Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Filter-Based Feature Selection Using Information Theory and Binary Cuckoo Optimisation Algorithm

Ali Muhammad Usman

School of Computer Sciences, University Sains Malaysia 11800 Pulau Pinang, Malaysia; Department of Computer Sciences, Federal College of Education (Technical) Gombe, Nigeria. E-mail: aliakko@yahoo.com

Umi Kalsom Yusof*

*Corresponding author, Assistant Professor, School of Computer Sciences, University Sains Malaysia 11800 Pulau Pinang, Malaysia. E-mail: umiyusof@usm.my

Maziani Sabudin

School of Computer Sciences, University Sains Malaysia 11800 Pulau Pinang, Malaysia. E-mail: maziani@usm.my

Abstract

Dimensionality reduction is among the data mining process that is used to reduce the noise and complexity of features in various datasets. Feature selection (FS) is one of the most commonly used dimensionalities that reduces the unwanted features from the datasets. FS can be either wrapper or filter. Wrappers select subsets of the feature with better classification performance but are computationally expensive. On the other hand, filters are computationally fast but lack feature interaction among selected subsets of features which in turn affect the classification performance of the chosen subsets of features. This study proposes two concepts of information theory mutual information (MI). As well as entropy (E). Both were used together with binary cuckoo optimization algorithm BCOA (BCOA-MI and BCOA-EI). The target is to improve classification performance (reduce the error rate and computational complexity) on eight datasets with varying degrees of complexity. A support vector machine classifier was used to measure and computes the error rates of each of the datasets for both BCOA-MI and BCOA-E. The analysis of the results showed that BCOA-E selects a fewer number of features and performed better in terms of error rate. In contrast, BCOA-MI is computationally faster but chooses a larger number of features. Comparison with other methods found in the literature shows that the proposed BCOA-MI and BCOA-E performed better in terms of accuracy, the number of selected features, and execution time in most of the datasets.

Keywords: Feature Selection; Filter-Based; Binary Cuckoo Optimization; Information Theory.

DOI: https://orcid.org/10.22059/jitm.2022.84900 Manuscript Type: Research Paper University of Tehran, Faculty of Management

Received: June 15, 2021 Accepted: July 23, 2021

Introduction

In the various fields of human endeavour such as e-commerce, education, health care, computing, online education, bioinformatics, and social media, among others, data have now become profuse. Thus, the exponential growth of the stored data has become a substantial problem for successful data management in many areas. As such, data mining together with machine learning approaches must be implemented to uncover secret information from these vast data pools (Jain, Sawhney, & Mathur, 2018; Liu & Wang, 2019). Classification is amongst the methods of data mining that is used to classify each instance of data into a set of groups. Feature space is the only problem that is downgrading a classifier's efficiency. Except there is an earlier knowledge of the best features, it is otherwise difficult to find the most useful and appropriate features, especially when the size of the feature is large (Russell, & Norvig, 2002). Based on that, the term feature selection (FS) is, therefore, introduced to select the most vital and appropriate features from these enormous volumes of the stored data. FS has two major problems (1. How to search for the best subsets and (2. How to evaluate the best subsets of featured being generated (Usman, Yusof, & Naim, 2020); (Li et al., 2017). To assess or evaluate the best subsets of the generated features depends on the method of the FS. FS can be either filter or wrapper.

To determine the accuracy or error rate of the chosen subset of features, the wrapper method of the FS uses a classification algorithm and selects the subsets with better accuracy. However, these processes are highly computationally expensive, particularly on high-dimensional datasets (Moslehi, & Haeri, 2020); (Samy, Hosny, & Zaied, 2020); (Ma & Gao, 2020). Filter methods, alternatively, are computationally fast and can scale speedily to large dimensional datasets. A lack of feature dependence or relationship between the selected features is one of its significant downsides (Jain, Sawhney, & Mathur, 2018); (Usman, Yusof, & Naim, 2020). Therefore, this study will address the issue of feature dependency among selected subsets of features.

Most of the existing algorithms cannot appropriately determine the enormous space of an FS without being deprived of being stuck in some local optima (Usman et al., 2018); (Fahad et al., 2020); (Hancer, et al., 2018); (Xue, et al., 2015); (Goswami et l., 2019). Evolutionary algorithms (EAs) are now being used as search methods to elucidate FS problems; nevertheless, several of them still grieve from early convergence. Cuckoo Optimization Algorithm (COA) introduced by Rajabioun, (2011), is one of the EAs mentioned in (Tavana et al., 2018); (Jain, Sawhney, & Mathur, 2018); (Li et al., 2017) that have qualified search operators and can contribute to the search space realization of the most promising area and converge more rapidly than many other EAs.

Information theory is a practical approach that can be able to measure the relevance within two or more features together with their class label in feature ranking. The most frequently used ones are information measures such as Information Gain (IG) (Muharram &

Smith, 2005), Information Gain Ratio (IGR) (Otero, et al., 2003); Correlation (Hart, et al., 2017) (Hall et al., 2009), mutual information (MI), and entropy (Dash & Liu, 1997) among others.

Researchers are now using the concepts of both MI and entropy to find the significance and redundancy of the selected features by combining them with various EAs. For instance, Cervante et al., (2012) used both entropy as well as MI as a fitness evaluation measure in Binary Particle Swarm Optimisation (BPSO). In the work of (Mlakar, Fister, & Brest, 2017), MI is being used along with PSO. Besides, Particle Swarm Optimisation (PSO) was used to enhance crowding features and clustering to obtain the best subset of features. Lately, Hancer, Xue, & Zhang, (2018) used differential evolution (DE) for feature ranking with the help of MI, Relief-F, and Fisher scores. The results obtained surpass both the single and multi-objective approaches presented. Recently, Ma & Gao (2020), used the concepts of a filter-based multiple feature construction method (FCM) using genetic programming GP (FCM) and a filter-based feature selection (FS) using GP (FS), to maintain the classification performance with a smaller number of features. Methods such as hill climbing are used along with GA in (Goswami et l., 2019) because of their property of easy scalability. All these previous works testified that the concept of information theory is successful in addressing the problems of FS. To address feature construction and feature selection problems, efficient global search algorithms are needed (Xue et al, 2015).

Thus, in this paper, the enhanced version of the COA, precisely the Binary COA (BCOA) developed by (Mahmoudi, Rajabioun, & Lotfi, 2013). that is suitable for handling FS is proposed as a search technique together with MI and information gained based entropy as the filter evaluation measures.

