International Journal of Bioinformatics and Biomedical Engineering
Vol. 1, No. 1, 2015, pp. 19-36
http://www.aiscience.org/journal/ijbbe

Optimal Feature Subset Selection Using Similarity-Dissimilarity Index and Genetic Algorithms

Muhammad Arif*

Department of Computer Science, College of Computer and Information Systems, Umm, Alqura University, Makkah, Kingdom of Saudi Arabia

Abstract

Optimal feature subset selection is an important pre-processing step for classification in many real life problems where number of dimensions of feature space is large and some features are may be irrelevant or redundant. One example of such a situation is genes expression profile data to classify among normal and cancerous samples. Contribution of this paper is five folds. Similarity-dissimilarity index (MSDI) is proposed which can estimate the class discrimination quality of the high dimensional feature space without using any kind of classifier. A framework to find out the best features subset from the *n*-dimensional feature space using genetic algorithm is proposed to select the minimum possible important features optimally using MSDI as fitness function to evolve the population. Similarity-dissimilarity plot is proposed to visualize the high dimensional data that can be used to extract important information about the class discrimination quality of the feature space. It is possible to predict the best classification accuracy using MSDI when an appropriate classifier is used. Another index called average differential of similarity and dissimilarity distances above similarity-dissimilarity line is proposed which gives information about how far each class instances or clusters are from other classes and the compactness of the classes in the feature space. Effectiveness of the methods is highlighted by using a large set of benchmark datasets in cancer classification and size of features subset and predicted classification accuracy is compared with the published results.

Keywords

Pattern Classification, Genetic Algorithm, Biomedical Datasets, Nearest Neighbor, Visualization

Received: May 15, 2015 / Accepted: June 10, 2015 / Published online: July 7, 2015

1. Introduction

In pattern classification, many features are extracted from the raw data coming from sensors, clinical test and other sources. These features may be relevant or irrelevant to the pattern classification problem or redundant in nature. Before designing an optimal classifier, it is very important to assess the discrimination quality of the feature space. Selecting relevant features and ignoring irrelevant or redundant features can reduce the computational cost, improve the classification accuracy and simplify the decision boundary among different classes.

Gene expression microarray data [1],[2]from oligonucleotide arrays or cDNA microarrays is used to classify different types of cancer tumors in the cancer research. This biotechnology is used to collect thousands of gene expressions. In gene expression microarray data, normally number of variables or genes may be huge and much greater than the number of instances [3]. This is “large dimension, small instances” problem in which number of features or variables (genes) are much larger than number of instances (like tumor samples). Feature selection is an important pre-processing step for microarray gene expression data after extraction of features and before designing an appropriate classifier. Feature selection is also applied to many other areas as pre-

* Corresponding author
E-mail address: mahamid@uqu.edu.sa

processing step like document processing [4], text categorization [5], spam filtering [6]etc. Reduction of features can be through dimension reduction techniques like principal component analysis, Fischer discriminant analysis; multidimensional scaling etc. Xu et al [7]applied two modified linear discriminant analysis (LDA) techniques to microarray classification datasets with limited sample size. The "large dimensions, small instances" datasets create problems of instability and singularity in the performance of LDA which is tried to solved by [8][9]. Xiong et al [10]proposed a hierarchical strategy in which all genes are evaluated first individually and near optimal set is defined. In the next step a subset of two or three genes are searched that can optimize the classification accuracy using stepwise and monte carlo methods. A good survey about dimension reduction techniques can be found in [11][12]. In these techniques, features are mapped or projected on a low dimensional space losing the understanding of the dimensions. Another method is to select a subset of the features that can describe the variance of the data in the best way possible.

Blum et al [21]divided the feature selection methods into three broader categories, embed, wrapper and filtering methods. In embedding methods, selection of feature subset and classifier are embedded together and both are optimized simultaneously. Chen et al [22]proposed flexible neural tree for feature selection and classification simultaneously. Flexible neural tree is constructed by using genetic programming and claimed that it is efficient for input feature selection and produced improved classification rates. In filtering methods, an evaluation criterion is defined and a subset of features is selected by using this criterion. Filtering methods use the statistical or some intrinsic properties of the data to define an appropriate evaluation criterion to select the relevant features. Advantages of filtering methods are computational simplicity, fast and independent of classifier [23]. After the selection of features subset, a classifier has to be optimized with this available subset only. Ando et al [24] used p-value calculated by Mann-Whitney test to rank the genes and selected top ten genes for the classification. Guoan et al [25]used t-statistics to find out relevant biomarkers. In mass spectrometry, many researchers used statistical classifiers. Satten et al [26]used random forest after de-noising and standardizing the whole mass spectra. Adam et al [27]used SELDI protein profiling on prostate cancer samples and best features are decided by peak detection and discrimination power of each peak by using area under the Receiver Operator Curve (ROC). Decision tree is used to classify the prostate cancer samples from the normal ones. One disadvantage of the filtering methods is their analysis of the features in isolation and they try to rank the features according to their solo performance in the discrimination of the classes. This is not true for all cases and it may happen that combination of features may produce better or worse classification results. Hence it is important to analyze the effectiveness of the features subsets collectively.

In wrapper methods, a learning method is defined and only those features are selected which show high classification or prediction accuracy by the learning method. In these types of methods, multiple feature subsets are selected and evaluated on a particular classifier iteratively. For high dimensional feature space, some heuristic search like genetic algorithm is required to define multiple features subset. Wrapper methods prove to be computationally very intensive especially for high dimensional space like microarray gene expressions datasets. It also has the risk of over-fitness. A good reference for the details of these types of methods can be found in ([23], table 2.).

Xie and Wang [28]combined the advantages of filtering method and wrapper method by defining improved F-score as selection criterion and sequential forward search is used to evaluate the wrapper method and support vector machine is used for evaluating the classification accuracy. Yang et al [29]reported that there is no single ranking scheme or statistics which is universally optimal for all datasets. So they proposed a mixture of measures to rank the genes and claimed that it showed better performance than a single statistics. Yang et al [30]in another paper proposed improved hybrid system using genetic ensemble system and different filtering techniques to enhance the performance of features subset.

In this paper, a new framework to select the feature subsets especially for "large dimension, small instances" cases is proposed to remove the irrelevant features and select best combinations of features. In this framework, concept of neighborhood is used and a new Similarity dissimilarity index (MSDI) is proposed which can assess the class discrimination quality of the features without using any classifier. Genetic algorithm is used to generate various feature subsets and their fitness is evaluated by the MSDI. A penalty function is defined to penalize subsets having large number of features. By using this framework it is possible to generate various features subsets having minimum sizes and maximum MSDI values.

Furthermore, visualization of high dimensional feature space is also very important in the context of patter classification in many real life applications. Lot of efforts has been done in the past to effectively visualize the data. Radviz [52][53]is another type of visualization technique that can be applied to visualize the data structure. It can be applied to feature space having continuous attributes normalized to interval [0,1].

Dimension anchors are placed evenly on the circumference of a unit circle. Dimension anchors attract every data point towards itself with the strength proportional to the value of data point in the dimension corresponding to the dimension anchor. Number of dimensions that can be placed on the circumference of the circle is limited due to limited space of the unit circle. Moreover, an optimization of the order of dimensions placed on the unit circle is very important to get some meaningful insight in the data. It is handled by Vizrank [53]and McCarthy et al [54]. Freeviz [55]is another extension of Radviz which allows dimension anchors to be placed anywhere in the unit circle. Correlated features are placed near to each other and less important features are placed near the center of unit circle. Vectorized Radviz [56] is proposed to better visualize the multi-clusters data by increasing the dimensions of the data through data flattening.

Similarity-dissimilarity plot proposed in [57][58]provide a good visualization tool to extract many useful information about the features discrimination quality in the high dimensional space. Independent of number of dimensions, the proposed plot can discriminate between good quality instances (producing good classification accuracy) and bad quality instances (creating confusion with other classes) on the feature space. In the sparse high dimensional feature space (like "large dimension, small instances" case in gene expressions datasets), sometimes it is possible that instances at the boundary of classes may considered as bad quality instance even though it is separable by a decision boundary among different classes. In this paper, a Similarity-dissimilarity plot is proposed in which neighborhood count is defined for every instance. This count explains that how many times an instance is considered in the neighborhood of other instances of similar class. High neighborhood counts means that a particular instance is at the boundary of its own class and should be considered as good instance. Furthermore outliers of different classes can also be identified by the Similarity dissimilarity plot. This framework is explained in detail in the coming sections and applied to many benchmark gene expressions datasets available online. In the end, results of the proposed framework are compared with the reported results in the literature.

2. Material and Methods

In this section, methodology of the proposed framework and description of the datasets will be elaborated. The methodology of the proposed framework is to search a single or multiple feature subsets which have the maximum value of MSDI and have minimum possible size of features. Block diagram of the proposed framework is given in Figure 1.

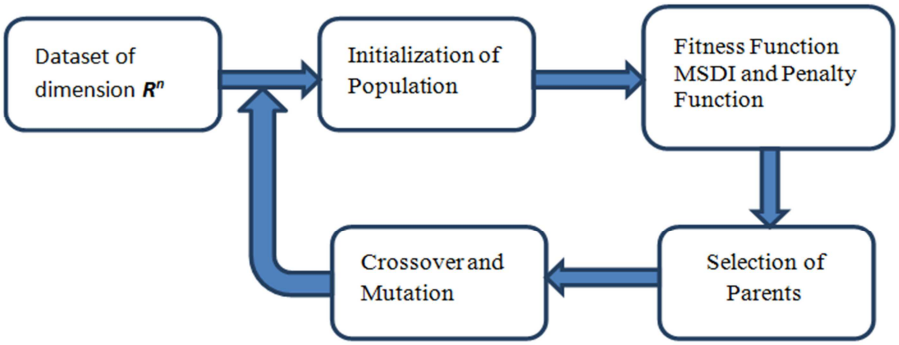

Figure 1. Proposed Framework to select optimal feature subsets.

