Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 4734-4741

Research Article

An Optimized Approach for Feature Selection and Clustering using Grasshopper Optimization Algorithm

Vivek Parganiha[1], Soorya Prakash Shukla[2], Lokesh Kumar Sharma[3]

[1]Ph.D. Scholar at Department of Computer Science & Engineering,
Bhilai Institute of Technology, Durg, 491001, Chhattisgarh State, India
vivekparganiha@gmail.com
[2]Professor at Department of Electrical Engineering,
Bhilai Institute of Technology, Durg, 491001 Chhattisgarh State, India
[3]Scientist-E at ICMR-National Institute of Occupational Health,
Department of Health Research, Ministry of Health and Family Welfare, Government of India,
New Delhi, India

Abstract

Clustering contains various significant applications on machine learning, image segmentation, data mining, pattern recognition. Hence, a clustering proper selection is more important in feature selection. In this manuscript, Feature Selection (FS), Clustering using Grasshopper Optimization Algorithm (GOA) is proposed and implemented in two dataset such as NSL-KDD and UNSW-NB15 for providing effective feature selection. These 2 datasets have many features, in which only a limited count of features contributes to clustering. Noise and unwanted data will be generated as the dimension of the space covering all the features will be huge as well as unclean, thus reducing accuracy of clustering. The effective feature selection technique will remove sound, redundant, redundant data, so the Grasshopper Optimization Algorithm with the feature selection method is proposed. The simulation process is executed in the MATLAB platform. In NSL-KDD data set, the proposed method attains high accuracy 25.6% and 20.45%, low mean square error (MSE) 70.55% and 71.33%, low Entropy 63.55% and 60.33%, low processing time 40.55% and 40.22% shows better performance when comparing with the existing method such as Feature Selection and Clustering Using hybrid GWO–GOA and Feature Selection and Clustering Using Quantum Whale Optimization Algorithm (QWOA). Finally, the proposed technique provides best clustering accuracy with low computational time.

Key Words: Feature Selection and Clustering, Grasshopper Optimization Algorithm, NSL-KDD and UNSW-NB15.

1. Introduction

Clustering approaches show a significant part on guiding user towards the goal in feature selection. Clustering contains several significant applications at machine learning, image segmentation, data mining, pattern recognition. Furthermore a use of feature selection shows an important part in setting efficient, enhanced decision schemes. Feature selection methods support to attain relevant, non-redundant types, which decrease the system difficulty based on time, space. Decrease the related types offer best efficiency. FS is significance in design recognition, data analysis, multimedia information retrieval, processing of medical data, machine learning, data mining.

The effective feature selection procedure lowers the price of measurement feature, improves efficiency, accuracy of clustering. A better clustering can be obtained if significant features are used. K-Means is simple in entire clustering methods and offers good clustering outcomes, but it's contains drawbacks like as sensitivity to early cluster center selection.

To overcome these issues, in this manuscript, FS, Clustering utilizing GOA is proposed and implemented in two dataset such as NSL-KDD and UNSW-NB15 for providing effective feature selection These 2 datasets have many features, in which only a limited count of features contributes to clustering. Noise and unwanted data will be generated as the dimension of the space covering all the features will be huge as well as unclean, thus reducing accuracy of clustering. The effective feature selection technique will remove sound, redundant, redundant data, so the Grasshopper Optimization Algorithm with the feature selection technique is proposed.

The major contributions of this work are

- In this manuscript Feature Selection and Clustering using Grasshopper Optimization Algorithm is proposed and implemented in two dataset such as NSL-KDD, UNSW-NB15 to providing effective feature selection.
- Here, 2 types of dataset that is set NSL-KDD, UNSW-NB15 data set are used. These 2 datasets have many features, in which only a limited count of features contributes to clustering.
- Noise and unwanted data will be generated as the dimension of the space covering all the features will be huge as well as unclean, thus reducing accuracy of clustering.
- The effective feature selection technique will remove sound, redundant, redundant data, so the Grasshopper Optimization Algorithm with the feature selection technique is proposed [10].
- The simulation process is executed in the MATLAB platform.
- Here, the evaluation metrics analyse Mean squared error (MSE), Entropy, accuracy, processing time etc. Then, the performance is likened with the proposed system using two existing methods such as Feature Selection and Clustering Using hybrid GWO–GOA [11] and Feature Selection and Clustering Using Quantum Whale Optimization Algorithm (QWOA)[12].
- Finally the proposed method provides best clustering accuracy with low computational time.

