# Bearing Fault Diagnosis using Hybrid Genetic Algorithm K-means Clustering

M. M. Ettefagh

Vibration and Modal analysis Laboratory, Mechanical Engineering Department, University of Tabriz, Tabriz 51666-14766, Iran
Corresponding Author:
ettefagh@tabrizu.ac.ir

M. Ghaemi

Department of Computer Science University of Tabriz, Tabriz 51666-14766, Iran
M_ghaemi89@ms.tabrizu.ac.ir

M. Yazdanian Asr

Vibration and Modal analysis Laboratory, Mechanical Engineering Department, University of Tabriz, Tabriz 51666-14766, Iran
yazdanian.mahsa@gmail.com

*Abstract*—**Condition monitoring and fault diagnosis of rotating machinery are very significant and practically challenging fields in industries for reducing maintenance costs. Fault diagnosis may be interpreted as a classification problem; therefore artificial intelligence-based classifiers can be efficiently used to classify normal and faulty machine conditions. K-means clustering is one of the methods applied for this purpose. In this paper, a new fault diagnosis method is proposed by applying Genetic Algorithm (GA) to overcome the drawback of K-means which it may be get stuck in local optima. For this purpose, the best solution of GA is chosen to be the initial point for K-means clustering. The proposed method is used in fault diagnosis of the scaled rotor-bearing system experimentally. Then the result of hybrid GA-K-means clustering is compared with classic K-means clustering.**

***Keywords- Condition Monitoring; Fault Diagnosis; K-means Clustering; Genetic Algorithm.***

## I. INTRODUCTION

Rolling element bearings are greatly used in most of industrial machine elements and have wide range applications in both heavy machineries and small devices. With the development in industry and manufacturing modern systems, machines are expected to operate continuously for long lasting hours. Therefore, condition monitoring and fault diagnosis of rotating machinery are very significant and practical challenges in industries for reducing maintenance costs, increasing machine availability, improving productivity, preventing unexpected downtimes, failures and avoiding personal casualties and also economical loss. Fault diagnosis is a kind of classification problem, and artificial intelligence techniques-based classifiers can be efficiently used to classify normal and faulty machine conditions. Traditional engineering approaches for rolling element bearing fault detection require a considerable amount of engineering experience. So, methods based on intelligent/expert systems are widely applied for automated defective rolling element bearing fault detection and also for machine condition monitoring. These methods include K-means clustering [1], Artificial Neural Networks (ANN) [2], different methods based on Support Vector Machines (SVM) [3-5], Fuzzy C-means (FCM) [6] and also methods based on evolutionary algorithms like Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and etc.

Among the abovementioned methods, K-means clustering has been widely studied and applied to automatic fault detection. The K-means clustering is an unsupervised pattern classification method, applied directly to industrial environments without the need to being trained by data measured on a machine under a fault condition. On the other hand, finding the initial centers prior to K-means clustering is one of the difficulties of applying K-means in applications and many researchers have addressed this problem. Yiakopoulos and et al. [7] proposed a method to detect fault in rolling element bearing based on K-means clustering. In order to overcome the sensitivity of the K-means method to the choice of the initial cluster centers, they selected the initial centers using features extracted from simulated signals, resulting from a well-established model for the dynamic behavior of defective rolling element bearings.

In this paper, GA is applied to overcome the drawback of K-means; because K-means clustering is probable to get stuck in local optima according to inappropriate starting point. So, this problem is avoided by suitably applying GA prior to K-means clustering. . In other words in order to escape from local optima, the best solution of GA is chosen to be the initial point for K-means clustering. As a case study, the proposed method is applied on an experimental rotor-bearing laboratory scale system for fault diagnosis of the bearings. Then the result of fault diagnosis based on hybrid GA-K-means clustering is compared with classic K-means clustering.

This article is structured as the following: Section II is devoted to introduce Hybrid GA-K-means clustering and the experimental set up are presented in section III. Vibration

978-1-4799-3020-3/14/$31.00 ©2014 IEEE

signal preprocessing is presented in section IV and section V presents the applying of the proposed method on the experimental data. Finally Section VI is the conclusion section.