2.1. Similarity-Dissimilarity Index (MSDI)

Similarity-dissimilarity index (MSDI) gives the fraction of instances from each class who are in the neighborhood of their correct classes in the high dimensional feature space. Therefore, it can predict the expected classification accuracy of every class based on the distribution of instances of that class in the high dimensional feature space without applying any classifier to the dataset.

Let a dataset $X = [X_1, X_2, X_3, \dots, X_N] \in R^n$ and X_i is an instance lies in an n-dimensional feature space. There are N_C classes in the dataset. Total number of instances in the dataset is $N = \sum_{i=1}^{N_c} nd_i$, and nd_i is the number of instances in the i^{th} class.

All features of the dataset are normalized such that their mean become zero and variance is set to one. Normalization can be done as follows,

$$X_i^j = \frac{X_i^j - \mu^j}{\sigma^j},\ i = 1, 2, \dots, N \text{ and } j = 1, 2, \dots, n \qquad (1)$$

Here μ^j is the mean and σ^j is the standard deviation of the j^{th} feature.

Algorithm of MSDI is explained in the following steps.

For each instance X_i of the dataset, do steps 1 to 3.

Step 1: Calculation of the Similarity Distance

For an instance X_i belonging to the class C_j, k nearest neighbors are found from the class C_j (class of X_i). This nearest neighbor set is called as similarity set NN_{sim} for an instance X_i. Similarity distance $d_{sim}(i)$ is calculated as the mean distance of the instance X_i from the similarity set NN_{sim},

$$d_{sim}(i) = \frac{1}{k}\sum_{l=1}^{k} d_p(X_i, X_l), \quad X_l \in NN_{sim} \tag{2}$$

Distance metric $d_p(X_i, X_l)$ used above is Minkowski distance metric of order p which is given as,

$$d_p(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{3}$$

Here $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$ are two n-dimensional instances in R^n. Euclidean distance is the Minkowski distance metric with $p = 2$ and Manhatten distance is another Minkowski distance metric with p = 1.

Step 2: Calculation of the Dissimilarity Distance

Similarly, for an instance X_i belonging to the class C_j, k nearest neighbors are searched from classes other than class C_j and these k nearest neighbors are called dissimilarity set NN_{dissim}. Dissimilarity distance $d_{dissim}(i)$ is calculated as the mean distance of the instance X_i from the dissimilarity set NN_{dissim}.

$$d_{dissim}(i) = \frac{1}{k}\sum_{l=1}^{k} d_p(X_i, X_l), \quad X_l \in NN_{dissim} \tag{4}$$

Step 3: Calculation of the Neighborhood Count

For each instance in the dataset, a neighborhood count η is defined. The neighborhood count η_m for an instance X_m is count of times the instance X_m is found to be in the neighborhood of other instances from the similar class. The neighborhood count η_m is calculated as follows,

The neighborhood count of all the instances in the dataset is initialized to zero. For the instance X_i, find out all instances of class C_j whose distances from X_i is less than $d_{dissim}(i)$ and call it as neighborhood set $NS(i)$.

$$NS(i) = \{X_i | X \in \forall_{k \neq j} C_k, d_p(X, X_i) \leq d_{dissim}(i)\} \tag{5}$$

Now neighborhood count η_m of all the instances belonging to the neighborhood set $NS(i)$ is incremented as follows,

$$\eta_m = \eta_m + 1, \ m \in NS(i) \tag{6}$$

This procedure is carried out for all the instances in the dataset.

The neighborhood count tells us about how many times the instance is listed in the neighborhood set $NS(i)$.

Step 4: Calculation of MSDI.

MSDI gives the fraction of instances which are good in the context of classification. So for every instance X_i of the dataset, a binary variable $Q(i)$ is defined as zero or one. If $Q(i)$ is one then instance X_i will be classified correctly and if $Q(i)$ is zero then instance X_i will most probably be misclassified and confused with some other class by the classifier. If the similarity distance $d_{sim}(i)$ is less than the dissimilarity distance $d_{dissim}(i)$, it suggests that the instance is near to its own class as compared to other classes and chances of classifying this instance X_i correctly by an optimal classifier are high. Lesser the similarity distance than the dissimilarity distance, easier is to classify this instance correctly. The variable $Q(i)$ is calculated as follows,

$$Q(i) = \begin{cases} 1 & if\ (d_{sim}(i) < d_{dissim}(i)) \vee (\eta_i > \varsigma) \\ 0 & Otherwise \end{cases} \tag{7}$$

Where ς is the lower threshold on the neighborhood count, based one which an instance is considered to be classifiable correctly. This is considered as an important factor because for all those instances which are on the border of the clusters of a particular class may have similarity distances larger than the dissimilarity distances. This is due to the fact that they are nearer to other classes as compared to their original class even though they can be classified correctly and a decision boundary exists among classes. This situation becomes more critical when the number of instances per class are few and the feature space is high dimensional which is very common in microarray genes expressions data for different kind of cancers. For such a sparse representation of classes in the high dimensional space, calculation of neighborhood count can improve the correct prediction of classification accuracy. This point will be further elaborated by some examples in the coming section.

Similarity-dissimilarity index (MSDI) for each class C_j is calculated as follows,

$$MSDI_j = \frac{\sum_i Q(i), i \in C_j}{nd_j} \tag{8}$$

Overall MSDI for the whole dataset is calculated as,

$$MSDI = \frac{\sum_j^{N_c} nd_j MSDI_j}{\sum_j^{N_c} nd_j} \tag{9}$$

Range of MSDI is from zero to one. Low value of MSDI corresponds to the overlapped clusters of different classes and classification accuracy is predicted to be very poor. High value of MSDI means different classes are well separated and high classification accuracy is possible by designing an

appropriate classifier.

Step 5: Similarity-Dissimilarity Plot.

To visualize the high dimensional data, Similarity-dissimilarity plot is proposed which is based on similarity-dissimilarity plot proposed in [57]. For every class C_j, a particular color and shape is defined. For the i^{th} instance of class C_j, if $d_{sim}(i) < d_{dissim}(i)$, this instance will be plotted over the similarity dissimilarity line by the shape and color of class C_j. If neighborhood count is greater than threshold value then the instance will be plotted below similarity-dissimilarity line will be plotted as shape of the class C_j but the color is black. If $d_{sim}(i) \geq d_{dissim}(i)$, this instance will be plotted as shape of the class C_j and the color of the shape will be decided by the class of majority of the instances in the neighborhood of i^{th} instance. Figure 2 explains the meaning of location of an instance on the Similarity-dissimilarity plot. If the instance is above the similarity-dissimilarity line then this instance is nearer to its own class as compared to other classes. As the instance moves upward, better is the quality of instance in the context of classification. Similarly, if instances are located on the left side of the plot, then the class is sparse (similarity distances are large). All the instances below the similarity-dissimilarity line may be considered as low quality instances. Some instances below the line may be considered as good points if their neighborhood count is high (points on the boundary of a class). This concept is further explained by example 1 as shown in Figure 3. The Similarity-dissimilarity plot is shows in Figure 4. In Figure 3, some data points of two class problem are plotted. It can be seen from the figure that both classes are linearly separable and 100% classification accuracy is achievable easily by any classifier.

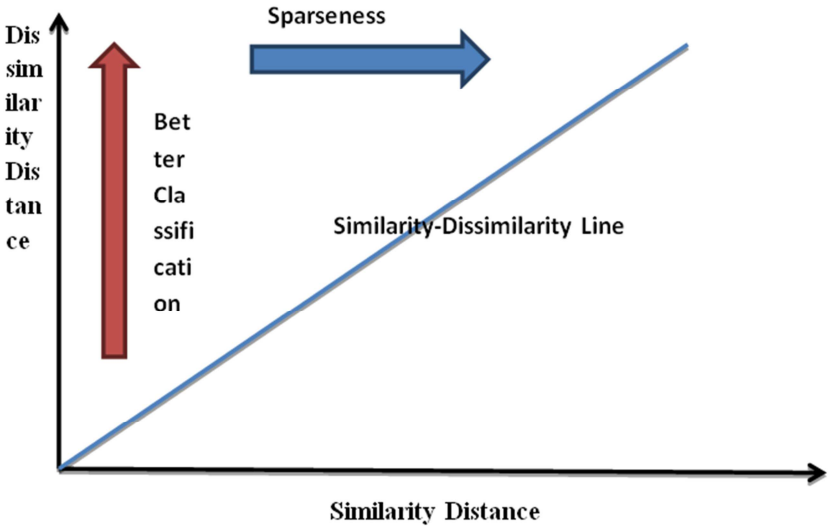

Figure 2. Similarity Dissimilarity Plot.

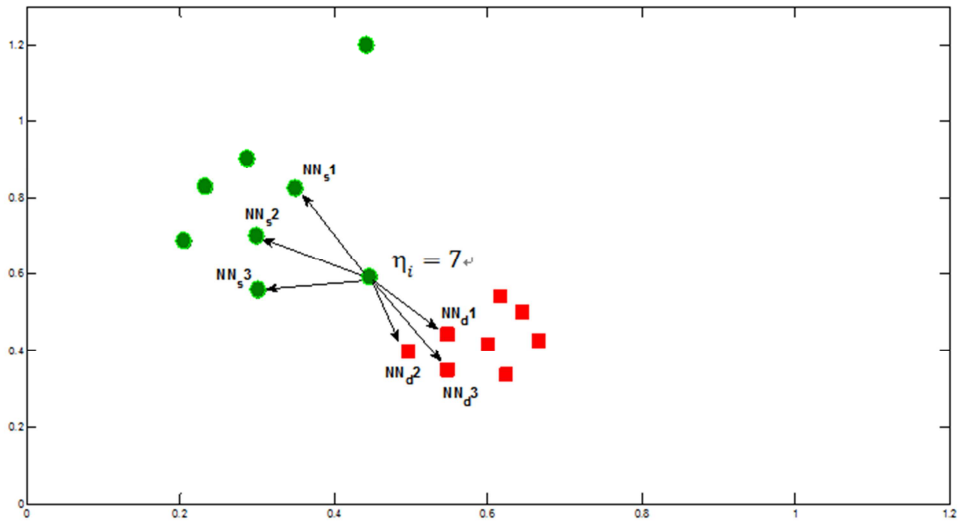

Figure 3. Example 1.

