International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:8, No:10, 2014

Statistical Measures and Optimization Algorithms for Gene Selection in Lung and Ovarian Tumor

C. Gunavathi, K. Premalatha

***Abstract*—**Microarray technology is universally used in the study of disease diagnosis using gene expression levels. The main shortcoming of gene expression data is that it includes thousands of genes and a small number of samples. Abundant methods and techniques have been proposed for tumor classification using microarray gene expression data. Feature or gene selection methods can be used to mine the genes that directly involve in the classification and to eliminate irrelevant genes. In this paper statistical measures like T-Statistics, Signal-to-Noise Ratio (SNR) and F-Statistics are used to rank the genes. The ranked genes are used for further classification. Particle Swarm Optimization (PSO) algorithm and Shuffled Frog Leaping (SFL) algorithm are used to find the significant genes from the top-m ranked genes. The Naïve Bayes Classifier (NBC) is used to classify the samples based on the significant genes. The proposed work is applied on Lung and Ovarian datasets. The experimental results show that the proposed method achieves 100% accuracy in all the three datasets and the results are compared with previous works.

***Keywords*—**Microarray, T-Statistics, Signal-to-Noise Ratio, F-Statistics, Particle Swarm Optimization, Shuffled Frog Leaping, Naïve Bayes Classifier.

I. Introduction

RAPID and recent advances in microarray gene expression technology have facilitated the simultaneous measurement of the expression levels of tens of thousands of genes in a single experiment at a reasonable cost. Gene expression profiling by microarray method has been appeared as a capable technique for classification and diagnostic prediction of tumor.

The raw microarray data are images that are transformed into gene expression matrices. The rows in the matrix correspond to genes, and the columns represent samples or trial conditions. The number in each cell signifies the expression level of a particular gene in a particular sample or condition [1], [2]. Expression levels can be absolute or relative. If two rows are similar, it implies that the respective genes are co-regulated and perhaps functionally related. By comparing samples, differentially expressed genes can be identified. The major limitation of the gene expression data is its high dimension which contains more number of genes and very few samples. A number of gene selection methods have been introduced to select the informative genes for tumor prediction and diagnosis. Feature or Gene selection methods remove irrelevant and redundant features to improve classification accuracy.

Particle Swarm Optimization (PSO) is one of the Swarm Intelligence (SI) optimization techniques simulates the behaviour of bird flocking. It is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems. Shuffled Frog Leaping (SFL) is swarm intelligence based sub-heuristic computation optimization algorithm used to solve discrete combinatorial optimization problem.

From the microarray gene expression data, the informative genes are identified based on the statistical measures like T-Statistics, Signal-to-Noise Ratio (SNR) and F-Statistics values. The initial candidate solutions of the PSO and SFL are obtained from top-m informative genes. The classification accuracy of kNN is used as the fitness function for the optimization algorithms. Naïve Bayes Classifier is used as the classifier to classify the given samples. The classification accuracy is obtained from 5-fold cross validation.

II. Related Work

In this section the works related with Gene selection and tumor classification using microarray gene expression data are discussed. An evolutionary algorithm is used to identify the near-optimal set of predictive genes that classify the data [3]. Self-organizing map for clustering tumor data which composed of important gene selection step is used by [4]. Rough set concept with depended degrees was proposed in [5]. In this method they screened a small number of informative single gene and gene pairs on the basis of their depended degrees.

A Swarm Intelligence feature selection algorithm was proposed based on the initialization and update of only a subset of particles in the swarm as in [6]. Gene Doublets concept was introduced based on the gene pair combinations [7]. A new Ensemble Gene Selection method was applied to choose multiple gene subsets for classification purpose, where the significant degree of gene was measured by conditional mutual information or its normalized form [8].

A hybrid method was proposed in which correlation-based feature selection and the Taguchi chaotic binary PSO is used for significant gene selection [9]. Hyper-Box Enclosure (HBE) method based on mixed integer programming for the classification of some tumor types with a minimal set of predictor genes is used in [10]. The use of single gene was explored to construct classification model in [11].

An efficient feature selection approach based on statistically defined effective range of features for every class termed as

C.Gunavathi is with the K. S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India (e-mail: sssguna@gmail.com).

