

## A Proposed Hybrid Feature Selection Method for Data Mining Tasks

Laith Abualigah<sup>1</sup>, Akram Jamal Ali Aldulaimi<sup>1</sup>, Mohammad Al Shinwan<sup>1</sup>, Mohammad Shehab<sup>2</sup>

<sup>1</sup>Faculty of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan.

lx.89@gmail.com

<sup>2</sup>Computer Science Department, Aqaba University of Technology, Aqaba, Jordan



### ABSTRACT

Feature selection technique is one of the important data pre-processing steps in data mining; it is used to find the important features subset in order to create a new subset of informative features. The model that used the informative subset such that a classification model built only with this subset would get better predictive accuracy than model that used a complete set of features. Hence, this technique can improve several data mining techniques by increasing its performance and reduce its computational time. In this study, we are going to propose a hybrid method based on Sine Cosine Algorithm (SCA) with Genetic algorithm (GA) that utilizes to select the best features in order to improve the performance of the feature selection problem. The performance of propose method will evaluate using 16 datasets from University California Irvine UCI Machine Learning repository such as (Breast cancer, Exactly, Waveform and others) and compare with original Sine Cosine Algorithm (SCA) and other related approaches in the literature such as Ant Lion optimization (ALO) and Particle Swarm Optimization (PSO).

**Key words:** Feature selection problem, Sine Cosine Algorithm, Optimization problems, Meta-heuristic optimization algorithms.

### 1. INTRODUCTION

Document Knowledge Discovery in database (KDD) is created for categorizing dataset so as to finding useful information from a big scale dataset. Discovery of knowledge contains four main types of process, namely, data warehouse, preprocessing, data mining and evaluation process [16, 15]. It is of attention to investigators in information acquisition, machine learning, pattern discovery, database, artificial intelligent, visualization of data, and statistic. Data Mining (DM) method is an important fragment in (KDD) as shown in Figure 1.

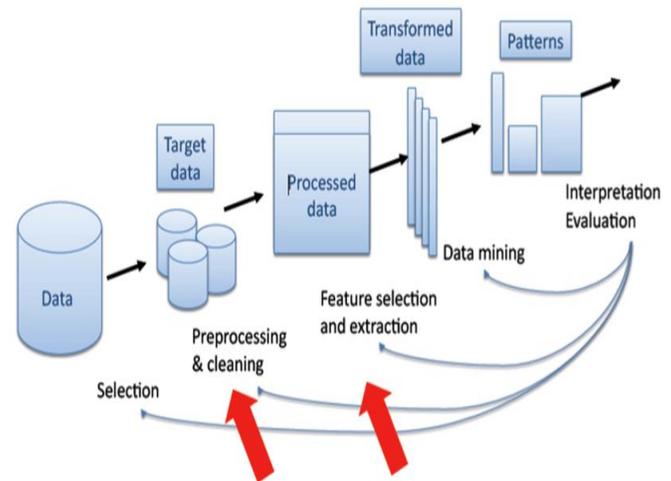


Figure 1: The main processes of KDD [16].

In DM, the worth of dataset or in what way to discovery beneficial info is a significant subject that possible affect classifications presentation and calculation typical. DM tries to collect dataset and explore it to catch beneficial knowledge, it ways contain reversion, categorizing, grouping, deviation discovery, dependence modeling, change discovery and summarization [1,2]. data mining has essential part is data pre-processing that includes reduction, transformation, normalization, discretization, integration, feature extraction, data cleaning and feature selection [4].

Feature selection problems create the learning work composite and computational expensive. In the literature some controversy about the term high-dimensionality. It happens while the datasets have the next things, very max numbers of sample or feature or both. The decrease approaches may be practical to enhance the prediction accuracy, ease of comments and speed of education. The unrelated and redundant features inside the original datasets are removed to produce a subset of features with enhanced discriminant power [31, 32].

Feature is also identified as attribute, property, dimension and characteristic. The increasing request of winning benefits from data that high dimensional kept a challenges for data mining. Feature selection consider one from the answers that's employ in order to make a predictive subset that reduces the mistakes of classifiers prediction through choice

informative features by eliminating redundant, unrelated and noise feature in original dataset [16].

