Finding Accuracy in Feature Selection Using Firefly Algorithm with Rough Set theory

A. Revathi, Dr. P. Sumathi

Abstract—Feature selection techniques play a vital role in bioinformatics applications. In addition to the large group of techniques that have already been developed in the machine learning and data mining fields, specific applications in bioinformatics have led to possess of newly proposed techniques. In this paper, a method for feature selection is based on Firefly Optimization (FFO) with Rough Set Theory(RST) is proposed. Data sets include a large volume of features with irrelevant and redundant features. Redundant and irrelevant features reduce accuracy. The main aim of this paper is to select a subset of relevant features. A statistical metric-based feature selection technique has been proposed in order to reduce the size of the extracted feature vector. The proposed method shows the improvement significantly in terms of performance measure metrics: accuracy, sensitivity, specificity, computation time and so on. FFO technique is applied to determine the features globally according to the light intensity. Then the selected features are grouped together to make a subset and applied RST to find the optimized feature. This optimized feature is used to analyze the protein information in the organisms and improve the feature selection accuracy and reduce the computation time in protein data analysis.

Keywords: Bioinformatics, Feature, Firefly, Optimization

I. INTRODUCTION

Extracting the biological data from the large amount of data is reducing the dimensionality reduction in bioinformatics data analysis. It plays a vital role to estimate the protein information. Proteins consist of one or more long chains of amino acid residues.

Manuscript received November 15, 2017.

A. Revathi, Assistant Professor, Department of Computer Science, New Prince Shri Bhavani Arts and Science College, Medavakkam, Chennai-100,Tamilnadu, India

P. Sumathi, Assistant Professor, Post Graduate and Research Department of Computer Science, Government Arts College, Coimbatore.

Proteins perform the duties specified by the information encoded in genes. Since the protein data set contains large amount of data, it is difficult to extract the relevant features from them.

Therefore, an optimal feature selection is performed to reduce the data dimensionality. Feature Selection (FS) is the pre-processing step in data mining, particularly when dealing with high dimensional space of features. Therefore, the selection process is used to select a relevant subset of features from the original set of features. It is necessary to decrease the irrelevant features by selecting the most relevant features. This is also used for reducing of feature space and training time hence it improves the accuracy. Several data mining methods have been provided the additional information about the protein sequence. Feature selection methods help to create an accurate predictive model. They help to choose features that give better accuracy though requiring less data. Feature selection methods such as Filter method, Wrapper method and Heuristic search method and soon have been used to identify and remove unwanted, irrelevant and redundant features from data that do not give the accuracy of a predictive model or decrease the accuracy of the model .

A multi-objective particle swarm optimization (PSO) was introduced in [1] for feature selection but it is failed to select the optimal feature subset. A novel random-forest-based predictor MePred-RF) was introduced in[2] using Minimum Redundancy and Maximum Relevance (MRMR) and Sequential Backward Search (SBS) approach to select the optimal feature subset from the ranked features set. But, it does not perform dynamic feature selection strategy.

Protein-protein Binding Site alignment technique was developed in [4] for Feature point and region selection. But it takes more computation time for feature selection. Analysis of variance (ANOVA) based method combined with incremental feature selection (IFS) was introduced in [5] to optimize the feature set. On the other hand, the accuracy of feature selection was not improved at a required level.

Random Forest-Recursive Feature Elimination (RF-RFE) method was introduced in [6] for determining the optimal features. But the method failed to have an ability to distinguish the protein features. A tri-gram feature extraction method was developed in [7] for protein fold recognition. But the accuracy of feature extraction was not improved.

In order to overcome such type of issues, Firefly Optimization technique with Rough Set Theory is introduced to find the accuracy and optimal feature selection. Contribution of the paper is as follows, Firefly Optimization based Rough set theory (FFO-RST) is introduced to improve the feature selection accuracy on protein sequence dataset and determine the optimized feature in the dataset using Rough Set Theory (RST**).**

II. METHODOLOGY

A. Firefly optimization based rough set theory for feature selection

In biological data analysis, the feature selection is also known as attribute selection to select optimal features from the large volume of data. When number of features increase, the number of risks will be increased and degraded the performance of data mining and machine learning tasks. The feature selection reduces the curse of dimensionality. Dimensionality reduction is the process of reducing the number of features which is missing and unwanted in the dataset. It improves the performance and reduces the complexity of the data set. To obtain such type of benefits, Firefly Optimization with Rough Set Theory is developed. The architecture of Firefly Optimization is shown in Figure 1.

