Research Journal of Applied Sciences, Engineering and Technology 5(5): 1726-1730, 2013
DOI:10.19026/rjaset.5.4928
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 27, 2012 Accepted: September 03, 2012 Published: February 11, 2013

Research Article

A Similarity Measure Method for Symbolization Time Series

Qiang Niu and Zhigang Li
Department of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

Abstract: Similarity measure is the base task of time series data mining tasks. LCSS measure method has obvious limitations in the two different length time series selection of a linear function. The ELCS measure method is proposed to normalize the sequence, which introducing the scale factor to limit the search path of the similarity matrix. Experiment in hierarchical clustering algorithm shows that the improved measure makes up for the shortcomings of LCSS, improves the efficiency and accuracy of clustering and improves time complexity.

Keywords: Hierarchical cluster, LCSS, similarity measure, time series

INTRODUCTION

Time series has always been an important and interesting research field due to its frequent appearance in different applications. Time series similarity measure that proposed by Agrawal *et al*. (1993) has become a hot research topic due to its wide application usages such as time series classification, clustering, abnormal findings on the basis of data mining, Many methods have been developed for searching time series measure method in large data sets and especially similarity measure of time series is a very important task in the process of data mining.

There is similarity measure methods of time series, such as Faloutsos *et al*. (1994) proposed a fast subsequence matching method based on the Euclidean distance metric, in which the similarity measure of the two time series is calculated as two points of the same dimension and it sets a threshold to judge whether the result is similar. Euclidean distance requires two sequences of equal length and ignored the temporal characteristics of time series, thus limiting its application in time series similarity measure. Chung *et al*. (2004) uses the weight method in the Euclidean distance method and eliminates transform offset, but there are parameters set by manual intervention.

Berndt and Cliford, (1994) introduce Dynamic Time Warping distance (DTW) to the time series similarity measure which performed well in the local characteristics comparation of the two unequal length sequences, but the time consumption of the algorithm is too expensive. In addition, DTW algorithm can't found two time series peaks between low point and inflection point, such as the corresponding relations between the feature points and the accuracy of the algorithm is low. Some researchers (Yi *et al*., 1998; Kim *et al*., 2001) improved DTW by introducing the index technology, making its time complexity reduced. An index-based approach for similarity search supporting time warping in large sequence databases (Kim *et al*., 2001) proposed the Segment-wise the Time Warping distance (STW), making the DTW time complexity decreased greatly, but making the similarity measure accuracy reduced too. Latecki *et al*. (2005) put forward a kind of minimum variance matching method to obtain the flexible similarity matching.

In 1994, the Longest Common Subsequence (LCS) measures (Paterson and Dancik, 1994) to the time series similarity measure. Bollobas *et al*. (1997) put forward LCSS on the basis of LCS, making a better similarity measure of time series which have amplitude translation, timeline stretching and bending deformation.

Some other researchers have proposed the slope-based, the model-based and the event-based similarity measure.

This research studies the similarity measure problem of symbolic time series. Firstly, this study introduces the definition and the classical similarity measure. Then, we propose a new similarity measure algorithm based on the LCSS algorithm: different to the LCSS algorithm, the new algorithm avoids the selection of a linear function effectively, improves the accuracy of measurement and improves time efficiency greatly compared to the DTW measure. Finally, experiments to verify the proposed algorithm.

LCS AND LCSS SIMILARITY MEASURE

LCS measure: There are time series samples $X, Y \in A$, their vector form is: $X=\{x_1, x_2, ..., x_n\}$, $Y=\{y_1, y_2, ..., y_n\}$, they

Corresponding Author: Qiang Niu, Department of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

satisfy the longest common subsequence of the following conditions were $X' = \{x_{i_1}, x_{i_2}, ..., x_{i_l}\}$ and $Y' = \{y_{j_1}, y_{j_2}, ..., y_{j_l}\}$, where l is the length of the Common subsequence, Similarity between time series X and Y is defined as $Sim(X,Y) = 1/n$.