K. Premalatha is with the Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India (e-mail: kpl_barath@yahoo.co.in).

Effective Range based Gene Selection (ERGS) was proposed in [12]. BioMarker Identifier (BMI), which identified features with the ability to distinguish between two data groups of interest, was suggested by [13]. Margin Influence Analysis (MIA) was an approach designed to work with SVM for selecting informative genes in [14]. A model for feature selection using Signal-to-Noise Ratio (SNR) ranking was proposed in [15].

An improved Semi-Supervised Local Fisher discriminant (iSELF) analysis for gene expression data classification is introduced in [16]. A method that relaxed the maximum accuracy criterion to select the combination of attribute selection and classification algorithm is introduced by [17]. A quantitative measure based on mutual information that incorporates the information of sample categories to measure the similarity between attributes was proposed by [18]. A feature selection algorithm which divides the genes into subsets to find the informative genes was proposed in [19].

III. Statistical Gene Selection Methods

Feature selection methods are used to rank the informative genes from the microarray data. The statistical feature selection methods used for significant gene selection are discussed here.

A. T-Statistics

Genes, who have considerably different expressions involving normal and tumor tissues, are candidates for selection. A simple T-statistic measure given in (1) is used by to find the degree of gene expression difference, between normal and tumor tissues [20]. The top-m genes with the largest T- statistic are selected for inclusion in the discriminant analysis.

$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{v1}{n1} + \frac{v2}{n2}}} \quad (1)$$

Here

$\overline{x1}$ - Mean of Normal samples
$\overline{x2}$ - Mean of Tumor samples
$n1$ - Normal Sample size
$n2$ - Tumor Sample size
$v1$ - variance of Normal samples
$v2$ - variance of Tumor samples

B. Signal-to-Noise Ratio

An important measure used to find the significance of genes is the Pearson Correlation Co-efficient. According to Golub et al. [1] it is changed to emphasize the 'Signal-to-Noise Ratio' in using a gene as a predictor. This predictor is shaped with the purpose of finding the Prediction Strength of a particular gene by [21]. The Signal-to-Noise ratio PS of a gene 'g' is calculated by (2).

$$PS(g) = \frac{\overline{x1} - \overline{x2}}{s1 + s2} \quad (2)$$

Here

$\overline{x1}$ – Mean of Normal samples
$\overline{x2}$ - Mean of Tumor samples
$s1$ - Standard Deviation of Normal samples
$s2$ - Standard Deviation of Tumor samples

This value is used to reveal the difference between the classes relative to the standard deviation within the classes. Large values of PS (g) indicate a strong correlation between the gene expression and the class distinction, while the sign of PS (g) being positive or negative corresponds to *g* being more highly expressed in class 1 or class 2. Genes with large SNR value are informative and are selected for tumor classification.

C. F-Statistics

F-Statistics is the ratio of the variances of the given two set of values which is used to test if the standard deviations of two populations are equal or if the standard deviation from one population is less than that of another population. In this work two-tailed F-Statistics value is used to check the variances of Normal Samples and Tumor Samples. Formula to calculate the F-Statistics value of a gene is given in (3). Top-m genes with the smallest F-Statistics value are selected for inclusion in the further analysis.

$$F = \frac{v1}{v2} \quad (3)$$

Here

$v1$ - Variance of Normal Samples
$v2$ - Variance of Tumor Samples

IV. Optimization Algorithms

A. Particle Swarm Optimization

PSO is one of the Swarm Intelligence techniques simulate the behavior of bird flocking [22]. It is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems.

PSO is inspired by the flocking behavior of birds. Particles in the swarm fly through an environment following the swarm members and biasing their movement toward historically good position of their environment. It is a population-based search algorithm and is initialized with a population of random solutions, called particles. Each particle in PSO is associated with a velocity. Particles fly through the search space with velocities which are dynamically adjusted according to their historical behaviors. Therefore, the particles have the tendency to fly towards the better search area over the course of search process.

In PSO, each single solution is like a 'bird' in the search space, which is called 'particle'. All particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. The particles fly through the problem space by following the particles with the best solutions so far.

The original PSO formulae define each particle as potential solution to a problem in N-dimensional space. The position of

particle i is represented as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$. Each particle also maintains a memory of its previous best position, represented as $P_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$. A particle in a swarm is moving; hence, it has a velocity, which can be represented as $V_i = (v_{i1}, v_{i2}, \ldots, v_{iN})$.

