24th COBEM - 2017

24th ABCM International Congress of Mechanical Engineering
December 3-8, 2017, Curitiba, PR, Brazil

COBEM-2017-0404
A GENETIC ALGORITHM BASED CLUSTERING APPLIED TO MULTIVARIATE TIME SERIES

Karine do Prado Ribeiro
Cristiano Hora de Oliveira Fontes
Federal University of Bahia, Politécnica School, Graduate Program in Industrial Engineering, Salvador, Brazil
karinepr@ufba.br, cfontes@ufba.br

Abstract. *This paper presents a method based on Genetic Algorithm (GA) and Fuzzy C-Means (FCM) for clustering multivariate time series. The method is applied to a real industrial case study which comprises pattern recognition for detecting operation failures in a gas turbine. The time series were collected from the Plant Information Management Systems (PIMS) and are associated with turbine starting events. In the proposed algorithm, each chromosome is an individual or solution, which encodes the clusters' centroids (patterns). A bi-criterion constrained clustering is proposed aiming to maximize both the similarity of objects in the same cluster (based on the SPCA metric) and the distance between the centers of the clusters. The proposed genetic algorithm obtained better results when compared to the traditional clustering method, the fuzzy c-means, according to the misclassification results. The recognized patterns (fault and normal operation) represent a potential for using in control systems or FDD (Fault Detection and Diagnostics) strategies, enabling the monitoring of the distance from the real process to the fault (or normal) operation condition.*

Keywords: *genetic algorithm, clustering, multivariate time series, pattern recognition, fault detection.*

1. INTRODUCTION

Data mining techniques aimed at acquiring knowledge from a database are widely used in several areas such as biology (Doyle et al., 2008), psychology (Henry and Tolan, 2005), image processing (Chen et al., 2008), climatology (Bisgin and Dalfes, 2008), information security (Kumar et al., 2011) and especially in industrial processes (Strachan et al., 2007, Bankó and Abonyi, 2012 , Izakian et al., 2015). Among these techniques, the data clustering comprises the recognition of similar unlabeled objects in a given sample and a representative pattern for each cluster of objects.

Among the non-hierarchical clustering, the rigid approach (such as the k-means algorithm) allows an object to belong totally to a given cluster and to have null membership degrees in relation to the others. Algorithms that allow an object to belong to different clusters with different membership degrees are called Fuzzy Clustering Techniques (Liao, 2005, Jain, 2010, Garai and Chaudhuri, 2004). Among these, the fuzzy c-means (FCM) algorithm (Bezdek, 2013) is the most used and represents the classic way to introduce uncertainty in the clustering problem.

A disadvantage of non-hierarchical clustering algorithms, based on classical optimization methods (k-means and fuzzy c-means), is the stabilization of the solution search process in a local minimum, which can result in unsatisfactory clustering quality. Another disadvantage is that the assessment of the quality of the clustering (or clusters) in each iteration is not carried out and, therefore, this information is not used in the following iterations (Rahman and Islam, 2014). In order to overcome these limitations, modified versions of classical algorithms (k-means, FCM) based on non-heuristic optimization, through the use of genetic algorithms (GA), have been proposed (Whitley, 1994, Maulik and Bandyopadhyay, 2000, Zhang et al., 2007, Bezdek and Hathaway, 1994, Maulik and Bandyopadhyay, 2003, Mukhopadhyay et al., 2009, Liu and Xie, 1995 and Sáez et al., 2008).

Inspired by natural evolution, GA has been shown to be an effective global optimization approach due to its ability to perform a probabilistic research in a large search space (Meng et al., 2002). The genetic algorithms aim to maintain and improve, over successive iterations, a set of individuals (chromosomes) that represent a population of possible optimal solutions. Each chromosome is evaluated according to a fitness function that comprises a metric for assessing the quality of the candidate solution. Through the fitness function it is possible to order the solutions by a quality criterion (fitness value). This fitness value is used to guide the selection of the most capable individuals who will be most likely to propagate their genetic information, that is, the best solutions will be used to generate even better solutions (Bandyopadhyay and Maulik, 2002).

