O. B. AKHIIEZER, G. L. GRINBERG, L. M. LYUBCHYK, K. S. YAMKOVYI

FAILURE RATE REGRESSION MODEL BUILDING FROM AGGREGATED DATA USING KERNEL-BASED MACHINE LEARNING METHODS

The problem of regression model building of equipment failure rate using datasets containing information on number of failures of recoverable systems and measurements of technological and operational factors affecting the reliability of production system is considered. This problem is important for choosing optimal strategy for preventive maintenance and restoration of elements of process equipment, which, in turn, significantly affects the efficiency of production management system. From a practical point of view, of greatest interest is the development of methods for regression models building to assess the impact of various technological and operational factors controlled during system operation on failure rate. The usual approach to regression models construction involves preselecting the model structure in the form of a parameterized functional relationship between failure rate and affecting technological variables followed by statistical estimation of unknown model parameters or training the model on datasets of measured covariates and failures. The main problem lies precisely in the choice of model structure, the complexity of which should correspond to amount of data available for training model, which in the problem of failure rate modeling is greatly complicated by lack of a priori information about its dependence on affecting variables. In this work, such a problem is solved using machine learning methods, namely, kernel ridge regression, which makes it possible to effectively approximate complex nonlinear dependences of equipment failure rate on technological factors, while there is no need to pre-select the model structure. Preliminary aggregation of data by combination of factor and cluster analysis can significantly simplify model structure. The proposed technique is illustrated by solving a practical problem of failure rate model building for semiconductor production equipment based on real data.

Keywords: dependability, factor analysis, failure rate, kernel methods, machine learning, production processes, preventive maintaince, ridge regression, regression model.

О. Б. АХІЄЗЕР, Г. Л. ГРІНБЕРГ, Л. М. ЛЮБЧИК, К. С. ЯМКОВИЙ

ПОБУДОВА РЕГРЕСІЙНОЇ МОДЕЛІ ІНТЕНСИВНОСТІ ВІДМОВ ЗА АГРЕГОВАНИМИ ДАНИМИ З ВИКОРИСТАННЯМ МЕТОДІВ ЯДЕРНОГО МАШИННОГО НАВЧАННЯ

Розглянуто задачу побудови регресійної моделі інтенсивності відмов обладнання з використанням наборів даних, що містять інформацію про кількість відмов відновлюваних систем та вимірювання технологічних і експлуатаційних факторів, що впливають на надійність виробничої системи. Ця проблема є важливою для вибору оптимальної стратегії профілактичного ремонту та відновлення елементів технологічного обладнання, що, у свою чергу, суттєво впливає на ефективність системи управління виробництвом. З практичної точки зору найбільший інтерес представляє розробка методів побудови регресійних моделей для оцінки впливу на інтенсивність відмов різноманітних технологічних та експлуатаційних факторів, контрольованих під час роботи системи. Звичайний підхід до побудови регресійних моделей передбачає попередній вибір структури моделі у формі параметризованого функціонального зв'язку між інтенсивністю відмов і технологічними змінними з наступною статистичною оцінкою невідомих параметрів моделі або навчанням моделі на наборах даних виміряних коваріат і відмов. Основна проблема полягає саме у виборі структури моделі, складність якої повинна відповідати кількості даних, доступних для навчання моделі, що в задачі моделювання інтенсивності відмов значно ускладнюється відсутністю апріорної інформації про її залежність від впливаючих змінних. У даній роботі така задача вирішується за допомогою методів машинного навчання, а саме ядерної гребневої регресії, що дає змогу ефективно апроксимувати складні нелінійні залежності інтенсивності відмов обладнання від технологічних факторів, при цьому немає необхідності попереднього вибору структури моделі. Попереднє агрегування даних шляхом поєднання факторного та кластерного аналізу може значно спростити структуру моделі. Пропонована методика проілюстрована розв'язанням практичної задачі побудови моделі інтенсивності відмов обладнання для виробництва напівпровідників на основі реальних даних.

Ключові слова: надійність, факторний аналіз, рівень відмов, ядерні методи, машинне навчання, виробничі процеси, прогнозне технічне обслуговування, хребтова регресія, регресійна модель.

Introduction. Among the problems of production management, an important place is occupied by preventive maintenance, which is an effective tool for reducing costs associated with failures and interruptions in production, its repair and restoration, as well as maintenance costs [1]. Preventive maintenance ensures to determine when maintenance should be carried out by estimations and predictions of degradation state of production items and expected time for occurrence of failures. Preventive maintenance is also play in important role for production and information and control system dependability ensuring.

