ORIGINAL RESEARCH

Active cluster replacement algorithm as a tool to assess bifurcation early-warning signs for von Karman equations

Vasilii A. Gromov,* Igor M. Voronin, Vladislav R. Gatylo, Evgenii T. Prokopalo

Oles Honchar Dnepropetrovsk National University, Dnepropetrovsk, Ukraine

Received: February 3, 2017 **Accepted:** April 17, 2017 **Online Published:** May 1, 2017
DOI: 10.5430/air.v6n2p51 **URL:** https://doi.org/10.5430/air.v6n2p51

ABSTRACT

The paper deals with a novel algorithm used to improve identification quality of clusters generated by predictive clustering algorithm as a tool to identify states preceding to bifurcations for a system governed by von Karman equations. To construct bifurcation precursors, solutions (of the equations) observed on bifurcation paths are clustered; centers of the clusters constitute a set of bifurcation precursors. To decrease identification error rate, quality of each precursor is assessed with the employment of an additional, validation set. The paper concerns with two approaches to this procedure; the first one employs a single number to assess identification value of a cluster in order to delete those with low identification values. The second approach uses proposed knowledge extraction procedure to ascertain rules of replacement of the precursors chosen by the algorithm (active) by more efficient one. A wide-ranging simulation reveals that the best variant (provided that the Wishart clustering algorithm is utilized) is the replacement of the active cluster in conjunction local normalization of data. The optimal parameters values for both algorithms, arriving at essentially decreased identification errors.

Key Words: Early-warning signs, Von Karman equations, Predictive clustering, Cluster identification value

1. INTRODUCTION

On-line assessment of vulnerability for various systems under rapidly changing conditions makes it necessary to develop algorithms able to identify the observed state preceding to loss of functionality. In the paper[1] the problem in question is formulated as the problem to develop early-warning signs for bifurcations;[2] the same paper reports that typical sequences of post-bifurcation solutions can be used as the early-warning signs. To construct these typical sequences, it is possible to apply clustering algorithms similar to those used (in the framework of predictive clustering[3]) to ascertain clusters employed to predict a time series.[1] Here a training set incorporates sequences (of fixed length) of post-bifurcation solutions; the centers of such clusters are the early-warning signs.

As far as the problem at issue is supposed to be solved on-line, it is highly reasonable to estimate clusters' identification value and thereby reduce the number of clusters. The present paper is concerned with novel algorithm to assess identification values of the clusters in the frameworks of the inverse bifurcation problem (the problem to construct a set of early-warning signs).

Conventional way to appreciate the identification value of a cluster is to estimate somehow its efficiency using additional validation set; the result is a single value assigned

*Correspondence: Vasilii A. Gromov; Email: stroller@rambler.ru; Address: Oles Honchar Dnepropetrovsk National University, post box 2718, post office 44, Dnepropetrovsk, 49044, Ukraine.

to each cluster (its identification value). Nevertheless, it is not necessary to present information about clusters identification values using scalars; quite contrary, it is possible to employ, for example, logical rules indicating practicability of utilizing the cluster in question to identify under some conditions.

To compare different methods we utilize (along with a root-mean-squared-error) the number of non-predictable observations for a testing set[4,5] that are the observations of the algorithm unable to identify due to the absence of the appropriate cluster.

The rest of the paper is organized as follows. The next section reviews recent advances in the field; the third section formally states the problems under study; the fourth one outlines clustering method and a method to estimate clusters' identification values as well; the fifth section provides results. Finally, the last section presents conclusions.

2. Related works

Conventionally, predictive clustering researchers (irrespective they would like to forecast a time series or identify observable dynamics) pursue two avenues of inquiry;[3] the first one states that a series is an integral unit and it is possible to develop a single model to identify its subsequences of observations. The second one seeks for typical dynamical patterns (variously known as typical sequences,[4–6] motifs, chunks,[7] shapelets,[8] patterns, subsequences,[3] *etc.*) in a time series observed. In what follows, we restrict our attention to the second line of investigation.