Each particle knows its best value so far (pbest) and the best value so far in the group (gbest) among pbests. This information is useful to know how the other particles around them have performed. Each particle tries to modify its position using the following information:

1. The distance between the current position and pbest
2. The distance between the current position and gbest

This modification can be represented by the concept of velocity. Velocity of each agent can be modified by (4). The inclusion of an inertia weight in the PSO algorithm was first reported in the literature [22].

$$V_{id} = w \times V_{id} + c_1 \times \text{rand}() \times (P_{id} - X_{id}) + c_2 \times \text{rand}() \times (P_{gd} - X_{id}) \quad (4)$$

where

- i - Index of the particle , $i \in \{1, \ldots, n\}$
- N - Population size
- d - Dimension, $d \in \{1, \ldots, N\}$
- rand( ) - Uniformly distributed random variable between 0 and 1
- V_{id} - Velocity of particle *i* on dimension *d*
- X_{id} - Current position of particle i on dimension *d*
- c_1 - Determines the relative influence of the cognitive component; Self confidence factor
- c_2 - Determines the relative influence of the social component; Swarm confidence factor
- P_{id} - Personal best or pbest of particle i
- P_{gd} - Global best *or* gbest of the group
- w - Inertia weight

The use of the inertia weight w has provided improved performance in a number of applications. As originally developed, w often is decreased linearly from about 0.9 to 0.4 during a run. Suitable selection of the inertia weight provides a balance between global and local exploration and exploitation, and results in less iteration on average to find a sufficiently optimal solution.

The constants c_1 and c_2 are known as learning factors. They represent the weighting of the stochastic acceleration terms that pull each particle towards the pbest and gbest positions. Thus, adjustments of these constants change the amount of stress in the system. Low values allow particles to travel far from target regions before being pulled back, while high values result in unexpected movement toward, the target regions. The cognitive parameter represents the tendency of individuals to duplicate past behaviors that have proven successful, whereas the social parameter represents the tendency to follow the success of others. Generally c_1 and c_2 are set to 2.0 which will make the search cover surrounding regions centered at pbest and gbest. Also, if the learning factors are equal, the same importance is given to social searching and cognitive searching.

The current position that is the searching point in the solution space can be modified by (5).

$$X_{id} = X_{id} + V_{id} \quad (5)$$

All swarm particles tend to move towards better positions; hence, the best position (i.e. optimum solution) can eventually be obtained through the combined effort of the whole population. The PSO algorithm is simple, easy to implement and computationally efficient.

PSO Algorithm

```
For each particle
   Initialize particle
End
Do
    For each particle
          Calculate fitness value
          If the fitness value is better than its personal best
          Set current value as the new pbest
   End
    Choose the particle with the best fitness value of all as
    gbest
   For each particle
          Calculate particle velocity according equation (4)
      Update particle position according equation (5)
   End
While maximum iterations or minimum error criteria is not attained
```

B. Shuffled Frog Leaping

SFL is swarm intelligence based sub-heuristic computation optimization algorithm proposed to solve discrete combinatorial optimization problem [23]. A group of frogs leaping in a swamp is considered and the swamp has a number of stones at distinct locations on to which the frogs can leap to find the stone that has the maximum amount of available food. The frogs are allowed to communicate with each other, so that they can improve their memes using other's information. An individual frog's position is altered by changing the leaping steps of each frog which improves a meme results.

The search begins with a randomly selected population of frogs covering the entire swamp. The population is partitioned into several parallel groups (memeplexes) that are permitted to evolve independently to search the space in different directions. Within each memeplex, the frogs are infected by other frog's ideas; hence they experience a memetic evolution.

Memetic evolution progresses the quality of the meme of an individual and enhances the individual frog's performance towards a goal. To ensure that the infection process is competitive, it is required that frogs with better memes (ideas) contribute more to the development of new ideas than frogs with poor ideas. During the evolution, the frogs may change their memes using the information from the memeplex best or the best of the entire population. Incremental changes in memotype(s) correspond to a leaping step size and the new meme corresponds to the frog's new position. After an individual frog has improved its position, it is returned to the

community. The information gained from a change in position is immediately available to be further improved upon.