2.2. Example 1: Graphical Illustration of MSDI

One data point of green class is at the boundary of the class and near to the red class. Let us consider this data point as query point. There are three data points of green class NN_s in the neighborhood of the query point and there are three data points NN_d of red class in the neighborhood of query point. When similarity and dissimilarity distances will be calculated for this query point, its similarity distance will be large as compared to the dissimilarity distance. But this data point will be included in the neighborhood of all other seven points when their similarity and dissimilarity distances will be calculated. Hence it is clear that this data point is at the boundary of the green class and should not be counted towards prediction of misclassification.

When MSDI will be calculated the value will be 1.0 showing 100% classification accuracy prediction. It is also clear from the similarity-dissimilarity plot of the dataset (Figure 4.) that all data points are above the similarity-dissimilarity line (SD Line) and only one data point is below the SD line. But since its neighborhood count is high so this data point will also be considered classifiable.

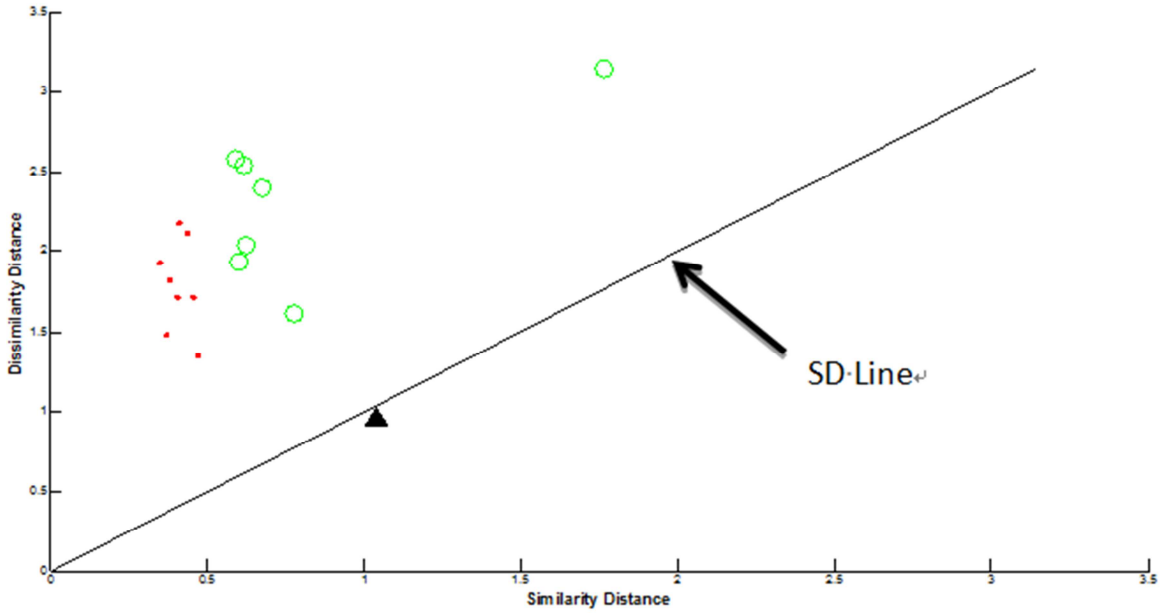

Figure 4. Similarity-Dissimilarity plot of Example 1.

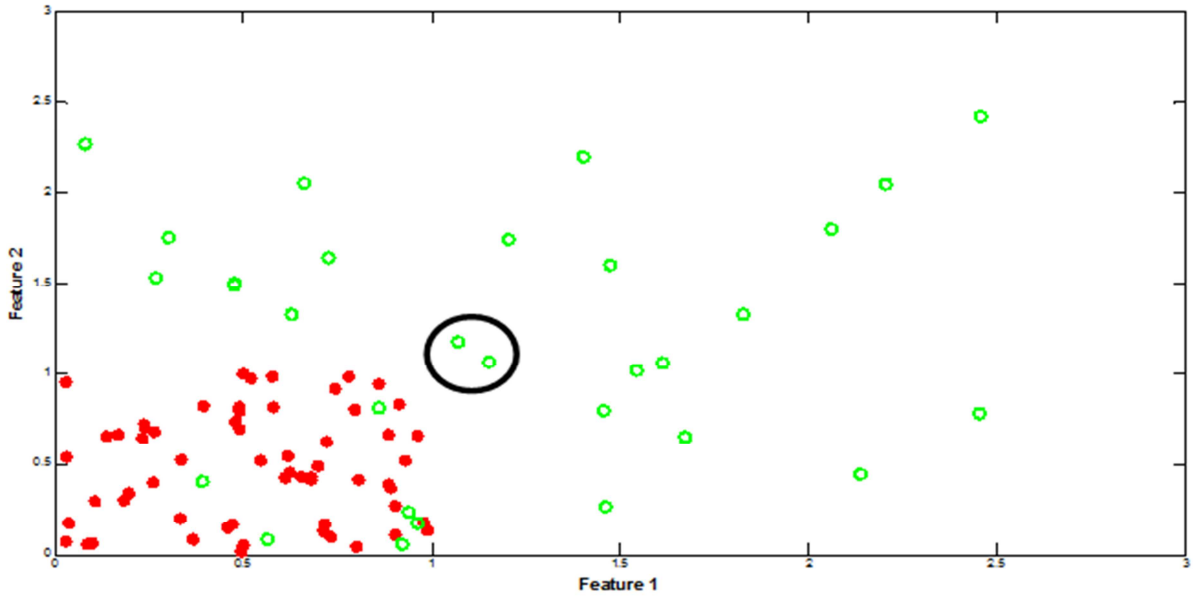

Figure 5. Example 2.

To further illustrate the concept of neighborhood count, dataset of example 2 is plotted in Figure 5. In this dataset, six data points of green class is within the red class which are not good data points in the context of classification. But there are two data points of green class in the black circle which are at the boundary and closer to the red class as compared to the green class. These data points are good data points and can be classified correctly. Figure 6 shows the similarity-dissimilarity plot of example 2 in which the two data points marked in black circle is shown by black markers and considered as good points. MSDI of dataset of example 2 is found to be 0.933 (84/90).

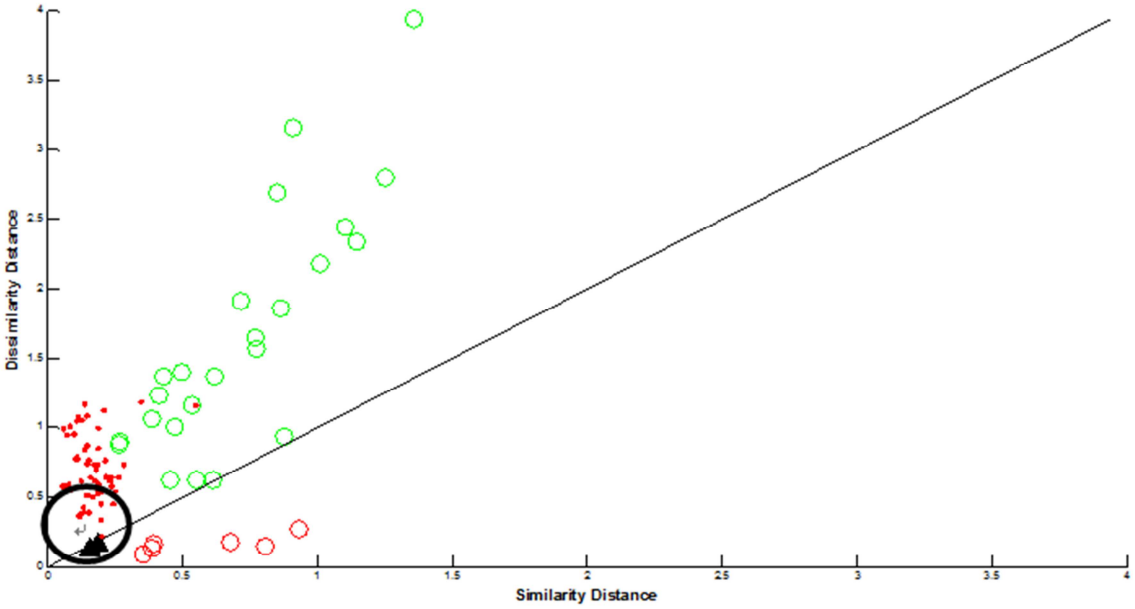

Figure 6. Similarity Dissimilarity plot of example 2.

2.3. Selection of the Optimal Feature Subset

In a high dimensional feature space, not all the features contribute in the correct classification of instances. It might be possible that optimal feature set has much lesser dimensions as compared to the total feature set. Let $\mathcal{F}$ is the total feature set and $\Phi(.)$ is the feature selection criterion function. Formally, the problem of selection of optimal feature subset for a particular dataset X is as follows,

$$\Phi(\mathrm{X}) = \max_{\mathcal{G}\subseteq\mathcal{F},|\mathcal{G}|=\mathrm{d}} \Phi(\mathcal{G}) \quad (10)$$

Here $\mathcal{G}$ is the feature subset and d is the size of this subset. For a particular dataset of n dimensions, there will be $\binom{\mathrm{n}}{\mathrm{d}}$ possible combinations of feature subsets. Since optimal value of d is not known before hand, number of possible combinations will increase exponentially as n grows. In cancer gene expressions datasets, value of n is normally in thousands. Hence, genetic algorithm [59]is used to explore all possible combinations of different sizes of feature subsets and to find out the optimal feature subset that can maximize the feature selection criterion Φ.

