Technology and Health Care 23 (2015) S501–S510
DOI 10.3233/THC-151002
IOS Press

Symbolic representation based on trend features for biomedical data classification

Hong Yin[a,c,*], Shuqiang Yang[a], Xiaoqian Zhu[a], Shaodong Ma[b] and Liqian Chen[a]
[a]*College of Computer, National University of Defense Technology, Changsha, Hunan, China*
[b]*School of Engineering, The University of Hull, Hull, UK*
[c]*Xiangyang School for NCOs, Xiangyang, Hunan, China*

Abstract.
BACKGROUND: The widespread access to portable medical devices or new personal devices is boosting the amount of biomedical data. These devices provide a growing massive data that far exceeds the analytical ability of a professional doctor. The computer-assisted analysis of biomedical data has become an essential tool in medicine diagnosis.
OBJECTIVE: Due to the advantages of discrete, noise elimination and dimensionality reduction, symbolic representation of biomedical data has attracted great interest. The symbolization results provide efficiently performing at data mining, such as pattern discovery, anomaly detection and association rules mining, so we want to use the method to improving the biomedical data classification.
METHODS: In this paper, we introduce a novel symbolic representation method, called Trend Feature Symbolic Approximation (TFSA).
RESULTS: The TFSA focuses on retaining most of the original series' trend features, and it also very suitable for subsequent mining work, such as association rules mining.
CONCLUSION: The TFSA provides the lower bounding guarantee and the experimental results show that comparing with some existing methods, its classification accuracy is improved.

Keywords: Symbolic, trend features, ECG series, classification, segments

1. Introduction

The increasing prevalence of long-term monitoring in remote medical, such as electrocardiograph (ECG), is boosting the amount of biomedical data. To healthcare professional, analysis of long-term recordings of the heart activity, is difficult to diagnose and can be highly error prone. Medical errors derived from "information overload" are not surprising so that medical relevant events are often missed [1]. Using the computer-assisted technology to analyze biomedical data has become an essential tool in medical diagnosis. For example, due to the useful information about heart rhythm, ECG signals can be used to research heart arrhythmias [4]. Extracting meaningful features to represent individual ECG series is also a research hotspot. Zadeh et al. [7] extracted morphological and timing interval features from ECG segments to classify heartbeats. Sepideh Babaei et al. [8] propose a identify ECG by BP neural, and five kinds of typical ECG feature extracted by Daub wavelet.

*Corresponding author: Hong Yin, College of Computer, National University of Defense Technology, Changsha, Hunan, China. Tel.: +86 18670361345; E-mail: yinhonggfkd@aliyun.com.

By extracting local temporal or frequency information to classify ECG series, most of the previous methods are very effective for short ECG series. However, these methods may not good at capturing the similarity of long ECG series for the reason of compute complexity. In order to capture the high-level information of ECG series, a bag-of-patterns (BoP) representation by converting an ECG series to a words string using the Symbolic Aggregate approximation (SAX) [3] is proposed [6]. The symbolic Representation of ECG series can efficiently facilitate classification, pattern discovery and anomaly detection, by converting the numerical form of ECG series into a series of discrete symbols. This discrete representation also reduces the dimension of the series, which makes the search space for classification of mean value long series more manageable. In this paper, a symbolic representation method for long-term ECG series – Trend Feature Symbolic Approximation (TFSA) is proposed, which focuses on preserving the trend features of original ECG series. These trend features can be further used for diagnostic knowledge discovery, such as discovering rhythms, transient patterns, abnormal changes in ECG, and clinically significant relationships among multiple streams of biomedical data.

