Clustering Crime News Headlines Using Grasshopper Optimization and K-Means

K.Dheenathayalan[1], K.K. Savitha[2]

[1]Assistant Professor, Department of Computer Science, Kamban College of Arts and Science, Coimbatore, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Applications, Bharathiar University PG Extension and Research Centre, Erode, Tamil Nadu, India.

[1]dheenamca08@gmail.com, [2]savitha.gopinath@gmail.com

Abstract – Nowadays security of a nation including its citizens which is regarded as a duty of government and given higher priority to reduce crime incidence. As data mining techniques are appropriate to apply on high volume crime dataset and knowledge extracted from data mining techniques will be beneficial to the police department to prevent crimes. So, in this paper crime data clustering is done by performing k-means clustering with grasshopper optimization algorithm on crime dataset using R-Studio tool.

Keywords: *clustering, optimization, crime, k-means, grasshopper.*

1. Introduction

The crime data available is in both the forms structured form and unstructured form. Structured form is written criminal record and unstructured forms are image, audio and video from surveillance. The all collected data will be put up together and the bulk of the data will contribute to the understanding of the events happened in past and based on the crime patterns what is more likely to happen in future. With such kind of crime prediction available, it can help the law enforcement agencies or respected police departments to take necessary actions by understanding the crime patterns.

Government invests significant money and time in developing reliable decision support systems in order to make timely decisions to prevent future crimes in a nation. The quality and availability of data have a direct impact on a decision-making approach and the outcome of its operations.

Crime in India has been recorded since the British period, with comprehensive statistics now collected annually by the National Crime Records Bureau (NCRB), under the Ministry of Home Affairs (India) (MHA). As the volume of data increasing day by day, it is necessary to organize those data to improve the quality and availability of data. The statistics of different types of crimes are maintained in the repository. Grouping those data under different crime head is an important and risky task. A crime analyst or police officer can analyze and understand crime statistics once they are categorized of clustered under respective crime head. There are some clustering and classification algorithms available in data mining technique to cluster and classify the crime data.

k-means clustering is one of clustering algorithms which is very familiar and most applied clustering technique to cluster large volume of data, but the drawback is due to the iterative nature of K-Means and random initialization of centroids, K-Means may stick in a local optimum and may not converge to global optimum. So, bio-inspired optimization algorithms can be hybridized with k-means to optimize the cluster centroids for better clustering performance. In this work, grasshopper optimization algorithm is used for cluster centroid optimization.

2. Methodology

The proposed methodology is to implement grasshopper optimization algorithm to optimize cluster centers so that the performance of k-means clustering will be improved and does achieve global optimum. At first, k-means clustering is performed on crime data to generate N number of sets and each set contains set of cluster centroids. Then, optimization algorithm is applied on this set of cluster centers to optimize the centers and end with global best among them. The final value from this optimization is referred as optimum solution. These optimized cluster centers are initialized in k-means clustering to group the crime data and this time k-means achieves global optimum.

