DOI: https://doi.org/10.24818/cike2024.41
UDC: 004.421:324

THE LEVEL OF SIMILARITY AS A FUNCTIONS CLASSIFICATION MEASURE

ILIE COANDA
Assoc. Prof., PhD
Department of Information Technology and Information Management
Academy of Economic Studies of Moldova
Chisinau, Republic of Moldova
coanda.ilie@ase.md
ORCID ID: 0000-0002-0010-1202

Abstract. Data processing in the essence of the notion of "Data Mining" may require procedures - algorithms for assessing the similarity between the resulting functions of the studied phenomena. A model for evaluating the level of similarity between two functions is proposed. According to the way of approaching the essence of the study regarding the differences between two functions, in this paper an algorithm is to be presented and discussed, based on which certain numerical values could be obtained. The respective values, following a synthesis, are to constitute a set of parameters included in a mathematical expression that numerically expresses the "distance" between two functions. The level of similarity of two functions will be considered to be a positive numerical value, and for the functions that coincide, according to the model, the respective value of the level will be "zero". The basic properties of functions will be considered through the lens of the fundamental notions involved in the procedures for researching functions in the field of mathematics. Certain numerical values (parameters) characteristics of the essence of some notions will be highlighted and used, such as: monotonicity intervals, critical points, inflection points, convexity, concavity, extreme values, values of first-order derivatives and second-order derivatives. The values obtained for each of the previously listed properties are supposed to be calculated for each function included in the similarity evaluation process. Depending on the set of values among those listed, various algorithms can be defined. For example, considering only the monotonic intervals, one algorithm could be created, and if the inflection points are also included, another algorithm will be obtained, with a different result.

Keywords: similarity, evaluation, algorithm, functions, distance, measure

JEL Classification: C63, I21, I23, I25, I29

1. Introduction

The process of analyzing a set of functions that reflects the essence of the behaviors of some phenomena in human social activity involves passing through several stages, among which we mention the following: data collection, initial pre-processing; choice – creation of field-specific analysis and synthesis models; development of an appropriate algorithm(s); development of a (some) soft application for organizing case studies; formulating conclusions. Regarding data collection, the material is dedicated (COANDĂ, Ilie, 2022), in which, as the main conclusion, the need to involve certain effective models in order to obtain; to an appropriate level of accuracy, continuous approximating functions based on the respective discrete functions. In (COANDĂ, Ilie, 2023) certain results are presented regarding the importance of the impact of the involvement of the preliminary processing process, in which an approach to the essence of "smoothing" out of the ordinary values of the studied phenomenon is proposed, a methodology specific to the phenomenon in the field respectively. This paper discusses the extension or concretization of the essence of the evaluation of the level of similarity of two functions, about which, in (COANDĂ, Ilie, 2022), only tangentially, a certain approach was presented.

The impact of the evaluation efficiency on the level of similarity of two functions, in some cases, can be particularly significant. Along with the effects conditioned by the intervals of monotonicity and convexity, respectively concavity, other values - parameters, conditioned by other effects, could be included in the evaluation algorithms, such as, for example: a) a function has a number of extreme points different from the other function, b) the ordinates of the extreme points and of the inflection points, respectively, differ substantially; c) the value of the largest difference over the entire definition interval of the function, between the maximum and minimum value, difference related to the distance between two neighboring extremes, for one function is significantly different from the respective value of the other function; d) the effect of concavity or convexity, as an impact of second-order derivative values, is significantly different; e) the values of the ordinates of the absolute maximum, respectively, of the absolute minimum are significantly different, etc.

Considering the above, the peculiarity of the complexity in defining an adequate model for the effects of all parameters, a model that would reasonably contain the essence of the impact of each property, becomes apparent.

