Multi Chaotic Flow Direction Algorithm for Feature Selection

Wy-Liang Cheng[1], Li Pan[1], Mohd Rizon Bin Mohamed Juhari[1], Abhishek Sharma[2], Hameedur Rahman[3], Chun Kit Ang[1], Sew Sun Tiang[1,*], Wei Hong Lim[1,*]

[1]*Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur 56000, Malaysia*

[2]*Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India*

[3]*Faculty of Computing and Artificial Intelligence, Air University, Islamabad Capital Territory 44000, Pakistan*

E-mail: 1001436889@ucsiuniversity.edu.my, 1002060534@ucsiuniversity.edu.my, mohdrizon@ucisuniversity.edu.my, abhishek15491@gmail.com, rhameedur@mail.au.edu.pk, angck@ucsiuniversity.edu.my, tiangss@ucsiuniversity.edu.my, limwh@ucsiuniversity.edu.my,

Abstract

Feature selection is a crucial pre-processing step used to remove redundant information from original datasets while preserving the accuracy and processing time of classifier. The feasibility of using metaheuristic search algorithms (MSAs) such as Flow Directional Algorithm (FDA) to solve feature selection problems is one of the active research topics. Similar with other MSAs, FDA also employs conventional initialization scheme that generates initial solutions in random basis. The absence of intelligent mechanisms in conventional initialize scheme tends to generate initial populations in local optima, hence compromising the performance of algorithm to handle datasets with complex features. In this paper, a modified algorithm known as Multi Chaotic Flow Directional Algorithm (MCFDA) is proposed to solve feature selection problems with enhanced performances by leveraging the strengths of multiple chaotic maps for population initialization. A total of 12 datasets from UCI Machine Learning Repository are selected for performance evaluation of MCFDA and another four peer algorithms to solve feature selection problems. The proposed MCFFA is revealed to deliver best performances by solving 7 out of 12 datasets with the best mean classification accuracy and 6 out of 12 datasets with the least numbers of selected features.

Keywords: chaotic map, feature selection, Flow Directional Algorithm (FDA), metaheuristic search algorithm

1. Introduction

Due to the emerging of big data era, data-driven methods such as machine learning and deep learning have become more popular to solve various real-world problems. The performances of these data-driven methods depend on the quality of datasets used for model training. In the real-world scenario, redundant and noisy information tend to be observed from the input datasets and their presence tends to degrade the performance of data-driven methods. Feature selection[1] is a popular pre-processing method used to remove the undesirable information from original datasets without compromising the accuracy and speed of data-driven methods. Some popular applications of feature selection are fault detection[2],[3] and automatic modulation recognition[4],[5]. Feature selection is considered as a challenging NP-type optimization problem because it involves the optimal selection of feature subset from the original input datasets.

Existing feature selection can be divided into two main approaches, i.e., filter- and wrapper-based methods. For filter-based methods, the potential useful feature subsets are determined based on statistical characteristics of datasets. For wrapper-based methods, a data-driven classifier is incorporated along to carefully evaluate the quality of potential feature subsets. Both approaches have their pros and cons. Filter-based methods are more computationally efficient with lower accuracy. Opposite behaviors are found from wrapper-based methods.

Motivated by their promising characteristics such as strong global search and easy implementation, many metaheuristic search algorithm (MSAs) inspired by various natural phenomena[6] (e.g., theory of evolution, animal forging behaviors, physics-based phenomena and human-based activities) are used to handle numerous complex optimization problems[7],[8],[9],[10],[11] including the feature selection. Flow Directional Algorithm (FDA) is a physic-based algorithm proposed in 2021 by emulating the flow direction within a drainage basin. Similar with other MSAs, FDA generates its initial population in random basis without considering any intelligent mechanisms that can prevent the generation of initial solution in local optima[12]. This undesirable behavior can compromise the performance of FDA due to premature convergence.

A modified algorithm known as Multi-Chaotic Flow Directional Algorithm (MCFDA) is designed to solve feature selection problems competitively. Multiple chaotic maps with the ergodicity and non-repetitive characteristics are incorporated into initialization phase of MCFDA, aiming to generate initial solutions in the more promising solution regions. The performance of MCFDA to solve feature selection problems is evaluated using 12 datasets obtained from UCI Machine Learning Repository and compared with 4 peer algorithms.

