Grenze International Journal of Engineering and Technology, Special Issue

Optimization Algorithms and the Clustering Problem

Nadia Abd-Alsabour
Cairo University, Cairo, Egypt
Email: nadia.abdalsabour@cu.edu.eg

***Abstract*—Optimization algorithms can play significant roles in enhancing and even changing the performance of clustering algorithms since the clustering problem is an optimization problem. This paper studies this point through an example of the k-means algorithm as it is the most widely utilized clustering method but it sometimes gets stuck at sub-optimal solutions. Meanwhile, optimization algorithms can avoid local optima besides their other merits. Therefore, many researchers have integrated several optimization algorithms with the clustering algorithms aiming at enhancing their performance and hence getting better clustering results. This work studies the utilization of genetic algorithms (as an example of optimization algorithms) besides the k-means algorithm for handling the clustering problem. It also demonstrates the clustering problem, its applications, etc.**

***Index Terms*— Clustering, optimization algorithms, the k-means, genetic algorithms.**

I. INTRODUCTION

Clustering is a fundamental data analysis procedure in a variety of fields and it provides a crucial role in a variety of their applications such as industry, astronomy, etc. [1]. For instance, assume that an organization needs to organize all of its clients into several clusters such that these groups can be allocated to various managers and the clients in each group are as similar as could be allowed. In addition, 2 given clients with different characteristics should not be put in the same cluster so as to create client relationship campaigns that particularly focus on each group in view of common features shared by the clients per group. Clustering is the method that achieves this task depending on the similarity and distance. It allows generating groups where each one comprises of items having common attributes. It aims to partition a dataset into clusters such that each one is similar within itself yet is as dissimilar as possible to other clusters. This is achieved through maximizing and minimizing the between-groups and the within-group distances respectively as shown in next figure. Some approaches require the number of clusters is determined as well as a seed to begin each cluster. Then, the items are clustered in view of self-similarity.

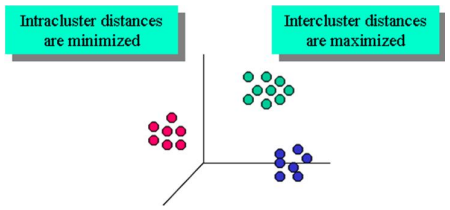


Figure 1. Intra-cluster and Inter-cluster

Grenze ID: 01.GIJET.4.3.35

It can be defined as the grouping of a collection of items to various groups such that items inside each one will be similar but are very dissimilar to those within the various groups. Based on the features' values characterizing the items, similarities and dissimilarities are assessed. It can also be defined as partitioning a collection of items into subsets such that items inside each one will be similar to each other and not similar to those in the various subsets. In this context, diverse clustering techniques may create distinctive clusters on a given dataset. It can find out unknown clusters in the given dataset [2]-[3].

It has various names as automatic classification. A crucial contrast between classification and clustering is that the latter can discover automatically the clusters which is considered its particular preferred standpoint. It is additionally known as segmentation by marketing specialists since it segments huge datasets into clusters. Besides, in machine learning, it is called unsupervised learning since the class name is not given. Therefore, it is a type of learning by observation. Moreover, it is sometimes called sorting by psychologists [4].

It is utilized for knowledge discovery since it gives an insight into the natural groupings within the given data as opposed to prediction. It has been widely utilized in numerous applications. It can be utilized to arrange a huge number of clients into clusters where they share common attributes within a group. This encourages creating business procedures for improved client relationship management. Also, it can be utilized to enhance projects' management since it can divide them into categories in view of similarity so that projects diagnosis and auditing (to enhance project outcomes and delivery) can be directed successfully. Moreover, it has numerous applications in the web search such as:

- Organizing documents into topics that is widely utilized in information retrieval.
- Clustering the search results (are often a huge number of pages relevant to the search) into clusters and displaying them easily and concisely.

Because of its flexibility and simplicity, the k-means iterative clustering algorithm has been recognized as one of the main ten data mining methods and it shows promising performance. It has been utilized for tackling different real-life problems like image processing, bioinformatics, and text mining [5].

