ISLAMIC UNIVERSITY OF TECHNOLOGY

Improved Technique for Cancerous Gene Selection Based on PSO and BW ratio

By:

Nazmus Saqib (094439)

Farhat Aman (094445)

Supervised by:

Prof. Dr. M. A. Mottalib

Head of the Department

Department of Computer Science and Engineering

Co-supervised by:

Shaikh Jeeshan Kabeer

Lecturer

Department of Computer Science and Engineering

A thesis submitted in partial fulfilment of the requirements
for the degree of Bachelor of Science in Computer Science and Engineering

Academic Year: 2012-2013

Department of Computer Science and Engineering

Islamic University of Technology.

A Subsidiary Organ of the Organization of Islamic Cooperation.

Dhaka, Bangladesh.

October 25, 2013

Declaration of Authorship

We, Nazmus Saqib & Farhat Aman, declare that this project titled,'Improved Technique for Cancerous Gene Selection Based on PSO and BW ratio' and the work presented in it are our own. We confirm that:

- This work was done wholly while in candidature for a Bachelor degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.

Submitted By:

Nazmus Saqib (094439)

Farhat Aman (094445)

Improved Technique for Cancerous Gene Selection Based on PSO and BW ratio

Approved By:

Prof. Dr. M.A. Mottalib
Supervisor,
Head of the Department,
Department of Computer Science and Engineering,
Islamic University of Technology.

Shaikh Jeeshan Kabeer
Co-supervisor,
Lecturer,
Department of Computer Science Of Technology,
Islamic University of Technology.

ISLAMIC UNIVERSITY OF TECHNOLOGY

Abstract

CSE

Department of Computer Science and Engineering

Bachelor of Science in Computer Science and Engineering

Improved Technique for Cancerous Gene Selection Based on PSO and BW ratio

by

Nazmus Saqib

Farhat Aman

The relentless development of Microarray technology have meant that the dimensionality of data that is produced by the Microarray chips have increased many folds over the years. Recognition of patterns and other subsequent analysis from the thousands of gene expression values is particularly difficult and primary role of an effective feature selection is to simplify this task. Removal of less informative genes helps to alleviate the effects of noise and redundancy, and simplifies the task of disease classification and prediction of medical conditions such as cancer. In this study the shortcoming of the current GA based approach for feature selection has been improved. A filter and wrapper models are put to use to take advantage of the facilities that each provides. As filter method exhibits some limitations, in this study an approach to filtering (BW ratio) has been employed. As a wrapper approach Particle Swarm Optimization (PSO) has been proposed.

Acknowledgements

Special thanks to:

Prof. Dr. M.A. Mottalib

Head of the Department

Department of Computer Science and Engineering

Special thanks to:

Shaikh Jeeshan Kabeer

Co-supervisor

Lecturer

Department of Computer Science and Engineering

Contents

List of Figures

Dedicated to our parents...

Chapter 1

Introduction

1.1 Overview

Modern technology has achieved a lot of innovations with the use of computer science. Bio-informatics is a part where computer scientists have given a lot of effort to. The study of genes is being done by the computer scientists to help in biological research. The use of DNA microarray technology has helped a lot in the progress of biological research. By using this, the scientists can measure the expression levels of thousands of genes within one experiment. Although this technology has taken us in a new era, interpreting the microarray data still remain a challenging issue for the reason of their high dimensional low sample sized data. Therefore robust and accurate feature selection methods are required to identify differentially expressed genes among varied samples.[1] Feature selection is the technique of selecting a subset of relevant features for use in model construction. Our thesis works on feature selection. Our initial work on this thesis was based on the study of the previously implemented evolutionary algorithms. We have been also analyzing the existing methods for our purpose. We had also the comparison of the advantages and disadvantages of various existing methods. Besides Bioinformatics, these feature selection techniques are also used in the pattern recognition, text categorization, robotics and many more. In Bio-informatics, feature selection techniques are used to reduce the dimensionality of the micro array data. There are many genes which are irrelevant and some genes which are redundant among other genes. By using feature selection techniques, we can reduce the microarray data set for our experimental use with the data set by removing the uninformative genes. Thus we can pick the genes which are most informative among samples.

1.2 Problem Statement

The high dimensional data has the main problem that it includes the noisy and redundant genes. As the dimensionality of the data increases, the resulting number of noisy and redundant genes also increases. If we apply various feature selection techniques, the reduction of the number of redundant and uninformative genes can be accomplished.

1.3 Research Challenges

The high dimensionality of data creates a big feature space so that we have to propose a method which can evaluate the features and give an optimal result. The desired output our method is where the number of features is reduced, as well as the predictive power of the classifiers should also be increased. Along with the removal of the redundant and uninformative genes, our proposed method should also be able to handle the correlation factor existing among the features and thus ensure the combine predictive power. Our study includes all these factors and tries to produce a better result by considering these factors.

1.4 Motivation

The main motivation for this study was to derive a better feature selection technique for bio-informatics. The reason of choosing PSO is that PSO is better than many other evolutionary approaches in many aspects. It takes less computation time than the other evolutionary approaches. PSO has its own memory. Here, all the particles tend to converge to the solution quickly. PSO doesn't have any overlapping and mutation calculation.[10] PSO seems to be somewhat less dependent of a set of initial points compared to other evolutionary methods, implying that convergence algorithm is robust.[11] There are also some drawbacks for PSO. In our proposed algorithm our main focus is to improve the "weak local search ability" problem of PSO. We wanted to reduce the drawbacks by using the combination of another approach with PSO. We wanted to get an optimal solution from PSO. This was our main motivation for proposing this evolutionary approach.