Feature selection (FS) is known also as variable selection or subset selection, these methods can be help us build a remarkably easy and high accurate mode that minimize the running time of a learning (learning high-dimensional data fast) and get a better classification performance [6]. FS techniques are applied several applications such as image processing [24], signal processing [7], pattern recognition [26], data mining [7], and machine learning [26].

Approaches of feature selection usually categorized into three wide groups are: wrapping model, filtering model, and embedded model [25, 27]. Filter model approaches are rankers; the attributes are assessed conferring to procedures direct from the data without predictor and use rank gathering of the attributes. It offers no number concerning the most noteworthy features.

A filter method is good generalization capacity and low computational cost; this method can also handle with high-scale dataset. The wrapping-based technique use method of learning as classifiers (part inside fitness function) in order to assess the informative of feature subsets and subsequently finding the prediction has better performance. Embedded approach works by linear classifier such as support vector machine, which the algorithms are embedded SVM as expand functional. It is also able to imprisonment dependence at a lower computation price than other methods [16].

## 2. RELATED WORKS

Feature selection (FS) technique is one of the answers that employed to make a predictive accuracy of the search space problem [14]. Search approaches can be summarized as shadows: thorough search, heuristic search, probabilistic search, and involuntary hybrid exploration algorithms. Metaheuristic algorithms plan is much quicker, time overwhelming and only search a particular path to finding the optimal solution [16].

Metaheuristic search typically practical to real world problem, and to exact varied series of computer science [12, 29, 30]. Also, heuristic are suitable to treat other parts of huge data, such as, diversity, and speed [8]. Heuristics-based exploration plan displays two criteria are exploitation to govern the better neighbor and exploration in the search area.

Different metaheuristics methods are applied to treat feature selection and a literature can be find in [10]. Genetic algorithm (GA) is the furthestmost inspect metaheuristics. Population and single-based metaheuristic algorithms are proposed. In metaheuristic algorithm that single based such as, Hill climbing, simulated annealing have been used [17,18], scatter search random search harmony search, hill climbing has main disadvantages, is very tricky for opening solutions and it often time dropping in local optima [9].

In populations-based, every method has been used from traditional memetic algorithms, genetic programming, particle swarm optimization (PSO), to bat algorithm [19]. The subset features have been selected by using the spider monkey optimization approach. The primary population algorithms have been given for the dataset, and the assessment of the fitness calculation was done using the SVM for classification accuracy. In order to continue or stop the process a stopping criterion is tested. The best subset final of attributes with large accuracy of classification is defining as the better optimal results [21].

One of the recent metaheuristic whale optimization algorithm is metaheuristic algorithm that mimics the whale-hunting appliance [3,20]. Coral reefs optimization as hybrid binary algorithm and simulated annealing for attribute selection (BCROSAT) can discovery the maximum accuracy and selects the minimum number of features for most data sets that it used [28]. The enhanced binary genetic algorithm and feature granulation (IBGAFG) to get feature space granular and find informative subset of feature [11]. Instance selection is a method that reduce the size of the original training data, instance selection and feature extraction as the combination reduces the large volume of computation time of training the classifier [22, 23].

## 3. RESEARCH MOTIVATION

Through studying the sine cosine algorithm and its results, we note weakness in the exploration of sine cosine algorithm. Therefore, the motivations for applying feature selection is the exploration. At first, features can be charge this study to obtain, preprocessing, computational reasons, storage and transfer, and need more training samples to achieve a better performance ability of a classifier. Feature selection is significant and necessary to treat the classification problems; reduction features and removes noisily features from data. These processing are given increase to speed of knowledge, simple of rules, imagines the data, the comprehensibility of the concepts and predictive accuracy of a classification task. Summary, dimensionality can also help to get better understanding of a given problem in data mining, machine learning and pattern acknowledgement etc.

## 4. PROBLEM STATEMENT

Metaheuristic algorithms process display two inconsistent criteria are exploration of the search space and exploitation to determining the best solution. In the native sine cosine algorithm, we note weakness in its exploration strategy that will lead to get weakness in its performance during the search space. However, enhancements or modify by hybridization technique can be made to introduce a new version of metaheuristic algorithms to enhance the

performance by balancing between exploration and exploitation of the search space. This stimulus underlies utmost our tries to create a model of predictive based on hybridization approaches for solving feature selection problems, through reducing the number of features, weak relevant and irrelevant features. Practically, a subset that optimal is likely to contain only strongly relevant features. This study compares the enhancement in classification accuracy with current methods that mimics the inbred behavior in the nature (such as particle swarm optimization and ant lion optimization) and discusses the improvement that can be obtained by applying a new metaheuristic algorithm.

