International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 4, August 2018, pp. 2338~2350
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i4.pp2338-2350

Feature Selection Approach based on Firefly Algorithm and Chi-square

Emad Mohamed Mashhour[1], Enas M. F. El Houby[2], Khaled Tawfik Wassif[3], Akram I. Salah[4]
[1]Computer Science Department, Modern Academy for Computer Science, Cairo, Egypt
[2]Systems & Information Department-Engineering Division, National Research Centre, Giza, Egypt
[3]Computer Science Department, Faculty of Computers and Information, Cairo University, Giza, Egypt
[4]Computer Science Department, Faculty of Computers and Information, Cairo University, Giza, Egypt

Article Info

Article history:

Received Jul 30, 2017
Revised Dec 22, 2017
Accepted Dec 29, 2017

Keyword:

Chi-square
Feature selection
Firefly algorithm
Fitness function
Swarm intelligence

ABSTRACT

Dimensionality problem is a well-known challenging issue for most classifiers in which datasets have unbalanced number of samples and features. Features may contain unreliable data which may lead the classification process to produce undesirable results. Feature selection approach is considered a solution for this kind of problems. In this paperan enhanced firefly algorithm is proposed to serve as a feature selection solution for reducing dimensionality and picking the most informative features to be used in classification. The main purpose of the proposedmodel is to improve the classification accuracy through using the selected features produced from the model, thus classification errors will decrease. Modeling firefly in this research appears through simulating firefly position by cell chi-square value which is changed after every move, and simulating firefly intensity by calculating a set of different fitness functionsas a weight for each feature. K-nearest neighbor and Discriminant analysis are used as classifiers to test the proposed firefly algorithm in selecting features. Experimental results showed that the proposed enhanced algorithmbased on firefly algorithm with chi-square and different fitness functions can provide better results than others. Results showed that reduction of dataset is useful for gaining higher accuracy in classification.

Corresponding Author:

Emad Mohamed Mashhour,
Computer science department,
Modern Academy for Computer Science, Cairo, Egypt.
Email: e_mashhour@hotmail.com

1. INTRODUCTION

A huge dimensionality problem is a kind of problem that appears in dataset swhich needs to be simplified or reduced. It contains a large number of features against small number of samples. A large number of features are considered a huge challenge for any classification process. Using the whole features will enforce the classifier to estimate unseen data with a pre-knowledge of undesirable features, which in turn will produce a poor performance for any classifier [1]. Feature selection can be used for reducing dimensionality of datasets, in order to reduce computation time, cost and classification error. Many researchers used statistical techniques for feature selection, but few of them apply swarm intelligence algorithms for feature selection. Applying swarm intelligence algorithms became a motivation for researchers to solve dimensionality problems due to its capability for selecting the most appropriate features used for classification.

Swarm intelligence approach appeared in 1989 by Gerrado and Jing wang [2]. It was inspired by the mutual behavior that appears on nature, including water and other creatures such as insects. In this kind of

approach each individual or insect is called an agent. Each agent works independently in a type of colony, but this behavior is controlled by certain rules. They cooperate with others in order to finish a certain task. These agents are considered to be a population that interacts with each other in different ways according to the type of the insect. For example, pheromone if among ants, waggle dance among bees to identify source of food and distance, intensity and flashing light among fireflies. Formulating and simulating swarm intelligence behavior depend on the nature of the problem being solved.

In this paper a well-known swarm intelligence algorithm called firefly [3] is used for feature selection. The firefly algorithm proved its capability to solve complex optimization problems. An enhanced firefly algorithm is proposed in this paper to reduce features and select the most informative features for the classification process. The modifications for the standard firefly algorithm are represented through considering the position of firefly as chi-square value assigned for each value in the feature vector. And a set of different fitness functions such as Rosenbrock, Sphere, Ackley [4], Xin-She yang, rastrigin, schwefel, and Salomon [5] are used to represent intensity of fireflies. The performance of the proposed model is tested using the K-Nearest Neighbor (K-NN) [6] and Discriminant Analysis (DA) [7] classifiers to measure the classification accuracy using the selected features.