2. Related work

The basic concept of ECG series symbolic representation is to convert the numerical form of ECG series into a series of discrete symbols according to designated mapping rules. With the characteristics of discrete, noise elimination and dimensionality reduction after symbolization, a reasonable symbolic method can improve the efficiency of ECG data mining. Based on these advantages, many researchers have proposed high-level representations of ECG series, including Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Piecewise Aggregate Approximation (PAA), and Singular Value Decomposition. As an effective analytic tool for non-linearity analysis, symbolic dynamics also has been applied to cardiology recently, especially in Premature Ventricular Contraction (PVC) analysis [2]. X. Zhang, et al. proposed a ventricular tachycardia (VT)/ventricular fibrillation (VF) detection algorithm based on symbolic dynamics, where they transformed the ECG series into a 2-symbolstringsequence and computed the complexity measure of the string sequence [5]. The Symbolic Aggregate Approximation (SAX) [3] becomes increasingly popular as for its simplicity and high computational efficiency in symbolic representation of time series. SAX is also reputed in presenting a reliable performance on the data mining tasks, such as clustering, classification, and anomaly detection. It is, nonetheless, not competent in discovering significant association rules among multiple streams of biomedical data. That is, SAX uses the mean value of the subsequence to represent original time series, so it ignores most of the trend features (such as uptrend, downtrend and the degree of the trend change). These trend features are the important information for acquiring knowledge from time series.

The contribution of this paper is to propose a novel symbolic representation method, and make series after symbolization can approximate the original series more closely. So that it can improve the classification accuracy of long biomedical series. The process of the symbolic representation consists of two steps. In the first step, the biomedical signal is segmented into subsequences based on trend features. The second step involves representing these subsequences using trend symbol.

3. Classification

3.1. The ECG series classification

As in many other applications, huge amounts of unlabeled data are often available. For example, the PhysioBank archive contains more than 40 gigabytes of freely available medical data, including ECG,

Table 1
Adaptive segment algorithm

Adaptive_Segment ()
Input: subsequence $Q_i = q_1, q_2, \ldots, q_j$
Output: change-points u_i
1: BEGIN
2: set the size of sliding window *fit* $= l$;
3: compute the slope of the first *fit* points of the series k_1;
4: slide over one point and compute the new slope k_2;
5: if $\lvert k_1 - k_2 \rvert >$ *tol*, a change point $u_i = q_{l+1}$ and k_1 is recorded, the window slide to the point q_{l+1}, the routine go to step 3;
7: if $\lvert k_1 - k_2 \rvert <$ *tol*, $k_1 = (k_1 + k_2)/2$, the routine go to step 4;
8: algorithm continues until the sliding window reaches the edge of the series.
9: END

EEG and gait data. Such large datasets are potential goldmines for building classifiers to mining these useful data. For concreteness, some definitions of biomedical data are given.

Definition 1. *ECG Series*: An biomedical series $T = t_1, t_2 \ldots t_m$ is a sequence of m samples collected at regular intervals over a period.

Definition 2. *ECG Series Classification*: Given a set of unlabeled ECG series, the task of ECG series classification is to map each ECG series to one of the predefined classes based on the distance measure.

The most common distance measure for time series is the Euclidean distance.

Definition 3. *Euclidean Distance*: Given two series Q and C with the length of n, the Euclidean distance between them is calculated by the following formula:

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

3.2. The segmentation of ECG series

As a preliminary step, segmenting ECG signals into basic quasi-periodic units is considered. The purpose of ECG series segmentation is to divide the original series into a set of independent subsequences. In order to preserve most of the trend features or patterns, the searching key point of ECG series is of great significant. The important key points include extreme points, local extreme points, important points and turning points, etc. This paper proposed a method to find key points, called adaptive segment algorithm based on trend features. This approach fully considers transformation information of the series, so it is useful to identify the turning points from one pattern to the next changed pattern.

The process of the adaptive segment algorithm is as follows: Given a sliding window length *fit* and an angle tolerance parameter *tol*, the slope of the first *fit* points of the series is computed using Least Squares Regression (LSR). Then, the window slides over one point, and a new series is obtained with the length of *fit* + 1. The slope of the new series is computed by LSR. Comparing the new slope with the older one, if the absolute difference between them exceeds the *tol*, the rightmost point of the previous window is recorded as a change point. Starting with the next point of the change point, the process of slope computing is repeated again. If the absolute difference of slopes does not exceed the *tol*, then the older slope will be updated by the new one. The segment algorithm continues execution until the sliding window reaches the end point of the series. The algorithm of adaptive segment is also described in the Table 1.

