INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 10(1):8-21, February, 2015.

Method for Visual Detection of Similarities in Medical Streaming Data

J. Bernatavičienė, G. Dzemyda, G. Bazilevičius,
V. Medvedev, V. Marcinkevičius, P. Treigys

Jolita Bernatavičienė*, Gintautas Dzemyda,
Gediminas Bazilevičius, Viktor Medvedev,
Virginijus Marcinkevičius, and Povilas Treigys
Vilnius University
Institute of Mathematics and Informatics
Lithuania, LT-08663 Vilnius, Akademijos, 4
Jolita.Bernataviciene@mii.vu.lt, Gintautas.Dzemyda@mii.vu.lt,
Gediminas.Bazilevicius@mii.vu.lt, Viktor.Medvedev@mii.vu.lt,
Virginijus.Marcinkevicius@mii.vu.lt, Povilas.Treigys@mii.vu.lt
*Corresponding author: Jolita.Bernataviciene@mii.vu.lt

Abstract: The analysis of medical streaming data is quite difficult when the problem is to estimate health-state situations in real time streaming data in accordance with the previously detected and estimated streaming data of various patients. This paper deals with the multivariate time series analysis seeking to compare the current situation (sample) with that in chronologically collected historical data and to find the subsequences of the multivariate time series most similar to the sample. A visual method for finding the best subsequences matching to the sample is proposed. Using this method, an investigator can consider the results of comparison of the sample and some subsequence of the series from the standpoint of several measures that may be supplementary to one another or may be contradictory among themselves. The advantage of the visual analysis of the data, presented on the plane, is that we can see not only the subsequence best matching to the sample (such a subsequence can be found in an automatic way), but also we can see the distribution of subsequences that are similar to the sample in accordance with different similarity measures. It allows us to evaluate differences among the subsequences and among the measures.

Keywords: Streaming data, similarity measures, multivariate time series, visualization, multidimensional scaling.

1 Introduction

Time series data are widely available in different fields including medicine, finance, and science. A time series is a collection of chronologically performed observations of the values of a feature that characterizes the behaviour of a particular object. There are many topics in time series data mining, i.e., similarity search, clustering, classification, anomaly detection, motif discovery, etc. The similarity problem can be defined as a comparison of two time series to determine whether they are similar or not. Usually, the choice of a similarity measure can affect the result of data mining tasks. By a similarity measure we mean a method, which compares two time series and returns the value of their similarity. If the object is characterized by several features, we have a multivariate time series (MTS) [1].

In this paper, we investigate the similarity search in multivariate physiological time series. A physiological time series is a series of some medical observations over a period of time. Such a type of data can be collected using devices (or sensors) that collect personal medical features, such as heart rate, blood pressure, etc. An example of such data can be the intensive care multivariate online-monitoring time series [2]. A sensor is an instrument that detects or measures a physical or environmental characteristic or state, transmits and/or records the reading in some form (e.g.,

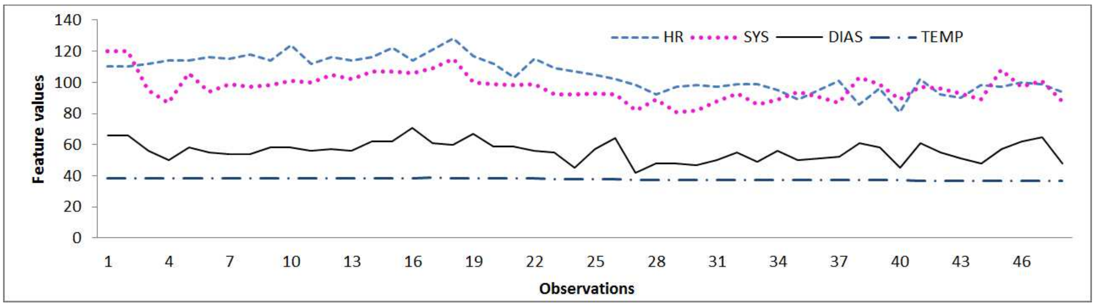


Figure 1: Example of the multivariate time series

a visual display, audio signal, digital transmission, etc.). A sensor converts the physical quantity to electric output. For example, a pressure sensor converts pressure to electric output. Remote monitoring of health parameters such as the pulse rate, oxygen level in blood or blood pressure can be very helpful for early detection of diseases, resulting in reduction of treatment time.

Most methods, used to analyse medical data, focus primarily on analysing the univariate time series. However, because of parameter dependences and variation over time, examination of all medical data together in a multivariate time series can provide more information about the data and patient, to make a better diagnosis and treat the patient [3].

Therefore, this paper deals with the multivariate time series analysis with a view to compare the current situation with that of in chronologically collected historical data, and to find subsequences of the multivariate time series most similar to the sample, corresponding, e.g. to the current situation. An example of MTS of four features (heart rate HR, non-invasive systolic arterial blood pressure SYS, non-invasive diastolic arterial blood pressure DIAS, temperature TEMP) is presented in Figure 1.

Let us have a multivariate time series of n features and T_a observations:

$$X^a = \begin{pmatrix} x^a_{11} & \cdots & x^a_{1T_a} \\ \vdots & \ddots & \vdots \\ x^a_{n1} & \cdots & x^a_{nT_a} \end{pmatrix}.$$

Denote the sample of n features and T_b observations as

$$X^b = \begin{pmatrix} x^b_{11} & \cdots & x^b_{1T_b} \\ \vdots & \ddots & \vdots \\ x^b_{n1} & \cdots & x^b_{nT_b} \end{pmatrix}.$$

Here $T_a > T_b$. In fact, X^a and X^b are matrices.