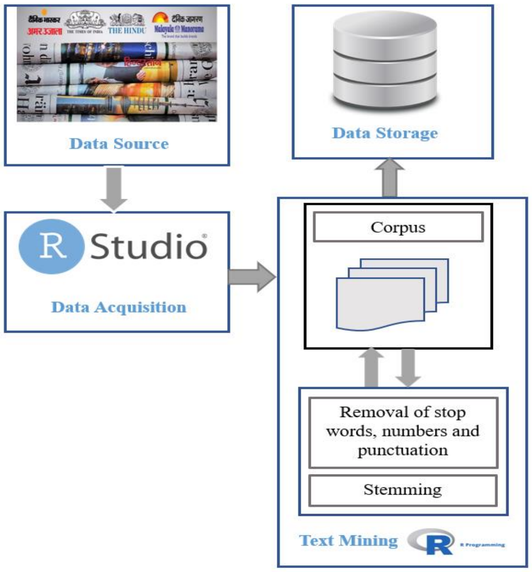


Fig.1. Framework of data extraction model

The crime data considered for this work is acquired from news websites like India Today, Times of India, The Hindu, etc. and specially concentrated on crimes against women in India. Fig.1, represents the framework of data extraction model. The data is in the form of text which is the headline of the crime incident. Initially there were more than 50,000 records gathered from various newspapers. Crime records are collected under three types violent crimes against women in India they are; rape, rape & murder and kidnap / abduction. Crime data are collected by using the web-scraping technique. After data preprocessing there were more than 5000 records under each crime head. Among them, only thousand records from each crime head is selected randomly to test the proposed clustering approach.

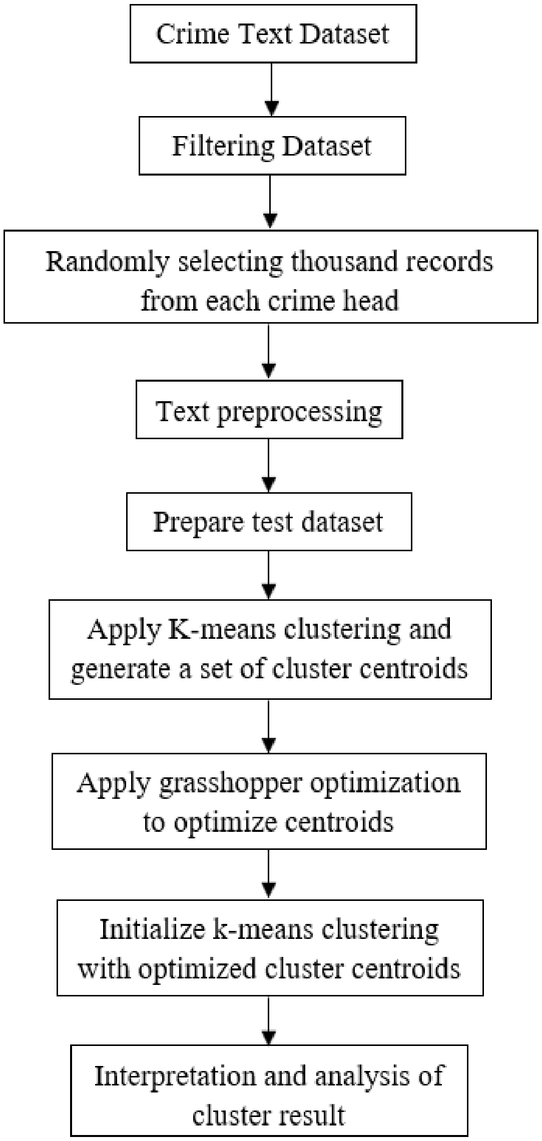

Crime Text Dataset
Filtering Dataset
Randomly selecting thousand records from each crime head
Text preprocessing
Prepare test dataset
Apply K-means clustering and generate a set of cluster centroids
Apply grasshopper optimization to optimize centroids
Initialize k-means clustering with optimized cluster centroids
Interpretation and analysis of cluster result

Fig.2. Architecture of proposed approach

Steps involved in proposed approach are given below,

- Crime text dataset contains news headlines of crime events occurred in India. Each row represents different crime event.
- This dataset is filtered to remove incomplete and inappropriate records.
- After that text preprocessing is carried out to remove stop words and punctuation, text tokenization and stemming.

- k-means clustering applied to generate initial population for grasshopper optimization. This initial population contains a set of cluster centroids.
- Grasshopper optimization is applied to optimize these cluster centers and find the global optimum.
- Again, k-means clustering is initialized with this optimized cluster center and clustered the data objects.
- Final cluster result is interpreted and analyzed.

