Examining High Performance Genetic Data Feature Selection and Classification Algorithms

G. Janani , G E. Gowsika , A. S. Bhaala Kumar, N. Sasipriyaa,
Department of Computer Science,
Kongu Engineering College

***Abstract*:- Data mining is the process of analyzing data from different perceptions and brief it into useful information . It is otherwise called as data or knowledge discovery.The biomedical research of today continuously provides a set of tough challenges for data analysis of large data sets.Selection of genes for classification is a common task in most gene expression , to identify the smallest possible set of genes that can achieve good predictive performance. Gene selection is the process of selecting significant genes via expressive pattern.**

In the existing system Particle Swarm Optimization (PSO) algorithm combined with a Support Vector Machine as the classifier is used. For selecting genes PSO algorithm is used. Support vector machine classifier is used for finding classification accuracy.

The goal is to achieve efficient gene selection that helps in identifying cancers. In the proposed work Firefly Algorithm (FA) is used for gene selection and Support vector machine classifier is used for finding classification accuracy. By comparing the results of the algorithms, it is found that Gene Selection using Binary Firefly Algorithm (GSBFA) provided a improved accuracy and minimal number of genes when compared with Gene Selection using Particle Swarm Optimization (PSO).

Keywords: Particle Swarm Optimization, Firefly algorithm, Support Vector Machine.

INTRODUCTION

Data mining is otherwise called as data or knowledge discovery. The data sources can include databases, data warehouses, the Web, data that are streamed into the system dynamically[1]. Data mining is widely used in areas such as Market based analysis, Education, Fraud detection, Intrusion detection, Lie detection, Financial Banking, Bioinformatics etc,. Gene selection is the process of selecting significant genes via expressive pattern.Methods are needed for choosing the important subset of genes with high classification accuracy.Three classes of gene selection are, the wrapper, the filter, and the embedded approaches.

EXISTING SYSTEM

Particle Swarm Optimization (PSO) algorithm combined with a support vector machine as the classifier used. PSO is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality[2]. It solves a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity[3]. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles[3]. This is expected to move the swarm toward the best solutions[3].

The PSO algorithm uses a population of individuals to find the best solutions. A particle represents a candidate solution to the problem being addressed[4]. Assume that a search space is D-dimensional, and the ith particle of a swarm can be a D-dimensional position vector $X_i = [x_{i1}, x_{i2}, . . ., x_{iD}]$. The particle velocity of particle i is denoted as $V_i = [v_{i1}, v_{i2}, . . .,v_{iD}]$. We also consider that the best visited position that yields the best fitness value for the particle is $PB_i = [pb_{i1}, pb_{i2}, . . .,pb_{iD}]$ and the best position explored thus far is $GB =[gb_1,gb_2,...,gb_D]$. Each particle is updated according to the velocity[4].

PSEUDOCODE FOR PSO

for each particle i=1,…,number of particles do Initialize the particles position with a uniformly distributed random vector:$\mathbf{x_i}$
Initialize the particles best known position to its initial position: $\mathbf{p_i} \leftarrow \mathbf{x_i}$
if$f(\mathbf{p_i}) < f(\mathbf{g})$ then
update the swarms best known position: $\mathbf{g} \leftarrow \mathbf{p_i}$
Initialize the particles velocity:$\mathbf{v_i}$
while termination criterion is not met do: for each particle i=1,…,number of particles for each dimension d=1,…,n do
Pick random numbers: r_p, r_g
Update the particles velocity:$\mathbf{v_{i,d}}$ Update the particles position: $\mathbf{x_i} \leftarrow \mathbf{x_i} + \mathbf{v_i}$
if$f(\mathbf{x_i}) < f(\mathbf{p_i})$ then
Update the particles best known position: $\mathbf{p_i} \leftarrow \mathbf{x_i}$
if$f(\mathbf{p_i}) < f(\mathbf{g})$ then
Update the swarms best known position: $\mathbf{g} \leftarrow \mathbf{p_i}$

A support vector machine constructs a hyperplane or set of hyperplanes in a high-or infinite-dimensional space,which can be used for classification,regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier[5].Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to

discriminate are not linearly separable in that space[5]. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in the terms of kernel function.The hyperplanes in the higher- dimensional space are defined as the set of points whose dot product with a vector in that space is constant[5].

PROPOSED WORK

The proposed work is Gene selection using binary firefly algorithm combined with Support vector machine as classifier.