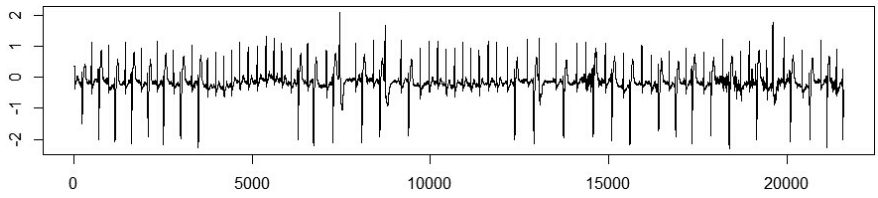


Fig. 1. A record of ECG signal from MIT-BIH database.

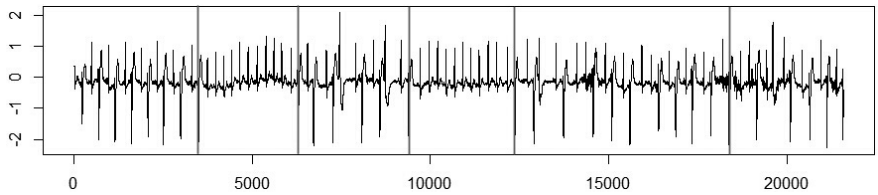


Fig. 2. The periodic feature obtained by adaptive segment algorithm. (The solid line indicates the segmentation points).

The adaptive segment algorithm not only can be used to find the key points of time series, but also is very suitable for recognizing the periodic feature. It is important for time series with the period, such as long-term ECG series. Taking the time series in Fig. 1 for example, the data is a record of ECG from MIT-BIH Arrhythmia Database.

Using the algorithm mentioned above to segment the series, the results are shown in Fig. 2. From the results, the key points to recognize the periodic feature are obtained, and it illustrated the series is composed of six series with different patterns.

4. The symbolic representation method TFSA

4.1. The symbolization of subsequence

After the segmentation for ECG series, the subsequences obtained become imbalanced in length but the features are retained as much as they can be. The next procedure will map the subsequences with a set of symbols. Some methods for time series symbolization neglect the trend feature of series, such as upward, downward, and horizontal. Many ECG series as they occur in practice are not stationary. Therefore, most of ECG series will show certain 'trend' as time progresses and these trends form important features of an ECG series. A trend should have a direction, that is, it has the higher or lower value at the end of the series, so that it will seem generally to increase or decrease throughout. Although recent researches based on SAX are supplemented with the local extreme points to express the trend information, this approach still needs to be improved for applications demanding better accuracy. In this paper, three kinds of trend are considered, which are *upward*, *downward* and *flat*.

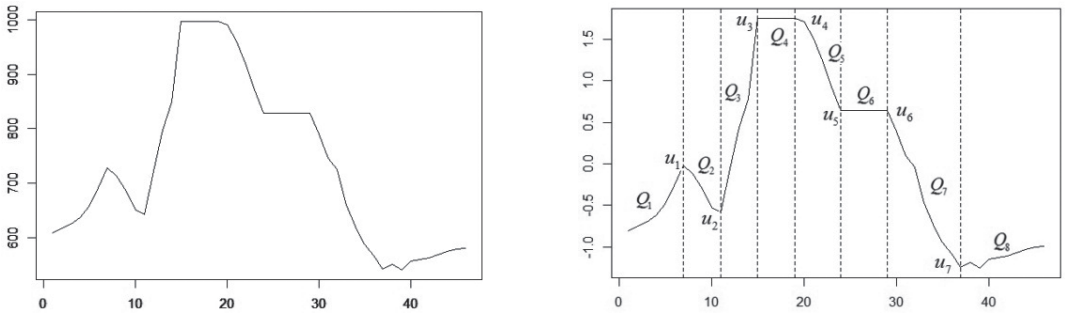
The Trend Feature Symbolic Approximation (TFSA) uses the trend symbols to represent the subsequences after segmentation, which also allows the subsequences can visually display these trend features. The TFSA provides specific symbolic representations of trends listed in Table 2.