Remaining manuscript is mentioned as below. Section 2 delineates that literature survey. Section 3 describes the feature Selection and Clustering Using Grasshopper Optimization Algorithm. Section 4 illustrates that result and discussion. At last, Section 5 concludes the manuscript

2. Literature Review

Among the frequent research work on Feature Selection and Clustering method; some of the latest investigations were assessed in this part.

In 2020, Purushothaman, et. al [11] have presented a functional selection technique aimed at increasing the text clustering approach's reliability and reducing the count of uninformed attributes. The suggested method provides a mature integration rate, needs less computational time, was trapped in local minima at lower dimensional space. Then text feature selection was performed via choosing a local optimization from in text file with selecting a better global optimization from local optimal using hybrid GWO–GOA. This algorithm raises a consistency, minimal a computational time cost. Compared to the GWO, GOA, the presented hybrid GWO–GOA approach suggested method exposes efficacy 87.6%.

In 2020, Agrawal et al [12] have presented Quantum Whale Optimization Approach (QWOA) for a combined feature selection of Quantum Concepts, Whale Optimization Algorithm (WOA). The suggested technique improves a exploration, the classical WOA exploitation power, using population characters' quantum bit representation, the quantum rotation gate operator as a difference operator. Investigational outcomes show the higher efficiency of recommended QWOA technique. Statistical experiments determine the significant efficiency of QWOA likened to 8 well-known meta-heuristic methods.

3. Proposed methodology for Feature Selection and Clustering using Grasshopper Optimization Algorithm

In this manuscript, FS as well as Clustering utilizing GOA is proposed and implemented in two dataset such as NSL-KDD and UNSW-NB15 for providing effective feature selection. These 2 datasets have many features, in which only a limited count of features contributes to clustering. Noise and unwanted

data will be generated as the dimension of the space covering all the features will be huge as well as unclean, thus reducing accuracy of clustering. The effective feature selection technique will remove sound, redundant, redundant data, so the Grasshopper Optimization Algorithm with the feature selection method is proposed.

3.1 Data set Acquisition

There are two set of data set are utilized to evaluate the proposed technique. This is NSL-KDD and UNSW-NB15 data set [13, 14].

3.2 Feature selection (FS) and clustering

Here, FS is more important before resolving the clustering issues. FS is used in design recognition, data analysis, multimedia information retrieval, processing of medical data, machine learning, data mining. The effective feature selection process decreases the price of feature measurement, increases efficiency, accuracy of clustering. A better clustering can be attained if significant features are utilized.

This manuscript attempts to locate a better data clustering by first performing a feature selection, selecting features to clustering. Hence, Grasshopper Optimization Algorithm with feature selection technique is proposed for providing effective feature selection which improves clustering accuracy.

3.2 Grasshopper Optimization Algorithm (GOA)

The proposed GOA approach mimics the performance of grasshopper swarms in nature to solve mathematically designs with optimisation issues. Grasshoppers are insects. It's considered pests due to damage to crop production, agriculture. The single feature of the grasshopper group is that swarm behaviour is established in nymph, adulthood. The major feature of swarm on a larval stage is the slow movement with the grasshoppers' of small steps. In difference, long- range as well as abrupt movement is a swarm in adulthood an essential feature. Searching of food is another important characteristic of the grasshoppers swarm.

GOA for Data Clustering Problem

Each element in GOA contains several elements as count of data points. Every particle maps to a data point, the element value provides a cluster belonging to element.