FIREFLY ALGORITHM

The firefly algorithm (FA) is a metaheuristic algorithm proposed by Xin-She Yang (2008) inspired by the flashing behaviour of fireflies[6]. The primary purpose for a firefly's flash is to act as a signal system to attract other fireflies .The interaction between fireflies is governed by the following rules:

- All the fireflies are unisex and are attracted by other fireflies independently from their sex.
- The attractiveness of a firefly is proportional to its brightness[7]. The brightness degree of a firefly perceived by another firefly is inversely proportional to the distance between them[8]. In this context, the less bright firefly moves towards the brighter one[10]. If no firefly is brighter than a given firefly, than the later moves randomly.
- The brightness of a firefly is determined by the value of the objective function to optimize[10].

BEHAVIOR OF FIREFLIES

There are near to two thousand firefly species, and most of them produce short and rhythmic flashes[11]. The pattern observed for these flashes is unique for most of the times for a specific species. The rhythm of the flashes, rate of flashing and the amount of time for which the flashes are observed are together forming a kind of a pattern that attracts both the males and females to each other. Females of a species respond to individual pattern of the male of the same species.
The intensity of light at a certain distance r from the light source conforms to the inverse square law. Additionally, the air keeps absorbing the light which becomes weaker with the increase in the distance. These two factors when combined make most fireflies visible at a limited distance, normally to a few hundred meters at night, which is quite enough for fireflies to communicate with each other.

MAIN FACTORS OF FIREFLY ALGORITHM

The main factors of firefly algorithms (Yang 2008) are:
Brightness: It depends on the objective function. In simple optimization problems, the brightness of a firefly x is reduced to the objective function for x : I(x)=f(x)

Attractiveness: The attractiveness of a firefly is proportional to the brightness perceived by the other neighbors. The attractiveness function may be any function which is monotonically decreasing with respect to the real distance r. A general form of such a function may be:
$\beta = \beta 0e^{-\gamma r2}$
where$r$ is the distance between two fireflies, $\beta 0$ is the attractiveness for $r = 0$ and γ is a constant (bright absorption coefficient).

Distance: Simply the Euclidean distance is considered to measure the distance between two fireflies $x_i$and x_j

$$r_{ij}= \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})2}$$

wherex_{ik}stands for the $k$thcomponent of the ith firefly

Movement : The movement of a firefly i attracted by another one j which is brighter is determined by:
$x_{i+1}=x_i+\beta 0e^{-\gamma r2}(x_j-x_i)^2+\alpha(rand-0.5)$
the first and the second terms are due to the attractiveness. Randomization is the third term.

PSEUDOCODEOFFIREFLYALGORITHM

Input:
Create an initial population of fireflies n within d-dimensional search space x_{ik},i = 1, 2, . . . , n and k = 1, 2, . . . , d
Evaluate the fitness of the population f (x_{ik}) Algorithm's parameter-$\beta 0$, γ
Output:
Obtained minimum location: $x_i$min
Begin repeat
for i = 1 to n for j = 1 to n if (Ij<Ii)
Move firefly i toward j in d-dimension using Equation
end if
Attractiveness varies with distance r via exp[−r2]
Evaluate new solutions and update light intensity using Equation
end for j end for i
Rank the fireflies and find the current best
until stop condition true
end
Where
i = 1, 2, . . ., N, [N-Number of fireflies]

j = 1, 2, . . ., d, [d-Number of dimension]

Position Updation of fireflies

Evaluated fitness values of all bats influence their movements.The movement of the dimmer firefly i towards the brighter firefly j in terms of the dimmer one's updated location is determined by the following equation:

$x_{i+1}=x_i+\beta_0 e^{-\gamma r^2}(x_j-x_i)^2+\alpha(rand-0.5)$

where β_0 is the degree of attractiveness of the firefly at r= 0 γ is a light absorption coefficient

r is the distance between firefly i and firefly j located at xi and xj is calculated by:

$$r=\sqrt{\sum(x_i^k-x_j^k)^2}$$

i,j is the position of any two fireflies

Fitness Evaluation

Fitness Evaluation is done by Support vector machine classifier.In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis[12].SVMs can efficiently perform a non-linear classification[12]. An advantage of Support vector machine is that it only requires a small amount of training data to estimate the parameters necessary for classification.The parameter for evaluation is classification accuracy. Classification accuracy is calculated using confusion matrix given in Figure 4.4. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. TP and TN tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong.