There are many qualitative questions of researcher that can be pointed out as follows:

- A. Does a native sine cosine algorithm adapt to solve the feature selection problem?
- B. Does the hybridization between population-based methods can act to improve the exploration ability to solve FS problems?
- C. How to evaluate the ability of hybrid feature selection approach for solving feature selection problem.

### 5. RESEARCH OBJECTIVES

The objectives of FS are reducing dimensionality and eliminating noise from the data. These are given increase to speed of learning, ease of rules, visualizes the data and predictive accuracy. The main aim of the study is to assess the capability of hybridization between metaheuristic methods to create a new FS approach in order to solve feature selection problem through rising the performance of search space.

That is doing by the following several contributions are planned in this thesis:

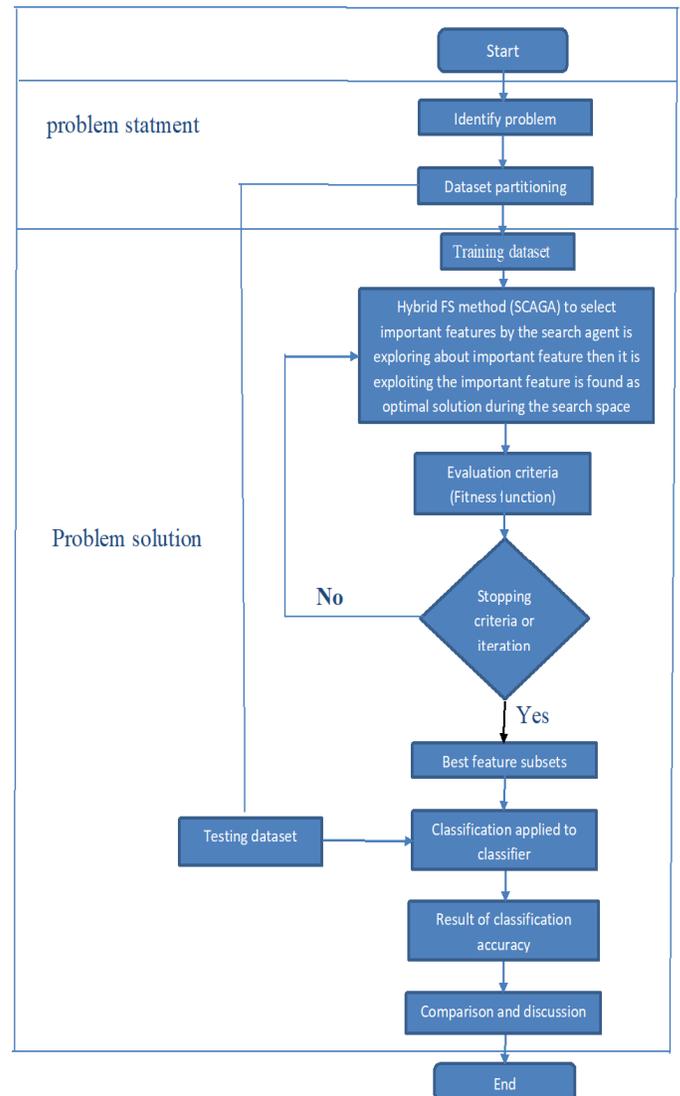
- A. In this study, a native sine cosine algorithm is adapted to solve the feature selection problem.
- B. To enhance the exploration strategy of sine cosine algorithm (SCA), a new feature selection method is suggested based on hybridization strategy between the Sine cosine algorithm with genetic algorithm (GA) called SCAGA.
- C. To evaluate the performance of the proposed approach, we will use; Classification accuracy, Statistical best, Statistical worst and Statistical mean fitness, Standard deviation (Std) and average number of features.

