Symbolic Time Series Representation for Stream Data Processing

Jakub Sevcech
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Ilkovičova 2, 842 16 Bratislava, Slovakia
jakub.sevcech@stuba.sk

Maria Bielikova
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Ilkovičova 2, 842 16 Bratislava, Slovakia
maria.bielikova@stuba.sk

***Abstract*—Over the past years, many representations for time series were proposed with the main purpose of dimensionality reduction and as a support for various algorithms in the domain of time series data processing. However, most of the transformation algorithms are not directly applicable on streams of data but only on static collections of the data as they are iterative in their nature. In this work we propose a symbolic representation of time series along with the method for transformation of the data into proposed representation. As one of the basic requirements for applicable representation is the distance measure which would accurately reflect the true shape of the data, we propose a distance measure operating on the proposed representation and lower bounding the Euclidean distance on the original data. We evaluate properties of the proposed representation and the distance measure on a number of different datasets.**

Keywords—time series representation; symbolic representation; stream processing; lower bound

I. INTRODUCTION

Many different time series representations were proposed over the past years. However, only small portion of them is applicable on stream data processing as most of the transformation procedures are iterative in their nature or they require some sort of statistical information about the whole dataset.

Our primary motivation is to propose a time series representation applicable in stream data processing. The prime requirement for such a time series representation is incremental procedure of the data transformation.

In our work, we are most interested in symbolic representations as they enable the application of methods that are not defined for real-valued data [1] such as Markov models, suffix trees or many algorithms from the domain of text processing. An example of such representation is SAX [1] – one of the most widely used time series representations. Similarly to the majority of other representations however, transformation into the SAX representation is iterative and cannot be directly applicable to stream data processing as it requires statistical information about the whole transformed dataset. Examples of other symbolic time series representation can be found in [1, 2, 3] but they all share the same limitation, stream data cannot be directly transformed into these representations.

The representation we propose is based on the symbolic time series representation used in [2] for rule discovery in time series. In this work authors use subsequence clustering for construction of clusters of similar sequences. They use cluster identifiers as symbols for transformation of the time series into sequences of symbols. This work influenced many researchers for several years, but it has two major limitations:

- It is iterative due to the K-means algorithm used for cluster formation.
- It has been proved that the transformation process produces meaningless clusters that do not reliably reflect the data they were formed from [4].

In our work, we address both of these limitations. To be able to transform the data into the proposed representation incrementally, we use an incremental greedy algorithm creating new clusters every time new sequence sufficiently distant from all other clusters occurs. In previous works multiple authors used various techniques to form meaningful subsequence clusters. Most of these methods limited the number of sequences used in the clustering process by using motifs [5] or perceptually important points [6]. We hypothesize, that by changing the clustering algorithm and not limiting the number of formed clusters, we will form meaningful clusters.

According to the authors of another study [1] many symbolic time series representations were proposed, but the distance measures on these representations show little correlation with the distance measures on original data. To show our representation is not the case, we propose the distance measure *SymD* that returns the minimum distance between time series in the representation and we show it lower bounds the Euclidean distance on the original time series. To evaluate the applicability of time series representation we use the tightness of lower bounds (TLB) [7] as it is the current consensus in the literature [8].

The rest of the paper is organized as follows. Section 2 introduces the symbolic time series representation. Section 3 defines the distance measure on the proposed representation and provides the proof it lower bounds the Euclidean distance on the original data. An experimental evaluation of properties of the proposed representation and distance measure on the number of datasets is presented in section 4. We conclude by summarizing obtained results and by hints on future work.

II. The Symbolic Representation

As a base for our time series representation we use an assumption presented in [9]. The authors state that frequent patterns extracted from time series data are more stable than the time series itself. We use this assumption to form the main idea of our representation as to represent time series data as a sequence of reoccurring patterns. We search for reoccurring similar subsequences in the course of the whole data stream by clustering subsequences. We transform them into sequences of symbols where every cluster identifier is transformed into a symbol similarly to the representation proposed in [2]. For the purpose of our work, we will refer the proposed representation as to *Incremental Subsequence Clustering* (ISC).