As a result, we need to find the optimal place of X^b on X^a. The place is defined by some time moment $T_* : 1 \leq T_* \leq T_a - T_b + 1$. Our procedure analyses the multivariate time series X^a by using the moving time window the width of which is adapted to the current situation X^b (width is equal to T_b) and comparing the content of this window with the sample, in the sense of several similarity measures, at the same time.

Visual method of finding the best subsequences matching to the sample is proposed in this paper. As it is indicated in [4], the goal of visual analytics research is to turn the information overload into an opportunity, i.e. decision-makers should be enabled to examine massive, multidimensional, multisource, time varying information stream to make effective decisions in time critical situations.

There are attempts to apply visual analysis for the streaming data. The example is the Visual Content Analysis of Real-Time Data Streams project [5] at the Pacific Northwest National Laboratory. Its goal was to allow users to quickly grasp dynamic data in forms that are intuitive and natural without requiring intensive training in the use of specific visualization or analysis tools and methods. The project has prototyped five different visualization prototypes that represent and convey dynamic data through human-recognizable contexts and paradigms such as hierarchies, relationships, time and geography.

In this paper, we suggest using the specific visualization tools and methods (multidimensional data visualization [6]) that are effective and also do not require intensive training of the users. Moreover, we show a possibility of making the decision, based on five criteria of similarity of the sample with the subsequence of real-time data stream by representing the similarity as a point on a plane. Dimensionality reduction and visual analysis of multidimensional data [6] have been applied when comparing the best found subsequences in X^a.

The proposed method is described in Section 2. Similarity measures for multivariate time series and comparative analysis of the measures are presented in Section 3. Multidimensional data visualization is reviewed in Section 4, where the emphasis is put on the multidimensional scaling. Comparative analysis of similarity measures for multivariate time series is presented in Section 5. An example, illustrating the proposed method, is presented in Section 6.

2 Visual Method for Finding the Best Subsequences Matching to the Sample

In our method, the multivariate time series X^a are analysed by using the moving time window. The width of this window is adapted to the current situation (sample) X^b and is equal to T_b. The content of this window is compared with the sample, in the sense of several similarity measures at the same time. It includes the dimensionality reduction procedure that allows us to observe multidimensional data visually.

The visual method for finding the best subsequences, matching to the sample, can be generalized as follows:

1. Let us have: - a multivariate time series X^a of n features and T_a observations; - sample X^b of n features and T_b observations; - m similarity measures $S_i, i = 1, \dots, m$.

2. The sample X^b is compared with all subsequences of X^a by using m similarity measures $S_i, i = 1, \dots, m$. The subsequences are obtained by moving the time window in the X^a from beginning to end. The content of such a window is a matrix of n rows. Denote it by X^c. The width of the window (the number of columns of X^c) is adapted to the current situation (sample) X^b (its width is equal to T_b). For each measure, k subsequences are chosen most similar to the sample. Therefore, the total number of subsequences for a further analysis is equal to km.

3. Each comparison of the sample with a subsequence, chosen in the way defined in the step above, produces a m-dimensional point $\overline{S}_q = (S_{q1}, S_{q2}, \dots, S_{qm})$, where, in our case, $q = 1, \dots, km$. Let us derive two additional points: - $\overline{S}_0 = (S_{01}, S_{02}, \dots, S_{0m})$ is the array of values of all similarity measures, computed for the subsequence, that is ideally coincident with the sample (the array of the best values of m similarity measures); - $\overline{S}_C = (S_{C1}, S_{C2}, \dots, S_{Cm})$ is the weight center of $\overline{S}_q = (S_{q1}, S_{q2}, \dots, S_{qm}), q = 1, \dots, km$. Therefore, the total number of m-dimensional points for discovering the most similar subsequences to the sample is equal to $km+2$. Afterwards, the normalization of the components

of these points is performed by z-score. Denote the obtained matrix of normalized points by Z. It consists of $km+2$ rows and m columns.

4. The points from matrix Z are mapped on the plane using the Multidimensional Scaling [6] (or are other algorithm of nonlinear projection of multidimensional points on the plane). Denote the resulting matrix by Y, that contains $km+2$ rows, corresponding to different comparisons of the sample with other subsequences, and 2 columns. Each row is coordinates of the point on the plane.

5. The investigator analyses the information presented graphically, where all m-dimensional points are represented as the points on a plane, and makes decisions. In general, the most similar subsequence to the sample can be the subsequence, the corresponding point of which on a plane is closest to the projection of $\overline{S}_0$ on the plane. However, more subsequences may be considered as similar to the sample. The points on the plane, corresponding to such subsequences, must be closer to the projection of $\overline{S}_0$ on a plane than to the projection of $\overline{S}_C$. These rules can be checked automatically in the program realization of this method, however participation of the investigator is valuable, because it gives a possibility to him to cognize the data deeper.

The advantage of this method is that the investigator can consider the results of comparison from the standpoint of several measures that may be supplementary to one another or contradictory among themselves. Therefore, the similarity of subsequences with the sample will be evaluated from different standpoints. The method is universal, because different sets of similarity measures can be chosen, depending on the problem, but the scheme of decision remains the same. Moreover, the involvement of the dimensionality reduction and visual analysis of multidimensional data in the proposed method renders the opportunity to the investigator to participate in the final decision, when comparing the best found subsequences of the multivariate time series with the sample. However, the decision on their similarity can also be made automatically.