2. Related Works

2.1. *Inspiration of FDA*

Excessive rainfall happens when the rainfall pours over the surface of ground and does not seep into the soil. Direct runoff refers to the additional water that is not absorbed after rainfall and losses. A method known as μ–index approach[13] is proposed based on this inspiration.

An index μ refers to the average number of water loss during precipitation with a unit of centimeter per hour, where this left-over water will change into runoff. The direct runoff value can be obtained by the difference of index μ at each interval of rainfall. The direct runoff w_d can be calculated as:

$$w_d = \sum_{g=1}^{G} (W_g - \mu \Delta t) \quad (1)$$

where Δt and G refer as the length of each time interval and total number of time steps, respectively.

The movement of runoff towards the outlet of basin is influenced by the angle of slope and can be emulated by splitting the drainage basin into several set of cells. The slope height and angle in a surrounding cell can also affect the amount of runoff in each cell moving to other cells. D8 method[14] can be used to verify the variation of runoff direction. Accordingly, each cell consists of eight neighbors with unique height and distance. Comparison between the distance and height of each cell with its adjacent cells is used to distinguish the flow direction. Referring to all cell slops obtained, the cell flows toward the adjacent cell with highest slope. The flow path is determined by using D8 for the basin. After the flow path is indicated, a variable corresponds to the number of cells that flow into that cell is defined. The greatest number is assigned to the outlet point of basin.

2.2. *Feature Selection Optimization Problem*

Feature selection is a bi-objective problem used to maximize the classification accuracy and minimize the numbers of selected features, simultaneously. The fitness function used to evaluate the quality of each solution X during feature selection process is represented as:

$$f(X) = kz + g\frac{|F_s|}{|F_o|} \quad (2)$$

where $k \in [0, 1]$ and $g = 1 - k$ are the parameters used to represent the weightage of classification error z and the length of selected feature subset $|F_s|$; $|F_O|$ refers to the length of original input dataset.

3. Proposed MCFDA

During the initialization phase, an initial chaotic variable $\sigma_0 \in [0,1]$ is randomly generated for each i-th MCFDA solution. At any k-th iteration, the chaotic variable σ_k can be updated with the chosen chaotic map, where $k = 1, \ldots, k_{max}$ and k_{max} refers to the maximum iteration numbers. In this paper, four chaotic maps are used to update σ_k. If $\sigma_k < 0.25$, a circle map is selected. For $0.25 \le \sigma_k < 0.5$, a Gauss map is used. Singer map is chosen to update chaotic variable if $0.5 \le \sigma_k < 0.75$. Finally, Sinusoidal map is selected when $\sigma_k \ge 0.75$. The Circle, Gauss, Singer and Sinusoidal maps used to this study are presented in Eqs. (3), (4), (5) and (6), respectively, as follows:

$$\sigma_{k+1} = mod\left(\sigma_k + 0.2 - \left(\frac{0.5}{2\pi}\right)\sin(2\pi\sigma_k), 1\right) \quad (3)$$

$$\sigma_{k+1} = \begin{cases} 1 & , \sigma_k = 0 \\ \dfrac{1}{mod(\sigma_k, 1)} & , otherwise \end{cases} \quad (4)$$

$$\sigma_{k+1} = 1.07(7086\sigma_k - 23.31\sigma_k^2 + 28.75\sigma_k^3 - 13.302875\sigma_k^4) \quad (5)$$

$$\sigma_{k+1} = c\sigma_k^2 \sin(\pi\sigma_k), \quad c = 2.3 \quad (6)$$

At the final iteration of k_{max}, the final chaotic variable $\sigma_{k(final)}$ is obtained and used to initialize the d-th dimension of i-th MCFDA solution as follow:

$$X_{i,d}^{flow} = X_d^{LB} + \sigma_{d(final)}(X_d^{UB} - X_d^{LB}) \quad (7)$$

where X_d^{LB} and X_d^{UB} represent as the minimum and maximum values of d-th decision variable. As compared to the conventional initialization scheme described in Eq. (8), the ergodicity and non-repetition characteristics of chaotic map can produce more thorough search in the solution space and prevent premature convergence. The initial population of MCFDA with the population size I and dimensional size D is generated using Eq. (7) for $i = 1,..,I$ and $d = 1,..,D$, before moving to next stage.

