Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 465–477
DOI : 10.47974/JSMS-1282

Mathematical enhancement of unsupervised machine learning algorithms for optimal lifeline donor knowledge extraction

P. Ramesh Naidu [@]
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore
Karnataka
India

Pratik Gite [†]
Department of Computer Engineering
St. John College of Engineering and Management
Palghar (East)
Palghar (Mumbai)
Maharashtra
India

Vaskar Deka [§]
Department of Information Technology
Gauhati University
Guwahati
Assam
India

K. D. V. Prasad [‡]
Department of Research
Symbiosis Institute of Business Management
Hyderabad
Telangana
India

and

[@] *E-mail:* ramesh.naidu@nmit.ac.in
[†] *E-mail:* pratikgite135@gmail.com
[§] *E-mail:* vaskardeka@gauhati.ac.in
[‡] *E-mail:* kdv.prasad@sibmhyd.edu.in

Department of Research
Symbiosis International (Deemed University)
Pune
Maharashtra
India

V. Dankan Gowda *
Department of Electronics and Communication Engineering
BMS Institute of Technology and Management
Bangalore
Karnataka
India

Mirzanur Rahman #
Department of Information Technology
Gauhati University
Guwahati
Assam
India

Abstract

Lifeline Donor groups are crucial in the field of humanitarian aid because they provide lifesaving aid in times of crisis. For effective resource management and aid coordination, it is crucial to draw relevant insights from donor data in a timely manner. Scalability, precision, and flexibility in responding to changing donor dynamics are all areas where traditional techniques of knowledge extraction fall short. This study offers a fresh method for optimizing knowledge extraction from Lifeline Donor databases by utilizing improved unsupervised machine learning methods. In order to extract, categorize, and prioritize donor knowledge from disparate data sources, this research presents a multi-faceted framework that includes cutting-edge machine learning algorithms. By merging cutting-edge developments in NLP and data mining, the proposed framework improves upon classic clustering and topic modelling methods. The program is able to capture subtle semantic links and underlying patterns in donor communication data by using methods like word embedding models and graph-based clustering.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Lifeline donor, Knowledge extraction, Unsupervised machine learning, Humanitarian assistance, Data mining and decision-making.*

* *E-mail:* **dankan.v@bmsit.in (Corresponding Author)**
# *E-mail:* **mr@gauhati.ac.in**

1. Introduction

Lifeline is a term used in the field of humanitarian aid. In times of disaster, the assistance provided by donor groups is critical. These groups provide crucial aid in the form of resources, funds, and expertise to help communities recover from disasters and other catastrophes [1]. Effective resource allocation, coordination of aid operations, and strategic decision-making depend on timely and accurate knowledge extraction from donor data. But conventional approaches to gleaning insights from donor interactions and communications have their limits, particularly when it comes to scale, precision, and flexibility[2]. Novel approaches are needed to extract useful insights that could guide humanitarian action in light of the ever-changing nature of crisis circumstances and the ever-increasing amount and complexity of donor data[3]. The field of computer science known as "data mining," which entails extracting valuable information from massive databases, is young but expanding quickly. One kind of data mining task involves making predictions about future attribute values based on past attribute values, while the other type involves describing existing patterns in the data[4]. The next step is data transformation, which involves reformatting the information so that it can be used by a computer. Knowledge discovery in databases (KDD) is a methodical procedure for gaining insight from massive data stores. As can be seen in Figure 1, it is a multi-stage process, the heart of which is data mining.

The following are the standard steps involved in KDD: Based on the nature of the problem and the objectives of the analysis, the appropriate data is chosen from the available sources[5]. The selected data is pre-processed by being cleaned, converted, and otherwise prepared to deal with problems like missing numbers and outliers. This study sets out to solve these problems by presenting a new, improved unsupervised

