Technical University of Varna Annual Journal
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30
ISSN 2603-316X (Online)

Processing of data from complex objects through pattern recognition methods

Mariya Ivanova Konsulova-Bakalova[1]

1-Technical University of Varna, Department of Mechanical Engineering and Machine Tools, 9010, 1 Studentska Street, Varna, Bulgaria

Corresponding author contact: mbakalova@tu-varna.bg

Abstract. In the description of complex objects, we need methods which could reflect the complex interconnections between components and sift out if possible those of them which are substantial for the specific application. It is offered in this publication the pattern recognition methods should be used as a unified method for processing of data from complex objects. The proposed algorithm may be used in the recognition of the condition of objects of various nature. The indicated examples prove the practical applicability of the methodology as they represent the solution of specific practical problems.

Keywords: pattern recognition, statistical processing, complex objects

1 Introduction

At the processing of data from scientific experiments as well as frequently at solution of purely practical problems, the processing of a multitude of data from the so called complex objects is usually imposed. To clarify the concept complex object, we may turn to the theory of the complex systems (**Bar-Yam, 2002**), (**Митев, Димитров § Узунов, 2004**). A system with sufficiently quite a few components, amongst which a multitude of interactions is observed, described with complex mathematical expressions, is accepted as a complex system. A similar definition may be given for a complex object – an object, whose description a multitude of mutually connected and interacting elements should be taken into consideration at. The human brain, the human body, an energy system, a machine and so on shall be a complex object.

The common thing amongst all the complex objects is that at the description of their condition, we need methods which could reflect the complex interconnections between components and sift out if possible those of them which are substantial for the specific application.

It is offered in this publication that the known pattern recognition methods should be used as a unified method for processing of data from complex objects.

2 Pattern recognition methods. Mathematical foundations.

Pattern recognition may be accepted as an aggregate of methods and technical means for indirect assessment of the condition of the object with regard to results from the measurement of a defined totality of secondary features (**Недев, 2012**). The measured secondary features may also be called classification features or information channels. The idea of the pattern recognition methods is to make an assessment of the condition of the complex object in conformity with the data from the information channels. The collected data usually consist of several groups of observations with various probability characteristics. The totality of values of the features describing the momentary condition of the object is meant as observations. It is appropriate to use a discriminatory analysis due to the heterogeneity of the data during their processing. This is the other name of classification or pattern recognition. It is suitable, however, to make an analysis of the collected information about the target of the investigation prior to the initiation of the procedures mentioned hereinabove. This is why the assessment of the information value of the features for recognition is accepted as the first task. In other words – selection of these information channels which bear substantial information about the solution of the specific problem and exclusion, if possible, of those which are not informative in the event.

Technical University of Varna
Annual Journal
ISSN 2603-316X (Online)
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30

At assumption for normal distribution of the features, the dispersion ratio (Fisher criterion) may be used as a criterion for assessment of the information value (**Божанов & Вучков, 1983**). We have the following sequence of calculations:

– Particularization of the mathematical expectations m and the dispersions σ_2 for each of the features (**Rumbos, 2009**):

$$\mu_k = \frac{1}{m_k} \sum_{X \in X_k} x \tag{1}$$

where x – vector of observation, μ_k - vector of means (mathematical expectations) per classes (к – in number), m_k - arithmetic mean.

The internal group dispersion may be calculated from:

$$S_k = \sum_{x \in w_k} (x - \mu_k)(x - \mu_k)^T \tag{2}$$

$$S_W = \sum_{k=1}^{c} S_k \tag{3}$$

And the intergroup dispersion is:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{4}$$

– Determination of the multidimensional normal density of distribution per classes;

$$f(x_1, x_2, ..., x_n) = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\left(\overline{X} - \overline{M_x}\right)^T V^{-1} \left(\overline{X} - \overline{M_x}\right)\right]\right\}, \tag{5}$$

where $\overline{X} = [x_1, x_2, ..., x_n]$ - vector of diagnostic features; $\overline{M_x}$ - vector of mathematical expectations; V^{-1} - inverse covariance matrix.