### 6. SIGNIFICANT OF STUDY

Significant of this study is to solve feature selection problem by a new hybrid feature selection approach (SCAGA) in which it is eliminating redundant, irrelevant, and noisy features from original dataset. These are given increase to speed of learning, ease of rules, visualizes the data and predictive accuracy of a classification task.

### 7. RESEARCH METHODOLOGY

The main steps of the propose feature selection approach is shown in Figure 2 using binary version of Sine Cosine algorithm (SCA) and Genetic algorithm (GA) as a hybrid approach. To apply feature selection task suitably, we do following steps:



**Figure 2:** Feature selection techniques for classification

- A. Dataset partitioning into (dataset for training and dataset for testing).
- B. The hybrid feature selection approach generate candidate feature subsets during select important

features by eliminating irrelevant, redundant and noise features from original dataset.

C. The informative feature subsets is evaluated by using the evaluation criteria (classification accuracy, statistical mean, statistical best and statistical worst fitness, average selection size standard deviation) so we get the best feature subsets then apply the classifier in which the result will be maximal accuracy of classification and minimal number of features.

## 8. DATASETS

The experimental results are assessed on 16 dataset with two high dimension from California University Irvine (UCI) machine learning repository <https://archives.ics.uci.edu/ml/datasets.html> [5]. All applied dataset details are represented in following Table:

**Table 1:** The datasets

Datasets	Features number	instances number	subject of datasets
Breast_EW	30	699	life
Breastcancer	9	699	life
Congress	16	535	social
Exactly	13	1000	n/a
Heart_Ew	13	270	life
Exactly-2	13	1000	n/a
Ionosphere	34	351	physical
lymphography	18	148	life
Lymphography	18	148	life
Sonar_EW	60	208	physical
m_of_n	13	1000	rules
tic_tac_toe	9	958	game
Spect_Ew	22	267	life
Waveform_EW	40	5000	physical
Zoo	16	101	life
Wine_EW	13	178	physical

These datasets are used by many researchers such as [13] in this domain and we shall compare our results with original Sine Cosine algorithm (SCA) and other related approaches published in the literature survey such as ant lion optimization and particle swarm optimization.

## 9. EVALUATION CRITERIA

To assess the performance of the suggested method with the other approaches such as ant lion optimization and particle

swarm optimization, we used the evaluation criteria as following, classification accuracy, statistical worst, statistical mean and statistical best fitness, average number of features and standard deviation.

The general evaluation criteria are explained as following:

- A. Classification accuracy: is a metric for evaluating Classification model. Maximal accuracy of classification is one of the contradictory goals for this study [13].
- B. Average number of features: is a measure of selected features by the suggested FS approach. Minimal number of features is also one of the contradictory goals for this study [13].
- C. Fitness function: work of fitness function for the SCAGA is increasing performance of classification over the dataset of validation given the dataset training, while it keeps a minimum number of features selected. So, the objective is to decrease the fitness value of equation,
 
$$f\Theta = \omega * E + (1-\omega) \sum_i \Theta_i / n$$
 Where fitness function  $f\theta$  give vector  $\theta$  sized  $n$  with 0/1 elements representing unselected/selected feature,  $n$  is numbers of features in the dataset.  $E$  is the error rate of classifier and  $\omega$  is a constant for controlling the accuracy of classification performance to the features number that selected [13].
- D. Standard deviation: is a measure of spread such as range and variance of data. Therefore, it gives information about wide or concentrated of dataset [13].

## REFERENCES

1. Abdullah, S., Shaker, K., & Shaker, H. (2011). Investigating a round robin strategy over multi algorithms in optimizing the quality of university course timetables. *International Journal of Physical Sciences*, 6(6), 1452-1462.
2. Abualigah, L. M., Khader, A. T., Al-Betar, M. A., & Alomari, O. A. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Systems with Applications*, 84, 24-36.
3. Aljarah & T. (2018, november 26). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. Springer.
4. Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for