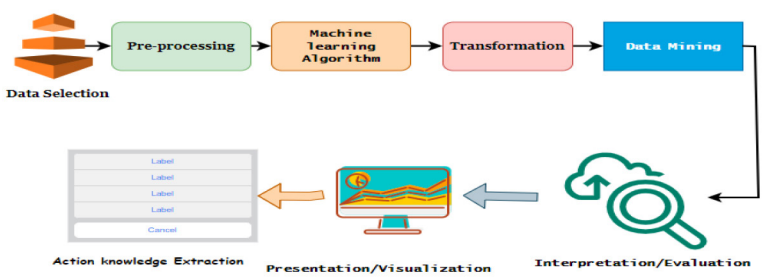

Figure 1

Process at Knowledge Discovery in Databases

machine learning approach with which to extract information on Lifeline Donors[6]. The goal of this algorithm is to provide Lifeline Donor organizations with a more accurate and flexible solution by utilizing recent developments in natural language processing and data mining[7]. Through the use of methods such as word embeddings and graph-based clustering, the system is able to reveal previously hidden patterns and semantic correlations in donor communication data.

2. Literature Survey

Researchers have been inspired to investigate novel approaches to knowledge extraction and decision assistance in crisis circumstances as a result of the development of humanitarian aid and the function of Lifeline Donor organizations. In this context, improved unsupervised machine learning algorithms have become increasingly important, and the following literature review will highlight some of the most important advances in this area[8]. There is an urgent need for a powerful analytical solution that can mine the wealth of information stored in an organization's databases and archives for actionable insights[9]. information Discovery in Databases (KDD) has emerged in response to this need, playing a crucial role in transforming raw data into refined information to aid in making well-informed decisions. Knowledge discovery in databases is a discipline that makes use of a methodical, iterative series of procedures that together form its operational operations[10]. The primary focus of our research was an evaluation of several data mining methods and how well they performed on blood donor data. This study elucidated, within the context of a given dataset[11], which classification methods perform best in terms of accuracy and error rates. For this purpose, we used a variety of data mining algorithms on five separate, authentic datasets that we obtained from the UCI Machine Learning Repository[12]. Our research builds on these prior investigations to fill in the gaps that have been observed in the existing literature. To improve knowledge extraction for Lifeline Donors, we propose a novel unsupervised machine learning algorithm that integrates natural language processing (NLP), graph-based clustering, and adaptive learning. In the face of ever-changing humanitarian needs, this novel method provides Lifeline Donor organizations with a potent instrument for data-driven decision-making and resource allocation.

3. Data Mining

Data mining, as we've established, is the process of uncovering hidden patterns and correlations inside databases that leads up to the creation of new information. A multitude of potentially game-changing hidden patterns sleeps within the depths of corporations' databases[13]. Only data mining can sift through these records and find these priceless regularities. Data mining makes it possible to examine such large databases and extract useful insights that would otherwise be impossible to obtain. Data mining is crucial for businesses because it helps them gain strategic insights from previously untapped data[14]. A large percentage of businesses use data mining to guide them through the full customer lifecycle, from initial contact through ongoing relationship building, revenue growth, and customer retention. Companies can better prepare for the actions of current and potential consumers by defining those customers' qualities based on past data[15]. Data mining is a field that depends heavily on the success of other data processing and statistical initiatives. In the sections that follow, we'll take a quick look at some of the most significant of these initiatives that complement data mining.

4. Clustering Algorithms

A cluster is a division of items chosen to minimize a given measure of dissimilarity, and clustering is a method for classifying related objects. It's a subfield of unsupervised learning by itself. Clustering is a way to arrange data sets by identifying groups of items that are quite different from one another yet contain many items that are very similar to one another[16]. The cost of performing such studies becomes prohibitive when applied to large databases. To solve this problem, we present a flexible framework for iterative clustering algorithms that may scale to large datasets. This architecture simply requires a single database scan, saving valuable system resources. K-Means, a popular clustering method[17], is used to concretely implement and verify the framework. The key to the approach is figuring out which parts of the data may be compressed, which need to be kept in memory, and which can be thrown away. This algorithmic method strikes a balance between computational performance and data volume while working within the limits of a limited memory buffer.