1.5 Scopes

Our study for the feature selection techniques has a big scope over the other techniques as we are using the evolutionary approach. The evolutionary approaches are being used in recent times for feature selection so that there are lots of scopes for betterment here. There can be possibly two types of research. One is to take the existing methods and do research on them. The other is to take approaches and generate new ones by combining them. In our case we have taken two approaches for feature selection and combined them to reduce the feature space. This approach will reduce the shortcomings of some existing features. We will discuss about our approach in later parts.

1.6 Research Contribution

In this thesis, our main focus is to reduce the shortcomings of PSO. PSO is a previously used approach for gene featuring. But there are various shortcomings of the PSO process. Our main focus is to improve the "weak local search ability" problem of PSO. To overcome the shortcomings of PSO, we have used the BW ratio approach as the preliminary approach for the feature selection. By implementing BW ratio on the existing data set, the size of the dataset is reduced and then the PSO is implemented on the reduced data set. Traditional PSO has some time and space complexity. Our target is to get an optimal solution by using BW ratio as our preliminary approach to PSO.

1.7 Thesis Outline

In chapter 1, we have given the introduction of our thesis by describing shortly what we are going to do. Chapter 2 deals with the basic feature selection methods and the brief description of the PSO method we are going to use. Chapter 3 deals with our proposed algorithm and an elaborate discussion about the BW ratio we are going to use as the preliminary approach.

Chapter 2

Background Review

2.1 Feature Selection

Identifcation of a set of genes that best discriminate biological samples of different types is a feature selection problem. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing the prediction accuracy of the classifier built by using only the selected features. As computer power grows and data collection technologies advance, a plethora of data is generated in almost every field where computers are used. Simply put as the processing speed and memory capabilities of computers continues to grow at an accelerated rate the amount of data generated is quite enormous and without the aid of computers itself it is almost certain that these huge amounts of data will never be examined, thus the use of feature selection to reduce the dimensionality for better understanding and knowledge extraction is imperative.

Feature selection techniques are widely used in statistics and machine learning. Our study focuses on the latter. Machine learning is the study of algorithms that automatically improve their performance with experience. At the heart of performance is prediction. Machine learning algorithms are organized into a taxonomy based on the desired outcome of the algorithm. There are several types of methods like- Supervised learning, Unsupervised learning, Reinforcement learning etc.

2.2 Categories of Feature Selection

2.2.1 Feature Ranking (FR)

This is also known as feature weighing which assesses individual features and assigns them weights according to their degree of relevance. Many researchers have been done with feature ranking as the base method

2.2.2 Feature Subset Selection (FSS)

This technique measures the goodness of each found feature subset. A great deal of work has also been done Feature subset selection Although, feature selection techniques have many benefits, but it also introduces extra complexity level, which requires thoughtful experiment design to address the challenging tasks yet provide fruitful results. Feature selection methods can be structured into three factions, which are filter methods, wrapper methods and embedded methods.

Filter method: Filter methods rank each feature according to some univariate metric, and only the highest ranking features are used while the remaining low ranking features are eliminated. This method also relies on general characteristics of the training data to select some features without involving any learning algorithm. Therefore, the results of filter model will not affecting any classification algorithm. Moreover, filter methods also provide very easy way to calculate and can simply scale to large- scale microarray datasets since it only have a short running time.[1]

Wrapper method: While the filter techniques handle the identification of genes independently, a wrapper method on the other hand, embeds a gene selection method within a classification algorithm. In the wrapper methods a search is conducted in the space of genes, evaluating the goodness of each found gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. The wrapper approach, which is very popular in machine learning applications, is not comprehensively used in DNA microarray tasks, and few works in the field make use of it. It is claimed that the wrapper approach obtains better predictive accuracy estimates than the filter approach; however, its computational cost must be taken into account. Wrapper methods can be divided into distinct groups, deterministic

and randomized search algorithm. Genetic Algorithm (GA) is a randomized search algorithm [1]

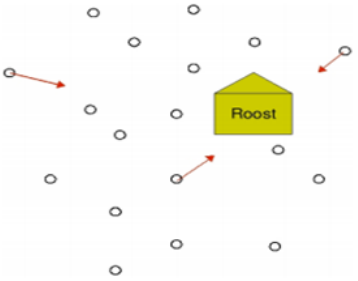
Embedded method: The third class of feature selection approaches is embedded methods. The different of embedded methods with others feature selection methods is the search mechanism is built into the classifier model. Identical to wrapper methods, embedded methods are therefore specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods [1]

2.3 Particle Swarm Optimization (PSO)

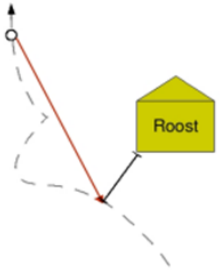
Particle swarm optimization (PSO) is a population-based stochastic optimization technique, and was developed by Kennedy and Eberhart in 1995. PSO simulates the social behavior of organisms, such as bird flocking and fish schooling, to describe an automatically evolving system. In PSO, each single candidate solution is "an individual bird of the flock", that is, a particle in the search space. Each particle makes use of its individual memory and knowledge gained by the swarm as a whole to find the best solution. All of the particles have fitness values, which are evaluated by a fitness function to be optimized, and have velocities which direct the movement of the particles. During movement, each particle adjusts its position according to its own experience, as well as according to the experience of a neighboring particle, and makes use of the best position encountered by itself and its neighbor. The particles move through the problem space by following a current of optimum particles. [2] The initial swarm is generally created in such a way that the population of the particles is distributed randomly over the search space. At every iteration, each particle is updated by following two "best" values, called pbest and gbest. Each particle keeps track of its coordinates in the problem space, which are associated with the best solution (fitness) the particle has achieved so far. This fitness value is stored, and called pbest. The best previously visited position of the i-th particle is denoted its individual best position pi = (pi1, pi2, , piD), a value called pbest. When a particle takes the whole population as its topological neighbour, the best value is a global "best" value and is called gbest. The best value of all the individual pbesti values is denoted as the global best position g = (g1, g2,, gD) and called gbest.