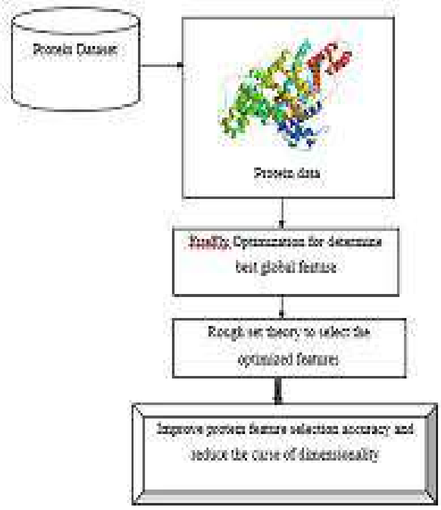
Protein Dataset
Protein data
Firefly Optimization for determine best global feature
Rough set theory to select the optimized features
Improve protein feature selection accuracy and reduce the curse of dimensionality

Figure 1: Architecture of the Firefly Optimization Rough Set Theory

Figure 1 shows that the architecture of the proposed Firefly Optimization based Rough Set Theory (FFO-RST) .The FFO-RST consists two processing steps for selecting an optimal protein feature subset. Initially, Firefly Optimization technique is applied to select the global feature and remove the missing and unwanted features from the dataset. Secondly, Rough Set theory is applied to select the optimized protein features from the dataset. The selected feature used for further sequence analyzing, which is explained in the following section.

B. Firefly optimization (FFO) for best global feature selection

The processing diagram of the Firefly Optimization to select the best feature. Initially, the population of fireflies generated. Then the intensity value of fireflies determined and the determined intensity values are updated. Finally, the fireflies are ranked accordingly. The diagrammatic representation of this technique is as follows: The processing diagram of the FFO is described as shown in figure 2.

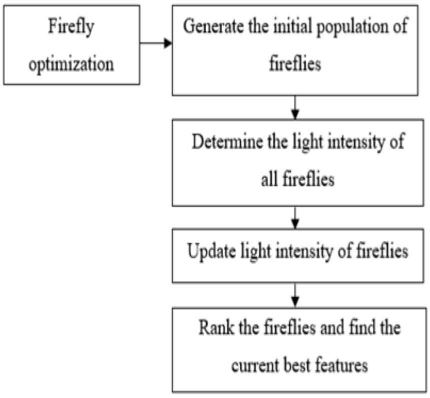
Firefly optimization
Generate the initial population of fireflies
Determine the light intensity of all fireflies
Update light intensity of fireflies
Rank the fireflies and find the current best features

Figure 2: Flow diagram of Firefly Optimization

Let us consider the number of fireflies $f_i = f_1, f_2, f_3......f_{n.}$ The fireflies are randomly positioned in feature space. The light intensity of firefly is denoted as (LI). The intensity is related with the objective function F(x).

Where, LI(f) = F(x) (1)

From (1), $LI\ (f)$ denotes the light intensity of the Firefly. Then the Firefly with higher intensity attracts the other one i.e $LI\ (f_i) > LI\ (f_j)$ where $j = 1,2,3, \dots n$ but i$\neq j$. The attractiveness changes with the degree of absorption as light intensity decreases with distance. The light intensity is a function of distance 'D' is formulated as,

$$LI\ (D) = L_o e^{-\gamma r} \qquad (2)$$

From (2), 'L_o' represents the actual light intensity and 'γ' denotes the light absorption coefficient. Firefly's attractiveness is proportional to the light intensity. Therefore, attractiveness varies with the distance between the two fireflies.

$$A_{ij} = D\,(f_i, f_j)$$

(3)

From (3), A_{ij} denotes the attraction of the two fireflies, $D\ (f_i, f_j)$ is the distance between the two fireflies. Therefore, the distance between the two fireflies is measured by using Euclidean distance measure.

$$D\ (f_i, f_j) = \sqrt{\sum_{i,j=1}^{n} (f_j - f_i)^2} \qquad (4)$$

C .Rough set theory for optimized protein feature selection

Once the best features are determined from the Firefly optimization technique, the selected features are grouped and applied the Rough Set approach to identify and select the optimized feature. Rough set theory (RST) is used to perform features dependencies and reduce the number of features.