After a certain number of memetic evolution time loops, the memeplexes are forced to mix and new memeplexes are formed through a shuffling process. This shuffling enhances the quality of the memes after being infected by frogs from different regions of the swamp. Migration of frogs accelerates the searching procedure sharing their experience in the form of infection and it ensures that the cultural evolution towards any particular interest is free from regional bias.

Here, the population consists of a set of frogs (solutions) that is partitioned into subsets referred to as memeplexes. The different memeplexes are considered as different cultures of frogs, each performing a local search. Within each memeplex, the individual frogs hold ideas, that can be influenced by the ideas of other frogs, and evolve through a process of memetic evolution. After a defined number of memetic evolution steps, ideas are passed among memeplexes in a shuffling process. The local search and the shuffling processes continue until defined convergence criteria are satisfied. An initial population of P frogs is created randomly. For S-dimensional problems (*S* variables), a frog *i* is represented as $x_i = (x_{i1}, x_{i2}, \ldots, x_{iS})$. Afterwards, the frogs are sorted in a descending order according to their fitness. Then, the entire population is divided into *m* memeplexes, each containing *n* frogs ($P_{m\times n}$). In this process, the first frog goes to the first memeplex, the second frog goes to the second memeplex, frog *m* goes to the m^{th} memeplex, and frog *m+1* goes back to the first memeplex, etc. Within each memeplex, the frogs with the best and the worst fitnesses are identified as X_b and X_w, respectively. Also, the frog with the global best fitness is identified as X_g. Then, a process similar to PSO is applied to improve only the frog with the worst fitness (not all frogs) in each cycle.

SFL Algorithm

Generate random population of P solutions (frogs);
Calculate fitness function f value of each frog;
Repeat for specific number of times
 Sort the population P in descending order of their fitness;
 Divide P into m memeplexes;
 Repeat for specific number of iterations
  For each memeplex determine the best and worst frogs X_b and X_w;
   Identify the best frog for the entire population X_g;
   Improve the worst frog position using
   $X_w(t+1) = rand( ) \times (X_b(t)- X_w(t))$
    If $f(X_w(t+1)) < f(X_w(t))$
     $X_w(t+1) = rand( )\times (X_g(t)- X_w(t))$
     if $f(X_w(t+1) < f(X_w(t))$
     generate the random solution for $X_w(t+1)$
    end;
   Combine the evolved memeplexes;
  end;
 Present the best frog X_g
end;

V. Naïve Bayes Classifier

Naïve–Bayes Classifier (NBC) is widely used for classification in machine learning. It is used mostly because of its simplicity and classification accuracy as compared to other supervised learning methods. The Naïve Bayes method (NB) is a simple approach to probabilistic induction that has been successfully applied in a number of machine learning applications [24]. It follows Bayes theorem with strong (Naïve) independence assumptions. NBC assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. This assumption is called class conditional independence. In this work NBC is used as the classifier.

VI. Proposed Method

The proposed approach is based on PSO and SFL along with NBC on the top-m genes (individuals). Fig. 1 gives the Schematic representation of the proposed method.

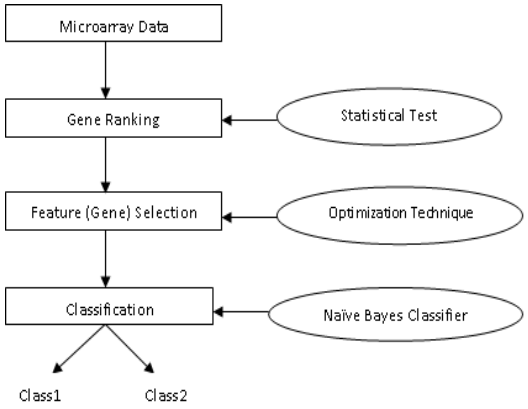

Fig. 1 Schematic representation of the proposed method

A. Particle (or) Frog Representation

The particle (or) frog should contain the information which represents the solution. Fig. 2 shows the candidate solution representation in PSO and SFL. The most used way of encoding the feature selection is a binary string in which '1' or '0' is used to mark whether the gene is selected or not. The random values are generated for gene position. The genes are considered when the value in its position is greater than 0.5, otherwise it is ignored.