2. A set of parameters included the model for quantitative evaluation (an example)

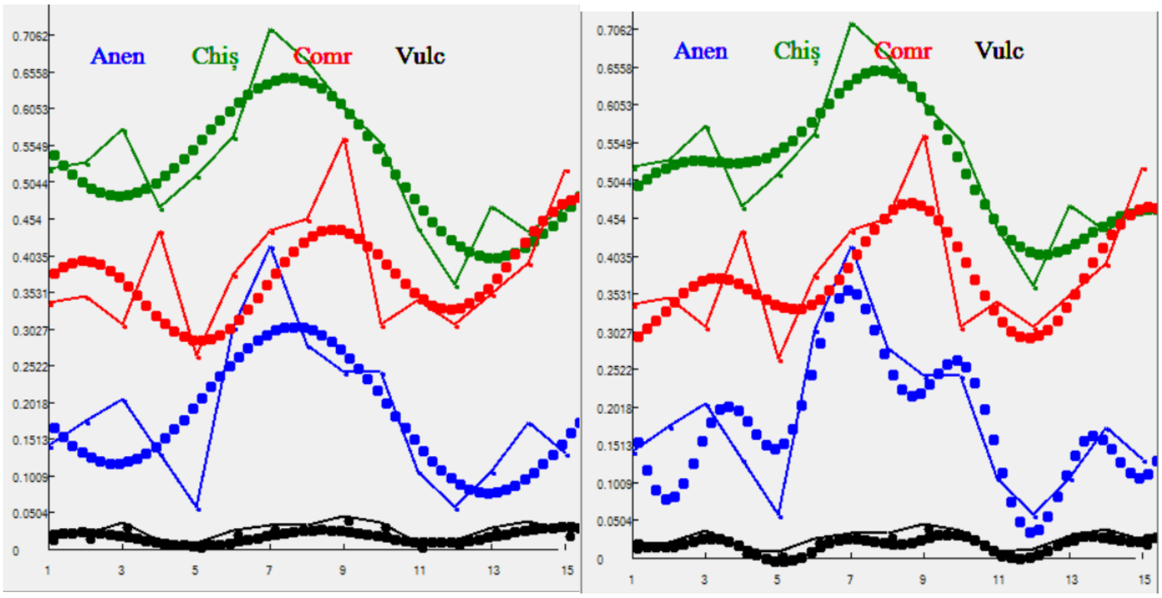
Based on the process of numerical evaluation of the level of similarity, some of the properties listed in the previous paragraph (Introduction) were included in the concrete model defined, as follows:

1. The intervals of monotonicity (the effect of the derivative of the first order);
2. The intervals of concavity/convexity (the effect of the second order derivative);
3. Positioning of the abscissas of the extreme points (the effect of the derivative of the first order equal to zero);
4. The positioning of the abscissas of the inflection points (the effect of the second-order derivative equal to zero);
5. The number of local minimum extreme points;
6. The number of local maximum extreme points;
7. Positioning on the abscissas of the absolute minimum and maximum values

In the given example, a set of properties consisting of 7 (seven) entities is proposed (see above). Each subset of elements (of which 7(seven) entities) could be considered as a base set for defining the quantitative value for evaluation. Therefore, the researcher is suggested the possibility of defining a general model based on all the items in the set. Thus there will be the possibility to organize the processes of obtaining the aggregate value based on the selected options. Certainly the results will be more or less different. Depending on the researched problem, as well as the need to ensure a certain level of accuracy of the assessment, the appropriate options can be chosen. For example, if, for a given problem, only the intervals of monotonicity and concavity/convexity are considered significant, then only the first two items are to be selected from the list. In this situation, the aggregated quantitative value includes the impact of only the intervals of monotonicity and concavity/convexity. Given the fact that the approximating functions are supposed to be obtained in analytical form, the respective parameters corresponding to any item in the set could be obtained analytically. It will just be necessary to derive the respective formulas for each subset of options. In addition, in the case of including a new item, a good part of the deduced formulas must be modified. For these reasons, the value for each selected property, are recommended to be obtained by numerical methods.

Figure 1 shows the results obtained for the approximating functions, in analytical form, with linear terms and trigonometric functions. A simple visual analysis explains quite clearly the effectiveness of the involvement of such expressions - functions in the information analysis process. The broken line represents the primary data, and the "continuous" ones - indicate the image of the obtained functions. In order to obtain a more appropriate broken line, we are suggested to involve several terms - trigonometric functions.