3. Experimental Setup

a) K-means Clustering

K-means clustering is one of the methods of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

b) Process

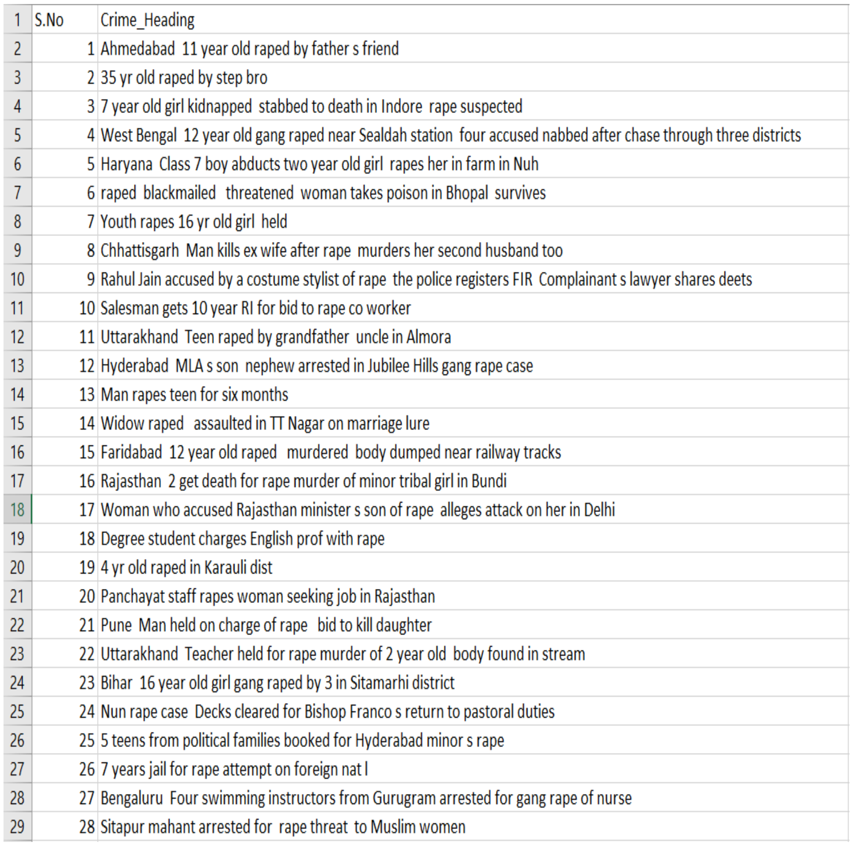
Initially, number of clusters must be known, here the dataset contains crime headlines under three crime heads, rape, rape and murder, and kidnap or abduction. Maximum number of iterations is 500. At first, k-means starts with randomly picked instances as cluster centroids and forms the cluster. Then, iteratively calculates cluster mean and reforms the clusters until it reaches maximum number of iterations or it converges. After the maximum number of iterations, all set of cluster centers are converted to data frame for optimization.

c) Grasshopper Optimization

Initial population of grasshopper optimization is the set of cluster centers which is in matrix form. Each row represents centers of each cluster k. Aim of grasshopper optimization is to find optimal clustering centers among population according to the objective function value (Euclidean distance measure). After that, optimal cluster centers are initialized with k-means clustering to cluster data objects. And finally, the performance of this approach is measured based on some cluster analysis metrics. They are, standard deviation, within cluster sum of square, between cluster sum of square and average silhouette score. The performance of proposed approach is compared with other three methods, simple k-means, FOA - k-means and MFO - k-means based on above said metrics.