Among the various types of objects that can be clustered, there is an increasing interest in the clustering of time series (Liao, 2005). This interest is due to the fact that several application areas such as industrial processes store information or knowledge through time series collected in historical databases (Abonyi et al., 2005). One of the intrinsic challenges to the time series clustering (univariate or multivariate) comprises the metric similarity choice. In the case of univariate series, the Euclidean distance is widely used (Xun and Zhishu, 2010, Wang et al., 2013). On the other hand, Euclidean distance is not recommended for multivariate time series (MTS) because these series are usually much more than a collection of univariate series (Bankó and Abonyi, 2012). Among the similarity metrics applied to multivariate series, the PCA (Principal Component Analysis) similarity metric (SPCA index) is a feasible and potential alternative that consists in quantifying the similarity between the directions of the principal components associated with different objects (Li and Wen, 2014).

There are still few works on the use of genetic algorithms in the clustering of time series. Baragona (2001), Liao et al. (2006) and Tseng et al. (2009) present different approaches for clustering of univariate series based on different representations of the chromosomes. However, there are no works on clustering and pattern recognition in multivariate time series based on genetic algorithms. All studies involving the use of genetic algorithms for clustering of time series are applied to univariate time series, regardless of the approach used for the chromosome coding.

This study proposes an approach based on genetic algorithm and fuzzy c-means for clustering multivariate time series. The approach is applied to a case study that comprises the detection of failures in a gas turbine of commercial scale (power generation capacity of 27 MW) which represents the main section of a thermoelectric unit belonging to the industrial park of the Brazilian Oil Company.

2. LITERATURE REVIEW

2.1 Fuzzy c-means

Fuzzy c-means (FCM) is a classical method suitable for clustering objects represented by univariate time series (Liao, 2005). The FCM algorithm uses the concept that each object belongs to the clusters with different degrees of membership (fractional value between 0 and 1) (Bezdek et al., 1984). The membership degrees of all objects are arranged in a partition matrix. FCM comprises an optimization-based clustering that consists of minimizing the sum of the distances of each object to each cluster center, weighted by its membership degree. In the case of univariate series, the Euclidean distance can be adopted as similarity metric (Liao, 2005).

Most clustering methods based on optimization and genetic algorithms are based on classical partitioning methods that include the k-means algorithm (Krishna and Murty, 1999, Meng et al., 2002), k-medoids algorithm (Estivill Castro et al., 1997) and FCM itself (Hall et al., 1999).

2.2 Genetic algorithm

Genetic algorithms (GAs) are part of a family computational models inspired by evolution. GAs encode a potential solution to a specific problem in a data structure called chromosome. In a broader use of the term, a genetic algorithm is any population-based model that uses selection and recombination operators to generate new possible solutions in a search space. Although the range of problems to which these algorithms have been applied is quite broad, a common application of GAs is as function optimizers (Whitley, 1994). GA has been shown to be an effective global optimization algorithm because of its ability of performing a probabilistic search in a large search space (Meng et al., 2002).

A genetic algorithm begins by creating an initial population of chromosomes and carries out the evolution towards an optimal solution through generations. Thus, the initial population is considered the first generation. A chromosome is formed by a collection of genes that represent parameters or decision variables. Genetic operators generate and alter the offspring composition while preserving some essential features. There are three of these operators: selection, crossover and mutation. The first operator (selection) makes the selection of the most capable individuals, based on the fitness values. The second operator (crossover) aims to mix the features of two individuals (parents) in order to generate two descendants (children), allowing a diversification in the solution space by generating different configurations. The last operator (mutation) changes one or more chromosome genes in a random way, usually with a small mutation rate, in order to avoid obtaining a local minimum (Baragona, 2001).

Once the genetic operators have been applied, the performance of the new individuals (new generation) is assessed after the crossover and mutation processes. The best individuals are selected and this process is repeated until a certain criterion is reached such as the maximum number of consecutive generations without solution change or the maximum number of generations. Each complete iteration is called generation (Chiou and Lan, 2001). Data clustering based on genetic algorithms explores the ability of GA to find clusters in the search space (Maulik and Bandyopadhyay, 2000). In order to implement the GA, regardless of application, it is necessary to choose the representation form of the chromosomes. Two basic coding formats have been commonly used in genetic algorithms. One of them is the binary scheme adopted by Liao et al. (2006), Hall et al. (1999), Chiou and Lan (2001), and the other is the codification by

using real numbers (Tseng et al., 2009, Wikaisuksakul, 2014, Maulik and Bandyopadhyay, 2000, Bandyopadhyay and Maulik, 2002, Liao, 2002).