g_1	g_2	g_3	g_4	…	g_{n-1}	g_m
0.25	0.56	0.12	0.98	---	0.43	0.112

Fig. 2 Particle (or) Frog representation

B. Fitness Function

The kNN is an instance-based classifier which works on the assumption that classification of unknown instances can be identified by relating the unknown to the known instances according to some distance or similarity measure. The accuracy of kNN is used as the fitness function as in (6) for

the optimization algorithms [25], [26]. The fitness function fitness(x) is defined as

$$fitness(x) = Accuracy(x) \quad (6)$$

Accuracy(x*)* is the test accuracy of testing data *x* in the kNN classifier which is built with the feature subset selection of training data. The classification accuracy of kNN is given in (7).

$$Accuracy(x) = (c/t) \times 100 \quad (7)$$

Here

c - Samples that are classified correctly in test data bykNN Technique

t - Total number of Samples in test data

C. K-Fold Cross-Validation

k-fold cross-validation is used for the result to be more valuable. In *k*-fold cross-validation, the original sample is divided into random *k*-subsamples; one among them is kept as the validation data for testing. The remaining *k*-1 sub-samples are used for training. The cross-validation process is repeated for *k*-times (the folds), with each of the *k* sub-samples used exactly once as the validation data. The average of *k* results from the folds gives the test accuracy of the algorithm. In order to achieve a reliable performance of the classifier, the 5-fold cross-validation method is used in this proposed method.

VII. Results and Discussion

In order to evaluate the performance of the proposed method, three datasets were analyzed. The datasets were collected from Kent Ridge Biomedical Data Repository. The details are given in Table I. In the columns (Class1 and Class2) of Table I, the number within the bracket denotes the number of samples. The Parameters and their values of PSO and SFL are shown in Table II.

TABLE I
MICROARRAY GENE EXPRESSION DATASETS

Dataset Name	Number of Genes	Class1	Class2	Total Samples
Lung Michigan	7129	Tumor (86)	Normal (10)	96
Ovarian Cancer	15154	Normal (91)	Cancer (162)	253
Lung Harvard2	12533	ADCA (150)	Mesothelioma (31)	181

From the microarray data, informative genes are identified based on T-statistics, Signal-to-Noise Ratio and F-Statistics values. The initial candidate solutions of PSO and SFL are obtained from the top-m informative genes. The m values are taken as 10, 50 and 100. The selected genes are used for further classification. The classification accuracy of kNN is used as the fitness function for PSO and SFL. By empirical analysis the value of k is assigned as 5. The 5-fold cross validation method is used to validate the classification accuracy.

TABLE III
PARAMETERS AND THEIR VALUES

Parameter	Value
Particle (or) Frog size	10, 50, 100
Number of memeplex (m)	10
Number of frogs in each memeplex (n)	5
population size	50
Maximum no. of Generations	200
Shuffling iteration	20
w	0.9
c_1	2.1
c_2	2.1
Distance Measure in kNN	Euclidean distance
k-value is kNN	5

Tables III and IV show the results obtained from PSO and SFL based methods. The tables contain the average classification accuracy with top-m genes using NBC.

TABLE IIIII
EXPERIMENTAL RESULTS OF PSO

Statistical Measure		Classification Accuracy (%)		
		Lung Michigan	Lung Harvard2	Ovarian Cancer
T-Statistics	Top-10	95.65	77.5	70.68
	Top-50	95.65	87.5	75.86
	Top-100	100	85	79.31
SNR	Top-10	95.65	85	98.27
	Top-50	95.65	100	100
	Top-100	100	100	98.27
F-Statistics	Top-10	95.65	97.5	96.55
	Top-50	100	97.5	100
	Top-100	100	100	100

TABLE IVV
EXPERIMENTAL RESULTS OF SFL

Statistical Measure		Classification Accuracy (%)		
		Lung Michigan	Lung Harvard2	Ovarian Cancer
T-Statistics	Top-10	91.3	80	70.69
	Top-50	91.3	85	70.69
	Top-100	95.65	87.5	70.69
SNR	Top-10	100	97.5	98.27
	Top-50	100	100	98.27
	Top-100	100	100	100
F-Statistics	Top-10	95.65	96.55	96.55
	Top-50	100	100	100
	Top-100	100	100	100

Figs. 3 and 4 show the results obtained from NBC through the feature selection methods PSO and SFL for top-10, top-50 and top-100 genes obtained from T-statistics, SNR and F-Statistics. These results show that for all the three datasets 100% accuracy is achieved by PSO and SFL. From the results it is inferred that the m value does not influence the accuracy of the classifier. So the value of m should be identified through empirical analysis.