5. Clustering Methods

Partitioning methods: The data is clustered into k groups according to the following requirements for this classification procedure: There must be at least one object in each cluster, and each object can only be in one cluster[18]. A partitioning technique kickstarts the procedure by generating the first partitioning based on a given value k that specifies the required number of clusters[19]. The goal of this evolving procedure is to achieve a clustering configuration that more accurately represents the underlying relationships and properties of the data.

Hierarchical methods: A hierarchical approach will methodically create a hierarchy from a set of data elements. Depending on how the hierarchy is initially established, this method might be labeled as either agglomerative or divisivep[20]. The core idea behind the divisive method is to break up the original coherent cluster into smaller, more manageable subclusters.

Grid-based methods: Subdividing the object space into a finite array of cells to form a structured grid is at the heart of grid-based approaches. All clustering processes occur inside this grid framework. Computing efficiency in handling huge datasets is provided by STING and related methods by discretizing the data space into discrete cells and then executing clustering operations inside this grid framework. This method allows users to take control of the clustering procedure and produce results that better fit their needs or preferences.

6. Cluster Analysis

Clustering algorithms that run entirely in main memory often make use of two fundamental data structures: the data matrix and the dissimilarity matrix. Object-by-variable data matrices: The essence of n objects, each of which is described by p variables (sometimes called attributes or measurements), is captured by this structure. The rows represent items, while the columns represent variables or characteristics of those objects. The dissimilarity matrix (also known as an object-by-object structure) is a data structure that keeps track of the similarities and differences between every pair of objects in a set of n. An nxn matrix is used to depict this structure. Since the dissimilarity between an object and itself is zero, it is common practice to set the diagonal elements where i = j to 0. The off-diagonal elements represent the measured differences or dissimilarities between objects i and j.

For example, in a dissimilarity matrix:

0	d(2, 1)	d(3, 1)	d(3, 2)	...
d(2, 1)	0	d(3, 1)	d(3, 2)	...
d(3, 1)	d(3, 1)	0	...	...
d(3, 2)	d(3, 2)	...	0	...
...	...	...	...	...

The distance between i and j in this case is denoted by d(i, j). To help clustering algorithms efficiently find and group similar objects, this dissimilarity matrix summarizes the relationships or differences between all pairs of objects in the dataset. Variables can be either interval-scaled, binary-scaled, categorical-scaled, ordinal-scaled, or ratio-scaled, allowing for the quantification of dissimilarity across a wide range of object kinds. This can be done in the case of a measured variable f in the following way: Determine the standard error of the mean (sf): This is less affected by extreme values than the standard deviation (f). Although more robust measures such as the median absolute deviation are available, using the mean absolute deviation keeps the benefit that outlier z-scores don't get too small. Standardization's usefulness is situation-dependent. The most widely used distance measure is the Euclidean distance, which is defined as follows: For two points in an n dimensional space, $p = (p_1, p_2, p_3, ..., p_n)$ and $q = (q_1, q_2, q_3, ..., q_n)$, the Euclidean distance ($d_{\text{Euclidean}}$) between these points is given by:

$$d_{\text{Euclidean}}(p,q) = \sqrt{\{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2\}} \quad (1)$$

The formula has n variables (dimensions) for the number of dimensions in the data space. In n-dimensional space, the Euclidean distance between two points is the shortest possible path between them. It's often used to quantify the degree to which two objects differ from one another, with smaller values indicating closer similarity and larger values indicating greater dissimilarity. Identity of Indiscernible: The distance between a point and itself is zero. For both Euclidean distance and Manhattan distance $d(p,q) = 0$. Symmetry: The distance between two points should be the same regardless of their order. For both Euclidean distance and Manhattan distance $d(p,q) = d(q,p)$. Triangle Inequality: The criteria are met by both the Euclidean distance and the Manhattan distance, making them reliable and popular distance metrics for usage in a wide range of mathematical and computer applications, such as clustering algorithms and distance-based data analysis. The Minkowski distance is a