d) Dataset Description

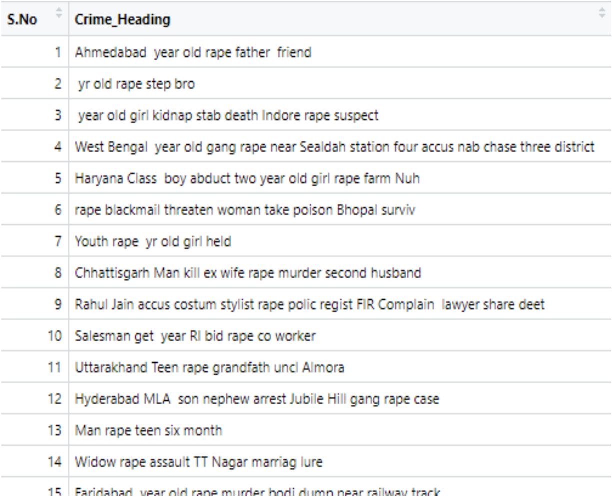
The dataset contains three thousand records where each line is a heading news of a crime. In Fig.3, it shows crime headings before text preprocessing. Text preprocessing is the step to remove stop words, punctuations, numbers, unwanted white spaces and word stemming. The concept here is to keep only important keywords of crime headings those are necessary for understanding the crime type.



1	S.No	Crime_Heading
2	1	Ahmedabad 11 year old raped by father s friend
3	2	35 yr old raped by step bro
4	3	7 year old girl kidnapped stabbed to death in Indore rape suspected
5	4	West Bengal 12 year old gang raped near Sealdah station four accused nabbed after chase through three districts
6	5	Haryana Class 7 boy abducts two year old girl rapes her in farm in Nuh
7	6	raped blackmailed threatened woman takes poison in Bhopal survives
8	7	Youth rapes 16 yr old girl held
9	8	Chhattisgarh Man kills ex wife after rape murders her second husband too
10	9	Rahul Jain accused by a costume stylist of rape the police registers FIR Complainant s lawyer shares deets
11	10	Salesman gets 10 year RI for bid to rape co worker
12	11	Uttarakhand Teen raped by grandfather uncle in Almora
13	12	Hyderabad MLA s son nephew arrested in Jubilee Hills gang rape case
14	13	Man rapes teen for six months
15	14	Widow raped assaulted in TT Nagar on marriage lure
16	15	Faridabad 12 year old raped murdered body dumped near railway tracks
17	16	Rajasthan 2 get death for rape murder of minor tribal girl in Bundi
18	17	Woman who accused Rajasthan minister s son of rape alleges attack on her in Delhi
19	18	Degree student charges English prof with rape
20	19	4 yr old raped in Karauli dist
21	20	Panchayat staff rapes woman seeking job in Rajasthan
22	21	Pune Man held on charge of rape bid to kill daughter
23	22	Uttarakhand Teacher held for rape murder of 2 year old body found in stream
24	23	Bihar 16 year old girl gang raped by 3 in Sitamarhi district
25	24	Nun rape case Decks cleared for Bishop Franco s return to pastoral duties
26	25	5 teens from political families booked for Hyderabad minor s rape
27	26	7 years jail for rape attempt on foreign nat l
28	27	Bengaluru Four swimming instructors from Gurugram arrested for gang rape of nurse
29	28	Sitapur mahant arrested for rape threat to Muslim women

Fig.3. Dataset before text preprocessing

In Fig.4, it shows crime headings after preprocessing. It clearly understandable that, all unwanted words and numbers are removed from original text and remaining are important keywords those are used to distinguish the crime type. Here, the clustering technique is applied to the given dataset to cluster them using the type of the crime.



S.No	Crime_Heading
1	Ahmedabad year old rape father friend
2	yr old rape step bro
3	year old girl kidnap stab death Indore rape suspect
4	West Bengal year old gang rape near Sealdah station four accus nab chase three district
5	Haryana Class boy abduct two year old girl rape farm Nuh
6	rape blackmail threaten woman take poison Bhopal surviv
7	Youth rape yr old girl held
8	Chhattisgarh Man kill ex wife rape murder second husband
9	Rahul Jain accus costum stylist rape polic regist FIR Complain lawyer share deet
10	Salesman get year RI bid rape co worker
11	Uttarakhand Teen rape grandfath uncl Almora
12	Hyderabad MLA son nephew arrest Jubile Hill gang rape case
13	Man rape teen six month
14	Widow rape assault TT Nagar marriag lure
15	Faridabad year old rape murder bodi dump near railway track

