

# Optimal Pragmatic Clustering for Wireless Networks

Suzan Basloom<sup>1</sup>, Nadine Akkari<sup>2</sup>, Ghadah Aldabbagh<sup>3</sup>  
Faculty of Computing and Information Technology  
Department of Computer Science  
King Abdulaziz University, Jeddah, Saudi Arabia

**Abstract**—Nodes' clustering in wireless networks is one of the solutions that used to improve network performance. This paper discusses the clustering in wireless networks. Then it presents a novel clustering algorithm named Pragmatic Genetic Algorithm (PGA). It combines two of the well known artificial intelligence techniques: K-means and Genetic algorithm. The proposed algorithm aims at minimizing the execution time of the clustering, especially in time-sensitive wireless networks applications. The performance of PGA has been compared with the classical clustering algorithms, namely, K-means and KGA. The experiments have been conducted using synthetic and real data from public repositories. PGA obtained excellent results in execution and stable accuracy even when the number of nodes was increased.

**Keywords**—Clustering; genetic algorithm; K-means; wireless networks

## I. INTRODUCTION

Machine-based decisions making is an important aspect that attracted the scientific research community. Artificial intelligence (AI) has proposed a plethora of machine learning algorithms to improve the decision-making process. These algorithms have been embedded in many modern devices such as cars, cameras, sensors, and networks, to make the proper decision and improve overall performance. Some of these decisions must be made in a real-time manner to maintain system stability. For example, in mobile networks, the decision of moving the current connection from one access point to another for the handover [1] must be done in a time faster than the node speed. The delay in such a process can break the connection which will affect the user satisfaction level (USL).

However, finding the optimum decision (solution) using machine learning algorithms needs long execution time to converge. This long execution makes any system manager choose the approaches that make the decision in a shorter time even if its accuracy was not the best. Clustering is one of the unsupervised machine learning algorithms. It classifies data elements into groups based on the similarity between them. There are many clustering algorithms such as K-means, DBSCAN, Mean-Shift, etc. K-means is the most common algorithm because of its short execution time and acceptable results, but it faces the problem of local minima [2].

Clustering has been used in wireless networks to improve network performance. It divides nodes into small groups based on specific criteria such as distance, energy, or route as shown in Fig. 1. Clustering has proved its ability to simplify the networks' task management. Its applications can be found in sensor networks [3] where the nodes are clustered to send their data to the sink node which is responsible for forwarding them to the server.

Also, it is used in dense networks to increase the networks' capacity and coverage area through tethering or relaying [4], [5]. Edge-nodes in such networks are tethered to the closer nodes to the access point, cluster-head, to forward their data. The cluster-head needs to have a certain amount of energy to forward the received packets or it must be replaced. To change the cluster-head, a real-time clustering must be conducted. This re-clustering process is critical and needs to be executed in a short time to maintain stable performance.

Moreover, the clustering is applied in proactive mobility management approaches [6]. In such approaches, the next attaching point of a mobile node is pre-predicted on the go, based on its mobility behavior. This type of clustering is time sensitive and need to be updated frequently.

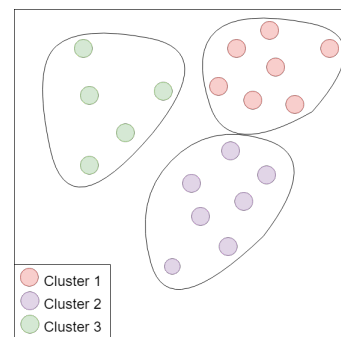


Fig. 1. Nodes Clustering

This paper's contribution is to propose an algorithm called Pragmatic Genetic Algorithm (PGA) to enhance the clustering execution time in wireless networks.

The rest of the paper is organized as follows. Section II a brief background is presented including the wireless network characteristics, definitions of K-means and genetic algorithm, and reasons for using them in this paper. Section III shows the related works that combine K-means and genetic algorithm for the clustering process. Section IV presents the proposed algorithm PGA. The simulation environment and results are discussed in Section V. The conclusion is in Section VI.

