Nonparametric algorithm of electronic components test data pattern recognition

IOP Conf. Series: Materials Science and Engineering **537** (2019) 042021 doi:10.1088/1757-899X/537/4/042021

Nonparametric algorithm of electronic components test data pattern recognition

N V Koplyarova[1], E A Chzhan[1], A V Medvedev[1], A A Korneeva[1], A V Raskina[1], V V Kukartsev[1,2] and V S Tynchenko[1,2]

[1] Siberian Federal University, Svobodny pr., 79, Krasnoyarsk, 660041, Russia
[2] Siberian State University of Science and Technology, Krasnoyarskiy Rabochiy Ave., 31, Krasnoyarsk, 660037, Russia

E-mail: ekach@list.ru

Abstract. The paper discusses the quality diagnostics of electrical radio components based on the results of non-destructive testing. A proposed clustering algorithm does not require preliminary information on the number of classes and the training sample. The algorithm allows to automatically determine the number of classes. The division into classes is due to the different characteristics of the measured variables, which correspond to different product quality ranges.

1. Introduction

The problem of automatic clustering of products according to real data is considered in the case when the number of classes is unknown [1]. The main problem in solving the problem is how to evaluate the results of the data set. The result is that each element has the number of elements in each class and in the list of class centers. The result of the clustering is adequate when the compactness hypothesis is satisfied for the classes obtained. Class centers are marked explicitly. However, it all depends on the quality of the selection and the nature of the data. Frequently set values cannot be explicit. In addition, there are many algorithms of pattern data recognition [2, 3].

The development of methods for checking the quality of electrical radio products as components of spacecraft is an actual problem of modern science and technology due to the fact that this industry requires that the products could effectively perform their functions. From the point of view of applications, the most important class is the problem of diagnosing products without the use of destructive testing [4].

2. Clustering Statement Problem

Let there be a collection of some objects O1, O2, ..., Os, the properties of which are defined in the feature space. It is required to break them into groups of objects, in a certain sense, close to each other. Information about objects is given in the form of an m × n matrix (m is the number of input parameters (features), s is the sample size of products). Thus, the task of clustering is to divide the feature space into disjoint regions.

Each element of the sample (object) is characterized by certain values of the vector of parameters $v = (v_1, ..., v_m)$ on the basis of which diagnostics are carried out (an object can be assigned to one of

the classes V_1 or V_2). Typical for diagnostic tasks is the presence of a cloud structure in the space of features that define a particular class (for example, satisfactory, of medium quality and high quality).

Pattern recognition (self-study) is learning without any instructions from the teacher about the correctness or incorrectness of the reaction of the system in various conditions. Suppose that a set of objects X consists of several non-intersecting subsets X_k, (k = 1, 2, …, l), corresponding to different classes of objects characterized by vectors $\bar{x} \in X$. Since an object $\bar{x} \in X$ appears in a particular set X_k, (k = 1, 2, …, l) randomly, it is natural to consider the probability of an object $\bar{x}$ appearing in a class X_k (we denote it as P_k) and the conditional probability density of a vector $\bar{x}$ $p_k(x) = P(x \mid k), k = \overline{1,l}$ inside the corresponding class,.

In this case, the maxima of the probability densities $P_k(x)$ are above the "centers" of the classes corresponding to the subsets X_k . However, when it is unknown to which class the object $\bar{x}$ belongs, it is impossible to determine these conditional probability densities. The joint probability density $P(x) = \sum_{k=1}^{l} P_k p_k(x)$ contains fairly complete information about the sets [5]. In particular, the maxima (modes) of the function will correspond to the "centers" of the classes. Therefore, the problem of self-learning is often reduced to the task of restoring the joint probability density $P(x)$ and determining the “centers” and then the class boundaries [6]. So for the case $\bar{x} = (x_1, x_2)$ a possible form of probability density is presented in figure 1.

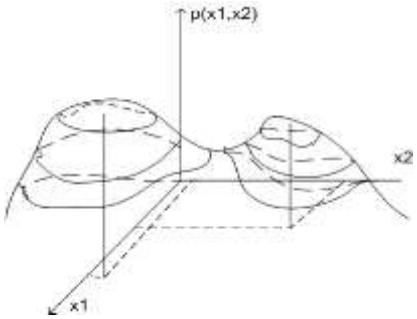

Figure 1. Probability distribution density.

