Experimental and Statistical Analysis on the Performance of Firefly based Predictive Association Rule Classifier for Health Care Data Diagnosis

Nandhini M.*, Rajalakshmi M.,**
Sivanandam S. N.***

** Computer Science, Government Arts College, Udumalpet, India, (e-mail: nandumano@yahoo.co.in).*
*** CSE & IT, Coimbatore Institute of Technology, Coimbatore, India,(e-mail:rajalakshmi@cit.edu.in)*
**** Computer Science and Engineering, Karpagam College of Engineering, Coimbatore, India,(e-mail:sns12.kit@gmail.com)*

Abstract: Health care data diagnosis deals with a prediction of the course of a disease by analyzing the information in health care systems. Analyzing healthcare datasets is one of the major challenges of recent times. Associative Classification (AC) is one of the data mining techniques commonly used for disease diagnosis. AC integrates the concept of Association Rule Mining (ARM) and classification. Though, AC is an efficient classification system, it often experiences poor accuracy as it generates huge volume of class rules in the 'rule generation' phase. This paper intends to address this issue by structuring an associative classifier using significant PARs (Predictive Association Rules) i.e. simply class rules. In this work, Firefly Algorithm (FA), a nature inspired metaheuristic optimization algorithm is adopted to fit into the 'rule generation' phase of existing CPAR (Classification based on Predictive Association Rule), an AC algorithm. This work acquires the essential inspiration of FA and CPAR to construct an associative classifier with significant PARs. FA with a customized fitness function specifically designed for the health care data diagnosis is proposed to find a small set of significant PARs. FA based Predictive Association Rule (FPAR) classifier thus built using significant PARs achieves high prognostic accurateness and interestingness value. Performance of FPAR and CPAR algorithms are analyzed over the six health care datasets from UCI machine learning repository. Based on the experiments, promising results in terms of classifier accuracy are provided by FPAR algorithm.

Keywords: Artificial intelligence; Optimization; Heuristics; Classifiers; Data association; Feature Selection; Diagnosis.

1. INTRODUCTION

Diagnosis of disease is significant as well as a complex task that needs to be completed precisely. Disease diagnosis might lead to false presumption due to physical examination of the patients based on the signs and symptoms. Data mining plays a vital role in discovering and analyzing the hidden knowledge in health care datasets. The diagnosis problem in medical field can be solved using data mining techniques such as ARM, classification and clustering. ARM and classification are the data analysis method used for easy recovery and effective usage of data. ARM is a descriptive data mining task commonly used to extract hidden knowledge in the form of association rules representing the frequent patterns from healthcare datasets. Classification is a predictive data mining task produces a model based on the historical data to predict unknown or future values of the variables of interest. AC is the supervised learning approach integrating ARM and classification techniques. In AC, ARM concentrates in generating significant rules whereas classification exploits the generated rules for classifying the test\unknown tuple. In literature, CBA, CMAR, CPAR etc., are the popular AC algorithms. Existing AC algorithms often experiences poor accuracy, as it generates too many insignificant rules thereby affecting the accuracy of the AC.

FA (Yang, 2010) is a nature inspired metaheuristic optimization algorithm inspired by flashing behaviour of fireflies. Fireflies flash act as a signal system to attract other flies. Generally, FA is formulated based on the three assumptions such as all fireflies are unisexual & individual fireflies will be attracted to all other fireflies, less bright fireflies will be attracted to the brighter one, if no fireflies brighter than given firefly, flies move randomly and the brightness of the firefly is influenced and determined by the objective function. In this paper, a novel fitness function is proposed to optimize FA to generate new and significant PARs. The proposed fitness function is encoded and associated to generate PARs based on the following three assumptions such as individual positive (negative) PARs will be attracted to other positive (negative) PARs. Lower fitness value PARs will be attracted to the higher fitness value PARs. PAR generation process is halted when only one PAR or no PAR in the population is reached. Fitness value of each PAR is influenced and determined by the proposed fitness function. Ultimately, this paper aims in generating significant PARs using FA in less number of iterations to formulate an efficient associative classifier. To conclude, an investigation was carried out to analysis the performance of FPAR and CPAR algorithms using statistical tests. From the investigation, it is found that FPAR outperforms CPAR algorithm in terms of classifier accuracy.

The rest of the paper is organized as follows. Section 2 discusses the literature survey. Section 3 discusses about FPAR. Section 4 describes the methodology followed in the proposed work. Experimental results are discussed in section 5 and section 6 concludes the paper.

2. BACKGROUND AND RELATED WORK

AC is one of the recent data mining techniques builds competitive classifiers with respect to accuracy when compared to classical classifiers such as decision tree, naive bayes and rule-based. It combines the concepts of association and classification. It is widely used for health care data diagnosis (Jabbar et al., 2012; Jabez, 2011; Natarajan and Murthy, 2011). AC algorithms often experience a number of notorious deficiencies as the generation of large quantity of class rules which makes it difficult for an end user to maintain and comprehend its outcome.

2.1 Associative Classification

There are many early versions of AC such as Classification Based on Association (CBA), Classification Based on Multiple Association Rules (CMAR), Classification based on Predictive Association Rule (CPAR). Classification Based on Association (CBA), the first AC algorithm employed Apriori to generate the complete set of class rules from the training dataset. CBA has high misclassification rate, since it uses the best first rule for classifying the test tuples. Classification Based on Multiple Association Rules (CMAR) was proposed to overcome the drawback of CBA. It uses FP Growth algorithm, a best variant of Apriori for ARM. It picks more than one class rule that best matches a test tuple. Though CMAR outperforms CBA in terms of classifier accuracy, it needs to spend more time in selecting the best rules among the huge volume of rules generated in the 'rule generation' phase. First Order Inductive Learner (FOIL), Predictive Rule Mining (PRM) (Yin and Han, 2003) and Classification based on Predictive Association Rule (CPAR) (Yin and Han, 2003) were proposed to generate significant predictive rules from the dataset. FOIL fails to achieve high accuracy because it generates only few significant rules. Shortly, PRM was proposed to attain better accuracy and efficiency than FOIL. CPAR, an extension of PRM, is one of the well known associative classifier yields better accuracy than its predecessors CBA, CMAR, FOIL and PRM. CPAR uses Laplace accuracy, an error estimate measure to evaluate class rules. Based on the Laplace accuracy value, the best k-rules are selected for classifier construction.

