposed algorithm performs better than clustering algorithm proposed in Acharya et al. (2014) with respect to two clustering performance metrics namely, adjusted rand index (ARI) (Mukhopadhyay et al., 2010) and classification accuracy (%CoA) (Mukhopadhyay et al., 2010). In the current work, PS-based distance is used for assigning points to different clusters as well as for computing the values of objective functions.

Gene markers are those genes which are significantly responsible for distinguishing one tumour 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. 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 tumour dataset.

The main contributions of the current paper are as follows:

- The cancer tissue classification problem is treated as a MOO problem. Thereafter, a modern MOO technique based on the concepts of simulated annealing (SA), namely AMOSA (Bandyopadhyay et al., 2008) in combination with PS-based distance, is utilised to develop a clustering technique to solve the particular problem of cancer tissue classification.
- Experimental results on three open access datasets show that PS-based clustering technique outperforms all the state-of-the-art clustering techniques including a recently introduced MOO with Euclidean distance-based clustering technique utilising the concepts of AMOSA (Acharya et. al, 2014), MOGASVM utilising the search capability of NSGA-II (Mukhopadhyay et. al., 2010), a genetic algorithm (GA)-based MOO technique. MOGASVM is a combination of MOO-based clustering along with a post processing 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 post processing technique, the proposed AMOSA-based clustering technique outperforms MOGASVM in terms of existing quality measurements.

- The proposed approach provides a way to obtain relevant partitioning of cancer tissues with less complexity.
- Gene markers identified by the proposed technique are also highly relevant.

2 Background

In this section, at first, we have discussed about the existing PS-based distance. After that we have thoroughly discussed about some existing works in the field of cancer classification.

2.1 PS-based distance

The PS distance or PS-based distance (Bandyopadhyay and Saha, 2007) $d_{ps}(x, c_i)$ associated with point x with respect to a cluster centre c_i of cluster C_i is described in this section. Let the dataset contain all distinct points, and let x is a point. The reflected or symmetrical point of x with respect to a particular cluster centre c_i is $2 * c_i - x$ and is denoted by x^*. If *knear* is the number of unique nearest neighbours of x^* according to Euclidean distances d_i, $i = 1, 2, \ldots, knear$, Then,

$$d_{ps}\left(x_i, c_i\right) = d_{sym}\left(x_i, c_i\right) \times d_e\left(x_i, c_i\right) \tag{1}$$

where

$$d_{sym}\left(x_i, c_i\right) = \frac{\sum_{i=1}^{knear} d_i}{knear} \tag{2}$$

Between points x and c_i, the Euclidean distance is measured by $d_e(x, c_i)$. In equation (2), *knear* should not be chosen as equal to 1, because if x^* exists in the dataset then the value of $d_{ps}(x, c_i) = 0$, and then there will be no impact of Euclidean distance. Again if the value of *knear* is large then also it will not be suitable because with respect to a particular cluster centre it may overestimate the symmetry amount of a point. So here we choose *knear* = 2. Computation complexity of $d_{ps}(x, c_i)$ is $O(n)$. Hence, for n points and K clusters, the complexity of assigning points to different clusters is $O(n^2K)$. In order to reduce the computational complexity of computing the PS-based distance, Kd-tree-based approximate nearest neighbour search approach is also used in Bandyopadhyay and Saha (2007).

2.2 Related work

Several previous works exist for classification of tissue samples of cancer datasets. But most of them are either supervised or semi-supervised classification (An and Doerge, 2012; Wang and Pan, 2014) techniques. These classification methodologies help in cancer diagnosis by classifying tumour samples as benign or malignant or any other sub types (Alizadeh et al., 2000; Yeung and Bumgarner, 2003; de Souto et al., 2008). But in many cases, it may be possible that labelled tissue samples are not available. For example, microRNA datasets used in Paul and Maji (2013) or real life gene expression datasets used in Saha et al. (2013) are some unlabeled datasets. No labelled data information is provided there. In those cases, role of unsupervised classification or clustering comes into play. In this article, we have proposed some unsupervised classification techniques to classify cancer tissue samples.

Evolutionary algorithms or GAs (Goldberg, 1989) are some widely used optimisation techniques utilised for unsupervised clustering (Maulik and Bandyopadhyay, 2000). A single fitness function or cluster quality measure is used in majority of the already existing GA-based clustering techniques (Dinger et al., 2012) in order to measure the goodness of the encoded partitions. In this article, we represent the problem of cancer tissue sample clustering as a MOO problem.

In Mukhopadhyay et al. (2010), a MOO-based clustering technique is developed using the search capability of non-dominated sorting genetic algorithm-II (NSGA-II) (Deb et al., 2002) 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) (Everitt et al., 2001). 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 post processing technique proposed in Mukhopadhyay et al. (2010) further increases the time complexity. It involves the training and testing time of SVM which would increase with the increase in sample size.

Recently, in Acharya et al. (2014), authors have proposed a MOO-based unsupervised clustering technique for classifying tissue samples. In this paper, they have used Euclidean distance in order to assign different data points to different clusters. But they have not considered symmetric nature of clusters present in real life datasets.