The transformation of stream data to the ISC representation can be divided into three steps:

1. Split incoming data into overlapping subsequences using running window.
2. Cluster subsequences by their similarity.
3. Use cluster identifiers as symbols, subsequences are transformed to.

The main difference of the proposed ISC representation to the representation Das et al. used [2] is the clustering algorithm we use for symbol formation. They used K-means, which is iterative in its nature and requires the number of formed clusters to be specified in advance. We use incremental greedy algorithm not limiting the number of cluster but limiting the maximal distance of instances in the cluster. The algorithm assigns subsequence into the cluster if its distance from the cluster centre is smaller than the predefined threshold. The algorithm forms new cluster with the subsequence in its centre if no cluster with the distance to the processed subsequence lower than the maximal distance exists. The proposed representation forms a dictionary of symbols (clusters) which grows with the amount of the data processed. We adopt the already mentioned assumption about frequent pattern stability presented by [9] and we assume the speed of growth of the dictionary of symbols will decrease with the amount of data processed. The experiments supporting this claim are presented in section 4.

The dictionary of symbols represent the main difference between the proposed ISC representation and SAX. The symbols formed by SAX represent equiprobable intervals of PAA coefficients [7] which in turn are results of an aggregate function (mean) performed on a sliding window of a time series. In the case of our representation, individual symbols represent repeating shapes and the alphabet of symbols represent a dictionary of all shapes occurring in the course of the time series.

The transformation uses three parameters: symbol length (size of the running window), step between two consecutive windows (typically equal to a fraction of symbol length), maximal distance of cluster centre and a subsequence in the cluster. Every symbol in the dictionary of symbols is represented by z-normalized subsequence forming the centre of the cluster and the cluster identifier. The transformed time series is formed by a sequence of triples: cluster identifier, mean and standard deviation of the original subsequence. Using these attributes we are able to approximately reconstruct the original time series.

III. Lower Bounding Similarity Measure

Having defined the symbolic time series representation, we now define the similarity measure on the transformed data and we prove it lower bounds the Euclidean distance on the original data. As the distance measure for the ISC representation we adapt the representation introduced in [1] where the authors proposed an adaptation of Euclidean distance called MINDIST. MINDIST uses table of distances between individual symbols in the SAX representation of the data to calculate the overall distance. In this representation, the distance table depends solely on the number of symbols used in the transformation process. As the ISC representation does not use stable alphabet of symbols and the distance between symbols depends on the shape of the data they are formed from, we have to calculate the distance table from the symbol alphabet. We define the symbolic distance measure (*SymD*) as an adaptation of MINDIST distance measure that returns the minimum distance between time series in the ICS representation.

The proposed distance measure builds on the most common time series distance measure - Euclidean distance. Eq. (1) shows the formula for Euclidean distance of two time series, Q and C of the length n.

$$ED(Q,C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2} \tag{1}$$

We show the lower bounding property of *SymD* by introducing an auxiliary distance measure as transition from Euclidean distance to the presented *SymD* distance measure. Among these distance measures we demonstrate the lower bounding property and transitively we extend the proof to the proposed *SymD* distance measure on the ISC representation (Eq. (2)). The auxiliary distance measure we introduce (for the explanation sake named *OverED*) is described in the following paragraphs.

$$SymD(\hat{Q},\hat{C}) \leq OverED(\bar{\bar{Q}},\bar{\bar{C}}) \leq ED(Q,C) \tag{2}$$

In Eq. (2), Q and C refers to two compared time series in their raw representation. $\bar{\bar{Q}}$ and $\bar{\bar{C}}$ refer to time series split into overlapping subsequences of length w and shift s. $\hat{Q}$ and $\hat{C}$ refers to time series in ISC representation.

The distance measure *OverED* refers to the adapted Euclidean distance, where we split the time series into overlapping subsequences of equal length w and shift s between two consecutive subsequences. The distance between two subsequences is calculated using Euclidean distance.