Fig.4. Dataset after text preprocessing

4. Results and Discussion

First and important step in clustering is to choose the number of clusters to be formed. Choosing optimum number of clusters can provide better clustering result. Here, elbow method is used for that with the help of within sum of squares (WSS) measure. Below Fig.5. represents graph produced by elbow method. The points in the graph are the intersection of number of clusters (K) on X axis and value of WSS on Y axis. The point from where the WSS goes down gradually is the elbow point and the number of clusters on that point is chosen as optimum number of clusters. In this graph the elbow point is 4 hence the optimum number of clusters K = 4.

In Fig.6. it represents word cloud produced from the text dataset, it shows the frequent terms or words of the text dataset. The word which is bigger in size is the most frequent word among other words, second bigger word is the second most frequent word among them. In this figure, rape is the most frequent word, murder is the second bigger word followed by abduct and kidnap. Cluster size is the number of data objects in that cluster. Here, cluster 1 contains 380 data objects, cluster 2 contains 376 objects, in cluster 3 there are 425 objects and in cluster 4 there are 1819 objects.

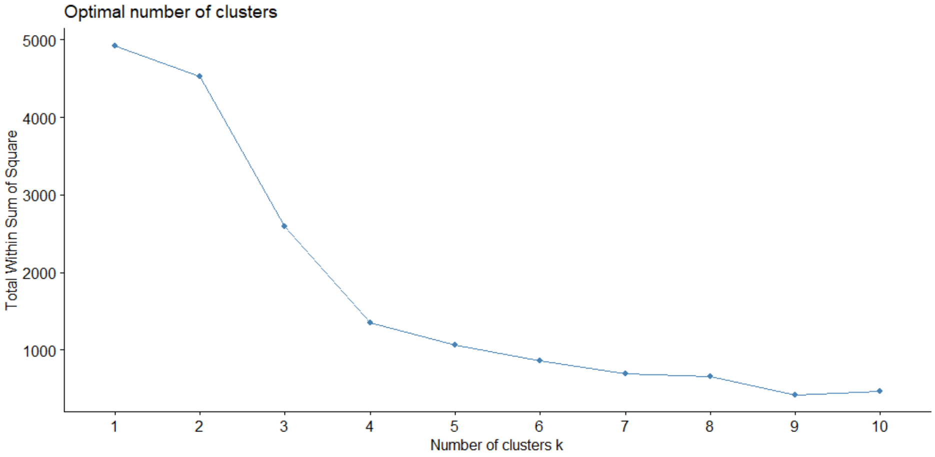


Fig.5. Elbow method- optimul number of clusters

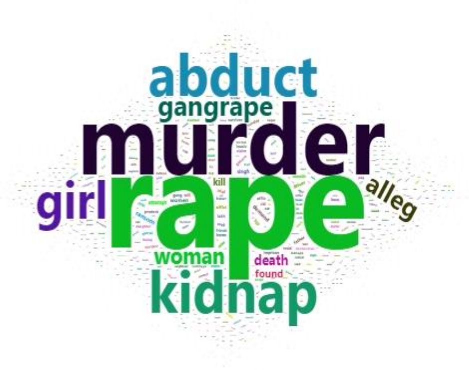


Fig.6. Word cloud

Table1. Terms and their frequencies with respect to the cluster

Terms frequency	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Rape	29	-	39	**1724**
Abduct	-	27	**425**	124
Gangrape	-	**377**	-	-
Kidnap	**380**	24	-	110
Murder	20	180	23	**1078**
Girl	36	56	55	**430**
Woman	23	41	20	177
Ransom	16	-	20	-