versatile distance metric that encompasses both the Euclidean distance and the Manhattan distance as special cases. It is defined as follows: For two points $p = (p_1, p_2, p_3, ..., p_n)$ and $q = (q_1, q_2, q_3, ..., q_n)$, in an n-dimensional space Minkowski distance ($d_{Minkowski}$) of order r between these points is given by: $d_{Minkowski}(p,q) = (\Sigma_{i=1}^{n} |p_i - q_i|^r)^{1/r}$. The formula uses the Minkowski distance order, denoted by r. Specifically, the Minkowski distance becomes the Euclidean distance at r=2, and the Manhattan distance at r=1. With the Minkowski distance, the "shape" of the distance metric may be adjusted precisely for any value of r. For bigger variations in coordinates (r>2), the distance metric becomes more sensitive, while for smaller differences (r<2), it becomes less so. It's a versatile distance measure that may be tweaked to fit the needs of your data and analysis. If distinct variables in a dataset are given varying degrees of significance, the notion of weighted Euclidean distance may be utilized to apply relative weights. The formula for the weighted Euclidean distance looks like this: The weighted Euclidean distance (dweighted Euclidean) between two points $p = (p_1, p_2, p_3, ..., p_n)$ and $q = (q_1, q_2, q_3, ..., q_n)$, in an n-dimensional space with weight factors w1, w2,..., wn allocated to each dimension is defined as follows.

$$d_{\text{weighted Euclidean}}(p,q) = \sqrt{\{w_1(p_1 - q_1)^2 + w_2(p_2 - q_2)^2 + \ldots + w_n(p_n - q_n)^2\}} \qquad (2)$$

To simplify, we may express this as where n is the total number of variables (dimensions) in the data space and w_i is the weight assigned to the i^{th} variable. By multiplying the square of the difference in each dimension's value between the two places, the weighted Euclidean distance takes into account the relative importance of each dimension.

7. Cluster Validation

A key issue in cluster analysis, cluster validity asks if the proposed cluster structure truly captures the essence of the data being clustered. However, there are many reasons why this is a difficult task: Clustering falls under the umbrella of unsupervised learning, which is used when the correct labels for the data are not accessible. Therefore, it becomes challenging to calculate accurate performance measures. These problems are addressed by using cluster validity indices. There are three broad groups into which these indexes fall: Internal indices measure how closely a clustering scheme corresponds to the underlying data. They offer help in counting clusters without any further data. Validity measures for clustering help researchers determine which algorithms, parameters, and number of

clusters will produce the best results. Due to the complexity of clustering analysis, however, the problem of assessing cluster validity has not yet been resolved.

8. Results and Discussion

In this section, we discuss the technical details of putting into practice our suggested system. In order to implement our concept, we must first retrieve the dataset from a specified file. A sequence of processes is carried out in response to user input, with parameters such as Blood Group and chosen location serving as guides. The main goal is to form groups according to how these elements interact with one another. Clusters are formed based on location and blood group, allowing for the identification of persons who are physically close to potential receivers.

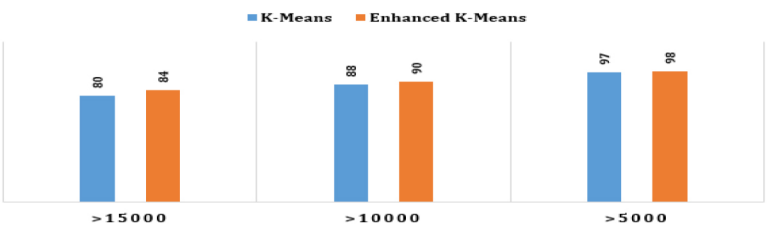

Figure 2

Performance Comparison (accuracy)

The Enhanced K-means algorithm, in comparison, takes a more deliberate strategy by determining initial centroids with appropriate formulas. We compare the experimental clusters to a dataset of blood donors who fit the criteria of blood group and location to evaluate the reliability of the clustering results. We give a thorough overview of the performance of the algorithms by quantifying the % accuracy of these clusters and analyzing the time spent for each trial. Figure 2 depicts the results of this investigation, allowing a clear depiction of the algorithms' accuracy and efficiency across different settings. Figure.2 is illuminating since it shows how the K-means and proposed Enhanced K-means algorithms compare in terms of accuracy when used on datasets with progressively more records. This improvement, which was accomplished with the help of mathematical formulas, greatly boosts the functionality of the system. When time is of the essence and it's necessary to locate possible blood donors who have the necessary blood group, this enhancement is