The remainder of the paper is standardized as follows: Section 2 describes BCOA, MI, as well as entropy. Section 3 is the proposed filter-based BCOA (BCOAMI and BCOA-E) along with the experimentation. Section 4 describes the results and discussion. Lastly, in Section 5, the conclusions were offered as well as further research directions.

Litrecher review

This section describes all the ingredients that are used to carry out this study. It includes the BCOA, MI, and gain ratio-based entropy together with the detailed formulas that are used in each case.

Binary Cuckoo Optimisation Algorithm

Binary Cuckoo Optimisation Algorithm (BCOA) was proposed in (Mahmoudi, Rajabioun, & Lotfi, 2013). since the original COA is meant to solve only a continuous optimization problem. The BCOA is the most suitable for solving FS problems than its COA counterpart. To calculate the X_G and X_{CP} of the habitat in the COA (Rajabioun, 2011), we use:

$$Y_{NH} = X_{CP} + rand(X_G - X_{CP}) \quad (1)$$

To create a new habitat X_{NH} suitable for discrete binary problems, a sigmoid function (*Sig*) in Eq.2 was used. The reason is to map X_{NH} into the range [0,1]. Then Eq.3 will alter the values in the habitat as 0 or 1. Whereby rand in Eq.3 is a random number, that is generated randomly.

$$Sig = \frac{1}{1+e^{-X_{NH}}} \quad (2)$$

$$If (Sig > rand\ Then\ X_{NH} = 1\ Else\ X_{NH} = 0 \quad (3)$$

Information Gain Based Entropy

The information gain-based entropy is calculated based on Eq.4. The higher values of the entropy signify the same probability of occurrence of each variable in contrast to the low entropy that means the different possibility of event of an incident for each variable.

$$H(X) = -\sum_i Px_i log_2 Px_i \quad (4)$$

X is the random variable and $P(x_i) = \Pr\{X = P(x_i), x_i \in X\}$ is the mass probability density of X.

Mutual Information

Mutual information (MI) is the measure of the relationship or dependence between two arbitrary variables by providing a means to assess or evaluate the relevance of the subset of the features. The MI between two features X and Y is defined as (Hancer, Xue, & Zhang, 2018); (Tavana et al., 2018):

$$I(X;Y) = -\sum_{i,j} P(x_i,y_j) log_2 P\left(\frac{P(x_i,y_j)}{P(x_i)\,P(y_j)}\right) \quad (5)$$

Eq.5 shows that the I (X; Y) will be large if the two features X and Y are so much related. Else, I(X; Y) = 0 if X and Y are not related at all.

Some Related Works

The concert of K nearest neighbour (KNN) and SVM based on current filters is presented by (Freeman, Kulić, & Basir 2015). The results have shown that MI can develop a better subset of functionality for SVM and KNN. Also, MI is capable of evolving useful subsets of functionality for the two classification algorithms. The idea of maximum relevance and minimum redundancy within the MI was presented by (Nogueira, Sechidis, & Brown, 2017). The objective was to find the subset of functionality with reduced redundancy and to improve the relevance with the class label. Based on that, researchers now use it to obtain the relationship or dependency between two pairs of features. But, due to the use of sequential search, it can quickly get trap in the local optima. Estevez et al. use a genetic algorithm (GA) in Est´evez, et al., 2009) to remedy the constraint of sequential search. Besides, a normalized FS-based MI (NMIFS) was proposed because MI favored characteristics with higher values. The NMIFS is an improvement of the MIFS, MIFS-U, and mRMR methods offered in

(Battiti, 1994). However, it is also limited to only one pair of features, and yet a non-optimal set of features are likely to be chosen.

This motivates many researchers to use other optimization algorithms, that can search for the best optimal subset of features with the best classification performance. For example, Cervante et. al., (2012) used a binary PSO together with entropy and MI as evaluation criteria. The results obtained on the datasets showed that BPSO with mutual information could evolve a set of features with a fewer number of features. Whereas BPSO with entropy has more classification accuracy using a DT compared to BPSO with MI. Moreover, Moghadasian & Hosseini (2014) used MI and entropy are used as evaluation criteria on some six high dimensional datasets. An artificial neural network was used to measure the classification accuracy and cuckoo search as the search technique. The experimental results displayed that around 90% of the main features were minimized and yet achieved better classification accuracy than using full-length features. In the work of Mlakar, Fister, & Brest, (2017), the concept of MI is being used along with PSO. Besides, the PSO is to enhances crowding features and clustering to obtain the best subset of features. Recently, Huda et al., (2019) use a group-based PSO by updating the Pbest along with the Gbest to get the relevant features while ignoring the redundant features. Moslehi & Haeri (2020) proposed a hybrid filter-wrapper FS by combining GA and PSO along with Artificial Neural Network on five different datasets.

Li et al., (2017) presented a survey paper on the optimization algorithm that has been used for FS. Out of the numerous algorithms, they conclude that there is still a chance to use other algorithms that are not fully explored in the FS domain. Recently, Usman et al., (2018) presented a comparative analysis among some nature-inspired algorithms for feature selection on some medical datasets. The results obtained showed that the binary flower pollination algorithm performed better than the standard flower pollination algorithm in terms of both the number of selected features and classification accuracy. Moreover, the proposed BPFA performed better than harmony search and particle swarm optimization that uses rough set and quick reduct, respectively. Recently Usman, Yusof, & Naim, (2018), use the concepts of BCOA for filter-based FS but its limited gain ratio-based entropy.

Other optimization algorithms are now becoming popular in dealing with FS problems. For example, Mafarja et al., (2017) hybridized Whale Optimisation Algorithm (WOA) together with Simulated Annealing (SA) to solve FS problems. The datasets used coincided with the datasets used in this study. Hence is used for comparison even though it is a wrapper-based approach. Similarly, Samy, Hosny, & Zaied (2020), introduced a new binary WOA for FS based on whales' behavior. The Optimum-Path Forest technique is used as an objective function. The results obtained were tested on five color image datasets. It's found that the process is much faster than the other classification techniques.

In another perspective, Arora & Anand (2019) presented two binary variants of the Butterfly Optimization Algorithm (BOA). Among these, two transfer functions are used to map the continuous search space to a discrete one. Twenty-one datasets are used in the experiments. The superior performance of the proposed binary variants is proved in the experiments. Moreover, Huda, & Banka, (2020), offer an enhanced binary version of the Gravitational Search Algorithm (GSA) is presented, which is based on the law of gravity and attraction of masses to address this problem of feature selection in medical data. The speed of a random forest classifier is combined with the optimization behavior of the GSA. A substantial improvement was recorded in terms of prediction accuracy. Furthermore, Hancer, et al., (2018), presented a new binary Grasshopper Optimisation Algorithm for FS. Whereby, the binarisation of continuous space transforms the continuous values of the continuous space into binary values 0 or 1 in the binary space was realized. Lately, Tahir et al., (2020) presented a novel Binary Chaotic GA for FS in healthcare. To conclude, Fahad et al, (2020) introduced an asymmetric uncertainty-based Ant Colony Optimisation Algorithm for streaming FS in high dimensional medical datasets.

