Hybrid HC-PAA-G3K for Novelty Detection on Industrial Systems

Yu Zhang, Jun Chen, Michael Gallimore

School of Engineering, University of Lincoln,
Lincoln, LN6 7TS, U.K.
{yzhang, juchen, mgallimore }@lincoln.ac.uk

***Abstract*— Piecewise aggregate approximation (PAA) provides a powerful yet computationally efficient tool for dimensionality reduction and feature extraction. A new distance-based hierarchical clustering (HC) is now proposed to adjust the PAA segment frame sizes. The proposed hybrid HC-PAA is validated by a generic clustering method 'G3Kmeans' (G3K). The efficacy of the hybrid HC-PAA-G3K methodology is demonstrated using an application case study based on novelty detection on industrial gas turbines. Results show the hybrid HC-PAA provides improved performance with regard to cluster separation, compared to traditional PAA. The proposed method therefore provides a robust algorithm for feature extraction and novelty detection. There are two main contributions of the paper: 1) application of HC to modify conventional PAA segment frame size; 2) introduction of 'G3Kmeans' to improve the performance of the traditional K-means clustering methods.**

***Keywords*— Piecewise aggregate approximation, Hierarchical clustering, G3Kmeans, Novelty detection, Industrial gas turbine.**

I. INTRODUCTION

Data feature extraction provides an essential tool in a wide range of signal processing systems, and often acts as a precursor technique for reducing the dimensionality of raw data whilst keeping informative features—which can be further utilized in fault/anomaly detection, pattern recognition and classification systems, for instance. Piecewise aggregate approximation (PAA), was proposed in [1] and [2] (independently) as a dimensionality reduction technique for large time-series datasets, and has since been adopted for use in medical, financial, engineering, and speech/image processing systems due to its low computation overhead [3]. However, the traditional practice of using equally distributed segments can lead to insufficient fidelity in some regions of interest, whilst providing over-segmentation in regions considered less information rich. Here then, the hybrid use of PAA and hierarchical clustering (HC) [4] is proposed. HC has been extensively used in data analysis and signal processing due to its simplicity and visual interpretation of the hierarchy structure [5], and is considered here as a means of optimizing PAA segment frame sizes according to sequence samples' similarity.

The performance of the modified feature extraction technique is further validated through use of a clustering method for pattern recognition. Whilst traditional K-means clustering algorithms only offer suboptimal solutions that are more likely to be trapped in regions of local optima, clusters are then less representative and possess lower separation of the identified data groups, and will degrade pattern recognition performance by hindering accurate classification. In view of these issues, [6] has proposed a novel clustering algorithm, termed 'G3Kmeans' (G3K), that uses a real-coded genetic algorithm, namely G3PCX [7] as its optimization method. G3Kmeans method is proven to be more robust to initial conditions, and is more efficient compared to other conventional partition-based clustering techniques [8]. Experimental trials for fault/novelty detection on industrial gas turbines [9] has already been investigated to demonstrate the efficacy of the hybrid HC-PAA-G3K for feature extraction and pattern recognition.

II. METHODOLOGIES

In the proposed methodology, HC is used to cluster the data series according to their similarities, PAA is applied to the clustered data series for data dimensionality reduction and feature extraction, and finally G3K is used for pattern recognition and fault detection.

A. Traditional PAA

PAA [1] or alternatively segmented means [2] has gained favor in data analysis because of its ease of application. The basic concept is to divide a sequence $\boldsymbol{x}$, which is a $1\times N$ vector, into n equal sized segments, and to use the mean of each segment as an extracted feature in the resultant sequence $\boldsymbol{y}$, giving an $1\times n$ vector

$$\boldsymbol{y} = [\text{mean}(\boldsymbol{g}_1), \cdots, \text{mean}(\boldsymbol{g}_n)], \quad (1)$$

where $\boldsymbol{g}_i$ is the ith segment with a length of (N/n). Because of its simple segmenting and means process, with a computational complexity of $O(n)$, PAA is extremely useful for large datasets [1].