D. Attribute dependency

Bioinformatics data analysis is used to discover which features are strongly related to other features. In RST, the concept of dependency is defined as follows: Let us consider two (disjoint) sets of features, and find out what degree of dependency obtains between them. Each attribute set makes an indiscernibility equivalence class structure. The attribute dependency is obtained as fellows,

$$\gamma_R(Q) = \frac{\sum_{i=1}^{n} \left| Pos_R\ (Q) \right|}{|\,U|} \quad (5)$$

From (5), Where Pos_R denotes a positive boundary region and |U| denotes the cardinality of set features during the feature selection process. R is a set of condition attributes and Q is the decision.

E. Rough set theory for optimized protein feature selection

Once the features are determined from the Firefly Optimization techniques, the selected features are grouped and applied Rough Set Theory to identify and select the optimized feature. The RST is used to identify an optimal feature set from the collection of data set in terms of the indiscernibility relationship between lower and upper approximation data. The relationship 'p' is defined as follows:

$$IND(R) = \{(u,v) \subset U^2 | \forall a \subset R, p(u) = p(v)\} \quad (6)$$

From (6), the relation $IND(R)$ is called as 'R' indiscernibility relation. u, v is the indiscernible value by attributes from R. R is the relationship between the lower approximation ($\underline{R}X$) and upper approximation ($\overline{R}X$) of set 'X' which is defined as follows,

$$\underline{R}X = \{u \in U : [u]_{IND(R)} \subseteq X\} \quad (7)$$

$$\overline{R}X = \{ u \in U : [u]_{IND(R)} \cap X \neq \emptyset \} \quad (8)$$

Where,

$$[u]_{IND(R)} = \{v \in U : p(u) = p(v), \forall a \in R\} \quad (9)$$

The boundary set of 'X' is defined as differentiation between the upper approximation and lower approximation. The boundary set is expressed as follows,

$$BND\ (X) = \underline{R}X - \overline{R}X \quad (10)$$

Let us consider R, Q $\subseteq Attr$ is the equivalence relations over U. Therefore, the positive boundary region and negative boundary region are defined as follows. A positive region of the Q is a set of all the features (i.e. attributes) which is classified with certainty to one class employing the attributes from R,

$$Pos_R\ (Q) - \cup_{u \in U/IND\ (Q)} RX \quad (11)$$

$$Neg_R\ (Q) = U - \cup_{u \in U/IND\ (Q)} \overline{R}X \quad (12)$$

From (11), (12), $Pos_R\ (Q)$ is the positive region and $Neg_R\ (Q)$ represents the negative region of the set. In the RST, the positive and negative region boundary value is evaluated and those values are used to calculate the degree of the features.

The following algorithms are used for finding the relevant and optimized feature.

Firefly Optimization Algorithm [8] describes the selection of relevant feature.

The Firefly Optimization algorithm is formulated as:

Input: Number of fireflies$f_i = f_1, f_2, f_3 \dots \dots f_n$, Light intensity LI, parameter controlling step size 'α_t', attractiveness of the Firefly 'A', light absorption coefficient γ'