The review of the related works of De Rezende, et al, (2014) shows that optimization algorithms are becoming more relevant in dealing with different kinds of FS problems. They are used explicitly as search techniques, to search for the most relevant subsets of features. On the other hand, the concepts of information theory play a vital role as a filter evaluation measure, specifically in the filter-based approach.

Guha et al., (2020) proposed a score-based filter FS approach known as Mutually Informed Correlation Coefficient (MICC) by combining two popular statistical dependence measures namely MI and Pearson Correlation Coefficient. The evaluated MICC on different variations of Local Binary Pattern-based feature vectors used for classifying the components of handwritten document images as text or non-text similar to the proposed work of (Peng, Long, & Ding, 2005). Moreover, Samuel et al., (2020) proposed a modified entropy MI feature selection to forecast medium-term load using a deep learning model in smart homes and a promising result was realized. In the same vein, the work in Usman et al., (2020) used the concepts of MI along with entropy together with Non-dominated Sorting GA III to purposely addresses the issues of multi-objective filter-based FS. But the work is limited to multi-objective FS, whereas, there is limited work in the single objective filter-based FS.

Rahman et al., (2020) introduced multiclass EEG signal classification utilizing Rényi min-entropy-based feature selection from wavelet packet transformation. The proposed method was tested on some EEG datasets and a better result was achieved.

A novel method for feature selection using the incorporation of copula-based multivariate dependency in mutual information was proposed in (Lall et al., 2021), which assists to remove the need to average out over multiple instances of bivariate dependencies. The method is unbiased against noisy datasets due to the scale-invariant property of the

copula. Hence, it can be applied to datasets, in which the ratio between sample size to class size is large enough, even though original marginal distributions are unknown. The method also satisfies the maximum relevance and minimum redundancy criteria of feature selection.

In another perspective, Lim, & Kim (2020) propose a method for generating the initial population of an EA-based multi-label feature selection method considering dependencies between features and labels. Whereas Sun et al., (2020) introduced a multilabel FS using the concepts of ML-ReliefF and neighborhood MI for multilabel neighborhood decision systems are proposed. Besides, Gonzalez-Lopez, Ventura & Cano. (2020) proposes a distributed model to compute a score that measures the quality of each feature for multiple labels on Apache Spark. They propose two different approaches that study how to aggregate the MI of multiple labels: Euclidean Norm Maximization and Geometric Mean Maximization. The former selects the features with the largest L2-norm whereas the latter selects the features with the largest geometric mean. Still, Shi et al., (2020) propose a multi-label FS method using MI and improved multilabel ReliefF (ML-ReliefF). Each label is calculated in label space and combined with the MI of features and labels to construct a novel correlation degree between features and label sets to preprocess multilabel datasets, which is used to reduce the runtime of ML-ReliefF. Then, the MI of label sets is introduced into improving the accuracy of the correlation degree among label sets. Furthermore, two types of correlation degrees for label sets based on ML-ReliefF are developed to divide similar and heterogeneous samples more clearly. Then, a divided method of heterogeneous neighbors is presented to effectively avoid the repeated calculation in ML-ReliefF, and a novel method of feature weighting based on ML-ReliefF is constructed to evaluate the importance of features. Finally, a multilabel FS algorithm based on MI and ML-ReliefF for multilabel classification is designed to improve the performance of multilabel classification. Promising results were realized in the experimental datasets used therein.

On another dimension, (Hart, et al., 2017) proposed a framework that first constructed multiple features using GP and then selected effective feature subsets for classification using GA. In the feature construction stage, the l best individuals are stored into the hall of fame in every generation, and randomly select m individuals to seed the population in the next generation.

Single-stage feature construction and FS method were proposed by Tran, Xue, & Zhang (2016), which used leaf nodes of a GP tree to select effective original features and used the GP tree to construct a new higher-level feature, i.e., performing feature construction and feature selection using GP at the same time. In the same vein, Tran, Zhang & Xue (2016) proposed another single-stage feature construction and feature selection method that constructed multiple features based on multiple tree representation and also chose terminal nodes as selected features. Later, Hall et al., (2009) proposed another feature construction and FS method that first constructed a predefined number of features by running the GP algorithm

multiple times, and then used GA to remove redundant features. The number of constructed features is the same as that of the GP runs.

The above-related works clearly show that EAs have gain popularity particularly in solving different FS problems with various information measures as filter evaluation. However, the use of other EAs specifically, BCOA to solve filter-based FS problems is not fully explored in the literature.

Methodology

In this section, the two filter evaluation measures are being used together with BCOA to form BCOA-MI and BCOA-E. The detail is explained below

BCOA Based MI for FS

The MI is used to measure the relationship between two pairs of features along with their target class. As such, it is used to measure the relevance and redundancy between two couple of features during the feature interaction between them. Based on that, BCOAMI is proposed containing both the relevance and redundancy as the fitness evaluation measure that guides the BCOA to hunt for the subset of features. It is indicated in the Eq.6:

$$F_{mi} = -\beta(Rel_{mi} + Red_{mi}) - Red_{mi} \tag{6}$$

$$Rel_{mi}(X;C) = max \sum_{i} I(x;c) \; and \; Red_{mi}(X;Y) = \min \frac{1}{|m|} \sum_{i,j} I(x_i; y_j)$$

C and X represent the target class and the discrete binary feature subsets, respectively. The Rel_{mi} uses a pairwise method to calculate the MI between every feature and its target class, which ultimately determines the relevancy of the chosen feature subsets to the target class. Red_{mi} evaluates the MI shared by each pair of the selected features, which means that there is redundancy inside the selected features. Thus, Eq. 6 F_{mi} is s

A maximization function because it maximizes the relevancy Rel_{mi} and simultaneously minimizes the Red_{mi} of the selected features.

