DOI 10.34185/1562-9945-6-137-2021-05
UDC 004.043: 004.67

L.I. Korotka

FORMATION OF A SET OF INFORMATIVE FEATURES WHEN DECIDING PROBLEMS OF PREDICTING THE DURABILITY OF STRUCTURES

Annotation. The paper proposes a method for extracting informative features for the training sample in the problems of predicting the durability of corroding structures. The work aims to analyze and evaluate the informative features that improve the quality of the data and to structure them. Kendall`s method is used to determine the value of each trait. It is proposed to use the training sample to work with a neural network or to build a fuzzy knowledge base only after the formation of a set of information attributes on their value.
Keywords: information features, extraction and formation of information features, Kendall's method

1. Introduction. Working with multidimensional data sets requires preliminary processing and analysis. This is due to many reasons, including the need to lower the dimension of space; identification of informative signs and their assessment; identification of abnormal or noisy elements in samples. These factors are generally aimed at improving the quality of data, at the same time, the extraction of informative features (IF) allows them to be structured [1, 2, 7, 8, 10].

The formation of a set of IF makes it possible to somewhat rank the variables of the set under consideration. As a rule, the features that may be insignificant or not too significant may be duplicated and/or even noisy. Therefore, the process of identifying and forming a set of information features when working with multidimensional data in any subject area can hardly be overestimated [13].

2. The problem. The work will consider a training sample, which has been repeatedly used in training neural networks of different types in the tasks of predicting the durability of corroding structures (CS) [4, 11]. However, for the first time, it is proposed to consider the informative features of these data and evaluate them.

This approach can be useful both for training neural networks of various architectures [5, 6], and for clustering (clear or fuzzy [4, 9, 11]), and for building a knowledge base when solving problems of this class.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Prediction of the durability of the CS (direct problem) is an integral part of the problem of optimal design of corrosive structures, in which the direct problem acts as the constraint functions (CF) of the inverse problem. The optimization problem statement is formulated and known [3, 11, 12]. Both problems describe the process of modeling the behavior of complex nonlinear dynamic systems that operate in an aggressive environment. Problems of increasing the efficiency of computational methods of such systems occupy a special place.

There are enough works devoted to these problems because the reduction of computational costs of direct and inverse problems is interconnected. The author thinks it expedient to formulate a sequence of problematic aspects that are interrelated and lead to a change in computational costs when solving this class of problems.

When setting the problem of predicting the durability, the CS is described by a set of independent parameters, each of which is characterized by a numerical value. The structure operates in an aggressive working environment that causes corrosive wear, which leads to a change in its initial geometric dimensions. At the same time, the structure must maintain its bearing capacity, that is, satisfy the limitations in strength and stability. The solution to the problem of predicting the resource of corroding structures involves determining the value of the durability of the system, during which it will function without failure.

In general, the mathematical formulation of the problem of predicting durability can be written as follows:

$$t^* = \min\{t_i\}; \quad i = \overline{1, N}. \tag{1}$$

Here t_i is the durability of the i-th structural element, determined by the conditions of strength and stability:

$$\begin{cases} [\sigma] - \sigma_i(t, v_0) = 0; & i = \overline{1, N} \\ \sigma_j^*(t) - \sigma_j(t, v_0) = 0; & j \in J \end{cases}, \tag{2}$$

Where $[\sigma]$ is permissible stress; $\sigma_i(t)$ is stress in i-th element; σ_j^* is critical buckling stress; v_0 is a corrosion rate in the absence of stresses; J is a set of compression elements; N is the number of elements in the system; t^* is definable durability.

There are a large number of mathematical models to describe corrosive wear, which makes it problematic to build a unified approach to solving the problem. Note

that the following assumptions are made for the considered corrosion process: 1) corrosion occurs by an electrochemical mechanism and consists of the spontaneous destruction of the metal in the electrolyte; 2) there is a case of general corrosion, the mechanism of which is described for the entire surface of the structure; 3) irregularity of corrosion is determined by the degree of inhomogeneity of the stress field over the area of the structure.