PSO was first found by Kennedy and Eberhart who included a 'roost' in a simplifed Reynolds-like simulation so that:

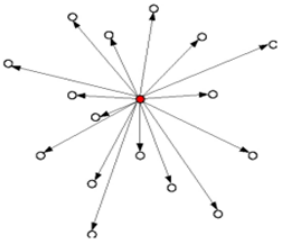
Each agent was attracted towards the location of the roost.



Each agent 'remembered' where it was closer to the roost.



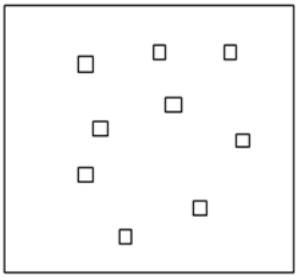
Each agent shared information with its neighbors (originally, all other agents) about its closest location to the roost.



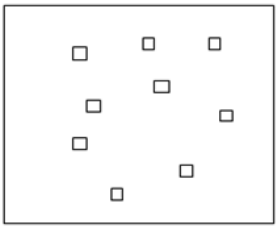
2.3.1 The PSO Procedure

We can visualize the procedure here how the particles move and how their position and velocities are updated.

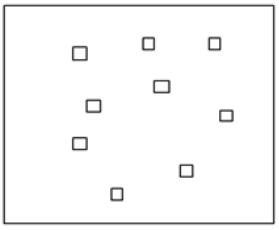
1.Create a population of particles uniformly distributed over the feature space



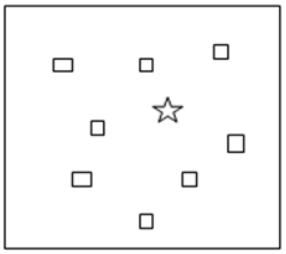
2.Evaluate each particle position by calculating fitness of the particle



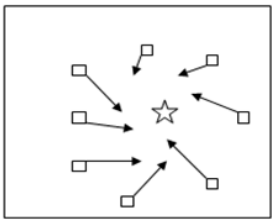
3.If a particles current position is better than its previous best position, update it



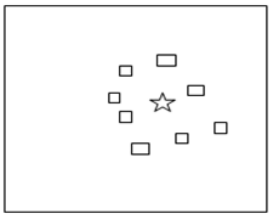
4.Determine the best particle according to the particle's previous best position



5. Update velocity according to the equation for new velocity



6.Move particle to their new position according to $x_{pd} = x_{pd} + v_{pd}$



7.Go to step 2 until stopping criteria are satisfied

2.3.2 Limitations of PSO

PSO sometimes face the problem of Local Optima, where it is trapped in the local optima. In this case the velocity of the particles become almost zero near the global optima which decrease the convergence rate. [4] Random selection often causes the appearance of irrelevant genes that causes increase in time and space complexity. Search process is non linear and complicated. After some generation the convergence rate in PSO decreases. So faster convergence in the refined state is a challenge in PSO. This is the problem of slow convergence in refined search space. [3] PSO is weak in local search ability. It mainly runs its search process globally.

2.3.3 Recent Works in PSO

PSO (feature selection using PSO-SVM) The paper implements the evolutionary approach by combining the particle swarm optimization approach with the SVM classifier. Here the classifier is used to evaluate the fitness function of PSO for classification problem. The proposed method is applied to five well known classification problems. Results show that their implementation has worked successfully with a higher classification accuracy. Here SVM is used by the one versus rest methods are used as evaluator for the fitness function of the PSO. ONE VERSUS REST METHODS: this method allows classifiers that distinguish one class from all the other classes. Firstly the output of the k classifiers is obtained, and then if there is a unique class label of a data point, which is consistent over all the k classifiers the data point is assigned into that unique class. Otherwise, one of the k classes is selected randomly. These methods are used in this paper and the results are being evaluated. Binary PSO is being used for the feature selection problem. [2]

PSO (using PSO with rough set theory) An evolutionary rough feature selection algorithm is proposed for classifying gene expression patterns. An initial redundancy reduction of the attributes is done first for the faster convergence. Rough set theory is used here to generate the distinction table which enables PSO to reduce the dataset by removing the redundant and irrelevant genes from the dataset. The results are experimented using three well known datasets. They are colon, lymphoma and leukemia using MOGA. Firstly the preliminary reduction technique, rough set theory is being applied for getting a reduct. Rough set theory provides an important and mathematically established tool, for

this sort of dimensionality reduction in large data. A basic issue addressed, in relation to many practical applications of knowledge databases. The task of finding reducts is reported to be NP-hard. The high complexity of this problem has motivated investigators to apply various approximation techniques to find near-optimal solutions. There are some studies reported in literature, where genetic algorithms (GAs) have been applied to find reducts. multi-objective GAs (MOGAs) provide an alternative, more efficient, approach to searching for optimal solutions. Each of the studies in employs a single objective function to obtain reducts. The essential properties of a reduct are (i) to classify among all elements of the universe with the same accuracy as the starting attribute set, and at the same time (ii) to be of small cardinality. A close observation reveals that these two characteristics are of a conflicting nature. Hence the determination of reducts is better represented as a two-objective optimization problem. Thus this procedure is applied and the results are compared among three cancer datasets in this paper. [5]