If the data points $Y=(y_1,y_2,....y_M)$ then the j^{th} element S'_j denoted as $S'_j=(x_1,x_2,....x_m)$, where x_i represents the cluster to which i^{th} data point y_i belongs. If there is K' clusters, then x_i is a value among $1, K'$. The swarm candidate denotes a count of solutions (clustering). Here, centroid of S'_j particle denotes the data clustering in equation (1)

$$n_i=\frac{\sum_{x\in D_i} x}{|D_i|} \quad i=1,.....k' \tag{1}$$

From equation (1) $|D_i|$ represents the total count of data points on cluster i. Here, fitness function is calculated every element as well as likened by its own better earliest fitness value, to entire particles' better fitness in a swarm. Then, the fitness value determined is in equation (2).

$$F'=\frac{\sum_{i=1}^{k'}[\frac{\sum_{\forall x\in D_i} c(x,n_i)}{|D_i|}]}{\sum_{i=1}^{k'} c(n_i,n)} \tag{2}$$

Where, $n = \frac{\sum_x X}{m}$, *and* $c(b,a) = \sqrt{z'_1 (b_1 - a_1)^2 + z'_2 (b_2 - a_2) + ...z_n (b_n - a_n)}$

Where $z_i \in (0,1)$, for this 0 means i^{th} feature is absent, 1 means i^{th} feature is current. A clustering with a low value of F' is an enhanced clustering of points. The numerator of F' is the sum of all squared error (SSE) of entire points from its centroid, denominators by, among distances of class.

GOA with Feature selection (FS)

Here, FS is a significant factor that influences clustering. There are many features in a database that only a limited count of features contributes to clustering. With completely these features, the outcomes in dimension of the space are large, unclean causing noise, unwanted data, therefore humiliating clustering accurateness. GOA with feature selection works similarly based on GOA with additional feature selection property with velocity and position functions, in which features are chosen in a random basis with fitness value is calculated in (2) only chosen features are consider. Here, step by step procedure of GOA with Feature selection is described as follows: Fig 1 displays the proposed GOA method's block diagram,