Genetic algorithm is used to select the best possible feature subset that can maximize MSDI value of a dataset. Following are the steps to apply genetic algorithm in selecting the optimal feature subset.

Step 1: Define genetic algorithm parameters: Parameters are listed in the Table 1.

Table 1. Genetic Algorithm parameters.

Parameter	Description
N_{pop}	Number of Individuals in a population
N_{gen}	Total number of generations
P_c	Probability of Crossover
P_m	Probability of Mutation
β	Penalizing factor

Step 2: Define initial population: First population is initialed randomly with variable length chromosomes. Each chromosome defines a subset of features and number of features to be included is selected randomly. Size of subset is also selected randomly. Counter of generation is initialized.

Step 3: Fitness Calculation: For i^{th} generation and fitness of every individual of populations is calculated as follows,

$$Fitness(j) = f(MSDI(j), g(\beta))$$

Fitness of j^{th} individual of i^{th} generation is calculated from MSDI. This fitness is penalized by $g(\beta)$ factor depending on

the size of the feature subset. Larger the size of the feature subset, more fitness will be penalized.

Step 4: Selection of Parents for Crossover: Based on the fitness values of i^{th} generation, a fixed number of parents will be selected by roulette wheel selection method.

Step 5: Crossover and Mutation: Crossover of parents will be performed based on single point crossover with the crossover probability P_c. Chromosomes of offspring will be corrected for duplicate entries of features by selecting only unique feature subset of the offspring as shown below (Figure 7.),

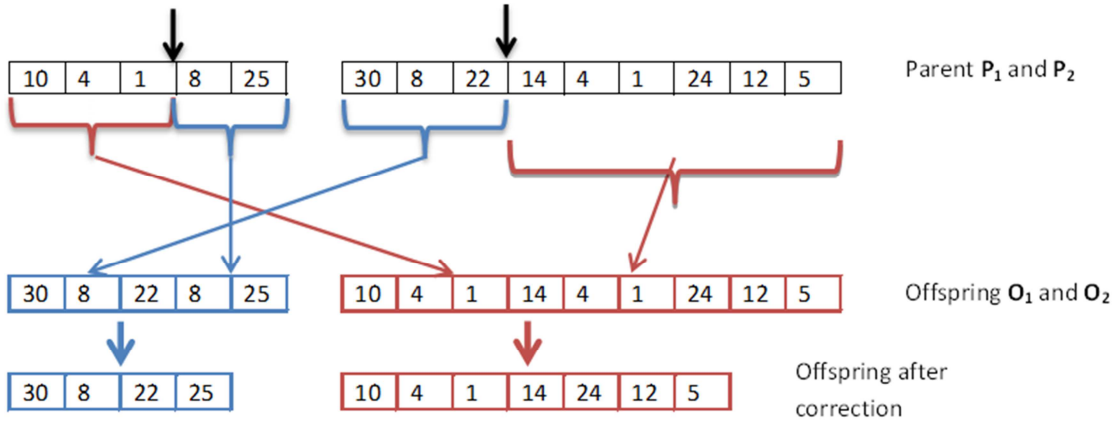

Figure 7. Crossover of chromosomes.

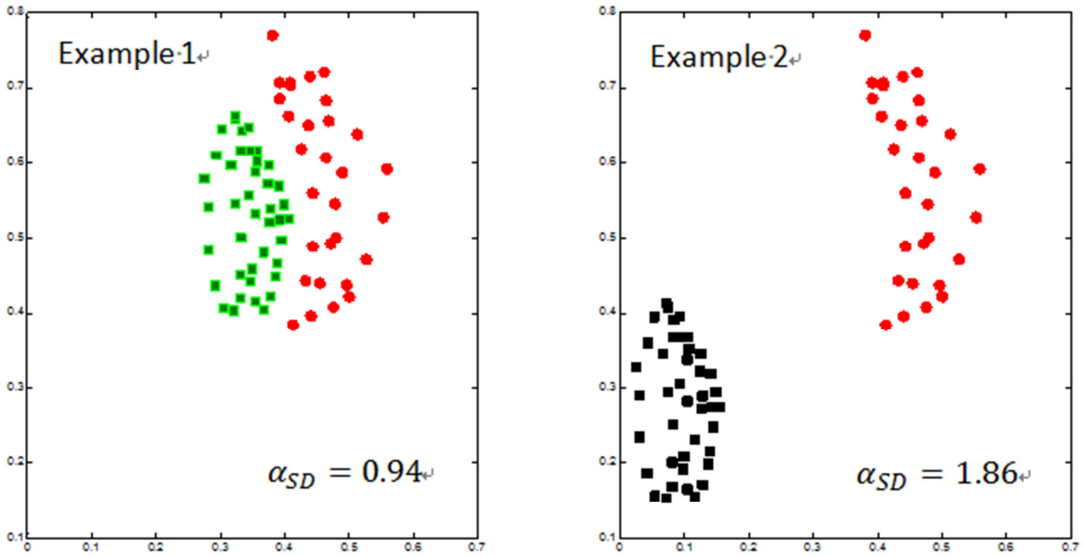

Figure 8. Value of α_{SD} for two examples.

Similarly, Mutation is performed on each individual with probability of P_m. Any location of the offspring is selected randomly and its values is changed to any values between 1 and maximum number of feature in the dataset. New population will be selected from previous population and offspring.

Step 6: Stopping criterion: Steps 2 to 5 will be performed iteratively until some stopping criterion is satisfied or max number of generations is reached.

Step 6: Average Differential of Similarity and Dissimilarity Distances above SD line: Genetic algorithm may return multiple feature subsets that have similar MSDI value. To differentiate among these feature subsets, another index is defined which is called average differential of similarity and dissimilarity distances above the SD line (α_{SD}). It is assumed that if inter-class distances among different classes are large, classification of classes will be easy. α_{SD} is defined as follows,

$$\alpha_{SD} = \frac{1}{N}\sum_{C_j} \sum_{\substack{i \in C_j \\ d_{sim}(i) < d_{dissim}(i)}} \left(d_{dissim}(i) - d_{sim}(i) \right)$$

In

Figure 8, two examples are drawn. In example 1, two classes

are very near to each other although they are classifiable. Similarly in example 2, there are two classes which are far away from each other. The value of α_{SD} is calculated for both examples and it can be seen from the values given in the figure that α_{SD} of example 2 is almost double than example 1. Hence this parameter can be used to assess the inter-class separation in the high dimensional feature set. Higher the value of α_{SD} is, better is the quality of instances above the similarity-dissimilarity line in the context of classification. The value of α_{SD} will be high when the distances among different classes are large or when the classes are very compact.

Table 2. Details of benchmark cancer datasets.

Datasets	Total Samples	Number of Genes	Class Labels	Class wise distribution
AML [60]	54	12625	Remission	28
			Relapse	26
CNS Tumor [62]	60	7129	Survivor	21
			Failures	39
Lung Cancer [64]	203	12600	lung adenocarcinomas (ADEN)	139
			Normal	17
			Squamous cell lung carcinomas (SQUA)	21
			pulmonary carcinoids (COID)	20
			small-cell lung carcinomas (SCLC)	6
Leukemia 1 [1]	72	7129	ALL	47
			AML	25
Leukemia 2 [69]	72	12582	ALL	24
			AML	28
			MLL	20

Table 3. Parameters used for genetic algorithm and MSDI.

Parameter	Description	Value
N_{pop}	Number of Individuals in a population	100
N_{gen}	Total number of generations	400
P_c	Probability of Crossover	0.8
P_m	Probability of Mutation	0.1
β	Penalizing factor	10
k	Number of Nearest neighbors	3
ς	Threshold for neighborhood count	2% of class instances[1]

3. Benchmark Datasets

In this set of benchmark data, we have selected some benchmark datasets representing gene expressions for different type of cancers which are famous in the machine learning community. All attributes are of numeric type. Further details of the dataset is given below,

AML: This dataset consists of 54 AML pediatric patients (age less than 15 years) probed by oligonucleotide microarray containing large number of probes so that genes associated with prognoses of AML patients may be identified. There are two classes, remission and relapsed. Remission class means patient survived more than three years after complete remission and relapsed means failure within one year after complete remission.

CNS: In this dataset, there are 60 patients samples out of each 21 are those patients who have survived after the treatment and 39 patients died after the treatment. There are 7129 probes from 6817 human genes.

Lung: A total of 203 snap-frozen samples are used including 186 lung tumors and 17 normal lung specimens. Expression levels of mRNA corresponding to 12,600 transcript sequences are studied.

Leukemia1: In this set, acute leukemia dataset is based on the analysis of bone marrow samples of adult patients. Whole dataset consists of total of 72 leukemia patients out of which 25 suffer from acute myeloid leukemia (AML) and 47 from acute lymphoblastic leukemia (ALL).

Leukemia2: This dataset is again related to leukemia patients. Whole dataset has 72 leukemia patients and 11225 genes expression profile is used to classify AML, ALL and MLL (Myeloid/Lymphoid Lineage Leukemia).

All of the above datasets are summarized in Table 2 with class distribution.

4. Results and Discussions

Similarity-dissimilarity index (MSDI) is used as fitness function in the evolution of genetic algorithm. Different setting related to Similarity-dissimilarity index and genetic algorithm is summarized in Table 3. Three types of information are extracted from the results, namely, trend of fitness evolution, minimization of features set along with the improvement in the MSDI and different sets of features showing equal performance. Accuracy of every class is predicted and Similarity-dissimilarity plot is analyzed. In the following sub-sections results of our proposed framework is analyzed for all benchmark datasets.

The fitness function for genetic algorithm is defined as,

$$Fitness(j) = f\big(MSDI(j), g(\beta)\big) = MSDI(j)g(\beta)$$

Where $g(\beta) = e^{-\beta\frac{\alpha}{n}}$ and α is the number of features in the subset defined by an individual in the population.

In the subsequent sections, results for different benchmark datasets given in table 2 will be explained and analyzed.