## II. HYBRID GA-K-MEANS CLUSTERING

K-means clustering is an unsupervised pattern classification method, applied directly to industrial environments without the need to being trained by data measured on a machine under a fault condition. Further advantage of this method is its ease of programming. The K-means algorithm seeks to partition the data into $K$ groups or clusters so that the within-group sum of squares is minimized; that is, it seeks the cluster centers $\{\mu j,\ j = 1,\dots,k\}$ that minimize:

$$M = \sum_{j=1}^{k} S_j \tag{1}$$

where the within-group sum of squares for group $j$ is:

$$S_j = \sum_{i=1}^{n} z_{ji} \left| x_i - \mu_j \right|^2 \tag{2}$$

in which $z_{ji} = 1$ if $x_i$ is in group $j$ (of size $n_j = \sum_{i=1}^{n} z_{ji}$ ) and zero otherwise; $\mu_j$ is the mean of group $j$,

$$\mu_j = \frac{1}{nj} \sum_{i=1}^{n} z_{ji} x_i \tag{3}$$

The first step in the application of K-means clustering is to find a set of initial centers. K-means clustering is an iterative hill-climbing algorithm and it is significantly sensitive to the initial randomly selected cluster centers. Varying the starting conditions can produce different stable cluster. Although the K-means algorithm had been applied to many practical clustering problems successfully, it may converge to a partition that is significantly inferior to the global optimum [8].

Clustering by the means of GA seems to overcome the problem of hill climbing algorithms in clustering. Commonly used hill climbing algorithms can only generate a local optimal solution; whereas, meta-heuristic algorithms such as GA are able to escape from local optima with the help of mutation operator. In the following, GA clustering, used in this paper, is described in details:

### *A. Population Initialization*

GA performs on the population of potential solutions. Each individual of this population is named as “chromosome”. Each chromosome of the GA represents the centers of the clusters and it is a vector with $K \times N$ columns; where $K$ is the number of clusters and $N$ is the dimension of each clusters. The $K$ cluster centers encoded in each chromosome are initialized to $K$ randomly chosen points from the data set. This process is repeated for each of the chromosomes in the population.

### *B. Fitness function*

The clustering metric $\mathcal{M}$ of each individual is calculated by (1), by which the sum of the distance of each data in the training set is calculated from its cluster center. In this article, Euclidean distance is used as the distance metric.

### *C. Crossover*

In this article, single point crossover with a fix crossover probability $p_c$ is adapted. For each pair of chromosomes that undergoes the crossover operation, a random integer, called the crossover point, is generated in the range [*1, K-1*]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

### *D. Mutation*

Each chromosome undergoes mutation with a fixed probability $p_m$. In order to perform mutation, for each selected individual, a mutation point is determined randomly and its value is exchanged with another randomly generated value in the variable’s range. Mutation operator performs the global search in the search space as an effort to escape from local optimum.

After finding the cluster centers from training data with either K-means or GA-K-means, the classification accuracy using K-Nearest Neighbor (KNN) method on the test data will be reported in addition to distance metric from (1). The accuracy measure is calculated as the following:

$$Accuracy = \frac{\sum_{i=1}^{n} \delta(y_i, c_i)}{n} \tag{4}$$

where $n$ is the number of samples in the testing data, $y_i$ and $c_i$ denote the true category label and the obtained cluster label with either methods, respectively. $\delta(y,c)$ is a function that equals 1 if $y=c$ and equals 0 otherwise [9].

Further improvements can be achieved by using the hybrid method of GA and K-means. In other words, in hybrid GA and K-means (GA-K-means), GA have applied for predefined iterations at first and then the best solution of GA is chosen to be the initial point for K-means clustering. The flowchart of the proposed method is shown as Fig. 1. The experimental results of applying proposed hybrid GA-K-means is presented in section V.