Different techniques have been applied for feature selection and classification in many literatures. A number of related researches which applied feature selection techniques and classification are highlighted. Sinatabakhi *et al*. [8] proposed a technique for reducing high dimensionality in datasets. An unsupervised gene selection technique was introduced to be applied on microarray datasets such as SRBCT, Colon, prostate tumor, leukemia and lung. The proposed technique utilized ant colony optimization algorithm to minimize the redundancy between genes and increase relevance of genes. They tried variant fitness functions that may improve the classification rate and select lower number of genes. They compared their results with different unsupervised and supervised gene selection methods, classification accuracy has been measured based on three different classifiers which are support vector machine, naïve Bayes and decision tree. Sharma Alok *et al*. [9] introduced a technique for feature selection based on fixed point algorithm. They applied their technique on human cancer datasets using microarray gene expression. The usage of fixed point algorithm incorporated with PCA (principal component analysis) doesn't need class labels for feature vectors. On contrary an eigenvector is computed by multiplying covariance matrix iteratively to select the desired genes. They applied their technique on three public datasets which are SRBCT, ALL and AML, and they used J4.8 and NB for classification. Chinnaswamy Arunkumar and Ramakrishnan Srinivasan [10] proposed a technique for developing feature selection process to reduce high dimensionality datasets. Their technique combined correlation coefficient with particle swarm optimization, in which correlation coefficient was used due to its capability to detect relationship between genes. Particle swarm optimization was used as a searching technique for the most valuable genes. They applied their technique on three microarray datasets which are SRBCT, Lymphoma and MLL. Extreme learning machines classifier was used as a classifier for evaluating the feature selection process. They compared their results with different classifiers such as j48, random forest, random tree, decision stump and genetic programming. Parveen Anisthana *et al*. [11] proposed a feature selection method to eliminate redundant and irrelevant features from datasets and improve classification accuracy. Principal component analysis (PCA), rough PCA, unsupervised quick reducts algorithm and empirical distribution ranking are used for feature selection process. Five datasets are tested for their techniques which are lung cancer, breast cancer, diabetes, heart and ecoli. A number of classifiers were used such as JRip, J48, RBFN, Naïve Bayes, decision table and k-star. KG Srinivasa *et al*. [12] introduced a technique for extracting informative features for classification using fuzzy c-means clustering. The cluster center is created such that it is closer to features with greater membership. Four datasets were tested using their technique, such as physics, sonar, dermatology, and waveform datasets. Two classifier were used which are SVM and Artificial Neural Network (ANN). Mei-Ling Huang *et al*. [13] introduced a framework for solving the problem of data dimensionality by applying feature selection process on datasets. They combined SVM with recursive feature elimination approach. A method called taguchi parameter optimization has been used for identifying the parameter value. They used two public datasets which are dermatology and zoo dataset. Hany M. Harb and Abeer S. Desuky [14] tried to invent a method for reducing features. They used particle swarm optimization for implementing feature selection method, three medical datasets were used: dermatology, breast cancer and heart statlog datasets. PSO is used a searching method for features and CFS is used for measuring the usefulness of each feature. Five classifiers were used for evaluating features which are NB, Bayesian, radial basis function network (RBF), decision Tree and K-NN. They compared their technique with genetic algorithm and different combination between PSO and different classifiers. Pinar yildirim [15] proposed different combinations of feature selection methods and classification techniques to select informative features from high dimensionality dataset. In this research feature selection methods such as Cfs Subset Eval, Principal Components, Consistency Subset Eval, Info Gain Attribute Eval, One R Attribute Eval and Relief Attribute Eval were compared. Hepatitis dataset was

used as a case study due to its serious health problem. Four different classifiers were used which are J48, NB, IBK and decision table. Nancy P *et al*. [16] introduced a study to explore a set of feature selection and classification methods applied on hepatitis dataset. For feature selection fisher filtering, relief filtering and step disc were used. For classification more than 10 classification algorithms has been used, number of features selected for three methods was 6 for fisher filtering, 9 for relief filtering and 4 for step disc. Smita Chormunge and Sudarson Jena [17] proposed a feature selection algorithm based on information gain.They applied a filter method and then applied information gain measure for the subset produced. Two different classifiers were used Naïve bayes and IBK andmedical datasets were used as a case study such as SRBCT. High percentage was produced by Information gain compared with Relief and CHI-square methods.

Researchers from [8] to [17] are using one or more of our datasets, therefore, a comparison of our results with the above researches will be conducted and demonstrated in section 5.1. Feature selection solution was applied on different kinds of datasets, for example in Singh and Chhikara [18] proposed a model for detecting features of images extracted from discrete wavelet transform (DWT) and discrete cosine transform (DCT) using firefly algorithm combined with SVM classifier. Whilst Long Zhang *et al*. [19] detected the most informative features in medical datasets using firefly algorithm based on distance with mutual information criterion. They used K-NN and SVM as classifiers to measure the performance of the proposed technique. V. Subha and D. Murugan [20] introduced a technique for solving the high dimensionality problem for cardiotocogram (CTG) data. Firefly algorithm was used with a novel approach called opposition base learning (OBL). Enny I Sela, *et al* [21] extract features from X-Ray images, researchers developed an algorithm to extract feature of images produced from human body. Samples extracted are for X-Ray dental bone to identify women with low skeletal BMD, J4.8 is used to evaluate the features extracted from feature selection algorithm, results proved that their technique achieve high accuracy, sensitivity, and specificity. Adi Suryaputra Paramita [22] proposed an algorithm for feature selection applied on internet traffic data. They used PCA for extracting discriminant features in data. Fuzzy c-mean is used to improve K-NN classifier performance. By distributing and grouping data into clusters. Results proved that when using PCA as a feature selection solution with K-NN and fuzzy C-Mean, it outperform other techniques.

The remainder of the paper is organized as follows: Section 2 reviews artificial firefly algorithm in general. Section 3 presents the proposed solution in this research followed by the enhanced firefly algorithm. Section 4 presents results and analysis. Section 5 presents experiment discussion followed by a comparative table that compares our approach with other researches. Section 6 presents conclusion & future work.

2. ARTIFICIAL FIREFLY ALGORITHM

Firefly algorithm is considered to be a meta-heuristic algorithm that was inspired by the behavior of flashing lights of real fireflies. The algorithm performance is based on the real behavior of fireflies that relies on the attraction between a firefly and another on basis of their brightness. Formulating the real firefly behavior into an algorithm must follow three rules which govern how the real fireflies act in real space. These rules are as follows:

a. The firefly is a unisex. So, all the fireflies will be attracted to each other regardless of their sex.
b. Attractiveness is proportional to brightness. Therefore, for any flash lighting between two fireflies, the less bright one will move to the brighter one. The attractiveness decreases as the distance increases between two fireflies. The fireflies will move randomly in case there is not a firefly that is brighter than the other.
c. Firefly brightness is influenced or determined by the landscape of the fitness function. In the maximization problem, brightness can simply be proportional to the value of the fitness function [23].