In preventive maintenance implementation, failure rate is widely used as important metric for evaluating equipment reliability, wherein failure rate prediction is an effective way to maintaince decision making [2].

The problem of estimating and predicting failure rate using statistical data of failure streams has been studied in detail in the literature. From a practical point of view, of greatest interest is the development of methods for predicting failure rate using their models, which take into account the influence of technological and operational factors controlled during the operation of the system. This approach is implemented by building multivariate regression models to estimate the relationship between affecting factors (predictors or covariates) and failure rate. A common approach to building regression models involves preselecting of model structure as a parameterized functional relationship between the output variable and the affecting variables, followed by statistical estimation of unknown model parameters or model training model on

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022

51

datasets of measured covariates and outputs. The main problem lies precisely in the choice of model structure, the complexity of which should correspond to the amount of data available for training model, which in the problem under consideration is greatly complicated by the lack of a priori information about the dependence of failure rate from affecting factors.

In this paper, the problem of failure rate regression model building is solved in machine learning framework using kernel regression method, which makes it possible to effectively approximate nonlinear dependences of equipment failure rate from technological factors. The advantage of proposed approach is the possibility of effective approximation of complex failure rate multivariable dependence from affecting factors, while there is no need to pre-select model structure, complexity of which under such an approach is determined only by amount of training data available.

Preliminary data aggregation using factor and cluster analysis can significantly simplify regression model structure. At the same time, the use of learning on clusters method made it possible to significantly reduce the complexity of resulting model and ensure its smoothing to prevent the effect of "spatial oscillation".

The proposed technique is illustrated by solving a practical problem of failure rate model building for semiconductor production process based on real data.

State of the art. The problems of failure rate model building have been considered in detail in literature; the main attention was paid to the development of methods for constant or slowly changing over time failure rate estimating using only data on failure statistics, building models of the change in the failure rate over time and its trend prediction using time series models [3].

A further development of such an approach was development of methods for failure rate predicting using their models, which take into account various affecting technological and operational factors controlled during production system operation. This approach is implemented by building regression models of failure rate, the foundations of which were laid in fundamental work [4]. Subsequently, methods for failure rate regression models building were studied in works [5–6]. Regression analysis provides a set of statistical methods to estimate the relationship between affecting (explanatory) variables and failure rate as a model outcome. Explanatory variables (predictors or covariates) are variables which might affect a response variable, in reliability context, failure rate.

The usual approach to building regression models includes a preliminary selection of model structure in the form of a parameterized functional dependence between output variable and explanatory variables (covariates), followed by unknown models' parameters statistical evaluation or model training using datasets of measured covariates and failures. The main problem lies precisely in the choice of model structure, the complexity of which must be consistent with amount of data available for model training, which in the problem under consideration is greatly complicated by the lack of a priori information about the dependence of failure rate on affecting variables. Recently, to building regression models, methods of

machine learning and neural networks have begun to be used [7]. In this paper, to solve this problem, we use kernel-based machine learning method [8] developed to build high-dimensional multivariate nonlinear regression models.

Problem statement. Consider the problem of building a regression model of failure rate using dataset, containing information on the number of failures of recoverable systems in certain time intervals and the results of measurements of technological and operational factors affecting system reliability. A set of of restored items are considered, moreover, on each time interval n with length τ the number of failures $\{f_n(\tau),\ n=0,1,2,...\}$ is observed. At the same time, a set of technological and/or operational factors (covariates in terminology of reliability regression analysis) $\boldsymbol{x}^{\mathrm{T}}=[x^1\ x^2\ ...\ x^N]$, that affect the failure rate $\lambda(\boldsymbol{x})$ is measured; measurements are made at successive times instances $\boldsymbol{x}(t_k)$ during a certain observation time interval.

It is assumed, that at each observation interval affected factors relative variations are insignificant and they can be replaced with a sufficient degree of accuracy by their average value over the interval, thus $\boldsymbol{x}_n^{\mathrm{T}}=[x_n^1\ x_n^2\ ...\ x_n^N]$ will be considered as feature vector.

We will interpret the number of failures at current observation interval as the results of "measurements" the failure rate for corresponding value of feature vector at this interval. In this way, the problem of failure rate model building is reduced to restoring unknown dependence $\lambda(\mathbf{x})$, using measurements dataset $\{f_n, \boldsymbol{x}_n\}_{n=1}^M$, where M is a number of observations intervals. In what follows, we will denote $\boldsymbol{X}_M^{\mathrm{T}}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ ...\ \boldsymbol{x}_M]$ the matrix of feature measurements (training dataset) over entire observation.