Currently available methods of this line can be grouped in compliance with theories of artificial intelligence they use with four resultant groups.[9,10] The first group comprises various neural networks models which able to reveal and approximate local tendencies in observed data.[11,12] The second one includes fuzzy and neuro-fuzzy approaches employed to construct robust and logically transparent models of identification.[13] The third group is associated with distributed artificial intelligence, namely, with genetic algorithms,[14] with swarm intelligence,[15] with ant colony optimization[16] and with other algorithm belonging to this group.[17] Also, distributed artificial intelligence methods can be applied to determine weights for neural networks identification models.[18]

Finally, the last group utilizes clustering technique in order to employ centres of clusters as the typical patterns. A paper[19] is concerned with k-means in order to adjust it to seek for similar sections in chaotic time series; the modified algorithm is dubbed with TSkmeans (Time Series k-means). The paper[20] also considers k-means to predict chaotic time series; it summarizes results by various investigators for forecasting Australia's national electricity market prices; an extended version of the results may be found in Gromov and Borisenko.[4] The papers[21,22] analyze spatio-temporal data using a clustering technique grounded on the modified Euclidean distance capable of taking into account hidden space and time patterns. The article[23] examines ways to extract typical patterns from series amassed by a generating company; it is aimed at designing algorithms of rational energy consumption; the authors use various modifications of k-means.

A common disadvantage of the above-mentioned algorithms[24] is that they depend heavily on the distance function used; besides that it is often required to know the number of clusters before clustering. In a sense, the clustering techniques (including the Wishart clustering[9]) based upon theory of graphs/complex networks are able to overcome this limitations. For example, the paper[9] is concerned with the algorithm that maps the sequence of a series to graph vertices and then (in the framework of complex networks theory[25]) attempt to find its cliques.

Conventionally, studies of such algorithms deal with techniques to generate learning samples and with clustering algorithms, these constituents of identification method correspond to the concepts of adaptation to data and adaptation to algorithm.[26] The present paper explores an additional constituent to design efficient identification algorithm – estimate identification values of clusters (adaptation to identification procedure).

Furthermore, the problem to estimate identification value of the clusters as an auxiliary problem to identify states preceding to bifurcations is solved mainly in such a way that each cluster is assessed by a single value, in the present paper, we consider the algorithm that extracts a number of logical rules to ascertain conditions under which each cluster should be used instead.

3. Problem statement

Identification error is associated with incorrect choice of the active cluster that is the cluster engaged to identify the current observation. Identification values assessment problem involves selecting subset of clusters such that the total identification error on additional validation set is either minimum (the first statement) or less than a predefined threshold (the second statement). Mathematically, the problem is formulated as follows. Let Λ is the set of clusters employed to identify se-quences of the time series at issue; $\Im \equiv \left\{G : \Lambda \to R^1\right\}$; $\widetilde{\Lambda}(G, \beta) = \{\lambda \in \Lambda : G(\lambda) \geq \beta\}$. The problem is to find the estimator $G^* \in \Im$ and the threshold value $\beta^* \in R^1$, $\beta^* > 0$ (the first statement) in order to mini-mize prediction error (on the testing set):

$$\min I(\widetilde{\Lambda}(G, \beta)) \tag{1}$$

The second statement implies that one minimizes the number of clusters belonging to $\widetilde{\Lambda}(G, \beta)$:

$$\min \left|\widetilde{\Lambda}(G, \beta)\right| \tag{2}$$

under constraint

$$\left|I(\widetilde{\Lambda}(G, \beta))\right| \leq \gamma \tag{3}$$

where γ is a parameter of the algorithm. In the framework of the first statement, one places emphasis on the minimum identification error, while the second statement is concerned primarily with speed to obtain identification results.

To solve the problem, an additional – the validation – set is introduced under the assumption that it differs from both the training and testing ones, and all three of them are drawn from the same universal set.

4. IDENTIFICATION ALGORITHM

The proposed algorithm is subdivided into two parts. The first part analyzes a group of time series at hand in order to cluster sequences made of its observations according to predefined patterns and then to use cluster centres as typical sequences. The second estimates clusters' identification values and deletes clusters with low values.