An example signal possessing a half bell shape, shown in Fig. 1, outlines the process, and highlights the underlying issues with the traditional method —it has 1000 samples that are separated into 10 equally spaced segments and which are represented by the mean of the data within each segment (traditional PAA). It can be seen that, from 700-1000 samples, which is considered an information rich region, PAA segments are too coarse to capture the important features, whilst from 1-700 (a region less information rich), adjacent PAA segments provide relatively little added detail, and could be reasonably combined to provide further dimensionality

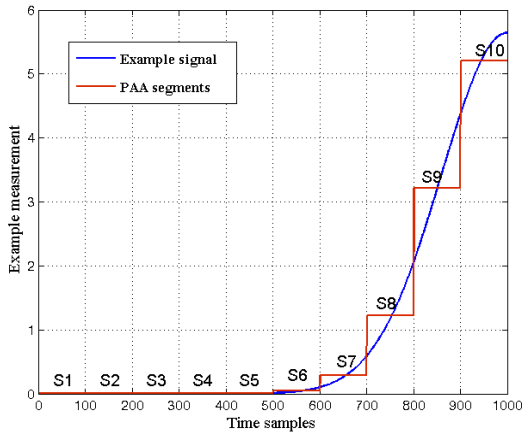


Fig. 1 Example signal and its PAA segmentation and representation (S=Segment)

reduction. HC is therefore considered here to address this issue.

B. Hierarchical Clustering (HC) for PAA segment frame optimization

HC is a very powerful tool for data analysis by providing a visual hierarchy of the datasets according to their similarities. The underlying concept of agglomerative HC is to assemble a set of objects into a hierarchical tree, where similar objects join in lower branches and these branches further join based on object 'similarity' [4].

Here, the most common measure is used, the Euclidean distance:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}\ , \tag{2}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two $1\times N$ vectors, i.e. the signals, $(x_1, x_2, ..., x_N)$ and $(y_1, y_2, ..., y_N)$. A cluster is formed when the data from two measurements has the minimum Euclidean distance. An average linkage measure is used to calculate the mean distance between all pairs of objects in clusters *m* and *n*:

$$D(m,n) = \frac{1}{N_m N_n}\sum_{j=1}^{N_m}\sum_{k=1}^{N_n} d(\boldsymbol{x}_{mj}, \boldsymbol{y}_{nk}). \tag{3}$$

where $j = 1,2,..., N_m$ and $k = 1,2,..., N_n$. $d(\boldsymbol{x}_{mj}, \boldsymbol{y}_{nk})$ is the distance between two objects in the two clusters. N_m is the number of objects in cluster *m*, and N_n is the number of objects in cluster *n*.

For the example shown in Fig. 1, HC is applied to the 1000 time samples, and the samples are clustered according to their similarities. The resulting dendrogram is shown in Fig. 2(a). The linkage distance threshold enters from above to capture the highest 10 clusters in the dendrogram. Now, PAA is applied to the resulting 10 unequal segments, with the frame sizes being dictated by the size of the respective dendrogram branches. Again, the mean of each segment is used to signify

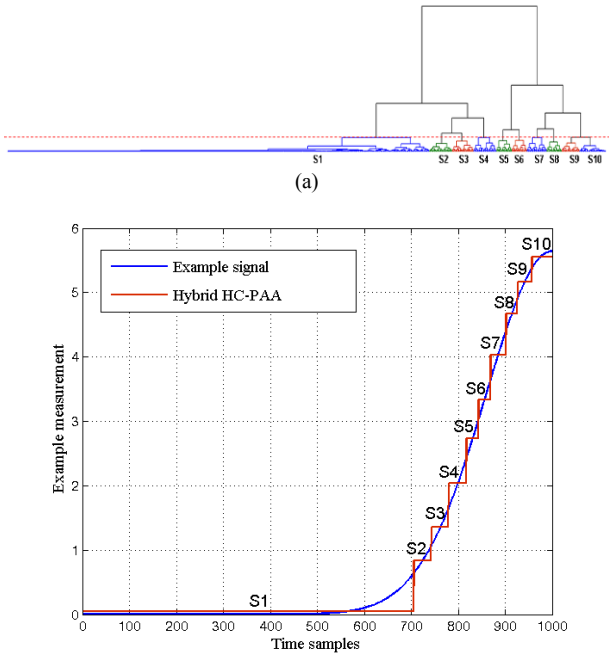


Fig. 2 (a) HC tree and 10 sub-clusters; (b) hybrid HC-PAA segments shown with the original example signal (S=Segment)

the underlying 'feature' of each segment according to (1). The resulting hybrid HC-PAA output, along with the original, is shown in Fig. 2(b). It is now evident that through application of the hybrid approach the regions that are information rich have a higher density of segments.