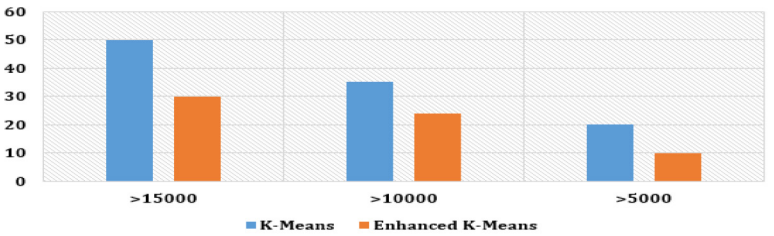

Figure 3

Performance Comparison (Time nano seconds)

invaluable. Figure 3 graphically depicts the vast difference between the K-means and Enhanced K-means algorithms in the amount of time needed to build clusters. As a result, the system improves its ability to locate nearby donors who possess the necessary blood group.

Because of this effectiveness, more potential donors can be contacted and asked to donate blood in a shorter length of time, which is of critical importance. The primary benefit of the Enhanced K-means algorithm is its capacity to maximize cluster formation, which in turn expedites and improves donor identification.

This methodical technique optimizes the blood donation process by promptly matching donors with the required blood group and proximity, ensuring the system can efficiently cater to the demands of recipients. The data used in the experiment is completely artificial and was created for the sole purpose of providing verification. A dataset consisting of 1500 records was chosen at random for this study. The obtained results (Figure.4) demonstrate the Active Cluster with Modified K-means method's superiority over the traditional K-Means approach in terms of cluster quality and computational efficiency. The inter-cluster distance versus

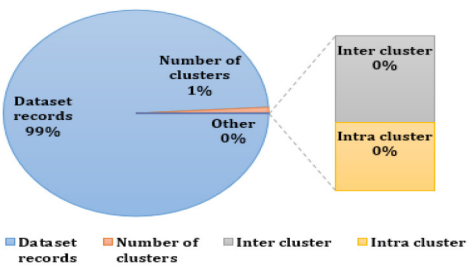

Figure 4

Comparison of Results for 1500 dataset records

intra-cluster distance balance is the crux of cluster quality. Specifically, the Modified K-means algorithm demonstrated superior performance in preserving cluster quality. The effectiveness of the method can be traced back to its capacity to reduce inter-cluster distances, guaranteeing the formation of reliable groups. The Modified K-means method successfully managed the distances between data points within and between clusters, demonstrating its ability to produce clusters that are both internally cohesive and outwardly distinguishable. The benefits of using the Active Cluster with Modified K-means algorithm are highlighted by the experiment's final results. Improved accuracy and efficacy of clustering outcomes may have far-reaching ramifications for data-driven applications because to its capacity to maximize cluster quality while retaining computing efficiency.

9. Conclusion

This study presents a new K-means clustering technique that makes use of the link between blood types and geographic regions. If you need to find a list of people who have the necessary blood group and where they are located, the proposed Enhanced K-means algorithm is a useful tool. In particular, it improves computing efficiency, which is a major issue with the classic K-means clustering approach. There is great promise for this method in modern settings, as it has the ability to save many lives by facilitating the effective matching of blood donors and patients in need. Predefining the number of clusters is an additional major difficulty with the K-means clustering technique. In practice, this is difficult since too few clusters could result in objects of different types being lumped together, and too many clusters could result in objects of the same type being separated. The Active Cluster with Modified K-means algorithm gets around this problem by calculating the right amount of clusters on the fly. This flexibility guarantees that the algorithm accurately reflects the data's underlying structure without relying on predetermined, potentially arbitrary cluster sizes. In conclusion, the suggested approach not only provides a novel solution to the problem of dynamic cluster determination, but it also meets the practical demand of effectively finding potential blood donors. This exemplifies the power of cutting-edge algorithms in the field of data-driven research to solve pressing societal problems.

