THE PROCEDURES OF VISUAL ANALYSIS FOR MULTIDIMENSIONAL DATA VOLUMES

A.E. Bondarev [1]

[1] Keldysh Institute of Applied Mathematics RAS, 125047 Miusskaya sq. 4, Moscow, Russia - bond@keldysh.ru

Commission II, WG II/10

KEY WORDS: Multidimensional Data, Visual Analysis, Elastic Maps, Cluster Structures

ABSTRACT:

The paper is devoted to problems of visual analysis of multidimensional data sets using an approach based on the construction of elastic maps. This approach is quite suitable for processing and visualizing of multidimensional datasets. The elastic maps are used as the methods of original data points mapping to enclosed manifolds having less dimensionality. Diminishing the elasticity parameters one can design map surface which approximates the multidimensional dataset in question much better. Then the points of dataset in question are projected to the map. The extension of designed map to a flat plane allows one to get an insight about the structure of multidimensional dataset. The paper presents the results of applying elastic maps for visual analysis of multidimensional data sets of medical origin. Previously developed data processing procedures are applied to improve the results obtained - pre-filtering of data, removal of separated clusters (flotation), quasi-Zoom.

1. INTRODUCTION

In the analysis of multidimensional data a special place is occupied by the task of classification. When solving classification problems, the approaches of visual analytics are very useful. They are the synthesis of several algorithms for reducing the dimension and the visual presentation of multidimensional data in manifolds of a lower dimension nested in the original volume. These algorithms include the display of the original multidimensional volume in elastic maps (Zinovyev, 2000), (Gorban et al, 2007), (Gorban, Zinovyev, 2010) with different properties of elasticity. These methods allow to get insight of the cluster structure contained in the initial multidimensional data volume under question.

Our team became interested in elastic maps in the process of implementing a project to develop computational technologies for building, processing, analyzing and visualizing multidimensional parametric solutions of CFD problems. Computational technology is implemented in the form of a single technological pipeline of algorithms for the production, processing, visualization and analysis of multidimensional data. Such pipeline can be considered as a prototype of a generalized computational experiment for non-stationary problems of computational gas dynamics. As a result, such a generalized computational experiment makes it possible to obtain a solution not for a single individual problem, but for a whole class of problems, defined by ranges of variation of the determining parameters. It should also be noted the universality of such approach. It can be applied to a wide range of problems of mathematical modeling of non-stationary processes. The description of the elements of the implemented computing technology is given in (Bondarev, Galaktionov, 2015a) and (Bondarev, Galaktionov, 2015b).

In practice, elastic maps turned out to be a useful and quite versatile tool, which made it possible to apply them to multidimensional data volumes of various types. This approach was applied to the tasks of analyzing textual information, where the frequencies of using words (Bondarev et al, 2016) were used as numerical characteristics, as well as to the tasks of analyzing mineral samples (Niedoba, 2014). In the process of working on these tasks, a number of procedures for processing the studied data were developed and tested, which made it possible to improve the results of visual analysis. These procedures include the preliminary filtering of data, which allows weeding out points with indistinctly defined values, the removal of separated clusters (flotation), quasi-Zoom. The latter procedure is designed to solve the problem of scalability, when the elastic map adapts both to the area of data points concentration and to separately located points of the data cloud, which complicates visual analysis. The essence of this technological approach is that for finer adjustment it is necessary to select large clusters in the studied volume of multidimensional data and build elastic maps for selected clusters separately, thus organizing an effect similar to the zoom function in modern phototechnics. The results of applying these procedures to multidimensional volumes of data of various origins are presented in (Bondarev et al, 2016), (Bondarev, 2017), (Bondarev, Bondarenko, Galaktionov, 2018).

This approach is generally universal, since it does not depend on the nature of the studied multidimensional data. This makes it possible to apply this approach and the developed procedures to the tasks of studying multidimensional medical data. This paper represents the results of applying the construction of elastic maps and procedures developed earlier for the visual analysis of multidimensional data volumes of medical origin.