It is essentially an optimization algorithm since it goes for minimizing the within cluster square error. It allocates membership to items by measuring the distance between the centroid of the assigned group and every pair of items. This process will be progressively refined until the best possible assignment is yielded—that is when the total inter-distances of the items across various clusters are maximized and the total intra-distances of the items inside a group are minimized. However, the quality of the results to a great extent depends on the initial centroids that are randomly created each time the clustering kick- begins that vary from time to time. Hence, the k-means can most likely plunge into local optima.

Although the significant advances that have been introduced recently in the clustering field (such as spectral clustering, subspace clustering, mixture models, etc.), no single method has succeeded to supplant the k-means algorithm and its versions in spite of its recognized disadvantages. This is a result of the troubles related with these new methods like large computational cost and extra parameters. Therefore, researchers compelled to utilize numerous diverse clustering algorithms in order to tackle difficult clustering applications such as those with large datasets and high dimensional ones [1], [6].

There is a variety of methodologies that combines the capabilities of k-means in partitioning data with capabilities of optimization techniques in executing the search to search for the closest optimal solutions [7]. Generally speaking in the past few years, various clustering methods related to genetic algorithms have been proposed [7].

This paper examines the roles optimization algorithms can provide for the clustering problem such as incorporating the genetic algorithms into the k-means for deciding the value of the initial centers since deciding on them is a crucial issue and requires detailed knowledge about the domain besides experience in data analysis as well. It additionally demonstrates the clustering as an optimization problem.

The rest of this paper is as follows. The following section introduces the clustering as an optimization problem. Section 3 demonstrates the k-means. Section 4 displays various approaches for integrating genetic algorithms with the k-means. The last section provides the conclusions and potential future work.

II. Clustering as an optimization problem

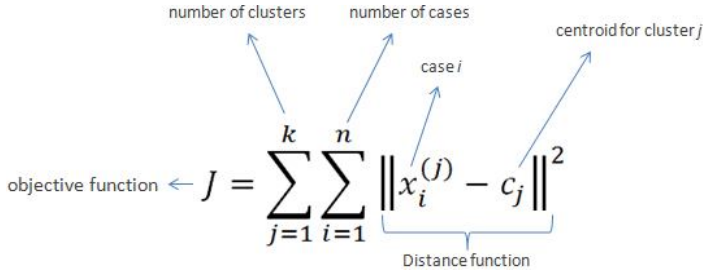
It is possible (but intractable) to accomplish a globally optimum clustering result by means of a brute-force approach by thoroughly trying all dividing possibilities. As the number of items and clusters increment, the combinatorial number of possible grouping arrangements escalates which prompts computationally restrictive. Consequently, heuristic methods are wanted for searching the global optima stochastically and enhancing the quality of the final results iteration by iteration. Genetic algorithms are ideal candidates of

these methods [8]. Recently, other recent meta-heuristics have been utilized successfully for tackling the clustering problem. Examples are artificial bee colony, and differential evolution. Moreover, the whale optimization method (one of the latest metaheuristics that was proposed on the basis of the foraging performance of the swarm of humpback whales) was utilized for tackling the clustering problem [9].
Since clustering aims to discover the best division of the given observations, it is essentially an optimization problem (maximizes or minimizes the condition we are searching for) and it is realized that discovering the ideal clustering solution is NP-hard [3]. This gives way to the use of optimization algorithms either for performing the clustering such as ant algorithms and particle swarm optimization [10]-[12] or to be used for optimizing the performance of clustering algorithms.
By having an objective function in the clustering problem, optimization algorithms can simply play significant roles not only by optimizing these functions but also by changing the behavior (performance) of their clustering algorithms. For instance, Shah and Koltun [1] made their clustering algorithm gets high accuracy when handling high dimensional datasets with a few modifications when handling various domains. This is via optimizing a smooth continuous function which allowed heavily mixed clusters to be untangled. Their contributions made their clustering algorithm do both dimensionality reduction besides performing the clustering.
The problem can be handled using integer programming but since having a substantial number of factors is time consuming, heuristic techniques (such as the k-means algorithm) that provide generally good solutions are often utilized [13].