BCOA Based Information Gain Entropy for FS

Unlike the F_{mi} that is considered as two-way relevance and redundancy, in FS, Feature interaction may happen in more than two ways; we may have a group of feature interactions. Therefore, BCOA-E is proposed to consider a group of features during feature interaction. Hence, the fitness function is clearly defined, as shown in Eq.7.

$$F_E = -\beta(Rel_E + Red_{mi}) - Red_E \tag{7}$$

$$Rel_E(X;C) = \max IG \sum_{i} I(x;c) \; and \; Red_E(X;Y) = \min \frac{1}{|m|} \sum_{i,j} IG(x\{X/x\})$$

Also, *RelE* evaluates the information gain of c given the information on the features in X, and this indicates the relevancy between the selected subset of features as well as the target class. On the other hand, *RedE* assesses the combined entropy of all the given features in X, and this shows that there is redundancy inside the chosen subsets of features. Therefore, Eq.7 *FE* is also considered as a maximization function that maximizes relevancy *RelE* and concurrently minimizes the redundancy *RedE* among the selected subset of features.

Algorithm 1 Proposed BCOA-MI and BCOA-E

```
Start
   Initialise each habitat with some features from a dataset
   Collect the features in their respective habitats
   Explain ELR for every single cuckoo
   Allow the cuckoos to lay their eggs in their matching ELR
   Destroy those cuckoos familiar by the multitude birds
   Allow egg to incubate and baby chicken raise
   Estimate the environment of every recently grownup cuckoo
   Limits cuckoos' highest number in location and abolish those that exist in poorer environments
   Group cuckoos and discover finest cluster and choose goal line environment
   Allow the new cuckoo populace to settle at the goal line environment
   Return the optimum solution (selected features)
   Evaluate the fitness function according to Equation 6 in BCOA-MI and Equation 7 in BCOA-E
   If the stop condition is satisfied Then stop, else go to 3
   End if
Stop
```

Relevance and Redundancy Weighted Values in BCOA-MI and BCOA-E

It can be discerned that both Eq. 6 and Eq. 7 have a β1 and β2 respectively. The essence of the β values is to see which one can significantly improve the relevance and consequently reduced redundancy. Based on that, we sum up the relevance and redundancy, then multiplied it with the values and deduct them from the outcome. The reason is that; relevance is needed the most than the redundancy for the optimal result as reported by (Hancer et al., 2018). The weighted values used by (Cervante et al., 2012) are adopted in this study.

Experimental Design

Table 1 depicts the datasets used in this study, and they can be found in (Frank & Asuncion, 2010). From the table, eight datasets are used in the experiments with the Sonar dataset having the highest number of features, while the Connect-4 dataset is having the highest number of instances. On the other hand, the Lympography dataset is having the least number of both features and instances from the entire dataset. The initial and maximum population of the BCOA are set to twenty and thirty; for the thirty different runs. SVM was used to measure the classification accuracy. The datasets are divided into a training set (70%) and a testing set (30%). Besides, ten-fold cross-validation was used on each of the eight datasets.

Table 1. Experimental Detests.

S/N	Detests	Features	Instances
1	Lymphography	18	148
2	SpectEW	22	267
3	KrvskpEW	36	3196
4	WaveformEW	40	5000
5	Dermatology	34	366
6	Connect-4	42	44473
7	Ionosphere	34	351
8	Sonar	60	208

Findings

Tables 2, 3, 4, and 5 show the results of the proposed methods. Firstly, BCOA-MI and BCOA-E results are displayed in Table 2. From the Table "Ave Size", "Ave Acc", "Best Acc", "Time" and "All" represent the average number of selected features, ave age accuracy, best accuracy, time, and all features, respectively.

Table 2. Experimental Results of the proposed (BCOA-MI) and (BCOA-E).

Detests	Approach	Ave-Size	Ave-Acc (Best Acc)	Time
Lymphography	All	18	0.875	
	BCOA-MI	**3**	0.840 (0.850)	1.68
	BCOA-E	4.8	**0.855 (0.859)**	52.08
SpectEW	All	22	0.851	
	BCOA-MI	**4**	0.881 (0.884)	1.85
	BCOA-E	4.2	0.888 (0.904)	54.21
KrvskpEW	All	36	0.892	
	BCOA-MI	**4.2**	0.920 (0.945)	56.11
	BCOA-E	13.9	**0.980 (0.984)**	1649.60
WaveformEW	All	40	0.771	
	BCOA-MI	**17.5**	0.660 (0.660)	172.62
	BCOA-E	20.2	**0.760 (0.760)**	5100.90
Dermatology	All	35	**0.892**	
	BCOA-MI	9.2	**0.922 (0.955)**	45.11
	BCOA-E	14.5	**0.982 (0.994)**	1234.50
Connect-4	All	42	**0.781**	
	BCOA-MI	19.5	**0.666 (0.756)**	182.55
	BCOA-E	21.2	**0.776 (0.770)**	4100.22
Ionosphere	All	34	**0.992**	
	BCOA-MI	4.2	**0.966 (0.977)**	65.11
	BCOA-E	11.5	**0.992 (0.995)**	1014.10
Sonar	All	60	**0.881**	
	BCOA-MI	5.5	**0.896 (0.926)**	242.11
	BCOA-E	12.2	**0.996 (0.996)**	4899.01

Results of BCOA-MI and BCOA-E

Table 2 shows the results of BCOA-MI along with BCOA-E without any weight function. It can be observed from the results that BCOA-MI performed much better on the average size

features selected in all the datasets where around 75% of the total features are reduced. In contrast to the BCOA-E which performed much better in terms of accuracy. Similarly, less computational time was recorded in the BCOA-MI, and this is due to the pair number of features it deals with compared to BCOA-E that used a group of features. The results clearly showed that both BCOA-MI and BCOA-E could significantly minimize the feature size and attain an improved or similar performance to using the full features.

Results of BCOA-MI and BCOA-E with βWeighted Values

From Table 3, it can be seen that the higher the β1 value in BOCA-MI, the better the accuracy in the entire datasets. Therefore, the relevance is more significant than the redundancy, which consequently leads to higher accuracy on the higher values of the β1. But looking at the WaveformEW dataset in the table when β1 = 0.9 and 0.8 the difference between the best values is not much they are 0.778 and 0.779 respectively. Moreover, the feature size got reduced to around 70%. On the other hand, the higher the β2 value in BCOAE depicted in Table 4, the higher the number of the selected feature. The number of features reduced by almost 40% compared to the full-length features. Also, the accuracy increases as the β2 increases in the majority of the datasets. Comparison between BCOA-MI with β1 in Table 3 along with BCOA-E with β2 in Table 4, one can notice that: (i. β1 is worse than β2 in terms of accuracy (ii. β2 is worse than β1 in terms of the number of selected features and (iii. β1 is computationally less expensive compared to β2. Employing both β1 and β2 values within the filter evaluation measures could significantly reduce the number of features and obtained appropriate classification accuracy than using the full-length features. The "Std" in both Table 3 and Table 4 represent the standard deviation in all the thirty different runs.

