Dimensionality Reduction of Data on Patients with Diabetes Mellitus by Multidimensional Scaling

Ievgen Meniailov[a], Serhii Krivtsov[b] and Tetyana Chumachenko[c]

[a] *V.N. Karazin Kharkiv National University, Kharkiv, Ukraine*
[b] *National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine*
[c] *Kharkiv National Medical University, Kharkiv, Ukraine*

Abstract

Diabetes Mellitus is a global public health problem. According to the World Health Organization, more than 6% of the world's population suffers from diabetes. In the context of the Russian invasion, the problem of diabetes is especially relevant for Ukraine. This is due to the difficulty of supplying medicines and obtaining medical care. Also, the stress caused by the war is one of the factors in the appearance and complications of diabetes. Automated models and information technologies for classifying patients with suspected diseases are practical decision support tools for making medical diagnoses in resource-limited settings. One of the problems with using such models is data redundancy. Therefore, this study uses multidimensional scaling to focus on dimensionality reduction in patients with suspected Diabetes Mellitus type II.

Keywords 1

Diabetes Mellitus, dimensionality reduction, multidimensional scaling

1. Introduction

Diabetes Mellitus is a disease characterized by increased blood sugar levels, leading to damage to the kidneys, and nervous system, impaired vision, and affecting the state of the nervous and vascular systems [1]. There are different types of diabetes, depending on which patient requires special treatment based on lifestyle changes, dietary choices, and medications. The disease can progress without symptoms for a long time, so many do not seek medical help promptly.

Diabetes is characterized by the following risk factors [2]:

- cardiovascular diseases;
- the predominance of carbohydrates in the food, leading to a violation of their metabolism;
- overweight and obesity;
- genetic predisposition;
- chronic stress;
- long-term use of drugs that contribute to the development of diabetes.

The main symptoms of the disease are:

- dry mouth and intense thirst;
- frequent and profuse urination;
- dry skin and mucous membranes;
- general weakness and fatigue;
- increased appetite;
- decreased vision;
- leg muscle cramps.

IDDM-2022: 5th International Conference on Informatics & Data-Driven Medicine, November 18-20, 2022, Lyon, France
EMAIL: evgenii.menyailov@gmail.com (IM); krivtsovpro@gmail.com (SK); tatalchum@gmail.com (TC)
ORCID: 0000-0002-9440-8378 (IM); 0000-0001-5214-0927 (SK); 0000-0002-4175-2941 (TC);

The most common is type II diabetes, which is characterized by high levels of insulin with low sensitivity of body cells to it [4]. This leads to damage to internal organs. The patient damages the retina, small vessels, nerves, and kidneys. As a result of malnutrition of the skin on the ankles, trophic ulcers form.

More than 400 million adults live with diabetes worldwide, which is growing yearly [5]. More than 60 million people have diabetes in the European Region [6]. To date, there are no official statistics on the incidence of diabetes in Ukraine. In 2017, 1.27 million people with diabetes were registered in Ukraine [7]. Among them, 200,000 patients need daily insulin.

One of the most effective tools to combat diabetes is its prevention and early detection. With the spread of the COVID-19 pandemic in the world, the number of studies aimed at applying information technology in healthcare has increased. Such studies were aimed at modeling the epidemic process [8, 9], analysis of medical data [10], analysis of medical images [11], analysis of factors in the spread of morbidity [12], analysis of the behavior of the virus [13], a study of the information content of factors affecting the dynamics morbidity [14], etc. Using mathematical modeling and information technologies to support doctors' decision-making when making medical diagnoses is practical. The problem in building models of medical diagnostics is the redundancy of data; therefore, reducing the dimensionality of data of patients with the suspected disease is an urgent task.

This study aims to develop a model to reduce the dimensionality of patient data on the incidence of diabetes based on the maximum likelihood method.

Research is part of a complex, intelligent information system for epidemiological diagnostics, the concept of which is discussed in [15].

2. Materials and Methods

The more information about the objects of study in the form of a set of characterizing features will be used to create a model, the better. However, too much information can reduce the efficiency of data analysis. It is important to note that non-informative features are a source of additional noise and affect the accuracy of model parameter estimation. In addition, datasets with a large number of features may contain groups of correlated variables. The presence of such groups of features means duplication of information, which can distort the model's specification and affect the quality of the estimation of its parameters. The higher the data dimension, the higher the size of calculations during their algorithmic processing [16].