Adaptive PSO (APSO): An adaptive particle swarm optimization (APSO) that features better search efficiency than classical particle swarm optimization (PSO) is performed. More importantly, it can perform a global search over the entire search space with faster convergence speed. The APSO consists of two main steps. First by evaluating the population distribution. Then an elitist learning strategy is performed when the evolutionary state is classified as convergence state. The strategy will act on the globally best particle to jump out of the likely local optima. The standard PSO algorithm can easily get trapped in the local optima. So accelerating convergence speed and avoiding the local optima have become the two most important and appealing goals in PSO research. To avoid possible local optima in the convergence state, combinations with auxiliary techniques have been developed elsewhere by introducing operators such as selection, crossover, mutation, local search, reset etc. into PSO. How to detect the different population distribution information and how to use this information to estimate the evolutionary state would be a significant and promising research topic in PSO. ESE perform this job. Here the PSO parameters are not only controlled by ESE but also taking the different effects of these parameters in different states into account. The ELS is proposed in this paper to perform only on the globally best particle and only in a convergence state. This is not only because the convergence state needs the ELS most but also because of a very low computational overhead. [3]

2.3.4 Some Improvements of PSO

There are some works to improve the performance of PSO. Here some of the techniques are briefly described:-

- **Inertia weight:** The bigger w is, the bigger the PSO's searching ability for the whole is, and the smaller w is, the bigger the PSO's searching ability for the partial. [4]
- **Acceleration co-efficient:**Varying the acceleration co-efficient may change the results.
- **No. of iteration:** Number of iterations has effect on PSO result.
- **Velocity clamping:** Velocity clamping will control the global exploration of the particle. If the velocity of a particle exceeds the maximum allowed speed limit, it will set a maximum value of velocity .[9]
- **CPSO:**This process is known as Complement PSO. In CPSO in every generation 50
- **Selection:**Here ESE enabled adaptive parameters are expected to bring faster convergence speed to the PSO algorithm. Nevertheless, when the algorithm is in a convergence state, for the gBest particle, it has no other exemplars to follow. So the standard learning mechanism does not help gBest escape from the current optimum if it is local. Hence, an elitist learning strategy (ELS) is developed to give momentum to the gBest particle. The ELS randomly chooses one dimension of gBest's historical best position and assigns it with momentum to move around [8]
- **Blending:**Blending with other algorithms make the performance better. [4]
- **Dimension of problem:** The change in dimensions of particles changes the result. [9]
- **Increase convergence factor:**A particle swarm optimization algorithm with convergence agents is introduced and a new formula is given to perform better. [4]

2.3.5 Flowchart for PSO

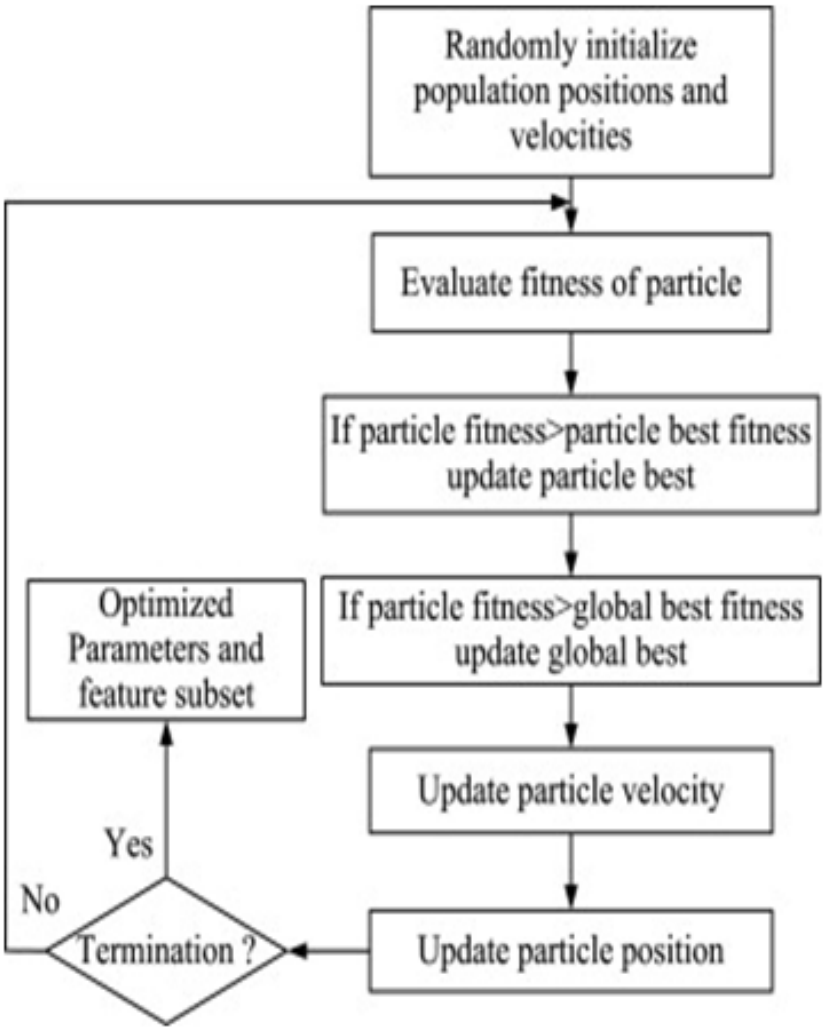


FIGURE 2.1: Flowchart for PSO

Chapter 3

Our Proposed Approach

3.1 Overall Concept

Our proposed approach contain PSO with the help of BW ratio. Firstly we are taking the dataset and applying BW ratio approach in our dataset. Then we are getting a reduced dataset in which we are applying the standard PSO approach. Then we are getting the final subset as our solution.