An illustration of time series transformed to overlapping subsequences is presented on Fig. 1.

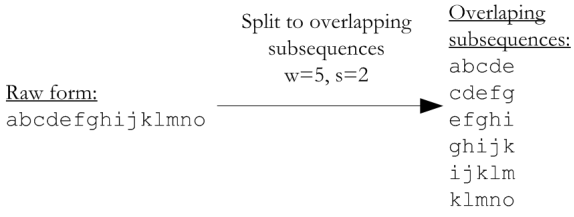

Fig. 1. Example of sequence split into overlapping subsequences

Fig. 1 shows a sequence of values in a time series **abcdefghijklmno** where every symbol refers to a different value. *OverED* operates on the time series split into overlapping subsequences of length *w* and shift *s*. We choose in our example *w=5* and *s=2* and we split the sequence.

As we can see from the example, some values are represented repeatedly in the transformed data (eg. **c**, **d**, **e** ...) and some are represented only once or with different frequencies (eg. **a**, **b**, **n** and **o**). The contribution of the time series value to the overlapping representation depends on its position in the processed time series. None of these values however is repeated more than $\left\lceil \frac{w}{s} \right\rceil$ times. We define the *OverED* as sum of squared distances between subsequences (similarly to Euclidean distance) divided by the maximal number of occurrences of individual values in the transformed representation. Eq. (3) shows the definition of *OverED* where $\overline{\overline{q}}_i$ and $\overline{\overline{c}}_i$ are *i*-th subsequences of time series $\overline{\overline{Q}}$and $\overline{\overline{C}}$, *n* is the total length of time series, *w* is the subsequence length, *s* is the shift between two subsequences and $\left\lceil \frac{n-w}{s} \right\rceil$ is the total number of symbols in the transformed representation.

$$OverED(\overline{\overline{Q}}, \overline{\overline{C}}) = \sqrt{\frac{\sum_{i=1}^{\left\lceil \frac{n-w}{s} \right\rceil} ED(\overline{\overline{q}}_i, \overline{\overline{c}}_i)^2}{\left\lceil \frac{w}{s} \right\rceil}} \quad (3)$$

An alternative notation for the *OverED* distance measure is based on the number of occurrences of individual time series values in the overlapping representation. To measure the contribution of individual values to the resulting representation, we can split the time series into three parts:

- Start - with increasing contribution of values to the overlapping representation.
- Centre - with constant contribution of different values to the representation.
- End - with decreasing contribution of different values.

The distance measure on such representation have to adjust to the variable contribution of values to the representation. We can define the contribution of for each part of the time series to the distance measure separately:

$$Start(\overline{\overline{Q}}, \overline{\overline{C}}) = \sum_{i=1}^{\left\lceil \frac{w}{s} \right\rceil} \sum_{j=1}^{\min(s, w-s(i-1))} i(q_{is+j-1} - c_{is+j-1})^2 \quad (4)$$

$$End(\overline{\overline{Q}}, \overline{\overline{C}}) = \sum_{i=1}^{\left\lceil \frac{w}{s} \right\rceil} \sum_{j=1}^{\min(s, w-s(i-1))} i(q_{n-is+j} - c_{n-is+j})^2 \quad (5)$$

$$Centre(\overline{\overline{Q}}, \overline{\overline{C}}) = \left\lceil \frac{w}{s} \right\rceil \sum_{i=w+1}^{n-w-1} (q_i - c_i)^2 \quad (6)$$

In Eq. (4), Eq. (5) and Eq. (6), q_i and c_i to *i*-th values of time series *Q* and *C*. Since every q_i and c_i from *Q* and *C* respectively is not repeated in the representation more than $\left\lceil \frac{w}{s} \right\rceil$ times, we can divide the sum of distances of three parts of the time series by $\left\lceil \frac{w}{s} \right\rceil$ and the resulting distance will be never greater than *ED(Q,C)* thus it satisfies the lower bounding property.