III. The K-Means Method

The term K-means was first utilized in 1967 by MacQueens. It is the most widely utilized clustering method based on partitioning data and has been utilized in numerous fields. It is a quite popular clustering technique because of its effectiveness and ability. Hence, it has attracted tremendous interests for a long time. It is a simple partitioning method that discovers *k* non overlapping groups [7].
However, its reliance on the initial data and its high computational cost that influences the convergence time have restricted its applications in expansive areas [5], [13]. This will be explained later.
It is being described by the simplicity of its implementation as follows. It involves determining the number of clusters (trial and error might be utilized and then choosing in light of between-clusters and within-cluster distances) and then partitioning the given *n* items. It involves additionally selecting *k* seeds randomly as beginning samples of these groups.

- Once the seeds have been determined, every item is allocated to a cluster nearest to it in view of some distance measure (similarity is computed using the mean values).
- Once all the items have been allocated, the mean of each group is calculated and become the new seeds.
- All the members are now re-allotted to the groups utilizing the mean value of each group. In most situations, some items will change groups unless the first guesses of seeds were great.
- The algorithm works iteratively to allocate each item to one of *k* groups in light of the given attributes as Equation 1 illustrates. In Figure 2, *k*=2, and there are 2 groups set from the given data.



Equation (1)

- This process proceeds until no changes occur to groups' memberships [2]-[3], [13]-[17]. This implies it has turned out to be steady and there is no requirement for any iteration anymore [7].

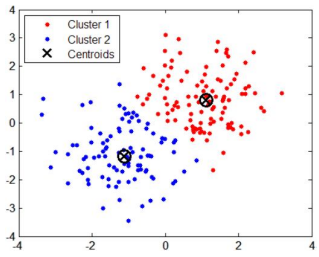


Figure 2. The k-means with k=2

IV. Combining the K-means with Genetic Algorithms for Handling the Clustering Problem

Recently, extensive research has been performed on integrating the k-means with genetic algorithms in different ways for tackling the clustering problem in order to improve its performance as follows.

Although it is one of the most well known and the most established clustering methods and is efficient computationally, it still has several basic shortcomings that absolutely depend on the initial centers as a bad selection to them may cause a local optimum and creating empty groups [13], [18]. An example of the use of genetic algorithms for handling this demerit is the work of Chittu and Sumathi [19]. They utilized a GA to find the value of k, where GA is first performed so as to provide the initial values to the k-means algorithm and this minimizes the quantity of cycles the k-means requires. Also, Al-Shboul, and Myaeng [20] proposed 2 methods to tackle the initialization problem; the k-means method to initialize a genetic algorithm and a genetic algorithm initializes the k-means method. In addition, Jahanbin et al. [21] proposed a hybrid approach of a particle swarm optimization and a genetic algorithm method to tackle this weakness (random begin and subsequently falling into the local optimum).

It is also sensitive to the outliers as they may distort significantly the items' locations [18] - [19]. Another way for using the genetic algorithm to enhance the k-means is to work around this demerit such as the recent work of Gupta et al. [22]. They handled the problem of its performance with outliers (as with real-world data) since it presumes all the patterns can be divided into a number of distinct groups which is often intractable because of the outliers to which the k-means is extremely sensitive. The authors introduced a simple local search-based method for it [22].

It is applicable only when the mean can be computed i.e., not in the case of categorical data [19]. Combining the k-means with the genetic algorithms can be beneficial when dealing with datasets with mixture of categorical and numerical attributes that is expected to produce better results in addition to optimizing the cluster centers [13]. The work of Saharan et al. [23] was inspired by the lack of clustering methods that particularly centre around binary data. Therefore, the authors combined a genetic algorithm with the incremental k-means method in order to achieve this objective.