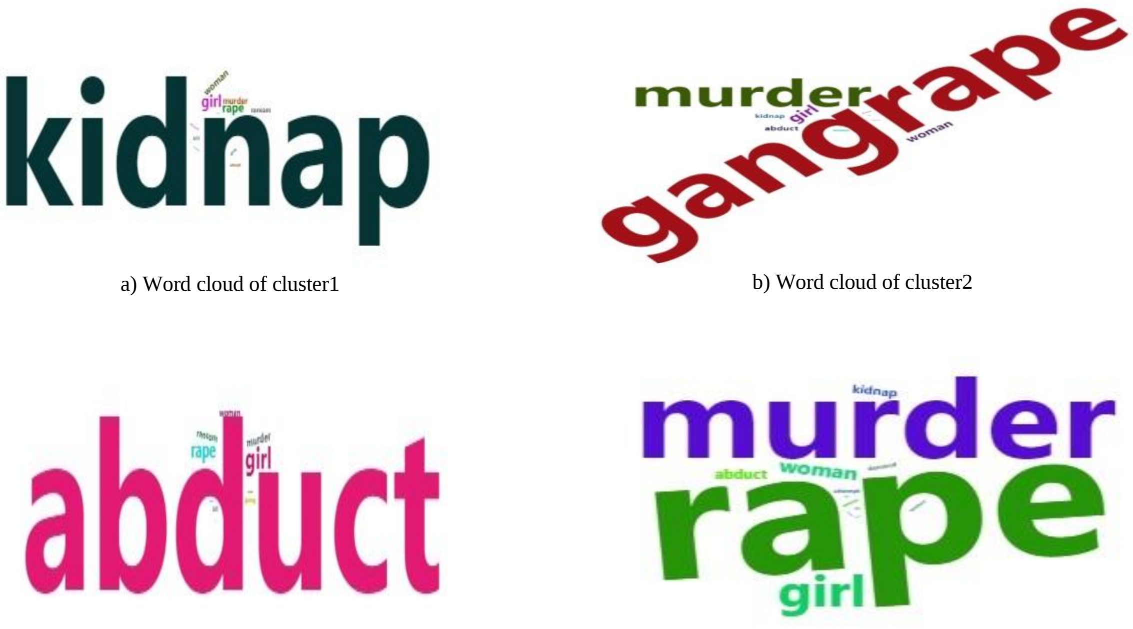
Initially the text dataset contains 3000 rows. Each row represents headline of a crime incident. After text preprocessing all stop-words are removed only keywords are kept for further analysis. These keywords are known as tokens; hence each row contains only the tokens. After that the dataset is copied to a data-frame and then converted to term document matrix where all tokens are represented in numerical value based on their frequency.

All these tasks are accomplished with the help of a R-library called as tm (text mining). Then, these matrix rows are clustered based on their frequency values. As said before, number of clusters (k) chosen was 4, hence 4 clusters are formed. Table1, consists terms and their frequencies in all clusters. Rape in cluster 4, abduct in cluster 3, gangrape in cluster 2 and kidnap in cluster 1 are frequent terms of the respected cluster.

Cluster quality metrics are the measures to obtain the quality of the clusters. There were three quality measures used here, WCSS, BCSS and cluster percentage. It clearly shows that k-means with grasshopper optimization gives better result than simple k-means for clustering crime text dataset. Table.2. represents cluster quality measures used to test the proposed clustering approach on news headlines data.

Table 2. cluster quality metrics

Metrics	KM-GOA	Simple K-Means
WCSS	**1353.12**	2134.64
BCSS	**3565.89**	2784.37
Percentage	**72.5 %**	56.6 %



a) Word cloud of cluster1

b) Word cloud of cluster2

c) Word cloud of cluster3

d)Word cloud of cluster4

Fig.7. word cloud of different clusters

Fig.7, is the pictorial representation of Table1, it displays the frequent terms of each cluster.