1 If 2% of class instances are less than 3 then ς is set to 3.

4.1. Analysis of AML Dataset

Genetic algorithm is run for 400 generations and trend of fitness is plotted in

Figure 9. Fitness value is increased sharply in the initial generations and then slowly moved towards the maximum fitness value of 1.0. Best fitness value for 400 generations is found to be 0.92 and best MSDI value is 0.94 Minimum and mean value of the size of the feature set is plotted in Figure 10 for all generations. In the initial generations, size of feature set is reduced drastically and then slowly converged to the feature size less than 10. Fitness and size of features subset trend for different generations of genetic algorithm is similar in all other datasets as well. In some datasets, the convergence is rapid while in some datasets, convergence is slow but the trend is same.

MSDI for Remission class is 0.964 whereas it is 0.923 for relapse class. It shows that predicted accuracies of remission and relapse classes are 96.4% and 92.3% respectively. Overall MSDI for both classes is found to be 0.944.

Similarity-dissimilarity plot (Figure 11) for AML dataset shows that most of the instances of both classes are above the SD line. Some instances of both classes are identified as boundary points. There are three instances (two from remission class and one from relapse class) below the SD line. These instances may be misclassified by any classifier.

Gene numbers of the best individual found by the genetic algorithm is given in Table 4. Eight genes are identified as the most important genes to classify two classes correctly. In the later generations, all individuals showing MSDI value of 0.94 have been converged to a single set of eight genes as given in table 4. The feature subset is corresponding to the gene number is given in table 4.

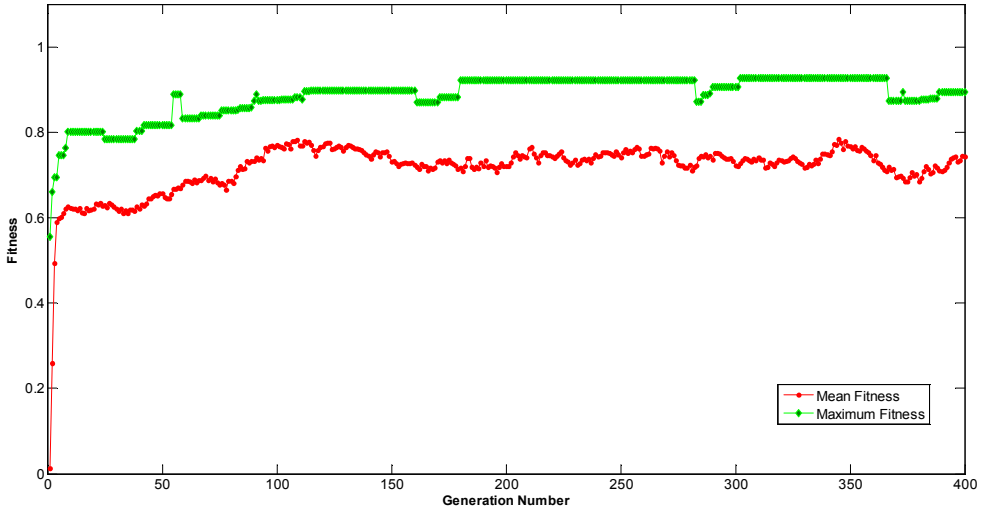

Figure 9. Trend of maximum and mean fitness values versus generation number for AML dataset.

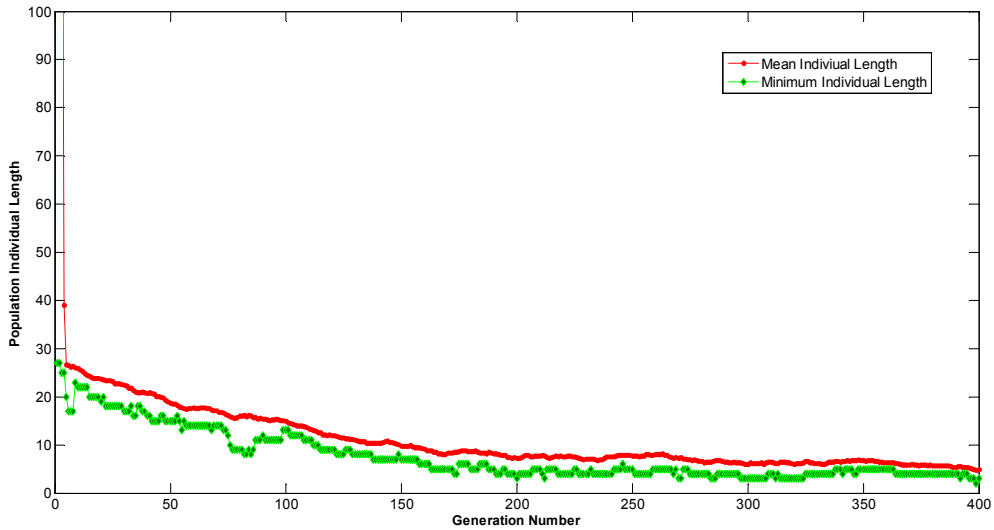

Figure 10. Size of feature set in different generations for AML dataset.

Table 4. Properties of best individuals after 400 generation.

Gen Number	Individual Number	Fitness	MSDI	Mink Distance	Feature set size
366	24	0.92	0.94	1	8

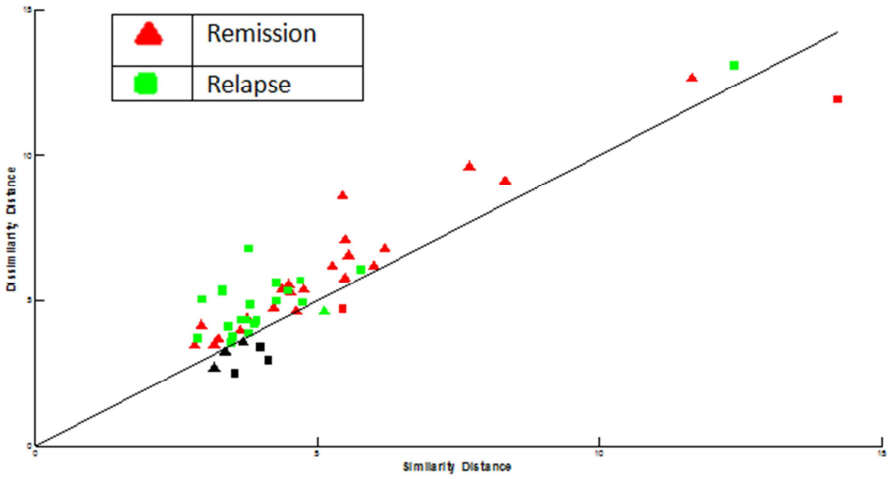

Figure 11. Similarity-Dissimilarity Plot for AML dataset (with 8 genes selected).

4.2. Analysis of CNS Tumor Dataset

SD plot is plotted in Figure 12 and it can be seen that some instances of class 1 (survivor) are below the SD line and values of MSDI are 0.809 and 1.0 for survivor and failure Classes respectively. The dataset contains less number of instances for survivor class (21 instances) as compared to the failure class (39 instances). So it may be assumed that classification accuracy may improve for the survivor class if the number of instances for survivor class is increased. Overall MSDI value is found to be 0.933 for the best feature subset. For Minkowski distance, value of p is found to be 1. For best feature subset overall MSDI value is 0.933 (Table 5). The value of α_{SD} is calculated as 1.08.

Table 5. Best individual after GA optimization for CNS tumor dataset.

Gen Number	Individual Number	Fitness	MSDI	Mink Distance	Feature set size
400	67	0.906	0.933	1	7

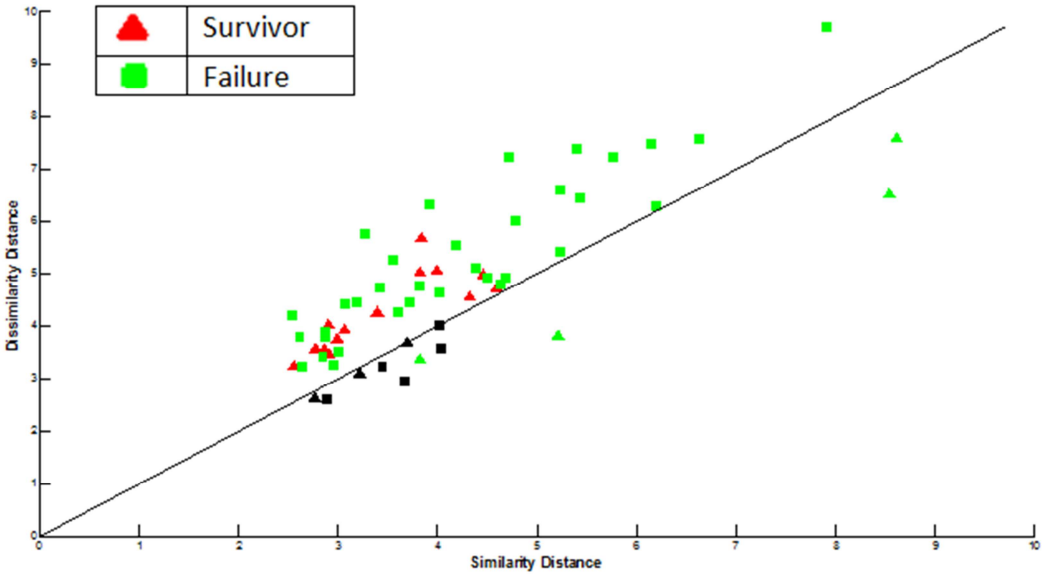

Figure 12. SD plot for CNS tumor dataset

4.3. Analysis of Lung Cancer Dataset

After 400 generations, best size of features subset is found to be 7 as given in the Table 6. SD plot for the dataset is plotted in Figure 13. MSDI values for all five classes are found to be 1.0, 0.94, 0.76, 1.0 and 1.0 for the best feature subset given in the Table 6. Class of SQUA showed low value of MSDI. All instances of SQUA which are below SD line are near to the ADEN class and hence in classification, these instances of SQUA will most probably be confused with ADEN class. One instance of NORMAL class is also near to the ADEN class. Some instances are near the boundary and shown as black on the SD plot.

