Volume: 15 Issue: 05 Pages: 6122 – 6125 (2023) ISSN: 0975-0290

# An Optimization-Based Data Search Clustering Approach for Multidimensional Datasets

**Pradeep Kumar Atulker**
Research Scholar, Department of CS & IT, Rabindranath Tagore University, Bhopal
Email : pradeepatulker@gmail.com
**Dr. Rajendra Gupta**
Associate Professor, Department of CS & IT, Rabindranath Tagore University, Bhopal
Email : rajendragupta1@yahoo.com

**ABSTRACT**

**Multidimensional data refers to datasets featuring with multiple columns, often referred to as features or attributes. The challenge in multidimensional data analysis is that clusters and outliers are often detected based on the dataset's features, which may not align well with ground truth in real-world scenarios (e.g., gene expression data). Efficiency is a critical consideration as optimized clustering algorithms must handle the growing size of multidimensional datasets. In this research paper, we have proposed a Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization for Clustering (SCIE_EOC) to data search in multidimensional datasets. The result shows the proposed method shows around 92-95 per cent accuracy for different datasets which is around 5 per cent better than the earlier methods.**

Keywords - **Multidimensional dataset, Clustering, Optimization, Data Search Capability.**

Date of Submission: January 10, 2024 Date of Acceptance: February 6, 2024

## I. INTRODUCTION

The Multidimensional data is a data set with many different columns, also called features or attributes. The more columns in the data set, the more likely discover hidden insights. Think of this data as being in a cube on multiple planes. It organizes the many attributes and enables users to dig deeper into probable trends or patterns. Queries can be interrogated rather than just submit them, as practiced in relational databases. It's a comparatively fast exercise, manipulating the different dimensions and perspectives by attribute [1.2].

The generation of multi-dimensional data has proceeded at an explosive rate in many disciplines with the advance of modern technology. Many new clustering, outlier detection and cluster evaluation approaches are presented in the last a few years. Nowadays a lot of real data sets are noisy, which makes it more difficult to design algorithms to process them efficiently and effectively [3].

## II. CLUSTERING FOR MULTIDIMENSIONAL DATASETS

Cluster analysis divides Multidimensional data into groups (clusters) for the purposes of summarization or improved understanding. For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, or as a means of data compression [4]. While clustering has a long history and a large number of clustering techniques have been developed in statistics, pattern recognition, data mining, and other fields, significant challenges still remain. In this part of study, the author provides a short introduction to cluster analysis, and then focus on the challenge of clustering multi-dimensional data [5].

The cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group should be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better the clustering [6].

The definition of what constitutes a cluster is not well defined, and in many applications, clusters are not well separated from one another. Nonetheless, most cluster analysis seeks, as a result, a crisp classification of the data into non-overlapping groups. Fuzzy clustering is an exception to this, and allows an object to partially belong to several groups [7, 23].

To illustrate the difficulty of deciding what constitutes a cluster, consider Figure 1, which shows twenty points and three different ways that these points can be divided into clusters. If we allow clusters to be nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three subclusters. However, the apparent division of the two larger clusters into three subclusters may simply be an artefact of the human visual system [8, 9-11]. Finally, it may not be unreasonable to say that the points form four

Volume: 15 Issue: 05 Pages: 6122 – 6125 (2023) ISSN: 0975-0290

clusters. Thus, the author again stress that the notion of a cluster is imprecise, and the best definition depends on the type of data and the desired results.

Moreover, cluster analysis is a classification of objects from the data, where by "classification" means a labelling of objects with class (group) labels. As such, clustering does not use previously assigned class labels, except perhaps for verification of how well the clustering worked. Thus, cluster analysis is sometimes referred to as "unsupervised classification" and is distinct from "supervised classification," or more commonly just "classification," which seeks to find rules for classifying objects given a set of pre-classified objects.

## III. DATA SEARCH OPTIMIZATION FOR MULTIDIMENSIONAL DATASETS

The researcher presents some basic concepts of multidimensional clustering and the importance of pre-processing in data mining.