Output: Determine the best global feature (Firefly)

```
Step 1: Begin
Step 2: Define the objective function F(x)
Step 3: Generate an initial population of fireflies
Step 4: Formulate light intensity LI associated with the objective function F(x) using (1)
Step 5:    Define light absorption coefficient 'γ'
Step 6:    while ($t < Max\ generation$)
Step 7:      for $i = 1 : n$ (all n fireflies)
Step 8:        for $j = 1 : n$ (all n fireflies)
Step 9:          Compute the distance $D$ between Firefly $f_i$, and Firefly $f_j$ using (4)
Step 10:         if ($LI_{f_i} > LI_{f_j}$)
Step 11:           attractiveness changes with distance $D$ via $exp(-\gamma r)$;
Step 12:           Move Firefly $f_j$ towards Firefly $f_j$ using (5)
Step 13:           Calculate new solutions and update light intensity using (5)
Step 14:         End if
Step 15:       end for $j$
Step 16:     end for $i$
Step 17:     Rank the fireflies according to the light intensity and determine best features
Step 18:   End while
Step 19: End
```

The brightness should be associated with the objective function.

RoughSetTheory algorithm is formulated as:

For finding the optimized features, the following algorithm is used.

```
Input:    Datasets, Number of features
$f_i = f_1, f_2, f_3 \dots\dots f_m$
Output:  Select optimized features
Step 1:   Begin
Step 2:   For each selected feature
Step 3:   Measure the indiscernibility relationship between lower and upper approximation (6)
Step 4:   Measure lower and upper approximation to identify the features within the boundary using (7) (8)
Step 5:   Identify positive and negative region to obtain the degree of features using (11) (12)
Step 6:   Measure the attribute dependency using (13)
Step 7:   If $(\gamma_R(Q) \leq 1)$ then
Step 8:   Select the optimized features
Step 9:   else
Step 10:  Features are not optimal
Step 11:  End if
Step 12:  End for
Step 13:  End
```

III. RESULTS AND DISCUSSION

The FFO-RST is compared against with the existing multi-objective particle swarm optimization (PSO) and Minimum Redundancy Maximum Relevance(MRMR) and Sequential Backward Search (SBS) (MRMR-SBS) approach. Firefly Optimization based Rough set theory (FFO-RST) is experimented using JAVA language with WEKA tool for selecting an optimal feature subset by using VariBench Dataset. Four different datasets namely Protein tolerance datasets, Protein stability datasets, mRNA splice site datasets and Transcription factor binding site dataset with four different parameters are taken from the VariBench Dataset. The experiment is conducted on the factors such as feature selection accuracy with number of features in protein tolerance dataset. Experimental results are compared and analyzed with the help of table and graph.

A. Impact of feature selection accuracy

Feature selection accuracy is defined as the ratio of number of optimized feature are correctly selected and incorrectly selected for protein data analysis to the total number of features. Feature selection accuracy is defined as follows,

$$FSA = \frac{\text{optimized feature are correctly selected} + \text{incorrectly selected}}{\text{No.of features}} * 100$$

Where FSA denotes the feature selection accuracy.

Table 1: Tabulation for feature selection accuracy

No. of features	Feature selection accuracy (%)		
	FFO-RST	PSO	MRMR-SBS approach
50	90.35	81.36	73.10
100	91.22	84.65	76.24
150	92.58	86.10	80.13
200	93.61	87.52	81.55
250	94.10	88.10	83.74
300	95.58	89.47	84.69
350	96.74	91.45	85.79
400	97.10	92.52	86.77
450	97.58	93.58	89.74
500	98.87	94.47	90.10

Table 1 describes the performance results of feature selection accuracy with different input features from 50 to 500. The feature selection accuracy is comparatively increased in proposed FFO-RST when compared to existing PSO and MRMR-SBS.

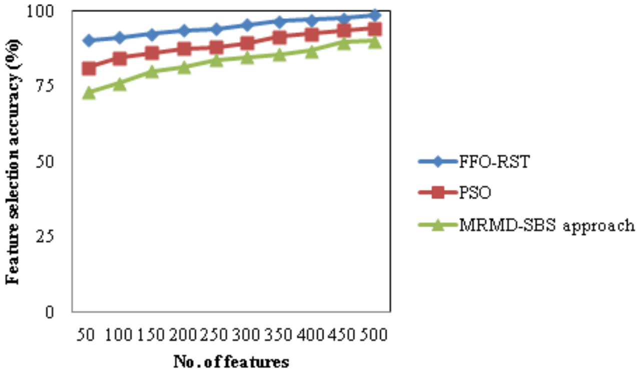

Figure 2 Measure of feature selection accuracy

Figure 2 shows the performance analysis of the feature selection accuracy with respect to number of features in dataset. Among the several features, the protein features are selected to analyze the protein sequence and protein functions. Firefly Optimization with Rough Set Theory helps to obtain the similar features to improve the protein functionality. Therefore, the feature selection accuracy is increased by 7% and 14% compared to existing PSO and MRMR-SBS respectively.