The paper proposes to use Dolinsky`s model:

$$\frac{d\delta_i}{dt} = v_0\left[1 + k\ \sigma_{eq}\right], \qquad (3)$$

Where σ_{eq} is the absolute value of the equivalent voltage; $k = 0.003\ MPa^{-1}$ is a coefficient taking into account the effect of stresses on the rate of the corrosion process; where $\overline{\delta}$ is a vector of depths of corrosion of all elements.

As noted in the works [3, 4, 11, 12] the solution of this system of differential equations is possible only numerically, since its right-hand sides are computational algorithms. The stresses in the right-hand sides (3) are calculated using the equations of mechanics of a deformed solid: systems of equations of equilibrium and compatibility of deformations, Cauchy relations, and physical relations (for elastic bodies - Hooke's law). Consequently, the choice of the parameters of the numerical procedures will significantly affect the error of the result obtained. With a decrease in the integration step, it is possible to achieve greater accuracy, but this will negatively affect the computational efficiency of the algorithm.

3. Analysis of requirements. Obviously, concerning the error of the obtained solution ε for the inverse problem, the following options are possible $D = D_1 \cup D_2 \supset D_3$:

1) the first subset of points of the solution space D_1, for which the error in calculating the CF will exceed the maximum permissible ε^*: $\varepsilon > \varepsilon^*$;

2) the second subset of points D_2 for which the error is less than the permissible $\varepsilon < \varepsilon^*$;

3) the third subset D_3, the power of which is incomparably less than the two previous ones, for which the error in calculating the CF turns out to be within the permissible value, that is, in a certain neighborhood γ: $|\varepsilon - \varepsilon^*| \le \gamma$, which can be considered as.

In the latter case, the conditions for the accuracy and efficiency of the computational algorithm are satisfied. It is obvious that the desired set of the solution space, which does not lead to excessive computational costs and satisfies $D \approx D_3$. If

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

the desired error of the numerical solution is given in advance, then such an approach requires a certain module for controlling the accuracy of the numerical result for a given error [ε]. The problem statement is formulated:

$$\begin{cases} h_t(\varepsilon,\bar{c}) \to \max \\ \varepsilon(\bar{c},h_t) \le [\varepsilon] \end{cases}. \quad (4)$$

Here $\bar{c} = \{v_0, k, \sigma_0, A_0, P_0\}$ is a vector of factors influencing the value of the integration step h_t; where σ_0 is initial stress; A_0, P_0 are geometric characteristics of structural elements (area and perimeter, respectively).

Since the direct problem has been solved several times, therefore, there are quite large amounts of data that are of an unordered nature. The author believes that if we carry out a preliminary analysis of them, which consists of extracting information about each feature, then it is possible to structure them in a certain way and, according to the author, influence the efficiency of the computational algorithm. This is the purpose of this work. Let us formulate the formulation of the problem of forming a set of information features based on their value for the class of problems under consideration.

Let us have input data in the form of a matrix $C = \{v_{0i}, \sigma_{0i}, A_{0i}, P_{0i}, [\varepsilon]\}$, $i = \overline{1,N}$, where N is the sample size. It is necessary, using clustering methods, to arrange the features in descending order of their informativeness. Since quantitative features are considered in the work, it is possible to use either the distance matrix approximation method or the Kendall method to solve the problem. It is the latter that will be considered.

For each characteristic, the following steps must be performed:

1. Construct a frequency matrix for each feature $C = \{v_{0i}, \sigma_{0i}, A_{0i}, P_{0i}, [\varepsilon]\}$, here $i = \overline{1,N}$.

2. Find the change interval of the characteristic values, which is common for all clusters.

3. Calculate Inf_j ($j = 1 \div 5$ is a number of features) as the number of objects of all classes, the values of which for a particular feature did not fall into the joint interval.

Further, the features are ordered in descending order Infj, which corresponds to their ranking of decreasing individual value. The quantities Inf_j ($j = 1 \div 5$) can be considered as weights of features: the larger the value of Infj, the more significant it is, and, therefore, more objects can be recognized by this feature. Therefore, this feature is more informative.

As you know, when the number of classes is more than two, Kendall's method is applied to each pair of classes, and the weights obtained are averaged over the number of pairs [7].