3 Similarity Measures for Multivariate Time Series

To detect events in real multivariate time series, it is necessary to compare time series using the appropriate similarity measure [7]. Different techniques and similarity measures are introduced and used for comparison of multivariate time series of different nature [8], [9]. Multivariate time series can be reduced to univariate time series and their similarity can be measured, using a univariate time series approach [10]. That may lead to a great loss of information, therefore, we concentrate on the multivariate time series approach here. Five similarity measures $S_i, i = 1, \ldots, 5$, used in this paper for multivariate time series, are presented below.

Let us compare two multivariate time series:

$$X^a = \begin{pmatrix} x^a_{11} & \cdots & x^a_{1T_a} \\ \vdots & \ddots & \vdots \\ x^a_{n1} & \cdots & x^a_{nT_a} \end{pmatrix} \text{ and } X^b = \begin{pmatrix} x^b_{11} & \cdots & x^b_{1T_b} \\ \vdots & \ddots & \vdots \\ x^b_{n1} & \cdots & x^b_{nT_b} \end{pmatrix}.$$

The Frobenius norm is often used in the matrix analysis [11]. This similarity measure is based on the Euclidean distance. The Frobenius norm of a matrix X^b is defined by the formula:

$$\left\| X^b \right\|_F = \sqrt{\sum_{p=1}^{n} \sum_{q=1}^{T_b} (x^b_{pq})^2} = \sqrt{tr((X^b)' X^b)}, \qquad (1)$$

where tr is the sum of elements on the diagonal of the square matrix. The Frobenius norm is used to compare the similarity of two matrices. The similarity of X^b and X^c is defined by the formula $Frob = \left\|X^b - X^c\right\|_F$. The best possible value of the Frobenius norm is 0.

The correlation coefficient between two matrices of the same size (*Matrix Correlation Coefficient*) can also be used as a similarity measure [12]:

$$r = \frac{\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^b - \bar{X}^b)(x_{pq}^c - \bar{X}^c)}{\sqrt{\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^b - \bar{X}^b)^2\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^c - \bar{X}^c)^2}}, \tag{2}$$

where $\bar{X}^b$ and $\bar{X}^c$ are the means of X^b and X^c, respectively. This measure is the Pearson correlation coefficient adapted to matrices and calculated using the MATLAB *corr2* function [12]. The best possible value of the matrix correlation coefficient is 1.The correlation coefficient between two matrices has found wide applications in the image analysis, molecular biology, etc.

The third similarity measure for multivariate time series is the *Principal Component Analysis (PCA) similarity factor* [9], [13]. PCA is a well-known and wide used technique for dimensionality reduction of data. It is a linear transformation that projects the original data to a new coordinate system with the minimal loss of information. In multivariate cases, the information is the structure of the original data, i.e. the correlation between the features and alteration of the correlation structure among them. To create a projection, PCA selects coordinate axes of the new coordinate system one by one according to the greatest variance of any projection. The PCA similarity factor is defined by the following formula:

$$S_{PCA}(X^b, X^c) = tr(L'MM'L), \tag{3}$$

where L and M are matrices that contain the first l principal components of X^b and X^c, respectively. It means that the principal components are computed by the standard algorithm using the matrices $(X^b)'X^b$ and $(X^c)'X^c$, and then l principal components with the highest eigenvalues are selected. The best possible value of the PCA similarity factor is l. In our experiments $l = 1$.

Dynamic time warping (DTW) [14] is the most widely used technique for comparison of time series data, where extensive a priori knowledge is not available. The Euclidean distance reflects the similarity in time, while the dynamic time warping (DTW) reflects the similarity in shape. DTW searches for the best alignment between two time series, attempting to minimize the distance between them. The advantage of DTW is that it can handle unequal series and distortions. *Multidimensional Dynamic Time Warping (MDTW)* is presented in [15]. Some distance matrix is defined: $\{d(p,q) = \sum_{k=1}^{n}\left(x_{kp}^b - x_{kq}^c\right)^2,\ p,q = 1,\ldots,T_b\}$. Then the matrix D of cumulative distances is calculated as in the traditional DTW algorithm [15]:

$$D(p,q) = \begin{cases} d(1,1), \text{if } p = 1, q = 1, \\ d(p,q) + D(p-1,q), \text{if } p = 2,\ldots,T_b, q = 1, \\ d(p,q) + D(p,q-1), \text{if } p = 1, q = 2,\ldots.,T_b, \\ d(p,q) + min\begin{cases} D(p-1,q) \\ D(p,q-1), \\ D(p-1,q-1) \end{cases} \quad \text{in other cases.} \end{cases} \tag{4}$$

(p,q) defines the pair of the pth observation in X^b and the qth observation in X^c. Finally, the minimal path and the distance along the minimal path are obtained using matrix D. The path must start at the beginning of each time series at $(1,1)$ and finish at the end of both time series at (T_b, T_b). See [13] for details. The best possible value of MDTW is 0. On the other hand, DTW can lead us to unintuitive alignments, where a single point on one time series maps onto a large subsection of another time series [16], [17]. Also, DTW can fail to find the obvious and natural alignments in two time series because of a single feature (i.e. peak, valley, infection point, plateau, etc.). One of the causes is due to the great difference between the lengths of the compared series.

