Fuzzy clustering of EEE components for space industry

V I Orlov[1], D V Stashkov[2], L A Kazakovtsev[1,3], A A Stupina[1,3]
[1] Reshetnev Siberian State Aerospace University 31 "Krasnoyarskiy Rabochiy" av., Krasnoyarsk, 660037, Russia
[2]JSC "Sinetic", 127 "3 Internacionala" street, Novosibirsk, 630009, Russia
[3]Department of Economics and Information Technology of Management Siberian Federal University, 3 Vuzovskii Lane, Krasnoyarsk, 660025, Russia

E-mail: levk@bk.ru

Abstract. One of the most important problems of the space industry is obtaining reliable methods of automatic grouping (clustering) of specialized EEE components for using in space systems. The main purpose of automatic grouping of EEE components on a set different parameters is the most legible splitting group of EEE components into several homogeneous production batches produced from a single bath of raw materials. The Expectation Maximization algorithm first time applied for the classification of EEE components.

Introduction

Particular belonging of EEE components to a single production batch and a single batch of raw materials is the major necessary condition to increase significance of further actions realization carried out within multistage tests of EEE components [1, 2, 3].

Determining the belonging of a set of EEE components to a set of production batches is a clustering problem clustering (unsupervised learning) [4], there is no reference selection for application of approaches to the problems of classification (supervised learning), a priori the number of clusters (batches of the raw materials) is unknown, however, this number is bounded above.

The problem is even more complicated because of the possible existence of outliers (i.e. copies of the EEE components that do not belong to the main production and batch of raw materials or made with essential aberrations).

At the solution of such clustering problems, it is possible to use two approaches:

1. Hard clustering. Each object belongs to only one cluster. In this case if there is no clear boundary between clusters in the experimental data, the result will not be satisfactory (numerical characteristics of belonging to a cluster will be indistinct).

2. Fuzzy clustering. A set of probabilistic characteristics of belonging of the object to each cluster is the result of the algorithms work of this type.

At application of both approaches, classical methods [5] demand that the number of clusters was set before an algorithm starts.

Thus it is necessary to touch several options with various estimated parameter (number of clusters) for the further choice of the best decision.

For decision making on the result quality of a clustering algorithm (i.e. analysis of compliance of sets of objects and clusters), various quality evaluation criteria [6] are used.

Probability assessment of a regularity of the object reference to each cluster is result of work of such criteria.

The result of use of fuzzy clustering methods is defining probability assessment of belonging of each object to each cluster.

Thus need to use additional criteria for checking and confirmation (deviation) results of the algorithms disappears.

At present, there are a lot of methods for data clustering and classification [5, 7, 8].

The review of the methods mostly often used in practice is provided in [5].

EM algorithm

An EM algorithm (Expectation Maximization) is among the most popular methods. It is successfully applied to statistical problems of analysis of inexact data for cases when some statistical data are absent or for cases when function of credibility has a form which is not allowing convenient research techniques but allowing serious simplifications at introduction of the additional "unobservable" ("hidden") variables [9].

Such problem definition (clustering of multi-dimensional data which have normal distribution with the hidden data) is used by us for solving the problem of splitting a set of EEE components into several production batches produced from a single batch of raw materials.

Let the density function on a set X have a form of a mixture k of distributions (we assume that these distributions are Gaussian):

$$\rho(x) = \sum_{j=1}^{k} \omega_j \rho_j(x), \ \sum_{j=1}^{k} \omega_j = 1, \ \omega_j \geq 0,$$

where $\rho_j(x)$ is a likelihood function of jth component of these mixture of distributions, ω_j is its prior probability.

Let likelihood functions belong to parametrical family of distributions of $\varphi(x;\theta)$ and differ in values of parameter $\rho_j(x) = \varphi(x;\theta_j)$ only.

The problem of unmixing (fuzzy or indistinct clustering) is to estimate a vector of parameters $\Theta = (\omega_1,...,\omega_k,\theta_1,...,\theta_k)$ having selection of X^m of random and independent observations of the mix $\rho(x)$, knowing number k and function φ.

The idea of an algorithm is as follows. We introduce an auxiliary vector of the latent (hidden) variables G having two remarkable properties:

- on the one hand it can be found if vector values Θ of parameters are known;
- on the other hand searching maximum of likelihood strongly becomes simpler if values of the latent variables are known.

The EM algorithm consists of iterative repetition of two steps:

On an E-step, we calculate the expected value (expectation) of a vector G of the latent variables based on the current approximation of a vector Θ of the parameters. Let us designate $\rho(x;\theta_j)$ the probability density of the fact that an object x is received from jth component of our mixture of distributions. The conditional probability is

$$\rho(x,\theta_j) = \rho(x)P(\theta_j \mid x) = \omega_j \rho_j(x).$$

Let us introduce the designation $g_{ij} \equiv P(\theta_j \mid x_i)$. It is an unknown posterior probability of the fact that the object x_i is received from the *j*th component of the mixture of distributions. These values are used as the latent variables.