Figure 1 shows the probability density $P(x)$, which has two maxima, and, consequently, two classes. There are several nonparametric methods to get estimation of density function [7, 8]. The centers of classes it is natural to assume the coordinates of the maxima (modes) of the distribution $(x_1(k), x_2(k)), k = 1, 2$.

Next, we consider the algorithm proposed for grouping data in the case where the initial number of classes is unknown.

3. Clustering Statement Problem

Let a product be attributed to one or another quality category in accordance with the requirements of state standard applicable to EC. This product is characterized by certain values of the parameters vector, on the basis of which the diagnostics of the latter is carried out. Consider a sample of observations $\left\{x_i, i = \overline{1,s}\right\}$ of a multidimensional variable $x \in R^m$, where m is the dimension of the variable x, s is the sample size. It is necessary to distinguish from the available sample classes of products that are similar in quality characteristics. For classes, the “compactness hypothesis” is valid, in other words, the arrangement of points in the feature space is such that they are grouped in different areas. The number of classes is unknown. The classification algorithm assumes the following sequence.

1. The normalization and centering of the original sample.
2. Find the distances between all points of the normalized sample of observations.

$$r_{k,p} = \sum_{j=0}^{m-1} (x^N{}_{k,j} - x^N{}_{p,j})^2, \tag{1}$$

where $r_{k,p}$ is the Euclidean distance between *k* and *p* elements of the sample.

Further, on the basis of the obtained sample of distances between points $r_{k,p}, k, p = \overline{1,m}$, a histogram of the distance distribution is plotted. If in the neighborhood of zero a histogram has a pronounced maximum, and outside this neighborhood there are usual tails of a distribution, this indicates the presence of classes.

3. Select a group of points for which the condition $r_{k,p} > \delta$ is satisfied, where δ is selected on the basis of the available a priori information and the experience of the researcher. For reasons of simplicity, we find 2 points located at a maximum distance from each other. These points are accepted as primary class centers $c_{k,j}, k = \overline{1,N}, j = \overline{1,m}$.

The following notation is introduced: $c_{k,j}, k = \overline{1,N}, j = \overline{1,m}$ are class centers, where *N* is the number of classes.

4. Setting the parameter Δ. The parameter Δ required for the operation of the algorithm is selected from the range of values Δmin <Δ <Δmax, where the values of Δmin and Δmax are calculated as follows:

$$\Delta\min = \min(r)/r_{mean}, \quad \Delta\max = \max(r)/r_{mean}, \tag{2}$$

where r_{mean} is the average value of the distances between all elements of the sample, $\min(r)$ is the minimum value of the distance, $\max(r)$ is the maximum value of the distance.

Thus, the researcher determines the boundary values of the parameter Δ that do not go beyond the permissible limits of the range of values Δmin <Δ <Δmax. The setting of the parameter Δ is carried out in such a way that in a given range, several values of Δ are selected and a grouping is performed for each of them and the number of classes is determined.

5. Classification is carried out for each element of the sample:

- take an arbitrary *k*-th element of the sample from the original training sample;
- calculate the distance from this element to each of the centers of primary classes:

$$r(c_k, x) = \sum_{j=0}^{m-1} \left(c_{k,j} - x^N{}_{i,j}\right)^2, \tag{3}$$

where $r_{k,p}$ is the Euclidean distance between *k* and *p* elements of the sample, $c_{k,j}, k = \overline{1,N}, j = \overline{1,m}$ are the centers of the classes.

- if the condition $r(c_k, x) < \Delta,$ is fulfilled, then the considered element belongs to the *k*-th class, otherwise, a new class with a center at the given point is created and the number of classes becomes N + 1. In this case, if the condition $r(c_k, x) < \Delta$ is satisfied for several centers of classes, the element joins the class, the distance to which is the smallest;
- the selected point $x^N_{i,j}$ joins the class *k* and is excluded from the original sample;
- the coordinates of the center of each class are recalculated as follows:

$$c_{k,j} = \frac{1}{N_k} \sum_{i=0}^{N_k - 1} x^N{}_{i,j}, \qquad x^N{}_{i,j} \in X_k, \tag{4}$$

where $x_{i,j}^{N}$ are the elements of the sample belonging to the class *k*, $c_{k,j}$ are the coordinates of the center of the class k in the *m*-dimensional space.