Construction of an AC involves two phases such as generation of class rules from the training tuples and the classification of the test tuples using class rules. In the 'rule generation' phase, voluminous rules are generated with class label value as consequent. In the classification phase, the class label of the given test\unknown tuple is predicted using the generated rules. In spite of the powerful mechanism, AC often results in poor accuracy because of generation of large number of insignificant class rules in the 'rule generation' phase. To overcome this drawback, evolutionary algorithms like GA, ACO, PSO, and FA are used in the 'rule generation' phase of AC to generate a optimal set of significant PARs from the dataset.

2.2 Evolutionary Algorithm

Most of the associative classification algorithms in the literature are futile for high dimensional datasets and stipulates optimization. An evolutionary GA (Chien and Chen, 2010) based associative classifier was built to discover trading rules from stock trading data. ACO was proposed to determine the optimal set of association rules to form an accurate rule classifier (Shahzad and Baig, 2011). Associative classifier was also built using Dynamic PSO in (Mangat and Vig, 2014).

Firefly Algorithm (FA) is one of the evolutionary algorithms, inspired by the behaviour of fireflies, attracting each other by flashing light. It is a metaheuristic optimization algorithm proposed by Yang (Yang, 2010). It had been widely applied and proved to be better technique for applications such as digital image compression (Horng and Liou, 2011), feature selection (Banati and Bajaj, 2011), clustering (Senthilnath et al., 2011), job scheduling (Aphirak et al., 2012) etc. It was also used in applications (Kazemzadeh and Kazemzadeh, 2011) which have non-linear and multimodal problems. FA yields better results than PSO and Genetic algorithm (GA) as its parameters can be changed dynamically and it provides optimal solution in less number of iterations.

Marichelvam proposed Discrete Firefly Algorithm (DFA) for flowshop scheduling (Marichelvam et al., 2014). Flowshop scheduling is a NP hard problem used to solve 'n' job in a series of 'm' stages. Discrete Firefly Algorithm (DFA), a variant of FA is proposed to solve this NP-hard problem. The experiments were conducted with the different parameter values and the results were compared with ACO, GA and Simulated Annealing (SA). It was found that the result of DFA provides better results than ACO, GA and SA. Metaheuristic algorithms such as FA and PSO for solving the noisy non-linear mathematical problems were discussed in (Pal et al., 2012). It was concluded that FA was able to find near optimum solution with reduced time. A hybrid filter-wrapper feature selection for load forecasting was proposed based on FA in (Hu et al., 2015). Fire Fly Algorithm(FFA) and Enhanced Artificial BEE Colony Optimization (EABC) were employed to diagnose brain tumor and breast cancer through mammograms along with image processing techniques in (Sahoo and Chandra, 2013; Karaboga and Akay, 2009). It was concluded that FFA outperforms than EABC. FA was employed to train the radial basis function network for data classification and disease diagnosis(Horng et al., 2012). FA had obtained satisfactory results than Gradient Descent, GA, PSO and Artificial Bee Colony(ABC) optimization. Dey in 2014 (Dey et al., 2014) proposed a novel approach to design a robust biomedical content authentication system. FA was applied to generate optimal scaling factors for image embedding. The performance of FA was compared with PSO. Based on the results, it was concluded that FA achieves better results than PSO. Modified FA(MFA) was used to develop the learning rule for identification of three benchmark Infinite Impulse

Response(IIR) and nonlinear plants (Shafaati and Mojallali, 2012). Performance of MFA's was compared with standard FA, GA and PSO. Based on the results, it was proved that MFA is superior in identifying dynamical systems. Younes in 2013 combined FA and ACO for solving economic power dispatch problem(Younes, 2013).

A hybrid model was proposed for heart disease diagnosis (Long et al., 2015). It combines rough set theory and chaos firefly algorithm. The proposed model used chaos FA to enhance the classification accuracy of the heart disease diagnosis with the reduced set of attributes obtained using rough sets. FA was used to extract an optimum set of high accuracy and interpretable fuzzy classifier rules to classify nine benchmark datasets in UCI machine learning repository(Pouyan et al., 2014). (Chao and Horng, 2015) used FA to train the parameters of support vector machine (SVM) classifier for diagnosing the ultrasonic supraspinatus images. Combination of FA-SVM yields better results in terms of classifier accuracy than original LibSVM.

3. FIREFLY BASED PREDICTIVE ASSOCIATION RULE (FPAR) CLASSIFIER

The proposed FPAR is an associative classifier, in which FA is employed to generate significant PARs. This section gives the formulation of the mathematical model of associative classification. It discusses a mathematical model to generate significant PARs using Firefly algorithm. It also gives the stepwise procedure for FPAR classification algorithm.

3.1 Problem definition

Given a finite set of tuples T, it is partitioned into two disjoint tupleset T_0 and T_1, where $T = T_0 \cup T_1$. Each tuple 'A_i' has a 'm' non-class attribute values $A_i = (A_{i1}, A_{i2}, \ldots, A_{im})$ and a class label B_i where $i = 1, 2, \ldots, n$

$$B_i = \begin{cases} 0 \text{ if } A_i \in T_0 \\ 1 \text{ if } A_i \in T_1 \end{cases} \tag{1}$$

In general, the goal of an associative classifier is to build an efficient classification system using significant class rules (i.e. PARs). The main objective of the proposed work is to generate significant PARs that have maximum fitness value. Hence the classification system developed from those significant PARs is able to correctly classify unknown tuples.

3.2 Firefly Algorithm (FA)

FA is an evolutionary algorithm that can be applied to various problems which desires optimization. In this work, it is applied to optimize the class rule generation process resulting with significant PARs in less number of iterations. A PAR which has the maximum fitness value is considered as significant. This optimization problem can be formulated as a mathematical model in order to apply FA.

A model $M = (S, f, C)$ of an optimization problem consists of:

- A search space S defined over a finite set of class rules (i.e. PARs) $A_{ij} \rightarrow B_{T_k}$, where, $A_{ij} \subseteq A_i$, $k = 0,1$
- An objective function is

$$\begin{aligned} &Maximise\ \ f(A_{ij} \rightarrow B_{T_k}) = \alpha * Cosine(A_{ij} \rightarrow B_{T_k}) + \\ &\beta * All - Confidence\ (A_{ij} \rightarrow B_{T_k}) + \gamma * Coverage(A_{ij} \rightarrow B_{T_k}) \end{aligned} \tag{2}$$

Subject to the constraints

$\alpha + \beta + \gamma = 1$, $0 \le \alpha \le 1$, $0 \le \beta \le 1$, $0 \le \gamma \le 1$,where α, β and γ are user specified significance values for Cosine-similarity, All-Confidence and Coverage measures. The significance values of α, β and γ depends on the type of the application and dataset.