4.1 Sample generation and clustering algorithms

The series are considered to be normalized. We used two different normalization techniques. The first one suggests that an entire time series is normalized with the employment of its maximum and minimum values, whereas the second technique implies that sample vectors are normalized separately, using their own maxima and minima. Hereinafter, we refer to these techniques as global (G) and local (L) respectively. The latter makes it possible to cluster not typical amplitudes (as it takes place for the former), but rather typical profiles.

To generate samples, we utilize a concept of pattern that is a preset sequence of distances between observations' positions which are to occupy the neighboring positions in a sample vector. The aforesaid algorithm is applied to generate samples with the employment of all possible patterns of four elements with the maximum (minimum) distance between neighboring positions in the pattern equal to 10. So the number of patterns used amounts to 10,000. Each pattern produces its own set of clusters, and then all sets of clusters are merged into a single set.

To cluster vectors, we employ the Wishart clustering[27] as modified by Lapko and Chentsov.[9] The method uses graph theory concepts and non-parametric probability density function estimator of k-nearest neighbors. Some problems associated with application of the algorithm are discussed in Gromov and Shulga.[6] The modification used is based upon non-parametric estimate of k-neighbour probability density function:

$$p(x) = \frac{k}{Vol_k(z)n} \tag{4}$$

Hereinafter $Vol_k(z)$ is a volume of the hypersphere of a radius $d_k(z)$ centred in z, which enclose at least k elements $z_i, i = \overline{1, n}$ of the sample to be clustered; $G(Z_n, U_n)$ is a similarity graph with vertices Z_n corresponding to sample's elements, while edges are defined as $U_n = \{d(z_i, z_j) \leq d_k(z_i), i \neq j\}$; $G(Z_i, U_i)$ is a subgraph such that $Z_i = \left\{z_j, j = \overline{1, n}\right\}$ and U_i is the subset of edges with final vertices belonging to Z_i.

A cluster c_l , $l > 0$ is called significant with respect to $h > 0$, if

$$\max \left\{|p(z_i) - p(z_j)| \, \forall z_i, z_j \in c_l\right\} \tag{5}$$

Algorithm 1 The Wishart clustering technique

```
1: Determine a distance $d_k(z_i)$ between each observation and its $k$-nearest neighbour and sort a sample in ascending order $d_k(z_i)$.
2: Let $w(z_i)$ is the number of class of observation $z_i$. Set $i = 1$.
3: The following variants are possible for a $G(Z_i, U_i)$:
   3.1. If $z_i$ is an isolated of $G(Z_i, U_i)$, then start form-ing a new cluster. Go to step 4.
   3.2. If $z_i$ is linked with vertices of the $l$-th cluster only, and the cluster is formed, then set $w(z_i) = 0$. Otherwise, set $w(z_i) = 1$. Go to step 4.
   3.3. If $z_i$ is linked with vertices of the clusters $l_1, l_2, \cdots, l_t, t > 1$.
      3.3.1. If all ($t$) clusters are formed, then set $w(z_i) = 0$. Go to step 4.
      3.3.2. If the number of significant clusters is $\xi(h) \leq t$.
         a) If $\xi(h) > 1$ or $l_1 = 0$, then set $w(z_i) = 0$, label significant clusters as formed, delete non-significant clusters, setting $w(z_i) = 0$ for all their elements.
         b) Else merge the clusters $l_2, \cdots, l_t$ with the cluster $l_1$, setting $w(z_i) = l_1$ for all their elements and $w(z_j) = l_1$.
4: Set $i = i + 1$. If $i \leq n$, then go to step 3.
```

4.2 The problem of estimating clusters' identification value

Two techniques to estimate the values in question are considered. The first (more conventional) one suggests that the identification value of k-th cluster is calculated as follows:

$$Q_k(\beta) = \sum_{i\in S_k} \frac{\overline{e_i}}{e_{ik}} \frac{1}{|V_i|},\ e_i = \frac{1}{|V_i|} \sum_{j\in V_i} e_{ij} \tag{6}$$

where V_i is a set of clusters able to predict i-th observation with an error less than β; S_k is the set of observations predicted by k-th cluster with an error less than β; e_{ij} is a prediction error for i-th observation if j-th cluster is used to identify.