### A.Particle Swarm Optimization (PSO)

The objective in an optimization problem is to find values for a set of parameters that maximize or minimize an objective function. PSO is a nature-inspired, population-based, stochastic heuristic for solving optimization problems by simulating the collective behavior of bird flocks [12]. A swarm in PSO is a population of interacting agents, called particles, representing each candidate solution. Once initialized, the *n* particles of a swarm collectively move in the search space, sharing information among them in order to collaboratively find the best solution. PSO uses particle interaction to find the optimal solution, although it converges slowly to the global optimum due to the high-dimensional search space.

### B.Grey Wolf Optimizer (GWO)

Grey Wolf Optimizer (GWO), which is inspired on the Canis lupus species. The COA has a different algorithmic structural setup and it does not focus on the social hierarchy and dominance rules of the animals, even though the alpha is employed as the leader of a pack. Further, the COA focus on the social structure and experiences exchange by the coyotes instead of only hunting preys as it happens in the GWO [13]. The Grey Wolf Optimisation (GWO) method has a few drawbacks, including a slow convergence rate, poor local searching ability, and low solving accuracy.

### C.k-Means

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function. Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen [14, 15]. Since, Data variation is not taken into consideration by K-Means, so it do not consider in large scale data clustering.

Following is the Entropy-based Elephant Herd Optimization technique which is employed for optimal clustering since the goal is to minimize the number of clusters.

## IV. SINUSOIDAL CHAOTIC AND INFORMATION ENTROPY BASED ELEPHANT-HERDING OPTIMIZATION FOR CLUSTERING (SCIE_EOC)

In general, elephants live in groups. An elephant herd is made up of several different clans. Different elephants in the same clan live together under the leadership of a matriarch. There are only adult females and calves in the herd. An elephant does not mate with members of its own family, so as adults; the male elephants will leave the herd to find mates [16, 19]. Inspired by the elephant herd behavior, the SCIE_EOC algorithm is proposed to solve the optimization problems. In SCIE_EOC, the behavior of the elephant group is idealized into two parts: clan updating operator and separating operator. In the clan updating process, each elephant carries on the clan updating operator by its current position and the matriarch's position in the herd.

Elephants, as social creatures, live in social structures of females and calves. An elephant clan is headed by a matriarch and composed of a number of elephants. Female members like to live with family members, while the male members tend to live elsewhere. They will gradually become independent of their families until they leave their families completely [17]. The basic idea of taking SCIE_EOC is to analyse multidimensional data for clustering, since the algorithm has shown the better performance in clustering in research data.

The following assumptions are considered in SCIE_EOC.

(1) Some clans with fixed numbers of elephants comprise the elephant population.
(2) A fixed number of male elephants will leave their family group and live solitarily far away from the main elephant group in each generation.
(3) A matriarch leads the elephants in each clan.
(4) In the process of clan updating, for ordinary elephants, their positions are updated according to their own positions and the positions of their matriarch. The position of matriarch is updated by the middle positions of the clan. It is easy to form into local optimum for one clan of an elephant group. In elephant groups, the matriarch should lead others to explore more suitable habitats.
(5) The process simulates the departure of an adult male elephant and the birth of a calf in the herd. As a

Volume: 15 Issue: 05 Pages: 6122 – 6125 (2023) ISSN: 0975-0290

young elephant, it will be protected by other elephants, and it will have a good position. Therefore, its position should be evaluated.

The location of succeeding elephants in clan $L_n$ is obstructed through matriarch. The location of $\alpha^{th}$ elephant is reformed in clan $L_n$ using eq (1).

$$X_{next,L_n,\alpha} = X_{L_n,\alpha} + \beta \times \left(X_{optimal,L_n} - X_{L_n,\alpha}\right) \times rm \quad (1)$$

Where,

$X_{next,L_n,\alpha}$ & $X_{L_n,\alpha}$ = The new and current locations of $\alpha^{th}$ elephant in clan $L_n$ respectively.