2. ELASTIC MAPS APPROACH

The ideology and algorithms for construction of elastic maps are described in detail (Zinovyev, 2000), (Gorban et al, 2007), (Gorban, Zinovyev, 2010). Elastic map is a system of elastic springs embedded in a multidimensional data space. This approach is based on an analogy with the problems of mechanics: the main manifold passing through the "middle" of the data can be represented as an elastic membrane or plate. The

method of elastic maps is formulated as an optimization problem, which assumes optimization of a given functional from the relative location of the map and data.

According to (Zinovyev, 2000), the basis for constructing an elastic map is a two-dimensional rectangular grid G embedded in a multidimensional space that approximates the data and has adjustable elastic properties with respect to stretching and bending. The location of the grid nodes is sought as a result of solving the optimization problem for finding the minimum of the functional:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \rightarrow min$$

where $|X|$ is the number of points in the multidimensional data volume X; m is the number of grid nodes, λ and μ are the elastic coefficients responsible for the stretching and curvature of the mesh. Here D_1, D_2, D_3 are the terms responsible for the properties of the grid. The term D_1 is a measure of the proximity of the grid nodes to the data. The term D_2 represents the measure of the stretching of the grid. The term D_3 represents the measure of the curvature of the grid.

The author of the approach (Zinovyev, 2000) has developed the software package (ViDaExpert, 2019), which allows the construction and visual presentation of elastic maps. The main functional features of this software are described in detail in (Zinovyev, 2000). The figures below in this article are created by means of this software package.

3. PROCEDURES FOR VISUAL ANALYSIS

Previously, to study multidimensional data, a number of procedures for processing the studied data were developed, which allowed to improve the results of visual analysis. These procedures include the preliminary filtering of data, which allows weeding out points with indistinctly defined values, the removal of separated clusters (flotation), quasi-Zoom. Below we briefly give examples of the application of these procedures to multidimensional volumes of data of different origin.

An example of constructing elastic maps for the volume of multidimensional data representing the characteristics of mineral resources, namely, three types of coal from Polish deposits (Niedoba, 2014), is given in (Bondarev, Bondarenko, Galaktionov, 2018). Multidimensional data are considered, representing points in the multidimensional feature space (characteristics of coal samples). The data set displays three grades of coal. The task of classifying coal by grade was considered. By combining the construction of elastic maps, the removal of fuzzy points and separated classes (filtering and flotation of data), it is possible to completely separate the samples specified in the initial volume into three classes corresponding to three types of coal.

Examples of the use of quasi-Zoom for analyzing the thematic proximity of the words of the Russian language are given in (Bondarev et al, 2016), (Bondarev, 2017), (Bondarev, Bondarenko, Galaktionov, 2018). The basis of the proposed method is the analysis of the environment of words. The main hypothesis is that similar words should occur in approximately the same context. In this regard, in the space of attributes, they will be located at a relatively close distance from each other, while the different words will be located at a distance more distant from each other. Text boxes from news sources were used as test data (news feeds for a certain period). For the primary tests, about 100 verbs with 353 nouns associated with them were selected. The data thus obtained was further considered as a multidimensional data volume, representing 100 points in 353-dimensional space. The numerical values of the resulting matrix are defined as frequencies of sharing. The data volume under study contained a region of high data density and points far enough from this region. In the study of the frequency of the joint use of verbs and nouns, the practical task was set as follows. It was necessary to separate the "stuck together" points. The use of filtering and two consecutive quasi-Zoom procedures allowed to solve this problem completely (Fig.1).

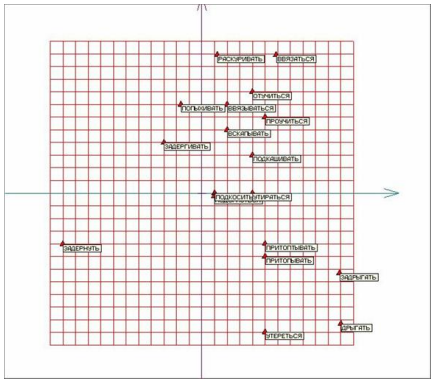

Figure 1. Extension of the elastic map after two consecutive quasi-Zoom applications

The use of a similar approach for the transposed data file allowed us to select among the set of nouns a number of semantic clusters (Fig.2). This opens up additional opportunities for the analysis and interpretation of semantic groups for specialists in this field.