Table 3. Results of the BCOA-MI with different weights of β1.

Detests	β1	Ave-Size	Ave-Acc (Best Acc)	Std	Time
Lymphography	0.9	7.8	0.860(0.888)	0.013	1.69
	0.8	5.2	0.840(0.850)	0.013	1.69
	0.7	4.9	0.834(0.834)	0.000	1.69
	0.6	4.1	0.800(0.800)	0.000	1.68
	0.5	3	0.780(0.799)	0.001	1.68
SpectEW	0.9	9.2	0.888(0.894)	0.012	1.87
	0.8	7.8	0.871(0.885)	0.012	1.87
	0.7	5.6	0.844(0.855)	0.011	1.86
	0.6	4.2	0.833(0.840)	0.011	1.86
	0.5	4	0.830(0.830)	0.000	1.85
KrvskpEW	0.9	17.2	0.942(0.946)	0.001	59.55
	0.8	16.7	0.935(0.940)	0.002	57.45
	0.7	15.2	0.930(0.937)	0.002	57.11

	0.6	14.2	0.924(0.925)	0.001	56.13
	0.5	12.2	0.920(0.923)	0.001	56.11
WaveformEW	0.9	21.4	0.775 (0.778)	0.001	179.9
	0.8	20.2	0.770 (0.779)	0.004	175.7
	0.7	19.2	0.760 (0.774)	0.003	174.0
	0.6	18.4	0.688 (0.727)	0.000	172.6
	0.5	16.5	0.660 (0.660)	0.000	172.6
Dermatology	0.9	15.2	0.942(0.966)	0.001	59.55
	0.8	14.7	0.955(0.960)	0.001	57.45
	0.7	14.2	0.950(0.957)	0.001	57.11
	0.6	14.2	0.944(0.945)	0.001	56.13
	0.5	10.2	0.940(0.933)	0.001	56.11
Connect-4	0.9	21.4	0.775 (0.778)	0.001	179.9
	0.8	20.2	0.770 (0.779)	0.004	175.7
	0.7	19.2	0.760 (0.774)	0.003	174.0
	0.6	18.4	0.688 (0.727)	0.000	172.6
	0.5	17.5	0.660 (0.660)	0.000	172.6
Ionosphere	0.9	4.2	0.992(0.966)	0.002	65.10
	0.8	4.7	0.985(0.960)	0.002	64.45
	0.7	5.3	0.980(0.957)	0.003	63.21
	0.6	9.5	0.984(0.945)	0.003	60.19
	0.5	9.9	0.980(0.933)	0.004	59.22
Sonar	0.9	5.1	0.876 (0.888)	0.001	241.0
	0.8	6.2	0.880 (0.889)	0.004	231.7
	0.7	9.2	0.860 (0.884)	0.003	198.3
	0.6	9.4	0.788 (0.857)	0.000	192.6
	0.5	10.5	0.868 (0.869)	0.000	192.6

Table 4. Results of the BCOA-E with different weights of β2.

Detests	β2	Ave-Size	Ave-Acc (Best Acc)	Std	Time
Lymphography	0.9	12.6	0.890 (0.890)	0.000	52.39
	0.8	10.5	0.880 (0.888)	0.000	52.39
	0.7	8.9	0.874 (0.879)	0.001	52.39
	0.6	6.4	0.860 (0.872)	0.001	52.08
	0.5	5.1	0.855 (0.859)	0.001	51.46
SpectEW	0.9	10.2	0.899 (0.914)	0.004	54.79
	0.8	8.7	0.891 (0.895)	0.001	54.79
	0.7	6.8	0.884 (0.889)	0.001	54.5
	0.6	5.2	0.871 (0.880)	0.002	54.5
	0.5	5	0.862 (0.869)	0.001	54.21
KrvskpEW	0.9	19.2	0.972 (0.976)	0.001	1750.8
	0.8	18.4	0.965 (0.980)	0.005	1689

	0.7	16.3	0.950 (0.977)	0.005	1679
	0.6	15.4	0.944 (0.945)	0.001	1650.2
	0.5	13.9	0.929 (0.933)	0.001	1649.6
WaveformEW	0.9	26.5	0.822 (0.888)	0.004	5315.2
	0.8	24.3	0.790 (0.819)	0.003	5190.5
	0.7	21.2	0.770 (0.785)	0.003	5140.2
	0.6	20.1	0.768 (0.769)	0.001	5100.9
	0.5	19.2	0.760 (0.760)	0.000	5100.9
Dermatology	0.9	19.2	0.972 (0.976)	0.001	1620.3
	0.8	18.4	0.965 (0.980)	0.004	1589.2
	0.7	15.3	0.955 (0.988)	0.003	1459.3
	0.6	15.4	0.940 (0.955)	0.003	1325.2
	0.5	12.9	0.929 (0.933)	0.002	1541.6
Connect-4	0.9	33.5	0.820 (0.888)	0.004	5315.2
	0.8	33.3	0.795 (0.833)	0.003	5290.2
	0.7	29.2	0.772 (0.780)	0.003	5240.1
	0.6	22.1	0.769 (0.762)	0.001	5200.2
	0.5	20.2	0.763 (0.750)	0.000	5200.3
Inosphere	0.9	16.9	0.992 (0.996)	0.011	1020.9
	0.8	16.4	0.965 (0.980)	0.010	1099.6
	0.7	16.4	0.955 (0.988)	0.011	1150.5
	0.6	12.4	0.940 (0.955)	0.009	1120.1
	0.5	11.5	0.929 (0.933)	0.008	1210.3
Sonar	0.9	20.5	0.990 (0.992)	0.004	5315.2
	0.8	19.3	0.895 (0.933)	0.003	5290.2
	0.7	19.8	0.882 (0.889)	0.003	5240.1
	0.6	12.1	0.879 (0.882)	0.001	5200.2
	0.5	12.1	0.883 (0.85a0)	0.000	5200.3

Average Fitness of BCOA-MI and BCOA-E

Table 5 shows that the proposed BCOA-E converged earlier with the least fitness value than the BCOA-MI on all eight datasets. Although, BCOA-MI recorded the highest fitness value it mostly obtained the best classification performance in terms number of selected features and computational time compared to its BCOA-E counterpart as shown earlier in Table 2, Table 3, and Table 4. Also, the values of the standard deviation in the table are within the required standard limit in all the iterations.

Table 5. Average Fitness For BCOA-MI and BCOA-E.