3.1.1 PSO

In PSO, a swarm of particles are represented as potential solutions, and each particle i is associated with two vectors, i.e., the velocity vector Vi =[v1i ,v2i ,...,vDi] and the position vector Xi =[x1i ,x2i ,...,xDi], where D stands for the dimensions of the solution space. The velocity and the position of each particle are initialized by random vectors within the corresponding ranges. During the evolutionary process, the velocity and position of particle i on dimension D are updated as

$$v_{pd}^{new} = \mathrm{w} \times v_{pd}^{old} + c_1 \times rand_1 \times (pbest_{pd} - x_{pd}^{old}) + c_2 \times rand_2 \times (gbest_d - x_{pd}^{old}) \quad (1)$$

$$\mathrm{S}\,(v_{pd}^{new}) = \frac{1}{1+e^{-v_{pd}^{new}}} \quad (2)$$

$$\text{If } (rand < \text{S}\,(v_{pd}^{new})) \text{ then } x_{pd}^{old} = 1; \text{ else } x_{pd}^{old} = 0 \qquad (3)$$

Where w is the inertia weight, c1 and c2 are the acceleration coefficients, and c1 and c2 are two uniformly distributed random numbers independently generated within [0, 1] for the Dth dimension . A pseudo-code description for PSOs is following: [6]

For each particle Initialize particle END

Do For each particle Calculate fitness value If the fitness value is better than the best fitness value (pbest) in history set current value as the new pbest End

Choose the particle with the best fitness value of all the particles as the gbest

For each particle Calculate particle velocity according equation (1) Update particle position according equation (3) End While maximum iterations or minimum error criteria is not attained.

3.1.2 BW Ratio

The BW ratio is the ratio of the between-treatment sum of squares and the within-treatment sum of squares of the expression values of the genes which we get from the dataset. We thus performed a preliminary selection of genes based on the ratio of the between group to within group sum of squares.[12] We have used this approach as the preliminary approach for our gene selection. The output genes which we get from this approach will be the input for the pso method. For a given gene j, the BW ratio for that gene will be :-

$$\text{BW}(j) = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}$$

Terminology

Here x_j and x_{kj} denote the average of the expression values of the genes of all the samples and the average across the samples belonging to class k. x_{ij} here is the expression value of a gene under the sample i.

Steps

1. Firstly take the genes from the ALL dataset and take the values needed for the equation of the BW ratio.
2. For every genes j, we have to calculate the average values, the average under class 1, the average under class 2, if there are 2 types of samples.
3. We have to calculate the BSS and WSS for every genes by following the equation for every samples.
4. Then we have to divide BSS and WSS for every genes and the result for every genes are stored, this is the BW ratio.
5. From the dataset, the number of genes from the dataset is being reduced by taking the genes with the higher BW ratio.
6. Then these genes are sent to the PSO algorithm as the input of PSO and the work of BW ratio is done.