The second method to perform quality assessment offers not to use a single characteristic, but rather to extract knowledge from data about prediction errors of algorithm for observations of the validation set.

We also define for j-th cluster (over the validation set):

d_{ij} is the minimum Euclidian distance between i-th observation and elements of j-th cluster; $S_j^{(d)}(\beta)$ is the number of observations with the distance less than β from the cluster j. m_j is the number of times the cluster has been active; n_j is the number of times the use of the cluster would lead to the minimum possible error.

Algorithm 2 The quality assessment routine with the replacement of the active cluster

```
1: Initialization: For each j, $S_i^{(d)}(\beta) \neq \varnothing$, $S_i(\beta) \neq \varnothing$,
   $m_j \leftarrow 0, n_j \leftarrow 0$.
2: $i \leftarrow 0$.
3: $j \leftarrow 0$.
4: If $d_{ij} < \beta$ then $S_i^{(d)}(\beta) = S_i^{(d)}(\beta) \cup x_i$.
5: If $e_{ij} < \beta$ then $S_i(\beta) = S_i(\beta) \cup x_i$.
6: Find $d_{imin} = d_{ik} = min_j\{d_{ij}\}; m_k = m_k + 1$.
7: Find $e_{imin} = e_{ip} = min_j\{e_{ij}\}$ and the distance of $d_{ip}$;
   $n_k = n_k + 1$.
8: $j = j + 1$ . If the list of clusters is not exhausted, then
   go to step 3.
9: $i = i + 1$. If the list of observations is not exhausted,
   then go to step 2.
```

In what follows, we refer to these algorithms as to 1st, and 2nd.

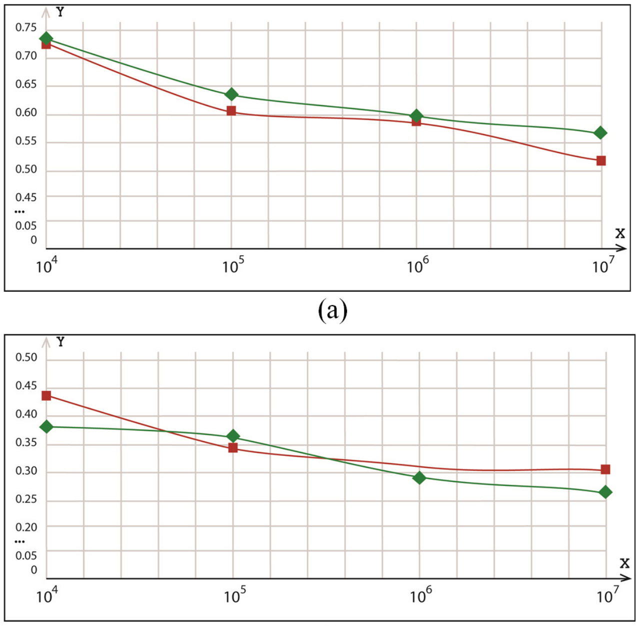
5. Numerical results

The method discussed in the previous section is applied to a time series constructed using solution sequences. To measure identification error, we used two measures. They are the percentage of uncorrected predictions and that of non-predictable observations. Both measures are averaged over the testing set which is used neither for training nor for quality assessment. The method under study was applied to the series observed on bifurcation paths (including primary and secondary ones) of non-linear boundary problem under study.