4. Results of numerical experiments. Let there be a sample of volume $N = 5000$ samples, which is characterized by the following features $C = \{v_{0i}, \sigma_{0i}, A_{0i}, P_{0i}, [\varepsilon]\}$. Based on the results of observations of objects using fuzzy clustering [4], five clusters are considered, which are considered hereinafter as class labels. We have a degeneration of the clustering problem into a classification problem. According to the steps described above, frequency matrices are built for each feature: corrosion rate, initial stresses, and geometric characteristics of structural elements (area and perimeter, respectively), errors.

The paper proposes to average the obtained weights not for each pair of classes, but for all at the same time. This approach presupposes more stringent requirements, however, according to the author, this allows taking into account the individual value of each trait for all classes at the same time.

The results of numerical experiments are shown in Table 1.

Here are informative signs: Inf_1 is for corrosion rate v_0; Inf_2 is for the informative sign for initial stresses σ_0; Inf_3 is for area A_0; Inf_4 is for the perimeter of the cross-section of the structural element P_0; Inf_5 is for a given error [ε].

Table 1

Results of numerical experiments

Informative features for elements of the training sample				
Inf_1	Inf_2	Inf_3	Inf_4	In_{f5}
0.7198	1,000	0.9936	0.9936	0.9698

Next, you can build informative features according to their value. The results obtained indicate that the least informative feature is the corrosion rate, and the most informative feature is the initial stresses. Therefore, it can be concluded that for the further use of the training sample, for example, when training neural networks [11] or for classification [4], attention should be paid to the least informative feature. It is possible, of course, to exclude it and not be considered as an input parameter when determining the parameters of numerical procedures. On the other hand, v_0 - is a parameter of an aggressive environment that makes a significant contribution to the operation of the structure. Therefore, according to the author, here the problem can be solved by reducing the number of classes for this input variable.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

5. Results. The approach considered in the work allows you to streamline the features and extract information about the information content of each of them. This will allow, during further work with training samples, to structure the data and improve the quality of their further use.

REFERENCES

1. Dy J.G. Feature Selection for Unsupervised Learning /J.G. Dy, C.E. Brodley //J. of Machine Learning Research. – 2004. – Vol. 5. – P. 845–889.

2. Kolesnikova S.I. Metody analiza informativnosti raznotipnyh priznakov [Methods for analyzing the informativeness of different types of signs] /S.I. Kolesnikova// Vesnik Tomskogo gosudarstvennogo universiteta: Upravlenie, vychislitel'naja tehnika i informatika. – 2009. - № 1(6), Obrabotka informacii. – S. 69-80. (in Russian)

3. Korotka L.I, Zelentsov D.G. Method of solving optimal design problems based on flexible tolerance strategy / L.I. Korotka, D.G. Zelentsov // International Journal of Mathematical Modelling and Numerical Optimisation. – 2020. – Vol.10, No.3. – pp. 255-269. DOI: 10.1504/IJMMNO.2020.108613

4. Korotka L.I. The use of fuzzy clustering in solving problem in predicting the durability of corrosive structures/ L.I. Korotka // Mathematical modeling. – 2020. – №2(43). С. 44-54. DOI: 10.31319/2519-8106.2(43)2020.219266

5. Levkin D.A., Berezhna N.H., Makarov O.A., Kutia O.V. Matematychne modeliuvannia tekhnichnykh system [Mathematical modeling of technical systems] /D.A. Levkin, N.H. Berezhna, O.A. Makarov, O.V. Kutia// Vcheni zapysky Tavriiskoho Natsionalnoho Universytetu imeni V.I. Vernadskoho. – 2021. – Vol. 32, Issue 71. – S. 98-102. (in Ukrainian)

6. Lohvin A.O. Typy heneratyvnykh neironnykh merezh [Types of generative neural networks]/ A.O Lohvin // Informatyka, obchysliuvalna tekhnika ta avtomatyzatsiia: Vcheni zapysky TNU imeni V.I. Vernadskoho. Seriia: Tekhnichni nauky. – 2021. – С. 103-109. DOI: 10.32838/2663-5941/2021.1-1/17 (in Ukrainian)