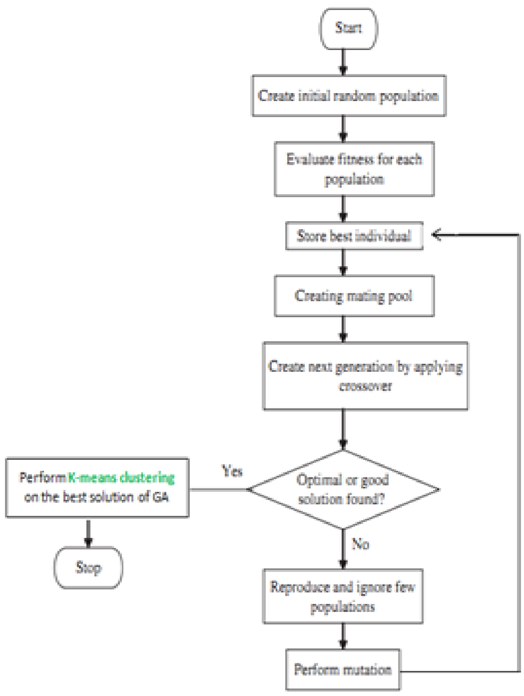


Fig. 1. Flowchart of hybrid GA-K-means

## III. EXPERIMENTAL SET UP

For investigating the ability of the proposed method in a real condition, the method is applied on a laboratory scaled rotor-bearing system for fault diagnosis of the beatings. To improve the diagnosis ability in identification of the bearing damages, selecting an acceptable data acquisition process is important for achieving vibration signals and features. In this case, to collect data set, the experimental set up consists of an electric motor, a shaft which is mounting on two ball bearings, multi-channel Pulse analyzer system (B&K Co.), a triaxial accelerometer (B&K Co.), and tachometer (B&K Co.) as shown in Fig. 2.

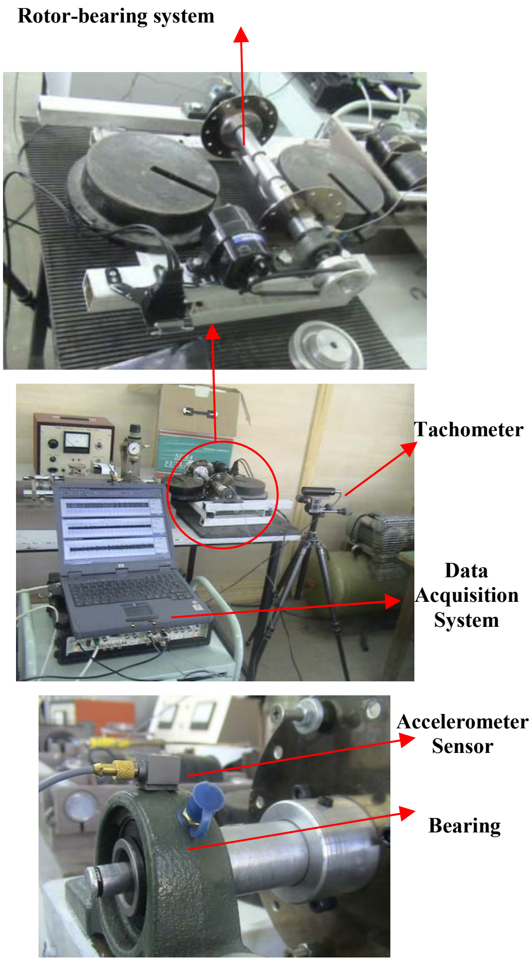


Fig. 2. Laboratory experimental set up

In order to reduce the undesired vibration and noise, four shock absorbers were installed under the base. All vibration signals were collected from the experimental testing using the accelerometer which was mounted on the outer surface of left-side bearing housing. The system was operated to reach steady state condition in order to fix motor rotational frequency and sampling frequency. After that, data set were collected with multi-channel Pulse analyzer system (B&K Co.). In this system output analog signals from accelerometer and tachometer were connected to Pulse data acquisition system and changed into digitals by A/D methods. Vibration signals were recorded from a machine with four bearing conditions including ball, cage, outer race damages, and the normal. These four groups are considered as four clusters. It should be noted that the fault type is not identifiable only using row vibration signals in this step. Therefore further signal processing and proposed method is necessary for fault diagnosis and identification. Raw signals recorded from Pulse analyzer system were imported to MATLAB for preprocessing.