Tables V-VII give the comparison of the proposed method with existing methods. Experimental results show that proposed method gives better performance compared to previous works.

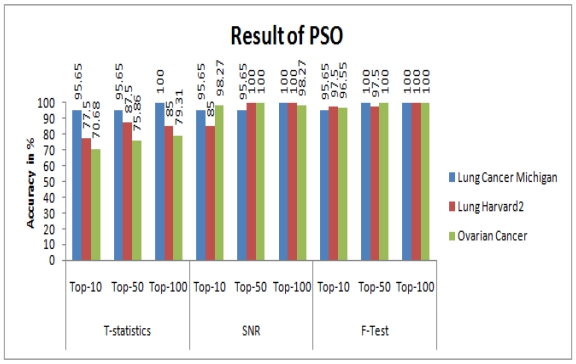

Fig. 3 Results of Particle Swarm Optimization Algorithm

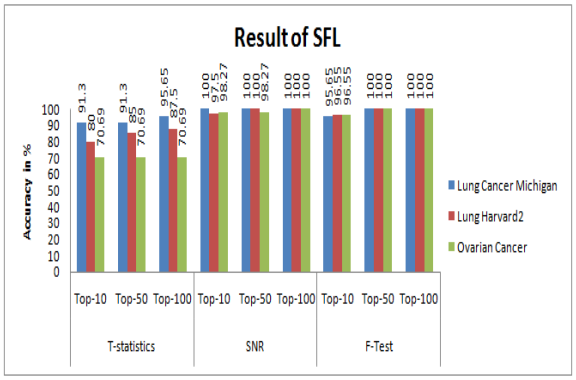

Fig. 4 Results of Shuffled Frog Leaping Algorithm

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR LUNG CANCER MICHIGAN

Reference	Methodology	Average Classification Accuracy in percentage
[17]	Combination of attribute selection and classification algorithm	100
[8]	EGS - Ensemble Gene Selection Method	89.58
This work	PSO + NBC	100
This work	SFL + NBC	100

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR LUNG HARVARD2

Reference	Methodology	Average Classification Accuracy in percentage
[17]	Combination of attribute selection and classification algorithm	99.63
[12]	Effective Range based Gene Selection	100
[11]	Univariate class discrimination with single gene	99
[7]	Based on Gene doublets	100
[5]	Rough sets	97.32
[4]	Gene selection step and clustering tumor data by using self-organizing map	100
This work	PSO + NBC	100
This work	SFL + NBC	100

TABLE VII
COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS FOR OVARIAN CANCER

Reference	Methodology	Average Classification Accuracy in percentage
[17]	Combination of attribute selection and classification algorithm	100
This work	PSO + NBC	100
This work	SFL + NBC	100

VIII. CONCLUSION

Tumor classification using gene expression data is an important task to deal with the problem of tumor prediction and diagnosis. For an effective and precise classification, investigations of feature selection methods are crucial. This article compares the performance of Particle Swarm Optimization (PSO) and Shuffled Frog Leaping (SFL) in microarray based tumor classification. T-statistics, Signal-to-Noise Ratio and F-Statistics are the feature selection methods used to rank the genes. PSO and SFL, with Naïve Bayes Classifier (NBC) method is applied on the top-m genes in this research work. Here the classification accuracy of kNN is considered as the fitness function. The performance of this hybrid method is tested with three different cancer datasets. The experimental results show that both PSO and SFL optimization algorithms works well. With the help of the selected genes, both PSO and SFL achieve 100% accuracy in all the three datasets.

The optimization techniques based feature selection methods are simple and can be easily combined with other statistical feature selection methods. The experiment results are demonstrated on well-known gene expression datasets. In the datasets, the proposed works using PSO and SFL perform well in distinguishing the classes of tumor. These simple models based on statistical measures and optimization techniques perform two level of feature selection to get the most informative genes for classification process.

REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, Vol. 286, no. 5439, pp. 531 - 537, 1999.
[2] E. Domany, "Cluster analysis of gene expression data," J Stat Phys, vol. 110, pp. 1117-1139, 2003.
[3] T. Umpai, A. Stuart, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," BMC Bioinformatics, vol. 6, no. 148, 2005.
[4] S. Vanichayobon, W. Siriphan, and W. Wiphada, "Microarray Gene Selection Using Self-Organizing Map," in Proceedings of the seventh WSEAS International Conference on Simulation, Modelling and Optimization, Beijing, China, 2007.
[5] X. Wang and O. Gotoh, "Accurate molecular classification of cancer using simple rules," BMC Medical Genomics, vol. 2, no. 64, 2009.
[6] E. Martinez, M.A. Mario, and T. Victor, "Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm," Computational Biology and Chemistry, vol. 34, pp. 244 – 250, 2010.
[7] P. Chopra, J. Lee, J. Kang, and S. Lee, "Improving Cancer Classification Accuracy Using Gene Pairs," PLoS ONE, vol. 5, no. 12, 2010.
[8] H. Liu, L. Lei, and H. Zhang, "Ensemble gene selection for cancer classification," Pattern Recognition, vol. 43, pp. 2763 – 2772, 2010.
[9] C. Li-Yeh, Y. Cheng-San, W. Kuo-Chuan, and Y. Cheng-Hong, "Gene selection and classification using Taguchi chaotic binary particle swarm

optimization," Expert Systems with Applications, vol. 38, pp. 13367 – 13377, 2011.
[10] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli, and M. Turkay, "Optimization Based Tumor Classification from Microarray Gene Expression Data," PLoS ONE, vol. 6, no. 2, 2011.
[11] X. Wang and R. Simon," Microarray-based cancer prediction using single Genes," BMC Bioinformatics, vol. 12, no. 391, 2011.
[12] B. Chandra and M. Gupta, "An efficient statistical feature selection for classification of gene expression data," Journal of Biomedical Informatics, vol. 44, pp. 529 – 535, 2011.
[13] I.H. Lee, H.L. Gerald, and V. Mahesh, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," Journal of Clinical Bioinformatics, vol. 1, no. 11, 2011.
[14] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, B.B. Tan, B.C. Deng, and C.C. Lin, "Recipe for Uncovering Predictive Genes Using Support Vector Machines Based on Model Population Analysis," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 6, pp. 1633 – 1641, 2011.
[15] D. Mishra and B. Sahu, "Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach," International Journal of Scientific & Engineering Research, vol. 2, no. 4, 2011.
[16] H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis," Expert Systems with Applications, vol. 39, pp. 2314 – 2320, 2012.
[17] G.C.J. Alonso, I.Q. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods," Expert Systems with Applications, vol. 39, pp. 7270 –7280, 2012.
[18] M. Pradipta, "Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 127 - 140, 2012.
[19] A. Sharma, I. Seiya, and M. Satoru, "A Top-R Feature Selection Algorithm For Microarray Gene Expression Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754 – 764, 2012.
[20] K. Yendrapalli, R. Basnet, S. Mukkamala, and A.H. Sung, "Gene Selection for Tumor Classification Using Microarray Gene Expression Data," in Proceedings of the World Congress on Engineering, vol. I, 2007.
[21] X. Momiao, W. Li, J. Zhao, J. Li, and B. Eric, "Feature (Gene) Selection in Gene Expression-Based Tumor Classification," Journal of Molecular Genetics and Metabolism, vol. 73, pp. 239–247, 2001.
[22] R.C. Eberhart, Y. Shi, "Comparison between Genetic Algorithms and Particle Swarm Optimization, Evolutionary Programming VII," Lecture Notes in Computer Science, Springer New York, vol. 1447, pp. 611-616, 1998.
[23] M. Eusuff, K. Lansey, "Optimization of Water Distribution Network Design Using Shuffled Frog Leaping Algorithm," Journal of Water Resources Planning and Management vol. 129, no. 3, pp 210 – 225, 2003.
[24] R.O. Duda, P.E. Hart, "Pattern Classification and Scene Analysis," New York: John Wiley and Sons, 1973.
[25] N.S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, no. 3, pp. 175-185, 1992.
[26] M.S. Mohamed, D. Safaai, and R.O. Muhammad, "Genetic Algorithms wrapper approach to select informative genes for gene expression microarray classification using support vector machines," in InCoB'04: Proceedings of Third International Conference on Bioinformatics, Auckland, New Zealand, 2004.