IOP Conf. Series: Materials Science and Engineering **864** (2020) 012066 doi:10.1088/1757-899X/864/1/012066

Feature Selection of High Dimensional Data Using Hybrid FSA-IG

Nur Fatin Liyana Mohd Rosely, Azlan Mohd Zain, Yusliza Yusoff

School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Johor, Malaysia.

E-mail: yanarosely@gmail.com

Abstract. Feature selection (FS) is a process of selecting a subset of relevant features depends on the specific target variables especially when dealing with high dimensional dataset. The aim of this paper is to investigate the performance comparison of different feature selection techniques on high dimensional datasets. The techniques used are filter, wrapper and hybrid. Information gain (IG) represents the filter, Fish Swarm Algorithm (FSA) represents metaheuristics wrapper and Hybrid FSA-IG represents the hybrid technique. Five datasets with different number of features are used in these techniques. The dataset used are breast cancer, lung cancer, ovarian cancer, mixed-lineage leukaemia (MLL) and small round blue cell tumors (SRBCT). The result shown Hybrid FSA-IG managed to select least feature that represent significant feature for every dataset with improved performance of accuracy from 4.868% to 33.402% and 1.706% to 25.154% compared to IG and FSA respectively.

1. Introduction

The main objective of feature selection (FS) is to discover the significant features from a minimal subset of feature from certain problem domain with a suitably high accuracy to represent the original data is the main goal of feature selection. Basically, there are four situations FS are implemented; (a) Simplify data so the data is easier to be used and interpreted, (b) Shorten time consuming is by selecting only important feature to process, (c) Avoid the curse of dimensionality, (d) Enhance generalization by reducing overfitting which generally refers to reduction of variance [1,2,3].

FS offered three main methods; filter, wrapper and embedded. Each method has different working mechanisms depend on requirement to tackle faced issues. As an example, filter is highly used to weight every feature and sort the features based on needed. Wrapper is very popular when the faced issues demand searching and learning algorithm during the selection process for the best feature. Meanwhile, embedded more concern on cutting time consuming [3].

Based on the Figure 1, let consider Set M is the data containing a number of features that need to undergo FS. Wrapper is good choice to obtain significant feature compared to filter because the features interact among others during searching process because learning algorithm is adapted in wrapper. As result wrapper takes longer completion time than filter. However, filter scalable to high dimensional problem. Meanwhile, embedded applied to a specific learning process that bind together with a classifier; classification is an example. Embedded promising least computation, but the development of embedded is the most complex among the three. Nowadays, a new FS method, hybrid that combine the advantages of filter and wrapper attract interest among researchers [3,4,5,6].

Therefore, this paper aim to introduce a hybrid FS method named as Hybrid Fish Swarm Algorithm-Info Gain (Hybrid FSA-IG).

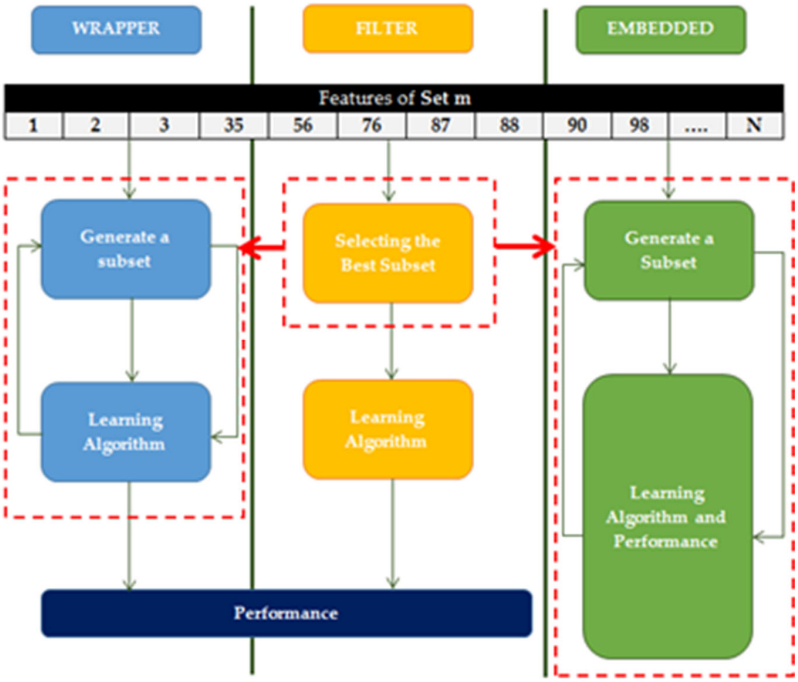

Figure 1. Working Mechanisms for Each Method.

2. Feature Selection Methods

This section will briefly describe the FS methods; IG as the Filter, FSA as the wrapper and Hybrid FSA-IG as the hybrid. Each of them will be described in Section 2.1, 2.2 and 2.3 respectively.