- A set C of constraints among the PARs is
  $$A_{ij} \cap B_{T_k} = \phi,\ A_{ij} \ne \phi,\ B_{T_k} \ne \phi \tag{3}$$
- A feasible solution $s \in S$ is a PAR that satisfies all constraints in C.
- A solution $s^* \in S$ is called a global optimum if and only if $f(s^*) \ge f(s)\quad \forall s \in S$.

3.3 Feature Set

A feature set F of 'm_1' non-class attributes which satisfy the user specified minimum support threshold (δ). It is formed for each tupleset T_0 and T_1 separately.

$$F_{T_k} = \left\{F_j \middle| F_j \in A_{ij}, A_i \in T_k\right\}_{i=1, j=1, k=0}^{n, m_1, 1} \tag{4}$$

$$\forall F_j\ \ \exists Supp(F_j) \ge \delta \text{, where } m_1 \le m \tag{5}$$

3.4 Firefly Representation

For each tupleset, from the corresponding feature set (i.e. F_0 and F_1), all possible 1-attribute PARs i.e. antecedent length of the PAR is one are generated. Generated 1-attribute PARs are taken as initial population 'C_k' of the FA which is represented as

$$C_k = \left\{(F_j \rightarrow B_{T_k}) \middle| (F_j \rightarrow B_{T_k}) \subseteq S, F_j \in F_{T_k}\right\}_{j=1, k=0}^{m_1, 1} \tag{6}$$

consisting of m_1, 1-attribute PARs as initial population for each tupleset. In general, initial population represents a potential feasible solution to the problem. With respect to the constraints defined, obtaining the feasible solution is a difficult task. These feasible solutions are evaluated using fitness function and obtain the optimum or near optimum solution with the help of the attraction\ absorption operation.

3.5 Fitness Function

The proposed research work aims to find the significant PARs by taking the advantage of objective function. Each

PAR in a population is considered as a firefly. Hence the feasibility of the fireflies is determined using fitness function. To achieve this goal, Cosine-similarity, All-confidence and Coverage measures are introduced to check the feasibility of the fireflies.

Fitness function $f(F_j \rightarrow B_{T_k}) = f(f_1, f_2, f_3)$ is a function of three variables f_1, $f2$ and f_3, where f_1 is a function used to find the Cosine-similarity of the PAR. Cosine-similarity is a measure used to determine correlation between the rule antecedents towards class attribute (i.e. consequent). f_2 is a function used to find the All-Confidence of the firefly. All-Confidence is used to measure the overall affinity\association among attributes of the antecedent and consequent within a PAR. It works well in skewed support distribution. f_3 is a function used to find the Coverage of the PAR. It is used to determine the comprehensiveness of a PAR.

3.6 Elitism in Firefly

In every generation, the fitness value of fireflies which is greater than equal to Local_Fitness_Threshold *(LFT)* is taken to next generation as best fireflies. Instead of selecting one highest fitness PAR as best PAR using attraction operator, FPAR selects more than one PAR as close-to-the-best PARs using *LFT*. The fitness value of PARs which are greater than or equal to *LFT* are considered as close-to-the-best PARs.

$$LFT = Max(f(F_j \rightarrow B_{T_k})) * FSR \tag{7}$$

Where, *FSR* represents user specified Fitness_Similarity_Ratio, $0 \leq FSR \leq 1$

$$best = \left\{ F_j \rightarrow B_{T_k} \;\middle|\; f(F_j \rightarrow B_{T_k}) \geq LFT \right\} \tag{8}$$

3.7 Attraction / Absorption

The purpose of attraction operator is to generate new PARs (fireflies) significantly different from its parents. To generate new fireflies (PARs), high bright fireflies (i.e. Fireflies with fitness value $\geq LFT$) are attracted by other low bright fireflies among the best fireflies. In this work, this operation is simulated by merging high fitness *(k - 1)* - attribute PAR to other *(k - 1)*-attribute PAR among best PARs to generate k-attribute PARs. Assume R_1 and R_2 are two 1-attribute PARs with fitness greater than the *LFT* are selected for attraction operation. The antecedents of two PARs are combined to form 2-attribute PARs. (i.e. length of the PAR antecedent is 2)

3.8 Algorithm Complexity

For each tupleset, the proposed FPAR algorithm (given in section 4.2) has one outer loop going through the '*m*' non-class attributes for creating initial population. It uses two inner loops and one outer loop when going through the population size '*l*' for '*k*' iteration to generate PARs. Hence the extreme case time complexity is $\boldsymbol{O(2(m+k(l^2+l))}$. As the '*l*' is small, the value of '*k*' becomes small, i.e. because of attraction operation, within few iterations population size reaches to ≤ 1. Hence the computational cost of FPAR is relatively economical as its time complexity is linear in terms of '*k*'. In general, for all metaheuristic algorithms, the evaluation of fitness functions plays a key role in computational cost.

4. PROPOSED METHODOLOGY

The methodology followed in FPAR consists of three phases such as PAR generation, associative classifier construction and the classification of test tuple. Similar to CPAR, FPAR requires the dataset to be partitioned into positive tupleset and negative tupleset. In first phase, significant PARs are generated from each tupleset using FA independently. The second phase involves the construction of associative classifier using the generated PARs. Generated PARs are ordered and ranked according to the Laplace accuracy *('La')* value. Laplace Accuracy is an error estimate measure used in CPAR to determine the significance of a PAR. In the final phase, the best k-PARs from each tupleset that satisfies the test tuple are selected according to the *'La'* value. Average *'La'* value for each best k-PARs is computed. The class of the best k-PARs which has the highest average *'La'* value is chosen as the predicted class label for the test tuple. The detailed work flow of proposed methodology is outlined in figure 1.

4.1 Data Pre-processing

Health care datasets consist of continuous valued attributes which cannot be directly taken for processing. With Weka 3.7 the continuous valued attributes are discretized using Discretize filter. Discretization is performed by simple binning with findNumBins set as False, and number of bins as 10. Each discretized value is represented in numeric format. Even the missing values are handled by replacing with the modes and means of the training data. Specialty of health care data lies in the fact that the attribute values can only be within certain ranges. All the six health care datasets are separately pre-processed using Weka 3.7 according to the requirements of the designed system.