## IV. VIBRATION SIGNAL PREPROCCESING

The preprocessing of vibration signals involves with synchronization of the signal, piecewise cubic spline interpolation and feature extraction are performed in the following three steps to extract the feature vector.

### *A. Synchronization of vibration signals*

It is important to average any raw data before use, since it contains noise. But because of the voltage fluctuations, which have been caused by AC power, the acceleration signal changes in each period of rotation. Therefore, synchronize averaging is necessary for reducing noise and fluctuation effect. For this purpose, the real rotational speed of the system was measured by tachometer. According to tachometer‘s diagram and synchronization with acceleration diagram and also finding the rotational speed, we can separate data into groups, that each group related to one period of rotation, see Fig. 3. The process can be accomplished by applying piecewise cubic Spline interpolation [10] method. This interpolation is a piecewise cubic function, whose cubic pieces join together to form a function with two continuous derivatives.

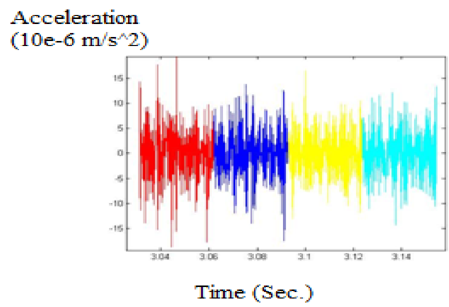


Fig. 3. Separation of row signals to recognize each period of rotation

### *B. Power spectrum of signal*

In this step the obtained time-domain signal is transformed to the frequency domain by deriving the power spectrum of the signal in each period of rotation. In signal processing, the frequency domain refers to the analysis of mathematical functions or signals with respect to frequency, rather than time [11]. A time-domain signal presents how a signal changes over time, whereas a frequency-domain signal presents how much of the signal lies within each given frequency band over a range of frequencies. Some of spectrum estimation methods include the Yule-walker, covariance, Burg, and Least Squares approaches. Each one uses a different technique to estimate power spectrum. In this paper the Burg method is applied in this research to represent the vibration signals in frequency domain for feature extraction. The Burg algorithm is based on parametric modeling of the signal that is a set of all-pole model parameters that minimize the sum of the squares of the forward and backward prediction errors. For the sake of stability, it performs the minimization sequentially with respect to the reflection coefficients [12].

### C. Feature extraction

For clustering purpose, proper features from input data should be extracted. To do so, after the transformation of the time-domain signals to frequency-domain by Burg method (as an example see Fig. 4), each period of rotation can be separated into 10 distinct sections. Then the peak of each section is chosen as a feature. So, 10 features for each period of rotation can be obtained, then clustering task will be performed.

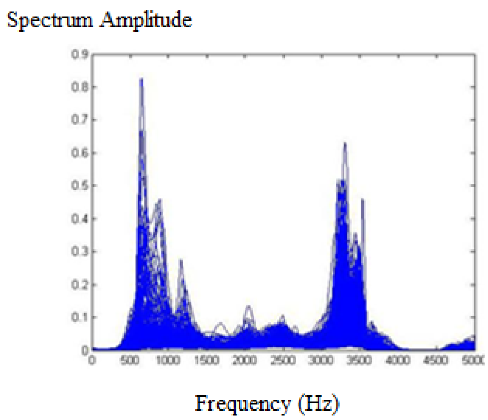


Fig. 4. Power spectrum of the vibration signals for normal condition of the rotor-bearing system (for all rotations)

## V. APPLYING CLUSTERING METHODS

In the experimental part of this work, 70% of the input data is training and the remaining is used as testing data. To avoid any bias in results, the mean of 30 individual runs for each method has been reported. Because the features are in different ranges, before applying any clustering method, the normalized value of each feature have been calculated by the following equation:

$$x' = (x - x_{min})/(x_{max} - x_{\min}) \quad (5)$$

where $x$ is the value of the feature before normalization and $x'$ is the normalized value of $x$. $x_{min}$ and $x_{max}$ are the lower and upper bounds of the features, respectively.