3 PS-based clustering technique

In this section, we have described the MOO-based clustering algorithm applied for classification of cancer tissue samples. In our chosen algorithm, we have used PS-based distance for measuring similarity between two data points. The proposed technique is a generalised unsupervised clustering algorithm which is capable of partitioning the given dataset based on the available unlabeled data. As the underline optimisation technique, we have chosen a SA-based MOO technique, namely archived multi-objective SA or AMOSA (Bandyopadhyay et al., 2008). To solve difficult optimisation problems, SA, a technique based on principles of statistical mechanics (Kirkpatrick, 1984), is utilised. SA has been widely used to solve several single objective combinatorial optimisation problems. But there are very few attempts in solving MOO problems using the search capability of SA. This is because of the difficulty in integrating multiple objective functions in the form of the acceptance probability in case of SA. In order to overcome this difficulty, an efficient multi-objective version of SA which is called AMOSA is developed in Bandyopadhyay et al. (2008). It has been shown in (Bandyopadhyay et al. (2008) that AMOSA performs better than the existing MOO techniques using the concepts of evolutionary computations for different benchmark test problems. Inspired by this, we have used AMOSA as the underline optimisation technique in the current paper.

3.1 Input pre-processing

Before application of any clustering algorithm, some pre-processing steps are required to be performed on input data to make it compatible to the proposed algorithm. In our work, we have performed some pre-processing 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 tumour samples having different classes than genes having lower variances. In the next pre-processing step, log transformation is done on the expression values of genes. At the end, each tissue sample is normalised to variance 1 and mean 0.

3.2 String representation of solution

AMOSA starts its execution after initialising the archive with some alternative random solutions. It utilises the concept of string to represent each individual solution. To encode the clustering problem in the form of a string, centre-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 the 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:

$$K_i = \left(rand() \bmod \left(K_{\max} - 1 \right) \right) + 2 \tag{3}$$

Here, *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 initialisation step, these K_i cluster centroids represented in a string are some randomly generated samples from the cancer dataset.

3.3 Assignment of points and computation of objective functions

After the initialisation 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 two cluster quality measures, PS-based XB index (Bandyopadhyay and Saha, 2012), and Sym index (Bandyopadhyay and Saha, 2012) which are used as two objective functions for each solution or string. Thereafter, using the search methodology of AMOSA, we simultaneously optimise the two 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 fuzzy C-means algorithm (Dembele and Kastner, 2003). To achieve this, for each data point its distance is measured with respect to each cluster centre separately using the PS distance measurement according to equation (1). A tissue sample or data point is placed to that cluster with respect to whose centroid the data point is having the minimum distance.

2 Objective functions:

   In our chosen MOO-based clustering technique, we have used two objective functions based on PS-based distance. Those are PS-based Xie Beni Index (Bandyopadhyay and Saha, 2012) and Sym index (Bandyopadhyay and Saha, 2012). All these objectives are functions of cluster compactness and separation. For computing cluster compactness we have used PS-based distance measurement [according to equation (1)]. The functional definitions of our chosen objective functions are given below,

   - *PS-based XB index*: It follows the definition of the popular Euclidean distance-based XB index (Bandyopadhyay and Saha, 2012). It is defined as follows:

$$PSym-XB(K)=\frac{\sum_{i=1}^{K}\left[\sum_{x\in i}d_{ps}^{*2}\left(x,c_i\right)\right]}{n\left[\min_{i,k=1,2,\ldots,K,i\neq k}d_e^2\left(c_i,c_k\right)\right]} \tag{4}$$

     $d_{ps}^{*}(x,c_i)$ is computed by equation (1). The reflected point x^* of the point x with respect to centroid of i^{th} cluster c_i and x should belong to the $^{i\text{th}}$ cluster. Most accurate partition (i.e., optimal K_{opt}) can be obtained by observing the minimal value of $PSym-XB$ index over $K = 2, 3, \ldots, K_{\max}$.

   - *Sym index*: The PS-based distance is used to define a cluster validity index (Bandyopadhyay and Saha, 2012) which measures overall average symmetry with respect to cluster centroids. It is defined as,

$$Sym(K)=\left(\frac{1}{K}\times\frac{1}{E_K}\times D_K\right) \tag{5}$$

     where

$$E_K=\sum_{k=1}^{K}\sum_{j=1}^{n_k}d_{ps}\left(c_k,x_j^k\right)$$

     represents total compactness of the partitioning in terms of symmetry. D_K is called the separation of the crisp K-partitioning. Where D_K is the maximum distance between any two cluster centroids, i.e.,

$D_K = \max_{i,j=1}^{K} \left\| c_i - c_j \right\|$

The objective is to maximise this index in order to obtain the actual number of clusters and the proper partitioning.

3.4 Search operators

In order to explore the search space, perturbation operations are used in AMOSA to generate new solutions from the current solution. In case of AMOSA-based clustering, we have used three different mutation operators. These are defined below: A clustering solution can be changed in three different ways,

- By increasing the number of encoded clusters in a solution by one. This is done by randomly selecting a point from the dataset as the new cluster centre and then inserting this in the solution.
- By decreasing the number of encoded clusters in a solution by one. This is done by deleting a randomly selected cluster centre from the given solution.
- By modifying the existing cluster centres encoded in the solution. By using Laplacian distribution, we have randomly selected some values near the old values of cluster centres and then updated the existing centres.

The proposed three mutation operators are described below:

1 *Mutation 1*: This is used to change each cluster centre by some small amount. Each cluster centre 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 the 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 centre is selected for mutation then for all of its dimensions mutation is applied.

