Approach to the Analysis of the Multidimensional Time Series Based on the UMAP Algorithm in the Problems of the Complex Systems Proactive Maintenance

Liliya Anatolievna Demidova[1] and Maksim Anatolievich Stepanov[2]

[1]Department of Intelligent Systems of Information Security, Institute of Integrated Safety and Special Instrumentation

Russian Technological University – MIREA
Vernadsky Avenue,78, Moscow, 119454, Russia

liliya.demidova@rambler.ru

[2]Department of Computational and Applied Mathematics, Faculty of Computing Engineering

Ryazan State Radio Engineering University named after V.F. Utkin

Gagarin Str., 59/1, Ryazan, 390005, Russia

smastefan@gmail.com

UMAP ALGORITHM

UMAP algorithm (Uniform Manifold Approximation and Projection) is an algorithm of machine learning which implements non-linear dimension reduction. This algorithm is based on three assumptions:

- data is evenly distributed over the Riemannian manifold;
- Riemannian metric is locally constant or can be approximated as such;
- Riemannian manifold is locally connected.

The basic parameters of the UMAP algorithm are the number of the closest neighbors *n*, minimal distance *d* between the points in the new metric space, metric and dimension of the new metric space.

PROPOSED APPROACH

We apply the UMAP algorithm to visualize the multidimensional TSs in the 2-dimensional space: the data point of the original multidimensional space, where the number of dimensions is equal to the number of time series, should correspond to the data point in the 2-dimensional space. Then the coordinates of a certain point in the multidimensional space will be defined by the values of the time series elements at the certain time moment. The general number of points in the original multidimensional space, as well as in the 2-dimensional space, will be equal to the number of the time moments.

APPROBATION OF THE APPROACH TO THE ANALYSIS OF MULTIDIMENSIONAL DATASET

We tested our approach on the basis of the Predictive Maintenance Dataset (PMD). This dataset is in the public domain of NASA Ames Research Center and includes the training set with the information of the multidimensional time series generated by the readings of various sensors till the failure; the test set with data without registration of the failure; the GTD set (Ground Truth Data set) with information on the cycles left till the failure for each engine of the test set.

The training and test samples are created on the basis of the training and test datasets by adding the values of the auxiliary characteristic label that defines the class label (“0”, if the engine is operating normally, “1”, if the failure of the engine is occurred). The values of the auxiliary characteristics of the training samples are defined with consideration for information, that the last cycle determines the engine failure point, and in the test samples the auxiliary characteristics are defined on the basis of the GTD set.

APPROBATION OF THE APPROACH TO THE ANALYSIS OF MULTIDIMENSIONAL DATASET

Table 1 – Fragment of the test sample

id	cycle	setting_1	setting_2	setting_3	s1	s2	s3	…	s19	s20	s21	label
1	1	0,0023	0,0003	100	518,67	643,02	1585,29		100	38,86	23,3735	0
1	2	-0,0027	-0,0003	100	518,67	641,71	1588,45		100	39,02	23,3916	0
…												
7	23	-0,0004	0	100	518,67	642,09	1578,49		100	39,19	23,4363	0
7	24	-0,0021	-0,0003	100	518,67	642,04	1577,27		100	38,99	23,3413	0
…												
100	198	0,0013	0,0003	100	518,67	642,95	1601,62		100	38,7	23,1855	1

The fragment of the test sample is shown in the Table 1, where **id** is the engine identification number (**id**=1,…,100); **cycle** is the number of the operation cycle of the specific engine; **setting_1**, **setting_2**, **setting_3** are the degrees of the initial wear and production differences; **s1** – **s21** are the readings of the sensors during the operating of the engine during the operation cycle.

APPROBATION OF THE APPROACH TO THE ANALYSIS OF MULTIDIMENSIONAL DATASET

During the experiments we used, particularly, the information on the engine with **id**=100 of the test sample, which describes the last 50 cycles of the operation, 11 of which were recorded as the emergency ones (we should mention that the size of the window, which is equal to 50, was proposed to apply when developing the classifiers on the basis of LSTM network in).

UMAP algorithm was applied to two subsamples of the test sample for the engine with **id**=100: the first subsample included the information on 39 cycles that were recorded as the normal ones, and the second subsample included the information on all the 50 last cycles of the engine operation.