Remark. Primary data are collected by the author from public sources, from the Internet, during the COVID pandemic - 30.09.2020 - 15.10.2020. (number of infected / 1000 inhabitants)



a) Y(x)=a+bx+csin(kx) b) Y(x)=a+bx+csin(kx)+dcos(tx)

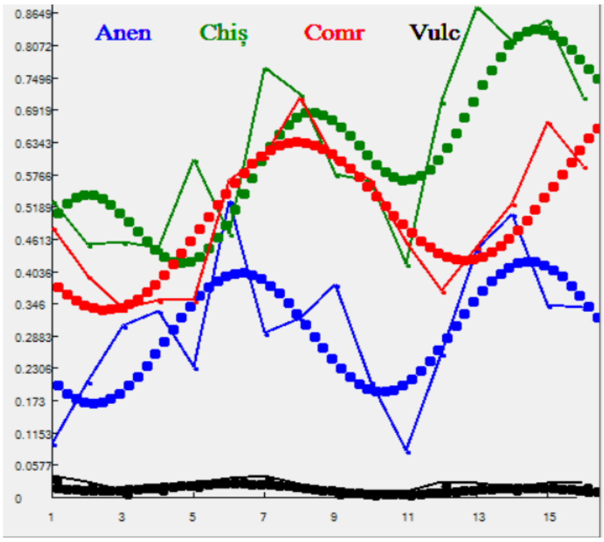
Figure 2 Impact of approximation's approach:

author's own study

In Figure 2, the images of the respective functions corresponds to the data collected in another time period, 06.10.2022 – 21.11.2020, which corresponds, as in Figure 1, to the involvement of two approximating function forms. And in this case, the efficiency of the involvement of several trigonometric terms is highlighted. For each of the displayed functions, the image on the right expresses the result of the approximation with a function of the form Y(x)=a+bx+csin(kx)+dcos(tx). (a, b, c, d, k, t – parameters).

The 7 (seven) criteria for comparing two functions listed in paragraph 2 express the essence of the differences, characteristics more pronounced in the image in Figure 2b). Thus the similarity between the second and third functions (green and red) is evident. Some level of similarity exists between the third and fourth functions (red and black).

However, the image in Figure 2a), left, the approximation results at a lower level, demonstrate a higher level of similarity, especially the first and third functions (green and blue), which, from a qualitative point of view, are almost identical. Thus, a dilemma arises: to accept the results according to the image Figure 2a) or those from Figure 2b). The solution rests on the researcher, who should also have additional information about the phenomenon, essence and the circumstances of data collection. Analyzing the data in Figure 2b) reveals a rather pronounced qualitative similarity between the second and fourth functions (respectively green, black). The values of the fourth function (black) are almost a constant, and the values of the second function (green) vary quite significantly relative to each other. If we were interested in both qualitative and quantitative behavior, then the second and third functions (green, respectively, red) could be considered the closest. Almost the same conclusions emerge from the analysis of the relative behavior of the functions presented in Figure 2a).



a) Y(x)=a+bx+csin(kx) b) Y(x)=a+bx+csin(kx)+dcos(tx)

Figure 2 Visual analyzing similarity between function:

author’s own study

3 Conclusions

- The comments regarding the information in Figures 2a) and 2b reveal the complexity of developing a model, within which the appropriate algorithms for moving from qualitative to quantitative analysis could be developed. Certain parametric and modeling techniques are to be defined according to the 7(seven) properties listed in chapter 2.
- If the researcher has the additional information that the localities "Anen" and "Comr" are high-level administrative centers, and in addition, that the respective data were collected during the second (final) round of the elections, he would formulate the conclusion that the intensity of communication between individuals was numerically comparable (the number of infected people depended on physical contact). So, in these localities the electoral agitation was just as intensive.

References

1. COANDA, Ilie. Evaluation of similarity of trend functions. In: Competitiveness and Innovation in the Knowledge Economy [online]: 26th International Scientific Conference: Conference Proceeding, September 23-24, 2022. Chişinău: ASEM, 2022, pp. 309-312. ISBN 978-9975-3590-6-1 (PDF). https://irek.ase.md/xmlui/handle/123456789/2607
2. COANDA, Ilie. The impact of data pre-processing on the assessment of the similarity of trend functions. In Annual international scientific conference “Competitiveness and Innovation in the Knowledge Economy”, [online]: September 22nd-23th, 2023, Chisinau, Republic of Moldova. DOI: https://doi.org/10.53486/cike2023.44, https://irek.ase.md:443/xmlui/handle/123456789/3096