## II. BACKGROUND

This section overviews wireless networks challenges and how the clustering is used to overcome their effects. After that, it justifies the reasons behind choosing to combine K-means and Genetic Algorithm (GA). Followed by a brief explanation of both approaches.

### A. Wireless Networks Characteristics

Wireless networks provide a location independent connection, where a node can be connected anywhere at anytime. The media in such networks is the air, which propagates a variety of electromagnetic waves in different frequencies and powers. These waves [7], in addition to the weather effects, can attenuate by distance, interfere and congest with each other. To control such effects many solutions have been proposed such as specifying the coverage area of the access points, determining the transmission power, and selecting frequencies for each type of wireless technology. Wireless networks have characteristics that challenge the network management processes which must be considered while clustering the network. Some of these characteristics summarized below.

1) *Architecture*: Wireless networks are classified into two architectures: infrastructure and infrastructure-less [8]. Infrastructure scheme has pre-constructed access points, that control the connection between the network's nodes. On the other hand, nodes in the infrastructure-less scheme can be connected directly as in the Ad hoc networks. Based on the network architecture the clustering algorithm could be done central-based or distributed-based.

2) *Coverage area*: Wireless signals [9] can propagate for a specific distance based on its band, which includes different frequency, phase, and amplitude. Hence, each wireless technology has a specific coverage area based on the used wireless band. For instance, *IEEE802.11g* works on 2.5 GHz band [10] with a coverage area up to 60  $m_2$ . On the other hand, LTE tower works on 698–787 MHz band and could cover an area up to 1  $Km_2$ .

3) *Energy*: Mobile nodes have limited power and in many cases cannot be recharged immediately which represent a challenge in network design. Nodes' clustering can minimize the power limitation and increase the network lifetime using cooperative communication.

4) *Mobility*: The most convenient feature in wireless networks is mobility. The mobility represents a hot topic in network management which plays a major role in providing seamless movement between a network access points. Clustering the network nodes or the access points can provide better mobility management in the network.

### B. Why K-means with GA

This paper studies the combination of Genetic Algorithm (GA) and K-means for the following reasons:

- K-means is a simple fast-executing clustering technique.
- GA is the most popular heuristic approach used based on Jones et al. [11] overview.
- GA can optimize multiple objectives which go along with wireless networks which have multi-conflicting parameters needed to be optimized.
- K-mean and GA are computationally simpler compared to other AI techniques such as neural networks. This is because they require only swapping and shifting of genes in chromosomes.

### C. K-means

K-means clustering algorithm [12] starts by generating random k centroids. The first clusters are created based on the Euclidian distance between the centroids and the nodes. After that the centroids are updated based on the calculated mean of each cluster. This process is repeated until the new centroids remain the same as the previous ones.

### D. Genetic Algorithm

Conventional Genetic Algorithm (GA) is inspired by the theory of evolution. In nature, weak organisms are faced with extinction. The stronger ones have the opportunity to pass their genes to next generations via reproduction. In GA, a

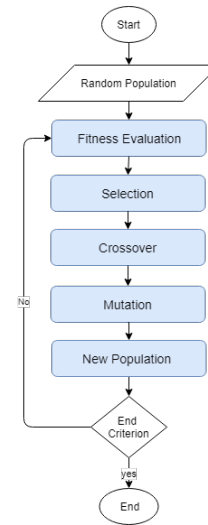


Fig. 2. Genetic Algorithm

chromosome represents a unique solution in the population. The population is usually randomly initialized. The GA uses two operators to create a new generation of chromosomes from previously existing chromosomes. These two operators are selection and reproduction. In the selection step, two chromosomes are selected that have the highest fitness value to be the parents for the next generation. After that, reproduction will start with two sub-steps: crossover and mutation. In the crossover, an offspring is generated by combining the two parent's chromosomes. This gives the offspring the opportunity to inherit good genes. Usually, the crossover rate is very high, typically 0.9 percent. In mutation, a random change is introduced into characteristics of chromosomes. The mutation rate is very small, typically less than 0.1 percent. After that the next generation is reproduced as depicted in the flowchart in Fig. 2.