2.1. Info Gain (IG)

The material used for the epoxidation are fatty acid palm oils as main solvent, acetic acid and hydrogen peroxide (30%) for peracetic formation , sulphuric acid as catalyst, and hydrogen bromide and crystal violet droplet as indicator for titration. For hydrolysis reaction to produce DHPA, distilled water and alumina as catalyst is used. IG is a filter FS method that did not use any search strategy for selecting feature subsets, but adds the features progressively based on their position raking [7]. In IG, ranking is defined as weight and be similar to conditional distribution P(C | F) where C is the class label and F is feature vector. Hence, IG usually used as a replacement for the conditional distribution [8]. Equation (1) and (2) are the key equation for IG.

$$IG\ (S_X, X_i) = H(S_X) - \frac{\left|S_{X_i=v}\right|}{\sum_{v=value\ (X_i)}^{S_X} H\left(S_{X_i=v}\right)} \quad (1)$$

$$H(S) = \ -p + (S)\log_2 p + (S) - \ p - (S)\log_2 p - (S) \quad (2)$$

Where

p±(S) = probability of a training example in the dataset to be of the +ve/-ve class

2.2. Fish Swarm Algorithm (FSA)

Generally, FSA is a natural schooling of fish inspired by Dr. Li in 2002 and consider as metahueristics of swarm intelligence (SI) technique [3]. The behaviour of fish handles information about surrounding and detect a high concentration of food level by control of tail and fin are translated into two main components of FSA,

Parameters and Functions respectively [3,9,10]. Parameters include the size of the movement of fish (Step), the visual distance of fish individual (Visual), the crowd factor of the fish (δ) and distance between two fish. Search, swarm and follow are the basic functions of FSA. Compared to others SI technique FSA work more flexible and easily adapted to various types of environment [3,6,9].

FSA parameters and behaviours work together in the FS environment by obeying four compulsory steps of adapting metaheuristics technique as wrapper; Initialization, Assessment of parameters, Augmentation of parameters and Output the optimal feature subset [6]. Figure 2 illustrates the implementation of FSA in FS wrapper development Following are the FSA step in FS with equation applied.

Step 1 - Initialization the value of feature to each fish by assigning feature as X.

$$X = (X_1, X_2, X_3, \dots, X_n) \quad (3)$$

$$X_v = (X_{v1}, X_{v2}, X_{v3}, \dots, X_{vn}) \quad (4)$$

$$Visual = \frac{\sum_{i=1}^{N} \sum_{k=1}^{N} Distance\ (X_i, FX_k)}{\sum Feature} \quad (5)$$

Step 2 - Assessment of the parameter.

$$X_{centre}(i) = \begin{cases} 0, \sum_{k=1}^{k} X_k(i) < \frac{k}{2} \\ 1, \sum_{k=1}^{k} X_k\ (i) \geq \frac{k}{2} \end{cases} \quad (6)$$

$$Crowd\ degree\ (X_i) = \frac{Neighbour\ of\ X_i}{Total\ number\ of\ feature} \quad (7)$$

Step 3 - Augmentation steps of fish swarm to identify feature position.

$$X_{next} = X + \frac{X_v - X}{\|X_v - X\|} \times Step \times Rand() \quad (8)$$

Step 4 - Output the optimal feature subset by undergoing FSA functions; Search, Swarm and Follow respectively.

$$X_i = X_i + Visual \times Rand() \quad (9)$$

$$X_i^{t+1} = X_i^t + \frac{X_j - X_i^t}{\|X_c - X_c^t\|} \times Step \times Rand() \quad (10)$$

$$X_i^{t+1} = X_i^t + \frac{X_j - X_i^t}{\|X_j - X_i^t\|} \times Step \times Rand\ () \quad (11)$$

2.3. Hybrid FSA-IG

Hybrid FSA-IG combined strength of filter and wrapper to enhance FS process performance. Hybrid FSA-IG allowed learning algorithm works together with filter weight mechanisms to find most significant feature with the least number of obtained features.

As mentioned previous section, to implement metaheuristics in FS, FSA must apply four main steps of adaptation so IG is applied in one of these steps. Figure 2 illustrates the implementation of IG in Hybrid FSA-IG development.Based on the Figure 2, IG is implemented in Hybrid FSA-IG at two parts; Step 1 and 2 (Initialization the parameter and assessment of the parameter). Visual in Step 1 was modified with IG (refer Equation (12)) to provide weight information for each feature and helped searching process more effective by shortening FS process time. Implementation of IG to Equation (7) in the assessment of the parameter helped learning algorithm and feature evaluation to select significant feature with least number easier as deduction of unimportant feature will reflect more doubtless result with high accuracy.