– Calculation of the optimal value of the dividing limiting line (**Vapnik, 2000**). An expression for calculation of the dividing limiting line in two state classes is given herein below:

$$Y_0 = \sigma_{cp}^2 \ln\left(\frac{PI}{PII}\right) \frac{1}{(m_{X_2} - m_{X_1})} + \frac{m_{X_2} + m_{X_1}}{2}, \tag{6}$$

where PI and PII – conditional densities of distributions per classes

– Calculation and analysis of the errors and the risks in the vicinity of Y_0:

$$\varepsilon(y_0) = \varepsilon_1(y_0) + \varepsilon_2(y_0), \tag{7}$$

where $\varepsilon_1(y_0)$, $\varepsilon_2(y_0)$ - probabilities of making errors of the first and second order.

– Calculation of the dispersion ratio (Fisher):

$$F = \frac{S_B^2}{S_w^2} \tag{8}$$

If $F<F_T$ (F_T is a tabular value) the influence of the factor may not be deemed as substantial. Otherwise the factor may be deemed as significant.

Technical University of Varna
Annual Journal
ISSN 2603-316X (Online)
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30

The information value of each of the features is determined at this stage with regard to the values of the minimal errors and the calculated dispersion ratio. The proposed calculations do not represent complex mathematical operations. They may be easily made in any of the frequently used software products of the type of Matlab, Excel and so on.

After the determination of the optimal totality of features the solution of the main part of the problem may be passed to – assessment or recognition of the status. Each algorithm for status assessment consists of two parts – training and recognition (**Недев & Тенекеджиев, 1994**). Here is the place to note down that each recognition algorithm is capable of recognizing only states which it was preliminarily trained for. In some methods of recognition, once set up for certain state classes, the recognition system may not be reset. This should be taken into consideration at the preliminary formation of the state classes and is the task of the narrow specialists from the sphere, which the recognition is made in. For instance, at assessment of the technical condition of a machine, the specialist who works with it, should determine which aggregate of features corresponds to a machine in a good working order and which to a faulty working machine.

The methods and the algorithms for recognition are divided into two groups – determined and statistical (stochastic). As it was also mentioned hereinabove, the discriminatory analysis is one of the most frequently used procedures for pattern recognition. We should learn through it about each observation from the heterogeneous extract of data how to determine its belonging to the class which it originates from.

Linear discriminant functions are created for each of the classes in the linear discriminatory analysis. The observation is classified to the class with maximal discriminant function. The convenience of the method consists of the fact that it may be used both for two state classes and for a bigger number. Depending on the a priori information, which we have, the discriminant functions may have a different kind. One of the most frequently used variants is:

$$g_i^{VI}(X) = \ln P(X / w_i) + \ln P(w_i) + \ln |c_i| = \max, \tag{9}$$

where $P(X/w_j)$ – conditional densities of the distributions of X per classes, c_{ij} – elements of the matrix of losses.

A decisive rule is obtained with normal distribution of the features, which is reduced to minimization of the Mahalonobis Distance (**Duda, 2001**):

$$r_i^2 = (X - \mu_i)^T \sum_i^{-1} (X - \mu_i) \Rightarrow \min \tag{10}$$

Then the current observation shall refer to the state class, which it is nearest located to within the Mahalonobis meaning. The following is obtained after replacement of the expression for normal density of distribution in the discriminant function:

$$g_i(X) = -\frac{1}{2}(X - \mu_i)^T \sum_i^{-1} (X - \mu_i) - \\ -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\sum_i\right| + \ln P(w_i) + \ln |c_i| = \max \tag{11}$$

We accept that some of the members of the function are inessential for the classification. After their disregard we have:

Technical University of Varna
Annual Journal
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30
ISSN 2603-316X (Online)

$$g_i(X) = -\frac{1}{2} X^T \sum\nolimits_i^{-1} X + (\sum\nolimits_i^{-1} \mu_i)^T X - \frac{1}{2} \mu_i^T \sum\nolimits_i^{-1} \mu_i - \\ -\frac{1}{2} \ln \left| \sum\nolimits_i \right| + \ln P(w_i) + \ln |c_i| = X^T A_i X + a_i^T X + a_{i0} = \max \quad (12)$$

The first two members of the equation express the quadratic Mahalonobis Distance, and the free member reflects the particularities of the strategy. We may enter coefficients in formula (12) as follows:

$$A_i = -\frac{1}{2} \sum\nolimits_i^{-1} \quad (13)$$

$$a_i = \sum\nolimits_i^{-1} \mu_i \quad (14)$$

$$a_{i0} = -\frac{1}{2} \mu_i^T \sum\nolimits_i^{-1} \mu_i - \frac{1}{2} \ln \left| \sum\nolimits_i \right| + \ln P(w_i) + \ln |c_i| \quad (15)$$

Depending on the available information, we have different varieties of recognition strategies – Bayes, maximum-likelihood (**Недев, 2012**). If any a priori information is missing, in coefficient a_0 the last two members shall be dropped out. In this case we have an absence of information about the losses and the errors from the first and second order.