The procedure of the basic genetic algorithm is given as follows [13]:

- 1) Set  $t=1$ . Randomly generate N solutions to form the first population  $P_1$ .
- 2) Fitness Assignment: Evaluate and assign a fitness value to each solution  $x \in P_1$  based on the objective function.
- 3) Selection: Select two solutions  $x$  and  $y$  from  $P_1$  which have the best fitness values to be used in the creation of the new population  $P_t$ .

- 4) Crossover: Using a crossover operator (one point or more) to generate the new offspring and add them to  $Q_t$ .
- 5) Mutation: based on predefined mutation rate (always a low rate) mutate some of the new offspring  $x \in Q_t$ .
- 6) If the stopping criterion is satisfied, terminate the search and return the current population, else, set  $t = t + 1$  go to Step 2.

### III. RELATED WORKS

This section presents the previous works that combine GA with K-means to enhance the clustering process. These works can be classified into three categories: improving the centroid location, reducing the local minima, and generating a unique solution at each run. A brief description of these categories presented below.

In [14], author added the GA algorithm to K-means to find the optimal clustered data. The clusters went through GA operations (selection and reproduction) to create generations. These generations recreated many times based on the distance fitness-function. The results show a better clusters formation compared to clustering by K-means. The main drawback of this solution is the high number of iterations needed to find the optimal solution.

On the other hand, a combination of K-means and the genetic algorithm has been used in [13], [15] to search for the optimal centroids. The clusters are initially created around random centroids of  $n$  dimensions. The new centroids are calculated based on the mean of the nodes that belongs to that cluster (the fitness function). When the newly calculated centroid has the same value as the previous centroid the process will be terminated. To evaluate the performance of GA-clustering, artificial and real-life data sets are used. GA-clustering outperformed K-means clustering results. The  $n$ -dimensional centroids could add complexity to the calculations.

Moreover, [16] also uses a K-means and GA as a solution to the multi-objective resource allocation problem. The K-means clustering algorithm used to divide the population into different-sized populations. The crossover and mutation are alternately applied to create the next generation. The aim of the work is to preserve diversity into the solutions, instead of having a similar solution each time.

### IV. PROPOSED ALGORITHM

This paper proposes a pragmatic clustering algorithm combines K-means and GA, because of the need for a fast executing clustering algorithm for the wireless networks.

The proposed PGA is different from the previous works by adding two fitness values to the clustering process. PGA flowchart is depicted in Fig. 3. The first fitness value is calculated between the centroids and the population, while the second fitness value is calculated between the centroids themselves. This fitness could eliminate unwanted crossovers. Hence reduces the execution time that is needed to define the clusters. The PGA algorithm steps are as follows:

- 1) Generated a number of  $k$  chromosomes based on K-mean algorithm.
- 2) Calculate the fitness value of each node to the chromosomes using distance Eq. 1.
- 3) Specify node's cluster such that the node belongs to the shortest distance centroid.
- 4) Calculate the fitness value between the chromosomes using distance value (proposed step) Eq. 1.
- 5) Starts the GA operations between chromosomes such that crossover will start if the crossover-centroids are apart away. This step could eliminate the creation of very close centroids which will result in unfitted individual. This will reduce the execution time needed if this crossover included.

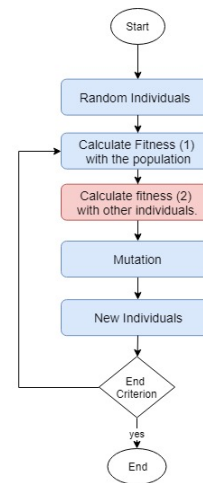


Fig. 3. Proposed Pragmatic GA Clustering PGA.