4.2 Generation of PARs using FA and FPAR Classifier Construction

The pre-processed dataset is taken as input for generating significant PARs. Six health care datasets considered in this research work are binary class datasets. Based on the class label, the tuples of the pre-processed dataset is partitioned into positive tupleset and negative tupleset. Since the datasets are discretized using 10 bins. Each attribute has at most 10 discretized values which make PAR generation tedious and time consuming. To address this issue, feature set for each tupleset is formed. A feature set consists of attribute values from each tupleset which satisfy the minimum support threshold*(δ)*. Initially, feature set of the positive tupleset is taken for generating 1-attribute positive PARs. 1-attribute positive PARs are class rules which have only one attribute value from the feature set as antecedent and positive class label as consequent. All possible 1-attribute positive PARs generated from the positive feature set are taken as the initial population of FA. Fitness value for each 1-attribute positive PARs are calculated using the fitness function given in (2).

The brief procedure involved in generation of PARs using FA is detailed in the following algorithm.

PAR Generation algorithm:

```
// $F_i[\,]$ - array to store features\attributes of the tupleset $T_i$
        that satisfy user specified minsupp
// $P_{i,k}[]$ - Population array consists of k-feature\attribute
        PARs generated from tupleset $T_i$
// PAR[ ]- array to store the close-to-the-best PARs (i.e.
        PARs whose fitness value >= $LFT$).
// Fitness( )- used to calculate the fitness value of a PAR
        using equation (2)
// Best_Fitness - parameter used to hold the highest fitness
        value.
// $LFT$- Local Fitness Threshold.
```

Input:

- $T[m]$-is the binary class dataset contains 'm' non-class attributes with class attribute B
- $T_0[m]$- set of positive tuples contains 'm' non-class attributes with class attribute B_0
- $T_1[m]$- set of negative tuples contains 'm' non-class attribute with class attribute B_1
- FSR- user specified Fitness Similarity Ratio

Output: Predictive Association Rules (PARs)

Procedure:

```
(1)  Initialization of parameters
     Best_Fitness = 0;
(2)  Iteration
   (a) Generate 1-attibute PARs from the feature set
       and create initial population

 for i= 0 to 1 do
     (i)    for j=1 to m do
     (ii)        Calculate support($T_i[j]$);
     (iii)       if ( support($T_i\,[j]$) >= minsupp) then
     (iv)            Include $T_i[j]$ to $F_i[\,]$;
     (v)             Add $F_i[j] \rightarrow B_i$ to the $P_{i,1}[\,]$;
     (vi)        end if
     (vii)  end for
     (viii) k=1;
   (b) Calculate fitness for each PAR in the population
       and include best PARs in the rule list

     while ( Size($P_{i,k}[\,]$) > 1 && k < Size($F_i[\,]$)) do
     (i)    for j=1 to Size($P_{i,k}[\,]$) do
     (ii)       Calculate Fitness for k-attribute PAR$_j$
                    in $P_{i,k}[\,]$;
     (iii)          Identify highest fitness value obtained
                    in that iteration;
     (iv)           Best_Fitness = highest fitness value;
     (v)            Calculate $LFT$=Best_Fitness * $FSR$;
     (vi)       if (Fitness(k-attribute PAR$_j$) >=$LFT$) then
     (vii)         Add k-attribute PAR$_j$ to PAR[ ];
     (viii)            else
     (ix)          Remove k-attribute PAR$_j$ from $P_{i,k}[\,]$;
     (x)        end if
     (xi)   end for
   (c) Perform attraction operation
     (i)    for x=1 to Size($P_{i,k}[\,]$) do
     (ii)       for y=1 to Size($P_{i,k}[\,]$) do
     (iii)          if (Fitness(k-attribute PAR$_x$) >
                        Fitness(k-attribute PAR$_y$)) then
     (iv)               Combine the rule antecedents of two
                        PARs to form (k+1)-attribute PARs
     (v)                Add (k+1)-attribute PARs to $P_{i,k+1}[\,]$;
     (vi)           end if
     (vii)      end for
     (viii) end for
     (ix)   k=k+1;
   end while
 end for
 return PAR[ ];
```

Based on the fitness value, PARs are ranked. Instead of selecting one best PAR having highest fitness value for the attraction\absorption, FPAR selects all the close-to-the-best CARs in each iteration using *FSR*. However, in a population there may be few PARs with fitness values similar or close to the highest fitness value. *LFT* is a parameter used to determine all possible close-to-the-best PARs in each iteration. *LFT* is calculated using (7). Fitness value of all the PARs which are close to the *LFT* (i.e. greater than or equal) are taken as the close-to-the-best PARs (8).

4.3 Evaluation and Ordering of PARs

'La' is one of the error estimate measure used to evaluate class rules in CPAR. This error estimate measure takes into account the coverage of each PAR in the training data set.

After evaluation, the PARs are ordered according to their coverage value. The error estimation of each PAR is calculated using (9).

$$Laplace\,Accuracy('\,La'\,) = \frac{(nc+1)}{(ntot+c)} \tag{9}$$

Where, '*c'* represents the number of classes.

'ntot' represents the total number of test tuples satisfying antecedent of the PAR.

'nc' represents the total number of test tuples satisfying both antecedent and consequent of the PAR.

4.4 Classification of test tuples

The PARs among the close-to-the-best PARs satisfying each test tuple are identified and grouped based on the class label. Among the entire set of PARs only the best k-PARs from each class\tupleset having the highest average 'La' value are considered to classify the test tuple thereby eliminating all the lower ranked PARs. The class label of the best k-PARs which has the maximum average 'La' value is taken as the predicted class label for the test tuple.