5. Conclusion

In this proposed work k-means clustering is optimized by grasshopper optimization algorithm for clustering crime dataset. Crime dataset used in this work contains text data which are web scraped from Indian news websites like Times of India, India Today and The Hindu. Crime incidents related to crime against women category only scraped from websites and stored in dataset. After applying necessary text preprocessing methods, the text document clustering is done. The optimum number of clusters k=4 is chosen by applying elbow method, hence the number clusters formed is four. These clusters contain frequent terms of crime text data. The four clusters are formed based on the frequency of terms in the dataset. The frequent terms in cluster1, cluster2, cluster3 and cluster4 are kidnap, gangrape, abduct and rape respectively. Clusters quality metrics show that k-means with grasshopper optimization is better than simple k-means for clustering crime dataset presented in this paper. This proposed method will be very useful in grouping the crime headlines with respect to the crime category.

Reference

[1] Almaw and Kadam K. Crime Data Analysis and Prediction Using Ensemble Learning, Second - ICICCS, pp. 1918-1923, doi: 10.1109/ICCONS. (2018).8663186.

[2] Bharati A. and Sarvanaguru K.Crime Prediction and Analysis Using Machine Learning, IRJET, Volume: 05 Issue: 09, (2018), e-ISSN: 2395-0056 , p-ISSN: 2395-0072.

[3] Kim S, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime Analysis Through Machine Learning," *(2018) IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 415-420, doi: 10.1109/IEMCON.2018.8614828.

[4] Mahmud S, Musfika Nuha and Abdus Sattar. Crime Rate Prediction Using Machine Learning and Data Mining. Soft Computing Techniques and Applications. Advances in Intelligent Systems and Computing, vol 1248. (2021), Springer, Singapore.

[5] Yassine Meraihi, Asma Benmessaoud Gabis, Seyedali Mirjalili and Amar Ramdane-cherif. Grasshopper Optimization Algorithm: Theory, Variants and Applications. DOI 10.1109 / ACCESS. (2021). 3067597, IEEE Access

[6] Doaa Abdullah, Hala Abdel-Galil and Ensaf Hussein. Enhancing K-Means Clustering with Bio-Inspired Algorithms. IJCSN - Volume 7, Issue 6, December (2018).

[7] Binitha S and S Siva Sathya. A Survey of Bio inspired Optimization Algorithms. (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May (2012).

[8] M. A. El-Shorbagy and A. Y. Ayoub. Integrating Grasshopper Optimization Algorithm with Local Search for Solving Data Clustering Problems. IJCIS Vol. 14(1), (2021), pp. 783‑793. ISSN: 1875- 6891.

[9] Simon Fong, Suash Deb, Xin-She Yang and Yan Zhuang. Towards Enhancement of Performance of K-Means Clustering Using Nature-Inspired Optimization Algorithms. The Scientific World Journal Volume (2014), Article ID 564829, 16 pages.

[10] Khushabu A. Bokde, Tiksha P. Kakade, Dnyaneshwari S. Tumsare and Chetan G. Wadhai Crime Analysis Using K-Means Clustering. (IJERT) ISSN: 2278-0181.IJERTV7IS040099 Vol. 7 Issue 04, April-(2018)

[11] Shahrzad Saremi, Seyedali Mirjalili and Andrew Lewis. Grasshopper Optimisation Algorithm: Theory and application. Advances in Engineering Software 105 (2017) 30–47.

Author's profile

Mr.K.Dheenathayalan, received B.Sc degree in Computer Science, M.C.A and M.Phil degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu in 2008, 2011 and 2014 respectively. He is currently working as Assistant Professor in Department of Computer Science, Kamban College of Arts and Science, Coimbatore. He has 7 years of experience as Assistant Professor. He is currently pursuing Ph.D in Computer Science in Bharathiar University. His current research interests include Data Mining and Bio-informatics.

Dr.K.K.Savitha received B.Sc, MCA and M.Phil degrees. She received Ph.D in Computer Science from Anna University, Chennai, in 2013. She is currently working as an Assistant Professor in the Department of Computer Applications, Bharthiar University PG Extension Centre, Erode. Her research interest includes mobile computing, Cloud Computing and Soft Computing. She is a Life member of ISTE, CSI.