If $1 \le k \le l$ for each k and if $i_k < i_{k+1}$ and $j_k < j_{k+1}$

If $1 \le k \le l$ for each k and $x_{i_k} = x_{j_k}$

LCSS measure: LCS measure can avoid the similar issues which brought by the time series of short-term mutation or intermittent. However, the time series of amplitude translation, timeline stretching and bending deformation can't get a good similarity measure results. LCSS measure is designed for the improvement of the above problems.

Let $\delta > 0$ be an integer constant, $0 < \varepsilon < 1$ a real constant. And $f \in L$, L a linear function set. Given two sequences $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, let $X' = \{x_{i_1}, x_{i_2}, ..., x_{i_l}\}$ and $Y' = \{y_{j_1}, y_{j_2}, ..., y_{j_l}\}$ be the longest subsequences in X and Y respectively such that:

- For $1 \le k \le l$, $i_k < i_{k+1}$ and $j_k < j_{k+1}$
- For $1 \le k \le l$, $|i_k - j_k| \le \delta$
- $1 \le k \le l$, $y_{i_k} / (1+\varepsilon) \le f(x_{i_k}) \le y_{j_k} (1+\varepsilon)$

Let $S_{f,\varepsilon,\delta}(X,Y) = l/n$. Then similarity between the time series is defined as formula (1):

$$Sim(X,Y) = \max_{f \in L} \{S_{f,\varepsilon,\delta}(X,Y)\} \quad (1)$$

EXTENDED LONGEST COMMON SUBSEQUENCE MEASURE (ELCS)

Although the LCSS measure has some advantages, there are still the following issues:

- LCSS measure derived from a solution set, for different time series data set, the selection of linear function f will different. In other words, only through the training data set for the corresponding linear function in advance, to further more accurate measure of the similarity of the sequence. Training and test set is always different, so the result is less than ideal.
- The LCSS can be applied with two different length sequence comparison, but because of $|i_k - j_k| \le \delta$, length difference of time series $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, that is $|m-n| \le \delta$. Otherwise, the sequence will undetect the candidate series (Keogh and Pazzani, 2001). Thus, the LCSS algorithm timeline stretching support is very limited.

For the existence problem of LCSS measure, this study presents an Extended Longest Common Subsequence (ELCS) measure:

Let $\mu > 1$ and $\theta > 0$ be a real constant, Given two sequences $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, The normalization that all sequence is located in between value [0,1], Get $X_{normal} = \{x'_1, x'_2, ..., x'_m\}$ and $Y_{normal} = \{y'_1, y'_2, ..., y'_m\}$, let $X'_{normal} = \{x'_{i_1}, x'_{i_2}, ..., x'_{i_l}\}$ and $Y'_{normal} = \{y'_{i_1}, y'_{i_2}, ..., y'_{i_l}\}$ be the longest subsequences in X and Y respectively such that:

- For $1 \le k \le l$, $i_k < i_{k+1}$ and $j_k < j_{k+1}$
- For $1 \le k \le l$, $\frac{1}{\mu} < \frac{i_k}{j_k} < \mu$ and $\frac{1}{\mu} < \frac{i_k - m}{j_k - n} < \mu$
- For $1 \le k \le l$, $\left| x'_{i_l} - y'_{i_l} \right| < \theta$

Let

$$S_{\theta,\mu}(X_{normal}, Y_{normal}) = \frac{2l}{m+n}$$

Then the similarity between the time series defined as the formula (2):

$$Sim(X,Y) = \max\{S_{\theta,\mu}(X_{norm}, Y_{norm})\} \quad (2)$$

Defined above, parameter μ makes the search path of the similarity measure matrix concentrated in a diamond area, not only to prevent the sequence of over match, while reducing the time complexity. And the selection of the search path area is related to each sequence length closely, not only appear undetected sequence, but also well adapted timeline stretching and deformation of the sequence match.

Parameter θ in the definition makes the similarity measurement algorithm, after normalization, get further flexibility to match the space.

Sequence normalized processing as the formula (3):

$$X_{norm} = \frac{X - \min_{i \in [1,m]} \{x_i\}}{\max_{i \in [1,m]} \{x_i\} - \min_{i \in [1,m]} \{x_i\}} \quad (3)$$

Which $x_i \in X$ avoid the linear function f selection of difficulties, at the same time retained the sequence of numerical trend information.