Let's take failure rate multivariable regression model in parametrized quasi-linear form

$$\begin{aligned}\lambda(\boldsymbol{w},\boldsymbol{x})&=\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{w},\\\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x})&=[\varphi^1(\boldsymbol{x}),\varphi^2(\boldsymbol{x}),...,\varphi^N(\boldsymbol{x})],\end{aligned}\tag{1}$$

where $\boldsymbol{w}^{\mathrm{T}}=[w^1\ w^2\ ...\ w^N]$ – model' parameters vector;

$\boldsymbol{\varphi}(\boldsymbol{x})$ – vector of coordinate functions,

T – sign of transposition.

Then measurement model taking into account (1) in turn can be represented as

$$y_n=\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_n)\boldsymbol{w}+\varepsilon_n,\quad n=\overline{1,M},\tag{2}$$

where $y_n=f_n/\tau$ – number of failures per time unit in n-th observation interval,

ε_n – failure rate measurement error, which is, in fact, the deviation of recorded number of failures on time interval from their predicted by model (1) value.

In matrix-vector form measurement equation (2) may be represented as

52

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022

$$y_M = \mathbf{\Phi}_M w + \varepsilon_M,$$
$$\mathbf{\Phi}_M^{\mathrm{T}} = [\varphi(\pmb{x}_1)\, \varphi(\pmb{x}_2)\,...\,\varphi(\pmb{x}_M)], \qquad (3)$$

where $y_M^{\mathrm{T}} = [y_1\, y_2\, ...\, y_M]$ and $\varepsilon_M^{\mathrm{T}} = [\varepsilon_1\, \varepsilon_2\, ...\, \varepsilon_M]$ are vectors of measurements and errors respectively.

Using the ordinary least squares (OLS) method, model (1) parameters estimate can be obtained as a solution of following optimization problem:

$$(y_M - \mathbf{\Phi}_M \cdot w)^{\mathrm{T}}(y_M - \mathbf{\Phi}_M \cdot w) \to \min_w, \qquad (4)$$

which is well known $\hat{w}_M = (\mathbf{\Phi}_M^{\mathrm{T}} \mathbf{\Phi}_M)^{-1} \mathbf{\Phi}_M^{\mathrm{T}} y_M$.

Meanwhile, the use of such traditional approach in the case under consideration encounters significant difficulties connected with the necessity of rational choice of coordinate functions $\varphi(\pmb{x})$ and their number, which, in the absence of a priori information about unknown restored dependence, is a difficult problem. In addition, the possible ill-conditionality of measurement's matrix $\mathbf{\Phi}_M^{\mathrm{T}} \mathbf{\Phi}_M$, which occurs with a relatively slow change in affected factors, leads to a significant increase in estimation error.

Materials and methods. In this regard, to solve the problem of failure rate model building, it is advisable to use kernel-based machine learning method [8], which removes the problem of choosing of a system of coordinate functions. Let us choose coordinate function such that their scalar products $\varphi^{\mathrm{T}}(\pmb{x}_i) \cdot \varphi(\pmb{x}_j) = \kappa(\pmb{x}_i, \pmb{x}_j), \quad i, j = \overline{1, M}$ are kernel functions; for definiteness, choose a system of kernels in the form of radial basis functions (RBF)
$\kappa(\pmb{x}, \pmb{x}^*) = \exp(-\sigma \cdot (\pmb{x} - \pmb{x}^*)^{\mathrm{T}}(\pmb{x} - \pmb{x}^*)), \ \sigma > 0.$

In accordance with kernel-based learning technique, introduce kernel matrix $\mathbf{K}_n = \mathbf{\Phi}_M \mathbf{\Phi}_M^{\mathrm{T}}$, which may be found directly without reference to coordinate vector functions, because $\mathbf{K}_n = [k_{ij}], \ k_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \ i, j = \overline{1, M}$.

We find model (1) parameters, in according to OLS kernel approach [8] by regularized functional minimization with linear constraints

$$\varepsilon_M^{\mathrm{T}} \varepsilon_M + \gamma w^{\mathrm{T}} w \to \min_{w, \varepsilon}, \quad \varepsilon_M = y_M - \mathbf{\Phi}_M w, \qquad (5)$$

where $\gamma > 0$ – regularization parameter.