High dimensionality can mean hundreds, thousands, or even millions of input variables. When dealing with high-dimensional data, it is often helpful to reduce the dimensionality by projecting the data onto a subspace of lower dimensions that retains the "essence" of the data. This is called dimensionality reduction [17]. More minor input data often means fewer parameters or a more straightforward structure in a machine learning model called degrees of freedom. A model with too many degrees of freedom is likely to overflow the training dataset and, therefore, may not work correctly on new data or not work at all.

The multidimensional scaling method is one of the well-known non-linear dimensionality reduction methods used to analyze the similarity (similarity or difference) of data by reducing data to a low-dimensional space [18]. It is also important to note that this method is one of the first fundamental teaching methods.

Multidimensional scaling (MDS) is a set of statistical methods dealing with the problem of constructing an n-point configuration in Euclidean space using dissimilarity information between n objects. It is not necessary to rely on differences between Euclidean distance objects; they can represent many types of differences. MDS aims to reflect objects before the configuration (or embedding) of points in such a way that the given differences are well approximated by the Euclidean distance [19].

MDS generally attempts to model data such as distances between points in geometric space. The main reason for this is that a graphical representation of the data structure is required, which is much easier to understand than an array of numbers, and, in addition, reflects essential information in the data, smoothing out the noise [20].

In MDS analysis, the data is typically embedded in a 2D or 3D map such that, given similarities or differences, the information matches the distances between points exactly. Objects of interest, such as objects, attributes, stimuli, respondents, etc., correspond to points in such a way that those nearby are empirically similar, and those far apart are considered different.

To evaluate the simulation result, two metrics were applied: Euclidean Distance [21] and Manhattan Distance [22].

The Euclidean Distance can be calculated from the Cartesian coordinates of points using the Pythagorean theorem, which is why it is sometimes called the Pythagorean distance. For observations a and b measured in multiple dimensions, this is $\sqrt{\sum_i ((a_i - b_b)^2)}$. It should be noted that even if you use zoom, normalize, or size weighting, the distance figure will still be the result. This is a good default distance measure if it makes sense to match the dimensions.

Manhattan or city-block distance is a distance introduced by Hermann Minkowski. According to this metric, the distance between two points equals the sum of the modules' differences in their coordinates $\sum_{i=1}^{N} |a_i - b_i|$. It is important to note that the Manhattan distance depends on the rotation of the coordinate system but does not depend on its mapping from the coordinate axis or offset.

3. Results

For the experimental investigation the Pima Diabetes dataset [23] has been used. Table 1 shows the parameters of the dataset. Distribution of the values by parameter is presented in Figure 1.

Table 1
Parameters of the dataset

Name	Scale type	Data range
Pregnancies	Metric	0…13
PG Concentration	Metric	44…197
Diastolic BP	Metric	0…110
Tri Fold Thick	Metric	0…60
Serum Ins	Metric	0…846
BMI	Metric	0…46.8
DP Function	Metric	0.134…2.288
Age	Metric	21…60
Diabetes	Nominal	Sick / Healthy

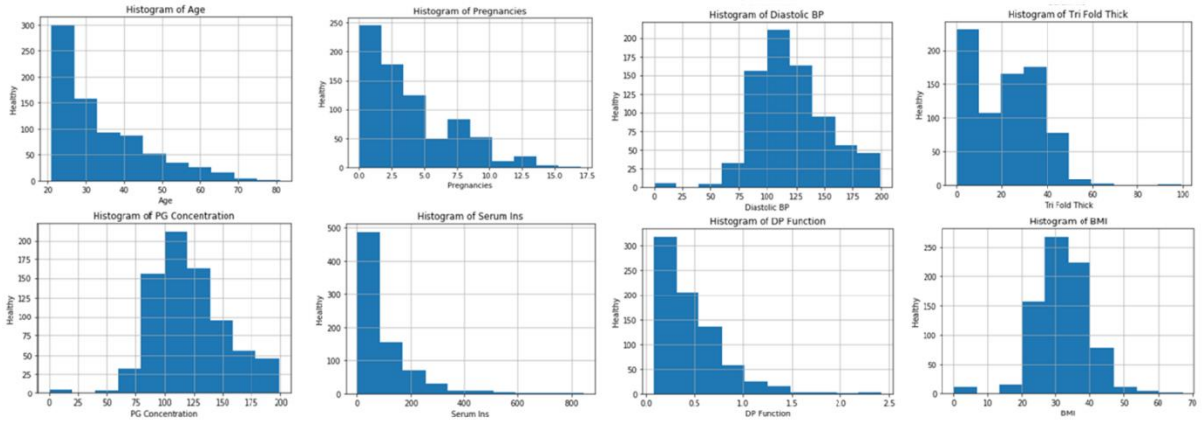

Figure 1: Distribution of parameters.