The following algorithm for training and recognition shall be formed with observation of this sequence of the calculations:

- Training – the coefficients of the discriminant functions shall be calculated. Their values shall be stored and shall form the discriminant function;
- Recognition – The recognition shall be performed making use of the already calculated in the preceding procedure coefficients, which are replaced in formula (12) and the determinant functions for each of the state classes shall be obtained. Classification shall be made to this class whose discriminant function is with the biggest value.

It is useful to introduce a criterion for assessment of the recognition of the use of static methods for data processing. The assessment as a rule is made in recognition of control extracts with known belonging. The use of discriminant functions provides the possibility for calculation of the a posteriori probability of each hypothesis. It is appropriate to use the following formula:

$$P(w_i / X) = \frac{e^{r_i(X)}}{\sum_{i=1}^{c} e^{r_i(X)}} \quad (16)$$

c – number of classes; r_i – discriminant function for the relevant class

A quantitative assessment of the authenticity of each classifying solution which is accepted or rejected may be made on the basis of this equation as well as the general algorithm for recognition per minimal risk may be realized in a completed kind.

The visual presentation of the data with their distribution per classes is one of the difficulties with the availability of a big number of state classes at the processing and the presentation of data from complex objects. Similar visualization provides a good idea for the classification. With two infor-

Technical University of Varna
Annual Journal
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30
ISSN 2603-316X (Online)

mation channels the presentation may also be made through the two-dimensional density of distribution (see Fig. 1).

