Forward Weighted Firefly for Feature Selection in Health Care Dataset

D. Livingston 1, Mrs. S. Vydehi 2 & Dr. M. Punithavalli 3

M. Phil Scholar, Dept. of Computer Science, Dr.SNS College of Arts and Science, Coimbatore, Tamil Nadu, India 1
Professor & Head Of the Department, Dept. of Computer Science, Dr.SNS College of Arts and Science, Coimbatore, Tamil Nadu, India 2
Associate Professor, Bharathiar University, Coimbatore, 3

Abstract: The process of selecting a subset of features from original data source is extensively hiring great importance in recent years in all scientific and technological fields. On account of considering the situation this paper proposes an optimal feature selection technique WFF (Forward Weighted Firefly) for feature selection in health care dataset.

1. Introduction

The Feature Selection in health care dataset is a significant area in data mining. The art of extracting a subset of features from is the core competent of this paper. The process of selecting optimal features from Health care dataset is a tedious task for further diagnosis and progression.

During, the feature selection process there is chances for eliminating the useful features and critical attributes which results in data loss. In order to minimize the data loss the optimality process is considered in this paper.

This paper focus on two issues, the former is extracting the subset of features and the latter is to ensure the optimality of the selected features.

The paper is organized as follows, the brief report of the supportive studies is explained in section II. The Methodology explanatory is discussed in section III. The Outcome of the research Experiment is presented in section IV. Conclusion and Future remarks are given in section V.

2. Literature Survey

Normally, the feature subsets space is exponentially grown with the number of features. The explored data size affects the feature selection process. Therefore, the need for incorporating heuristic search optimization methods is essential to guide the search for an optimal feature selection. (W. Daelemans et al., 2003 proposes an approach that instead of optimizing the full feature subset, the existing work focuses to find the optimal model parameters for selecting the features. But the outcome of the experimental results of this work shows there is no guarantee for the subset selected to be optimal.

[Y. Deng et al, 2004] proposed a K-means clustering based on PSO. The PSO helps to gain the optimality in inter cluster similarity measures. Herewith, in this paper PSO process is preceded with two parameters namely force and velocity.

Alba et al., 2007 have introduced an enhanced version of PSO for gene selection, called Geometric PSO [GPSO]. However, the experimental results are less significant because the geometric PSO is more about generalizing optimizers based on a notion of distance where different distance metrics give rise to different operators with regards to the predefined geometric operators.

Also, Binary PSO [BPSO] an enhanced version of the optimized algorithm PSO have been proposed by Chuang et al., 2008. The methodology produced 100% accuracy in classification for several data sets. The limitation behind this algorithm is that it involves large number of selected genes to achieve the high accuracy.

Through several studies it is better understood that nature inspired algorithms are more potential for problems related to optimization solution. X.-S. Yang, 2009 proposed a novel approach for reducing the dimensionality by incorporating firefly algorithm (FA). Thereby, the Firefly algorithm works by flashing behavior with the objective function to produce optimized Solution for the feature problem.

(Hassanzadeh and Mohammad) proposed a new hybrid algorithm called K-FA for clustering the dataset. The core job of K-FA is to find the centroid of the cluster.
Hence, through the background study it is crystal clear that the optimization algorithm is more essential for feature selection in large dataset. The next section briefly describes about the methodology.

3. Forward Weighted Firefly

3.1. Firefly

Firefly algorithm is naturally inspired by fireflies based on biochemical and social aspects. This firefly attracts the other fireflies by producing a short and rhythmic flash. It also produces as protective warning mechanism to its kind. The firefly algorithm is also termed as meta-heuristic search algorithm and it was proposed by Xin-She Yang. The firefly depends on the intensity of flash in order to attract the other fireflies. The higher intensity fireflies are more attracted and hence in the feature selection process the intensity refers to the best position for acquiring the optimal feature. In this work the FA is enhanced to select the optimal features.

3.2. Working principle of Firefly

The firefly algorithm stresses on the intensity parameter of flash which attracts the other fireflies. If the intensity is higher the other fireflies are more attracted. Hence, the intensity is termed as the best position in the feature selection process. The parameter also helps in acquiring the optimal feature from the large dataset.

The firefly algorithm works with the help of variable called global optima. The variable is derived by calculating the brightness of a firefly. The firefly is compared with the neighbors, the brighter firefly is more attracted by its neighbors. The firefly that has maximum influence over its neighbors is referred as global optima.

The proposed Forward Weighted Firefly algorithm (WFF) incorporates the basic behavior of Firefly Algorithm (FA).

The algorithm is initiated with n number of features denoted as $d_i$ wherein i ranges from 1 to n. The corresponding function in conditional feature set $c_f$ is denoted by

$$ (f d_i = 1, 2, 3, \ldots, n) $$

The intensity of the features in firefly $I_i$ of each firefly $f d_i$ is set with its dependency function $D f_i$, which also denotes the set of decision features.

Generate initial population with $c_f$.

The light intensity of each firefly is derived at step 4, when the conditional feature is not equal to the decision feature then each firefly i start moving towards its best solution j that procures minimum distance with $r_{ij}$ and this movement results in increased dependency of the feature.