The software implementation of the data dimensionality reduction model by the multidimensional scaling method was carried out in the Python programming language in the Anaconda programming environment.

Table 2 shows the import of the data.

Table 2

Input data

#	Pregnancies	PG Concentration	Diastolic BP	…	DP Function	Age	Diabetes
0	6	148	72	…	0.627	50	Sick
1	1	85	66	…	0.351	31	Healthy
2	8	183	64	…	0.672	32	Sick
3	1	89	66	…	0.167	21	Healthy
4	0	137	40	…	2.288	33	Sick
…	…	…	…	…	…	…	…
763	10	101	76	…	0.171	63	Healthy
764	2	122	70	…	0.340	27	Healthy
765	5	121	72	…	0.245	30	Healthy
766	1	126	60	…	0.349	47	Sick
767	1	93	70	…	0.315	23	Healthy

After that, the console will display information about the dissimilarity matrices (distance), new data sets, stress indicators for the multidimensional scaling method based on two metrics, Manhattan and Euclidean. The dissimilarity matrices are shown in tables below. Table 3 shows Manhattan MDS, Table 4 shows Euclidean MDS.

Table 3

Manhattan MDS

[[0	312	199	…	432	249	299]
[312	0	335	…	206	119	83]
[199	335	0	…	441	320	402]
…	…	…	…	…	…	…
[432	206	441	…	0	239	213]
[249	119	320	…	239	0	118]
[299	83	402	…	213	118	0]]

Table 4

Euclidean MDS

[[0	178.4320599	106.465957	…	256.4488253	161.78689687	192.82893974]
[178.4320599	0	201.86876925	…	113.91224693	59.4726828	46.4865572]
[106.465957	201.86876925	0	…	271.97977866	194.12882321	230.32151441]
…	…	…	…	…	…	…
[256.4488253	113.91224693	271.97977866	…	0	116.02154972	102.32790431]
[161.78689687	59.4726828	194.12882321	…	116.02154972	0	54.55272679
[192.82893974	46.4865572	230.32121441	…	102.32790431	54.55272679	0]

Figure 2 shows a visual representation of the Manhattan distance dissimilarity matrix. Figure 3 shows a visual representation of the Euclidean distance dissimilarity matrix. On graphical representations, you can see that each is symmetrical and contains zero values on the diagonals.

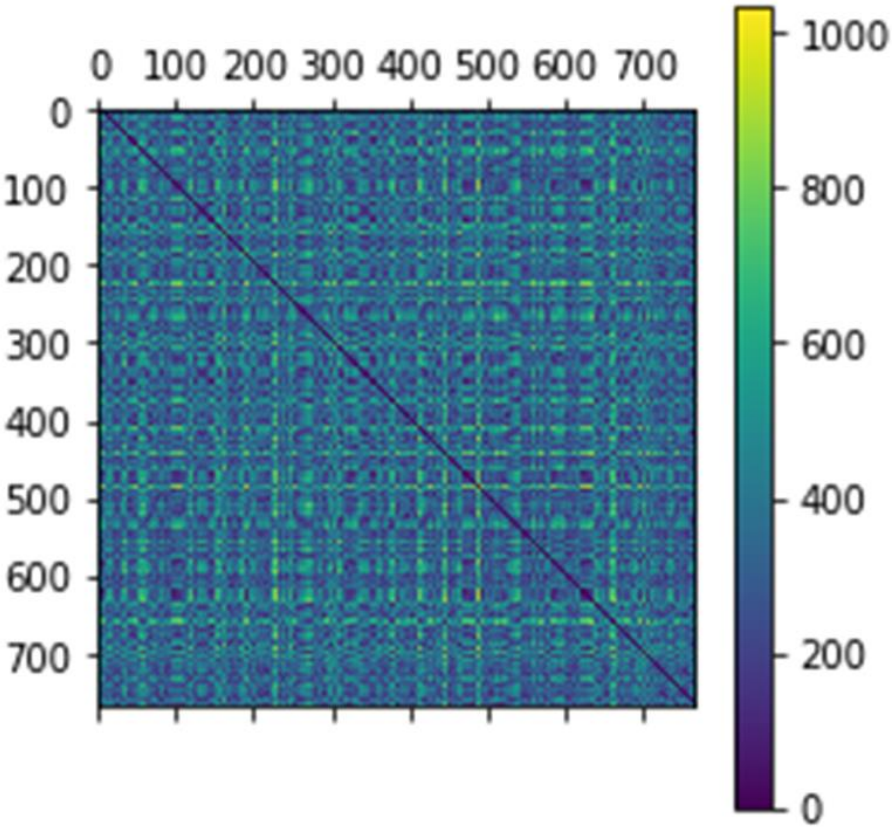

Figure 2: Visualization of Manhattan distance.