Among the few approaches related to the use of genetic algorithms in the clustering of time series (Baragona, 2001, Liao et al., 2006, Tseng et al., 2009), different representations for the chromosomes are proposed. Baragona (2001) considers that the chromosome has the same number of genes equal to the number of objects and each gene receives a positive integer that establishes the cluster to which the respective object belongs in a given solution. Liao et al. (2006) propose that the chromosomes represent the centers of the clusters. The chromosome length is obtained by the predetermined number of clusters together with the number of digits used to represent each center (binary encoding). Tseng et al. (2009) store in each chromosome the result of a possible segmentation for a given time series. All these approaches use some measure of distance as a criterion to evaluate the similarity between two series and the Euclidean distance is the most commonly used criterion (Tseng et al., 2009).

Clustering and pattern recognition studies that use genetic algorithms in multivariate time series are not commonly found. This work proposes an approach for clustering of multivariate time series using genetic algorithm and based on the SPCA (PCA Similarity Factor) as similarity metric.

3. THE PROPOSED GENETIC ALGORITHM

The work comprised two main steps: obtaining the database and recognizing patterns of operation (clustering itself). The data were obtained from the Process Information Management System (PIMS) in the period from 2008 to 2010. The data are organized in multivariate time series. It has been analyzed 70 series (or objects) associated with turbine starting events. Three process variables were considered, namely, the flow of natural gas, the inlet temperature of the natural gas and the temperature of the exhaust gas. One of these objects (MTS) is presented in Fig. 1.

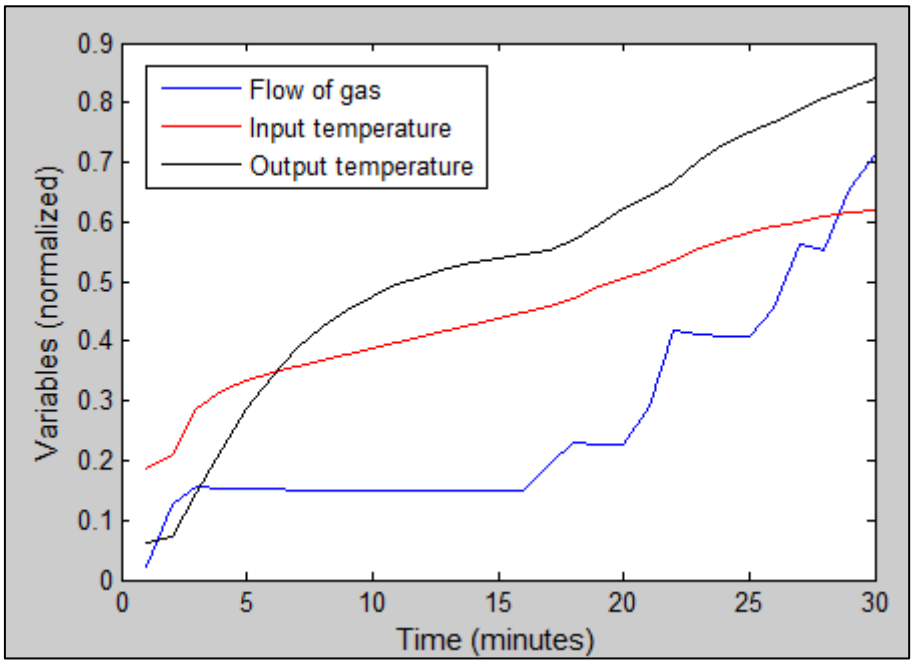

Figure 1. Multivariate time series

In order to perform the turbine data clustering, a genetic algorithm and fuzzy c-means based method was developed for clustering multivariate time series. The algorithm starts by the random generating of an initial population of 6 chromosomes (2 clusters, 3 times series associated with each one). Each chromosome is an individual, or solution, formed by a collection of genes, which represents the centers of the obtained clusters. The chromosome is coded by real numbers. The pre-specified number of clusters, along with the number of digits used to represent each cluster will determine the length of the chromosome. At each iteration, based on the fitness value of each chromosome, GA selects the best solutions to generate even better solutions, so that the next generation is always better than the previous one. According to the objective function of the classical FCM, the following fitness function was adopted:

$$J_m = \sum_{i=1}^{D}\sum_{j=1}^{N}\mu_{ij}^{m}\,||x_i - c_j||^2 + \frac{1}{\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j>i}}^{N}\|c_i - c_j\|^2} \tag{1}$$

Where D is the number of objects, N is the number of clusters, m is the fuzzification coefficient that controls the degree of overlap between clusters ($m > 1$), x_i is the *i*th object, c_j is the *j*th cluster and μ_{ij} is the membership degree of the *i*th object to the *j*th cluster.