Introduce Lagrange function

$$\begin{aligned} \Lambda(\varepsilon_M, w, \pmb{\mu}) = \\ = \varepsilon_M^{\mathrm{T}} \varepsilon_M + \gamma w^{\mathrm{T}} w + \pmb{\mu}^{\mathrm{T}}(\varepsilon_M - y_M + \mathbf{\Phi}_M w), \end{aligned} \qquad (6)$$

we obtain optimization problem (5) solution in the form

$$\hat{w}_M = \mathbf{\Phi}_M^{\mathrm{T}} (\gamma \mathbf{I}_M + \mathbf{\Phi}_M \mathbf{\Phi}_M^{\mathrm{T}})^{-1} y_M. \qquad (7)$$

Taking into account that

$$\begin{aligned} \mathbf{\Phi}_M \mathbf{\Phi}_M^{\mathrm{T}} = \pmb{K}_n = [\kappa(\pmb{x}_i, \pmb{x}_j)], \\ \varphi^{\mathrm{T}}(\pmb{x}) \cdot \mathbf{\Phi}_M^{\mathrm{T}} = [\kappa(\pmb{x}, \pmb{x}_1)\, \kappa(\pmb{x}, \pmb{x}_2)\, ...\, \kappa(\pmb{x}, \pmb{x}_M)] = \pmb{\kappa}_M^{\mathrm{T}}(\pmb{x}), \end{aligned} \qquad (8)$$

we get final expression for failure rate model $\hat{\lambda}_M(\mathbf{x})$ obtained by training data $\{f_n, \mathbf{x}_n\}, n = \overline{1, M}$

$$\hat{\lambda}_M(\pmb{x}) = \pmb{\kappa}_M^{\mathrm{T}}(\pmb{x})(\gamma \pmb{I}_M + \pmb{K}_M)^{-1} y_M. \qquad (9)$$

Note that regression model (9) depends only from elements of kernel matrix $\mathbf{K}_M$ and vector $\pmb{\kappa}_M(\pmb{x})$, calculated at the points of training sample and does not require a preliminary choice of coordinate functions. Regularization ensures stability of computational procedure even for ill-conditioned kernel matrix $\mathbf{K}_M$.

Using kernel approach allows approximating complex multidimensional dependencies, however, since the number of kernels used is determined by number of points in training dataset, the resulting restored dependency may have the property of "spatial oscillation". To prevent this undesirable effect, it is advisable to carry out additional smoothing of restored dependence on data cloud.

To do this, we use an additional regularization of optimization problem (5), using *data graph model* [9], described by adjacent graph, consists of nodes, that are all observed data points, and matrix $\pmb{V} = [v_{ij}], \ i, j = \overline{1, M}$ determines edge weights v_{ij} in adjacency graph, wherein the edge, connecting two data graph nodes $(\pmb{x}_i, \pmb{x}_j)$, is weighted by kernel function

$$v_{ij} = \kappa(\pmb{x}_i, \pmb{x}_j) = \exp\{-\omega \|\pmb{x}_i - \pmb{x}_j\|^2\}, \quad \omega > 0, \quad (10)$$

defined over graph nodes [10] and reflect data cloud geometric structure.

In this way, additional regularization term, smoothing the evaluated function $\lambda(\pmb{w}, \pmb{x})$ on training dataset with edge weights v_{ij} is the following:

$$\pmb{\Omega}(\pmb{w}, X_M) = \sum_{i, j=1}^{n} v_{ij}(\lambda(\pmb{w}, \pmb{x}_i) - \lambda(\pmb{w}, \pmb{x}_j))^2. \qquad (11)$$

Thus, graph regulariser (10) ensures to find restored function $\lambda(\pmb{w}, \pmb{x})$ which has similar values for close data points from training dataset.

In accordance with [11] approach, let's take manifold regularization term as data depended semi-norm

$$\begin{aligned} \pmb{\Omega}(\pmb{w}, X_M) = \pmb{\lambda}^{\mathrm{T}}(\pmb{w}, X_M) \pmb{L} \lambda(\pmb{w}, X_M), \\ \lambda(\pmb{w}, X_M) = [\lambda(\pmb{w}, \pmb{x}_1), \lambda(\pmb{w}, \pmb{x}_2), ..., \lambda(\pmb{w}, \pmb{x}_M))]^{\mathrm{T}}, \end{aligned} \qquad (12)$$

where $\pmb{L}$ – graph Laplacian matrix,

$$\pmb{L} = \pmb{D} - \pmb{V},$$

$$\pmb{D} = \mathrm{diag}[d_{11}\, d_{22}\, ...\, d_{MM}], \ d_{jj} = \sum_{i=1}^{M} v_{ij}. \qquad (13)$$

At that, to construct a smoothing regulariser, we will use data-depended kernels, corresponding to semi-norm $\pmb{\Omega}(\pmb{w}, X_M)$; explicit form of transformed reproducing kernel $\tilde{\kappa}(\pmb{x}, \pmb{x}')$, derived in [11], are given by:

$$\begin{aligned} \tilde{\kappa}(\pmb{x}, \pmb{x}') = \kappa(\pmb{x}, \pmb{x}') + \\ + \pmb{k}_M^{\mathrm{T}}(\pmb{x})(\pmb{I}_M + \pmb{L}\pmb{K}_M)^{-1}\pmb{L}\pmb{k}_M(\pmb{x}), \end{aligned} \qquad (14)$$

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022

53

where $\boldsymbol{K}_M, \boldsymbol{k}_M(\boldsymbol{x})$ defined by (8).