$$X_{i,d}^{flow} = X_d^{LB} + r_1 * (X_d^{UB} - X_d^{LB}) \quad (8)$$

During the iterative searching process, it is assumed that there are β neighborhoods around every flow, where the position is represented as shown:

$$X_j^{flow} = X_i^{flow} + r_2 * \Delta \quad (9)$$

where X_j^{flow} is refer the position of each j-th neighbour for every i-th MCFDA solution; X_i^{flow} is current position of i-th flow; r_2 is a random number between 0 and 1 generated by uniform distribution; Δ is a parameter used to control the behavior of MCFDA, i.e.,

$$\Delta = \left(r_3 X_i^r - r_4 X_i^{flow}\right) * \left\|X^{best} - X_i^{flow}\right\| * W \quad (10)$$

where r_3 and r_4 are the random numbers between 0 and 1 generated by uniform distribution; X_i^r is a randomly flow obtained in related with Eq. (1); X^{best} is the flow position with the best (i.e., lowest) fitness; W is a nonlinear weight with random number, i.e.,

$$W = \left(\left(1 - \frac{\gamma}{\gamma_{max}}\right)^{(2*r_5)}\right) * \left(\overline{r_6} * \frac{\gamma}{\gamma_{max}}\right) * \overline{r_7} \quad (11)$$

where r_5 is a random number between 0 and 1 generated by uniform distribution; $\overline{r_6}$ and $\overline{r_7}$ refer as a random number generated by normal distribution; γ and γ_{max} represent the current fitness evaluation number and maximum fitness evaluation number, respectively.

The velocity of flow V towards the neighbour with the lowest fitness and is related to the slope as shown:

$$V = r_8 * S_0 \quad (12)$$

where r_8 is a random number between 0 and 1 generated by uniform distribution; S_0 represents the slope vector between the i-th flow and its j-th neighthour, where

$$S_0 = \frac{f\left(X_i^{flow}\right) - f\left(X_j^{flow}\right)}{\left\|X_{i,d}^{flow} - X_{j,d}^{flow}\right\|} \quad (13)$$

where $f\left(X_i^{flow}\right)$ and $f\left(X_j^{flow}\right)$ represent the fitness value of i-th flow and its j-th neighbour with position vectors of X_i^{flow} and X_j^{flow}, respectively. The new position X_i^{newf} of each i-th flow is calculated as:

$$X_i^{newf} = X_i^{flow} + V * \frac{X_i^{flow} - X_j^{flow}}{\left\|X_i^{flow} - X_j^{flow}\right\|} \quad (14)$$

When a randomly selected r-th neighbour has better fitness the current i-th flow, it will follow the r-th neighbour flow. Otherwise, the i-th flow will move in the dominant slope direction as shown below:

$$X_i^{newf} = \begin{cases} X_i^{flow} + \overline{r_9}\left(X_r^{flow} - X_i^{flow}\right), \text{if } f\left(X_r^{flow}\right) < f\left(X_i^{flow}\right) \\ X_i^{flow} + r_{10}\left(X^{best} - X_i^{flow}\right), \quad \text{Otherwise} \end{cases} \quad (15)$$

The pseudocode of MCFDA to solve feature selection is explained in Fig. 1. The algorithm is repeated until the stopping condition of $\gamma > \gamma_{max}$ is attained, where γ and γ_{max} represents current fitness evaluation number and maximum fitness evaluation number, respectively.