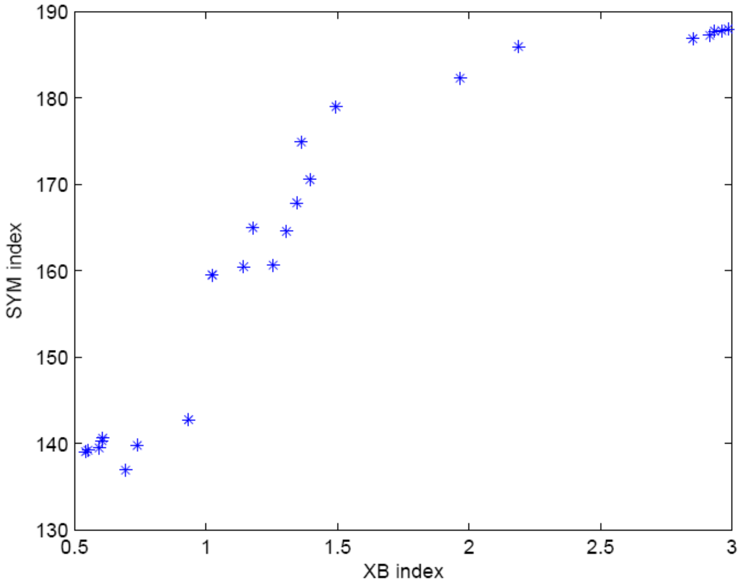
2 *Mutation 2*: In this type of mutation the size of the string is decreased by one. From the string, a cluster centre is chosen randomly and then deleted. As each cluster centre is considered to be indivisible, so by deleting a cluster centre all of its dimensional values are removed.

3 *Mutation 3*: This mutation is used to increase the size of the string by one. This is performed by inserting a new centre in the string. Similar to 2nd type mutation here also each centre is considered to be indivisible.

3.5 Selecting best solution from the set of non-dominating solutions

A set of non-dominated solutions is produced by any MOO technique (Deb, 2009) on its Pareto front. We have plotted final Pareto front obtained by our proposed approach for SRBCT dataset. This is shown in Figure 2. Each point on the final Pareto front represents one complete clustering solution. Each of these non-dominated 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 the external cluster validity index, ARI measure (Mukhopadhyay et al., 2010). The solution having the highest ARI value is selected as the best solution.

Figure 2 Pareto front obtained by the AMOSA-based approach for SRBCT dataset (see online version for colours)



4 Experimental setup

4.1 Datasets we choose

In this article, we have chosen three publicly available datasets, brain tumour (http://algorithmics.molgen.mpg.de/Static/Supplements/), adult malignancy (http://algorithmics.molgen.mpg.de/Static/Supplements/) and small round blood cell tumour (SRBCT) (http://www.ailab.si/supp/bi-cancer/projections/info/SRBCT.htm.). Brain tumour dataset contains total 42 number of tumour samples and five different classes. Adult malignancy dataset contains 190 number of tumour samples and total 14 classes. For SRBCT dataset, there are total 63 number of samples and four classes.

4.2 Evaluation metrics

We have chosen two metrics for evaluating clustering solutions with respect to actual or true clustering solutions. Those are *ARI* (Mukhopadhyay et al., 2010) and %*CoA*

(Mukhopadhyay et al., 2010). Higher values of these two indices indicate better compatibility of clustering output with actual or true clustering output.

4.3 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 tumour samples) can be identified from the clustering outcome of PS-based AMOSA. In this paper, we have identified gene markers for brain tumour dataset from its clustering output. For this, at first, clustering solutions are collected by executing AMOSA on preprocessed Brain tumour dataset. The dataset is clustered into five tumour 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 tumour classes. Now after taking into consideration both of these classes, a statistic called SNR ratio (Golub et al., 1999) is calculated for each of the genes. It is defined below:

$$SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \times 100 \tag{6}$$

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 10 genes are selected (among 10, 5 are up regulated and other 5 are down regulated) for each subtype of a particular dataset for example MGLIO subtype. For other subtypes, 10 gene markers for each type are selected similarly. It has been observed that the final set of 10 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.

5 Results and discussions

The performance of the proposed PS-based clustering technique using the concepts of AMOSA is compared with other state-of-the-art clustering algorithms like MOGASVM (Mukhopadhyay et al., 2010), EM clustering (Jain et al., 1999), K-means clustering (Jain and Dubes, 1988), hierarchical average linkage clustering (Jain and Dubes, 1988), SiMM-TS clustering (Bandyopadhyay et al., 2007), SOM clustering (Tamayo et al., 1999) and consensus clustering (Strehl and Ghosh, 2002). 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. We have also compared performance of our proposed algorithm with the clustering performance of Euclidean distance-based AMOSA (Acharya et al., 2014).

5.1 Input parameters

We have executed AMOSA-based clustering technique on three gold standard datasets: adult malignancy, brain tumour and SRBCT. The proposed algorithm is executed with the following parameter combinations:

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

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 Bandyopadhyay et al. (2008), initial value of the temperature should be so chosen that it allows the SA to perform a random walk over the landscape. As in Bandyopadhyay et al. (2008), we have set the initial temperature to achieve an initial acceptance rate of approximately 50%. The geometrical cooling schedule α is chosen in the range between 0.5 and 0.99 according to Bandyopadhyay et al. (2008). 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 11 selected state-of-the-art clustering algorithms for all three data-sets. The results are shown in Table 1.