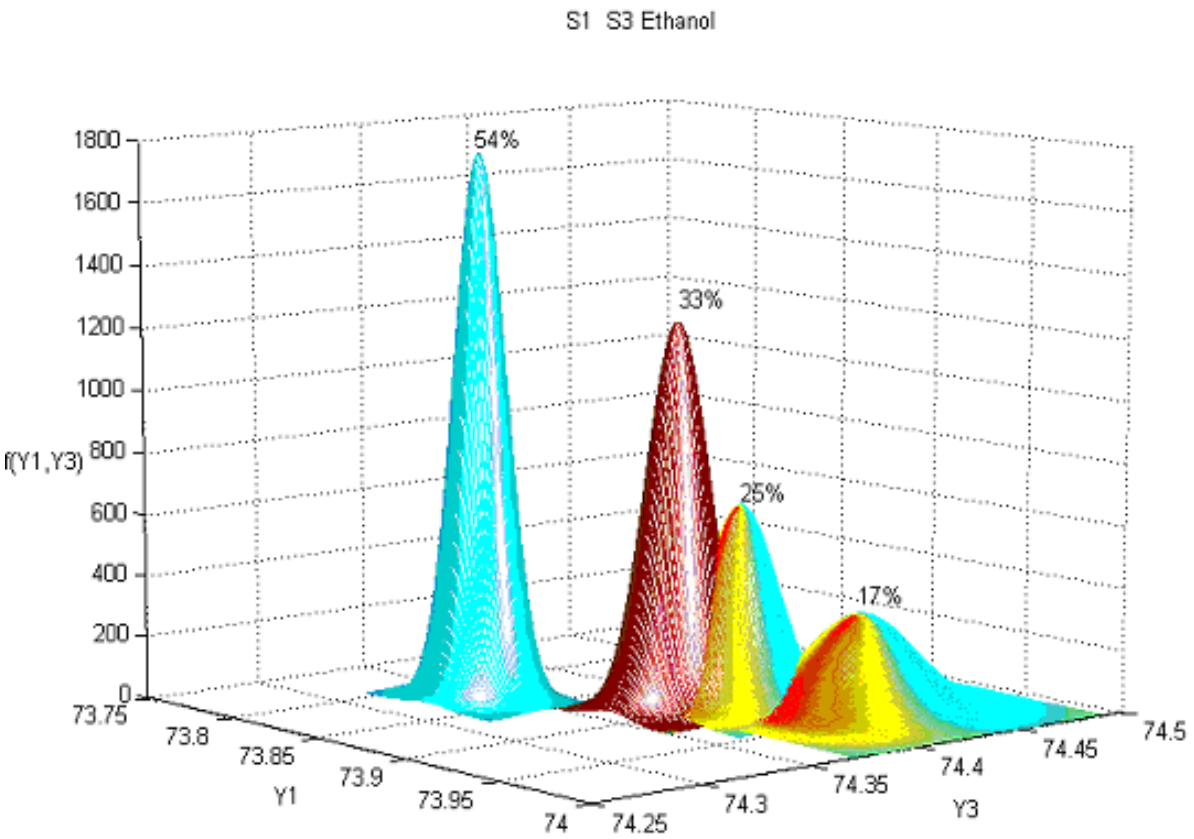

Fig. 1. Two-dimensional density of distribution with four state classes (four different gases).

With more information channels the density of distribution may not be presented in a Descartes Cartesian Coordinate System. With a multi-dimensional vector of observation the so called Principal Component Analysis method may be used (**Bro & Smilde, 2014**). It includes mathematical procedures, which transform a certain number of correlated variables into a fewer in number non-correlated variables called Principal Components. Each axis or base vector is qualitatively independent of all the rest. The first axis was selected so that it shall present the direction with the biggest dispersion, the second indicates the second biggest dispersion and so on. Said otherwise, the principal components are arranged in conformity with the degree, by which they describe the behavior of the initial variables – the first principal components are of the biggest significance, each following one explains a smaller and smaller part of the "non-described" dependencies. The advantage of the procedure is that it provides a possibility to present in the two-dimensional or three-dimensional space a multitude of state classes. The complete set of principal components is as big as the number of the original variables is (Fig. 2). In most events, however, it is possible to present over 80 per cent of the information making use of only two directions (**Jolliffe, 2002**).

Technical University of Varna
Annual Journal
ISSN 2603-316X (Online)
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30

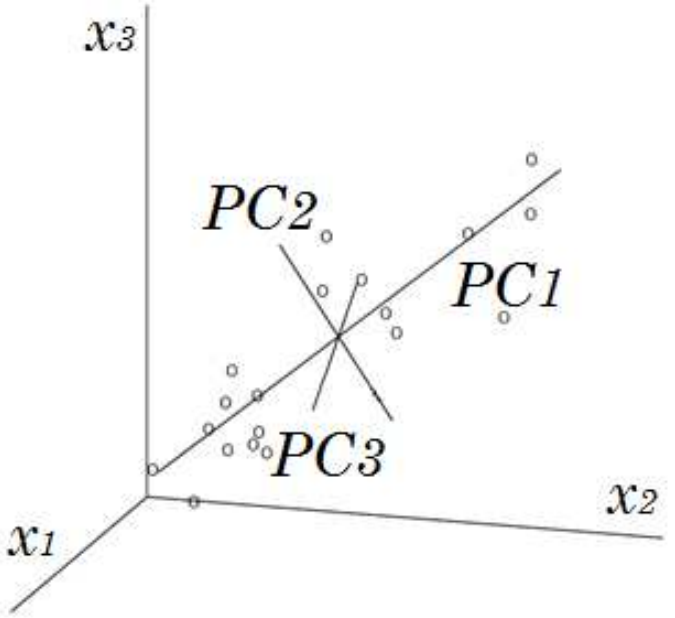

Fig. 2. Feature space and principled components.

The flow of the calculations with the principal components method is presented in Fig. 3. The essence of the analysis is the singular decomposition of the type X = UDV. The latent values of the covariance matrix of X are in matrix D. Matrix V gives us the principal components (PC) (formula 17). The new basis of data (score) is obtained at designing matrix X over V. Their drawing provides presentation in the space of the main directions.