EXPERIMENT

Similarity measure is other data mining process foundation, the measure veracity directly affect other process treatment results. Instead, we can use the clustering results to estimate the accuracy of the different similarity measure.

Experimental environment and the data: The experimental environment is 2.20 GHz E4500CPU, memory for the 1024M and Window XP Professional system.

The experimental data sets use Synthetic Control Chart Time Series (SCC) in the UCI of KDD Archive and CBF dataset. The number of experimental data in the SCC is 600, every time the sequence's length is 60, divided into six categories. The CBF dataset contains Cylinder (C), Bell (B), Funnel (F), it is typical of synthetic data sets.

Experiment process: In cluster analysis, time series of the same group resemble each other, different sets of time series are not similar. This study uses the bottom-up hierarchical clustering. Set the initial data for the $C_1, C_2, ..., C_n$, the algorithm steps are:

Step 1: Each time series as a class C_i

Step 2: Calculate the similarity between any two categories, get a similarity matrix

Step 3: Merge the two categories which are similar, then go to Step 2 loop, until the class number is equal to the predetermined number of clusters

The distance between the clusters uses ELCS similarity measure computation.

The results of the clustering are standard $C = C_1, C_2, ..., C_k$ and the clustering results of each measure are $C' = C'_1, C'_2, ..., C'_k$, the clustering accuracy is computed by the following formula (4) and (5):

$$Sim\left(C_i, C'_j\right) = 2\frac{|C_i \cap C'_j|}{|C_i| + |C'_j|} \tag{4}$$

$$Sim\left(C, C'\right) = \frac{\sum_i \max_j Sim(C_i, C'_j)}{k} \tag{5}$$

The calculation of sim (C', C) and sim (C', C) is same. Because $Sim(C', C)$ and $Sim(C, C')$ is asymmetry, so $\frac{Sim(C', C) + Sim(C, C')}{2}$ is used as a final evaluation criteria.

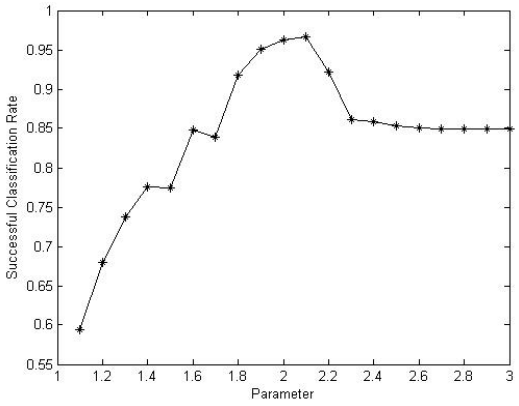


Fig. 1: Successful classification rate

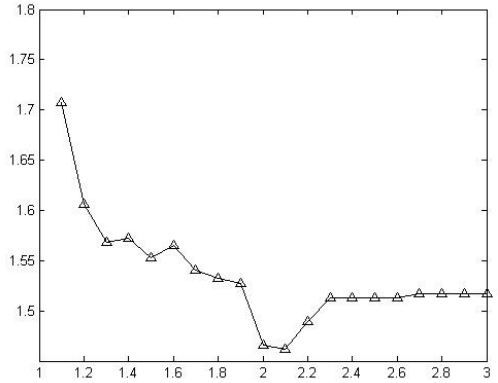


Fig. 2: Average internal class distance

EXPERIMENTAL RESULTS AND ANALYSIS

Parameter determination: The experiment using the SCC dataset is to analyses the influence of the algorithm. The ELCS measure contains the parameters μ and θ, the μ in the performance of the algorithm is very significant. With the changes of the parameter μ, the clustering accuracy rate is showed in Fig. 1, the clustering average internal class distance and average among class distance are shown in Fig. 2 and 3.