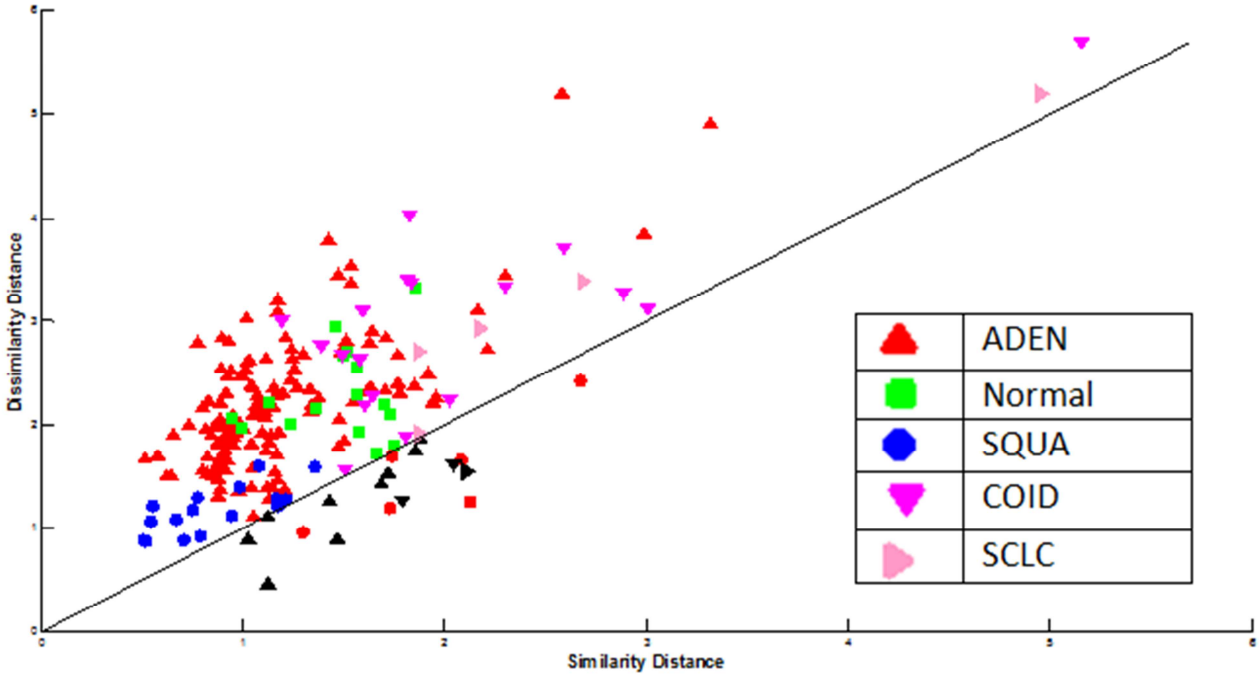

Figure 13. SD plot for Lung Cancer dataset.

Table 6. Best features subset for Lung cancer dataset.

Gen Number	Individual Number	Fitness	MSDI	Mink Distance	Feature Set Size	α_{SD}
399	32	0.9544	0.97	2	7	0.53
400	97	0.9544	0.97	2	7	0.51

Table 7. Best features subset for Leukemia1 dataset.

Gen Number	Individual Number	Fitness	MSDI	Mink Distance	Feature Set Size	α_{SD}
400	100	0.99	1	2	3	0.64

One instance of COID and SCLC each are very far away from their own classes and also from the rest of classes. Overall MSDI for the best feature subset is 0.97 and size of the feature subset is seven. It is predicted that classification accuracy for this subset with an optimal classifier will be about 97%. Minkowski distance parameter is optimized as 2 for the distance metric.

4.4. Analysis of Leukemia 1 Dataset

Leukemia dataset is a very famous dataset and many researchers have used this dataset in proving the classification capabilities of their classifiers. Selection of the best features subset is important. Genetic algorithm converged to fitness value of 0.99, MSDI value of 1.0 and size of features subset is found to be 3 as given in Table 7. SD plot of Best features subset is shown in Figure 14. MSDI values of the individual classes are 1.0 and 1.0 and overall MSDI is also 1.0. In the figure, quality of ALL instances is better as they have lesser similarity distances and larger dissimilarity distances. Very few are away from their own classes and other classes as well.

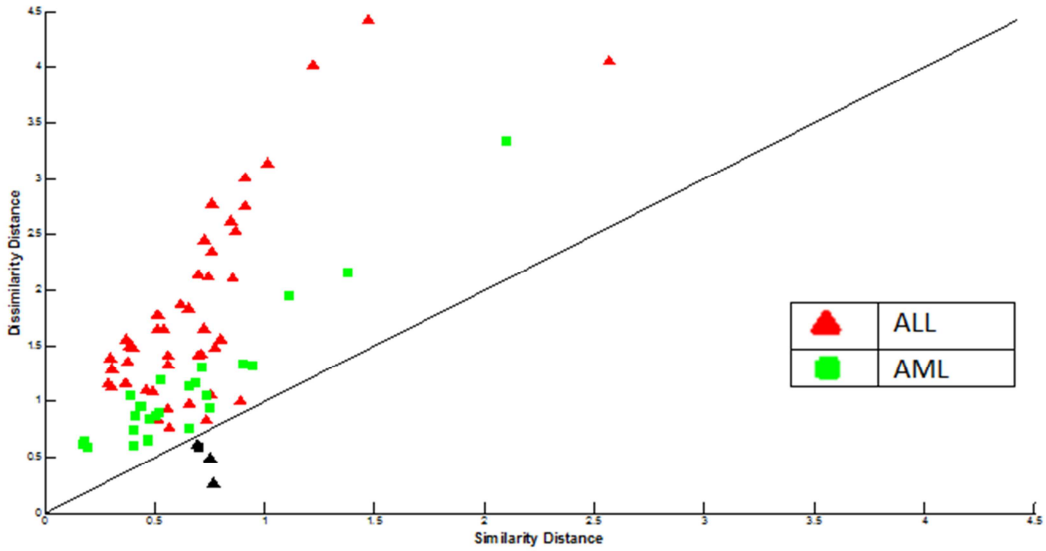

Figure 14. SD plot for Leukemia1 dataset.

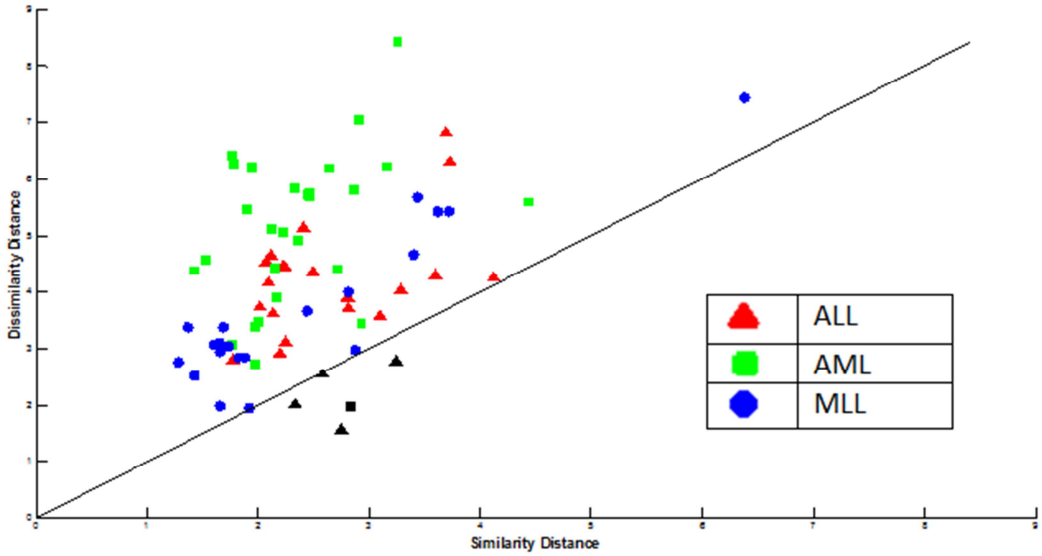

Figure 15. SD plot for Leukemia2 dataset.

Table 8. Best features subset for Leukemia2 dataset.

Gen Number	Individual Number	Fitness	MSDI	Mink Distance	Feature Set Size	α_{SD}
384	1	0.985	1	1	6	1.18
370	38	0.985	1	1	6	1.02
361	98	0.985	1	1	6	1.17
400	2	0.985	1	1	6	1.22

4.5. Analysis of Leukemia 2 Dataset

Leukemia2 dataset consists of three classes, ALL, AML and MLL. Genetic algorithm found four subsets of features, each having five features with same fitness value of 0.985. MSDI is found to be 1.0 and Minkowski distance parameter is 1.0 for this dataset. Details are given in Table 18. The value of α_{SD} is calculated for all these four subsets. The values are almost similar and above 1.0. SD plot of best feature subset is plotted in Figure 15. Most of the instances of three classes are above the SD line and individual MSDI for all three classes are 1.0 predicting 100% classification accuracy with the feature subset of four features. Four instances are below SD line but they are considered as boundary points because of their neighborhood counts.

4.6. Comparison with Published Results

Optimized MSDI results are compared with the published results and tabulated in 9. Classification accuracy is predicted from MSDI value which reflects number of instances above the similarity-dissimilarity line or has high neighborhood counts. For the AML Leukemia dataset, there is only one reported result found in the literature who has reported 94% and 100% classification accuracy considering all features respectively. Predicted accuracy for AML dataset using MSDI is calculated as 94% as well using a feature subset of eight features with Minkowski distance parameter equals to 1. In the dataset of CNS tumor predicted accuracy is 93.3% with 7 features which is consistent with the reported accuracies in the literature. The best accuracy is found to be 98% with 8 features. In case of Lung cancer dataset, predicted accuracy with feature subset of size seven is similar to the range of accuracies and features subset sizes in the literature. Similar trend can be seen for the case of prostate cancer dataset in which best predicted accuracy by genetic algorithm is found to be 98% with only four features subset. Only one reference [98]reported accuracy of 100% using average feature subset value as 3.1 (exact size of feature subset is not known).