$\beta$ = A magnitude term attaining the influence of matriarch $L_n$ on $X_{Ln,\alpha}$. $\{\beta \in [0,1]\}$

$X_{optimal,L_n}$ = The matriarch (an elephant having highest fitness value in clan $L_n$.

$rm$ = Random number. $\{rm \in [0,1]\}$

The location of $\alpha^{th}$ elephant having highest fitness value is reformed in every clan using eq. (2). In this, reformation is not performed using eq. (1), i.e., $X_{L_n,\alpha} = X_{optimal,L_n}$

$$X_{next,L_n,\alpha} = \delta \times X_{mid,L_n} \quad (2)$$

Where,

$\delta$ = A term attaining the rein of $X_{mid,L_n}$ on $X_{next,L_n,\alpha}$. $\{\delta \in [0,1]\}$

$X_{mid,L_n}$ = The center value of Clan $L_n$.

**Table 1 : Accuracy Calculation for K-Means, PSO, GWO and SCIE_EOC**

| Dataset | k-Means | PSO | GWO | SCIE_EOC |
|---|---|---|---|---|
| HCV | 79 | 84 | 85 | 92 |
| Blood Transfusion | 80 | 82 | 86 | 90 |
| Thoraic Surgery | 78 | 82 | 80 | 92 |
| HTRU | 77 | 85 | 84 | 92 |
| Diabetic Retinopathy | 79 | 81 | 85 | 93 |
| Cancer | 80 | 86 | 87 | 95 |

The SCIE_EOC algorithm is executed on six separate datasets like HCV, Blood Transfusion, Thoraic Surgery, HTRU, Diabetic Retinopathy and Cancer obtained from UCI repository [18, 20-22]. The data in Table 1 and Fig. 1 reveal that the SCIE_EOC yielded the extreme value of Accuracy for the six comprehensive datasets. According to Accuracy, the EHO exceeds better than GWO, GWO, PSO and K-Means by 12%, 16% and 22% respectively, across all six datasets.

In the proposed technique, only focus is given on accuracy in clustering since the earlier research show the less accuracy results.

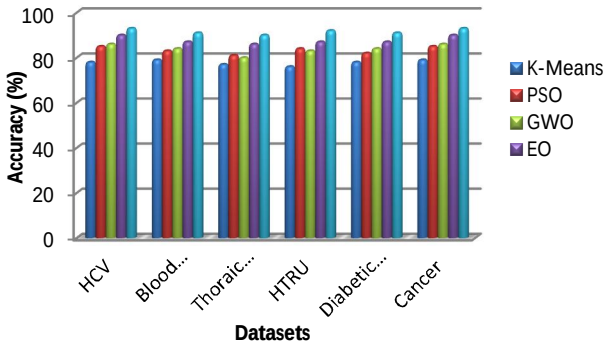


**Fig. 1 : Accuracy Exploration of K-Means, PSO, GWO and SCIE_EOC**

The utilization of information entropy in conjunction with the EO scheme for clustering ensures precise and proficient dispersal of data values within the dataset.

## V. CONCLUSION

Data clustering plays a crucial role across diverse industries, including education, healthcare, and decentralized business environments. Utilizing various classification techniques, different datasets are effectively managed and analyzed to enhance the quality of data sharing in decentralized multidimentional datasets. In this research paper, a Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization based Clustering (SCIE_EOC) is introduced, aiming at efficiently classifying diverse datasets and enhancing data search capabilities. SCIE_EOC is shown great promise in accurately clustering diverse datasets and improving optimization performance. The proposed scheme is predicted to forecast and validate with vast databases within the domain of multi-dimensional data.

The position variable is adopted in EHO, and the moving speed of the elephant is ignored. These issues affect the convergence speed of the algorithm. In the future, research can be carried out to solve problems of clustering such as optimal controller selection in wireless sensor networks, controller placement problems in a software-defined network, clustering in image segmentation by using the proposed approach. The suggested approach can also be extended to solve multidimensional optimization problems.

## REFERENCES

[1] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[2] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, pp. 1–66133, 2004.