$$OverED(\overline{\overline{Q}}, \overline{\overline{C}}) = \sqrt{\frac{Start(\overline{\overline{Q}}, \overline{\overline{C}}) + Centre(\overline{\overline{Q}}, \overline{\overline{C}}) + End(\overline{\overline{Q}}, \overline{\overline{C}})}{\left\lceil \frac{w}{s} \right\rceil}} \leq ED(Q, C) \quad (7)$$

The last step of the proof is to show that clustering of similar subsequences using Euclidean distance into clusters, defined by its centre and maximal distance of the subsequence from the centre, lower bounds the *OverED* distance measure. The sole difference between *SymD* and *OverED* is, that the *SymD* does not compute the distance using the raw time series subsequences, but rather centres of cluster every subsequence is attached to. To calculate the distance of time series in ISC representation, we have to substitute the distance of overlapping subsequences by the distance of clusters centres. However, the substitution by these clusters introduces some error as they are only approximate representation of the original overlapping subsequence. To use the cluster centres instead of the original subsequences we have to define the relation of Euclidean distance of the individual subsequences and the Euclidean distance of cluster centres. For the purpose of this proof $\tilde{a}$ and $\tilde{b}$ refer to the cluster centres time series *a* and *b* respectively are associated with. The cluster diameter or maximal distance between cluster centre and time series associated to this cluster is denoted *r*. We start the proof using the equality of Euclidean distance of cluster centres to itself in Eq. (8).

$$ED(\tilde{a}, \tilde{b}) = ED(\tilde{a}, \tilde{b}) \quad (8)$$

Using triangular inequality (Eq. (9)) of *ED* twice on the right side of Eq. (8), we obtain Eq. (10)

$$ED(a, b) \leq ED(a, c) + ED(c, b) \quad (9)$$

$$ED(\tilde{a}, \tilde{b}) \leq ED(a, \tilde{a}) + ED(a, b) + ED(b, \tilde{b}) \quad (10)$$

As $ED(a, \tilde{a}) \leq r$ and $ED(b, \tilde{b}) \leq r$ we can transform the Eq. (10) to:

$$ED(\tilde{a}, \tilde{b}) - 2r \leq ED(a, b) \quad (11)$$

The geometrical illustration of this proof is on Fig. 2.

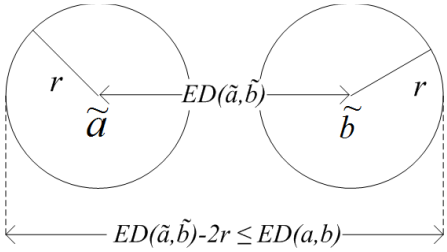

Fig. 2. Geometrical illustration of the relation between distance and distance of cluster centres.

By applying the Eq. (11) on OverED distance measure from Eq. (3), we show that:

$$\sqrt{\frac{\sum_{i=1}^{\left\lceil \frac{n-w}{s} \right\rceil} ED(\hat{q}_i, \hat{c}_i)^2}{\left\lceil \frac{w}{s} \right\rceil} - 2r \left\lceil \frac{n-w}{s} \right\rceil} \leq \sqrt{\frac{\sum_{i=1}^{\left\lceil \frac{n-w}{s} \right\rceil} ED(\overline{\overline{q}}_i, \overline{\overline{c}}_i)^2}{\left\lceil \frac{w}{s} \right\rceil}} \quad (12)$$

And thus:

$$SymD(\hat{Q},\hat{C}) = \sqrt{\frac{\sum_{i=1}^{\left\lceil\frac{n-w}{s}\right\rceil} ED(\hat{q}_i,\hat{c}_i)^2}{\left\lceil\frac{w}{s}\right\rceil} - 2r\left\lceil\frac{n-w}{s}\right\rceil} \leq OverED(\bar{\bar{Q}},\bar{\bar{C}}) \quad (13)$$

where n is the total number of values in the time series, $\hat{q}_i$ and $\hat{c}_i$ refers to i-th symbol time series $\hat{Q}$ and $\hat{C}$ in ISC representation, r is the radius of the clusters forming the symbols, w is the length of the symbol and s is the shift between two symbols. Using the Eq. (13), we prove *SymD* lower bounds *OverED* and thus we complete the proof of Eq. (2). We show that the proposed distance measure *SymD* operating on time series transformed into ISC representation lower bounds the Euclidean distance on raw form of the time series.