Datasets	BCOA-MI		BCO-E	
	Fitness	StdDev	Fitness	StdDev
Lymphography	0.158	0.001	**0.131**	0.001
SpectEW	0.179	0.001	**0.137**	0.002
KrvskpEW	0.064	0.000	**0.055**	0.002
WaveformEW	0.279	0.000	**0.274**	0.000
Dermatology	0.034	0.000	**0.141**	0.002
Connect-4	0.173	0.000	**0.161**	0.000
Ionosphere	0.014	0.000	**0.110**	0.001
Sonar	0.079	0.001	**0.139**	0.000

Convergence Trends of BCOA-MI and BCOA-E

Figure 1 shows the convergence of the proposed BCOA-MI and BCOA-E. At the top of the chart is the name of the dataset, while the fitness and number of iterations are represented on the x-axis and y-axis, respectively. From the curve on each graph, it can be observed that BCOA-E is at the bottom compared to the BCOA-MI, this means that BCOA-E converges to the best fitness compare to the BCOA-MI. Perhaps, it can be due to interaction among a group of features in the BCOA-E. On the other hand, BCOA-MI has limited feature interaction, since it interacts with only a pair of features at a time.

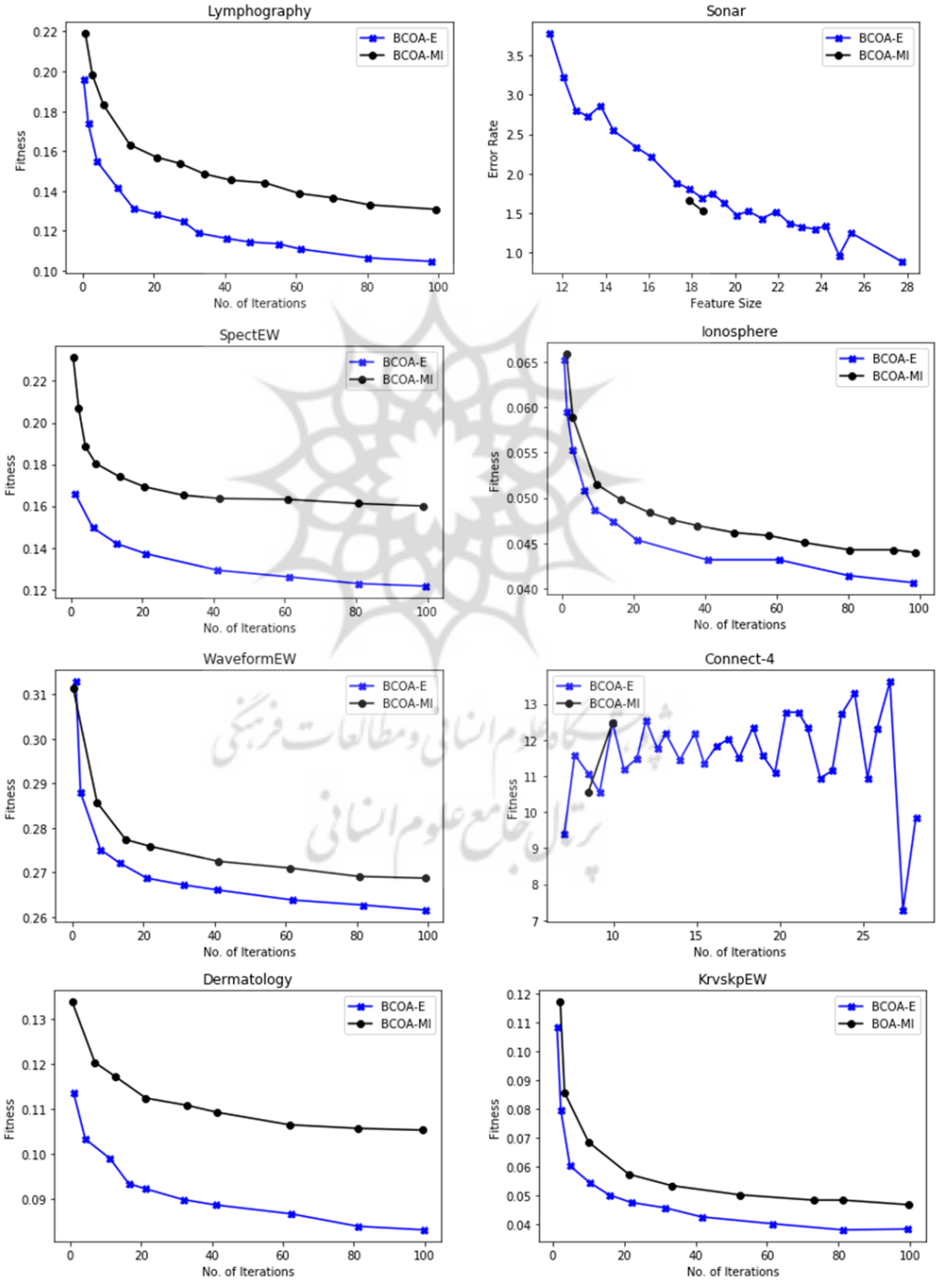

Fig. 1. Convergence Trends of BCOA-MI and BCOA-E.

Comparison with Other Existing Approaches

The results obtained are compared with the existing work, that works with similar datasets, for example, BPSO-MI and BPSO-E in (Cervante et al., 2012), WOA-SA in (Mafarja et al., 2017), and FCMFS in the work of (Ma & Gao, 2020). The detailed comparison is depicted in Table 6.

The proposed results were compared with the existing works in terms of numbers of selected features, classification accuracy, and the time it takes to finish its execution on each dataset during the thirty independent runs. In all aspects, our proposed methods performed better than the BPSOMI and BPSOE. Whereas, in terms of the computational time, our approaches performed better than WOA-SA excepts on the Lymphography dataset where WOA-SA recorded the least time. In terms of accuracy, WOA-SA achieved the best accuracy in two of the datasets, while our proposed methods achieved the best accuracy on the remaining two datasets. Comparing with the recent work of FCMFS (Ma & Gao, 2020), one can be observed that both Sonar and Ionosphere datasets obtained competitive results compare to the FCMFS. Our proposed methods recorded the least number of features in all datasets compared to the other approaches. Therefore, one can conclude that the proposed methods performed better than the existing works in terms of the number of selected features, computational time as well as classification accuracy.

Table 6. Comparison of the proposed algorithms with other existing approaches.