[3] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," *Proceedings of the National Academy*

Volume: 15 Issue: 05 Pages: 6122 – 6125 (2023) ISSN: 0975-0290

*of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[4] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, Article ID 066111, 6 pages, 2004.

[5] Y. Ou and C.-Q. Zhang, "A new multi-membership clustering method," *Journal of Industrial and Management Optimization*, vol. 3, no. 4, pp. 619–624, 2007.

[6] X. Qi, K. Christensen, R. Duval et al., "A hierarchical algorithm for clustering extremist web pages," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM '10)*, pp. 458–463, August 2010.

[7] P. Zhao and C. Zhang, "A new clustering method and its application in social networks," *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2109–2118, 2011.

[8] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp. 191–218, 2006.

[9] S. V. Dongen, *Graph clustering by flow simulation [Ph.D. dissertation]*, University of Utrecht, 2000.

[10] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z.Di, "Community detection by signaling on complex networks," *Physical Review E*, vol. 78, no. 1, Article ID 016115, 2008.

[11] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, 2004.

[12] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of SIAM International Conference on Data Mining*, pp. 76–84, 2005.

[13] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, "Detecting communities in large networks," *Physica A*, vol. 352, no. 2-4, pp. 669–676, 2005.

[14] F. Wu and B. A. Huberman, "Finding communities in linear time: a physics approach," *European Physical Journal B*, vol. 38, no. 2, pp. 331–338, 2004.

[15] Z. Shi, Y. Liu, and J. Liang, "PSO-based community detection in complex networks," in *Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling (KAM'09)*, pp. 114–119, December 2009.

[16] C. Shao, W. Lou, and L. Yan, "Optimization of algorithm of similarity measurement in high dimensional data," *Computer Technology and Development*, vol. 20, no. 2, pp. 1–4, 2011.

[17] H. Luo and H.Wei, "Clustering algorithm for mixed data based on clustering ensemble technique," *Computer Science*, vol. 37, no. 11, pp. 234–238, 2010.

[18] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems*, vol. 2096 of *Lecture Notes in Computer Science*, pp. 309–318, 2001.

[19] Belal M. and Daoud A., 2005. A new algorithm for cluster initialization, *World Academy of Science, Engineering and Technology*, Vol. 4, pp. 74-76.

[20] Chao Shi and Chen Lihui, 2005. Feature dimension reduction for microarray data analysis using locally linear embedding, *3rd Asia Pacific Bioinformatics Conference*, pp. 211-217.

[21] Davy Michael and Luz Saturnino, 2007. Dimensionality reduction for active learning with nearest neighbour classifier in text categorization problems, *Sixth International Conference on Machine Learning and Applications*, pp. 292-297.

[22] Deelers S. and Auwatanamongkol S., 2007. Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance, *International Journal of Computer Science*, Vol. 2, No. 4, pp. 247-252.

[23] Kummathi Chenna Reddy, Venkatesh S N, Manjula K B, Boban Mathews, Lifetime Enhancement of Wireless Sensor Networks Using Energy Centric Clustering Algorithm, *International Journal of Advanced Networking and Applications – IJANA*, Vol. 13, No. 5, pp. 5095-5101, 2022

**Authors Biography**

**Mr. PRADEEP KUMAR ATULKAR** earned his M. Phil. and pursing Ph.D. from Rabindranath Tagore University, Bhopal. His research work focuses on Data Mining, Machine Learning and Data Analytics. He has 10 years of teaching experience and 5 years of research experience in the field of Computer Science.

**Dr. RAJENDRA GUPTA** is an Associate Professor in Department of Computer Science at Rabindranath Tagore University, Bhopal, India having Doctoral degree in Computer Science. He is awarded as Distinguished Research Professional Award, Best Promising Trainer Award, Best Promising Facilitator Award by recognized agencies. His teaching and research areas belong to Data Mining, Machine Learning, Data Mining, Statistical Analysis etc. He has 4 patents and 88 research papers in International & National Journals of reputed.