IV. Evaluation

To evaluate properties of the proposed representation we performed a series of experiments on the UCR datasets [10]. We focused on the evaluation of tightness of lower bound as one of the most widely used metrics for evaluation of time series representations [8]. The second metric we chose for evaluation of proposed representation is the size of symbol dictionary formed during the transformation as it determines the memory requirements of the representation and its applicability in stream data processing.

Since the transformation into the ISC representation requires three parameters to be set, in the following figures we provide several examples of relationship between these attributes, tightness of lower bound and symbol alphabet size. Fig. 3, Fig. 4 and Fig. 5 display the data obtained by processing the Symbols dataset from the UCR [10] repository. Similar results were obtained for other datasets from the repository, but they are omitted due the limited length of this paper.

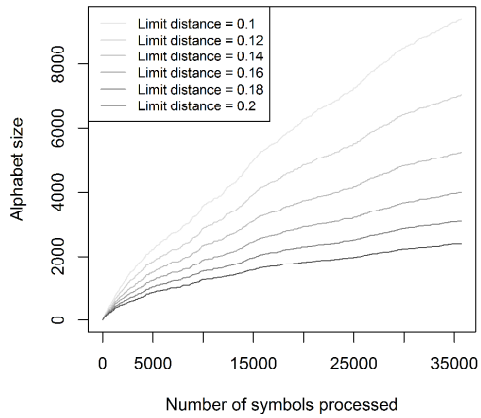

Fig. 3. The relationship between alphabet size and number of data processed with different settings of maximal distances of subsequence to the centre of associated cluster. Data for UCR [10] dataset Symbols.

Fig. 3 shows the relationship between the amount of data processed and the size of the symbol alphabet. The figure displays the evolution of alphabet size with increasing portion of the dataset processed and for different settings of the limit distance used in cluster formation. We can see that the speed of formation of new symbols decreases with the amount of processed data in accordance with our assumption about stability of frequent patterns introduced in the section 2. The differences in the course of alphabet size for distinct limit distance settings indicate the increasing number of clusters formed when size of the cluster is small.

The relation between the size of alphabet formed after transformation of the whole dataset and the size of cluster created during the transformation is displayed on Fig. 4. One can see that the relation is not linear, but with the increasing size of the clusters the number of symbols decreases slower.

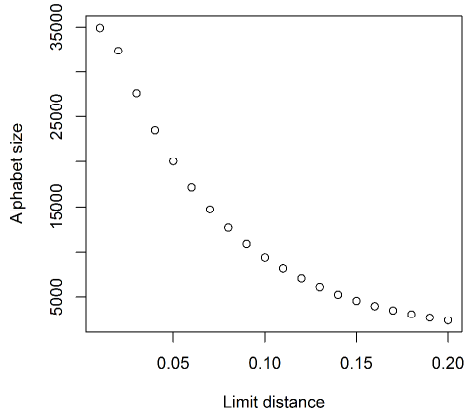

Fig. 4. Relationship between the final alphabet and size of created clusters. Data for UCR [10] dataset Symbols.

With the increasing size of the clusters, more similar subsequences are grouped to the same cluster centre. This should result in decreased accuracy of reconstruction of the representation to the original time series data. The accuracy of reconstruction is reflected in the tightness of lower bound metric as it indicates the ratio between the similarity of two transformed time series calculated using the *SymD* distance measure and the distance calculated using Euclidean distance on the original time series. The relation between tightness of lower bound and cluster size is presented on Fig. 5.