With the μ increases, the clustering accuracy rate is changed from low to high. When $\mu = 2.2$, clustering accuracy rate is the highest, the average internal class distance is the smallest; the average among class distance is largest. This result means each one of ELCS measure in the sequence satisfies the length μ. While $m = n$ is too large, not well qualified the position of the test sequence corresponds to the information, get meaningless similar sequence segments; While μ is too small, the search range of the similarity matrix is

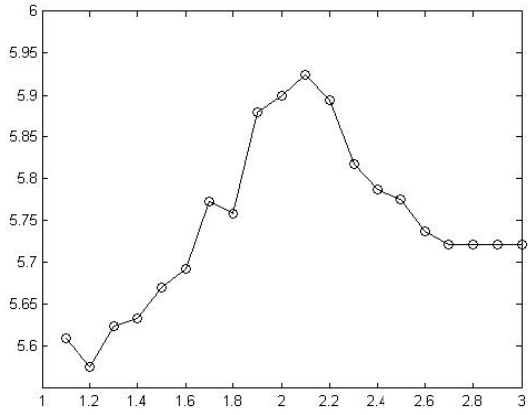
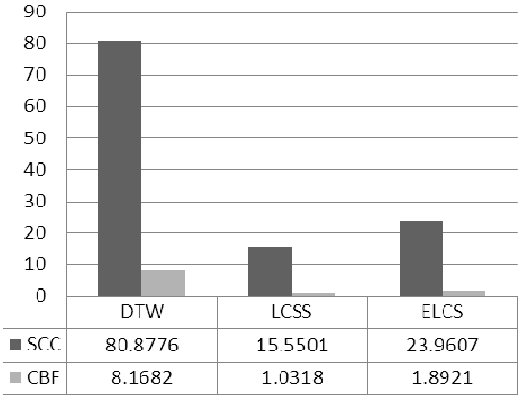
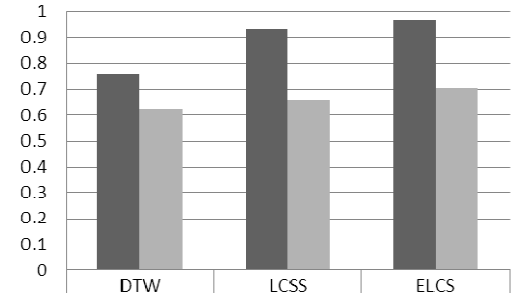


Fig. 3: Average among class distance



	DTW	LCSS	ELCS
SCC	80.8776	15.5501	23.9607
CBF	8.1682	1.0318	1.8921

Fig. 4: Time-consuming comparisons of three distance measures

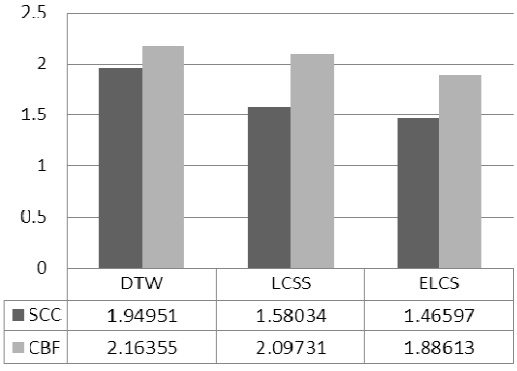


	DTW	LCSS	ELCS
SCC	0.76194	0.93042	0.96667
CBF	0.62514	0.66127	0.70371

Fig. 5: Clustering accuracy rate for the three distance measures

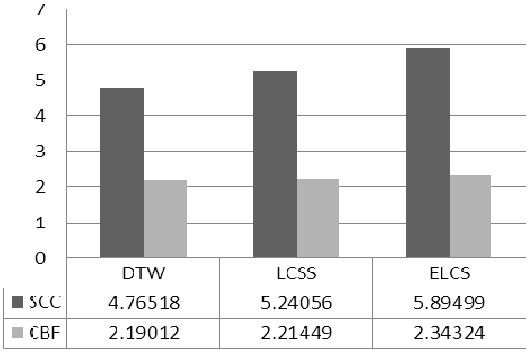
very limited, a lot of data is discarded to be missed. With the decrease of μ, the classification accuracy dropped sharply.

Three kinds of measure-based clustering comparison: To comparison of DTW, LCSS and the ELCS three kinds of distance measures' time consuming, set them as the similarity metric of hierarchical clustering. LCSS and the ELCS algorithm is selected the appropriate parameters, making the final classification accuracy is their best. The results are shown in Fig. 4. DTW algorithm consuming significantly higher than other measures, as ELCS measure's condition $\frac{1}{\mu} < \frac{i_k}{j_k} < \mu$ and $\frac{1}{\mu} < \frac{i_k - m}{j_k - n} < \mu$ is complex than LCSS measure's $|i_k - j_k| \le \delta$, so spend more time.