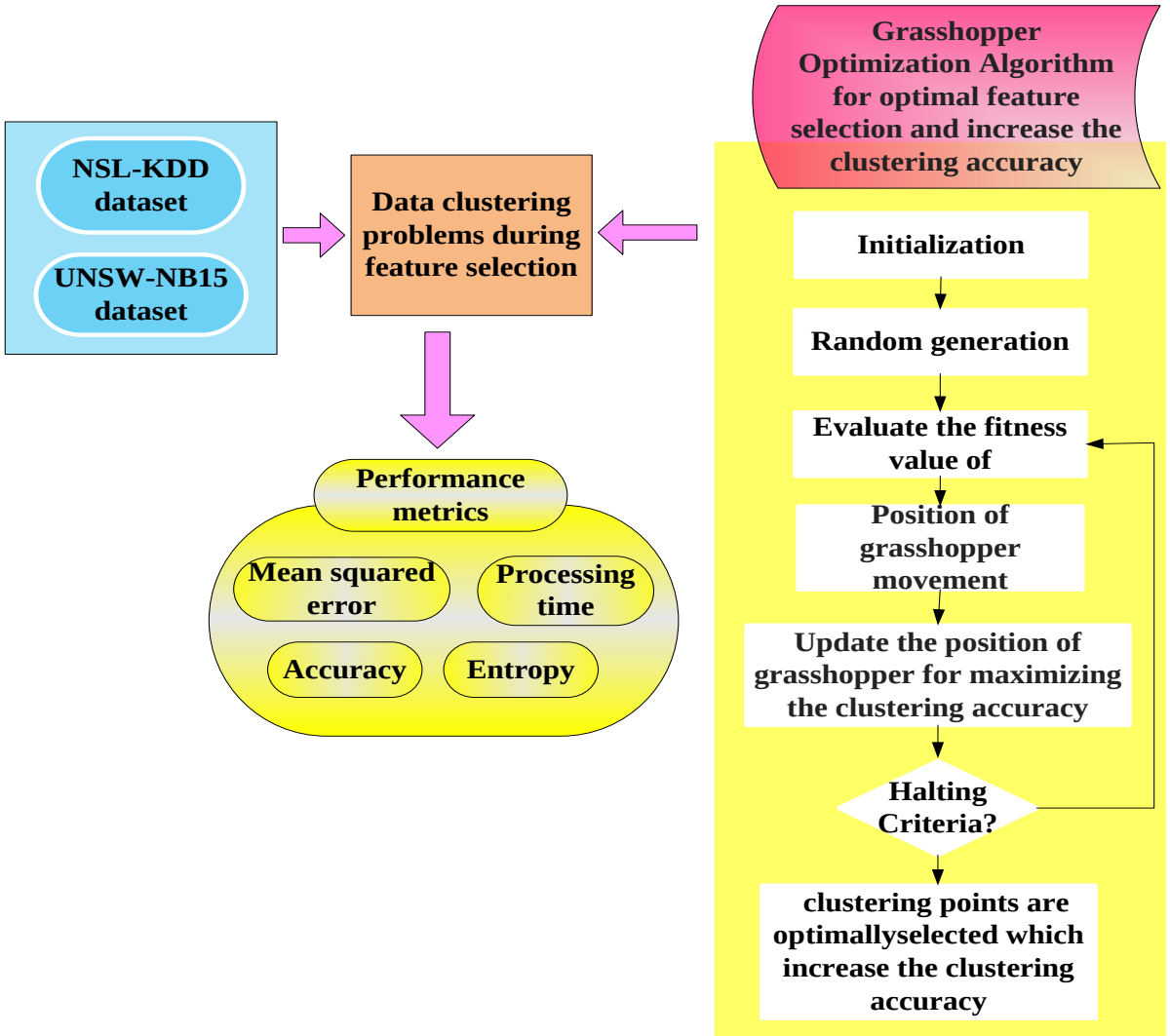

Figure 1: Block diagram of proposed GOA method

Step 1: Initialization and random Generator

The initial M particles are formed randomly with features that are selected randomly. Let's keep, the dataset contains m data points or instances with c features or dimensions to every particle, and then the random behaviour is given in equation (3).

$$Y'_j = r_1 d^{k'}{}_j + r_2 m^{k'}{}_j \quad (3)$$

From equation (3) Y'_j represents the i^{th} grasshopper position, r_1 *and* r_2 are random counts that is [0, 1], $d_j^{k'}$ *and* $m_j^{k'}$ are the features of the K' cluster.

Step 2: Fitness function

In this, the fitness function is evaluated to get the objective function to maximize the efficiency and clustering accuracy. In this, the clustering accuracy and clustering efficiency maximization, that is given in equation (2).

Step 3: Position of grasshopper movement

In the grasshopper movement's process, social interaction J'_1 plays an important part. Then can be formulated in equation (4).

$$J'_1 = \sum_{g=1, g\neq 1}^{M'} j(d'_{g1}) \hat{d'}_{g1} \quad (4)$$

From equation (4) $\hat{d'}_{1g} = \frac{x'_g - x'_1}{d'_{1g}}$ is a unit vector of two grasshoppers. Euclidean distances between two grasshoppers are denoted as $d'_{g1} = \left| x'_1 - x'_g \right|$.

Step 4: Position searching of grasshopper

For search the position searching, the social interaction of grasshopper is more important. The intensity of social interaction j' can be calculated in equation (5).

$$j'(r') = f' e'^{\frac{-r'}{l'} - e^{-r'}} \quad (5)$$

From equation (5) the intensity attraction is represented as f', attractive length scale can be indicated as l'.

Step 5: Update the position of grasshopper for maximizing the clustering accuracy

Grasshoppers quickly reach their comfort zone in equation (4). The formula has improved to make the algorithm close to the optimum solution of clustering accuracy, to integrate at a certain point. Let $J'^{d}{}_{1h}$ represent the position of grasshopper h in d^{th} dimension. The improved formula is given in equation (6)

$$J'^{d}{}_{1h} = c'_1 \left(\sum_{g=1, g\neq 1}^{M'} c'_2 \frac{upperbound - lowerbound}{2} g'_{g1} \right) + T'_d \quad (6)$$

From equation (6) c'_1 *and* c'_2 utilized to simulate the grasshoppers' slowdown process, which accessed the position of food gradually, finally consume a food, T'_d represents a d^{th} dimension of the search agent. Using this equation (6) the maximization of clustering accuracy and efficiency can be attained.