APPROBATION OF THE APPROACH TO THE ANALYSIS OF MULTIDIMENSIONAL DATASET

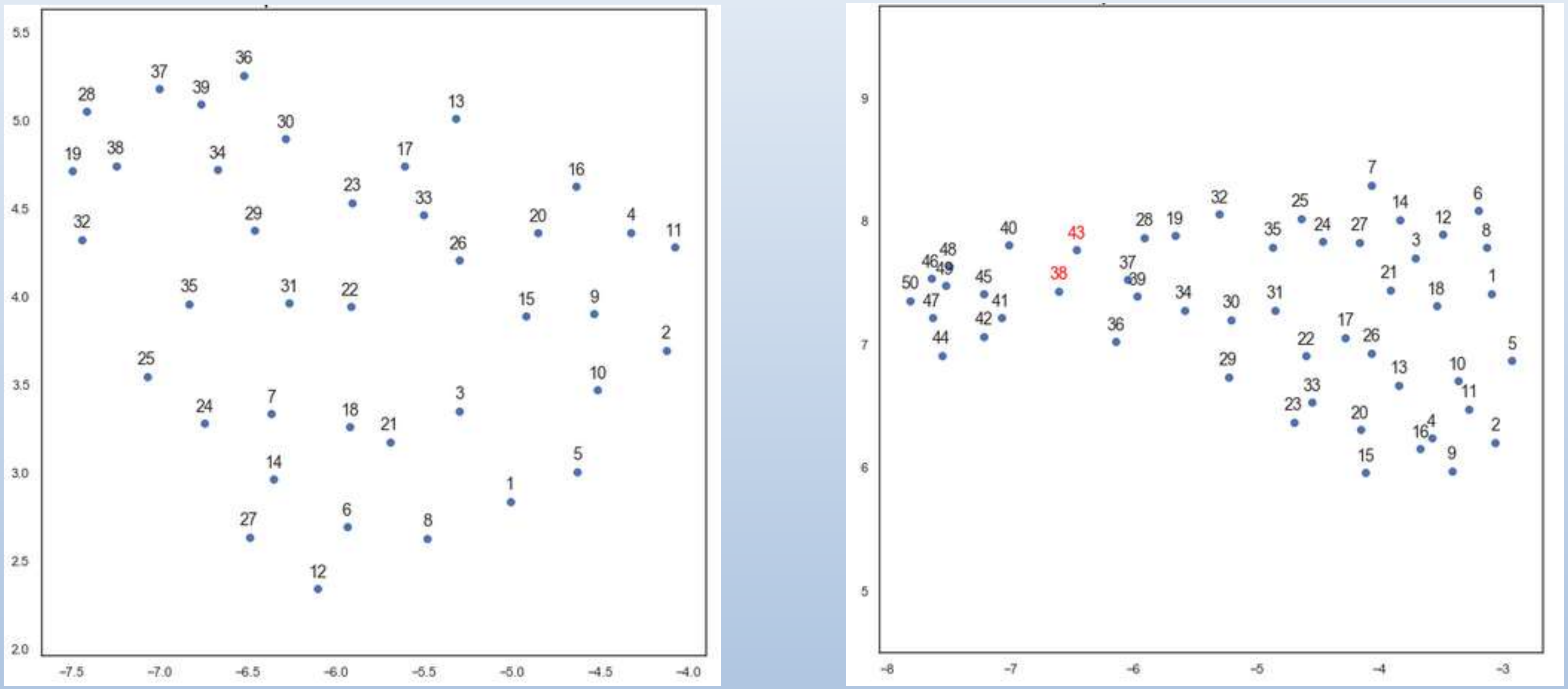

Figure 1 – Visualization of the aero engine states when *id*=100 by cycles:
a – when the number of cycles is equal to 39;
b – when the number of cycles is equal to 50

CONCLUSION

The proposed approach allows to execute the early prediction of abnormal states of the systems and can be recommended to apply while creating of the training and test datasets, which are necessary for development of the intellectual classifiers, which are required to analyze the multidimensional time series, arising in the problems of the technical and medical diagnosis, in the problems of the predictions of the social and economic systems’ states.