Such a choice of kernel functions implements the smoothness assumption with respect to training dataset geometric structure.

Obtained transformed kernels can be used to modify failure rate kernel-based regression model:

$$\tilde{\boldsymbol{w}}_M = \boldsymbol{\Phi}_M^{\mathrm{T}} (\gamma \boldsymbol{I}_M + \tilde{\boldsymbol{K}}_M)^{-1} \boldsymbol{y}_M ,$$

$$\tilde{\lambda}_M(\boldsymbol{x}) = \tilde{\boldsymbol{\kappa}}_M^{\mathrm{T}}(\boldsymbol{x})(\gamma \boldsymbol{I}_M + \tilde{\boldsymbol{K}}_M)^{-1} \boldsymbol{y}_M , \qquad (15)$$

$$\tilde{\boldsymbol{\kappa}}_M^{\mathrm{T}}(\boldsymbol{x}) = [\tilde{\kappa}(\boldsymbol{x}, \boldsymbol{x}_1)\ \tilde{\kappa}(\boldsymbol{x}, \boldsymbol{x}_2)\ ...\ \tilde{\kappa}(\boldsymbol{x}, \boldsymbol{x}_M)],$$

where $\tilde{\boldsymbol{K}}_M = [\tilde{\kappa}(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ is transformed kernel matrix.

Thus, dependence (15) describes obtained smoothed kernel-based regression model, that determines relationship between failure rate and affecting factors.

Case study. The proposed technique was investigated using real data from semiconductor manufacturing industry, which was provided by Michael McCann and Adrian Johnston [12]. The study uses Semiconductor Manufacturing (SECOM) dataset available from University of California, Irvine Machine Learning Repository UCI Machine Learning Repository: SECOM Data Set. The SECOM dataset contains data collected during semiconductor manufacturing process and represents a series of monitoring signals from sensors at process measurement points. Each example of SECOM dataset represents a single production entity with associated measured features (affected technological factors), the labels represent a simple pass/fail Yield in production quality control (Yield = −1, pass, no failure, and Yield = +1, fail), and associated date time stamp.

Training dataset contains 1567 examples, each with 590 measured features, and only 104 themes represent yield failures.

Two datasets are provided [12]:

– secom_data.csv, contains time series of features measurements from sensors;

– secom_labels.csv, contains time units and for each time unit output result (yield pass/fail).

– Calculations for numerical experiments were carried out using free software machine-learning library Scikit https://scikit-learn.org/stable/.

The presence of a large number of observations and their heterogeneous structure and statistical properties predetermine the need of preliminary data cleaning and preprocessing.

Some results of preliminary analysis of SECOM data are illustrated in fig. 1. Proportion of failures in total set of observations (fig. 1, *a*) indicates, that dataset is unbalanced, there are only 7% "failed" items. Analysis of statistics of failures distribution at fixed time intervals (fig. 1, *b*) confirms the hypothesis of its Poisson distribution, and founds the assumption of failure rate constancy at certain time intervals.

Analysis of empirical distributions of affecting technological factors showed that in the most cases features are distributed approximately normally, for example, histograms visualization of measurements from first 9 sensors are presented at fig. 2.
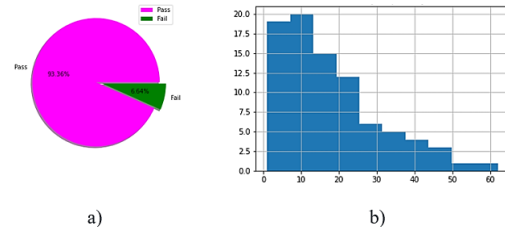

Fig. 1. Training dataset properties:
a – diagram of proportion of failures in total observations;
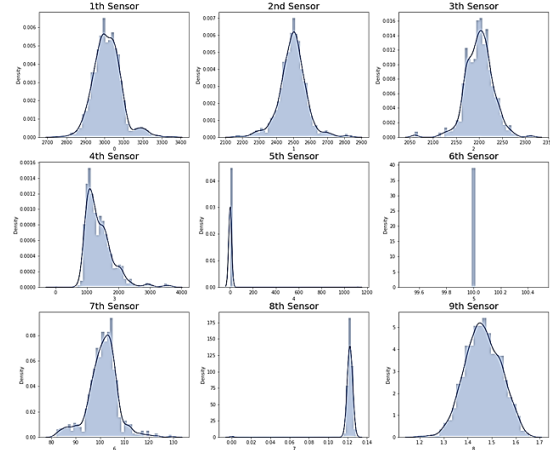*b* – histogram of the daily distribution of failures