Table 9. Comparison of Optimized MSDI results with published results.

Datasets	Authors	Published Results Accuracy (# of features)	Predicted Accuracy (# of features)	Minkowski Distance parameter
AML [Yagi2003]	[71]	94 (All)	94(8)	1
	[72]	73.29 (53)	93.3 (7)	1
	[73]	86 (200)		
CNS Tumor	[74]	90 (36)		
	[75]	77 (2)		
	[76]	98 (8)		
	[98]	95.7 (7)	97 (7)	2
	[77]	97.3 (1)		
Lung Cancer [Bhattacharjee 2001]	[77]	96.6 (2)		
	[83]	99.3 (6)		
	[84]	98 (2)		
	[97]	100 (4), 98 (3)	100 (4)	2
Leukemia 1 [Galoub 1999]	[96]	99.7 (7)		
	[78]	98 (10)		
	[81]	100 (5)		
	[69]	90 (100)	100 (6)	1
Leukemia 2 [Armstrong 2002]	[77]	93 (1)		
	[94]	100 (26)		
	[69]	95 (40)		

For leukemia 1 and Leukemia 2 datasets, proposed framework produced subsets of four and six features respectively. It is worth mentioning that Leukemia1 dataset is a two class classification problem whereas Leukemia2 dataset is a three class classification problem. These subsets of features also expected to produce 100% classification accuracy if Manhattan distance is used as distance metric. All the results reported in the Table 9, showed good agreement with the reported results and in some cases better results were also produced. Some optimized subsets of features are found for almost all the datasets. It shows the effectiveness of the proposed framework in this paper where MSDI along with a penalty function is used as a fitness function in the genetic algorithm. Moreover, different features subsets having similar fitness are also compared by average differential of similarity-dissimilarity distances over SD line index to find out best feature subset.

5. Conclusions

In this paper, a framework is proposed to search an optimal set of features subset that can maximize the quality of feature discrimination in the subspace of selected features in the context of classification using MSDI and a penalizing function. Furthermore, Similarity-dissimilarity plot is also proposed to study the distribution of instances of different classes in the dataset. Important information can be obtained from this plot related to the quality of instances in discriminating different classes. Sparseness or compactness of instances of classes can be predicted. Outliers if present in the dataset can also be identified. The proposed methodology is applied to various datasets comprising of different type of cancers. Effectiveness of the method is proved by comparing our results with some reported results in the literature. Hence it is suggested that this method can be used effectively to find out multiple optimal features subsets according to some pre-

defined criterion or MSDI proposed in this paper. In the future work, this methodology will be extended to mixed type of attributes including numeric, categorical, nominal or binary attributes.

Acknowledgements

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University (project # 43408032) for the financial support.

References

[1] Galoub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloom eld C. D., and Lander E. S. (1999), Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, 286, 531-537.

[2] Alon A., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D., Levine A.J., (1999), Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, Proc. Natl. Acad. Sci. USA, 96, 6745-6750.

[3] Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. (2002), Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97, no. 457 77-87, 2002.

[4] Chen, Zhiping, and Kevin Lü. (2006), A preprocess algorithm of filtering irrelevant information based on the minimum class difference. Knowledge-Based Systems 19, no. 6 422-429.

[5] Jieming Yang, Yuanning Liu, Xiaodong Zhu, Zhen Liu, Xiaoxu Zhang, (2012) A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization, Information Processing & Management, Volume 48, Issue 4, 741-754.

[6] Jieming Yang, Yuanning Liu, Zhen Liu, Xiaodong Zhu, Xiaoxu Zhang, (2011) A new feature selection algorithm based on binomial hypothesis testing for spam filtering, Knowledge-Based Systems, Volume 24, Issue 6, 904-914.

[7] Xu, Ping, Guy N. Brock, and Rudolph S. Parrish. (2009) Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis* 53, no. 5, 1674-1687.

[8] Guo, Y., Hastie, T., Tibshirani, R., (2007) Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8 (1), 86–100.

[9] Hastie, T., Tibshirani, R., (2004) Efficient quadratic regularization for expression arrays. Biostatistics 5 (2), 329–340.

[10] Xiong, Momiao, Wuju Li, Jinying Zhao, Li Jin, and Eric Boerwinkle. (2001) Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism* 73, no. 3, 239-247.

[11] Van der Maaten, L. J. P., E. O. Postma, and H. J. Van den Herik. (2009) Dimensionality reduction: A comparative review. *Journal of Machine Learning Research* 10, 1-41.

[12] Antoniadis, Anestis, Sophie Lambert-Lacroix, and Frédérique Leblanc. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* 19, no. 5, 563-570.

[13] Notterman, D. A., et al., (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 61, 3124–3130.

[14] Bouras, T., et al., (2002) Stanniocalcin 2 is an estrogen-responsive gene coexpressed with the estrogen receptor in human breast cancer. Cancer Res. 62, 1289–1295.

[15] Mootha, V. K., et al., (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. Nat. Genet. 34, 267–273.

[16] Bushel, P. R., et al., (2002) Computational selection of distinct class- and subclass-specific gene expression signatures. J. Biomed. Inform. 35, 160–170.

[17] Haseeb Ahmad Khan, (2013) A novel gene expression index (GEI) with software support for comparing microarray gene signatures, Gene, Volume 512, Issue 1, 82-88.

[18] Su, Yang, T. M. Murali, Vladimir Pavlovic, Michael Schaffer, and Simon Kasif. (2003) RankGene: identification of diagnostic genes based on expression data. *Bioinformatics* 19, no. 12, 1578-1579.

[19] Wu, Baolin, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19, no. 13, 1636-1643.

[20] Levner, Ilya. (2005) Feature selection and nearest centroid classification for protein mass spectrometry. *BMC bioinformatics* 6, no. 1, 68.

[21] A. L. Blum, P. Langley, (1997) Selection of relevant features and examples in machine learning, Artificial Intelligence 97, 245–271.

[22] Chen, Yuehui, Ajith Abraham, and Bo Yang. (2006) Feature selection and classification using flexible neural tree. *Neurocomputing* 70, no. 1, 305-313.

[23] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, no. 19, 2507-2517.

[24] Ando, Tatsuya, Miyuki Suguro, Takeshi Kobayashi, Masao Seto, and Hiroyuki Honda. (2003) Selection of causal gene sets for lymphoma prognostication from expression profiling and construction of prognostic fuzzy neural network models. *Journal of bioscience and bioengineering* 96, no. 2, 161-167.

[25] Chen, Guoan, Tarek G. Gharib, Chiang-Ching Huang, Dafydd G. Thomas, Kerby A. Shedden, Jeremy MG Taylor, Sharon LR Kardia et al. (2002) Proteomic analysis of lung adenocarcinoma identification of a highly expressed set of proteins in tumors. *Clinical Cancer Research* 8, no. 7 (2002): 2298-2305.

[26] Satten, Glen A., Somnath Datta, Hercules Moura, Adrian R. Woolfitt, Maria Da G. Carvalho, George M. Carlone, Barun K. De, Antonis Pavlopoulos, and John R. Barr. (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics* 20, no. 17, 3128-3136.

[27] Adam, Bao-Ling, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, O. John Semmes et al. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* 62, no. 13, 3609-3614.

[28] Xie, Juanying, and Chunxia Wang. (2011) Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications* 38, no. 5, 5809-5815.

[29] Yang, Yee Hwa, Yuanyuan Xiao, and Mark R. Segal. (2005) Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* 21, no. 7, 1084-1093.

[30] Yang, Pengyi, Bing B. Zhou, Zili Zhang, and Albert Y. Zomaya. (2010) A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC bioinformatics* 11, no. Suppl 1, S5.

[31] Liu, Zhenqiu, Feng Jiang, Guoliang Tian, Suna Wang, Fumiaki Sato, Stephen J. Meltzer, and Ming Tan. (2007) Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology* 6, no. 1.

[32] Tibshirani, Robert. (1997) The lasso method for variable selection in the Cox model. *Statistics in medicine* 16, no. 4, 385-395.

[33] Krishnapuram, Balaji, Lawrence Carin, Mario AT Figueiredo, and Alexander J. Hartemink. (2005) Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27, no. 6, 957-968.

[34] Wang, Xiaoming, Taesung Park, and K. C. Carriere. (2010) Variable selection via combined penalization for high-dimensional data analysis. *Computational Statistics & Data Analysis* 54, no. 10, 2230-2243.

[35] Ma, Shuangge, and Jian Huang. (2008) Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics* 9, no. 5, 392-403.

[36] Emmanoullidis C., Hunter A., Macintyre J., (2000) A multi objective evolutionary setting for feature selection and a commonality-based crossover operator, in: 2000 Congress on Evolutionary Computation (CEC' 2000), San Diego, California, July 2000. IEEE Service Center.

[37] P. Baraldi, N. Pedroni, E. Zio, (2009) Application of a niched Pareto genetic algorithm for selecting features for nuclear transients classification, International Journal of Intelligent Systems 24 (2) 118–151.

[38] Ying Li and Keiichi Horio, (2010) Visualization and Analysis of Mental States Based on Photoplethysmogram, ICIC Express Letters, vol.4, no.3(B), pp.923 -928.

[39] Qing Ma and Toshiyuki Kanamaru, (2010) Extraction and Visualization of Numerical and Named Entity Information from a Very Large Number of Documents Using Natural Language Processing, Information and Control, vol.6, no.3(B), pp.1549-1568.

[40] Kenichi Kawasaki, (2009) Study on the Visualization of the Impression of Button Sounds, International Journal of Innovative Computing, Information and Control, vol.5, no.11(B), pp.4189-4204.