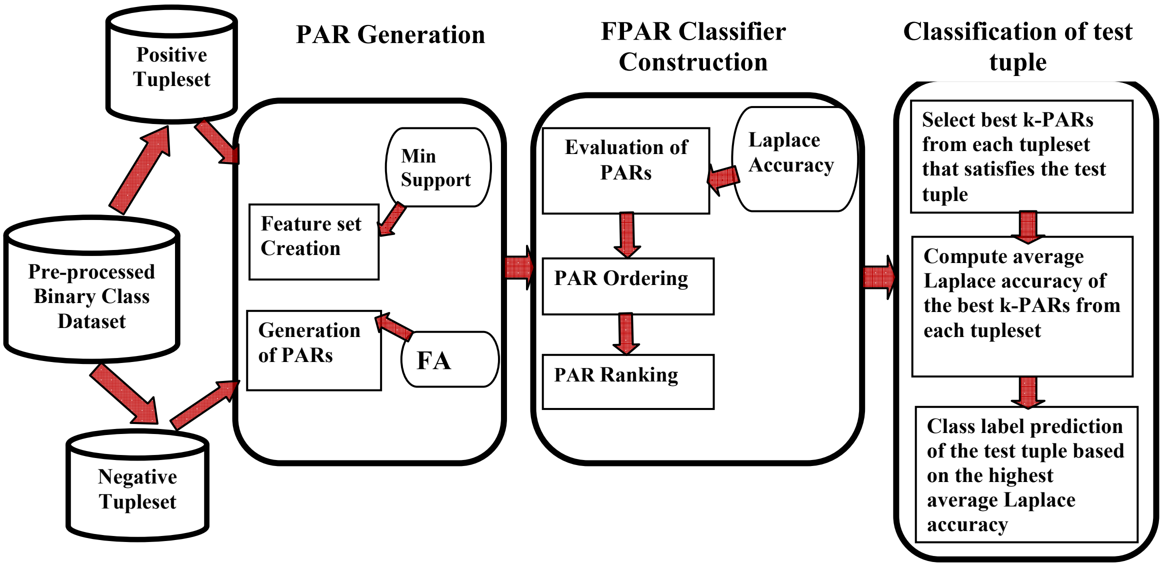
Positive Tupleset
Pre-processed Binary Class Dataset
Negative Tupleset
PAR Generation
Min Support
Feature set Creation
Generation of PARs
FA
FPAR Classifier Construction
Evaluation of PARs
Laplace Accuracy
PAR Ordering
PAR Ranking
Classification of test tuple
Select best k-PARs from each tupleset that satisfies the test tuple
Compute average Laplace accuracy of the best k-PARs from each tupleset
Class label prediction of the test tuple based on the highest average Laplace accuracy

Fig. 1. Workflow of the FPAR Algorithm.

5. EXPERIMENTS AND RESULTS

5.1 Dataset

The experiments are conducted using the six health care datasets such as Breast Cancer, Cleve, Heart, Hepatitis and Pima from UCI machine learning repository. Breast Cancer dataset is taken as a sample for the detailed discussion. This dataset consist of 11 attributes where 11th attribute is the class attribute. The attributes of Breast Cancer dataset is shown in Table 1.

Initially the given dataset is pre-processed as discussed in section 4.1 by categorizing the attribute values based on the domain. The medical dataset should not exceed the certain range of values so pre-processing should be done carefully. The final attribute of the Breast Cancer dataset determines the class label of the dataset, which determines the severity of the disease. “Class” attribute has 2 values ‘benign’ and ‘malignant’. The remaining non class attribute values are also transformed as per the requirements of the algorithm. This pre-processed dataset is then taken as the input for the PAR generation process. Similar pre-processing technique is applied for all other medical datasets such as Cleve, Heart, Hepatitis, Pima and Sick.

Table 1. Attributes of Breast Cancer dataset.

Attribute	Domain values(Integer)
Sample Code Number	Numeric value
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	(0 for benign, 1 for malignant)

5.2 Experimental setup

The pre-processed dataset\tupleset is split into positive and negative tupleset based on the class label. Existence of many features in each tupleset causes generation of numerous 1-attribute PARs in the initial population. Feature selection is a pre-processing step commonly applied before any data mining task. It eliminates insignificant features by keeping good ones without information loss. In this work, the significant feature set is formed from each tupleset using Support. Features satisfying the user-specified minimum support*(δ)* are considered as significant features to be included in the feature set.

In this work, *δ, α, β* and *γ* are set with default values 0.5, 0.5, 0.34 and 0.16 respectively. Using the significant features in the feature set, 1-attribute PARs are generated. Using 1-attribute PARs, higher attribute PARs are generated by means of FA as explained in section 4.2. The PAR generation process is continued until no PARs can be formed or only one PAR is retained in a generation. After the generation of all possible positive PARs from positive tupleset, the same procedure is applied in negative tupleset to generate negative PARs.

Both positive and negative PARs are taken to PAR evaluation and ordering. When the test tuple is considered for classification, the best k-PARs from each class that matches the test tuple are selected. In this work, k is set as 5. The average ‘*La*’ of best 5-PARs that satisfy the test tuple from each class is calculated and the class with highest average *‘La’* is chosen as predicted class label of the test tuple. In this work, the performance of FPAR is compared with existing CPAR under 10CV (10 X fold validation) and 50/50 (50: 50 split) test options. Experiments were performed over six datasets for the varying values of *FSR* (0.5-0.9) and *GSR* (Gain similarity Ratio) (0.5-0.9) within FPAR and CPAR algorithms respectively.

5.3 Results and Discussion

The classifier accuracy obtained by the FPAR and CPAR classifiers under two test options are shown in Table 2 and Table 3. From the Table 2 and 3, it is found that 50/50 test option yields better results than 10CV. 50/50 test option brings considerable improvement in the FPAR classifier accuracy than 10CV test option for all datasets using different *FSR* values (0.5-0.9). Results show that 0.9 and 0.8 are the best values for *FSR* used within FPAR algorithm. In particular, FPAR under 50/50 with *FSR*=0.9 value has achieved a classification accuracy of 97.61%, 96.68%, 96.08% and 94.1% approximately 2%,14%,11% and 14% higher the accuracy achieved using CPAR with *GSR*=0.9 for Breast Cancer, Cleve, Heart and Pima datasets respectively.

From the results, it is found that FPAR has not achieved better accuracy for Sick and Hepatitis dataset than CPAR. As part of analyzing the factors, it is identified that datasets such as Breast Cancer, Cleve, Heart, Hepatitis, Pima and Sick after pre-processing has obtained 10, 13, 13,19,8 and 28 non-class attributes. The number of significant attributes retained in the feature set is high for Sick and Hepatitis compared to other datasets. Feature set size is large for these datasets hence there is no effective dimension reduction happened for these datasets. The number of 1-attribute PARs generated always depends on the feature set size. Since the feature set size of Sick and Hepatitis is relatively large compared to other datasets, FPAR generates many 1-attribute PARs from these datasets thereby making the initial population size large which affects the classifier accuracy. The accuracy of the FPAR and CPAR over the six health care datasets under two test options is represented as line graphs in figures 2 and 3.