C. G3K Clustering for pattern recognition

After feature extraction, a novel clustering technique, termed 'G3K' [6] is applied for pattern recognition. A real-coded generic algorithm, G3PCX [7] is incorporated into K-means clustering, which only encodes cluster centers in the chromosome. The hybridized clustering technique is described by the following procedure:

1) Initialisation: the randomly generated '*k*' cluster centres are encoded in each chromosome in a concatenated form. *P* initial chromosomes are generated from the initial population.

2) Assigning data points: Each data point is assigned to one cluster with center, C_i calculated using:

$$X_m \in C_i : if \begin{cases} \|X_m - C_i\| < \|X_m - C_l\| \\ m = 1,2,..., N; i, l = 1,2,..., k; l \neq i \end{cases} \tag{4}$$

where, $\|\ \|$ is the Euclidean norm, X_m is the *m*th data point, C_i is the *i*th cluster centre, *k* is the pre-specified number of cluster centers, and *N* is the number of data samples. After the assignment, cluster centers encoded in the chromosome are updated by calculating the mean value of each cluster.

3) Fitness computation: the fitness of each individual is calculated using:

$$\varpi(C_1, C_2, ..., C_k) = \sum_{l=1}^{k} \sum_{X_m \in C_l} \|X_m - C_l\|^2, l = 1,2,..., k \qquad (5)$$

where, ϖ is a within-cluster-distance metric to be optimized (minimized) and $C_1, C_2, ..., C_k$ are k cluster centers.

4) Parent-Centric Crossover (PCX): Generate λ offspring from the μ parents using the PCX recombination [7].

5) Fitness Calculation: the cluster centres and fitness values of the offspring are updated and re-calculated, as in Steps 2 and 3.

6) Parents to be replaced: choose two parents at random from the population *P*.

7) Replacement: from the combined subpopulation of two chosen parents and λ created offspring, choose the 'best' two solutions and replace the chosen two parents (Step 6) with the new solutions.

8) Iteration: Step 2 onwards is repeated for a pre-determined number of generations or until the standard deviation of the fitness values of the last five iterations becomes less than some threshold (*stable*), and the final solution is that with the smallest fitness value at the end of the execution.

Of note in the following case studies, is that all user-specified parameters are set to those suggested by [7] unless otherwise stated. Hence, P=100, λ=2, μ=3, *stable*=0.001.

A comparison of the performance between G3K and K-means clustering has been provided by [8], using a case study of a series of rundown vibration signatures from industrial gas turbines with feature extraction by the original PAA. A measure of the separation between the clusters, defined as the sum of the distance between any two identified cluster centres, is used, where a higher index indicates better clustering performance and will reduce the scope for misclassification. The results in Table 1 summarize the maximum, minimum, mean and standard deviation of the objective values obtained by these two clustering techniques, where 20 independent runs of each algorithm are used. It can be seen that G3K consistently produces the lowest objective values with no standard deviation, whilst the results of K-means are highly dependent on the initial conditions. Notably, G3K has converged in 8 generations, whilst only 1 out of the 20 runs has K-means correctly classified the rundowns associated with a known fault condition.

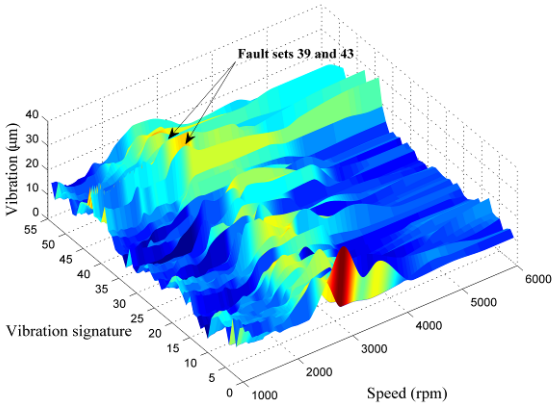
TABLE I
COMPARISON OF THE OBJECTIVE VALUES AND SEPARATION MEASURE FOR G3KMEANS AND K-MEANS CLUSTERING [8]