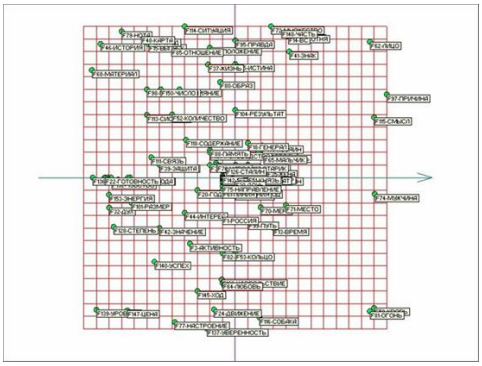

Figure 2. Extension of the elastic map for the transposed data set after applying quasi-Zoom

Also, the construction of elastic maps was applied to the study of multidimensional arrays of errors of different solvers compared to the etalon solution. We considered the numerical results of comparing the accuracy of the work of various solvers of the OpenFOAM software package using the example of the well-known inviscid flow problem around a cone at zero angle of attack. The results obtained using various OpenFOAM solvers were compared with the well-known numerical solution of this problem with the variation of the free-stream Mach number and the angle of the cone. Four solvers of OpenFOAM software package - *rhoCentralFoam*, *pisoCentralFoam*, *sonicFoam*, *rhoPimpleFoam* participated in the comparison. All these solvers have different approximation and computational properties. Figure 3 shows the elastic map for pressure, obtained as a result of parametric calculations, in the space of the first principal components. The yellow circles show the results for *rhoCentralFoam* solver, the red ones for

pisoCentralFoam, the green ones for *sonicFoam* and the blue ones for *rhoPimpleFoam*.

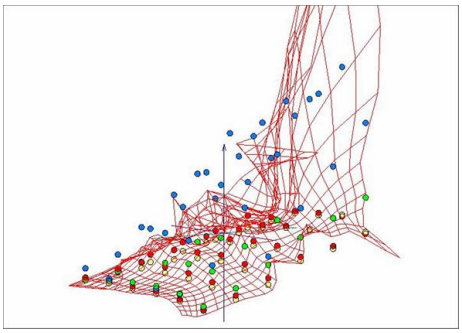

Figure 3. Elastic map for the array of errors for different OpenFOAM solvers

The results of the visual analysis showed that the errors for rhoCentralFoam and for pisoCentralFoam can be roughly approximated by a plane reflecting the dependence of the error on the Mach number and the cone angle.

4. PROCESSING OF MEDICAL DATASETS

The attempt of applying elastic maps to medical data was made in (Bondarev, 2017). . For this purpose the data from (Rocha Neto, A., Barreto, G., 2009) were used. This data set contains values for six biomechanical features used to classify orthopaedic patients into 2 classes (normal or abnormal). Each patient is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis. The data set contains 310 points in 6-dimensional space. Unfortunately, elastic maps didn’t give good results from the point of view of classification.

Below are the results for the three other volumes of multidimensional data that involve the solution of the classification problem. All data sets were taken from (UCI Machine Learning Repository, 2019).

The first data set considers variability of impedivity in normal and pathological breast tissue (Jossinet, 1996) and tasks of classifying various types of diseases (Silva, de Sá, Jossinet, 2000). This dataset contains 106 points placed in 9-dimensional attribute space. Also each point has its class attribute corresponding to the type of disease - carcinoma, fibro-adenoma , mastopathy, glandular, connective, adipose. According to (Silva, de Sá, Jossinet, 2000), the dataset can be used for predicting the classification of either the original 6 classes or of 4 classes by merging together the fibro-adenoma, mastopathy and glandular classes whose discrimination is not important (they cannot be accurately discriminated anyway).

Further, we use the following notation and color scheme for the classes studied: *car* (carcinoma) - red, *adi* (adipose) - yellow, *con* (connective) - green, *fad* + (fibro-adenoma + mastopathy + glandular) - blue. We use the combined *fad* + class because of the above remark by the authors of the volume of data that these classes are not separated exactly.