In the case of univariate time series, the Euclidean distance is adopted as a metric similarity. However, this metric could not be used in this case study since considering that the objects are multivariate time series with different lengths (the time windows referring to the objects are not the same). The chosen index to measure the distance between each object and the center of a given cluster was the SPCA (Eq. 1) and, therefore, the similarity among the objects (MTS) is quantified by comparing the directions of its principal components (Li and Wen, 2014).

The fitness function presented in Eq. (1) comprises a second criterion that consists of maximizing the distances between the clusters centers. Therefore, the fitness value of each solution is based on the minimizing intra-cluster distances and maximizing the extra-cluster distance. The fitness value of each solution allows classifying the different solutions by a quality criterion.

Each object is classified into a given cluster based on the nearest centroid. Then the selection operator is applied. Such as in Liao et al. (2006) and Tseng et al. (2009), the roulette wheel selection is used to select the best individuals in order to create a new generation that fits better than the current one and the new generation always has the same amount of individuals as the current one. In this type of selection, each chromosome has a slot in the roulette with a size proportional to its suitability. Roulette is then "turned" and the higher fitness chromosomes are more likely to be propagated to the next generation. The next step comprises the crossover operator. Pairs of chromosomes are chosen randomly and two different solutions (parents) are merged to generate two descending solutions (children). In this way, the crossover diversifies the space of solutions by generating different configurations with a probability equal to 0.6. In order to perform this operation randomly, a single cut point p for the chromosomes is defined. To generate the first offspring (Chromosome 1'), the genes from the beginning of the chromosome up to p are copied from the first parent (Chromosome 1) and the remainder copied from the second parent (Chromosome 2). For the second child (Chromosome 2') the order is changed, as can be seen in Fig. 2.

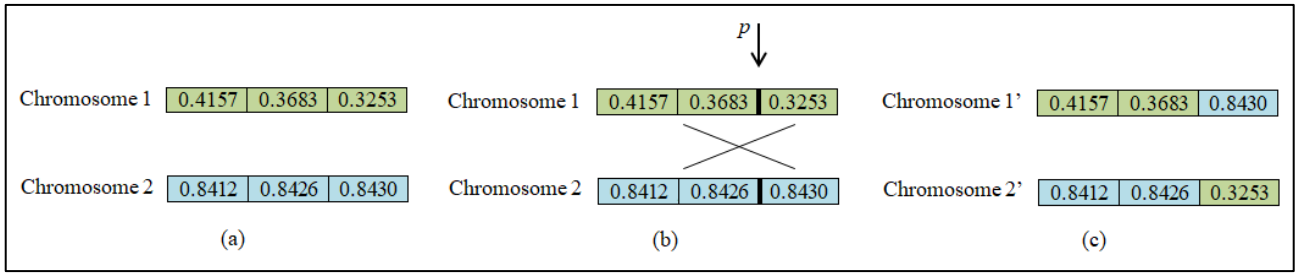

Figure 2. One-point crossover
(a) two individuals are chosen. (b) a crossover cut point is chosen.
(c) the features are merged, generating two new individuals.

At the end of the crossover, parents are replaced by their children in the population. The last operator (mutation) alters some genes of the chromosome in a random way with a probability equal to 0.02. This probability has been chosen so that the next generation is different from the current one, but it also preserves the main features of its parents. In addition, each gene was altered at a rate of 1% over the current value. In this way, it is avoided that the algorithm stays in a local minimum and, at the same time, there is a consistency of the obtained solutions in relation to the dynamic behavior of the variables. After the mutation, the fitness function is again applied to evaluate the new generation and then this is used as input to the next algorithm iteration. These processes are repeated until the solution remains unchanged by successive iterations. Figure 3 is a graphical outline of GA procedures developed.