In this paper, the fifth similarity measure for multivariate time series is *Eros (Extended Frobenius norm)* [9]. Eros is based on the principal component analysis and computes the similarity between two MTS items by measuring how close the corresponding principal components are using the eigenvalues as weights. In our case, X^b and X^c are two multivariate time series items of n features and T_b observations. $V^b = [v_1^b, \dots, v_n^b]$ and $V^c = [v_1^c, \dots, v_n^c]$ are two right eigenvector matrices obtained by applying a singular value decomposition (SVD) to the covariance matrices M^b and M^c of features in X^b and X^c respectively. The *Eros similarity* of X^b and X^c is defined as follows:

$$Eros(X^b, X^c, w) = \sum_{i=1}^{n} w_i |\langle v_i^b, v_i^c \rangle|, \tag{5}$$

where $\langle v_i^b, v_i^c \rangle$ is the inner product of v_i^b, and v_i^c, w is a weight vector, based on eigenvalues of the MTS data set (see more in detail in [9]), $\sum_{i=1}^{n} w_i = 1$.

4 Multidimensional Data Visualization

The method, proposed in Section 2 for finding the best subsequences matching to the sample, is based on the visual presentation and analysis of multidimensional points the coordinates of which are the values of similarity measures, computed for a pair of subsequences. The visualization technology is introduced below.

For an effective data analysis, it is important to include a human into the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and computational power of today's computer. Visual data mining aims at integrating the human in the data analysis process, applying the human's perceptual abilities to the analysis of large data sets, available in today's computer systems. Visualization finds a wide application in the medical data analysis, too [18], [19].

The goal of the projection method is to represent the input data items in a lower-dimensional space so that certain properties of the structure of the data set were preserved as faithfully as possible. The projection can be used to visualize a data set, if rather a small output dimensionality is chosen. One of these methods is the principal component analysis (PCA). The well-known principal component analysis [6] can be used to display the data as a linear projection on a subspace of the original data space such that best preserves the variance in the data. PCA cannot embrace nonlinear structures, consisting of arbitrarily shaped clusters or curved manifolds, since it describes the data in terms of a linear subspace. Therefore, several methods have been proposed for reproducing nonlinear higher-dimensional structures on a lower-dimensional display: multidimensional scaling and its modifications [6], [20], [21], [22], Isomap [23], locally linear embedding [24], etc. Various neural network approaches are used for this aims as well (see e.g. [25], [26], [27]).

Multidimensional scaling (MDS) is a group of methods that project multidimensional data to a low (usually two) dimensional space and preserve the interpoint distances among data as much, as possible. Let us have the m-dimensional points $\overline{S}_q = (S_{q1}, S_{q2}, \dots, S_{qm}), q = 1, \dots, t, (\overline{S}_q \in$

R^m). The pending problem is to get the projection of these points onto the plane R^2. Two-dimensional points $Y_1, Y_2, \dots, Y_t \in R^2$ correspond to them. Here $Y_q = (y_{q1}, y_{q2}), q = 1, \dots, t$. Denote the distance between the points $\overline{S}_q$ and $\overline{S}_p$ by d^*_{qp}, and the distance between the corresponding points Y_q and Y_p on the projected space by d_{qp}. In our case, the initial dimensionality is m, and the resulting one is 2 (2-D). Naturally, 1-D and 3-D projections could be considered, too. However, in the 1-D case, we lose knowledge that can be obtained from 2-D or 3-D views. Advantages of 3-D can be achieved when special means to present such data on the screen are applied. Therefore, 2-D projections of the multidimensional data are commonly used.

There exists a multitude of variants of MDS with slightly different so-called stress functions. In our experiments, the raw stress is minimized $E_{MDS} = \sum_{q<p}^{t} \left(d^*_{qp} - d_{qp}\right)^2$, seeking to find the optimal coordinates of points $Y_1, Y_2, \dots, Y_t$.

5 Comparative Analysis of Similarity Measures for Multivariate Time Series

The data from PhysioNet/Computing in Cardiology Challenge are used for the experimental analysis (*http://www.physionet.org/challenge/2012/*). The records were collected in the Intensive Care Unit. In the experiments we used a set X^a, containing multivariate time series of 50 patients of the same age, i.e. if to follow the notation of Section 1, we have 50 different multivariate time series $X^a_i, i = 1, \dots, 50$, each consisting of 48 observations (columns) of $n = 4$ features (rows: non-invasive diastolic arterial blood pressure, non-invasive systolic arterial blood pressure, heart rate, temperature). $X^a = \{X^a_i, i = 1, \dots, 50\}$.

In general, the sample X^b (the current situation) does not belong to X^a. However, seeking for more precise conclusions in this research, we have chosen X^b in X^a at random of the length $T_b = 8$. Moreover, we have chosen X^b such that its contents belongs to the same patient, i.e. X^b does not consist of the parts of different patients' records.

The goal is to go through time windows X^c (of the size of X^b) in X^a and compare them with X^b. For each similarity measure the optimal place of X^b on X^a has been found. By the optimal place of X^b, in accordance with the chosen similarity measure, we assume X^c such that produces the best value of this similarity measure when comparing X^b and X^c. Then the values of remaining measures were computed for the same X^c. Due to the specific character of data (50 patients), the place of X^c on X^a may be denoted as follows: $i[T_{start}; T_{end}]$, where i is the order number of a patient, $i = 1, \dots, 50, T_{start}$ and T_{end} are start and end positions of X^c on X^a. The illustration results are presented in Table 1.