Table 1. Identification errors

Size	N	QA	Uncorr (%)	Non (%)
10^4	G	1	8.2	0.73
10^5	G	1	5.89	0.61
10^6	G	1	4.81	0.59
10^4	G	2	7	0.74
10^5	G	2	6. 18	0.64
10^6	G	2	3. 75	0.6
10^4	L	1	7.73	0.43
10^5	L	1	5.69	0.34
10^6	L	1	3.52	0.31
10^4	L	2	4.72	0.38
10^5	L	2	3.63	0.36
10^6	L	2	2.51	0.29

Note. Size is a size of training set; N is a normalization technique; QA is quality assessment algorithm; Uncorr is the percentage of uncorrected predictions; Non is the percentage of non-predictable observations.



(a)

(b)

Figure 1. The percentage of the non-predictable observations for series of solution norms for von Karman equations; (a) for the global normalization technique; (b) for the local normalization technique; red line with squares stands for the quality assessment method based upon a scalar estimate of clusters' identification value; green line with rhombi stands for the one based upon a replacement of active cluster.

The results obtained are presented in the following way. After introductory information about the sample, we present the identification errors for different method versions both graphically and in the form of a table. Table 1 shows identification errors corresponding to various choice of normalization, clustering, and quality assessment routines. The first column indicates a size of the validation set (sizes of training sets are identical and equal to 10^5 – the total number of clusters amounts to 263); the next two columns present information about the method used. Namely, the second and third columns correspond to a normalization technique (G is global and L is local), and a method to estimate clusters' identification values (quality assessment; 1 is the quality assessment method based upon a scalar estimate; 2 is the one based upon a replacement of the active cluster) respectively. The next two columns present the percentage of uncorrected predictions and that of non-predictable observations.

Figure 1 presents the same data graphically. Subfigures (a) and (b) show the percentage of the non-predictable observations (for global and local normalization technique) using the same notation.

The Wishart clustering technique in conjunction with local normalization routine and the quality assessment method based upon a scalar estimate of clusters' identification values proves the most efficient; however, it also proves the most time-consuming. Another point of interest is the fact that the percentage of the clusters to be discarded to obtain the best identification converges to a certain limit (around 19%) as size of the validation set increases.

6. Conclusions

Quality assessment procedure aimed at estimating clusters' identification values and at deleting clusters with low ones (in the framework of predictive clustering) decreases essentially identification error for series of solution norms for von Karman equations. The best variant appears to be the Wishart clustering algorithm in conjunction with local normalization and replacement of the active cluster.

References

[1] Obodan NI, Adlutskii VY, Gromov VA. Inverse bifurcation problem as a tool for rapid of progressive collapse for thin-walled systems. Proceeding of 5-th international conference "Nonlinear Dynamics". 2016: 356-61.

[2] Scheffer M, Bascompte J, Brock WA, et al. Early-warning signals for critical transitions. Nature. 2009; 461: 53-9. PMid:19727193. https://doi.org/10.1038/nature08227

[3] Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering – A decade review. Information Systems. 2015; 23: 16-38. https://doi.org/10.1016/j.is.2015.04.007

[4] Gromov VA, Borisenko EA. Chaotic time series prediction and clustering methods. Neural Computing and Applications. 2015; 2: 307-15.

[5] Gromov VA, Konev AS. Precocious identification of popular topics on Twitter with the employment of predictive clustering. Neural Computing and Applications. 2016 (In Press). https://doi.org/10.1007/s00521-016-2256-1

[6] Gromov VA, Shulga AN. Chaotic time series prediction with employment of ant colony optimization. Expert Systems with Applications. 2012; 39 (9): 8474-78. https://doi.org/10.1016/j.eswa.2012.01.171

[7] Palit AK, Popovich D. Computational intelligence in time series forecasting. Theory and engineering applications. NY: Springer; 2005.

[8] Widiputra H, Kho H, Pears R, et al. A novel evolving clustering algorithm with polynomial regression for chaotic time-series prediction. Neural Inf. Process. 2009; 5864: 114121. https://doi.org/10.1007/978-3-642-10684-2_13

[9] Lapko AV, Chentsov SV. Nonparametric information processing systems. Moscow: Science; 2000.