	Maximum	Minimum	Mean	Stand Deviation	Separation (mean)
G3K	8.9541	8.9541	8.9541	0	167.4564
K-means	18.5550	11.1193	14.4565	2.0177	114.5850

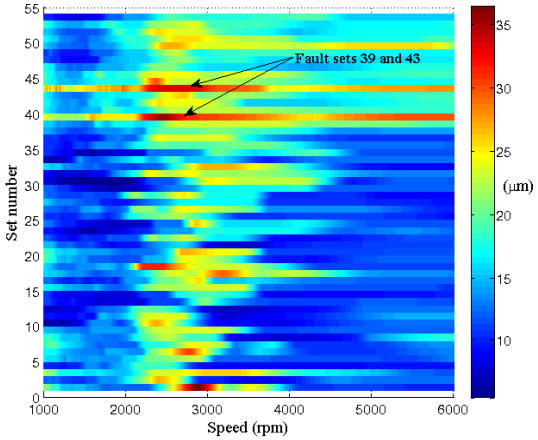
G3K is thereby shown to be more robust than original K-means clustering for this type of application, and is chosen as the pattern recognition technique to produce the final clustering results in this paper.

III. CASE STUDY

Vibration measurements taken from the rundowns of industrial gas turbines are considered information-rich for determining the health of the underlying units. To provide a consistent reference datum between the rundown series, only data for rotational speeds between 1000 rpm and 6000rpm are considered for comparison with [9]. Dimensionality reduction techniques, PAA and the hybrid HC-PAA, are used as a pre-processing step for subsequent clustering techniques for fault detection. A series of 54 rundown characteristics from a sub-15MW gas turbine is used in this case, which includes 52 'normal' rundowns and 2 rundowns that are associated with a known fault condition, sets 39 and 43, as shown in Fig. 3 (a) and (b). Vector representations of the 54 data sets are obtained using PAA and the hybrid HC-PAA separately, and then are applied to G3Kmeans to investigate the performance of the algorithms for fault/novelty detection. 10 segments for PAA and the hybrid HC-PAA, and a nominal cluster number of 4 for G3Kmeans are used in this case.



(a)

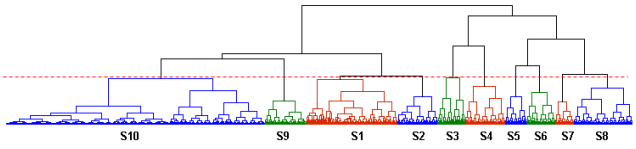


(b)

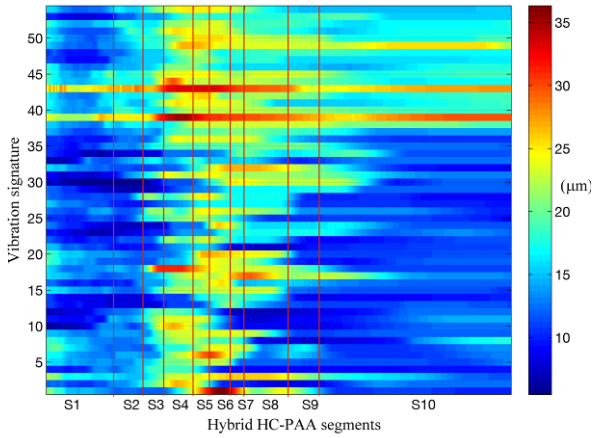
Fig. 3 (a) 3D plot and (b) 2D overview of 54 sets of vibration signatures

The HC results of the 5000 rotational speed samples are shown in Fig. 4(a), where the largest 10 clusters are selected according to a threshold of the HC distance measure. Each of the clusters is referred back to the original dataset, so that 10 segments for the 5000 speed samples can be found according to the HC dendrogram threshold, as shown in Fig. 4(b). It is seen that, through the HC-based modification, the regions that are of interested (the peak regions) are divided with higher density, whilst the regions of less interest, such as S10 in this case are connected up as a single segment. The samples of rotational speed in the segments of the original PAA, and the hybrid HC-PAA, are listed in Table 2 for comparison.