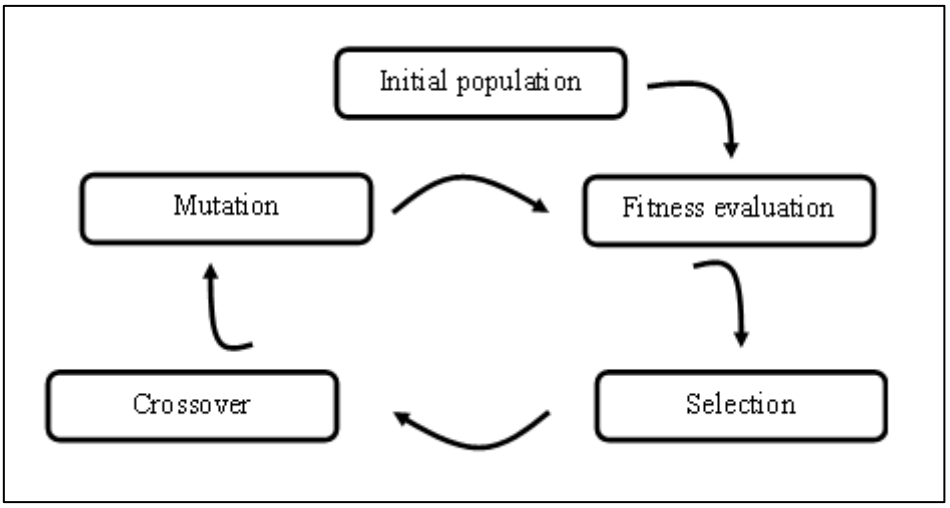

Figure 3. GA procedure

4. RESULTS AND DISCUSSION

Trips (failures) may occur in the turbines and they can be caused by specific factors such as surge, vibration and temperature dispersion. This work focuses on the detection of faults during turbine start-ups caused by temperature dispersion. Among the 70 multivariate time series (sample objects), 10 refer to the turbine failure and the others (60) refer to normal operation, which essentially characterizes an unbalanced sample with little information regarding fault behavior. When applying the genetic algorithm in the 70 time series, the result is reached when there is a stabilization in the centroids obtained, that is, when the centroids remain unchanged by successive iterations.

In order to avoid obtaining centers with excessive oscillations, change in each gene was limited to ± 1% in relation to the current value, even adopting a mutation rate equal to 0.02. This constraint allowed to obtain centers (patterns) more consistent with the dynamic behavior of the process (Figure 4).

The sample was partitioned into training (40 objects, 30 normal and 10 fault objects) and test (30 normal operation objects) data (Tab. 1). The clusters and centers were obtained using the training samples and the test data were used to validate the classification results.

Table 1. Data partitioning

Sample	Total failed operation	Total normal operation
Original	10	60
Training	10	30
Test	0	30

The results obtained with the proposed approach (GA-based clustering) were compared with the classical FCM, adapted to cope with multivariate time series, and based on classical (non-heuristic) optimization method. Tables 2 and 3 present the results of misclassifications in each case. Although 2 of the 10 fault objects can be considered as a possible second fault pattern (Fontes and Budman, 2017), the GA-based clustering was able to obtain better classification results for normal operation data, showing the reach of a local minimum with better clustering quality. This result was achieved even considering that the clustering problem is an unsupervised learning as there are no pre-labeled objects.

Table 2. Percentage of misclassifications – GA-based FCM

Sample	Operation with failure	Normal operation
Training	20%	**10%**
Test	—	3,3%

Table 3. Percentage of misclassifications – classical optimization

Sample	Operation with failure	Normal operation
Training	20%	**16,7%**
Test	—	3,3%

The recognized patterns (fault and normal operation) represent a potential for being used in control systems or FDD (Fault Detection and Diagnostics) that are able to monitor, in real time, the probability of equipment failure over time.