4. Performance Evaluation on MCFDA

4.1. *Simulation Settings*

The performance of MCAVOA to solve feature selection problem is evaluated with 12 datasets taken from UCI Machine Learning Repository[15], i.e., (a). Breast Cancer Wisconsin (Diagnostic), (b) Dermatology, (c) Statlog (Heart), (d) Ionosphere, (e) Iris, (f) Arrhythmia, (g) Echocardiogram, (h) Haberman's Survival, (i) Diabetes, (j) MONK 1 Problem, (k) Zoo and (l) Letter Recognition. The performance of MCFDA is compared with four feature selection algorithms developed using Bezier Search Differential Evolution Algorithm (BeSD)[16], Coronavirus Herd Immunity Optimizer Algorithm (CHIO)[17], Flow Direction Algorithm (FDA)[18] and Improved Flame Generation Mechanism

(ODSFMFO)[19]. The performances of all algorithms are measured with mean accuracy Acc^{mean} and average number of selected features $N^{feature}$. Same population size of $I = 20$ and maximum fitness evaluation of $\tau^{max} = 2000$ are set for all algorithms. Each algorithm is simulated for 30 independent runs to solve datasets.

Algorithm: MCFDA-based Feature Selection Algorithm
Inputs: $i, j, \gamma, \gamma_{max}, X^{LB}, X^{UB}, k$

```
Define fitness functions and input variables;
for each i-th flow do
    for each d-th dimension do
    while k < k_max do
        if σ_k < 0.25 then
            Select Circle map using Eq. (3);
        else if 0.25 ≤ σ_k < 0.5 then
            Select Gauss map using Eq. (4);
        else if 0.50 ≤ σ_k < 0.75 then
            Select Singer map using Eq. (5);
        else if σ_k ≥ 0.75 then
            Select Sinusoidal map using Eq. (6);
        end if
        Update σ_{k+1} and k ← k + 1;
    end while
    Initialize X_{i,d}^{flow} based on Eq. (7);
    end for
end for
Evaluate fitness of flows and select the best flow
with lowest fitness value;
while γ < γ_max do
    for each i-th flow do
        Update Q based on Eq. (11);
        for each j-th neighbour do
            Update Δ and X_j^{flow} with Eqs. (9) and (10);
            Evaluate f(X_j^{flow}) using Eq. (2);
            γ ← γ + 1;
        end for
        Sort all X_j^{flow} from best to worst fitness;
        if f(X_j^{flow}) < f(X_i^{flow}) then
            Calculate W_i using Eq. (13);
            Update each velocity flow V with Eq. (12);
            Limit velocity for V within [v^{low}, v^{up}];
            Calculate X_i^{newf} using Eq. (14);
        else
            Generate random integer number of r;
            Calculate X_i^{newf} using Eq. (15);
        end if
        Evaluate f(X_i^{newf}) using Eq. (2);
        γ ← γ + 1;
        Update current flow and best flow.
    end for
end while
```

Output: X_{best} (i.e., optimal feature subset)

Fig. 1. Feature selection with MCFDA.

4.2. *Comparisons between competing algorithms*

It is desirable for a compared algorithm to produce higher Acc^{mean} and lower $N^{feature}$ when solving feature selection problem. Boldface style value in Table 1 and Table 2 is categorized as the best result, while the second-best result is presented in underlined and italic fronts. Table 1 shows that MCFDA has the highest accuracy by producing seven best Acc^{mean}, followed by BeSD and CHIO with four best Acc^{mean}, ODSFMFO with three best Acc^{mean} and FDA with one best Acc^{mean}. Table 2f shows that the proposed MCFDA solve 6 out of 12 datasets with best $N^{feature}$, followed by FDA with three best $N^{feature}$, BeSD and CHIO with two best $N^{feature}$ and ODSFMFO without any best $N^{feature}$.

Table 1. Comparison of Acc_{mean}.

Dataset	BeSD	CHIO	FDA	ODSF MFO	MCFDA
(a)	*0.972*	0.968	0.930	0.955	**0.973**
(b)	0.981	**1.000**	0.996	*0.997*	**1.000**
(c)	0.838	0.854	*0.910*	0.810	**0.912**
(d)	0.901	**0.951**	*0.944*	0.900	0.925
(e)	**1.000**	**1.000**	0.857	**1.000**	*0.967*
(f)	0.687	*0.736*	0.673	0.605	**0.772**
(g)	**1.000**	**1.000**	*0.979*	**1.000**	**1.000**
(h)	**0.836**	0.828	*0.830*	0.787	0.705
(i)	0.771	*0.781*	0.742	0.765	**0.789**
(j)	**0.933**	0.906	*0.913*	*0.913*	0.875
(k)	0.888	0.955	**0.995**	*0.987*	0.850
(l)	0.946	*0.951*	0.950	**0.958**	**0.958**