The following example illustrates the TFSA method, and the original time series used for the experiment is shown in Fig. 3(a). To illustrating more clearly, the ECG series is not used in the example.

Numerous studies have realized the importance of standardizing the time series before clustering, classification, and comparison of the similarity, etc. Accordingly, our method will normalize the series

Table 2
The trend symbols of TFSA

The symbol	The meaning of the symbol
01_b^a	The symbol indicates a subsequence with upward trend. The superscript a is the slope of subsequence, which is used to describe the degree of change. The subscript b is the last point value of the subsequence after standardizing.
10_b^a	The symbol indicates a subsequence with downward trend. The superscript a is the slope of subsequence, which is used to describe the degree of change. The subscript b is the last point value of the subsequence after standardizing.
$00_b^0, 11_b^0$	The two symbols indicate the subsequence with a flat trend. The symbol 11_b^0 is used to represent flat trend after 01_b^a model, and the symbol 00_b^0 to represent flat trend after 10_b^a model. The slope is zero, b is defined as above.



(a) (b)

Fig. 3. (a) The original time series; (b) The segmentation of the normalization series.

before symbolization into a standard sequence. Using the segmentation algorithm introduced in Section 3 to split the series, the results are shown in Fig. 3(b). In the Fig. 3(b), the u_i represents the change points, and the Q_i represents the subsequences after segmentation.

In the segmentation of series, the slopes of subsequences a and the positions of split points b can be easily retained, so according to the symbols defined in Table 2, the symbolic results of subsequences are:

$$Q_1 = 01_{-0.02}^{1.28}, \;\; Q_2 = 10_{-0.59}^{-1.29}, \;\; Q_3 = 10_{1.75}^{5.32} \ldots.$$

Hence the symbolic representation of the series is:

$$Q_1Q_2Q_3Q_4Q_5Q_6Q_7Q_8 = 01_{-0.02}^{1.28}10_{-0.59}^{-1.29}01_{1.75}^{5.32}11_{1.75}^{0}10_{0.64}^{-2.02}00_{0.64}^{0}10_{-1.25}^{2.15}01_{-0.99}^{0.26}.$$

From the results of symbolization, the trend of the time series initially tends to move upwards, and then drops downwards to -0.59, after that there is a drastic increase that ascends to 1.75 in a short time. After a flat section, the time series declines until another flat region. The descent of the time series continues before reaching a turning point at -0.99, and is followed by a slow rise to finish. The trend is characterized and symbolized using the TFSA and is shown in Fig. 4(a).

Note that the number of segments by SAX is the same as the number obtained using TFSA, and the basis of character set is 4 (The number 4 means that using four characters to represent time series, and here is a, b, c, d). So the time series is mapped into the string {bbddcbaa} using SAX. It is noted that the results obtained using the SAX has poor performance than the results obtained by TFSA, and many key points are discarded. From the symbolic results of SAX, only a fewer trends or patterns of the original series are retained.

