СРАВНЕНИЕ МНОГОМЕРНЫХ ДАННЫХ СТЕПЕНИ СХОДСТВА В ИССЛЕДОВАНИИ ФИЗИЧЕСКОЙ АКТИВНОСТИ ЛЮДЕЙ

Г. Базилевичус, В. Мартинкевичус

Институт математики и информатики, Вильнюсский университет
Вильнюс, Литва
e-mail: gediminas.bazilevicius@mii.vu.lt, virginijus.marcinkevicius@mii.vu.lt

В последнее время люди приобретают различные технические приборы для наблюдения за своей физической активностью. Чтобы получить информацию из данных, собранных с помощью этого оборудования, необходимо знать, как сравнивать и классифицировать их. Измерения степени сходства используются для сравнения объектов, однако различные инструменты измеряют различные атрибуты объектов и часто трудно определить, какие из них наиболее точные.

Ключевые слова: степени сходства; классификация; многомерные временные ряды; кластеризация.

THE COMPARISON OF MULTI-DIMENSIONAL DATA SIMALARITY MEASUREMENTS ON THE RECONGNITION OF HUMAN ACTIVITIES

G. Bazilevičius, V. Marcinkevičius

Institute of Mathematics and Informatics, Vilnius University
Akademijos 4, LT-08663,
Vilnius, Lithuania

Recently a tendency dominates that people acquire various technological equipment to observe their physical activities. In order to get information from the data collected by this equipment, it is necessary to know how to compare or classify it. The similarity measurements are used for the comparison of objects, however, different measurements evaluate different attributes of objects and it is often hard to determine which of them is the most accurate. In this article, the influence of similarity measure analysis is introduced, in order to find out how the similarity measurements impact the time series segment classification. The object of the research is which similarity measurements allow to determine the type of physiological activity the most accurately, when evaluating time series that define human physiological attributes and the attributes of their position in a space.

Keywords: similarity measures; classification; multivariate time series; clusterization.

INTRODUCTION

People that participate in sports, that perform active physical activities at work (Patel, Park, Bonato, Chan, & Rodgers, 2012), (Alemdar & Ersoy, 2010), in a hospital or during their leisure time observe their health activities, road or body position coordinates, by sensors. They receive various data about the nature of their activities. This data is multi-dimensional, because the physiological activities, state or actions are defined by many various attributes and variables that are usually independent and changing in time (Esling & Agon, 2012), (Hamilton, 1994). Time is the reading that is important when investigating data and especially when such data is necessary to be investigated at the same moment when they are received (Gaber, Zaslavsky, & Krishnaswamy, 2005). Time series is a set of values of data collected in time, which defines the behavior, state of a specific object. Such data can be received from sensors by continuously resending them to a more powerful calculation device: a mobile phone, computer, server (Rodríguez, Goñi, & Illarramendi, 2005).

One of the challenges of time series is the search of similar time series segments (Bernatavičienė et al., 2015). A specific chosen similarity determination technique has a great impact when solving tasks, the purpose of which is to find data of a similar nature, in such areas. When methodically comparing two time series, the numeral estimate is calculated, i.e. a measure that shows how much the segment is a similar to a specific sample.

In this article it is investigated what influence do the similarity measures that are capable to estimate the similarity of two time series (segments) on the classification. When classifying (Hu, Shao, & Tan, 2011), (Attar, Sinha, & Wankhade, 2010) the states of activity we can investigate the available measures and tell which measure is the best when grouping the most similar segments of the states of activity to separate respective groups. In our case, the segments of the state of the same activity should be assigned to a respective group, which would mean one and the same state of activity, e.g. running, walking, etc. Such a classification could help to describe a person's body position that could serve the manifestation of epilepsy, heart attack, stroke in accordance with body movements typical for a respective illness (Tasoulis et al., 2011), at the same time warning the family doctor about the disorder in human health.

SIMILARITY MEASURES FOR MULTIDIMENSIONAL TIME SERIES

In order to find similarities in the multi-dimensional time series, the overall estimate of one measure or measure groups can be used. In this research only 4 measures of similarity are introduced, with which the best results of the classification were obtained. The similarity measures that compare the segments of the multi-dimensional time series of the same size are presented in detail in the article of J. Bernatavičienė (Bernatavičienė et al., 2015).