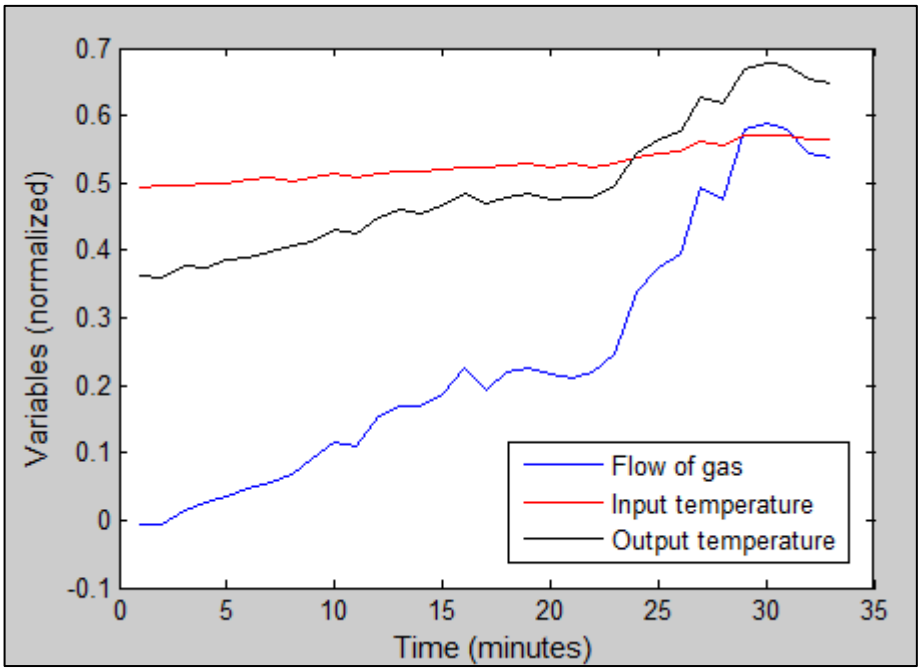

Figure 4. Pattern obtained by GA-based clustering (fault operation)

5. CONCLUSIONS

The potential of genetic algorithms for pattern recognition was investigated in multivariate time series. A database extracted from a real process was used and classification results demonstrated that the GA-based clustering was able to provide a better local minimum compared to the traditional FCM approach based on classical optimization (non-heuristic) methods.

Considering the lack of work coping with the application of GA-based clustering in multivariate times series, this paper presents a feasible proposal for chromosome coding, formulation of the fitness function and use of similarity metrics, among others.

The proposed strategy can be useful to support decision making at the operational level by predicting, in real time, possible occurrence of process failures, contributing to increase the production efficiency since prior detection of an undesirable operating condition allows corrective action to be taken.

6. REFERENCES

Abonyi, J., Feil, B., Nemeth, S., & Arva, P. (2005). “Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series”. *Fuzzy Sets and Systems*, 149(1), 39-56.

Bandyopadhyay, S., & Maulik, U. (2002). “Genetic clustering for automatic evolution of clusters and application to image classification”. *Pattern recognition*, 35(6), 1197-1208.

Bankó, Z., & Abonyi, J. (2012). “Correlation based dynamic time warping of multivariate time series”. *Expert Systems with Applications*, 39(17), 12814-12823.

Baragona, R. (2001). “A simulation study on clustering time series with metaheuristic methods”. *Quaderni di Statistica*, 3, 1-26.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). “FCM: The fuzzy c-means clustering algorithm”. *Computers & Geosciences*, 10(2-3), 191-203.

Bezdek, J. C., & Hathaway, R. J. (1994, June). “Optimization of fuzzy clustering criteria using genetic algorithms”. *In Evolutionary Computation*, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on (pp. 589-594). IEEE.

Bezdek, J. C. (2013). “Pattern recognition with fuzzy objective function algorithms”. *Springer Science & Business Media*.

Bisgin, H., & Dalfes, H. N. (2008). “Parallel clustering algorithms with application to climatology”. *In Geophysical Research Abstracts* (Vol. 10).

Chen, T. W., Chen, Y. L., & Chien, S. Y. (2008, October). "Fast image segmentation based on K-Means clustering with histograms in HSV color space". *In Multimedia Signal Processing*, 2008 IEEE 10th Workshop on (pp. 322-325). IEEE.

Chiou, Y. C., & Lan, L. W. (2001). "Genetic clustering algorithms". *European journal of operational research*, 135(2), 413-427.

Deng, X., & Tian, X. (2013). "Nonlinear process fault pattern recognition using statistics kernel PCA similarity factor". *Neurocomputing*, 121, 298-308.

Doyle, S., Agner, S., Madabhushi, A., Feldman, M., & Tomaszewski, J. (2008, May). "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features". *In Biomedical Imaging: From Nano to Macro*, 2008. ISBI 2008. 5th IEEE International Symposium on (pp. 496-499). IEEE.

Estivill-Castro, V., & Murray, A. T. (1997). "Spatial clustering for data mining with genetic algorithms". *Australia: Queensland University of Technology*.

Fontes, C., Budman, H., "A hybrid clustering approach for multivariate time series – A case study applied to failure analysis in a gas turbine", *ISA Transactions*, https://doi.org/10.1016/j.isatra.2017.09.004, 2017.