Fig. 2. Measured signals from first 9 sensors histograms

Features covariance matrix heatmap is presented at fig. 3, the number of factors having positive or negative covariance values of more than 70% is equal to 262. At the stage of data cleaning, pairs of features with a covariance of more than 70% were removed; after this manipulation, we get a total of 212 features.
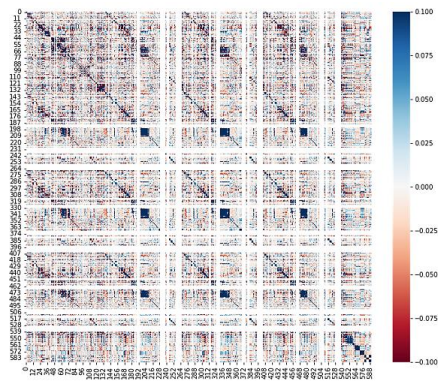
Fig. 3. Features covariance matrix heatmap

Preliminary analysis showed that training dataset has significant redundancy and contains many collinear feature vectors. Thus, there is a need to aggregate data, which is a prerequisite to reduce model complexity and training time and increase model performance.

Data aggregation was carried out in three stages. At *the first stage*, the entire set of observations was divided into successive intervals, for each of which the observed number of failures was fixed, and the feature vectors were replaced by their current average value over the intervals.

54

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022

At the *second stage*, the reduction in the number of features was performed by method of factor analysis, which performs a linear transformation of original feature vectors into the space of hidden factors of a smaller dimension. Experimentally, the number of hidden factors, which aggregate original features, was chosen as 5, the structure of correlations between the initial features and found hidden factors is illustrated by corresponding covariance matrix, which is shown in fig. 4.
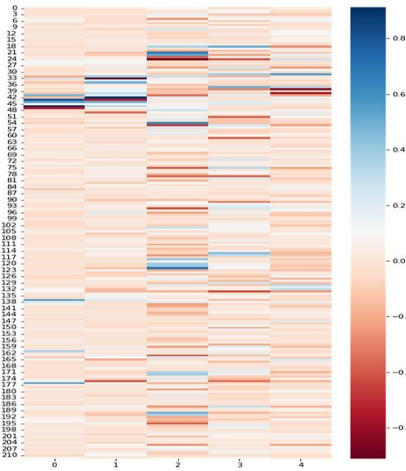

Fig. 4. Features-factors covariance matrix heatmap

Fig. 4 clearly shows, that each of hidden factors correlates well with a certain group of initial features, while obtained factors themselves do not correlate (fig. 5).
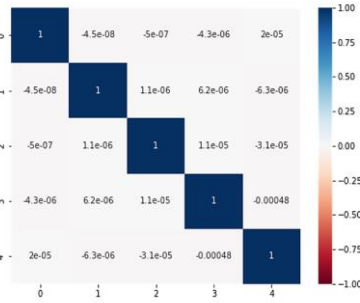

Fig. 5. Hidden factors covariance matrix

At the *third stage* of preprocessing, the clustering of reduced dataset in factor space was carried out in order to reduce the number of kernel functions in resulting failure rate model, while centers of clusters were chosen as centers of the kernel components of regression model, and average values of number of failures for each cluster were used as measured response function. At that, the construction of a regression model on clusters in hidden factors space provides its additional smoothing as a result of number reduction of kernel functions used.

On fig 6, 7 the results of model training are presented, regression model surface is given in projections on axes corresponding to hidden factors (a – X1-X2 projection, b – X1-23 projection, c – X1-X4 projection, d – X1-X5 projection). Fig 6 shows visualization of base regression model, trained on dataset constructed from original