Let us choose the sample at random. In our case, $X^b = 1[23 - 30]$. It was compared with all the other available subsequences X^c of the analysed data set X^a. Considering that the sample taken for the analysis is selected from X^a and striving for the independence of investigation results on this case, we choose X^c only such that has no more than 5 common observations with X^b.

Five subsequences $X^c \neq X^b$, that are most similar to the sample, were selected according to each similarity measure. Therefore, we get five collections of five most similar subsequences. The total number of the selected subsequences is 25 and some of them coincident according to different measures. For each collection, the values of all similarity measures are computed and presented in Table 1. The best values obtained for all similarity measures are marked in bold.

Table 1 shows that different measures often mark different (not the same) subsequences as similar. All the measures acquire their best values in the case of identical coincidence of X^b and X^c. Thus, a further task is to summarize the obtained results and to develop a method for selecting of the most similar subsequence X^c to the sample X^b. The analysis of subsequences, found by

Table 1: Most similar subsequences according to each similarity measure; sample $X^b = 1[23-30]$

	No.	X^c	Similarity measures				
			r	*Eros*	*Frob*	*MDTW*	S_{PCA}
Ideal coincidence	0	1[23-30]	1	1	0	0	1
r	1	1[16-23]	**0.9639**	0.1745	4.4224	**16.9434**	0.7175
	2	48[37-44]	0.9575	0.1302	**2.5551**	19.2153	0.6874
	3	1[17-24]	0.9504	**0.6067**	3.9967	17.0556	**0.8480**
	4	42[16-23]	0.9455	0.3440	3.8527	30.8482	0.1680
	5	1[19-26]	0.9430	0.2969	3.1215	18.8935	0.8474
Eros	6	10[30-37]	0.0013	**0.9423**	6.5985	59.1889	**0.9847**
	7	25[31-38]	-0.3214	0.9099	7.1453	62.7647	0.6705
	8	49[35-42]	-0.2209	0.9064	8.0460	41.6136	0.5461
	9	17[9-16]	**0.6266**	0.8979	**4.6676**	**39.8056**	0.4329
	10	40[16-23]	0.1164	0.8948	7.7252	55.2800	0.5772
Frob	11	48[37-44]	**0.9575**	0.1302	**2.5551**	19.2153	0.6874
	12	48[36-43]	0.9366	**0.4445**	2.8003	**18.6496**	**0.9789**
	13	1[36-43]	0.8931	-0.2531	2.9396	31.2548	0.9206
	14	48[35-42]	0.9169	-0.2479	3.0346	23.2913	0.9148
	15	48[38-45]	0.8943	0.1853	3.0886	25.6391	0.5152
MDTW	16	34[39-46]	-0.0158	0.4709	7.2486	**8.8982**	0.3182
	17	22[14-21]	0.1724	0.6900	6.5888	9.6141	0.5765
	18	29[4-11]	0.5061	-0.0177	4.9202	9.6744	0.6974
	19	12[12-19]	**0.6511**	-0.3166	**4.8690**	10.1095	**0.9553**
	20	15[3-10]	0.5351	0.2744	7.4371	11.0668	0.2524
S_{PCA}	21	10[32-39]	-0.1145	0.0688	7.5452	96.4585	**0.9920**
	22	16[33-40]	-0.0749	**0.2350**	9.2904	67.8260	0.9896
	23	10[33-40]	-0.0695	0.0074	7.6335	93.8404	0.9896
	24	10[34-41]	0.0036	-0.4202	7.5849	70.1514	0.9892
	25	4[5-12]	**0.4171**	0.1522	**6.9828**	**57.2766**	0.9890

different measures shows that matrix correlation coefficient and Frobenius norm measures try to find the most similar subsequence according to the values of the subsequence elements. MDTW tends to compare the data change dynamics: the scales of features of the subsequences can be different, but MDTW can indicate these sequences as similar. S_{PCA} and *Eros* measures do not also depend on the scale and are less sensitive to abrupt signal changes. It is very important in the medical data analysis.

The correlation analysis (see Table 2) has the depicted a strong inverse correlation between the Frobenius norm and the matrix correlation coefficient. For the investigation, 50 randomly selected samples $X_i^b, i = 1, \ldots, 50$ consisting of $T_b = 8$ successive observations were selected; from each X_i^a one subsequece X_i^b was selected as a sample. For each selected sample and for each similarity measure 25 most similar subsequences were identified in the whole X^a. Just like above, we choose X^c only such that has no more than 5 common observations with X^b. The total number of comparisons of X^b and X^c is 1250. According to these data, the correlation matrix of similarity measures has been computed. It is presented in Table 2. As the results shows, there is a very strong inverse correlation (-0.7355) between the Frobenius norm *Frob* and the matrix correlation coefficient *r*.

Because of the strong inverse correlation between the Frobenius norm and the matrix corre-

Table 2: Correlation matrix of similarity measures

	r	$Eros$	$Frob$	$MDTW$	S_{PCA}
r	1.0000	-0.1917	**-0.7355**	-0.2127	-0.1652
$Eros$	-0.1917	1.0000	0.1819	0.1731	-0.1748
$Frob$	-0.7355	0.1819	1.0000	0.1071	0.1055
$MDTW$	-0.2127	0.1731	0.1071	1.0000	0.0454
S_{PCA}	-0.1652	-0.1748	0.1055	0.0454	1.0000

lation coefficient, the Frobenius norm was casted away. It has been done with a view to reduce the general impact of the correlated parameters on the investigation results. In such a way, outcasting of the Frobenius norm measure evens the impact of the rest measures.