After passing through this procedure, if $k < s$, then $k = k + 1$ and go back to the beginning of step 5. This process continues for all points of the original sample. As a result, the initial sample is divided into *N* classes. Thus, the number of classes *N* for each particular batch of products is determined at the end of the study.

4. Processing Electronic Components Test Data

As an example, the clustering of electronic components (EC) is considered. One of the main tasks of the modern space industry is to complete the onboard equipment of a high-reliability EC spacecraft. First of all, it is necessary to prevent products that do not satisfy reliability requirements from entering the equipment. As part of solving this problem, it is necessary to ensure the purchase of electronic components from verified suppliers, as well as conducting input control, additional screening tests and destructive physical analysis of EC.

The following are the results of numerical calculations for EC diagnostics based on real data obtained when measuring the parameters of transistors. Data provided by test center of ISS-Reshetnev Company.

To diagnose transistors, we will classify all available observations in order to identify groups of transistors in the space of diagnostic indicators. It is required to assemble the specified 78 cases of transistors into clusters according to 16 variables characterizing their quality.

The results of clustering by the proposed method are given. According to the results of the classification, it can be said that the optimal solution is where the whole sample is divided into 2 classes corresponding to transistors of different quality. We also obtain the table of the belonging of each element to specific clusters, in which one can see that 35 transistors belong to the first cluster and 43 to the second cluster. Below there are the average values of the variables for each cluster (cluster centers).

Table 1. The average values of the variables for each cluster

Variable	Cluster 1	Cluster 2
x_1	0,0040	0,0040
x_2	0,0404	0,0603
x_3	0,0008	0,0007
x_4	0,7065	0,7063
x_5	0,7568	0,7576
x_6	0,8322	0,8346
x_7	0,7560	0,7613
x_8	0,8837	0,8940
x_9	1,2403	1,2615
x_{10}	0,0359	0,0520
x_{11}	0,1056	0,1082
x_{12}	0,2571	0,2531
x_{13}	0,6890	0,6858
x_{14}	45,9140	26,1286
x_{15}	72,4023	41,4343
x_{16}	50,0600	32,2189

From the table 1 it can be seen that in the classification the greatest difference is observed in the values of attributes x8, x9, x14-x16.

Thus, the sampling data of the quality measurements of transistors were divided into clusters, which differ in values of most parameters, which can correspond to different levels of quality of the

products under consideration (transistors), and the products belong to different manufacturing batches. The proposed algorithm shows a rather high classification efficiency and can be applied in real-life tasks to improve the quality of EC diagnostics.

5. Conclusion

When analyzing various real data, there is often a need for the task of grouping data, which leads to the appearance of clusters in the space of parameter characterizing the quality of the product. In article an algorithm is proposed for solving the clustering problem, which does not require knowledge of the number of classes. Based on the analysis of the results of model and real data, conclusions can be made about the validity and quality of the classification carried out.

References

[1] Medvedev A V 2017 Some remarks on the theory of non-parametric systems *Applied Methods of Statistical Analysis* pp 72-81
[2] Dalitz C, Ayyad Y, Wilberg J, Aymans L, Bazin D and Mittig W 2019 Automatic trajectory recognition in active target time projection chambers data by means of hierarchical clustering *Computer Physics Communications* **235** 159-68
[3] Ozturk C, Hancer E and Karaboga D 2015 Dynamic clustering with improved binary artificial bee colony algorithm *Applied Soft Computing Journal* **28** pp 69-80
[4] Soman R R, Davidson E M, McArthur S D, Fletcher J E and Ericsen T 2012 Model-based methodology using modified sneak circuit analysis for power electronic converter fault diagnosis *IET Power Electronics* **5** 813-26
[5] Cho H, Venturi D and Karniadakis G E 2016 Numerical methods for high-dimensional probability density function equations *Journal of Computational Physics* **305** 817-37
[6] Jackson Q and Landgrebe D A 2001 An adaptive classifier design for high-dimensional data analysis with a limited training data set *IEEE Transactions on Geoscience and Remote Sensing* **39** 2664-79
[7] Chzhan E A 2017 Non-parametric dual control algorithms of discrete-continuous processes with dependent input variables *Applied Methods of Statistical Analysis* pp 82-7
[8] Medvedev A V and Chzhan E A 2017 On nonparametric modelling of multidimensional noninertial systems with delay *Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software* **10** 124-36