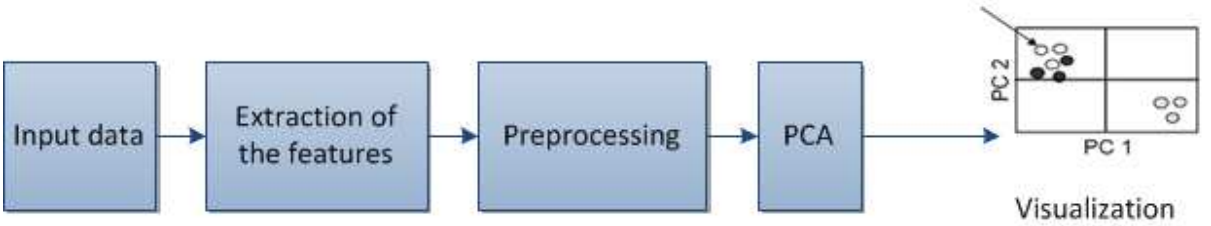

Fig. 3. Method of the principal components.

$$X_{M\times N} = U_{M\times N} D_{N\times N} V^T_{N\times N} = (u_1, u_2, \dots, u_N) \begin{bmatrix} \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{NN} \end{pmatrix} \end{bmatrix} (\vartheta_1, \vartheta_2, \dots, \vartheta_N) \quad (17)$$

Usually the data in the new orthogonal coordinate system are presented through the so called "scatter plot" as a multitude of points (Fig. 4). The distances between the points represent the known Mahalonobis Distance.

Technical University of Varna
Annual Journal
ISSN 2603-316X (Online)
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30

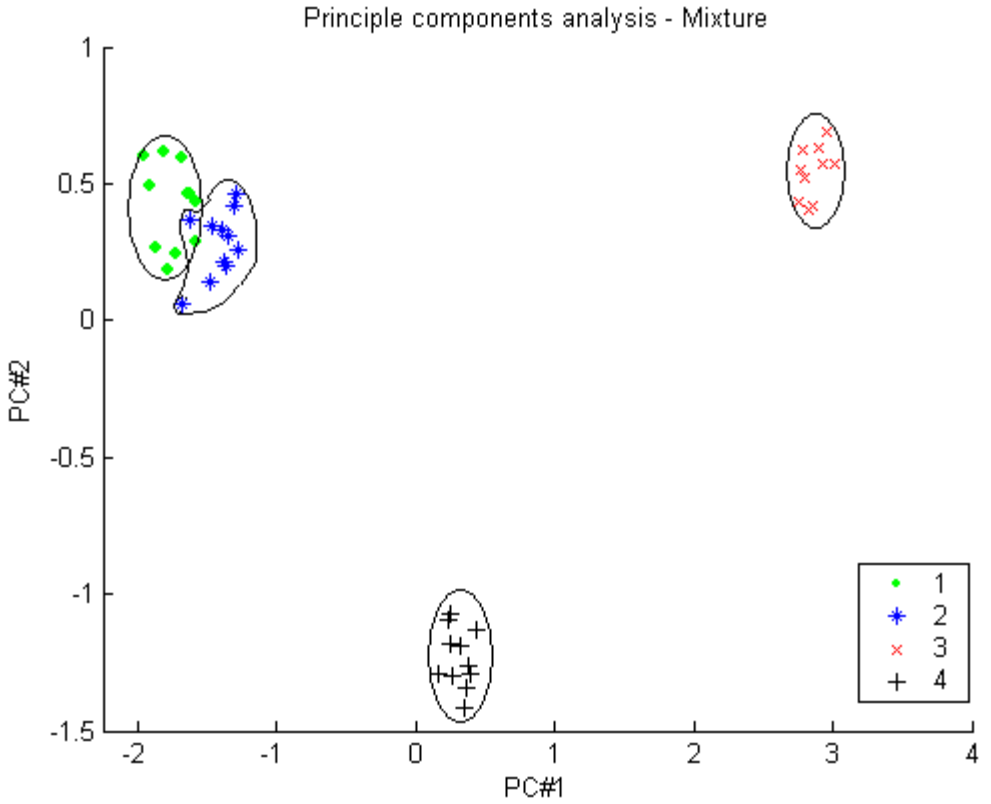

Fig. 4. Four state classes in the space of principal components.

The general algorithm for processing of data from complex objects may be seen in Fig. 5. These workflows have been developed in Matlab and a working software product has been created. It is anticipated that the assessment of the information value of the features for recognition should be made in conformity with the wish of the user. It is possible to conduct training not only one-time but also at a later stage if it is deemed as needed.

A similar algorithm may easily be realized through various software products because it includes comparatively simple calculation methods.