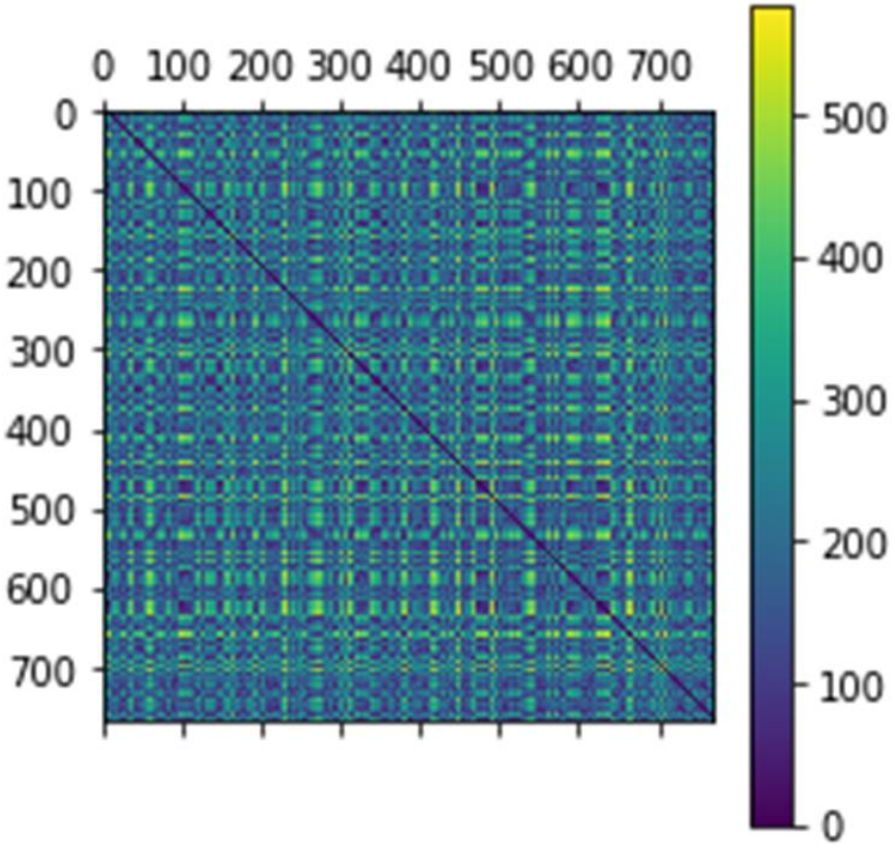

Figure 3: Visualization of Euclidean distance.

Table 5 shows new dataset according to Manhattan MDS. Table 6 shows new dataset according to Euclidean MDS. Stress indicator value of Manhattan distance is 0.17852952329291213. Stress indicator value of Euclidean distance is 0.11104963752850103.

Table 5

New database (Manhattan MDS)

[[140.03126997	111.65116183]
[36.81845208	-142.76590475]
[305.34294034	-75.87439157]
...	...
[-98.88342103	-157.0404145]
[-2.90571001	-94.96936076]
[-32.55505262	-106.65190576]]

As we can see, for multidimensional scaling based on the Manhattan distance, the stress factor is 0.17, which is sufficient reason to doubt the results' reliability. Understandably, the number of features set for new data is not optimal for data dimensionality reduction. It is better to set the data dimension to more than two to avoid such a situation for a given set.

Table 6

New database (Euclidean MDS)

[[-108.25344951	-59.01160171]
[61.24171203	-47.18164425]
[-80.4182441	-160.9898698]
...	...
[126.87588531	7.0477588]
[49.04460745	-15.20916602]
[72.37191565	-4.4528538]]

In turn, the stress factor for multidimensional scaling using Euclidean distance is 0.11, which is also not ideal, but acceptable to rely on the results obtained, but do not forget that the data is still built with possible errors.