Detests	Approach	Ave-Size	Ave-Acc (Best Acc)	Std-Acc	Time
Lymphography	All	18	0.875		
	BCOA-MI	3	0.780 (0.799)	0.001	1.66
	BCOA-E	5.1	0.855 (0.859)	0.001	52.08
	All	18	0.755		
	BPSO-MI	3	0.711 (0.711)	0.000	3.89
	BPSO-E	6.3	0.740 (0.778)	0.017	61.45
	WOA-SA	7.2	0.890		1.66
SpectEW	All	22	0.851		
	BCOA-MI	**4**	0.830 (0.830)	0.000	1.85
	BCOA-E	4.2	0.862 (0.869)	0.001	54.21
	All	22	0.809		
	BPSO-MI	3.1	0.783 (0.794)	0.002	2.13
	BPSO-E	4.5	0.812 (828)	0.010	62.89
	WOA-SA	6	0.880		313.38
KrvskpEW	All	36	0.892		
	BCOA-MI	4.2	0.920 (0.945)	0.001	56.11
	BCOA-E	13.9	0.980 (0.984)	0.001	649.60
	All	36	0.985		
	BPSO-MI	4.7	0.797 (0.902)	0.027	76.23
	BPSO-E	15.7	0.970 (0.977)	0.011	203.67
	WOA-SA	12.8	0.980	641.0	641.01
WaveformEW	All	40	0.771		
	BCOA-MI	**17.5**	0.660 (0.660)	0.000	172.62
	BCOA-E	20.2	0.760 (0.760)	0.000	5100.90
	All	40	0.696		

	BPSO-MI	19.4	0.620 (0.649)	0.011	1497.9
	BPSO-E	20.9	0.688(0.698)	0.002	6102.76
	WOA-SA	20.6	0.770		1770.48
Dermatology	All	35	0.992		
	BCOA-MI	4.2	**0.990 (0.995)**	0.000	55.10
	BCOA-E	13.9	0.980 (0.989)	0.005	529.45
	All	36	0.985		
	BPSO-MI	4.7	0.898 (0.962)	0.022	78.21
	BPSO-E	15.7	0.965 (0.995)	0.023	213.11
	WOA-SA	12.8	**0.995**	621.0	631.01
Connect 4	All	42	0.881		
	BCOA-MI	**12.5**	0.880 (0.880)	0.000	113.10
	BCOA-E	17.2	**0.960 (0.960)**	0.000	3099.55
	All	40	0.896		
	BPSO-MI	12..9	0.820 (0.809)	0.011	1221.9
	BPSO-E	18.9	0.888(0.898)	0.002	5101.11
	WOA-SA	13.3	**0.960**		1250.22
Ionosphere	All	60	0.892		
	BCOA-MI	4.2	0.920 (0.945)	0.001	59.14
	BCOA-E	13.9	0.980 (0.984)	0.001	719.33
	All	36	0.985		
	FCMFS	4.6	0.941		
Sonar	All	36	0.771		
	BCOA-MI	**17.5**	0.660 (0.660)	0.000	172.62
	BCOA-E	20.2	0.760 (0.760)	0.000	5100.90
	FCMFS	5.5	(0.850)		

Conclusion

The aim of this paper has been achieved by developing two filter-based evaluation measures based on entropy and MI, together with BCOA. The results demonstrated that BCOA-MI is capable of evaluating the relevance and redundancy of the pair features. In comparison, BCO-E shows its priority in assessing both the relevance and redundancy when dealing with a group of features. In either case, weighted values are employed. And it is found that the higher the values, the higher the number of features and the accuracy. BCOA-MI recorded the least accuracy compared with BCOA-E. Perhaps, it might be due to the feature interaction among a group of features by the BCOA-E.

On the other hand, BCOA-E is computationally expensive compared with the BCOA-MI. BCOA-MI interacts with only pair features that make it computationally faster. Apart from using different newer optimization algorithms to solve similar problems for competitive results, in the future, we will investigate the use of the nondominated sorting mechanism together with BCOA to solve the conflicting issues in FS rather than using the weighted values. On the other hand, using other emerging evolutionary algorithms may likely provide competitive results. Similarly, working with proposed methods on some medical datasets may likely assist medical practitioners in selecting the most relevant subsets of features, thereby

reducing the cost of laboratory tests. And, the time wasted on consultation, lab, etc will be drastically reduced with the help of the proposed method.

In the future, we will extend our work to multi-label FS using the same concepts of MI and entropy. On the other end, the application of information gain-based entropy is reported to be problematic especially on datasets with a large number of features. As such working with high dimensional datasets with the concepts of gain ratio-based entropy is theorized to solve the problem much better and obtain the best subsets of features with better classification performance and computationally less expensive.

Acknowledgements

This document is the results of the research project funded by the Universiti Sains Malaysia via Research University Grant (RUI) (1001/PKOMP/8014084).

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Arora, S., & Anand, P. (2019). Binary butterfly optimization approaches for feature selection. *Expert Systems with Applications*, *116*, 147-160.

Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5(4), 537–550 (1994)

Cervante, L., Xue, B., Zhang, M., Shang, L.: Binary particle swarm optimization for feature selection: A filter-based approach. In: 2012 IEEE Congress on Evolutionary Computation (CEC). pp. 1–8. IEEE (2012)

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(1-4), 131-156.

De Rezende, L. F. M., Lopes, M. R., Rey-López, J. P., Matsudo, V. K. R., & do Carmo Luiz, O. (2014). Sedentary behavior and health outcomes: an overview of systematic reviews. *PloS one*, *9*(8), e105620.

Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on neural networks*, *20*(2), 189-201.

Fahad, L. G., Tahir, S. F., Shahzad, W., Hassan, M., Alquhayz, H., & Hassan, R. (2020). Ant Colony Optimization-Based Streaming Feature Selection: An Application to the Medical Image Diagnosis. *Scientific Programming*, *2020*.

Frank, A., & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. *School of information and computer science*, *213*, 2-2.

Freeman, C., Kulić, D., & Basir, O. (2015). An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recognition*, *48*(5), 1812-1826.

Gonzalez-Lopez, J., Ventura, S., & Cano, A. (2020). Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Systems*, *188*, 105052.

Goswami, S., Chakraborty, S., Guha, P., Tarafdar, A., & Kedia, A. (2019). Filter-Based Feature Selection Methods Using Hill Climbing Approach. In *Natural Computing for Unsupervised Learning* (pp. 213-234). Springer, Cham.

Guha, R., Ghosh, K. K., Bhowmik, S., & Sarkar, R. (2020, February). Mutually Informed Correlation Coefficient (MICC)-a New Filter Based Feature Selection Method. In *2020 IEEE Calcutta Conference (CALCON)* (pp. 54-58). IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Hancer, E., Xue, B., & Zhang, M. (2018). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, *140*, 103-119.