In our article, we will mark the entire multi-dimensional matrix (time series), of which the segment of a fixed size that is called a sample and we will mark as X^b is chosen, as X^a. By using the sliding window, the sample X^b is compared to the all other segments of the X^a multi-dimensional matrix X^b length segments that we will mark in the work as X^c. The sizes of the compared X^b, X^c matrices are the same.

$$X^a = \begin{pmatrix} x_{11}^a & \cdots & x_{1T_a}^a \\ \vdots & \ddots & \vdots \\ x_{n1}^a & \cdots & x_{nT_a}^a \end{pmatrix}; X^b = \begin{pmatrix} x_{11}^b & \cdots & x_{1T_b}^b \\ \vdots & \ddots & \vdots \\ x_{n1}^b & \cdots & x_{nT_b}^b \end{pmatrix}. \quad (1)$$

Here $T_a > T_b$. Denote the sample of n features and T_b observations.

Frobenius norm. Frobenius norm – is a similarity measure based on the Euclidean discovery that is used for the analysis of matrices (Moon & Stirling, 1999), (Yang & Shahabi, 2004). The Frobenius norm of one multi-dimensional matrix X^b is defined as:

$$\|X^b\|_F = \sqrt{\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^b)^2} = \sqrt{trace((X^b)^{\text{transpose}}\ X^b)}. \quad (2)$$

The $trace$ is the sum of elements on the diagonal of the square matrix. We define the two multi-dimensional matrices comparison as: $D_{Frob}(p,q) = \|X^b - X^c\|_F$.

Correlation coefficient. The correlation coefficient is calculated according to this formula:

$$D_{Corr}(p,q) = \frac{\left(\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^b - \overline{X^b})(x_{pq}^c - \overline{X^c})\right)}{\left(\sqrt{\sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^b - \overline{X^b})^2 \sum_{p=1}^{n}\sum_{q=1}^{T_b}(x_{pq}^c - \overline{X^c})^2}\right)}, \quad (3)$$

where $\overline{X^b}$ and $\overline{X^c}$ are the averages of X^b and X^c matrices. $x_{p,q}^b$ and $x_{p,q}^c$ is a respective element from X^b and X^c matrices.

MDTW. The measure of dynamic time warping similarity (DTW) (Berndt & Clifford, 1994), (Ten Holt, Reinders, & Hendriks, 2007) is often applied for the research of the financial, medicinal data. This measure of similarity defines the dynamics of time series change. The modification of DTW similarity measure for the multi-dimensional time series is the measure of MDTW (Moon & Stirling, 1999).

MDTW matrix is calculated the same as in the traditional DTW algorithm (Sanguansat, 2012):

$$D_{MDTW}(p,q) = \begin{cases} d(1,1), & if\ p=1, q=1, \\ d(p,q) + D_{MDTW}(p-1,q), & if\ p=2,\dots,T_b, q=1, \\ d(p,q) + D_{MDTW}(p,q-1), & if\ p=1, q=2,\dots,T_b, \\ d(p,q) + min\begin{cases} D_{MDTW}(p-1,q), \\ D_{MDTW}(p,q-1), \\ D_{MDTW}(p-1,q-1), \end{cases} & in\ other\ cases. \end{cases} . \quad (4)$$

Where some distance $d(p,q)$is defined as follows:

$$d(p,q) = \sum_{k=1}^{n}\left(x_{kp}^b - x_{kq}^c\right)^2, \text{where } p,q = 1,\dots,T_b. \quad (5)$$

The (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_{MDTW}.

Hamming measurement. Essentially, Hamming measurement represents the number of dimensions between two d-dimensional vectors that contain different elements. Hamming is described formally like that where i is index of segment element; d is size of segment (Dashiell Kolbe, 2004), (D. Kolbe, Zhu, & Pramanik, 2007), (Hamming, 1950):

$$D_{Hamm}(p,q) = \sum_{i=1}^{d}\begin{cases} 1, & if\ p[i] \neq q[i] \\ 0, & if\ p[i] = q[i] \end{cases}. \quad (6)$$

DATA

In this investigation we used "PAMAP2" Physical Activity Monitoring Data Set (Reiss & Stricker, 2012) that contains data of 18 different physical activities, performed by 9 subjects wearing 3 inertial measurement units and a heart rate monitor.

In the experiments, such of the physical activity states that were performed by all of the surveyed people were investigated: lying, sitting, standing, walking, running, cycling, Nordic walking, ascending stairs, descending stairs, vacuum cleaning, ironing, rope jumping.

DATA PRE-PROCESSING AND NORMALIZATION

Scheme of data pre-processing is presented in Fig. *1*. At first, we have to fill missing values that appeared during data collection process. We replaced missing data values with values taken from same time series that comes next to missing values.

In the following step, we remove the linearly correlated data. The attributes of data was used for this purpose.

In the final step, before classifying data, considering to the measurement, the data normalization was applied.