Prediction outcome

P	**N**	**total**
True Positive	**False Negative**	**P'**
False Positive	**True Negative**	**N'**

Confusion matrix

Accuracy= (TP+TN)/(P+N)

Where,

TP- True Positive- These refer to the positive tuples that were correctly labelled by the classifier. Let TP be the number of true positives

TN-True Negative-These are the negative tuples that were correctly labelled by the classifier. Let TN be the number of true negatives

P-Positive- These refer to the positive tuples

N-Negative-These refer to the negative tuples

BINARY FIREFLY ALGORITHM

In firefly algorithm positions are updated using the brightness of fireflies. Instead of brightness, in binary firefly algorithm 0s and 1s are used for representing the position of fireflies .Random numbers are generated. If it is less than 0.5 it is converted to 0 otherwise it is converted to 1. Then according to fitness function positions are updated. The updated positions of fireflies are obtained as decimal number and it is converted to binary numbers using a sigmoid function.

Sigmoid function

In discrete binary spaces, position updating means switching between ''0'' and ''1'' values[13].In order to do this, a transfer function is necessary to map velocity values to

probability values for updating the positions. In other words, a transfer function defines the probability of changing a position vector's elements from 0 to 1 and vice versa[14]. Needless to say, transfer functions force particles to move in a binary space[14].

$S(x_i^k(t))=\frac{1}{1+e^{-x_i^k(t)}}$

$$x_i^k(t+1)=\begin{cases}0 If rand < S\left(x_i^k(t)\right)\\ 1 If rand > S\left(x_i^k(t)\right)\end{cases}$$

where

x_i^k indicate the position i-th particle at iteration t in k-th dimension.

PSEUDOCODE FOR BINARY FIREFLY ALGORITHM

Input:

Create an initial population of fireflies n within d-dimensional search space x_{ik},i = 1, 2, . . . , n and k = 1, 2, . . . , d

β0, alpha, gamma, delta

Output: Number of selected genes and accuracy

Begin

Initialize the population

Evaluate the fitness value

while(max iteration is not met)

for i = 1 to n

for j = 1 to d

update the positions of firefly

Convert the positions to binary using sigmoid function

if rand < sigmoid(x_i^{old})

x_i^{new}=0

else

x_i^{new} =1

end if

next d

Evaluate the fitness value for new solutions using Equation

next n

Rank the fireflies and find the current best solution

end while

Find the best solutions

End

RESULTS AND CONCLUSION

In this section thorough evaluation is made with microarray datasets for both and the comparison between GSBPSO and GSBFA are made to find out the improvement of proposed algorithm.

Dataset Details

Microarray datasets with diverse sizes, features, and classes used in this work are given below in Table 1.

Dataset	**Feature size**	**Sample size**	**Class size**
Colon cancer	2001	62	2
Brain_Tumor	5921	90	9
Lung cancer	12601	203	3
Prostate_Tumor	10510	102	2

Table 1:Dataset Details

Comparison of PSO and FA

It is found that accuracy obtained from GSBFA is better than or equal to the accuracy obtained from GSBPSO and the minimal number of genes are selected in GSBFA when compared with GSBPSO.

Colon cancer dataset:

firefly	99.677
pso	99.465

Lung cancer dataset:

firefly	96.55
pso	94.47

Brain tumor dataset:

firefly	100
pso	100

prostate tumor:

firefly	82.6
pso	78.4

FUTURE WORK

Moreover, the study can be extended to the applications of the selection and classification algorithms that have demonstrated the practical values in studying an expanded range of cancer datasets other than the colon cancer only[15].

REFERENCES

[1] http://hanj.cs.illinois.edu/cs412/bk3/01.pdf
[2] https://docs.rapidminer.com/8.0/studio/operators/modeling/predictive/support_vector_machines/support...
[3] https://www.mathworks.com/matlabcentral/fil eexchange/62214-pso-feature-selection-andoptimization
[4] https://pdfs.semanticscholar.org/ea0c/44680e47dcc2e4e4258b7f8a851f259341af.pdf
[5] https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/supp…
[6] https://en.wikipedia.org/wiki/Firefly_algorithm
[7] https://doi.org/10.1155/2013/283919
[8] https://pdfs.semanticscholar.org/3bf3/8fb45e4e024b6d8a9fa41faa2cb49ec894ab.pdf
[9] http://www.mecs-press.org/ijitcs/ijitcs-v6-n6/IJITCS-V6-N63.pdf
[10] http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0163230
[11] http://download.atlantispress.com/php/download_paper.php?id=5012
[12] https://en.wikipedia.org/wiki/Support_vector_machine
[13] https://pdfs.semanticscholar.org/140f/c48ed4641864cdb93034c0d5c0dfea6b66cf.pdf
[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5373580/
[15] 1-s2.0-S0169260716304163-main.pdf
[16] https://en.wikipedia.org/wiki/Particle_Swarm _Optimization