Note that $\sum_{j=1}^{k} g_{ij} = 1$ for any $i = 1,..,m$ since it means the total probability of belong of an object x_i of one of the k formulation constituent. From Bayesian formula

$$g_{ij} = \frac{\omega_j \rho_j(x_i)}{\sum_{s=1}^{k} \omega_s \rho_s(x_i)} \text{ for all } i, j.$$

On the M-step the problem of maximizing likelihood (maximization) is solved and there is the following approximation of Θ vector on the current values vectors of G and Θ.

Let us maximize a logarithm of the complete likelihood:

$$Q(\Theta) = \ln \prod_{i=1}^{m} \rho(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} \omega_j \rho_j(x_i) \to \max_{\Theta}$$

Solving an Lagrange optimization problem with restriction ω_j, we have:

$$\omega_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, j = 1,..,k$$

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i, \theta), j = 1,..,k.$$

Thus the M-step comes down to calculation of components weights ω_j as arithmetic averages and to a parameter estimation θ_j a solution of k independent optimization problems. Note that the separation of variables was possible thanks to successful introduction of the latent variables.

The repetitive process stops according to a stop criterion which is set in advance (we choose a metric $\rho(\theta_1, \theta_2)$ and number ε in advance). Process stops on m step if $\rho(\theta^{(m)}, \theta^{(m-1)}) < \varepsilon$.

It is experimentally established that the main EM algorithm has the strong instability according to input data. For example in case for four-component mix of normal distributions at a sample of 200-300 observations the replacement of only one observation by another can cardinally change the total estimates received by means of the EM algorithm [9].

The main EM algorithm as well as the method based on greedy heuristics [10] and *k*-means model [6] applied at present for splitting electronic components into the homogeneous production batches does not allow us to define the quantity of mixture components (quantity of clusters). This size has to be set before algorithm starts operating. Otherwise a series of problems has to be solved with various estimated number of clusters.

It is possible to carry to advantages of the EM algorithm [8]:

- selection of metrics is not required;
- a possibility of use it in combination with algorithms of data processing;
- it works on small volumes of data;
- it allows to allocate noise emissions and a rarefied hum noise [7];

- a number of the main algorithm modifications is developed [9]: so, a number of modifications of the main algorithm (median modifications, a SEM algorithm, a CEM algorithm, MCEM and SAEM algorithms) are developed for partial elimination of the disadvantages listed above [9].

Thus at application of more widespread *k*-means model to a problem of selection of the homogeneous production batches of EEE products the choice of a metrics or measure of distance [6, 11] is the complex problem which does not have the unique decision: the choice of the Euclidean distance square gives more compact clusters, the choice of a rectangular metrics makes the model less sensitive to the presence of outliers.

In practice, it is necessary to use the special methods of data normalization [6, 12].

The problem of the outliers selection is also very complicated. In [6], at application of the *k*-means model it is solved by introducing a special criterion. The threshold value of this criterion is rather difficult to define experimentally.

As well as the classical *k*-means algorithm, the EM algorithm is the prime two-step procedure. It contains a step of determination clusters parameters and a step of splitting a set of objects into clusters that allows us to use in the future the same methods of increase in accuracy and stability of solutions as those successfully applied for solving problems of the automatic grouping of EEE components applying *k*-means, *k*-medoid and *k*-medians models [6].

The methods based on the specified models remarkably proved on rather larger selections [6] from several hundred copies of EEE.

At the same time some the most expensive types of EEE components bearing the greatest functional loading in spacecrafts are shipped in lots from several pieces that makes the application of the EM algorithm and algorithms on its basis very perspective in relation to our problem.

The known versions of the EM algorithm (for example, in the R software environment [13]) include EMCluster [14] package and a mclust [15] package. In these packages, versions of the EM algorithm and various ways of initialization for a normal distribution of mixes are realized. The possibility of visualization of the work results is provided.

We applied a realization of the EM algorithm in the R software environment for automatic grouping of EEE parties from 50 to 620 pieces (see Fig. 1). In addition to the definition of belonging each EEE component (a point on the chart shown in Fig.1) to a cluster (which are highlighted in the color), the algorithm gives the table with probabilistic characteristics of belonging to a cluster and shows a cluster form, visualizing possible correlations of the EEE parameters.