The new data sets contain information about 768 patients, but not with 20 features, as initially, but with only two. This is due to the specified data dimension. These received features include a geometric justification. Each data pair represents x, y coordinates. These coordinates will be used to visually represent the result of data dimensionality reduction.

It is important to note that the axes in the resulting plots alone do not make sense and that the figures' orientations are arbitrary.

Figure 7 shows a visual representation of the results, the graph called MDS (Manhattan distances) is a reflection of multidimensional scaling using Manhattan distance, and called MDS (Euclidean distances) is a multidimensional scaling method using Euclidean distance. In the resulting graphs, each point corresponds to a patient, which means that the graph shows information about 768 patients, but this information only shows the dissimilarity between patients. This can be explained as follows: if two points are near, this means that they have similar indicators, but if two points are far apart, this means that these input features presented at the beginning in these patients are very different.

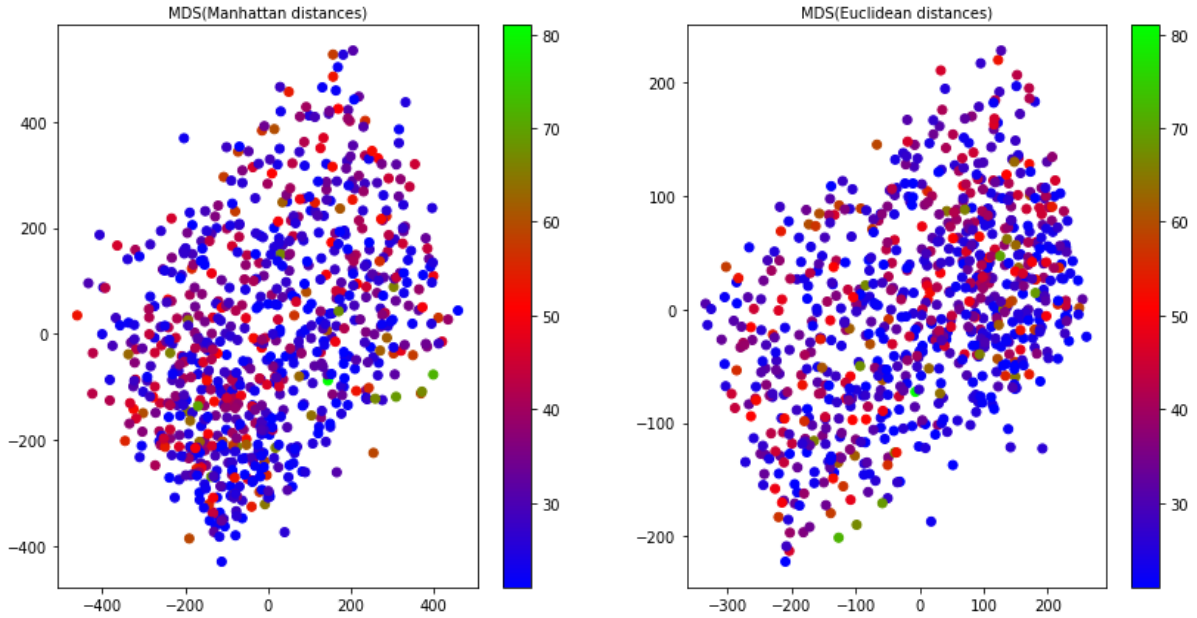

Figure 7: Visualization of new data samples.

4. Conclusions

The task of dimensionality reduction is relevant for the application of mathematical modeling methods and information technologies to support doctors' decision making when making diagnoses in conditions of limited resources.

Within the framework of this study, a model for reducing the dimensionality of medical data was built based on the multidimensional scaling method. An information system for automated data processing has been developed in the Python language.

Diabetes Mellitus Type II was chosen as the object of study, the containment of which is especially relevant in the context of the escalation of the Russian war in Ukraine.

As a result of the study, the Pima Indians Diabetes dataset was processed, consisting of 768 records and 9 attributes. After processing, the new dataset consists of 2 attributes. Manhattan distance is 0.17, Euclidean distance is 0.11.

5. Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”.

