# COMPARATIVE ANALYSIS OF SIMILARITY MEASURES FOR MEDICAL STREAMING DATA

Jolita Bernatavičienė, Gediminas Bazilevičius,
Gintautas Dzemyda, Viktor Medvedev
*Vilnius University, Institute of Mathematics and Informatics*
*Akademijos St. 4, LT-08663, Vilnius, Lithuania*
e-mail: Jolita.Bernataviciene@mii.vu.lt

**Abstract**

In this paper, we analyse the similarity measures for comparison of medical streaming data (MSD). In general, the real time streaming data can be interpreted as the multivariate time series. In medicine, the comparison of patient physiological MSD data can be useful in disease detection. Four similarity measures Correlation Coefficient, Frobenius norm, Principal Component Analysis similarity factor, Multidimensional Dynamic Time Warping are presented and the experiments of the comparison of these measures have been performed.

# 1 Introduction

Measuring the similarity between objects described by several features is very important in data mining and decision-making process. Moreover, values of features vary in time. In this case, we have some multivariate time series that characterizes the behaviour of particular object in time.

In this paper, we investigate the comparison problem of multivariate physiological time series. Different similarity measures are used and compared. A physiological time series is a series of some medical observations over a period of time. Such type of data can be collected using devices (or sensors) that collect personal medical features, such as heart rate, blood pressure, etc. When the observed space is multidimensional, the time series becomes multivariate [1].

In general, the problem is to detect events in real time streaming data that can be interpreted as the multivariate time series. Decisions are made in accordance on the previously detected and estimated streaming data. Therefore, the problem is to compare the new real time streaming data with the previously detected one. The comparison problem may be formulated as the search of most similar segment in the previously detected and estimated streaming data to the new real time streaming data.

Denote the previously detected and estimated streaming data and the new real time streaming data as

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

Here $T_a$ and $T_b$ are the numbers of observations, $T_a > T_b$, $n$ is the number of measured features. In a result, we need to find the optimal place of $X^b$ on $X^a$. The place is defined by some time moment $T_* : 1 \leq T_* \leq T_a - T_b + 1$. Similarity measures for multivariate time series are used as the criterion of optimality.

## 2 Similarity Measures for Multivariate Time Series

Different techniques and similarity measures are introduced and used for comparison of multivariate time series of different nature: [4], [8]. Four similarity measures for MTS are presented here.

*The Frobenius norm* is often used in matrix analysis [6]. This similarity measure is based on Euclidean distance. Frobenius norm of a matrix $X^b$ is defined by formula:

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

here $tr$ is the sum of the elements on the diagonal of the square matrix. Frobenius norm is used to compare the similarity of two matrices $X^b$ and $X^c$. The similarity is defined by formula $\left\|X^b - X^c\right\|_F$. In our case, the matrix $X^c$ is a segment of length $T_b$ in the previously detected and estimated streaming data $X^a$. The best possible value of the Frobenius norm is 0.

*Correlation coefficient* between two matrices of the same size also can be used as similarity measure [2]:

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

where $\bar{X}^b$ and $\bar{X}^c$ are the means of the $X^b$ and $X^c$, respectively. The best possible value of correlation coefficient is 1.

The third similarity measure for multivariate time series is *Principal Component Analysis (PCA) similarity factor* [5], [8]. PCA similarity factor is defined by formula: $S_{PCA}(X^b, X^c) = tr(L^{'}MM^{'}L)$, where $L$ and $M$ are matrices that contained the first $k$ principal components of $X^b$ and $X^c$. The best possible value of the PCA similarity factor is $k$. In our experiments $k = 1$.

*Multidimensional Dynamic Time Warping* (MDTW) is presented in [3]. Some distance matrix is defined: $\{d(p,q) = \sum_{k=1}^{n}(x_{kp}^b - x_{kq}^c)^2,\ p,q = 1,...,T_b\}$. Then the matrix $D$ of cumulative distances is calculated as in the traditional DTW algorithm [7]:

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

$(p,q)$ defines the pair of $p$th observation in $X^b$ and $q$th observation in $X^c$. Finally, the minimal path and distance along minimal path is obtained using the 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 [7] for details. The best possible value of the MDTW is 0.