The details of PGA steps follow the main five steps of the GA as follows:

#### A. Representation and Initial Population

The initial individuals  $P_1$  is randomly chosen and stored in a list of decimal numbers. Each individual (chromosome) in the population represent  $k$  centroids.

#### B. First Fitness Value Calculation

The mean of the distances between the each individual and nodes is calculated based on equation 1. The result is saved in the last digit of each chromosome, which represents the first fitness value.

$$d = \sqrt{\frac{x_2 - x_1}{y_2 - y_1}} \quad (1)$$

#### C. Selection

There are many selection strategies used for the GA. PGA uses the Roulette wheel selection strategy [17] which selects two individuals (parents), that have the highest first-fitness values to create the new individuals (children), for the next generation  $P_t$ .

D. Second Fitness Value Calculation

The second fitness value is calculated based on the distance between the individuals themselves based on the equation 1. This step is added in order to eliminate the crossover between adjacent-centroids. The crossover between adjacent centroids will create non-uniformed clusters. Also, it will add unwanted calculations to the first-fitness step, where the newly created clusters will have low fitness values. The second fitness step will calculate the fitness of all combinations between the selected individuals (centroids). Then the highest fitness combination will be used in the crossover process.

E. Crossover

Crossover exchanges the genes of the individuals (parents) to generate the offspring. The crossover in GA could be done on a single point or more. In this paper, a One-point crossover strategy is used as appeared in Fig. 4. The probability of crossover is usually high, but for this paper, it depends on the second fitness value. If the fitness value is high, the crossover will be conducted otherwise it will not.

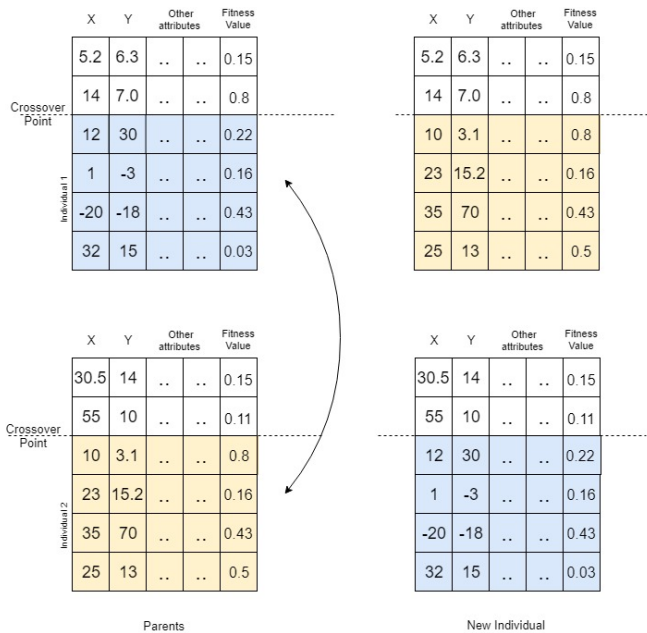


Fig. 4. Crossover in PGA

F. Mutation

Mutation is an occasional alternative of a character in a chromosome. In mutation, a character will be changed if a randomly generated number is equal or less than the probability of the mutation. The probability of mutation in GA is usually very small.

G. Elitism

The purpose of the elitism strategy is to move the good genes from the previous iteration to the next. So the centroids that has the best fitness values can be used in the creation of the new population.

H. Stopping Criterion

The algorithm will stop when it reaches to the predefined number of generations.

V. EXPERIMENTS AND ANALYSIS

To validate the proposed PGA algorithm, we use synthetic and real data sets. The experiments and their results are described below. The initial values of the GA algorithm, for all experiments are shown in Table I. Two metrics are used to evaluate the performance of the proposed algorithm as follows:

- **Execution time:** It is the time the program takes to reach to the final clusters.
- **Accuracy:** The percentage of the right created clusters by the algorithm, compared to pre-known results of the used data set.