Table 2. Comparison of $N_{feature}$

Dataset	BeSD	CHIO	FDA	ODSF MFO	MCFDA
(a)	15.93	9.17	*5.93*	18.33	**5.57**
(b)	16.57	14.53	**12.20**	22.90	*13.07*
(c)	5.67	**5.10**	*5.63*	8.10	5.77
(d)	15.90	13.80	*10.27*	16.87	**9.30**
(e)	2.40	2.03	*2.00*	3.73	**1.00**
(f)	134.73	**112.17**	131.47	155.87	*126.07*
(g)	4.37	2.60	**1.00**	3.63	*2.00*
(h)	*1.87*	6.13	5.33	2.30	**1.00**
(i)	**3.47**	*4.07*	4.97	6.20	4.53
(j)	3.67	*3.10*	**3.00**	5.33	**3.00**
(k)	7.63	6.27	*5.00*	9.97	**4.20**
(l)	**7.87**	*10.57*	11.60	15.33	11.83

5. Conclusion

In this paper, a modified algorithm known as MCFDA is introduced to solve feature selection problems more competitively. The initial population of MCFDA with better solution quality is produced by using the ergodicity and non-repetition characteristics of multiple chaotic maps are used in the initialization to improve the initial position. This is followed by the search processes with balanced exploration and exploitation searches as inspired by the rainfall behavior. A total of 12 datasets are selected to evaluate the performance of MCFDA and its peer algorithms to solve feature selection problems. Based on the simulation results, it is concluded that MCFDA can surpass other competing feature selection algorithms with better classification accuracy and lesser numbers of selected features when solving majority of selected datasets.

Acknowledgements

This research is supported by Ministry of Higher Education Malaysia (MOHE) under the Fundamental Research Grant Scheme with the project codes of FRGS/1/2019/TK04/UCSI/02/1 and FRGS/1/2020/ TK0/UCSI/02/4. This research is also supported by UCSI Research Excellence & Innovation Grant (REIG) with project code of REIG-FETBE-2022/038.

References

1. Visalakshi, S. & Radha, V. A literature review of feature selection techniques and applications: Review of feature selection in data mining. *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 1-6 (2014).
2. Alrifaey, M., Lim, W. H. & Ang, C. K. A Novel Deep Learning Framework Based RNN-SAE for Fault Detection of Electrical Gas Generator. *IEEE Access* **9**, 21433-21442 (2021).
3. Alrifaey, M. *et al.* Hybrid Deep Learning Model for Fault Detection and Classification of Grid-Connected Photovoltaic System. *IEEE Access* **10**, 13852-13869 (2022).
4. Jdid, B., Hassan, K., Dayoub, I., Lim, W. H. & Mokayef, M. Machine Learning Based Automatic Modulation Recognition for Wireless Communications: A Comprehensive Survey. *IEEE Access* **9**, 57851-57873 (2021).
5. Jdid, B., Lim, W. H., Dayoub, I., Hassan, K. & Juhari, M. R. B. M. Robust Automatic Modulation Recognition Through Joint Contribution of Hand-Crafted and Contextual Features. *IEEE Access* **9**, 104530-104546 (2021).
6. Ahmad, M. F., Isa, N. A. M., Lim, W. H. & Ang, K. M. Differential evolution: A recent review based on state-of-the-art works. *Alexandria Engineering Journal* **61**, 3831-3872 (2022).
7. Hassan, C., Durai, V., Sapuan, S., A.A, N. & Mohamed Yusoff, M. Z. Mechanical and Crash Performance of Unidirectional Oil Palm Empty Fruit Bunch Fibre-reinforced Polypropylene Composite. *Bioresources* **13**, 8310-8328 (2018).
8. Tee, Z. Y., Yeap, S. P., Hassan, C. S. & Kiew, P. L. Nano and non-nano fillers in enhancing mechanical properties of epoxy resins: a brief review. *Polymer-Plastics Technology and Materials* **61**, 709-725 (2022).
9. Jamaludin, F. A. *et al.* Considering the effects of a RTV coating to improve electrical insulation against lightning. *2016 33rd International Conference on Lightning Protection (ICLP)*, 1-5 (2016).
10. Shaari, M. *et al.* Supervised evolutionary programming based technique for multi-DG installation in distribution system. *IAES International Journal of Artificial Intelligence (IJ-AI)* **9**, 11 (2020).
11. Berghout, T., Benbouzid, M., Muyeen, S. M., Bentrcia, T. & Mouss, L. H. Auto-NAHL: A Neural Network Approach for Condition-Based Maintenance of Complex Industrial Systems. *IEEE Access* **9**, 152829-152840 (2021).
12. Ahmad, M. F., Isa, N. A. M., Lim, W. H. & Ang, K. M. Differential evolution with modified initialization scheme using chaotic oppositional based learning strategy. *Alexandria Engineering Journal* **61**, 11835-11858 (2022).
13. Chow, V. T. Handbook of applied hydrology. *New York: Mc Graw Hilld* (1959).
14. Huang, P.-C. Analysis of Hydrograph Shape Affected by Flow-Direction Assumptions in Rainfall-Runoff Models. *Water* **12** (2020).
15. Dua, D. & Graff, C. *{UCI} Machine Learning Repository*, 2019).
16. Civicioglu, P. & Besdok, E. Bezier Search Differential Evolution Algorithm for numerical function optimization: A comparative study with CRMLSP, MVO, WA, SHADE and LSHADE. *Expert Systems with Applications* **165**, 113875 (2021).
17. Al-Betar, M. A., Alyasseri, Z. A. A., Awadallah, M. A. & Abu Doush, I. Coronavirus herd immunity optimizer (CHIO). *Neural Computing and Applications* **33**, 5011-5042 (2021).
18. Karami, H., Anaraki, M. V., Farzin, S. & Mirjalili, S. Flow Direction Algorithm (FDA): A Novel