5.2 Clustering performance

In Table 1, 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 tumour and SRBCT datasets.

From Table 1, we can conclude that PS-based AMOSA performs much better than all of SOO-based clustering algorithms provided in Table 1 (all algorithms in Table 1 except

MOGASVM and AMOSA with Euclidean distance) for clustering tissue samples of Brain tumour and SRBCT datasets in terms of *ARI* and %*CoA*. For adult malignancy dataset, the proposed approach attains slightly less value with respect to *ARI* and %*CoA* scores compared to Euclidean distance-based AMOSA algorithm. Results are as desired because MOO is expected to perform better than SOO as well as PS-based distance measures are expected to perform well in order to identify clusters of real life datasets. But the interesting part of our obtained results is that our clustering algorithm performs better than other MOO-based clustering algorithms, MOGASVM and Euclidean distance-based AMOSA. MOGASVM clustering algorithm (Mukhopadhyay et al., 2010) is a combination of NSGA-II and SVM (after getting clustering solutions using NSGA-II, those are combined using majority voting concept following the principles of SVM). But without taking advantage of SVM, AMOSA solely performs better than MOGASVM. Also, Euclidean distance-based AMOSA (Bandyopadhyay et al., 2008) fails to get better results than our proposed PS-based AMOSA. It is because PS-based distance is more capable than Euclidean distance to identify clusters from real life datasets.

Table 1 The average *ARI* and %*CoA* scores for the adult malignancy, brain tumour and SRBCT datasets generated by 20 consecutive runs of different algorithms

Algorithms	*Adult-ARI*	*Malignancy % CoA*	*Brain tumour ARI*	*% CoA*	*SRBARI*	*CT % CoA*
AMOSA (with PS distance)	*0.8411*	*96.978*	*0.76594*	*92.2833*	*0.7237*	*89.5643*
AMOSA (with Euclidean distance)	*0.84830*	*97.673120*	*0.755430*	*91.742230*	*0.700913*	*87.7112*
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

Results on all the datasets show that the proposed clustering technique performs much better than Euclidean distance-based AMOSA in terms of both the performance measurements.

5.3 Statistical significance test

In Table 1, 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 (MOGASVM, K-means, EM, SGA, average linkage, SOM, SiMM-TS, CSPA, HGPA, MCLA and Euclidean-based AMOSA) for all of three chosen datasets.

The improved performance by the proposed techniques is due to the following reasons:

1 better searching capability of AMOSA which utilises SA-based search technique.

2 utilisation of symmetry-based distance in AMOSA instead of Euclidean distance.

3 consideration of multiple objectives instead of single objective for clustering.

To prove the superiority of AMOSA statistically, a statistical significance test is conducted (also known as t-test) at 5% significance level. Twelve groups, corresponding to the 12 algorithms (PS-based AMOSA, MOGASVM, K-means, EM, SGA, average linkage, SOM, SiMM-TS, CSPA, HGPA, MCLA, and Euclidean-based AMOSA) are created for all datasets.

Now between each two groups (a group corresponding to AMOSA and another group corresponding to any algorithm among 11 selected algorithms) the p-values produced by t-test are reported in Table 2. As null hypothesis we assume that there are insignificant differences 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 2 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.

Table 2 The p-values produced by t-test comparing PS-based AMOSA with other algorithms

	p-values					
Datasets	*Euclidean-AMOSA*	*K-means*	*MOGASVM*	*SGA*	*EM*	*SOM*
Brain tumour	3.1E-243	4.9E-264	1.39E-266	8.3E-261	1.3E-262	3.0E-258
SRBCT	7.8E-253	3.5E-271	2.6E-018	6.4E-271	6.6E-025	1.6E-270
Adult malignancy	1.89E-242	6.19E-257	5.7E-240	2.24E-249	1.253E-251	1.88E-256

Datasets	*Avg. linkage*	*CSPA*	*SiMM-TS*	*MCLA*	*HGPA*
Brain tumour	6.3E-266	2.2E-262	7.9E-257	4.4E-259	7.0E-262
SRBCT	2.3E-282	2.1E-270	5.0E-027	2.0E-270	5.5E-265
Adult malignancy	1.036E-258	2.658E-250	5.245E-245	1.99E-249	1.112E-253

5.4 Gene markers for brain tumour dataset

In Figure 3, the identified relevant gene markers are shown by heatmap of sample vs. gene matrix of brain tumour 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 colour. High expression level is represented as red colour, 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 tumour class.

Figure 3 Heatmap for brain tumour dataset (see online version for colours)