- the text clustering. *The Journal of Supercomputing*, 73(11), 4773-4795.  
<https://doi.org/10.1007/s11227-017-2046-2>
5. Bache, K., & Lichman, M. (2015). UCI Machine Learning Repository Online.
  6. Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. *International Journal of Computer Science, Engineering and Applications*, 5(1), 19.
  7. Choi, S. I., Oh, J., Choi, C. H., & Kim, C. (2012). Input variable selection for feature extraction in classification problems. *Signal Processing*, 92(3), 636-648.
  8. Dhaenens, C., & Jourdan, L. (2016). *Metaheuristics for big data*. John Wiley & Sons.
  9. Diao, R., & Shen, Q. (2012). Feature selection with harmony search. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 42(6), 1509-1523.
  10. Diao, R., & Shen, Q. (2015). Nature inspired feature selection meta-heuristics. *Artificial Intelligence Review*, 44(3), 311-340.  
<https://doi.org/10.1007/s10462-015-9428-8>
  11. Dong, L., T., & Sun (2018, January 7). A novel hybrid genetic algorithm with granular information for feature selection and optimization. Elsevier, pp. 33-46.
  12. Du, K. L., & Swamy, M. N. S. (2016). *Search and optimization by metaheuristics* (p. 434). New York City: Springer.
  13. Hafez, M. Zawbaa & Emary, H. (2016). Sine Cosine Optimization Algorithm for Feature Selection. *IEEE*.
  14. Ibrahim, E. O., A. & Lu. (2018, September 11). Improved salp swarm algorithm based on particle swarm optimization for feature selection. Springer.
  15. Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
  16. Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, Springer Science & Business Media, (Vol. 454).
  17. Abualigah, L. M., Khader, A. T., Hanandeh, E. S., & Gandomi, A. H. (2017). A novel hybridization strategy for krill herd algorithm applied to clustering techniques. *Applied Soft Computing*, 60, 423-435.
  18. Abualigah, L. M. Q. (2019). *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*. Springer.
  19. Nakamura, R. Y., Pereira, L. A., Costa, K. A., Rodrigues, D., Papa, J. P., & Yang, X. S. (2012, August). BBA: a binary bat algorithm for feature selection. In *Graphics, Patterns and Images (SIBGRAPI)*, 2012 25th SIBGRAPI Conference on (pp. 291-297). IEEE.
  20. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science*, 25, 456-466.  
<https://doi.org/10.1016/j.jocs.2017.07.018>
  21. Rania, D. (2018). *Microarray Cancer Gene Feature Selection Using Spider Monkey Optimization Algorithm and Cancer Classification using SVM*. Elsevier, pp. 108-116.
  22. Abualigah, L. M., Khader, A. T., Al-Betar, M. A., Alyasseri, Z. A. A., Alomari, O. A., & Hanandeh, E. S. (2017, May). Feature selection with  $\beta$ -hill climbing search for text clustering application. In *2017 Palestinian International Conference on Information and Communication Technology (PICICT)* (pp. 22-27). IEEE.
  23. Suganthi & V. Karunakaran. (2018, January 30). Instance selection and feature extraction using cuttle fish optimization algorithm and principal component analysis using decision tree. Springer.
  24. Tsai, J. S., Huang, W. B., Kuo, Y. H., & Horng, M. F. (2012). Joint robustness and security enhancement for feature-based image watermarking using invariant feature regions. *Signal Processing*, 92(6), 1431-1445.
  25. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence*, 48(11), 4047-4071.
  26. Wang, Y., Dahnoun, N., & Achim, A. (2012). A novel system for robust lane detection and tracking. *Signal Processing*, 92(2), 319-334.
  27. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A combination of objective functions and hybrid Krill herd algorithm for text document clustering analysis. *Engineering Applications of Artificial Intelligence*, 73, 111-125.
  28. Yan, M. & Luo, W. (2018). A Hybrid Algorithm Based on Binary Chemical Reaction Optimization and Tabu Search for Feature Selection of High-Dimensional Biomedical Data. *Tsinghua Science and Technology*. 23(6), pp.733-743.
  29. Al-Sai, Z. A., & Abualigah, L. M. (2017, May). Big data and e-government: a review. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 580-587). IEEE.
  30. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A hybrid strategy for krill herd algorithm with harmony search algorithm to improve the data clustering. *Intelligent Decision Technologies*, (Preprint), 1-12.
  31. Zhang, Z., & Ning, Y. (2010). Effective semi-supervised nonlinear dimensionality reduction for wood defects recognition. *Computer Science and Information Systems*, 7(1), 127-138.
  32. Zhao, Z. A., & Liu, H. (2011). *Spectral feature selection for data mining*. CRC Press.