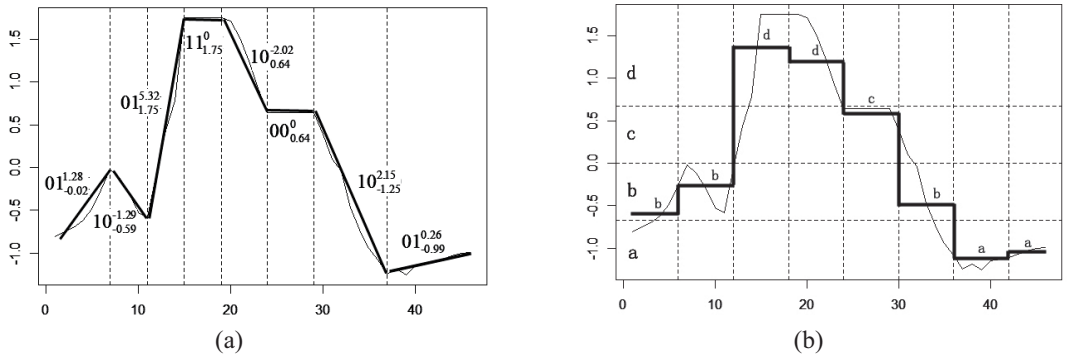


Fig. 4. (a) The symbolic representation of time series using TFSA; (b) The symbolic representation by SAX.

4.2. The distance measure for TFSA

All symbolization methods need to be evaluated to how close the approximated symbol can represent the features of the original series. Faloutsos [9] drew an important conclusion that to ensure no false dismissals in the distance measure between the symbolic string and the true distance, and the following condition must satisfy:

$$D_{symbolic}(Q, C) \leqslant D_{true}(A, B) \tag{1}$$

Where in In Eq. (1), A and B are the original time series which is measured by the true distance D_{true}, and Q, C are symbolic sequences of A, B with reduced dimensions measured by $D_{symbolic}$. This theory is also known as the Lower Bounding. The distance after symbolization should not exceed the true distance. In this paper, the Euclidean distance is used to measure the true distance.

$$D_{ture}(A, B) = D_{Euclidean}(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{2}$$

According to the symbolic representation method of TFSA, the $D_{symbolic}(Q, C)$ is defined in Eq. (3):

$$D_{symbolic}(Q, C) = D_{TFSA}(Q, C) = \sqrt{\frac{n}{w} \sum_{i=1}^{w} T_i \times (qb_i - cb_i)^2 \times \frac{||qa_i| - |ca_i||}{\max(|qa_i|, |ca_i|)}} \tag{3}$$

Where, n is the length of original series; w is the number of subsequences after segment; T_i is the distance coefficient between different trends which is a penalty coefficient, its value is shown in Eq. (4); qb_i, cb_i are the last point of the i^{th} subsequence from series Q and C; qa_i, ca_i are the slopes of the i^{th} subsequence from Q and C; $\overline{Q}$ and $\overline{C}$ are the mean values of time series Q and C.

$$T_i = \left[\frac{\overline{Q} - \overline{C}}{qb_i + f \times cb_i} \right]^2, \quad f = \begin{cases} 0, & cb_i \geqslant 0 \\ -1, & cb_i < 0 \end{cases} \tag{4}$$

4.3. Association rules mining

Due to the trend features, one of the TFSA's differences from the previous symbolization methods is that it facilitates the subsequent data mining research, such as frequent item sets mining, association

Table 3
A comparison of different methods in efficiency

Degree of trends	Angle range	Angel range
1 level	$(0°, 10°)$	$(-10°, 0°)$
2 level	$(10°, 20°)$	$(-20°, -10°)$
3 level	$(20°, 30°)$	$(-30°, -20°)$
4 level	$(30°, 40°)$	$(-40°, -30°)$
5 level	$(40°, 50°)$	$(-50°, -40°)$
6 level	$(50°, 60°)$	$(-60°, -50°)$
7 level	$(60°, 70°)$	$(-70°, -60°)$
8 level	$(70°, 80°)$	$(-80°, -70°)$
9 level	$(80°, 90°)$	$(-90°, -80°)$

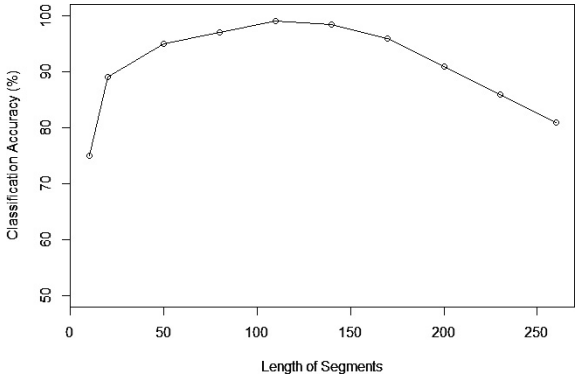


Fig. 5. The classification accuracy with different length of segments.