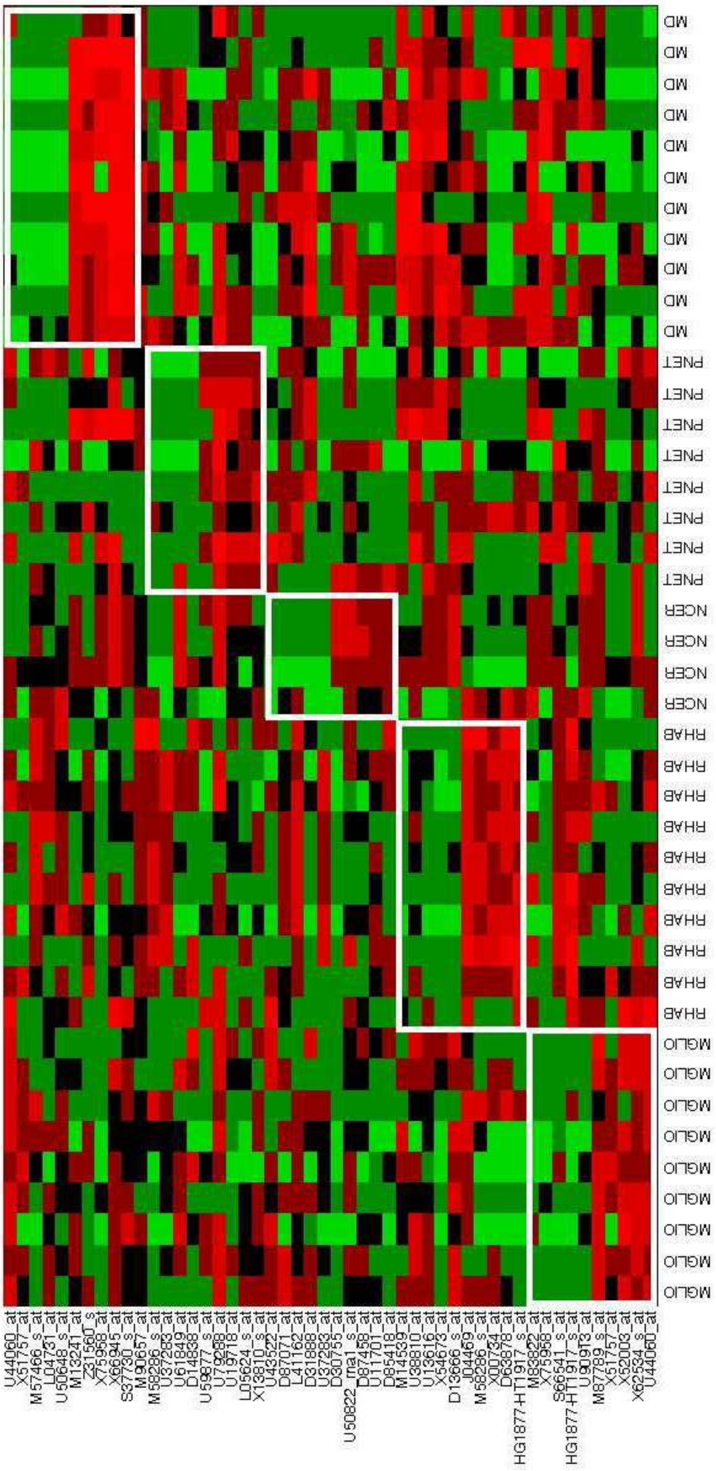


Table 3 The gene markers of the brain tumour data for the MGLIO class, their IDs, selection frequencies, and up/down regulation natures

MGLIO			*RHAB*			*NCER*		
Gene ID	*Frequency*	*Up/down*	*Gene ID*	*Frequency*	*Up/down*	*Gene ID*	*Frequency*	*Up/down*
U44060_at	100	up	HG919-HT919_at	99	up	D85418_at	100	Up
X62534_s_at	100	up	D63878_at	100	up	U11701_at	98	Up
X52003_at	98	up	X00734_at	99	up	D87463_at	97	Up
X51757_at	100	up	M58286_s_at	100	up	U50822_rna1_s_at	100	Up
M87789_s_at	97	up	J04469_at	98	up	D30755_at	99	Up
U90913_at	99	Down	D13666_s_at	100	Down	U37283_at	100	Down
HG919-HT919_at	100	Down	X54673_at	100	Down	D31888_at	100	Down
S66541_s_at	100	Down	U13616_at	96	Down	L41162_at	99	Down
X75958_at	100	Down	U38810_at	100	Down	D87071_at	100	Down
M83822_at	100	Down	M14539_at	99	Down	U43522_at	100	Down

PNET			*MD*		
Gene ID	*Frequency*	*Up/down*	*Gene ID*	*Frequency*	*Up/down*
X13810_s_at	100	up	S37730_s_at	98	Up
L05624_s_at	98	up	X66945_at	100	Up
U19718_at	97	up	X75958_at	99	Up
U79288_at	100	up	Z31560_s	100	Up
U59877_s_at	98	up	M13241_at	100	Up
D14838_at	100	Down	U50648_s_at	100	Down
U61849_at	100	Down	L04731_at	100	Down
U37283_at	99	Down	M57466_s_at	100	Down
M58286_s_at	100	Down	X51757_at	98	Down
M90657_at	100	Down	U44060_at	99	Down

In Table 3, we have reported the top 10 gene markers along with their descriptions and up/down regulation states for the MGLIO, RHAB, NCER, PNET and MD tumour classes. Alsom the frequencies of selection of each gene over 20 runs of AMOSA are also reported. For the MGLIO class, the genes U44060_at, X62534_s_at, X52003_at, X51757_at,M87789_s_at are up regulated and U90913_at, HG919-HT919_at, S66541_s_at, X75958_at, M83822_at are down regulated. Interesting thing to observe is that these genes behave almost oppositely in the remaining tumour classes (Figure 3). For RHAB class, the genes HG919-HT919_at, D63878_at, X00734_at, M58286_s_at, J04469_at are up regulated and D13666_s_at, X54673_at, U13616_at, U38810_at, M14539_at are down regulated. For the NCER class, the genes D85418_at, U11701_at, D87463_at, U50822_rna1_s_at, D30755_at are up regulated and U37283_at, D31888_at, L41162_at, D87071_at, U43522_at are down regulated.