[41] D. F. Andrews, (1972) Plot of high dimensional data, Biometrics, 29, 125-136.

[42] J. M. Chambers, W. S. Cleveland , B. Kleiner, P. A. Tukey, (1976) Graphical methods for data analysis, Chapman and Hall.

[43] J. J. van Wijk, R. van Liere, (1993) HyperSlice, Proc. of IEEE Visualization '93, Nielson, G. M., Bergeron, R. D., editors, IEEE Computer Society Press, Los Alamitos, 119-125.

[44] B. Alpern, L. Carter, (1991) Hyperbox, Proc. of IEEE Visualization '91, 133-139.

[45] R. Spence, L. Tweedie, H. Dawkes, H. Su, (1995) Visualisation for Functional Design, Proc of IEEE Visualization '95, 4-10.

[46] A. Inselberg, (1985) The plane with parallel coordinates, The Visual Computer, 69-92.

[47] Inselberg, B. Dimsdale B., (1990) Parallel coordinates: A tool for visualization high dimensional geometry, Proc. of IEEE Visualization, 361-378.

[48] Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, Baoquan Chen, (2008) Visual Clustering in Parallel Coordinates IEEE-VGTC Symposium on Visualization, 27.

[49] W. Peng, M.O. Ward, E.A. Rundensteiner, (2004) Cluster reduction in multi-dimensional data visualization using dimension reordering, Proc of IEEE symposium on Information visualization, 89-96.

[50] J. Johansson, P. Ljung, M. Jern, M. Cooper, (2000) Revealing structures within clustered parallel coordinates display, Proc. of IEEE symposium on Information visualization, 125-132, 2005.

[51] H. Siirtola, Direct manipulation of parallel coordinates, Proc of IEEE 4th International Conference on Information visualization, 373-378.

[52] Brunsdon, A. S. Fotheringham, M. E. Charlton, (1998) An Investigation of Methods for Visualising Highly Multivariate Datasets, In Case studies of Visualization in Social Sciences, 55-80.

[53] G. Leban, I. Bratko, U. Petrovic, T. Curk, B. Zupan, (2005) Vizrank: finding informative data projections in functional genomics by machine learning, Bioinformatics, 21/3, 413-414.

[54] J. F. McCarthy, K.A. Marx, P.E. Hoffman, A. G. Gee, P. O'Neil, M. L. Ujwal J. Hotchkiss, (2004) Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis and management, Annals of New York Academy of Sciences, 1020, 239-262.

[55] B. Zupan, (2007) FreeViz-an intelligent multivariate visualization approach to explorative analysis of biomedical data, Journal of biomedical informatics, 40/6, 661-671.

[56] John Sharko, Georges Grinstein, and Kenneth A. Marx, (2008) Vectorized Radviz and Its Application to Multiple Cluster Datasets, IEEE Transactions on Visualization and Computer Graphics, 14(6), pp 1444-1451.

[57] Arif M, (2012) Similarity-Dissimilarity Plot for Visualization of High Dimensional Data in Biomedical Pattern Classification, Journal of Medical Systems, Journal of Medical Systems, Volume 36, Issue 3, pages 1173–1181, 2012.

[58] Arif M and Saleh Basalamah, (2012) Similarity-Dissimilarity Plot for High Dimensional Data of Different Attribute types in Biomedical Datasets, International Journal of Innovative Computing, Information and Control, Vol 8, No 2, 1275-1298.

[59] Holland, John H. (1992) Genetic algorithms. Scientific American 267, no. 1, 66-72.

[60] Yagi, Tomohito, Akira Morimoto, Mariko Eguchi, Shigeyoshi Hibi, Masahiro Sako, Eiichi Ishii, Shuki Mizutani, Shinsaku Imashuku, Misao Ohki, and Hitoshi Ichikawa. (2003) Identification of a gene expression signature associated with pediatric AML prognosis. Blood 102, no. 5, 1849.

[61] Crossman LC, Mori M, Hsieh YC, Lange T et al. (2005) In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures. Haematologica, 90(4):459-64.

[62] Pomeroy, Scott L., Pablo Tamayo, Michelle Gaasenbeek, Lisa M. Sturla, Michael Angelo, Margaret E. McLaughlin, John YH Kim et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, no. 6870, 436-442.

[63] Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, no. 6769, 503-511.

[64] Bhattacharjee, Arindam, William G. Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd et al. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences 98, no. 24, 13790-13795.

[65] Singh, Dinesh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer cell 1, no. 2, 203-209.

[66] Khan, Javed, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7, no. 6, 673-679.

[67] Dyrskjøt L, Thykjaer T, Kruhøffer M, Jensen JL et al. (2003) Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet, Jan;33(1):90-6.

[68] Hippo Y, Taniguchi H, Tsutsumi S, Machida N et al. (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res, 62(1):233-40.

[69] Armstrong, Scott A., Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. den Boer, Mark D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub, and Stanley J. Korsmeyer. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics30, no. 1, 41-47.

[70] Guyon, Isabelle, Steve Gunn, Asa Ben-Hur, and Gideon Dror. (2004) Result analysis of the nips 2003 feature selection challenge. Advances in Neural Information Processing Systems 17, 545-552.

[71] Gasparovica, Madara, Ludmila Aleksejeva, and Valdis Gersons. (2012) The Use of BEXA Family Algorithms in Bioinformatics Data Classification. Information Technology and Management Science 15, no. 1, 120-126.

[72] Zhu, Zexuan, Yew-Soon Ong, and Manoranjan Dash. (2007) Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 40, no. 11, 3236-3248.

[73] Huang, Chenn-Jung, and Wei-Chen Liao. (2004) Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. Neural Processing Letters 19, no. 3, 211-226.

[74] Piao, Yongjun, Minghao Piao, Kiejung Park, and Keun Ho Ryu. (2012) An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics 28, no. 24, 3306-3315.

[75] Wang, Xiaosheng. (2012) Robust two-gene classifiers for cancer prediction. Genomics, 99(2):90-5.

[76] Liu, Huawen, Lei Liu, and Huijie Zhang. (2010) Ensemble gene selection for cancer classification. Pattern Recognition 43, no. 8, 2763-2772.

[77] Wang, Xiaosheng, and Osamu Gotoh. (2009) Accurate molecular classification of cancer using simple rules. BMC medical genomics 2, no. 1, 64.

[78] Sounak Chakraborty. (2009) Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Comput. Stat. Data Anal.* 53, 12, 4198-4209.

[79] J. H. Cho, D. Lee, J. H. Park, I. B. Lee, (2004) Gene selection and classification from microarray data using kernel machine, FEBS Lett. 571, 93–98.

[80] K. Deb, A. R. Reddy, (2003) Reliable classification of two-class cancer data using evolutionary algorithms, Biosystems 72, 111–129.

[81] Lee, Chien-Pang, and Yungho Leu. (2011) A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing 11, no. 1, 208-213.

[82] L. Li, C. R. Weinberg, T. A. Darden, L. G. Pedersen, (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA / KNN method, Bioinformatics 17, 1131–1142.

[83] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R: (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*, 62(17):4963-4967

[84] Momin BF, Mitra S: (2006) Reduct generation and classification of gene expression data.In *proceedings of the First International Conference on Hybrid Information Technology: 9-11 November 2006; Jeju Island*. Edited by Szczuka MS, Howard D, Slezak D, Kim HK, Kim TH, Ko IS, Lee G, Sloot PMA. Berlin/Heidelberg: Springer; 699-708.

[85] Geman D, d'Avignon C, Naiman DQ, Winslow RL: (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, 3:Article 19.

[86] Tan, Jun-Yan, Chun-Hua Zhang, and Nai-Yang Deng. (2010) Cancer related gene identification via p-norm support vector machine. In The 4th international conference on computational systems biology, pp. 101-108.

[87] Aksu, Yaman. (2012) A Fast SVM-based Feature Selection Method, Combining MFE (Margin-Maximizing Feature Elimination) and Upper Bound on Misclassification Risk. arXiv preprint arXiv:1210.4460.

[88] Chang, Fu, and Chan-Cheng Liu. (2012) Ranking and selecting features using an adaptive multiple feature subset method. number TR-IIS-12-005, Institute of Information Science, Academia Sinica.

[89] J. Deutsch, (2003) Evolutionary algorithms for finding optimal gene sets in microarray prediction, Bioinformatics 19 (1), 45–52.

[90] Seo, Minseok AND Oh, Sejong CBFS, (2012) High Performance Feature Selection Algorithm Based on Feature Clearness, PLoS ONE, 7, e40419.

[91] Cohen, Shay, Gideon Dror, and Eytan Ruppin. (2007) Feature selection via coalitional game theory. Neural computation 19, no. 7, 1939-1961.

[92] Gaudel, Romaric, and Michele Sebag. (2010) Feature selection as a one-player game. In International Conference on Machine Learning, pp. 359-366.

[93] Neal, Radford, and Jianguo Zhang. (2006) High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees. Feature Extraction, 265-296.

[94] Wang Y, Makedon FS, Ford JC, Pearlman J: HykGene: (2005) A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, 21(8):1530-1537.

[95] Lin, Tsun-Chen, Ru-Sheng Liu, Chien-Yu Chen, Ya-Ting Chao, and Shu-Yuan Chen. (2006) Pattern classification in DNA microarray data of multiple tumor types.Pattern Recognition 39, no. 12, 2426-2438.

[96] Asim Roy, Patrick D. Mackin, Somnath Mukhopadhyay, (2013) Methods for pattern selection, class-specific feature selection and classification for automated learning, Neural Networks, ISSN 0893-6080, 10.1016/j.neunet.2012.12.007.

[97] Shu-Lin Wang, Xueling Li, Shanwen Zhang, Jie Gui, De-Shuang Huang, (2010) Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction, Computers in Biology and Medicine, Volume 40, Issue 2, Pages 179-189.

[98] Hui-Ling Huang, Fang-Lin Chang, (2007) ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data, Biosystems, Volume 90, Issue 2, Pages 516-528.