The Basic Flowchart for BW Ratio

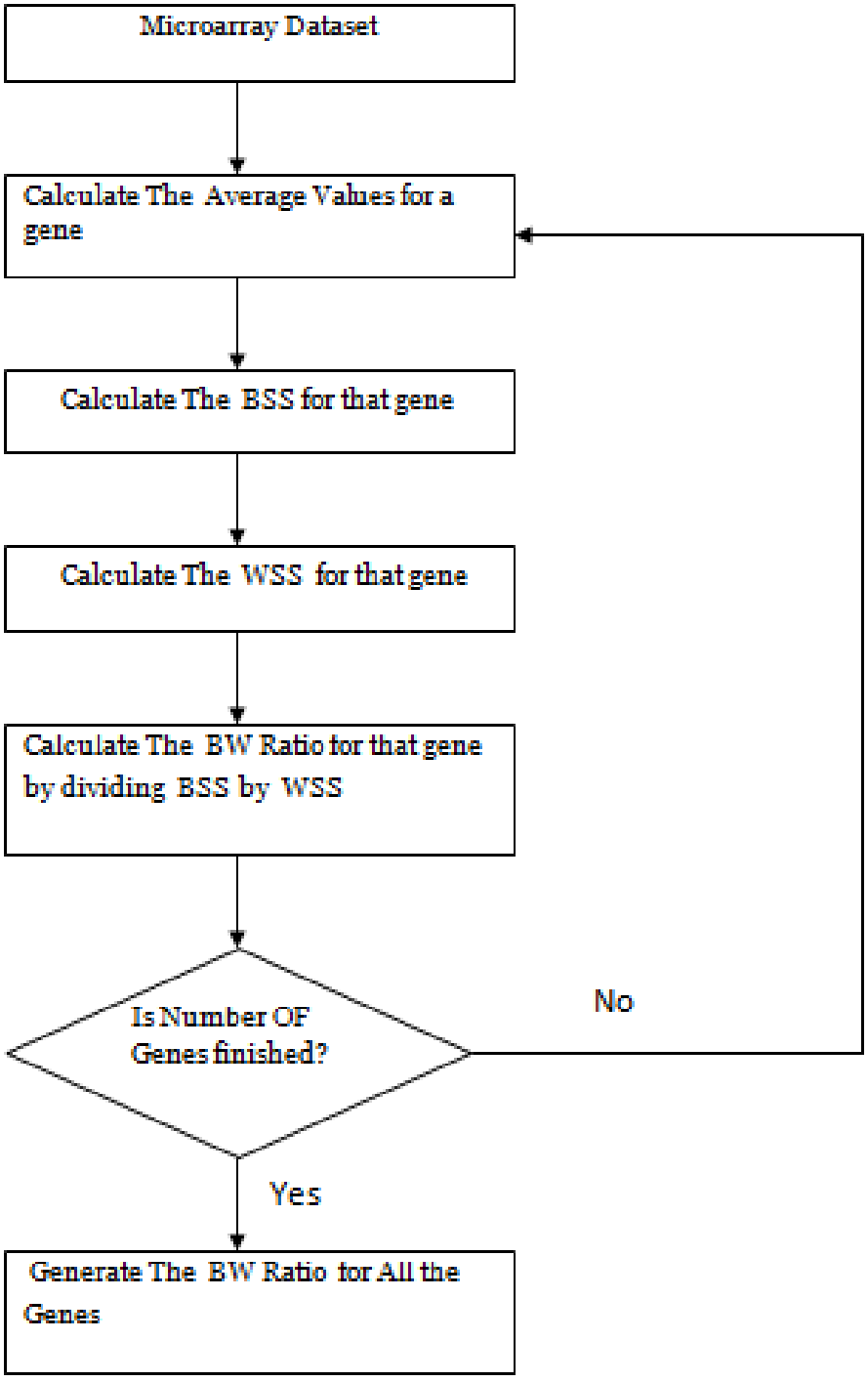


FIGURE 3.1: The Basic Flowchart for BW Ratio

3.2 Overall Procedure in Flow-chart

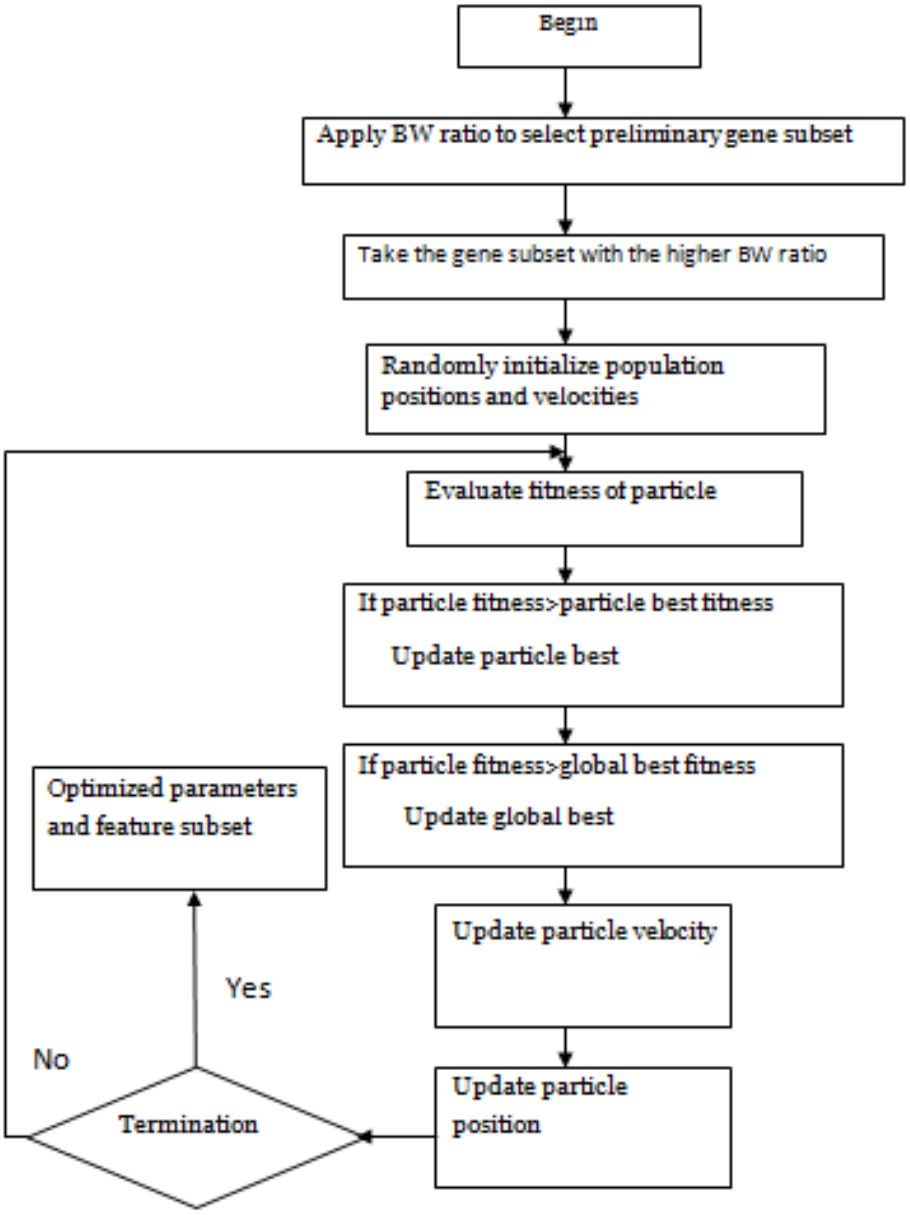


FIGURE 3.2: Overall Procedure in Flow-chart

Chapter 4

Experimental Analysis and Result Comparison

We have used matlab 2010a as the main platform of our work and used several function references under bio-informatics toolbox.

4.1 Dataset Details

These microarray datasets on which we have applied our proposed approach are: Acute Lymphoblastic leukemia cancer (ALL) and colon cancer. Table 1 summarizes the data sets. In the ALL dataset there are 72 tissue samples (47 B-cell and 25 T-cell). Colon dataset contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. 2000 genes out of around 6500 genes were selected based on the confidence in the measured expression levels

Table 4.1: Datasets

Dataset	Number of classes	Number of samples in datasets	Number of genes
ALL	2	72	7129
COLON	2	62	2000

From Table 4.1 we can see that dataset for Colon cancer contains the lowest number of genes of the two datasets, exposing higher possibilities of misclassifications and over fitting. It is because the more the number of samples the more we can train classifiers to classify test samples.