The firefly algorithm relies on two important factors: the light intensity and the attractiveness between fireflies [24]. Light intensity varies in each source according to the brightness of the firefly, which is represented and calculated with a kind of fitness function. Brightness that relies on light intensity determines attractiveness. Attractiveness of each firefly is calculated using the following Equation (1) [24].

$$\beta(r) = \beta_0 e^{-\gamma r^2} \quad (1)$$

Where β_0 represents the attractiveness at distance $(r) = 0$ and sometimes for mathematical computation is considered as 1. γ symbol represents how much the light absorption is.r is the distance between any two fireflies i and j at different positions. Fireflies are always in moving status from position to position. According to the fact of attractiveness between fireflies is related to the distance between them. Hence, the

distance between any two fireflies *i* and *j* is computed through a well-known distance law called Euclidean, which is calculated as follows [24]:

$$rij = \|xi - xj\| = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^{2}} \quad (2)$$

Where *d* represents the dimensionality of the problem, and $x_{i,k}$ is the k^{th} component of the position of firefly *i*. After calculating the distance between the two fireflies, suppose the firefly *i* is less brightness than firefly *j*, so the attractiveness between them occurs while moving the firefly *i* to the firefly *j*. The following Equation (3) [24] controls this kind of movement and it is represented as follows:

$$x_i^{t+1} = x_i^t + \beta_0 e^{-\left(\gamma r_{ij}^2\right)} * (x_j^t - x_i^t) + \alpha * (rand - 1/2 \quad (3)$$

where *t* represents the number of iterations, and the coefficient α represents a random number controlling the size of the random walk, and *rand* represents a random number generator which falls between [0,1]. The firefly with low brightness moves to the higher one after considering three terms [24]. The first term is the current position of the low brightness firefly. Second term is the movement toward the firefly with higher brightness by the attraction coefficient β. Finally, the last term is a kind of random walk calculated by a random generator multiplied by α.

3. PROPOSED METHODS

3.1. Proposed framework of firefly based feature selection

In this research, a framework has been developed to select the most informative subset of features from different datasets based on firefly algorithm. The used firefly algorithm has been modified and combined with different techniques to improve feature selection process which in turn may achieve highest classification accuracy as possible. The modified firefly algorithm is based on an assumption that each dataset contains a number of features *n* and a number of samples *d*. Each feature *i* is a vector of values ($V_{i,k}$), where $(k=1,2,3,....d)$ for different samples $S=(s_1,s_2,s_3,.....,s_d)$, $(i=1,2,3,...,n)$ *for different features*. Modeling features to fireflies is represented by creating *n* fireflies f_1, f_2, f_3, $f_4 ... f_n$. For each created firefly $(f_i),(i=1,2,3,....,n)$, a vector ($x_{i,k}$) of chi-square values is calculated as a mapping vector to the corresponding vector ($V_{i,k}$) in the original dataset to represent firefly position, where $(k=1,2,3,....d)$ to represent a set of different positions for a firefly/feature in different samples. This research aims to apply the firefly framework on microarrays datasets and other kind of datasets by simulating existing features as a number of fireflies, each firefly (feature) has its own position and intensity. Dynamic parameters such as γ, α, β, number of iterations and population size (npop) have been determined by different experiments to achieve the highest performance for feature selection and classification.The proposed framework in this research contains six phases which are as follow (1) pre-processing phase is responsible for dataset filtration from noisy data, searching for missed values in datasets and filling it with reliable values; (2) ranking phase is responsible for sorting the original dataset in descending order according to its evaluation value; (3) firefly position calculation phase is for determining firefly position values for different fireflies; (4) firefly intensity calculation phase is responsible of calculating intensity values for different fireflies; (5) firefly processing phase is used for selecting the highest informative features from the ranked features through applying the modified firefly algorithm, and finally (6) classification phase for evaluating the ability of selected features in classification. If the classification accuracy is acceptable, then a set of features are suitable to classify future unseen data, otherwise the process is repeated with other criteria such as other ranking methods, or fitness function to improve the accuracy. The process continues until reaching the criteria that achieve the highest possible accuracy. The following sections discuss different framework phases:

3.1.1. Pre-processing phase

Huge datasets often suffer from noisy and missed data values that may affect any classifier negatively. In this research the used datasets suffer from missed values; this problem may lead any classifier to unreliable results. In this phase, each feature has been scanned for different datasets searching for missed values; and filling it by considering the average value of the whole feature vector.

3.1.2. Ranking phase

In this phase, different statistical approaches have been used to rank the features. A value is calculated for each feature according to a specific criterion to rank them. In this research, T-test and relieff

techniques were used to rank the pre-processed dataset to pick the highest ranked features first. The ranked features by the two different techniques are introduced to the subsequent phases. In case of huge dataset such as microarrays, the highest ranked features which are the most informative features are selected as candidate for firefly processing.

3.1.2.1. T-test method

T-test is considered as a well-known ranking feature method. T-test is used to measure the difference between two Gaussian distributions. The standard T-test is used to rank datasets with two classes; in the case of this research the datasets may vary to be multi class datasets. A modification was done by [25] in order to calculate the difference between one class and the center of all classes. Calculations are formulated through Equations (4)-(8) [25].

$$TS_i = max\left\{\left|\frac{\bar{x}_{ik}-\bar{x}_i}{m_k s_i}\right|, k = 1,2,\dots,k\right\} \quad (4)$$

Where

$$\bar{x}_{ik} = \sum_{j\in c_k} \bar{x}_{ij} / n_k \quad (5)$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n \quad (6)$$

$$S_i^2 = \frac{1}{n-k}\sum_k \sum_{j\in C_k}\left(x_{ij} - \bar{x}_{ik}\right)^2 \quad (7)$$

$$m_k = \sqrt{1/n_k + 1/n} \quad (8)$$

3.1.2.2. RELIEFF method

Relieff is a kind of method which can be used for feature ranking. It evaluates each feature and assigns a kind of weight for each feature. This weight is assigned according to the capability for this feature to distinguish between classes. RELIEFF technique is used for binary and multiclass problems. This method is more robust and can deal with incomplete and noisy data [26].