Only with the correlation measure, the normalized multi-dimensional data was compared, where the same states of different people, the parts of the multi-dimensional data were normalized according to the formula: $x_{ij} = (x_{ij} - \bar{x}_j)/\sqrt{\sigma_j^2}$.

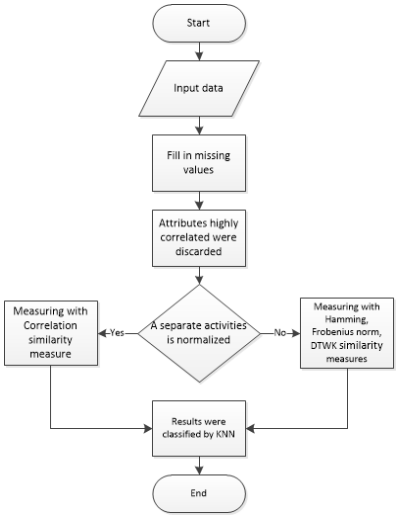

Fig. 1. Data preprocessing scheme

SIMILARITY MEASUREMENT AND KNN

After data pre-processing we executed series of classification experiment with different segment size. These results are presented in Fig. *2*. The size of the compared segments varied from 50 to 1000 time readings, the chosen segment step was equal to 50 time readings.

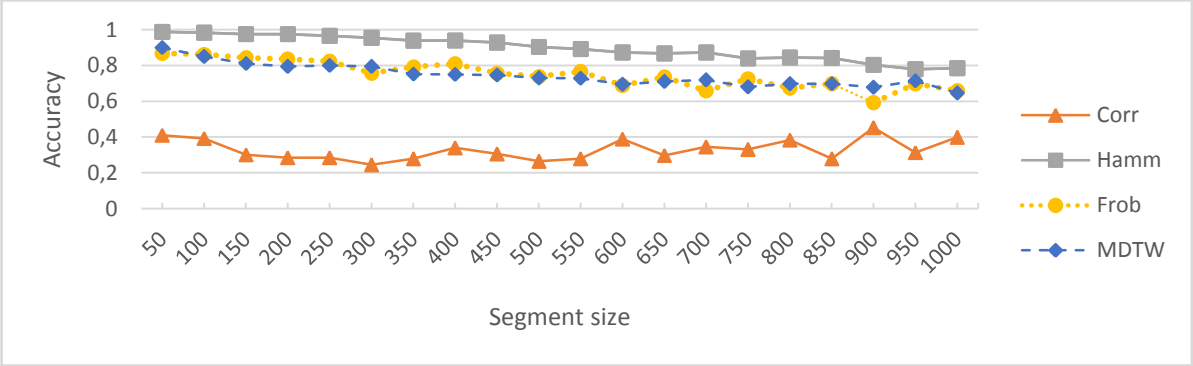

Fig. 2. Dependence between KNN classifier accuracy, segment size and similarity measurement

Dependence between KNN classifier accuracy, segment size and similarity measurement

	Corr	Hamm	Frob	MDTW
50	0,4103	**0,9877**	**0,8673**	**0,9002**
100	0,3926	0,9831	0,8623	0,8495
150	0,3011	0,9751	0,8421	0,8107
200	0,2852	0,9750	0,8341	0,7937
250	0,2853	0,9657	0,8225	0,7996
300	0,2444	0,9545	0,7572	0,7935
350	0,2785	0,9389	0,7924	0,7514
400	0,3399	0,9395	0,8074	0,7500
450	0,3056	0,9283	0,7572	0,7464
500	0,2648	0,9036	0,7390	0,7297
550	0,2792	0,8918	0,7652	0,7281
600	0,3889	0,8729	0,6889	0,6936
650	0,2976	0,8677	0,7341	0,7097
700	0,3449	0,8729	0,6574	0,7184
750	0,3303	0,8394	0,7242	0,6790
800	0,3833	0,8446	0,6742	0,6967
850	0,2789	0,8418	0,6973	0,6979
900	***0,4527***	0,8037	0,5917	0,6780
950	0,3140	0,7791	0,6977	0,7143
1000	0,3981	0,7854	0,6574	0,6458

KNN classifier with the available data tried to classify 12 classes as accurately as possible. The KNN classifier of MATLAB package with such settings was used in experiments: 10-fold cross-validation, 2 neighbors, distance weight 1/d, where d is the distance between this neighbor and the point being classified.