5.4 Statistical Validation

Wilcoxon signed-rank test

The Wilcoxon signed-rank test (Wilcoxon, 1945) is a non-parametric statistical hypothesis test used to compare new algorithms with existing successful algorithms. It can be computed using (10).

$\sum R_{+}$ represents the sum of positive ranks, $\sum R_{-}$ represents the sum of negative ranks.

$T = \min(\sum R_{+}, \sum R_{-})$ is used to calculate Z-statistic.

Where, '*n*' represents the number of datasets considered for experiments. When the sample size is less than 30, Z-distribution can be used.

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \qquad (10)$$

If the value of Z is ≤ -1.96, or ≥ 1.96, null hypothesis H_0: there is no significant difference between the proposed and existing algorithms is rejected. Otherwise alternative hypothesis H_a is accepted.

The course of action followed in performing above described Wilcoxon signed rank test to validate the significant differences in the performances between proposed FPAR and existing CPAR classifier for six datasets under two test options (Table 2 & 3) are illustrated as follows:

In this work, experimental results under two test options show better accuracy by FPAR and CPAR algorithms only if the *FSR* and *GSR* values are set with 0.9 and 0.8. The accuracy obtained by FPAR using 0.9 and 0.8 *FSR* values for each dataset under two test options are taken for Wilcoxon signed rank test. As a sample, Wilcoxon ranks are computed for the accuracy obtained by FPAR and CPAR algorithms under 50/50 test option using 0.9 similarity ratio values. Table 4 shows the computed Wilcoxon ranks for six health care datasets. Using the ranks calculated in the Table 4, the Z- statistic is calculated as follows:

$T = \min(\sum R_{+}, \sum R_{-}) = \min(20, 1) = 1$

Since the number of datasets used for experimentation is six less than 30, the Z-distribution is calculated as follows:

$$Z = \frac{1 - \frac{6(6+1)}{4}}{\sqrt{\frac{6(6+1)(2*6+1)}{24}}} = -1.99 \leq -1.96,$$

hence the null hypothesis is rejected thereby alternate hypothesis there is significant difference in the performances between FPAR and CPAR is accepted.

One-way Anova Test

Anova test (DeCoster, 2002) is commonly used statistical test to inspect the significant difference in the performances between the classifiers. The classifier accuracy obtained by FPAR and CPAR classifiers for the 0.9 similarity ratio values under 50/50 test option is taken for performing one-way Anova test. It is performed to assess whether FPAR is significantly different from CPAR at 95% confidence level. Table 5 shows the results of one-way ANOVA test for the classifier accuracy obtained by FPAR and CPAR over six health care datasets under 50/50 test option using 0.9 as *FSR* and *GSR* values respectively. The probability *(p)* value denotes the probability under the null hypothesis. From Table 5, it is found that F=5.431516 > F_{cric}=4.964603 and smaller *'p'* value (i.e. p=0.04201< 0.05) indicates the rejection of null hypothesis, which means that performance of FPAR is significantly different from CPAR. Since the null hypothesis is rejected, post-hoc tests are performed to identify the significant difference between the classifiers.

In this work, paired t-test is conducted for post-hoc analysis. Results of the paired t-test performed over FPAR and CPAR is shown in Table 6. From the Table 6, it is evident that one-tailed p-value is 0.035036681, which is less than the level of significance (0.05), hence the null hypothesis is rejected thereby alternate hypothesis is accepted, stating that FPAR with 0.9 *FSR* value produces better accuracy than CPAR under 50/50 test option over six health care datasets. Results show that FPAR with *FSR*=0.9 yields better accuracy than CPAR with *GSR*=0.9 under 50/50 test option at 95% confidence level.

Table 2. Classifier Accuracy (%) obtained by FPAR for six health care datasets with varying values of *FSR* (0.9 to 0.5).

	10CV					50/50				
Dataset	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**
Breast Cancer	96.58±1.09	96.67±1.33	97.15±1.84	96.9± 0.94	96.91±1.19	**97.61±0.95**	91.27±0.27	90.76±1.13	94.42±1.03	91.17±3.68
Cleve	96.39±1.61	96.3±1.55	96.59±0.85	96.89±0.52	96.68±0.43	**96.68±1.1**	97.26±0.22	96.4±0.58	98.17±0.43	97.6±0.33
Heart	96.19 ±0.9	94.47±2.69	96.35±0.87	96.18±0.78	96.17±0.63	96.08±2.17	**98.51±0.02**	96.99±0.32	95.56±0.54	92.44±0.59
Hepatitis	82.96±3.83	83.38±4.49	83.16±3.6	82.11±3.73	82.62±3.53	80.78±3.32	**85.68±4.12**	82.32±3.7	83.89±4.56	82.12±4.35
Pima	93.08±2.45	93.06±2.34	88.57±3.04	88.63±3.32	88.7±4.41	**94.1±3.24**	90.05±0.65	92.82±1.01	86.92±1.26	88.33±1.09
Sick	83.04±1.78	79.34±1.34	78.27±0.98	79.12±1.2	78.45±0.56	**83.68±1.57**	80.27±1.1	79.54±0.78	78.78±1.43	78.92±1.98

Table 3. Classifier Accuracy (%) obtained by CPAR for six health care datasets with varying values of *GSR* (0.9 to 0.5)

	10CV					50/50				
Dataset	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**
Breast Cancer	95.5±2.02	86.8±2.44	80±3.27	74.7±3.2	71.1±1.66	**95.7±1.06**	91.7±2.83	85.1±3.11	82.3±2.26	77±2.79
Cleve	77.82±4.02	74.2±2.66	73.4±2.72	69±1.83	71.8±2.15	**82.78±1.75**	77.1±2.85	74.6±2.32	72±2.26	69.7±1.89
Heart	78.89±1.58	76.3±1.57	72.6±1.43	69.1±1.85	68.3±2.58	75.6±1.51	**79.7±1.25**	76.5±1.43	71.7±1.7	71.4±2.37
Hepatitis	82.5±4.08	80.5±5.58	79.4±5.07	76.3±6.16	72.3±7.57	**84.24±5.55**	83.6±6.07	79.9±2.29	78.5±6.65	75.3±4.95
Pima	74.95±3.89	71.32±4.32	66.6±2.99	64.7±3.74	61.8±2.1	70.31±2.37	**75.93±1.65**	68.58±1.35	66.28±1.74	66.68±1.92
Sick	83.1±1.45	80.6±1.58	78.5±1.25	80.3±0.95	79.6±1.17	**84.56±1.07**	81.9±1.2	81.1±0.99	79.7±1.16	79.2±1.23