3.1.3. Firefly position calculation phase

The changing of position in fireflies relies on the intensity and the movement of the low intensity firefly to the highest intensity one. In this research, the firefly position will be represented using cell chi-square [27]. In this phase, determining positions for different features fi (where i=1, 2…, n) is done by calculating n cell chi-square vectors xi. Each feature vector value $V_{i,k}$ is assigned a relevant vector value $x_{i,k}$ obtained by calculating chi-square for each feature value in vector $V_{i,k}$. The created chi-square vectors are to represent the position values of the fireflies. This is done by measuring each table cell and tests whether it is different from its expected value throughout the whole dataset using Equation (9) [27].

$$x^2 = \Sigma\left((V_i - E_i)^2 / E_i\right) \quad (9)$$

Where (Vi) is the observed value in the feature vector, and (E) stands for the expected value for each cell or value in the feature vector.

3.1.4. Firefly intensity calculation phase

In this phase, each firefly (f_i) is assigned a light intensity value (L_i) calculated by a fitness function. Intensity is used to compare between fireflies in order to decide which have the lower intensity to move with a controlled movement using Equation (3). The firefly with lower intensity updates it's intensity after each movement through a set of iterations. In this research seven different fitness functions have been tried to represent intensity, the goal of utilizing more than fitness function is to search for the best fitness function that can simulate firefly intensity to help in selecting the most informative features that can be used to minimize classifications errors, Table 1 reviews the used fitness functions.

Table 1. Different Fitness Functions used for Simulating Firefly Intensity

Fitness Function	Equation	
Rosenbrock [4]	$f_{\text{Rosenbrock}}(x_1,\ldots,x_n)=\sum_{i=1}^{n-1}\left((1-x_i^2)+100(x_{i+1}-x_i^2)^2\right)$	(10)
Ackley [4]	$f(\langle x_1,\ldots,x_n\rangle)$=20+e-20exp$\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)-exp\left(\frac{1}{n}\sum_{i=1}^{n}cos(2\pi x_i)\right)$	(11)
Sphere [4]	Sphere $(x_1,\ldots,x_n)=\sum_{i=1}^{n}x_i^2$	(12)
Rastrigin [5]	$f(x)$ =10n+$\sum_{i=1}^{n}[x_i^2-10cos(2\pi x_i)]$	(13)
Salomon [5]	$f(x)$ =-cos$\left(2\pi\sqrt{\sum_{i=1}^{n}x_i^2}\right)+0.1\sqrt{\sum_{i=1}^{n}x_i^2}$+1	(14)
Schwefel [5]	$f(x)=418.9829d-\sum_{i=1}^{d}x_i\,sin(\sqrt{\lvert x_i\rvert})$	(15)
Xin-She Yang [5]	$f(x)$ =-$(\sum_{i=1}^{n}\lvert x_i\rvert)exp(-\sum_{i=1}^{n}x_i^2)$	(16)

3.1.5. Firefly processing phase

In this phase, the firefly algorithm is applied as a feature selection process. The process begins with comparing two random fireflies intensity with each other, the one with lower light intensity will move to the higher firefly, distance (r) between them will be calculated using Equation (2), the attraction value is calculated using Equation (1), the new position (x_i) for the lower firefly is calculated using Equation (3), and finally new intensity will be updated through calculating fitness function, this task relies on two important factors:

a. Firefly intensity, the light intensity produced from each firefly in space.
b. Firefly position, the position of the firefly in space, it keeps changing according to some factors

This process continues for a number of iterations or generations specified by the user. In these iterations, the firefly are always in moving status from position to position where the lower light intensity will move to the higher firefly. After that the highest ranked features are introduced to the classifier incrementally starting from the higher intensity (most informative) until reaching the highest possible accuracy.

3.1.6. Classification phase

In the classification phase the fireflies (features) produced from the previous phase are exposed to a classifier. Different machine learning techniques can be used as classifiers; in this research K-nearest neighbor (KNN) and discriminant analysis (DA) are used.

3.1.6.1. K-nearest neighbor classifier

K-nearest neighbor (K-NN) approach is considered as a non-parametric learning algorithm. Non parametric algorithm means it doesn't need to assume any data distribution [6]. The K-NN algorithm is one of the simplest machine learning algorithms and it is considered as instance-based learning, where the unseen data has been classified based on training dataset stored before. The algorithm relies on the distance between the training dataset and the unseen or the testing dataset, the distance is calculated by a kind of similarity measure, such as the Euclidean distance, cosine similarity or the Manhattan distance.

3.1.6.2. Discriminant analysis classifier

Discriminant analysis is an approach used for classification, where two or more groups are known as *a priori* and one or more new observations are classified into one of the known groups based on the measured characteristics. It is used to predict the membership of a sample to a group based on a set of independent variables. The process of discriminant analysis relies on computing the relationship of variables by minimizingdistance the within class distance and maximizing the between class distance simultaneously, to earn the highest class discrimination rate [7].

The framework including different phases for selecting the highest best informative subset of features using firefly is illustrated in Figure 1.

Figure 2 shows pseudo code that integrates the different steps for selecting features and finding the best informative features subset using the proposed firefly framework.

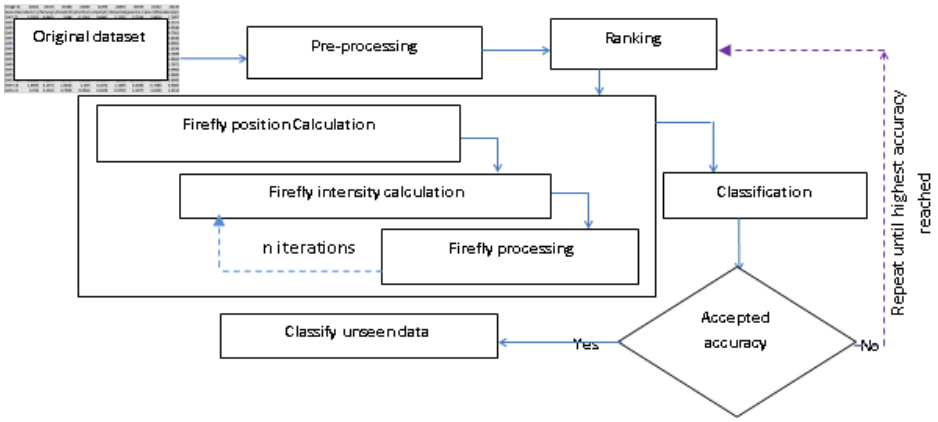

Figure 1. The proposed firefly framework for picking informative features

Input: Matrix M (d, n), M is the original dataset matrix, where d number of samples & n number of features
Output: Matrix S (d, r), S is a reduced matrix dataset, where d number of samples & r reduced number of features
Step 1: Rank features using t-test / relieff for the original dataset M.
Step 2: Initialize parameters:

- Number of iterations t
- Light absorption coefficient γ
- Attraction coefficient β
- Randomization parameter α
- (npop) number of fireflies considered in space

Step 3: Simulate features as fireflies, where each firefly is represented by a vector of values (original values for each feature)
Step 4: Assign position values for each firefly, by calculating cell Chi-square for each value existed in each feature in the dataset.
Step 5: Initialize intensity value for each firefly, through calculating one of the following fitness functions:

- Rosenbrock
- Sphere
- Ackley
- Rastrigin
- Schwefel
- Salomon
- Xin-She yang

Initialize distance r=0. (Distance between fireflies).
Step 6: While (loop < number of iterations)
If intensity (fireflyi) < intensity (fireflyj)
Calculate distance (r) between them.
Calculate attraction based on distance.
Calculate the movement of the low intensity firefly using.
Calculate new chi-square values as new positions for the moved firefly.
After changing position, intensity is recalculated for the moved firefly using fitness function.
Rank fireflies according to the intensity.
Consider the next fireflies to compare.
End while
Step 7: Repeat
Repeat (step 1 to step 6) with different combination until accuracy is acceptable.

Figure 2. Shows pseudo code 1 of the modified firefly algorithm

4. RESULTS AND ANALYSIS

The proposed algorithm has been tested using four public benchmark datasets. The description of these datasets is shown in Table 2. Datasets used are SRBCT microarray cancer dataset and it has been obtained from the GEMS website (www.gems-system.org). The three other datasets (Lung, Hepatitis, and Dermatology) have been obtained from the University of California at Irvine (UCI) machine learning repository [28]. Datasets arerandomly divided into 75% for training and 25% for testing.

After applying the preprocessing and ranking phase on the input dataset, a pool containing a set of features for each dataset ranked by t-test or relieff is constructed first, and then exposed to the modified firefly algorithm experimented with different fitness functions. The output of the firefly processing phase is the highest ranked features subset. These features are passed feature by feature to the classifier in order to evaluate features. Classification rate is monitored until the classification accuracy had been improved as

possible with the most informative fireflies (features). Two ranking methods have been tested which are t-test and relieff, seven different fitness functions have been tried with the firefly processing to represent intensity, and two different classifiers K-NN and discriminant analysis (DA) are used to evaluate selected features. The next sections will demonstrate results produced from applying firefly framework with different combinations of ranking methods, fitness functions, and classifiers on datasets.Our experiment has been applied on four different datasets, such as small round blue cell tumors (SRBCT) which contains 4 different tumors. Lung dataset contains three different class labels. Hepatitis dataset contains two classes (live, die). Dermatology dataset is a kind of dataset that suffer from differential diagnosis of erythemato-squamous diseases.It contains 6 different class labels.

Table 2. Different Datasets used in the Experiments

Dataset	Dataset type	#classes	#features	Samples	Resource
The small round blue cell tumors (SRBCT)	Microarray	4	2308	83	GEMS website (www.gems-system.org)
Lung	Medical	3	56	32	https://archive.ics.uci.edu/ml/datasets/Lung+Cancer
Hepatitis	Medical	Binary class	19	155	https://archive.ics.uci.edu/ml/datasets/Hepatitis
Dermatology	Medical	6	35	366	https://archive.ics.uci.edu/ml/datasets/Dermatology

4.1. Results for SRBCT dataset

This section demonstrates the results of applying the proposed framework on SRBCT dataset. Table 3 shows a comparison among different fitness functions with both classifiers K-NN and DA, and with both ranking method t-test and relieff. It shows thatusing ranking method t-test with rosenbrock function evaluated by classifier K-NN is the best combination for improving classification accuracy. In which it uses 4 genes only toclassify unseen data, with classification accuracy reached its higher value 100%. While usingrelieff with K-NN, the highest classification accuracy reached 95% by 7 genes using Xin-she yang fitness function.

Table 3. Classification Accuracy for each fitness Function with different Classifiers using T-Test and Relieff for SRBCT dataset in (%)

Ranking method	Fitness Function / classifier / #feature	Ackley		Rosenbrock		Sphere		Xin-she yang		Rastrigin		Salomon		Schwefel	
		DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN
T-test	1	55	55	55	55	55	55	50	60	55	60	55	60	45	35
	2	90	80	90	80	90	80	75	85	55	65	55	65	45	30
	3	90	80	90	95	90	80	80	80	75	75	60	80	50	50
	4	85	90	95	100	85	90	80	75	80	80	85	95	65	85
Relieff	5	65	65	65	65	65	65	90	90	65	65	65	65	70	85
	6	65	65	65	70	65	65	90	85	65	65	65	65	75	80
	7	65	65	85	90	65	65	90	95	85	85	65	65	90	85
	8	85	85	85	90	85	85	90	95	85	85	65	70	95	80

4.2. Results for lung dataset

In this section the results of applying the proposed framework on lung dataset will be demonstrated. A set of different combinations of techniques was applied. As shown in Table 4 lung dataset was tested with different combinations of fitness functions, ranking methods and classifiers. Using ranking method t-test withsalomon function and evaluated using classifier K-NN yields 80% with 4 features. Whilstusing DA classifier with ranking method relieff, and sphere function with 4 features increased to higher accuracy 90%.