As mentioned before, the main drawback of K-means is that it may get stuck in local optima. So, it have been tried to avoid this problem by applying GA prior to K-means clustering in this paper. The parameters for GA are listed in Table I. As reported by Deb [13], a crossover probability in the range 0.75-0.95 and the mutation rate in the range 0.005 to 0.1 perform well. One point crossover and selection with elitism is used in this experiments. Also K-means clustering in both classic K-means and hybrid GA-K-means will continue until the cluster center doesn't change for 50 consequent iterations.

TABLE I. PARAMETERS FOR GA

| Population size | Probability of crossover | Probability of mutation |
|---|---|---|
| 50 | 0.8 | 0.1 |

The results for both classic K-means and hybrid GA-K-means are shown in Table II and Table III, based on testing data. As it is obvious, if K-means starts from a proper point in the search space, it can reach to the optimum solution and we can expect results near to 100% accuracy in the test dataset. Otherwise it will be get stuck in the local optimum and clustering on the test dataset will end with low accuracy in later experiments. In order to deal with this drawback, we have employed GA prior to K-means, which is an attempt to overcome this problem. Among the statistic parameters, Standard Deviation (STD) has a good demonstration of dissimilarity between data; so this parameter is used to show the superiority of our proposed method to classic K-means.

TABLE II. CLASSIFICATION ACCURACY (CA) AND STANDARD DEVIATION (STD) FOR CLASSIC K-MEANS AND HYBRID GA-K MEAMNS ( AVERAGE OF 30 ITERATION WERE REPORTED)

| | CA/STD (cluster 1) | CA/STD (cluster 2) | CA/STD (cluster 3) | CA/STD (cluster 4) |
|---|---|---|---|---|
| **Classic K-means** | 97.1333/ 8.5610 | 91.8667/ 8.3861 | 76.7333/ 34.4993 | 64.3333/ 30.2636 |
| **Hybrid GA-K-means** | 100/0 | 95.0345/ 2.3065 | 99.10345/ 1.1447 | 92.8965/ 3.0041 |

The results of Table II show the Classification Accuracy (CA) and Standard Deviation (STD) of KNN classifier of 30 individual runs on the testing data after the cluster centers are obtained with both classic K-means and hybrid GA-K-means. CA is defined as the correct classifications that the KNN classifier makes and it is calculated as (4) from Section II. KNN is member of lazy learning algorithms [14] and it depends on some near neighbors according to the value of $K$ in order to classify the testing data. The value of $K$ is chosen to be 1 for KNN classification algorithm in our experiments.

TABLE III. DISTANCE METRIC OF CLASSIC K-MEANS AND HYBRID GA –K-MEANS (STD AND MEAN OF 30 ITERATIONS)

| | Mean | STD |
|---|---|---|
| **Classic K-means** | 316.646 | 19.1851 |
| **Hybrid GA-K-means** | 299.3608 | 9.4330 |

The results of Table III represent the clustering metric for both methods according to (1), which is the sum of the distance of each data from its corresponding center. Also, if the STD value of both methods are compared in Table III, it can be concluded that, although K-means can perform better than GA-K-means in some runs, GA-K-means has stable behavior than K-means due to its low STD. To sum up GA-K-means has high efficiency than K-means alone in clustering dataset. It should be mentioned that results are reported in this research are based on over 30 runs on the experimental test dataset.

According to Table II and Table III, applying GA before K-means, improves the results in both classification accuracy and sum of distance of each data with its corresponding center (distance metric). Thus hybrid GA-K-means outperforms classic K-means. Also, as it is obvious from Tables II and III, classic K-means has large STD in comparison with Hybrid GA-K-means. This large STD is predictable according to the fact that classical K-means may have gotten stuck in local optima in some iterations over 30 runs and this will lead to low performance.

For better representation of the obtained results of clustering, the 3D plot of the testing data with hybrid GA-K-means have been illustrated as Fig. 5. In this figure, three dimensions are chosen among 10 features (here features 1, 4 and 9 are selected) and the members of each cluster are plotted with different colors. The center of each cluster is shown with a star symbol. As it is obvious, our proposed method is able to classify four conditions of the system more accurately and the results are more stable than classic K-means.