4.2 Experimental Settings

In our procedure,we have used some values for the parameters.We have taken these values:

1. **Inertia Weight**: for Colon-0.5, for Leukemia-0.3,0.5,0.8,1.1
2. **Accelaration Factor**:for Colon-0.2, for Leukemia-0.2,2
3. **BW Ratio**: for colon-0.02, for leukemia-0.15

c1 and c2 are acceleration factors that controls how far a particle will move in a single generation.Usually c1= c2 =2. With this value of the acceleration factor the velocity of the particle int each dimension increases rapidly and tends to select higher number of genes for obtaining high classification accuracy. While for excessive adjustment particle movement will be excessive, causing the algorithm to weaken early, so that a useful feature set cannot be obtained. Hence, suitable parameter adjustment enables particle swarm optimization to increase the efficiency of feature selection. for BWratio , we have selected the genes by pre-processing with the dataset before running the PSO. We have selected the genes exceeding the minimum number we have set for BWratio. From table of above, we can see the values for which values of BWratio, the results taken are good. **Local Optima** - Convergence rate decreases at later stages of evolution; when closer to a near optimal solution the algorithm stops optimizing and accuracy becomes limited. Velocity of particle rapidly approaches 0.

4.3 Performance Analysis

Our implementation begins with BWratio where we select a particular number of genes using BWratio method. The main objective to perform BWratio is to provide PSO with a better initial population rather than generating it randomly thus avoiding early convergence and over-fitting problems. In case of implementing BWratio we have used BWratio of each gene to calculate the scoring for each gene . BWratio process is to calculate the ratio of between group and within group distance for each gene. Genes with the greater BWratio are taken for PSO as the best scored gene.

We applied BWratio on the two important available microarray datasets and got the reduced number of genes to apply as the input for PSO. These are shown in the table below-

Table 4.2: BWratio output

Dataset	Original number of genes	BWratio output
ALL	7129	2036
COLON	2000	1105

We have taken roughly 30 percent of the existing dataset by applying BWratio for leukemia cancer dataset. And we have taken 50 percent of the existing dataset for colon cancer. For these PSO will take 2036 and 1105 genes as initial population.

In the below two tables,the accuracies and the number of genes selected were shown for the two datasets.For ALL dataset,we have found 100 percent accuracy with 9 genes selected.

Table 4.3: Output of ALL Dataset

Number of runs	Accuracy	Number of selected genes
1	100%	6
2	97%	3
3	100%	13
4	97%	12
5	97%	9
6	97%	7
7	97%	8
8	97%	4
9	100%	8
10	100%	9
Average	98.2%	8

TABLE 4.4: Output of COLON Dataset

Number of runs	Accuracy	Number of selected genes
1	90%	4
2	100%	3
3	97%	3
4	100%	3
5	93%	5
6	100%	10
7	100%	2
8	100%	6
9	94%	6
10	94%	7
Average	96.8%	5

4.4 Comparative Analysis

In this below table, the accuracies of different approaches are compared. We can see that the PSO-BW approach has the greater accuracy in average.

TABLE 4.5: Comparison of Accuracies

Dataset	BWratio	PSO	GA-SVM	CGA-SVM	BW-PSO
ALL	83.89%	84.22%	88.24%	96.12%	98.2%
COLON	76.87%	76.89%	86.27%	89.74%	96.8%

Now we are gonna compare the number of genes selected in different approaches in this below table:-

TABLE 4.6: Selected Genes in Different Methods

Dataset	Original genes	GA-SVM	CGA-SVM	BPSO-SVM	BW-PSO
ALL	7129	23	17.90	71.5	8
COLON	2000	23	25.5	85.8	5

From these above two tables, we can surely say that our BW-PSO approach is giving better results in every sense. The accuracies in our approach is better than the other previous approaches. Also for the number of genes selected, our approach is giving better result which we can see from the table.