TABLE I. THE USED GA PARAMETERS

Parameter	Value
Crossover Probability	0.8
Mutation Probability	0.01
Number of Individuals	20
Generations	100

A. Experiment 1

In this experiment, “Iris” data set [18] is used, which is commonly used for testing machine learning algorithms performance [19], [20]. This data set represents different categories of irises with four feature values: the sepal length, sepal width, petal length and the petal width in centimeters. It has 150 records which are categorized into three classes with 50 samples each. Therefore the chosen number of clusters (k) is three. The aim of this experiment is to compare the performance of the proposed PGA algorithm with the K-means and KGA. KGA [16] is an algorithm that uses the GA algorithm to improve the K-means centroid. The accuracy is calculated by comparing the clustering results with the pre-known correct answers of the Iris data set.

It can be seen in Table II that PGA has the best accuracy compared to K-means and KGA. On the other hand, K-means has the best execution time but with the lowest accuracy. PGA is better than KGA by 5% in execution time and accuracy in this experiment. Fig. 5 shows the result of clustering the “Iris” data set by PGA.

TABLE II. EXPERIMENT 1 RESULTS

Algorithm	Iteration	accuracy
K-means	0.03	49%
KGA	28	79%
PGA	21	81%

B. Experiment 2

In this experiment 200-400 points in  $R^2$  are randomly generated to from four classes. The classes are taken in such a way that the distance from a point to its class-centroid is less than the distance of that point to other classes centroids as appeared in Fig. 6.



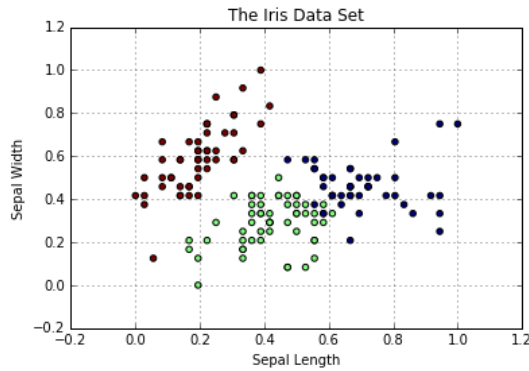


Fig. 5. Iris Clustering by PGA

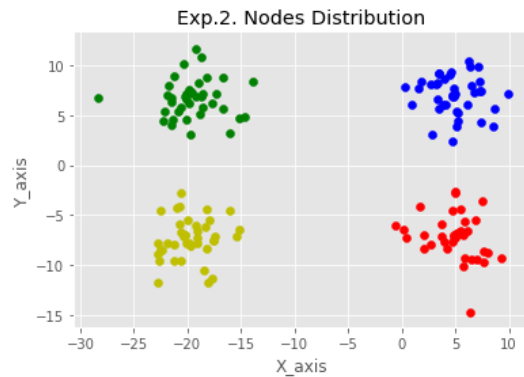


Fig. 6. Exp.2 Nodes Distribution.

TABLE III. EXPERIMENT 2 RESULTS

Number of Nodes	200		300		400	
	Algorithm	Time	Accuracy	Time	Accuracy	Time
K-means	0.1	50%	0.3	25%	0.5	24%
KGA	78	65%	81	68%	108	75%
PGA	51	75%	56	71%	76	75%

The results of experiment 2 are depicted in Tables III. PGA-clustering attains the best value of accuracy compared to  $K$  –  $means$  and  $KGA$ .  $PGA$  has stable accuracy even when the number of nodes increased.  $K$ -means, on the other hand, attains the best execution time but with the lowest accuracy especially when the number of nodes is increased.

The addition of the second fitness value to the  $PGA$  eliminate the unwanted calculations. These calculations could be added if the incorrect clusters (clusters with close centroids) are generated by the crossover process. This is can be seen in the result of  $PGA$  in the former experiments.