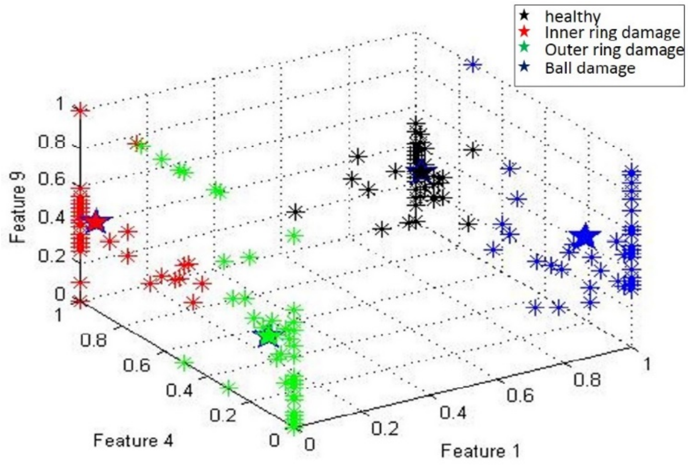


Fig. 5. 3D plot of clustering result for testing data

## VI. CONCLUSION

A hybrid GA-K-mean clustering method is proposed in this paper by applying GA for predefining iterations at first and then the best solution of GA is chosen to be the initial point for K-means clustering. The application of the proposed method is an experimental laboratory scale rotor-bearing setup for identifying healthy and three faulty condition of the machine. The result of the clustering was compared with classic K-means clustering method. It was concluded from the experiments that the proposed method is able to identify the rotor condition with more accuracy.

## REFERENCES

[1] J.B. MacQueen , "Some methods for classification and analysis of multivariate observations," Fifth Berkley symposium on mathematical statistics and probability, vol. 1, pp. 281,1967.

[2] A.C. McCormick, A.K. Nandi, "Classification of the rotating machine condition using artificial neural networks," Journal of Mechanical Engineering Science, vol. 211, no. 6, pp. 439–450, 1997.

[3] B. Samanta, "Gear fault detection using artificial neural networks and support vector machines with genetic algorithms," Mechanical Systems and Signal Processing, vol. 18, pp. 625–644, 2004.

[4] L.B. Jack, A.K. Nandi, "Fault detection using support vector machines and artificial neural networks augmented by genetic algorithms," Mechanical Systems and Signal Processing, vol. 16, pp. 373–390, 2002.

[5] H. Qiao, H. Zhengjia, Z. Zhousuo, Z. Yanyang, "Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble," Mechanical Systems and Signal Processing, vol. 21, pp. 688–705,2007.

[6] X. Wang, Y. Wang, L. Wang, " Improving fuzzy c-means clustering based on feature-weight learning," Pattern Recognition Letters, vol. 25, pp. 1123–1132, 2004.

[7] C.T. Yiakopoulos , K.C. Gryllias, I.A. Antoniadis, "Rolling element bearing fault detection in industrial environments based on a K-means clustering approach," Expert Systems with Applications, vol. 38, pp. 2888–2911, 2011.

[8] Y. Liu , X. Wu , Y. Shen, "Automatic clustering using genetic algorithms," Applied Mathematics and Computation, pp. 1267-1279, 2011.

[9] S.N. Sivanandam, S.N. Deepa, "Introduction to Genetic Algorithms," Springer-Verlag, Berlin, Heidelberg, 2008.

[10] F.N. Fritch, R.E. Carlson, "Monotone piecewise cubic interpolation," Society for Industrial and Applied Mathematics, vol. 17, no. 2, 1980.

[11] S.A. Broughton, K. Bryan, "Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing," John Wiley & Sons, 2008.

[12] A. Zaknich , "princples of adaptive filters and self-learning systems," advanced textbook in control and signal proccesing , pp.216-217, 2005.

[13] k. Deb, "Genetic algorithm in search and optimization: the technique and application," Proceedings of the 4th International Conference on Genetic Algorithms, San Diego, 1998.

[14] T. Isaac, J. Derrac, S. Garcia, F. Herrera, "Integrating a differential evolution feature weighting scheme into prototype generation", Elsevier Neurocomputing, vol 97, pp. 332-343 2012.