rules mining, etc. If the angle space containing the slopes of subsequences can be mapped into the set of $(-90°, 0°)$, the space can be divided into a series of non-overlapping intervals. Each interval corresponds to a number (from 1 to 9) that indicates the change in steepness or degree of trends. The angle space and the corresponding numbers are defined in Table 3. According to Table 3, the slopes of the subsequences can be transformed from infinity into a limited space. In other words, all the trend features can be represented by finite symbol.

By this method, time series can be described by trend feature item sets, and each of them has specific meaning that is not only a symbol. Take the series illustrated in Section 4.1 for example, the results of symbolization can be expressed as:

$$Q_1Q_2Q_3Q_4Q_5Q_6Q_7Q_8 = 01^6 10^6 01^8 11^0 10^7 00^0 10^7 01^2.$$

As a result, the frequent item sets and association rules mining algorithm can be applied in medical diagnosis with the results from TFSA. For example, gathering the symbolic representations of ECG and EEG series by TFSA, it is very likely to find a potential relationship between cardiovasculardiseaseand-cerebrovasculardisease using association rules mining.

5. Experimental results

5.1. Length of ECG segments

In Section 3, the length of segments (sliding window) in adaptive segment algorithm is considered. It is an important parameter *fit* which can determine the effects of segmentation. Too short length of segments will cause to end up with a string keeping too much of the original series and will not simplify the series, however, too long segments will cause considerable amount of information loss. On the other hand, considering the characteristic of ECG, the initialization of sliding window length is based on the average distance between two peak points or two 0-near points. In order to select the most appropriate length, the experiment is tested by varying the length of ECG segments between 10 and 260. The data used to test the parameter *fit* is the ECG-40 dataset from the Fantasia ECG database, which consists of 40 classes. The ECG series are collected from 40 healthy persons monitored for about 2 hours with a rate of 250 Hz. From each of ECG with more than 100,000 data points, 50 shorter series are extracted

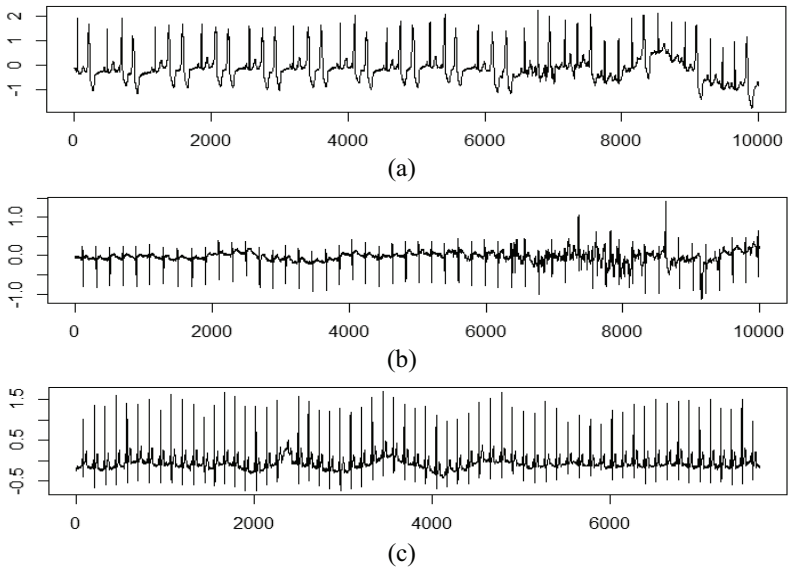


Fig. 6. (a) Arrhythmia ECG; (b) Atrial Fibrillation ECG; (c) Normal Sinus Rhythm ECG.

with the length of 2048. So the ECG-40 dataset contains 2000 series of length 2048, distributed in the 40 classes.