References

[1] Zubair, M., Iqbal, M.A., Shil, A. et al. An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. Ann. Data. Sci. (2022). https://doi.org/10.1007/s40745-022-00428-2

[2] G. Feng, M. Fan and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," in *IEEE Access*, vol. 10, pp. 19558-19571 (2022), doi: 10.1109/ACCESS.2022.3151652.

[3] L. Dolder and B. De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," 2016 SAI Computing Conference (SAI), London, UK, pp. 523-531 (2016), doi: 10.1109/SAI.2016.7556030.

[4] O. Baker, J. Liu, M. Gosai and S. Sitoula, "Twitter Sentiment Analysis using Machine Learning Algorithms for COVID-19 Outbreak in New Zealand," 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), Shah Alam, Malaysia, pp. 286-291 (2021), doi: 10.1109/ICSET53708.2021.9612431.

[5] K. Rameshkumar, M. Sambath and S. Ravi, "Relevant association rule mining from medical dataset using new irrelevant rule elimination technique," 2013 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, pp. 300-304 (2013), doi: 10.1109/ICICES.2013.6508351.

[6] Z. Ge, Z. Song, S. X. Ding and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," in *IEEE Access*, vol. 5, pp. 20590-20616 (2017), doi: 10.1109/ACCESS.2017.2756872.

[7] Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science 2(5) : 1–22 (2021).

[8] W. Zhu and M. Zahid, "Integration of data mining clustering approach in the personalized E-learning system", *IEEE Access*, vol. 6, pp. 72724-72734 (2018).

[9] Ravindra, R.; Rathod, R.D.G. Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data. *Int. J. Energ. Sect. Manag. 2*, 295–310 (2017).

[10] C. Angeli, S. K. Howard, J. Ma, J. Yang and P. A. Kirschner, "Data mining in educational technology classroom research: Can it make a contribution?", *Comput. Educ.*, vol. 113, pp. 226-242, Oct. (2017).

[11] Pham DT, Dimov SS, Nguyen CD. An incremental k-means algorithm. Proceedings of the Institution of Mechanical Engineers, Part C: *Journal of Mechanical Engineering Science* 218(7) : 783–795 (2004).

[12] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey", *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 1, May (2020).

[13] T. Wang, K. Lu, K. P. Chow and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model", *IEEE Access* (2020).

[14] Sandeep Dwarkanath. Development in intelligent autonomous agents for the high-level cognitive functions like reasoning, planning, learning and abstraction, *Journal of Interdisciplinary Mathematics*, 26:3, 301-310 (2023), DOI: 10.47974/JIM-1661.

[15] Santosh Kumar & Gupta, Anand Kumar. A novel approach of unsupervised feature selection using iterative shrinking and expansion algorithm, *Journal of Interdisciplinary Mathematics,* 26:3, 519-530 (2023). DOI: 10.47974/JIM-1678.

[16] Naziya & Chinamuttevi, Veera Sivakumar. A novel RF-SMOTE model to enhance the definite apprehensions for IoT security attacks, *Journal of Discrete Mathematical Sciences and Cryptography*, 26:3, 861-873 (2023), DOI: 10.47974/JDMSC-1766.

[17] Selvakumar, C. , Shekhar, R. & Chaturvedi, Abhay. Securing networked image transmission using public-key cryptography and identity authentication, *Journal of Discrete Mathematical Sciences and Cryptography*, 26:3, 779-791 (2023), DOI: 10.47974/JDMSC-1754.

[18] H. G. Govardhana Reddy & K. Raghavendra. Vector space modelling-based intelligent binary image encryption for secure communication, *Journal of Discrete Mathematical Sciences and Cryptography*, 25:4, 1157-1171 (2022), DOI: 10.1080/09720529.2022.2075090.

[19] Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*, 226-235 (2019). https://doi.org/10.3390/j2020016.

[20] sulaiman, Barah M., Hassan, Basim A. & Sulaiman, Ranen M. A new secant method for minima one variable problems, *Journal of Interdisciplinary Mathematics*, 26:6, 1015–1021 (2023), DOI: 10.47974/JIM-1601.