It may be easily caught in a local minimum besides being time consuming when handling huge datasets [23]. It may only discover linearly separable groups [24]. To decrease the computational time and enhance the estimation accuracy for nonlinear optimization, Min-Yuan and Kuo-Yu [25] integrated chaos genetic method and the k-means where the initial population was created via chaos mapping then enhanced through competition. In every cycle, besides the evolution of genetic method, the k-means algorithm is executed so as to accomplish speedier convergence. Bouhmala et al. [26] employed a genetic algorithm with the k-means greedy method. This is to utilize the genetic search approach for producing new groups utilizing the two-point crossover and execute then the k-means to enhance the quality of the generated clusters further in order to quicken the search process. Haraty et al. [27] proposed a k-means method on large datasets utilizing a greedy approach to generate the preliminary centroids and then takes k or lesser passes over the dataset in order to adapt these centre items. Arkeman et al. [28] utilized a multi-objective genetic algorithm with Pareto rank method to enhance the k-means performance by maximizing the variants between the clusters and limiting variations in every group as well.

V. Conclusions and Future Work

The main drawbacks of the clustering algorithms can be overcome utilizing the use of the optimization algorithms as most of these drawbacks are basically optimization problems. For instance the main drawback of the k-means method; determining the initial centers; The initial beginning points often prompt results reaching the local optima that may be enhanced by computing more than an iteration which leads to getting

better results. While for accomplishing this, it is hard to decide the computation limit, the characteristics of modern PCs can overcome this. Moreover, randomly generating the initial clusters does not prompt unique solution. The result is sensitive to the initial seeds. This is called the cluster initialization problem. Hence, it experiences the local minima issue. Therefore and as shown in the previous section, this demerit (the sensitivity to the initial centroids) can be overcome utilizing optimization algorithms as this drawback is basically an optimization problem. This clarified the widely utilization of the genetic algorithms (as they are one of the well-established optimization methods) for handling this issue due to their numerous merits which encourages a variety of researchers in various fields to use them as shown in the previous section.

However, the traditional genetic algorithms may experience the premature convergence when handling some optimization problems. This urges specialists to upgrade them more through integrating them with local search methods and/or via the utilization of parallel versions that have not been investigated much with the k-mean algorithm. A good example is the work of Abd-Alsabour [29]. The author utilized the parallel approaches for the knapsack problem which can be utilized simply for optimizing several aspects of the clustering algorithms in general and in particular the k-means algorithm.

Besides, the efforts of producers of software tools to introduce methods for handling the initial centers as happened with the R language group that introduced a method for achieving that which encourages increasing the use of the k-means. These efforts because this drawback is a fundamental difficulty in clustering in general and in the k-means algorithm in particular as it affects the performance of the used clustering algorithm and because of the deserved advantages of the k-means algorithm.

Other clustering methods should be investigated in future. Besides, more metaheuristics other than the genetic algorithms can also be utilized in the future work.

REFERENCES

[1] Shah SA, Koltun V. Robust continuous clustering. Proceedings of the National Academy of Sciences. 2017 Sep 12; 114(37): 9814-9819.

[2] Han J. Kamber M. and Pei J. Data Mining, Concepts and Techniques, 3rd Edition, Morgan Kaufmann, 2012.

[3] Exploring K-Means clustering analysis in R, available at: https://en.proft.me/2016/06/18/exploring-k-means-clustering-analysis-r/. Last visited on: 30-7-2018.

[4] Äyrämö S, Kärkkäinen T. Introduction to partitioning-based clustering methods with a robust example. Reports of the Department of Mathematical Information Technology. Series C, Software engineering and computational intelligence. 2006(1/2006).

[5] Liu W, Shen X, Tsang I. Sparse Embedded Means Clustering. In Advances in Neural Information Processing Systems 2017 (pp. 3319-3327).

[6] Jain AK. Data clustering: 50 years beyond K-means. Pattern recognition letters. 2010 Jun 1; 31(8):651-666.

[7] Zeebaree DQ, Haron H, Abdulazeez AM, Zeebaree SR. Combination of K-means clustering with Genetic Algorithm: A review. International Journal of Applied Engineering Research. 2017; 12(24):14238-14245.

[8] Fong S, Deb S, Yang XS, Zhuang Y. Towards enhancement of performance of K-means clustering using nature-inspired optimization algorithms. The Scientific World Journal. 2014 August 18; 2014: 564829.