Below one can see the illustrations of the construction of elastic maps for the studied data volume. Figure 4 shows the source data in the space of the first three principal components. Figures 5 and 6 show the elastic map and its development for a given amount of data.

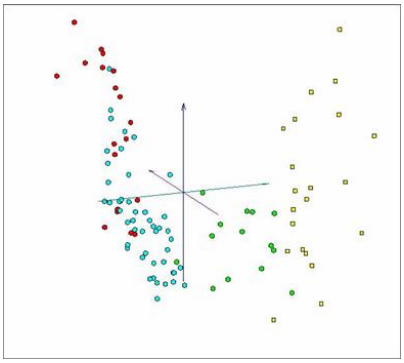

Figure 4. Source data in the space of the first principal components

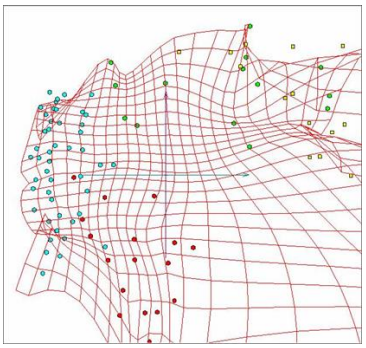

Figure 5. Elastic map for source data

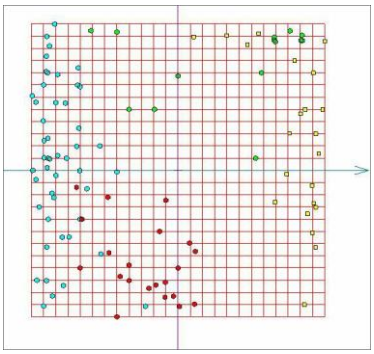

Figure 6. Extension of elastic map for source data

Figures show that (*car* + *fad* +) and (*con* + *adi*) pairs of classes are well separated. However, within the pair, data from different classes are mixed. To improve the picture of the separation, use flotation and remove *fad* +. The results of building an elastic map for this case are shown in Figure 7. In this case, the car class was fully distinguished.

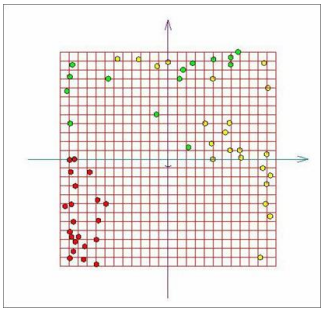

Figure 7. Extension of elastic map for classes *car*, *con*, *adi*

Now remove the car class and consider separately the remaining pair of classes - *con* and *adi*. After constructing the elastic map and its development, we obtain the picture presented in Figure 8. In this case, a satisfactory separation of classes was achieved.

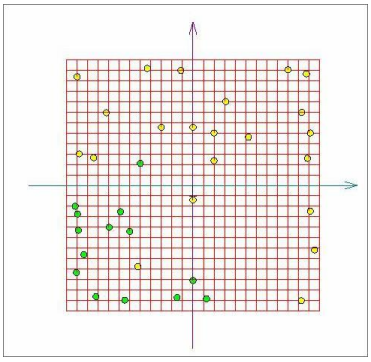

Figure 8. Extension of elastic map for classes *con*, *adi*

Next, consider together a couple of classes - *car* and *fad* +. Figure 9 presents the extension of the elastic map for these classes. There is also a satisfactory separation. The use of q-Zoom in order to improve the separation in the center of the picture did not lead to success. Also, the attempt to divide the mixed *fad* + class into the *fad*, *mas*, *gla* classes was not successful. The comment in (Silva, de Sá, Jossinet, 2000) about the inseparability of these classes turned out to be true.

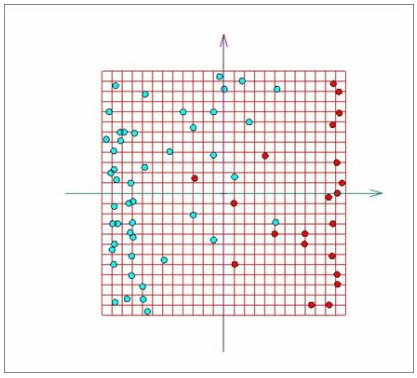

Figure 9. Extension of elastic map for classes *car*, *fad*+.