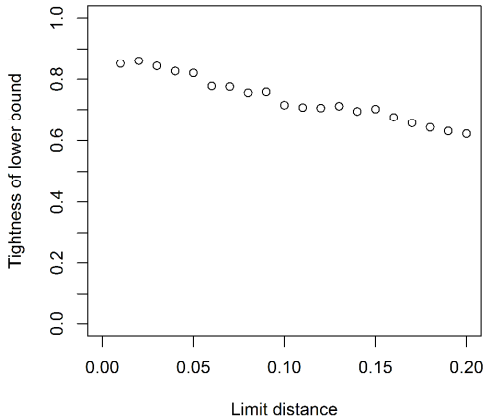

Fig. 5. The relationship between the tightness of lower bound and size of created clusters. Data for UCR [10] datasets Symbols.

To evaluate the tightness of lower bound we performed an experiment where we took a sample of 200 time series from the Symbols dataset and we calculated the average tightness of

lower bound for every pair of these time series. We performed the experiment for different sizes of formed clusters. The results are presented in Fig. 5. The relationship between the tightness of lower bound and cluster size is almost linear with small variability caused by the size of the used sample. These results indicate there is a tradeoff between the size of the created symbol alphabet and the tightness of lower bound obtained by the ISC representation and associated SymD distance measure. When one will choose the settings for the transformation he/she have to decide on the basis of the application at hand.

The relation between the tightness of lower bound and limit distance used in cluster formation for other datasets from the UCR repository [10] is displayed on Fig. 6. The graph shows the TLB increases with the decreasing size of the clusters for every used dataset. The value of the maximal obtained tightness for the used settings, however, is variable between datasets. For some datasets the limit distance have to be smaller to obtain the same TLB. This is caused by the shape of the time series in the dataset.

To compare the proposed representation to other time series representations such as SAX, PAA or DFT, we can use the results presented in [8]. This comparison however, provides only limited informative value as these representations use different parameters and majority of them is iterative in their nature in contrast to the proposed representation. The authors evaluated various time series representations with different transformation settings on EEG dataset from the UCR repository [10]. The obtained tightness of lower bound varied from 0.258 to 0.782. The results for ISC representation in combination with SymD distance measure varied from 0.268 to 0.601 with different settings of the transformation. The proposed representation thus obtained comparable results with possible improvements if smaller limit distance was used in the transformation process.

To evaluate the clustering meaningfulness we had to adapt the formula used in [4]. The clustering meaningfulness is a measure defined on two distinct datasets as a fraction of mean minimal cluster distances within dataset over mean minimal clusters distances between datasets [4]:

$$meaningfulness(\hat{X},\hat{Y}) = \frac{within\_set\_\hat{X}\_distance}{between\_set\_\hat{X}\_and\_\hat{Y}\_distance} \quad (14)$$

The original definition of $within\_set\_\hat{X}\_distance$ presented in [4] calculates the mean minimal distance of cluster centres formed by multiple runs of K-means algorithm on the dataset. Since our clustering algorithm does not use random initialization, the minimal distance of clusters formed by multiple executions of the algorithm would be zero. We simplify the meaningfulness formula to be equal to the mean minimal distance between sets.

To evaluate the meaningfulness of subsequence clusters formed during the transformation of time series into the ISC representation we performed an experiment on several datasets from UCR repository [10]. We clustered pairs of datasets and compared mean distance of formed clusters for different settings of cluster formation. We used whole time series to form the clusters and fractions of the time series as symbols in the ISC representation. As the lengths of the formed symbols we used 1/2, 1/4 and 1/8 of the sequence length. As for other transformation settings, the step between symbols was set for one half of the symbol size (not in the case of whole clustering, where the step was not used) and limit distance between cluster centre and associated subsequences was set to 0.2. The results are presented on Fig. 7 where values of mean minimal distance between two datasets for whole sequence clustering and different lengths of subsequence clustered are showed. One can see the mean distance decreases when the size of the symbol is decreasing for every examined combination of datasets. The change in distance approximately follows the size of the time series fraction used as symbol. This is caused by the space of similar sequences filling up when the length of clustered subsequences is decreasing and when the radius of clusters is fixed. This results in more formed clusters, closer together.