Step6: Termination

In this step, if the optimal solution is achieved, then iteration is stopped if not the step 1 to 5 are repeated until the criteria are met. Finally the GOA algorithm provides maximal clustering efficiency and accuracy.

4. Result and discussion

Here, Feature Selection and Clustering using GOA is proposed and implemented in two dataset such as NSL-KDD and UNSW-NB15 for providing effective feature selection which improves the clustering accuracy. The simulations are performed in PC with Intel Core i5, 2.50 GHz CPU, 8GB RAM, Windows 7. The simulations executed in MATLAB. Then, the calculation metrics like MSE, Entropy, accuracy, processing time are analysed. Here, the performance is likened with the proposed system using two existing methods such as Feature Selection and Clustering Using hybrid GWO–GOA and Feature Selection and Clustering Using QWOA. The simulation parameters of the proposed approach are exposed in Table 1

Table1: Parameters used in simulation

Simulation parameters	values
Population size	$M = 10$ particles
inertia weight ω	0.79
No of iteration	650
Constant values	c'_1 *and* c'_2=2

4.1 Data set description

The proposed model having the comparison analysis is carried out into two dataset. The two datasets are NSL-KDD, UNSW-NB15 dataset. The data set description is shown below table2.

Table2: Data set description of NSL-KDD, UNSW-NB15 data set

Data set	Count of instances	Count of features	Count of attack
NSL-KDD	125,973	41 features	5
UNSW-NB15	2,540,044	45 features	10

4.2 Performance Comparison

4.2.1 Performance metrics For NSL-KDD Dataset

Here, the performance of NSL-KDD dataset is likened with two existing methods, like Feature Selection and Clustering Using hybrid GWO–GOA and Feature Selection and Clustering Using Quantum Whale Optimization Algorithm (QWOA). Here assessment metrics like MSE, Entropy, accuracy and processing are analysed. Table 3 shows the performance metrics of NSL-KDD dataset.

Table 3: Performance metrics of NSL-KDD dataset (selected features=13)

Methods	Features selected	Fitness values	Mean squared error (MSE)	Entropy	Accuracy	Processing time
hybrid GWO–GOA	All(41)	1.202	1.007	1.456	.68	2.45sec
QWOA	All(41)	0.382	0.302	1.102	.60	1.45sec
GOA(proposed)	13	0.063	0.072	1.302	.94	.63sec

In table 3 the fitness values of the proposed method are 73.55% and 60.33% lesser than the existing techniques. The MSE of the proposed technique is 70.55% and 71.33% smaller than the existing techniques. The Entropy of the proposed technique is 63.55% and 60.33% smaller than the existing methods and processing time of the proposed method is 40.55% and 40.22% lesser than the existing techniques, and Accuracy of the proposed method 25.6% and 20.45% high when compared to the two existing methods such as hybrid GWO-GOA and QWOA.

4.2.2 Performance metrics For UNSW-NB15 Dataset

The performance of UNSW-NB15dataset is likened with two existing methods, like Feature Selection and Clustering Using hybrid GWO–GOA and FS and Clustering utilizing QWOA. Here assessment metrics like MSE, Entropy, accuracy, processing are analysed. In table 4 tabulates the performance metrics of NSL-KDD dataset.