The results of the experiment that were presented in Fig. *2* and show the influence of the similarity measure and segment size on the classification of the states of activity. In this experiment, the algorithm of Hamming measurement showed very good results in all cases and especially correct class assignment accuracy was equal to **0,9877** in, when the length of the segment was 50 time readings, it is an optimal segment size to this measure. The results of the experiment showed that a small segment length is a significantly better variant for the classification of data than a larger segment length, i.e. the smaller the multi-dimensional matrix (less information) the easier it is to find the most similar matrix to the latter. The results gained by the Frobenius norm and the MDTW measure are very similar, the accuracy of classification is adequately good, only the MDTW measure conducts calculations especially slow. The best gained accuracy results were these: Frobenius = **0,8673**, MDTW = **0,9002**, when the segment length was equal to 50 time readings. Which is why after the Hamming algorithm, the most suitable measure for the finding of segment similarities and later classifying of that data is the Frobenius norm measure. The segment similarity results gained with the help of this measure are adequately accurate and calculated fast. The worst measure in our case is the correlation coefficient measure, because the similarity results of this measure are the worst even though the calculation speed is equivalent to the calculation of the Hamming distance. The best-achieved correlation coefficient measure result was **0,4527**, when the segment length was equal to 900 time readings.

CONCLUSIONS

In this paper, the influence of the similarity measures on the classification on human physiological states is investigated. After conducting the experiments, these conclusions were achieved:

When classifying the PAMAP2 Physical Activity Monitoring Data Set data, the best case when the compared matrix segment size is 50 time readings.

Frobenius norm and MDTW similarity measure meaning classification showed the similar classification accuracies. The MDTW measure is not suitable when the segment length is very small, because the calculations take a very long period or a much more powerful computer is needed for such a measure.

When using the correlation method when classifying data, small accuracy is achieved.

LITERATURE

1. Alemdar H., Ersoy C. Wireless sensor networks for healthcare: A survey. Computer Networks, 2010. № 54(15), P. 2688–2710.

2. Attar V., Sinha P., Wankhade K. A fast and light classifier for data streams. Evolving Systems, 2010. № 1(3). P. 199–207.

3. Method for visual detection of similarities in medical streaming data / J. Bernatavičienė [et al.] // International J. of Computers, Communications and Control, 2015. № 10(1). P. 8–21.

4. Berndt D., Clifford J. Using dynamic time warping to find patterns in time series. Workshop on Knowledge Knowledge Discovery in Databases, 1994. 398. P. 359–370.

5. Esling P., Agon C. Time-series data mining. ACM Computing Surveys, 2012. № 45(1). P. 1–34.

6. Gaber M. M., Zaslavsky A., Krishnaswamy S. Mining data streams: A Review. ACM Sigmod Record, 2005. № 34(2). P. 18–26.

7. Hamilton J. Time Series Analysis, 1994.

8. Hamming R. W. Error Detecting and Error Correcting Codes. Bell System Technical Journal, 1950.

9. Hu S., Shao Z., Tan J. A Real-Time Cardiac Arrhythmia Classification System with Wearable Electrocardiogram. 2011 International Conference on Body Sensor Networks, 2011. P. 119–124.

10. Kolbe D. Efficient k-Nearest Neighbor Searching in Non-Ordered Discrete Data Spaces, 2004.

11. Kolbe D., Zhu Q. Z. Q., Pramanik S. On k-Nearest Neighbor Searching in Non-Ordered Discrete Data Spaces. 2007 IEEE 23rd International Conference on Data Engineering, 2007. P. 426–435.

12. Moon T. K., Stirling W. C. Mathematical Methods and Algorithms for Signal Processing, 1999. P. 937.

13. A review of wearable sensors and systems with application in rehabilitation. J. of NeuroEngineering and Rehabilitation / S. Patel [et al.]. 2012. № 9(1). P. 21.

14. Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. Proceedings – International Symposium on Wearable Computers, ISWC, 2012. P. 108–109.

15. Rodríguez J., Goñi A., Illarramendi A. Real-time classification of ECGs on a PDA. IEEE Transactions on Information Technology in Biomedicine, 2005. № 9(1). P. 23–34.

16. Sanguansat P. Multiple Multidimensional Sequence Alignment Using Generalized Dynamic Time Warping. WSEAS Transaction on Mathematics, 2012. № 11(8). P. 668–678.

17. Statistical Data Mining of Streaming Motion Data for Fall Detection in Assistive Environments / S. K. Tasoulis [et al.]. 2011. P. 3720–3723.

18. Ten Holt G. a, Reinders M. J. T., Hendriks E. a. Multi-Dimensional Dynamic Time Warping for Gesture Recognition. Time, 2007. 5249. P. 23–32.

19. Yang K., Shahabi C. A PCA-based similarity measure for multivariate time series. Proceedings of the 2nd ACM International Workshop on Multimedia Databases MMDB ’04, 65, 2004.