7. Matsuha, O.M. Informatsiini tekhnolohii rozpiznavannia obraziv: Navchalnyi posibnyk do vyvchennia kursu ["Information technologies of pattern recognition"] /O.M. Matsuha, Yu.M. Arkhanhelska, N.M. Yereshchenko. – D.:RVV DNU, 2016.–60s. (in Ukrainian)

8. Paklin N.B. Biznes-analitika: ot dannyh k znanijam: ucheb. posobie [Business analytics: from data to knowledge: textbook. Allowance] /N.B. Paklin, V.I. Oreshkov. – SPb: Piter, 2013. – 704 s. (in Russian)

9. Vjatchenin D.A. Nechetkie metody avtomaticheskoy klassifikacii [Fuzzy methods of automatic classification] /D.A. Vjatchenin. – Mn.: UP «Tehnoprint», 2004. – 219 s. (in Russian)
10. Zagorujko N.G. Prikladnye metody analiza dannyh i znanij [Applied methods of data and knowledge analysis] /N.G. Zagorujko. – Novosibirsk: IM SO RAN, 1999. – 270 s. (in Russian)
11. Zelentsov D.G., Korotkaya L.I. Tehnologii vyichislitelnogo intellekta v zadachah modelirovaniya dinamicheskih sistem: monografiya [Technologies of Computational Intelligence in Tasks of Dynamic Systems Modeling: Monograph] / D.G. Zelentsov, L.I. Korotkaya,. – Balans-Klub, Dnepr, 2018. – 178 p. DOI: 10.32434/mono-1-ZDG-KLI (in Russian)
12. Zelentsov D.G., Korotka L.I., Denysiuk O. R. The Method of Correction Functions in Problems of Optimization of Corroding Structures /D.G. Zelentsov, L.I. Korotka, O.R. Denysiuk// Advances in Computer Science for Engineering and Education III (ICCSEEA 2020), 2020. – pp 132-142. DOI: 10.1007/978-3-030-55506-1_12
13. Zelentsov D., Us S., Koryashkina L., Stanina O. Solving Continual Two-Stage Problems of Optimal Partition of Sets / D. Zelentsov, S. Us, L. Koryashkina, O. Stanina// International Journal of Research Studies in Computer Science and Engineering (IJRSCSE). 2017. – Volume 4, Issue №4. - P. 72-80.

Received 13.09.2021.
Accepted 15.09.2021.

Формування набору інформативних ознак
при розв'язанні задач прогнозування довговічності конструкцій

У роботі запропоновано спосіб здобуття інформативних ознак для навчальної вибірки у задачах прогнозування довговічності кородуючих конструкцій. Метою роботи є аналіз та оцінка інформативних ознак, які дозволяють покращити якість даних та структурувати їх. Для здобуття цінності кожної ознаки використовується метод Кендалла. Пропонується користуватися навчальною вибіркою для роботи з нейронними мережами або для побудови нечіткої бази знань тільки після формування набору інформативних ознак за їх індивідуальною цінністю.

Формирование набора информативных признаков
при решении задач прогнозирования долговечности конструкций

В работе предложен способ извлечения информативных признаков для обучающей выборки в задачах прогнозирования долговечности корродирующих конструкций. Целью работы является анализ и оценка информативных признаков, которые позволяют улучшить качество данных и структурировать их. Для определения ценности каждого признака используется метод Кендалла. Предлагается применять обучающую выборку для

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

работы с нейронной сетью или для построения нечеткой базы знаний только после формирования набора информационных признаков по их индивидуальной ценности.

Коротка Лариса Іванівна - к.т.н., доцент, доцент кафедри інформаційних систем ДВНЗ «Український державний хіміко-технологічний університет».

Короткая Лариса Ивановна - к.т.н., доцент, доцент кафедры информационных систем ГВУЗ «Украинский государственный химико-технологический университет».

Korotka Larysa Ivanivna - Candidate of Technical Sciences, Docent, Associate Professor at the Department of Information Systems Ukrainian State University of Chemical Technology.