Table 4: Performance metrics of UNSW-NB15 dataset (selected features=10)

Methods	Features selected	Fitness values	Mean squared error (MSE)	Entropy	Accuracy	Processing time
hybrid GWO–GOA	All(45)	1.302	1.007	1.456	.68	3.45sec
QWOA	All(45)	0.392	0.302	1.102	.70	2.45sec
GOA(proposed)	10	0.073	0.072	1. 402	.95	.73sec

In table 4 the fitness values of the proposed method is 70.55% and 62.33% lesser than the existing techniques. The MSE of the proposed technique is 63.55% and 50.33% smaller than the existing techniques. The Entropy of the proposed technique is 63.55% and 50.33% smaller than the existing methods and processing time of the proposed method is 43.55% and 60.33% lesser than the existing techniques, and Accuracy of the proposed method 22.6% and 21.45% high when compared to the two existing methods such as hybrid GWO-GOA and QWOA.

5. Conclusion

In this manuscript, FS and Clustering utilizing GOA is proposed and implemented in two dataset such as NSL-KDD and UNSW-NB15 for providing effective feature selection. These 2 datasets have many features, in which only a limited count of features contributes to clustering. Noise and unwanted data will be generated as the dimension of the space covering all the features will be huge as well as unclean, thus reducing accuracy of clustering. The effective feature selection technique will remove sound, redundant, redundant data, so the Grasshopper Optimization Algorithm with the feature selection technique is proposed. Finally the proposed technique provides best clustering accuracy with low computational time. The simulation process is executed in the MATLAB platform. The UNSW-NB15 data set of the proposed method attains high accuracy 22.6% and 21.45%, low mean square error (MSE) 63.55% and 50.33%, low Entropy 63.55% and 60.33%, low processing time 43.55% and 60.33% shows

better performance when comparing with the existing method such as Feature Selection and Clustering Using hybrid GWO–GOA and Feature Selection and Clustering Using Quantum Whale Optimization Algorithm (QWOA).

References

1) Zhou, P., Chen, J., Fan, M., Du, L., Shen, Y.D. and Li, X., 2020. Unsupervised feature selection for balanced clustering. *Knowledge-Based Systems*, *193*, p.105417.
2) Zhu, X., Zhang, S., Zhu, Y., Zhu, P. and Gao, Y., 2020. Unsupervised spectral feature selection with dynamic hyper-graph learning. *IEEE Transactions on Knowledge and Data Engineering*.
3) Song, X.F., Zhang, Y., Guo, Y.N., Sun, X.Y. and Wang, Y.L., 2020. Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. *IEEE Transactions on Evolutionary Computation*, *24*(5), pp.882-895.
4) Zhang, X., Fan, M., Wang, D., Zhou, P. and Tao, D., 2020. Top-k feature selection framework using robust 0-1 integer programming. *IEEE Transactions on Neural Networks and Learning Systems*.
5) Hancer, E., 2020. A new multi-objective differential evolution approach for simultaneous clustering and feature selection. *Engineering Applications of Artificial Intelligence*, *87*, p.103307.
6) Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-pour, H., 2020. MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. *Expert Systems with Applications*, *142*, p.113024.
7) Nasir, V. and Cool, J., 2020. Intelligent wood machining monitoring using vibration signals combined with self-organizing maps for automatic feature selection. *The International Journal of Advanced Manufacturing Technology*, *108*, pp.1811-1825.
8) Rostami, M., Forouzandeh, S., Berahmand, K. and Soltani, M., 2020. Integration of multi-objective PSO based feature selection and node centrality for medical datasets. *Genomics*, *112*(6), pp.4370-4384.
9) Alazzam, H., Sharieh, A. and Sabri, K.E., 2020. A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. *Expert systems with applications*, *148*, p.113249.
10) Abualigah, L. and Diabat, A., 2020. A comprehensive survey of the Grasshopper optimization algorithm: results, variants, and applications. *Neural Computing and Applications*, pp.1-24.
11) Purushothaman, R., Rajagopalan, S.P. and Dhandapani, G., 2020. Hybridizing Gray Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for text feature selection and clustering. Applied Soft Computing, 96, p.106651.
12) Agrawal, R.K., Kaur, B. and Sharma, S., 2020. Quantum based whale optimization algorithm for wrapper feature selection. Applied Soft Computing, 89, p.106092.
13) https://www.kaggle.com/hassan06/nslkdd
14) https://research.unsw.edu.au/projects/unsw-nb15-dataset