In case, if any of the firefly is unable to find any best solution, then the intensity of firefly i is absorbed and it will be invisible to all other fireflies in the space.

The weight parameters $w_i$ is defined to each firefly to accelerate the outcome. Furthermore, the mean and variance value for each $f d_i$ in each subset is calculated for satisfying the criteria at step 19-27.

The result derived at step 27 is treated as updated light intensity which then proceeds to find new solution for movement of i towards j for each firefly with feature selection.

Among the solution derived as best solution which has minimal and required dependency are selected.

The algorithm is iterated with the same procedure for new groups of fireflies generated in previous iteration and determines the intensity $I_{ij}$ of each group $f d_i$ until the above condition is satisfied at step 33. The algorithm for Weighted firefly proceeds as follows

STEP 1: Initialize the set of all conditional features $c_f$

STEP 2: Initialize the set of decision features $D f_i$

STEP 3: Initialize the objective function $R = \{ x : x \subseteq c_f, y_x(D f_i) = y_{ca}(D f_i) \}$ rate an initial population with fireflies $(f d_i = 1, 2, 3, \ldots, n)$ corresponding to conditional features $c_f$.

STEP 4: Light intensity $I_i$ at $f d_i$ is determined by $I_{xi} = y_{xi}(D f_i)$

STEP 5: Set $F_c = c_f$

STEP 6: While $y_{xi}(D f_i)! = y_{cf}(D f_i)$

STEP 7: $F = [ \} ] c_f$

STEP 8: $F = F_c$

STEP 9: for i=1: F’ fireflies

STEP 10: for j=1: F’ fireflies

STEP 11: Find the best solution j for i that satisfies the following conditions

STEP 12: Intensity of j is greater than intensity of i.

STEP 13: if $(I_j > I_i)$, where j=1,2,……n $j \neq i$
STEP 14: Distance between i and j should be minimum in terms of distance between 
\[ y_{X_iX_j}(Df_i) \] and \[ y_{CF}(Df_i) \] i.e
\[ r_{ij} = \min(y_{CF}(Df_i) - y_{X_iX_j}(Df_i)) \]

STEP 15: Movement of i towards j increases the light intensity of j i.e 
\[ y_{X_iX_j}(Df_i) = I_{X_j} \]

STEP 16: Define weight parameters \( w_i \)

STEP 17: Set \( \sum_{i=1}^{n} w_i = 1 \) where \( w_{i+1} < w_{i+2} < w_{i+3} \)

STEP 18: Find mean best position between two fireflies

STEP 19: \( x_i^{t+1} = x_i^t + r_{ij} \cdot w_i \)

STEP 20: Display the mean best position

STEP 21: Evaluate new solutions and update light intensity with best position

STEP 22: Calculate the mean value for each \( x_i \) position at STEP 20

STEP 23: For each position calculate variance

STEP 24: Calculate variance 
\[ \text{variance} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \text{mean})^2} \]

STEP 25: Light intensity is updated with variance and mean

STEP 26: \( x_i^{t+1} = x_i^t + \text{variance} \cdot w_i \)

STEP 27: Return current best position

STEP 28: Select the best position among best solutions.

STEP 29: End for j

STEP 30: Move firefly i towards j;

STEP 31: \( I_{ij} = y_{X_iX_j}(Df_i) \)

STEP 32: \( F = F \cup I_{ij} \)

STEP 33: end for i

STEP 34: end if

STEP 35: Evaluate each \( x_{ij} \) in feature F for dependency i.e.

The above procedure helps in retrieving the optimal features from the original feature space. The next section depicts the outcome of the algorithm.

4. Experimental Results and Discussion

The experiment is carried out on health care dataset like liver and cancer dataset. The outcome of the experiment is evaluated using the parameters Precision, Sensitivity, Accuracy and specificity and it is compared with the existing technique called Genetic algorithm (GA).

- **Specificity** – measures the proportion of negatives that are correctly identified.
- **Sensitivity** – measures the proportion of positives that are correctly identified
- **Accuracy** – Determines the correctness
- **Precision** – Repeated process same result

<table>
<thead>
<tr>
<th>Parameters</th>
<th>GA</th>
<th>WFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.9219</td>
<td>0.9389</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.9331</td>
<td>0.9438</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.9233</td>
<td>0.9408</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.0669</td>
<td>0.0562</td>
</tr>
</tbody>
</table>

Table 1. Liver Dataset
The result in table 1 and figure 1 depicts the better performance of proposed WFF in liver dataset.

### Table 2. Cancer Dataset

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Genetic</th>
<th>WFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.9303</td>
<td>0.9547</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.9419</td>
<td>0.9538</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.9333</td>
<td>0.9442</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.0581</td>
<td>0.0462</td>
</tr>
</tbody>
</table>

Figure 2. Comparative result in Cancer dataset

The result in table 2 and figure 2 depicts the excellent performance of the proposed algorithm WFF.

### 5. Conclusion

Thus the Weighted firefly algorithm helps in selecting the optimal features from the health care dataset. In future the paper can be extended with other optimization algorithm.

### 6. References