For PNET class, the genes, X13810_s_at, L05624_s_at, U19718_at, U79288_at, U59877_s_at are up regulated and D14838_at, U61849_at, U37283_at, M58286_s_at, M90657_at are down regulated.

For MD class, the genes S37730_s_at, X66945_at, X75958_at, Z31560_s, M13241_at are up regulated and U50648_s_at, L04731_at, M57466_s_at, X51757_at, U44060_at are down regulated. 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 tumour dataset with different existing literature. For example, in case of brain tumour dataset, genes U44060_at, X62534_s_at and M83822_at are reported to belong to MGLIO tumour class in Pomeroy et al. (2002) as similar to our obtained results. Similarly for RHAB tumour class D63878_at, X54673_at genes are reported in Pomeroy et al. (2002). In literature (Pomeroy et al., 2002; Tsai et al., 2008), D85418_at, D30755_at and D31888_at genes are reported to belong to NCER class as we obtained in our work. Gene U79288_at is reported in Tsai et al. (008) to belong to PNET class. Z31560_s and U50648_s_at genes are reported to belong to MD class in Wang et al. (2008) and Tsai et al. (2008), respectively. These findings are similar to the observations derived from our present study.

6 Conclusions

In this article, we have formulated the problem of clustering of cancer tissue samples as a MOO problem and solved it with the help of a MOO-based clustering approach. We have considered symmetry property of real life objects and inspired by that we have proposed PS version of our chosen MOO clustering algorithm. Two PS-based cluster quality measures, XB, Sym indices are used as two objective functions. The performance of the PS-based clustering technique is evaluated on three gold standard datasets, brain tumour, adult malignancy and SRBCTs. The results show that for most of the datasets proposed algorithm not only outperforms the existing single objective-based clustering techniques but also achieves better results than a recently developed MOO-based clustering technique namely AMOSA and MOGASVM. 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 the relevancy of them is shown visually with the help of heatmap.

In future, we are planning to apply our proposed PS distance-based clustering algorithm on some gene expressing datasets. Also, we are working on applying some post processing technique like majority voting technique in order to combine non-dominated solutions which we obtain as clustering output of our proposed algorithm.

References

Acharya, S., Thadisina, Y. and Saha, S. (2014) 'Multi-objective clustering of tissue samples for cancer diagnosis', *2014 International Conference on advances in Computing, Communications and Informatics*, 24–27 September, pp.1059–1064.

Alizadeh, A.A. et al. (2000) 'Distinct types of diffuse large b-cell lymphomas identified by gene expression profiling', *Nature*, Vol. 403, No. 6769, pp.503–511.

An, L. and Doerge, R.W. (2012) *Dynamic Clustering of Gene Expression*, ISRN Bioinformatics, Article ID 537217, 12pp.

Bandyopadhyay, S. and Saha, S. (2007) 'GAPS: a clustering method using a new point symmetry based distance measure', *Pattern Recognit.*, Vol. 40, No. 12, pp.3430–3451.

Bandyopadhyay, S. and Saha, S. (2012) *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*, Springer Science & Business Media, Artificial Intelligence, 248pp.

Bandyopadhyay, S., Mukhopadhyay, A. and Maulik, U. (2007) 'An improved algorithm for clustering gene expression data', *Bioinformatics*, Vol. 23, No. 21, pp.2859–2865.

Bandyopadhyay, S., Saha, S., Maulik, U. and Deb, K. (2008) 'A simulated annealing-based multiobjective optimization algorithm: AMOSA', in *IEEE Transactions on Evolutionary Computation*, pp.269–283.

de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B. and Schliep, A. (2008) 'Clustering cancer gene expression data: a comparative study', *BMC Bioinformatics*, Vol. 9, No. 1, p.497.

Deb, K. (2009) *Multi-Objective Optimization Using Evolutionary Algorithms*, Vol. 16, John Wiley & Sons.

Deb, K., Pratap, A., Agrawal, S. and Meyarivan, T. (2002) 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE Transactions on Evolutionary Computation*, Vol. 6, No. 2, pp.182–197.

Dembele, D. and Kastner, P. (2003) 'Fuzzy C-means method for clustering microarray data', *Bioinformatics*, Vol. 19, No. 8, pp.973–980.

Dinger, S.C., Van Wyk, M.A., Carmona, S. and Rubin, D.M. (2012) 'Clustering gene expression data using a diffraction-inspired framework', *Biomedical Engineering Online*, Vol. 11, No. 1, p.85.

Everitt, B.S., Landau, S. and Leese, M. (2001) *Cluster Analysis*, E. Arnold, London, UK.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York.

Golub, T.R. et al. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, No. 5439, pp.531–537.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) 'Data clustering: a review', *ACM Computing Surveys*, Vol. 31, No. 3, pp.264–323.

Kirkpatrick, S. (1984) 'Optimization by simulated annealing: quantitative studies', *Journal of Statistical Physics*, Vol. 34, Nos. 5–6, pp.975–986.

Maulik, U. and Bandyopadhyay, S. (2000) 'Genetic algorithm based clustering technique', *Pattern Recognition*, Vol. 33, No. 9, pp.1455–1465.

Mukhopadhyay, A., Bandyopadhyay, S. and Maulik, U. (2010) '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, DOI: 10.1371/journal.pone.0013803.