Table 4. Classification Accuracy for Each Fitness Function with different Classifiers using T-Test and Relieff for Lung Dataset in (%)

Ranking method	Fitness Function / classifier / #feature	Ackley		Rosenbrock		Sphere		Xin-she yang		Rastrigin		Salomon		Schwefel	
		DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN
T-test	1	30	30	50	30	50	30	30	30	40	30	30	30	40	30
	2	30	30	40	30	10	50	50	70	30	20	50	60	50	30
	3	40	40	40	40	20	10	50	60	10	10	50	70	40	30
	4	20	40	40	40	70	30	50	70	50	30	50	80	40	30
Relieff	1	30	30	30	30	30	30	30	30	30	30	30	40	20	30
	2	60	50	60	50	60	50	60	50	60	50	40	50	50	50
	3	60	50	60	50	60	50	60	50	50	50	40	50	30	50
	4	60	60	60	60	90	50	40	60	60	50	20	50	50	60

4.3. Results for hepatitis dataset

A different kind of dataset called hepatitis is tested with the proposed framework. It is considered as a binary class dataset with status die or live. Table 5 shows a number of experiments applied on this dataset with different combinations. Using K-NN classifier with different fitness functions and T-test ranking methods, the best fitness function for this combination wasXin-she yang which gives classification accuracy 79% with 2 features, while usingrelieff with the K-NN, very poor results was produced. Usingrelieff ranking methodand DA achieved the best classification accuracy with both fitness functions rastrigin and Xin-She yang with 85% by 2 features.

Table 5. Classification Accuracy for Each Fitness Function with different Classifiers using T-Test and Relieff for Hepatitis Dataset in (%)

Ranking method	Fitness Function / classifier / #feature	Ackley		Rosenbrock		Sphere		Xin-she yang		Rastrigin		Salomon		Schwefel	
		DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN
T-test	1	60	55	68	50	68	50	81	42	10	21	63	42	68	50
	2	55	66	11	40	74	61	73	79	10	29	55	42	73	60
	3	60	68	18	63	76	71	71	58	26	60	58	60	76	71
	4	52	55	50	71	71	76	76	55	50	71	74	78	71	76
	5	58	55	45	74	45	73	76	47	44	71	71	55	45	73
Relieff	1	76	21	76	21	73	21	73	21	73	21	73	21	81	42
	2	76	71	76	21	73	60	85	50	85	50	10	21	73	60
	3	26	60	71	68	68	66	73	55	71	68	13	18	10	55
	4	50	60	23	60	50	60	68	63	68	63	50	60	10	55
	5	50	60	50	60	50	60	50	60	50	60	50	60	50	60

4.4. Results for dermatology

In this section dermatology dataset is introduced with a comparative Table 6 showing the results of applying the proposed framework. Dermatology is a kind of skin cancer that contains six different classes. Different combinations of techniques are tested for the best performance. The best classification accuracy was obtained through applying K-NN classifier with t-test and using Ackley fitness function, the classification accuracy was 97% by14 features. While using relieff, results were disappointed because it decreases with higher percentage. DA has been used with t-test, the highest classification accuracy 91% was obtained with 9 features using Rosenbrock fitness function, and 95% with 10 features using Schwefel fitness function.

Table 6. Classification Accuracy for Each Fitness Function with different Classifiers using T-Test and Relief for Dermatology Dataset in (%)

Ranking method	Fitness Function / classifier / #feature	Ackley		Rosenbrock		Sphere		Xin-she yang		Rastrigin		Salomon		Schwefel	
		DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN	DA	K-NN
T-test	9	72	75	91	80	85	72	79	70	46	52	72	52	90	72
	10	80	77	89	88	83	81	91	88	64	74	66	54	95	77
	11	80	78	91	89	86	85	91	88	77	74	66	60	95	79
	12	80	78	91	89	95	94	91	87	77	74	78	66	92	82
	13	94	89	91	89	95	95	91	88	77	74	78	66	92	82
	14	95	97	92	86	94	94	93	90	77	74	89	78	93	83
	15	95	97	92	86	94	94	93	90	74	74	89	78	93	83
Reli eff	1	20	35	25	36	54	54	40	36	54	54	24	37	36	36
	2	27	35	38	40	46	58	59	54	36	54	24	37	54	54
	3	53	33	70	58	50	58	72	54	63	54	24	37	61	54
	4	68	51	70	58	50	58	82	58	63	54	35	40	61	54
	5	70	51	73	61	64	56	82	60	63	54	41	40	60	54
	6	70	54	73	63	64	56	75	64	63	54	49	48	63	54
	7	65	50	69	63	71	58	79	66	63	54	77	63	63	54

5. EXPERIMENT DISCUSSION

The proposed hybrid framework describes how firefly algorithm has been used as a feature selection tool, the algorithm was assessed using well-known datasets, and classification error rates produced by the selected features were monitored. It is found that utilizing chi-square for simulating firefly position and different fitness functions for simulating firefly light intensity has improved the firefly performance for feature selection and gives promising results. The classification performance represents how our model succeeds in reducing number of features and selecting the most informative features for classification. In every running trial, different fitness functions have been experimented in order to gain the most suitable fitness function to represent intensity. Each fitness function has been applied on different datasets with two different ranking approaches t-test and relieff. The highest ranked features have been given to different classifiers one by one for evaluation. Testing is done by selecting first feature from the firefly pool of features for classification, then results are evaluated, if classification percentage not accepted another feature from the pool is added to the previous one, and then pass both of them to classifier, and check for classification accuracy percentage, the process is repeated until the highest possible accuracy has been achieved with the lowest number of features. Inside the framework there are a set of parameters must be initialized and tuned for running the modified firefly algorithm, the parameters to be considered such as light absorption coefficient γ, attraction coefficient β, randomization parameter α, number of iterations t and number of firefly population (npop). Number of iterations inside the firefly framework may vary, considering the computation time and cost. After running 500 trails we conclude that the best range of iterations for the proposed firefly model may fall between 150 and 400 iterations. Outside this range may lead the firefly model to pick low informative features that may lead the classifier to poor performance. Number of population chosen for firefly processing relies on number of features picked from the ranking phase, as stated before ranking phase produce as much as possible the most descriptive features ready for firefly algorithm. γ, β and α are three different parameters which may control the behaviour of firefly in space, tuning these parameters needs more than one experiment.