3 Application

The proposed algorithm for processing of data from complex objects may be used in the recognition of the condition of objects of various nature. A lot of tasks for determination of the condition of complex objects are shown in (**Недев и др., 2012**), whose basis the pattern recognition methods stand in. The spheres of application are navigation, energy efficiency, diagnostics of chemical equipment, protection of environment, medicine and management of academic and medical structures. The indicated examples indisputably prove the practical applicability of the methodology as they represent the solution of specific practical problems and are on the basis of successfully defended doctoral dissertations of the authors.

Recognition of the layers of soil is conducted in (**Naskova, 2017**) and the condition of the soil microbiological activity is investigated with regard to indirect features in (**Konsulova, Naskova & Malcheva, 2017**).

The common thing between the proposed examples is the availability of blurred data and the need that the recognition or the assessment of the condition should be made in conformity with indirect features. After the initial collection of the data at each of the problems we get rid of the specifics of the concrete sphere and making use of a discriminatory analysis we enter the static data processing.

Technical University of Varna
Annual Journal
ISSN 2603-316X (Online)
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30

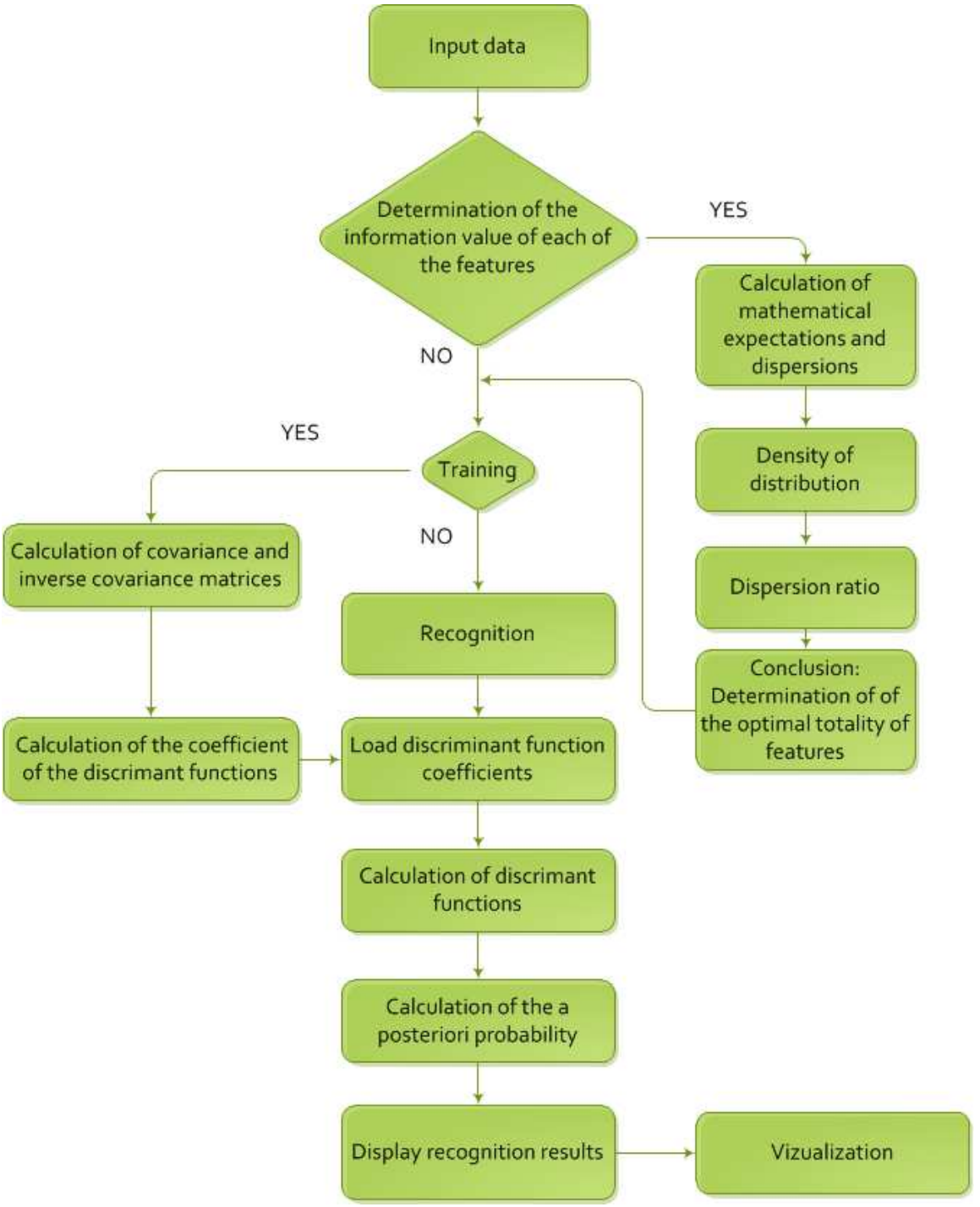

Fig. 5. Algorithm for processing.

4 Conclusions

An overall algorithm for analysis and processing of data from complex objects was developed. The first stage of assessment of the capability for division of each of the multitude of features for recognition is useful in laboratory conditions, in the stage of selection of an aggregate of information channels.

Procedures for training and recognition are proposed with the multi-dimensional analysis. It is also possible that an assessment of the quality of recognition in conformity with the a posteriori error from recognition is obtained. These results would be sufficiently indicative and useful both for narrow specialists and for a broader circle of users.

Technical University of Varna Annual Journal
DOI: 10.29114/ajtuv.vol2.iss1.69
Vol 2 Issue 1 (2018)
Published: 2018-06-30
ISSN 2603-316X (Online)