## VI. CONCLUSION

In this paper, we have presented a new clustering technique called  $PGA$ . It aims at improving the execution time that is needed to cluster group of nodes. Clustering is an important class of unsupervised learning techniques that have attracted the attention of the research community over the last few years.

The proposed approach is a novel contribution, which adds a second fitness value to the clustering process.  $PGA$

combines the benefits of Genetic algorithm and  $K$ -means. This combination can omit some of the drawbacks of each algorithm such as the local minima of the  $K$ -means and the long execution time of  $GA$ .

We have tested the proposed algorithm in different synthetic and real clustering problems. The results of our testing of execution times clearly show that  $PGA$  outperformed  $K$ -means and  $KGA$  algorithms while maintaining good accuracy results.

## REFERENCES

- [1] R. Ahmad, E. A. Sundararajan, N. E. Othman, and M. Ismail, "Handover in lte-advanced wireless networks: state of art and survey of decision algorithm," *Telecommunication Systems*, pp. 1–26, 2017.
- [2] J. Qin, W. Fu, H. Gao, and W. X. Zheng, "Distributed  $k$ -means algorithm and fuzzy  $c$ -means algorithm for sensor networks based on multiagent consensus theory," *IEEE transactions on cybernetics*, vol. 47, no. 3, pp. 772–783, 2017.
- [3] S. Basloom, N. Akkari, and G. Aldabbagh, "Wireless sensor networks: Cross-layer optimization performance," in *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on*. IEEE, 2016, pp. 950–955.
- [4] S. Basloom, A. Nazar, G. Aldabbagh, M. Abdullah, and N. Dimitriou, "Resource allocation using graph coloring for dense cellular networks," in *2016 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2016, pp. 1–5.
- [5] M. Hajjar, G. Aldabbagh, and N. Dimitriou, "Using clustering techniques to improve capacity of lte networks," in *Communications (APCC), 2015 21st Asia-Pacific Conference on*. IEEE, 2015, pp. 68–73.
- [6] P. Kuila and P. K. Jana, "Energy efficient clustering and routing algorithms for wireless sensor networks: Particle swarm optimization approach," *Engineering Applications of Artificial Intelligence*, vol. 33, pp. 127–140, 2014.
- [7] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012, vol. 34.
- [8] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE access*, vol. 3, pp. 1206–1232, 2015.
- [9] S. S. Rakib and R. Hadani, "Signal modulation method resistant to echo reflections and frequency offsets," Jul. 14 2015, uS Patent 9,083,595.
- [10] V. Ssorin, A. Artemenko, A. Sevastyanov, and R. Maslennikov, "Compact bandwidth-optimized two element mimo antenna system for 2.5–2.7 ghz band," in *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP)*. IEEE, 2011, pp. 319–323.
- [11] D. F. Jones, S. K. Mirrazavi, and M. Tamiz, "Multi-objective metaheuristics: An overview of the current state-of-the-art," *European journal of operational research*, vol. 137, no. 1, pp. 1–9, 2002.
- [12] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [13] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern recognition*, vol. 33, no. 9, pp. 1455–1465, 2000.
- [14] C. A. Murthy and N. Chowdhury, "In search of optimal clusters using genetic algorithms," 1996.
- [15] S. Bandyopadhyay and U. Maulik, "An evolutionary technique based on k-means algorithm for optimal clustering in rn," *Information Sciences*, vol. 146, no. 1-4, pp. 221–237, 2002.
- [16] A. Mousa, M. El-Shorbagy, and M. Farag, "K-means-clustering based evolutionary algorithm for multi-objective resource allocation problems," *Appl. Math*, vol. 11, no. 6, pp. 1681–1692, 2017.
- [17] A. Shukla, H. M. Pandey, and D. Mehrotra, "Comparative review of selection techniques in genetic algorithm," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*. IEEE, 2015, pp. 515–519.
- [18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

- [19] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015.
- [20] T. Gupta and S. P. Panda, "A comparison of k-means clustering algorithm and clara clustering algorithm on iris dataset," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 4766–4768, 2018.