For the second investigation 50 randomly selected samples $X_i^b, i = 1, \ldots, 50$ consisting of $T_b = 8$ successive observations have been selected. One sample X_i^b is selected from each X_i^a. For each selected sample and for each similarity measure (4 similarity measures are investigated now), 20 most similar subsequences were identified in the whole X^a. Just like previously, we choose X^c only such that has no more than 5 common observations with X^b. The total number of comparisons of X^b and X^c is 1000. According to the obtained data, for each sample X_i^b and for four measures, a 4-dimensional point is constructed: $\overline{S}_q^i = (S_{q1}^i, S_{q2}^i, S_{q3}^i, S_{q4}^i), q = 1, \ldots, 20$. Coordinates of the point are the values of different measures. For example, the ideal subsequence match is assumed as $\overline{S}_0 = (1, 1, 0, 1)$. Further, the Euclidean distance is calculated between the ideal match $\overline{S}_0$ and all the rest $\overline{S}_q^i$ of similar subsequences. Then the subsequences are sorted according to the shortest distance and the first 5 subsequences are treated as similar (from the total 5x50 subsequences). Table 3 summarizes the investigation results, i.e. shows the most often found similar subsequences.

Table 3: Most often found similar subsequences according to each similarity measure

Frequencies	r	$Eros$	$MDTW$	S_{PCA}
Total 250	93	32	91	34
Percentage from 250 subsequences	31	11	30	11

As can be seen from Table 3, the matrix correlation coefficient and MDTW measures are the best.

In addition, the Frobenius norm measure can be considered together with these two measures because of its high correlation with the matrix correlation coefficient measure. Moreover, a high correlation of the Frobenius norm with the matrix correlation coefficient does not mean that they yield very similar results. The experiment below illustrates this fact. Ten most similar subsequences were found for the chosen sample, using these two similarity measures for multivariate time series. The obtained subsequences are presented in Table 4 in decreasing order of their goodness, depending on the similarity measure. Coincident subsequences are presented in different tint of grey colour. We see three coincident subsequences only with quite different order numbers.

Table 4: Comparison of the Frobenius norm with the matrix correlation coefficient

	No.	r	*Frob*
Sample 1[23-30]	1	1[16-23]	48[37-44]
	2	48[37-44]	48[36-43]
	3	1[17-24]	1[36-43]
	4	42[16-23]	48[35-42]
	5	1[19-26]	48[38-45]
	6	48[36-43]	1[35-42]
	7	1[18-25]	20[15-22]
	8	42[35-42]	1[19-26]
	9	42[17-24]	1[34-41]
	10	1[7-14]	48[40-47]

6 Experimental Illustration of the Visual Method for Finding the Best Subsequences, Matching to the Sample

The performance of the proposed method is illustrated by the example. Like in Section 5, the data from PhysioNet/Computing in Cardiology Challenge (http://www.physionet.org/challenge) are used for the experimental analysis.

1. Sample $X^b = 1[23-30]$ has been chosen.

2. The sample is compared with all the subsequences of X^a by using 5 similarity measures, given in Section 3. In accordance with each measure, 10 subsequences, most similar to the sample, are chosen. Therefore, the total number of subsequences for a further analysis is equal to 50.

3. Each comparison of a sample to a subsequence, chosen in the way defined in the step above, produces the 5-dimensional point $\overline{S}_q = (S_{q1}, S_{q2}, \dots, S_{q5})$, where $q = 1, \dots, 50$. Let us derive two additional points:

   - $\overline{S}_0 = (S_{01}, S_{02}, \dots, S_{05})$ is the array of values of all the similarity measures computed for the subsequence that is ideally coincident with the sample (the array of the best values of m similarity measures); in our case, $\overline{S}_0 = (1, 1, 0, 0, 1)$;
   - $\overline{S}_C = (S_{C1}, S_{C2}, \dots, S_{Cm})$ is the weight center of $\overline{S}_q = (S_{q1}, S_{q2}, \dots, S_{q5}), q = 1, \dots, 50$; in our case, $\overline{S}_C = (0.39298, 0.31968, 5.98348, 35.26159, 0.715413)$.

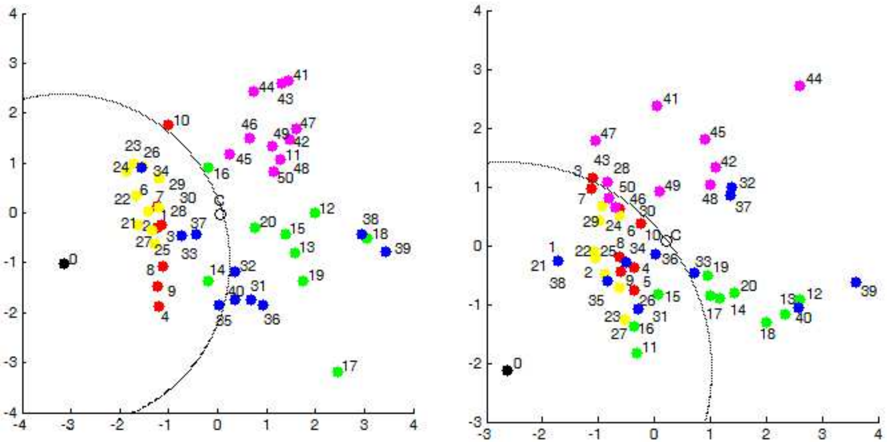
   Therefore, the total number of 5-dimensional points for discovering the most similar subsequences to the sample is equal to 52. Afterwards, the normalization of the components of these points is performed by z-score. The obtained matrix Z of normalized points consists of 52 rows and 5 columns.