Paul, S. and Maji, P. (2013) 'City block distance for identification of co-expressed MicroRNAs', *SEMCCO*, No. 2, pp.387–396.

Pomeroy, S.L. et al. (2002) 'Prediction of central nervous system embryonal tumour outcome based on gene expression', *Nature*, Vol. 415, No. 6870, pp.436–442.

Saha, S. and Bandyopadhyay, S. (2011) 'On principle axis based line symmetry clustering techniques', *Memetic Comp*., Vol. 3, No. 2, pp.129–144.

Saha, S. et al. (2013) 'Gene expression data clustering using a multiobjective symmetry based clustering technique', *Computers in Biology and Medicine*, Vol. 43, No. 11, pp.1965–1977.

Strehl, A. and Ghosh, J. (2002) 'Cluster ensembles – a knowledge reuse framework for combining multiple partitions', *J. Machine Learning Research*, Vol. 3, pp.583–617.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q. and Kitareewan, S. (1999) 'Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation', in *Proc. Nat. Academy of Sciences*, USA, Vol. 96, pp.2907–2912.

Tsai, Y.S. et al. (2008) 'Discovery of dominant and dormant genes from expression data using a novel generalization of SNR formulti-class problems', *BMC Bioinformatics*, Vol. 9, No. 1, p.425.

Wang, Y. and Pan, Y. (2014) 'Semi-supervised consensus clustering for gene expression data analysis', *BioData Mining*, Vol. 7, No. 1, pp.1–13.

Wang, Y. et al. (2008) *Integrative Methods for Gene Data Analysis and Knowledge Discovery on the Case Study of KEDRI's Brain Gene Ontology*, Doctoral dissertation, Auckland University of Technology.

Yeung, K.Y. and Bumgarner, R.E. (2003) 'Multiclass classification of microarray data with repeated measurements: application to cancer', *Genome Biology*, Vol. 4, No. 12, pp.R83–R83.

Abbreviations

Multi-objective-optimisation	MOO
Point symmetry	PS
Line symmetry	LS
Small round blood cell tumours	SRBCT
Adjuster rand index	ARI
Classification accuracy	CoA
Archived multi-objective simulated annealing	AMOSA

Appendix

A.1 Illustrating the clustering algorithm with some examples

Below, we have provided an example illustrating our proposed algorithm. Let us assume that we want to classify six data points. These data points are two-dimensional in nature and are shown as follows: (1, 2), (2, 3), (4, 6), (5, 9), (3, 3), (8, 4).

Step 1 String representation of solutions, let us assume that we choose population size = 3, so three archive-elements or solutions are needed to be represented as strings. They are shown as follows:

For solution 1: rand() function is applied for choosing string size or the number of clusters in that solution. Let the random number returns the value 2. So, two clusters will be there in solution 1. Now centres of two clusters are randomly chosen from data points. Let us assume that for 1st cluster, (1, 2), and for 2nd cluster (3, 3) are chosen as cluster centroids.

It is shown in Figure A1.

For other two solutions according to the upper specified way, we perform the string representation.

Figure A1 String representation of solutions

Solution 2	(1,2)	(4,6)	(8,4)

Solution 3	(3,3)	(8,4)

Step 2 Assignment of points and computation of objective functions.

Assignment of points: Once the string representation is done, in the next part for each solution its membership matrix is computed. For example, for solution 1, members of two clusters are computed. For this computation calculation of PS-based distance is needed and it is done in the following way,

To find the members of cluster 1 of solution 1.

PS-based distances between all other points (2, 3), (4, 6), (5, 9) and (8, 4) and centre of cluster 1, i.e., (1, 2) are computed in the following way,

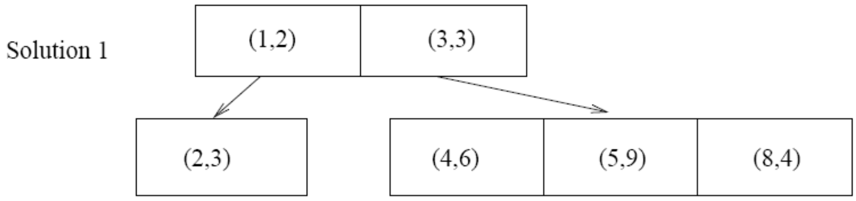
a suppose we are computing PS-based distance between (1, 2) and (2, 3) denoted as $d_{ps}(x, c)$. So, reflected point of (2, 3) or 'x' with respect to cluster centre (1, 2) or 'c' is denoted by 'x^*', $\{2 \times (1, 2) - (2, 3)\} = (2, 4) - (2, 3) = (0, 1) = x^*$ $d_e(x, c) = \sqrt{(1-2)^2 + (2-3)^2} = 1.41$.

To compute $d_{sym}(x, c)$ we have to find knear number of nearest neighbours of reflected point x^*. In this work we have fixed knear = 2. So, we compute Euclidean distance between reflected point x^* or (0, 1) and rest points of datasets, i.e., (4, 6), (5, 9) and (8, 4). We found that (4, 6) and (8, 4) are nearest neighbours of (0, 1) and their corresponding Euclidean distances are 6.4 and 8.5, respectively.

so, $d_{sym}(x, c) = (6.4 + 8.5) / 2 = 7.47$.

So, overall PS-based distance between (1, 2) and (2, 3) or $d_{ps}(x, c) = d_{sym}(x, c) \times d_e(x, c) = 7.47 \times 1.41 = 10.5$.