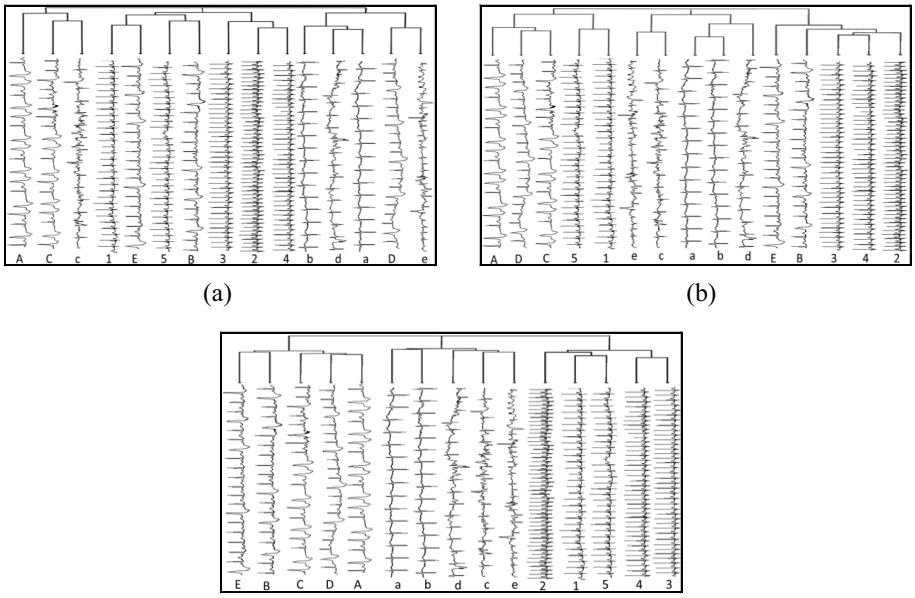
The length of segments from 10 to 260 depends on the fact that biomedical signals are relatively flat. The result of classification accuracy on the ECG-40 dataset is illustrated in Fig. 5. It can be seen from the result that the accuracy of classification is relatively stable when the length of segments is from50 to 180. The accuracy will fall abruptly when the length below 20, and it also occurs with the length longer than 180. It is also illustrated that the segments with too short or too long length cannot capture the trend features of the series. In the following experiments, the length of segments is set as 110 empirically.

5.2. The accuracy of TFSA compared with other methods

The accuracy of the TFSA is examined by the classified results of ECG data. Three classes of data from the Physionet MIT-BIH database are used as the test data, which are Arrhythmia ECG, Atrial Fibrillation ECG and Normal Sinus Rhythm ECG. Some examples for the three ECG series are shown in Fig. 6.

The TFSA is compared with Euclidean distance and SAX methods. Figure 7 shows some example of the Arrhythmia, Atrial Fibrillation and Normal Sinus Rhythm classes which are classified by the three methods. The ECG series “A” to “E” are members of the Arrhythmia, and the ECG series “a” to “e” are members of the Atrial Fibrillation. The numbers 1 to 5 represent the members of the Normal Sinus Rhythm class. In Fig. 7, (a) is the classification using Euclidean distance, and (b) is the classification using SAX algorithm, so (c) is the classification of TFSA algorithm.

Despite the Euclidean distance can be calculated quickly, the Euclidean distance approach is incapable of fully distinguishing the three classes due to the different length. The SAX can convert the ECG series very easily, but it only uses the mean value of each subsequence to represent the original series. Lots of information about cardiopathy is ignored, that will reduce the classified accuracy of SAX, and it only classified the Atrial Fibrillation class correctly. Owing to consider more trend features, the classification results obtained by TFSA are more acceptable. TFSA preserves most of trend information, in this experiment, with the length of segments set as110, the classification results of TFSA are the best.



(a) (b)

(c)

Fig. 7. (a) Classification by Euclidean distance; (b) Classification by SAX with segments length as 110; (c) Classification by TFSA with segments length as 110.