features, averaged over the intervals, fig. 7 shows cluster-based regularized smoothed model.
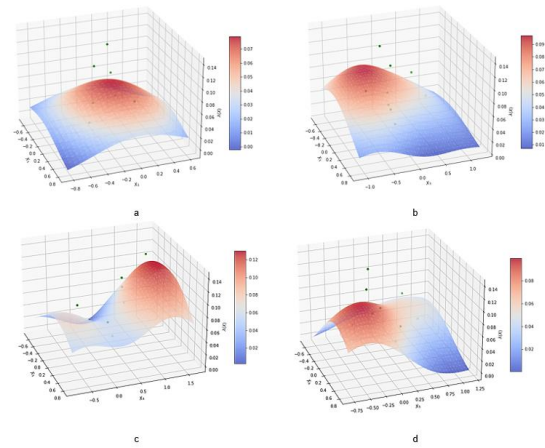
Fig. 6. Failure rate base regression model visualization
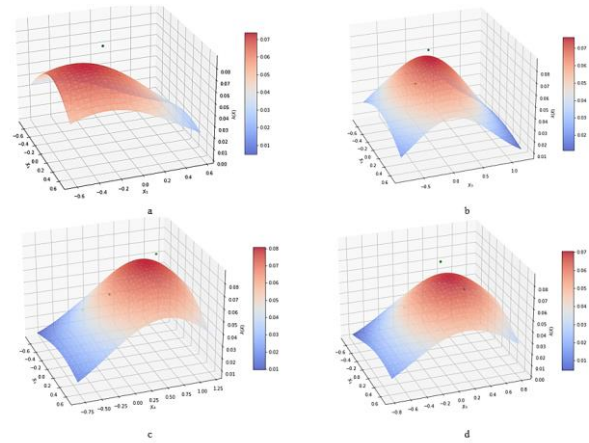
Fig. 7. Failure rate smoothed regression model visualization

As a performance metric for built regression models mean absolute mean absolute error (MAE) was chosen. The results of computational experiments show that cluster-based smoothed model demonstrates slight decrease in accuracy (MAE=0.0211 compared to MAE=0.0199 for base model), but avoids "spatial oscillation" as well as overfitting, and is more stable to outliers and noise.

Conclusions. The results obtained indicate, that proposed method for failure rate kernel-based regression model building using machine learning technique provides an opportunity to analyze the influence of a large number of technological and operational affecting factors and to successful approximate complex dependences of reliability indicators with high accuracy. Further development of proposed approach is expedient in terms of developing methods and algorithms for failure rate and failures themself forecasting based on time series of affecting factors measurements in real time. To build a predictive time-series model, it is also advisable to use kernel machine learning methods.

References

1. Boardman N., Hu W., Mishra R. Optimal Maintenance Design for a Simple Reparable System. *Proceedings of the 58th IEEE Conference on Decision and Control*. 2019. P. 3098–3103.

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022

55

2. Wang J., Yin H. Failure Rate Prediction Model of Substation Equipment Based on Weibull Distribution and Time Series Analysis. *IEEE Access*. 2019. Vol. 7. P. 298–309.

3. Jones H. A Method and Model to Predict Initial Failure Rates. *Annual Reliability and Maintainability Symposium*. 2020. P. 1–6.

4. Ястребенецкий М. А., Соляник Б.А. *Определение надежности аппаратуры промышленной автоматики в условиях эксплуатации*. Москва: Энергия, 1968. 128 с.

5. Jürgensen J. *Individual Failure Rate Modelling and Exploratory Failure Data Analysis for Power System Components*. *Doctoral Thesis*. Stockholm, KTH, 2018. 79 p.

6. Jürgensen J., Nordström L., Hilber P. Estimation of Individual Failure Rates for Power System Components Based on Risk Functions. *IEEE Trans. on Power Delivery*. 2019. Vol. 34, no. 4. P. 1599–1607.

7. Susto G., Schirru A., Pampuri S., McLoone S., Beghi A. Machine Learning for Predictive Maintenance. *IEEE Transactions Industrial Information*. 2015. Vol. 11. P. 812–820.

8. Scholkopf B., Smola A. *Learning with Kernels*. Cambridge: MA, MIT Press. 2002. 640 p.

9. Guo X., Uehara K. Graph-based Semi-Supervised Regression and Its Extensions. *International Journal of Advanced Computer Science and Applications*. 2015. Vol. 6, no. 6. P. 260–269.

10. Pozdnoukhov A., Bengio S. Semi-supervised kernel methods for regression estimation. *Proc. of IEEE Intern. Conference on Acoustics, Speech and Signal Processing*. 2006. P. 577–580.

11. Sindhwani V., Niyogi P., Belkin M. Beyond the Point Cloud: from Transductive to Semi-supervised Learning. *Proc. of the 22nd international conference on Machine learning*. 2005. P. 824–831.

12. McCann M., Johnston A. *SECOM Data Set 2008*. URL: https://archive.ics.uci.edu/ml/datasets/secom (accessed 20.08.2021).

References (transliterated)

1. Boardman N., Hu W., Mishra R. Optimal Maintenance Design for a Simple Reparable System. *Proceedings of the 58th IEEE Conference on Decision and Control*. 2019, pp. 3098–3103.