$$\frac{\sum IG \cdot Distance\ (X_i, X_k)}{\sum X} \quad (12)$$

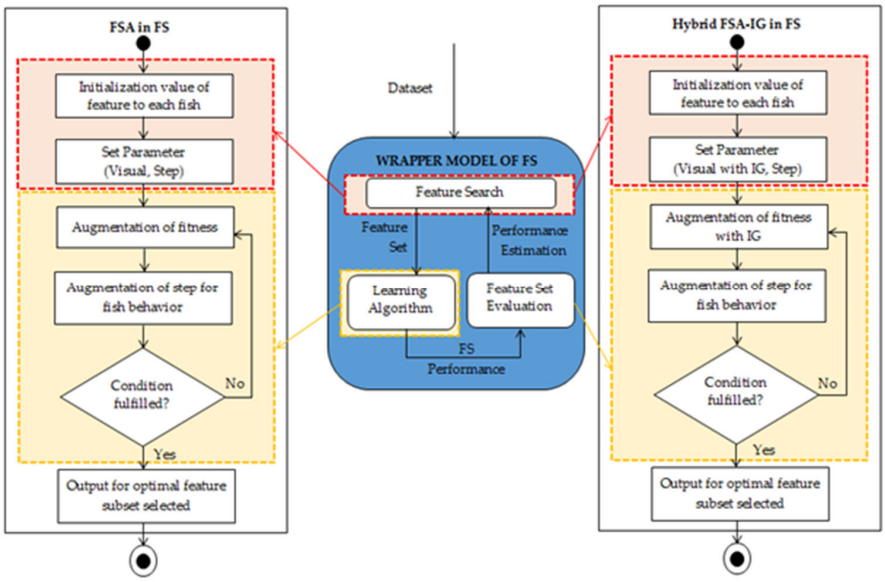

Figure 2. FSA and Hybrid FSA-IG Mapping in Wrapper Model.

3. Development and Result

3.1. Experimental Setting

Development of IG, FSA and Hybrid FSA-IG are built in MATLAB R2018b. Five different data sets (refer Table 1) are used to test all three FS methods.

Table 1. Dataset Description

Dataset	No. of Features	No. of Samples	No. of Classes
Breast	24,481	97	2
Lung	12,533	181	2
Ovarian	15,154	253	2
MLL	12,582	72	3
SRBCT	2308	83	4

3.2. Experimental Result

Experimental results of FS are divided into the total number of selected feature after FS process and accuracy of FS performance. Results of IG, FSA and Hybrid FSA-IG are presented in both result parts and are compared to each other. Table 2 shows the obtained result.

Based on Table 2, number of selected features of IG are the total number of features that have a weight value between 0 and not equal to 0. Meanwhile, FSA is based on best value of subset feature. Finally, Hybrid FSA-IG is based on weight level among the best subset feature. Hybrid FSA-IG record the least number of features with the highest accuracy performance compared to IG and FSA. The highest accuracy is Ovarian with 10 total number of selected features and 97.358% accuracy.

Table 2. FS Method Result Summary.

Dataset	No. of Feature	No. of selected features Temperature, (°C)			FS Performance accuracy			FSA-IG Performance Different	
		IG	FSA	FSA -IG	IG	FSA	FSA-IG	IG	FSA
Breast	24 481	819	138	48	54.639	62.887	88.041	33.402	25.154
Lung	12 533	9561	548	133	80.788	93.103	95.567	14.779	2.264
Ovarian	15 154	6236	35	10	92.490	95.652	97.358	4.868	1.706
MLL	12 582	5235	149	49	85.833	84.722	93.222	7.387	8.500
SRBCT	2 308	669	111	55	78.795	84.337	92.795	14.00	8.458

4. Conclusions

Section 1, FS definition and methods offered are defined. From the Section 2, the brief description of FS method used in the experiment are described. Through this, the readers able to notify the implementation and development of IG, FSA and Hybrid FSA-IG. Then the obtained results are recorded in Section 3. Among the three, Hybrid FSA-IG prove able to select the least number of features from the best feature subset with the highest performance accuracy. The result is also proven hybridization of filter and wrapper show better performance than individual FS methods.

References

[1] Dash R 2018 *Journal of King Saud University-Computer and Information Sciences.*
[2] Qi C, Zhou Z, Sun Y, Song H, Hu L and Wang Q 2017 *Neurocomputing* 220 pp 181-190.
[3] Rosely N F L M, Salleh R and Zain A M 2019 *Journal of Physics:Conference Series* **1** 012068.
[4] Jain L, Jain V K and Jain R 2018 *Applied Soft Computing* 62 pp 203-215.
[5] Apolloni J, Leguizamón G and Alba E 2016 *Applied Soft Computing* 38 pp 922-932.
[6] Diao R and Shen Q 2015 *Artificial Intelligence Review* 44(3) pp 311-340.
[7] Cilia N D, De Stefano C, Fontanella F, Raimondo S and Scotto di Freca A 2019 *Information* 10(3) p 5.
[8] Luan X Y, Li Z P and Liu T Z 2016 *Neurocomputing.* 174 pp 522-529.
[9] Kusairi R M, Moorthy K, Haron H, Mohamad M S, Napis S and Kasim S 2017 *Int. J. on Adv. Sc. Eng. and Inf. Tech.* 7 (4-2) pp 1595-1600.
[10] X Zhu Z, Ni Z, Cheng M, Jin F, Li J and Weckman G. 2018 *Aplied Intelligence* 48(7) pp 1757-1775.