Graphical Representation of Comparison for Various Datasets

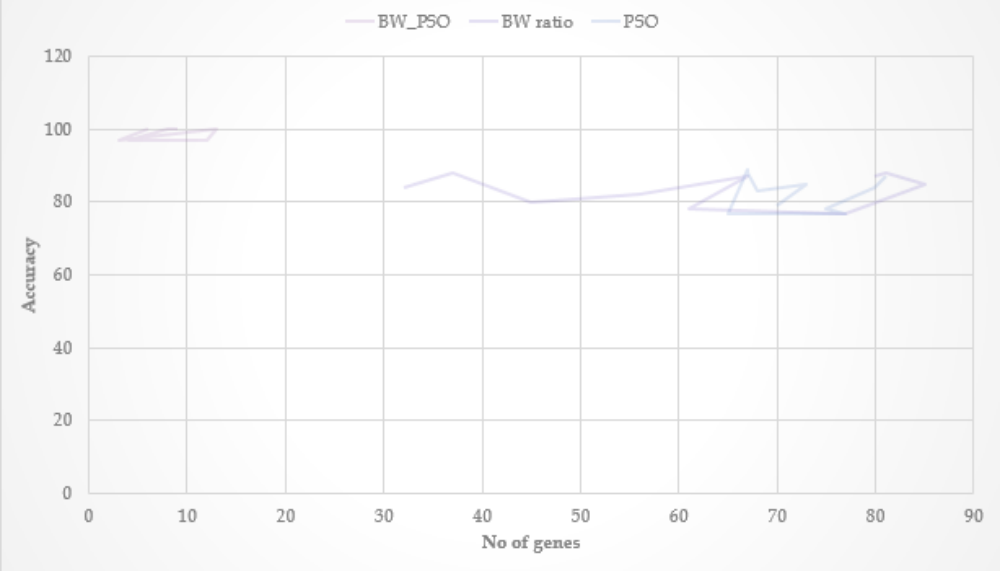


FIGURE 4.1: graphical representation of comparison for ALL dataset

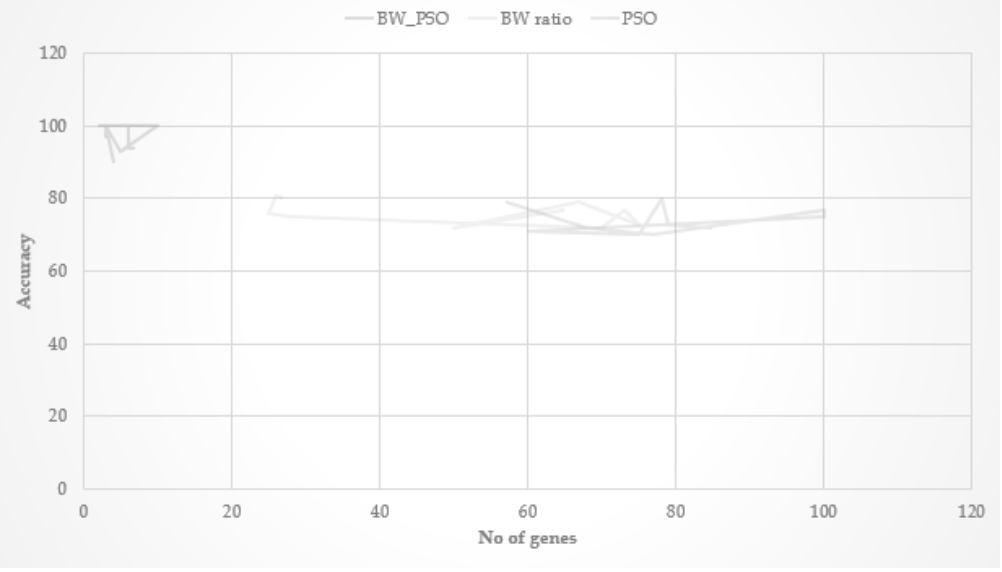


FIGURE 4.2: graphical representation of comparison for Colon dataset

From the above graphs we can see that our approach for either SVM or KNN is giving better results from two different perspectives. In our approach, we got the best accuracies for selecting less number of genes thus giving a better result.

In these above two tables, we have shown the index numbers of the genes which are selected for the two different datasets.

Top Most Informative Genes Selected for Leukemia Dataset

TABLE 4.7: Index of gene selected (ALL)

Number	Index
1	823
2	902
3	1017
4	1098
5	1205
6	1321
7	1509
8	1524
9	1822

Top Most Informative Genes for Colon Cancer Dtasets

TABLE 4.8: Index of gene selected (COLON)

Number	Index
1	74
2	229
3	263
4	348
5	403
6	525
7	586
8	723
9	986

Here we have shown that the above selected genes for the two datasets,are the most informative ones.

Chapter 5

Conclusion

In this study, we have seen that our approach works well on microarray datasets other than existing approaches. Traditional PSO suffered from the problem of initialization as firstly genes are selected randomly. BWratio solves this problem as we preprocess the dataset using BWratio. BWratio scores individual gene but does not consider the collective predictive power of genes. Thus we feed the output of BWratio to PSO to maintain correlation. For future development this framework can also be used for other high dimensional data used in other fields such as archeology, geography, climate study, data mining, image processing and many others. The data analysis is expected to produce good results for these other datasets. Even in our field of specialization we have not used our framework on all the microarray datasets such as brain cancer, bone cancer, stomach cancer etc. Also there are many classifiers available; in our study we have used SVM classifier. Other classifiers such as C4.5,KNN, Nave Bayes Classifier can also be integrated with our framework for an enhanced comparative analysis.

Chapter 6

Reference

[1] Farzana Kabir Ahmad, Prof. Dr. Safaai Deri, Assoc. Prof. Dr. Norita Md. Norwawi, Prof. Dr. Nor Hayati Othman "A Review of Feature Selection Techniques via Gene Expression Profiles",2008.
[2] Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, and Cheng-Hong Yang "Feature Selection using PSO-SVM",2009.
[3] Zhi-Hui Zhan, Jun Zhang, Yun Li, Henry Shu-Hung Chung "Adaptive Particle Swarm Optimization",2009.
[4] Qinghai Bai "Analysis of Particle Swarm Optimization Algorithm",2010.
[5] Haider Banka, Suresh Dara "Feature Selection and Classification for GeneExpression Data using Evolutionary Computation",2012.
[6] Baiyi Xie, Shihong Chen, Feng Liu "Biclustering of Gene Expression Data Using PSO-GA Hybrid",2007.
[7] Li-Yeh Chuang, Hua-Fang Jhang, Cheng-Hong Yang, "Feature Selection using Complementary Particle Swarm Optimization for DNA Microarray Data"
[8] Dian Palupi Rini, Siti Mariyam Shamsuddin, Siti Sophiyati Yuhaniz "Particle Swarm Optimization: Technique, System and Challenges",2013.
[9] V.Selvi, Dr.R.UMARANI "Comparative Analysis of Ant Colony and Particle Swarm Optimization Techniques",2010.
[10] Kwang Y. Lee, Jong-Bae Park "Application of Particle Swarm Optimization to Economic Dispatch Problem: Advantages and Disadvantages",2006.
[11] Sandrine Dudoit, Jane Fridlyand, and Terence P. Speed "Comparison of Discrimination Methods for the Classication of Tumors Using Gene Expression Data",2012.