IV. CONCLUSION

The FFO-RST consists of two processing steps to improve the accuracy of the feature selection. Initially, the Firefly Optimization (FFO) is applied to rank the fireflies according to the light intensity for selecting the feature set from the whole feature set. This helps to select the relevant features and removes the irrelevant features from the dataset. After that, optimized features are determined and selected by using rough set theory. This helps to obtain the curse of dimensionality reduction. Therefore, an optimal feature is selected based on the attribute dependency measure and improve the accuracy with minimum computation time. The performance result shows that the proposed FFO-RST significantly improves the feature selection accuracy. In future work, we can compare the accuracy with some other evolutionary algorithms and get the accuracy in disease prediction.

REFERENCES

[1] Bing Xue, Mengjie Zhang , Will N. Browne, “Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach, “IEEE Transactions on Cybernetics, Volume 43, Issue 6, 2013, Pages 1656 – 1671

[2] Leyi Wei , Pengwei Xing , Gaotao Shi , Zhi-Liang Ji , Quan Zou,” Fast prediction of protein methylation sites using a sequence-based feature selection technique”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume PP, Issue 99, Pages 1-12

[3] Nancy Yu Song and Hong Yan, “Autoregressive and Iterative Hidden Markov Models for Periodicity Detection and Solenoid Structure Recognition in Protein Sequences”, IEEE Journal of Biomedical and Health Informatics, Volume 17, Issue 2, March 2013, Pages 436 – 441

[4] Jamal Ahmad, Faisal Javed, Maqsood Hayat, “Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods”, Artificial Intelligence in Medicine, Elsevier, Volume 78, 2017, Pages 14–22

[5] Mohamed F. Ghalwash, Xi Hang Cao1, Ivan Stojkovic and Zoran Obradovic, “Structured feature selection using coordinate descent optimization”, BMC Bioinformatics, Volume 17, Issue 158, 2016, Pages 1-14

[6] Alok Sharma , Seiya Imoto and Satoru Miyano, “A Top-r Feature Selection Algorithm for Microarray Gene Expression Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics , Volume 9, Issue 3, May-June 2012, Pages 754 – 764

[7] Bin Pang, David Schlessman, Xingyan Kuang, Nan Zhao, Daniel Shyu, Dmitry Korkin, and Chi-Ren Shyu, “An Integrated Approach to Sequence-Independent Local Alignment of Protein Binding Sites”, IEEE/ACM Transactions On Computational Biology and Bioinformatics, Volume 12, Issue 2, 2015, Pages 298-308

[8] https://en.wikipedia.org/wiki/Firefly_algorithm

A. REVATHI is working as an Assistant Professor, Department of Computer Science, New Prince Shri Bhavani Arts and Science College, Medavakkam, Chennai-100. She is pursuing PhD in the area of Datamining, Bharathiar University, Coimbatore. She did her M.Phil.(CS) in Alagappa University, Karaikudi. She has completed her MCA at Alamelu Angappan College, Komarapalayam. She has presented papers in various National and International conferences and published papers in International Journals. She has twenty years of experience in teaching and guided MCA projects. Her research interests include Data Mining, Computer Networks and Software Engineering.

Dr.P.SUMATHI is working as an Assistant Professor in the post Graduate and Research Department of Computer Science, Government Arts College, Coimbatore. She did her PhD in the area of Grid Computing in Bharathiar University. She has done her M.Phil in the area of Software Engineering in Mother Teresa Women’s University. She did her MCA degree at Kongu Engineering College, Perundurai. She has published many national and International journals. She has about twenty years of teaching and research experience. Her research interests include Data Mining, Distributed Computing and Software Engineering. .