2. Wang J., Yin H. Failure Rate Prediction Model of Substation Equipment Based on Weibull Distribution and Time Series Analysis. *IEEE Access*. 2019, vol. 7, pp. 298–309.

3. Jones H. A Method and Model to Predict Initial Failure Rates. *Annual Reliability and Maintainability Symposium*. 2020, pp. 1–6.

4. Yastrebeneckij M. A., Solyanik V.A. *Opredelenie nadezhnosti apparatury promyshlennoj avtomatiki v usloviyah ekspluatacii* [Determination of the reliability of industrial automation equipment in operating conditions]. Moscow, Energiya, 1968. 128 p.

5. Jürgensen J. *Individual Failure Rate Modelling and Exploratory Failure Data Analysis for Power System Components*. *Doctoral Thesis*. Stockholm, KTH, 2018. 79 p.

6. Jürgensen J., Nordström L., Hilber P. Estimation of Individual Failure Rates for Power System Components Based on Risk Functions. *IEEE Trans. on Power Delivery*. 2019, vol. 34, no. 4, pp. 1599–1607.

7. Susto G., Schirru A., Pampuri S., McLoone S., Beghi A. Machine Learning for Predictive Maintenance. *IEEE Transactions Industrial Information*. 2015, vol. 11, pp. 812–820.

8. Scholkopf B., Smola A. *Learning with Kernels*. Cambridge, MA, MIT Press. 2002. 640 p.

9. Guo X., Uehara K. Graph-based Semi-Supervised Regression and Its Extensions. *International Journal of Advanced Computer Science and Applications*. 2015, vol. 6, no. 6, pp. 260–269.

10. Pozdnoukhov A., Bengio S. Semi-supervised kernel methods for regression estimation. *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. 2006, pp. 577–580

11. Sindhwani V., Niyogi P., Belkin M. Beyond the Point Cloud: from Transductive to Semi-supervised Learning. *Proc. of the 22nd international conference on Machine learning*. 2005, pp. 824–831.

12. McCann M., Johnston A. *SECOM Data Set 2008*. Available at: https://archive.ics.uci.edu/ml/datasets/secom (accessed 20.08.2021).

Відомості про авторів / About the Authors

Ахієзер Олена Борисівна – кандидат технічних наук, доцент, Національний технічний університет «Харківський політехнічний інститут», завідувачка кафедри комп'ютерної математики і аналізу даних; м. Харків, Україна; ORCID: https://orcid.org/0000-0002-7087-9749; e-mail: olena.akhiiezer@khpi.edu.ua

Грінберг Галина Леонідівна – кандидат технічних наук, доцент, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри маркетингу; м. Харків, Україна; ORCID: https://orcid.org/ 0000-0002-0774-5414; e-mail: galyna.grinberg@khpi.edu.ua

Любчик Леонід Михайлович – доктор технічних наук, професор, Національний технічний університет «Харківський політехнічний інститут», професор кафедри комп'ютерної математики і аналізу даних; м. Харків, Україна; ORCID: https://orcid.org/ 0000-0003-0237-8915; e-mail: leonid.liubchyk@khpi.edu.ua

Ямковий Клим Сергійович – Національний технічний університет «Харківський політехнічний інститут», асистент кафедри комп'ютерні математики і аналізу даних; м. Харків, Україна; ORCID: https://orcid.org/ 0000-0001-9512-4150 e-mail: klym.yamkovyi@cs.khpi.edu.ua

Akhiiezer Olena Borisivna – Candidate of Technical Sciences, Docent, National Technical University "Kharkiv Polytechnic Institute", Head of Computer Mathematics and Data Analysia Department; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0002-7087-9749; e-mail: olena.akhiiezer@khpi.edu.ua

Grinberg Galyna Leonidivna – Candidate of Technical Sciences, Docent, National Technical University "Kharkiv Polytechnic Institute", Associate Professor at Department of Marketing; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0002-0774-5414; e-mail: galyna.grinberg@khpi.edu.ua

Lyubchyk Leonid Mykhailovych – Doctor of Technical Sciences, Professor, National Technical University "Kharkiv Polytechnic Institute", Professor of Computer Mathematics and Data Analysia Department; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0003-0237-8915; e-mail: leonid.liubchyk@khpi.edu.ua.

Yamkovyi Klym Serhiyovych – National Technical University "Kharkiv Polytechnic Institute", Assistant Professor of Computer Mathematics and Data Analysia Department; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0001-9512-4150; e-mail: klym.yamkovyi@cs.khpi.edu.ua

56

Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології, № 2 (8)'2022