	DTW	LCSS	ELCS
SCC	1.94951	1.58034	1.46597
CBF	2.16355	2.09731	1.88613

Fig. 6: Average internal class distance for the three distance measures



	DTW	LCSS	ELCS
SCC	4.76518	5.24056	5.89499
CBF	2.19012	2.21449	2.34324

Fig. 7: Average among class distance for the three distance measures

Figure 5 is a comparison of DTW, LCSS and the ELCS measure of clustering accuracy rate. Each measure for the SCC data set has good results, because of the obvious characteristics of SCC dataset of data and the data has a little noise. CBF dataset is a randomly generated dataset; each time series has a lot of glitches that increase the difficulty of the clustering. But no matter to which dataset, ELCS have shown good results, that is the correct rate of clustering is the highest.

The dataset differences above-mentioned, can be seen in Fig. 6 and 7 easily. Clustering results of the CBF dataset average distance internal class is greater than the

SCC dataset, while the average among class distance is smaller. Due to LCSS and ELCS are based on LCS algorithm, do not exist DTW algorithm point corresponds to a multi-point problems, local noise can be ignored.

CONCLUSION

Based on the LCS measure, by introducing parameters which standardizes similarity matrix search path, this study improves the accuracy of the similarity measure and overcomes the traditional similarity measure based on Euclidean distance which lack of dealing with noise interference. By the experiment on two different types of data sets, ELCS measure gets higher clustering correctness than the existing similarity measures, but the time expense is higher. In short, the measure can be applied effectively to a variety of time series similarity measure.

ACKNOWLEDGMENT

This study was supported by Doctoral Program Foundation of Ministry of Education of China (20100095110003) and Fundamental Research Funds for the Central Universities (2011QNB23).

REFERENCES

Agrawal, R., C. Faloustos and A. Swami, 1993. Efficient similarity search in sequence database [c]. Proceedings of 4th International Conference on Foundations of Data Organization and Algorithms. Springer, Berlin, pp: 69-84.

Berndt, D. and J. Clifford, 1994. Using Dynamic Time Warping to Find Patterns in Time Series. AAAI-94 Workshop on Knowledge Discovery in Databases, AAAI Press, Seattle, Washington.

Bollobas, B., G. Das, D. Gunopulos and H. Mannila, 1997. Time-series similarity problems and well-separated geometric sets [A]. Proceedings of the 13th Annual Symposium on Computational Geometry [C]. ACM Press, New York, pp: 454-456.

Chung, L., T.C. Fu and R. Luk, 2004. An evolutionary approach to pattern-based time series segmentation. IEEE T. Evolut. Comput., 8(5): 471-489.

Faloutsos, C., M. Ranganathan and Y. Manolopoulos, 1994. Fast subsequence matching in time-series databases [J]. SIGMOD Rec., 23(2): 417-429.

Keogh, E.J. and M.J. Pazzani, 2001. Derivative Dynamic Time Warping [DB/OL]. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.3383&rep=rep1&type=pdf.

Kim, S.W., S. Park and W. Chu, 2001. An index-based approach for similarity search supporting time warping in large sequence databases [A]. Proceedings of the International Conference on Data Engineering [C]. IEEE Computer Society, Heidelberg, pp: 207-614.

Latecki L.J., V. Megalooikonomou, Q. Wang, R. Lakaemper, C.A. Ratanamahatana, *et al*., 2005. Partial elastic matching of time series [A]. 5th IEEE International Conference on Data Mining [C]. Nov. 27-30, Philadelphia.

Paterson, M. and V. Dancik, 1994. Longest common subsequences [J]. Lect. Notes Compu. Sc., 841: 127-142.

Yi, B.K., H.V. Jagadish and C. Faloutsos, 1998. Efficient retrieval of similar time sequences under time warping [A]. Proceedings of the International Conference on Data Engineering [C], IEEE Computer Society, Orlando, pp: 201-208.