The aim of this research is to focus on improving feature selection process using firefly algorithms, and achieving highest classification rates with lowest number of features. The lower number of features can be selected as long as it keeps the accuracy high. The features extracted from the original dataset, may serve as a feature\gene markers that can recognize and differentiate classes. The following tables show the names of the dominant features\genes that improve the classification accuracy. For SRBCT dataset, Table 7 represents selected genes that may help in successful diagnosis, for lung dataset, there is no a proper description for the features selected, Table 8 shows the dominant features for the hepatitis dataset, while Table 9 shows the important features in the dermatology dataset.

Table 7. Selected Gene Description for SRBCT

Gene Description
Gene 1 : insulin-like growth factor 2 (somatomedin A)
Gene 2 : microtubule-associated protein 1B
Gene 3 : high-mobility group (nonhistone chromosomal)
Gene 4 : ESTs

Table 8. Selected Feature Description for Hepatitis

Feature Description
Malaise : no & yes
Bilirubin

Table 9. Selected Feature Description for Dermatology

Feature Description	Feature type
scalp involvement	Clinical Attributes
scaling	Clinical Attributes
melanin incontinence	Histopathological Attributes
follicular papules	Clinical Attributes
knee and elbow involvement	Clinical Attributes
koebner phenomenon	Clinical Attributes
erythema	Clinical Attributes
PNL infiltrate	Histopathological Attributes
oral mucosal involvement	Histopathological Attributes

Searching for the lowest number of features in any optimization problem for feature selection depends on many factors such as the kind of dataset, its structure, multiclass or binary class, the feature selection model and the classification technique. Our simulation model improved firefly algorithm performance in feature selection which in turn improve the classification process.

5.1. Comparative analysis

For proving how the proposed framework based on firefly improves the feature selection process and also to what extend can compete with other researches. A comparison has been made with other researches that applied algorithms for feature selection. The same datasets used in our research are used in their researches. We claim that our technique outperform other techniques and gives better results, compared with other researches. A comparative table is shown in Table 10.

Table 10. Comparative Table Showing the Proposed Technique over other Techniques with different Datasets

	Number of features\ genes	Technique	Dataset	Percentage
Proposed approach	4	FFA-Chi-square-K.NN-Rosenbrock	SRBCT	K-NN 100%
[8]	10	Microarray gene selection-ant colony optimization	SRBCT	SVM 60%, DT 79% NB 80%
[9]		Fixed-point algorithm, eigenvector, as a classifier: Naive bayes,J4.8	SRBCT	NB 70%,J4.8 70%
[10]	63	Correlation coefficient with particle swarm optimization. Extreme Learning Machines Classifier	SRBCT	ELM93.7%
[17]		Information gain based FeatureSelection Algorithm,Naive bayes and IBK	SRBCT	NB 98%, IBK 98%
Proposed approach	4	FFA-Chi-square-DA-Sphere	Lung	DA 90%
[11]	17	Classifier RBFN, PCA for feature selection	Lung	K-star 90%, RBFN 90%
Proposed approach	2	FFA-Chi-square-DA-Yang	Hepatitis	DA 85%
[15]	12	Naïve Bayes\ConsistencySubsetEval	Hepatitis	Naïve Bayes 85%
[16]	4,6,9	Step disc, fisher filtering, relief filtering	Hepatitis	K-NN, C4.5 85%
Proposed approach	9	FFA-Chi-square-DA-Rosenbrock	Dermatology	DA 91%
Proposed approach	10	FFA-Chi-square-DA-Schwefel	Dermatology	DA 95%
[12]	15	Fuzzy C - Means Clustering	Dermatology	NN 85%
[13]	15-20	SVM-RFE-Taguchi	Dermatology	SVM 95%
[14]	15-22	NB+PSO+CFS Bayesian+PSO+CFS	Dermatology	NB99.45 % Bayesian 99.45%

6. CONCLUSION AND FUTURE WORK

In this paper a novel hybrid framework for feature selection was proposed based on firefly algorithm, the purpose of the framework was to improve the feature selection process. Feature selection process is done by reducing or eliminating irrelevant and noisy features that may have a negative effect on

the classification process. Simulating position with chi-square was suitable representation for firefly position in the space, where every feature/firefly was weighted with a chi-square value in space. Firefly position is an important parameter for controlling distance among fireflies which in turn controls the firefly attractiveness. Simulating and representing light intensity with different fitness functions give the ability of testing a set of different functions that may improve the performance, by selecting the most appropriate fitness function with the highest intensities. Therefore the process of firefly attractiveness was successfully done by attracting the highest intensity features to others. Results of applying the proposed framework on different datasets are promising, and they showed that firefly algorithm can compete successfully as a feature selection tool. Results proved that firefly have the ability to search for the lowest informative bio-marker features that may help in medical diagnosis. The proposed work has been compared with other techniques. It outperforms these techniques in reducing features and achieving the highest classification accuracy using low number of features. The research proved that Rosenbrock, Xin-she yang, Sphere and Ackley are the best fitness functions suitable for simulating intensity. Combining these functions with ranking methods such as t-test and relieff are a good decision for solving problems.

A future suggested model for this research is to develop a full solution or a model using firefly techniques, i.e. both of feature selection and classification processes will be based on firefly. Another suggested work is to create an algorithm for firefly based on parallel processing, in order to l*et al*l fireflies be processed in parallel, which in turn willreduce time processing.