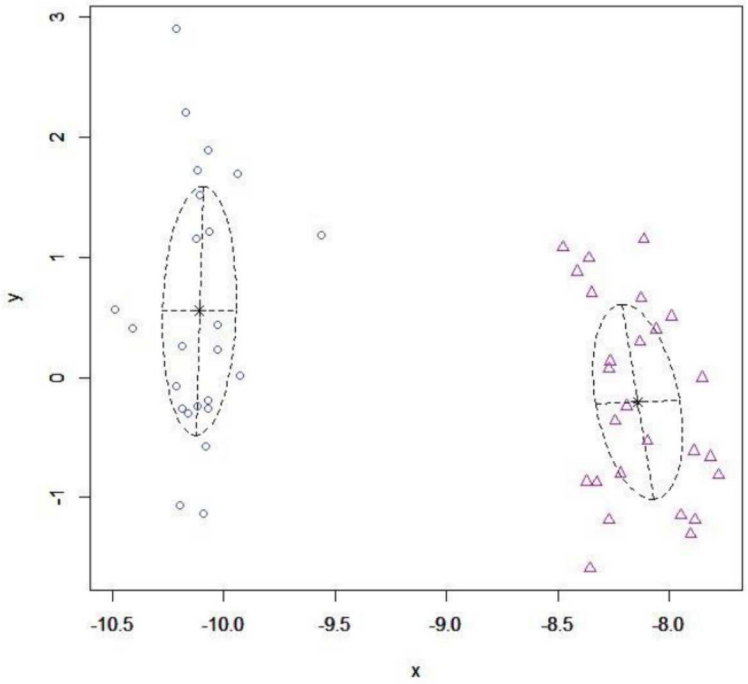

Figure 1. A projection (2 parameters from 32 are shown) of result of automatic group of amplifiers 140UD25AS1VK.

Conclusion

The application of the EM algorithm and algorithms on its basis opens new prospects in the solution of a problem of automatic group of EEE components on production batches, in particular, at the small volume of input data. At the same time, it is required to solve a problem of increase in stability of result of an algorithm.

Thus the application of the EM algorithm and algorithms on its basis opens the considerable prospects in the solution of a problem of automatic group of EEE on production parties, in particular, at the small volume of input data. However the problem of instable results of this algorithm must be solved for successful implementation in the space industry. The equivalence of the main steps of the EM algorithms and steps of the *k*-means algorithm allows to use the same methods for increasing the stability and preciseness of its results.

References

[1] Kazakovtsev L A, Antamoshkin A N, Fedosov V V 2016 Greedy heuristic algorithm for solving series of EEE components classification problems *IOP Conference Series: Materials Science and Engineering* Vol. **122** Article ID 012011. 7 P. DOI: 10.1088/1757-899X/122/1/012011.

[2] Fedosov V V, Orlov V I Minimal necessary extent of examination of microelectronic products at inspection test stage *Izvestiya Vuzov. Priborostroenie*, Volume **54(4)** pp. 62-68.

[3] Koplyarova N V, Orlov V I 2014 About a research of computer system of diagnostics of EEE on the basis of these tests *Vestnik SibGAU* Volume **1(53)**, pp.24-30.
[4] Kazakovtsev L A, Orlov V I, Stupina A A and Masich I S 2014 Problem of electronic components classifying *Vestnik SibGAU* **4(56)** pp 55-61
[5] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. Top 10 algorithms in data mining. URL http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf
[6] Kazakovtsev L A Greedy Heuristic Method for Systems of Automatic Object Groupping: Doctoral Dissertation. — Krasnoyarsk, SFU, 2016.
[7] Vorontsov K V Lectures on algorithms of a clustering and multidimensional scaling http://www.machinelearning.ru
[8] Cherezov D S, Tyukachev N A 2009 Review of the main methods of classification and clustering of data. *Voronezh State University. Vestnik VSU. Series: System analysis and information technologies* Vol. **2**
[9] Korolev V Y 2007 EM algorithm, his modifications and their application to a problem of division of mixes of probabilistic distributions. *Theoretical review*
[10] Kazakovtsev L A, Antamoshkin A N, Masich I S 2015 Fast deterministic algorithm for EEE components classification. *IOP Conference Series: Materials Science and Engineering* Vol. **94** Article ID 012015. 10 P. doi:10.1088/1757-899X/94/1/012015.
[11] Kazakovtsev L A, Orlov V I, Stupina A A 2015 [On Distance Metric for the System of Automatic Classification of the EEE Devives by Production Batches] // Programmnye produkty I sistemy Issue **2** P. 124-129. doi: 10.15827/0236-235X.110.124-129. In Russian.
[12] Fedosov V V, Kazakovtsev L A, Gudyma M N 2016 Problem of Normalization of Input Test Data of Space Industry EEE Components for Automatic Groupping Algorithm *Informatsionnye tekhnologii modelirovaniya I upravleniya* Issue **4** P. 263-268. In Russian.
[13] Software environment for statistical computing and graphics R. URL https://cran.r-project.org/
[14] Chen W.-C., Maitra R 2015 EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution. R Package, URL http://cran.r-project.org/package=EMCluster Chen, W.-C., Maitra, R. (2015) A Quick Guid for the EMCluster Package (Ver. 0.2-5). R Vignette, URL http://cran.r-project.org/package=EMCluster
[15] Chris Fraley, Adrian E, Raftery T Brendan Murphy and Luca Scrucca (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington Chris Fraley and Adrian E. Raftery (2002) Model-based Clustering, Discriminant Analysis and Density Estimation Journal of the American Statistical Association 97:611-631