Figure A2 Assignment of data points to different clusters



In the upper specified way PS-based distance between centre (1, 2) and other points (4, 6), (5, 9) and (8, 4) are calculated and stored.

For cluster 2 similarly PS-based distances of four same points (2, 3), (4, 6), (5, 9) and (8, 4) from centre of cluster 2, (3, 3) are calculated and stored.

Now each point of (2, 3), (4, 6), (5, 9) and (8, 4) is placed in any one of two clusters depending on the minimum distance-based criterion. A particular point is assigned to that cluster with respect to which it is having the minimum distance.

According to this, we found that (2, 3) resides in cluster 1 and (4, 6), (5, 9) and (8, 4) move to in cluster 2. The final clustering is shown in Figure A2.

Similarly for other two solutions, assignments of data points to the respective cluster centres are performed.

Computation of objective functions: Now for each solution, values of its two objective functions XB and Sym are calculated.

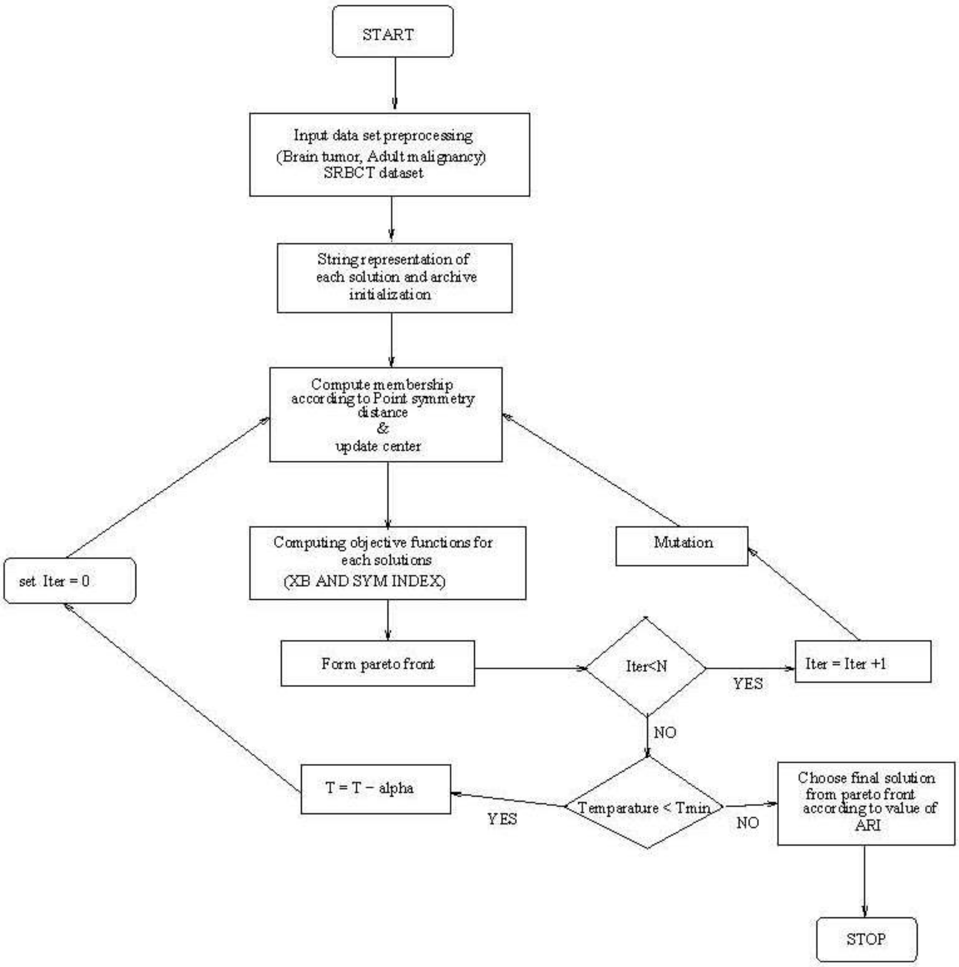
- suppose for Solution 1: XB = *xin*, *Sym*= sin,
- for solution 2: XB = $xin + \alpha$, Sym = sin + γ
- for solution 3: XB= $xin + \alpha$, Sym = sin – γ.

From these values our algorithm chooses non-dominating solutions. As we can see from these three solutions that solution 1 and solution 2 are non-dominating to each other. So, two solutions are selected and dominated one (3rd solution) is discarded.

Step 3 Search operators: Mutation is performed on selected solutions of Step 2. Any one of the three types of mutation is performed on each solution as shown in main manuscript. Then for each solution, Step 2 is repeated.

Step 4 Step 2 and Step 3 are repeated for each iteration of the proposed algorithm. At the end, from the set of non dominated solutions, one solution is chosen as final solution whose *ARI* value is the maximum.

Figure A3 Flowchart of our proposed clustering algorithm


START
Input data set preprocessing
(Brain tumor, Adult malignancy)
SRBCT dataset
String representation of
each solution and archive
initialization
Compute membership
according to Point symmetry
distance
&
update center
Computing objective functions for
each solutions
(XB AND SYM INDEX)
Mutation
set Iter = 0
Form pareto front
Iter<N
YES
Iter = Iter +1
NO
T = T − alpha
Temparature < Tmin
YES
NO
Choose final solution
from pareto front
according to value of
ARI
STOP