[9] Nasiri J, Khiyabani FM. A Whale Optimization Algorithm (WOA) approach for Clustering. Cogent Mathematics & Statistics. 2018 Jun 5, 5:1483565.

[10] Subekti R, Sari ER, Kusumawati R. Ant colony algorithm for clustering in portfolio optimization. Journal of Physics: Conference Series 2018 Mar (Vol. 983, No. 1, p. 012096). IOP Publishing.

[11] Menéndez HD, Otero FE, Camacho D. Medoid-based clustering using ant colony optimization. Swarm Intelligence. 2016 Jun 1; 10(2):123-145.

[12] Gong C, Chen H, He W, Zhang Z. Improved multi-objective clustering algorithm using particle swarm optimization. PloS one. 2017 Dec 5; 12(12):e0188815.

[13] Khotimah BK, Irhamni F, Sundarwati T. A Genetic Algorithm for Optimized Initial Centers K-Means Clustering In SMEs. Journal of Theoretical and Applied Information Technology. 2016 August 1; 90(1):23-30.

[14] Hand DJ, Mannila H, Smyth P. Principles of data mining (adaptive computation and machine learning). Cambridge, MA: MIT press; 2001 Aug.

[15] Kantardzic M. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons; 2011 Jan 5.

[16] Gupta GK. Introduction to data mining with case studies: Prentice Hall of India. New Delhi. 2006.

[17] Raghupathi K. 10 Interesting Use Cases for the K-Means Algorithm. · AI Zone · Analysis Mar. 27, 2018. Available at: https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-alg. Last visited on: 30-7-2018

[18] Al Malki A, Rizk MM, El-Shorbagy MA, Mousa AA. Hybrid genetic algorithm with K-means for clustering problems. Open Journal of Optimization. 2016 Jun 15;5(02):71-83.

[19] Chittu V, Sumathi N. A modified genetic algorithm initializing K-means clustering. Global Journal of Computer Science and Technology. 2011 Feb 3; 11(2):55-62.

[20] Shboul BA, Myaeng SH. Initializing k-means using genetic algorithms. In International Conference on Computational Intelligence and Cognitive Informatics (ICCICI 09) 2009 (pp. 114-118).
[21] Jahanbin K, Rahmanian F, Rezaie H, Farhang Y, and Afroozeh A. K-means Optimization Clustering Algorithm Based on Hybrid PSO/GA Optimization and CS validity index. International Journal of Scientific Study. July 2017; 5(4): 232-242.
[22] Gupta S, Kumar R, Lu K, Moseley B, Vassilvitskii S. Local search methods for k-means with outliers. Proceedings of the VLDB Endowment. 2017 Mar 1; 10(7):757-768.
[23] Saharan S, Baragona R, Nor ME, Salleh RM, Asrah NM. Clustering for Binary Data Sets by Using Genetic Algorithm-Incremental K-means. In Journal of Physics: Conference Series 2018 Apr (Vol. 995, No. 1, p. 012038). IOP Publishing.
[24] Prabha K, Saranya R. Refinement of k-means clustering using genetic algorithm. Journal of Computer Applications (JCA). 2011; 4(2):40-44.
[25] Min-Yuan C, Kuo-Yu H. K-means clustering and chaos genetic algorithm for nonlinear optimization. In The 26th International Symposium on Automation and Robotics in Construction (ISARC) 2009.
[26] Bouhmala N, Viken A, Lønnum JB. Enhanced Genetic Algorithm with K-Means for the Clustering Problem. International Journal of Modeling and Optimization. 2015 Apr 1; 5(2):150-154.
[27] Haraty RA, Dimishkieh M, Masud M. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of distributed sensor networks. 2015 Jun 1; 2015:1-11.
[28] Arkeman Y, Wahanani NA, Kustiyo A. Clustering K-Means Optimization with Multi-Objective Genetic Algorithm. International Journal of Electrical & Computer Sciences IJECS-IJENS. 2012; 12(05):61-66.
[29] Abd-Alsabour N. Local search for parallel optimization algorithms for high dimensional optimization problems. In MATEC Web of Conferences 2018 (Vol. 210, p. 04052). EDP Sciences.