# 3 Comparative Analysis of Similarity Measures for Multivariate Time Series

Data from PhysioNet/Computing in Cardiology Challenge (*http://www.physionet.org/challenge/2012/*) is used for the experimental analysis. The records were collected in the Intensive Care Unit. In the experiments we used multivariate time series dataset of 5 patients of the same age, i.e. if to follow the notations of Section 1, we have 5 different matrices $X_i^a, i = \overline{1,5}$, consisting of $T_a = 47$ observations of $n = 4$ features (non-invasive diastolic arterial blood pressure, non-invasive systolic arterial blood pressure, heart rate, temperature).

As the new real time streaming data $X^b$, we have chosen first ten observations from one more patient record from PhysioNet data base. For each similarity measure, the optimal place of $X^b$ on $X^a$ has been found. Then the values of remaining measures were computed for the same place of $X^b$ on $X^a$. The place of $X^b$ on $X^a$ may be denoted as follows: $i[T_{start}; T_{end}]$, where $i$ is the order number of patient, $i = \overline{1,5}$, $T_{start}$ and $T_{end}$ are start and end positions of $X^b$ on $X^a$. The results are presented in Table. The best-found places of $X^b$ on $X^a$ for different measures are given in bold.

Table : Values of similarity measures

| $i$ | Place of $X^b$ on $X^a$ | $\|X^i\|_F$ | $r$ | $S_{PCA}$ | MDTW |
|---|---|---|---|---|---|
| 1 | 1[6;12] | 5.9039 | **0.7635** | 0.4986 | 39.6146 |
| | 1[11;17] | **5.3568** | 0.6873 | 0.4650 | **17.0764** |
| | 1[4;10] | 5.7248 | 0.6423 | **0.5902** | 51.0581 |
| 2 | 2[6;12] | **3.4326** | **0.7935** | 0.4113 | **18.2572** |
| | 2[31;37] | 8.59732 | -0.4569 | **0.8039** | 26.7820 |
| 3 | 3[37;43] | **2.4704** | **0.9279** | 0.8029 | 29.7421 |
| | 3[36;42] | 2.8350 | 0.9260 | 0.8179 | **23.2644** |
| | 3[41;47] | 2.5146 | 0.9139 | **0.8205** | 23.5281 |
| 4 | 4[36;42] | **6.5803** | **0.6427** | 0.7029 | 20.5356 |
| | 4[11;17] | 7.0432 | 0.3517 | 0.0058 | **17.5301** |
| | 4[20;26] | 6.8849 | -0.1647 | **0.8860** | 24.5282 |
| 5 | 5[29;35] | 8.9804 | **-0.9298** | 0.4815 | 60.3518 |
| | 5[24;31] | 8.6541 | -0.7769 | 0.1075 | **24.3271** |
| | 5[15;21] | **7.4975** | -0.4360 | 0.2291 | 27.4271 |
| | 5[7;13] | 9.8165 | -0.0824 | **0.6498** | 64.0362 |

# 4 Conclusions

Four similarity measures for comparison of medical streaming data are analysed. The results of this paper may be considered as the preliminary ones. For more comprehensive conclusions, the wide statistical estimations must be done.

Application of the Frobenius norm, Correlation coefficient between matrices and Multidimensional dynamic time warping gives similar results - these measures indicate often the same optimal place of $X^b$ on $X_i^a$. PCA similarity factor gives much more different results as compared with the first three measures. This is advantage of application of measures of different nature.

Data are very specific for the individual patient. Therefore, in all the experiments, the values of all the measures are far from the best possible ones, given in Section 2. This may cause the problem of the reliability of the decision when the problem is to detect events in real time streaming data in accordance on the previously detected and estimated streaming data of various patients. Some threshold of measure values is necessary to be fixed for proper decision.

**Acknowledgement.** The research has been inspired and funded partly by joint stock company "Algoritmų sistemos" and Agency for Science, Innovation and Technology (MITA), Lithuania.

# References

[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.

[2] Escoufier Y. (1973). Le Traitement des Variables Vectorielles. Biometrics (International Biometric Society). Vol. **29(4)**, pp. 751-760.

[3] ten Holt G., Reinders M., Hendriks E. (2007). Multidimensional dynamic time warping for gesture recognition. Annual Conference on the Advanced School for Computing and Imaging.

[4] 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, 2008, USA. pp. 253-259.

[5] Krzanowski W. (1979). Between-groups comparison of principal components. *JASA* . Vol. **74(367)**.

[6] Moon T., Striling W. (2000). *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall.

[7] Sanguansat P. (2012). Multiple Multidimensional Sequence Alignment Using Generalized Dynamic Time Warping. WSEAS Transactions on Mathematics. Vol. **11(8)**, pp. 668-678.

[8] 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*, pp. 65-74.