Final representations of the original PAA and the hybrid HC-PAA for the rundown vibration signatures are shown in Fig. 5(a) and (b) respectively. And the G3Kmeans clustering results for the extracted features from the original PAA and the hybrid HC-PAA are shown in Fig. 6(a) and (b) respectively, with the set numbers included in the clusters are also listed in the figures. The separation measure is the sum of the distance between any two cluster centres, which is calculated to be 167.46 for original PAA extracted features, and 182.04 for the hybrid HC-PAA extracted features. Although in both cases the faulted sets are clustered correctly, the hybrid HC-PAA features provide a higher separation index, which indicates improved clustering performance compared to original PAA and will reduce the scope for misclassification.

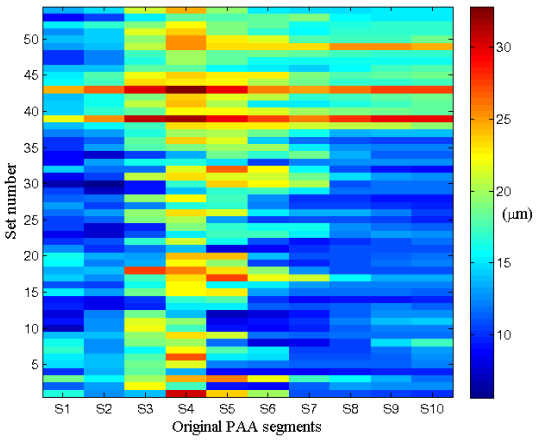


(a)

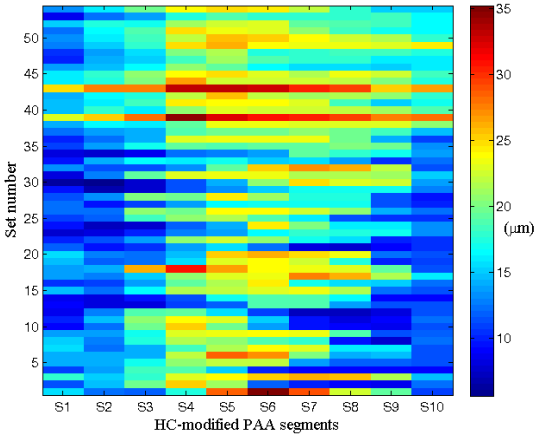


(b)

Fig. 4 (a) HC tree and 10 sub-clusters; (b) hybrid HC-PAA segments applied to the contour map of the vibration signatures (S=Segment)

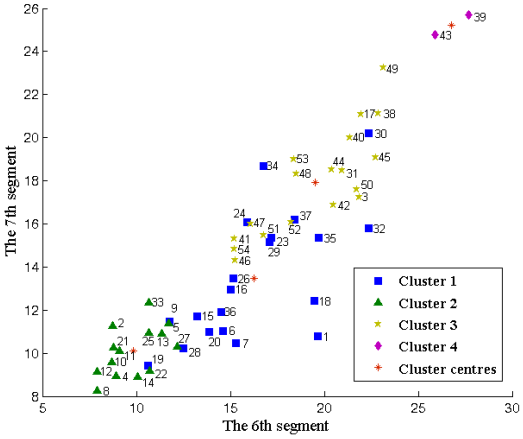


(a)

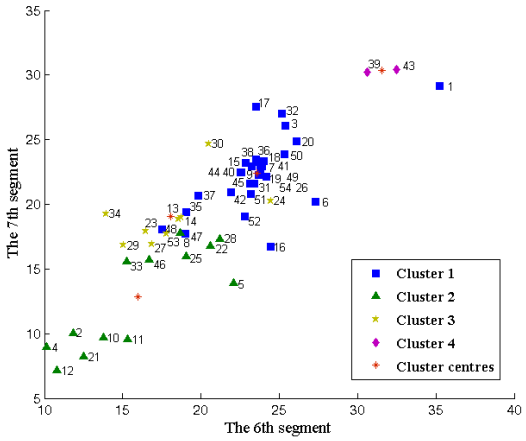


(b)