[10] Wang J, Chi D, Wu J, et al. Chaotic time series method combined with particle swarm optimization and trend adjustment for electricity demand forecasting. Expert Systems with Applications. 2011; 38: 8419-29. https://doi.org/10.1016/j.eswa.2011.01.037

[11] Fu YY, Wub CY, Jeng JT, et al. ARFNNs with SVR for prediction of chaotic time series with outliers. Expert Systems with Applications. 2010; 37: 37. https://doi.org/10.1016/j.eswa.2009.12.067

[12] Gan M, Peng H, Peng X, et al. A locally linear RBF network-based state-dependent AR model for nonlinear time series modeling. Information Sciences. 2010; 180: 4370-83. https://doi.org/10.1016/j.ins.2010.07.012

[13] Singh P, Borah B. High-order fuzzy-neuro expert system for time series forecasting. Knowledge-Based Systems. 2013; 46: 12-21. https://doi.org/10.1016/j.knosys.2013.01.030

[14] Mirzaee H. Linear combination rule in genetic algorithm for optimization of finite impulse response neural network to predict natural chaotic time series. Chaos, Solitons and Fractals. 2009; 41: 2681-89. https://doi.org/10.1016/j.chaos.2008.09.057

[15] Hong WC. Application of chaotic ant swarm optimization in electric load forecasting. Energy Policy. 2010; 38: 5830-39. https://doi.org/10.1016/j.enpol.2010.05.033

[16] Pan Y, Jiang JC, Wang R, et al. Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization. Journal of Loss Prevention in the Process Industries. 2011; 24: 85-9. https://doi.org/10.1016/j.jlp.2010.11.001

[17] Qin AK, Huang VL, Suganthan PN. Differential Evolution Algorithm with Strategy Adaptation for Global Numerical Optimization. IEEE Transactions on Evolutionary Computation. 2009; 13(2): 398-417. https://doi.org/10.1109/TEVC.2008.927706

[18] Donate JP, Li X, Sánchez GG, et al. Time series forecasting by evolving artificial neural networks with genetic algorithms, differential evolution and estimation of distribution algorithm. Neural Computing and Applications. 2013; 22(1): 11-20. https://doi.org/10.1007/s00521-011-0741-0

[19] Huang X, Ye Y, Xiong L, et al. Time Series k-Means: A New k-Means Type Smooth Subspace Clustering for Time Series Data. Information

Sciences. 2016; 367(368): 1-13. https://doi.org/10.1016/j.ins.2016.05.040

[20] Martınez-Alvarez F, Troncoso A, Riquelme JC, et al. Energy time series forecasting based on pattern sequence similarity. IEEE Trans. Knowl. Data. 2011; 23(8): 1230-43.

[21] Izakian H, Pedrycz W. Agreement-based fuzzy c-means for clustering data with blocks of features. Neurocomputing. 2014; 127: 266-80. https://doi.org/10.1016/j.neucom.2013.08.006

[22] Izakian H, Pedrycz W, Jamal I. Clustering spatiotemporal data: an augmented fuzzy c-means. IEEE Trans. Fuzzy Syst. 2013; 21(5): 855-68. https://doi.org/10.1109/TFUZZ.2012.2233479

[23] Benítez I, Díezb JL, Quijanoa A, et al. Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. Electric Power Systems Research. 2016 (In Press). https://doi.org/10.1016/j.epsr.2016.05.023

[24] Ferreira LN, Zhao L. Time series clustering via community detection in networks. Information Sciences. 2016; 326: 227-42. https://doi.org/10.1016/j.ins.2015.07.046

[25] Barrat A, Barthelemy M, Vespignani A. Dynamical processes on complex networks. ISBN: 9780521879507, Cambridge University Press; 2008. https://doi.org/10.1017/CBO9780511791383

[26] Liao TW. Clustering of time series data-a survey. Pattern Recogn. 2005; 38(11): 1857-74. https://doi.org/10.1016/j.patcog.2005.01.025

[27] Wishart D. A numerical classification methods for deriving natural classes. Nature. 1969; 221: 97-8. PMid:5782630. https://doi.org/10.1038/221097a0