Hancer, E., Xue, B., Zhang, M., Karaboga, D., & Akay, B. (2018). Pareto front feature selection based on artificial bee colony optimization. *Information Sciences*, *422*, 462-479.

Hart, E., Sim, K., Gardiner, B., & Kamimura, K. (2017, July). A hybrid method for feature construction and selection to improve wind-damage prediction in the forestry sector. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 1121-1128).

Hichem, H., Elkamel, M., Rafik, M., Mesaaoud, M. T., & Ouahiba, C. (2019). A new binary grasshopper optimization algorithm for feature selection problem. *Journal of King Saud University-Computer and Information Sciences*.

Huda, R. K., & Banka, H. (2020). A group evaluation based binary PSO algorithm for feature selection in high dimensional data. *Evolutionary Intelligence*, 1-15.

Jain, R., Sawhney, R., & Mathur, P. (2018, March). Feature selection for cryotherapy and immunotherapy treatment methods based on gravitational search algorithm. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-7). IEEE.

Lall, S., Sinha, D., Ghosh, A., Sengupta, D., & Bandyopadhyay, S. (2021). Stable feature selection using copula-based mutual information. *Pattern Recognition*, *112*, 107697.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 1-45.

Lim, H., & Kim, D. W. (2020). MFC: Initialization method for multi-label feature selection based on conditional mutual information. *Neurocomputing*, *382*, 40-51.

Liu, W., & Wang, J. (2019, May). A brief survey on nature-inspired metaheuristics for feature selection in classification in this decade. In *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 424-429). IEEE.

Ma, J., & Gao, X. (2020). A filter-based feature construction and feature selection approach for classification using Genetic Programming. *Knowledge-Based Systems*, *196*, 105806.

Mafarja, M. M., & Mirjalili, S. (2017). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, *260*, 302-312.

Mahmoudi, S., Rajabioun, R., & Lotfi, S. (2013). Binary cuckoo optimization algorithm. *nature*.

Mlakar, U., Fister, I., & Brest, J. (2017, June). Hybrid Multi-objective PSO for Filter-Based Feature Selection. In *23rd International Conference on Soft Computing* (pp. 113-123). Springer, Cham.

Moghadasian, M., & Hosseini, S. P. (2014). Binary cuckoo optimization algorithm for feature selection in high-dimensional datasets. In *International conference on innovative engineering technologies (ICIET'2014)* (pp. 18-21).

Moslehi, F., & Haeri, A. (2020). A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. *Journal of Ambient Intelligence and Humanized Computing*, *11*(3), 1105-1127.

Muharram, M., & Smith, G. D. (2005). Evolutionary constructive induction. *IEEE transactions on knowledge and data engineering*, *17*(11), 1518-1528.

Nogueira, S., Sechidis, K., & Brown, G. (2017). On the stability of feature selection algorithms. *J. Mach. Learn. Res.*, *18*(1), 6345-6398.

Otero, F. E., Silva, M. M., Freitas, A. A., & Nievola, J. C. (2003, April). Genetic programming for attribute construction in data mining. In *European Conference on Genetic Programming* (pp. 384-393). Springer, Berlin, Heidelberg.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226-1238.

Rahman, M. A., Khanam, F., Ahmad, M., & Uddin, M. S. (2020). Multiclass EEG signal classification utilizing Rényi min-entropy-based feature selection from wavelet packet transformation. *Brain informatics*, *7*(1), 1-11.

Rajabioun, R. (2011). Cuckoo optimization algorithm. *Applied soft computing*, *11*(8), 5508-5518.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Samuel, O., Alzahrani, F. A., Hussen Khan, R. J. U., Farooq, H., Shafiq, M., Afzal, M. K., & Javaid, N. (2020). Towards modified entropy mutual information feature selection to forecast medium-term load using a deep learning model in smart homes. *Entropy*, *22*(1), 68.

Samy, A., Hosny, K. M., & Zaied, A. N. H. (2020). An efficient binary whale optimization algorithm with optimum path forest for feature selection. *International Journal of Computer Applications in Technology*, *63*(1-2), 41-54.

Shi, E., Sun, L., Xu, J., & Zhang, S. (2020). Multilabel Feature Selection Using Mutual Information and ML-ReliefF for Multilabel Classification. *IEEE Access*, *8*, 145381-145400.

Sun, L., Yin, T., Ding, W., Qian, Y., & Xu, J. (2020). Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems. *Information Sciences*, *537*, 401-424.

Tahir, M., Tubaishat, A., Al-Obeidat, F., Shah, B., Halim, Z., & Waqas, M. (2020). A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare. *Neural Computing and Applications*, 1-22.

Tavana, M., Shahdi-Pashaki, S., Teymourian, E., Santos-Arteaga, F. J., & Komaki, M. (2018). A discrete cuckoo optimization algorithm for consolidation in cloud computing. *Computers & Industrial Engineering*, *115*, 495-511.

Tran, B., Xue, B., & Zhang, M. (2016). Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing*, *8*(1), 3-15.

Tran, B., Zhang, M., & Xue, B. (2016, December). Multiple feature construction in classification on high-dimensional data using GP. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-8). IEEE.

Tsanas, A., Little, M. A., & McSharry, P. E. (2010). A simple filter benchmark for feature selection. *Journal of Machine Learning Research*, *1*(1-24).

Usman, A. M., Abdullah, A. U., Adamu, A., & Ahmed, M. M. (2018). Comparative Evaluation of Nature-Based Optimization Algorithms for Feature Selection on Some Medical Datasets. *i-manager's Journal on Image Processing*, *5*(4), 9.

Usman, A. M., Yusof, U. K., & Naim, S. (2018). Cuckoo inspired algorithms for feature selection in heart disease prediction. *International Journal of Advances in Intelligent Informatics*, *4*(2), 95-106.

Usman, A. M., Yusof, U. K., & Naim, S. (2020). Filter-Based Multi-Objective Feature Selection Using NSGA III and Cuckoo Optimization Algorithm. *IEEE Access*, *8*, 76333-76356.

Usman, A. M., Yusof, U. K., Naim, S., Musa, N., & Chiroma, H. (2020). Multi-objective Filter-based Feature Selection Using NSGAIII With Mutual Information and Entropy. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)* (pp. 1-7). IEEE.

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, *20*(4), 606-626.

Bibliographic information of this paper for citing:

Usman, Ali Muhammad, Yusof, Umi Kalsom & Sabudin, Maziani (2022). Filter-Based Feature Selection Using Information Theory and Binary Cuckoo Optimisation Algorithm. *Journal of Information Technology Management*, Special Issue, 203-222.