The following data set is also devoted to the problems of forecasting breast diseases (Crisóstomo et al., 2016), (Patrício et al, 2018). The data set contains 116 points in a 10 - dimensional attribute space. Each point also contains a binary variable indicating the presence or absence of the disease. Attribute space contains ten predictors. According to (Patrício et al, 2018) the predictors are anthropometric data and parameters which can be gathered in routine blood analysis.
Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.
For this data volume, an elastic map was constructed. Dots with the absence of the disease are shown in green, and the presence of the disease is marked in red.
Figures 10 and 11 represent the constructed elastic map and its extension. As one can see, the green and red dots are strongly mixed. This caused some confusion, since by construction this picture represents points that have to be close to each other in the multidimensional attribute space.

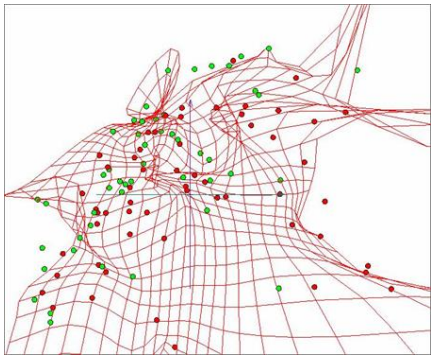

Figure 10. Elastic map for 10-dimensional attribute space

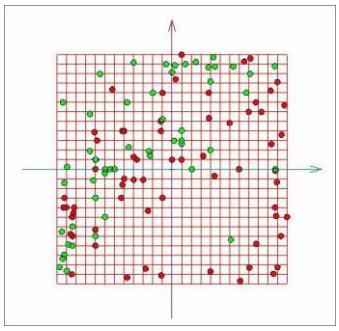

Figure 11. Extension of elastic map for 10-dimensional attribute space

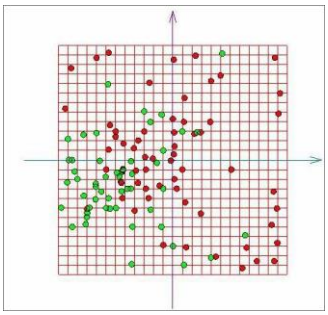

Figure 12. Extension of elastic map for 4-dimensional attribute space

However, in the original article (Patrício et al, 2018) a picture was given from which it was possible to conclude that only for 4 parameters (glucose, Insulin, Resistin, HOMA-homeostasis model assessment) there is a significant difference between patients and healthy people. From the data space, only these 4 dimensions were left, and the elastic map was re-constructed. The results are shown in Fig. 12. The separation between the green and red dots has improved significantly, however, in the center of the picture there is an area where the dots are mixed.

The following dataset is for the early diagnosis of the Autistic Spectrum Disorder (ASD) (Thabtah, 2017). The data set consists of 692 points originally defined in the 21-dimensional attribute space. The diagnostic approach is based on the analysis of the questionnaire data consisting of 10 questions. About half of the attributes are patient data. Therefore, it was decided to leave 12 attributes - 10 answers to the questionnaire, the age of the patient and the total score according to the results of the questionnaire. The results are presented in Figures 13 and 14 in the form of an elastic map and its scan.