Optimization Approach for Solving Optimization Problems. *Computers & Industrial Engineering* **156**, 107224 (2021).
19. Li, Z., Zeng, J., Chen, Y., Ma, G. & Liu, G. Death mechanism-based moth–flame optimization with improved flame generation mechanism for global optimization tasks. *Expert Systems with Applications* **183**, 115436 (2021).

Authors Introduction

Mr. Wy – Liang Cheng

He received the B.Eng. degree in Mechatronic Engineering with Hnours from UCSI University, Malaysia in 2020. He is currently pursuing Master of Philosophy in Engineering in UCSI University, Malaysia. His research interests are swarm intelligence and feature selection.

Ms. Li Pan

She received her Master of Engineering in Computer Technology degree from Huazhong University of Science and Technology, China in 2008. She is currently a Doctoral research student in UCSI University, Malaysia.

Prof. Dr. Mohd Rizon Mohamad Juhari

He is a Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Engineering from Oita University, Japen in 2002. His research interests are face analysis, pattern recognition and vision for mobile robot.

Dr. Abhishek Sharma

He is a Research Assistant Professor at Graphic Era Deemed to be University in India. He received his PhD from University of Petroluem & Energy Studies in 2022. His research interests are artificial intellgence and power electronics.

Dr. Hameedur Rahman

He is an Associate Professor in Faculty of Computing and AI at AIR University in Pakistan. He received his PhD in Computer Science from Universiti Kebangsaan Malaysia in 2018. His research interests are Virual/Augmented Rality, Image Processing, Data Mining, Artificial Intelligence, Natural Language Processing and CyberSecurity.

Dr. Chun Kit Ang

He is the Dean and Associate Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Mechanical and Manufacturing Engineering from Universiti Putra Malaysia in 2014. His research interests are artificial intelligence, soft computing, robotics and mechatronics.

Dr. Sew Sun Tiang

She is an Assistant Professor in Faculty of Engineering at UCSI University in Malaysia. She received her PhD in Electrical and Electronic Engineering from Universiti Sains Malaysia in 2014. Her research interests are optimization and antenna design.

Dr. Wei Hong Lim

He is an Associate Professor in Faculty of Engineering at UCSI University in Malaysia. He received his PhD in Computational Intelligence from Universiti Sains Malaysia in 2014. His research interests are optimization and artificial intelligence.