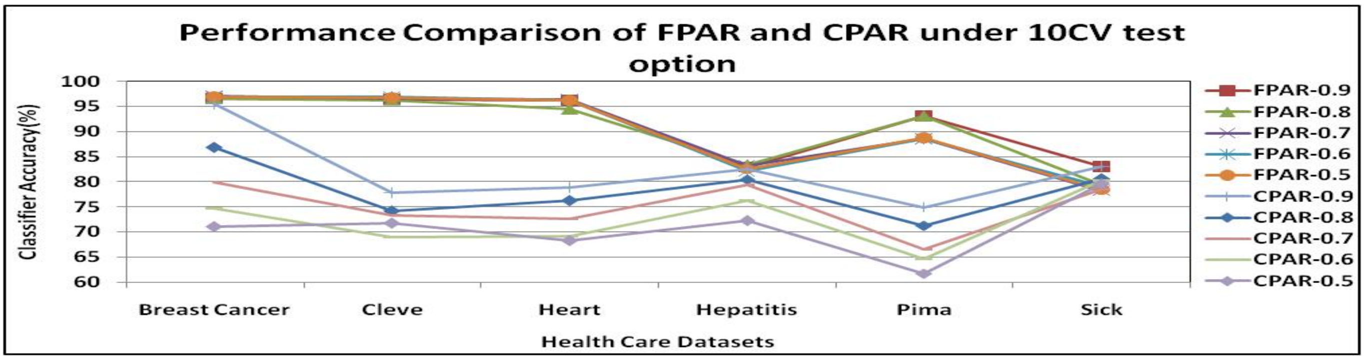

Fig. 2. Performance of FPAR and CPAR algorithms over six health care datasets under 10CV test option.

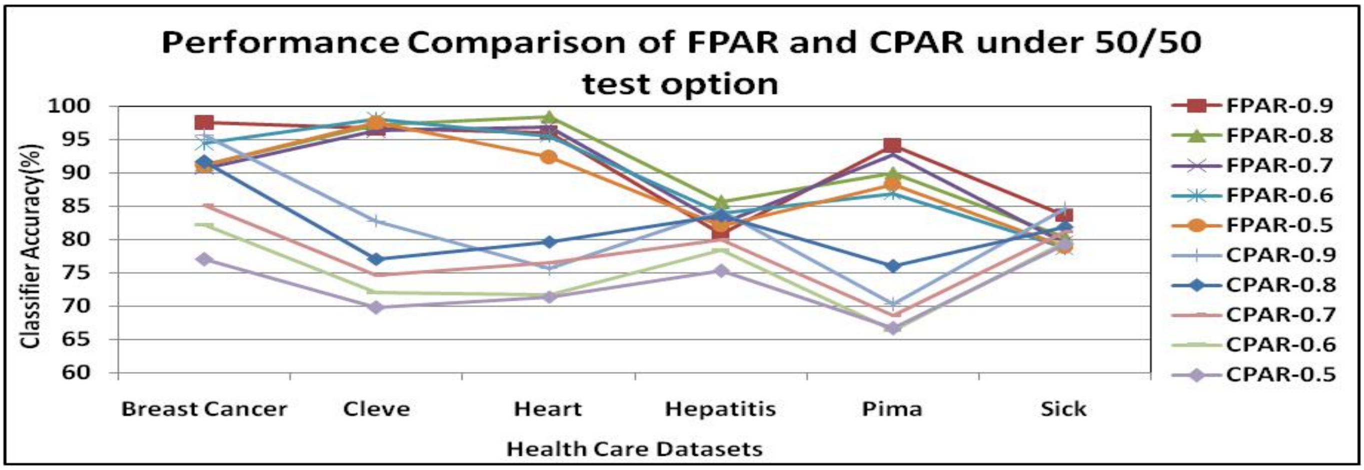

Fig. 3. Performance of FPAR and CPAR algorithms over six health care datasets under 50/50 test option.

Table 4. Wilcoxon Signed Rank Test.

Dataset	Accuracy(%) obtained by CPAR	Accuracy(%) obtained by FPAR	Accuracy Difference	Absolute Difference	Rank	Positive rank (R_+)	Negative rank (R_-)
Breast Cancer	95.7	97.61	1.91	1.91	3	3	
Cleve	82.78	96.68	13.9	13.9	4	4	
Heart	75.6	96.08	20.48	20.48	5	5	
Hepatitis	84.24	85.78	1.54	1.54	2	2	
Pima	70.31	94.1	23.79	23.79	6	6	
Sick	84.56	83.68	-1.18	1.18	1		1
						$\sum R_+$ =20	$\sum R_-$ =1

Table 5. Statistical validation using one-way ANOVA test over six health care datasets.

Source of Variation	Sum of Squares(SS)	Degrees of freedom(df)	Mean square error(MS)	F-value(F)	Probability(p)-value	F critical value(F crit)
Between Groups	304.4161	1	304.4161	5.431516	0.04201	4.964603
Within Groups	560.4626	10	56.04626			
Total	864.8787	11				

6. CONCLUSIONS

In this work, FPAR had been devised to develop an associative classifier with eminent PARs that are valuable for decision-making in health care diagnostic system. From the Table 2 and 3, it was observed that generation of significant PARs using FA encourages the classifier accuracy than CPAR for almost all datasets except Sick. From the experimentation, it was inferred that Sick dataset has many significant attributes in the feature set which leads to the generation of huge volume of 1-attribute PARs. Generation of numerous 1-attribute PARs makes the initial population size large thereby affecting the classifier accuracy. Experimental results also signified that promising results could be obtained only if *FSR* and *GSR* were set with 0.9 or 0.8 values within FPAR and CPAR algorithms respectively. Wilcoxon signed rank test, one-way ANOVA followed by post-hoc paired t-tests (Table 4- 6) were also performed to show that FPAR has brought the considerable significant difference in performance than existing CPAR. In this paper, FPAR algorithm is designed to generate PARs from binary class datasets, further it can be enhanced to generate class rules from multiclass datasets. Suitable feature selection techniques can be explored to reduce the feature subset size. In addition, identifying suitable evolutionary algorithm for generating significant PARs leads to exploration research in future.