4. The points from matrix Z are mapped on the plane using the Multidimensional Scaling. The resulting matrix is Y, the rows of which correspond to different comparisons of the sample with other subsequences and are the coordinates of points on the plane.

5. Figure 2 presents graphically all 52 5-dimensional points from Y on a plane for decision making. In general, the subsequence best matching to the sample can be the subsequence the corresponding point of which on a plane is closest to the projection of $\overline{S}_0$ on a plane.

However, more subsequences are similar to the sample. The points on the plane, corresponding to such subsequences, must be closer to the projection of $\overline{S}_0$ on the plane than to the projection of $\overline{S}_C$. If we draw a circle with the center on the projection of $\overline{S}_0$ on the plane and the radius that is equal to the distance between the projections of $\overline{S}_0$ and $\overline{S}_C$, we can visually detect subsequances most similar to the sample. In Figure 2, the colour of the point corresponds to the similarity measure according to which the subsequence falls among the most similar ones:

- Red points - matrix correlation coefficient,
- Green points - Eros,
- Yellow points - Frobenius norm,
- Blue points - MDTW,
- Magenta points - S_{PCA}.



(a)

(b)

Figure 2: Results for the visual analysis: a) sample 1[23-30]; b) sample 5[15-22]

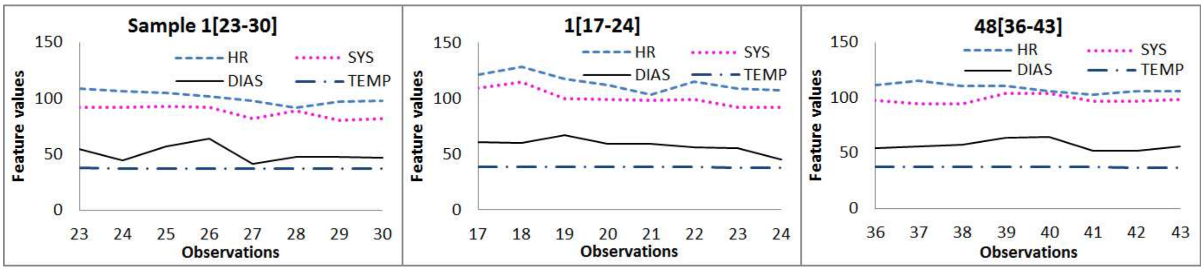


Figure 3: Two subsequences most similar to the sample 1[23-30]

In Figure 2a, two best similarity measures, in accordance with which the subsequence falls among the most similar ones, are the matrix correlation coefficient and Frobenius norm measures. In Figure 2b, these two measures are exceptional again, however MDTW and Eros have influenced the appearance of subsequences in the most similar subsequence list. A relatively small number of the most similar subsequences is produced by S_{PCA}, only. Therefore, all the measures are important and can be successfully used jointly .

Finally, the sample and two found most similar subsequences are presented, as an example, in Figure 3.

7 Conclusions

This paper deals with the multivariate time series analysis seeking to compare the current situation (sample) with that in chronologically collected historical data and to find subsequences of the multivariate time series most similar to the sample. The visual method for finding the best subsequences matching to the sample has been proposed. Using this method, the investigator can consider the results of comparison of the sample and some subsequence of the series from the standpoint of several measures that can be supplementary to one another or contradictory among themselves.

The analysis of medical streaming data is quite a difficult problem. The data are very specific to an individual patient. It may cause the problem of reliability of the decision if the problem is to estimate the health-state situations in real time streaming data in accordance with the previously detected and estimated streaming data of various patients. The usage of a larger number of similarity measures (not a single fixed measure) can increase the efficiency of decisions on the health state of the patient based on the streaming data of other patients.

The advantage of the visual analysis of the data, presented on the plane, is that we can see not only the subsequence best matching to the sample (such a subsequence can be found in the automatic way), but also we can see the distribution of subsequences that are similar to the sample in accordance with different similarity measures. It allows us to evaluate differences among the subsequences and among the measures.

Five similarity measures were integrated in this research. Note that the correlation coefficient between two matrices is quite effective and easily interpreted among other measures, that are usually used for multivariate streaming data analysis. However, the best efficiency of applications of this measure is achieved when combining it with other measures.

This method is universal and can be used in the analysis of streaming data of various nature (not only medical data). It is necessary to select the proper set of similarity measures depending on the problem solved only.

Bibliography

[1] Batal I., Sacchi L., Bellazzi R., Hauskrecht M. (2009); Multivariate Time Series Classiffcation with Temporal Abstractions, *Florida Artificial Intelligence Research Society Conference, Twenty-Second International FLAIRS Conference*, ISBN 978-1-57735-419-2, 344–349.

[2] Borowsky M., Imhof M., Schettlinger K., Gather U. (2008); Multivariate Signal Filtering from Intensive Care Online-Monitoring Time Series, avialable at `https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/MSind/SFB_475/C/2008_-_Borowski_Imhoff_Schettlinger_Gather_-_Multivariate_Signal_Filtering_from_Intensive_Care_Online_Monitoring_Time_Series_-_Biosignalverarbeitung_2008.pdf`.