An analysis is made in the sphere of the principal components. The presentation of the data in the coordinate system of the principal components provides a clear idea of the remoteness of the classes and a good termination of the general algorithm for presentation and recognition of data from multi-channel gas analyzers.

The procedures are efficient and may be used for processing of data from complex objects with various spheres of application. One of the possible areas of application is in the case of precision diagnostics of machinery and equipment.

References

Bar-Yam, Yaneer (2002). *General Features of Complex Systems*. Encyclopedia of Life Support Systems. EOLSS UNESCO Publishers, Oxford, UK. Retrieved 16 September 2014

Duda R.O., P.E. Hart, D.G. Stork (2001), *Pattern Classification*, Wiley Interscience.

Jolliffe I.T. (2002), *Principal Component Analysis*, Second Edition, Springer Series in Statistics. https://doi.org/10.1007/b98835

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical Methods, 6(9), 2812-2831. https://doi.org/10.1039/c3ay41907j

Rumbos A.J. (2009). *Statistical Theory*. Lecture Notes. http://pages.pomona.edu/~ajr04747/Fall2009/Math152/Notes/Math152NotesFall09.pdf

Vapnik, V. N. (2000). The Nature of Statistical Learning Theory. "Springer", N.Y. doi:10.1007/978-1-4757-3264-1

Naskova, P. (2017). Mathematical Model for Evaluation of the Content of Heavy Metals in Soil by Indirect Plant Signatures. New knowledge Journal of science, 6(3), 149-160.

Божанов, Е. С., & Вучков, И. Н. (1983). Статистически методи за моделиране и оптимизиране на многофакторни обекти. Техника.

Konsulova, M., Naskova, P., Plamenov, D., & Malcheva, B. (2017). Recognition and probability of migrant microbiological activity by indirect signature. Pochvoznanie, agrokhimiya i ekologiya/Bulgarian Journal of Soil Science, Agrochemistry and Ecology, 51(3/4), 12-20. (In Bulgarian)

Митев, Д. Г., Д. Димитров, Р. Узунов (2004). *Управление на сложни системи*. Изд. на Шуменски унив. Епископ Константин Преславски. Център за дистационно обучение. ISBN 954-577-229-8

Недев А. (2012). *Разпознаване на образи и оптимално стохастичеко управление, I част*, ИК „Геа-Принт“, Варна, ISBN 978-954-9430-80-6

Недев А., К. Тенекеджиев (1994). *Техническа диагностика и разпознаване на образи*, Издателство ТУ-Варна.

Недев А., М. Бакалова, Г. Антонов, Б. Андреев, С. Сезгин, Д. Камберов (2012). *Разпознаване на образи и оптимално стохастичеко управление. Приложение на методите за разпознаване на образи в управлението на стопански, биологични и обществени системи*, ИК „Геа-Принт“, Варна, ISBN 978-954-9430-91-2