Garai, G., & Chaudhuri, B. B. (2004). "A novel genetic algorithm for automatic clustering". *Pattern Recognition Letters*, 25(2), 173-187.

Hall, L. O., Ozyurt, I. B., & Bezdek, J. C. (1999). "Clustering with a genetically optimized approach". *IEEE Transactions on Evolutionary computation*, 3(2), 103-112.

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). "Cluster analysis in family psychology research". *Journal of Family Psychology*, 19(1), 121.

Izakian, H., Pedrycz, W., & Jamal, I. (2015). "Fuzzy clustering of time series data using dynamic time warping distance". *Engineering Applications of Artificial Intelligence*, 39, 235-244.

Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". *Pattern recognition letters*, 31(8), 651-666.

Krishna, K., & Murty, M. N. (1999). "Genetic K-means algorithm". *IEEE Transactions on Systems*, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.

Kumar, P., Sehgal, V., Shah, K., Shukla, S. S. P., & Chauhan, D. S. (2011, December). "A novel approach for security in cloud computing using hidden markov model and clustering". *In Information and Communication Technologies (WICT)*, 2011 World Congress on (pp. 810-815). IEEE.

Li, S., & Wen, J. (2014). "Application of pattern matching method for detecting faults in air handling unit system". *Automation in Construction*, 43, 49-58.

Liao, T. W., Bolt, B., Forester, J., Hailman, E., Hansen, C., Kaste, R. C., & O'May, J. (2002, December). "Understanding and projecting the battle state". *In 23rd Army Science Conference*, Orlando, FL (Vol. 25).

Liao, T. W. (2005). "Clustering of time series data—a survey". *Pattern recognition*, 38(11), 1857-1874.

Liao, T. W., Ting, C. F., & Chang, P. C. (2006). "An adaptive genetic clustering method for exploratory mining of feature vector and time series data". *International Journal of Production Research*, 44(14), 2731-2748.

Liu, J., & Xie, W. (1995, March). "A genetics-based approach to fuzzy clustering". *In Fuzzy Systems*, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int (Vol. 4, pp. 2233-2240). IEEE.

Maulik, U., & Bandyopadhyay, S. (2000). "Genetic algorithm-based clustering technique". *Pattern recognition*, 33(9), 1455-1465.

Maulik, U., & Bandyopadhyay, S. (2003). "Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification". *IEEE Transactions on geoscience and remote sensing*, 41(5), 1075-1081.

Meng, L., Wu, Q. H., & Yong, Z. Z. (2002). "A genetic hard c-means clustering algorithm". DYNAMICS OF CONTINUOUS DISCRETE AND IMPULSIVE SYSTEMS SERIES B, 9, 421-438.

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2009). "Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes". *IEEE transactions on evolutionary computation*, 13(5), 991-1005.

Rahman, M. A., & Islam, M. Z. (2014). "A hybrid clustering technique combining a novel genetic algorithm with K-Means". *Knowledge-Based Systems*, 71, 345-365.

Sáez, D., Cortés, C. E., & Núñez, A. (2008). "Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering". *Computers & Operations Research*, 35(11), 3412-3438.

Strachan, S. M., Stephen, B., & McArthur, S. D. (2007, June). "Practical applications of data mining in plant monitoring and diagnostics". *In Power Engineering Society General Meeting*, 2007. IEEE (pp. 1-7). IEEE.

Tseng, V. S., Chen, C. H., Huang, P. C., & Hong, T. P. (2009). "Cluster-based genetic segmentation of time series with DWT". *Pattern Recognition Letters*, 30(13), 1190-1197.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). "Experimental comparison of representation methods and distance measures for time series data". *Data Mining and Knowledge Discovery*, 1-35.

Whitley, D. (1994). "A genetic algorithm tutorial". *Statistics and computing*, 4(2), 65-85.

Wikaisuksakul, S. (2014). "A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering". *Applied Soft Computing*, 24, 679-691.

Xun, L., & Zhishu, L. (2010, October). “The similarity of multivariate time series and its application”. I*n Management of e-Commerce and e-Government (ICMeCG)*, 2010 Fourth International Conference on (pp. 76-81). IEEE.

Zhang, J., Chung, H. S. H., & Lo, W. L. (2007). “Clustering-based adaptive crossover and mutation probabilities for genetic algorithms”. *IEEE Transactions on Evolutionary Computation*, 11(3), 326-335.

7. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.