[3] Ordonez P., des Jardins M., Feltes C., Lehmann C., Fackler J. (2008); Visualizing Multivariate Time Series Data to Detect Specific Medical Conditions. *Proceedings of AMIA (American Medical Informatics Association) 2008 Annual Symposium*, ISBN 978-1-61567-435-0, 6: 530–534.

[4] Keim D.A , Mansmann F., Schneidewind J., Ziegler H. (2006); Challenges in Visual Data Analysis, *Proc. Intl Conf. Information Visualization (IV)*,ISBN 0-7695-2602-0/06, 9–16.

[5] Chin G., Singhal M., Nakamura G., Gurumoorthi V., Freeman-Cadoret N. (2009); Visual analysis of dynamic data streams, *Information Visualization*, ISSN 1473-8716, 8(3): 212-229.

[6] Dzemyda G., Kurasova O., Žilinskas J. (2013); *Multidimensional Data Visualization: Methods and Applications (Springer Optimization and Its Applications, 75)*, Springer, ISBN 978-1-4419-0235-1.

[7] Ye N. (2003); *The Handbook of Data Mining*, Mahwah, NJ: Lawrence Erlbaum, ISBN 0-8058-4081-8.

[8] Karamitopoulos L., Evangelidis G., Dervos D. (2008); Multivariate Time Series Data Mining: PCA-based Measures for Similarity Search, *Proceedings of The 2008 International Conference on Data Mining*, USA, ISBN 1-60132-062-0, 253–259.

[9] Yang K., Shahabi C. (2004); A PCA-based Similarity Measure for Multivariate Time Series, *MMDB '04 Proceedings of the 2nd ACM international workshop on Multimedia databases*, ISBN 1-58113-975-6, 65–74.

[10] Xun L., Zhishu L. (2010); The Similarity of Multivariate Time Series and Its Application, *2010 Fourth International Conference on Management of e-Commerce and e-Government (ICMeCG)*, ISBN 978-0-7695-4245-4, 76–81.

[11] Moon T., Striling W. (2000); *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, ISBN 978-0201361865.

[12] *MATLAB R2014a and Image processing toolbox*, Natick, Massachusetts: The MathWorks Inc., 2014, avialable at `http://www.mathworks.se/help/images/ref/corr2.html`.

[13] Krzanowski W. (1979); Between-groups Comparison of Principal Components, *JASA*, ISSN 0162-1459, 74(367): 703–707.

[14] Berndt D., Clifford J. (1994); Using Dynamic Time Warping to Find Patterns in Time Series, *KDD Workshop*, 359–370.

[15] Sanguansat P. (2012); Multiple Multidimensional Sequence Alignment Using Generalized Dynamic Time Warping, *WSEAS Transactions on Mathematics*, e-ISSN 2224-2880, 11(8): 668–678.

[16] Fu T.-C. (2011); A Review on Time Series Data Mining, *Engineering Applications of Artificial Intelligence*, ISSN 0952-1976, 24: 164–181.

[17] Keogh E., Pazzani M. (2001); Derivative Dynamic Time Warping, *First SIAM International Conference on Data Mining (SDM2001)*, Chicago, USA, ISBN 978-0-89871-495-1, 1–11.

[18] Bernatavičienė J., Dzemyda G., Kurasova O., Marcinkevičius V., Medvedev V. (2007); The Problem of Visual Analysis of Multidimensional Medical Data, *Springer optimization and its applications 4, Models and algorithms for global optimization*, New York, Springer, ISBN 0-387-36720-9, 277–298.

[19] Klawonn F., Lechner W., Grigull L. (2013); Case-Centred Multidimensional Scaling for Classification Visualisation in Medical Diagnosis, *Health Information Science,Lecture Notes in Computer Science*, ISBN 978-3-642-37898-0, 7798: 137–148.

[20] Borg I., Groenen P. (1997); *Modern Multidimensional Scaling: Theory and Applications*, Springer, ISBN 0-387-94845-7.

[21] Bernatavičienė J., Dzemyda G., Marcinkevičius V. (2007); Conditions for Optimal Efficiency of Relative MDS, *Informatica*,ISSN 0868-4952, 18(2): 187-202.

[22] Bernatavičienė J., Dzemyda G., Marcinkevičius V. (2007); Diagonal majorization algorithm: properties and efficiency, *Information technology and control*, ISSN 1392-124X, 36(4): 353–358.

[23] Tenenbaum J.B., de Silva V., Langford J.C. (2000); A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, ISSN 0036-8075, 290(5500): 2319–2323.

[24] Karbauskaitė R., Dzemyda G., Marcinkevičius V. (2010); Dependence of Locally Linear Embedding on the Regularization Parameter, *TOP*, ISSN 1134-5764, 18(2): 354–376.

[25] Medvedev V., Dzemyda G. (2006); Optimization of the local search in the training for SAMANN neural network, *Journal of global optimization*, ISSN 0925-5001, 35(4): 607–623.

[26] Bernatavičienė J., Dzemyda G., Kurasova O., Marcinkevičius V. (2006); Optimal Decisions in Combining the SOM with Nonlinear Projection Methods, *European Journal of Operational Research*, ISSN 0377-2217, 173(3): 729–745.

[27] Medvedev V., Dzemyda G., Kurasova O., Marcinkevičius V. (2011); Efficient Data Projection for Visual Analysis of Large Data Sets Using Neural Networks, *Informatica*, ISSN 0868-4952, 22(4): 507–520.