Fig. 5 (a) Original PAA representation and (b) HC-modified PAA representation of the rundown vibration signatures in a 10-dimensional hyperspace (S=Segment)



(a)



(b)

Fig. 6 G3Kmeans clustering results with (a) original PAA extracted features and (b) HC-modified PAA extracted features for rundown vibration signatures

TABLE II
SPEED SAMPLES (RPM) INCLUDED IN THE ORIGINAL PAA AND THE HYBRID HC-PAA SEGMENTS (S=SEGMENTS)

	S1	**S2**	**S3**	**S4**	**S5**
Original PAA	1001-1500	1501-2000	2001-2500	2501-3000	3001-3500
PAA-HC	1001-1725	1726-2045	2046-2265	2266-2585	2586-2757

	S6	**S7**	**S8**	**S9**	**S10**
Original PAA	3501-4000	4001-4500	4501-5000	5001-5500	5501-6000
PAA-HC	2758-2985	2986-3133	3134-3602	3603-3936	3937-6000

IV. CONCLUSIONS

This paper has presented a new method to improve the performance of traditional PAA by modifying the segment frame sizes through the use of HC. Using the resulting hybrid HC-PAA as a feature extraction methodology, pattern recognition is subsequently accomplished using a novel G3Kmeans clustering method due to performance benefits compared to K-means clustering. Experimental trials on an industrial gas turbine for fault detection have been used to demonstrate the efficacy of the technique. Results show that the hybrid HC-PAA provides improved the PAA feature extraction performance by increasing the cluster separation distances in order to reduce the chance of misclassification. The proposed HC-PAA-G3K methodology has been shown to provide a computationally efficient and robust method of feature extraction and novelty detection on datasets for a diverse spectrum of applications.

ACKNOWLEDGMENT

The authors would like to thank Siemens Industrial Turbomachinery, Lincoln, U.K., for providing access to on-line real-time data to support the research outcomes.

REFERENCES

[1] E. Keogh, M.J. Pazzani. “A simple dimensionality reduction technique for fast similarity search in large time series databases”. In *Proc. of Pacific Asia Conf. on Knowledge Discovery and Data Mining*, pp. 122-133. 2000.

[2] B.K. Yi, C. Faloutsos. “Fast time sequence indexing for arbitrary lp norms”. In *Proc. of Very large Data Bases*, pp. 385-394, San Francisco, CA, USA, 2000.

[3] J. Austin, R. Davis, M. Fletcher, T. Jackson, M. Jessop, B. Liang, A. Pasley. “DAME: searching large data sets within a grid-enabled engineering application”. In *Proc. of the IEEE*, vol. 93, No. 3, pp. 496-509, 2005.

[4] T. Hastie, R. Tibshirani, J. Friedman. “14.3.12 hierarchical clustering”. The elements of statistical learning, 2nd ed. New York: Springer. 2009.

[5] Y. Zhang, C.M. Bingham, M. Gallimore. “Applied sensor fault detection, identification and data reconstruction”. *Advances in Military Technology*, vol. 8, No. 2, pp. 13 – 26, 2013.

[6] J. Chen. Biologically inspired optimisation algorithms for transparent knowledge extraction allied to engineering materials processing, PhD. Thesis, University of Sheffield. 2009.

[7] K. Deb, A. Anand, D.A. Joshi. “Computationally efficient evolutionary algorithm for real-parameter optimization”, Evolutionary Computation, MIT Press, vol. 10, No. 4, pp. 371-395, 2002.

[8] J. Chen, M. Gallimore, C.M. Bingham, M. Mahfouf, Y. Zhang. “Intelligent data compression, diagnostics and prognostics using an evolutionary based clustering algorithm for industrial machines”, In: Fault detection: classification, techniques and role in industrial systems. Nova Science Publishers, Inc., NY, USA. 2014.

[9] M. Gallimore, M.J. Riley, C.M. Bingham, Z. Yang, Y. Zhang. “Intelligent diagnostics and prognostics for industrial machines using an optimization approach”. In *Proc. of 4th Int Conf on Integrity, Reliability and Failure*, pp. 223-224, 23-27 Jun. 2013.