REFERENCES

[1] Zena M. Hira and Duncan F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data", *Advances in Bioinformatics*, vol. 2015, Article ID 198363, 13 pages.

[2] Monali Pat Il, *et al.*, "Survey of different swarm intelligence algorithms", *International Journal of advance Engineering and Research Development*, vol. 1, no. 12, December -2014.

[3] Xin-She Yang, "Firefly algorithm, stochastic test functions and design optimization", *International Journal of Bio-Inspired Computation archive*, vol. 2, no. 2, March 2010.

[4] Galina Merkuryeva, Vitalijs Bolshakovs, "Benchmark Fitness Landscape Analysis", *IJSST*, 2005.

[5] Jian Xie, *et al.*, "A Novel Bat Algorithm Based on Differential Operator and Lévy Flights Trajectory", *Computational Intelligence and Neuroscience*, vol. 2013, March 2013.

[6] P. Bhuvaneswari and Dr. A. Brintha Therese, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm", *2nd International Conference on Nanomaterials and Technologies* (CNT 2014), Materials Science 10 (2015), pp. 433-440.

[7] Desheng Huang, *et al.*, "Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data", *Journal of Experimental & Clinical Cancer Research*, 2009, vol. 28, no. 1, p. 149.

[8] Tabakhi, Sina, Ali Najafi, Reza Ranjbar and Parham Moradi, "Gene Selection for Microarray Data Classification Using a Novel Ant Colony Optimization", *Neurocomputing*, 168, (2015), pp. 1024-1036.

[9] Sharma, Alok, Kuldip K Paliwal, SeiyaImoto, Satoru Miyano, Vandana Sharma and Rajeshkannan Ananthanarayanan, "A Feature Selection Method Using Fixed-Point Algorithm for DNA Microarray Gene Expression Data", *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 18, no. 1 (2014), pp. 55-59.

[10] Chinnaswamy, Arunkumar and Ramakrishnan Srinivasan, "Hybrid Feature Selection Using Correlation Coefficient and Particle Swarm Optimization on Microarray Gene Expression Data", In *Innovations in Bio-Inspired Computing and Applications*, pp. 229-239: Springer, 2016.

[11] Parveen, A Nisthana, H Hannah Inbarani and EN Sathish Kumar, "Performance Analysis of Unsupervised Feature Selection Methods", In *Computing, Communication and Applications (ICCCA), 2012 International Conference on*, pp. 1-7: IEEE, 2012.

[12] KG, Srinivasa, KR Venugopal and LM Patnaik, "Feature Extraction Using Fuzzy C-Means Clustering for Data Mining Systems", *IJCSNS* 6, no. 3A (2006), p. 230.

[13] Huang, Mei-Ling, Yung-Hsiang Hung, WM Lee, RK Li and Bo-Ru Jiang, "Svm-Rfe Based Feature Selection and Taguchi Parameters Optimization for Multiclass Svm Classifier", *The Scientific World Journal*, 2014.

[14] Harb, Hany M and Abeer S Desuky, "Feature Selection on Classification of Medical Datasets Based on Particle Swarm Optimization", *International Journal of Computer Applications*, vol. 104, no. 5, 2014.

[15] Yildirim, Pinar, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", *International Journal of Machine Learning and Computing*, vol. 5, no. 4, 2015, 258.

[16] Nancy.P, Sudha.V, Akiladevi. R, "Analysis of feature Selection and Classification algorithms on Hepatitis Data", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 6, no. 1, January 2017.

[17] Smita Chormunge, Sudarson Jena, "Efficient Feature Subset Selection Algorithm for High Dimensional Data", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 4, August 2016, pp. 1880-1888.

[18] Latika Singh and Rita Rana Chhikara, "An Improved Discrete Firefly and t-Test based Algorithm for Blind Image Steganalysis", *6th International Conference on Intelligent Systems*, Modeling and Simulation, 2015.

[19] Long Zhang, *et al.*, "Optimal feature selection using distance-based discrete firefly algorithm with mutual information criterion", *J. Neural Comput& Applications*, pp. 1-14, 2016.

[20] V. Subha, D. Murugan, "Opposition-Based Firefly Algorithm Optimized Feature Subset Selection Approach for Fetal Risk Anticipation", *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 3, no. 2, June 2016.
[21] Enny I Sela, *et al*., "Feature Selection of the Combination of Porous Trabecular with Anthropometric Features for Osteoporosis Screening". *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 1, February 2015, pp. 78-83.
[22] Adi Suryaputra Paramita, "Improving K-NN Internet Traffic Classification Using Clustering and Principle Component Analysis", *Bulletin of Electrical Engineering and Informatics*, BEEI, vol. 6, no. 2, June 2017, pp. 159-165.
[23] Nadhirah Ali, *et al*., "A Review of Firefly Algorithm", *ARPN Journal of Engineering and Applied Sciences*, vol. 9, no. 10, October 2014 ISSN 1819-6608.
[24] Bin Wang, *et al.*, "A modified firefly algorithm based on light intensity difference", *Journal of Combinatorial Optimization*, April 2016, vol. 31, no. 3, pp. 1045-1060.
[25] Chu, Feng and Lipo Wang, "Applications of Support Vector Machines to Cancer Classification with Microarray Data", *International journal of neural systems*, vol. 15, no. 6, pp. 475-484, 2005.
[26] Robnik-Šikonja, Marko and Igor Kononenko, "Theoretical and Empirical Analysis of Relieff and Rrelieff", *Machine learning*, vol. 53, no. 1-2, pp. 23-69, 2003.
[27] Mary L. McHugh, "The Chi-square test of independence", *Biochem Med (Zagreb)*, 2013 Jun, vol. 23, no. 2, pp. 143-149.
[28] K. Bache and M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 013.<http://archive.ics.uci.edu/ml>.