Table 6. Post-hoc Analysis using Paired t-test over Six Health Care Datasets.

t-Test: Paired Two Sample for Means		
	CPAR	**FPAR**
Mean	82.24833333	92.32166667
Variance	75.74833667	36.34417667
Observations	6	6
Pearson Correlation	-0.031869716	
Hypothesized Mean Difference	0	
df	5	
t Stat	-2.296553491	
P(T<=t) one-tail	0.035036681	
t Critical one-tail	2.015048372	
P(T<=t) two-tail	0.070073361	
t Critical two-tail	2.570581835	

REFERENCES

Aphirak, K., Sirikarn, C., Thatchai, T., Peeraya, T., Warattapop, C. and Pupong, P. (2012). Application of Firefly Algorithm and its Parameter Setting for Job Shop Scheduling. *The Journal of Industrial Technology*, vol. 8, no. 1, pp.89-97.

Banati, H. and Bajaj, M. (2011). Firefly Based Feature Selection Approach. *International Journal of Computer Science Issues*, vol 8,no.2, pp. 473– 480.

Chao, C.F. and Horng, M.H.(2015). The construction of support vector machine classifier using the firefly algorithm. *Computational intelligence and neuroscience*, pp.2.

Chien, Y. W. C. and Chen, Y. L. (2010). Mining Associative Classification Rules with Stock Trading Data–A GA-Based Method. *Knowledge Based Systems*, vol. 23, no.6, pp.605-614.

DeCoster, J. (2002). Using ANOVA to examine data from groups and dyads. <http://www.stat-help.com/notes.html>.

Dey, N., Samanta, S., Chakraborty, S., Das, A., Chaudhuri, SS. and Suri, JS. (2014). Firefly Algorithm for Optimization of Scaling Factors during Embedding of Manifold Medical Information: an Application in Ophthalmology Imaging. *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 3, pp. 384-394.

Horng, M.H and Liou, R, J. (2011). Multilevel Minimum Cross Entropy Threshold Selection Based on the Firefly Algorithm. *Expert Systems with Applications*, vol. 38, no.12,pp. 14805–14811.

Horng, M.H., Lee, M.C., & Liou, R.J. (2012). Firefly Algorithm for Training the Radial Basis Function Network for Data Classifications. *Advanced Science Letters*, vol. 11, no.1, pp. 755-758.

Hu, Z., Bao, Y., Xiong, T. and Chiong, R. (2015). Hybrid Filter–Wrapper Feature Selection for Short-Term Load Forecasting. *Engineering Applications of Artificial Intelligence*, vol.40, pp.17-27.

Jabbar, M.A., Deekshatulu, B.L. and Chandra, P. (2012). Heart Disease Prediction System using Associative Classification and Genetic Algorithm', *Proceedings of the international conference on emerging trends in electrical, electronics and Communication* technologies, pp.183-192.

Jabez, C.J. (2011). A Statistical Approach for Associative Classification. *European Journal of Scientific Research*, vol. 58,no.2, pp.140-147.

Karaboga, D. and Akay, B. (2009). A Comparative Study of Artificial Bee Colony Algorithm. *Journal of Applied Mathematics and Computation*, vol. 214, no.1, pp. 108–132.

Kazemzadeh, A.S. and Kazemzadeh, A.S. (2011). Optimum Design of Structures using an Improved Firefly Algorithm. *International Journal of Optimization in Civil Engineering*, vol. 1, no. 2; pp. 327-340.

Long, N.C., Meesad, P. and Unger, H.(2015). A highly accurate firefly based algorithm for heart disease prediction. *Expert Systems with Applications*, vol.42,no.21, pp.8221-8231.

Mangat, V. and Vig, R. (2014). Dynamic PSO-based Associative Classifier for Medical Datasets. *IETE Technical Review*,vol.31, no.4, pp. 258-265.

Marichelvam, M. K., Prabaharan, T. and Yang, X. S. (2014). A Discrete Firefly Algorithm for the Multi-Objective Hybrid Flowshop Scheduling Problems. *IEEE Transactions on Evolutionary Computation*, vol. 18, no.2, pp. 301-305.

Natarajan, S and Murthy, KNB (2011). A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction, *IJCA Special Issue on Artificial Intelligence Techniques – Novel Approaches & Practical Applications*, AIT, no. 3, pp.18-23.

Pal, S. K., Rai, C. S. & Singh, A. P. (2012). Comparative Study of Firefly Algorithm and Particle Swarm Optimization for Noisy Non-Linear Optimization Problems. *International Journal of Intelligent Systems and Applications (IJISA)*, vol.4,no.10, pp.50.

Pouyan, M.B., Yousefi, R., Ostadabbas, S. and Nourani, M., 2014, May. A Hybrid Fuzzy-Firefly Approach for Rule-Based Classification. *In The Twenty-Seventh International Flairs Conference*.

Sahoo, A. and Chandra, S. (2013). L'evy-Flight Firefly Algorithm based Active Contour Model for Medical Image Segmentation. *In Contemporary Computing (IC3), 2013 Sixth IEEE International Conf.*, pp. 159-162.

Senthilnath, J., Omkar, S. N. and Mani, V. (2011). Clustering Using Firefly Algorithm: Performance Study. *Swarm and Evolutionary Computation*, vol.1,no.3,pp.164-171.

Shafaati, M. and Mojallali, H. (2012). Modified Firefly Optimization for IIR System Identification. *Journal of Control Engineering and Applied Informatics*, vol.14,no.4, pp.59-69.

Shahzad, W and Baig, A. (2011). Hybrid Associative Classification Algorithm using Ant Colony Optimization. *International Journal of Innovative Computing, Information and Control*, vol.7, no.12, pp.6815-6826.

Wilcoxon.F (1945). Individual Comparisons by Ranking Methods. *Biometrics,* vol.1,pp. 80–83.

Yang, X.S.(2010).*Nature-Inspired Metaheuristic Algorithms*. Luniver press.

Yin, X. and Han, J. (2003). CPAR: Classification Based on Predictive Association Rules. In*: Proc. Int. Conf. on Data Mining*, pp. 331–335.

Younes, M. (2013). A Novel Hybrid FFA-ACO algorithm for Economic Power Dispatch. *Journal of Control Engineering and Applied Informatics*, vol.15, no.2, pp. 67-77.