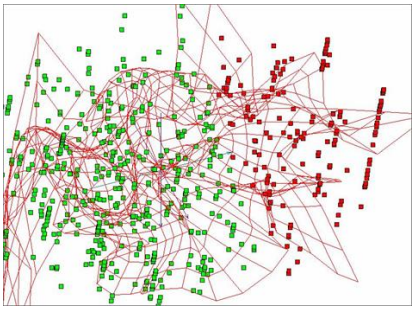

Figure 13. Elastic map for 12-dimensional attribute space when diagnosing ASD

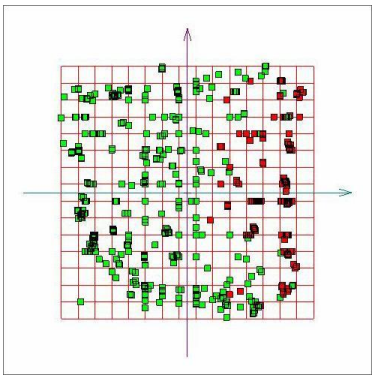

Figure 14. extension of elastic map for 12-dimensional attribute space when diagnosing ASD

These results show that the separation between diagnoses about the presence or absence of ASD is quite satisfactory on the studied data set.

5. CONCLUSIONS

For the analysis of structures in multidimensional data volumes, technologies for constructing elastic maps are used, which are methods for mapping points of the original multidimensional space to nested manifolds of lower dimension. A number of data processing techniques that can improve the results are considered - pre-filtering of data, removal of separated clusters (flotation), quasi-Zoom. Examples of the construction of elastic maps and the use of these procedures for multidimensional data of medical origin are given. The results showed that the construction of elastic maps together with the procedures of accompanying data processing can serve as a useful tool for visual data analysis and complement other methods for studying multidimensional data volumes.

ACKNOWLEDGEMENTS

Author acknowledges the support by grant of RFBR № 17-01-00444A.

REFERENCES

Bondarev, A.E. et al, 2016. Visual analysis of clusters for a multidimensional textual dataset. *Scientific Visualization*. 8(3), 1-24.

Bondarev, A.E., 2017. Visual analysis and processing of clusters structures in multidimensional datasets. *ISPRS Archives*, XLII-2/W4, 151-154.

Bondarev, A.E., Bondarenko, A.V., Galaktionov, V.A., 2018. Visual analysis procedures for multidimensional data. *Scientific Visualization* 10(4), 109 - 122, doi.org/10.26583/sv.10.4.09.

Bondarev, A.E., Galaktionov, V.A., 2015a. Analysis of Space-Time Structures Appearance for Non-Stationary CFD Problems. *Procedia Computer Science*, 51, 1801–1810.

Bondarev, A.E., Galaktionov, V.A. , 2015b. Multidimensional data analysis and visualization for time-dependent CFD problems. *Programming and Computer Software*, 41(5), 247–252, doi.org/10.1134/S0361768815050023.

Crisóstomo, J. et al., 2016. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. *Endocrine,* 53(2), 433-442, doi.org/10.1007/s12020-016-0893-x

Gorban, A. et al, 2007. *Principal Manifolds for Data Visualisation and Dimension Reduction*, Springer, Berlin – Heidelberg – New York, 2007.

Gorban A., Zinovyev A., 2010. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems*, 20(3), 219–232.

Jossinet, J., 1996. Variability of impedivity in normal and pathological breast tissue. *Med. & Biol. Eng. & Comput*, 34, 346-350.

Niedoba, T., 2014. Multi-parameter data visualization by means of principal component analysis (PCA) in qualitative evaluation of various coal types / *Physicochemical Problems of Mineral Processing*, 50(2), 575-589.

Patrício, M., et al 2018. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, 18(1), doi.org/10.1186/s12885-017-3877-1.

Rocha Neto, A., Barreto, G., 2009. On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis, *IEEE Latin America Transactions*, 7(4), 487-496.

Silva, J.E., Marques de Sá, J.P., Jossinet, J., 2000. Classification of Breast Tissue by Electrical Impedance Spectroscopy. *Med & Bio Eng & Computing*, 38, 26-30.

Thabtah, F., 2017. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. · *Informatics for Health and Social Care*, doi.org/ · 10.1080/17538157.2017.1399132

UCI Machine Learning Repository, 2019. archive.ics.uci.edu/ml/ (01 March 2019).

ViDaExpert, 2019. bioinfo.curie.fr/projects/vidaexpert (01 March 2019).

Zinovyev, A., 2000. *Vizualizacija mnogomernyh dannyh [Visualization of multidimensional data]*. Krasnoyarsk, publ. NGTU. 2000. 180 p. [In Russian]