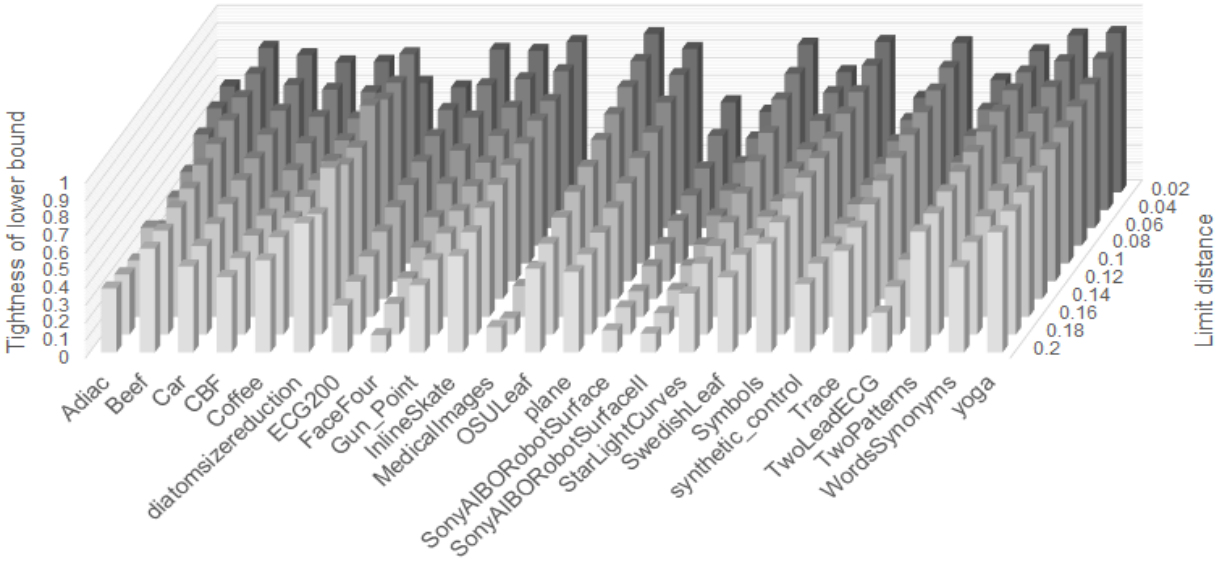

Fig. 6. Tightness of lower bound for different datasets from the UCR repository [10] and different sizes of formed clusters

When we shrink the size of symbols even more, the normalized symbols are reduced into a small alphabet of basic shapes as seen on Fig. 8. The decrease in mean minimal cluster centre distance is not caused by the randomness of formed cluster but by the shrinking subsequence space as the centres are formed from the original time series shapes.

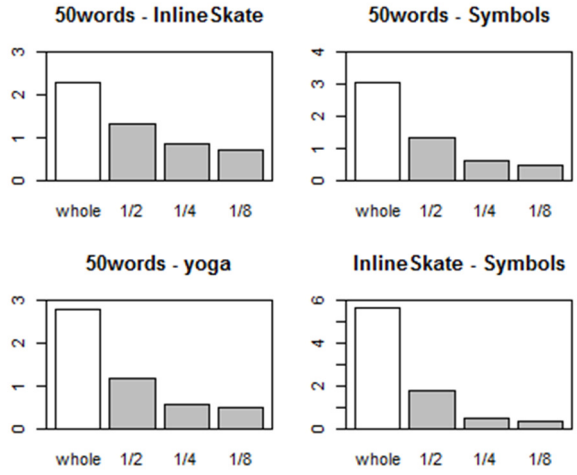

Fig. 7. The meaningfulness evaluation for multiple dataset combinations and different settings of symbol lengths used for the transformation. Diagrams show mean shortest distance between clusters of two datasets when whole sequences were clustered and when ISC transformation was used with symbol sizes of 1/2, 1/4 and 1/8 of time series length.