6. Conclusions

Huge amounts of biomedical data continue to emerge, and how to discover useful medical diagnosis knowledge from the massive data poses a serious challenge to healthcare professionals. For biomedical series, this paper proposes a symbolic representation method, called Trend Feature Symbolic Approximation. The contribution of the TFSA is that it focuses on retaining most of the original series trend features and patterns, and it represents biomedical series using trend symbols, which can be used to improve the classification accuracy of symbolic methods. The experimental results also show that compared to other methods, with the same length of segments, the classification accuracy of TFSA is improved.

The motivation of this paper is to provide a new classification method of biomedical series, and in Section 4.3, we simply introduced how to use the TFSA for association rules mining, but not deep enough. The next step for the work is to study the association rules mining for the biomedical series after symbolization using TFSA.

Acknowledgement

This work was supported in part by National High-tech R&D Program of China under Grants No. 2012AA012600, 2012AA01A401, 2012AA01A402.

References

[1] Kopec D, Kabir M H, Reinharth D, et al. Human errors in medical practice: systematic classification and reduction with automated information systems[J]. Journal of Medical Systems, 2003, 27(4): 297-313.

[2] Zhao L, Wiggins M, Vachtsevanos G. Premature ventricular contraction beat detection based on symbolic dynamics analysis[C]. *Proc. IASTED*. 2003: 48-50.
[3] Lin J, Keogh E, Lonardi S, et al. A symbolic representation of time series, with implications for streaming algorithms[C]. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, 2003: 2-11.
[4] Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of ECG signals[J]. Biomedical Engineering, IEEE Transactions on, 2009, 56(5): 1415-1426.
[5] Zhang X S, Zhu Y S, Thakor N V, et al. Detecting ventricular tachycardia and fibrillation by complexity measure[J]. Biomedical Engineering, IEEE Transactions on, 1999, 46(5): 548-555.
[6] Lin J, Khade R, Li Y. Rotation-invariant similarity in time series using bag-of-patterns representation[J]. Journal of Intelligent Information Systems, 2012, 39(2): 287-315.
[7] Zadeh A E, Khazaee A, Ranaee V. Classification of the electrocardiogram signals using supervised classifiers and efficient features[J]. Computer Methods and Programs in Biomedicine, 2010, 99(2): 179-194.
[8] Babaei S, Geranmayeh A. Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals[J]. Computers in Biology and Medicine, 2009, 39(1): 8-15.
[9] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases[M]. ACM, 1994.
[10] Sarkar S, Mukherjee K, Ray A. Symbolic dynamic analysis of transient time series for fault detection in gas turbine engines. *Journal of Dynamic Systems, Measurement, and Control*. 2013: 135, 014506-1.
[11] Wong D F, Chao L S, Zeng X D. A supportive attribute-assisted discretization model for medical classification[J]. Bio-Medical Materials and Engineering, 2014, 24(1): 289-295.
[12] Lkhagva B, Suzuki Y, Kawagoe K. New time series data representation ESAX for financial applications[C]. *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006: x115-x115.
[13] Pratt K B, Fink E. Search for patterns in compressed time series[J]. International Journal of Image and Graphics, 2002, 2(01): 89-106.
[14] Sun L, Xu J. Feature selection using mutual information based uncertainty measures for tumor classification[J]. Bio-medical Materials and Engineering, 2014, 24(1): 763-770.
[15] Vullings H, Verhaegen M H G, Verbruggen H B. ECG segmentation using time-warping[M]. *Advances in Intelligent Data Analysis Reasoning about Data*. Springer BerlinHeidelberg, 1997: 275-285.
[16] Yairi T, Kato Y, Hori K. Fault detection by mining association rules from house-keeping data[C]. *Proc. of International Symposium on Artificial Intelligence, Robotics and Automation in Space*. 2001: 3(9).
[17] Tanantong T, Nantajeewarawat E, Thiemjarus S. Toward continuous ambulatory monitoring using a wearable and wireless ECG-recording system: A study on the effects of signal quality on arrhythmia detection[J]. Bio-Medical Materials and Engineering, 2014, 24(1): 391-404.