6. References

[1] W. Kerner, J. Bruckel, Definition, classification and diagnosis of diabetes mellitus, Experimental and Clinical Endocrinology & Diabetes 122 (7) (2014): 384-6. doi: 10.1055/s-0034-1366278

[2] D. Glovaci, W. Fan, N.D. Wong, Epidemiology of Diabetes Mellitus and Cardiovascular Disease, Current Cardiology Reports 21 (4) (2019): 21. doi: 10.1007/s11886-019-1107-y

[3] L. Cloete, Diabetes mellitus: an overview of the types, symptoms, complications and management, Nursing Standard 37 (1) (2022): 61-66. doi: 10.7748/ns.2021.e11709

[4] F. Zaccardi, D.R. Webb, T. Yates, M.J. Davies, Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective, Postgraduate Medical Journal 92 (1084) (2016): 63-9. doi: 10.1136/postgradmedj-2015-133281
[5] D. Lovic, et. al., The growing epidemic of diabetes mellitus, Current vascular pharmacology 18 (2) (2020): 104-109. doi: 10.2174/1570161117666190405165911
[6] M.S. Paulo, N.M. Abdo, R. Bettencourt-Silva, R.H. Al-Rifai, Gestational diabetes mellitus in Europe: a systematic review and meta-analysis of prevalence studies, Frontiers in Endocrinology 12 (2021): 691033. doi: 10.3389/fendo.2021.691033
[7] R.M. Stuart, et. al., Diabetes care cascade in Ukraine: an analysis of breakpoints and opportunities for improved diabetes outcomes, BMC Health Services Research 20 (1) (2020): 409. doi: 10.1186/s12913-020-05261-y
[8] D. Chumachenko, On intelligent multiagent approach to viral Hepatitis B epidemic processes simulation, Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018 (2018): 415-419. doi: 10.1109/DSMP.2018.8478602
[9] D. Chumachenko, et. al., Development of an intelligent agent-based model of the epidemic process of syphilis, International Scientific and Technical Conference on Computer Sciences and Information Technologies (2019): 42-45. doi: 10.1109/STC-CSIT.2019.8929749
[10] I. Izonin, R. Tkachenko, N. Shakhovska, N. Lotoshynska, The additive input-doubling method based on the SVR with Nonlinear Kernels: small data approach, Symmetry 13 (4) (2021): 4. doi: 10.3390/sym13040612
[11] R. Radutniy, et. al., Automated measurement of bone thickness on SCT sections and other images, Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing (2020): 222-226. doi: 10.1109/DSMP47368.2020.9204289
[12] N. Davidich, et. al., Monitoring of urban freight flows distribution considering the human factor, Sustainable Cities and Society 75 (2021): 103168. doi: 10.1016/j.scs.2021.103168
[13] D. Chumachenko, K. Chumachenko, S. Yakovlev, Intelligent simulation of network worm propagation using the code red as an example, Telecommunications and Radio Engineering 78 (5) (2019): 443-464. doi: 10.1615/TELECOMRADENG.V78.I5.60
[14] O. Skitsan, I. Meniailov, K. Bazilevych, H. Padalko, Evaluation of the informative features of cardiac studies diagnostic data using the Kullback method, CEUR Workshop Proceedings 2917 (2021): 186-195.
[15] S. Yakovlev, et. al., A. The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020): 265-274.
[16] R. Xiang, et. al., A comparison for dimensionality reduction methods of single-cell RNA-seq data, Frontiers in Genetics 12 (2021): 646936. doi: 10.3389/fgene.2021.646936
[17] T. Isomura, T. Toyoizumi, Dimensionality reduction to maximize prediction generalization capability, Nature Machine Intelligence 3 (2021): 434-446. doi: 10.1038/s42256-021-00306-1
[18] M.A.A. Cox, T.F. Cox, Multidimensional Scaling, Handbook of Data Visualization (2008): 315-317. doi: 10.1007/978-3-540-33037-0_14
[19] J. Tzeng, H.H.S. Lu, W.H. Li, Multidimensional scaling for large genomic data sets, BMC Bioinformatics 9 (2008): 179. doi: 10.1186/1471-2105-9-179
[20] C. Becavin, et. al., Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition, Bioinformatics 27 (10) (2011): 1413-1421. doi: 10.1093/bioinformatics/btr143
[21] I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, Euclidean distance matrices: essential theory, algorithms, and applications, IEEE Signal Processing Magazine 32 (6) (2015): 12-30. doi: 10.1109/MSP.2015.2398954
[22] R. Shahid, S. Bertazzon, M.L. Knudtson, W.A. Ghali, Comparison of distance measures in spatial analytical modeling for health service planning, BMC Health Services Research 9 (2009): 200. doi: 10.1186/1472-6963-9-200
[23] J.W. Smith, et. al., Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, Proceedings of the symposium on computer applications and medical care (1988): 261-265.