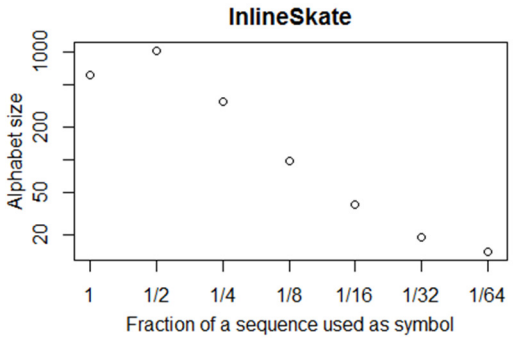

Fig. 8. The alphabet size when different symbol length are used. Logarithmic scale used on both axes.

V. Conclusions

We proposed a symbolic representation of time series (ISC) using clusters of similar subsequences as symbols. The clusters are formed using incremental, greedy algorithm which differs the representation from the representation used in [2] and makes it applicable on stream data processing. As we showed, the subsequence clustering decreases the mean minimal cluster centre distance but it is caused by the shrinking space and not the randomness of formed sequences as they are formed from the basic shapes of the original time series. The major difference of the proposed representation to the SAX representation is the meaning of individual symbols as they represent repeating shapes in the course of the time series.

The similarity metric on the proposed representation (*SymD*) is introduced along with the proof that it lower bounds the Euclidean distance. We performed several experiments on UCR collection of datasets [10] to show the properties of the representation.

As the tightness of lower bound of the representation depends on the settings of the cluster formation process, the potential user of the representation has to make a trade-off between the accuracy of the representation and the size of the alphabet of symbols created during the transformation. One of the limitations of the proposed representation are three parameters of the transformation (symbol length, between symbol step and cluster radius). The representation is applicable in domains where symbols of stable length are repeating over time and where we process large amounts of data continuously. We use the representation for example for short term prediction of electricity consumption or anomaly detection and application monitoring. The other direction of our future work is in the management of the ever growing alphabet of symbols during data stream processing.

Acknowledgment

This work was partially supported by grant VG 1/0646/15 and was created with the support of the Research and Development Operational Programme for the project International centre of excellence for research of intelligent and secure information-communication technologies and systems, ITMS 26240120039, co-funded by the ERDF.

References

[1] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series,” Data Mining and knowledge discovery, vol. 15, no. 2, pp. 107-144, 2007.

[2] G. Das, K. I. Lin, H. Mannila, G. Renganathan, and P. Smyth, “Rule Discovery from Time Series,” in KDD, vol. 98, pp. 16-22, 1998.

[3] M. G. Baydogan, G. Runger, “Learning a symbolic representation for multivariate time series classification,” Data Mining and Knowledge Discovery, pp. 1-23, 2014.

[4] E. Keogh, and J. Lin, “Clustering of time-series subsequences is meaningless: implications for previous and future research,” Knowledge and information systems, vol. 8, no. 2, pp. 154-177, 2005.

[5] J. R. Chen, “Useful clustering outcomes from meaningful time series clustering”, in Proceedings of the sixth Australasian conference on Data mining and analytics, vol. 70, pp. 101–109, 2007.

[6] T. Fu, F. Chung, R. Luk, and C. Ng, “Preventing meaningless stock time series pattern discovery by changing perceptually important point detection”, in Fuzzy Systems and Knowledge Discovery, Springer, pp. 1171–1174, 2005.

[7] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, “Dimensionality reduction for fast similarity search in large time series databases,” Knowledge and information Systems, vol. 3, no. 3, pp. 263-286, 2001.

[8] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann and E. Keogh, “Experimental comparison of representation methods and distance measures for time series data,” Data Mining and Knowledge Discovery, vol. 26, no. 2, pp. 275-309, 2013.

[9] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu, “Mining frequent patterns in data streams at multiple time granularities,” Next generation data mining, vol. 212, pp. 191-212, 2003.

[10] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei and C. A. Ratanamahatana, “The UCR Time Series Classification/Clustering Homepage”, www.cs.ucr.edu/~eamonn/time_series_data/, 2011.